Universidad de Granada

Departamento de Ciencias de la Computación

e Inteligencia Artificial

PREDICTING PROKARYOTIC AND EUKARYOTIC GENE NETWORKS BY FUSING DOMAIN KNOWLEDGE WITH CONCEPTUAL CLUSTERING ALGORITHMS

Tesis Doctoral
Oscar Marcos Harari

Granada, Diciembre de 2008

Editor: Editorial de la Universidad de Granada
Autor: Oscar Marcos Harari
D.L.: GR. 2855-2008
ISBN: 978-84-691-8342-7

Universidad de Granada

Predicting prokaryotic and eukaryotic gene networks by fusing domain knowledge with conceptual clustering algorithms

Memoria que presenta
Oscar Marcos Harari
Para optar por el grado de doctor en informática
Diciembre de 2008

Director
Igor Zwir

Departamento de Ciencias de la Computación e Inteligencia Artificial

La memoria titulada ``**Predicting prokaryotic and eukaryotic gene networks by fusing domain knowledge with conceptual clustering algorithms** ', que presenta Oscar Marcos Harari para optar al grado de doctor, ha sido realizada dentro del programa de doctorado ``Diseño, Análisis y Aplicaciones de Sistemas Inteligentes''del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección del doctor D. Igor Zwir

Granada, Diciembre de 2008

El Doctorando

El director

Fdo. Oscar Marcos Harari

Fdo: Igor Zwir

Acknowledgments

Gracias totales.

Tesis Doctoral parcialmente subvencionada por la Agencia Española de Cooperación Internacion, Ministerio de Asuntos Exteriores de España; el proyecto TIN2006-12879 (2006-2009) del Ministerio de Ciencia y Tecnología de España; y el proyecto TIC02788 de la Consejería de Innovacion, Investigación y Ciencia de la Junta de Andalucía.

Contents

Front Matter

Contents iv
List of Tables ix
List of Figures xi

Introducción 15
I. Problemática 15
II. Objetivos 17
III. Resumen 19

1 Fundamentos Biológicos y Bioinformática 21
1.1 Material Genético en la Célula 21
1.2 ADN y Evolución 28
1.3 Biología y Avances Tecnológicos 29
1.4 Microarrays de ADN 31
1.5 Biología computacional y Bioinformática 32
1.5.1 Objetivos de la Bioinformática 32
1.6 Introduction to Microarray Technology 34
1.6.1 Spotted Arrays 35
1.6.2 Oligonucleotide Arrays 36
1.6.3 Microarray Scope of Application 39

2 Enfoque Computacional 41
2.1 Modelado de sistemas 41
2.1.1 Lógica y conjuntos difusos 44
2.2 Clustering 47
2.2.1 Clustering Jerárquico 49
2.2.2 Subtractive Clustering 50
2.2.3 Clustering difuso *c*-medias (Fuzzy *c*-means) 51
2.2.4 Árboles de decisión 52
2.2.5 Clases latentes 54
2.2.6 Nonnegative matrix factorization 54
2.2.7 Clustering conceptual 56

2.2.8 Descubrimiento de subgrupos 57
2.3 Optimización multiobjetivo 59
2.4 Test de Coincidencia 63
2.5 Inferencia 64
2.5.1 Algoritmo de los *k*-vecinos más cercanos. 64
2.5.2 Sistemas Basados en Reglas Difusas 65
2.6 Algoritmos evolutivos 67
2.6.1 Algoritmos genéticos 68
2.6.1.1. Algoritmo genético básico 70
2.6.1.2. Algoritmos genéticos para funciones multimodales: *Nichos* 73
2.6.1.3. Elitismo 75
2.6.2 Problemas multiobjetivo y algoritmos genéticos multiobjetivo 75
2.7 Metodology 75

3 Delimiting plasticity of transcription factor binding sites by disassembling DNA consensos sequences 81

3.1 Results 83
3.1.1 A single motif encoding TFBS sequences does not describe the entire binding dataset 83
3.1.2 A Divide & Conquer (*D&C*) approach decomposes transcription factor binding sites into a family of motifs 84
3.1.3 A multi-classifier based on submotifs overcomes limitations of a single consensus 86
3.1.4 *Cis*-features employed for the "*divide*" phase overcomes limitations of a single consensus 87
3.1.5 *D&C* uncovers the PhoP regulon 89
3.1.6 Sensitivity analysis of parameters demonstrates the robustness of the *D&C* for the PhoP regulon 94
3.1.7 Submotifs distinguish functional PhoP binding sites in Genome-wide analysis 95
3.1.8 Submotifs reflect distinct DNA physical constraints of the protein-binding interactions 96
3.1.9 Submotifs provide a model of evolution of the PhoP regulon 98
3.2 Methods 102
3.2.1 Clustering TFBS sequences. 102
3.2.2 Voting multi-classifier 103
3.2.3 Fuzzy membership functions. 103
3.2.4 Fuzzy IF-THEN rules 104
$\mu_1(X)\, AND\, \mu_2(X) = \mu_1 \cap \mu_2 = PROD(\mu_1,\mu_2)$
$= \{PROD(d_{11},d_{21})/x_1,\ldots,PROD(d_{1n},d_{2n})/x_n\}$ 104
3.2.5 Performance measurements 104
3.2.6 Genetic algorithms 105
3.2.7 Rate of evolution for binding sites 106
3.2.8 Gamma enterobacteria othologs 106
3.2.9 Statistical significance 106
3.2.10 Microarray analysis 107

3.3 Concluding remarks 107

4 Dissecting network motifs by identifying promoter features that govern differential gene expression 109

4.1 Results 111

4.1.1 Approach 111

4.1.2 Transcription factor binding site submotifs 112

4.1.3 Transcription factor binding site orientation 114

4.1.4 RNA polymerase site 116

4.1.5 Activated/Repressed promoters 116

4.1.6 Binding sites for other transcription factors 118

4.1.7 Regulation of orthologous genes 119

4.2 Materials and Methods 121

4.2.1 Network motifs 123

4.2.2 Binding site submotifs and orientation 123

4.2.3 RNA polymerase sites 124

4.2.4 Activated/repressed 125

4.2.5 Binding sites for other transcription factors 126

4.2.6 Fuzzy logic expressions 126

4.2.7 Fuzzy membership functions 127

4.2.8 Performance Measurement 127

4.2.9 Genetic algorithm 127

4.2.10 Dataset 128

4.2.11 Programming resources 128

4.2.12 Bacterial strains, plasmids and growth conditions 129

4.3 Concluding remarks 129

5 Gene promoter scan methodology for identifying and classifying co-regulated promoters 131

5.1 Identifying the promoter features governing gene transcription 135

5.2 The GPS methodology as an integrated algorithm 135

5.3 Exploring the targets of regulation of a response regulator using GPS. 138

5.3.1 GPS built-in features. 138

5.3.2 GPS initialization strategy. 140

5.3.3 GPS grouping strategy. 142

5.3.4 GPS evaluation strategy. 145

5.3.5 GPS validation strategy. 146

5.3.6 Predicting new members 147

5.4 Technical specifications of GPS 147

5.4.1 Programming resources. 147

5.4.2 User interface 148

5.4.3 Parallel execution 148

5.4.4 GPS input. 148

5.4.5 Feature specifications. 148

5.4.6 GPS output. 152

5.5 Uncovering promoter profiles regulated by the response regulator PhoP using GPS 153

5.5.1 Experimentally validated profiles ... 153
5.5.2 GPS performance ... 156
5.5.3 Comparision to other methods ... 156
5.5.3.1. Bayesian Networks. ... 157
5.5.3.2. Association rules. ... 157
5.5.3.3. Decision Tree. ... 157
5.5.3.4. Comparisons among validated profiles ... 158
5.5.4 PhoP profiles as discriminators of kinetic behaviour ... 160
5.6 Concluding remarks ... 161

6 Learning robust dynamic networks in prokaryotes by Gene Expression Networks Iterative Explorer (GENIE) ... 165
6.1 Introduction ... 165
6.2 Problem: Computational and Biological Challenges ... 167
6.2.1 Modeling genetic networks ... 167
6.2.2 Two-component systems ... 168
6.3 Discovering genetic networks using GENIE ... 169
6.3.1 System components ... 170
6.3.1.1. Network architecture ... 171
6.3.1.2. Network constrains ... 171
6.3.2 System identification ... 172
6.3.2.1. Learning network parameters and species ... 172
6.3.2.2. Evaluating networks using probability measurements ... 172
6.3.3 Sensitivity analysis of parameters ... 173
6.4 Application of GENIE ... 174
6.4.1 Concluding remarks ... 179

7 Plotting schizophrenia risk factor function by learning phenotypic/genotypic relations ... 181
7.1 Introduction ... 181
7.1.1 Genotypic characteristics of schizophrenia ... 182
7.1.2 Phenotypic characteristics of schizophrenia ... 183
7.1.3 Learning phenotypic genotypic relations ... 184
7.2 Datasets ... 186
7.2.1 Study Population. ... 186
7.2.2 Ascertainment. ... 186
7.2.3 Evaluation of the phenotype. ... 186
7.2.4 Genotyping. ... 187
7.3 Data Analysis. ... 187
7.3.1 Exploring phenotypic commonalities ... 187
7.3.2 Exploring genotypic commonalities ... 189
7.3.3 Fusing distinct domain of knowledge ... 191
7.3.4 Learning schizophrenia risk factor of the relations ... 193
7.4 Results ... 194
7.4.1 Informativeness of the Kolla sample compared to HapMap data. 197
7.4.2 Phenotypic/genotypic relations. ... 197

7.4.3 Reaction surface of risk based on shared SNPs and the quantitative phenotype. 202
7.5 Methods 204
7.5.1 Decision trees 204
7.5.2 Latent Classes 204
7.5.3 Nonnegative matrix factorization 205
7.6 Concluding remarks. 206

8 Comentarios finales 211
8.1 Resumen y conclusiones 211
8.2 Trabajos futuros 212
8.3 Publicaciones derivadas de esta tesis 215

Back Matter

Appendix A 217
Appendix B 225
Bibliography 239

List of Tables

Table 3-1 Scores of the CRP Single Motif 83
Table 3-2 CRP Single motif vs. submotif optimized by SCC performance comparison. 87
Table 3-3 CRP Single motif vs. submotif optimized by CC performance comparison. 87
Table 3-4 CRP Single motif vs. *cis*-features classifiers *by SCC* performance comparison. 88
Table 3-5 CRP Single motif vs. Submotif & distances by SCC performance comparison. 89
Table 3-6 PhoP submotifs crossvalidation. 91
Table 3-7 Distinguishing physical properties of PhoP submotifs........ 98
Table 5-1Confusion matrix for GPS 156
Table 5-2 GPS vs. other Conceptual clustering techniques comparison........ 159
Table 6-1 Patterns of input/output* for the PhoP/PhoQ-PrmA/PrmB systems. 176
Table 6-2 Evaluation of the performance of the GA. 177
Table 6-3 Performance comparison (Random walk vs. GA) 177
Table 7-1 Clustering methods & parameters 188
Table 7-2 Summary of descriptive statistics for the sample. 195
Table 7-3 Relations summary........ 198
Table 7-4 Relevant SNPs analyzed for identified for the relations (First batch, pending further analysis). 200
Table 7-5 Description of SNPs listed in Table 7-4 201
Table 9-1 PhoP regulated promoters raw data........ 221
Table 10-1 Equations that model the final refined model.........227
Table 10-2 Parameters that fulfill less than the 50% of their biological range .. 236
Table 10-3 Predicted patterns of behavior........ 236
Table 10-4 Predicted patterns of behavior. Correlation of predictions to GFP experiments. 237

List of Figures

Figure 1.1 Célula eucariota y detalle de sus orgánulos 22
Figure 1.2 Esquema de un nucleótido 22
Figure 1.3 Azúcares 23
Figure 1.4 Bases nitrogenadas 24
Figure 1.5 Replicación de las hebras de AND 24
Figure 1.6 Estructura en doble cadena del ADN 25
Figure 1.7 Polinucleótidos de ADN y ARN 26
Figure 1.8 Estructura química de 4 de los 20 aminoácidos que componen las proteínas 27
Figure 1.9 Algunas estructuras tridimensionales de proteínas 27
Figure 1.10 Portada de la revista Times dedicada a la clonación de la oveja Dolly, primer clon de mamífero obtenido a partir de una célula de animal adulto. 30
Figure 1.11 Proceso de creación de un microarray de ADN. 32
Figure 1.12 Diagram of typical dual color microarray experiment. 36
Figure 1.13 Affymetrix Chips 37
Figure 1.14 Probe Set structure in Affymetrix GeneChips® 38
Figure 1.15 Scanned image of an Affymetrix array. 39
Figure 2.1 Diferentes formas de los clusters 48
Figure 2.2 Clustering jerarquico aplicado aun conjunto de 20 puntos generados al azar 50
Figure 2.3 Subtractive clustering aplicado aun conjunto de 20 puntos generados al azar 51
Figure 2.4 Árbol de decisión 53
Figure 2.5 Estructura del modelo de clases latentes 54
Figure 2.6 NMF learns parts-based representation of objects. 56
Figure 2.7 Diferencia entre cercanía y cohesión conceptual 57
Figure 2.8 Dominancia entre soluciones 60
Figure 2.9 Ejemplo de clasificación mediante *k*-NN 65
Figure 2.10 Aplicación del operador de Implicación 66
Figure 2.11 Población 69
Figure 2.12 Genotipo vs. Fenotipo 70
Figure 2.13 Generaciones 70
Figure 2.14 Ejemplo de aplicación del mecanismo de selección 72
Figure 2.15 Ejemplo de aplicación del operador de cruce simple de un punto .. 72
Figure 2.16 The five phases of the proposed the methodology 77
Figure 3.1 The Divide & Conquer Method. 85
Figure 3.2 Families of submotifs describe the Phop regulon. 90

Figure 3.3 IF-THEN rules encompassing submotifs and distances between PhoP and RNAP BSs. 93
Figure 3.4. Sensitivity analysis of *D&C* parameters 94
Figure 3.5 Genome-wide analysis of *Salmonella* using PhoP submotifs. 95
Figure 3.6 Phop submotifs characterized by the physical properties of DNA-binding protein interaction 97
Figure 3.7 Evolution of the PhoP-PhoQ two component system analyzed by PhoP submotifs 100
Figure 3.8 Submotifs as a model of evolution in PhoP BSs. 102
Figure 4.1 The PhoP/PhoQ system employs a variety of network motifs to regulate gene transcription 111
Figure 4.2 The PhoP protein achieves differential expression using the single-input network motif by controlling genes that differ in their binding site submotifs. 113
Figure 4.3 The PhoP binding site submotifs. 114
Figure 4.4 Expression of PhoP-regulated promoters that differ in the orientation of the PhoP-binding site 115
Figure 4.5 Learning and prototyping the relationships between the proximity of a transcription factor binding site and the RNA polymerase site. 117
Figure 4.6 Expression of PhoP-regulated promoters that differ in the RNA polymerase sites. 118
Figure 4.7 Learning promoter features. 120
Figure 4.8 Expression of PhoP-regulated promoters that use the bi-fan network motif. 121
Figure 4.9 PhoP-regulated promoters are described on the basis of five types of features 122
Figure 5.1 The PhoP/PhoQ system controls the expression of a large number of genes, in a direct or indirect fashion. 134
Figure 5.2 . The GPS method 137
Figure 5.3 Schematics of PhoP-regulated promoters harboring different features analyzed by GPS 141
Figure 5.5 GPS navigates through the feature-space lattice generating profiles. 143
Figure 5.6 Using GPS to build promoter profiles 144
Figure 5.7 Pareto optimal frontier. 146
Figure 5.8 Selection of the most representative profiles. 155
Figure 5.9 . Independent validation of profiles using kinetic classes 161
Figure 6.1 The PhoP/PhoQ-PmrA/PmrB functional scheme in *Salmonella enterica serovar Typhimurium*. 169
Figure 6.2 . Flowchart of the GENIE methodology. Each phase is decomposed into different task that are implemented in the methodology. 170
Figure 6.3 . Final refined model. 176
Figure 6.4 . Predicted and experimentally validated gene expression level. 178
Figure 6.5 . Robustness analysis. 178
Figure 6.6 . Predicted expression patterns. 179
Figure 6.7 . Phop regulated genes growth kinetics for GFP. 179
Figure 7.1 Clusters learned for the clinical dataset 188
Figure 7.2 Decision tree for learning the features that best characterize a cluster 189
Figure 7.3 Synthesis of QTL analysis 190

Figure 7.4 Clusters learned for the SNPs dataset 191
Figure 7.5 Heatmap of probability of intersection between phenotypes and genotypes. 192
Figure 7.6 Matrix of phenotypic/genotypic relations analyzed by quadrants. 193
Figure 7.7 show transcranial ultrasound images. 195
Figure 7.8 . The oral version of the Trail Making test without letters and presumably independent of language 197
Figure 7.9 Distribution of subjects among the relations 198
Figure 7.10 Hierarchical organization of the rules. 200
Figure 7.11 Surface of reaction of the risk of schizophrenia. 203
Figure 7.12 Structure of LC Model 205
Figure 7.13 NMF learns parts-based representation of objects. 206
Figure 9.1 CRP Submotifs 217
Figure 9.2 ". Set of optimal solutions for CRP submotifs 218
Figure 9.3 Learning and prototyping the relationships between the proximity of a transcription factor binding site and the RNA polymerase site 219
Figure 9.4 IF-THEN rules encompassing CRP single motif and distances between its BS and RNAP. 220
Figure 9.5 IF-THEN rules encompassing PhoP single motif and distances between BS and RNAP evaluated for genes activated by PhoP. 222
Figure 9.6 Physical properties for PhoP BS 223
Figure 9.7 PhoP Submotifs for *Salmonella* and *E. Coli* are extended to encompass *Yersinia* 224
Figure 10.1 Modeling genetic interactions by analyzing Transcription Factor binding sites. 225
Figure 10.2 Reduced model. 226
Figure 10.3 Robustness of parameters 229

Introducción

I. Problemática

Uno de los desafíos más importantes de la llamada era post-genómica es lograr identificar las piezas de información, aún casi completamente desconocidas, que especifican cuándo, dónde y por cuánto tiempo los genes son activados o reprimidos (Brenner 2000)]. Esta afirmación la comprendemos al considerar que organismos de las formas más diferentes están construidos de una misma batería de genes; y que la diversidad de formas de vida existentes resultan producto de pequeños cambios en los sistemas reguladores que gobiernan la expresión de estos genes (Jacob 1998) .

Los avances en biología molecular y nuevas tecnologías computacionales nos permiten investigar sistemáticamente los procesos moleculares complejos que subyacen debajo de los sistemas biológicos (Durbin 1998). En particular, el continuo desarrollo de grandes repositorios de conocimiento e información han facilitado el acceso a una gran cantidad de datos provenientes tanto de *microarray*; *chromatin immuno-precipitation* (ChIP); *green fluorescente protein* (GFP) ; y de *single nucleotide polymorphism* (SNP). Su disponibilidad abre nuevas oportunidades para el estudio de cómo un genotipo, que es el contenido genético o genoma específico de un individuo, da a lugar a un fenotipo, que son las características morfológicas, de desarrollo, propiedades bioquímicas y/o fisiológicas de un organismo. Entender las bases genéticas de un fenotipo es imprescindible para estudiar casos como la virulencia de una bacteria (Zwir, Shin et al. 2005), o los factores de riesgo de una enfermedad genética compleja (Gottesman and Shields 1973). La comprensión de las relaciones genotipo/genotipo posibilita el desarrollo de nuevas técnicas de diagnósticos así como de drogas mas eficientes y con menor efecto lateral. Un estudio de los casos más evidentes o generales de estas relaciones genera un entendimiento sesgado y probablemente erróneo, ya que no todos los organismos con un mismo genotipo tienen una misma característica observable o actúan de la misma manera. Asimismo, no todos los organismos que tienen un rasgo similar tienen un mismo genotipo.

La genética cuantitativa por lo general carece de la resolución requerida para identificar las diferencias en las secuencias de ADN responsables de un fenotipo particular (Wray, Hahn et al. 2003). Sin embargo, al ser combinadas con nuevos test experimentales se puede identificar variaciones específicas en las secuencias de ADN (Frazer, Ballinger et al. 2007). Los SNPs son el tipo de variación más común, en la cual un solo nucleótido difiere para los miembros de una misma especie. Por ejemplo, en los seres humanos se estima una ocurrencia de diez millones de SNPs. Debido a los SNPs estan correlacionados según la región del ADN en que ocurren (i.e. *haplotypes*), seiscientos mil SNPs son considerados como marcadores suficientes para identificar regiones potencialmente involucradas en un fenotipo diferencial (Frazer, Ballinger et al. 2007). Una vez localizadas estas regiones es posible realizar análisis sobre la expresión de los genes contenidos en éstas, así como de las proteínas codificadas por estos genes.

La expresión genética es central para el entendimiento de la relación genotipo/fenotipo en todos los organismos (Wray, Hahn et al. 2003), y es un componente importante de las bases genéticas que dan lugar a los cambios evolutivos en los diferentes aspectos del fenotipo. Sin embargo, la regulación de la expresión de los genes aun no está completamente entendida. En particular, la regulación transcripcional es el mecanismo que controla en que momento y en que cantidad el ADN de un gen es transcripto a ARN. Secuencias de ADN cercanas a los genes, llamadas elementos *cis*, son clave en el proceso de regulación transcripcional (Elemento, Slonim et al. 2007). Resulta muy dificultoso obtener información de la función aproximada de los elementos *cis* ya que su caracterización requiere de técnicas laboriosas, no siempre dominadas en los laboratorios (Wray, Hahn et al. 2003). Asimismo, la información comparativa de su funcionamiento continua siendo limitada, debido a que el análisis bioquímico y funcional de los elementos *cis* se limita a pocos casos y a una fracción de ellos, y más aún si se tiene en cuenta que estos elementos son fuertemente dependientes del contexto en que se encuentran.

Un estudio minucioso de la relación genotipo/genotipo así como la expresión genética requiere técnicas computacionales que cumplan requerimientos específicos muchas veces aún no satisfechos: *i)* deben poder buscar y recuperar modelos cuantitativos, cualitativos e interpretables por los expertos; *ii)* deben poder generar un conjunto de hipótesis optimales; incluyendo aquellas más especificas (i.e. soportadas por un menor número de observaciones) así como las más generales (i.e. soportadas por un mayor número de observaciones); *iii)* los modelos aprendidos deben poder ser organizados jerárquicamente, facilitando su compresión; *iv)* los modelos deben describir las observaciones desde distintas ópticas o puntos de vista, dado que se desconoce que características pueden ser importantes o no, resultando a priori todas válidas; *v)* las observaciones deben poder dar soporte a mas de un modelo; *vi)* los modelos deben reflejar la organización propia de cada dominio de información, permitiendo su agregación en forma independiente (i.e. no sesgada); *vii)* los modelos deben poder predecir el funcionamiento de otras observaciones, para así poder inferir nuevo conocimiento.

II. Objetivos

La presente memoria corresponde a un trabajo interdisciplinar, que involucra a la biología de sistemas, la biología molecular, medicina, bioinformática y ciencias de la computación, por lo que los objetivos pueden considerarse desde la perspectiva de la problemática de cada disciplina. Por este motivo nos planteamos objetivos en cuanto a la solución de los problemas biológicos a investigar, así como en lo relativo a los modelos computacionales a emplear para la solución de dichos problemas.

Desde el punto de vista biológico, el objetivo general de la presente memoria es encontrar las características intrínsecas de un genotipo que dan lugar a un fenotipo, incluyendo el estudio de la expresión genética. Puntualmente, las preguntas que planteamos en referencia a esta último son:

- ¿De qué manera un gen regula a otro gen? ¿ Es esta regulación directa o mediante algún otro regulador? ¿Cuales son y cómo están organizados los sitios de unión al ADN usados por una proteína? ¿Qué otros elementos *cis* influyen en la regulación de los genes? ¿Qué relación existe entre los elementos *cis*?
- ¿Cuándo se expresa un gen? ¿Con que intensidad? ¿Qué factores influyen para que un gen se exprese en forma diferencial de otro co-regulado?
- ¿Cómo evolucionan los reguladores y los genes co-regulados? ¿Se conservan los elementos *cis* en los distintos organismos? ¿Se puede plantear un modelo de evolución en base a los elementos *cis*?

La respuesta a estas incógnitas permitirá echar luz a los mecanismos de regulación transcripcional en organismos procariotas y entender la relación existente genotipo/fenotipos.

Para el análisis de regulación transcripcional estudiaremos las redes genéticas denominadas sistemas de dos componentes en procariotas. Particularmente, estudiaremos el sistema PhoP/PhoQ presente en las gamma enterobacterias cuyos genes están involucrados en la virulencia las bacterias, así como su respuesta ante antibióticos (Zwir, Shin et al. 2005). El proyecto lo realizamos en colaboración con el laboratorio del Dr. Groisman del Howard Hughes Medical Institute, Departamento de Microbiología molecular, Washington University, St. Louis, MO, Estados Unidos de América. Para lograr nuestro propósito consideraremos los siguientes sub-objetivos:

- Analizar y agrupar la expresión de los genes regulados por PhoP en *Escherichia coli* y *Salmonella* enterica serovar Typhimurium en base a resultados obtenidos a partir de técnicas basadas en *microarray* (Nimblegen tiling arrays) y GFP (VICTOR, Perkin Elmer).
- Aprender patrones genotípicos de las regiones promotoras de genes regulados por PhoP según sus características *cis*. Validar los sitios de unión de PhoP al ADN mediante experimentos ChIP (Nimblegen ChIP-chip arrays).

- Relacionar los perfiles de regulación con los perfiles genotípicos, para inferir mecanismos que la célula usa para obtener una expresión diferencial de genes co-regulados.
- Utilizar este nuevo conocimiento para predecir nuevos genes regulados por PhoP en *E. Colli* y *Salmonella,* así como otras gamma enterobacterias.

Asimismo, abordamos el estudio de enfermedades con componentes genéticos en organismos eucariotas, concentrándonos en encontrar relaciones entre genotipos y genotipos relevantes en pacientes con esquizofrenia. Este proyecto lo realizamos en colaboración con el laboratorio del Dr. de Erausquin del Departamento de Psiquiatría, Washington University, St. Louis, MO, Estados Unidos de América. Los subobjetivos planteados son los siguientes:

- Identificar y agrupar individuos de una población de nativos americanos (i.e. Collas habitantes del norte de la Argentina) en base a sus características genotípicas. El estudio se realiza a un conjunto de 72 individuos, distribuidos uniformemente entre pacientes que sufren de esquizofrenia, familiares de éstos e individuos de control, a los cuales se les realiza el análisis de SNPs (*Affymetrix GeneChip Human Mapping 10k Array v2*).
- Analizar y agrupar esta misma población mediante sus características cognitivas, de comportamiento, estructurales y motoras (i.e. perfiles fenotípicos),
- Relacionar los perfiles genotípicos y fenotípicos para identificar el riesgo de la enfermedad y los posibles orígenes genéticos para diferentes estratos de la población en estudio.
- Construir una función de riesgo de la esquizofrenia para la población de estudio en base a las relaciones aprendidas. Construir predictores de fenotipos en base a genotipos y viceversa.

En consecuencia, para cumplimentar los anteriores objetivos biológicos, planteamos objetivos metodológicos que nos permitan abordar los mencionados problemas. Emplearemos técnicas de análisis inteligente de datos y de descubrimiento de conocimiento, que puedan ser generalizadas y extendidas a sistemas similares, ya sea el estudio de otros reguladores o bien otras enfermedades. Los objetivos de nuestro marco de trabajo computacional para el entendimiento, interpretación y predicción de relaciones genotípicas/fenotípicas, incluyendo la regulación transcripcional, son los siguientes:

- Detectar la información relevante proveniente de la literatura, bases biológicas y la experimentación de laboratorio. Codificar esta información en forma adecuada, creando bases de información independientes para cada dominio.
- Descubrir patrones en los distintos dominios de información (e.g. secuencias de ADN correspondientes a genes co-expresados; expresión de genes; SNPs; y fenotípicos). Las técnicas a emplear han de ser capaces de manipular información incompleta, imprecisa y ambigua; también

han de proveer modelos optimales en cuanto al numero de observaciones que recuperan y la cantidad de características que las distingue.

- Integrar los modelos descubiertos, generando hipótesis alternativas que puedan explicar desde diferentes puntos de vista las relaciones genotipo/fenotipo (e.g. elementos *cis* y expresión diferencial; SNPs y características cognitivas, motoras, etc.).
- Estudiar la ocurrencia y/o las posibles transformaciones de estas hipótesis en distintos organismos y plantear un modelo de evolución.
- Predecir nuevas observaciones en base a los modelos aprendidos.

III. Resumen

Para desarrollar los objetivos planteados, esta memoria está organizada en siete capítulos, una sección de comentarios finales y un apéndice. A continuación describimos brevemente cada una de estas partes:

En el Capítulo 1 introducimos los conceptos básicos de biología molecular necesarios para la adecuada comprensión de los capítulos posteriores. Comenzamos con una breve descripción de los componentes principales de los organismos vivientes, continuamos describiendo los procesos necesarios para la supervivencia de la célula, y finalizamos con una breve reseña acerca de los métodos biológicos para el estudio de secuencias de ADN. Adicionalmente, hacemos una introducción a la *Bioinformática* y los problemas tratados.

En el Capítulo 2 presentamos las diferentes técnicas y métodos computacionales sobre los cuales se basa la metodología que proponemos en esta memoria. Los temas que desarrollamos en este capítulo son: el modelado o la identificación de sistemas, el uso de *clustering conceptual* de datos para la detección de patrones en cada dominio de información; la *lógica difusa*, para la fusión de la información, la *optimización multiobjetivo* para la selección de los modelos optimales; y los *algoritmos evolutivos* para la optimización de los predictores. Finalizamos este capítulo con una descripción de la metodología, que hace uso de estas técnicas, y nos permite abordar los diferentes problemas biológicos

Los siguientes cuatro capítulos detallan nuestro estudio de la redes de regulación transcripcional en organismos procariotas:

En el Capítulo 3 estudiamos de los sitios de unión al DNA. Para ello aplicamos la técnica de "Divide y Vencerás" (*Divide & Conquer*) a los sitios de unión de dos reguladores maestros (i.e. PhoP y CRP) en *E. coli* y *Salmonella* para obtener familias de motivos (submotivos). En adición a las ventajas computacionales que obtenemos para la clasificación de estas secuencias, mostramos como los submotivos alivian el problema de determinar si los sitios de unión al ADN predichos son funcionales o no; permiten revelar propiedades físicas de la interacción proteína-ADN; y establecen un modelo de evolución mediante la adquisición y perdida modular de submotivos.

En el Capitulo 4 extendemos el estudio a otros elementos *cis* involucrados en la expresión diferencial de los genes. Con una aproximación basada en expresiones de lógica difusa (*fuzzy logic expressions*) analizamos los genomas de bacteria, lo que nos permite considerar la variabilidad de las secuencias, posicionamientos y topologías de regiones clave para la expresión diferencial de los genes. Aplicamos este método para caracterizar los genes inmersos en las distintas arquitecturas de redes de regulación (*network motifs*) que son regulado por PhoP en *Escherichia coli* y *Salmonella* enterica serovar Typhimurium. Identificamos rasgos clave que permiten a PhoP producir distintos patrones de expresión de los genes regulados.

En el Capitulo 5 integramos los patrones aprendidos en los capítulos 3 y 4, generando modelos que agregan los elementos *cis* meditante un método que denominamos *Gene Promoter Scan* (GPS) basado en *clustering conceptual*, el cual selecciona los modelos optimales mediante la el uso de optimización *multi-objetivo* y *multimodales*. La aplicación de este método al sistema regulador de dos componentes PhoP/PhoQ de *E. coli* y *Salmonella* nos posibilita descubrir nuevos genes directamente regulados por esta proteína. Los hallazgos son validados experimentalmente para verificar que PhoP utiliza distintos mecanismos de regulación transcripciónal.

En el Capitulo 6 analizamos la respuesta dinámica de los genes regulados por el sistema de dos componentes PhoP/PhoQ. Exploramos los diferentes arquitecturas posibles de ser generadas. Utilizamos *algoritmos genéticos* y *random walk* para aprender que tan robustas, reales y flexibles pueden llegar a ser. Finalizamos contrastando las predicciones con los resultados experimentales obtenidos mediante análisis de GFP. La aplicación de este método a la red de regulación genética de *Salmonella* revela los mecanismos que posibilitan la interconexión de los sistemas de dos componentes PhoP/PhoQ y PmrA/PmrB. La validación experimental demuestra que tanto la regulación transcripcionales como la post-transcripcional son empleados en la célula para conectar ambos sistemas.

En el Capítulo 7 estudiamos las relaciones genotipo/fenotipo que caracterizan a una población de enfermos de esquizofrenia. Si bien ésta es un desorden altamente hereditable, aún no se ha logrado identificar los genes involucrados en la misma. Analizamos datos genotípicos (SNPs) de una población de individuos que padecen la enfermedad, parientes de éstos y individuos de control, como así también datos fenotípicos (datos clínicos) para luego identificar las relaciones más significativas y cualitativas entre ambos dominios.. Las relaciones aprendidas nos permiten realizar predicciones sobre fenotipos basados en genotipos y viceversa. Asimismo, las relaciones nos permiten modelar la función de riesgo de padecer esquizofrenia para la población estudiada. Ésta función fue planteada teóricamente pero a nuestro conocimiento es la primera vez que se plantea una superficie de riesgo basada en múltiples causas genéticas.

Chapter 1

Fundamentos Biológicos y Bioinformática

Todos los seres vivos están formados por células que comparten una maquinaria común para sus funciones más básicas. Los seres vivos, aunque infinitamente diversos por fuera, son muy similares por dentro (Figure 1.1). En este capítulo expondremos las características universales de todos los seres vivos, analizando brevemente la diversidad celular, y veremos cómo, gracias a un código común en el que están escritas las especificaciones de todos los organismos, es posible leer, medir y desentrañar estas especificaciones para alcanzar un conocimiento coherente de todas las formas de vida, de las más simples a las más complejas. Luego de esta introducción al dominio de estudio, si podremos definir el problema de interés que se aborda en este trabajo.

1.1 Material Genético en la Célula

Se calcula que las células llevan evolucionando y diversificándose más de tres mil millones y medio de años (Berg, Tymoczko et al. 2003). Todas las células vivas, sin ninguna excepción conocida, guardan su información hereditaria en el material genético: moléculas de ADN (abreviatura de ácido desoxirribonucleico) de doble cadena -dos largos polímeros paralelos no ramificados formados por cuatro tipos de monómeros (el material esencia o unidad con la cual se construye un polímero.)-. Estos monómeros están unidos entre sí formando una larga secuencia lineal que codifica la información genética de la célula (Alberts, Johnson et al. 2003; Berg, Tymoczko et al. 2003)

Los organismos vivos pueden clasificarse en dos grupos atendiendo a su estructura: los organismos **eucariotas** y los **procariotas**. Los eucariotas guardan su ADN en un compartimiento intracelular denominado núcleo. Los procariotas no presentan un comportamiento nuclear diferenciado para almacenar su ADN. Las plantas, los hongos y los animales son eucariotas; las bacterias son procariotas (Alberts, Johnson et al. 2003).

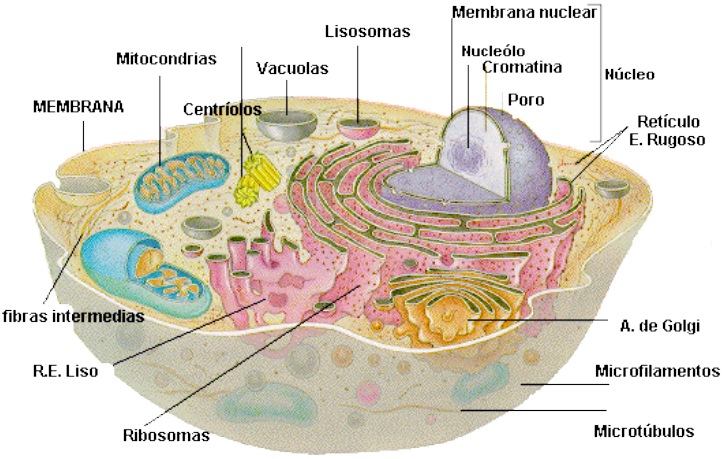


Figure 1.1 Célula eucariota y detalle de sus orgánulos

Para comprender los mecanismos biológicos, primero tenemos que conocer la estructura de la molécula de ADN. Cada monómero de una de las cadenas sencillas del ADN -denominado **nucleótido** (Figure 1.2)- tiene dos partes: un azúcar (la desoxirribosa, (Figure 1.3) con un grupo fosfato unido y una *base* que puede ser adenina (A), guanina (G), citosina (C) o timina (T) (Figure 1.4). Cada azúcar está unido al siguiente azúcar de la cadena por el grupo fosfato mediante un enlace fosfodiéster, formando un polímero cuyo eje central está compuesto por los azúcares fosfato y del cual sobresalen las bases. El polímero de ADN crece por la unión de monómeros a uno de sus extremos. En el caso de una cadena sencilla de ADN, los monómeros pueden incorporarse al polímero de forma aleatoria, sin un orden preestablecido, ya que todos los nucleótidos pueden unirse entre sí en el sentido del crecimiento del polímero de ADN.

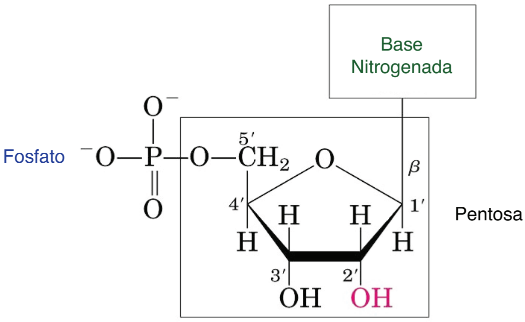


Figure 1.2 Esquema de un nucleótido

Por el contrario, en la célula viva existe una limitación, ya que el ADN no se sintetiza como una cadena libre aislada sino sobre un patrón o molde de ADN de otra cadena preexistente. Las bases contenidas en la cadena patrón se unen a las bases de la nueva cadena siguiendo una estricta norma de complementariedad: A se une a T, y C se une a G (Figure 1.5). Este emparejamiento controla la selección del monómero que se añade a la cadena. De esta forma, una estructura de doble cadena consiste en dos secuencias complementarias de A, C, G y T. El orden de la secuencia es muy importante, ya que en él reside la información contenida en el ácido nucleico. La orientación viene dada en el sentido 5'-3' o 3'-5', donde el 5' representa el extremo terminal del fosfato y el 3' el extremo final del átomo de carbono de la desoxirribosa. Además, las dos cadenas de nucleótidos se enrollan una sobre la otra generando una doble hélice (Figure 1.6).

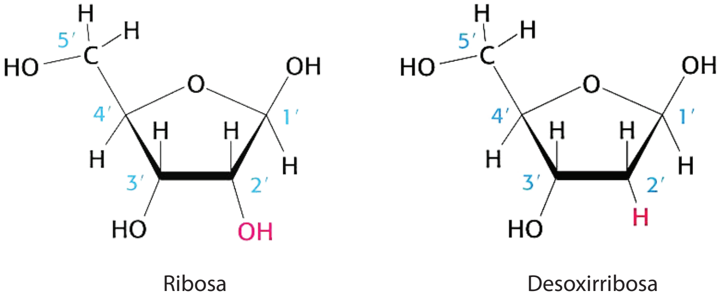


Figure 1.3 Azúcares

El ADN tiene la capacidad de expresar su información para gobernar el comportamiento de otras moléculas de la célula. El mecanismo responsable de este proceso es el mismo en todos los organismos vivos y se inicia con la síntesis secuencial de dos tipos de moléculas: el ácido ribonucleico (ARN) y las proteínas. El proceso comienza con la polimerización sobre un patrón, denominada **transcripción**, proceso en el que diferentes segmentos de la secuencia de ADN se utilizan como molde para la síntesis de moléculas cortas de un polímero muy relacionado con el ADN: el **ácido ribonucleico** o **ARN**. Después de un proceso complejo denominado **traducción**, muchas de estas moléculas de ARN se utilizan para dirigir la síntesis de polímeros de una clase química radicalmente diferente: las *proteínas*.

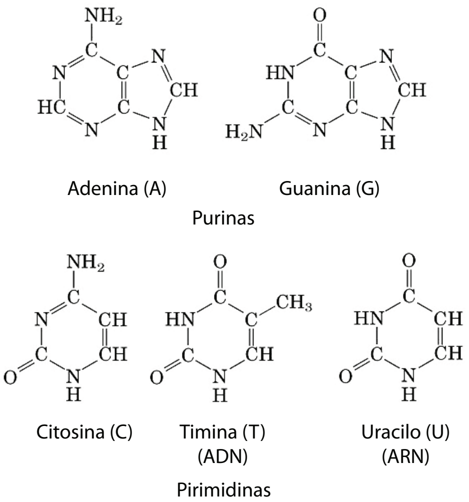


Figure 1.4 Bases nitrogenadas

Los ennlaces establecidos entre las bases son débiles si se comparan con las uniones azúcar-fosfato del resto del esqueleto. Esta debilidad permite separar las dos cadenas de ADN sin forzar la rotura de su esqueleto. Cada una de las cadenas puede comportarse como un molde para la generación de su pareja mediante la formación de pares de bases específicos. Es precisamente esta capacidad para la generación de nuevas hebras de ADN la que le permite crear nuevas células con idéntico material genético a la célula replicada.

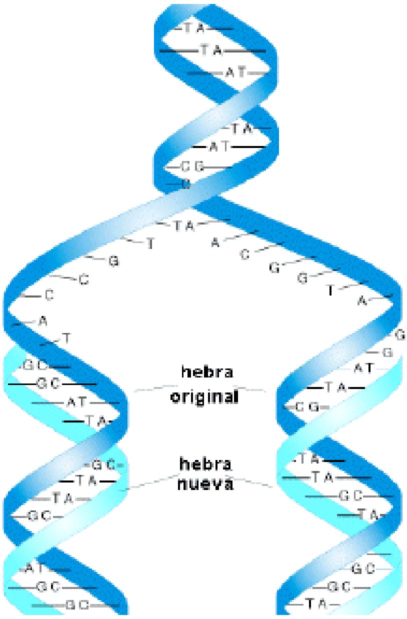


Figure 1.5 Replicación de las hebras de AND

En el ARN, el esqueleto del polímero está formado por azúcares ligeramente diferentes a los del ADN -ribosa en lugar de desoxirribosa- y, además, una de las cuatro bases es diferente -uracilo (U) (Figure 1.4) en el lugar de la timina (T)-, pero las otras tres bases -A, C, G- son las mismas y se emparejan con su complementaria, como en el ADN -la A, la U, la C y la G del ARN se unen con la T, la A, la G y la C del ADN, respectivamente. Durante la transcripción, los monómeros de ARN se seleccionan para la polimerización del ARN sobre una cadena molde de ADN, de la misma manera que se seleccionan los monómeros de ADN durante la replicación del ADN.

El resultado de la transcripción es un polímero de ARN que contiene una parte de la información genética de la célula, aunque escrita en un alfabeto diferente de monómeros de ARN en lugar de monómeros de ADN.

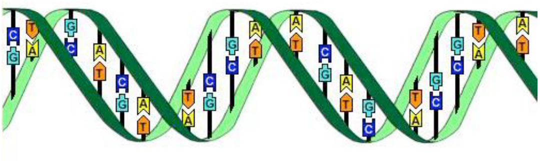


Figure 1.6 Estructura en doble cadena del ADN

El papel principal de muchas secuencias de ADN es el de codificar secuencias de las **proteínas**, el componente mas activo de la célula, que participan en todos los procesos esenciales. Al igual que el ADN y el ARN, las proteínas son polímeros no ramificados formadas por monómeros, los **aminoácidos**, muy diferentes de los del ADN o el ARN y de los que existen veinte tipos diferentes en lugar de tan sólo cuatro (Figure 1.8). Los aminoácidos tienen una estructura central semejante por la que pueden unirse entre ellos. Junto a esta estructura central, se encuentra un grupo lateral que confiere a cada aminoácido su carácter químico característico. Cada una de las moléculas proteicas o *polipéptidos*, formadas por la unión de varios aminoácidos siguiendo una secuencia determinada, se pliega en una estructura tridimensional elaborada y muy bien definida que está determinada por la secuencia de aminoácidos de su cadena .Esta capacidad de auto-ensamblarse de las proteínas es la responsable de su papel primordial en bioquímica. Las proteínas tienen muchas funciones -ser catalizadores de reacciones (enzimas), mantener estructuras celulares, generar movimientos, traducir señales, etc.- y cada una cumple una función específica según su secuencia de aminoácidos, determinada genéticamente.

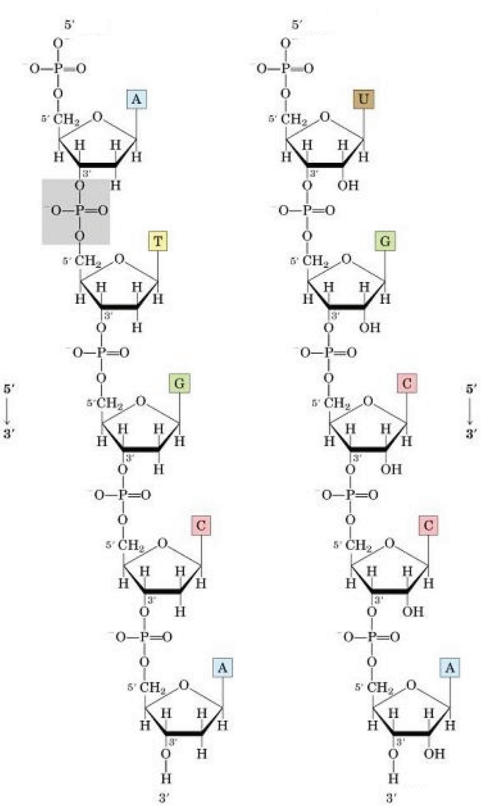


Figure 1.7 Polinucleótidos de ADN y ARN

Un mismo fragmento de la secuencia del ADN se puede usar varias veces para guiar la síntesis de muchos transcritos de ARN idénticos. Así, mientras que el archivo de información de la célula es fijo -el ADN-, los transcritos de ARN se producen en gran número y son desechables. La función de la mayoría de estos transcritos es servir de intermediarios en la transferencia de la información genética, actuando como un **ARN mensajero** (ARNm) que dirige la síntesis de proteínas según las instrucciones almacenadas en el ADN.

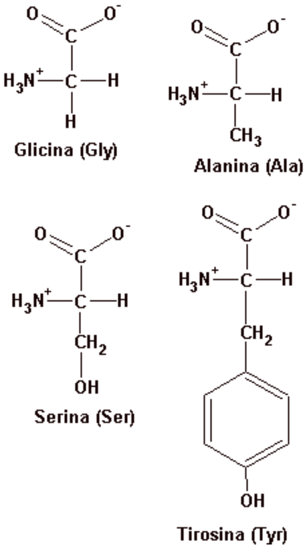


Figure 1.8 Estructura química de 4 de los 20 aminoácidos que componen las proteínas

La información contenida en la secuencia de ARNm se lee en grupos de tres nucleótidos; cada triplete de nucleótidos o *codón* especifica (codifica) un aminoácido de una proteína. Debido a que hay 64 posibles codones, pero sólo veinte aminoácidos, necesariamente hay muchos casos en los que varios codones corresponden a un mismo aminoácido. El código se lee por una clase especial de pequeñas moléculas de ARN, el ARN de transferencia (ARNt). Cada tipo de ARNt une en uno de sus extremos un aminoácido y tiene una secuencia específica de tres nucleótidos en su otro extremo -un *anticodón*- que le permite reconocer un codón o subgrupo de codones del ARNm por emparejamiento de bases.

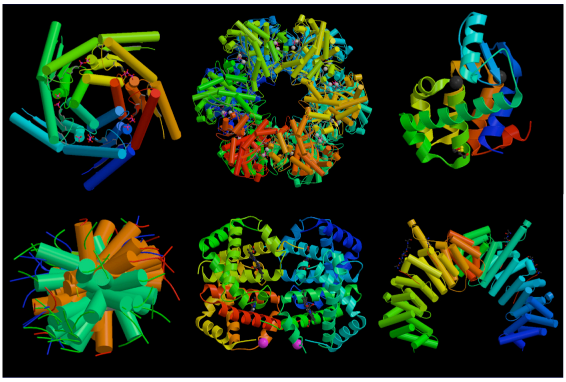


Figure 1.9 Algunas estructuras tridimensionales de proteínas

Para la síntesis de proteínas, un conjunto de moléculas de ARNt cargadas con sus aminoácidos respectivos se une a un ARNm por emparejamiento de sus anticodones con cada uno de los codones sucesivos del ARNm. Después, los aminoácidos se van uniendo de forma que la proteína naciente va creciendo y cada ARNt, relegado de su carga, se libera.

Las moléculas de ADN son muy largas y contienen la especificación de miles de proteínas. Por tanto, fragmentos de esta secuencia completa de ADN se transcriben en diferentes moléculas de ARNm, cada uno de los cuales codifica una proteína diferente. Un **gen** se define como un fragmento de la secuencia de ADN que corresponde a una sola proteína (o a una molécula de ARN catalítica o estructural, para los genes que producen ARN pero no proteína).

En todas las células, la expresión de determinados genes está regulada: en lugar de sintetizar el catálogo completo de posibles proteínas en todo momento, la célula ajusta la velocidad de transcripción y de traducción de diferentes genes de forma independiente y de acuerdo con sus necesidades. En el ADN celular existen secuencias de ADN no codificantes -denominadas *ADN regulador*- que están distribuidas entre las regiones codificantes de proteínas, y estas regiones no codificantes se unen a proteínas especiales que controlan la velocidad local de transcripción. Existen también otras regiones no codificantes, algunas de las cuales actúan como elementos de puntuación, indicando el inicio y el final de la información de una proteína. La región del ADN donde se establece cómo y cuándo se expresará el gen que se codifica en la región codificante inmediatamente adyacente se conoce como *región promotora*. En este sentido, el **genoma** de una célula -la totalidad de la información genética incluida en su secuencia completa de ADN- dicta no sólo la naturaleza de las proteínas celulares, sino también cuándo y dónde se sintetizarán.

1.2 ADN y Evolución

El material básico de la evolución es la secuencia de ADN que ya existe. No hay ningún mecanismo natural por el que se generen grandes cadenas de ADN de secuencia nueva aleatoria. Así, ningún ADN es completamente nuevo.Tanto durante el almacenamiento como durante el copiado del material genético se pueden producir accidentes y/o errores aleatorios que pueden alterar la secuencia de nucleótidos -es decir, generar **mutaciones**-. Como consecuencia de ello, cuando una célula se divide, a menudo sus dos células hijas no son idénticas entre sí o a su progenitora. Algunas veces poco frecuentes, el error puede representar un cambio favorable; más probablemente, el error no supondrá diferencias importantes en las capacidades de la célula; y en muchos casos, el error causará daños importantes -por ejemplo, alterando la secuencia de una proteína clave-. Cambios debidos a errores del segundo tipo pueden ser o no perpetuados, dependiendo de si la célula o sus familiares tienen o no éxito en la competencia por los recursos limitados del ambiente donde viven. Los cambios que causan daños importantes no conducen a la célula a ninguna parte, por lo general provocan su muerte, y por tanto, no dejan descendencia. Mediante la repetición de este ciclo de ensayo y error -de *mutación* y *selección natural*- los organismos van evolucionando y sus especificaciones genéticas van cambiando, proporcionándoles nuevas vías de aprovechamiento del entorno más eficaces para poder sobrevivir en competencia con otros organismos, reproduciéndose con más éxito. Las variaciones en fragmentos de ADN pueden ser generadas por varios métodos: (Berg, Tymoczko et al. 2003)

- *Mutación intragénica:* un gen ya existente puede ser modificado por mutaciones en su secuencia de ADN.

- *Mezcla de fragmentos:* dos o más genes existentes pueden romperse y reagruparse generando un gen híbrido formado por segmentos de ADN que originariamente pertenecían a genes independientes.

- *Transferencia horizontal:* un fragmento de ADN puede ser transferido desde el genoma de una célula al de otra célula, incluso de una especie diferente, si ambos organismos comparten el mismo ambiente. Este proceso contrasta con la *transferencia vertical* de información genética, habitual entre los progenitores y la progenie.

Una célula ha de duplicar todo su genoma cada vez que se divide en dos células hijas. Sin embargo, algunos accidentes pueden causar la duplicación de una parte del genoma, manteniendo el genoma original. Cuando un gen se ha duplicado por esta vía, una de las dos copias queda libre para mutar y especializarse en la realización de una función diferente en la misma célula. Repetidos ciclos de este proceso de duplicación y divergencia, durante millones de años, han permitido que algunos genes generen una familia completa de genes en un mismo genoma.Cuando los genes se duplican y divergen de esta manera, los individuos de una especie resultan dotados de diferentes variantes del gen inicial. Este proceso evolutivo ha de distinguirse de la divergencia genética que ocurre cuando una especie se separa en dos líneas de descendencia diferentes en una bifurcación del árbol de la vida. En este punto, los genes se vuelven diferentes en el curso de la evolución, pero continúan teniendo funciones correspondientes en las dos especies hermanas. A los genes que están relacionados de esta forma -es decir, genes de dos especies separadas que derivan de un mismo gen ancestral presente en el último ancestro común de ambas especies- se los denomina **ortólogos**. A los genes relacionados que derivan de una duplicación en el mismo genoma -y que posiblemente divergirán en sus funciones- se los denomina **parálogos** (son parálogos, por ejemplo los genes que determinan las distintas clases de hemoglobinas que se producen a lo largo de la vida fetal y adulta). A los genes que están relacionados por una descendencia de cualquier tipo se los denomina **homólogos**, un término general que se utiliza para englobar ambos tipos de relación.

Cabe destacar que los intercambios horizontales de la información genética juegan un papel muy importante en la evolución bacteriana en el mundo actual. La reproducción sexual genera una transferencia horizontal de información genética a gran escala entre dos linajes celulares inicialmente separados -los de los progenitores-. Independientemente de si esto ocurre entre especies o dentro de una misma especie, la transferencia horizontal de genes deja una huella característica: genera individuos que están más relacionados entre sí con un grupo de parientes con respecto a determinados genes y con otros con respecto a otro grupo de genes.

1.3 Biología y Avances Tecnológicos

Hasta principios de los años setenta, el ADN era la molécula de la célula que planteaba más dificultades para su análisis bioquímico. Actualmente, el ADN ha pasado a ser la macromolécula más estudiada. Ahora podemos separar una región determinada del ADN, obtener un número de copias casi ilimitado y determinar su secuencia de nucleótidos.

Estos adelantos técnicos en la ingeniería genética han tenido un impacto espectacular en la biología celular, permitiendo el estudio de las células y de sus macromoléculas mediante sistemas que antes eran inimaginables. La tecnología del ADN recombinante constituye un conjunto variado de técnicas, algunas de las cuales son nuevas y otras han sido adoptadas de otros campos de la ciencia, como la genética microbiana. Las más importantes son:

- La rotura específica del ADN mediante nucleasas de restricción, que facilita enormemente el aislamiento y la manipulación de los genes.
- La clonación del ADN, con el uso de vectores de clonación o de la reacción en cadena de la polimerasa, de tal forma que una molécula sencilla de ADN puede ser reproducida generando muchos miles de millones de copias idénticas (Figure 1.10).
- La hibridación de los ácidos nucleicos, que hace posible localizar secuencias determinadas de ADN o de ARN con una gran exactitud y sensibilidad, utilizando la capacidad que tienen estas moléculas de unirse a secuencias complementarias.
- La secuenciación rápida de todos los nucleótidos de un fragmento purificado de ADN, que hace posible identificar genes y deducir la secuencia de aminoácidos de las proteínas que codifican.
- El seguimiento simultáneo del nivel de expresión de cada uno de los genes de una célula, utilizando microchips de ADN (microarrays) que permiten efectuar simultáneamente decenas de miles de reacciones de hibridación.

A continuación describiremos en más detalle este último ítem, que un elemento fundamental en el desarrollo de esta tesis.

Figure 1.10 Portada de la revista Times dedicada a la clonación de la oveja Dolly, primer clon de mamífero obtenido a partir de una célula de animal adulto.

1.4 Microarrays de ADN

Las técnicas clásicas para el análisis de secuencias permiten examinar la expresión de un número muy limitado de genes simultáneamente. Los *microarrays*, desarrollados en los años noventa, han revolucionado la forma en la que actualmente se estudia la expresión génica, al permitir el estudio de los productos de ARN de miles de genes a la vez. Esto ha permitido la identificación y el estudio de los patrones de expresión génica que subyacen a la fisiología celular: podemos ver qué genes se encuentran activados (o reprimidos) bajo distintas condiciones o ante la presencia de agentes externos.

Un microarray o biochip es una colección de pequeños fragmentos de genes unidos a la superficie de pequeños cristales, o dicho con otras palabras, es un dispositivo de pequeño tamaño que tiene inmovilizado material biológico, que permite la automatización simultánea de miles de ensayos encaminados a conocer en profundidad la estructura y funcionamiento de nuestra dotación genética. En ellos se integran decenas de miles de fragmentos de material genético, de secuencia conocida y de diferente tamaño, ordenados sobre un sustrato sólido, de manera que forman una matriz de secuencias en dos dimensiones. Si las secuencias son cortas, se denominan microarrays de oligonucleótidos, si tienen mayor tamaño, chips de ADNc (ADN complementario, sintetizado a partir de ARNm). A los fragmentos inmovilizados en el soporte, se les denomina sondas. Los ácidos nucleicos de las muestras a analizar se pueden marcar por diversos métodos (enzimáticos, fluorescentes, etc.), incubándose posteriormente sobre la matriz de sondas, produciéndose una hibridación entre las secuencias homólogas, es decir, sólo las cadenas complementarias a las del chip se hibridan. Después de la hibridación entre las secuencias del microarray y la muestra marcada con fluorescencia, los chips son leídos en un escáner, originándose un patrón de luz característico y una cuantificación de la intensidad de hibridación de cada punto, los datos obtenidos son interpretados mediante un ordenador. Esto permite una identificación y cuantificación del ADN o ARN presente en la muestra, así como conocer la estructura y función de la dotación genética, tanto en los diferentes estados de desarrollo normal como patogénicos del paciente.

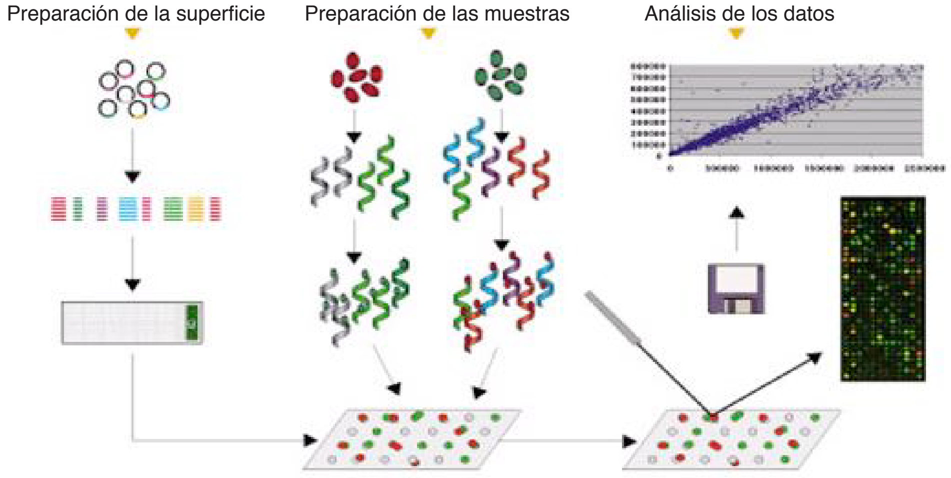


Figure 1.11 Proceso de creación de un microarray de ADN.

1.5 Biología computacional y Bioinformática

En las últimas décadas, los avances en la biología molecular y el equipamiento disponible para la investigación en este campo han permitido la rápida secuenciación de grandes porciones de genomas de diversas especies. En la actualidad, varios genomas de bacterias, tales como *Saccharomyces cerevisiae*, y algunos eucariotas simples ya han sido secuenciados por completo. El proyecto Genoma Humano (Collins, Morgan et al. 2003), diseñado con el fin de secuenciar los 24 cromosomas del ser humano, también está progresando. Las bases de datos de secuencias más populares, como GenBank (Benson, Karsch-Mizrachi et al. 2005)y EMBL (Kanz, Aldebert et al. 2005) están creciendo de forma exponencial. Esta gran cantidad de información necesita de un alto nivel de organización, indexado y almacenamiento de las secuencias. Es por ello que la Informática ha sido aplicada a la Biología para producir un nuevo campo de investigación llamado *Bioinformática* que permita ayudar a esta organización (Attwood and Parry-Smith 2002).

1.5.1 Objetivos de la Bioinformática

El término bioinformática ha sido adoptado por varias disciplinas diferentes . En su sentido más amplio, puede considerarse que el término significa tecnología de la información aplicada a la gestión y análisis de datos biológicos. Esto tiene implicaciones en diversas áreas, desde la inteligencia artificial y la robótica al análisis de genomas. En el contexto de los proyectos genoma, el término se aplicó originalmente a la manipulación computacional y al análisis de datos de secuencias biológicas (ADN o proteínas). Sin embargo, a la vista de la rápida y reciente acumulación de estructuras de proteínas disponibles, el término ahora tiende a emplearse abarcando también la manipulación y análisis de datos de estructuras tridimensionales (3D).

Las tareas más simples de la Bioinformática conciernen la creación y mantenimiento de bases de datos de información biológica. Secuencias nucleotídicas (y las secuencias proteicas que derivan de las mismas) componen la mayoría de la información que está almacenada en estos repositorios. Mientras que el almacenamiento y organización de millones de nucleótidos está muy lejos de ser una tarea trivial, el diseño de una base de datos y el desarrollo de una interfaz con la cual los investigadores puedan tanto acceder a la información existente como agregar nuevas instancias, es simplemente el comienzo.

Tal vez, la tarea más apremiante sea la que involucra el análisis de la información de secuencias. *Biología Computacional* es el nombre dado a este proceso e incluye las siguientes tareas:

- Encontrar genes en secuencias de ADN pertenecientes a varios organismos.
- Desarrollar métodos para la predicción de la estructura y/o la función de nuevas proteínas y secuencias estructurales de ARN.

- Agrupar secuencias de proteínas en familias de secuencias relacionadas y el desarrollo de modelos de proteínas.
- Alinear proteínas similares y generar árboles filogenéticos para examinar las relaciones de la evolución.

El proceso de evolución ha producido secuencias de ADN que codifican proteínas con funciones muy específicas. Es posible predecir la estructura tridimensional de una proteína usando algoritmos derivados de nuestros conocimientos en el campo de la Física, la Química y, en mayor medida, del análisis de otras proteínas con secuencias de aminoácidos similares.

La mayoría de las bases de datos biológicas consisten en largas secuencias nucleotídicas y/o secuencias de aminoácidos. Cada secuencia representa un gen o proteína particular (o una sección de la misma), respectivamente. Mientras que la mayoría de las bases de datos biológicas contienen este tipo de información, también existen otros repositorios que incluyen información taxonómica tales como características estructurales o bioquímicas de los organismos.

En las últimas tres décadas, las contribuciones al área de la Biología y de la Química han facilitado el aumento en la velocidad del proceso de secuenciación de genes y proteínas. El advenimiento de la tecnología de clonación ha permitido que secuencias de ADN foráneas sean introducidas en bacterias. De esta manera fue posible la rápida producción de secuencias de ADN particulares, un preludio necesario para la determinación de secuencias. La síntesis de oligonucleótidos dio a los investigadores la habilidad de construir pequeños fragmentos de ADN con secuencias elegidas por ellos mismos. Estos oligonucleótidos son luego utilizados como parte de bibliotecas de ADN y permiten la extracción de genes que contengan esta secuencia. Estos fragmentos de ADN también pueden ser utilizados en reacciones en cadena de polimerización para amplificar secuencias de ADN o modificar estas secuencias. Mediante estas técnicas, el progreso de la investigación biológica ha crecido exponencialmente.

Sin embargo, para que los investigadores puedan beneficiarse de esta información, es necesario cumplir con dos requisitos: (1) tener acceso inmediato al conjunto de secuencias coleccionadas y (2) tener una forma de extraer de este conjunto solamente aquellas secuencias que interesen al investigador. La simple colección, de forma manual, de toda la información necesaria para un proyecto dado a partir de un artículo de revista publicado puede convertirse rápidamente en una tarea epopéyica. Después de obtener los datos, es necesario organizarlos y analizarlos. La búsqueda manual de genes y proteínas relacionadas puede llevar semanas e incluso meses para un investigador.

La tecnología informática ha proporcionado la solución a este problema. Los ordenadores, no solo pueden acumular y organizar la información de secuencias en bases de datos, sino que también pueden analizar los datos de las secuencias muy rápidamente. La evolución del poder computacional y la capacidad de almacenamiento ha logrado lidiar con la creciente cantidad de información de secuencias que está siendo creada. Los científicos teóricos han desarrollado sofisticados algoritmos que permiten comparar secuencias mediante teoría de probabilidades. Estas comparaciones se han convertido en la base de la determinación de la función de genes, desarrollando relaciones filogenéticas y simulando modelos de proteínas.

La colección, organización e indexado de la información de secuencias en una base de datos es una tarea desafiante por sí misma y ha generado una gran cantidad de información pero de uso limitado. El poder de una base de datos no proviene de la colección de información que tenga, sino de su análisis. Una secuencia de ADN no necesariamente constituye un gen, puede constituir solamente un fragmento de un gen o contener varios genes.

La investigación científica actual, de acuerdo con los principios de la evolución, muestra que todos los genes tienen elementos comunes. Para muchos elementos genéticos es posible construir secuencias consenso, las cuales representan de la mejor manera posible la norma de una clase dada de organismo. Algunos elementos genéticos comunes incluyen promotores, reforzadores, señales de poliadenización y sitios de binding de proteínas. Para estos elementos también se conocen algunas características de sus subelementos. Los elementos genéticos comunes comparten secuencias similares, siendo éste el hecho que permite la aplicación de algoritmos al análisis de secuencias biológicas.

1.6 Introduction to Microarray Technology

Advances in molecular biology and new computational techniques are enabling us to systematically investigate the complex molecular process underlying biological systems (Durbin 1998. To take full advantage of the large and rapidly increasing body of sequence information, new technologies are required. Among the most powerful and versatile tools for genomics are high-density arrays of oligonucleotides or complementary DNAs. Also known as microarrays, they have revolutionized modern biological research by its capacity of monitoring the expression level of thousands of genes simultaneously {Brown, 1999 #561), while traditional methods could only handle the one-gene at a time approach.

A microarray works by exploiting the ability of a given mRNA molecule to bind specifically to, or hybridize to, the DNA template from which it originated. By using an array containing many DNA samples, scientists can determine, in a single experiment, the expression levels of hundreds or thousands of genes within a cell by measuring the amount of mRNA bound to each site on the array. With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.

Microarrays are therefore useful for rapid surveying large number of genes or when the sample to be studied is small. Microarrays may be used to assay gene expression within a single sample or to compare gene expression in two different cell types or tissue samples, such as in healthy and diseased tissue.

DNA Microarrays are small, solid supports onto which the sequences from thousands of different genes are immobilized, or attached, at fixed locations. The supports themselves are usually glass microscope slides, the size of two side-by-side pinky fingers, but can also be silicon chips or nylon membranes. The DNA is printed, spotted, or actually synthesized directly onto the support.

The whole microarray experiment process is based on hybridization probing, a technique that uses fluorescently labeled nucleic acid molecules as ``mobile probes'' to identify complementary molecules, sequences that are able to base-pair with one another. Each single-stranded DNA fragment is made up of four different nucleotides, adenine (A), thymine (T), guanine (G), and cytosine (C), that are linked end to end. Adenine is the complement of, or will always pair with, thymine, and guanine is the complement of cytosine. Therefore, the complementary sequence to G-T-C-C-T-A will be C-A-G-G-A-T. When two complementary sequences find each other, such as the immobilized target DNA and the mobile probe DNA, cDNA, or mRNA, they will lock together, or hybridize.

Two main types of DNA chips can be discerned, either oligonucleotides or complementary DNAs (cDNA). Both are based on the same principle, however the method of addition of the nucleotide stretches to the chip differs. We now briefly describe each of the technologies.

1.6.1 Spotted Arrays

In spotted microarrays (or two-channel or two-colour microarrays), the probes are cDNA or small fragments of PCR products that correspond to mRNAs (messenger RNA) and are spotted onto the microarray surface. This type of array is typically hybridized with cDNA (complementary DNA) from two samples to be compared (e.g. diseased tissue versus healthy tissue) that are labeled with two different fluorophores (e.g. Rhodamine (Cyanine 5, red) and Fluorescein (Cyanine 3, green)). The two samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores (see Figure 1.12). Relative intensities of each fluorophore are then used to identify up-regulated and down-regulated genes in ratio-based analysis. Absolute levels of gene expression cannot be determined in the two-colour array, but relative differences in expression among different spots (=genes) can be estimated with some oligonucleotide arrays. Examples of providers for such microarrays includes Agilent with their Dual-Mode platform, Eppendorf (company) with their DualChip platform and ArrayIt.

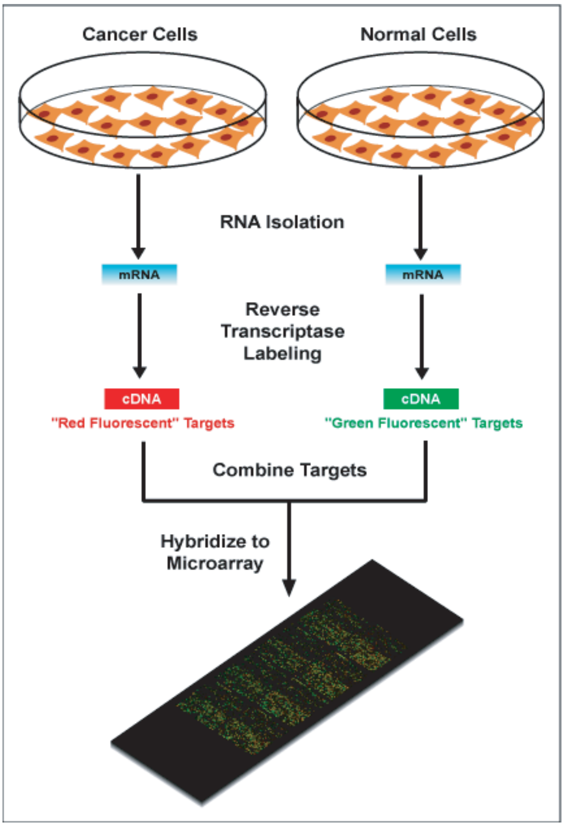


Figure 1.12 Diagram of typical dual color microarray experiment.

1.6.2 Oligonucleotide Arrays

In oligonucleotide microarrays (or single-channel microarrays), the probes are designed to match parts of the sequence of known or predicted mRNAs. There are commercially available designs that cover complete genomes from companies such as GE Healthcare, Affymetrix, Ocimum Biosolutions, or Agilent. These microarrays give estimations of the absolute value of gene expression and therefore the comparison of two conditions requires the use of two separate microarrays (see Figure 1.13).

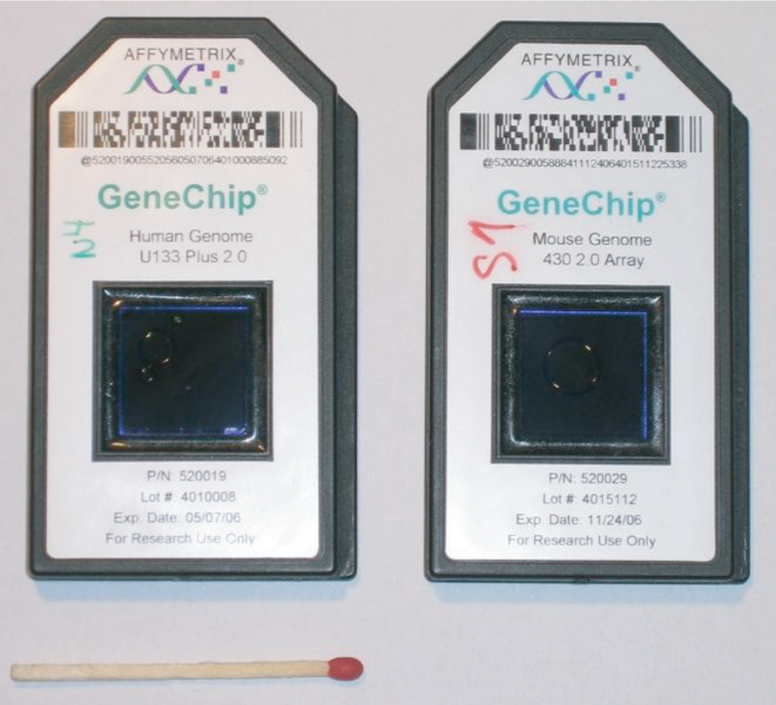


Figure 1.13 Affymetrix Chips

In this type of arrays, the oligonucleotides are synthesised directly onto the chip. The solid surface is prepared such, that there are 3`-OH ends sticking out, to which nucleic acids can be attached in sequence. These are the arrays used for the main experiment related to this work, in particular the Affymetrix GeneChip HG133A, comprised of more than 22,000 probe sets and 500,000 distinct oligonucleotide features including 14,500 well characterized human genes. Probe sets are the ``basic unit'' that Affymetrix uses for its array (see Figure 1.14). Each *probe set* is made up of a number of *probe pairs*, between 11 and 20. Each of this probe pairs is made of two positions, a ``Perfect Match'' (PM) and a ``Miss Match'' (MM) which are complementary. The PM is made out of 25 oligonucleotides, designed to be perfectly compatible with the RNA sequence. The MM is made out of 25 oligonucleotides compatible with the RNA sequence it hybridizes, except in its central position, number 13, which serves as a control for the specific hybridation, since MM hybridation should always be less than PM hybridation. Therefore, some probe sets are compatible with intragenic regions, and thus the probe set will be associated to an specific gene, some other are compatible with intergenic region etc. Note than two or more probe sets might correspond to different regions of the same gene.

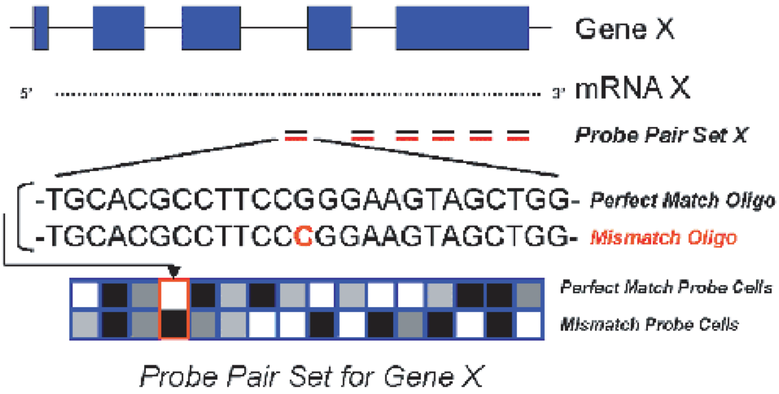

Gene X
5'
3'
mRNA X
Probe Pair Set X
-TGCACGCCTTCCGGGAAGTAGCTGG-
Perfect Match Oligo
-TGCACGCCTTCCCGGAAGTAGCTGG-
Mismatch Oligo
Perfect Match Probe Cells
Mismatch Probe Cells
Probe Pair Set for Gene X

Figure 1.14 Probe Set structure in Affymetrix GeneChips®

Affymetrix provides for each chip several files associated to it in order to process the information acquired. The raw image data (see Figure 1.15) from chip scanner is saved in .DAT file. The information about the expression leves of individual probe sets is extracted from the image data, .DATA file, and stored in a .CEL file. The probe set information in the .CEL file by itself is not particularly useful as there is no indication in the file as to which probe set a probe belongs. This information is stored in the .CDF library file associated with a GeneChip type. The Affymetrix probe set IDs are not particularly descriptive (e.g. 200008_s_at, 200015_x_at or 200035_at). The mapping between the IDs and the gene names is stored in the .GIN file. Affymetrix also provides for each particular GeneChip an annotation file, for use by any interested party to understand what biological entities are represented on Affymetrix arrays. In such files probe sets are related to its sequence source, UniGene ID, Gene title and symbol, RefSeq protein ID, SwissProt entry, Gene Ontology Association, Pathway information and much more information. The .CHP file contains the results of the experiment. These include the average signal measures for each probe set as determined by the Affymetrix software and information about which probe sets are called as present, absent or marginal and the *p*-values for these calls.

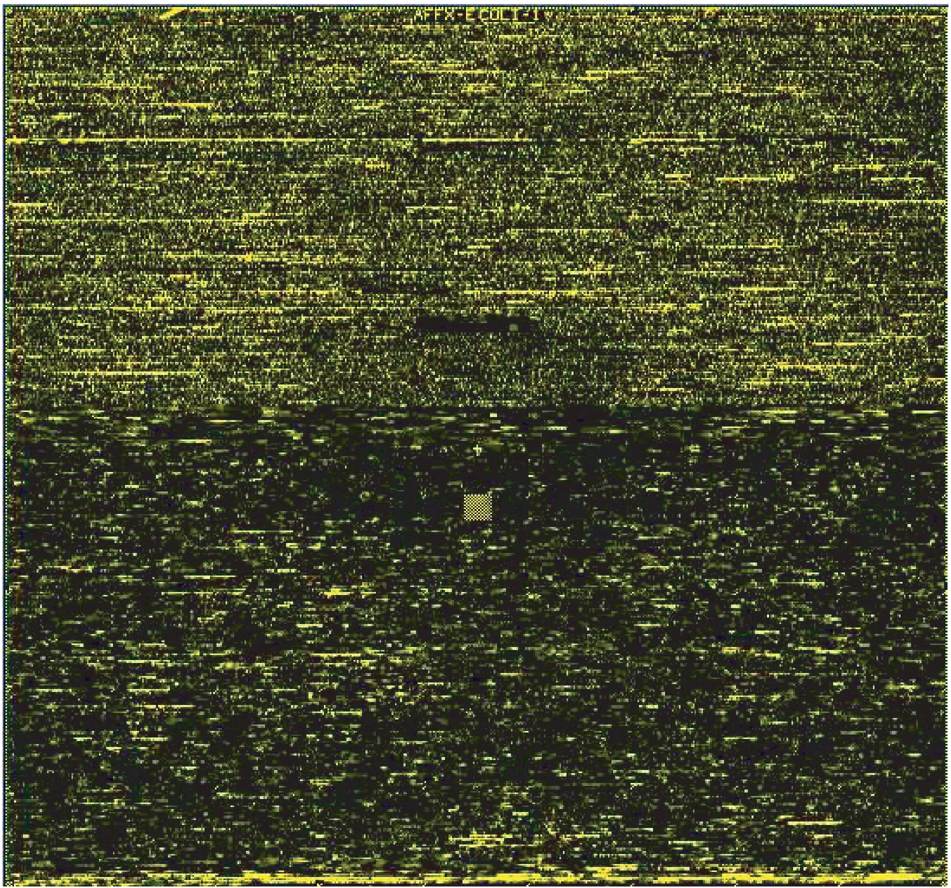


Figure 1.15 Scanned image of an Affymetrix array.

1.6.3 Microarray Scope of Application

One of the most important applications for arrays so far is the monitoring of gene expression (mRNA abundance). The collection of genes that are expressed or transcribed from genomic DNA, sometimes referred to as the expression profile or the ``transcriptome'', is a major determinant of cellular phenotype and function. The transcription of genomic DNA to produce mRNA (messenger RNA) is the first step in the process of protein synthesis, and differences in gene expression are responsible for both morphological and phenotypic differences as well as indicative of cellular responses to environmental stimuli and perturbations. Unlike the genome, the transcriptome is highly dynamic and changes rapidly and dramatically in response to perturbations or even during normal cellular events (Lockhart and Winzeler 2000) such as DNA replication and cell division (Cho, Campbell et al.). In terms of understanding the function of genes, knowing *when, how and for how long a gene is turned on/off* is central to understanding the activity and biological roles of its encoded protein. In addition, changes in the multi-gene patterns of expression can provide clues about regulatory mechanisms and broader cellular functions and biochemical pathways. In the context of human health and treatment, the knowledge gained from these types of measurements can help determine the causes and consequences of disease, how drugs and drug candidates work in cells and organisms, and what gene products might have therapeutic uses themselves or may be appropriate targets for therapeutic intervention.

Gene expression profiles as ``fingerprints". An often overlooked aspect of measurements of global gene expression is that the sequence or even the origin of the arrayed probes does not need to be known to make interesting observations - the complex profiles, consisting of thousands of individual observations, can serve as transcriptional ``fingerprints`". These fingerprints are extremely interesting to be known. They can be used, for instance, for classification purposes or as tests for relatedness, in a similar manner to the way in which DNA fingerprints are used in paternity testing. Many papers have been published, where these ``fingerprints" are used as classification features for different phenotypes, specially in cancer classification: (Alizadeh, Eisen et al. 200; Ben-Dor, Bruhn et al. 2000).

Transcriptional fingerprints have been also used to determine the target of an specific drug (Marton, Bennett et al. 1998). The basic idea is that if a drug interacts with and inactivates a specific cellular protein, the phenotype of the drug-treated cell should be very similar to the phenotype of a cell in which the gene encoding the protein has been genetically inactivated, usually through mutation. Thus, by comparing the expression profile of a drug-treated cell to the profiles of cells in which single genes have been individually inactivated, specific mutants can be matched to specific drugs, and therefore, targets to drugs (Lockhart and Winzeler 2000).

Finally, expression profiles can be used to classify drugs and their mode of action. For example, the functional similarity and specificity of different purine analogues have been determined by comparing the genome-wide effects on treated yeast, murine and human cells (Rosania, Chang et al. 2000).

Chapter 2
Enfoque Computacional

En este capítulo se introducirán los conceptos básicos sobre diferentes métodos y algoritmos computacionales que constituirán la base sobre la cual se apoyará nuestro trabajo. Los temas que se explicarán en las próximas secciones corresponden a los conceptos de *modelado de sistemas*, *lógica difusa*, *algoritmos de clustering (agrupamiento)*, *técnicas de optimización multiobjetivo*, *,métodos de inferencia*, *algoritmos evolutivos*, en particular *algoritmos genéticos* y finalmente presentaremos la metotología marco con que abordamos los problemas a resolver.

2.1 Modelado de sistemas

Cada vez que resolvemos un problema correspondiente al mundo real, debemos comprender que, en realidad, solamente estamos encontrando la solución de un modelo del problema (Michalewicz and Fogel 2000). Todos los modelos son simplificaciones del mundo real, de otro modo serían tan complejos e intrincados como lo es el escenario natural. El problema que existe con los modelos es que cada uno de ellos debe hacer una serie de asunciones respecto del problema real, en otras palabras, dejar algo de éste a un lado. En consecuencia, el error más frecuente que existe cuando tratamos con modelos es olvidar dichas asunciones. En este sentido, un problema puede tener varias soluciones correctas posibles, tan solo dependientes de cómo fue construido el modelo del mismo.

El proceso de resolución de un problema consiste básicamente en dos pasos generales que se dan por separado:

- Creación de un modelo del problema.

- Utilización de ese modelo para la generación de una solución:

$$\text{Problema} \Rightarrow \text{Modelo} \Rightarrow \text{Solución}$$

Tal como hemos dicho, la solución de un problema está dada solamente en términos del modelo y depende de su grado de fidelidad. Esto es, confiamos en que el mejor modelo produzca la mejor solución. Sin embargo, algunas veces los modelos más perfectos resultan inútiles para que un método de solución preciso decida qué hacer, es decir, permita derivar una solución basada en dicho modelo. En esta memoria, trataremos con modelos *(Modelos$_a$)*y soluciones (*Soluciones$_a$*) y aproximadas, con el propósito de representar y resolver de la mejor manera posible cierta clase de problemas biológicos para los cuales, debido a su nivel de complejidad e imprecisión, resulta difícil obtener una solución adecuada.

En este sentido, la construcción de modelos basados en la lógica difusa (ver Sección 2.1.1) ha adquirido un gran interés gracias a que proporciona mayor generalidad, poder expresivo y tolerancia frente a la imprecisión.

El desarrollo de modelos matemáticos de sistemas reales es un tópico central en diferentes disciplinas como la ingeniería y las ciencias. Como hemos mencionado, dichos modelos sirven para resolver problemas reales a través de simulaciones, análisis del comportamiento de un sistema, diseño de nuevos procesos para controlar sistemas, predicciones, etc. El desarrollo inadecuado de dichos modelos puede conducir a la obtención de resultados no exitosos y erróneos de los sistemas en cuestión.

Gran parte de estos sistemas actuales comparten una serie de características que dificultan su modelado con técnicas tradicionales, tales como: necesidad de una importante precisión, tratamiento y respuestas en tiempo real, carencia de conocimiento formal sobre su funcionamiento, posesión de un comportamiento fuertemente no lineal, con alto grado de incertidumbre y características con variaciones en el tiempo. Ejemplos de esta clase de problemas son los procesos complejos de los sistemas de ingeniería aeroespacial o bioquímica, aunque también se encuentran en sistemas ecológicos, sociales o correspondientes al área económico-financiera.

A continuación plantearemos al menos tres paradigmas diferentes en cuanto a la tarea de modelado, es decir, de la comprensión de la naturaleza y el comportamiento de un sistema, y la consecuente tarea de creación de un modelo que pueda ser posteriormente utilizado (Babuska 1998):

- *Modelado de caja blanca.* Tradicionalmente, el modelado de datos se ha visto como una conjunción de una actividad que permite la comprensión de la naturaleza del sistema y su comportamiento, y de un tratamiento matemático adecuado que conduce hacia la realización de un modelo utilizable. Este enfoque, denominado *modelado de ``caja blanca''*, está limitado en la práctica cuando se emplean sistemas complejos y poco comprensibles. Como consecuencia directa de este planteamiento, los sistemas no lineales se simplifican, pasando a ser tratados como sistemas lineales, debido a la imposibilidad de comprender su mecanismo de funcionamiento. En vista del diagrama previamente empleado, observamos esta situación de modelado como:

$$Problema \Rightarrow Modelo_{\alpha} \Rightarrow Solución_{p}$$

lo que corresponde a solucionar un modelo simplificado (α) de un problema por medio de un método (p) que produce soluciones óptimas y precisas.

- *Modelado de caja negra. Este es un planteamiento diferente que consiste en estimar los parámetros del modelo en base a los datos de entrada disponibles. A pesar de que estos modelos generalmente pueden ser desarrollados de forma bastante fácil, la estructura identificada no posee un significado físico claro. Estos modelos resultan difícilmente escalables y el comportamiento real del sistema puede ser solamente analizado a través de la simulación numérica. Un inconveniente común de la mayoría de estas técnicas de modelado es que no permiten el uso efectivo de información adicional, tal como conocimiento y experiencia de expertos, y operaciones previas, las cuales son imprecisas y cualitativas por naturaleza. Nuevamente, esto puede ser visto como:*

$$Problema \Rightarrow Modelo_{p} \Rightarrow Solución_{\alpha}$$

Es decir, un modelo muy preciso solucionado por un método cercano al óptimo, por tanto, una solución aproximada.

- *Modelado de caja gris*. Los hechos descritos anteriormente motivan el desarrollo de técnicas de modelado híbrido o de ``caja gris", que combinan las ventajas de los enfoques de tipo caja blanca, utilizando conocimiento físico para modelar las partes conocidas del sistema, y de los de caja negra, para aproximar las partes más inciertas. Para realizar esto, se han introducido algunas metodologías inteligentes, las cuales emplean técnicas motivadas por sistemas biológicos e inteligencia humana (lenguaje natural, reglas, redes semánticas o modelos cualitativos) para desarrollar modelos y controladores para sistemas dinámicos.

En consecuencia, obtenemos una solución aproximada de un problema en base a un modelo aproximado, con el propósito de ser interpretable y tan preciso como se pueda. Podemos distinguir así varias clases de modelos aproximados, esto es, con distinto grado de precisión (β):

$$Problema \Rightarrow Modelo_{\beta} \Rightarrow Solución_{\alpha}$$

Debemos señalar que un modelo aproximado de un problema no necesariamente es una simplificación del mismo, sino que por el contrario puede ser una descripción aún más precisa que la dada por un modelo de caja blanca. Esto se debe a que estos modelos son capaces de captar la imprecisión e incertidumbre de los problemas del mundo real, que muchas veces son asunciones o excepciones no tenidas en cuenta en el otro caso.

2.1.1 Lógica y conjuntos difusos

La lógica difusa es una teoría de reciente aparición que se atribuye al investigador Lofti Zadeh. En lo que se considera como el trabajo seminal de la teoría de conjuntos difusos (Zadeh 1965) y su posterior aportación sobre el concepto de variable lingüística (Zadeh 1975), este investigador proporcionó las bases para el entendimiento y tratamiento de la incertidumbre de forma cualitativa o mediante términos lingüísticos.

Con carácter general, la representación del conocimiento ha sido una de las áreas de mayor interés investigadas en la disciplina de las ciencias de la computación y la inteligencia artificial. Uno de los principales aspectos tratados es el de cómo representar el conocimiento que es lingüísticamente impreciso, para cuya representación se han mostrado ineficaces las técnicas convencionales. Debemos ser conscientes de que los métodos que tradicionalmente se utilizaban para tratar la información hacían uso de datos precisos (cuantitativos), intentando aportar una visión predictiva basada en procesos deterministas. Sin embargo, no resultan útiles cuando se aplican como apoyo de procesos de razonamiento que cuentan con información incierta o definida de manera imprecisa. En este sentido, el desarrollo de la lógica difusa se vio motivado por la necesidad de un marco conceptual que pudiese aplicarse con éxito al tratamiento de la información en entornos de incertidumbre e imprecisión léxica(Zadeh 1992).

La lógica difusa pude ser vista como una extensión a la lógica clásica, donde se incorporan nuevos conceptos para trabajar con el problema de representación en un ambiente de incertidumbre e imprecisión. La lógica difusa, como su nombre sugiere, es una forma de lógica cuya forma de razonamiento subyacente es más aproximada que exacta. La diferencia fundamental entre las proposiciones de la lógica clásica y las proposiciones difusas está en el rango de valores de verdad. Mientras que en las proposiciones clásicas sólo existen dos posibles valores de verdad (verdadero o falso), el grado de verdad o falsedad de las proposiciones difusas puede tomar distintos valores numéricos. Asumiendo que la verdad y la falsedad se representan con 1 y 0 respectivamente, el grado de verdad de cada proposición difusa se expresa como un valor en el intervalo [0,1]. La lógica difusa es, en realidad, una forma de lógica multivaluada. Su finalidad última es proveer de una base para el *razonamiento aproximado* con proposiciones imprecisas utilizando la teoría de conjuntos difusos como herramienta principal.

El paso de la lógica clásica a la difusa que acabamos de comentar tiene serias implicaciones, como no puede ser de otra manera, sobre la teoría de conjuntos. Si un conjunto se utiliza para clasificar los elementos de un universo de estudio, en determinadas situaciones (aquellas no deterministas o definidas de manera vaga) no se deben obtener los mismos resultados si se utilizan los principios de la lógica clásica a si se utilizan los de la lógica difusa.

De forma previa a la introducción de los conjuntos difusos, partamos de lo que se entiende por un conjunto clásico, también denominados como conjuntos *crisp*, mediante la siguiente definición (Pedrycz, Bonissone et al. 1998): un conjunto clásico A del universo de discurso o dominio X, haciendo uso de la siguiente función característica $\mu_A : X \to [0,1]$ del conjunto A si y solo si para x:

$$\mu_A(x) = \begin{cases} 1 & si \quad x \in A \\ 0 & si \quad x \notin A \end{cases}$$

donde X es el universo de discurso y A un conjunto definido en dicho discurso. Como se puede observar, la función característica que define los conjuntos clásicos es un caso típico de función booleana, de verdadero o falso, expresado numéricamente como 1 ó 0. También se denomina como función discriminante porque discrimina los elementos del universo de discurso entre aquellos que pertenecen al conjunto definido y aquellos que no. Por tanto, se tratan de conjuntos de elementos cuyos límites están fijados de forma determinista.

Por tanto, de manera previa a la resolución de la pregunta que planteamos, definamos lo que se entiende por un conjunto difuso (Pedrycz, Bonissone et al. 1998): un conjunto difuso A en el universo de discurs*o* X es un conjunto de pares ordenados de un elemento genérico x y su correspondiente grado de pertenencia $\mu_A(x)$ de manera que $A = \{(x, \mu_A(x)) / x \in X\}$. Por tanto, como puede apreciarse, la definición de un conjunto difuso es similar a la de un conjunto clásico, con la sustancial salvedad de que cada elemento perteneciente al universo de discurso tiene asociado un *grado de pertenencia* a dicho conjunto difuso. Por tanto, los conjuntos crisp se pueden considerar como un caso específico de conjuntos difusos, en tanto que $0,1 \in [0,1]$

Existen siete tipos de funciones de pertenencia típicas: Triangular, Gamma, S, Gaussiana, Trapezoidal, Pseudo-exponencial y de Trapecio extendido. Las funciones más utilizadas son de la clase triangular, trapezoidal y gaussiana. Las dos primeras se basan en un comportamiento lineal de la función de pertenencia, mientras que la gaussiana se caracteriza por un comportamiento no lineal de dicha función.

Es incorrecto asumir que un conjunto difuso indica, de alguna manera, alguna forma de probabilidad. A pesar del hecho que pueden llegar a tomar valores en un mismo intervalo de definición, es importante comprender que los grados de pertenencia *no* son probabilidades. Una diferencia aparentemente inmediata es que la suma de las probabilidades en un conjunto universal debe ser 1, mientras esto no es un requisito para los conjuntos difusos.

Ya hemos destacado que los conjuntos difusos son extensiones o generalizaciones de los conjuntos clásicos o crisp de la lógica bivaluada. Por tanto, las mismas operaciones que se determinan para los conjuntos crisp pueden determinarse igualmente para los conjuntos difusos, con el añadido de que existen unos grados de pertenencia asociados. Asimismo, los conjuntos resultantes de las operaciones con conjuntos difusos son también difusos. Las principales operaciones básicas que se van a presentar son: la igualdad, la inclusión, la intersección o conjunción, la unión, y el complemento.

- *Igualdad.* Para que se considere que dos conjuntos difusos son iguales, se deberá satisfacer la condición siguiente:

$$A = B \Leftrightarrow \forall x \in X : \mu_A(x) = \mu_B(x)$$

- *Inclusión.* Se considera que un conjunto difuso *A* en un subconjunto de *B* si:

$$A \subseteq B \Leftrightarrow \forall x \in X : \mu_A(x) \leq \mu_B(x)$$

- *Intersección.* La intersección entre dos conjuntos difusos se puede obtener mediante el mínimo de ambos:

$$\mu_A(x) \cap \mu_B(x) = \min\{\mu_A(x), \mu_B(x)\}$$

- *Unión.* Por el contrario, para la unión de dos conjuntos difusos se considera el máximo en los grados de pertenencia de los elementos del universo de discurso:

$$\mu_A(x) \bigcup \mu_B(x) = \max\{\mu_A(x), \mu_B(x)\}$$

- *Complemento.* El complemento, $\overline{A}$, de un conjunto difuso *A* con respecto al conjunto universal *X* se define para todo elemento $x \in X$ como:

$$\overline{A} = 1 - \mu_A(x)$$

Por otro lado, existen generalizaciones de las operaciones anteriores, puesto que tanto las funciones de pertenencia de los conjuntos difusos como sus operaciones dependen del contexto en el que se apliquen. En este respecto, para poder aplicar la lógica difusa en un sistema informático basado en reglas es preciso que se pueda trabajar con los operadores ``Y'' y ``O'', es decir. la intersección y la unión respectivamente. La familia de funciones que se utilizan para tal fin se conocen como *T-normas* y *T-conormas*.

Una *T-norma* generaliza el concepto de intersección de forma que

$$T : [01]x[0,1] \rightarrow [0,1]$$
$$\mu_A \bigcap \mu_B = T[\mu_A(x), \mu_A(x)]$$

y además satisface las siguientes propiedades:

- Conmutativa: *T(a; b) = T(b; a)*

- Asociativa: *T(a; T(b; c)) = T(T(a; b); c)*

- Monotónica: *T(a; b) ≥ T(c; d); si a ≥ c y b ≥ d*

- Condiciones frontera: *T(a; 1) = a*

En segundo lugar, la *T-conorma*, también conocida como S-norma, generaliza el concepto de unión de forma que

$$S:[01]x[0,1]\rightarrow[0,1]$$
$$\mu_A \cup \mu_B = S[\mu_A(x),\mu_A(x)]$$

y además satisface las siguientes propiedades:

- Conmutativa: *S(a; b) = S(b; a)*

- Asociativa: *S(a; S(b; c)) = S(S(a; b); c)*

- Monotónica: *S(a; b) ≥ S(c; d); si a ≥ c y b ≥ d*

- Condiciones frontera: *S(a; 0) = a*

En resumen, la utilización de funciones triangulares para la intersección y/o unión de conjuntos difusos ofrece un abanico bastante extenso para realizar estas operaciones, y, por tanto, para calcular las *T-normas* y *T-conormas* en las que se basarán las funciones de pertenencia resultantes. Asimismo, su flexibilidad permite que el investigador proponga sus propias fórmulas con el objeto de adaptarse mejor a las características de los conjuntos difusos con los que trabaje.

2.2 Clustering

El objetivo del agrupamiento de datos (*clustering*) es la clasificación de objetos de acuerdo a similitudes entre ellos, para luego organizar estos datos en grupos. Las técnicas de clustering están incluidas entre los métodos de aprendizaje *no supervisado*, debido a que no utilizan conocimiento de identificadores de clases. La mayoría de los algoritmos de clustering tampoco se basan en asunciones comunes a los métodos estadísticos convencionales, tales como la distribución estadística subyacente de los datos, y por ello son útiles en situaciones donde existe poco conocimiento sobre los mismos. El potencial de los algoritmos de clustering para revelar las estructuras subyacentes de los datos puede ser explotado no sólo para la clasificación y reconocimiento de patrones sino también para la reducción de la complejidad en modelado y optimización.

Las técnicas de clustering pueden aplicarse a datos que sean cuantitativos (numéricos), cualitativos (categóricos) o una mezcla de ambos. Los datos son típicamente observaciones de algún proceso físico.

Pueden formularse varias definiciones de cluster, dependiendo del número de objetivos del clustering. Generalmente, uno puede aceptar la visión de que un cluster es un grupo de objetos, los cuales son más similares entre sí que los miembros de otros clusters. El término ``similaridad'' debería ser entendido como la similaridad matemática, medida formalmente. En espacios métricos, la similaridad está definida comúnmente como una *norma* de distancia. La distancia puede medirse entre los vectores de datos en sí, o como la distancia de un vector de datos a algún objeto prototípico del cluster. Los prototipos no son usualmente conocidos de antemano y los algoritmos de clustering los buscan simultáneamente con la partición de los datos. Los prototipos pueden ser vectores de igual dimensión que la de los objetos de datos, pero se pueden también definir como objetos geométricos de ``alto nivel", tales como subespacios lineales y no-lineales, o funciones.

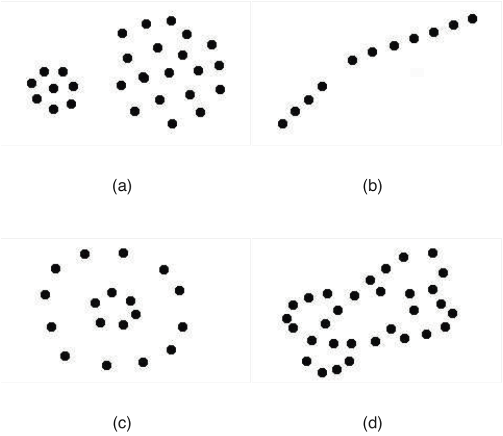


Figure 2.1 Diferentes formas de los clusters

Los datos pueden revelar clusters de diferentes formas geométricas, tamaños y densidades, como se puede ver en la Figure 2.1. Mientras que los clusters (a) son esféricos, los clusters (b), (c) y (d) pueden caracterizarse como subespacios lineales y no-lineales del espacio de datos. El rendimiento de la mayoría de los algoritmos de clustering está influenciado no solo por la forma geométrica y densidad de los clusters individuales, sino también por las relaciones espaciales y distancias entre ellos. Los clusters pueden estar bien separados, conectados en forma continua, o solapados entre ellos. La separación de los clusters está influenciada por el factor de escala y la normalización de los datos.

Se han propuesto muchos algoritmos de clustering en la literatura. Dado que los clusters pueden verse formalmente como subconjuntos del conjunto de datos, una posible clasificación de los métodos de clustering puede ser de acuerdo a si los subconjuntos son difusos o crisp. Los métodos de clustering crisp están basados en la teoría clásica de conjuntos y requieren que un objeto pertenezca o no a un cluster dado. Un clustering crisp significa particionar los datos en un número especificado de subconjuntos mutuamente excluyentes. Los métodos de clustering difuso, sin embargo, permiten a los objetos pertenecer a varios clusters simultáneamente, con distinto grado de pertenencia. En muchas situaciones, los clusters difusos son un concepto más natural que los clusters crisp, dado que los objetos en las fronteras de algunas clases no están forzados a pertenecer completamente a una de ellas. En lugar de ello, se les asignan grados de pertenencia entre 0 y 1 indicando su pertenencia parcial.

A continuación se mostrarán los metodos de clustering utilizados en este manuscrito.

2.2.1 Clustering Jerárquico

Este método crea una jerarquía de clusters, usualmente representada por un dendrograma (Figure 2.2), a partir de observaciones individuales mediante la progresiva agregación de sub-jerarquías. Dada norma de similaridad (e.g. distancia euclídea) selecciona los dos objetos (i.e. observaciones suministradas o clusters) mas cercanos, los cuales son agrupados. Este nuevo cluster reemplaza a los objetos seleccionados y el algoritmo continúa obtener un único grupo. Existen diferentes opciones para medir la distancia entre dos objetos (*linkage*). Como ejemplo podemos describir los siguientes:

- *Single likage*: es también llamada vecino más cercano ya que utiliza la menor distancia entre objetos de los dos clusters

$$d(r,s) = \min(dist(x_{ri},x_{sj})),\ i \in (1,...,n_r),\quad j \in (1,...,n_s)$$

- *Complete linkage*: es también llamada vecino mas lejano ya que utiliza la mayor distancia entre objetos de los dos clusters

$$d(r,s) = \max(dist(x_{ri},x_{sj})),\ i \in (1,...,n_r),\quad j \in (1,...,n_s)$$

- *Average linkage*: utiliza la media de las distancias entre todos los pares de objetos en el cluster *r* y el cluster *s*

$$d(r,s) = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} dist(x_{ri},x_{sj})$$

- *Centroid linkage*: evalúa la distancia euclídea entre los centroides de los dos clusters, donde

$$d(r,s) = \left\| \overline{x_r} - \overline{x_s} \right\| \qquad \overline{x_r} = \frac{1}{n_r}\sum_{i=1}^{n_r} x_{ri} \qquad \overline{x_s} = \frac{1}{n_s}\sum_{i=1}^{n_s} x_{si}$$

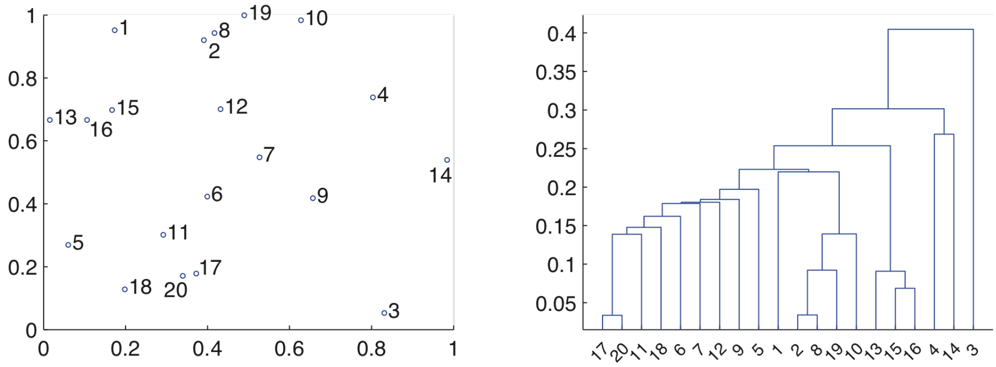


Figure 2.2 Clustering jerarquico aplicado aun conjunto de 20 puntos generados al azar. En el panel de la izquierda graficamos y etiquetamos los valores de 20 puntos pertenecientes a R^2 . En el panel de la derecha graficamos el dendrograma producto de aplicar el clustering jerarquico, utilizando la distancia euclidea y el *single linkag*e. En el eje de las X se detalla la etiqueta de cada punto, y en el eje de las Y las distancias entre ellos.

2.2.2 Subtractive Clustering

El método de clustering substractivo asume que cada observación es un centroide potencial de un cluster por lo que calcula una medida de verosimilitud de que cada observación pueda definir el centroide, basado en la densidad de las observaciones circundantes. El algoritmo realiza los siguientes pasos:

- Selecciona la observación con mayor potencialidad para definir el primer centroide
- Remueve todas las observaciones próximas a una vecindad de la observación seleccionada. La vecindad esta determinada mediante un radio suministrado como parámetro.
- Itera los pasos anteriores hasta que todas las observaciones pertenecen a un cluster.

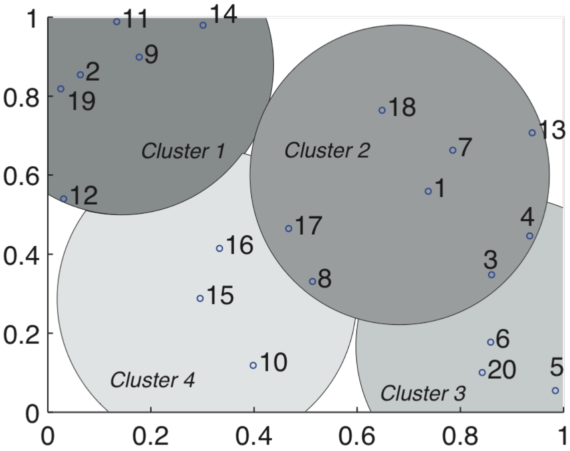


Figure 2.3 Subtractive clustering aplicado aun conjunto de 20 puntos generados al azar. Graficamos y etiquetamos los valores de 20 puntos pertenecientes a R^2 e indicamos los 4 clusters generados con un radio=0.5.

2.2.3 Clustering difuso *c*-medias (Fuzzy *c*-means)

Este método de clustering es una extensión del método de clustering tradicional k-medias, donde los elementos pueden pertenecer a más de un cluster con distinto grado de pertenencia. Por ejemplo, el grado de pertenencia de una instancia *k* con un valor x_k a un cluster en particular v_i se calcula como:

$$\mu_{i,k} = \left[1 + \left(\frac{\left\| x_k - V_i \right\|_A^2}{w_i} \right)^{\frac{1}{m-1}} \right]^{-1} \quad \forall i,k;1<m$$

donde las *c*-particiones de los datos *X* se suelen almacenar en una matriz dimension *cxn q*ue contiene el vector donde se representan los grados de similaridad entre las *n* instancias y las *c*-particiones de un tipo de característica $\mu_{i,k}$ corresponde al grado de pertenencia del valor x_k en la partición difusa *i*-ésima de *X*; *m* representa el grado de imprecisión (*fuzzification degree*); *A* determina el tipo de norma utilizada, como puede ser la norma euclidiana (*A*=2); y w_i es un peso para penalización de los términos, el cual se sustituye por 1 en ausencia de información externa.

Si el enfoque es probabilístico, $u_{i,k}$ generalmente corresponde a la probabilidad a posteriori $p(i \mid x_k)$ de que, dado x_k, provenga de la clase *i* siguiendo la regla de Bayes (Bezdek, Pal et al. 1992; Mitchell 1997; Bezdek 1998). Si el enfoque es difuso, x_k puede provenir de más de una clase.

El centroide del cluster de la partición se calcula como:

$$V_i = \frac{\sum_{k=1}^{n} u_{i,k}k^{m} x_k}{\sum_{k=1}^{n} u_{i,k}k^{m}} \quad \forall i$$

fórmula basada en el uso de la distancia Euclídea como función de similaridad:

$$\|x - V\|_2 = \sqrt{(x - V)^T (x - V)}$$

En resumen, los pasos a seguir se muestran en el Algoritmo:

```
Clustering difuso $c$-medias
C-MEDIAS(T número máximo de iteraciones, $c$ número de particiones)
1: Inicializar $V_o = \{v_1, ..., v_c\}$
2: mientras t < T y $\|V_t - V_{t-1}\| > \varepsilon$ hacer
3:     Calcular $U_t$ en base a $V_{t-1}$
4:     Actualizar $V_t$ en base a $V_{t-1}$ y $U_t$
5: fin mientras
```

2.2.4 Árboles de decisión

Los árboles de decisión son modelos predictivos que permite relacionar observaciones sobre un objeto *S* a una conclusión o variable de clase C. El método construye estructuras en las que las hojas representan clasificaciones y las ramas (u nodos internos) representan conjunciones de distintas características que guían estas clasificaciones (Figure 2.4). La técnica de aprendizaje automatizado que permite inducir árboles de decisión a partir de *S* y las variables de clase *C* se llama aprendizaje de árboles de decisión.

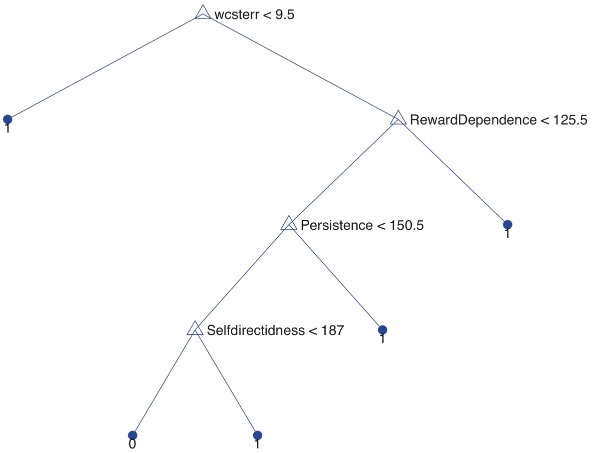


Figure 2.4 Árbol de decisión
El nodo raíz del árbol separa el conjunto de datos de entrenamiento mediante la expresión wcsterr<9.5. La rama izquierda recupera las observaciones que satisfacen esta condición. El nodo destino clasifica a estas observaciones con el valor de clase 1. La rama izquierda recupera las observaciones que no satisfacen esta condición. Los árboles de decisión tienen una estructura recursiva donde los nodos internos (ramas) son adicionados para separar las observaciones, con el fin de obtener nodos hoja que recuperan observaciones con el mismo valor de clase.

Un árbol puede ser "aprendido" al separar el conjunto de observaciones S en dos subconjuntos basados en la expresión de un atributo. Este proceso se repite en cada subconjunto derivado de forma recursiva. La condición de parada para la recursividad esta determinada por la imposibilidad de generar dos nuevos subconjuntos o bien cuando una clasificación singular (mismo valor de clase) es obtenida para cada elemento del conjunto. Por ejemplo, la implementación ID3 usa un política de búsqueda *voraces* (*greedy).* Comienza por el nodo raíz del árbol y aprende que atributo A debe ser utilizado en la condición de test evaluando cada uno de los atributos mediante un test estadístico llamado *information gain* (Mitchell 1997):

$$Gain(S,A) = Entropy(S) - \sum_{v \in Vaues(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(S) = \sum_{v \in Values(A)} -p_v \log_s(p_v)$$

donde *values(A)* es el conjunto de valores posibles para el atributo A y S_v es el subconjunto de S que recupera las observaciones para las cuales A tiene el valor v. El segundo término de la ecuación es la suma de las entropías de cada subconjunto S_v pesado por la fracción de observaciones que recupera. *Gain(S,A)* puede ser interpretada como la reducción esperada de la entropía causada por conocer el valor para el atributo A.

2.2.5 Clases latentes

Los modelos de clases latentes (LC) son utilizados para analizar clusters de datos categóricos. El modelo parte de la suposición que las variables manifiestas son mutualmente independientes, dada la variable de clase.

La variable de clase es oculta (latente) que discrimina a subgrupos de los casos, por lo que los casos dentro de un valor para la clase latente son homogeneos para cierto criterio, mientras que los casos pertenecientes a distintos valores de clase latente son distintos entre si para ese mismo criterio.

Formalmente, un modelo LC incorpora una variable latente *X* a un número de variables manifiestas (observables) Y_1, Y_2, $\ldots Y_n$. Todas las variables (X, Y_1, Y_2, $\ldots Y_n$) son categóricas y las relaciones entre ellas son descriptas por una red Bayesiana (Figure 2.5.). Por lo general, al aplicar este modelo se considera a la variable *X* como la representación de un concepto. Los estados (o valores) de la variable latente corresponden a clases de casos dentro de un universo en estudio. Las variables manifiestas *Yi* representan manifestaciones observables de un concepto latente. El aprendizaje de un modelo LC involucra los siguientes pasos:

1- Determinar la cardinalidad de la variable *X* (i.e. el número de clases latentes)

2- Estimatar los parámetros del modelo *P(X)* y *P(Yi | X)*. Por lo general los parametros son aprendidos usando el algoritmo *Expectation Maximization* (Dempster et al., 1977).

La cardinalidad de *X* es determinada mediante la comparación de alternativas mediante índices o métricas de que tan bien el modelo representa los datos (*goodness-of-fit*). El más comúnmente utilizado es el denominado BIC(Schwarz, 1978). Al igual que el índice *Maximum Description Lenght* (Lanterman, 2001) el índice BIC es una aproximación de la verosimilitud marginal derivada en una configuración en la que todas las variables son observadas (Zhang 2004).

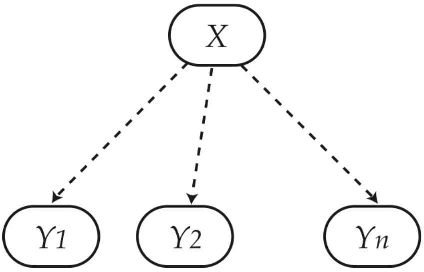


Figure 2.5 Estructura del modelo de clases latentes

2.2.6 Nonnegative matrix factorization

Este método ha sido empleado exitosamente para detectar información latente en relaciones de datos experimentales (Mejia-Roa, Carmona-Saez et al. 2008).

Nonnegaitve matrix factorization (NMF) es similar a otras técnicas estadísticas de minería de datos, como Análisis de componentes principales (PCA), en cuanto el método construye una factorización que aproximada los datos de la forma V~WH, o $V_{i\mu} \sim (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$, donde $V \in \Re^{mxn}$ es una matriz de datos positivos con m variables y n objetos; y k el número de factores. Los métodos difieren en las restricciones que impuestas: MNF solo impone que las matrices V,W y H contentan valores no negativos mientras que las columnas de W y las filas de H debes ser ortogonales al realizar PCA. MNF solo reduce efectivamente la dimensionalidad de la matriz V, sino que también provee un modelo de la información mas interpretable (Mejia-Roa, Carmona-Saez et al. 2008) (En la Figure 2.6 se muestran dos esquemas representativos del módelo)

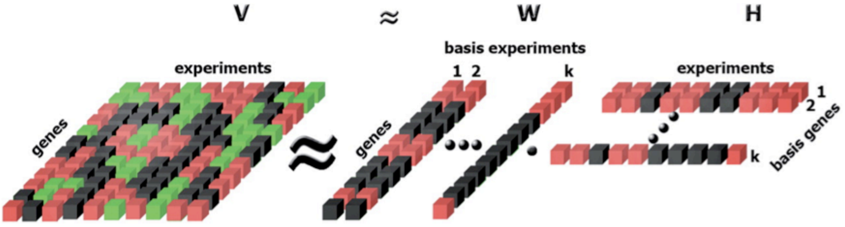
La implementación del algoritmo NMF consiste en iterativamente actualizar las matrices W y H de la siguiente manera:

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \qquad W_{ia} \leftarrow \frac{W_{ia}}{\sum_{j} W_{ja}} \qquad H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$$

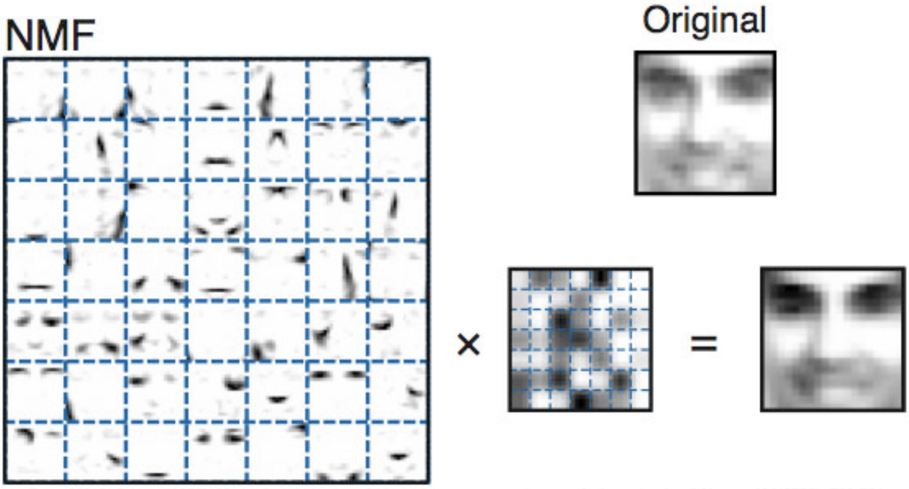
Se ha demostrado que esta iteración hace converger a un máximo local a la función objetivo definida como:

$$F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} \left[V_{i\mu} \log(WH)_{i\mu} - (WH)_{i\mu} \right]$$

sujeta a las restricciones de valores no negativos (Lee and Seung 1999). Esta formula es derivada al interpretar NMF como un algoritmo para la construcción de un modelo probabilístico.



Mejia-Roa, E. et al. Nucl. Acids Res. 2008 36:W523-W528; doi:10.1093/nar/gkn335



Lee, D. and H.-S. Seung (1999). "Learning the parts of objects by non-negative matrix factorization." Nature 401(6755): 788-791.

Figure 2.6 NMF learns parts-based representation of objects.

2.2.7 Clustering conceptual

El agrupamiento conceptual (conceptual clustering) es similar al considerado en el análisis de clusters tradicional, pero está definido de una manera diferente. Dados un conjunto de descripciones en base a atributos de ciertas entidades, un lenguaje de descripción para caracterizar clases de estas entidades y un criterio de calidad de clasificación; el problema consiste en particionar las entidades en clases de tal manera que se maximice el criterio de calidad de clasificación y, simultáneamente, en determinar descripciones generales de estas clases en el lenguaje de descripción dado. Por ello, un método de clustering conceptual busca no sólo una clasificación de las entidades, sino también una descripción simbólica de las clases propuestas (clusters). Un aspecto importante que distingue al clustering conceptual es que, a diferencia del análisis de clusters clásico, las propiedades de las descripciones de clases se toman en consideración en el proceso de determinación de las clases.

Para clarificar la diferencia entre el clustering conceptual y el clustering tradicional, nótese que un método de clustering convencional típicamente determina clusters en base a una medida de similaridad, que es una función definida únicamente utilizando las propiedades (valores de los atributos) de las entidades que están siendo comparadas:

Similaridad(A,B)=f(propiedades(A), propiedades(B))

donde *A* y *B* son entidades a ser comparadas.

En contraste con esta metodología, las técnicas de clustering conceptual agrupan entidades en base a la *cohesión de conceptos*, la cual es una función no sólo de las propiedades (valores de los atributos), sino también de otros dos factores: el *lenguaje de descripción L*, usado por el sistema para describir las clases de entidades, y el *entorno E*, el cual es el conjunto de ejemplos vecinos:

CohesiónConceptual(A,B)=f(propiedades(A),propiedades(B),L,E)

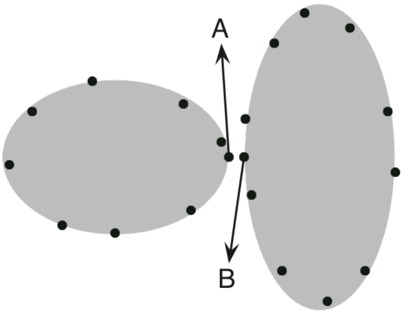


Figure 2.7 Diferencia entre cercanía y cohesión conceptual

Por tanto, dos objetos pueden ser similares, es decir, cercanos de acuerdo a una cierta medida de distancia (o similaridad), pero pueden tener una baja cohesividad conceptual, y viceversa. Un ejemplo de la primera situación se puede ver en la Figure 2.7. Los puntos negros *A* y *B* son ``cercanos'' entre sí y, por ello, serían ubicados en el mismo cluster por cualquier técnica basada únicamente en las distancias entre los puntos. Sin embargo, estos puntos tienen una cohesividad conceptual pequeña debido a que pertenecen a configuraciones que representan diferentes conceptos. Si se dispone de un lenguaje de descripción apropiado, un método de clustering conceptual permitirá agrupar los puntos de la Figure 2.7 en dos ``elipses'', como lo haría la mayoría de las personas.

Un criterio de calidad de clasificación utilizado en clustering conceptual puede involucrar una variedad de factores, tales como el *ajuste* de la descripción de un cluster a los datos, la *simplicidad* de una descripción u otras propiedades de las entidades o conceptos que los describen.

2.2.8 Descubrimiento de subgrupos

La extracción de modelos para la toma de decisiones a partir de conjuntos de datos en un proceso básico en la minería de datos (Mitchell 1997). Los modelos, dependiendo del dominio al cual se deseen aplicar, podrán ser o bien predictivos o bien descriptivos. En los modelos predictivos, el objetivo principal es la capacidad de clasificación del modelo así como su interpretabilidad, mientras que en los descriptivos se busca encontrar relaciones o patrones de comportamiento.

Como hemos visto en la Sección 2.1, existen tres tipos de modelos; de caja blanca, de caja negra y de caja gris, en función de su interpretabilidad.

Una clase de modelos descriptivos son los generados para el descubrimiento de subgrupos, donde se pretende obtener modelos descriptivos empleando mecanismos predictivos. Los modelos descriptivos habitualmente están destinados al aprendizaje de reglas de asociación (Agrawal, Mannila et al. 1996). El aprendizaje de reglas de asociación es un mecanismo de inducción descriptiva que se dedica a descubrir reglas individuales que definen patrones interesantes en los datos.

Por otro lado, se puede definir el descubrimiento de subgrupos como un tipo de modelo descriptivo que se sitúa en la intersección entre la inducción predictiva y descriptiva. Fue formulado inicialmente por Klösgen, con su propuesta de aprendizaje de reglas EXPLORA, y Wrobel con MIDOS (Klosgen 1996){S., 2000 #739. En ellos, el problema del descubrimiento de subgrupos se define de la siguiente forma:

> *``Dada una población de individuos y una propiedad de esos individuos en la que estamos interesados, buscar subgrupos en esa población que sean estadísticamente ``más interesantes'', siendo tan grandes como sea posible y ofreciendo el mayor valor de atipicidad estadística con respecto a la propiedad en la que estamos interesados''.*

En el descubrimiento de subgrupos, las reglas son de la forma *Cond* →Clase, donde la propiedad de interés es el valor de la clase que aparece en el consecuente de la regla. El antecedente de la regla estará compuesto por una conjunción de características (pares atributo-valor) seleccionadas de entre las características que definen las instancias de entrenamiento. Dado que las reglas se han obtenido a partir de prototipos de entrenamiento etiquetados, el proceso de descubrimiento de grupos se centra en encontrar las propiedades de un conjunto determinado de individuos de la población que satisfacen la propiedad de interés dada. El descubrimiento de subgrupos se puede considerar como un mecanismo de inducción descriptiva que persigue encontrar patrones interesantes en los datos. Debido a esta circunstancia, algunas consideraciones estándar llevadas a cabo por los algoritmos de clasificación basados en reglas, tales como el que ``las reglas inducidas deben presentar tanta precisión como sea posible'' o que ``las reglas deben ser tan diferentes como sea posible, para cubrir diferentes porciones de la población'', deben de ser relajadas.

En el descubrimiento de subgrupos, el objetivo es encontrar reglas individuales o patrones de interés, los cuales deben ofrecerse en una representación simbólica adecuada de tal forma que puedan ser utilizados con efectividad por potenciales usuarios de esa información. La interpretabilidad de las reglas es por tanto un factor clave en el descubrimiento de subgrupos.

Ésta es la razón por la que a menudo se considera diferente el descubrimiento de subgrupos de las tareas propias de clasificación. El descubrimiento de subgrupos se centra en encontrar subgrupos de población interesantes en vez de maximizar la precisión del conjunto de reglas inducido.

Para evaluar el éxito en descubrimiento de subgrupos, se estudiarán medidas descriptivas sobre el interés de cada regla obtenida. Las medidas de calidad propuestas consistirán en el valor medio del conjunto de reglas obtenido, lo cual nos permite comparar diferentes algoritmos.

2.3 Optimización multiobjetivo

Muchos problemas reales se caracterizan por la existencia de múltiples medidas de actuación, las cuales deberían ser optimizadas, o al menos ser satisfechas, simultáneamente. Como el propio nombre sugiere, el problema de la optimización multiobjetivo consiste en el proceso de optimización simultánea de más de una función objetivo {Cohon, 1978 #573}.

La falta de metodologías para resolver este tipo de problemas llevó a que, en un principio, se resolvieran como problemas mono-objetivo. Sin embargo, no es correcto tomar esta determinación ya que existen diferencias entre los principios en que se basan los algoritmos que tratan un solo objetivo y los que trabajan con varios. De esta forma, al trabajar con problemas mono-objetivo nos enfrentamos con la búsqueda de una solución que optimice esa única función objetivo, tarea distinta a la que se nos plantea al trabajar con problemas multiobjetivo. En este último caso, no pretendemos encontrar una solución óptima que se corresponda a cada una de las funciones objetivo, sino varias soluciones que satisfagan todos los objetivos a la vez de la mejor manera posible. Como en un problema de optimización con un solo objetivo, también suele existir un número de restricciones que debe satisfacer cualquier solución factible.

La forma general de un problema de optimización multiobjetivo es la siguiente (Deb 2001):

Definición 1 *Un problema de optimización multiobjetivo se define como la maximización/minimización de f:*

$$\begin{array}{ll} f_m(x) & m=1,2,...,M \\ g_j(x)\geq 0 & j=1,2,....,J \\ h_k(x)=0 & k=1,2,....,K \\ x_i^{(L)}\leq x_i\leq x_i^{(U)} & i=1,2,....,N \end{array}$$

donde M corresponde al número de objetivos que tiene el problema, J al número de restricciones de desigualdad, K al número de restricciones de igualdad, y, finalmente, N al número de variables de decisión. Una solución x es un vector de N variables de decisión: $x=(x_1., x_2.,\ldots. x_n.)^T$.

El último conjunto de la ecuación restringe cada variable de decisión a tomar un valor en el intervalo $[x_i^{(L)},x_i^{(U)}]$. Si alguna solución satisface todas las restricciones y los límites de las variables se la conoce como solución factible.

La mayoría de los algoritmos de optimización multiobjetivo usan el concepto de dominancia es su búsqueda del óptimo. A continuación describimos con detalle este concepto (Deb 2001). En los algoritmos de optimización multiobjetivo, la preferencia entre dos soluciones se especifica en función de que una domine a la otra.

Definición 2 *Se dice que una solución x domina a otra solución y ($x \prec y$) cuando se cumplen las siguientes condiciones:*

- *La solución x no es peor que y en todos los objetivos:* $f_i(x)\overline{\triangleright}f_i(y)$ *para todo i=1,2,..,M*
- *La solución x es estrictamente mejor que y en, al menos, un objetivo:* $f_i(x) \triangleleft f_i(y)$ *para al menos i.*

Si alguna de las condiciones anteriores es violada, la solución *x* no domina a la solución *y*. Si *x* domina a la solución *y* también es común escribir que *x* es *no dominada* por *y.*

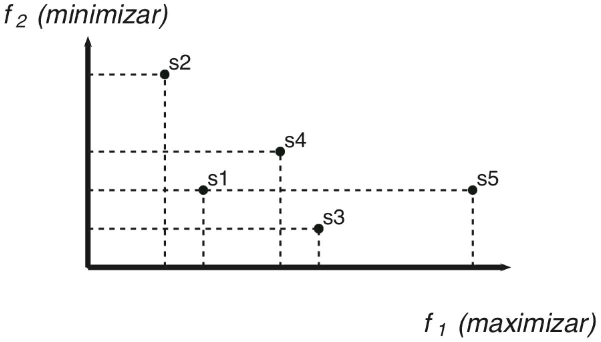


Figure 2.8 Dominancia entre soluciones.

Supongamos el problema con dos funciones objetivo representado en la Figure 2.8: la primera de las funciones f_1 será una función a maximizar, mientras que f_2 será una función a minimizar. Considerando la optimización conjunta de las dos funciones objetivo, es difícil encontrar una solución que sea la mejor respecto a ambas. Utilizando la definición de dominancia podremos decidir qué solución es la mejor de dos soluciones dadas en términos de ambos objetivos. Por ejemplo: si comparamos las soluciones s1 y s2, observamos que la solución s1 es mejor que la s2 en las dos funciones objetivo. Esto supone que se cumplen las dos condiciones de dominancia, luego podremos afirmar que la solución s1 domina a la solución s2.

De forma intuitiva podemos decir que si una solución x domina a otra y, entonces x es mejor que y para la optimización multiobjetivo.

El concepto de dominancia proporciona una forma de comparar soluciones con múltiples objetivos. Como ya hemos comentado, la mayoría de los métodos de optimización multiobjetivo usan este concepto para buscar soluciones no dominadas.

Si, en el ejemplo que mostramos antes, comparamos la solución s3 con la s5, observamos que la solución s5 es mejor que la solución s3 en el objetivo f_1 mientras que es peor en el segundo objetivo, f_2 . Vemos que la primera condición no se cumple para ambas soluciones, lo que simplemente nos dice que no podemos concluir que la solución 5 domine a la solución s3 ni tampoco que la s3 domine a la s5. Cuando esto ocurre, es costumbre decir que las soluciones s3 y s5 son no dominadas una respecto de la otra. Teniendo en cuenta ambos objetivos, no podemos decidir cuál de las dos soluciones es la mejor.

Para un conjunto de soluciones dado, podemos realizar todas las posibles comparaciones de pares de soluciones y encontrar cuáles dominan a cuáles y qué soluciones son no dominadas con respecto a las otras. Al final esperamos conseguir un conjunto de soluciones tales que cualesquiera dos no dominen una a la otra, es decir, un conjunto en el que todas las soluciones son no dominadas entre sí. Este conjunto se conoce como conjunto de soluciones no dominadas.

Este conjunto también tiene otra propiedad: dada cualquiera solución que no pertenezca a él, siempre podremos encontrar una solución del conjunto que la domine. Así, este conjunto particular tiene la propiedad de dominar a todas las soluciones que no pertenecen al mismo. En términos simples, esto significa que las soluciones de este conjunto son mejores comparadas con el resto de soluciones.

Una dificultad común en la optimización multiobjetivo es la aparición de un conflicto entre objetivos (Hans 1988), es decir, el hecho de que ninguna de las soluciones factibles sea óptima simultáneamente para todos los objetivos. En este caso, la solución matemática más adecuada es quedarse con aquellas soluciones que ofrezcan el menor conflicto posible entre objetivos. Estas soluciones pueden verse como puntos en el espacio de búsqueda que están colocados de forma óptima a partir de los óptimos individuales de cada objetivo, aunque puede que dichas soluciones no satisfagan las preferencias del experto que quiera establecer algunas prioridades asociadas a los objetivos.

Para encontrar tales puntos, todas las técnicas clásicas reducen el vector objetivo a un escalar, es decir, a un único objetivo. En estos casos, en realidad, se trabaja con un problema sustituto buscando una solución sujeta a las restricciones especificadas.

A continuación, vamos a repasar tres de las técnicas clásicas más comunes para afrontar problemas con múltiples objetivos. Posteriormente, dedicaremos una sección a analizar los inconvenientes que presentan.

Optimización Mediante Ponderación de los Objetivos. Ésta es probablemente la más simple de todas las técnicas clásicas. En este caso, las funciones objetivo se combinan en una función objetivo global, F , de la siguiente manera (Gass and Saaty 1955):

$$F(x) = \sum_{i=1}^{M} w_i f_i(x)$$

donde los pesos w_i cumplen la condición $\sum w_i = 1$.

En este método, la solución óptima se controla mediante un vector de pesos *w* de forma que la preferencia de un objetivo puede cambiarse modificando dichos pesos. En muchos casos, primero se optimiza cada objetivo individualmente y después se calcula el valor de la función objetivo completa para cada uno de ellos. Así podemos conseguir evaluar la importancia que ejerce cada objetivo y encontrar un vector de pesos adecuado. Después, la solución final aceptada se calcula optimizando *F* según los pesos establecidos.

Las únicas ventajas de usar esta técnica es que se puede potenciar a un objetivo frente a otro y que la solución obtenida es normalmente Pareto-optimal.

Optimización Mediante Funciones de Distancia. Con esta técnica, la reducción a un escalar se lleva a cabo usando un vector de nivel de demanda, $\overline{y}$, que debe especificar el experto. Por esta razón, suele denominarse ``programación por metas'' (goal programming) {Charnes, 1961 #567}. En este caso, la función *F* se obtiene por medio de la siguiente fórmula:

$$F = \left(\sum_{i=1}^{M} f_i(x) - \overline{y}_i^{\,r} \right)^{1/r} \qquad 1 \leq r \leq \infty$$

Normalmente, se elige una métrica Euclídea *r=2* , considerando como óptimos de los objetivos individuales. Es importante recalcar que la solución obtenida depende enormemente del vector de nivel de demanda establecido, de modo que si éste es malo, no se llegará a una solución Pareto-optimal. Como la solución no está garantizada, el experto debe tener un conocimiento profundo de los óptimos individuales de cada objetivo para establecer adecuadamente. De esta forma, el método busca la meta indicada (representada por $\overline{y}$) para cada objetivo.

Esta técnica es similar a la anterior. La única diferencia es que ahora se requiere saber la meta de cada objetivo mientras que en el enfoque de ponderación era necesario conocer su importancia relativa.

Optimización Mediante Formulación Min-Max. Esta técnica intenta minimizar las desviaciones relativas de cada función objetivo a partir de óptimos individuales, esto es, intenta minimizar el conflicto entre objetivos (Osyczka 1978). El problema Min-Max se define como:

$$Minimizar\ f(x)=max[Z_j\,(x)];\ \ j=1,..,M$$

donde $Z_j(x)$ se define para el valor óptimo positivo $\overline{f_j} > 0$ como:

$$Z_j(x) = \frac{\left| f_j - \overline{f_j} \right|}{\overline{f_j}}; \qquad j = 1,...,M$$

Esta técnica puede obtener la mejor solución cuando los objetivos a optimizar tienen igual prioridad. Sin embargo, la prioridad de cada objetivo puede alterarse utilizando pesos en la fórmula. También es posible introducir un vector de nivel de demanda.

Inconvenientes de las Técnicas Clásicas. Todas las técnicas clásicas que se han utilizado para resolver problemas multiobjetivo tienen graves inconvenientes que han dado lugar a que no sean adecuadas en muchas ocasiones. A continuación, mencionamos los más significativos:

- Dado que los distintos objetivos se combinan para formar uno único, sólo podremos obtener una solución Pareto-optimal simultáneamente. En situaciones reales, los expertos necesitan con frecuencia varias alternativas para decidir, pero estas técnicas no pueden ofrecerlas.

- Además, para realizar esta combinación, muchas veces es necesario tener un conocimiento sobre el problema que generalmente es difícil de obtener.

- No funcionan bien cuando los objetivos no son fiables o tienen un espacio de variables discontinuas.

- Si las funciones objetivo no son determinísticas, la elección de un vector de pesos o de niveles de demanda entraña gran dificultad.

- Son muy sensibles y dependientes de los pesos o niveles de demanda usados.

Para solucionar estos problemas, han surgido técnicas avanzadas de optimización multiobjetivo basadas en Enfriamiento Simulado (Lee and Wang 1992; Bennage and Dhingra 1995), AGs (Deb 2001; Coello-Coello, Veldhuizen et al. 2002), EEs (Kursawe 1991), etc. En concreto, los AEs multiobjetivo han demostrado un muy buen comportamiento.

2.4 Test de Coincidencia

Los test de coincidencia permiten evaluar la superposición entre dos grupos de observaciones, lo cual permite comparar diferentes cluster. Por ejemplo, estos tests permiten comparar los resultados de aplicar diferentes métodos de clustering a una misma base de datos de observaciones.

Un método frecuentemente utilizado es el valor de probabilidad (*p-value*) para identificar las relaciones entre pares de observaciones conjuntos sobrerepresentadas en el conjunto de datos. En otras palabras, identificar las relaciones que ocurren con más frecuencia que la esperada por puro azar, asumiendo la independencia de los conjuntos a evaluar. Esta información puede ser utilizada para descubrir hechos interesantes y crear hipótesis sobre la relaciones entre el par de los conjuntos siendo evaluado.

La probabilidad de intersección (*PI*) , basada en la distribución hipergeométrica permite calcular el *p-value* de la relación entre dos subconjuntos (Tavazoie, Hughes et al. 1999). Específicamente, la probabilidad de observar al menos *k* elementos en común entre dos clusters uno con *h* elementos y el otro con *n* elementos esta definida como:

$$PI = 1 - \sum_{q=0}^{k-1} \frac{\binom{h}{q}\binom{g-h}{n-q}}{\binom{g}{h}}$$

donde *g* es el universo de observaciones totales. A menor *p-value*, mayor la coincidencia. La probabilidad de intersección es una métrica sensitiva al contexto, que considera el tamaño del dominio de los clusters donde se calcula la intersección. Esto es evidente al compararla con otras métricas, como ser Jaccard (Saporta 1996) definida por :

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

donde *A* y *B* son los clusters o subconjuntos. Por ejemplo, dados dos cluster, uno con 15 y 12 elementos respectivamente, 7 de los cuales en común, Jaccard siempre evalúa la misma coincidencia (J=0.35) mas allá del universo de observaciones. Por otra parte, en un universo de 30 observaciones la *PI* obtenido es 0.3552. Al ser evaluado en un universo con 60 observaciones, la *PI* obtenido es de 0.0064. Por lo general, se toma como significativos valores inferiores a 0.05 o 0.01. Esta diferencia, se puede interpretar que la relación entre los dos clusters no se aparte de lo puramente azaroso en el universo de 30 observaciones, cuando si lo hace en el universo de 60 observaciones, por lo tanto si es significativa en este segundo caso.

2.5 Inferencia

2.5.1 Algoritmo de los *k*-vecinos más cercanos.

Los *k*-vecinos más cercanos (*k*-nearest neighbor, (*k*-NN)) es un método de clasificación de objetos basado en la similaridad a los ejemplos de entrenamiento. Éste es un método de los llamados *"lazy learning"* en el cual la función de clasificación es solo aproximada localmente, ya que todos los cómputos son retrasados hasta el momento de clasificación, sin casi existir una etapa de entrenamiento.

k-NN es uno de los métodos mas sencillos para el reconocimiento de patrones. Un objeto es clasificado en base al vote mayoritario de sus vecinos: la clase mayoritaria de los vecinos es la asignada al objeto. El aprendizaje consiste en averiguar el numero *k* de vecinos a considerar (Figure 2.9), así como la medida de similaridad para evaluar cuales son las instancias mas cercanas al objeto. Al seleccionar numero impares para *k* se evitan los empates de las votaciones, pero otros mecanismos de resolución mas sofisticados pueden ser utilizados.

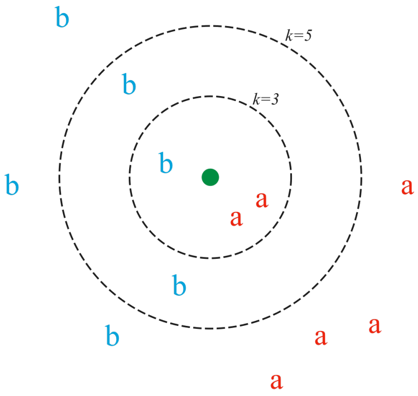


Figure 2.9 Ejemplo de clasificación mediante *k*-NN
El test (objeto a clasificar) es indicado en verde. Si el parametro k=3 es clasificado como clase a, ya que de los 3 vecinos mas cercanos, 2 tienen este valor. Cuando k=5, es clasisicado como clase b, ya que 3 de los 5 vecinos mas cercanos votan por este valor.

2.5.2 Sistemas Basados en Reglas Difusas

Los sistemas basados en reglas difusas (SBRDs) fueron inicialmente propuestos por Mamdani y Assilian (Mamdani 1974), que plasmaron las ideas preliminares de Zadeh (Zadeh 1973) en el primer SBRD concreto en una aplicación de control (i.e. control automatizado; clasificación de datos; sistemas expertos).

La inferencia difusa plantea la formulación de la predicción de un valor esperado a partir de observaciones utilizando la lógica difusa. Involucra los conjuntos difusos, la lógica difusa y las reglas difusas:

$$\textbf{SI } X_1 \text{ es } A_1 \text{ y.... y } X_n \text{ es } A_n \textbf{ ENTONCES } Y \text{ es } B$$

siendo $X = (X_1, \ldots, X_n)$ las variables difusas de entrada, Y la variable difusa de salida, y $A_1,\ldots, A_n$ y B las funciones de pertenencia asociadas a las variables de entrada y salida respectivamente. Simplificando esta expresión con $A = A_1 x \ldots x\ A_n$ (donde el símbolo x denota el producto cartesiano) tenemos

$$\textbf{SI } X \text{ es } A \textbf{ ENTONCES } Y \text{ es } B,$$

A la variable difusa A se la denomina *antecedente* o *premisa* mientras que a la variable difusa B es llamada *consecuente* o *conclusión*. Estas reglas se interpretan de la siguiente manera: Si el antecedente es cierto con cierto valor de pertenencia entonces en cierta medida el consecuente también lo es. El proceso de inferencia difusa está basado en la aplicación del *modus ponens generalizado*, extensión del *modus ponens* de la lógica clásica, propuesto por Zadeh según la siguiente expresión (Zadeh 1973):

SI *X* es *A* **ENTONCES** *Y* es *B*
X es *A′*

Y es *B′*

El operador de implicación modifica el conjunto difuso consecuente según el valor de pertenencia del antecedente. Una de las formas mas habituales de modificar *B* es truncarlo mediante el uso e la función *min* (Figure 2.10)

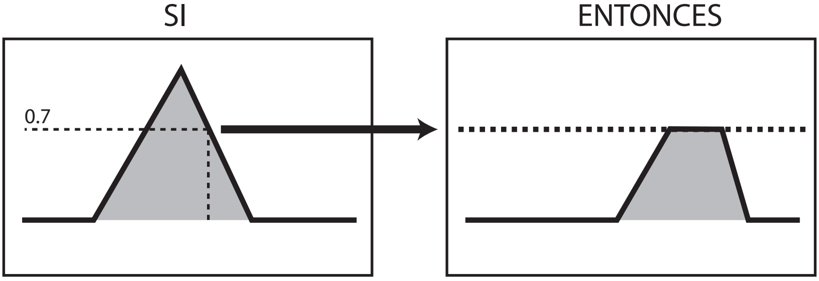


Figure 2.10 Aplicación del operador de Implicación

La agregación ADEMÁS, que se emplea para combinar las distintas salidas individuales de un conjunto de reglas en una final en conjunción con un método de defuzzificación. La composición de este operador depende del tipo de defuzzificación que el SBRD emplee.

Puesto que el sistema debe devolver una salida precisa, la interfaz de defuzzificación debe asumir la tarea de agregar la información aportada por cada uno de los conjuntos difusos individuales y transformarla en un valor preciso. Existen dos formas de trabajo diferentes para efectuar esta agregación:

Agregar primero, defuzzificar después:. En este primer caso, la interfaz de defuzzificación lleva a cabo las siguientes tareas:

- Agrega los conjuntos difusos individuales inferidos *B′ᵢ*, para obtener un conjunto difuso final *B′*, empleando para ello un operador de agregación difuso *G* que, modelando el operador ADEMÁS, relaciona las reglas:

$$\mu_B(y) = G\left\{\mu_{B'_i}(y),\ldots,\mu_{B'_m}(y)\right\}$$

- Mediante un método de defuzzificación *D*, transforma el conjunto difuso *B′* obtenido en un valor preciso y_0, que será proporcionado como salida global del sistema:

$$y_0 = D(\mu_{B'}(y))$$

Defuzzificar primero, agregar después: Este segundo modo de trabajo considera individualmente la contribución de cada conjunto difuso inferido y el valor preciso final se obtiene mediante una operación (una media, una suma ponderada o la selección de uno de ellos, entre otras) sobre un valor preciso característico de cada uno de los conjuntos difusos individuales.

De este modo, se evita el cálculo del conjunto difuso final *B'*, hecho que ahorra una gran cantidad de tiempo computacional. Este modo de operación supone una aproximación distinta al concepto representado por el operador ADEMÁS.

Inicialmente fue propuesto el modo *Agregar primero, defuzzificar después*, el cual fue empleado por Mamdani en su primera aproximación al control difuso (Mamdani 1974). En los últimos años, la modalidad *Defuzzificar primero, agregar después* ha sido muy empleada (Driankov, Phil et al. 1993; Sugeno and Yasukama 1993; Cordon, F. et al. 1997), sobre todo en sistemas de tiempo real, donde se requieren tiempos de respuesta rápidos.

Cuando se trabaja en el primero de estos modos, la función del operador de agregación ADEMÁS es unir todos los conjuntos difusos, resultantes de la inferencia de cada regla, en un único conjunto difuso global. Para definir matemáticamente este operador, se emplean distintos operadores, principalmente *t-normas* y *t-conormas*, los cuales están descritos en (Bardosey and Ducjstein 1995), donde también se analizan sus propiedades en detalle.

En cuanto a la definición matemática del modo de defuzzificación a emplear para transformar el conjunto difuso global resultante del proceso de inferencia en un valor preciso de salida, encontramos que los más habituales son: el *centro de gravedad*, el *centro de sumas* (aproximación al centro de gravedad computacionalmente más rápida de obtener) y la *media de los máximos* (Driankov, Phil et al. 1993). Por otro lado, en caso de emplear el modo de *defuzzificar primero y agregar después*, los operadores de agregación más utilizados son la *media*, la *media ponderada* o la *selección de algún valor característico* de los conjuntos difusos en función del grado de importancia de la regla que los ha generado en el proceso de inferencia (Cordon, F. et al. 1997). Como métodos para extraer valores representativos se suelen emplear el *centro de gravedad* y el *punto de máximo criterio;* y como grados de importancia de la regla, el *área* y la *altura* del conjunto difuso inferido o el grado de emparejamiento de los antecedentes de la misma con la entrada al sistema. El operador más empleado dentro de este grupo es la media ponderada por el grado de emparejamiento, que se suele combinar con el *centro de gravedad* como valor característico del conjunto difuso (Sugeno and Yasukama 1993; Cordon, F. et al. 1997) (Hellendoorn and Thomas 1993).

2.6 Algoritmos evolutivos

La *Computación Evolutiva* (CE) se basa en el empleo de modelos de procesos evolutivos para el diseño e implementación de sistemas de resolución de problemas. Los distintos modelos computacionales que se han propuesto dentro de esta filosofía suelen recibir el nombre genérico de *Algoritmos Evolutivos* (AEs) (1997). Existen cuatro tipos de AEs bien definidos que han servido como base a la mayoría del trabajo desarrollado en el área: los *Algoritmos Genéticos* (AGs), las

Estrategias de Evolución (EEs), la *Programación Evolutiva* (PE) y la *Programación Genética* (PG).

Un AE se basa en mantener una población de posibles soluciones del problema a resolver, llevar a cabo una serie de alteraciones sobre las mismas y efectuar una selección para determinar cuáles permanecen en generaciones futuras y cuáles son eliminadas. Aunque todos los modelos existentes siguen esta estructura general, existen algunas diferencias en cuanto al modo de ponerla en práctica. Los AGs se basan en operadores que tratan de modelar los operadores genéticos existentes en la naturaleza, como el cruce y la mutación, los cuales son aplicados a los individuos que codifican las posibles soluciones. En cambio, las EEs y la PE aplican transformaciones basadas en mutaciones efectuadas sobre los padres para obtener los hijos, lo que permite mantener una línea general de comportamiento del individuo en su descendencia. Finalmente, la PG codifica las soluciones al problema en forma de programas, habitualmente codificados en una estructura de árbol, y adapta dichas estructuras empleando operadores muy específicos.

Cada individuo de la población recibe un valor de una medida de adaptación que representa su grado de adecuación al entorno. La selección hace uso de estos valores y se centra en los individuos que presentan mayor valor en la media. Los operadores de recombinación y/o mutación alteran la composición de dichos individuos, guiando heurísticamente la búsqueda a través del espacio. Aunque simples desde un punto de vista biológico, este tipo de algoritmos son suficientemente complejos para proporcionar mecanismos de búsqueda adaptativos muy robustos. Los mismos procedimientos pueden ser aplicados a problemas de distintos tipos, sin necesidad de hacer muchos cambios (Goldberg and Richardson 1987).

En las siguientes secciones, se explicarán con más detalle los algorítmos evolutivos, que son los utilizados en este trabajo.

2.6.1 Algoritmos genéticos

Los AGs (Goldberg and Richardson 1987) son una técnica *metaheurística* para la solución de problemas de optimización.

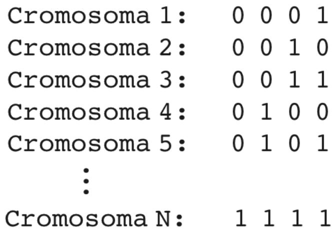
> Una metaheurística es un proceso iterativo que guía a una heurística subordinada combinando inteligentemente diferentes conceptos para explorar y explotar el espacio de búsqueda, usando estrategias de aprendizaje para estructurar la información con el objetivo de encontrar eficientemente soluciones cercanas al óptimo (Osman 1995).

Los AGs se basan en una analogía de la teoría biológica de la evolución de las especies y toman esta idea para buscar una o más soluciones óptimas entre un conjunto de posibles soluciones.

La *Teoría de la Evolución* explica el origen y la transformación de los seres vivos como el producto de la acción de dos principios fundamentales: la selección natural y el azar. La selección natural regula la variabilidad de la recombinación y mutación aleatorias de los genes: toda la variedad que observamos en la naturaleza se basa en la capacidad de los seres vivos de producir copias de sí mismos, en que el proceso de reproducción actualiza muchas variantes, y en que, en la interacción con el ambiente, algunas de ellas son seleccionadas para sobrevivir y producir las copias subsiguientes (Darwin 1859).

En lugar de buscar de general a específico o de simple a complejo, los AGs generan descendientes de soluciones realizando repetidas mutaciones y cruces de las mejores soluciones de un conjunto. A cada paso, una colección de soluciones es actualizada reemplazando una fracción de la población por la descendencia de las soluciones más adaptadas al medio. Entonces se puede ver que estas soluciones serán las que tendrán mayor probabilidad de pasar a la próxima generación.

Para poder mostrar el algoritmo propuesto y explicar como funcionan, en general, los AGs es necesario primero comprender la terminología utilizada. En las Figure 2.11, Figure 2.12 y Figure 2.13 se pueden ver los conceptos en forma gráfica tomando como ejemplo el problema de encontrar el número máximo entre 1 y 15 usando una representación binaria.



```
Cromosoma 1:   0 0 0 1
Cromosoma 2:   0 0 1 0
Cromosoma 3:   0 0 1 1
Cromosoma 4:   0 1 0 0
Cromosoma 5:   0 1 0 1
      ⋮
Cromosoma N:   1 1 1 1
```

Figure 2.11 Población

Población. Se denomina población al conjunto de individuos que representan las soluciones a optimizar.

Cromosoma - Gen. Se denomina cromosoma a cada individuo de la población. A su vez, se conoce con el nombre de gen a cada parte del cromosoma que tiene significado por sí misma.

Genotipo - Fenotipo. En la naturaleza, un genotipo es la información genética que, al desarrollarse, crea un fenotipo o ser vivo. En el ámbito de los AGs, se denomina genotipo al conjunto de parámetros representado por un cromosoma particular que contiene toda la información necesaria para construir una solución (organismo), a la cual se la denomina fenotipo.

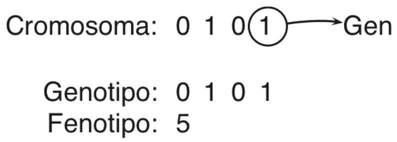
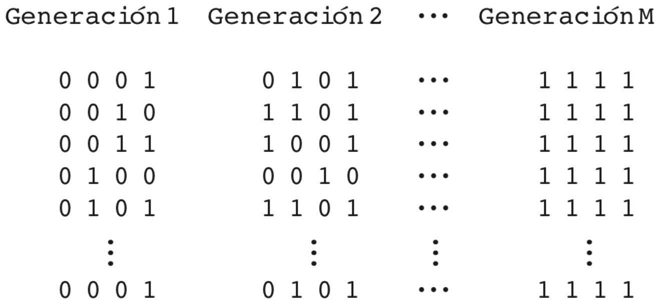


Figure 2.12 Genotipo vs. Fenotipo

Generación. Se denomina generación a cada iteración del algoritmo.



Generación 1	Generación 2	⋯	Generación M
0 0 0 1	0 1 0 1	⋯	1 1 1 1
0 0 1 0	1 1 0 1	⋯	1 1 1 1
0 0 1 1	1 0 0 1	⋯	1 1 1 1
0 1 0 0	0 0 1 0	⋯	1 1 1 1
0 1 0 1	1 1 0 1	⋯	1 1 1 1
⋮	⋮	⋮	⋮
0 0 0 1	0 1 0 1	⋯	1 1 1 1

Figure 2.13 Generaciones

2.6.1.1. Algoritmo genético básico

Los AGs exploran un espacio de soluciones candidatas en busca de la mejor solución. Cuando decimos la mejor solución, nos referimos a aquella que optimice una cierta función relevante para el problema tratado, a la que se conoce con el nombre de función de *fitness* o función de aptitud. A pesar que existen clases muy diversas de AGs, todas mantienen una estructura en común:

> En cada iteración, todos los miembros de la población se evalúan de acuerdo a la función de fitness. Se genera una nueva población seleccionando de forma probabilística los mejores individuos de la población actual. Algunos de estos individuos pasan a la próxima generación automáticamente, mientras que otros se utilizan para procrear nuevas soluciones por medio de cruces de dos individuos, o bien se mutan antes de pasar a la próxima generación.

El algoritmo itera hasta cumplir con un criterio de parada. Éste puede ser la cantidad de generaciones o evaluaciones de la función de fitness realizadas o la obtención de una solución que esté dentro de un cierto umbral de aceptación. El pseudocódigo del algoritmo básico para un algoritmo genético simple se muestra en la Figure 2.12 (Goldberg and Richardson 1987).

Representación de los cromosomas

La representación de los cromosomas depende del problema a tratar e influye directamente sobre los resultados del AG. Existen distintos esquemas generales de codificación entre los que destacan los siguientes:

- La *codificación binaria*: Es la más antigua de todas las existentes. La representación de los cromosomas está definida como cadenas de bits de modo que, dependiendo del problema, cada gen puede estar formado por una subcadena de uno o varios bits.
- La *codificación real*: La codificación binaria presenta algunos inconvenientes cuando se trabaja con problemas que incluyen variables definidas sobre dominios continuos: excesiva longitud de los cromosomas, falta de precisión, etc. Una posible manera de evitar estos inconvenientes es considerar un esquema de representación real. Aquí, cada variable del problema se asocia a un único gen que toma un valor real dentro del intervalo especificado, por lo que no existen diferencias entre el genotipo y el fenotipo.
- La *codificación basada en orden*: Este esquema está diseñado específicamente para problemas de optimización combinatoria en los que las soluciones son permutaciones de un conjunto de elementos determinando.

Estos ejemplos de posibles representaciones nos dan una idea genérica del tipo de esquemas que se utilizan más comúnmente. Esto no implica que sean los únicos, o que no se puedan crear esquemas propios, que no tengan relación alguna con los comentados, si es que se adaptan mejor a un problema en particular.

Mecanismo de Selección

El mecanismo de selección es el encargado de seleccionar la población intermedia de individuos la cual, una vez aplicados los operadores de cruce y mutación, formará la nueva población del AG en la siguiente generación. De este modo, el mecanismo de selección se encarga de obtener una población intermedia formada por copias de los cromosomas de la población original como se muestra en la Figure 2.14.

Dos son los métodos de selección más comunes:

- *La Ruleta.* Consiste en crear una ruleta en la que cada cromosoma tiene asignada una fracción proporcional a su aptitud. Esta ruleta se gira varias veces para determinar qué individuos se seleccionarán. Debido a que a los individuos más aptos se les asignó un área mayor de la ruleta, se espera que sean seleccionados más veces que los menos aptos.
- *El torneo.* La selección por torneo se ha popularizado debido a que sólo utiliza información local para elegir a los mejores candidatos, por tanto reduciendo la complejidad de cálculo en poblaciones de gran tamaño. El torneo consiste en seleccionar un conjunto de individuos de la población al azar, dependiendo del tamaño del torneo, para luego comparar entre sí dichos individuos y elegir aquél con mejor valor de aptitud. Este proceso se realiza tantas veces como elementos existan en la pobla-

ción. En el caso de un torneo binario, la competencia se realiza entre dos individuos, seleccionando aquél con mejor función de aptitud.

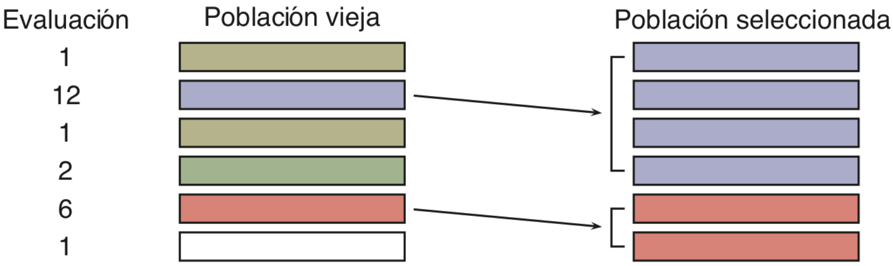


Figure 2.14 Ejemplo de aplicación del mecanismo de selección

Operadores genéticos

El operador de cruce constituye un mecanismo para compartir información entre cromosomas. Combina las características de dos cromosomas padre para obtener dos descendientes, con la posibilidad de que los cromosomas hijo, obtenidos mediante la recombinación de sus padres, estén mejor adaptados que éstos. No suele aplicase a todas las parejas de cromosomas de la población intermedia, sino que se lleva a cabo una selección aleatoria en función de una determinada probabilidad de aplicación, la *probabilidad de cruce, PC*

El operador de cruce cumple un papel fundamental en el AG. Su tarea consiste en **explotar** el espacio de búsqueda combinando las soluciones obtenidas hasta el momento mediante la recombinación de las buenas características que presenten.

Un ejemplo del cruce más conocido, llamado *cruce simple en un punto*, se muestra de forma gráfica en la Figure 2.15. El cruce simple se basa en seleccionar aleatoriamente un punto de cruce e intercambiar el código genético de los cromosomas padre a partir de dicho punto para formar los dos hijos.

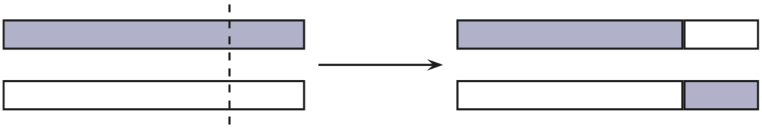


Figure 2.15 Ejemplo de aplicación del operador de cruce simple de un punto

La mutación, en cambio, pretende **explorar** el espacio de búsqueda alterando una de las componentes del código genético de un individuo. La mutación altera localmente el genotipo esperando obtener un individuo mejor.

Debido al efecto de la selección, se sabe que sólo serán elegidas las buenas soluciones para pasar a la próxima generación, mientras que las malas soluciones serán eliminadas.

2.6.1.2. Algoritmos genéticos para funciones multimodales: *Nichos*

Como el nombre sugiere, las funciones multimodales tienen múltiples soluciones óptimas, de las cuales varias pueden ser óptimos locales. Como ya hemos comentado, los AGs son conocidos por su capacidad para llevar a cabo procesos de búsqueda en espacios complejos. A pesar de ello, un AG clásico puede no trabajar de modo adecuado cuando el espacio de búsqueda es multimodal y presenta varios óptimos locales. En estos casos, los AGs simples se caracterizan por converger hacia la zona del espacio donde se encuentran los mejores óptimos locales, abandonando la búsqueda en las zonas restantes (proceso conocido con el nombre de ``deriva genética") (Goldberg and Richardson 1987).

Se han propuesto varios métodos para tratamiento de funciones multimodales en AGs. Veremos a continuación algunos de ellos:

- **Diversidad a través de la mutación**. Encontrar soluciones cercanas a diferentes soluciones óptimas en una población y mantenerlas por varias generaciones son dos temas diferentes. En las etapas iniciales del AG, las soluciones están distribuidas por todo el espacio de búsqueda. Las soluciones cercanas a los múltiples óptimos se enfatizan en las primeras generaciones. Eventualmente, cuando la población contiene buenos clusters de soluciones cercanas a los óptimos, comienza la competencia entre los diferentes óptimos. Este proceso competitivo continúa hasta que la población converge a un solo óptimo. Para lograr mantener las soluciones encontradas en las primeras generaciones, es necesario un operador explícito de preservación de diversidad. El operador de mutación se utiliza, usualmente, con esa función. Este operador tiene una función tanto constructiva como destructiva. Debido a esto, se suele utilizar una baja probabilidad de mutación. Esto produce que sea insuficiente como único operador de preservación de diversidad.

- **Preselección**. En (Cavicchio 1970), se utiliza un mecanismo conocido como *preselección*. El concepto principal de este operador es reemplazar con un individuo *parecido*. Cuando se genera un hijo a partir de dos soluciones padre, se compara automáticamente con ambos progenitores. Si el hijo está mejor adaptado que el peor de sus padres, entonces reemplaza al padre. Como un hijo suele ser similar a sus padres, el reemplazo permite que diferentes soluciones co-existan en la población, manteniendo de esta forma la diversidad.

- **Modelo de Crowding**. En (De Jong 1975) se desarrolló este método para introducir diversidad en la población de un AG. En este esquema, se emplea una población solapada donde los individuos reemplazan a aquellos individuos de la población original de acuerdo a su similaridad. Un individuo se compara con una subpoblación de miembros de la población solapada determinada al azar. El individuo con la mayor similaridad es reemplazado por este nuevo individuo.

Sin embargo, la forma más habitual de diseñar AGs multimodales se basa en los conceptos de *nicho* y *especie*. Ambos conceptos fueron introducidos con objeto de mantener múltiples óptimos. Una de las formas más habituales para provocar la formación de especies y la creación de nichos se basa en el esquema de *sharing* o proporción (Goldberg and Richardson 1987). El proceso de sharing permite mantener en la población de cada generación una cantidad proporcional de individuos en distintas zonas del espacio de búsqueda, manteniendo de esta manera una buena diversidad de soluciones. Cada zona del espacio de búsqueda estará representado en el algoritmo como un *nicho*. En cada una se encontrará un conjunto de soluciones cercanas de acuerdo a una cierta distancia. Estas subpoblaciones crean nichos en dos espacios posibles: el genotípico y el fenotípico. Al primero se lo conoce también con el nombre de *espacio de variables*. Además del espacio de variables, existe también el *espacio de objetivos*. Este espacio está determinado por los objetivos del problema a optimizar. Para medir distancias en un espacio o en el otro, utilizaremos medidas diferentes, las que resulten más adecuadas al espacio y al problema.

Como ocurre en la naturaleza, los individuos de cada nicho comparten la recompensa asociada a dicho nicho entre ellos. Para esta tarea se define una función conocida como *fitness sharing* (función de proporción). Este método permite distribuir la población sobre diferentes picos (máximos o mínimos locales dependiendo del tipo de función de aptitud utilizada) del espacio de búsqueda, donde la cantidad de individuos que recae en cada pico es proporcional a la calidad del pico si se trata de optimizar una función.

Entonces, la selección de individuos debe permitir que elementos pertenecientes a distintos nichos sean mantenidos en las futuras generaciones. Para poder hacerlo, se genera un torneo. Dos individuos compiten entre sí para determinar cual de los dos pasará a la próxima generación. Para decidir cual de los dos competidores será el ganador, se calcula la cantidad de elementos que existen en los nichos a los cuales pertenecen. La cantidad de elementos en cada nicho se define como:

$$nicho(x_i) = \sum_{x_j \in P} sh(d(x_i, x_j))$$

donde *d* es la medida de distancia entre dos elementos, *P* es el conjunto de individuos de la población y *sh* es la función de sharing. Esta función se define por lo general como:

$$sh(v) = \begin{cases} 1 - v/\sigma_{share} & v \leq \sigma_{share} \\ 0 & v > \sigma_{share} \end{cases}$$

En este caso, σ_{share} es el radio del nicho que debe ser especificado por el usuario y determina la mínima separación deseada entre picos. La función de proporción (función de fitness modificada) sobre un individuo x_i se define entonces como $f(x_i)/nicho(x_i)$, donde *f* es el valor de su función de fitness. Este proceso permite evitar que toda la población converja a una única solución y, en lugar de ello, permite que los individuos de cada nicho converjan independientemente. Es deseable obtener nichos igualmente poblados en las distintas generaciones de forma tal que:

$$\frac{f(x_i)}{nicho(x_i)} = \frac{f(x_j)}{nicho(x_j)} \quad \forall x_i, x_j \ \ individuos$$

2.6.1.3. Elitismo

El ciclo de nacimiento y muerte de los individuos está muy relacionado con el manejo de la población. El tiempo de vida de un individuo es típicamente de una generación, aunque en algunos AGs puede ser mayor. La estrategia de elitismo relaciona la vida de los individuos con su aptitud. Estas estrategias son técnicas utilizadas para mantener las buenas soluciones más de una generación. Una estrategia de elitismo habitualmente utilizada en AGs es mantener una copia del mejor individuo encontrado hasta el momento en cada generación. Esto se realiza porque en el cruce los padres suelen ser reemplazados por sus hijos y, por ello, no hay seguridad de que los individuos con mayor aptitud sobrevivan a la próxima generación.

2.6.2 Problemas multiobjetivo y algoritmos genéticos multiobjetivo

Una diferencia notable entre los métodos de búsqueda y optimización clásicos y los AGs es que, en estos últimos, se procesa una población de soluciones en cada iteración. Esta característica, por sí sola, le da a los AGs una tremenda ventaja para su uso en problemas de optimización con múltiples objetivos.

Existen, al menos, dos dificultades asociadas a los métodos clásicos de resolución de estos problemas:

- Para obtener varias soluciones de la frontera del Pareto con un método clásico de optimización, es necesario ejecutarlo varias veces, comenzando cada vez de una solución inicial diferente.
- Sin embargo, las diferentes ejecuciones de los métodos clásicos sobre distintas soluciones iniciales no garantizan encontrar diferentes soluciones óptimas. Este escenario sólo ocurre si la solución inicial elegida es atraída siempre por el óptimo.

2.7 Metodology

In this dissertation we are concerned with the problem of discovering interesting qualitative structures in complex biological systems and the associated problem of determining interesting relations between these structures. It is assumed that the notion of interestingness, which is problem-dependent, is formally defined by means of a family of parameterized models $M = \{M_\alpha\}$ and by a set of relations between them that are provided beforehand by domain experts.

The models contained in the collection *M* are approximate or qualitative in the sense that they measure the degree of matching between substructures—corresponding to a subset of the dataset representing the object—and prototypical instances of the interesting feature.

Cluster analysis—or simply clustering—is a data mining technique often used to identify various groupings or taxonomies in databases. Most existing methods for clustering are designed for grouping instances based on the whole set of the linear features that describe the instances of the dataset. However, sometimes we need to represent data from distinct sources of information and thus distinct domains. Moreover, beforehand we do not know which features are relevant; or might expect that these features are significant to some clusters and result non-informative for some others. Therefore, from our perspective mining into biological databases entails addressing both the uncertainty of which observations should be placed together, and also which distinct relationships among features best characterize different sets of observations. Typical clustering techniques (Der and Everitt 1996) are not designed to achieve this, even when combined with global filter feature selection methods such as principal component analysis or stepwise descendent methods (Kohavi and John 1997; Yeung and Ruzzo 2001).

In contrast, conceptual clustering techniques have been successfully applied to uncover concepts that are embedded in subsets of data or substructures (Cheeseman and Oldford 1994; Cook, Holder et al. 2001) (Cooper and Herskovits 1992). Consequently, conceptual learning can be formulated as the problem of searching through a predefined space of potential profiles (i.e. non-disjoint substructures or associations of features and observations) for those observations that best fit the training examples.

In this work, we propose a methodology for applying machine learning techniques to biological databases with the foundations of the conceptual clustering methods (Figure 2.16). Our framework encompass a variety of metrics, heuristics or probability interpretations (Cheeseman and Oldford 1994; Cook, Holder et al. 2001) (Cooper and Herskovits 1992) that are used depending on the context of problem, the quality of the data, its granularity and its own constrains. Moreover, its global procedure is generalized, or abstracted in five clearly defined phases that are summarized in Figure 2.16.

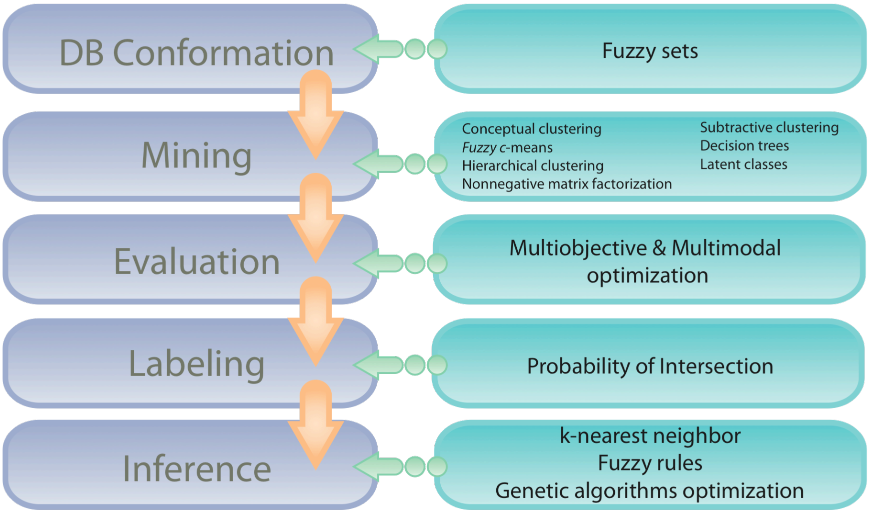


Figure 2.16 The five phases of the proposed the methodology

Database conformation. Data can be efficiently organized by taking advantage of a naturally occurring structures over feature space. The objectives of this representation are both the annotation of complex objects in terms of features meaningful to database users and the eventual analysis of the underlying systems by knowledge-discovery techniques (Ruspini and Zwir 2001). We learn initial family of models $M = \{M_\alpha\}$ of the data and represent instances by the similarity to these prototypes.

> *Example: We decompose transcription factor binding sites motifs into a family of models, termed submotifs. These are learnt by applying a "Divide and Conquer" strategy to the input sequence dataset, and let uncover intrinsic properties of subsets of binding sites, concealed by the single motif.*

Mining. We employ an approach that generates a set of profiles by:

a) Searching through the feature space for potential relation among substructures, by aggregating the models that represent distinct attributes, or features, and obtaining a lattice of profiles. The we return either the best one found or an optimal sample of them.

b)Learning profiles by employing distinct clustering methods, and a set of distinct parameters for each one (i.e. number of cluster and distance measures -Euclidian, Pearson correlation-), that highlight different commonalities.

Instead of reducing the clusters to certain granularity or selecting a subset that covers the whole dataset, we let non-disjoint cluster to exist (allowing an observation to be member of more than one cluster). By this way, we are able to cluster the dataset by the different optimal number of partitions indicated by the validity indices: although a full battery of these has been developed (accounting on *crisp*, *fuzzy* and *probabilistic* validity indices), they usually indicate contradictory partition number. Another alternative to face this problem is provided by the *dynamic* cluster validation, that proposes the integration of validity criteria into the clustering scheme itself (Bezdek 1998). In this approach clusters are continuously evaluated and those that do not conform a specific criteria ("weak" clusters) are merged and the data reclustered after adjusting the parameters. Unfortunately this approach restricts the employment of those methods that do not contemplate this strategy and are already proved to be useful (e.g. hierarchical clustering, k-means).

To be able to make use of any available method, we learn the most significant features among clustered observations, when is not provided by the method itself, by combining supervised/unsupervised methods (i.e. local feature selection). As might be expected this approach generates an excessive and sometimes unmanageable number of profiles. However, at this phase we only eliminate those clusters that show a high degree of overlap of observations among them, obtaining a non-compact set of profiles.

> *Example: a) To uncover the PhoP regulon we navigate the complete set of profiles that describe promoter regions of target genes by using cis-acting elements as describing features. We are able to detect genes that differ from the canonical PhoP-regulated promoters. These new detected genes were previously though to be indirectly regulated by PhoP. b) We employ a complete battery of clustering methods to learn phenotypic profiles of patients suffering Schizophrenia, relatives and controls (i.e. status). The particularities of this dataset (i.e. a total number of a hundred individuals and +60 features) let us conform a set of profiles that group individuals independently of their status. A coincidence analysis to the genotypic data revels that some of these profiles are far more qualitative than the ones obtained by barely using the status. Moreover these profiles let us build a risk function, until now only theoretically proposed.*

Evaluation. The formulation of the clustering problem in a lattice or non-compact set of profiles would result in the generation of many substructures with small extent, as it is easier to explain or substructure-match smaller data subsets than those that constitute a significant portion of the dataset. For this reason, any successful methodology should also consider additional criteria to extract broader or more comprehensive substructures based on their size, the number of retrieved substructures, and their diversity and extent of overlap (Cook et al., 2001; Ruspini, 2001). These are conflicting criteria that can be formulated as a multi-objective optimization problem, analogous to minimum description-length methods (Rissanen, 1989), based on the combination of the individual criteria or objectives into a global measure of cluster quality. The basic challenge with this approach, however, is its potential bias and inflexibility caused by weighting of the objectives (Ruspini, 2001).

> *Example: The lattice of profiles for the PhoP regulon promoters is evaluated by two objectives in conflict: the number of promoters described by each profile and the similarity among observations. We learn the Pareto optimal frontier, of those non-dominated solutions (i.e. t there is no other solution that is superior*

> *to them in at least one objective, and as good in the others), which are used to plan the experiments that validate our findings.*

Labeling. The above phases of our methodology propose an unsupervised approach of machine learning. Thus the methodology is neither constrained by a dependent variable (Mitchell 1997; Beer and Tavazoie 2004), nor requires pre-existing data. As a result, one or more external classes (i.e. independent data, usually product of experimentations) can be incorporated into the analysis. We evaluate the coincidence of the observations of the learned profiles to external, or output classes, by employing a context-sensitive metric, which takes into account the extent of the profile being evaluated, the external class, and the universe of observations. This approach allows the identification of distinct profiles that formulate valid hypothesis about the underlying mechanism or origins of the external classes.

> *Example: The coincidence of independently learned profiles of the PhoP regulated promoters and GFP assays for a subset of these genes is evaluated. This let us hypothesize about the different mechanisms that the cell employs to obtain differential gene expression patterns.*

Inference. We employ the previously learned profiles to predict new observations output class, by evaluating their distance, or degree of similarity to profiles prototypes. We employ heuristics (i.e. Genetic algorithms) to learn optimal prototype similarity thresholds that let them maximize the global overall classification performance, constraining the influence of profiles and eliminating the redundant ones.

> *Example: We use submotifs of PhoP binding sites to feed tools that encode collection of sequences into models useful to screen genomes and detect new instances (i.e. Consensus, MEME, AlignAce). After learning their optimal thresholds we are able to improve the sensitivity of these tools.*

Chapter 3

Delimiting plasticity of transcription factor binding sites by disassembling DNA consensos sequences

Whole genome sequences, microarray and ChIP data that examine genome-wide gene expression patterns provide the raw material for the characterization and understanding of the underlying regulatory systems. However, it is still challenging to discern the sequence elements relevant to differential gene expression, such as those corresponding to the binding sites for transcriptional regulators and RNA polymerase that are embedded in the background genomic DNA sequences (Tompa, Li et al. 2005). It is known that certain regulators, such as LacI and MelR from the bacterium *Escherichia coli*, control transcription of a single gene or operon (Martinez-Antonio and Collado-Vides 2003). By contrast, global regulators such as the *E. coli* ArcA and Crp proteins can govern expression of hundreds of genes (Zheng, Constantinidou et al. 2004). This raises the question: how does a single regulator distinguish promoter sequences, which affinities are a major determinant of differential expression? This problem is exacerbated though the evolution pathway, where non-monotonic co-evolution of regulators and targets (Alm, Huang et al. 2006) constitute a more intriguing scenario for discerning the DNA scope recognized by a single regulatory protein (Moses, Pollard et al. 2006), and thus, its target regulon.

To circumvent the limitation of consensus methods (Tompa, Li et al. 2005), we present a computational framework, termed *Divide & Conquer* (*D&C*), a classic approach in the machine learning literature, specifically developed for extracting the maximal amount of useful genomic information through the effective handling of the biological and experimental variability inherent in the data. We decomposed the binding site motif of a transcription factor into families of motifs (i.e., "submotifs") and then combined them using a multi-classifier (Bauer and Kohavi 1999), thereby increasing the sensitivity to weak sites without losing specificity. We incorporate other *cis* features like the relative distance between a binding site and the RNA polymerase, allowing extracting the maximum information encoded in the DNA sequences. We demonstrate that the approach works well using different grouping methods used to partition sequences into submotifs (Gasch and Eisen 2002; Hering, Innocent et al. 2004), and distinct position weight matrix (PWM) methods used to encode these groups into models. Indeed, this approach is independent of performance measurement used (Tompa, Li et al. 2005). Sensitivity analysis reveals that the approach is robust with minimum influence of the method-specific parameters.

In the first part of this paper we apply our method, based on families of motifs, to classify promoters controlled by CRP, a well characterized master regulator of >100 genes in *Escherichia coli*. Our results can be easily extrapolated to other well documented (e.g., RegulonDB) regulators. Then, we evaluate the applicability of the method to provide insights of other transcriptional regulators, which are not so well described in *E. coli* databases, and even less in other species. Thus, we analyze the promoters controlled by the PhoP/PhoQ regulatory system starting from of *E. coli* and *Salmonella* enterica serovar Typhimurium, and propagating the study to other gamma enterobacteria genomes. The PhoP/PhoQ system is an excellent test case because it controls the expression of a large number of genes (e.g., ca. 3% of the genes in the case of *Salmonella*). We explicitly demonstrate that the simple *D&C* approach improves at least 20% of the computational classification of the binding sites of these master regulators.

However, the computational usefulness of the submotifs raises the question: are the families of motifs a computational artifact or do they provide insights on the regulatory process carried out by a regulator and its targets? To address this question we evaluate three hypotheses. First, we hypothesize that biologically meaningful families of motifs should facilitates the process of distinguishing functional from non-functional binding sites in genome-wide searches. Therefore, we apply the submotifs to discern functional binding using from a combination of genome-wide chromatin inmunoprecipitation followed by array hybridization (ChIP-chip), which is often characterized by high false negative rates (Buck and Lieb 2004), and custom expression microarray analysis (Nimblegen tiling arrays). Second, we conjecture that binding constraints encoded in families of motifs may be interpreted with respect to physical constraints imposed by the DNA-binding protein (Pedersen, Jensen et al. 2000) interactions. Thus, they probably exhibit sequence dependent conformation and deformability attributes that can shed light on the geometrical interactions with other regulatory elements (e.g., RNA polymerase, other TFSs).

Third, and probably more conclusive, we evaluate the different rates of evolutions of these families of motifs (Moses, Pollard et al. 2006) and map gain and loss events along the phylogenetic tree of gamaenterobacterias. The evolutionary dynamics of the submotifs in a complex scenario of coding and/or non-coding turnovers (e.g., laterally acquired genome regions) can shed light on the biologically significance of the proposed approach, and thus, would provide a model of evolution of binding sites. In this work, we explicitly demonstrate that the proposed analysis is concealed while using a single consensus model.

3.1 Results

3.1.1 A single motif encoding TFBS sequences does not describe the entire binding dataset

A diverse collection of useful tools have been developed to analyze DNA sequences bound by a TF and to discover recurrent patterns of nucleotides that differ from the genome background (Tompa, Li et al. 2005). These tools vary their searching algorithms and measurements used to evaluate candidate motifs. For example, AlignACE employs Gibbs sampling to guide the alignment of sequences, which are evaluated by a maximum posterior score that measures its quality relative to the random expectation (Hughes, Estep et al. 2000). Another method termed MEME (Bailey, Williams et al. 2006), searches for parameters of mixture models composed of alignments of sequences and background distributions, which are evaluated by the probability of being different from random models. Finally, the Consensus method (Stormo 2000) uses a greedy searching approach to find optimal sequence alignments, which are evaluated by their information content. Although these methods partially differ in the representations of the motifs, ultimately, all of them use a position weight matrix (PWM) that describes the independent frequencies of nucleotides in a motif.

We apply these three methods to uncover the targets of the cyclic AMP receptor protein (CRP), which is a global TF that regulates over 100 *E. coli* genes. We consider 148 BS of CRP as positive examples and 622 BS of other TF as negatives examples in promoter sequences reported in RegulonDB database (Salgado, Santos-Zavaleta et al. 2001). The classification was based on method-specific thresholds (Robison, McGuire et al. 1998). We compare the obtained results using three different and/or complementary statistical indices including the Correlation Coefficient (CC) and the Standardized CC (SCC), which tends to "equilibrate" unbalanced true positive and negative datasets. We find that AlignACE employs 102 from a total of 148 BSs to build a motif and recovers 97 BS (SCC=0.61); MEME uses and recovers 60 BSs (SCC=0.45); and Consensus employs all available BS and recovers 86 BSs (SCC=0.52) (see Table 3-1 for CC results).

Table 3-1 Scores of the CRP Single Motif

Method	**TP**	**TN**	**FP**	**FN**	**SP**	**SN**	**PPV**	**CC**	**SCC**
Dataset	*148*	*622*							
Consensus	86	586	56	62	0.91	0.58	0.61	0.50	0.52
MEME	60	620	22	88	0.97	0.41	0.73	0.47	0.45
AlignACE	97	596	46	51	0.93	0.66	0.68	0.59	0.61

3.1.2 A Divide & Conquer (*D&C*) approach decomposes transcription factor binding sites into a family of motifs

Motifs discovery tools are designed to find unknown, relatively short sequence patterns located primarily in the promoter regions of the genomes (Tompa, Li et al. 2005). Because these searches are performed in a context of short signals embedded in high statistical noise, current tools tend to discard important number of samples that only weakly resemble the consensus pattern (Wade, Struhl et al. 2007). Moreover, because the consensus is a pattern from averaged DNA sequences, it often conceals subsets of these sequences that share particular sub-patterns, which might define distinct regulatory mechanisms (Browning and Busby 2004). Here we propose a method that does not disregard any preexisting motif-finding methods, but work on top of them, utilizing their intrinsic advantages.

We use a *divide & conquer* (*D&C*) approach that decomposes a transcription factor binding site motif into a family of patterns, termed submotifs, and then combine them, by a voting multi-classifier (Figure 3.1). First, we group DNA sequences using clustering methods (i.e. hierarchical (Gasch and Eisen 2002), substracting (Hering, Innocent et al. 2004) and fuzzy clustering (Bezdek 1998), but other methods can be used) and/or *cis*-promoter features constraints (e.g., genome location, orientation). Second, we encode these groups into submotifs using any of the above described PWM methods. Third, we combine them into a classifier, where the individual and the cooperative contribution of each submotifs are optimized using Genetic Algorithms (GA) (Gertz, Riles et al. 2005) to identify thresholds that increase the sensitivity to weak sites without losing specificity.

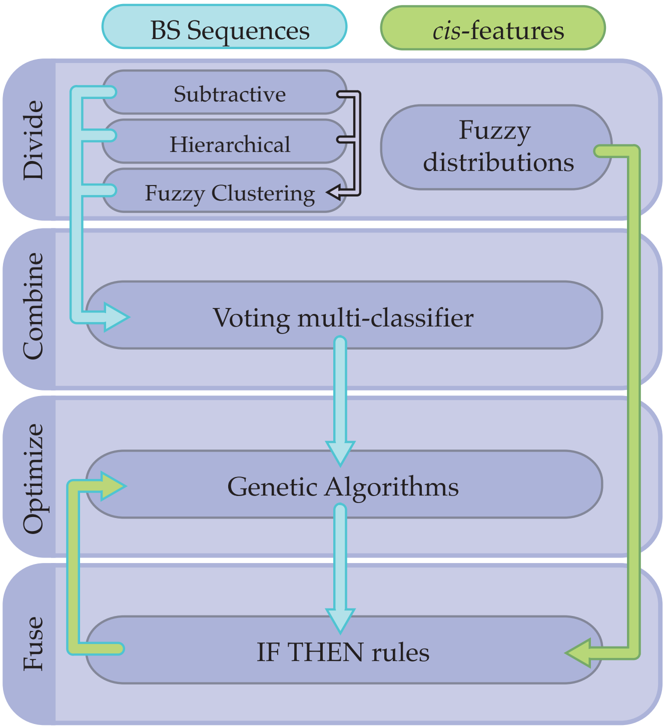


Figure 3.1 The Divide & Conquer Method.
D&C consists of four phases. 1) Divide BS sequences by using different clustering methods, including substractive, hierarchical and fuzzy, to generate submotifs. The submotifs are encoded into PWM by different methods including Consensus, Meme and AlignAce. Other *cis*-regulatory elements (e.g., distance from a TFBS to the RNAP) are divided into distinct data distributions. 2) Combine PWMs from submotifs into a multi-classifier, where each PWM classify a query sequence as a TFBS based on an individual threshold. 3) Optimize the global performance of the PWMs (CC or SCC) and eliminating their redundancy by using GA. (Other optimization methods can be used, see Methods). 4) Fuse different features into IF THEN rules, where the antecedents are the conjunction of the individual features, and the consequents are the prediction of a TFBS. To do so, all data distributions are converted into Fuzzy Sets and encoded following Fuzzy Logic theory to improve the interpretability of the system. (Other methods like Naïve Bayes can also be applied, see Methods).

Different clustering methods employed for the "*divide*" phase recover distinct position-dependent conserved patterns

The hierarchical clustering (HS) method (Gasch and Eisen 2002) subdivides the CRP binding sites into 7 subsets, and the subtracting clustering (SS) method (Hering, Innocent et al. 2004) subdivides it into 5 subsets. The number of clusters results from optimizing validity indices (see Methods). We found both coincidences and differences between the two clustering methods (Appendix A Figure 1). For example, the SS5 and the HS3 clusters exhibit high levels of coincidence (*p-value* < 5.90E-04), and thus, represent the same submotif (see also SS3 and HS4 (*p-value* < 2.60E-03), and SS2 and HS7 (*p-value* < 3.70E-03)). However, the SS1 cluster, describing 58 BSs, is split into two disjoint sets: the HS1 cluster, harboring 47 BSs (*p-value*=8.40E-04), and the HS2 cluster, with 31 BSs (*p-value* < 6.10E-03) (Appendix A Figure 1). These differences exhibit qualitative aspects of the submotifs. For example, SS1 encodes a general pattern that shows a balanced conservation between both CRP tandems. This is in contrast to the more specific patterns exhibited by the HS1 cluster, which emphasizes the conservation of the second tandem; and the HS2 cluster, which shows a more conserved first tandem repeat.

One of the major difficulties that arise when clustering short length sequences, as is the case of TFBS, is that crisp clustering methods assign sequences to single but unstable clusters (i.e., low bootstrap values). In other words, the clusters are highly overlapping because one sequence can belong to more than one group. To solve this problem we introduce a second grouping phase, based on fuzzy clustering (Bezdek 1998) that encodes memberships to multiple clusters (see Methods). The fuzzy approach combined with the hierarchical clustering (Figure. S1) preserves the original well-defined clusters (e.g., HS1 and HSF3 (*p-value*< 4.64E-10); HS4 and HSF2 (*p-value*<3.51E-14); and HS5 and HSF4 (*p-value*<=4.28E-09). Indeed, it recovers more informative submotifs (Appendix A Figure 1) as measured by the information content (IC, (Hertz and Stormo 1999)) (e.g., HSF8 (IC=13.91) includes sequences from HS1 (IC=10.68), HS6 (IC=10.22), and HS7 (IC=8.52)). We obtained similar results by combining the fuzzy and substracting clustering (Appendix A Figure 1D) (e.g., SSF3 (IC=11.84) and SSF4 (IC=12.35) include sequences of SS2 (IC=8.86)).

3.1.3 A multi-classifier based on submotifs overcomes limitations of a single consensus

The submotif approach provides a set of alternative and complementary models describing TFBS, which requires a coordinating strategy to allow their usage as predictors while screening DNA sequences. We integrate these models into a voting multi-classifier (Bauer and Kohavi 1999), where each model votes above an individual threshold that reflects the similarity of a querying sequence with the corresponding submotif. This simple concept exploits the specialization of each submotif to recognize particularities of its subset. However, this strategy can result in an excessive number of overfitted clusters, each supported by few observations (Ruspini and Zwir 2002). Moreover, several models can collaboratively produce a poor classification performance (Zwir, Shin et al. 2005). Therefore, we design a multiobjective genetic algorithm (GA) that optimizes thresholds for each model; considers the cooperation between them; and eventually, constrains the influence of any model (Figure 3.1).

We find that the use of submotifs increases the SCC on an average of 22.57%, obtaining an improvement up to 17% for AlignACE (i.e. 0.692 vs. 0.592); 43.7% for MEME (i.e. 0.682 vs. 0.475) and 22.69% for Consensus (i.e. 0.734 vs. 0.598). This enhancement is mostly due to submotifs sensitivity gain (average of 27.08%) and a slight increase of specificity (2.54%) (Table 3-2). Comparable improvements were obtained optimizing CC (

Table 3-3).

Table 3-2 CRP Single motif vs. submotif optimized by SCC performance comparison.

	Consensus			**MEME**			**AlignAce**		
	SCC	SP	SN	SCC	SP	SN	SCC	SP	SN
Single Motif	0.598	0.827	0.770	0.475	0.966	0.405	0.592	0.928	0.655
Non fuzzy clustering									
Subtractive	0.683	0.907	0.770	0.592	0.949	0.608	0.662	0.945	0.696
Hierarchical	0.677	0.907	0.764	0.639	0.936	0.682	0.665	0.933	0.716
Method Specific							**0.735**	0.925	0.804
Fuzzy clustering									
Subtractive	0.729	0.908	0.818	0.634	0.945	0.662	0.692	0.903	0.784
Hierarchical	**0.734**	0.919	0.811	**0.682**	0.938	0.730	0.665	0.938	0.709

Table 3-3 CRP Single motif vs. submotif optimized by CC performance comparison.

	Consensus			**MEME**			**AlignAce**		
	CC	SP	SN	CC	SP	SN	CC	SP	SN
Single Motif	0.516	0.827	0.770	0.475	0.966	0.405	0.592	0.928	0.655
Non-fuzzy codification									
Subtractive	0.642	0.925	0.730	0.599	0.949	0.608	0.659	0.945	0.696
Hierarchical Specific	0.633	0.913	0.750	0.638	0.952	0.649	**0.661**	0.935	0.730
Fuzzy clustering									
Subtractive	0.688	0.933	0.770	0.643	0.958	0.635	0.654	0.924	0.750
Hierarchical clustering	**0.706**	0.935	0.791	**0.677**	0.953	0.696	0.654	0.945	0.689

The results can be also evaluated by their complexity, when the GA also optimizes the number of submotifs. For example, the GA can build a multiclassifier using Consensus PWMs that employs only 5 of the original 8 submotifs derived from the hierarchical clustering, obtaining a performance improvement of the SCC by 13%. It can also reduce the usage of up to 2 subtractive submotifs from the original 7, obtaining an improvement of 23% when using the MEME method; and it can construct a multiclassifier based on 5 submotifs utilizing AlignAce subtractive submotifs, which still outperforms the single motif in 14% (Appendix A Figure 2). This suggests that *D&C* improves the classification performance even by employing simplified models.

3.1.4 *Cis*-features employed for the "*divide*" phase overcomes limitations of a single consensus

The interaction between a TF and the RNA polymerase (RNAP), a critical determinant of gene expression [Barnard, 2004 #337], can be represented by the distance between their binding sites (Cox, Surette et al. 2007; Elemento, Slonim et al. 2007). Moreover, gene expression data allows distinguishing between activated and repressed genes, which often correlate with the BS location relative to the transcription start site (TSS) (Browning and Busby 2004). Therefore, we explore the possibility of grouping binding site sequences using these *cis*-regulatory constraints.

To integrate the distance between CRP and RNAP binding sites into the motif models, we compile 136 reported CRP distances in RegulonDB (Salgado, Santos-Zavaleta et al. 2001) whose locations are between 110bp upstream the TSS and 10bp downstream. Then, we represent their distributions as histograms, and encode these distributions into fuzzy sets (Appendix A Figure 3). We identify three sets (i.e. *far*, *medium* and *close* distances) representing different activation distributions, and another three sets corresponding to repression distributions [Collado-Vides, 1991 #28]. We also encode the BS motifs (PWM) into a fuzzy set by using the scores as measurements of similarity with a prototype (Zwir, Huang et al. 2005). Then, we connected the fuzzy sets corresponding to the motif and the distance distributions using the fuzzy logic AND-operator, and encoded the resulting relationships into fuzzy IF-THEN rules (Cordon, Herrera et al. 2002). The consequents of these rules are determined by the CRP bound and not-bound classes constrained by a lower-bound threshold. We end up with 6 rules, resulting from the optimization of both the shape of the fuzzy sets (distance distributions) and the threshold of each rule (see Methods). The optimization process is carried out by a global fitness based on the performance of the whole set of rules, and the inference process allows firing of more than one concurrent rules (Appendix A Figure 4) (Cordon, Herrera et al. 2002).

The identified rules improve the prediction of CRP BS by 15% using the SCC criterion (Table 3-4). We also evaluate the influence of a more generic form of the interaction between a TFS and the RNAP by considering the complete set of TFSs reported in the *E. coli* genome (Salgado, Santos-Zavaleta et al. 2001). Even this less-specific distance information produces a 11.71% of SCC improvement over the single motif approach. This suggests that this procedure can be extended to other TFs even if the specific distances from their BS to the TSS are not available.

Table 3-4 CRP Single motif vs. *cis*-features classifiers *by SCC* performance comparison.

	Activators			**Repressors**			**Activators & Repressors**		
	SCC	SP	SN	SCC	SP	SN	SCC	SP	SN
Single Motif (SM)	0.589	0.783	0.806	0.550	0.837	0.708	0.560	0.809	0.750
Approx. distances & SM	0.656	0.783	0.871	0.617	0.898	0.708	0.625	0.838	0.787
CRP distances & SM	**0.710**	0.803	0.903	**0.664**	0.906	0.750	**0.644**	0.842	0.801

Finally, we encode the submotifs into fuzzy sets and integrate them with the distances as described above. The result is an optimized multi-classifier that employs 35 of the 48 possible rules that provides an improvement of 34.59% (i.e. 0.753 vs. 0.56) of the SCC with respect to the single motif (Table 3-5). Using the less-specific distances from all TFs (39 rules) still improves SCC by 33.50% (i.e. 0.747 vs. 0.56). Notably, these multi-classifiers that aggregate different kind of information outperforms our previous models, even while using general TF distances. These results explicitly avoids the strong assumption that the regulatory elements are independent features, thus, providing a more realistic representation of the encoded genome information (Barash 2003).

Table 3-5 CRP Single motif vs. Submotif & distances by SCC performance comparison.

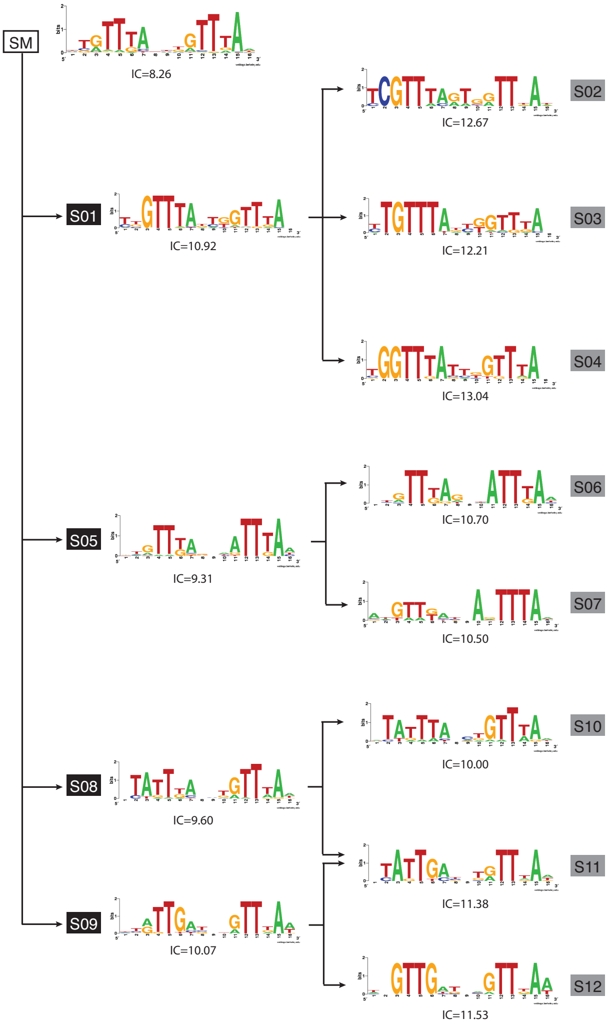
CRP	**Activators**			**Repressors**			**Activators & Repressors**		
	SCC	SP	SN	SCC	SP	SN	SCC	SP	SN
Single Motif	0.589	0.783	0.806	0.585	0.902	0.667	0.560	0.809	0.750
Approximated distances	**0.745**	0.874	0.871				0.747	0.894	0.853
CRP distances	**0.748**	0.843	0.903	**0.845**	0.929	0.917	**0.753**	0.913	0.838

3.1.5 *D&C* uncovers the PhoP regulon

We apply our approach to study genes regulated by the PhoP two component system well conserved through bacterial genomes, and which controls expression of a large number of genes that mediate adaptation to low Mg^{2+} environments, and/or virulence in several bacterial species including *Salmonella enterica, Shigella flexneri, Yersinia pestis, Erwinia carotovora, Neisseria meningitidis, Photorhabdus luminescens* and *Escherichia coli* (see (Groisman 2001) for a review). The PhoQ protein is a sensor for extracytoplasmic Mg^{2+} that modifies the phosphorylated state of the DNA-binding protein PhoP {Castelli, 2000 #330;Montagne, 2001 #331;Chamnongpol, 2003 #332}.

It has been proposed that the PhoP protein recognizes the direct hexanucleotide repeat (T/G)GTTTA separated by five nucleotides, which has been termed the PhoP box (Kato, Tanabe et al. 1999). Indeed, experiments carried out with the PhoP-activated *mgtA* promoter of *E. coli* demonstrated the critical role that certain PhoP box nucleotides play in *mgtA* expression, and that the purified PhoP protein and RNA polymerase were sufficient to promote *mgtA* transcription *in vitro* (Yamamoto, Ogasawara et al. 2002). However, there is uncertainty about what constitutes a PhoP binding site because many PhoP-regulated promoters do not conform to the *mgtA* promoter model (Lejona, Aguirre et al. 2003; Minagawa, Ogasawara et al. 2003). Moreover, the identification of PhoP-regulated targets is confounded by the fact that many genes are indirectly regulated by PhoP, which controls other two-component regulatory systems at the transcriptional (e.g., RstA/RstB) (Minagawa, Ogasawara et al. 2003), posttranscriptional (e.g., SsrB/SpiR) (J. Bijlsma and EAG, unpublished results), and posttranslational (e.g., PmrA/PmrB) (Kato and Groisman 2004) levels.

We use a dataset of known *E. coli* and *Salmonella enterica* PhoP BSs including those reported in the literature (Minagawa, Ogasawara et al. 2003; Groisman and Mouslim 2006) and our previous work (Zwir, Huang et al. 2005; Zwir, Shin et al. 2005). We identify 12 PhoP submotifs by applying the hierarchical fuzzy clustering (Figure. 2, Appendix A Table 1). The logo representation of the submotifs reveals several differences between them. For example submotif S05, including the *proP*, *ybjX* and *mig-14* BSs, has a strong pattern for the second tandem that differs from the canonical S01 submotif, including *mgtA* and *phoP* BSs, which harbors a strong conservation of both direct repeats. The PhoP submotifs are hierarchically organized into families by inclusion. For example, the sequences assigned to submotif S01, are also assigned to more specific submotifs S02, S03 and S04, which have particular commonalities (Figure 3.2).



```
proP:   TCGTTTAG---TCATTGAT  E. Coli
rsta:   TCGTTTAG---AGATTTAT  Salmonella
slyb:   TCGTTTAA---TGGTTAAT  Salmonella
yrbl:   TCGTTTAG---TTGTTTAA  Salmonella
nmpc:   TCGTTTAA---TGGTTGAG  Salmonella
irap:   CCGTTACG---TGGTTTAA  Salmonella

rstA:   TTGTTTAG---CGATTGAT  E. Coli
slyB:   TTGTTTAT---TGGTTGAT  E. Coli
yrbL:   TTGTTTAG---TTGTTTAA  E. Coli
ompT:   CTGTTTAT---TTGTTTAG  E. Coli
ybcU:   TTGTTTAG---TTGTTTAA  E. Coli
yeaF:   TTGTTTAA---ATGATTAA  E. Coli
nagA:   CTGTTTAT---CGGTGTAG  E. Coli
udg:    ATGTTTAA---CGGTTTAA  Salmonella
irap2:  TTGTTTTT---TGGTTTAT  Salmonella
ybjx:   TTGTTTAG---CGGTTTAC  Salmonella
naga:   CTGTTTAT---CGGCGTCG  Salmonella
pagP:   CTGTTTAT---TTGTTAAG  Salmonella
mgtc:   CTGTTTAA---TGTTTGAT  Salmonella

treR:   TGGTTTAT---TGGTTTAG  E. Coli
mgtA:   TGGTTTAT---TGGTTTAG  E. Coli
phoP:   TGGTTTAT---ATGTTTAC  E. Coli
ompX:   TGGTTTAT---TGGTTTAG  E. Coli
mgta:   TGGTTTAT---TGGTTTAA  Salmonella
phoP:   TGGTTTAT---CTGTTTAT  Salmonella
ybjx:   CGGTTTAC---CTGGTTAA  Salmonella
hila:   GGGTTTAT---TTTTTAAC  Salmonella
ompx:   CGGTTGAG---TCGTTGAA  Salmonella

proP:   TCGTTTAG---TCATTGAT  E. Coli
ybjG:   TTCTTTAA---TTATTTAA  E. Coli
hemL:   ATGTTTGA---GTATTTAA  E. Coli
proP:   AAGTTGAT---AAATTTAA  E. Coli
rsta:   TCGTTTAG---AGATTTAT  Salmonella
ugtl:   CGGTTGAG---CTATTTAC  Salmonella
pagc:   GTGTTTAG---GAATTTAC  Salmonella
mig-14:ATGTTTAG---GTATTTAA  Salmonella
pipd:   TTATTGAG---GTATTGAT  Salmonella
pagK:   CCATTTAT---ATATTTAA  Salmonella
orgb:   TTATTGAG---GCATTGAA  Salmonella

pmrD:   CCGTTGAT---GAGTTTAT  E. Coli
proP:   GTGTTTAG---GAATTTAC  E. Coli
pagc:   ATGTTTCC---TATTTTAA  Salmonella
mgtC:   AATAATAC---TAGTTTAA  Salmonella
ugtl:   ATGTTGTT---CAGTTTAT  Salmonella
virK:   AAGTTGAT---AAATTTAA  Salmonella

ybjX:   ATATTTCG---AAGTTAAT  E. Coli
crcA:   CTATTAAG---ATGTTAAT  E. Coli
yhiE:   CTATTTAC---CTGATAAC  E. Coli
nmpC:   TTATTTTT---CGGTTTAA  E. Coli
yeaF:   TTAATTAT---AAATTTAT  E. Coli
b2833: ATATATAA---CGGTTTAT  E. Coli
b1826: ATAGTTAG---CTGTTTAA  E. Coli
trs:    GTTTTTAA---CTGTTTAA  E. Coli
prop:   ATATTTAA---CTGTTAGG  Salmonella
pdgl:   ATTTTAAC---CTGTTTAA  Salmonella
pagc:   TTATTTAC---GTGTTTAA  Salmonella
yobg:   ACAGTTAC---TGGTTTAA  Salmonella

ompT:   ATATTGCT---CTGTTTAT  E. Coli
ybjX:   CTATTGAT---TGGTTAAG  E. Coli
pagK:   CCATTTAT---ATATTTAA  Salmonella
pipd:   TTATTGAG---GTATTGAT  Salmonella
orgb:   TTATTGAG---GCATTGAA  Salmonella
mgtc:   TTAGTGAC---CTGTTTAA  Salmonella
ybjx:   GTATTGAC---TGGTTAAT  Salmonella
virk:   CCATTGAT---CTGTTTAA  Salmonella
pmrd:   CTATTGCC---TTGTTTAT  Salmonella

pmrD:   CCGTTGAT---GAGTTTAT  E. Coli
proP:   AAGTTGAT---AAATTTAA  E. Coli
ompX:   TGGTTGAG---TTGTTGAA  E. Coli
ybjX:   TTGTTGAT---TCGTTGAA  E. Coli
yiaG:   TTGTTGTT---TTGTTAAA  E. Coli
virK:   ATGTTGTT---CAGTTTAT  Salmonella
virK:   TCGTTGCC---ACGTTTAA  Salmonella
ompX:   CGGTTGAG---TCGTTGAA  Salmonella
```

Figure 3.2 Families of submotifs describe the Phop regulon.
The hierarchical and fuzzy organization of the PhoP BSs is represented by logos, where the three nucleotides between the direct repeat tandems are omitted. Left panel denotes general, while right panel indicates specific submotifs. The root of the tree represents the single motif, while each branch emphasizes distinct nucleotide patterns. The support sequences for each specific submotifs (gray boxes) and their genomic source are listed on the right side. The information content of each submotif is displayed below the logos (i.e., the higher the more informative).

Both general (S01, S05, S08 and S09) and specific submotifs (S02-4, S06-7, S10-12) contribute to the classification performance. For example, S01 is a generalization of its dependent submotifs, however, it does not recover BSs for *Salmonella* genes like *iraP*, *nagA* and *ybjX*, which are detected by the more specific S02, S03 and S04 respectively. Similarly, the S05 submotif neither recognizes the BS for the *mgtC* and the *virK* genes of *Salmonella* nor the BS of the *ybjX* gene of *E. coli* , which are only recovered by the more specific S06 and S07 submotifs. This family of submotifs is necessary and a subset is sufficient to describe the PhoP regulon, since a leave-one submotif-out analysis reveals that the remaining submotifs are incapable of retrieving the binding site sequences that conform the omitted submotif (Table 3-6).

Table 3-6 PhoP submotifs crossvalidation.

		S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
		28	6	13	9	17	11	6	21	17	12	9	8
S01	**28**	-	-	-	-	7 *0.48*	3 *0.87*	0 *1.00*	0 *1.00*	7 *0.48*	1 *0.99*	0 *1.00*	3 *0.65*
S02	**6**	-	-	2 *0.28*	1 *0.55*	0 *1.00*	1 *0.63*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	1 *0.51*
S03	**13**	-	0 *1.00*	-	6 *0.00*	5 *0.14*	3 *0.30*	0 *1.00*	0 *1.00*	4 *0.34*	1 *0.92*	0 *1.00*	2 *0.43*
S04	**9**	-	0 *1.00*	4 *0.05*	-	1 *0.92*	0 *1.00*	0 *1.00*	0 *1.00*	2 *0.66*	0 *1.00*	0 *1.00*	0 *1.00*
S05	**17**	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	-	-	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*
S06	**11**	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	-	-	1 *0.63*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*
S07	**6**	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	-	0 *1.00*	-	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*
S08	**21**	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	-	-	-	-	0 *1.00*
S09	**17**	1 *0.99*	0 *1.00*	0 *1.00*	0 *1.00*	2 *0.95*	0 *1.00*	0 *1.00*	-	-	1 *0.97*	-	-
S10	**12**	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	-	-	-	1 *0.82*	0 *1.00*
S11	**9**	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	0 *1.00*	-	-	1 *0.82*	-	0 *1.00*
S12	**8**	1 *0.98*	0 *1.00*	0 *1.00*	0 *1.00*	2 *0.59*	0 *1.00*	0 *1.00*	0 *1.00*	-	0 *1.00*	0 *1.00*	-

We learn optimal configurations of submotifs, which are necessary and sufficient to describe the PhoP regulon. We obtain a multi-classifier that uses 9 of the 12 submotifs, by discarding two general (i.e. S05 and S09) and 1 specialized submotifs (i.e. S11). This multi-classifier improves both SCC, by 29% compared to the single motif results (i.e. 0.835 vs. 0.547), and CC by 23 % (i.e. 0.885 vs. 0.653). This improvement is due to the recovery of 25 BSs that were not detected by the single motif approach (i.e. 57 vs. 32 BS), at the expense of predicting only one more false positive. Moreover, we were able to capture more than one BS per gene (Figure 3.2), including those ones exhibiting low affinities but bound by the PhoP protein (manuscript in preparation, IZ et al.).

To analyze the influence of the interaction between the PhoP protein and the RNAP, we integrate the distances of activation sites between -90 bp upstream and 10 bp downstream the TSS into 3 rules. They provide an improvement of 21% of the SCC (i.e. 0.63 vs 0.83) compared to the single motif model (Appendix A Figure 5). Indeed, the incorporation of the submotifs results in a multi-classifier that employs 24 of the 36 possible rules (Figure 3.3), providing an improvement of 45%

of the SCC (i.e. 0.63 vs. 0.91) compared to the single motif. Interestingly, two sets of distances peak close to the TSS at characteristic positions of the typical class II activation (Browning and Busby 2004). The location of the closest set, which peaks at -34 bps from the TSS, suggests that the PhoP binding site completely overlaps the -35 hexamer. This configuration is often found in promoters lacking sequences with good matches to the -35 hexamer and bound by a different sigma^{70} subdomain (Browning and Busby 2004). The *far* set of distances may correspond to Class I activation, where the activator targets a sequence located upstream of the -35 hexamer and recruits RNAP by interacting with the α–CTD (Browning and Busby 2004).

We found that the PhoP submotifs do not discriminate between *E. coli* and *Salmonella* genomes, since few of them are built based on a prevalent genome (Figure 3.2). Similarly, the interaction of the PhoP and the RNAP BSs, exemplified by their distance, suggests that three associations between submotifs and distances (i.e., IF-THEN rules) are similar for both genomes (i.e., S1 and *close*, S8&9 and *close*; and S8&9 and *medium* (Figure 3.3). However, we identify one subset of rules that is *E. coli* specific (i.e., S5 and *close*) and three subsets that are *Salmonella* specific (S1 and *medium*; S5 and *far*; and S8&9 and *far*) (Figure 3.3B). This finding together with the existence of different activation classes (see above) indicate that there should be regulatory differences even in close related species, and these differences can be identified by incorporating other *cis*-features to the submotif classification. For example, the orthologous *pagP* and the *crcA* genes of *Salmonella* and *E. coli*, respectively, are recognized by two IF-THEN rules that differ in their distance (i.e., *medium* vs *close*). The PhoP BS at the *pagP* promoter region is located at 44 bp from the TSS, corresponding to typical class II activation. However, the PhoP BS at the *crcA* promoter is located at 32 bp (Minagawa, Ogasawara et al. 2003), corresponding to an atypical class II activation often found in promoters bound by a different sigma^{70} subdomain (data not shown).

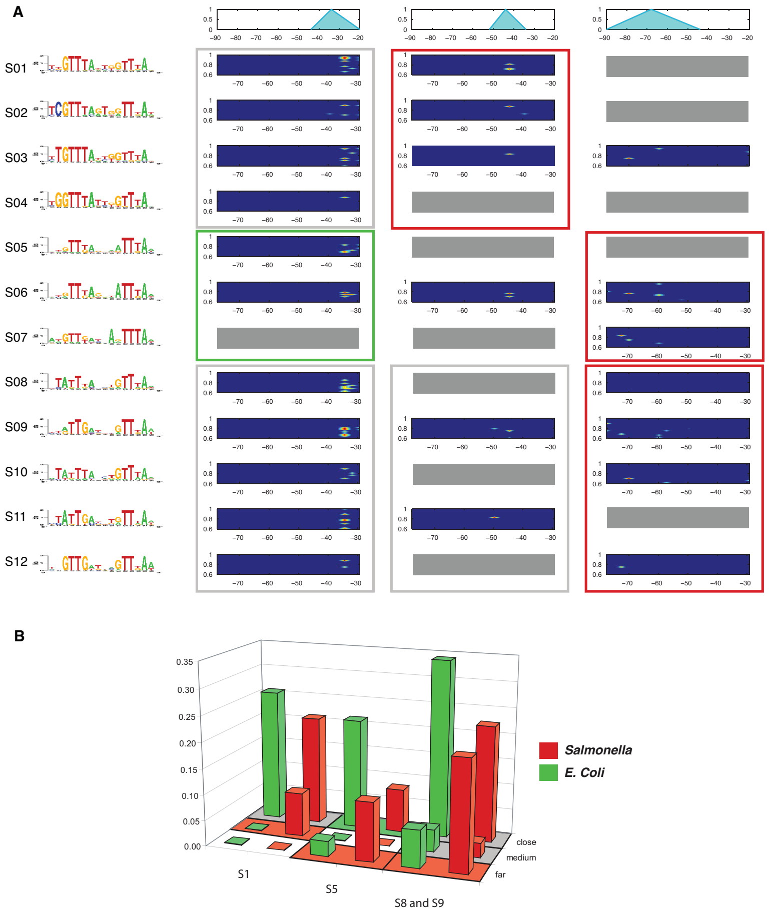
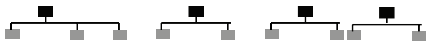


Figure 3.3 IF-THEN rules encompassing submotifs and distances between PhoP and RNAP BSs. A)Cells correspond to IF-THEN rules, where the vertical axis represents submotifs, and the horizontal axis represents the distances between PhoP and RNAP BSs, which are approximated by data distributions (*close*, *medium*, and *far* distances from left to right). These features are encoded into fuzzy sets and combined by using the fuzzy *AND* operator (i.e., Product), constituting the antecedents of rules. The consequent of a rule classifies PhoP BSs in the unit interval as a function of the antecedents (1: high, 0: low). One or more rules can be fired concurrently when this function exceeds each rule specific threshold. The isobars show the degree of membership of the training set to the rules (red: high; blue: low). B) Synthesis of the most representative IF-THEN rules that distinguish *E. coli* (green) from *Salmonella* (red) genes. The bars characterize the percentage of BSs recognized by each rule for each genome. We compact the submotifs into their most general families (i.e., S01, S05, and S08 & S09) for simplicity. Grey boxes highlight three subsets of rules similarly distributed in both genomes (i.e., S01 and *close*, S8&9 and *close*; and S8&9 and *medium*). Green boxes illustrate *E. coli* specific rules (i.e., S05 and *close*). Red boxes correspond to *Salmonella* specific rules (S01 and *medium*; S05 and *far*; and S8&9 and *far*).

3.1.6 Sensitivity analysis of parameters demonstrates the robustness of the *D&C* for the PhoP regulon

Although more accurate, our method is still a more complex model (i.e., more than one submotif) than previous tools used to identify TFBSs. However, the complexity is tempered by the possibility of optimizing the number of submotifs/rules and imposing higher boundary thresholds for them. To test this we incorporate different "constrains" to the size of the classifier (number of submotifs used) in the optimization process, impose a higher lower-bound threshold for each submotifs/rules, and learn an optimal set of distinct configurations of the multi-classifier (Figure 3.4). All 8 resulting configurations bettered the single motif performance, allowing the identification of at least 22 BSs that were otherwise undetected. Moreover, we found that increased complexity is partially correlated with accuracy. For example, the 7th configuration performes better than the 8th configuration that employs 2 less submotifs. Overall, all configurations suggests a remarkable robustness behavior of the proposed method.



CF	OO	SN	TP	TN	FP	FN	CC	SCC	#Sub	S2	S1	S3	S4	S6	S5	S7	S10	S8	S11	S9	S12	Min Th.
	SM		32	771	1	37	0.6537	0.5474	1													0.65
1	CC	0	57	770	2	12	0.8848	0.8359	12	●	●	●	●	●	●	●	●	●	●	●	●	0.6641
2		0.1	57	770	2	12	0.8848	0.8359	9	●	●	●	●	●		●	●	●			●	0.6641
3		0.2	52	772	0	17	0.8587	0.7776	7	●		●	●	●			●		●		●	0.6734
4		0.3	52	772	0	17	0.8587	0.7776	6		●	●		●			●		●		●	0.6734
5	SCC	0	58	767	5	11	0.8695	0.844	11	●	●	●	●	●	●	●	●	●	●	●		0.6614
6		0.1	56	768	4	13	0.8598	0.8203	9	●	●		●	●		●	●	●		●	●	0.6614
7		0.2	54	770	2	15	0.8586	0.7987	8		●			●	●	●	●	●	●		●	0.6523
8		0.3	54	768	4	15	0.842	0.7955	6	●	●			●			●	●			●	0.6613

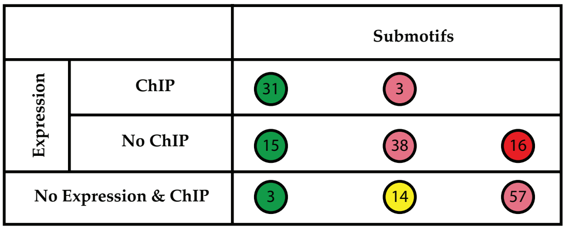
Figure 3.4. **Sensitivity analysis of *D&C* parameters.**
Submotifs are encoded into PWM by using the Consensus method, and employed to screen the PhoP training dataset (69 PhoP BSs and 772 BS of other TF as negative examples). A genetic algorithm learns which submotifs are used in the multi-classifier and their optimal thresholds (i.e. Configurations (CF)). One of the objectives of optimization (OO) is either the SCC or the CC measurements, and another objective is the number of submotifs employed (Pr). TP/TN stands for true positive/negative and FP/FN for false positive/negatives predicted values. #Sub indicates the number of submotifs effectively employed and columns S2 to S12 show the final configuration conformation. Min Th. Corresponds to the minimum learned threshold. SM shows the results obtained by the single motif.

The different metrics used to evaluate the configurations reveal distinct selection strategies. The SCC tends to equilibrate the unbalanced dataset conformed by 69 PhoP BS vs. 772 BS of other TF used as negative examples. It is biased towards the recovery of true BS, at the expense of predicting some negative samples as PhoP BS. In contrast, the CC equally weights both positive and negative samples, reducing the number of FP at the expense of increasing the number of FN. Therefore, the selection of the SCC or the CC metrics can drive the optimization process toward selecting the more general or more specific set of submotifs. For example, the 3rd configuration is optimized by the CC and only recovers specific submotifs without FP (Figure 3.4). On the other hand, the 5th to 8th configurations use the SCC and tend to include the more general submotifs, resulting in multi-classifiers that recover the highest number of PhoP BS employing less number of submotifs but have the highest number of FP (Figure 3.4).

3.1.7 Submotifs distinguish functional PhoP binding sites in Genome-wide analysis

We investigate the ability of our approach to analyze whole genome sequences by screening the regulatory regions of *Salmonella enterica* serovar Typhimurium LT2 to detect PhoP BSs. We compare our predictions with gene expression measured by microarray assays of wild-type and *phoP* mutated strains, and promoter occupancy measured by chromatin inmunoprecipitation assays (ChIP). We subdivided the experimental results into three subsets containing expressed genes harboring a significant peak in the regulatory region; expressed genes without that peak; and genes harboring a peak without exhibiting significant expression (Appendix A Table 2).

Functional binding sites should be present in genes harboring significant expression and ChIP scores (Figure 5). We detect PhoP BSs in 31 of the 34 genes comprising that subset. The remaining three genes have low peak scores (<0.38), which are only detectable in one of the three ChIP replicas, and probably their expression reflects read-through transcription (e.g., *ycfD*).



		Submotifs		
Expression	ChIP	31	3	
	No ChIP	15	38	16
No Expression & ChIP		3	14	57

Figure 3.5 Genome-wide analysis of *Salmonella* using PhoP submotifs.
Genes have been sorted into three categories based on the tiling expression array and ChIP experiments: 1) Genes with significant expression and ChIP peaks; 2) Genes with significant expression and without ChIP peaks; and 3) Genes without significant expression and ChIP peaks (Appendix A Table 2). We recognize functional BSs in most of the genes reported in 1). We also distinguish 15 functional BSs from genes without a ChIP peak in 2) (i.e., ChIP FN), which were reported as PhoP directly regulated genes ((Zwir, Huang et al. 2005; Zwir, Shin et al. 2005) Appendix A Table 2). Indeed, we do not detect submotifs in 54 genes (i.e., ChIP TN), where 11 genes are known to be indirectly regulated by PhoP (Groisman and Mouslim 2006) and 5 genes are part of expressed operons. Finally, we identify 3 genes with functional BSs in 3) (i.e., expression FN) that are directly regulated by PhoP (Zwir, Huang et al. 2005). We also find 14 genes with putative PhoP BSs (yellow) that requires further experimental validation. We do not detect submotifs in the remaining non-functional 57 genes.

Non-existing binding sites can be understood as expressed genes without significant ChIP peaks, which suggest an indirect regulation. We do not detect BSs in 54 of the total 69 genes showing the former constraints (Figure 5, Appendix A Table 2). Eleven of these genes, organized in three operons, are known to be indirectly regulated by PhoP (Groisman and Mouslim 2006). As expected, we do not detect submotifs in them. Similarly, we do not detect submotifs in 5 genes, where the head of the operon is functional, as well as in 38 remaining genes of this set. In contrast, we detect BSs 15 genes and consider these BSs as functional (Figure 5) and ChIP false negative results. Although this has led to the proposal that PhoP regulates some of these genes indirectly, via another regulatory protein(s) (Lejona, Aguirre et al. 2003), PhoP directly regulates 8 of these genes (Zwir, Shin et al. 2005). The remaining 8 genes are predicted to be regulated in the same fashion.

Non-functional binding sites can be also characterized by genes without significant expression but harboring ChIP peaks. We do not detected submotifs in 57 of these genes (Figure 5), which are presumably false positive ChIP results (Wade, Struhl et al. 2007). However, we identify submotifs in 3 genes, which are directly regulated by PhoP (Zwir, Huang et al. 2005; Zwir, Shin et al. 2005)and expression false negative results. Indeed, we identified 14 genes, which require a further experimental verification. This set includes 9 genes divergently located with respect to other expressed genes with functional BSs, which use to provide false negatives in this type of array implementation, and 3 genes with low expression, but containing several probes expressed at higher levels than 2 folds.

We detect several genes harboring more than one BS. These BSs primarily correspond to S05, S08 and S09 families of submotifs, often associated with laterally acquired genes in *Salmonella* (Figure 2). These BSs are rarely detected by the single motif approach (5%), but at least 19 of them are functional (manuscript in preparation IZ et al.). Thirteen of these binding sites were located at much shorter distances than the minimum length of a ChIP peak. Therefore, they become indistinguishable for this technique. Overall expression and binding data validates the sensitivity and specificity of the approach. Indeed, the use of submotifs in them improves the identification of the members of the PhoP regulon while reducing the false positives due to the limitations of the experimental assays.

3.1.8 Submotifs reflect distinct DNA physical constraints of the protein-binding interactions

It was suggested that the quality of binding constraints can be interpreted directly with respect to physical constraints imposed by the DNA-binding protein (Moses, Chiang et al. 2003). Therefore we study the different properties recovered by each submotif. Transcription factors recognize target DNA BS by hydrogen-bond interaction and sequence dependent conformation and deformability (Pedersen, Jensen et al. 2000; Goni, Perez et al. 2007). These attributes are given by the local geometry of di-nuclotides (*shift, slide, twist, roll, and tilt*); and structural properties (*bendability* –measures the major groove compressibility, that is considered as an indication of DNA flexibility [Pedersen 2000]; *propeller twist* – the angle between the planes of two aromatic bases, also shown to be related to the DNA flexibility (Pedersen, Jensen et al. 2000); *protein-induced deformability* –measures how easily di-nucleotides are deformed by proteins (Pedersen, Jensen et al. 2000); *stacking energy* – the strength with which planar aromatics bases interact, interpreted as the DNA meltability, or how easy the two DNA strands can be separated (Pedersen, Jensen et al. 2000)).

We employ 10 properties to analyze PhoP BS sequences and find that 7 submotifs show distinguishing values for up to 6 different properties (*p-value* < 7.1E-04), while 4 properties result uninformative (Appendix A Figure 6). For example, four position specific bendabiliy values are equivalent for all submotifs (i.e. low: positions 5 and 16; high: positions 7 and 18)(Figure 3.6A). However, high bendability values at position 6 distinguishes sequences from submotif S09; and low bendability at position 4 and low stacking energy at positions 2 and 4 characterize submotif S02 (Table 3-7, Figure 3.6B).).

Particularly, two examples illustrate these properties. The *yrbL* gene of *Salmonella* harbors an S02 submotif and is inserted in a promoter region with a clear pattern of bendability (Figure 3.6C). This BS overlaps the -35 box of the RNAP and its downstream region, spacer comprised between the -35 and -10 hexamers of RNAP, contains the sequence AA-TT-TTT that is easy bendable in an intrinsically narrow minor groove (Liu, Tolstorukov et al. 2004). Indeed, its upstream region contains a clear pattern of bendability (Figure 3.6C). Therefore, it is not surprising that the PhoP BS in this promoter contains the –CG- sequence that resists such bending (Liu, Tolstorukov et al. 2004). In addition, the *rstA* gene of *Salmonella* exhibits an S02 submotif, which is inserted in a non-stable promoter region: it harbors a C+G rich region upstream the -10 hexamer of the RNAP BS; lacks an extended -10 element; overlaps of the -35 hexamer of the RNAP; and presents a discriminator sequence (Haugen, Berkmen et al. 2006) (Figure 3.6D). Not surprisingly, the BS exhibits a differential low stacking region at position 3 and 4, which characterizes stable DNA sequences (Table 3-7).

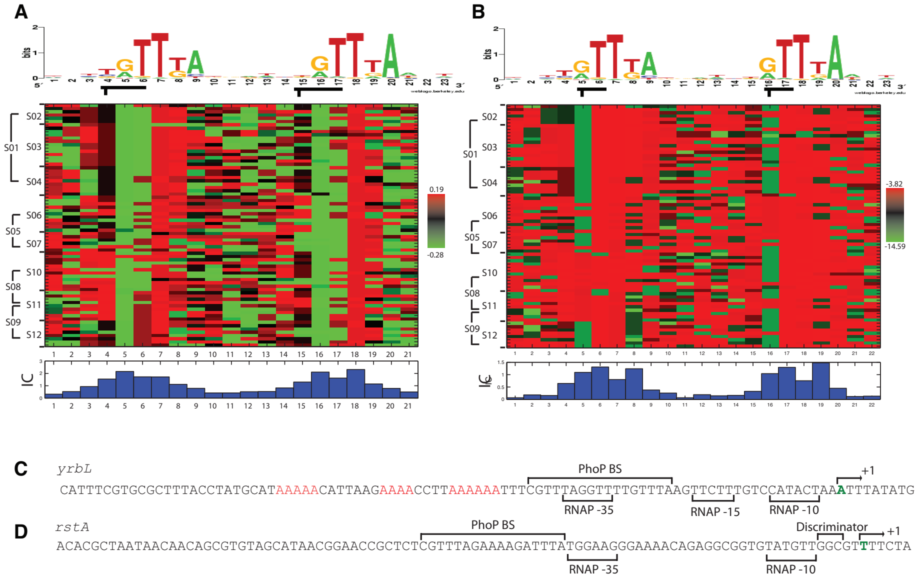


Figure 3.6 Phop submotifs characterized by the physical properties of DNA-binding protein interaction.

A) Heatmap representing the bendability of the PhoP submotifs (red: high, green: low) calculated based on tri-nucleotides bending propensity deduced from DNase I digestion (Brukner, Sanchez et al. 1995). Columns are aligned with respect to the logo, where column *i* corresponds tri-nucleotide values beginning at position *i*. All submotifs have similar values for columns 5, 7, 16 and 18. The other columns indicate differences among submotifs. For example, column 6 exhibits high values of bendability for S09 and low values for the remaining submotifs (*p-value*: 1.48E-07). And, column 15 shows lower bendability values for S07 than the other submotifs (*p-value:* 8.35E-15). The information content is shown at the bottom of the chart. B) Heatmap (green: low; red: high staking energy) representing the stacking energy of the PhoP submotifs calculated for every di-nucleotide based on physical scales reported in (Baldi, Chauvin et al. 1998). Columns 5 and 16 show distinguishable high values for submotif S06 (*p-value*: 5.27E-07) and S08 (*p-value*: 1.0E-35), respectively. C) PhoP BS (S02) of the *Salmonella yrbL* gene exhibits low bendability at column 4 (a)) and is inserted in a high bendable promoter region. -10 and -35 hexamers, -15 sequence of RNAP, +1, PhoP BS, and high bendability upstream region (red) are indicated. D) PhoP BS (S02) of the *Salmonella rstA* gene exhibits a low stacking region at columns 3 and 4 (b)) and is inserted in a non-stable promoter region. -10 and -35 hexamers, discriminator sequence of RNAP, +1, and PhoP BS are indicated.

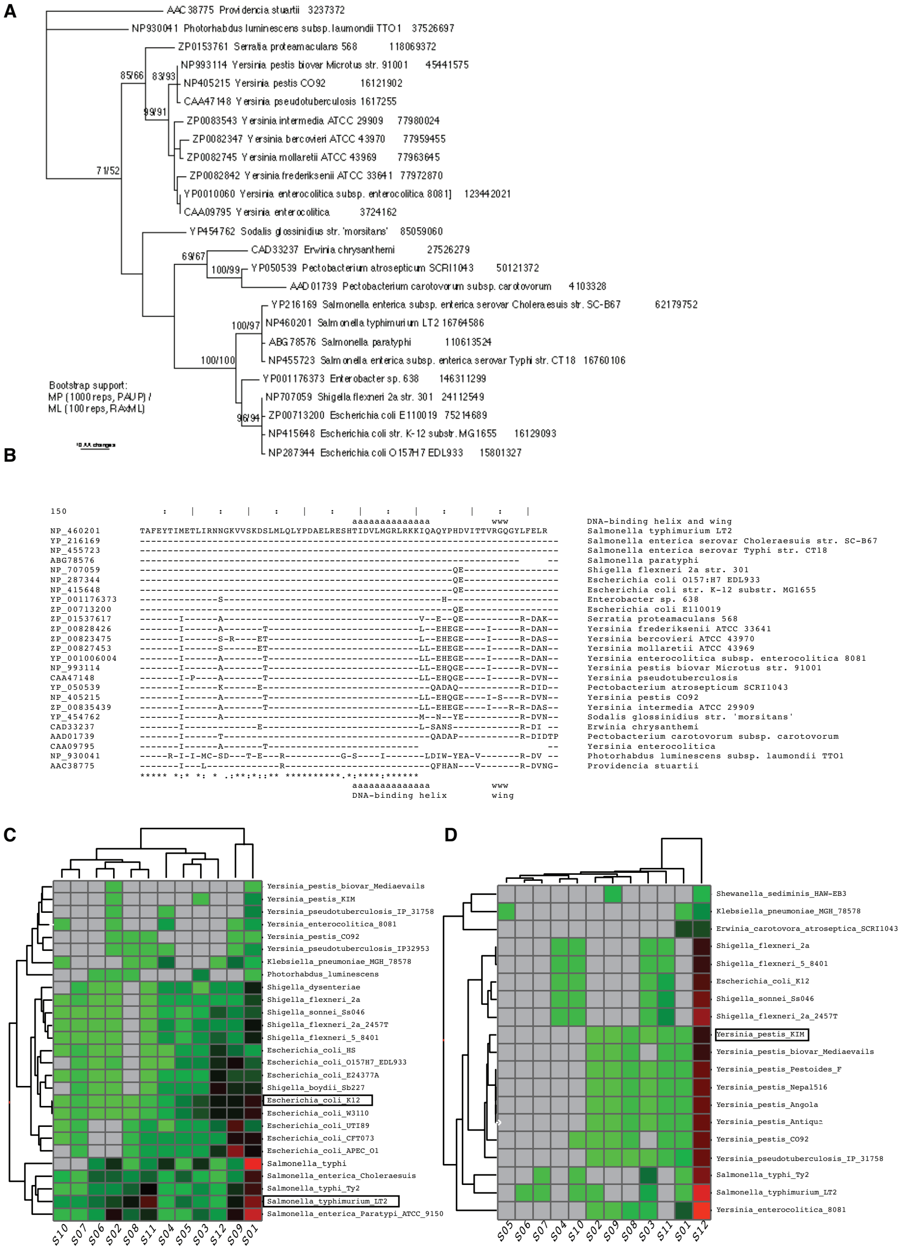
Table 3-7 Distinguishing physical properties of PhoP submotifs

	Sequence	Position	Property	Value	p-value
S02	--**TC**GTTTAG---TGGTTTAT--	D3	Stacking Energy	Low	3.60E-04
	--T**CG**TTTAG---TGGTTTAT--	D4	Stacking Energy	Low	4.74E-05
		T4	Bendability	Low	1.90E-04
S04	--**TG**GTTTAT---T-GTTTA---	D3	Propeller Twist	High	5.64E-06
			Protein Deformability	High	1.90E-12
			Shift	Low	7.10E-04
	--T**GG**TTTAT---T-GTTTA---	D4	Rise	High	2.81E-10
			Shift	High	6.12E-06
		T4	Bendability	Low	1.90E-04
S06	--TTGTTTAG---GT**AT**TTAA--	D16	Stacking Energy	High	5.27E-07
			Shift	Low	4.26E-08
S07	--ATGTT-AT----AGTTTAA--	T15	Bendability	Low	8.35E-15
	--ATGTT-AT----A**GT**TTAA--	D16	Propeller Twist	Low	4.98E-05
S08	--AT**AT**TTAC---CTGTTTAA--	D5	Stacking Energy	High	1.00E-35
			Propeller Twist	Low	4.55E-15
S09	--TTATTGAT---TTGTTTAA--	T6	Bendability	High	1.48E-07
	--TTAT**TG**AT---TTGTTTAA--	D7	Propeller Twist	High	2.71E-10
			Protein Deformability	High	3.03E-11
			Shift	Low	1.57E-06
	--TTATT**GA**T---TTGTTTAA--	D8	Rise	High	1.75E-09
			Shift	High	7.31E-05
S10	--CT**AT**TGAT----TGTTTA---	D5	Shift	Low	3.00E-04

3.1.9 Submotifs provide a model of evolution of the PhoP regulon

The evolution of the regulon can be devoted to 1) the evolution of the regulatory protein, and 2) the evolution of the regulatory machinery by which regulators regulate its target genes (i.e., orthologous and non-orthologous), which in transcriptional regulation may correspond to gain and loses of BSs. It is known that Two-component systems like PhoP-PhoQ can evolve slowly (Alm, Huang et al. 2006). Particularly, we study the evolution of the DNA binding domain of the PhoP regulator and found that is well conserved in gama enterobacterias (Figure 3.7A). This raises the question: how does a conserved protein domain recognizes its target sequences along the phylogenetic evolution, given that there exists a variety -but finite— set of submotifs identified in *E. coli* and *Salmonella* (Figure 3.7B). We hypothesize that computationally found submotifs would be biologically significant if they constitute a model of evolution, where they differentially evolve and do not completely disappear.

We address this question by studying the evolution rate of the submotifs (Moses, Pollard et al. 2006). It was theoretically suggested that the rate of evolution at each position in a motif is a function of the frequencies in the position weight matrix (Moses, Chiang et al. 2003). We first sought to verify that functional non-coding regions from submotifs evolve more slowly than 'background sequences' (Moses, Pollard et al. 2006). To do so, we compared the rate of evolution of the different submotifs to that of the promoter regions in which they were found (Figure 3.8). Representative plots substitutions per site and information content reveal a correspondence between positions of high information content and slower rates of evolution (Figure 3.8A-C). We found that different submotifs have distinct rates of evolution (8AB), which would be concealed by just looking at a single motif (8C). These differences are increased when the submotifs are evaluated in distinct backgrounds. For example, S01 exhibits a lower rate of evolution than S05 (Figure 3.8D-E), when the background is represented from promoter regions of laterally acquired genes. This allows us to predict that some submotifs have more chances to be present in other genomes than others, and thus, to keep track of the direction of the evolution. For example, we predict that the S05 family of submotifs has more chances of disappear than S01. Again, this prediction would be concealed by just looking to a single motif (6F).



```
150                 :    |    :    |    :    |    :    |    :    |    :    |    :    |
                                                      aaaaaaaaaaaaaa           www                    DNA-binding helix and wing
NP_460201       TAFEYTIMETLIRNNGKVVSKDSLMLQLYPDAELRESHTIDVLMGRLRKKIQAQYPHDVITTVRGQGYLFELR       Salmonella typhimurium LT2
YP_216169       ---------------------------------------------------------------------------     Salmonella enterica serovar Choleraesuis str. SC-B67
NP_455723       ---------------------------------------------------------------------------     Salmonella enterica serovar Typhi str. CT18
ABG78576        --------------------------------------------------------------------     --     Salmonella paratyphi
NP_707059       --------------------------------------------------------QE-----------------     Shigella flexneri 2a str. 301
NP_287344       --------------------------------------------------------QE-----------------     Escherichia coli O157:H7 EDL933
NP_415648       --------------------------------------------------------QE-----------------     Escherichia coli str. K-12 substr. MG1655
YP_001176373    --------------S---------------------------------------H--------------------     Enterobacter sp. 638
ZP_00713200     --------------------------------------------------------QE-----------------     Escherichia coli E110019
ZP_01537617     -------I------A-----------------------------------V--E--QE----------R-DAK--     Serratia proteamaculans 568
ZP_00828426     -------I------A-------T---------------------------L--EHEGE---I-----R-DAN--      Yersinia frederiksenii ATCC 33641
ZP_00823475     -------I------S-R----ET---------------------------L--EHEGE---I-----R-DAS--      Yersinia bercovieri ATCC 43970
ZP_00827453     -------I------S------ET---------------------------LL-EHEGE---I-----R-DAN--      Yersinia mollaretii ATCC 43969
YP_001006004    -------I------A-------T---------------------------LL-EHEGE---I-----R-DAN--      Yersinia enterocolitica subsp. enterocolitica 8081
NP_993114       -------I------A-------T---------------------------LL-EHQGE---I-----R-DVN--      Yersinia pestis biovar Microtus str. 91001
CAA47148        -------I-P----A-------T---------------------------LL-EHQGE---I-----R-DVN--      Yersinia pseudotuberculosis
YP_050539       -------I------K------E------------------------------QADAQ-----------R-DID--    Pectobacterium atrosepticum SCRI1043
NP_405215       -------I------A-------T---------------------------LL-EHQGE---I-S---R-DVN--      Yersinia pestis CO92
ZP_00835439     -------I------A-------T---------------------------LL-EHEGE---I---R-R-DAS--      Yersinia intermedia ATCC 29909
YP_454762       -------I------A-----------------------------------M--N--YE----------R-DVN--     Sodalis glossinidius str. 'morsitans'
CAD33237        -------I-------------E----------------------------L-SANS------------R-DI --     Erwinia chrysanthemi
AAD01739        -------I------T-------------------------------------QADAP-----------R-DIDTP     Pectobacterium carotovorum subsp. carotovorum
CAA09795        -------I------A-------T------------------------                          --     Yersinia enterocolitica
NP_930041       -----R-I-I-MC-SD---T-E---R----------G-S----I-------LDIW-YEA-V-------R-DV --     Photorhabdus luminescens subsp. laumondii TTO1
AAC38775        -------I---L-------------R-------------------------QFHAN---V-------R-DVNG-      Providencia stuartii
                ***** *:* *: * .:**:*::** **********.*:****:******
                                                      aaaaaaaaaaaaaa           www
                                                      DNA-binding helix        wing
```

Figure 3.7 Evolution of the PhoP-PhoQ two component system analyzed by PhoP submotifs
A) Phylogenetic tree indicating the evolution of the PhoP protein in gama enterobacterias. B) Sequence conservation of the PhoP protein, individualizing the DNA-binding domain. C) Distributions of submotifs along promoter regions of genes orthologous of PhoP regulated genes in *E. coli* and *Salmonella*. D) Idem c) considering orthologous of PhoP regulated genes in *Yersinia Pestis KIM*.

To examine the occurrence of the PhoP submotifs among different species, we first examine the intergenic regions corresponding to the orthologous genes of the PhoP training dataset across the gamma enterobacteria (p-value< 1E-5, Fig.) using *E. coli* and *Salmonella* as origins (Janky and van Helden 2007). We observe that not all submotifs are present in all species and that their conservation is mostly correlated with the phylogenetic distance (Figure 3.7FC). Presumably, the decreasing number of BS in the more distant species is caused by the dynamic interchanging of the laterally acquired genes, which are of the sources of the diversity in the BS motifs (Figure 3.7C). However, there are differences even within strains of the same specie. For example, *E. coli* UTI89, CFT073 and APEC 01 do not contain submotifs S02 and S06; *E. coli* 157H7 harbors these submotifs but lacks submotifs S08 and S10; and K12 recognizes all of them. Moreover, different distributions of submotifs are also present in close related species where all submotifs are recognized (e.g., *E. coli* and *Salmonella*). For example, a subset of *E. coli* and *Salmonella* genomes effectively recognize submotifs S01 and S09, but they differ in the recognition of S03, S05 and S12. And, within the *Salmonellas*, the *Salmonella* typhi also exhibits a completely distinct pattern (Figure 3.7C). Curiously, all *Shigellas* lack the submotif S08, but still recognize most of the other submotifs. As predicted by the high rate of evolution (Figure 3.8B and E), more distant species like *Yersinias* lack a complete family of submotif (i.e., the S05, S06, and S07 submotifs). Moreover, one family of submotifs with low rates of evolution (Figure 3.8A and D) is the only one just preserved (i.e., the S01, S02 and S03 submotifs) in orthologous genes in *Yersinia* KIM. This suggests that the occurrence of the motifs can be modular, and a family of binding motifs instead of a single sequence is sensitive to changes in evolution.

Interestingly, the submotifs are not completely lost through evolution since non-orthologous genes regulated by PhoP in *Yersinia* KIM (JCP et al. manuscript in preparation) exhibit submotifs S08-S12, but still lack the S05-S07 submotifs as the other *Yersinias* (Appendix A Figure 7). Particularly, the non-disjoint S08-S09 submotifs are specialized. For example, the fuzzy S11 submotif, containing sequences in both S08 and S09, disappears (Figure 3.7D). Moreover, the general S08 submotif is replaced by it most crisp derivative S10, while S9 and its crisp derivative S12 are both preserved. This suggests that the uncovered submotifs even harboring different rates of changes are conserved through evolution. It is noteworthy that all of these differences and conservations have been observed in species where the PhoP protein has not significantly change (Figure 3.7B) (E-value <1E-5).

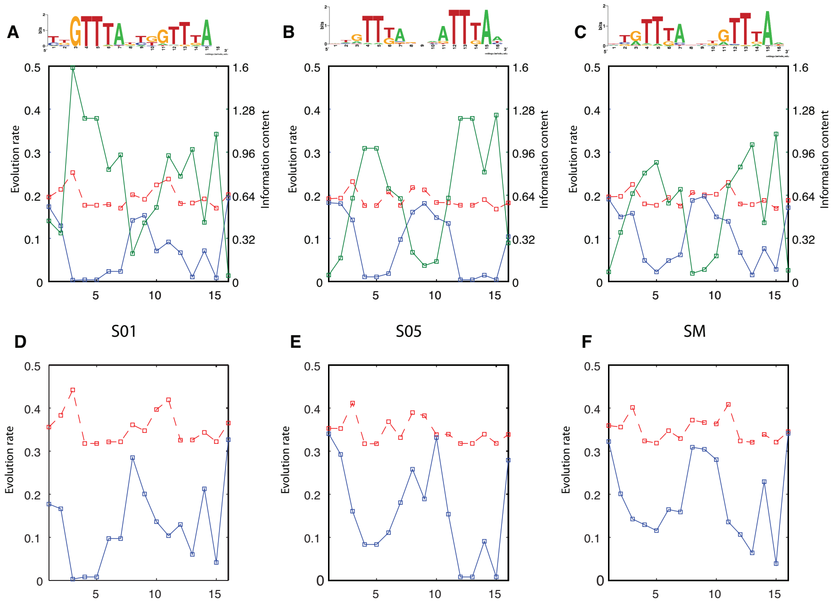


Figure 3.8 Submotifs as a model of evolution in PhoP BSs.
A-C) Plots variations in the rate of evolution at each position in the PhoP BSs using the HB model (blue), compared to variations in a *Salmonella* background (nucleotide distributions from promoter regions) using the HKY model (red), and information content (green) corresponding to submotifs S01 and S05, and single motif. The motifs evolve slower and at different rates (e.g., S01 vs S05) than the background (Moses, Pollard et al. 2006), because of the purifying pressure, and are inversely correlated to the information content. D-E) Idem A-C) but in a *Salmonella* background estimated from promoter regions corresponding to laterally acquired genes (i.e., AT rich background). Submotif S01 follows a similar to A) distribution, while S05 and single motif exhibit higher rates of evolution than B-C) (i.e., closer to the background). Submotif S05 shows a higher rate of evolution than S01, and thus, a lower chance of conservation. These differences are concealed by analyzing a single motif.

3.2 Methods

3.2.1 Clustering TFBS sequences.

Hierarchical clustering (Gasch and Eisen 2002): We transform nucleotides from BS sequences into dummy variables (Everitt and Der 1996) and calculate the Euclidean distance matrix to create a dendrogram by employing the single linkage method (i.e., nearest-neighbor). We use the inconsistency index implemented in the Statistic Toolbox of Matlab (V6.0) to detect the clusters that maximize the similarity.

Subtractive clustering (Hering, Innocent et al. 2004): This method iteratively selects submotifs that exhibit the highest density of sequences recovered by a PWM tool (e.g. Consensus, MEME, AlignAce). The retrieved true positives sequences (i.e. those above a threshold optimized by the SCC/CC measures) are removed from the training set to conform a cluster. This process is exhausted while the used measure is above 0.5.

Fuzzy c-means: We evaluate the complete set of TFBS sequences using the submotifs obtained the subtractive or hierarchical clustering methods (submotifs are previously encoded into PWM by using Consensus, MEME or AlignACE). Then we conform a new database where each sequence is represented by the obtained scores. This representation of the TFBS is the input to the traditional Fuzzy c-means algorithm (Bezdek 1998; Gasch and Eisen 2002): (i) Initialize $L_0 = \{\overline{V}_1,..,\overline{V}_c\}$; (ii) while (*s*<*S* and $\|L_s - L_{s-1}\| > \varepsilon$), where *S* is the maximum number of iterations; (iii) calculate the membership of U_s in L_{s-1}as in (equation (3)); (iv) update L_{s-1} to L_s with U_s and $\overline{V}_i = \sum_{k=1}^{n} \mu_{ik} x_k / \sum_{k=1}^{n} \mu_{ik}$; (v) iterate.

Xie-Beni validity index (Bezdek 1998): The minimization of this index through different number of clusters (i.e., $c = 2$ to $c = \sqrt{n}$) detects compact representations of Fuzzy c-means partitions:

$$XB(U,L) = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c} u_{i,k}^2 \left\| x_k - \overline{V}_i \right\|^2}{n\left(\min_{i \neq j} \left\{ \left\| \overline{V}_i - \overline{V}_j \right\|^2 \right\} \right)}$$

3.2.2 Voting multi-classifier

The set of submotifs are encoded as PWMs, where each one classifies a query sequence as positive TFBSs if the corresponding score is above a threshold. We use a single vote strategy, where all PWMs vote but one positive classification is sufficient to predict a TFBS. This process is analogous to a hierarchical Naïve Bayes (Mitchell 1997): $v_{MAP} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$ where v_{MAP} (maximum a posteriori probability) denotes the target value output by the Naïve Bayes classifier and v_j, a_i correspond to the features and attributes or variables, respectively.

3.2.3 Fuzzy membership functions.

They can be viewed as approximation of data distributions, where the degree of matching in the [0,1] scale is calculated using triangular functions (Klir and Folger 1988):

$$\mu(x) = \begin{cases} 0 & \text{if } x < a_0 \text{ or } x > a_2 \\ (x - a_0)/(a_1 - a_0) & \text{if } x < a_1 \\ (a_2 - x)/(a_2 - a_1) & \text{if } x > a_1 \\ 1 & \text{otherwise} \end{cases} \qquad (1)$$

where $a_0,...,a_2$ are learned from the projection of the histograms onto the variable domains by simple regression and minimum squared methods (Sugeno and Yasukama 1993; Everitt and Der 1996). This process is analogous to fitting histograms to a distribution, and assigning probability values based on a density function. Our approach, however, adopts a distribution-independent and non-parametric fitting process by projecting data (Sugeno and Yasukama 1993; Everitt and Der 1996) into triangular functions. We employ the Matlab Fuzzy Logic Toolbox version V2.2.5.

3.2.4 Fuzzy IF-THEN rules

We encode the PWM scores and distances between TFBS and RNAP into fuzzy sets (see above). Given a dataset $X=\{x_1,...,x_n\}$, the feature that characterizes it can be best described as a set $\mu_1(X)=\{d_{11}/x_1,...,d_{1n}/x_n\}$, where $\{d_{11},...,d_{1n}\}\in\{0,1\}$ in classical set theory and $[0,1]$ in fuzzy set theory. These fuzzy values represent the degree of matching between an observation of the dataset and a fuzzy set. The degree of matching is defined in the unit interval and can be obtained from evaluating the membership function of the corresponding fuzzy set (see above). Then, given $\mu_2(X)=\{d_{21}/x_1,...,d_{2n}/x_n\}$ and the *product* (Klir and Folger 1988; Bezdek 1998) as an AND operator, we define the expression corresponding to the antecedent of the rule:

$$\begin{aligned}\mu_1(X)\,AND\,\mu_2(X) &= \mu_1\cap\mu_2 = PROD(\mu_1,\mu_2)\\ &= \{PROD(d_{11},d_{21})/x_1,...,PROD(d_{1n},d_{2n})/x_n\}\end{aligned}$$

Then, a fuzzy IF-THEN rule is defined as $IF\ \ \mu_1(X)\ \ AND\ \ \mu_2(X)\ \ THEN\ \ C$, where C predicts a TFBS. We use a defuzzify-then-combine strategy (Berenji and Khedkar 1992), where each rule is fired if it overcomes its own learned threshold. The rules are combined by using the *maximum* operator. Again, this process can be implemented as a Bayesian classifier (see above).

3.2.5 Performance measurements

Correlation coefficient (CC): This measure is based on the Pearson product-moment coefficient of correlation that indicates the relation between predicted and observed values, and is suitable for balanced datasets:

$$CC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)\times(TN+FN)\times(TP+FN)\times(TN+FP)}}$$

where *P= positive, N=negative, T = true and F=false* (Benitez-Bellon, Moreno-Hagelsieb et al. 2002).

Standardized correlation coefficient (SCC) : The standardized version considers the magnitude of the positive and negative examples, resulting an appropriate choice where the dataset presents an unbalanced number of positive and negative examples. It extends the CC (equation(X)) replacing its parameters:

$$TP' = \frac{TP.100}{TP+FN} \quad TN' = \frac{TN.100}{TN+FP} \quad FP' = \frac{FP.100}{TN+FP} \quad FN' = \frac{FN.100}{TP+FN}$$

3.2.6 Genetic algorithms

This method learns optimal thresholds for PWM and IF-THEN rules. It represents sets of random solutions into data structures termed chromosomes and iteratively searches for their optimal values (Gertz, Riles et al. 2005). To do so, it systematically evaluates the solutions according to a fitness function (CC and SCC measures), and continues searching by applying genetic operators (i.e. crossover and mutation). We use the Matlab implementation (i.e. Genetic Algorithm and Direct Search toolbox, Version 2.1), which supports real codification of the chromosomes. Other optimization methods can also be used with lower-equal performance (i.e. Matlab Optimization Toolbox V3.1.1)

Optimizing PWM multi-classifiers. Each allele in the chromosome is implemented as a tupple, where the first element represents the presence/absence of a submotif (i.e., >0.5 or <0.5), and the second element is its threshold. We employ the "Max Min arithmetical" crossover (Herrera, Lozano et al. 1995), which given two solutions C_v and C_w to be crossed generates four offspring and picks the one with best fitness:

$$O_1 = aC_w + (1-a)C_v \qquad O_2 = aC_v + (1-a)C_w$$
$$O_3 \;\; with \;\; o_{3i} = \min\{c_{vi}, c_{wi}\} \quad O_4 \;\; with \;\; o_{4i} = \max\{c_{vi}, c_{wi}\}$$

where $a \in [0,1]$ is chosen randomly following an uniform distribution. The mutation operator switches the presence or absence of a submotif with probabilities p=0.05 and p=0.005, respectively, in the first element of the tupple. Indeed, it increases or decreases the corresponding threshold in the second element of the tupple up to 10% of its value following a uniform distribution. We left the default values for the remaining parameters

Optimizing IF-THEN rules. The distance distributions between BS and RNAP are encoded as triangular membership functions (see above) $D = \{(a_{i0}, a_{i1}, a_{i2})...(a_{n0}, a_{n1}, a_{n2})\}$. Each allele in the chromosome is implemented as a 3-tupple $(th, \partial a_0, \partial a_2)$, where *th* is the threshold of the rule, ∂a_0 and ∂a_2 are correction factors applied to the left and right extremes of the triangular membership function $(a_0 + \partial a_0, a_1, a_2 + \partial a_2)$, respectively. We employ the "Max Min arithmetical" crossover (Herrera, Lozano et al. 1995). The mutation operator randomly stretches or compacts the distances between points (a_0, a_1) and (a_1, a_2) up to 50% of the membership function range. Indeed, it increases or decreases the corresponding threshold (*th*) up to 10% of its value following a uniform distribution. We left the default values for the remaining parameters.

3.2.7 Rate of evolution for binding sites

We employ the model of Halpern and Bruno (Halpern and Bruno 1998) and follow the procedures as indicated in (Moses, Chiang et al. 2003): calculate the position-independent mutation matrix (*Q*) employing the HYPHY package (Pond, Frost et al. 2005) by learning the HKY85 model based on a set of 1000 sequences of 19bp length belonging to non-coding regions of *Salmonella.* We also learn this model using 19 bp length sequences belonging to promoter regions of laterally acquired genes (i.e., AT rich regions). The rate of evolution (*R*) from base *a* to base *b* for the frequencies of the submotifs *f* at position *p* is:

$$R_{pab} = Q_{ab} x \frac{\ln\left(f_{pb}Q_{ba} / f_{pa}Q_{ab}\right)}{1 - f_{pa}Q_{ab} / f_{pb}Q_{ba}}$$

The evolution rate for the submotif is defined as $K_p = \sum_a \sum_{a \neq b} f_{pa} R_{pab}$ and the background evolution rate as $K_p = \sum_a \sum_{a \neq b} f_{pa} Q_{ab}$.

3.2.8 Gamma enterobacteria othologs

Given a gene or a list of genes from a query organism sequences (*E. coli*, *Salmonella* and *Yersinia Pestis KIM*), and a reference taxon (gamma enterobacteria) we return the orthologous of the query gene(s) in all the organisms belonging to the reference taxon (Janky and van Helden 2007) (http://rsat.ulb.ac.be/rsat/get-orthologs_form.cgi)

3.2.9 Statistical significance

The coincidence between submotifs is evaluated by using the hypergeometric distribution that gives the chance probability (i.e. probability of intersection PI) of observing at least *p* candidates binding sites from a submotif V_i within another submotif V_j of size *n*:

$$PI(V_{i,j}) = 1 - \frac{\sum_{q=0}^{p} \binom{h}{q} \binom{q-h}{n-q}}{\binom{g}{h}}$$

where *h* is the total number of elements within V_i and *g* is the total number of binding sites, such that the lower the *p-value* the better the association (Tavazoie, Hughes et al. 1999).

P-values for the physical properties are calculated by the function stepwise implemented in the Statistics Toolbox of Matlab (V 6.0).

3.2.10 Microarray analysis

The experiments were conducted in triplicates to determine the error due to technical aspects of the process. Systematic error (Nadon and Shoemaker 2002) was treated by a the Moderated t-Test (Smyth 2004), which is similar to the Student's t-Test in that it is used to compare the means of probe expression values for replicates for a given gene. The Student's t-Test calculates variance from the data that is available for each gene, while the Moderated t-Test uses information from all of the selected probes to calculate variance.

To correct for multiple test (i.e., false positives within a large dataset) we used the Benjamini Hochberg (Benjamini and Hochberg 1995) method, which is not as conservative as the Bonferroni approach. This method aims to reduce what is called the False Discovery Rate (FDR) and is used when the objective is to reduce the number of false positives and to increase the chances of identifying all the differentially expressed genes. In this method, the p-values are first sorted and ranked. The smallest value gets rank 1, the second rank 2, and the largest gets rank N. Then, each p-value is multiplied by N and divided by its assigned rank to give the adjusted p-values. In order to restrict the false discovery rate to 0.05, all the probes with adjusted p-values less than 0.05 are selected.

Probes that exhibit differential expression all through the six experiments were selected. Overall, 1463, 1998, 2319 probes were identified at 99%, 95%, and 90% of confidence, respectively; and 2285 and 1273 show 4 and 8 fold changes, respectively. Altogether, 1195, 1263, 1268 probes at 99%, 95% and 90% confidence exhibit 8-fold changes; and 1148, 1930 and 2072 99%, 95% and 90% confidence exhibit 4-fold changes. The significant expressed ORFs were identified by collating the extracted probe locations with the *Salmonella enterica* serovar Typhimurium genome. [LINK FIG]

3.3 Concluding remarks

Unlike regulators such as the LacI (Martinez-Antonio and Collado-Vides 2003) and MelR (D. C. Grainger, T. W. Overton et al. 2004) proteins of *E. coli* that govern expression of single promoters, many transcriptional regulators control multiple promoters that express products required in different amounts or for different extents of time. This is clearly the case for the regulatory protein CRP and PhoP, which control transcription of a large numbers of genes that are differentially regulated. Our findings argue that understanding a cell's behavior in terms of differential expression of genes controlled by a transcription factor requires a detailed analysis of the promoter's regulatory features. As a single nucleotide difference in the binding site for a transcription factor can dictate the requirement for co-activator proteins (Leung, Hoffmann et al. 2004), we feel that by considering multiple models (as opposed to the relying on a single consensus) it will be possible to detect and uncover subtle differences between regulatory targets and to capture the salient properties of co-regulated promoters.

We describe a flexible computational framework to encode and subdivide BS sequences into families of submotifs that may have differential affinities for the regulator. We computationally demonstrated that the *D&C* approach improves the recognition of functional BS, differentiating them from a background of variable DNA sequences that do not play a direct role in gene regulation. The proposed method is independent of the clustering approaches used to group sequences, as well as of the methods used to encode sequences into PWMs and the metrics utilized to characterize their performance. The proposed framework can incorporate other *cis*-regulatory features like the interaction with the RNAP, which results in an improvement of the computational performance. Moreover, the existence of rules that associate specific submotifs with certain distances to the RNAP binding site suggests that these features are not independently organized in the genome. Thus, a comprehensive understanding of the regulatory elements should treat them together.

In addition to the computational usefulness of the submotif approach, we show that these families of motifs also characterize the evolution of the PhoP regulon and differentiate among the targets of regulation even within the same species. Although some families are lost during the evolution of the orthologous targets, some of them may be resurrected resurrect when new genes are acquired in the regulon. This suggests that the sequences grouped into submotifs are not fortuitous computational convenience. Furthermore, the submotifs approach can complement experimental assays in the genome-wide analysis of the regulon. Particularly, it sheds light on cases of regulation that otherwise would be undetected or misinterpreted.

Chapter 4

Dissecting network motifs by identifying promoter features that govern differential gene expression

Whole genome sequences and genome-wide gene expression patterns (usually in the form of microarray data) provide the raw material for the characterization and understanding of transcription regulatory networks. These networks can be represented as directed graphs in which a node stands for a gene (or an operon in the case of bacteria) and an edge symbolizes a direct transcriptional interaction. Recurrent patterns of interactions, termed network motifs, occur far more often than in randomized networks, forming elementary building blocks that carry out key functions. This is a convenient representation of the topology of a set of regulatory Boolean (i.e. ON-OFF) networks, in which each gene is either fully expressed or not expressed at all, or that it has a binding site for a transcriptional regulator or lacks such a site. However, this approach has serious limitations because most genes are not expressed in a simple Boolean fashion. Indeed, genes that are co-regulated by the same transcription factor are often differently expressed with characteristic expression levels and kinetics. Therefore, a deeper understanding of regulatory networks demands the identification of the key features used by a transcriptional regulator to differentially control genes that display distinct behaviors despite belonging to networks with identical motifs.

The identification of the promoter features that determine the distinct expression behavior of co-regulated genes is a challenging task because: first, there are difficulties in discerning the sequence elements relevant to differential expression patterns (e.g., the binding sites for transcriptional regulators and RNA polymerase) from a background of variable DNA sequences that do not play a direct role in gene regulation. Second, the sequences recognized by a transcription factor may differ from promoter to promoter within and between genomes and may be located at various distances from other *cis*-acting features in different promoters (Winfield and Groisman 2004; Zwir, Shin et al. 2005). Third, similar expression patterns can be generated from different or a mixture of multiple underlying features, thus, making it more difficult to discern the causes of analogous regulatory effects.

In this study, we present a method specifically aimed at handling the variability in sequence, location and topology that characterize gene transcription. Instead of using an overall consensus model for a feature, where important differences are often concealed because of intrinsic averaging operations between promoters and even across species, we decompose a feature into a family of models or building blocks. This family of models can be arranged to collaboratively classify promoters (i.e., multi-classifier) that maximizes the sensitivity of detecting those instances that weakly resemble a consensus (e.g., binding site sequences) without decreasing the specificity. In addition, features are considered using fuzzy assignments, which allow us to encode how well a particular sequence matches each of the multiple models for a given promoter feature. Individual features are then linked into more informative composite fuzzy expressions that can be used to explain the kinetic expression behavior of genes. We applied our method to analyze promoters controlled by the PhoP/PhoQ regulatory system of *Escherichia coli* and *Salmonella* enterica serovar Typhimurium. This system responds to the same inducing signal (i.e. low Mg2+) in both species (Minagawa, Ogasawara et al. 2003; Zwir, Shin et al. 2005). Moreover, the *E. coli phoP* gene could complement a *Salmonella* phoP mutant (Groisman 2001). The DNA-binding PhoP protein appears to recognize a tandem repeat sequence separated by 5 bp (Minagawa, Ogasawara et al. 2003; Zwir, Shin et al. 2005), consistent with being a dimer. The PhoP/PhoQ system is an excellent test case because it controls the expression of a large number of genes, amounting to ca. 3% of the genes in the case of Salmonella. Furthermore, the PhoP/PhoQ regulon has been shown to employ a variety of network motifs including the single-input module (Figure 4.1A), the multi-input module (Figure 4.1B), the bi-fan (Figure 4.1C), the chained (Figure 4.1D), and the feedforward loop (Figure 4.1E) (Zwir, Huang et al. 2005; Zwir, Shin et al. 2005). Our analysis uncovered the salient features that distinguish genes co-regulated by PhoP even if they belong to similar network motif. This approach has been extensively applied to other transcriptional regulators such as the master regulator cAMP receptor protein (CRP). Gene transcription measurements provided experimental support for the investigated predictions.

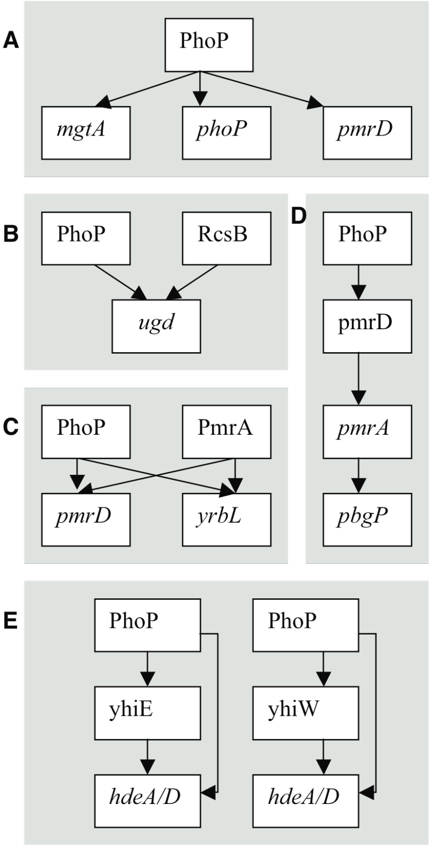


Figure 4.1 The PhoP/PhoQ system employs a variety of network motifs to regulate gene transcription
(a) In the single-input module, PhoP as a single transcription factor regulates a set of genes (i.e. *mgtA*, *phoP* and *pmrD*). (b) In the multi-input module, two or more transcription factors (e.g., PhoP and RcsB) regulate a target gene (i.e. *ugd*). (c) In the bi-fan module, a set of genes (i.e. *pmrD* and *yrbL*) are each regulated by a combination of transcription factors (i.e. PhoP and PmrA). (d) In the chained motif, genes are regulated in an ordered cascade. (e) In the feedforward loop, a transcription factor (i.e. PhoP) regulates the expression of a second transcription factor (i.e. YhiW), and both jointly regulate one or more genes (i.e. *hdeA*/*D*).

4.1 Results

4.1.1 Approach

We investigated five types of *cis*-acting promoter features by extracting the maximal amount of useful information from datasets and then creating models that describe promoter regulatory regions. This entailed applying three key strategies: first, we conducted an initial survey of the data provided from different available sources, capturing and distinguishing between broad and easily discernable patterns, which allowed the detection of those instances where a binding site sequence resembles the consensus only weakly or where the distances between the transcription factor and the RNA polymerase are unusual. Second, we utilized fuzzy clustering methods (Bezdek 1998; Gasch and Eisen 2002) to encode how a promoter matches each of the multiple models for a given promoter feature, which avoided having to make premature categorical assignments, thus producing an initial classification of the promoters into multiple subsets. Finally, we applied fuzzy logic (Klir and Folger 1988) to link basic features into more informative composite models that explain the distinct expression behavior of genes belonging to similar networks. These models were optimized by using a genetic algorithm (GA) that takes an input feature and iteratively optimizes its performance (see Methods). As a result we obtained a method that did not compromise the specificity and has an overall improved sensitivity.

4.1.2 Transcription factor binding site submotifs

Many genes are controlled by a single-input network motif where the affinity of a transcription factor for its promoter sequences is a major determinant of gene expression (Figure 4.2A). Thus, co-regulated genes displaying distinct expression patterns are likely to differ in the binding site for such a transcription factor. Methods that look for matching to a consensus sequence have been successfully used to identify promoters controlled by particular transcription factors (Stormo 2000). However, the strict cutoffs used by such methods increase specificity but decrease sensitivity (Stormo 2000), which makes it difficult to detect binding sites with weak resemblance to a consensus sequence.

To circumvent the limitation of consensus methods (Tompa, Li et al. 2005), we decomposed the binding site motif of a transcription factor into several submotifs (Figure 4.3.) and then combined the submotifs into a multi-classifier (see Methods), which increased the sensitivity to weak sites without losing specificity. In the case of PhoP, we identified four submotifs, and used them to search both strands of the intergenic regions of the *E. coli* and *Salmonella* genomes.

This allowed the recovery of promoters, such as that corresponding to the *E. coli hdeA* gene or the *Salmonella pmrD*, that had not been detected by the single consensus position weight matrix model (Stormo 2000) despite being footprinted by the PhoP protein (Zwir, Huang et al. 2005; Zwir, Shin et al. 2005).

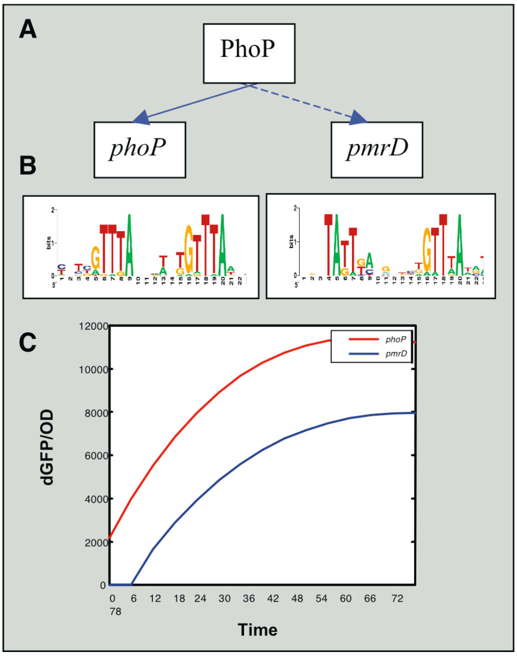


Figure 4.2 The PhoP protein achieves differential expression using the single-input network motif by controlling genes that differ in their binding site submotifs.
(a) PhoP regulates several promoters (i.e. *phoP* and *pmrD*) using a single-input network motif. (b) The PhoP protein recognizes a binding site motif consisting of a hexameric direct repeat separated by 5 bp, but distinguishes between different submotifs. (c) Transcriptional activity of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene and the *Salmonella phoP* (red color) or *pmrD* (blue color) promoters. The activity of each promoter is proportional to the number of GFP molecules produced per unit time per cell [$dG_i(t)/dt]/OD_i(t)$], where $G_i(t)$ is GFP fluorescence from wild-type *Salmonella* strain 14028s culture and conditions described in Methods, and $OD_i(t)$ is the optical density. The activity signal was smoothed by a polynomial fit (sixth order). The results are not normalized. Faster and earlier GFP expression was observed when transcription was driven by the *phoP* promoter, which has the M$_2$ submotif, than by the *pmrD* promoter, which has the M$_1$ submotif.

To test the notion that PhoP binding to promoters with different PhoP box submotifs is a determinant promoter activity, we compared the gene expression patterns of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene to different PhoP-activated promoters. Faster GFP expression kinetics were observed when transcription was driven by the *phoP* promoter, which has the M$_2$ submotif, than when it was driven by the *pmrD* promoter, which has the M$_1$ submotif, (Figure 4.2.B-C). Thus, the binding site for a transcriptional regulator is a key determinant in gene expression.

The use of submotifs instead of a single consensus increased the sensitivity for PhoP binding sites from 66% to 91%; yet, the specificity remained essentially the same (i.e., 98% in a consensus model versus 97% in the case of submotifs

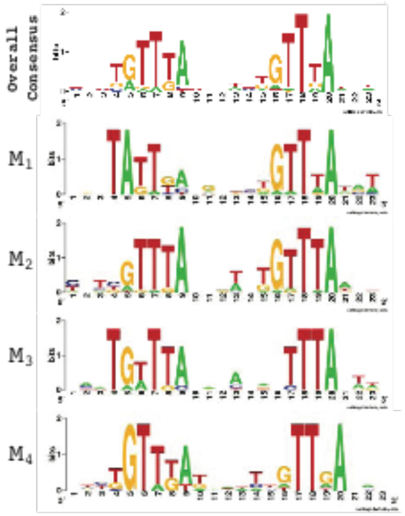


Figure 4.3 The PhoP binding site submotifs.
We built a model for the PhoP binding site by applying an extension of the Consensus/Patser program using promoters exhibiting coherent expression patterns. This initial model was used for identifying promoter motifs, which were clustered into four subsets based on their degree of sequence similarity to one another (M1- M4). All four submotifs have an hexameric direct repeat separated by 5 bp and preserved those conserved positions critical for PhoP-promoted transcription of the mgtA promoter (Yamamoto, Ogasawara et al. 2002). By distinguishing classes of submotifs each supported by many promoters, and clarifying how they differ from the original consensus, we could increase the specificity of the PhoP binding site model as well as its sensitivity to weak sites. Therefore, this approach has the potential of revealing promoters with weak degree of matching to a consensus sequence.

4.1.3 Transcription factor binding site orientation

Functional binding sites for a transcription factor may be present in either orientation relative to the RNA polymerase binding site. This is due to the possibility of DNA looping and to the flexibility of the alpha subunit of the bacterial RNA polymerase in its interactions with transcriptional regulators (Barnard, Wolfe et al. 2004).

Analysis of PhoP-regulated promoters revealed that the PhoP box could be found with the same probability in either orientation in the intergenic regions of the *E. coli* and *Salmonella* genomes. For example, the *E. coli ompT* and *yhiW* promoters and the *Salmonella mig-14*, *pipD*, *pagC* and *pagK* promoters harbor putative PhoP binding sites in the opposite relative orientation to that described for the prototypical PhoP-activated mgtA promoter (Zwir, Shin et al. 2005). Yet other promoters (i.e. those of the *ybjX*, *slyB*, *yeaF* genes in *E. coli* and the *virK*, *ybjX*, and *mgtC* genes in Salmonella) contain sequences resembling the PhoP box in both orientations. The demonstration that PhoP does bind to the *mgtC*, *mig-14* and *pagC* promoters (Zwir, Shin et al. 2005), which harbor the PhoP binding site in the opposite orientation as in the *mgtA* promoter, validates our predictions and argues against alternative network designs where these promoters would be regulated by PhoP only indirectly (Lejona, Aguirre et al. 2003).

To assess the contribution of PhoP box orientation to gene expression, we determined the fluorescence of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between promoterless gfp genes to PhoP-regulated promoters that differed in the orientation of the PhoP box. Promoters with the PhoP box in the direct orientation, such as those corresponding to the *yobG* and *slyB* genes, were transcribed earlier and faster than the *pagK* and *pagC* promoters in which the PhoP box is in the opposite relative orientation (Figure 4.4A-C). This is in spite of the fact that *yobG* and *pagK* promoters are equally divergent from the PhoP binding site consensus (60% and 66% of the consensus information content). Furthermore, promoters sharing the same PhoP binding site submotif but arranged in different orientations (e.g. the *ugd* and *mig-14* promoters) produced distinct rise times and expression levels (data not shown).

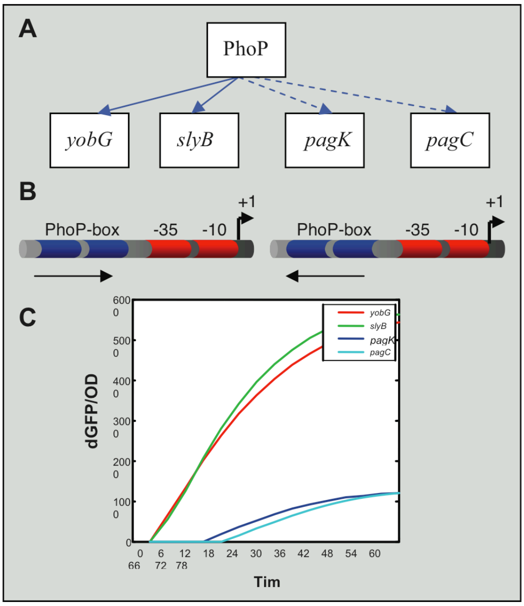


Figure 4.4 Expression of PhoP-regulated promoters that differ in the orientation of the PhoP-binding site.
(a) PhoP regulates a set of promoters including those of the Salmonella yobG, slyB, pagK and pagC genes using a single-input network motif. (b) We established that when Salmonella experiences low Mg2+, the PhoP protein binds to both the archetypal directly oriented yobG and slyB promoters as well as the oppositely oriented pagK and pagC promoters using chromatin immunoprecipitation (ChIP) in vivo. (c) Transcriptional activity of wild-type Salmonella harboring plasmids with a transcriptional fusion between a promoterless gfp gene and the Salmonella yobG (red color) or slyB (green color) promoters reveals a much earlier an higher levels of activity than the isogenic strains with fusions to the pagK (blue color) and pagC (cyan color) promoters. Promoter activity was determined as described in the legend to **Figure 4.2**. Thus, the orientation of the binding site for a transcriptional regulator contributes to the kinetic behavior as well as the maximum expression levels achieved by the promoters

4.1.4 RNA polymerase site

The distance of a transcription factor binding site to the RNA polymerase binding site(s) and the class of sigma 70 promoter are critical determinants of gene expression (Barnard, Wolfe et al. 2004). These classes correspond to the different types of contacts that can be established between a transcription factor and RNA polymerase.

We identified seven patterns among PhoP-regulated promoters of *E. coli* and *Salmonella* that combine promoter class and distance between the PhoP box and the RNA polymerase site (Figure 4.5). These patterns may correspond to different kinetic behaviors within a network motif (Barnard, Wolfe et al. 2004). For example, the *ugtL* and *pagC* promoters share the orientation of the PhoP box but differ in the distance of the PhoP box to the RNA polymerase binding site (Figure 4.6A-B). This may account for the different dynamic behavior of these promoters when tested in a wild-type strain harboring plasmids with promoter fusions to the promoterless gfp gene (Figure 4.6 C).

In addition, some PhoP-regulated promoters (e.g. the *hemL* and *phoP* promoters of *E. coli*) contain several putative RNA polymerase binding sites located at different positions and belonging to different classes, suggesting that such promoters may be regulated by additional signals and/or transcription factors (Minagawa, Ogasawara et al. 2003).

4.1.5 Activated/Repressed promoters

Gene expression data normally allow clear separation of genes into those that are activated and those that are repressed by a regulatory protein. Because the expression signal is sometimes absent or too low to be informative, we considered the location of a transcription factor binding site relative to that of the RNA polymerase to separate promoters into activated and repressed subsets (Collado-Vides, Magasanik et al. 1991).

We determined that the location of binding sites functioning in activation is different from that corresponding to sites functioning in repression, being centered ~40 and ~20 bp upstream of the transcription start site, respectively. This allowed us to distinguish among PhoP-regulated promoters that have apparently similar network motifs For example, we identified a PhoP binding site at a relative distance to the RNA polymerase consistent with repression in the promoter region of the *hilA* gene, which encodes a master regulator of *Salmonella* invasion and had been known to be under transcriptional repression by the PhoP/PhoQ system (Groisman 2001). Several promoters, including those of the *Salmonella pipD* and *nmpC* genes, were classified as candidates for being both activated and repressed, because the distance between the predicted transcription start site and the PhoP box is consistent with either activation or repression. Gene expression experiments conducted in *E. coli* indicate that *nmpC* is a PhoP-repressed gene (Minagawa, Ogasawara et al. 2003; Zwir, Shin et al. 2005). Other promoters were predicted to have more than one PhoP box (e.g., those of the PhoP-activated *mgtC* and *pagC* genes), where one could correspond to an activation site and the other to a repression site. Indeed, this appears to be the case of the PhoP-activated *iraP* gene (X. Tu, T. Latifi et al. 2006).

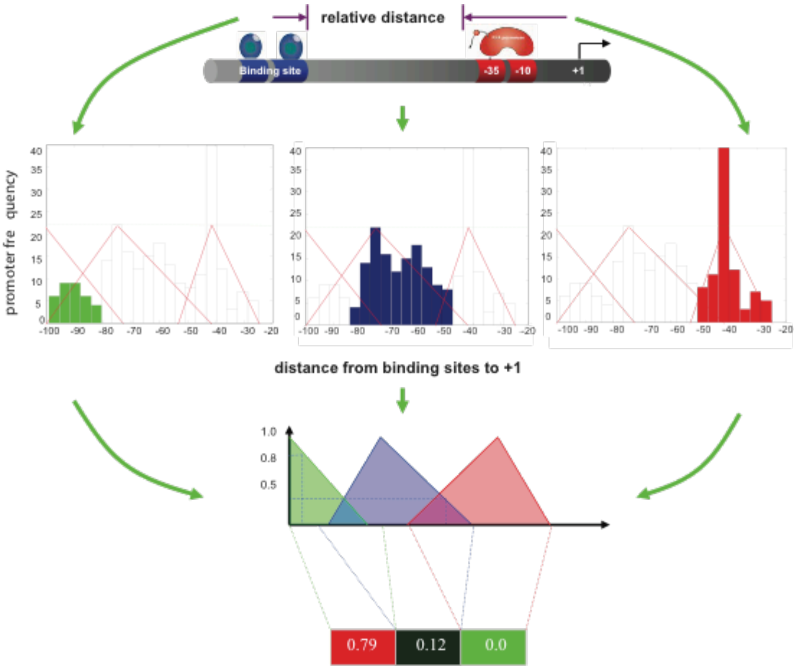


Figure 4.5 Learning and prototyping the relationships between the proximity of a transcription factor binding site and the RNA polymerase site.
We learned models corresponding to the distances between transcription start sites (+1) and the binding sites of regulators from examples in databases (e.g., RegulonDB). These distances were grouped into three histograms and codified as elastic (fuzzy)unit-interval functions, which can be interpreted as the membership degrees (μ)by which subsets of the dataset can embrace this property. This process is analogous to fitting data from a parametric or non-parametric distribution and then assigning probabilities of membership to such distributions. We labeled these distance models as *close* (red), *medium* (blue) and *remote* (green). The similarity of a hypothetical promoter to two fuzzy triangular membership functions, corresponding to the labels *medium* and *remote*, illustrate the degree by which distances from the transcription factor binding site to transcription start site (+1) (X-axis) are encoded (Y-axis). In this example $\mu(-70)=0.12$ while $\mu(-90)=0.79$, revealing more confidence that the promoter belongs to the *remote* model than the *medium* model). We used these models to characterize the relationships between binding sites for the PhoP protein and the RNA polymerase binding site in the genome. Relationships were classified according to their similarity (fuzzy membership) with the prototypes to obtain a similarity vector of expression values, in a manner analogous to that described in the legend to **Figure 4.7**

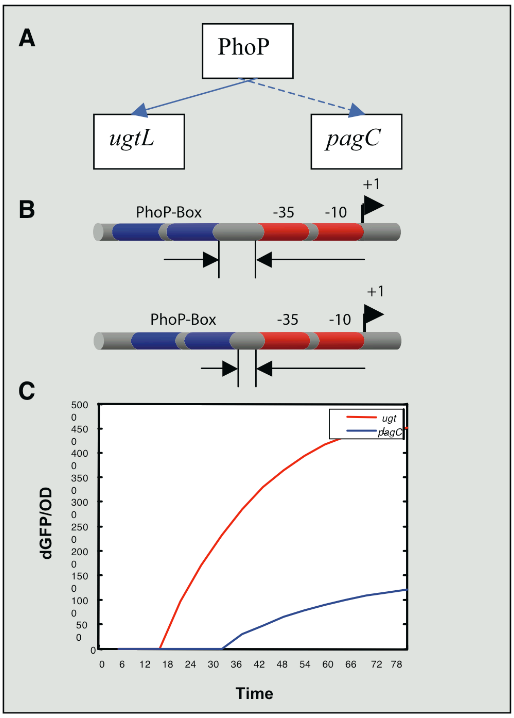


Figure 4.6 Expression of PhoP-regulated promoters that differ in the RNA polymerase sites. The PhoP-activated *ugtL* and *pag*C promoters share the orientation of the PhoP-binding site as well as the class I sigma 70 promoter, but differ in the distance between the PhoP box and the RNA polymerase site. (c) Transcriptional activity of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene and the *Salmonella ugtL* (red color) and *pagC* (blue color) promoters. Promoter activity was determined as described in the legend to **Figure 4.2**. The *ugtL* promoter is transcribed earlier than the *pag*C promoter, which also exhibits a PhoP box in the opposite orientation but more distant from the RNA polymerase site.

4.1.6 Binding sites for other transcription factors.

Certain promoters harbor binding sites for more than one transcription factor. This could be because transcription requires the concerted action of such proteins, or because the promoter is independ¬ently activated by individual transcription factors, each responding to a distinct signal.

We analyzed the intergenic regions of the E. coli and Salmonella genomes for the presence of binding sites for 54 transcription factors (Salgado, Gama-Castro et al. 2004). We then investigated the co-occurrence of 24 sites with the binding site of the PhoP protein in an effort to uncover different types of network motifs involving PhoP-regulated promoters. For example, the *Salmonella pmrD, ugd* and *yrbL* promoters and the *E. coli yrbL* promoter harbor PhoP- and PmrA-binding sites, consistent with the experimentally-verified regulation by both the PhoP and PmrA proteins that can be described by the bi-fan network motif (Kato and Groisman 2004; Zwir, Shin et al. 2005) (Figure 4.8A). In addition, the relative position of transcription factor binding sites (Figure 4.7D) can play a critical role because the PmrA-box in the *Salmonella pmrD* and *yrbL* promoters is located closer to the PhoP-box (~38 bp and ~24 bp, respectively) than in the *udg* promoter (~65 bp), which could account for the different expression patterns exhibited by their respective genes (Figure 4.8B-C). By analyzing both the binding site quality and the location of transcription factor binding sites, we increase the chances of identifying co-regulated promoters.

By considering the presence of binding sites for multiple transcription factors, it is possible to generate hypotheses about potential network motifs. For example, the promoters of the PhoP-activated *gadA, dps, hdeA, yhiE* and *yhiW* genes of *E. coli* also have binding sites for the regulatory proteins *YhiX* and *YhiE* (Zwir, Shin et al. 2005), raising the possibility that some of these genes might be regulated by feedforward loops where both the PhoP protein and either the *YhiW* or the *YhiE* proteins would bind to the same promoter to activate transcription. This notion was experimentally verified (Zwir, Shin et al. 2005), validating our prediction.

4.1.7 Regulation of orthologous genes

A distinguishing characteristic of our approach is that promoters for orthologous genes are considered individually. This is in contrast to some phylogenetic footprinting methods (McCue, Thompson et al. 2001) that often ignore regulatory differences among closely-related organisms due to their strict reliance on the conservation of regulatory motifs across bacterial species. Thus, we could uncover cases of phenotypic differences between closely-related species resulting from the differential regulation of homologous genes. For example, the *ugd* and *iraP* promoters of *Salmonella* harbor functional PhoP boxes that are footprinted by the purified PhoP protein (Zwir, Huang et al. 2005; Zwir, Shin et al. 2005; X. Tu, T. Latifi et al. 2006). By contrast, the PhoP boxes are missing from the *ugd* and *iraP* promoters of *E. coli*, preventing it from expressing these genes under the same conditions as *Salmonella* (X. Tu, T. Latifi et al. 2006) (Mouslim and Groisman, unpublished results). Likewise, there is a PhoP box in the *pmrD* promoters of both *E. coli* and *Salmonella* (albeit of different submotifs) but only the *Salmonella pmrD* promoter has a PmrA box that functions as a repression site (Winfield and Groisman 2004; Zwir, Shin et al. 2005). This demonstrates that the detailed analysis of *cis*-features can shed light on different network motif design among closely-related bacterial species.

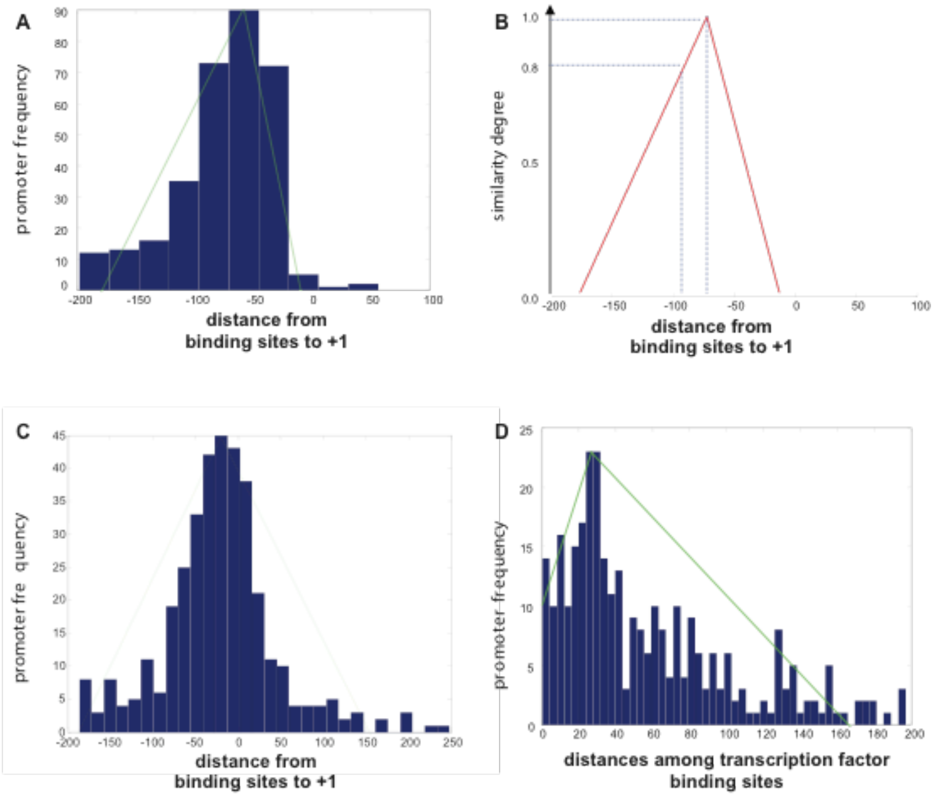


Figure 4.7 Learning promoter features.

Promoter features were learned as models from examples in databases (e.g., RegulonDB) and then used to describe the intergenic regions of the *E. coli* and *S. enterica* genomes. (a) Promoters were classified into activated, repressed or both, based on the location and the distance of a regulatory protein binding site to the RNA polymerase site. Different distributions are observed for activated, repressed and activated/repressed genes. The property that characterizes activated genes was learned from distances between the transcription start sites (+1) and the binding sites of different transcription factors. These distances were grouped in histograms and codified as elastic (fuzzy) functions, which can be interpreted as the membership degrees (in a unit interval) by which subsets of the dataset can embrace this property. (b) A fuzzy triangular membership function (μ) corresponding to (a) is used to encode the degree by which binding site distances to transcription start sites (+1) (X-axis) for *activated* promoters (e.g., μ is maximal at -70 and μ(-90)=0.8, revealing less confidence). (c) The histogram and membership function corresponding to repressed promoters. μ is maximal at much closer distances. Thus, the promoter distances can be probabilistically interpreted as the posterior probability *p(close|activated)* that given an *activated* gene, the regulator binding site is at a *close* distance from the transcription start site, following Bayes' rule. (d) The histogram illustrates the distances for binding sites of different regulators sharing the same promoter regions. The resulting membership functions, which were learned from such distributions, allows evaluating the putative relationship between a transcription factor motif and a PhoP box based both on motif quality and physical location.

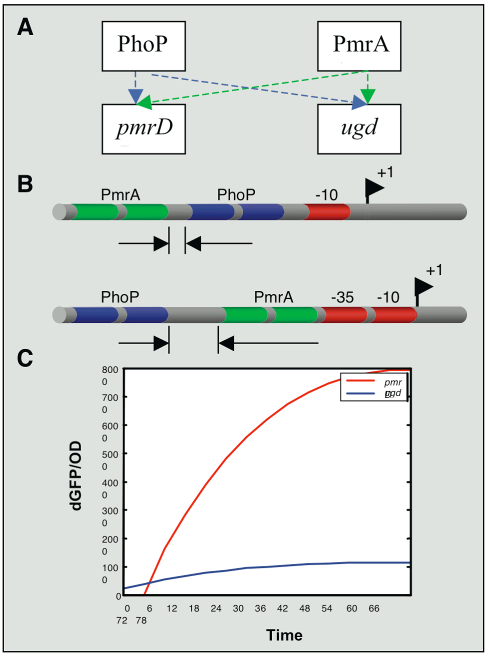


Figure 4.8 Expression of PhoP-regulated promoters that use the bi-fan network motif. (a) The *Salmonella pmrD*, and *ugd* promoters harbor experimentally verified PhoP- and PmrA-binding sites that can be described by the bi-fan network motif. (b) The distance between the PhoP and PmrA boxes in the *Salmonella pmrD* and *ugd* promoters are different (~38 bp and ~65 bp, respectively). (c) Transcriptional activity of wild-type *Salmonella* harboring plasmids with a transcriptional fusion between a promoterless *gfp* gene and the *Salmonella pmrD* and *ugd* promoters. Promoter activity was determined as described in the legend to Fig. 2. The two promoters confer different expression and kinetic patterns.

4.2 Materials and Methods

Our method consists of five phases: first, encoding the available information into preliminary model-based features, which includes identifying *cis*-features from DNA sequences and information from available databases; performing initial modeling of each individual feature, allowing the process of multiple occurrences of a feature and using relaxed thresholds and permitting missing values.

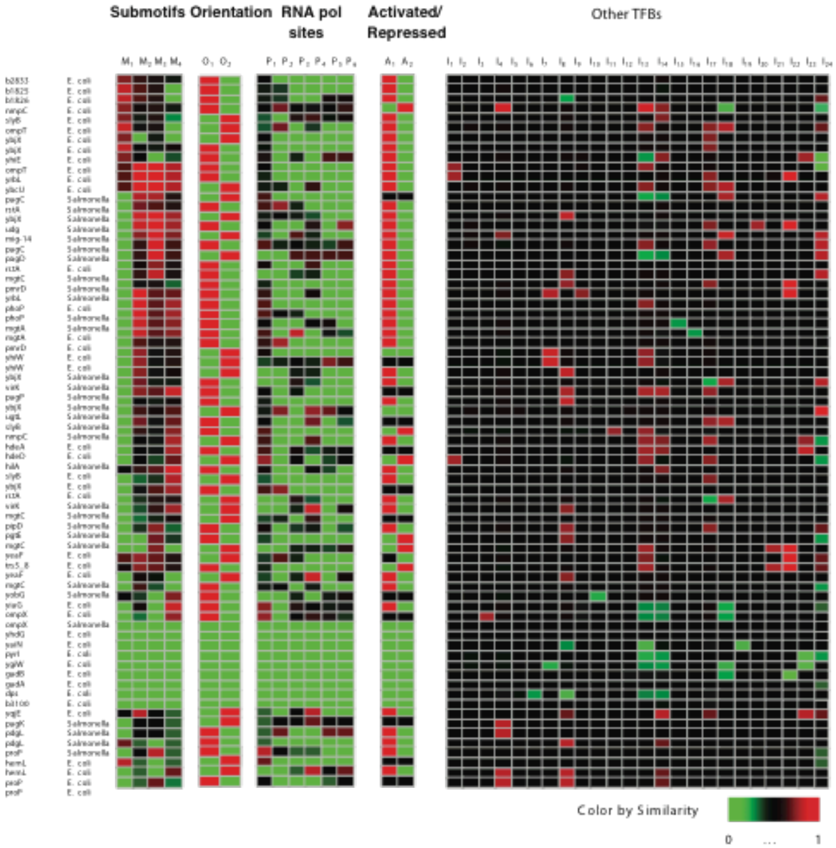


Figure 4.9 PhoP-regulated promoters are described on the basis of five types of features
PhoP-regulated promoters are described on the basis of five types of features: PhoP box submotifs (submotifs), the orientation of the PhoP box (orientation), the distance of the PhoP box relative to the RNA polymerase site and the class of sigma 70 promoter (RNA polymerase sites), whether the position of the PhoP box suggests that a promoter is activated or repressed (activated/repressed), and the presence of potential binding sites for 24 transcription factors in the PhoP-regulated promoters (Other TFBs). The identification of a feature in a promoter is based on measuring the degree of match between a promoter instance and a model that represents that feature, which results in a vector of [0, 1] values where 1 (red) corresponds to maximum matching and 0 (green) corresponds to the absence of the feature. Individual genes are allowed to have more than one promoter because more than one candidate PhoP box can be identified in an intergenic region. In addition, promoters for the same gene in different genomes are considered separately in the *E. coli* and *Salmonella* genomes. Submotif analysis of the PhoP box resulted in four groups (M_1-M_4), which are detailed in **Figure 4.3**. The PhoP box could be present in the opposite (O_1) or the same (O_2) orientation as the regulated open reading frame. RNA pol sites analysis revealed six groups (P_1-P_6) corresponding to types and location of sigma 70 promoters: (1) *close class II*, (2) *close class I*, (3) *medium class II*, (4) *medium class I*, (5) *remote class II* and (6) *remote class I*. Activated/repressed analysis discriminates among three groups (A_1-A_2) corresponding to activated, and repressed genes, respectively. The presence of other transcription factor binding sites in PhoP-regulated promoters includes: (1) OxyR, (2) FruR, (3) DeoR, (4) MalT, (5) MelR, (6) CytR, (7) GlpR, (8) ArcA, (9) FNR, (10) RcsB, (11) Fur, (12) ArgR, (13) RhaS, (14) AraC, (15) CRP, (16) DnaA, (17) YhiW, (18) Lrp, (19) NarL, (20) FIS, (21) IHF, (22) OmpR, (23) PmrA and (24) SlyA.

A model-based feature is generated by the identification of a feature in a subset of observations (F) in the dataset, based on measuring the degree of match (Q) between an observation and a model, or a family of models (M={ M_{α} }), at some degree (α) defined in a unit-interval scale (i.e., fuzzy values, Q(F, M_{α})) {Ruspini, 1999 #390;Ruspini, 2002 #427;Zwir, 2002 #1}. Second, grouping the results into subsets, thus, decomposing the preliminary models into a family of models or building blocks by using fuzzy clustering. Third, composing the building blocks by either combining the same or different types of features by using fuzzy logic expressions. Fourth, learning the optimal manner by which these building blocks can cooperate to classify promoters, because aggregated independent rules not necessarily improve the results. And fifth, describing new promoters using the resulting models.

4.2.1 Network motifs

In theory, the term "network motifs" is related to a statistical significant subgraph; however, in practice, they are treated as an over represented subgraph (see (17-19)). For example, a motif termed "single input motif " of three/four nodes in the *E. coli* (20) (e.g., mfinder1.2 p-value < 34.7+-8.5) or Saccharomyces cerevisiae network (21) is not recognized as significant, while the only motif that exceeds the standard threshold is the "feed forward motif".

4.2.2 Binding site submotifs and orientation

(1) We built an initial model for the PhoP binding site by learning a position weight matrix (*E-value* < 10E-12) based on the upstream sequences of genes corresponding to the training set of the *E. coli* and *Salmonella* genomes. **(2)** We searched the intergenic regions of the genes in both orientations, using low thresholds corresponding to two standard deviations below the mean score obtained with the initial model (Robison, McGuire et al. 1998). Multiple PhoP binding site candidates were allowed in a given promoter operator region. **(3)** After transforming nucleotides into dummy variables, we grouped sequences matching the PhoP position weight matrix using the fuzzy C-means clustering method with the Xie-Beni validity index (see below) to estimate the number of clusters (Bezdek, Pal et al. 1992; Bezdek 1998). **(4)** We built models for these clusters using position weight matrices (*E-value* < 10E-22) and searched the *E. coli* and *Salmonella* genomes to characterize each gene according to its similarity to each model as a fuzzy partition (Figure 4.3).

Performance. To evaluate the ability of the resulting models to describe PhoP-regulated promoters, we extended the dataset by including 772 promoters (RegulonDB V3.1 database (Salgado, Gama-Castro et al. 2004)) that are regulated by transcription factors other than PhoP (see "Search known transcription factor motifs" in gps-tools.wustl.edu), by selecting the promoter region corresponding to the respective transcription factor binding site ± 10 bp.

We considered the compiled list of PhoP regulated genes as true positive examples and the binding sites of other transcriptional regulators as true negative examples to evaluate the performance of the submotif feature.

We used a leave-one-out crossvalidation process (Crossvalind, Matlab r2006a), which is appropriate for reduced datasets, as a procedure to estimate the variance error on the training set (correct test estimation of 94% vs. 75% between submotifs and single position weight matrices, respectively). The thresholds of the matrices were optimized for classification purposes by using a GA (see Methods) based on the extended dataset. (see the complete evaluation of genomes in gps-tools.wustl.edu). We found that the PhoP-binding site model increases its sensitivity from 66% to 91% when submotifs are used instead of a single consensus, while its specificity went from 98% to 97% (correlation coefficient 73% vs. 87%). We also obtained substantial improvements for other transcription factors from RegulonDB. For example, by considering the CRP regulator, we used 130 promoters regulated by this protein in RegulonDB as the true positive values and 642 regulated by other proteins than CRP as negative examples. We found that the sensitivity of the CRP model for binding sites increases from 29% to 50%, by using submotifs instead of a single consensus, while the specificity remains the same at 98% (correlation coefficient 39% vs. 62%). Overall, by considering transcription factors with more than ten reported binding sequences in the RegulonDB data base (including CRP, Lrp, FIS, IHF, FNR, ArcA, NarL, GlpR, PurR, OmpR, TyrR, AraC, Fur, CytR, FruR,Hns, ArgR, DnaA, PhoB, and LexA), we could increase the sensitivity in an average of 35%, while retain almost the same sensitivity than a single position weight matrix (average correlation coefficient 87%).

4.2.3 RNA polymerase sites

(1) We gathered sigma 70 class I and class II promoters (Salgado, Gama-Castro et al. 2004) from the RegulonDB database. Then, we built models of the RNA polymerase site using a neuro-fuzzy method (see HPAM in gps-tools.wustl.edu (Cotik, Zaliz et al. 2005)), and used the resulting models to perform genome-wide descriptions of the intergenic regions of the *E. coli* and *Salmonella* genomes with a false discovery rate <0.001 (see Promoter search in gps-tools.wustl.edu). **(2)** We used an intelligent parser to differentiate class I and class II promoters that evaluate the quality of the -35 motif (Ishihama 1993; Barnard, Wolfe et al. 2004), based on fuzzy logic and GAs techniques (see MOSS in gps-tools.wustl.edu (Romero Zaliz, Zwir et al. 2004)). **(3)** To characterize the distance relationship between transcription factors binding sites and RNA polymerase binding sites, we built models of such distances from the examples reported in the RegulonDB database. **(3.1)** We modeled activated and repressed promoters (see below *Activated or repressed* feature). **(3.2)** We re-built histograms for each group of distances (i.e. activated and repressed), distinguishing three overlapping distributions for each of them (Figure 4.5). **(3.3)** We built models for distances by fitting their distributions into models based on fuzzy membership functions (Klir and Folger 1988), which were termed close, medium and remote distances for each set of activated and repressed genes. The initial models of the distances were optimized for classification purposes by using GAs (see Methods). Finally, to characterize the distance relationship between the PhoP box and putative RNA polymerase binding site, we connected (2) and (3) by using fuzzy logic-based operations (see below).

This process allowed us to retrieve the most representative RNA polymerase binding site candidates for each promoter region relative to the PhoP binding site (e.g., best class II RNA polymerase site, which is located close to the PhoP box in an activated promoter), which were arrayed and constituted the value of the RNA polymerase site feature.

Performance. The RNA polymerase site feature was evaluated using 721 RNA polymerase sites from RegulonDB as positive examples and 7210 random sequences as negative examples. We obtained an 82% sensitivity and 95% specificity for detecting RNA polymerase sites. These values provide an overall performance measurement (see below) of 92% corresponding to a false discovery rate <0.001 and a correlation coefficient of 82%. In addition, we selected 34 examples of RNA polymerase sites reported to be of class II, which all differ from the typical class I promoter by exhibiting a degenerate -35 sequence motif (Minagawa, Ogasawara et al. 2003; Barnard, Wolfe et al. 2004), and obtained 74% sensitivity and 95% specificity.

4.2.4 Activated/repressed

We modeled PhoP-regulated promoters as activated or repressed based on examples reported in the RegulonDB database (Salgado, Gama-Castro et al. 2004). **(1)** We separately grouped activated and repressed promoters, and developed histograms for each group corresponding to the distances between transcription factor binding sites and the transcription initiation (+1) site. **(2)** We distinguished two non-disjoint distributions in each group and built models for these distances by fitting histograms with fuzzy membership functions (Klir and Folger 1988) (Figure 4.7A-D), which do not force promoters to be exclusively Activated or Repressed. **(3)** Finally, we connected **(2)** and sigma 70 promoters previously detected to select the most representative candidate for each promoter condition (e.g., best promoter that characterize the activated condition) by using fuzzy logic-based operations as described above to characterize relationships between predicted PhoP and RNA polymerase binding sites detected in candidate promoters. Simple features, such as activated and repressed can be combined in more complex composite models to represent divergently transcribed genes (e.g., two adjacent genes, one repressed, the other activated, both sharing the same putative PhoP box in different orientations) using *fuzzy logic* expressions.

4.2.5 Binding sites for other transcription factors

We developed models for different transcription factor binding sites from the RegulonDB database as follows: **(1)** We built position weight matrices for each transcription factor using the Consensus/Patser (Stormo 2000) program, choosing the best final matrix for motif lengths between 14-30 bps if the corresponding length had not been previously specified (see "Consensus matrices" in gps-tools.wustl.edu). We accounted for the motif symmetry (e.g., asymmetric, direct, inverted (Salgado, Gama-Castro et al. 2004)) if available (see "Search known transcription factor motifs" in gps-tools.wustl.edu). **(2)** We searched the intergenic regions of the *E. coli* and *Salmonella* genomes with these models, using the *overall performance* measure (see below) and additional 772 promoters from the RegulonDB database (Salgado, Gama-Castro et al. 2004) to establish a threshold (average *E-value* < 10E-10) for each matrix (Benitez-Bellon, Moreno-Hagelsieb et al. 2002) (see "Thresholded consensus" in gps-tools.wustl.edu). **(3)** We accounted for the distances between distinct transcription factors binding sites occurring in the same promoter region (e.g., the distance between the CRP and FIS sites in the *proP* promoter) in promoters reported in RegulonDB database and built a histogram with the obtained results (Figure 4.7D). **(4)** We fitted the histogram using a fuzzy membership function (see below) and used this model as a fuzzy cluster to characterize the distances between a putative PhoP box and another putative transcription factor binding site detected in the same region. **(5)** Finally, we connected **(2)** and **(4)** by using fuzzy logic-based operations as described above, which can also have a probabilistic interpretation (e.g., *p(CRP,FIS|appropriate distance*) upstream of the *proP* open reading frame of *E. coli*), to characterize PhoP regulated candidates promoters.

4.2.6 Fuzzy logic expressions

Propositional calculus logic expressions can be extended by incorporating predicates having fuzzy variables, which are manipulated using various theorems/axioms and methods {Klir, 1988 #67}. This approach, which has been widely used in several fields including decision-making, artificial intelligence and electrical engineering {Klir, 1988 #67} for many years, was applied to model related features that describe different regulatory objects. **Thus, given a dataset** $X=\{x_1,...,x_n\}$, the feature that characterizes it can be best described as a set , where $\{d_{11},...,d_{1n}\}\in\{0,1\}$ in classical set theory and $[0,1]$ in fuzzy set theory. These fuzzy values represent the degree of matching between an observation of the dataset and a fuzzy set. The degree of matching is defined in the unit interval and can be obtained from evaluating the membership function of the corresponding fuzzy set (see below). **Then, given** $F_2(X)=\{d_{21}/x_1,...,d_{2n}/x_n\}$ and the Minimum as an intersection operator, we define the expression:

$$F_1(X)\ \ AND\ \ F_2(X) = F_1 \cap F_2 = MIN(F_1, F_2) = \{MIN(d_{11}, d_{21})/x_1, ..., MIN(d_{1n}, d_{2n})/x_n\}$$

Fuzzy logic-based operations, such as T-norms/conorms, include operators like *MINIMUM, PRODUCT,* or *MAXIMUM,* which are used as basic logic operators, such as AND or OR, or their set equivalents *INTERSECTION* or *UNION* {Klir, 1988 #67;Bezdek, 1998 #43}. We used in this work the Minimum and Maximum as T- and Tconorms, respectively.

4.2.7 Fuzzy membership functions

They can be viewed as approximation of data distributions, where the degree of matching in the [0,1] scale is calculated using triangular functions (Klir and Folger 1988). These functions were learned from the projection of the histograms onto the variable domains (Figure 4.7) by simple regression and minimum squared methods (Sugeno and Yasukama 1993).

4.2.8 Performance Measurement

We use a correlation coefficient implementation to establish best local thresholds for transcription factor binding site motifs. That is, from a range of possible thresholds applied over a particular motif, we choose the one that maximizes this coefficient defined as:

$$CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}},$$

where *specificity = TN/(TN + FP)* and *sensitivity = TP/(TP + FN); P= positive, N=negative, T = true and F=false*(Benitez-Bellon, Moreno-Hagelsieb et al. 2002). We constrained the sensitivity of the selected threshold to be above the 60%. The false positive rate for binding site analysis was calculated by detecting binding sites from other transcription factors different from the one being evaluated (RegulonDB database).

4.2.9 Genetic algorithm

Our proposed method employs GAs to optimize both: individual features as well as composite fuzzy expressions that connect more than one feature.

We constructed a GA based on "Genetic algorithm and direct search toolbox" of Matlab. Individual features like submotifs, were optimized by tuning the thresholds of the position weight matrices using real coded chromosomes. The triangular fuzzy membership functions, representing distances of the RNA polymerase feature, were encoded as triplets in the same type of chromosome.

The fitness of each individual was evaluated based on the correlation coefficient. The recombination was produced using the Max-Min-Arithmetical crossover operator as described in (Herrera, Lozano et al. 1995), which combines the features of two parent structures to form two similar offspring. We used a non-uniform mutation strategy, proposed by Michalewicz (Michalewicz 1994), which arbitrarily alters one or more components of a selected structure so as to increase the structural variability of the population.

4.2.10 Dataset

We initially used the intergenic regions of *E. coli* and *Salmonella* operons from -800 to +50 because >5% are larger than 800 bp in bacterial genomes (as described in the RegulonDB database or generously provided by H. Salgado); however, predictions have been performed in whole coding and non coding regions (see gps-tools.wustl.edu). The promoter and transcription factor information was taken from RegulonDB database. We compiled from the literature and our own lab information genes whose expression (using microarrays) differed statistically between wild-type and *phoP E. coli* strains experiencing inducing conditions for the PhoP/PhoQ regulatory system (Zwir, Shin et al. 2005), as well as a list of genes known/assumed to be PhoP regulated. However, this information did not explicitly indicate whether these genes were regulated directly or indirectly by the PhoP protein. The learned features were used to make genome-wide predictions in the *E. coli* and *Salmonella* genomes.

4.2.11 Programming resources

The scripts and programs used in this work, some of which are accessible from gps-tools.wustl.edu web site, were based on Perl, Matlab r2006a and C++ interpreters/languages, and the visualization routines were performed on Spotfire DecisionSite software 8.2. Data and predictions for *E. coli* and *Salmonella* genomes will be available at gps-tools.wustl.edu.

CRP	**Consensus**			**Meme**		
	CC	**Sp**	**Ss**	**CC**	**Sp**	**Ss**
Single sequence motif	0.43	0.93	0.45	0.42	0.97	0.33
Submotifs	0.54	0.90	0.66	0.47	0.97	0.39
Motif + Dist. to RNA Pol.	0.51	0.95	0.49	0.51	1.00	0.32

Table. 1. The Correlation Coefficient (CC), specificity (Sp) and sensitivity (Ss) obtained by using a single CRP motif, several submotifs and sequence motifs combined with distance to the RNA Polymerase site. Sequence motifs were build using Consensus (Stormo 2000) and MEME (Bailey and Elkan 1994) methods.

4.2.12 Bacterial strains, plasmids and growth conditions

Bacterial strains and plasmids used in this study were obtained and constructed at Groisman Lab. Washington University, School of Medicine, St. Louis, MO. C. (GFP reporter vector plasmid, a gift from Alon, U. (Mangan and Alon 2003)). *Salmonella* strain harboring GFP reporter plasmid was measured in parallel using automated microplate reader (VICTOR^3, Perkin Elmer) (Mangan and Alon 2003). The raw GFP and OD signals were used to calculate the promoter activity as $[dG_i(t)/dt]/OD_i(t)$. The activity signal was then smoothed by a shape-preserving interpolant (Piecewise Cubic Hermite Interpolating Polynomial, Matlab r2006a) fitting algorithm that finds values of an underlying interpolating function at intermediate points that are not described in the experimental assays. Then, we applied a polynomial fit (sixth order, Matlab r2006a) on each expression signal. This smoothing procedure captures the dynamics well, while removing the noise inherent in the differentiation of noisy signals.

4.3 Concluding remarks

We demonstrated that a transcription factor can mediate differential expression of genes that are described by the same network motif. This is because of the functional significance of variability in sequence, location and topology that exists among promoters that are co-regulated by a given transcription factor. We developed a flexible computational framework to encode and to combine these promoter features, which allows matching of *cis*-observations to multiple models for a given promoter feature. This enables the description of regulatory elements from different angles and the generation of composite models that can be used to explain the different kinetic behavior of co-regulated genes.

Unlike regulators such as the LacI and MelR (D. C. Grainger, T. W. Overton et al. 2004) proteins of *E. coli* that govern expression of single promoters, many transcriptional regulators control multiple promoters that express products required in different amounts or for different extents of time. This is clearly the case for the regulatory protein PhoP, which controls transcription of a large numbers of genes that can be described by a variety of network motifs (Figure 4.1).

Notably, this approach is not exclusive to regulatory features recognized by the PhoP protein, as the sensitivity for sites recognized by other regulators such as the CRP protein was also increased when submotifs and/or distances were considered instead of a single consensus motif; yet, the specificity remained the same. Indeed, our approach is a meta-method that the proposed is not constrained by a particular sequence analysis method (Table 1).

Our findings argue that understanding a cell's behavior in terms of differential expression of genes controlled by a transcription factor requires a detailed analysis of a promoter's regulatory features. As a single nucleotide difference in the binding site for a transcription factor can dictate the requirement for co-activator proteins (Leung, Hoffmann et al. 2004), we feel that by considering multiple models (as opposed to the relying on consensuses) it will be possible to uncover subtle differences between regulatory targets and to capture the salient properties of co-regulated promoters.

Chapter 5

Gene promoter scan methodology for identifying and classifying co-regulated promoters

The two-component system constitutes a major form of bacterial signal transduction. Typically, a two-component system consists of a sensor kinase that responds to a specific signal by modifying the phosphorylated state of a cognate response regulator. The majority of response regulators are DNA-binding proteins that modulate gene transcription. Because the phosphorylated form of the response regulator binds to target promoters with higher affinity than the unphosphorylated one, sensor-promoted changes in the phosphorylated state of a response regulator can have a profound impact in the gene expression profile of an organism.

Genomic analysis has revealed that there is a direct correlation between genome size and the number of two-component systems present in a given bacterial species. In addition, organisms that live in varied environments tend to have a larger number of two-component systems than those that occupy a single environment. For example, the aphid endosymbiont *Buchnera aphidicola* has a genome size of approximately 640 kb that does not encode two-component systems {Shigenobu, 2000 #495}. In contrast, *Escherichia coli* has a genome size of 4.5 Mb encoding 30 such systems {Blattner, 1997 #494}, and the environmental microbe and opportunistic pathogen *Pseudomonas aeruginosa* with a genome size of 6.3 Mb, harbors 118 two-component system proteins {Stover, 2000 #496}.

The number of targets that a response regulator controls varies among the different systems found in a given bacterial species, and between homologous systems in related bacterial species. In *E. coli*, for example, the response regulator KdpD appears to govern transcription of a single promoter whereas the response regulator ArcA modulates expression of >30 operons {Georgellis, 1999 #493;Salgado, 2004 #344}. Because the products encoded by the multiple targets of regulation of a response regulator such as ArcA are likely required in different amounts and/or for different extents of time, the corresponding genes must differ in their *cis*-acting promoter sequences responsible for the distinct gene expression patterns of individual members of a regulon (i.e. a group of genes that is coordinately regulated by a regulatory protein). The analysis of co-regulated genes is complicated by the fact that two-component systems can control gene expression indirectly, by modulating the expression and/or activity of other two-component systems, transcriptional regulators and sigma factors. Moreover, the targets of regulation of orthologous response regulators overlap only partially in closely related species such as *Salmonella* and *E. coli*, suggesting that small changes in the amino acid sequence of a response regulator and/or in *cis*-acting promoter features can have a big impact on gene regulation. Cumulatively, these issues highlight the need for methods that identify the critical elements of a promoter determining gene expression and that are not heavily dependent on sequence conservation such as phylogenetic footprinting methods (Manson McGuire and Church 2000).

The material required for analyzing the promoter features governing bacterial gene expression is widely available. It consists of genome sequences (often of multiple isolates of a given bacterial species), genome-wide transcription data (typically obtained using microarrays), and biological databases containing examples of previously explored cases. However, it is not yet possible to scan a bacterial genome sequence and readily predict the expression behavior of genes belonging to a regulon. In principle, co-regulated genes could be differentiated by incorporating into the analysis quantitative and kinetic measurements of gene expression (Ronen, Rosenberg et al. 2002) and/or considering the participation of other transcription factors (Bar-Joseph, Gerber et al. 2003; Conlon, Liu et al. 2003; Beer and Tavazoie 2004). However, there are constraints in such analyses due to systematic errors in microarray experiments, the extra work required to obtain kinetic data and the missing information about additional signals impacting on gene expression. These constraints hitherto only allow a relatively crude classification of gene expression patterns into a limited number of classes (e.g., up- and down-regulated genes (Oshima, Aiba et al. 2002; Tucker, Tucker et al. 2002)).

Here, we discuss a methodology designed to identify and classify promoters that are co-regulated by a bacterial transcriptional regulator, such as the response regulators of two-component systems. This methodology, termed GPS for *G*ene *P*romoter *S*can, groups promoters sharing distinct sets of promoter features to generate groupings that may reflect biological properties of a system under investigation such as the time and place that a promoter is activated or silenced.

We have applied the GPS method to investigate the targets of regulation of the response regulator PhoP, which together with the sensor kinase PhoQ form a two-component system that is a major regulator of virulence and of the adaptation to low Mg^{2+} environments in several Gram-negative species (Groisman 2001). The PhoQ protein responds to the levels of extracytoplasmic Mg^{2+} by modifying the phosphorylated state of the DNA-binding protein PhoP {Castelli, 2000 #497;Montagne, 2001 #498;Chamnongpol, 2003 #499}. The PhoP/PhoQ system is a particularly interesting case study because: first, it controls the expression of a large number of genes, amounting to ca. 3% of the genes in the case of *Salmonella* (Kato, Latifi et al. 2003). Second, promoters harboring a binding site for the PhoP protein may differ in the distance and orientation of the PhoP box relative to the RNA polymerase binding site as well as in other promoter features. And third, PhoP also controls gene expression indirectly by regulating the expression/and or activity of other two-component systems at the transcriptional (e.g., RstA/RstB) (Minagawa, Ogasawara et al. 2003), posttranscriptional (e.g., SsrB/SpiR) {Bijlsma, 2005 #490}, and posttranslational (e.g., PmrA/PmrB) (Kato and Groisman 2004) levels. In addition, PhoP regulates the levels of the alternative sigma factor RpoS {Tu, 2006 #482} and participates of a feed-forward loop with the regulatory protein SlyA (Shi, Latifi et al. 2004) (Figure 5.1).

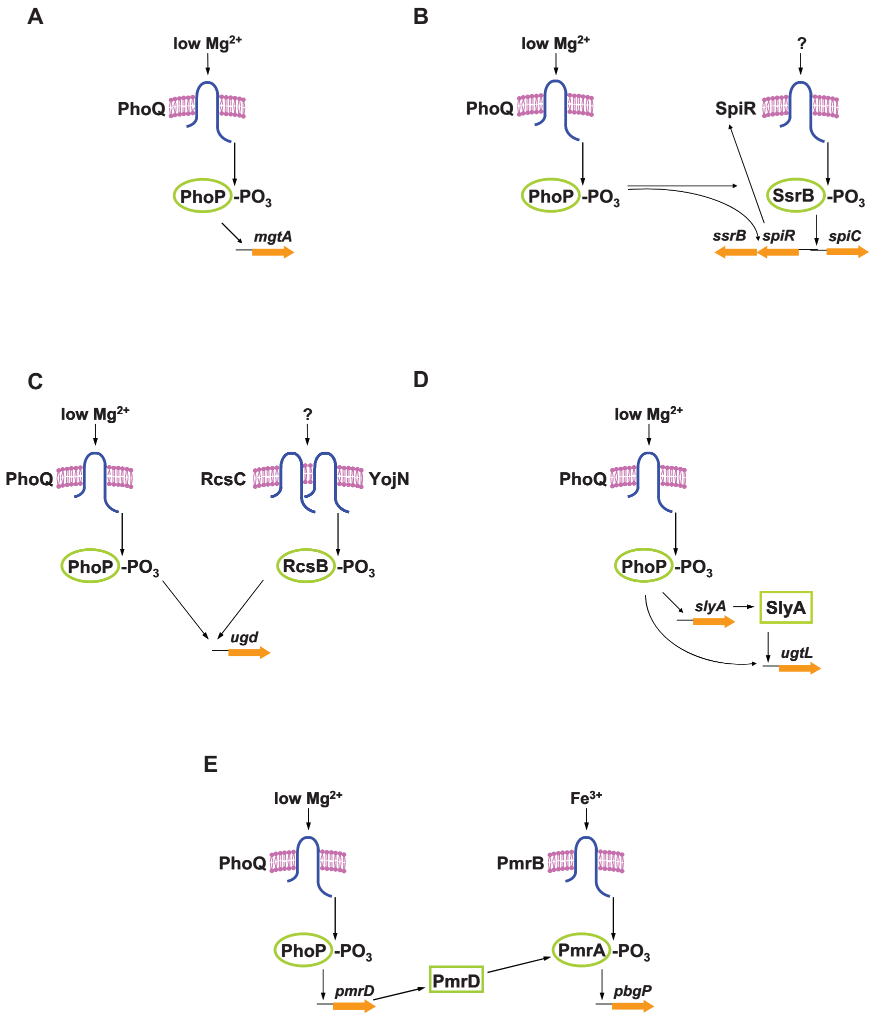


Figure 5.1 The PhoP/PhoQ system controls the expression of a large number of genes, in a direct or indirect fashion.

(a) The PhoP protein recognizes a direct hexanucleotide repeat separated by five nucleotides, which has been termed the PhoP box, activating the *mgtA* promoter of *Salmonella*. (b) The PhoP/PhoQ uses a transcriptional cascade mediated by the SsrB/SpiR two-component system to regulate the *spiC* promoter. (c) The PhoP/PhoQ system cooperatively works with the RcsB/RcsC system to activate the *ugd* promoter. (d) The PhoP/PhoQ system utilizes a feed-forward loop mediated by the SlyA protein to activate the *ugtL* promoter in *Salmonella*. (e) The PhoP/PhoQ system controls the *pbgP* promoter at the posttranslational level, where the PhoP-dependent PmrD protein activates the regulatory protein PmrA.

5.1 Identifying the promoter features governing gene transcription

The identification of the promoter features that determine the distinct expression behavior of co-regulated genes is a challenging task because of the difficulty in ascertaining the role that subtle differences in shared *cis*-acting regulatory elements of co-regulated promoters play in gene transcription. Therefore, approaches that homogenize features among promoters (e.g., relying on consensuses to describe the various promoter features) and even across species can hamper the discovery of the key differences that distinguish promoters that are co-regulated by the same transcriptional regulator. For example, methods that look for matching of a sequence to a consensus have been successfully used to identify promoters controlled by particular transcription factors (Bailey and Elkan 1995; Stormo 2000; Martinez-Antonio and Collado-Vides 2003). Although these methods often increase specificity, their strict cutoffs decrease sensitivity (Hertz and Stormo 1999; Stormo 2000), which makes it difficult to detect binding sites with a weak resemblance to a consensus sequence. The complexity of the analysis is exacerbated by the need to consider other sequence elements relevant to differential expression patterns including the class and location of the RNA polymerase, the presence of binding sites for other transcription factors and their topological location in the DNA (Beer and Tavazoie 2004; Pritsker, Liu et al. 2004). Indeed, similar expression patterns can be generated from different or a mixture of multiple underlying features, thus, making it more difficult to discern the molecular basis for analogous gene expression.

5.2 The GPS methodology as an integrated algorithm

The increased availability of biological information, such as genome sequences, microarray gene expression, as well as text data stored in public databases, and knowledge-discovery techniques (or data mining) used to analyze this information, creates a need to evaluate this information in order to increase the confidence of an uncovered hypothesis. For example, when groups of co-regulated transcripts are identified by clustering the expression patterns generated by a series of microarray experiments, the promoter sequences for each transcript in a cluster may be fed to a motif discovery algorithm to find common elements implicated in transcriptional regulation among co-expressed genes. These approaches incorporate knowledge in a cascade of a decision-making process that can be summarized as follows: find genes with similar expression patterns, and then, see if they have similar promoters {Holmes, 2000 #488}.

Most of the available algorithms implemented by the approaches described above base their decision in cutoffs that constrain one analysis stage on the previous one. Therefore, the analysis is hampered by the need to decide whether to consider first the gene expression data or the promoter features. In addition, the analysis is complicated due to the noisy nature of microarray data, the possibility of cryptic promoter elements contributing to gene expression, the potential for interaction among regulatory proteins, and the existence of alternative modes of transcription regulation, which remain poorly understood.

There are simple algorithms that ignore the constraints listed above, which appear to generate interesting results and be of practical benefits (Tavazoie, Hughes et al. 1999). However, the identification of the promoter features that determine the distinct expression behavior of co-regulated genes within a regulon requires a more detailed and integrated analysis of the regulatory features. Why is it useful to have an integrated model? One reason is that the cascade algorithms and the integrated algorithms are solving subtly different problems. In contrast to the cascade algorithms, the integrated algorithms can be summarized as {Holmes, 2000 #488}: find clusters of genes that have (a) similar expression patterns AND (b) similar promoters.

GPS is a machine learning method (Cooper and Herskovits 1992; Cheeseman and Oldford 1994; Cook, Holder et al. 2001) that identifies, differentiates and groups sets of co-regulated promoters by simultaneously considering multiple *cis*-acting regulatory features and gene expression (Figure 5.2). GPS carries out an exhaustive description of *cis*-acting regulatory features including the orientation, location and number of binding sites for a regulatory protein, the presence of binding site submotifs, and the class and number of RNA polymerase sites (Figure 5.3).

The GPS method specifically aimed at handling the variability in sequence, location and topology that characterize gene transcription. Instead of using an overall consensus model for a feature, where potentially relevant differences are often concealed because of intrinsic averaging operations between promoters and even across species, we decompose a feature into a family of models or building blocks. This approach maximizes the sensitivity of detecting those instances that weakly resemble a consensus (e.g., binding site sequences) without decreasing the specificity. In addition, features are considered using fuzzy assignments (i.e., not precisely defined) instead of categorical entities (Bezdek 1998; Gasch and Eisen 2002; Ruspini and Zwir 2002), which allow us to encode how well a particular sequence matches each of the multiple models for a given promoter feature. Individual features are then linked into more informative composite models that can be used to explain the kinetic expression behavior of genes.

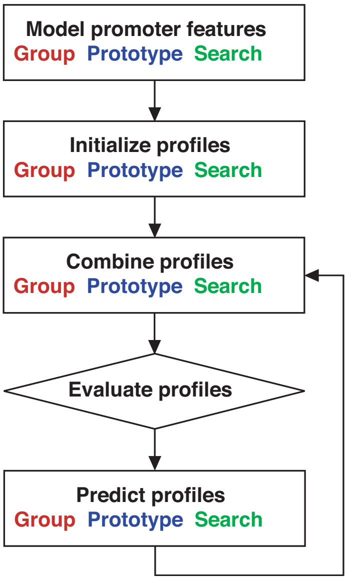


Figure 5.2 . The GPS method

GPS is a machine learning technique that *models* promoter features as well as relations between them, uses them to describe promoters, *combines* such characterized promoters into groups termed profiles, *evaluates* the resulting profiles to select the most significant ones, and performs genome-wide *predictions* based on such profiles. To accomplish this task, GPS carries out three basic operations: *grouping* observations from the dataset; *prototyping* such groups into their most representative elements (centroid); and *searching* in the set of optimal solutions (i.e. Pareto optimal frontier) to retrieve the most relevant profiles, which are used to describe and identify new objects by similarity with the prototypes.

Moreover, GPS treats each of the promoter features with equal weight because it is not known beforehand which features are important. To circumvent limitations imposed by relatively few classes of gene expression levels to *cis*-acting features, the GPS method treats gene expression data as one feature among many. The various features are analyzed concurrently and recurrent relations are recognized to generate profiles, which are groups of promoters having features in common. GPS uses an unsupervised strategy (i.e., pre-existing examples are not required), as well as multiobjective optimization techniques, which enhance the likehood of recovering all optimal feature associations rather than potentially biased subsets (Deb 2001; Ruspini and Zwir 2002). The resulting profiles group promoters that may share underlying biological properties.

5.3 Exploring the targets of regulation of a response regulator using GPS

5.3.1 GPS built-in features.

GPS performs an integrated analysis of promoter regulatory features to identify profiles. We have initially focused on six types of features for describing a training set of promoters (Li, Rhodius et al. 2002; Bar-Joseph, Gerber et al. 2003; Beer and Tavazoie 2004; Zwir, Shin et al. 2005): *"submotifs"*, which models the studied transcription factor binding motifs; *"orientation"*, which characterizes the binding boxes as either in direct or opposite orientation relative to the open reading frame; *"RNA pol sites"*, which characterizes the RNA polymerase motif (Cotik, Zaliz et al. 2005), the class of sigma 70 promoter (Romero Zaliz, Zwir et al. 2004) that differentiates *class I* from *class II* promoters, and (3) the distance distributions (*close*, *medium*, and *remote*) between RNA polymerase and transcription factor binding sites in activated and repressed promoters (Salgado, Gama-Castro et al. 2004); *"activated/repressed"*, where we learn activation and repression distributions by compiling distances between binding sites for RNA polymerase and a transcription factor; *"interactions"*, where we evaluate motifs for several transcription factor-binding sites and model the distance distributions between motifs co-located in the same promoter regions; and *"expression"*, which considers gene expression levels.

DNA Binding Site Motifs: (a) Fix-length Hierarchical Motifs: we modeled the PhoP box motifs by using position weight matrices[1] (Stormo 2000) (see *Consensus matrices* in *gps-tools.wustl.edu*). Then, we used these preliminary models to describe promoters by using low thresholds corresponding to two standard deviations below the mean score obtained with the initial model (Robison, McGuire et al. 1998). We grouped the retrieved observations into subsets by using the possibilistic implementation of fuzzy C-means (PCM) (Bezdek 1998) and re-built matrix models for each one (E-value < 10E-22), thus obtaining several more refined models, and increasing the sensitivity to departures from the consensus. These multiple matrices constitute the prototypes of the feature:

$$M_i(x_1,..,x_K) = \prod_{k=1}^{K} M(x_k) \qquad \textbf{(1)}$$

where $M(x_k)$ is the marginal probability of each nucleotide x_k in the k'th position on motifs of length K, and i indexes a family of prototypes M_i {Barash, 2003 #372}. The degree of matching between an observation and a feature is

[1] A matrix of log-odd score $\log \frac{P(x_k)}{P_0(x_k)}$ where $P_0(x_k)$ is a background distribution.

measured by its similarity with the prototype by using the informational content scores normalized as fuzzy values in the unit interval. The prototypes can be combined and arranged as a multiclassifier (see Bagging consensus in gps-tools.wustl.edu).

(b) Variable-length Motifs: we gathered *sigma* 70 promoters (Salgado, Gama-Castro et al. 2004) from the RegulonDB database and built models of the RNA polymerase site using a neuro-fuzzy method (see Promoter search (CPR-MOSS) in gps-tools2.wustl.edu), and used the resulting models to perform genome-wide descriptions of the intergenic regions of the *E. coli* and *Salmonella* genomes with a false discovery rate <0.001. The time delay neural network constitutes the feature prototype (Ruspini and Zwir 2002) and the scores were also normalized.

Transcription Factor Binding Site Orientation: categorical data. We classified PhoP boxes as either in direct or opposite orientation relative to the open reading frame (Fig. 1d), and the prototype is a simple Boolean function.

RNA Polymerase Distances: data distributions modeled as fuzzy sets. We built histograms with the distance between RNA polymerase and transcription factor from information available in RegulonDB database {Salgado, 2004 #344}. We encoded these distributions by using fuzzy set representations {Ruspini, 2002 #427} into close, medium, and remote sets. These fuzzy sets constitute the prototypes of the feature, and can be viewed as approximation of data distributions:

$$D_i(x) = \begin{cases} 0 & if\ x < a_0\ or\ x > x_2 \\ (x - a_0)/(a_1 - a_0) & if\ x < a_1 \\ (a_2 - x)/(a_2 - a_1) & if x > a_2 \\ 1 & otherwise \end{cases} \qquad \textbf{(2)}$$

where x is any distance between the transcription start site of an RNA polymerase binding site and the center of a transcription factor binding site, and *i* indexes a family of distances D_{close}, D_{medium} and D_{remote}. Initial partitions are learned from the projection of the histograms onto the variable domains by simple regression and minimum squared methods {Sugeno, 1993 #60}. The degree of matching between an observation and a prototype is calculated by specializing a value in a triangular fuzzy membership functions {Klir, 1988 #67}.

Microarray Expression Data: collection of fuzzy sets encoded as a fuzzy centroid. We clustered PhoP-regulated gene expression levels (Fig. 1f) by using PCM and built models for each cluster by calculating its centroid. These models represent the prototypes, where the values of the expression feature for each promoter in E. coli is calculated by its similarity to the centroids $\overline{V_i}$ as a vector of fuzzy sets:

$$E_i\,(x)=\left[1+\left(\left\|x-\overline{V_i}\right\|_A^2\Big/w_i\right)^{1/m-1}\right]^{-1} \tag{3}$$

where $x=\{x_1,...,x_k\}$ corresponds to the expression of a gene in k microarray experiments; w_i is the "bandwidth" of the fuzzy set E_i; m is the degree of fuzzification which is initialized as 2; the type of norm, determinated by A, is Pearson correlation coefficient; and i indexes a family of prototypes E_i.

Composite Features. We combine several features with dependencies between each other into more informative models by using AND-connected fuzzy predicates:

$$C(F_i\;,F_j\,)=F_i{}^{\;i}\;ANDF_j\;=F_i\;\cap F_j \tag{4}$$

where F_i and F_j are previously defined features. Fuzzy logic-based operations, such as *T-norm/T-conorm*, include operators like MINIMUM, PRODUCT, or MAXIMUM, which are used as basic logic operators, such as AND or OR, or their set equivalents INTERSECTION or UNION {Bezdek, 1998 #43;Klir, 1988 #67}. In this work we used the MINIMUN and MAXIMUM as *T-norm* and *T-conorm*, respectively. For example, the RNA polymerase motif, learned by using a neural network method, it *sigma class*, identified by using an intelligent parser that differentiates *class* I from *class* II promoters, and the distance distributions $(D_{close}, D_{medium}, D_{remote})$ between RNA polymerase and transcription factor binding sites, learned by using fuzzy set representations {Ruspini, 2002 #427}, are normalized and combined into a single fuzzy vector (e.g., $P_i(x)=R_j\;\;AND\;\;D_k\;\;AND\;\;T_l$).

5.3.2 GPS initialization strategy.

GPS takes a list of candidate genes obtained from the literature, gene expression experiments (e.g., microarray, ChiP, or RT-PCR) or user-based hypothesis, and generates initial profiles of each individual type of feature. Then, it generates a set of single-type input profiles from this information. The generation of initial profiles increases the sensitivity of a feature without decreasing its specificity (Zwir, Shin et al. 2005). This distinguishes GPS from methods relying a single consensus, which often fail to describe and retrieve potentially interesting candidates that exhibit a weak resemblance to an average consensus pattern, which may be construed as a gene being indirectly regulated by a transcription factor (Zwir, Shin et al. 2005).

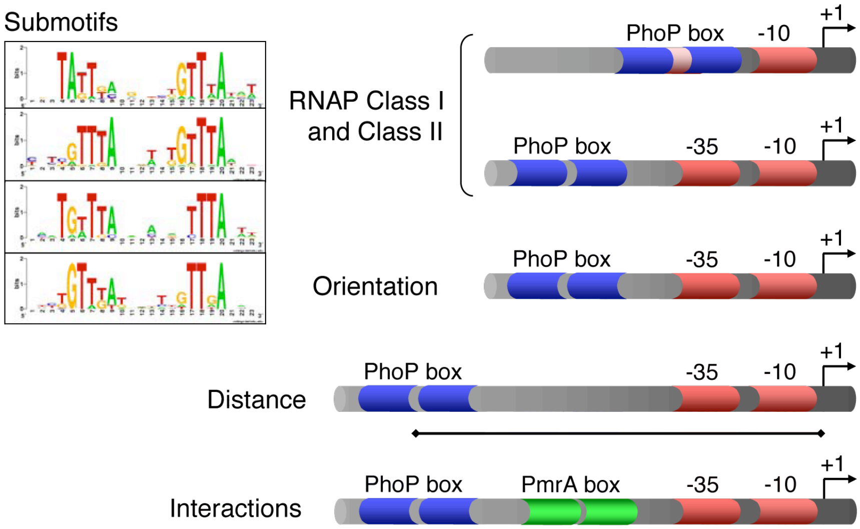


Figure 5.3 Schematics of PhoP-regulated promoters harboring different features analyzed by GPS GPS performs an integrated analysis of promoter regulatory features initially focusing on six types of features for describing a training set of promoters: *"submotifs"*, which models the studied transcription factor binding motifs; *"RNA pol sites"*, which characterizes the RNA polymerase motif, the class of sigma 70 promoter that differentiates *class I* from *class II* promoters, and the distance distributions (*close*, *medium*, and *remote*) between RNA polymerase and transcription factor binding sites in activated and repressed promoters; *"activated/repressed"*, where we learn activation and repression distributions by compiling distances between binding sites for RNA polymerase and a transcription factor; *"interactions"*, where we evaluate motifs for several transcription factor-binding sites and model the distance distributions between motifs co-located in the same promoter regions; and *"expression"*, which considers gene expression levels.

The input data should be specified according to each type of feature, that is, DNA sequences for the "*submotifs*", and if available, the gene expression levels for the "*expression*" (see online manual at http://gps-tools2.wustl.edu). The initial models for each feature can also be provided by the user. For example, GPS uses position weight matrices generated by the Consensus/Patser method (Stormo 2000); but also can accept any other built-in matrix generated from other methods (Tompa, Li et al. 2005). Indeed, the number of profiles can be a priori specified or automatically calculated by using the Xie-Beni index (Cotik, Romero-Zaliz et al. 2005). Although the specification of these initial conditions are crucial for clustering algorithms (Bezdek, Pal et al. 1992), they are not critical for GPS and can be later solved by the dynamic approach followed by the method (Zwir, Shin et al. 2005).

Two or more promoter regions containing different binding sites for a given transcription factor are considered as distinct instances, which can be later associated by the method as more features become incorporated into the analysis. Indeed, GPS considers promoters independently of phylogenetic conservation. Therefore, after dissecting direct and indirect regulation, each instance in the database is constrained to a promoter region where a binding site motif of the studied transcription factor is found.

One of the most salient properties of the strategy followed by GPS to encode features is the use of metadata. Thus, GPS can encode features as fuzzy data (i.e., not precisely defined) instead of categorical entities {Gasch, 2002 #38;Ruspini, 2002 #427;Bezdek, 1998 #43}, where a promoter instance can be related to more than one model. This captures the variability that exists in biological systems and delays the grouping of promoters until more information (i.e., features) are added. For example, a sequence corresponding to a transcription factor binding site can be initially similar to both "*submotif*" M_1 and "*submotif*" M_2. Later, it could be assigned to a profile containing M_1 after adding the "*orientation*" and the "*RNA pol site*" features. Moreover, if M_1 or M_2were initially designed inadequately, the partial matching of several promoters with both profiles can generate new intermediate profiles by taking advantage of the implementation of the profiles as fuzzy clusters {Bezdek, 1998 #43}.

GPS also uses metadata to analyze composite features. It could be the case, for example, that two or more features would not be independent of each other. Thus, GPS joins them by using fuzzy predicates (i.e., such as P(A and B) in a probabilistic interpretation). Indeed, the distance between the binding sites for RNA polymerase and for a transcription factor is meaningless if one does not consider the occurrences of the sites.

5.3.3 GPS grouping strategy.

GPS groups profiles by navigating in a lattice corresponding to the feature searching space {Cook, 2001 #95;Cooper, 1992 #346} and systematically creating compound higher level profiles (i.e., offspring profiles) based on combining parental profiles, by taking the fuzzy intersection (Fig. 1h). For example: level-1: (E_1^1 ,M_2^1 and P_3^1) $\mapsto$ level-2: ($E_1^2 M_2^2$, $M_2^2 p_3^2$ and $E_1^2 P_3^2$) $\mapsto$ level-3: ($E_1^3 M_2^3 P_3^3$), where level-3-profiles are obtained from intersection of the promoter members of level-2- profiles (e.g., $E_1^2 M_2^2$, $M_2^2 P_3^2$ and $E_1^2 P_3^2$) and not between those belonging to the initial profiles (E_1^1 ,M_2^1 and P_3^1). This is because our approach dynamically re-discretizes the original features at each level and allows re-assignations of observations between sibling profiles. In this hierarchical process, each level of the lattice increases the number of features shared by a profile (Figure 5.4). After searching through the whole lattice space, the most specific profiles (i.e., the most specific hypothesis (Mitchell 1997)) are found. As a result of this strategy, one promoter observation can contribute to more than one profile in the same or a different level of the lattice, with different degrees of membership (Figure 5.5). This differentiates our approach from a hierarchical clustering process where, once an observation is placed in a cluster, it can only be re-assigned into offspring clusters. In contrast, our approach is similar to optimization clustering methods (Falkenauer 1998) in that it allows transfers among sibling clusters in the same level.

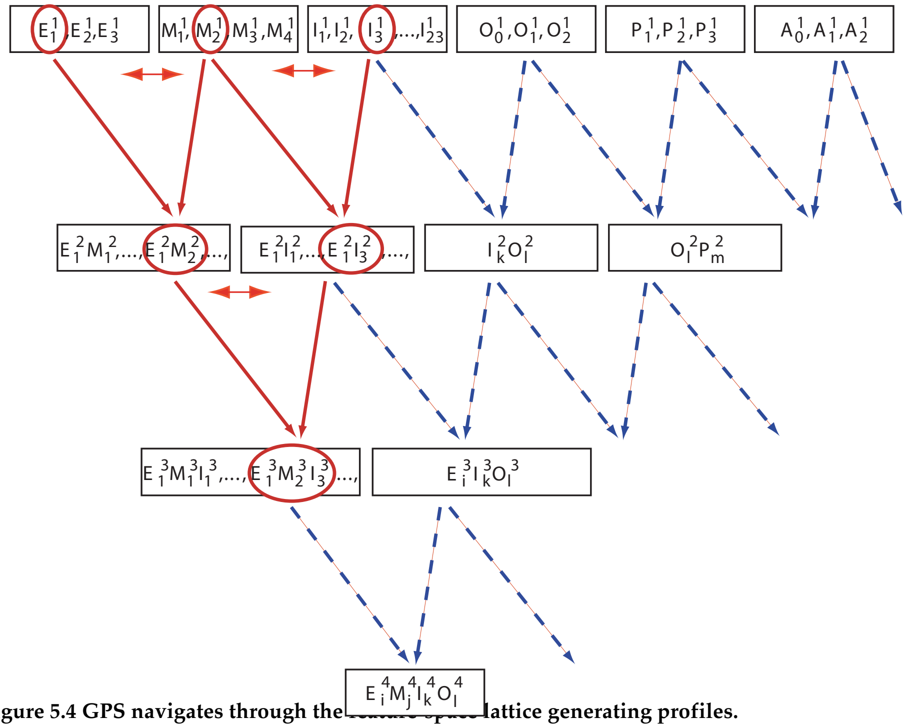

E_1^1, E_2^1, E_3^1
$M_1^1, M_2^1, M_3^1, M_4^1$
$I_1^1, I_2^1, I_3^1, \ldots, I_{23}^1$
O_0^1, O_1^1, O_2^1
P_1^1, P_2^1, P_3^1
A_0^1, A_1^1, A_2^1
$E_1^2M_1^2, \ldots, E_1^2M_2^2, \ldots,$
$E_1^2I_1^2, \ldots, E_1^2I_3^2, \ldots,$
$I_k^2O_l^2$
$O_l^2P_m^2$
$E_1^3M_1^3I_1^3, \ldots, E_1^3M_2^3I_3^3, \ldots,$
$E_i^3I_k^3O_l^3$
$E_i^4M_j^4I_k^4O_l^4$

Figure 5.4 GPS navigates through the feature space lattice generating profiles.

For analysis of promoters regulated by the PhoP protein, we identified up to five models for each type of feature, which are used to describe the promoters. Then, GPS generates profiles, which are groups of promoters sharing common sets of features. (The subscripts denote the different profiles for each feature, the superscripts denote the level in the lattice of the profile). For example, E_1^1 is a particular "expression" profile that differs from E_2^1 and E_3^1. These level-1 profiles of each feature are combined to identify level-2 profiles, and similarly, level-2 profiles are combined to create level-3 profiles. In addition, because of the fuzzy formulation of the clustering, any promoter that was initially assigned to a specific profile E_i^t, can participate in profile of level-t where E_j^t is involved (i.e. indicated as a double-headed arrow). Thus, observations can migrate from parental to offspring clusters (i.e. hierarchical clustering), and among sibling clusters (i.e. optimization clustering). Here, we show a small part of the complete lattice, where the part that is highlighted in red is also described in (**Figure 5.5**).

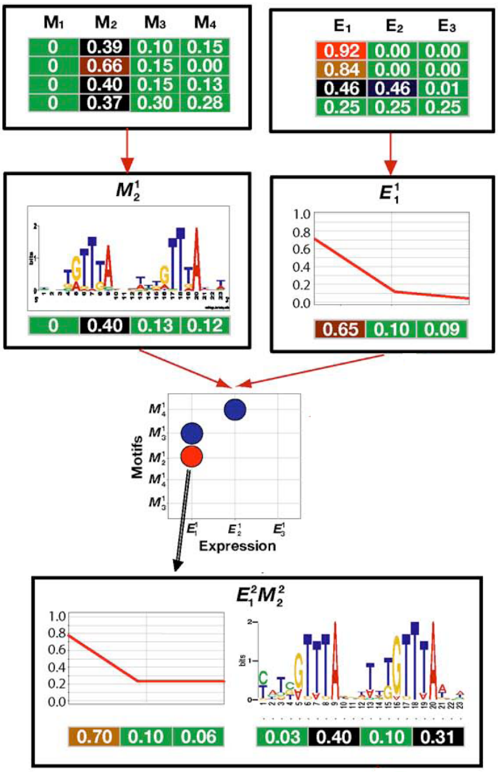


Figure 5.5 Using GPS to build promoter profiles

GPS generation of the $E_1^2M_2^2$ profile is shown here. It partially corresponds to the highlighted substructure of the lattice in (**Figure 5.4**). GPS starts by using information from databases and microarray data to construct a family of models for each feature (e.g., expression levels E1 to E3, PhoP box submotif M1 to M4, as well as other features (not shown). The promoters are described using the modeled features, the degree of matching between features and promoters being encoded as a vector of independent values, where 1 (red color) corresponds to maximum matching and 0 (green color) corresponds to the absence of the feature. For each feature, the promoters are then grouped into subsets that share similar patterns using fuzzy clustering. Each subset shown in the initial panel is prototyped by locating the centroid that best represent the group, to generate the initial, level-1 profiles (e.g., E_1^1, M_2^1, and E_3^1). The centroids are encoded as a vector, and also visualized by graphical plots for the "expression" and the "interactions" features, and by a sequence logo {Crooks, 2004 #356} for the "submotifs" feature. These level-1 profiles are combined to generate level-2 profiles (e.g., $E_1^2M_2^2$ and $M_2^2I_3^2$ (red circles)), by the intersection of the ancestor profiles and then prototyped. (Blue circles represent profiles containing other subsets of promoters. The absence of a circle signifies that no promoters are classified into these profiles). Further navigation through the feature-space lattice generates the level-3 profiles, e.g., $E_1^3M_2^3I_3^3$ after incorporating the "interactions" feature (**Figure 5.4**). Note that the vectors of the daughter profiles are built anew from the constituent promoters, and slightly different than those of their ancestors, due to the refinement that takes place during the profile learning process.

5.3.4 GPS evaluation strategy.

The profile searching and evaluation process is carried out as a multiobjective optimization problem (Rissanen 1989; Deb 2001; Ruspini and Zwir 2002), which must consider conflicting criteria: the extent of the profile, the quality of matching among its members and the corresponding features, and its diversity (Cook, Holder et al. 2001; Ruspini and Zwir 2002). This strategy allows the identification of sets of optimal - instead of single or maximum estimated - profiles as models of alternative hypotheses describing distinct regulatory scenarios.

The extent of a profile is calculated by using the hypergeometric distribution that gives the chance probability (i.e., probability of intersection (PI)) of observing at least *p* candidates from a set V_i of size *h* within another set V_j of size *n*, in a universe of *g* candidates:

$$PI(V_{i,j}) = 1 - \sum\nolimits_{q=0}^{p} \binom{h}{q}\binom{q-h}{n-q} \Big/ \binom{g}{h} \qquad (5)$$

where V_i is an alpha-cut of the offspring profile and V_j is an alpha-cut of the union of its parents. The PI (Tavazoie, Hughes et al. 1999) is a more informative measure than the number of promoters belonging to the profile, such as the Jaccard coefficient, in being an adaptive measure that is sensitive to small sets of examples, while retaining specificity with large datasets.

The quality of matching between promoters and features of a profile (i.e., similarity of intersection (SI)) is calculated using the following equation:

$$SI(V_i\ ) = \frac{1}{f}\left(1 - \frac{\sum\nolimits_{k \in U_\alpha} \mu_{ik}}{n_\alpha}\right) U_\alpha = \{\mu_{ik} : \mu_{ik} > \alpha\} \qquad (6)$$

where n_α is the number of elements in an arbitrary *alpha*-cut U_α.

The tradeoff between the opposing objectives (i.e., PI and SI) is estimated by selecting a set of solutions that are non-dominated, in the sense that there is no other solution that is superior to them in all objectives (i.e., Pareto optimal frontier) (Deb 2001; Ruspini and Zwir 2002). The dominance relationship in a minimization problem is defined by:

$$a \prec b\, iif\ \forall i\ O_i\,(a) \le O_i\,(b)\ \exists j\, O_j\,(a) < O_j\,(b) \qquad (7)$$

where the O_i and O_j are either PI or SI. The method applies the non-dominance relationship only to profiles in the local neighborhood or niche (Deb 2001) (Figure 5.6) by using the hypergeometric metric (equation (5)) between profiles and selecting an arbitrary threshold; in this way combining both multi-objective and multimodal optimization concepts (Deb 2001)

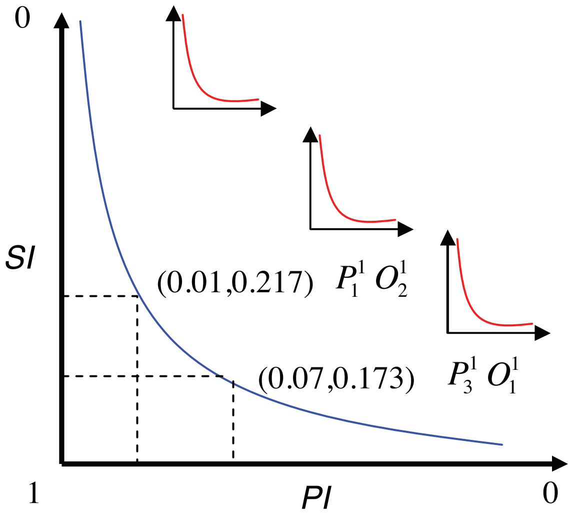


Figure 5.6 Pareto optimal frontier.

GPS evaluates profiles based on a tradeoff between the opposing objectives (i.e. PI and SI), which is estimated by selecting a set of solutions that are non-dominated (i.e. Pareto optimal frontier (blue line)), in the sense that there is no solution that is superior to the others in all objectives (e.g., in a minimization optimization of the objectives, $P_1^2O_2^2$ and $P_3^2O_1^2$ profiles are retained as solutions because although ($PI(P_1^2O_2^2) = 0.01 < PI(P_3^2O_1^2) = 0.07$)), $SI(P_1^2O_2^2) = 0.215 > SI(P_3^2O_1^2) = 0.173$). GPS retrieves solutions that are locally nondominated (red lines) in a neighborhood of solutions or niches, by combining multiobjective and multimodal optimization concepts. Therefore, by maintaining several niches in different zones of the feature searching space with optimal solutions, GPS can describe the system from different points of view.

5.3.5 GPS validation strategy.

GPS is an unsupervised method that does not need the specification of output classes, which is in contrast to supervised approaches (Zwir, Shin et al. 2005). Thus, the discovered profiles can be used for independently explaining external classes as a process often termed labeling (Mitchell 1997). These classes can be introduced as a control in GPS, which automatically correlates them with the obtained profiles. For example, GPS uses the expression as one feature among many often derived from constrained microarray gene expression experiment that just distinguishes between up and down-regulated genes (e.g., mutant vs. wild-type conditions). However, the posterior availability of more discriminating classes, such as those derived from time-dependent ChiP experiments, can be used as an external phenomenon to be explained by the learned profiles.

5.3.6 Predicting new members

GPS uses a fuzzy k-nearest prototype classifier (FKN) to predict new profile members using an unsupervised classification method (Bezdek 1998) applied to regulatory regions of genomes described by regulatory features. First, we determine the lower-boundary similarity threshold for each non-dominated profile. This threshold is calculated based on the ability of each profile to retrieve its own promoters and to discard promoters from other profiles (Benitez-Bellon, Moreno-Hagelsieb et al. 2002). Second, we calculate the membership of a query observation x_q to a set of k profiles previously identified and apply a fuzzy OR logic operation:

$$FKN(x_q, V_1, ..., V_k) = i,\ i \in \{1,..,k\} \qquad \textbf{(8)}$$

where $\mu_{i,q} = OP_{OR}\{\mu_{1,q},...,\mu_{k,q}\}$, μ is calculated based on (equation **(4)**) in which w_i (equation **(3)**) is initialized as:

$$w_i = \frac{r_1\, PI(V_i) + r_2\,(f/t')SI(V_i)}{r_1 + r_2} \qquad \textbf{(9)}$$

with t' being the number of distinct features observed in x_q and V_i, and f is the number of features in common between x_q and V_i, which are combined to obtain a measure of belief or rule weight (Cooper and Herskovits 1992); r_1 and r_2 are user-dependent parameters, initialized as 1 if no preference exists between both objectives; and OP_{OR} is the Maximum fuzzy operator (Bezdek 1998; Gasch and Eisen 2002).

5.4 Technical specifications of GPS

5.4.1 Programming resources.

The GPS system has been implemented to be a platform-independent method, with a flexible and fast performance. It combines various machine learning techniques, implemented in cohesive programming languages and frameworks to satisfy these non-functional requirements. The software consists of a core application, which executes both sequentially as well as in parallel fashion on a cluster of computers; and two remote interfaces: a light web front end user interface developed in php, which accepts user's input and e-mails results; and a web service interface coded in java.

5.4.2 User interface

Data definition and parameters are specified to the system as a single XML document {Wang, 2005 #489}. This standard provides the required flexibility and readability to allow the specification of the database, features and initial profiles. An XML schema is provided (http://gps-tools2.wustl.edu/gps/gps.xsd) to verify the document and to facilitate its editing. Apache Tomcat and Apache Axis are used to provide Web service interface, allowing application-to-application interaction in a standardized fashion (http://gps-tools2.wustl.edu:8080/gps).

The core system is coded in java independent platform. Advanced java virtual machines with adaptive and just-in-time compilation and other techniques now typically provide performance up to 50% to 100% the speed of C++ programs {Lindsey, 2005 #491}. We also encapsulated the execution of existing position weight matrices software by developing ad-hoc scripts in perl scripting language.

5.4.3 Parallel execution

Components that require a large amount of processing power are executed in parallel in a High-Performance Computing environment provided by Condor High throughput computing workload management system {Basney, 1999 #492}; which administers batch jobs on clusters of dedicated computing resources.

5.4.4 GPS input.

GPS captures the input specifications by an XML file that contains two parts. The first part correspond to the specifications of the features, while the second corresponds to the database composed of the promoter values for the features.

5.4.5 Feature specifications.

Here we describe examples of several features. The complete manual is on line at http://gps-tools2.wustl.edu.

Purpose: representing DNA binding site submotifs

Syntax

<Feature type="sequence" name="submotif" >

Indicates that the input data can be a DNA sequence or a position weight matrix containing a motif, and its name, which must be unique.

<Bin name="M_1" membershipFile="gps_data/M_1.mat" fileType="mat" />

Specifies one input bin (i.e., submotif) that was previously clustered and pre-processed as a position weight matrix and stored in a file termed *M_1* with extension "mat" located in a user defined directory.

<Bin name="M_2">

<InitialMember name="mgtC_681

<InitialMember name="mgtC_718"/>

<InitialMember name="mgtC_925"/>

</Bin>

Specifies a bin containing the name of the promoters belonging to a desired submotif, which are used to automatically calculate the initial matrices if they are not available.

<Bin cluster="n">

If a single bin is proposed, GPS automatically clusters the instances into (*"n"*) bins. If the number of clusters is null (""), GPS uses the Xie-Beni index to calculate the initial number of clusters.

Description: GPS takes initial lists of candidate promoter sequences for a specific transcription factor and clusters them by using Fuzzy C-Means algorithm into bins. Each of these bins are further encoded as position weight matrices by using the Consensus/Patser method and used as a single-type initial profiles. If the number of clusters is not specified, GPS uses the Xie-Beni index to provide the corresponding number. Matrices provided by other methods (e.g., MEME) can be also directly incorporated as input data. The initial bins would be dynamically re-formulated when new features were aggregated.

Purpose: representing microarray gene expression

Syntax

<Feature type="expression" name="Expression" >

Indicates that the input data can be a vector of continuous values corresponding to levels of gene expression resulting from one or more experiments. The name must be unique.

<Bin name="E_1" membershipFile="gps_data/E_1.exp" fileType="exp" />

Specifies one input expression bin, where columns are distinct experimental or time conditions, that was previously clustered and pre-processed as a prototype (i.e., centroid or array of real numbers) and stored in a file termed *E_1* with extension "exp" located in a user defined directory.

<Bin name="E_2">

<InitialMember name="mgtC_681

<InitialMember name="mgtC_718"/>

<InitialMember name="mgtC_925"/>

</Bin>

Specifies a bin containing the name of the promoters belonging to a desired expression profile, which is used to automatically calculate the corresponding initial prototype.

<Bin cluster="n">

If a single bin is proposed, GPS automatically clusters the instances into (*"n"*) profiles. If the number of clusters is null (""), GPS uses the Xie-Beni index to calculate the initial number of profiles.

Description: GPS takes lists of candidate genes, where columns indicate different or time dependent experiments. GPS clusters them by using Fuzzy C-Means algorithm into bins. Each of these bins are further encoded as a centroid and used as a single-type initial profiles. If the number of clusters is not specified, GPS uses the Xie-Beni index to provide the corresponding number. The initial profiles would be dynamically re-formulated when new features were aggregated.

Purpose: representing the orientation or topological order of a regulatory element

Syntax

<Feature type="value" name="Orientation" deviation_factor="0.5">

Indicates that the input data can be a continuous/integer value, which represents continuous or discrete events, respectively. For example, the orientation of a binding site relative to the open reading frame (e.g., direct or opposite) or the topological order of a regulatory element regarding another (e.g., in front off or behind). The prototypes are uniformly discretized according to the *deviation factor*. For example, choosing a partition with three values p_0, p_1 and p_2, GPS establishes that p_1 will be the central value and $p_{0,2} = p_1 \pm df \times stdev$.

Description: GPS takes lists of promoters characterized by a discrete or continuous values (e.g., direct=0 and indirect =1; repressed = 0, activated = 1, and fuzzy activated or repressed = 0.5). The method clusters them by using Fuzzy C-Means algorithm into bins. Each of these bins are further encoded as prototypes calculated as specified in the syntax section.

Purpose: representing fuzzy features

Syntax

<Feature type="fuzzy" name="Fuzzy_Motif" input type="sequence" interpretation= "possibilistic|fuzzy">

Indicates a metadata that encodes the degree of matching between an instance and several profiles (i.e., the similarity between instances and the prototypes that represent the profiles). The input can accept different types of features: "sequence", "expression", etc. The encoding method could be fuzzy or possibilistic (i.e., membership to all profiles do not have to sum 1).

<Bin name="MF_1" centroid="[0.652 0.036 0.160 0.528]" wi="0.5339"(default=1)/>

<Bin name="MF_2" centroid="[0.808 0.284 0.348 0.174]" wi="0.1264"/>

<Bin name="MF_3" centroid="[0.433 0.229 0.422 0.625]" wi="0.1015"/>

The initial profiles (e.g., submotifs) are encoded as vectors of continuous values (*centroids*) that represent the averaged similarity of their members to a feature submodel (e.g., the position weight matrix of submotif *M_1*). Indeed, the vector contains the similarity values of the profile members to all other single-type profiles. The *wi* values correspond to the amplitude of the fuzzy clusters defined for each centroids.

<Bin name="MF_4" centroid="" wi="" membershipFile= "gps_data/M_1.mat" ... membershipFile= "gps_data/M_4.mat"/>

If the centroids are not specified, GPS calculates them based on the profiles defined in *membershipFile*, and adjust the amplitude of the fuzzy cluster based on the *wi* parameter.

Description: GPS allows each promoter instance to belong to multiple profiles in parallel, by encoding into a metadata its degree of similarity to all profile-prototypes. Then, these membership values can be considered by GPS, instead of the original data, during the learning phase of the method. This codification allows to represent different types of input data (e.g., expression, sequences) into the same framework composed of numeric vectors. Moreover, this approach allows to encode intermediate classes that were not initially specified (e.g., the expression class representing the concept "between high and medium" corresponding to those genes which expression is consistent with both levels of expression: high and medium).

5.4.6 GPS output.

The output of the program is composed of four main sections: the XML file submitted by the user, the list of explored profiles, the selected non-dominated profiles, and a snapshot matrix designed easily export the results into a typical spreadsheet or into the Spotfire environment (Wilkins 2000).

List of profiles. This section is identified by the tag "+Profiles:" and enumerates all profiles in the lattice of potential hypothesis. The name of a profile corresponds to the abbreviated names of the features contained in that profile (e.g. the profile named orientation_i.expression_j.motif_k is composed of the re-discretized version of the original features and i, j and k correspond to the initial single-type profiles. The values corresponding to the evaluation of the profiles is dumped as the probability of intersection (i.e., profile extent evaluated by the probability of the features intersection (PI)) and similarity degree of matching between promoters and the prototypes of the profile (SI). Finally, we describe each profile by listing its features, its prototype or centroid and the recovered promoters, indicating name, feature values and evaluation score.

Dominance relationship: The start of this section is indicated by the "+Dominance tag", where each profile is described by name, PI and SI scores, and a tag indicating if it is either dominated or non-dominated. Profiles containing unique promoters are not considered. Dominated profiles also contain a list of the their dominating profiles.

Snapshot matrix. This section is identified by the "+Matrix tag", where columns represent profiles and rows correspond to the profile name, the domination status, the PI and SI values, the number of promoters recovered by the profile, the number of features that characterize the profile, and finally, the membership value of all of the promoters to the profile.

5.5 Uncovering promoter profiles regulated by the response regulator PhoP using GPS

We examined the genome-wide transcription profile of wild-type and *phoP E. coli* strains experiencing low Mg^{2+}, and identified genes whose expression differed statistically between the two strains (Li and Wong 2001; Tusher, Tibshirani et al. 2001). We used these genes, as well as *Salmonella enterica* promoters suspected to be regulated by PhoP, which were provided from our own lab knowledge and the literature to generate the initial list of promoter candidates.

We utilized this list to make the initial models of the features, which were used with relaxed thresholds (Hertz and Stormo 1999) to describe promoters with weak matching to consensus. For example, GPS clustered these genes by their expression similarity: E_1 and E_2, consisting of up-regulated genes; and E_3, harboring down-regulated genes. Then we classified all candidates based on the similarity of their expression to that of models built for each of the three expression groups, permitting individual genes to belong to more than one group (i.e., E_1 and/or E_2) (Bezdek 1998; Gasch and Eisen 2002). This enabled us to recover weakly expressed genes that would have otherwise gone undetected using strict statistical filters (Li and Wong 2001; Tusher, Tibshirani et al. 2001). GPS applied the same strategy to the other features. The initial submotifs corresponding to the PhoP binding site were dissected by GPS, allowing the recovery of PhoP-regulated promoters with weak matching to the PhoP box consensus, such as the *Salmonella pmrD* promoter, that could not be detected using consensus cutoffs (Hertz and Stormo 1999; Stormo 2000) despite being regulated and footprinted by the PhoP protein (Kox, Wosten et al. 2000; Kato, Latifi et al. 2003).

5.5.1 Experimentally validated profiles

We use several features for the initial profiles including: discrimination of PhoP box submotifs (M_1- M_4), the orientation (O_1- O_2) and distance of the PhoP box relative to the RNA polymerase site (P_1- P_3), the class of sigma 70 promoter (because sigma 70 is responsible for transcription of PhoP-regulated genes (Yamamoto, Ogasawara et al. 2002)) (P_1- P_3), the presence of potential binding sites for 60+ transcription factors (Salgado, Santos-Zavaleta et al. 2001) (I_0- I_4) and whether the position of the PhoP box suggests a promoter is activated or repressed (A_1- A_3). Then, GPS applied its *g*rouping, *p*rototyping and *s*earching strategy and uncovered several optimal profiles, which were experimentally validated (Zwir, Shin et al. 2005).

One of the profiles identifies profiles with canonical PhoP-regulated promoters. This profile, $P_1^4 E_1^4 M_2^4 I_3^4$ (PI=0.39, SI=0.07), encompasses promoters (e.g., those of the *phoP, mgtA, ybcU* and *yhiW* genes of *E. coli* and the *slyB* gene of *Salmonella*) that share the same RNA polymerase sites, expression patterns, PhoP box submotif, and the same pattern for other transcription factor binding sites. The profile includes not only the prototypical *phoP* and *mgtA* promoters (Minagawa, Ogasawara et al. 2003), but also the promoters of the *yhiW* gene, which was not known to be under PhoP control.

Another profile describe promoters with PhoP boxes in the opposite orientation of the canonical PhoP-regulated promoters. This profile, $P_3^2 O_1^2$ (*PI*=0.07, *SI*=0.17), includes promoters also with the PhoP box in the opposite orientation (e.g., those of the *slyB* and *yhiW* genes of *E. coli* and the *ybjX, mig-14, virK, mgtC,* and *pagC* genes of *Salmonella*) but differs from the former profile in that the PhoP box is located further upstream from the RNA polymerase site than the typical PhoP-regulated gene. Notably, these promoters could be assigned to a profile even in the absence of expression data. Despite the unusual orientation of the PhoP box in these promoters, the identified PhoP boxes are *bona fide* PhoP-binding sites (Shi, Latifi et al. 2004; Shin and Groisman 2005; Zwir, Shin et al. 2005). Curiously, it had been suggested that PhoP regulates these genes of *Salmonella* indirectly, because a PhoP binding site could not be identified at a location typical of other PhoP-activated genes (Lejona, Aguirre et al. 2003).

By using gene expression as one feature among many, GPS could distinguish between promoters of the acid resistance genes (Tucker, Tucker et al. 2002; Masuda and Church 2003) that, otherwise, would have stayed undifferentiated within the same expression group. These promoters were found to belong to one of three distinct profiles: $E_2^3 M_0^3 I_1^3$ (PI=0.11, SI=0.03), includes promoters for acid resistance structural genes lacking a recognizable PhoP box (e.g., those of the *dps* and *gadA* genes of *E. coli*); $E_2^2 M_4^2$ (PI=0.25, SI=0.10), comprises promoters of a different set of structural genes that include *hdeD* and *hdeAB*; and $E_2^2 P_3^2$ (PI=0.419, SI=0.185), harbors promoters of the acid resistance regulatory genes *yhiE* and *yhiW* (also termed *gadE* and *gadW*, respectively. The promoters in the latter two profiles harbor PhoP boxes but these profiles differ in the RNA polymerase sites and their distance to the PhoP box. The promoters in the latter two profiles harbor PhoP boxes but these profiles differ in the RNA polymerase sites and their distance to the PhoP box. These findings enabled the prediction that PhoP uses at least two modes of regulation to control transcription of acid resistance genes: a feedforward loop and classical transcriptional cascade (Zwir, Shin et al. 2005).

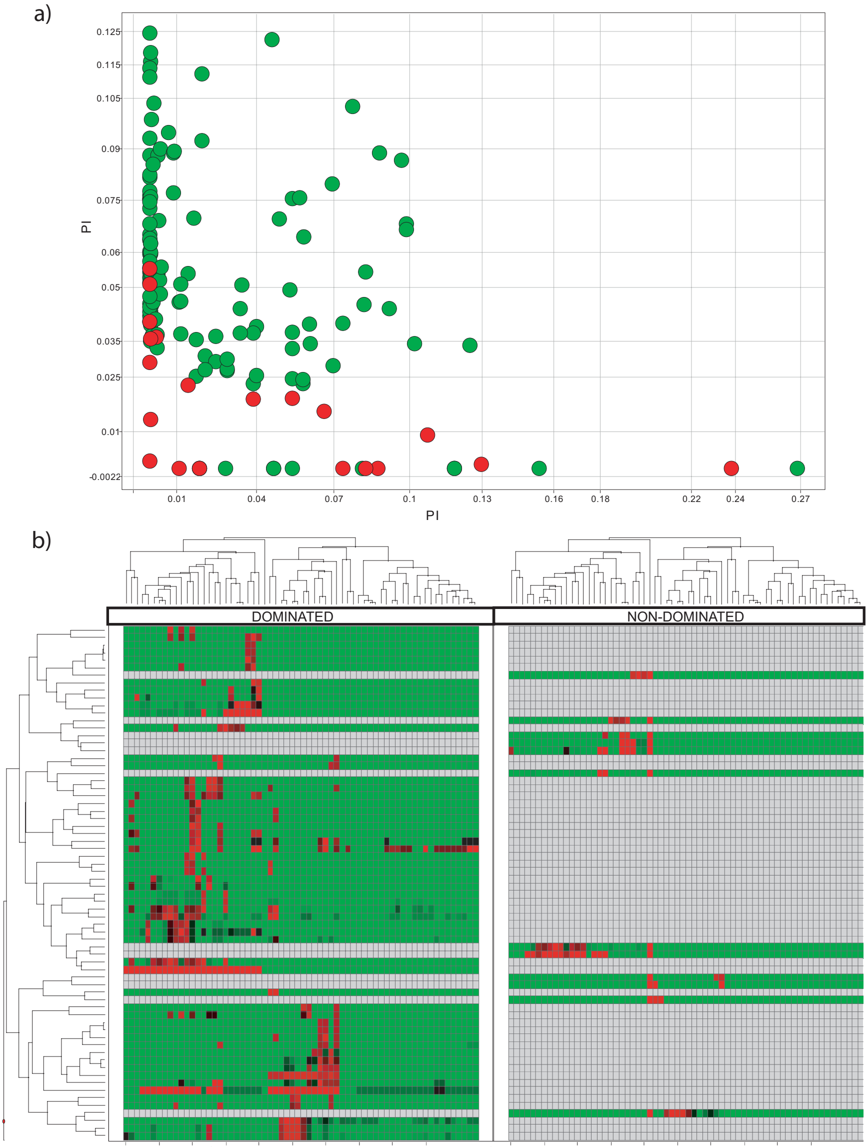


Figure 5.7 Selection of the most representative profiles.

a) Non-dominance optimization approach (Non-dominated solutions red; Dominated ones in green) between two conflicting objectives PI and SI. This guideline is applied in local neighbourhood to support diversity. b) Heat map corresponding to promoters (columns) recognized at different degrees of matching (green: low; red: high) by the profiles (rows) divided in neighbourhood (clusters). These localities are dominated by a representative profile (left columns). This guideline prevents the population of solutions to converge to a single region and obtains optimal and diverse solutions.

5.5.2 GPS performance

To evaluate the ability of GPS to retrieve PhoP-regulated promoters, we analyzed the statistical significance of GPS predictions in comparison with random classifications, and then, we evaluated the ability of GPS to discriminate between promoters regulated by PhoP and by other transcription factors.

We compared GPS prediction of the test set with a typical statistical approach consisting of randomly assigning two classes to 100.000 sets of observations with the same size of the test partition (Beer and Tavazoie 2004). This experiment retrieved an expected ca. 50% of 'correct' classifications, following a distribution close to normal and providing a standard deviation of 9.76%. Therefore, GPS prediction of 92% for the test set is 4.3 standard deviations away from the mean obtained by random assignment, which corresponds to a *p-value* <10E-5, determined by using paired t-test with Bonferroni correction (Matlab statistical toolbox). These results are in agreement with a sample size >23, a power of 92% and significance level given by the stated *p-value* (Rosner 1986).

We extended the test set by including 487 promoters from the RegulonDB database (Benitez-Bellon, Moreno-Hagelsieb et al. 2002; Salgado, Gama-Castro et al. 2004 that are regulated by transcription factors other than PhoP, by selecting the promoter region corresponding to the respective transcription factor binding site ±10 bp, its corresponding RNA polymerase site ±10 bp and expression levels from our own experiments. GPS had a false positive rate of 5.3% and a 93.92% of overall performance measurement) as a particular correlation coefficient implementation, with a 94 and 92% specificity and sensitivity on the extended set, respectively (Table 5-2).

Table 5-1Confusion matrix for GPS

		Predicted Class	
		Yes	**No**
Actual Class	**Yes**	92% *(TP)*	7.7% *(FN)*
	No	5.3% *(FP)*	94.6% (TN)

5.5.3 Comparision to other methods

We compare validated profiles (Zwir, Shin et al. 2005) and profiles provided by GPS, Bayesian networks (BN), Association Rules (AR), and Decision Trees (DT). These profiles are detailed in [TABLA]

5.5.3.1. Bayesian Networks.

Undirected graphical models, also called Markov Random Fields or Markov networks, have a simple definition of independence: two (sets of) nodes A and B are conditionally independent given a third set, C, if all paths between the nodes in A and B are separated by a node in C. By contrast, *directed* graphical models, also called Bayesian Networks or Belief Networks (BNs), have a more complicated notion of independence, which takes into account the directionality of the arcs. The most important advantage of this representation is that one can regard an arc from A to B as indicating that A ``causes'' B (i.e., causality). This can be used as a guide to construct the graph structure and the parameters of the model. For a directed model, we must specify the conditional probability distribution at each node. If the variables are discrete, this can be represented as a table, which lists the probability that the offspring node takes on each of its different values for each combination of values of its parents. We applied several structure learning algorithms implemented in (Consortium 2002): K2 (Cooper and Herskovits 1992), PC (Spirtes, Glymour et al. 1991), K2SN (de Campos 2001), and DVNS, based in a local search optimization approach (de Campos 2001) combined with several scoring metrics: K2 (Cooper and Herskovits 1992), BDE (Heckerman, Geiger et al. 1995) and BIC (Schwarz 1978). We selected the most interconnected network provided by the combination K2SN/DVNS. Other combinations provide more dominated solutions.

5.5.3.2. Association rules.

Among the best known algorithms for association rule induction is the Apriori algorithm (Agrawal and Shafer 1996). This algorithm works in two steps: In a first step the frequent item sets (often misleadingly called large item sets) are determined. These are sets of items that have at least the given minimum support (i.e. occur at least in a given percentage of all transactions). In the second step association rules are generated from the frequent item sets found in the first step. We applied the Apriori implementation of (Borgelt 2002) in our experiments.

5.5.3.3. Decision Tree.

A decision tree (Quinlan 1993) is a way of explaining the behavior of one target variable as a function of other source variables in a data set. The output takes the form of a tree structure, where each node represents the subset remaining after a sequence of conditions has been applied. Solution evaluation is performed based on the information gain ratio criterion, which essentially ensures that the amount of information gained about a target variable is maximized at each split. We used the C4.5 version of decision trees (Quinlan 1993) in our experiments.

5.5.3.4. Comparisons among validated profiles

We analyze the coincidence between profiles retrieved by GPS and previously listed methods: our comparison is based on profiles that are recovered by GPS and any other method (*p-value<0.5*). Neither profiles harboring promoters retrieved by another method different than GPS nor profiles found by GPS that are not present in any other method are analyzed.

We considered the set of promoters included in profile $P_3^2 O_1^2$. The AR and BN methods grouped these promoters together, however, BN included additional promoters in its retrieved group exhibiting downregulated values for the "expression" feature. (The group slightly differs from the profile detected by GPS because the latter locally re-discretizes the features). The DT method could not group together all promoters included in the profile, and also included additional promoters with PhoP boxes in the opposite orientation without discriminating between their RNA polymerase sites. This happens because it re- discretizes the "RNA pol sites" into an excessively general feature.

We considered the set of promoters included in profile $P_1^3 I_4^3 O_2^3$. BN retrieved the promoters of this profile, however, it was unable to describe them by the "interactions" feature, even though this feature was crucial for identifying promoters regulated by both PhoP and PmrA proteins. As a consequence, BN retrieved a large list of other promoters that do not specifically addressed the biological mechanism described by the $P_1^3 I_4^3 O_2^3$ profile. AR retrieved the promoters in the $P_1^3 I_4^3 O_2^3$ profile, but only via the "interaction" feature. Thus, by missing the "RNA pol sites" and "orientation" features in the group, AR produced an unspecific group less informative about the regulatory mechanism. DT did not grouped these promoter when its default parameters were used; however, it did after customizing them.

We considered the set of promoters included in profile $E_2^3 M_0^3 I_1^3$. BN specifically identified together the set of promoters included in this profile. AR and DT combined the promoters in the profile with those in $E_2^2 M_0^2$ profile, because neither methods were able to distinguish between the two sets of promoters by the "interactions" feature, that identifies the acid resistance regulatory genes that regulate those in the $E_2^3 M_0^3 I_1^3$ but not those in the $E_2^2 M_0^2$ profile.

We considered the set of promoters included in profile $E_2^2 M_4^2$. Only GPS characterized them by both their "expression" and "submotifs" features, which are the most relevant features distinguishing these promoters from other acid resistance genes, as their expression only slightly differs from that of the canonical PhoP regulated genes, and they are directly regulated by PhoP, by the presence of PhoP binding site motifs.

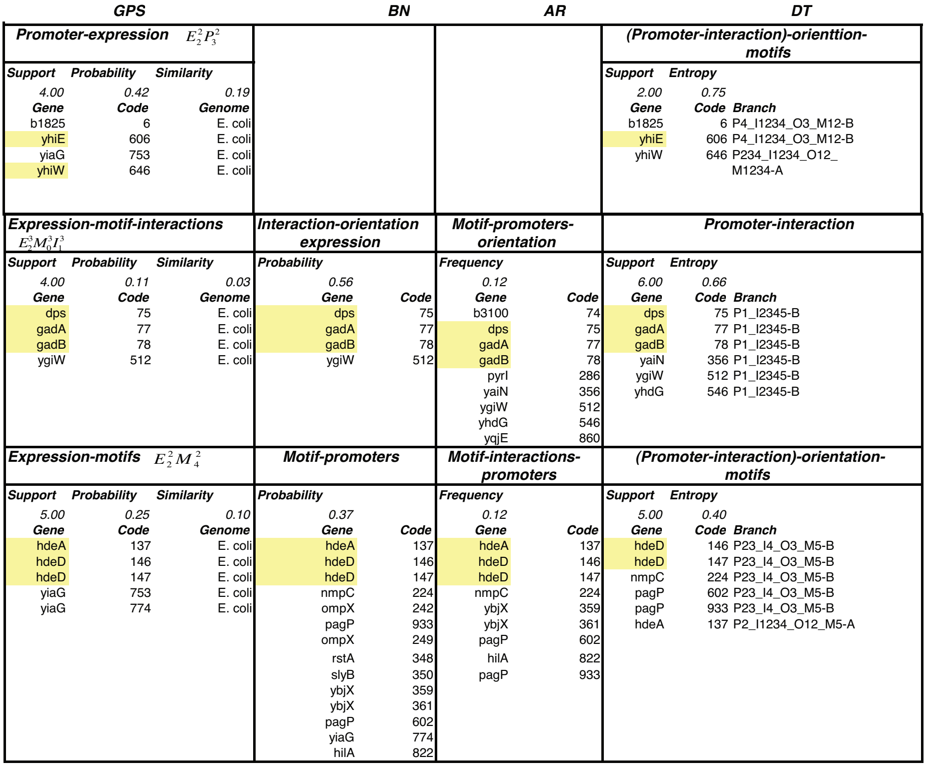
We considered the set of promoters included in profile $E_2^2 M_4^2$. Neither BN nor AR identified the promoters included in the profile, which are crucial for inferring the architecture of the regulatory network that control acid resistance genes. DT did not group these promoter using its default parameters.

Table 5-2 GPS vs. other Conceptual clustering techniques comparison
Specificity and sensitivity of results obtained by GPS in comparison with those retrieved by Bayesian Networks (BN), Association Rules (AR) and Decision Trees (DT). Shadow cells indicate promoter coincidence based on previously validated profiles. The probability, similarity, frequency, support and entropy are evaluation measures used in each column method. The control column indicates an internal code to discriminate among distinct promoters of a same gene.

GPS			BN		AR		DT		
Promoter-orientation $_3^3 O_1^3$			PPromoter-orientation		Promoter-orientation-expression		Promoter-interaction-orientation-motifs		
Support	Probability	Similarity	Probability		Frequency		Support	Entropy	
8.00	0.07	0.17	0.36		0.11		15.00	1.00	
Gene	Code	Genome	Gene	Code	Gene	Code	Gene	Code	Branch
slyB	351	E. coli	slyB	351	slyB	351	yhiW	743	P234_I1234_O12_M1234-A
ybjX	417	Salmonella	ybjX	417	ybjX	417	yhiW	646	P234_I1234_O12_M1234-A
mig-14	568	Salmonella	yeaF	469	mig-14	568	mgtC	718	P234_I1234_O12_M1234-A
virK	643	Salmonella	mig-14	568	virK	643	mig-14	568	P234_I1234_O12_M1234-A
yhiW	646	E. coli	virK	643	yhiW	646	ompT	226	P234_I1234_O12_M1234-A
mgtC	681	Salmonella	yhiW	646	mgtC	681	ompT	230	P234_I1234_O12_M1234-A
mgtC	718	Salmonella	mgtC	681	mgtC	718	pagC	931	P234_I1234_O12_M1234-A
pagC	1027	Salmonella	mgtC	718	pagC	1027	pagC	1026	P234_I1234_O12_M1234-A
			pagC	1027			pagC	1027	P234_I1234_O12_M1234-A
							pipD	847	P234_I1234_O12_M1234-A
							pipD	935	P234_I1234_O12_M1234-A
							slyB	351	P234_I1234_O12_M1234-A
							ybjX	358	P234_I1234_O12_M1234-A
							ybjX	417	P234_I1234_O12_M1234-A
							ybjX	438	P234_I1234_O12_M1234-A

Promoter-interactions orientation $_1^3\ _4^3O_2^3\ I\ P$			Promoter-orientation		Interactions		(Promoter-interaction-orientation)-motif		
Support	Probability	Similarity	Probability		Frequency		Support	Entropy	
6.00	0.21	0.17	0.67		0.11		4.00	0.75	
Gene	Code	Genome	Gene	Code	Gene	Code	Gene	Code	Branch
yeaF	464	E. coli	b2833	73	trs5_8	354	mgtA	599	P234_I5_O3_M3-A
pmrD	513	Salmonella	hdeD	146	yeaF	464	yeaF	464	P234_I5_O3_M3-A
mgtA	599	Salmonella	hdeD	147	yeaF	469	yrbL	929	P234_I5_O3_M3-A
udg	729	Salmonella	mgtA	149	pmrD	513	yrbL	943	P234_I5_O3_M3-A
yrbL	929	E. coli	nmpC	224	mgtA	599	udg	729	P234_I5_O3_M45-C
yrbL	943	Salmonella	ompX	242	udg	729	pmrD	513	P234_I5_O3_M12-B
			ompX	248	yrbL	929			
			ompX	249	yrbL	943			
			phoP	277					
			rstA	330					
			rstA	348					
			slyB	350					
			slyB	352					
			slyB	353					
			ybcU	357					
			ybjX	360					
			ybjX	402					
			yeaF	464					
			pmrD	513					
			mgtA	599					
			pagP	602					
			pagD	704					
			pgtE	720					
			udg	729					
			yiaG	774					
			mgtC	925					
			yrbL	929					
			pagP	933					
			yrbL	943					

Table 5-2 GPS vs. other Conceptual clustering techniques comparison (Continued)



GPS BN AR DT

Promoter-expression $E_2^2P_3^2$

Support	Probability	Similarity
4.00	0.42	0.19
Gene	**Code**	**Genome**
b1825	6	E. coli
yhiE	606	E. coli
yiaG	753	E. coli
yhiW	646	E. coli

(Promoter-interaction)-orienttion-motifs

Support	Entropy	
2.00	0.75	
Gene	**Code**	**Branch**
b1825	6	P4_I1234_O3_M12-B
yhiE	606	P4_I1234_O3_M12-B
yhiW	646	P234_I1234_O12_M1234-A

Expression-motif-interactions $E_2^3M_0^3I_1^3$

Support	Probability	Similarity
4.00	0.11	0.03
Gene	**Code**	**Genome**
dps	75	E. coli
gadA	77	E. coli
gadB	78	E. coli
ygiW	512	E. coli

Interaction-orientation expression

Probability	
0.56	
Gene	**Code**
dps	75
gadA	77
gadB	78
ygiW	512

Motif-promoters-orientation

Frequency	
0.12	
Gene	**Code**
b3100	74
dps	75
gadA	77
gadB	78
pyrI	286
yaiN	356
ygiW	512
yhdG	546
yqjE	860

Promoter-interaction

Support	Entropy	
6.00	0.66	
Gene	**Code**	**Branch**
dps	75	P1_I2345-B
gadA	77	P1_I2345-B
gadB	78	P1_I2345-B
yaiN	356	P1_I2345-B
ygiW	512	P1_I2345-B
yhdG	546	P1_I2345-B

Expression-motifs $E_2^2M_4^2$

Support	Probability	Similarity
5.00	0.25	0.10
Gene	**Code**	**Genome**
hdeA	137	E. coli
hdeD	146	E. coli
hdeD	147	E. coli
yiaG	753	E. coli
yiaG	774	E. coli

Motif-promoters

Probability	
0.37	
Gene	**Code**
hdeA	137
hdeD	146
hdeD	147
nmpC	224
ompX	242
pagP	933
ompX	249
rstA	348
slyB	350
ybjX	359
ybjX	361
pagP	602
yiaG	774
hilA	822

Motif-interactions-promoters

Frequency	
0.12	
Gene	**Code**
hdeA	137
hdeD	146
hdeD	147
nmpC	224
ybjX	359
ybjX	361
pagP	602
hilA	822
pagP	933

(Promoter-interaction)-orientation-motifs

Support	Entropy	
5.00	0.40	
Gene	**Code**	**Branch**
hdeD	146	P23_I4_O3_M5-B
hdeD	147	P23_I4_O3_M5-B
nmpC	224	P23_I4_O3_M5-B
pagP	602	P23_I4_O3_M5-B
pagP	933	P23_I4_O3_M5-B
hdeA	137	P2_I1234_O12_M5-A

5.5.4 PhoP profiles as discriminators of kinetic behaviour

We explore the correlation of sets of genes exhibiting similar gene expression patterns to the profiles that describe the distinct *cis*-elements that control the transcription initiation. We find that these profiles can be used to effectively explain the different kinetic behaviour of co-regulated genes measured by GFP reporter strains with high-temporal resolution (Figure 5.8). As our GFP assays are performed in vitro, we omit the profiles that incorporate the "interactions" features.

We detected the profile $E_1^3M_2^3P_2^3$ (PI=1.95E^{-6}; SI=0.006) that corresponds to the canonical PhoP-regulated promoters (e.g., those of the *phoP, mgtA, rstA, slyB, yobG and yrbL* genes). These promoters share the class II RNA polymerase sites situated close to the PhoP boxes, high expression patterns, and typically PhoP box submotif., produced the earlier rise times and the higher levels of transcription (Figure 5.8).

Another uncovered profile $O_1^3E_2^3P_3^3$ (PI=1.95E-5, SI=0.05), includes promoters that share a PhoP box in the opposite orientation of the canonical PhoP-regulated promoters as well as a class I RNA polymerase sites situated at medium distances from the PhoP boxes (e.g., those of the *mgtC*, *mig-14*, *pagC*, *pagK*, and *virK* genes of *Salmonella*). We tested this profile by GFP and found that effectively differs from the previous canonical profile, exhibiting the latest genes with the lowest levels of expression (Figure 5.8).

We also uncovered another slightly different profile $O_1^2P_2^2$ (PI=0.033, SI=0.044), which includes promoters (e.g., those of the *ompT* gene of *E. coli* and the *pipD, ugtL* and *ybjX* genes of *Salmonella*) that exhibit a PhoP binding site in the opposite orientation, but preserves the RNA polymerase of the canonical PhoP regulated promoters. We tested the kinetic behaviour of genes in this profile and found that present an intermediate value between previously described regulatory profiles (Figure 5.8).

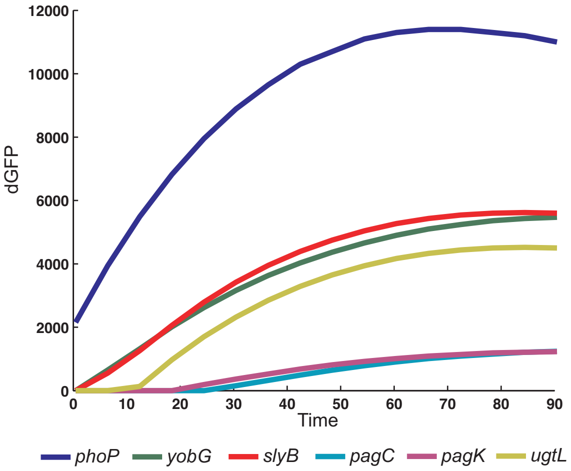


Figure 5.8 . Independent validation of profiles using kinetic classes.
Transcriptional activity of wild-type Salmonella harbouring plasmids with a transcriptional fusion between a promoterless gfp gene and the promoters.. The activity of each promoter is proportional to the number of GFP molecules produced per unit time per cell [dGi(t)/dt]/ODi(t)], where Gi(t) is GFP fluorescence from wild-type Salmonella strain 14028s, and ODi(t) is the optical density. The activity signal was smoothed by a polynomial fit (sixth order). The genes are evaluated by their rise time and levels of transcriptions

5.6 Concluding remarks

We have described an unsupervised machine learning method, termed GPS for *G*ene *P*romoter *S*can, that discriminates among co-regulated promoters by simultaneously considering both *cis*-acting regulatory features and gene expression. The GPS method encode regulatory features that specifically aimed at handling the variability in sequence, location and topology that characterize gene transcription. Then, the method uses an integrated approach for discovering promoter profiles, thereby uncovering an unsuspected complexity in the regulatory targets that are under direct and indirect transcriptional control of the regulatory protein.

Several characteristics of GPS contribute to its power. First, it considers gene expression as one feature among many, thereby allowing classification of promoters even in its absence (Conlon, Liu et al. 2003; Beer and Tavazoie 2004). Particularly, GPS differs from supervised learning methods (Mitchell 1997) that group features and observations based on explicitly defined dependent variables (Quinlan 1993; Conlon, Liu et al. 2003; Beer and Tavazoie 2004). Second, GPS performs a local feature selection for each profile because not every feature is relevant for all profiles (Kohavi and John 1997), and, a priori, we do not know which feature is biologically meaningful for a given promoter. This is in contrast to approaches that filter or reduce features for all possible clusters (Yeung and Ruzzo 2001). Third, GPS finds all optimal solutions among multiple criteria (Pareto optimality) (Deb 2001), which avoids the biases that might result from using any specific weighing scheme (Rissanen 1989). This can detect cohesion within a small number of promoters that would remain undetected by methods that emphasize the number of promoters in a profile (Agrawal and Shafer 1996). Fourth, GPS has a multimodal nature that allows alternative descriptions of a system by providing several adequate solutions (Deb 2001; Ruspini and Zwir 2002), thus recovering locally optimal solutions, which have been shown to be biologically meaningful (Azevedo, Lohaus et al. 2005; Cotik, Zaliz et al. 2005). This differentiates GPS from methods that are focus on a single optimum (Gutierrez-Rios, Rosenblueth et al. 2003; Martinez-Antonio and Collado-Vides 2003). And fifth, GPS allows promoters to be members of more than one profile by using fuzzy clustering (Bezdek 1998; Cordon, Herrera et al. 2002; Gasch and Eisen 2002), thus explicitly treating the profiles as hypotheses, that are tested and refined during the analysis (Mitchell 1997). This distinguishes GPS from clustering approaches that prematurely force promoters into disjointed groups (Qin, McCue et al. 2003). In addition, GPS recognizes that not every profile is meaningful (Bezdek 1998), which avoids the constraints of methods that force membership even to uninteresting groups because the sum of membership is required to be one (Cooper and Herskovits 1992).

The GPS method, termed *G*ene *P*romoter *S*can here, can be generalized to a method for *G*rouping, *P*rototyping, and *S*earching in the lattice space of hypotheses, which can be used in different structural domains. For example, we have described the analysis of the targets of regulation of the response regulator PhoP (Zwir, Shin et al. 2005), and it is now being applied to describe other two-components systems (e.g., PmrA/PmrB) and general regulators (e.g., CRP) in different genomes (e.g., *Yersinia pestis* and *Vibrio cholerae*). Moreover, it is being used to mine the Gene Ontology database (Ashburner, Ball et al. 2000) to discover and annotate profiles across biological processes, cellular components and molecular functions, and to identify molecular pathways that provide insight into the host response over time to systemic inflammatory insults (Calvano, Xiao et al. 2005).

Chapter 6

Learning robust dynamic networks in prokaryotes by Gene Expression Networks Iterative Explorer (GENIE)

6.1 Introduction

Gene expression is determined by protein-protein interactions among regulatory proteins and with RNA polymerase(s), and protein-DNA interactions of these transacting factors with cis-acting DNA sequences in the promoters of regulated genes (Kærn 2003). These interactions define complex genetic networks, whose designs have motivated researchers to draw direct analogies with established techniques in electrical engineering (Guet, Elowitz et al. 2002; Hasty, McMillen et al. 2002). As with the construction of electrical circuits, the gene circuit approach uses mathematical and computational tools in the construction and posterior analysis of a proposed network diagram. The qualitative agreement between model and experiment in a series of studies depends both on the design of the network topology, which most of the times includes uncertain connections between genes, as well as on the dynamic behavior of the network, which is affected by the ambiguity inherent to the biological processes (e.g., monomer or dimmer binding of promoters, enzymes having kinase and/or phosphatase activities, etc.) and the mathematical models used to represent them (e.g., Boolean or continuous models; reverse or forward algorithms) (van Someren, Wessels et al. 2002). Moreover, the number of genes considered in the networks is usually large compared to the number of the available measurements (e.g., time-point expression), thus, more than one possible model may be consistent with the subjacent data (Wahde, Hertz et al. 2001). Finally, the data always contains a substantial amount of noise (McAdams and Arkin 1999; Li and Hung Wong 2001; Li and Wong 2001) provided by the systematic variability of the experiments, which in addition to previous problems, makes it difficult to deduce the implications of the underlying logic of genetic networks through experimental techniques alone.

We propose a methodology termed GENIE, for Gene Expression Networks Iterative Explorer, which embraces the uncertainty inherent to the biological problem and the imprecision of their underlined mathematical models by using an iterative approach. First, GENIE proposes a network topology based on DNA sequence analysis of transcription factor interactions, which, together with previous knowledge from the literature, constitute the raw material for the architecture design. Second, it transform the hypothesis provided by the network topology, by means of its possible chemical reactions and physical constraints, into a system of nonlinear ordinary differential equations, whose variables are concentrations of proteins, mRNA, etc (von Dassow, Meir et al. 2000; Meir, Munro et al. 2002). Rather than advocating a single and definitive model of the genetic network, we describe a variety of optimal models learned by random walk (Zwir, Traverso et al. 2003) and improved by genetic algorithm (Herrera and Lozano 2005) techniques. Third, the network non-linear models are evaluated by testing their ability to reproduce the biological behavior observed in vivo including time-dependent changes of the concentrations of the system components (e.g., kinase, phosphatase, and transcription activities). Fourth, the successful models are tested by considering different emergent properties, such as flexibility to reproduce all possible functional patterns, and robustness to changes in parameters and initial conditions. Fifth, we revisit the original topology and iterate, developing adaptive models of genetic networks. Finally, a decision making process reveals the most realistic models.

We apply GENIE to uncover regulatory networks in the bacteria *Salmonella enterica serovar Typhimurium* by focusing on the PhoP/PhoQ and PmrA/PmrB two-component systems, which govern virulence and the adaptation to low Mg^{2+} and high Fe^{3+} environments, respectively (Hoch 2000; Kato, Latifi et al. 2003). The study of the PhoP regulon constitutes a special challenge due to the multiplicity of PhoP-controlled targets, and the connectivity of the PhoP/PhoQ system with other two-component systems, such as the PmrA/ PmrB system. We verified our predictions by measuring time-dependent gene expression using Green Fluorescence Protein (GFP) techniques.

6.2 Problem: Computational and Biological Challenges

6.2.1 Modeling genetic networks

The scientific community has put a considerable amount of effort into designing approaches to model genetic networks (Wahde, Hertz et al. 2001; Milo, Shen-Orr et al. 2002). Most of the models define species as nodes, and interaction between them as links of a graph. They differ in the values assigned to the nodes (i.e. initial concentration) and links (i.e. the value of the interaction between the species), generating alternative models. Indeed, the interactions between elements can be considered as static or dynamic, and the entire model can be studied in a stochastic or deterministic context (e.g. Boolean, discrete or continues) (Rubio-Escudero, Harari et al. 2007).

The usage of continuous values to determine the level of gene expression and relationships among them results the most expressive model, because it allows capturing biological properties that can be experimentally observed. Ordinary Differential Equations (ODE's) are good approximation to continues models: ODE's capture the system by equations that calculate the difference of concentration of species (i.e. RNA, proteins) along the time. Statical ODE's (Batchelor and Goulian 2003) model the systems when they reach their steady state (i.e. the system has reached an equilibrium in which the difference of concentrations of species in function of time is equal to zero).

In contrast, dynamic models (Meir, Munro et al. 2002) do not necessarily consider this equilibrium, enabling the observation of the gene expression behavior over time. This important characteristic allows the temporal simulation of the system and results critical when studying biological systems like PmrA/PmrB-PhoP/PhoQ, in which is possible to experimentally observe the dynamics for different sets of stimuli. An interesting concept of dynamic ODE's models is that the actual values of the parameters are not a priori estimated. Instead, the model can be evaluated by employing different sets of parameters to test if it follows certain macroscopic patterns previously known. As a result, the quality of the obtained network is not only determined by the chosen model, but also by the design of the inference method (i.e. learning strategy) that estimates the parameters of the network. These methods have to deal with a high dimensional problem, and researchers have proposed several strategies to bypass this problem by reducing the number of elements modeled (clustering and thresholding), by increasing the number of samplings (i.e microarrays) (Mjolsness, Mann et al. 1999), or by simplifying the complexity of the model (i.e. limited conectivity) (Wahde, Hertz et al. 2001). Optimization techniques based on genetic algorithms have been successfully applied for system identification (Kimura, Kawasaki et al. 2004), constituting a promising strategy that narrows the former constraints mostly evidenced in random walk learning methods (Zwir, Traverso et al. 2003).

6.2.2 Two-component systems

The discovery and understanding of the underlying mechanisms employed by a cell to integrate multiple input signals and generate a response is a key field of study in biology. For every set of signals, the cell uses genes (not always fully known) that codify distinct proteins, which are able to sense these signals; perform the signal transduction required to activate the response cellular machinery; and generate the adequate output response, which is often manifested in a variation of the expression level of another set of genes. All of these genes constitute true biological circuits of cellular control (Alon 2007).

In prokaryotes organisms, the "two-component systems" are small networks that control an important amount of cellular functions, constituting the main mechanism of signal transduction that allows the bacteria to modify its cellular behavior in response to environmental stimuli. These systems include a sensor protein that responds to specific signals and phosphorylates its cognate regulators. The response regulators are mostly transcription factors proteins (TF) that once they become phosphorylated, bind the DNA, and then, activate or repress their target genes. These systems can be autoregulated, and some of them have cross-talks to other two-component systems. Although there are between 30 to 60 different two-component systems identified in bacterial genomes, they are not completely understood and some of them can be also preserved in eukaryotic genomes (Zwir, Harari et al. 2007).

The PhoP/PhoQ two-component system constitutes a master regulator in *Salmonella enterica serovar Typhimurium*, regulating the transcription of more than 2% of the genes in response of a low extra cellular Mg^{2+} (i.e. genes related to the low environmental Mg^{2+} cell survival and mouse virulence). Another two component system present in *Salmonella* is the PmrA/PmrB system, which is related to the polymyxin B antibiotic inducted resistance; resistance to cell death mediated by Fe^{3+}; cell soil growth, mouse virulence; and intramacrophage infection. The target genes regulated by this system independently respond to two signals: high level of extra cellular Fe^{3+}, sensed by the PmrB protein; and low levels of Mg^{2+}, sensed by the PhoQ protein. This cross-talk between both two-component systems is mediated by the *pmrd* gene, which resulting protein PmrD can bind the PmrA protein probably in a posttranscriptional or posttranslational fashion. Curiously, *pmrD* harbors a PmrA binding site that results in a negative feedback that closes the regulatory loop. Although, this system has been widely studied (Zwir, Shin et al. 2005), the exact mechanisms that defines the system dynamics is still unknown (see Figure 4.1).

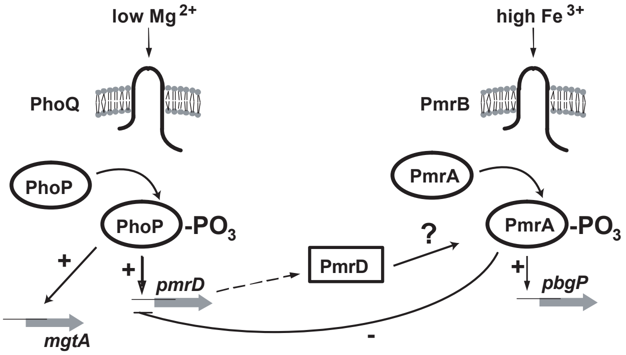


Figure 6.1 The PhoP/PhoQ-PmrA/PmrB functional scheme in *Salmonella enterica serovar Typhimurium*.

The PhoQ protein senses low Mg^{2+} and the PmrB protein high Fe^{3+} concentrations from the environment and both proteins phosphorylate their cognate response regulators PhoP and PmrA, respectively. Although each of these proteins control the expressions of their target genes in response to their own signal, an alternative cross-talk suggest that some genes regulated by the PmrA protein can be regulated by PhoP in low Mg^{2+} conditions via the PmrD protein. Indeed, a transcriptional negative feedback has been detected in the *pmrD* gene.

6.3 Discovering genetic networks using GENIE

In this work, we propose a method termed GENIE devoted to infer genetic regulatory networks. This method consists of three main phases (see Figure 6.2): (1) *discovery of the components of the studied system*, where we analyze the literature, databases and experimental evidence of *cis*-features (e.g., TF binding sites) to formulate alternative architectures for a genetic network and encode these models as continuous ODE's; (2) *identification of the desired system*, where we learn the parameters of the network, simulate its dynamics, and evaluate the performance of different models both by probabilistic measures and correlation with experimental results; and (3) *sensitivity analysis of the system parameters*, where we evaluate the robustness of the learned system and extract emergent properties from the evaluated architectures that may uncover biological significance (e.g., gene expression diversity).

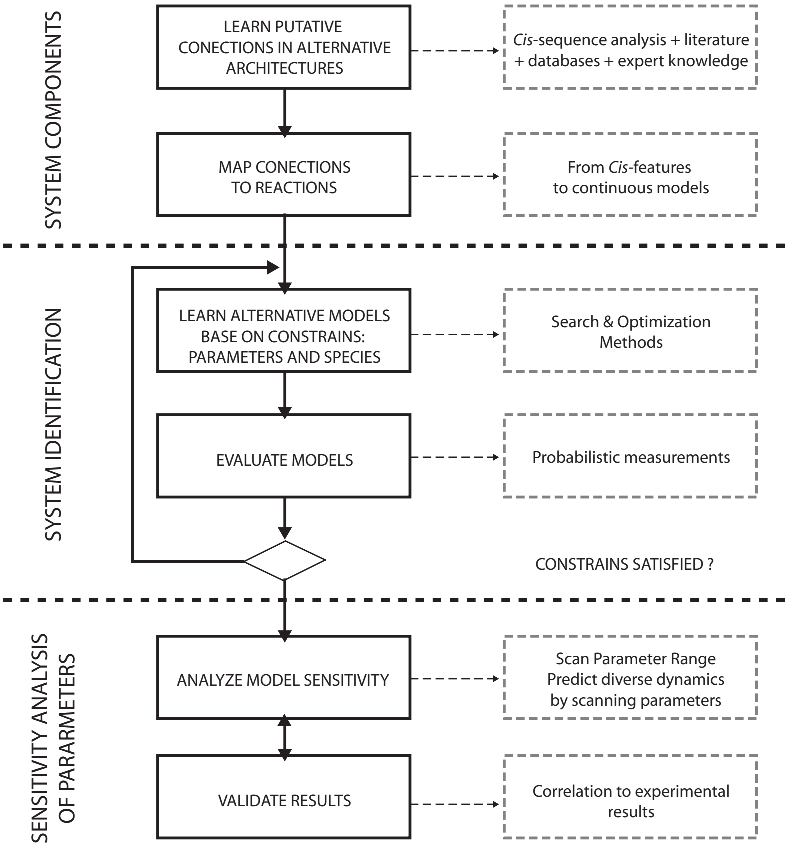

SYSTEM COMPONENTS
LEARN PUTATIVE CONECTIONS IN ALTERNATIVE ARCHITECTURES
Cis-sequence analysis + literature + databases + expert knowledge
MAP CONECTIONS TO REACTIONS
From Cis-features to continuous models
SYSTEM IDENTIFICATION
LEARN ALTERNATIVE MODELS BASE ON CONSTRAINS: PARAMETERS AND SPECIES
Search & Optimization Methods
EVALUATE MODELS
Probabilistic measurements
CONSTRAINS SATISFIED ?
SENSITIVITY ANALYSIS OF PARAMETERS
ANALYZE MODEL SENSITIVITY
Scan Parameter Range Predict diverse dynamics by scanning parameters
VALIDATE RESULTS
Correlation to experimental results

Figure 6.2 . Flowchart of the GENIE method.
Each phase is decomposed into different task that are implemented in the methodology.

6.3.1 System components

GENIE discovers genetic regulatory networks by formulating hypothetical architectures, representing them as continuous models encoded as biochemical reactions that capture the dynamics of the system under different constrains.

In spite of the fact that continuous model require a proper parameter configuration (i.e. system identification), they offer the advantage that the parameterized components on which they are constructed can model a complete set of analogous gene dynamic (i.e. expression intensity and rise time or order), thus these models can be customized to predict gene expression of a complete cluster of genes.

6.3.1.1. Network architecture

Cis-sequence analysis methods. Transcriptional regulation evidence can be found in sequences. We employ machine learning techniques (Zwir, Huang et al. 2005) that analyze genome sequences and databases (Salgado, Gama-Castro et al. 2004) to uncover initial hypothesis about architectures.

Expert knowledge: regulation evidence can be reinforced by microarrays experiments, however, the constraints in such analyses hitherto allow a relatively crude classification of gene expression patterns into a limited number of classes (e.g., up- and down-regulated genes (Oshima, Aiba et al. 2002; Tucker, Tucker et al. 2002))

Mapping sequence-based circuits into continuous models: the equations obtained for the differentials, in function of time, calculate the concentration of species (i.e. nodes) based on the values of the species directly connected to them, allowing temporal simulations that capture the dynamic behavior of the system. GENIE relays on Ingeneue software (Meir, Munro et al. 2002) which provides the Cash Karp method to integrate ODE's (see (Meir, Munro et al. 2002) for a comparative analysis of Cash Karp, SEBE, SEAPC, and Adams-Bashforth-Moulton alternatives applied to biological problems such ours).

Developing incremental models. our methodology incrementally formulates network architectures to find the minimal one, according to the number of species and interactions, that exhibits the experimentally observed properties. It starts with a model that reflects the recovered information and postulates the most general possible hypothesis for the unknown interactions. We express the rules that determine the behavior of genetic regulatory networks by decomposing the network into an aggregation of functional modules (e.g. negative/positive gene autoregulation, gene direct regulation, (des)phosphorylation of a protein) (Lee, Rinaldi et al. 2002; Milo, Shen-Orr et al. 2002; Kærn 2003), which in turn are translated into a system of ODE's (Meir, Munro et al. 2002).

6.3.1.2. Network constrains

Optimization methods are always based on constraints that should be satisfied in order to obtain feasible solutions (Deb 2001). The optimization of genetic networks has to consider at least two kind of constraints: Input/output constrains, where input signals activate the system and produce a desired output gene expression; and temporal constrains, which impose that the genes have to be ON and OFF at certain times with a specific order (Alon 2007). For example, in the previously described two-component systems the genes *pmrD* and *mgtA* have to be on in the present of Mg^{2+}. Indeed, the *pmrD* gene is constrained by the presence of Fe^{3+} (e.g., repressed) but not the *mgtA* gene. Moreover, *mgtA* should be early activated because of its role as a magnesium transporter gene (Zwir, Shin et al. 2005).

6.3.2 System identification

We formalize regulatory genetic networks by employing continuous models that enable the representation of gene expression dynamics. We apply optimization strategies to learn the parameters and initial values of the species contained in the models, some of which are difficult to be experimentally identified.

6.3.2.1. Learning network parameters and species

GENIE employs both random walk and genetic algorithms (GA) strategies to search and optimize for parameters that identify the system. The random walk approach is a formalization of the intuitive idea of taking successive steps, each in a random direction) (Meir, von Dassow et al. 2002). GA provide a learning method motivated by an analogy to biological evolution (Mitchell 1997): it iteratively updates a pool of hypothesis, called population, to identify the best one. On each iteration, all members of the population (represented as chromosomes) are evaluated according to the fitness function. A new population is then generated by applying genetic operators (i.e. crossover and mutation) to the most fit individuals.

Chromosome representation: we encode the parameters of the solution as a vector of real numbers

Fitness function: It considers the value of every specie for each constrain at simulated time=300 seconds (see Evaluation below).

Selection: we employ bit tournament to select the population that breed the new generation.

Crossover: new individuals are generating by applying both two-point crossover and arithmetical crossover operators

Mutation: we select a fraction of the vector of parameter with an uniform distribution and replace these entries with random numbers selected uniformly from the range for that entry.

Elitism: we retain the 3 solutions with best score in the elite set.

6.3.2.2. Evaluating networks using probability measurements

We determine the capability of network architectures and their related parameters to reproduce the behavior of the living organism by applying a score function which evaluates the predicted concentration of distinguished species (equation (6.1)

$$
\begin{aligned}
score &= \frac{\sum_i T(x_i)}{1+\sum_i T(x_i)} \\
T_{off} &= \alpha_{\max}\left(\left(x_i/x_t\right)^3 \Big/ \left(1+\left(x_i/x_t\right)^3\right)\right) \\
T_{on} &= \alpha_{\max}\left(1-\left(\left(x_i/x_t\right)^3 \Big/ \left(1+\left(x_i/x_t\right)^3\right)\right)\right)
\end{aligned}
\tag{6.1}
$$

where i represents each specie; x_t represents the threshold for each specie; and $\alpha_{\max}$ is the worst possible value (i.e. 0.5). The functions T_{on} and T_{off} calculate a scalar value based on the half-maximal-activity threshold, according to the constrains (i.e. activated/repressed). A score value close to 0 indicates a high similitude with the constrained (we consider that a solution represents the expected pattern if its score is below 0.3).

Moreover, based on this score, we can compute the frequency of feasible solutions, and estimate the corresponding probability of randomly finding a configuration for a genetic network that fulfill the constrains (equation (6.2)). A high probability of finding configurations that reproduce the expected pattern can indicate that the functionality is more related to the network architecture itself than to the parameters:

$$
\begin{aligned}
p^n &= f \\
p &= 10^{\frac{\log f}{n}}
\end{aligned}
\tag{6.2}
$$

where p is the probability of randomly choosing a feasible solution (i.e. a configuration that allows the architecture to reproduce the expected patterns); f is the frequency of feasible configurations; and n is the number of parameters. In this way, we consider architectures as a "black box" modules which are expected to reproduce the functionality in proportion to the configuration of parameters explored.

6.3.3 Sensitivity analysis of parameters

Different approaches have been proposed to evaluate the quality of the genetic regulatory network models (e.g. robustness; and flexibility) (Meir, von Dassow et al. 2002). However these approaches partially evaluate the fidelity of model while representing a biological system. In this work, we propose a global quality measure based on: *Realism,* the model should be able to reproduce the experimentally observed behavior, relatively independent of its parameters; *Robustness*, network architectures should preserve the functional characteristic of the system when one or more parameters are perturbed. The models should tolerate variations without lost of realism of the link parameters (relations between species), because of the biological property of network resistance to subtle mutations of the participating genes; and node parameters (concentrations), because of the intrinsic noise of molecular systems. Finally, the *flexibility* criteria evaluates the capability of networks to simultaneously reproduce distinct patterns of behavior (i.e. constrains) of the system under study.

We evaluate the *Robustness* of a network architecture by randomly choosing a solution that fulfill all of the imposed constrains (i.e. it complies with the *realism* and *flexibility* criteria) and observe its behavior when we independently sample each parameter value, within a biological significant range, and fix the original configuration values for the other parameters. Thus, we determine a feasible solution range for the parameters, indicating possible alternatives to adapt the network to reflect the behavior of other genes.

Differential gene expression can be obtained from two distinct sources. A variety of network motifs integrated in the network architecture produce distinct expression in the target genes including the single-input (PhoP → *mgtA*), the chained (PhoP → *pmrD* → *pmrA* → *pbgP*) and the multi-component motif (PhoP → *pmrD* → *pmrA* → *pmrD*). Indeed, even within a particular network motif we can obtain differential expression in distinct target genes (e.g., PhoP → mgtA; PhoP → *mgtC*) by scanning the range of feasible solutions. Thus, allowing making predictions about diversity of unseen gene expression.

Furthermore, our methodology helps the evaluation of the biological significance of the results, by comparing the predictions to experimentally obtained results. We measure the promoter activity and growth kinetics for GFP reporter strains with high-temporal resolution (Zwir, Shin et al. 2005); smooth the activity signal by a polynomial fit (sixth order), and then we calculate the Pearson's coefficient (equation 6.3) to estimate the correlation of the experimental results and the values predicted by the model:

$$d = \frac{\left(x_r - \overline{x_r}\right)\left(x_s - \overline{x_s}\right)}{\sqrt{\left(x_r - \overline{x_r}\right)\left(x_r - \overline{x_r}\right)^T}\sqrt{\left(x_s - \overline{x_s}\right)\left(x_s - \overline{x_s}\right)^T}}, \quad \overline{x_i} = \frac{1}{n}\sum_i x_{ij} \tag{6.3}$$

where *d=1* shows a perfectly correlated samples and *d=-1* indicates a negatively correlated values.

6.4 Application of GENIE

Learning Phop/PhoQ-PmrA/PmrB putative architecture: literature and TF binding site evidence (Zwir, Huang et al. 2005) indicate that PhoP/PhoQ system is activated by a low Mg^{2+} extra cellular level and the PmrA/PmrB system is activated by a high Fe^{3+} extra cellular level (see Appendix B Figure 1 for the analysis of the TF evidence). The PhoQ and PmrB membrane proteins sense these conditions respectively and in the presence of the signal change to an activated state in which they act as kinase and phosphatase of their cognate response regulators PhoP and PmrA. These proteins behave as transcription factors regulating several target genes. Both PhoP/PhoQ and PmrA/PmrB constitute operons, each one positively regulated by PhoP and PmrA respectively

We considered the crosslink of these two systems by a "forward" conection, from PhoP/PhoQ two component system to the PmrA/PmrB system and a "backward" connection in the opposite direction. Binding sites evidence and CHIP experiments show that PmrA represses the expression of *prmD* gene (i.e. "backward" connection). Given the fact that there is no binding site evidence of regulation of *pmrA* gene by the product of the *pmrD* gene we assume the "forward" connection of the systems is post-trancriptional (i.e.. PmrD protects the PmrA phosphatated form from the phosphating activity of PmrB) (see Figure 6.3 for the final refined network architecture; and Appendix B Figure 2 for the initial reduced model).

Reactions and input concentrations of species: We translate architectures into a system of ODE's, by employing the Ingeneue library, which allows simulating the dynamic behavior of the network architecture (see Appendix B Table 1 for the list of equations that model the final refined model).

The Fe^{3+} and Mg^{2+} concentration correspond to the "input" of the PhoP/PhoQ-PmrA/PmrB two component systems, while the values of *mgta, pbgP* and *pmrD* correspond to the "output" of the system. High values for the *mgta* and *pbgP* indicate the activation of the PhoP/PhoQ and PmrA/PmrB system respectively. A high value of *pmrD* shows the activation of the "forward" connection between the two systems, and a low one the activation of the "backward" connection. (see Table 6-1for a list of expected patterns of behavior)

Learning parameters: We test the inference method by executing our GA using different configurations (i.e. population size, number of generations) and observe that both the population size and the maximum number of executions independently improve the quality of the results (see Table 6-2).

Moreover, we compare the solutions obtained by the GA to the solutions obtained by the random walk approach (see Table 6-3) (Zwir, Traverso et al. 2003). A score difference above 0,20 supports the notion that GA strategy is adequate for learning genetic regulatory networks.

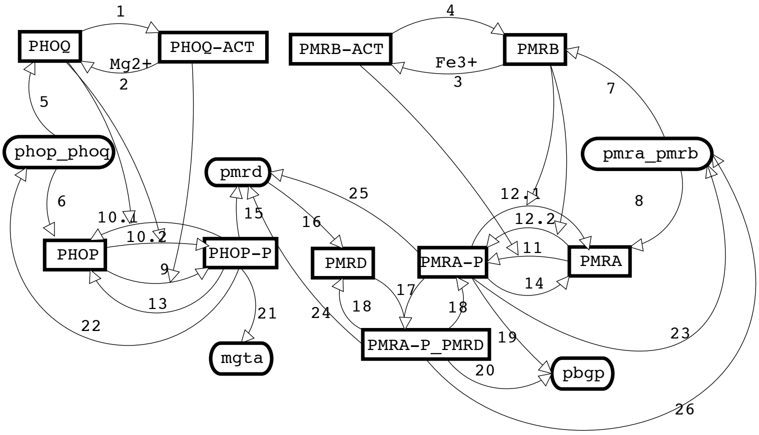


Figure 6.3 . Final refined model.
The species interact as follows: 1/2- Low/High Mg^{2+} level favors the PHOP-ACT(ivated)/PHOP state in equilibrium. 3/4- High/Low Fe^{3+} level favors PMRB-ACT(ivated)/PMRB state in equilibrium. 5/6- phop_phoq is translated into PHOQ/PHOP proteins. 7/8- pmra_pmrb is translated into PMRB/PMRA proteins. 9- PHOP is phosphorilated (PHOP-P) by PHOQ-ACT kinase activity. 10.1- PHOP-P is desphosphorilated to PHOP by PHOQ phosphatase activity. 10.2- PHOP is phosphorilated to PHOP-P by PHOQ kinase activity. 11- PMRA is phosphorilated to PMRA-P by PMRB-ACT kinase activity. 12.1- PMRA-P is desphosphorilated to PMRA by PMRB phosphatase activity. 12.2- PMRA is phosphorilated to PMRA-P by PMRB kinase activity. 13/14- PHOP-P/PMRA-P is spontaneous desphosphorilated to PHOP/PMRA. 15- PHOP-P activates the *pmrD* transcription. 16- *pmrD* is translated into PMRD. 17- PMRD binds PMRA-P (constituting PMRD_PMRA-P) which activates *pbgP* and represses *pmrD* genes, but it is not affected by the phosphatase activity of PMRB-ACT. 18- PMRA-P_PMRD unbinds into PMRD and PMRA-P. 19/20- PMRA-P/ PMRA-P_PMRD activates the transcription of *pbgP* gene. 21/22- PHOP-P activates the transcription of *mgta*/*phoP_phoQ*. 23- PMRA-P activates the transcription of *pmrA_pmrB*. 24/25- PMRA-P_PMRD/PMRA-P represses the transcription of *pmrD*. 26- PMRA-P_PMRD activates the transcription of *pmrA_pmrB*.

Table 6-1 Patterns of input/output* for the PhoP/PhoQ-PrmA/PrmB systems**.**

	Input		Output		
Constrain	**Mg^{2+}**	**Fe^{3+}**	***mgta***	***pmrD***	***pbgP***
1	1	1	1	0	1
2	1	0	1	1	1
3	0	1	0	0	1
4	0	0	0	0	0

* A 1 in columns Mg^{2+} and Fe^{3+} indicates the presence of the signal (a low/High concentration of the corresponding input parameter), and a 0 its absence. For the output parameters columns, a 0 indicates the repression of genes *mgta*, *pmrD* and *pbgP*, and a 1 the expression of the corresponding gene.

Table 6-2 Evaluation of the performance of the GA.

Population size	Nbr. Generations	Evaluations	Best score	Best solution generation
50	100	5,000	0.1914	20
200	100	20,000	0.0522	9
50	250	12,500	0.0473	22

Table 6-3 Performance comparison (Random walk vs. GA)

Population size	Evaluations	Best score
Random Walk	100,000	>0.25
GA	1,100/12,500*	0.0473

* The GA obtained the best score after 1,100 evaluations. Heuristics like stall time can decrease the number of evaluations by indicating possible algorithm's stop condition.

Evaluating models: we initially propose a reduced model (Appendix B Figure 2) designed as a test bed for our methodology: for sake of simplicity it lacks of the "forward connection" between the PhoP/PhoQ and PmrA/PmrB systems. We formalize this lack of realism by not specifying the expression of *pbgp* gene in a low Mg^{2+} and Fe^{3+} environment concentrations. The good probability measure obtained by this initial model (*p*=0.8341) in a flexible configuration (i.e. it satisfies all the constrains) gives us a solid foundation to evolve it towards the final refined model (see Figure 6.3), which reflect the "forward" connection. Along the process, we adapt the constrains to expect the expression of *pbgp* in the above conditions and relaxed the expression of *pmrD* in a low Mg^{2+} and Fe^{3+} environment concentrations (*pmrD* can be either activated by PhoP or repressed by PmrA). This final refined architecture that is more complex than the initial, thus requiring more parameters, actually obtains slightly better probability measure (*p*=0.8354).

Furthermore, we measure the promoter activity and growth kinetics for GFP reporter strains with high-temporal resolution (Zwir, Shin et al. 2005) for our distinguished genes *phoP, mgta* and *pmrD*, smooth the activity signal and then calculate the correlation to the predictions of the model. Pearson's coefficient indicates a correlation of 0.997 for *pmrD* gene; 0.983 for the *mgta* gene; and 0.991 for the *pbgp* gene, which reflects a highly correlated behavior between our predictions and the experimentally obtained values (see Figure 6.4) (The random walk approach obtains a coefficient of 0.12 for *pmrD*, 0.474 for *mgtA* and 0.44 for *pbgp).*

Sensitivity of the model: Our analysis of the sensitivity of the final refined network architecture for the PhoP/PhoQ-PrmA/PrmB system shows a tolerance of different magnitude order for distinct set of parameters. (Appendix B Figure 3). for a detail description). Indeed, the final network architecture behaves according to the expected pattern when parameters (e.g nu_phop_mgta) take the entire biological meaningful range. Moreover, the architecture has only 3 parameters (i.e. a 4.5% of the 66 parameters) that can accept less than 25% of their entire range, what shows the robustness quality of our final refined network. (see Figure 6.5 and Appendix B Table 2)

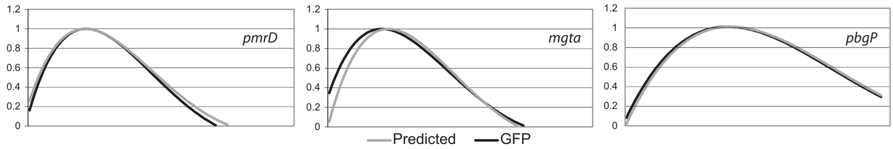


Figure 6.4 . Predicted and experimentally validated gene expression level.
This charts reflects the high correlation between the predicted behavior (blue) and the experimentally obtained values (red) (i.e. promoter activity and growth kinetics for GFP reporter strains with high-temporal resolution).

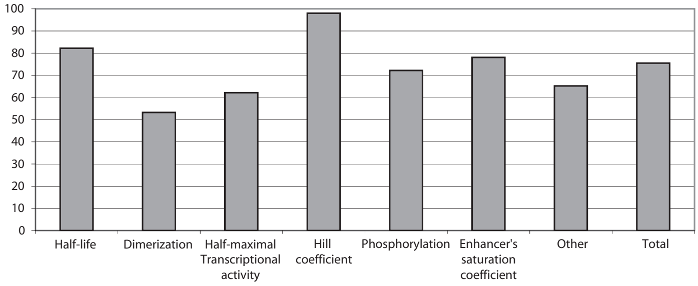


Figure 6.5 . Robustness analysis.
This chart shows the percentage of fulfillment for the biological meaningful range for parameters. (The parameters are grouped by their type; the values represent the obtained average for 10 solutions).

Predicting by scanning ranges of feasible solutions: we hypothesize about the different kinetic behavior that genes co-regulated by PhoP might exhibit by scanning the parameters related to the *mgtA* specie (i.e. the distinguished specie that represents Phop regulated genes) in the previously learnt range of values. We observe that the simulation of the model can produce different patterns of rise time and level of expression, what is desirable for the operon of a master regulator like PhoP (see Figure 6.6 and Appendix B Table 3 for the obtained results).

Validating results: we perform GFP experiments to evaluate the rise time and level of expression of PhoP regulated genes (e.g. *rstA, mgtA, phoP, slyB. pmrD, pagP, mig-14, pagC, pcgL, mgtC*) (see Figure 6.7), and calculate the correlation (*c*) between these experimentally obtained results to the patterns already predicted (See Appendix B Table 4 for a detail correlation results). Our analysis shows that pattern 12 predicts the dynamics of genes with early rise time and high level of transcription (i.e. *phop* – *c*=0.913, *pmrD* – *c*=0.981, and *mgtA* – *c*=0.975); pattern 13 correlates to genes with a late rise time and low level of expression (i.e. *pagC* – *c*=0.917 and *mgtC* – *c*=0.919); and finally that pattern 8 predicts genes with an intermediate kinetic behavior (i.e. *rstA* – *c*=0.946, *mig-14* – *c*=0.922, *pcgL* – *c*=0.932)

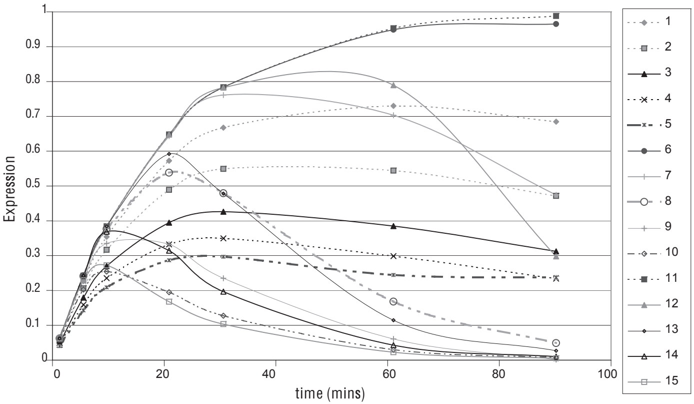


Figure 6.6 . Predicted expression patterns.
Multiple patterns result from scanning parameters of the single-input network motif controlled by the PhoP protein. Different symbols indicate distinct temporal order and intensity dynamics of the target genes.

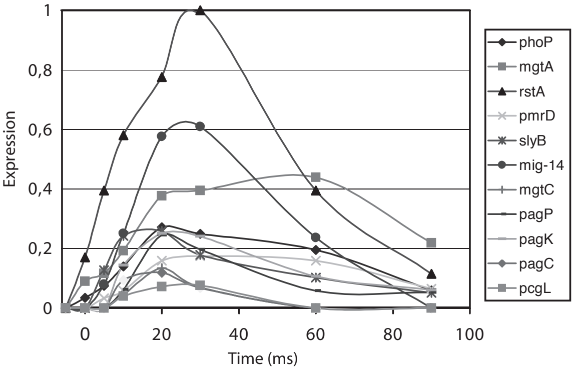


Figure 6.7 . Phop regulated genes growth kinetics for GFP.
The promoter activity and growth kinetics for GFP reporter strains with high-temporal resolution show different kinetic behavior of PhoP regulated genes

6.4.1 Concluding remarks

The experiments and simulations for the genetic regulatory network PhoP/PhoQ-PmrA/PmrB allowed us to extract several conclusions about the method shown in this work and the network under study: *(1) GENIE predicts interactions that explain experimentally observed behaviors*, the final refined architecture let predict the *in vivo* observed interaction between the two-component systems PhoP/PhoQ and PmrA/PmrB; *(2)The final refined architecture exhaustively predicts the network functionality*, proposing new hypothesis for the PmrD and PmrA phosporilated interaction as well as new hypothesis for the phosphatase-kinase action; *(3) PmrA/PmrB and PhoP/PhoQ constitute a robust and flexible genetic network*, our final refined model satisfies all of the constrains and only 4.5% of its parameters are limited to accept values that cover 25% of their biological meaningful range; *(4) PhoP/PhoQ and PmrA/PmrB show a coordinated functionality,* both the forward and backward connections (transcriptional and post-transcriptional) between the two systems are required to let the systems interact. When we did not model these interactions, the resulting architectures could not satisfy the biological patterns. *(5) GA approach is adequate for inferring regulatory genetic networks*, the heuristic produces a better proportion of feasible solutions and better numerically evaluated results (i.e. scores) for the predictions which are highly correlated to experimental values.

Moreover, GENIE results can be used to predict the behavior of other set of genes (i.e. clusters) regulated by PhoP/PmrA, because it indicates the range of values for parameters of distinct network motifs, to scan and obtain the expected behavior of not modeled species.

Finally, we would like to remark that the decisions regards the architecture enhancement (i.e. adding or not new elements) are based on the definition of conditions (i.e. realness, robustness and flexibility) to be fulfilled by the models, all of them satisfied by our final refined model for the PhoP/PhoQ-PmrA/PmrB genetic regulatory network. Our approach to model regulatory genetic networks provides a framework to explore genetic regulatory networks, including biochemical elements (i.e. different equations to model the reactions), biological (i.e. constrains imposed to the networks), and computational (i.e. simulations and a learning strategy that tackles the high dimensional search space).

Chapter 7

Plotting schizophrenia risk factor function by learning phenotypic/genotypic relations

7.1 Introduction

Schizophrenia is a disease diagnosed most commonly in the third decade of life, upon the onset of psychosis and functional deterioration. Medications can suppress psychotic symptoms more or less successfully over a period of weeks to months, but functional improvement is very limited and usually leaves the affected person unable to lead an autonomous life. The mechanism of antipsychotic efficacy depends on antagonism of dopamine receptors, and the kinetics of receptor binding may determine the severity of drug induced parkinsonian motor impairment (a combination of clinically significant rigidity, bradykinesia, akinesia and/or tremor). Notably, motor impairment is the best predictor of long term functional impairment and poor quality of life in subjects with schizophrenia (Peralta, Cuesta et al. 2000). Yet, parkinsonian motor impairment has been consistently observed in at least a fraction of newly diagnosed patients with schizophrenia before they receive any medication, (Caligiuri, Lohr et al. 1993) (Chakos, Mayerhoff et al. 1992) (Wolff and O'Driscoll 1999) posing the paradox of simultaneous deficit (represented by parkinsonism) and excess (represented by antipsychotic-responsive symptoms) dopaminergic function. One possible explanation for this paradox provides also a window to explore the elusive neurobiology of this extremely complex disease; a developmental injury to the dopaminergic system could indeed be causal in the late generation of psychosis. In addition to the presence of parkinsonism, a number of observations suggest loss of dopaminergic neuronal function in schizophrenia:

i. Neurons in the ventral tegmental area are probably reduced in number and have dystrophic appearance (Kolomeets and Uranova 1999).

ii. The dorsolateral prefrontal cortex (DLPFC) of schizophrenic brains has a marked reduction in density of dopaminergic terminals, and in expression of the dopamine-regulated protein DARP-32 (Akil, Pierri et al. 1999).

iii. Reduced availability of dopamine in the DLPFC is associated with increased risk of schizophrenia, and with poor performance on a working memory task (Egan, Goldberg et al. 2001; Goldberg, Egan et al. 2003) and cortical metabolic inefficiency (Egan, Goldberg et al. 2001).

These findings resemble the observations in an experimental model of lesions to dopaminergic neuronal projections during development; (Carter and Pycock 1980) a similar developmental lesion in humans could underlie schizophrenia as shown by multiple imaging studies (Breier, Su et al. 1997; Egan, Goldberg et al. 2001). We hypothesize that subjects with chronic untreated schizophrenia express a phenotype dependent upon loss of dopaminergic neurons in the mesencephalon during development and whose measurable manifestations include: *i*) parkinsonism, *ii*) hyperechogenicity of the substantia nigra, *iii*) deficits in working memory and executive function, and *iv)* deviance of specific temperament dimensions. If this phenotype is a core feature of schizophrenia its presence should be sufficient to identify reliably those affected without additional reference to psychotic symptoms. Additionally, if the same deficits express a genetic predisposition or vulnerability to the disease, they should be also present, albeit in a less severe form, in unaffected relatives of affected subjects.

7.1.1 Genotypic characteristics of schizophrenia

Up to 80% of the variance in liability to schizophrenia can be attributed to genes, and yet these genes have remained elusive (Harrison and Weinberger 2005). Linkage studies have shown support for loci on regions 6p24-22, 1q21-22, and 13q32-34, and promising findings on regions 8p21-22, 6q16-25, 22q11-12, 5q21-q33, 10p15-p11, and 1q42 (Craddock, O'Donovan et al. 2005). A metaanalysis found the strongest signals on regions 8p, 13q and 22q (Badner and Gershon 2002). Consensus interpretation of these findings suggests that the contribution of genes is likely to be polygenic and unlikely to be sufficient by itself to allow expression of the syndrome. Specific genes or loci have been implicated in schizophrenia susceptibility with some replicability; association with the genes for neuregulin 1 (NRG1), dysbindin 1 (DTNBP1), D-aminoacid oxidase (DAO) and D-aminoacid oxidase activator (DAOA), and regulator of G-protein signaling 4 (RSG4) have all been replicated (albeit not on every sample) (Craddock, O'Donovan et al. 2005). A number of facts complicates interpretation of genetic findings. First, susceptibility associated with all of the listed genes or loci overlaps with other forms of psychosis, notably bipolar affective disorder (Craddock, O'Donovan et al. 2005) to a greater degree than clinical overlap in epidemiological samples would have predicted. Population studies also show that schizophrenia is more likely to be inherited from the father than from the mother, that increasing paternal age confers increased risk, and that prenatal exposure to famine or to viral infections, and hypoxic obstetric complications also increase the risk of disease.(Abel 2004) Thus, schizophrenia is a complex trait that may receive contributions from a variety of sources including genes with small effect sizes, locus heterogeneity, epistasis, genetic imprinting, and environmental influences (Abel 2004).

7.1.2 Phenotypic characteristics of schizophrenia

Differences in the overt clinical characteristics of affected members of the families such as the type and severity of symptoms do not appear to be useful in defining subgroups of families that are segregating particular sets of predisposing genes. Also, the risk of schizophrenia amongst the offspring of affected and unaffected monozygotic co-twins is equivalent indicating that the transmission of predisposing genes is not dependent on the phenotypic expression of the disorder. Endophenotypes have been found particularly useful to attempt to fill the gap between the genes and the disease manifestations. Sensory motor gating deficiency, eye-tracking dysfunction, and working memory deficits have been used as endophenotypes of schizophrenia with some success (Horan, Braff et al. 2008). Sensory gating-deficiency appears to be heritable (Freedman, Olincy et al. 2003), and has been linked to a susceptibility locus on chromosome 15 including the gene for the α7 nicotinic receptor (Horan, Braff et al. 2008). Patients with schizophrenia also have deficits in smooth pursuit eye movements when compared to healthy subjects, and the trait is also present more frequently in their first degree relatives (Horan, Braff et al. 2008). A locus of susceptibility to smooth pursuit abnormalities has been found in chromosome 6, but the finding has not been replicated, and more importantly is not specific to schizophrenia (Lencer, Trillenberg et al. 2004). Deficits in the performance of tasks that require working memory or executive function are a central part of the phenotype of patients with schizophrenia, are heritable (Cannon, Huttunen et al. 2000), and have been specifically linked to a locus in chromosomes 1 (Gasperoni, Ekelund et al. 2003), 2 and 4 (Paunio, Tuulio-Henriksson et al. 2004). Association and physiological studies have also shown that variations in the genotype of the enzyme catecol-o-methyl transferase (COMT) with a small increase in the risk of schizophrenia (Egan, Goldberg et al. 2001), and with poor performance on a working memory task (regardless of diagnosis) (Carter and Pycock 1980; Egan, Goldberg et al. 2001). Variation in the genotype of COMT results in a fourfold change in its catabolic activity over dopamine, and has a measurable impact on cognitive performance and DLPFC metabolic efficiency during a working memory task (Egan, Goldberg et al. 2001), implicating a dopaminergic deficit in the DLPFC as the underlying mechanism for the impairment in schizophrenia. Likewise, a haplotype of SNPs within a metabotropic glutamate receptor gene (mGluR3) was strongly associated with schizophrenia, with poorer performance on tests of prefrontal and hippocampal function, and abnormal activation on functional imaging.

7.1.3 Learning phenotypic genotypic relations

The central hypothesis of this work is that a reduction in dopaminergic function (that is, a hypodopaminergic state) present in subjects with schizophrenia and (to a lesser extent) in at-risk relatives is an endophenotype dependent on multiple susceptibility genes.

We also hypothesized that if developmental injury to dopamine neurons is a central part of the pathogenesis of schizophrenia, a vulnerability to it is likely to be present in unaffected at-risk first degree relatives of patients. To test this possibility we also looked for evidence of dopaminergic deficits in the siblings of untreated patients with schizophrenia.

Our approach is based on independently learning phenotypic and genotypic profiles (i.e. overlapped or non-disjoint clusters of individuals sharing a set of clinical features and genetic variations respectively) and extracting the most qualitative and quantitative relations. To allow the exploration and retrieval of the most significant clusters of individuals for each domain we do not incorporate subject status into the searching process. Thus, our approach emphasizes on the relevant relations existing in the population of study, which results critical to find multiple variations that occur to only a subset of the population. In this way, and in contrast to supervised methods where the subject status is determinant to learn the clusters, we obtain a set of unbiased and qualitative profiles. This same reasoning explains our method of learning the genotypic profiles independently of the phenotypic ones.

Two general non mutually exclusive models are postulated to study the genetic implications of Schizophrenia: *"a common disease and multiple common genetic variations with small effects"* and *"a common disease and multiple rare variations with large effects"*. Our approach of learning qualitative phenotypic/genotypic relations does not assume the prevalence nor discard any of these models, as is the case of the traditional quantitative traits locus (QTL) approach.

The data reviewed indicate that endophenotypes provide a model to identify genes contributing to the vulnerability of schizophrenia with small effect sizes. Since liability to schizophrenia appears to vary on a continuous scale, quantitative measures of phenotypic variance (particularly those indicating a pathophysiologic process) may be more likely to be successful in identifying contributing genes (Gottesman and Shields 1973). Quantitative traits also extract maximal information from the study sample resulting a powerful strategy for gene identification. The contribution of each QTL to the total variance of a behavioral measure is almost certainly small, of the order of 10% or less based on the results of exhaustive animal studies. A caveat is necessary here, as an alternative explanation for the available data has been proposed; according to this hypothesis the findings would result not from the cumulative effect of many common mutations with small effect sizes as discussed above, but rather from many unrelated, individually rare, highly penetrant mutations with large effect sizes (McClellan, Susser et al. 2007). Indeed, two very recently published studies of copy number variations in schizophrenia revealed that microdeletions or duplications (in 15q13.3 and 1q21.1) that have only nominal association in large samples have indeed large effect sizes in a very small number of subjects.

The above findings implies that reducing the phenotypic variance to a single scale, should only decrease the likelihood of identifying rare mutations, because of a "dilution" of the individual effects (McClellan, Susser et al. 2007). Interactions between (or among) QTLs contributing to the same behavior are most probably complex, including negative interactions and non-additive facilitation (epistasis), making detection very difficult (Purcell, Neale et al. 2007). On the other hand, by applying QTL mapping to multiple tests presumed to measure the same physiological process, one can predict that a QTL affecting all measures in the expected direction (that is, a QTL acting pleiotropically) would be a locus that influences the physiological process under study. One can further predict that the same QTL would not have an effect on measures unrelated to the same physiological process, that is, that its effect would be specific. One way to establish both the pleiotropy and specificity of a QTL effect is to exploit the predicted direction of the allelic effects. Ideally, assessment of the function under study should include measures that are predicted to change in opposite directions under the influence of the same QTL, circumventing the need for large effect sizes in favor of a robust direction of effect on measures changing in opposite directions (Flint 2003). In addition, multivariate analysis of the diverse measures allows identification of the common source of variance presumably representing the underlying function regulated by the QTL (Purcell, Neale et al. 2007). Using this approach QTLs can be calculated for complex traits using sib pair datasets. The addition of non-stratified population controls should allow, in combination of unbiased clustering of phenotypic traits, a broader range of expression of the endophenotype and increase the ability to detect the effect of the hypothesized QTL. Also, the use of endophenotypes in combination with a case control approach may allow segregation of traits into those shared by affected and unaffected relatives (which may reflect genetic sources) and those specific to individuals with the clinical phenotype (which may reflect involvement of non-genetic etiological factors or factors secondary to the illness itself).

Finally, because of the unsupervised machine learning approach chosen, we obtain a collection of phenotypic/genotypic relations that group subjects exhibiting distinct status distributions (i.e. affected, relatives and control subjects). These relations let us plot the risk factor surface for the population of study and eventually might be the foundation of predictors that infer genotypes based on phenotypes, and vice versa.

7.2 Datasets

7.2.1 Study Population.

The population of Jujuy, Argentina, numbers 650,000, of which about a third are Kolla. Access to the Mental Health system is limited by transportation difficulties and geographical isolation. We trained 650 outreach workers to detect symptoms of severe mental illness. Outreach workers visit every village/town twice a year and are familiar with the culture, language, and health attitudes of the indigenous population. Prospective cases were referred for assessment using the provincial epidemiology reporting system and the referral information included the availability of siblings or parents as well as neighbors of the same age to act as controls. Treatment was offered free of charge independently of participation in the study. Researchers contacted prospective subjects and obtained consent for participation.

7.2.2 Ascertainment.

Prospective subjects were interviewed by one of three psychiatrists (MC, EP, GG) certified on the use of the Schedules for Clinical Diagnosis in Neuropsychiatry (SCAN, WHO) Spanish Version 2.1, and these interviews were videotaped. Following face to face interview, one of three psychiatrists (SD, CL, AG) certified on the use of SCAN carried out blind reviews of the videotaped interviews and re-scored the SCAN; we then ran associated computer algorithms (CATEGO) that generate categories of Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV). Disagreement would lead to a third revision. Subjects with DSM-IV schizophrenia were included as index subjects only when at least one unaffected relative (by SCAN) agreed to participate as well. Normal volunteers (controls), usually a neighbor of the index subject matched by ethnic background, age, sex, and education were also recruited and interviewed with SCAN. Training and certification of psychiatrists on the use of SCAN was carried out previously by two WHO certified trainers (GdE, JE).

7.2.3 Evaluation of the phenotype.

Following ascertainment of diagnosis with SCAN, the index subject, his/her sibling and the normal control were assigned a number randomly adjudicated (three numbers placed in a container and drawn after shaking). Subject identification numbers were written in colored wristbands and the entire assessment (movement disorder, neuropsychological assessment, temperament and character inventory, and transcranial ultrasound) was carried out by evaluators blind to diagnosis. The same number was used to label the videotaped SCAN examinations (for blind re-scoring), videotaped motor exams, and all other research materials.

i) Parkinsonism was evaluated from videotaped exams obtained by research physicians following standard guidelines for the Unified Parkinson's Disease Rating Scale (UPDRS) and scored blindly by one of two movement disorders specialists (GdE, FM). *ii*) Cognitive performance was evaluated on tests for executive function, working memory, visual/spatial attention, impulsivity and dexterity by one of four trained neuropsychologists (GGA, MB, CA, XG) blind to diagnosis. Tasks were selected for relative independence from literacy (none of the tasks requires reading or writing) to minimize distortion based on educational level. Tasks, in order of administration, were: Purdue pegboard task, Word List I (Weschler Memory Scale III), manual sequencing task of Luria, Symbol Digit Modalities Test, Word List II (Weschler Memory Scale III), Five Digit Test, Raven's Progressive Matrices, Wisconsin Card Sorting Task, Spatial Span (Weschler Memory Scale III), Trail Making Test– Oral version, Symbol Cancellation Test. *iii*) Transcranial ultrasound performed by an expert sonographist (NF) blind to subject condition with a 2.5 MHz transducer (s3-1, Philips EnVisor, Koninklijke Philips Electronics, N.V.) through a temporal bone window (penetration depth = 16 cm, dynamic range = 45 dB). The substantia nigra was reliably identified within the butterfly-shaped mesencephalon. Quantification of echogenic area was carried out on saved images by a blind evaluator (GdE) using a region of interest approach. *iv*) Personality traits were measured with the Temperament and Character Inventory (Spanish version) (TCI). TCI is self administered. To ensure that translation variants were not interfering with the answers we provided assistance by one of the investigators while the subjects were filling the answering sheet.

7.2.4 Genotyping.

We used the Data Transfer Tool (DTT) and the Affymetrix GeneChip Genotyping Analysis Software (GTYPE) to read raw data and generate SNP calls from the Affymetrix Mapping 10K 2.0 Array (http://www.affymetrix.com/products/software/specific/gtype.affx). Arrays not meeting a 93% call rate threshold were not used for analysis. Furthermore, SNPs failing Hardy-Weinberg equilibrium (p-value < 0.00001) and SNPs with excessive missingness (call rate < 90% for a particular SNP) were excluded from analysis. Additionally, the minor allele frequency for each SNP was calcualated. Minor allele frequency (MAF) was computed.

7.3 Data Analysis.

7.3.1 Exploring phenotypic commonalities

We employ a full battery of clustering methods to learn phenotypic clusters, without imposing any constrain beyond the ones specified by each method (see Table 7-1 for a list of applied methods). Indeed, each method is configured with a broad sample of input parameters (e.g. Fuzzy *c*-means method initialized to learn different number of *c* clusters). Thus, we obtain an extensive collection of non-disjoint cluster, where observations belong to more than one cluster. In this way, we are able to cluster the dataset by the different optimal number of partitions indicated by the validity indices. Although this approach to the exploration of the variable space generates clusters that might exhibit high degree of overlap, which result an extra difficulty to cope, we maximize the probability of finding the commonalities recovered by each method and their respective set of parameters.

Table 7-1 Clustering methods & parameters

Method	Parameters	Values
Fuzzy *c*-means [	Partition number (*c*)	*c* between 2 and 8
Hierarchical latent classes	Score Metric (*SM*)	*SM*= {LS, AIC, BIC, L2, MC, CS}
Latent Classes	Discrete factor (*d*)	*d*=3
NMF	factor (*f*)	*f*: between 2 and 14

An evaluation of clusters and consequent selection of optimal ones might be applied at this stage (e.g. by applying multi-objective and multimodal optimizations techniques). This would be suitable for a strategy that would aggregate distinct sources of information in this learning stage. However, this mining process is an initial step of a multi-phase learning strategy. Thus, we neither compact nor reduce the collection of learnt clusters. Instead we opt for a "lazy" approach, retaining all of the distinct cluster as valid options for subsequent learning phases (Figure 7.1.)

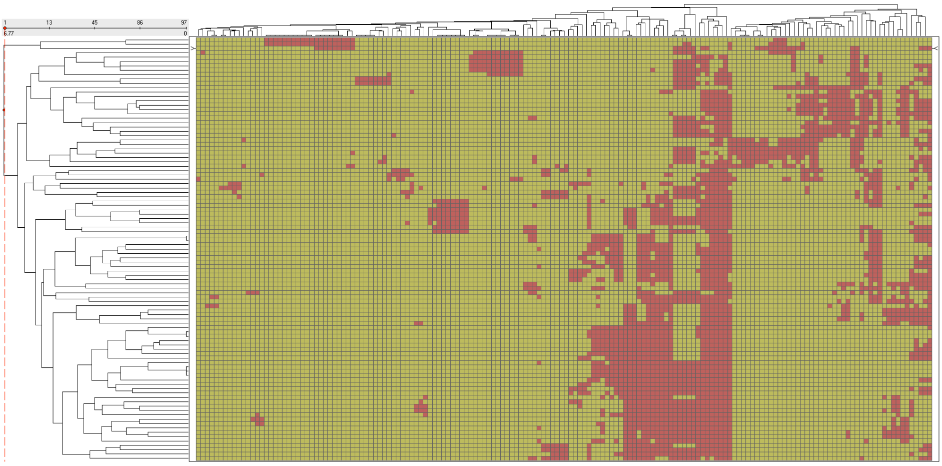
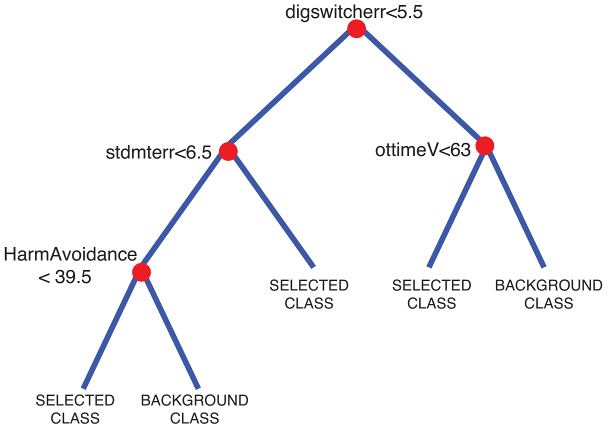


Figure 7.1 Clusters learned for the clinical dataset
A subset of the clusters (columns) learned of the clinical dataset by showing the subject membership (orange; belongs; yellow; do not) to each cluster.

Building phenotypic profiles: Our choice of using a set of clustering methods, without imposing non-overlaping constrains, gives us the chance of use almost any unsupervised clustering available method. Because not every method incorporates a feature selection, we learn the set of variables that best characterize each cluster, and obtain profiles interpreted easily by the experts. We employ decision trees (Mitchell 1997), a supervised machine learning method, to detect the variables that best discriminates among the subjects retrieved by each cluster. We label subjects belonging to the cluster as "selected class" and the remaining ones as "background class" and obtain a local feature selection for each cluster, based on the information gain of each variable (Figure 7.2). Similarly, we can employ stepwise regression (Hocking 1976) as feature wrapping. These two methods, select the minimum set of variables that best discriminates observations. We note that an opposite viable alternative is to employ metrics, that instead of selecting a minimal set of determinant variables would learn each feature weight. For example, choosing the relative entropy (Cover and Thomas 2006) we would weight the relevance of the features by their information gain (Mitchell 1997).



Subject ID	Status	
10004	2	TP
10013	1	TP
10015	1	TP
10017	2	TP
10065	1	TP
10066	2	TP
10114	1	TP
10119	1	FP

Figure 7.2 Decision tree for learning the features that best characterize a cluster
The tree recovers all of the subject belonging to the cluster and one false positive (i.e. subject 10119 which is not label as selected class). Branches at inner nodes (red circles) correspond to cut-off values for the determinant features

7.3.2 Exploring genotypic commonalities

Schizophrenia is a complex trait that may receive contributions from a variety of sources, giving support to the model *"multiple rare variations with large effects"*. Traditional QTL analysis facilitates the study of polygenic inheritance (also known as quantitative or multifactorial inheritance), which refers to inheritance of a phenotypic characteristic that is attributable to two or more genes and their interaction with the environment. Unlike monogenic traits, polygenic traits do not follow patterns of Mendelian inheritance (qualitative traits). Instead, their phenotypes typically vary along a continuous gradient depicted by a bell curve

(Pharoah, Antonio et al. 2008). A reference implementation of QTL is the Plink software package, specifically designed to accept a set of subject, each one described by their respective collection of SNPs and an attribute that reflect the phenotypic value within a meaningful scale. The method then calculates the mean phenotype value for each genotype (i.e. AA, AB, BB), and the corresponding *p-value* of the SNP (Figure 7.3). This supervised method is suitable for the model *"multiple common genetic variations"*, but is unable to discriminate among different SNP variations that only occur in a subpopulation.

```
SNP_A-1513804                                          p-value:      0.007993
                    B/B          B/A          A/A
      COUNTS                16           21            18
      FREQ              0.2909       0.3818        0.3273
      MEAN               -2.72       -1.467         2.988
      SD                 5.509        5.883         6.841
.
.
.
.
SNP_A-1511565                                          p-value:      0.008988
                    A/A          A/B          B/B
      COUNTS                11           29            12
      FREQ              0.2115       0.5577        0.2308
      MEAN              -4.134      -0.7796         2.743
      SD                 5.098        6.401         6.277
```

Figure 7.3 Synthesis of QTL analysis
We exemplify QTL analysis results by showing two significant SNPs (*p-value*<0.01). Because the phenotype single value (*ph*) is learned by using *Linear Discriminat Analysis,* and our clinical dataset exhibits null values, the analysis is reduced to only 52 subjects (ph_{patient}=-7; ph_{relative}=0; ph_{control}=7) .

Our approach to learn genotypic profiles is designed to accomplish two goals: Firstly, it should detect SNPs occurring to only a subset of the subjects under study; and secondly, it should detect a collection of SNPs that might map to *multiple* genes or differential chromosome regions. Thus we opt to apply a unsupervised approach, making use of available software able to handle this special set of data, where the number of variables is far extent than the number of observations. Indeed this kind of dataset is usually processed when analyzing microarray data, with the exemption that SNPs are categorical values where expression data reflect continuous values. To circumvent this difference and adapt our problem to already proved mining tools, we transform the SNPs dataset by converting each SNP values into dummy variables which preserve Hamming distance (i.e. AA is codified as vector [0 0 1]; AB as [0 1 0]; and BB as [1 0 0]). Then we cluster SNPs dataset by using the software package bioNMF (Mejia-Roa, Carmona-Saez et al. 2008), which implements the non-negative matrix factorization to learn form microarray assays. Instead of learning the factor value by using a validity index(i.e. cophenetic index) we learn clusters by using a range of values (i.e. between 2 and 14 factors) (Figure 7.4).

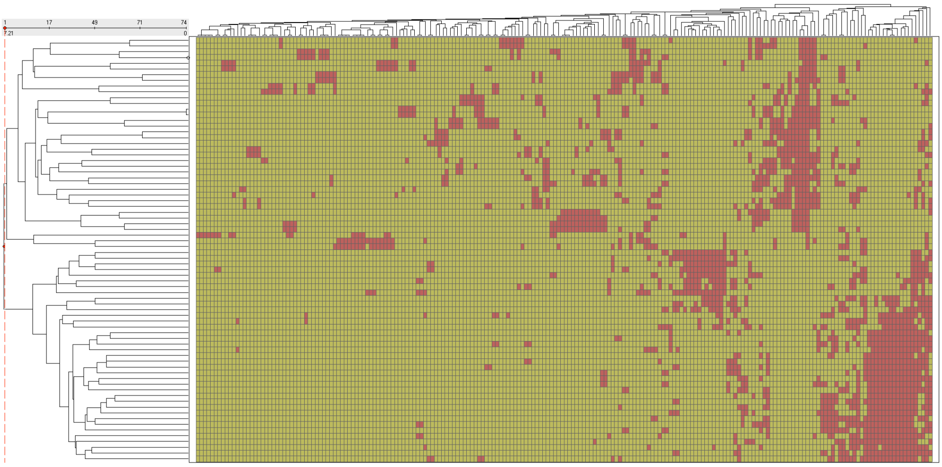


Figure 7.4 Clusters learned for the SNPs dataset
The clusters (columns) learned of the SNPs are represented by subject membership (orange; belongs; yellow; do not) to each one. The different number of subjects included in the clusters possibilities the analysis of both the "*multiple common genetic variations*" and the "*multiple rare variations*" models.

Selecting relevant SNPs for clustered subjects: We calculate the *p-value* for every SNP by using the hypergeometric distribution that gives the chance probability (i.e. probability of intersection *PI*) of observing at least *p* subjects belonging to the cluster (of size *h*) that have a common value for the SNP within the total number (*n*) of subjects exhibiting this same SNP value:

$$PI(h,n,g,p) = 1 - \sum\nolimits_{q=0}^{p} \binom{h}{q}\binom{q-h}{n-q} \Big/ \binom{g}{h} \tag{1}$$

where *g* is the total number of subjects, such that the lower the *p-value* the better the association (Tavazoie, Hughes et al. 1999). We termed "*pseudohaplotypes*" to those SNPs whose *p-values* are below 0.01.

7.3.3 Fusing distinct domain of knowledge

We learn relations among phenotypic and genotypic profiles by evaluating the degree of matching between the subjects group by profiles independently learned for each domain. To accomplish it, we make use of the probability of intersection metric (equation(1)), applied to the subjects in phenotypic profiles (*h*) and subjects in genotypic profile (*n*) where *g* is the total number of subjects included in both datasets (Figure 7.5). In this way, we can identify interesting candidates relations that highlight commonalities of each domain for the approximate same subjects.

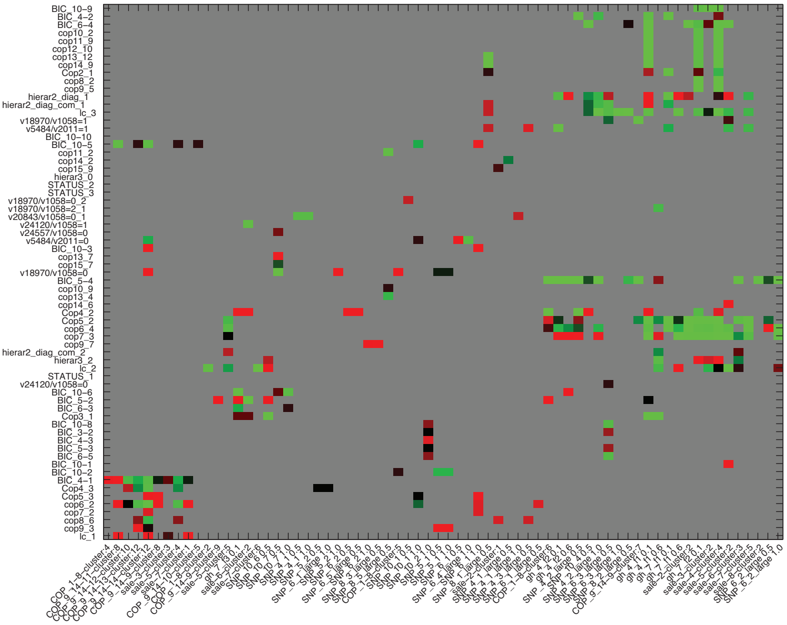


Figure 7.5 Heatmap of probability of intersection between phenotypes and genotypes. A matrix of correlations between SNP clusters (pseudohaplotypes) vs phenotypic profiles The hypergeometric function is used to calculate the *p-value* of every pair of profiles, and those with the lowest probability (red: low; green high) are selected.

This learning strategy results in a collection of relations (learnt by the subject correspondence), which encompass both the clinical features and relevant SNPs for the subjects grouped. Indeed, because we do not compact neither phenotypic nor genotypic profiles, we enable the finding of specific relations, at the expense of obtaining a redundant set of overlapped relations. Thus, an evaluation process is required to compact the relations set: we apply a multimodal optimization by partitioning the matrix of relations by "quadrants" (Figure 7.6) and selecting a subset of the most representing relations for each one. We learn quadrants boundaries by applying the hierarchical clustering method to the subjects grouped by the phenotypic profiles and repeat this clustering of subjects to those gather by the genotypic profiles (i.e. double clustering). We select the top level branches of each domain as quadrants boundaries. Thus, a set of subjects that exhibit clinical commonalities can be view, analyzed or "explained" by distinct sets of SNPs. Indeed, we apply a second level of multimodal optimization, by picking those relations belong to the same quadrant but do not fall in the same subject local neighborhood or *niche* (Deb 2001)

In this way, we apply a two level multimodal optimization of relations: we allow a profile from one domain to be related to more than one profile in the other domain (quadrants) and allow more than one relation to be selected from the same quadrants if the subjects that support the relation do not belong to the same niche (subject level).

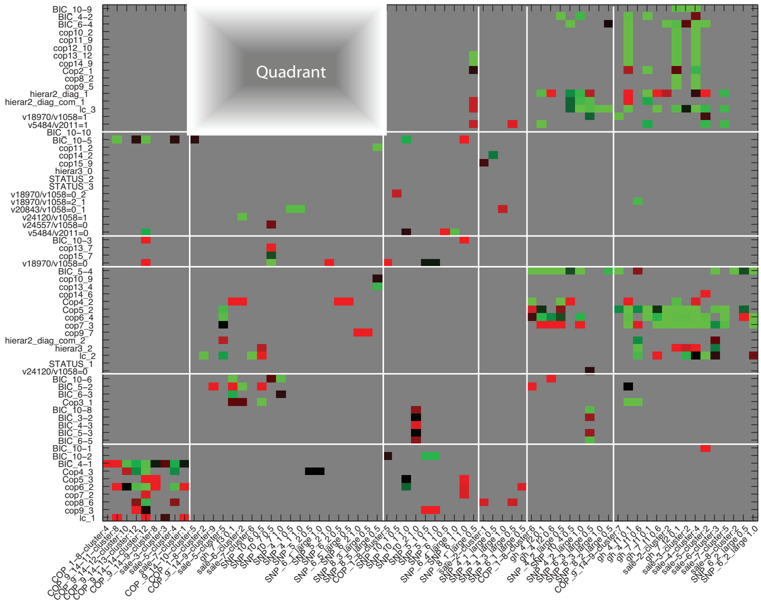


Figure 7.6 Matrix of phenotypic/genotypic relations analyzed by quadrants.
We cluster phenotypic (rows) profiles independently of genotypic profiles (columns) to learn quadrants boundaries.

Finally, we note that the domain independent learning strategy let us analyze the available dataset, handling missing values and missing records. For example, if the SNP analysis for a subject is discarded because a low recall rate, we make use of its corresponding clinical data to reinforce the learning of phenotypic profiles.

7.3.4 Learning schizophrenia risk factor of the relations

Our unsupervised machine learning approach does not incorporate the status of subjects to find profiles. Thus, it obtains a collection of phenotypic/genotypic relations that have a broad distribution of subject status. To quantify the risk factor of each relation we apply as a weighted sum of cases (equation (2)) :

$$Risk = \frac{\sum |ST_i| W_i}{\sum |ST_i|} \tag{2}$$

where the weights are given by epidemiologic risk of schizophrenia (0.01 for controls, 0.1 for relatives, and 1 for affected subjects).

7.4 Results

We collected complete data of 105 subjects (35 with chronic untreated schizophrenia, 35 first degree relatives and 35 population controls). The average age of affected subjects was 28.9 ±11.7, of relatives 34.7 ±15.1, and of controls 31.4 ±12.5. Educational level was similar in all three groups (schizophrenia=1.92 ±0.7, relatives=2.1 ±1.1, controls=2.7 ±1.1, on a scale where completed elementary education =1, high school =2, associate's level degree=3, and academic degree=4). Mean duration of untreated psychosis was 6.83 years (±9.79) with extremes of 3 months and 49 years. SCAN proved valid and highly reliable as a diagnostic tool. We found no disagreements on CATEGO diagnoses, or between clinical diagnosis after semistructured interview (using SCAN) and blind CATEGO diagnoses. Average interrater reliability (Cohen's κ) scores for individual items (41 videotaped interviews scored independently by three blind raters) were: section 1=0.659, section 15=0.833, section 16 =1, section 17=0.702, section 18=0.867, section 19=0.646, section 22=0.432, and section 24=0.765. Neuropsychological assessments were more difficult on subjects with severe symptoms and could not be completed on two subjects. Interrater reliability for cognitive tests was calculated by repeated assessments of unaffected subjects with a short version of the battery; average κ was 0.55 ± 0.02 (highest Trail Making=0.82, lowest WCST=0.23-0.29). Most κ scores were above 0.7. Table 7-2 provides descriptive statistics for each variable and results of the ANOVA comparing schizophrenia, relatives and controls across all phenotypic features.

Subjects with schizophrenia had significant impairment on UPDRS motor subscores. Scores reflect primarily bradykinesia and akinesia, because tremor was relatively rare, and rigidity could not be assessed on videotape and was not scored (thus, the maximum possible score is 88 instead of 108). Because subjects with schizophrenia commonly exhibited rigidity on clinical exams, reported scores underestimate motor impairment. Relatives had significantly more motor impairment than controls. Subjects with schizophrenia had significant impairment on the Purdue pegboard task, a timed task requiring repetitive, accurate motion of the hand and arm involving flexion and extension. Relatives had significantly more impairment than controls on this task. A similar pattern was evident on the manual sequencing tasks of Luria, traditionally used to assess the effects of prefrontal damage in neurological patients.

The area of echogenicity on the right substantia nigra subjects with schizophrenia had a twofold increase compared to controls ($p<0.05$), and in relatives it was intermediate in size and significantly different from the latter (Figure 7.7). Echogenicity was increased on the left substantia nigra of schizophrenia, but this difference did not reach statistical significance.

Regardless of the magnitude of individual differences, an overall tendency was found such that subjects with schizophrenia displayed worst performance on cognitive tasks and controls were best, with relatives scoring in between. Between groups comparisons reached statistical significance for parts of the Five Digits Test, the visuospatial attention task, the Symbol Digit Modalites Test, the word lists task, the Trail Making test, and the symbol cancellation test. The difference between controls and relatives, and between relatives and subjects with schizophrenia progressively increased as the complexity of the task increased in the Trail Making test (Figure 7.8). Scores of neuropsychological testing generate two kinds of results, which could be characterized as dimensional (represented by continuous variables, usually time to completion), and counting or discontinuous variables such as error number. Simple inspection of the mean values for each kind of result suggests that dimensional variables display a graded impairment: controls<first degree relatives<schizophrenia; mean values for counting variables (errors or categories completed) could not distinguish between controls and first degree relatives, but rather separated these two categories from schizophrenia, therefore reflecting a threshold for state (i.e., disease vs. non-disease).

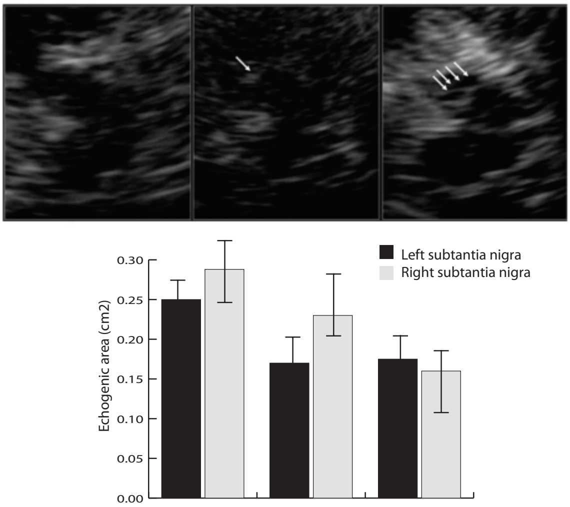
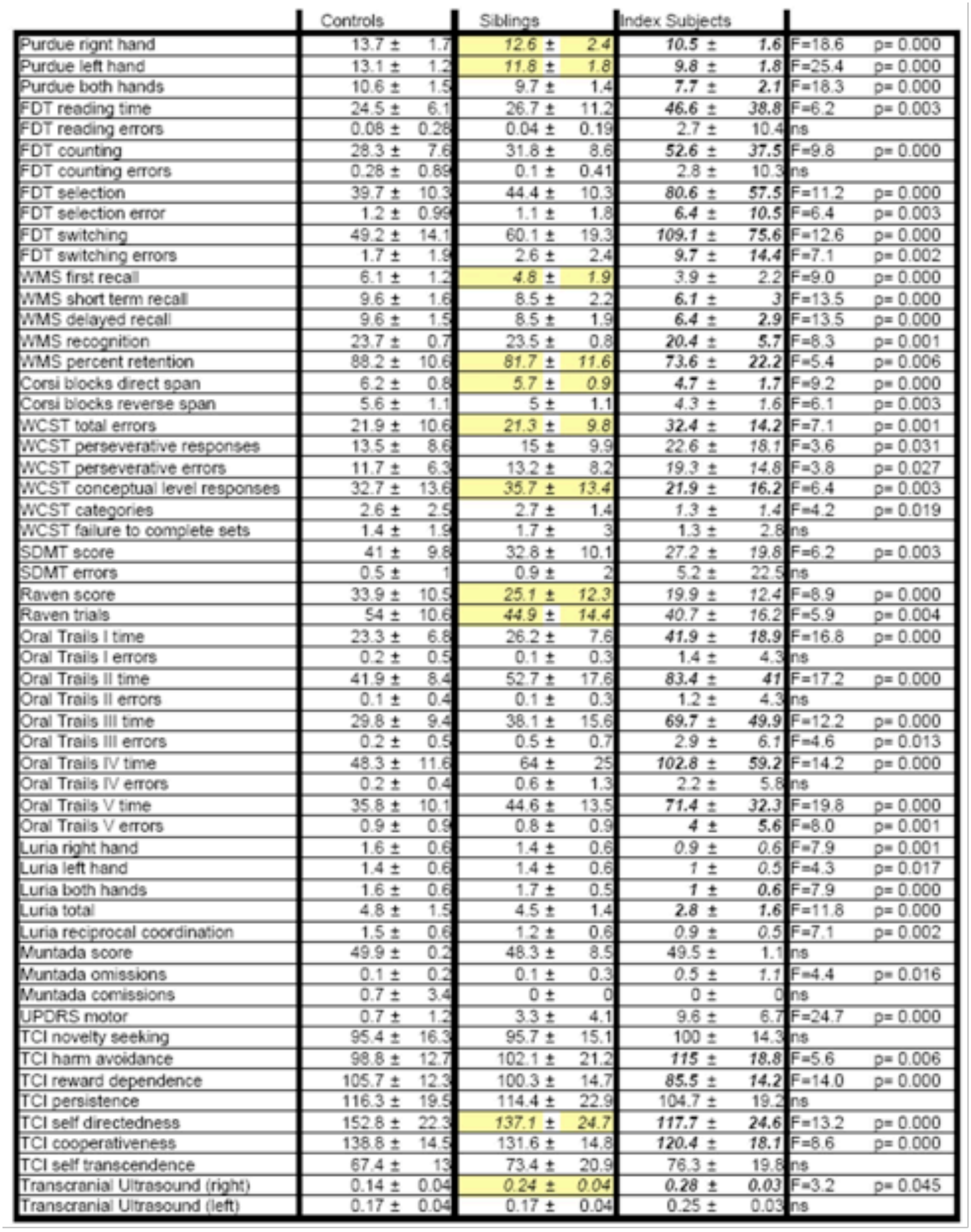


Figure 7.7 show transcranial ultrasound images.
Top panels show transcranial ultrasound images from a set of subjects obtained through the right temporal bone window. Quantitation of echogenic area was only performed from the ipsilateral side. Arrows indicate the hyperechogenic area in the substantia nigra of the affected subject and his relative. The control had no signal in the right brainstem. Bottom panel shows average echogenic areas of the left (black) and right (gray) substantia nigra in subjects with schizophrenia, their first degree relatives, and normal population controls. Bars represent means ± SEM, * p = 0.05, ** p <0.01 compared with control by ANOVA followed by Bonferroni post hoc comparisons.

Table 7-2 Summary of descriptive statistics for the sample.

F represents main effects of one-way ANOVA. Italics represent p≤0.05 vs. control. Bold represents p≤0.05 vs. siblings



	Controls		Siblings		Index Subjects			
Purdue rignt hand	13.7 ±	1.7	*12.6* ±	*2.4*	***10.5*** ±	***1.6***	F=18.6	p= 0.000
Purdue left hand	13.1 ±	1.2	*11.8* ±	*1.8*	***9.8*** ±	***1.8***	F=25.4	p= 0.000
Purdue both hands	10.6 ±	1.5	9.7 ±	1.4	***7.7*** ±	***2.1***	F=18.3	p= 0.000
FDT reading time	24.5 ±	6.1	26.7 ±	11.2	***46.6*** ±	***38.8***	F=6.2	p= 0.003
FDT reading errors	0.08 ±	0.28	0.04 ±	0.19	2.7 ±	10.4	ns	
FDT counting	28.3 ±	7.6	31.8 ±	8.6	***52.6*** ±	***37.5***	F=9.8	p= 0.000
FDT counting errors	0.28 ±	0.89	0.1 ±	0.41	2.8 ±	10.3	ns	
FDT selection	39.7 ±	10.3	44.4 ±	10.3	***80.6*** ±	***57.5***	F=11.2	p= 0.000
FDT selection error	1.2 ±	0.99	1.1 ±	1.8	***6.4*** ±	***10.5***	F=6.4	p= 0.003
FDT switching	49.2 ±	14.1	60.1 ±	19.3	***109.1*** ±	***75.6***	F=12.6	p= 0.000
FDT switching errors	1.7 ±	1.9	2.6 ±	2.4	***9.7*** ±	***14.4***	F=7.1	p= 0.002
WMS first recall	6.1 ±	1.2	*4.8* ±	*1.9*	*3.9* ±	*2.2*	F=9.0	p= 0.000
WMS short term recall	9.6 ±	1.6	8.5 ±	2.2	***6.1*** ±	***3***	F=13.5	p= 0.000
WMS delayed recall	9.6 ±	1.5	8.5 ±	1.9	***6.4*** ±	***2.9***	F=13.5	p= 0.000
WMS recognition	23.7 ±	0.7	23.5 ±	0.8	***20.4*** ±	***5.7***	F=8.3	p= 0.001
WMS percent retention	88.2 ±	10.6	*81.7* ±	*11.6*	***73.6*** ±	***22.2***	F=5.4	p= 0.006
Corsi blocks direct span	6.2 ±	0.8	*5.7* ±	*0.9*	***4.7*** ±	***1.7***	F=9.2	p= 0.000
Corsi blocks reverse span	5.6 ±	1.1	5 ±	1.1	*4.3* ±	*1.6*	F=6.1	p= 0.003
WCST total errors	21.9 ±	10.6	*21.3* ±	*9.8*	***32.4*** ±	***14.2***	F=7.1	p= 0.001
WCST perseverative responses	13.5 ±	8.6	15 ±	9.9	*22.6* ±	*18.1*	F=3.6	p= 0.031
WCST perseverative errors	11.7 ±	6.3	13.2 ±	8.2	*19.3* ±	*14.8*	F=3.8	p= 0.027
WCST conceptual level responses	32.7 ±	13.6	*35.7* ±	*13.4*	***21.9*** ±	***16.2***	F=6.4	p= 0.003
WCST categories	2.6 ±	2.5	2.7 ±	1.4	*1.3* ±	*1.4*	F=4.2	p= 0.019
WCST failure to complete sets	1.4 ±	1.9	1.7 ±	3	1.3 ±	2.8	ns	
SDMT score	41 ±	9.8	32.8 ±	10.1	*27.2* ±	*19.8*	F=6.2	p= 0.003
SDMT errors	0.5 ±	1	0.9 ±	2	5.2 ±	22.5	ns	
Raven score	33.9 ±	10.5	*25.1* ±	*12.3*	*19.9* ±	*12.4*	F=8.9	p= 0.000
Raven trials	54 ±	10.6	*44.9* ±	*14.4*	*40.7* ±	*16.2*	F=5.9	p= 0.004
Oral Trails I time	23.3 ±	6.8	26.2 ±	7.6	***41.9*** ±	***18.9***	F=16.8	p= 0.000
Oral Trails I errors	0.2 ±	0.5	0.1 ±	0.3	1.4 ±	4.3	ns	
Oral Trails II time	41.9 ±	8.4	52.7 ±	17.6	***83.4*** ±	***41***	F=17.2	p= 0.000
Oral Trails II errors	0.1 ±	0.4	0.1 ±	0.3	1.2 ±	4.3	ns	
Oral Trails III time	29.8 ±	9.4	38.1 ±	15.6	***69.7*** ±	***49.9***	F=12.2	p= 0.000
Oral Trails III errors	0.2 ±	0.5	0.5 ±	0.7	*2.9* ±	*6.1*	F=4.6	p= 0.013
Oral Trails IV time	48.3 ±	11.6	64 ±	25	***102.8*** ±	***59.2***	F=14.2	p= 0.000
Oral Trails IV errors	0.2 ±	0.4	0.6 ±	1.3	2.2 ±	5.8	ns	
Oral Trails V time	35.8 ±	10.1	44.6 ±	13.5	***71.4*** ±	***32.3***	F=19.8	p= 0.000
Oral Trails V errors	0.9 ±	0.9	0.8 ±	0.9	***4*** ±	***5.6***	F=8.0	p= 0.001
Luria right hand	1.6 ±	0.6	1.4 ±	0.6	*0.9* ±	*0.6*	F=7.9	p= 0.001
Luria left hand	1.4 ±	0.6	1.4 ±	0.6	*1* ±	*0.5*	F=4.3	p= 0.017
Luria both hands	1.6 ±	0.6	1.7 ±	0.5	***1*** ±	***0.6***	F=7.9	p= 0.000
Luria total	4.8 ±	1.5	4.5 ±	1.4	***2.8*** ±	***1.6***	F=11.8	p= 0.000
Luria reciprocal coordination	1.5 ±	0.6	1.2 ±	0.6	*0.9* ±	*0.5*	F=7.1	p= 0.002
Muntada score	49.9 ±	0.2	48.3 ±	8.5	49.5 ±	1.1	ns	
Muntada omissions	0.1 ±	0.2	0.1 ±	0.3	*0.5* ±	*1.1*	F=4.4	p= 0.016
Muntada comissions	0.7 ±	3.4	0 ±	0	0 ±	0	ns	
UPDRS motor	0.7 ±	1.2	3.3 ±	4.1	9.6 ±	6.7	F=24.7	p= 0.000
TCI novelty seeking	95.4 ±	16.3	95.7 ±	15.1	100 ±	14.3	ns	
TCI harm avoidance	98.8 ±	12.7	102.1 ±	21.2	***115*** ±	***18.8***	F=5.6	p= 0.006
TCI reward dependence	105.7 ±	12.3	100.3 ±	14.7	***85.5*** ±	***14.2***	F=14.0	p= 0.000
TCI persistence	116.3 ±	19.5	114.4 ±	22.9	104.7 ±	19.2	ns	
TCI self directedness	152.8 ±	22.3	*137.1* ±	*24.7*	***117.7*** ±	***24.6***	F=13.2	p= 0.000
TCI cooperativeness	138.8 ±	14.5	131.6 ±	14.8	***120.4*** ±	***18.1***	F=8.6	p= 0.000
TCI self transcendence	67.4 ±	13	73.4 ±	20.9	76.3 ±	19.8	ns	
Transcranial Ultrasound (right)	0.14 ±	0.04	*0.24* ±	*0.04*	***0.28*** ±	***0.03***	F=3.2	p= 0.045
Transcranial Ultrasound (left)	0.17 ±	0.04	0.17 ±	0.04	0.25 ±	0.03	ns	

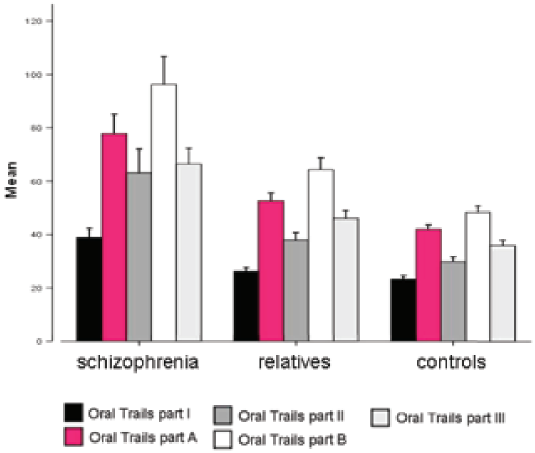


Figure 7.8 . The oral version of the Trail Making test without letters and presumably independent of language

Subjects must name numbers and fruits in the order established by the numbers, from small to large. Objects are initially organized in the sheet (baseline) and later randomized. On the second condition the examiner provides feedback based on colors instead of fruits. Subjects must avoid naming of the fruit and respond to the color. Bars represent means ± SEM, * p = 0.05, ** p <0.01 compared with control by ANOVA followed by Bonferroni post hoc comparisons.

7.4.1 Informativeness of the Kolla sample compared to HapMap data.

Chromosome wise bulk download of data for all the four populations (CEU, CHB, JPT and YRI) from the HapMap project was done and the overlapping set of SNPs with the Kolla population was identified (n=9137). MAF of these SNPs was compared with the allele frequency obtained form the Kolla control samples. A comparison of the allele frequency across chromosome in this preliminary analysis suggests that the Kolla population may be similar to the CHB and the JPT compared to the Caucasian population. Most notably, the proportion of informative SNPs was not different in the Kolla and the HapMap samples.

7.4.2 Phenotypic/genotypic relations.

We identified 30 key relations (Figure 7.9) hierarchically organized in 6 major classes, leading to the affected subjects alone, affected subjects and their relatives, relatives alone, relatives and controls, controls alone (Table 7-3). These relations are based on subject coincidence in both the clinical dataset analysis and SNPs analysis. Figure 7.10 shows the hierarchical structure of the relations of affected subjects as well as affected subjects and their relatives (i.e., the two categories of high risk).

Id.	ST	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26	R27	R28	R29	R30
10005	1	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN
10008	1	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN
10020	1	TN	TP	TP	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	FP	TP	TP	TN	TN	TN	TN	TN	TN	TN
10028	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TP	FP	FP	TN	TN	TN	TN	TN	TN	TN
10031	1	TP	TN	TP	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10095	1	TN	TP	TP	FN	TN	TP	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TP	TP	TN	TN	TN	TN	TN	FP	TN
10100	1	TN	TP	TP	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN
10105	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	TN	TN	TN	TN	TN	TN	TN
10116	1	TP	TN	TP	TN	TN	TN	TN	FP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	TP	TN	TP	TN	TN	TN	TN	TN	TN	TN
10001	1	TP	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN
10002	2	TP	TN	TP	TN	TN	TN	TN	TP	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10003	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	FN	TP	TP	TN
10004	2	TN	TP	TN	TP	TP	TP	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN
10006	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	FN	TN	FP	TN	TN	TN	TN	TN	TN	TN	TN	FP	TP	TP	TN	TP	TN	TP	TP
10009	2	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN
10010	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TP	TP	TP
10011	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TP	TN	TN	TP
10012	2	TP	TN	TP	TP	TN	TP	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TP	TN	TN	TN	FP	TN	TN	TN	TN
10013	1	TP	TN	FN	TP	TP	TP	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10015	1	FN	TN	TP	TN	TP	TP	TN	TP	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP
10016	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP
10017	2	TN	TN	TN	TN	TP	TP	TN	TN	TN	TP	TN	TN	TN	FN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	FP	TN	TN	TN	TN
10021	2	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TP	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TP	TN	TN	TN	TP
10022	1	TP	TN	TP	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TP	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN
10023	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	TN	TN	TP	TP
10025	2	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	FP	TN
10026	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TP	TN	TN	TN	TN	TN	TN	TN	FP
10027	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	TN
10030	2	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TN	TN	TN
10032	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TP	TN	TN	TN
10035	1	TN	TN	TP	TN	TN	TN	FP	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	FP	TN	TN	TN	TN	TN
10037	3	TN	TN	TN	TN	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP
10038	1	TP	TN	TP	TN	TN	TN	TP	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN
10040	2	TN	TN	TP	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN
10042	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TP	TN	TP	TP
10047	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN
10048	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TP	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	FP	TP	TN
10050	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN
10056	1	TP	TP	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN
10057	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TP	TN
10058	2	TP	TP	FN	TN	TN	TN	FN	TN	TN	TN	TN	TN	FN	FN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	FP
10059	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TP	TP
10062	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	FP	FP	TN	TN	TP	TP	TP
10065	1	TP	TN	TP	TP	TP	FN	TN	TN	FN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10066	2	TN	TN	TN	TN	TP	TP	TN	TN	TN	TP	TN	TN	TN	TP	TN	TN	FP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP
10067	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TP	TP
10068	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP
10069	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN
10071	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	FN	TN	TN	TP	TP	TP
10072	1	TN	TN	TP	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN
10075	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TP	TP	TP	TP
10082	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN
10083	1	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	FN
10085	3	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	FN	FN	TN	FP	TN	TP	TP
10088	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TP	TN
10089	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TP	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TP	TP
10091	1	TN	FP	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TP	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10092	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TP	TN	TP	TN	TN	TN	TN	TN	TN	TN
10093	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	FN	TP	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN
10096	2	FN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10099	2	TN	TN	TN	TN	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TP	TP
10101	2	TP	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TP
10102	1	TP	FN	TP	TN	TN	TN	TN	TN	FN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN
10103	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	FN	TP	TN
10110	1	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	FN
10114	1	TN	TN	TN	TP	TP	TP	TN	TN	TN	FP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FP	TN	TN	TN	TN	TN	TN	TN
10115	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10119	1	TN	TN	TN	TN	FP	TN	FP	TN	TN	TP	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10120	2	TN	TN	TN	TN	TN	TN	TN	TP	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10121	3	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	FN	TN	TP	TN	TP	TN
10122	2	TP	TN	TP	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
10123	1	TP	FP	TP	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TN	TN	TN	TN	FP	FP	TN	TN	TN	TN	TN

Figure 7.9 Distribution of subjects among the relations

First column indicates subject identification and is followed by their corresponding status (1 patient; 2 relative; 3 control subject). Remaining columns correspond to relations 1 to 30. Subjects recognized by determinant features learned for the phenotypic profiles are indicated in green if they correspond to the relation (TP); red if they are selected but do not correspond (FP) and yellow if they belong but are not retrieved.

Table 7-3 Relations summary

Relation	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
p-value	4.69E-14	1.90E-07	4.01E-13	4.29E-07	5.43E-09	7.52E-10	4.75E-04	4.48E-08	1.47E-12	7.93E-14
Risk	0.68	0.66	0.70	0.70	0.61	0.60	0.33	0.70	0.76	0.64
# Expected	17	8	24	6	7	9	7	6	15	15
# Predicted	15	9	19	5	8	8	6	7	13	17
#TP	15	7	19	5	7	8	4	6	13	15
#FP	0	2	0	0	1	0	2	1	0	2
#FN	2	1	5	1	0	1	3	0	2	0
Status 1	10	6	13	3	5	4	4	5	9	11
Status 2	5	3	6	2	3	4	1	2	4	6
Status 3	0	0	0	0	0	0	1	0	0	0
Relation	**R11**	**R12**	**R13**	**R14**	**R15**	**R16**	**R17**	**R18**	**R19**	**R20**
p-value	3.76E-09	1.23E-10	1.00E+00	7.79E-05	1.18E-10	3.35E-04	4.48E-08	2.44E-08	5.43E-09	7.01E-13
Risk	0.13	0.18	0.55	0.29	0.25	0.01	0.22	1.00	0.89	0.70
# Expected	8	12	6	8	9	6	6	7	8	15
# Predicted	10	10	0	7	10	3	7	9	7	15
#TP	8	10	0	5	9	3	6	7	7	14
#FP	2	0	0	2	1	0	1	2	0	1
#FN	0	2	6	3	0	3	0	0	1	1
Status 1	1	1	0	1	2	0	1	9	6	11
Status 2	0	1	0	2	3	0	4	0	1	4
Status 3	9	8	0	4	5	3	2	0	0	0
Relation	**R21**	**R22**	**R23**	**R24**	**R25**	**R26**	**R27**	**R28**	**R29**	**R30**
p-value	7.52E-10	3.98E-12	6.67E-14	4.98E-07	2.00E-06	5.38E-07	4.69E-11	6.67E-09	0.00E+00	1.87E-13
Risk	0.88	0.92	0.60	0.09	0.01	0.24	0.11	0.12	0.17	0.21
# Expected	8	11	19	12	11	6	12	9	25	22
# Predicted	9	12	22	12	11	9	12	9	27	22
#TP	8	11	19	9	8	6	11	8	25	20
#FP	1	1	3	3	3	3	1	1	2	2
#FN	0	0	0	3	3	0	1	1	0	2
Status 1	8	11	13	3	2	1	1	1	5	3
Status 2	0	1	3	0	0	7	2	0	2	5
Status 3	1	0	6	9	9	1	9	8	20	14

Columns description: *p-value* of the relation between phenotypic and genotypic profiles; Risk of the relation calculated based on the retrieved subjects; #expected and #Predicted stand for the number of expected and predicted subjects respectively; #TP, #FP and #FN stand for True positive, False positive and False negative. Finally Status 1, Status 2 and Status 3 indicate the number of affected, relatives and controls subjects expected.

Remarkably, inspection of the learned cut-off points for phenotypic selected features revealed that these were consistent with normative values of normal performance on cognitive tasks, and with two standard deviations of normative data of temperament and character dimensions, motor measures and transcranial ultrasound. Figure 7.5 shows the correlation matrix of genotypic profiles to the phenotypic profiles used to select those relations with very low *p-value*.

The selected relations include between 75 and 825 SNPs. Some of the SNPs result of interest, mapping chromosome regions known implicated in schizophrenia (Table 7-4 and Table 7-5). Two candidate genes that were identified by the strategies are DCN1 and FRYL.

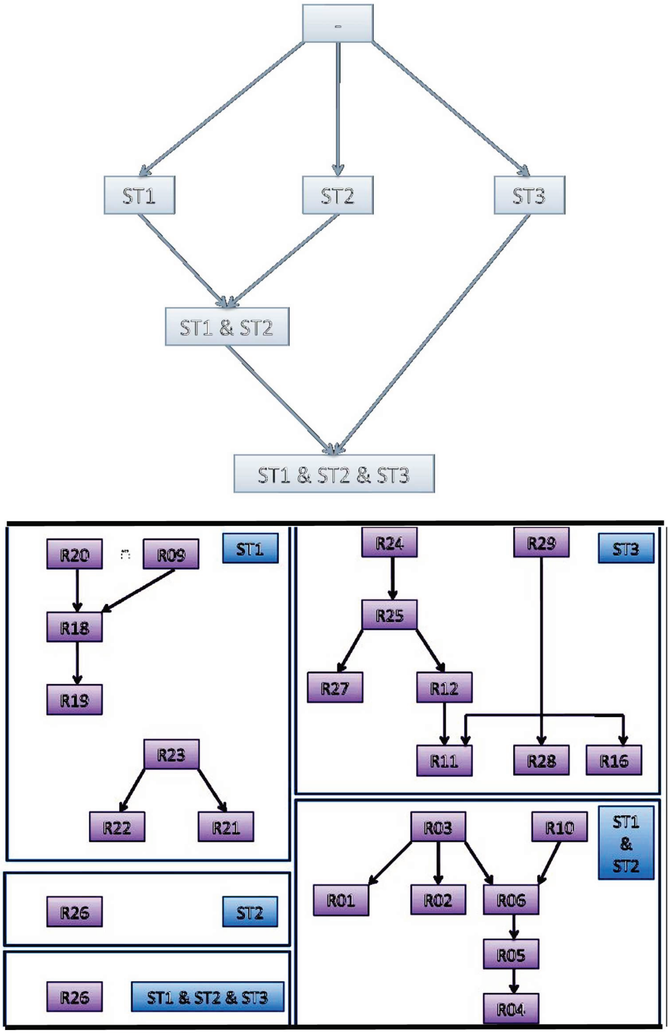
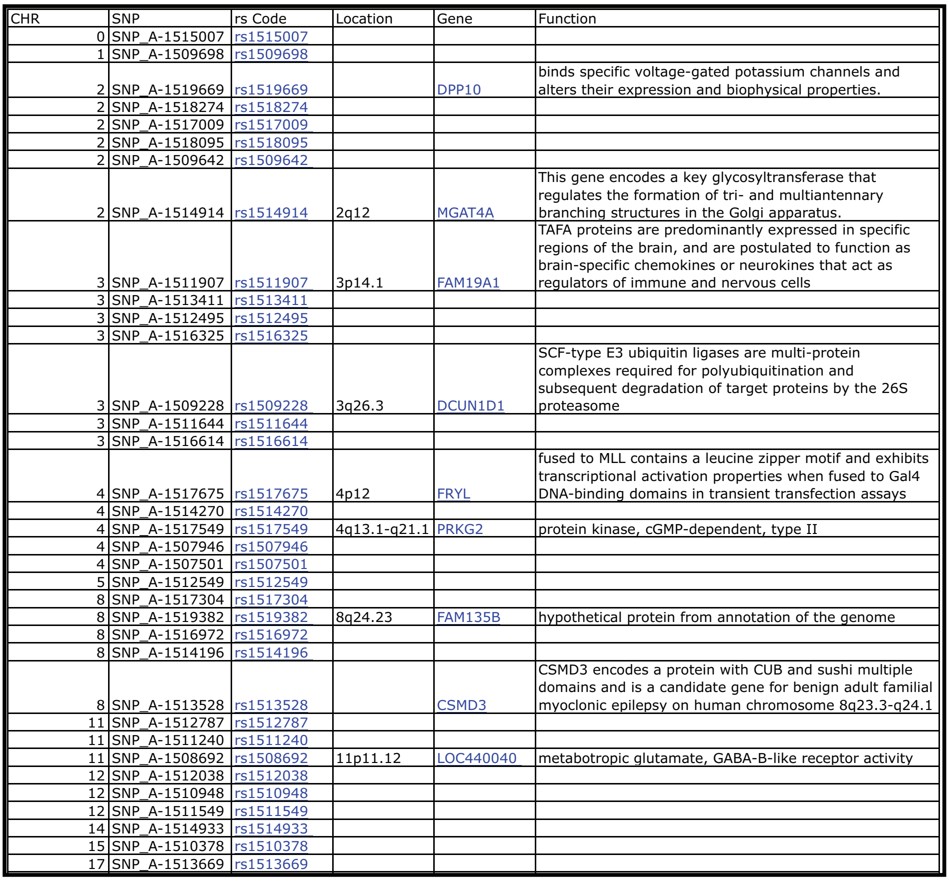


Figure 7.10 Hierarchical organization of the rules.
Decision trees describing the hierarchical structure of the classification. The top tree displays the top relationships found by the classifier. Status 1 corresponds to affected subjects, status 2 to first degree relatives, and status 3 to controls. The rules found assign subjects to categories including only affected subjects, only relatives, a combination of all subjects at risk (affected and relatives), controls, and a residual category with common features of the entire sample. The bottom panel displays the major hierarchical relations between rules for each classification group.

Table 7-4 Relevant SNPs analyzed for identified for the relations (First batch, pending further analysis).

p-values		Patients			Patients & Controls			
SNP		R09	R18	R21	R01	R04	R05	R10
SNP_A-1507501	AB	5.07E-04						1.73E-03
	BB		1.46E-03	8.73E-04				
SNP_A-1507946	AB	6.17E-05						1.58E-04
SNP_A-1508692	AA	3.35E-04		9.45E-03				7.81E-04
SNP_A-1509228	BB	2.82E-05	1.24E-03	8.69E-03	7.60E-05			5.60E-05
SNP_A-1509642	BB	1.02E-10						2.89E-10
SNP_A-1509698	AA	4.40E-06		5.72E-04				9.74E-06
	AB		9.33E-03					
SNP_A-1510378	BB	3.66E-03						2.40E-03
SNP_A-1510948	AA	2.98E-08		1.08E-04				7.62E-08
SNP_A-1511240	AA	2.23E-03						1.61E-03
SNP_A-1511549	AA	7.42E-04						2.83E-03
SNP_A-1511644	AA		1.71E-03					
SNP_A-1511907	AB	3.75E-04						8.28E-04
SNP_A-1512038	BB	2.74E-08	2.80E-04					5.95E-08
SNP_A-1512495	AA	1.00E-03						2.12E-03
SNP_A-1512549	AB			4.42E-03				
	BB	7.49E-03						
SNP_A-1512787	AA		5.65E-03					
	AB	4.18E-04						1.06E-03
SNP_A-1513411	AB	9.83E-05						2.56E-04
SNP_A-1513528	BB	2.57E-03						1.97E-03
SNP_A-1513669	AB			6.81E-03				
	BB	1.12E-05						2.63E-05
SNP_A-1514196	AA		2.73E-03					
	AB	9.38E-03						6.00E-03
	BB			1.08E-03				
SNP_A-1514270	AA	2.65E-07						6.94E-07
SNP_A-1514914	BB	5.91E-08	9.52E-04					1.38E-08
SNP_A-1514933	AA	6.31E-05						2.41E-04
	AB		7.46E-03					
SNP_A-1515007	AA	2.00E-06						7.33E-07
SNP_A-1516325	AB	1.66E-03						3.35E-03
SNP_A-1516614	AB			1.00E-03				
	BB	3.35E-04	4.70E-04					1.58E-04
SNP_A-1516972	AA	1.09E-04						3.59E-04
	AB			1.07E-04				
SNP_A-1517009	AA	4.43E-08						1.20E-07
SNP_A-1517304	AA	7.49E-03						5.42E-03
	AB			3.31E-03				
SNP_A-1517549	AB					5.23E-03	5.23E-03	
	BB			4.95E-03				
SNP_A-1517675	AA	4.43E-08						1.20E-07
	AB			4.59E-03				
SNP_A-1518095	AB	3.04E-05						1.41E-04
	BB		1.39E-04	1.98E-04				
SNP_A-1518274	AA			1.20E-03				
	AB	6.93E-03						
SNP_A-1519382	BB	2.67E-03		4.95E-03				1.36E-03
SNP_A-1519669	BB	4.13E-05						1.41E-05

Table 7-5 Description of SNPs listed in Table 7-4



CHR	SNP	rs Code	Location	Gene	Function
0	SNP_A-1515007	rs1515007			
1	SNP_A-1509698	rs1509698			
2	SNP_A-1519669	rs1519669		DPP10	binds specific voltage-gated potassium channels and alters their expression and biophysical properties.
2	SNP_A-1518274	rs1518274			
2	SNP_A-1517009	rs1517009			
2	SNP_A-1518095	rs1518095			
2	SNP_A-1509642	rs1509642			
2	SNP_A-1514914	rs1514914	2q12	MGAT4A	This gene encodes a key glycosyltransferase that regulates the formation of tri- and multiantennary branching structures in the Golgi apparatus.
3	SNP_A-1511907	rs1511907	3p14.1	FAM19A1	TAFA proteins are predominantly expressed in specific regions of the brain, and are postulated to function as brain-specific chemokines or neurokines that act as regulators of immune and nervous cells
3	SNP_A-1513411	rs1513411			
3	SNP_A-1512495	rs1512495			
3	SNP_A-1516325	rs1516325			
3	SNP_A-1509228	rs1509228	3q26.3	DCUN1D1	SCF-type E3 ubiquitin ligases are multi-protein complexes required for polyubiquitination and subsequent degradation of target proteins by the 26S proteasome
3	SNP_A-1511644	rs1511644			
3	SNP_A-1516614	rs1516614			
4	SNP_A-1517675	rs1517675	4p12	FRYL	fused to MLL contains a leucine zipper motif and exhibits transcriptional activation properties when fused to Gal4 DNA-binding domains in transient transfection assays
4	SNP_A-1514270	rs1514270			
4	SNP_A-1517549	rs1517549	4q13.1-q21.1	PRKG2	protein kinase, cGMP-dependent, type II
4	SNP_A-1507946	rs1507946			
4	SNP_A-1507501	rs1507501			
5	SNP_A-1512549	rs1512549			
8	SNP_A-1517304	rs1517304			
8	SNP_A-1519382	rs1519382	8q24.23	FAM135B	hypothetical protein from annotation of the genome
8	SNP_A-1516972	rs1516972			
8	SNP_A-1514196	rs1514196			
8	SNP_A-1513528	rs1513528		CSMD3	CSMD3 encodes a protein with CUB and sushi multiple domains and is a candidate gene for benign adult familial myoclonic epilepsy on human chromosome 8q23.3-q24.1
11	SNP_A-1512787	rs1512787			
11	SNP_A-1511240	rs1511240			
11	SNP_A-1508692	rs1508692	11p11.12	LOC440040	metabotropic glutamate, GABA-B-like receptor activity
12	SNP_A-1512038	rs1512038			
12	SNP_A-1510948	rs1510948			
12	SNP_A-1511549	rs1511549			
14	SNP_A-1514933	rs1514933			
15	SNP_A-1510378	rs1510378			
17	SNP_A-1513669	rs1513669			

7.4.3 Reaction surface of risk based on shared SNPs and the quantitative phenotype.

Our unsupervised approach of learning phenotypic/genotypic relations without making any reference to subject status produces groups of subject exhibiting distinct status distribution. Moreover, we incorporate the status into the relation by learning the relation risk. These relations support the construction of a surface risk of schizophrenia by plotting the learned risk of each selected relation (Figure 7.11).

The reaction surface of risk of schizophrenia in the sample is determined by shared SNPs and shared features of the quantitative endophenotype described by the relations. As it is easily apparent from the figure, shared SNPs and shared features of the hypodopaminergic state correctly predicted risk including intermediate clusters (such as clusters of combined relatives and subjects, or combinations of relatives and controls).

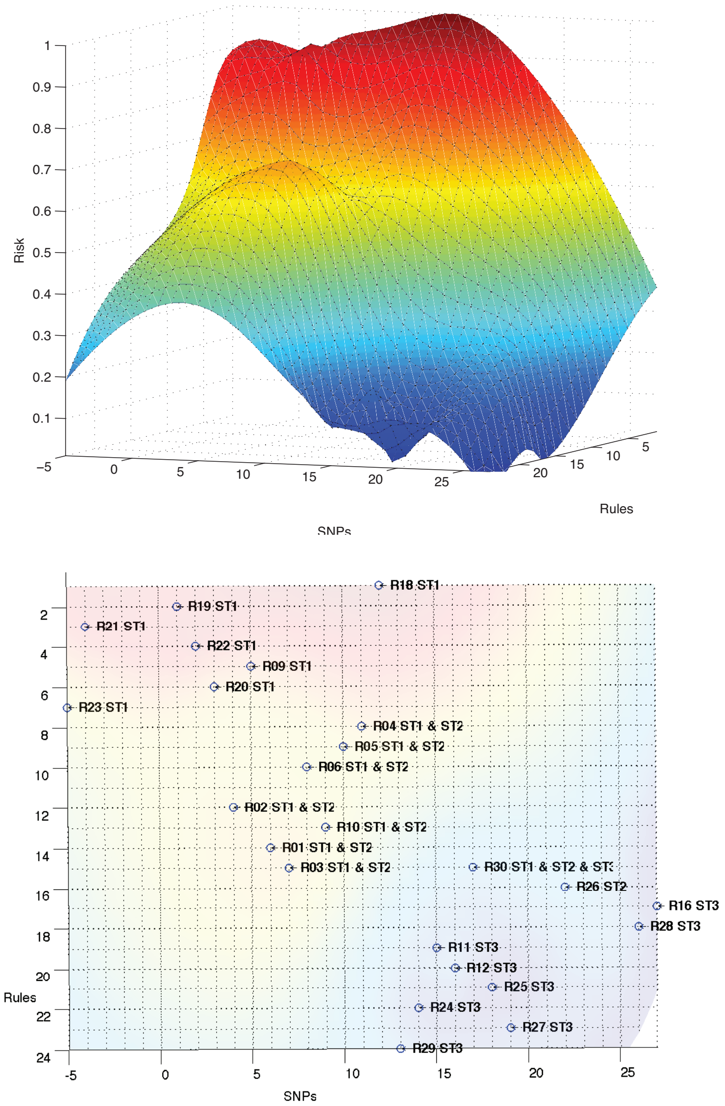


Figure 7.11 Surface of reaction of the risk of schizophrenia.
Weights are given by epidemiologic risk of schizophrenia (0.01 for controls, 0.1 for relatives, and 1 for affected subjects), given the dendrogram of distances for phenotypic rules (based on number of individuals) and the distance between “pseudohaplotypes” calculated as the degree of inclusion of SNP clusters within each other.

7.5 Methods

7.5.1 Decision trees

Decision tree is a predictive model; which allows mapping from observations about an item *S* to conclusions about its target value. This method builds structures where the leaves represent classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when splitting is either non-feasible, or a singular classification can be applied to each element of the derived subset. For example, the ID3 implementation uses a greedy search. It starts at the root of the tree and learns which attribute A should be use as test condition by evaluating every attribute using a statistical test called *information gain* (Mitchell 1997):

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(S) = \sum_{v \in Values(A)} -p_v \log_s(p_v)$$

where *values(A)* is the set of all possible values for attribute *A*, and S_v is the subset of *S* for which attribute *A* has value *v*. The second term of the equation is the sum of the entropies of each subset S_v weighted by its fraction of examples. *Gain(S, A)* is the expected reduction of entropy cause by knowing the value of attribute *A*.

7.5.2 Latent Classes

Latent class (LC) models are used for cluster analysis of categorical data. Underlying such a model is the assumption that the observed variables are mutually independent given the class variable.

LCs are unobservable (latent) subgroups or segments. Cases within the same latent class are homogeneous on certain criteria, while cases in different latent classes are dissimilar from each other in certain important ways.

A LC model involves a latent variable *X* and a number of manifest variables *Y1 , Y2 , . . . , Yn* . All the variables are categorical and the relationships among them are described by the simple Bayesian network shown in Figure 2.5. In applications, the latent variable *X* represents concepts. States of the latent variable correspond to classes of individuals in a population. The manifest variables *Yi* represent manifestations of the latent concept. Learning an LC model from data means to (1) determine the cardinality for variable X , i.e., the number of latent classes; and (2) estimate the model parameters *P(X)* and *P(Yi|X)*. Parameters are usually estimated using the Expectation Maximization algorithm (Dempster et al., 1977). The cardinality of *X* is determined by comparing alternatives using goodness-of-fit indices or scoring metrics. The most commonly used scoring metric is BIC (Schwarz, 1978). Equivalent to the Maximum Description Lenght score (Lanterman, 2001), the BIC score is an approximation of the marginal likelihood that is derived in a setting when all variables are observed (Zhang 2004).

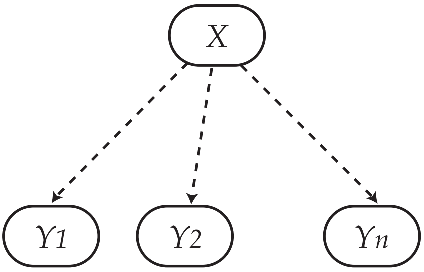


Figure 7.12 Structure of LC Model

7.5.3 Nonnegative matrix factorization

This method has been successfully employed to reveal information about latent relationships in experimental data sets (Mejia-Roa, Carmona-Saez et al. 2008). Nonnegative matrix factorization (NMF) is similar to other statistical data mining techniques, such as Principal components Analysis (PCA), in the sense that it constructs approximate factorization of the form V~WH, or $V_{i\mu} \sim (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$, where $V \in \Re^{mxn}$ is a positive matrix with m variables and n objects, and *k* the number of factors. The difference between NMF arises from the constrains imposed on the matrix *M* and *H*. While NMF constrains *V*, *W* and *H* to have nonnegative values, PCA constrains the columns of *W* to be orthogonal and the rows of *H* to be orthogonal to each other. MNF not only effective dimensionality reduction but also a more interpretable information (Mejia-Roa, Carmona-Saez et al. 2008) (Figure 2.6 shows a graphic representation of the model).

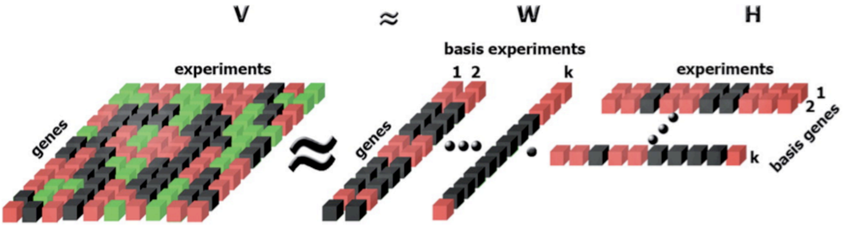
The implementation of NMF algorithm consist on iteratively updating *W* and *H* by:

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \qquad W_{ia} \leftarrow \frac{W_{ia}}{\sum_{j} W_{ja}} \qquad H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$$

It can be shown that iterations of these update converges to a local maximum of the objective function:

$$F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} \left[V_{i\mu} \log(WH)_{i\mu} - (WH)_{i\mu} \right]$$

subject to the nonnegative constrains (Lee and Seung 1999). This objective formula can be derived by interpretating NMF as algorithm for constructing a probabilistic model.



Mejia-Roa, E. et al. Nucl. Acids Res. 2008 36:W523-W528; doi:10.1093/nar/gkn335

Figure 7.13 NMF learns parts-based representation of objects.

7.6 Concluding remarks.

Schizophrenia is a highly heritable disorder but identification of individual genes has remained elusive (Abel 2004; Harrison and Weinberger 2005). Where specific genes or loci have been implicated in schizophrenia susceptibility, interpretation is far from simple because of overlap with bipolar affective disorder is more frequent than clinical and epidemiological data would predict. Environmental influences such as prenatal viral infections or obstetric complications also increase the risk, and non-genetic contributions must exist because up to one-third of monozygotic co-twins of schizophrenics are not affected with a schizophrenia-spectrum disorder .

Transmission of predisposing genes is not always expressed on the clinical phenotype, and risk of schizophrenia amongst the offspring of affected and unaffected monozygotic co-twins is identical. Differences in the psychiatric symptoms of affected members within families are not useful in defining particular sets of predisposing genes. Thus, it is essential to describe phenotypes that may be useful to fill the gap between genes and the disease processes. Heritability of a trait is usually suggested by twin studies showing higher concordance of disease among monozygotic than dizygotic twins, or as in the present study by case-control designs showing higher risk among relatives of patients with disease than among the general population. No exhaustive evaluation of the heritability of parkinsonism in schizophrenia has been published, but a recent study showed statistically significant heritability estimates for abnormalities of rapid alternating movements and manual sequences, both of which are evaluated in the present study.

On the basis of our experimental results we proposed a specific vulnerability of dopaminergic neurons in the mesocortical projections during development as a substrate for gene-environment interactions (de Erausquin 2004). At least two known functional polymorphisms in candidate genes for schizophrenia could increase vulnerability of dopaminergic neurons to known environmental risk factors: (i) a functional polymorphism in the orphan receptor gene NURR1 increases susceptibility of dopaminergic neurons to mitochondrial toxins, and (ii) expression of a mutant form of the calcium-sensitive potassium channel SK3 identified in schizophrenia should also increase susceptibility of dopaminergic neurons to excitotoxicity (de Erausquin 2004). If a similar mechanism is operational in at least a fraction of the patients with schizophrenia, an endophenotype measuring deficits in dopaminergic function would be useful in segregating such patients from syndromes with other pathophysiological origin.

Our data show that left sided motor impairment and hyperchogenicity of the right substantia nigra are significantly more common in subjects with chronic untreated schizophrenia than in their first degree relatives or in controls, and significantly more prominent in the relatives than in controls. Echogenicity in the substantia nigra is increased in patients with Parkinson's disease (Berg, Siefker et al. 2001), and correlates inversely with uptake of the dopamine precursor fluoroDOPA in striatum, and directly with severity of motor involvement in patients (Berg, Siefker et al. 2001), in unaffected relatives, and in undiagnosed elderly subjects with parkinsonian motor abnormalities. Hyperechogenicity is exceedingly rare in normal healthy volunteers, but when present it predicts decreased fluoroDOPA uptake. In schizophrenia, hyperechogenicity of the substantia nigra correlates with decreased striatal fluoroDOPA uptake and predicts susceptibility to parkinsonian symptoms induced by antipsychotic drugs in newly diagnosed, untreated subjects (Berg, Jabs et al. 2001); in chronically treated patients, hyperechogenicity correlates with age and severity of parkinsonian symptoms but not with dose or type of medication.

Certain personality traits also appear to reflect the state of dopamine transmission. Patients with Parkinson's disease have lower novelty seeking and higher harm avoidance than matched controls (Kaasinen, Nurmi et al. 2001); novelty seeking was not associated with disease duration, current motor symptoms, or medication, whereas harm avoidance was significantly correlated with severity of bradykinesia and depression. These findings have been replicated in transcultural settings (Kaasinen, Nurmi et al. 2001). Low novelty seeking correlated with right sided motor impairment and with low uptake of fluoroDOPA in the left caudate in one study, but not in another (Kaasinen, Nurmi et al. 2001). Conversely, increased harm avoidance correlates with left sided motor impairment and with low fluoroDOPA uptake in the right caudate nucleus (Kaasinen, Nurmi et al. 2001). Consistent with the hypothesis of a dopaminergic deficit, subjects with schizophrenia in our sample had significantly higher harm avoidance than controls and relatives; the latter two were indistinguishable. We found no differences in novelty seeking scores, but reward dependence was significantly lower in schizophrenia, and relatives had also significantly lower scores than controls. TCI is indicator of heritable personality features, strengthening its use in this family study. A novel finding of this study is the significantly lower scores of self-directedness in schizophrenia and in first degree relatives. The results of the discriminant analysis indicate that both low self-directedness and high UPDRS scores are correlated to the same underlying biological function. This is particularly noteworthy because both parkinsonism and self-directedness (Craddock, O'Donovan et al. 2005) have been shown to predict quality of life in schizophrenia.

As reported by others (Goldberg, Egan et al. 2003), cognitive performance was significantly different between the three groups. However, the contribution of cognitive dysfunction to correct classification of subjects with schizophrenia and their relatives was small compared to parkinsonism and temperament traits. This may be explained by the intrinsic variability of cognitive performance in the general population, and the relative complexity of the underlying neurobiological processes. Alternatively, parkinsonism may reflect a core feature of the pathophysiologic process. Indeed, our data conclusively demonstrate that most of the variance between diagnostic groups can be explained by a single discriminant function correlated with motor impairment, temperament traits, and a smaller (but significant) contribution from right substantia nigra hyperechogenicity. Furthermore, and more strikingly, unbiased classification based on latent variables, and without any reference to psychopathology, correctly groups in a single cluster most cases with schizophrenia, most normal controls in a second cluster, and leaves a third cluster composed with most at-risk relatives, a few cases of schizophrenia and a few controls. These data make a most compelling case in favor of considering dopaminergic deficits (including prominently parkinsonian motor impairment) a core feature of schizophrenia.

Unsupervised and supervised machine learning methods are used to generate predictive models and systems of classification; that is, a mapping from observations about an item to conclusions about its target value or its class pertenence. We used this approach to generate relations of phenotypic and genotypic features defining subgroups with the highest likelihood in our sample, and found that this unbiased classification system used the core variables in the endophenotype of dopaminergic deficit to define the expected risk groups (as latent classes), including an intermediate class containing both affected subjects and their first degree relatives (Figure 7.10). The informative content of this rules will not be discussed in detailed here, but suffice it to say that they had high face validity; indeed, the cut-off values for individual variables learned by the algorithm were nearly identical to normative values for all of the cognitive tests, temperament and character dimensions, movement disorder scales and ultrasound imaging (not shown).

When the phenotypic profiles were intersected to the genotypic profiles independently obtained by cluster analysis we found associations with a very high probability between them, and most strikingly, the association was predictive of risk status (Figure 7.11), without making any reference to subject status. This is, to our knowledge, the first demonstration of an endophenotype driven by polygenic risk with multiple small effects successfully used to predict the reaction surface or risk.

As previously mentioned, the literature has been recently divided between the more prevalent model of a common disease and multiple common genetic variations with small effects and a newly proposed model of a common disease and multiple rare variations with large effects (McClellan, Susser et al. 2007). Two recent publications have shown that at least a fraction of the cases of schizophrenia are likely to be explained by very rare microdeletions or duplications with large effect sizes. The two models are not mutually exclusive, however, and our data suggest that the polygenic, common variation model may account for a significant amount of the genetic variance at least in our sample.

We found SNPs associated with the quantitative measure of the endophenotype with a $p < 0.005$; all of the SNPs map to chromosomal locations already implicated (or close to regions implicated) in schizophrenia risk by genome wide scans. These SNPs can be mapped to known genes suggested highly functionally pertinent candidates, including mechanisms (such as modulation of the *wnt* transcription pathway or regulation of neuronal potassium channels function) that have already been repeatedly implicated in the genetic basis of schizophrenia. Also of interest is the suggestion that risk may be associated with variation in the glycosyltransferase gene that has been found to increase the risk of type 2 (adult onset) diabetes mellitus, because there is increasing evidence that subjects with schizophrenia have increased risk of metabolic syndrome.

The major weakness of our current data is the small sample size. Replication of our findings in a second independent sample of the same population is currently underway, and should provide insight into the degree of rubustness of the present findings.

Chapter 8
Comentarios finales

Dedicaremos esta sección a resumir brevemente los resultados obtenidos y a destacar las conclusiones que esta memoria puede aportar. Además, plantearemos algunos aspectos sobre trabajos futuros que siguen la línea aquí expuesta.

8.1 Resumen y conclusiones

En este trabajo hemos propuesto una metodología que permite abordar diferentes problemas en la identificación de relaciones genotipo/fenotipo en procariotas y eucariotas. La nuestra es una metodología que integra distintas fuentes de conocimiento y extrae hipótesis cualitativas e interpretables que echan luz sobre los mecanismos de regulación transcripcional en organismos procariotas y relaciones genotipo/fenotipo en enfermedades hereditarias en eucariotas.. Asimismo hemos provisto de un marco computacional para la fusión de información proveniente de bases de datos biológicas que hemos empleado para obtener resultados validados experimentales.

Nuestro trabajo se ha enmarcado bajo la perspectiva de obtener una metodología robusta y a la vez, flexible y adaptable a diferentes problemas biológicos. El eje central de la propuesta se basa en la representación de la información mediante una familia de modelos, la agregación de las diferentes fuentes de información para la exploración del espacio de las hipótesis; la utilización de la optimización multiobjetivo para la selección de las soluciones optimales; y la representación multimodal para la generación de hipótesis que describan al problema desde distinto punto de vista.

En concreto, hemos considerado el uso de conjuntos difusos y lógica difusa como marco general de representación y agregación de la información; clustering conceptual para el aprendizaje de perfiles que agrupan las observaciones mediante las características mas representativas; algoritmos evolutivos como técnica de optimización; y clasificadores basados en la similaridad de una observación nueva a un conjunto de prototipos previamente aprendidos.

Algunos de los resultados que hemos obtenido mediante la aplicación de la metodología a lo largo de esta tesis son:

- Presentamos un método basado la técnica "Divide y Vencerás" para desensamblar un motivo de sitios de unión de un regulador al ADN en una familia de modelos. Hemos demostrado que este enfoque mejora el reconocimiento de sitios de uniones al ADN funcionales. Además de las ventajas computacionales que el método ofrece, hemos demostrado que esta familia de modelos permite estudiar la evolución de los genes regulados por PhoP en las gamma enterobacterias; y que describen aspectos físicos de la interacción proteína-DNA.

- Hemos demostrado como genes co-regulados e inmersos en una misma arquitectura de regulación (*network morif*) pueden tener una expresión diferencial y que esto se debe tanto a la variabilidad en las secuencias de unión así como otras elementos *cis*: ubicación del sitio de unión respecto a la RNA polimerasa, hebra del AND, y distancia al sitio de unión de otros factores de transcripción.

- Hemos realizado predicciones y validado experimentalmente que el sistema de dos compomemntes PhoP/PhoQ utiliza múltiples mecanismos para controlar la transcripción de los genes regulados. Asimismo hemos detectado genes directamente regulados por PhoP cuyos promotores difieren del canónico, resultando PhoP el elemento central de una red de regulación altamente conectada.

- Hemos realizado simulaciones de la interconexión de las redes de regulación genéticas de PhoP/PhoQ-PmrA/PmrB que predicen que tanto mecanismos transcripcionales como post-transcripcionales son empleados para conectar los dos sistemas de dos componentes.

- Hemos identificado conjuntos de individuos con distinto riesgo de padecer esquizofrenia, caracterizándolos por la relación genotipo/fenotipo. El valor cualitativo de las variables clínicas ha sido validado por los expertos y 40 de los SNPs que dan soporte a los estos grupos se corresponden con áreas del cromosoma implicadas en la esquizofrenia

8.2 Trabajos futuros

A continuación, mencionamos algunas extensiones sobre los problemas los problemas biológicos y metodológicos que hemos tratado en esta memoria:

- Un factor clave para el entendimiento de la regulación de la expresión de los genes es poder analizar la evolución de un regulador y los genes por él regulados en especies relacionadas (Wray, Hahn et al. 2003). En tal sentido, nosotros hemos provisto un modelo de evolución de PhoP en las gamma enterobacterias y utilizado los submotivos de *E. Coli* y *Salmonella* para el análisis de los genomas de éstas especies mediante la predicción de sitios de union de PhoP al ADN. Así como hemos contrastado nuestras predicciones mediante experimentos de *microarray* y ChIP para *Yersinia Perstis KIM*, resulta esclarecedor hacerlo para otras especies. En tal sentido, hemos de disponer en un futuro próximo inmediato resultados de *microarray* y ChIP para *Shigella Flexneri*, *Erwinia Carotovara atroseptica* y *Escherichia Coli*. El validación experimental de nuestras predicciones en nuevos genomas abre la posibilidad de extender la metodología propuesta:
    - La incorporación de técnicas de aprendizaje incremental permitirá extender, refinar y adaptar los modelos e hipótesis a las nuevas especies en estudio.
    - Desarrollar un modelo de evolución basado en técnicas de *clustering poblacional* (de la misma manera en que son empleadas en ciencias sociales). Esta técnica nos permite estudiar como y con que política evolucionan los cluster, a diferencia de las técnicas hasta ahora empleadas por nosotros, que nos permite analizar que individuos pertenecen a cada cluster.
- Un componente clave de la regulación de la expresión es la afinidad con que una proteína se une al ADN (Ptashne and Gann 2002). Dada la evidencia que se acumula respecto al papel que juegan las propiedades estructurales del ADN en diversos procesos biológicos (Baldi, Chauvin et al. 1998), y debido a que la estructura del ADN es un factor influyente en la unión proteína-DNA proponemos el estudio de esta interacción mediante las propiedades estructurales. El objetivo que planteamos es poder predecir *in silico* la afinidad de un regulador al ADN. Para ello deberemos:
    - Desarrollar las técnicas que nos permitan aprender un modelo que recupere las propiedades físicas determinante para predecir la afinidad de un regulador al ADN. El ADN presenta diferentes propiedades a ser analizadas para determinar cuales de ellas pueden ser determinantes de la afinidad de un regulador al sitio de unión. Planeamos la recuperación de modelos con el menor numero de parámetros posible y que prediga la afinidad utilizando el menor número de propiedades estructurales. Para ello, planeamos estudiar el contenido de información (i.e. *Information content*) de cada característica así como la cantidad de información que una característica arroja sobre otra (i.e. *Mutual information*) que guié el proceso de agregación de información. Ésto posibilitará la construcción de predictores basados en dife-

rentes métodos, como *Support Vector Machines*, o árboles de decisión.

- La dinámica de la expresión permite entender cuando y con que intensidad un gen se expresa o no, y es clave para el estudio de la relación genotipo/fenotipo (Wray, Hahn et al. 2003). Es por este motivo que creemos importante continuar su estudio, analizando la relación de los resultados experimentales *in vivo* y *in vitro*, lo cual requerirá:
  - Desarrollar modelos que permitan una comparación de los datos arrojados mediante *Gel shift assays* (EMSA) realizados *in vivo* y *in vitro*.
  - Desarrollar modelos de la dinámica de la expresión de los genes basados en las características *cis*.
- Para el análisis de las redes de regulación genética en procariotas hemos estudiado con particular atención el sistema de dos componentes PhoP/PhoQ. Los sistemas de dos componentes controlan un importante numero de funciones celulares, constituyendo el mecanismo principal de transducción de señales, que permite a la bacteria modificar su comportamiento celular en respuesta a estímulos ambientales (Alm, Huang et al. 2006). [Consideramos importante extender nuestro enfoque, aplicándolo a otros sistemas (e.g. PmrA/PmrB). Para abordar el estudio en una dimensión mayor de datos consieramos:
  - Algunos de los métodos de aprendizaje empleados en esta memoria son altamente demandante de tiempo y/o recursos. Eventualmente, aplicando heurísticas sería posible reducir los requerimientos de hardware y los tiempos de procesamiento. Por ejemplo, el método GPS hace una búsqueda exhaustiva en el espacio de las variables, el cual puede ser reemplazada por alguna heurística (i.e. algoritmos genéticos).
- Nuestro estudio de la esquizofrenia nos ha permitido encontrar relaciones genotipo/fenotipo asociadas a un factor de riesgo para los individuos recuperados. Las técnicas de aprendizaje utilizadas no nos han permitido validar nuestros resultados mediante medidas estadísticas típicas (i.e. *Leave one out*). Debido a que las observaciones disponibles al comenzar el estudio eran limitadas (72 individuos), no fue posible reservar una muestra para la validación de nuestros hallazgos. Al momento de la redacción de este último capítulo de la memoria, se nos informó de la incorporación de un nuevo grupo de muestras al estudio, lo cual nos permitirá validar o refutar las relaciones aprendidas. Asimismo nos proponemos:
  - Construir un predictor de fenotipo, en base al genotipo utilizando las relaciones aprendidas.
  - Construir un predictor del genotipo en base al genotipo, utilizando estas mismas relaciones.

8.3 Publicaciones derivadas de esta tesis

Para concluir, mencionamos que la mayor parte de este trabajo ha sido publicado o enviado para su revisión. A continuación los enumeramos:

International Journals

Oscar Harari, Henry Huang, Igor Zwir, "Delimiting plasticity of transcription factor binding sites by disassembling DNA consensus sequences". *PLoS Computational Biology*. Submited 2008

Igor Zwir, Oscar Harari, and Eduardo A. Groisman. "Gene promoter scan (GPS) methodlogy for identifying and classifying co-regulated promoters". *Methods in Enzymology - Two-Component Signaling Systems, Part A* Vol. 422. 2007. Academic Press – Elsevier Inc.

Oscar Harari, Igor Zwir. "Dissecting network motifs by identifying promoter features that govern differential gene expression". Simulation: Transactions of The Society for Modeling and Simulation Internation. Submited 2008.

Christopher Previti, Oscar Harari, Igor Zwir, Coral del Val. "Profile analysis and prediction of tissue-specific methylation classes". Genome Biology. Submited 2008.

Book Chapters

Oscar Harari, Cristina Rubio Escudero, Patricio Traverso, Marcelo Santos and Igor Zwir. "Learning robust dynamic networks in prokaryotes by gene expression networks iterative explorer (GENIE)". 2007. Nature Inspired Cooperative Strategies for Optimization (NICSO 2007). Springer. London, UK.

Oscar Harari, Igor Zwir. "A hybrid promoter analysis methodology for prokaryotic genomes". 2008. Fuzzy Systems in Bioinformatics, Bioengineering and Computational Biology. To be published as Springer's Series on Studies in Computational Intelligence

Lectures Notes

Rocio Romero Zaliz, Oscar Harari, Cristina Rubio Escudero, and Igor Zwir. "Identifying the promoter features governing differential kinetics of co-regulated genes using fuzzy expressions". Proceedings of IEEE International Conference on Fuzzy Systems 2007. pp 1167-1173. London, UK.

Christopher Previti, Oscar Harari, and Coral del Vall. "Mining and Predicting CpG islands". Proceedings of IEEE International Conference on Fuzzy Systems 2007. pp 1216-1221. London, UK.

Oscar Harari, Igor Zwir. “Dissecting network motifs by identifying promoter features that govern differential gene expression”. Proceedings of Summer Computer Simulation Conference 2007 (SCSC’07). pp 817-826. San Diego, California, USA.

Oscar Harari, Cristina Rubio-Escudero, Igor Zwir. “Targeting differentially co-regulated genes by multiobjective and multimodal optimization”. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics - Lecture Notes in Computer Science.* Vol. 4447 – pp. 68-77. 2007. Springer – Verlag Berlin, Heidelberger.

Cristina Rubio-Escudero, Oscar Harari , Oscar Cordón, Igor Zwir. “Modeling genetic networks: comparison of static and dynamic models”. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics - Lecture Notes in Computer Science.* Vol.4447 –pp. 78-89. 2007. Springer – Verlag Berlin, Heidelberger

Oscar Harari, Coral del Val, and Igor Zwir. “Fusing genetic knowledge by dynamic learning regulatory profiles: visualizing the strategy”, Current Research in Information Sciences and Technologies Multidisciplinary approaches to global information systems, Vol. 1 – 47-51. 2006. Open Institute of Knowledge.

Cristina Rubio-Escudero, Coral del Val, Oscar Cordón, Oscar Harari, Igor Zwir. “Decision making association rules for recognition of differential gene expression profiles”. *Intelligent Data Engineering and Automated Learning - Lecture Notes in Computer Science.* Vol.4224 – pp. 1137-1149. 2006. Springer – Verlag Berlin, Heidelberger

Oscar Harari, Rocio Romero-Záliz, Cristina Rubio-Escudero, Igor Zwir. “Fusion of domain knowledge for dynamic learning in transcriptional networks”. *Intelligent Data Engineering and Automated Learning - Lecture Notes in Computer Science.* Vol.4424 –pp.1067-1078. 2006. Springer – Verlag Berlin, Heidelberger

R. Romero-Zaliz, C. Rubio-Escudero, O. Cordón, O. Harari, C. del Val, I. Zwir. “Mining structural databases: an evolutionary multi-objetive conceptual clustering methodology”. *Applications of Evolutionary Computing - Lecture Notes in Computer Science.* Vol.3907 – pp. 159-171. 2006. Springer – Verlag Berlin, Heidelberger.

C. Rubio-Escudero, R. Romero-Zaliz, O. Cordon, O. Harari, C. del Val, and I. Zwir, " Optimal selection of microarray analysis methods using a conceptual clustering algorithm ". *Applications of Evolutionary Computing, Lecture Notes in Computer Science.* Vol. 3907 - pp. 172-183. 2006. Springer-Verlag Berlin, Heidelberger

Appendix A

Additional Figures and tables for Chapter 3 Delimiting plasticity of transcription factor binding sites by disassembling DNA consensus sequences

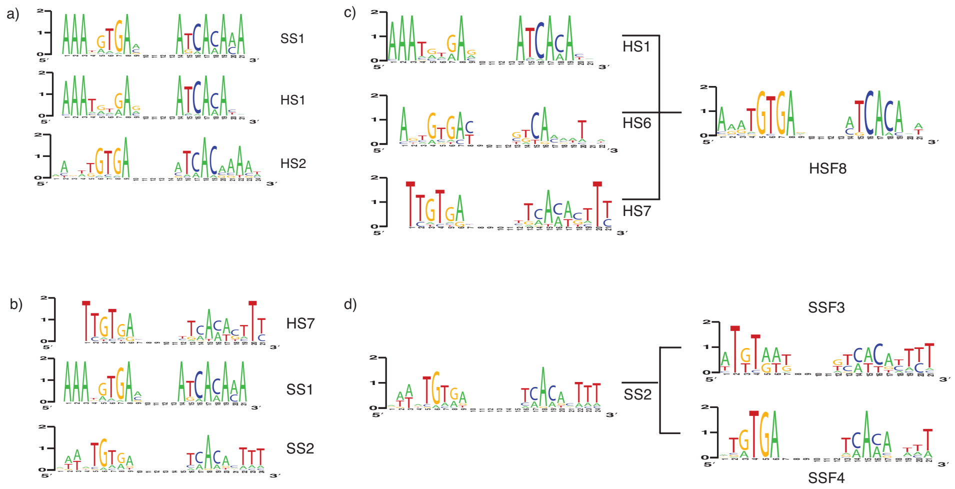


Figure 9.1 CRP Submotifs

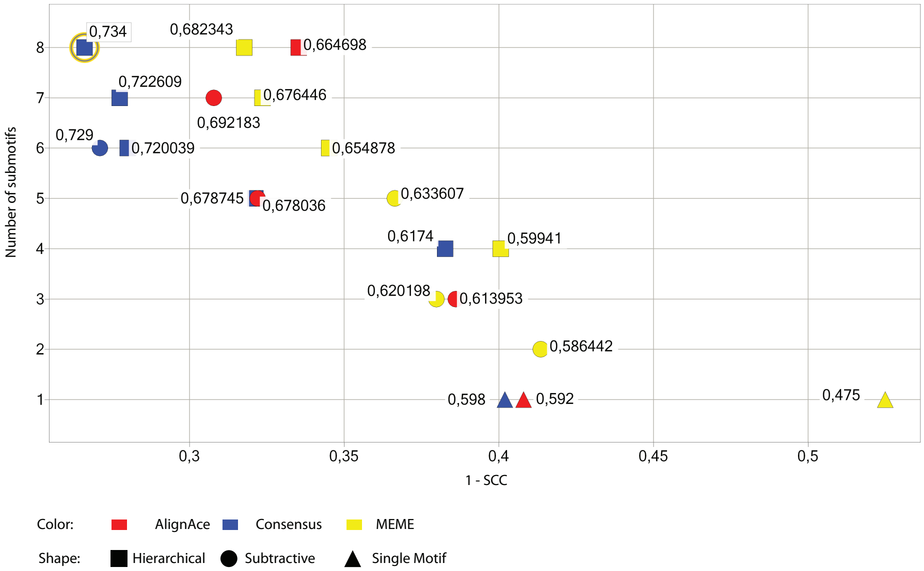


Figure 9.2 ″. Set of optimal solutions for CRP submotifs
PWM that encode the submotifs are indicated by color (AlignACE red; MEME yellow; and Consensus red); clustering method by shape (Hierarchical square; Subtractive circle; and Single motif triangule); and each configuration label reflects the SCC obtained.

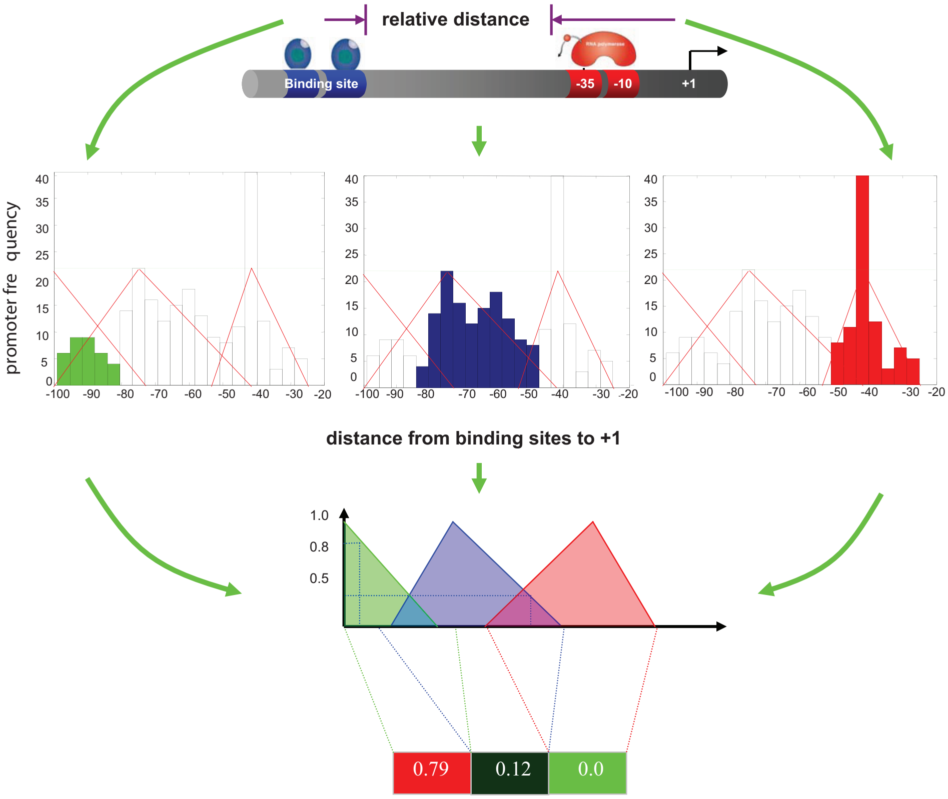


Figure 9.3 Learning and prototyping the relationships between the proximity of a transcription factor binding site and the RNA polymerase site

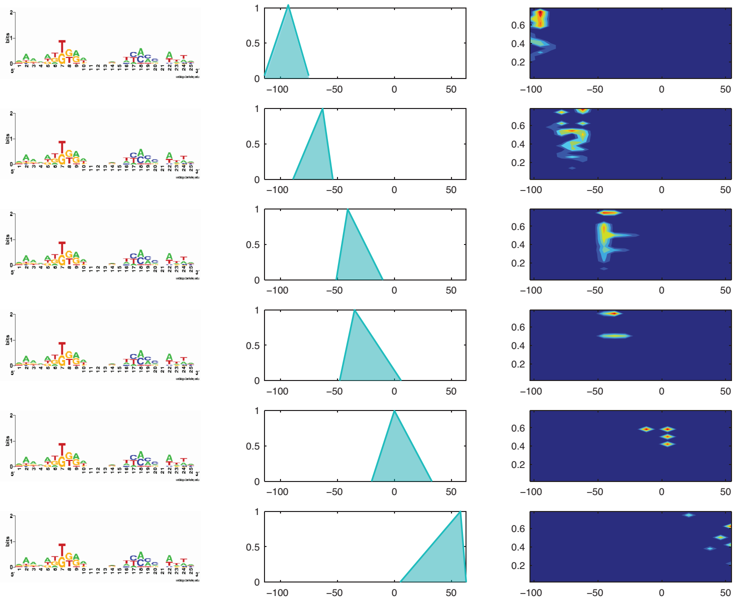


Figure 9.4 IF-THEN rules encompassing CRP single motif and distances between its BS and RNAP.

Table 9-1 PhoP regulated promoters raw data

Species	Gene	Sequence	Dist. to +1
E. Coli	b1826	ACATAGTTAGGCGCTGTTTAACT	33
E. Coli	b2833	AGATATATAACGTCGGTTTATAA	30
E. Coli	crcA	GTCTATTAAGGTTATGTTAATTG	32
E. Coli	hdeA	TTCTGTATATGTCATGTTGATGG	32
E. Coli	hemL	TGATGTTTGACGAGTATTTAACT	31
E. Coli	mgtA	TCTGGTTTATCGTTGGTTTAGTT	34
E. Coli	nagA	ATCTGTTTATGGGCGGTGTAGGT	31
E. Coli	ompT	CACTGTTTATATTTTGTTTAGTA	59
E. Coli	ompT	ACATATTGCTCCACTGTTTATAT	48
E. Coli	ompX	GCTGGTTGAGCATTTGTTGAAAA	33
E. Coli	phoP	GCTGGTTTATTTAATGTTTACCC	33
E. Coli	pmrD	TGCCGTTGATAAAGAGTTTATCT	32
E. Coli	proP	TGAAGTTGATCACAAATTTAAAC	28
E. Coli	proP	TTTCGTTTAGGACTCATTGATGT	59
E. Coli	rstA	ACTTGTTTAGAAACGATTGATAG	44
E. Coli	slyB	TTTTGTTTATAATTGGTTGATCC	33
E. Coli	ybcU	CATTGTTTAGGGTTTGTTTAATT	33
E. Coli	ybjG	CTTTCTTTAAGTTTTATTTAACC	32
E. Coli	ybjX	CACTATTGATGTTTGGTTAAGAT	32
E. Coli	ybjX	TGATATTTCGTTGAAGTTAATGA	79
E. Coli	yhiW	CAGCGTATAGCTTATGTTTATAA	31
E. Coli	yiaG	ATTTGTTGTTTCATTGTTAAAAA	28
E. Coli	yrbL	CATTGTTTAGGTTTTGTTTAAGT	28
Salmonella	heml	AAATGCGTAAAACTTTCATAACC	31
Salmonella	irap	TGCCGTTACGATATGGTTTAAAT	39
Salmonella	mgtA	TCTGGTTTATCGTTGGTTTAATT	34
Salmonella	mgtC	TTCTGTTTAAGTTTGTTTGATAT	68
Salmonella	mgtC	GTTTAGTGACGTTCTGTTTAAGT	56
Salmonella	mig-14	AAATGTTTAGCTTGTATTTAATG	58
Salmonella	naga	ATCTGTTTATGGGCGGCGTCGGC	30
Salmonella	ompX	GGCGGTTGAGGGTTCGTTGAAAA	43
Salmonella	orgb	ATTTATTGAGGAGGCATTGAAGC	33
Salmonella	pagC	CTGTGTTTAGAGAGAATTTACAT	68
Salmonella	pagK	AACCATTTATAAAATATTTAACT	58
Salmonella	pagP	CTCTGTTTATAGTTTGTTAAGAT	44
Salmonella	pdgL	TTATTTTAACCATCTGTTTAAGC	31
Salmonella	phoP	TCTGGTTTATTAACTGTTTATCC	34
Salmonella	pipD	CTTTATTGAGGTTGTATTGATAA	77
Salmonella	pmrD	CGCTATTGCCGTTTTGTTTATCC	32
Salmonella	proP	ACATATTTAAACCCTGTTAGGGT	69
Salmonella	rstA	TCTCGTTTAGAAAAGATTTATGG	43
Salmonella	slyB	CTTCGTTTAAGATTGGTTAATTA	33
Salmonella	udg	AAATGTTTAAGCCCGGTTTAATA	93
Salmonella	ugtL	CACGGTTGAGCAACTATTTACTT	52
Salmonella	virK	CTTCGTTGCCTTTACGTTTAACT	34
Salmonella	virK	CGATGTTGTTAAACAGTTTATCA	71
Salmonella	virK	CGCCATTGATAAACTGTTTAACA	77
Salmonella	ybjX	CTGTATTGACGATTGGTTAATGT	32
Salmonella	ybjX	GTTTGTTTAGATACGGTTTACTT	79
Salmonella	yobG	CTACAGTTACTCCTGGTTTAAGT	32
Salmonella	yrbL	TTTCGTTTAGGTTTTGTTTAAGT	28

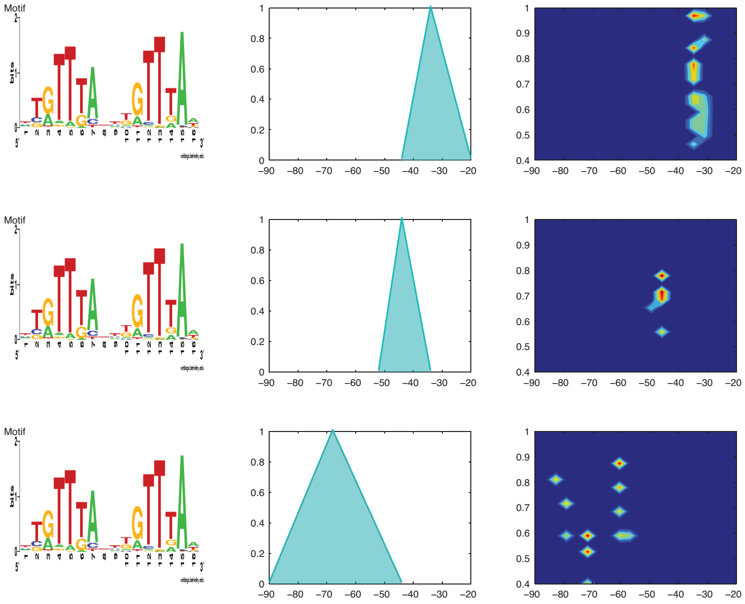


Figure 9.5 IF-THEN rules encompassing PhoP single motif and distances between BS and RNAP evaluated for genes activated by PhoP.

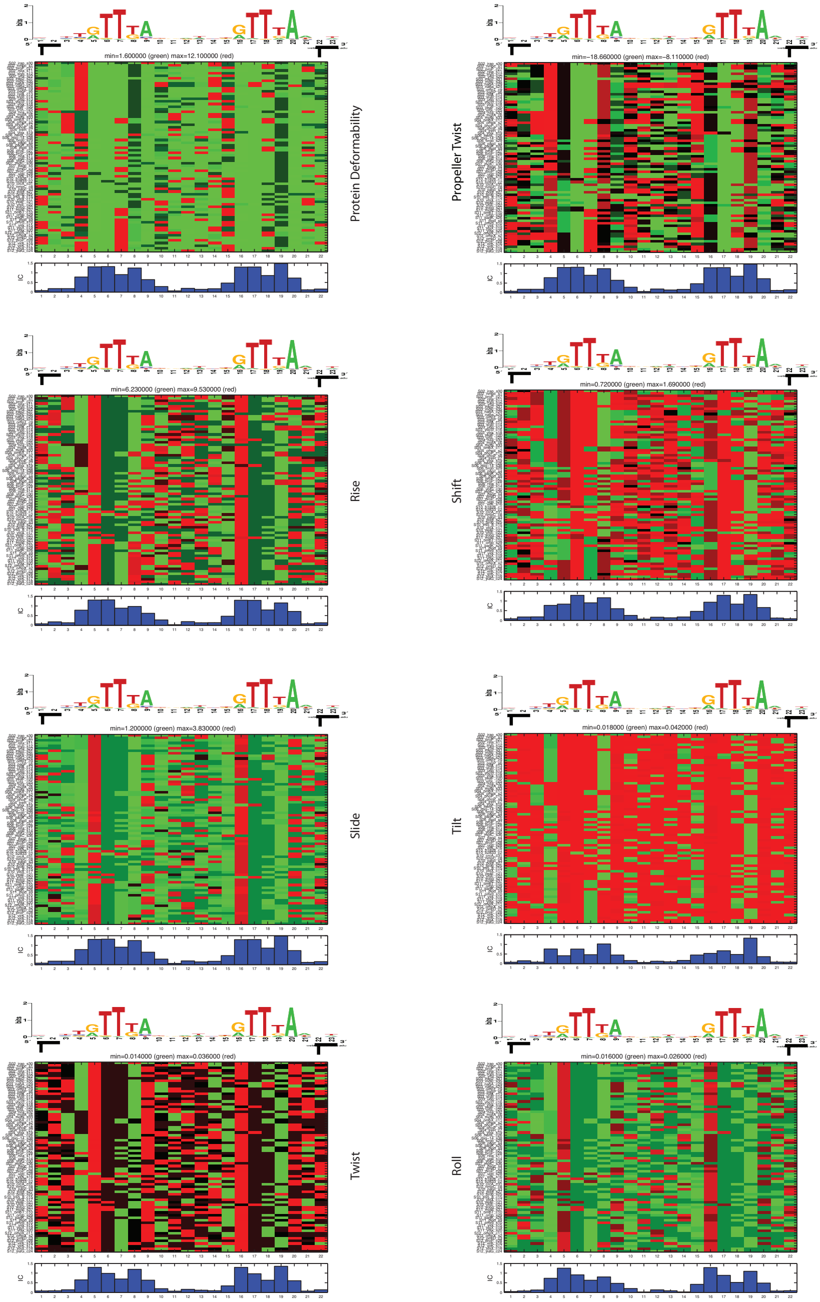


Figure 9.6 Physical properties for PhoP BS

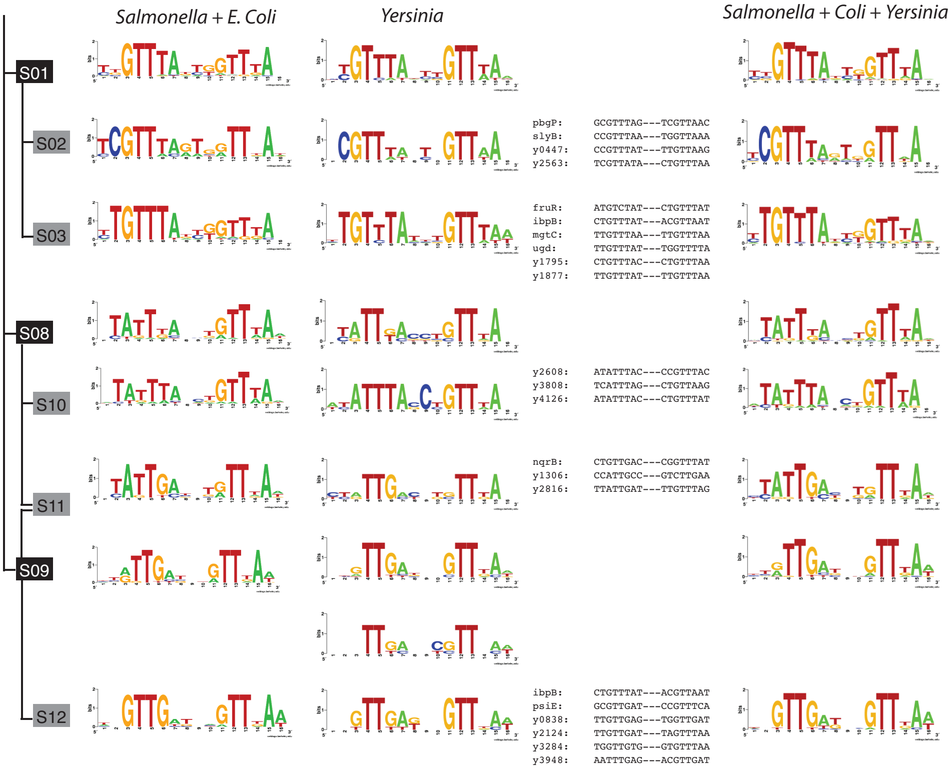

Salmonella + E. Coli
Yersinia
Salmonella + Coli + Yersinia
S01
S02
S03
S08
S10
S11
S09
S12
pbgP: GCGTTTAG---TCGTTAAC
slyB: CCGTTTAA---TGGTTAAA
y0447: CCGTTTAT---TTGTTAAG
y2563: TCGTTATA---CTGTTTAA
fruR: ATGTCTAT---CTGTTTAT
ibpB: CTGTTTAT---ACGTTAAT
mgtC: TTGTTTAA---TTGTTTAA
ugd: TTGTTTAT---TGGTTTTA
y1795: CTGTTTAC---CTGTTTAA
y1877: TTGTTTAT---TTGTTTAA
y2608: ATATTTAC---CCGTTTAC
y3808: TCATTTAG---CTGTTAAG
y4126: ATATTTAC---CTGTTTAT
nqrB: CTGTTGAC---CGGTTTAT
y1306: CCATTGCC---GTCTTGAA
y2816: TTATTGAT---TTGTTTAG
ibpB: CTGTTTAT---ACGTTAAT
psiE: GCGTTGAT---CCGTTTCA
y0838: TTGTTGAG---TGGTTGAT
y2124: TTGTTGAT---TAGTTTAA
y3284: TGGTTGTG---GTGTTTAA
y3948: AATTTGAG---ACGTTGAT

Figure 9.7 PhoP Submotifs for *Salmonella* and *E. Coli* are extended to encompass *Yersinia*

Appendix B

Additional Figures and tables for Chapter 6 Learning robust dynamic net-works in prokaryotes by Gene Expression Networks Iterative Explorer (GENIE)

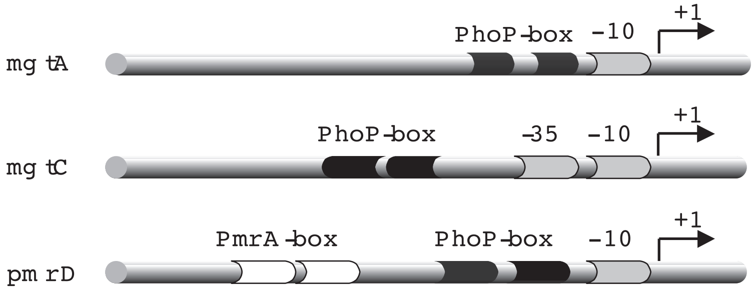


Figure 10.1 Modeling genetic interactions by analyzing Transcription Factor binding sites. This schema shows the transcriptional interaction between the PhoP/PhoQ two component system and PmrA/PmrB two component system (pmrD gene), obtained by studying the TFBS of both PhoP and PmrA. We also exemplify the *mgtA* and *mgtC* genes, regulated by PhoP.

We learn models of the transcription factor binding sites (TFBS) by clustering samples taken from RegulonDB and prototyping them by using positional weight matrices (i.e. motives). We also measured the distances between known TFBS and PhoP boxes; clustered these distances; and approximated them by fuzzy distributions (i.e. fuzzy membership functions). We search the entire intergenic region of *Salmonella enterica serovar Typhimurium*, employing these learnt motives, and detect the co-occurrence of distinct TFBS motives and PhoP putative boxes (i.e. those distances that have a significant membership value according to the learnt fuzzy functions.)

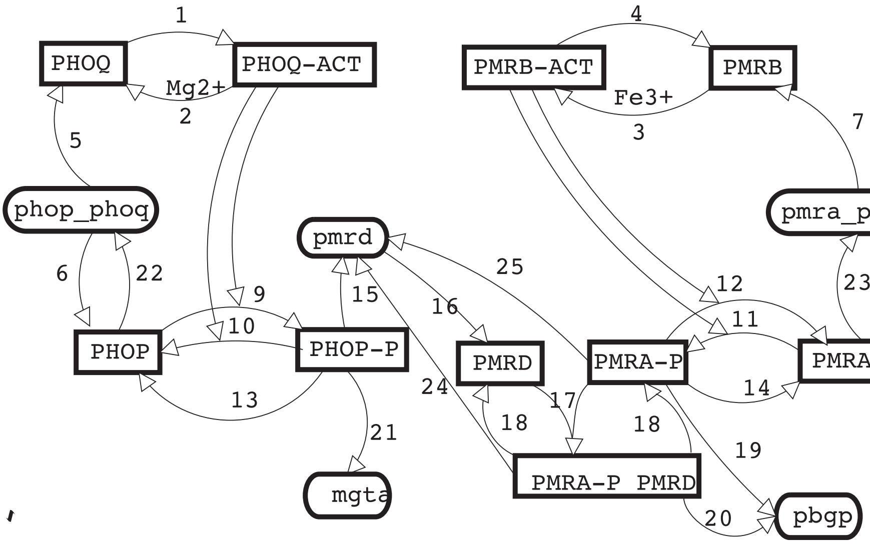


Figure 10.2 Reduced model.

The species interact as follows: 1/2- Low/High Mg^{2+} level favors the PHOP-ACT(ivated)/PHOP state in equilibrium. 3/4- High/Low Fe^{3+} level favors PMRB-ACT(ivated)/PMRB state in equilibrium. 5/6- phop_phoq is translated into PHOQ/PHOP proteins. 7/8- pmra_pmrb is translated into PMRB/PMRA proteins. 9- PHOP is phosphorilated (PHOP-P) by PHOQ-ACT kinase activity. 10- PHOP-P is desphosphorilated to PHOP by PHOQ-ACT phosphatase activity. 11- PMRA is phosphorilated to PMRA-P by PMRB-ACT kinase activity. 12- PMRA-P is desphosphorilated to PMRA by PMRB-ACT phosphatase activity. 13/14- PHOP-P/PMRA-P is spontaneous desphosphorilated to PHOP/PMRA. 15- PHOP-P activates the *pmrD* transcription. 16- *pmrD* is translated into PMRD. 17- PMRD binds PMRA-P (constituting PMRD_PMRA-P) which activates *pbgP* and represses *pmrD* genes, but it is not affected by the phosphatase activity of PMRB-ACT. 18- PMRA-P_PMRD unbinds into PMRD and PMRA-P. 19/20- PMRA-P/ PMRA-P_PMRD activates the transcription of *pbgP* gene. 21/22- PHOP-P activates the transcription of *mgta/phoP_phoQ*. 23- PMRA-P activates the transcription of *pmrA_pmrB*. 24/25- PMRA-P_PMRD/PMRA-P represses the transcription of *pmrD*

Table 10-1 Equations that model the final refined model.

Where **K** is the Coefficient of half-maximal activation; **H** is the Half-life (inverse to the degradation rate); **nu** is the Hill coefficient (values close to 1 produce near lineal curves that take sigmoid shape with higher values); **alpha** is the enhancer's saturation coefficient; r is the transformation rate (e.g. the transformation reaction of PHOQ into PHOQ_ACT); **P** is the phosphorylation rate; and ***T₀*** is characteristic constant of time, it relates the dimensional time (*t*) and the adimensional time (l): *t*= *T₀* l

$$
\begin{aligned}
\frac{\mathbf{d[PHOQ]}}{\mathbf{dt}} &= T_0(\frac{1}{H_{PHOQ}}[phop\_phoq]^{(5)} + r_{PHOQACT}max\_Mg2[PHOQACT][Mg2]^{(2)} \\
&\quad -[PHOQ]r^{(1)}_{PHOQ2PHOQACT} - \frac{[PHOQ]}{H_{PHOQ}}^{(dec)})
\end{aligned}
$$

$$
\frac{\mathbf{d[phop\_phoq]}}{\mathbf{dt}} = T_0(\frac{1}{H_{phop\_phoq}}\frac{[PHOPP]^{\nu PHOPP\_phop\_phoq}}{K^{\nu PHOPP\_phop\_phoq}_{PHOPP\_phop\_phoq} + PHOPP^{\nu PHOPP\_phop\_phoq}}^{(22)} - \frac{[phop\_phoq]}{H_{phop\_phoq}}^{(dec)})
$$

$$
\begin{aligned}
\frac{\mathbf{d[PHOQ\_ACT]}}{\mathbf{dt}} &= T_0([PHOQ]r^{(1)}_{PHOQ2PHOQACT} - r_{PHOQACT}max\_EMg2[PHOQACT][EMg2]^{(2)} \\
&\quad -\frac{[PHOQACT]}{H_{PHOQACT}}^{(dec)})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathbf{d[PHOP]}}{\mathbf{dt}} &= T_0(\frac{1}{H_{PHOP}}[phop\_phoq]^{(6)} \\
&\quad -P_{PHOQACTK}[PHOP]\frac{[PHOQACT]^{\nu PHOQACT\_PHOP}}{K^{\nu PHOQACT\_PHOP}_{PHOQACT\_PHOP} + [PHOQACT]^{\nu PHOQACT\_PHOP}}^{(9)} \\
&\quad +P_{PHOQP}[PHOPP]\frac{[PHOQ]^{\nu PHOQ\_PHOPP}}{K^{\nu PHOQ\_PHOPP}_{PHOQ\_PHOPP} + [PHOQ]^{\nu PHOQ\_PHOPP}}^{(10,1)} \\
&\quad -P_{PHOQK}[PHOP]\frac{[PHOQ]^{\nu PHOQ\_PHOP}}{K^{\nu PHOQ\_PHOP}_{PHOQ\_PHOP} + [PHOQ]^{\nu PHOQ\_PHOP}}^{(10,2)} \\
&\quad +r_{PHOPP}[PHOPP]^{(13)} - \frac{[PHOP]}{H_{PHOP}}^{(dec)})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathbf{d[PHOPP]}}{\mathbf{dt}} &= T_0(P_{PHOQACTK}[PHOP]\frac{[PHOQACT]^{\nu PHOQACT\_PHOP}}{K^{\nu PHOQACT\_PHOP}_{PHOQACT\_PHOP} + [PHOQACT]^{\nu PHOQACT\_PHOP}}^{(9)} \\
&\quad -P_{PHOQP}[PHOPP]\frac{[PHOQ]^{\nu PHOQ\_PHOPP}}{K^{\nu PHOQ\_PHOPP}_{PHOQ\_PHOPP} + [PHOQ]^{\nu PHOQ\_PHOPP}}^{(10,1)} \\
&\quad +P_{PHOQK}[PHOP]\frac{[PHOQ]^{\nu PHOQ\_PHOP}}{K^{\nu PHOQ\_PHOP}_{PHOQ\_PHOP} + [PHOQ]^{\nu PHOQ\_PHOP}}^{(10,2)} \\
&\quad -r_{PHOPP}[PHOPP]^{(13)} - \frac{[PHOPP]}{H_{PHOPP}}^{(dec)})
\end{aligned}
$$

$$
\frac{\mathbf{d[mgta]}}{\mathbf{dt}} = T_0(\frac{1}{H_{mgta}}\frac{[PHOPP]^{\nu PHOPP\_mgta}}{K^{\nu PHOPP\_mgta}_{PHOPP\_mgta} + [PHOPP]^{\nu PHOPP\_mgta}}^{(21)} - \frac{[mgta]}{H_{mgta}}^{(dec)})
$$

$$
\begin{aligned}
\frac{\mathbf{d[pmra\_pmrb]}}{\mathbf{dt}} &= T_0(\frac{1}{H_{pmra\_pmrb}}(1-(1-alpha_{PMRAP}\frac{1}{H_{pmra\_pmrb}} \\
&\quad \frac{[PMRAP]^{\nu PMRAP\_pmra\_pmrb}}{K^{\nu PMRAP\_pmra\_pmrb}_{PMRAP\_pmra\_pmrb} + [PMRAP]^{\nu PMRAP\_pmra\_pmrb}}) \\
&\quad (1-alpha_{PMRAP\_PMRD}\frac{1}{H_{pmra\_pmrb}} \\
&\quad \frac{[PMRAP\_PMRD]^{\nu PMRAP\_PMRD\_pmra\_pmrb}}{K^{\nu PMRAP\_PMRD\_pmra\_pmrb}_{PMRAP\_PMRD\_pmra\_pmrb} + [PMRAP\_PMRD]^{\nu PMRAP\_PMRD\_pmra\_pmrb}}))^{(23,26)} \\
&\quad -\frac{[pmra\_pmrb]}{H_{pmra\_pmrb}}^{(dec)})
\end{aligned}
$$

Table B.1: Equations that model the final refined mode (Continued)

$$\begin{aligned}
\frac{\mathbf{d[PMRB]}}{\mathbf{dt}} &= T_0(\frac{1}{H_{PMRB}}[pmrd]^{(7)} + r_{PMRBACT}[PMRBACT]^{(4)} \\
&\quad -r_{PMRBACT}max\_Fe3[PMRBACT][FE3+]^{(3)} - \frac{[PMRB]}{H_{PMRB}}^{(dec)})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathbf{d[PMRBACT]}}{\mathbf{dt}} &= T_0(r_{PMRBACT}max\_Fe3 + [PMRB][Fe3+]^{(3)} - [PMRBACT]r_{PMRBACT}^{(4)} \\
&\quad -\frac{[PMRB]}{H_{PMRB}}^{(dec)})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathbf{d[PMRA]}}{\mathbf{dt}} &= T_0(\frac{1}{H_{PMRA}}[pmra\_pmrb]^{(8)} + r_{PMRAP}[PMRAP]^{(14)} \\
&\quad -P_{PMRBACTK}[PMRA]\frac{[PMRBACT]^{\nu_{PMRBACT\_PMRAP}}}{K_{PMRBACT\_PMRAP}^{\nu_{PMRBACT\_PMRAP}} + [PMRBACT]^{\nu_{PMRBACT\_PMRAP}}}^{(11)} \\
&\quad +P_{PMRBK}[PMRAP]\frac{[PMRB]^{\nu_{PMRB\_PMRAP}}}{K_{PMRB\_PMRAP}^{\nu_{PMRB\_PMRAP}} + [PMRB]^{\nu_{PMRB\_PMRAP}}}^{(12,1)} \\
&\quad -P_{PMRBK}[PMRA]\frac{[PMRB]^{\nu_{PMRB\_PMRA}}}{K_{PMRB\_PMRA}^{\nu_{PMRB\_PMRA}} + [PMRB]^{\nu_{PMRB\_PMRA}}}^{(12,2)} \\
&\quad -\frac{[PMRA]}{H_{PMRA}}^{(dec)})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathbf{d[PMRAP]}}{\mathbf{dt}} &= T_0(P_{PMRBACTP}[PMRA]\frac{[PMRBACT]^{\nu_{PMRBACT\_PMRA}}}{K_{PMRBACT\_PMRA}^{\nu_{PMRBACT\_PMRA}} + [PMRBACT]^{\nu_{PMRBACT\_PMRA}}}^{(11)} \\
&\quad -P_{PMRBK}[PMRAP]\frac{[PMRB]^{\nu_{PMRB\_PMRAP}}}{K_{PMRB\_PMRAP}^{\nu_{PMRB\_PMRAP}} + [PMRB]^{\nu_{PMRB\_PMRAP}}}^{(12,1)} \\
&\quad P_{PMRBK}[PMRA]\frac{[PMRB]^{\nu_{PMRB\_PMRA}}}{K_{PMRB\_PMRA}^{\nu_{PMRB\_PMRA}} + [PMRB]^{\nu_{PMRB\_PMRA}}}^{(12,2)} \\
&\quad +\frac{[PMRAP\_PMRD]}{H_{PMRAP\_PMRD}}^{(18)} \\
&\quad -r_{PMRAP}[PMRAP]^{(14)} - r_{PMRAP\_PMRD}max\_PMRD[PMRAP][PMRD]^{(17)} \\
&\quad -\frac{[PMRAP]}{H_{PMRAP}}^{(dec)})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathbf{d[pbgp]}}{\mathbf{dt}} &= T_0(\frac{1}{H_{pbgp}}(1-(1-alpha_{PMRAP}\frac{1}{H_{pbgp}}\frac{[PMRAP]^{\nu_{PMRAP\_pbgp}}}{K_{PMRAP\_pbgp}^{\nu_{PMRAP\_pbgp}} + [PMRAP]^{\nu_{PMRAP\_pbgp}}}) \\
&\quad (1-alpha_{PMRAP\_PMRD}\frac{1}{H_{pbgp}} \\
&\quad \frac{[PMRAP\_PMRD]^{\nu_{PMRAP\_PMRD\_pbgp}}}{K_{PMRAP\_PMRD\_pbgp}^{\nu_{PMRAP\_PMRD\_pbgp}} + [PMRAP\_PMRD]^{\nu_{PMRAP\_PMRD\_pbgp}}}))^{(19,20)} \\
&\quad -\frac{[pbgp]}{H_{pbgp}}^{(dec)})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathbf{d[PMRD]}}{\mathbf{dt}} &= T_0(\frac{1}{H_{PMRD}}[pmrd]^{(16)} - r_{PMRAP\_PMRD}max\_PMRD[PMRAP][PMRD]^{(17)} \\
&\quad -\frac{[PMRD]}{H_{PMRD}}^{(dec)})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathbf{d[PMRAP\_PMRD]}}{\mathbf{dt}} &= T_0(r_{PMRAP\_PMRD}max\_PMRD[PMRAP][PMRD]^{(17)} \\
&\quad -\frac{[PMRAP\_PMRD]}{H_{PMRAP\_PMRD}}^{(18)})
\end{aligned}$$

$$\begin{aligned}
\frac{\mathbf{d[pmrd]}}{\mathbf{dt}} &= T_0(\frac{1}{H_{pmrd}}\frac{[PHOPP]^{\nu_{PHOPP\_pmrd}}}{K_{PHOPP\_pmrd}^{\nu_{PHOPP\_pmrd}} + [PHOPP]^{\nu_{PHOPP\_pmrd}}} \\
&\quad (1-\frac{[PMRAP]^{\nu_{PMRAP\_pmrd}}}{K_{PMRAP\_pmrd}^{\nu_{PMRAP\_pmrd}} + [PMRAP]^{\nu_{PMRAP\_pmrd}}}) \\
&\quad (1-\frac{[PMRAP\_PMRD]^{\nu_{PMRAP\_pmrd}}}{K_{PMRAP\_pmrd}^{\nu_{PMRAP\_pmrd}} + [PMRAP\_PMRD]^{\nu_{PMRAP\_pmrd}}})^{(15,24,25)} - \frac{[pmrd]}{H_{pmrd}}^{(dec)})
\end{aligned}$$

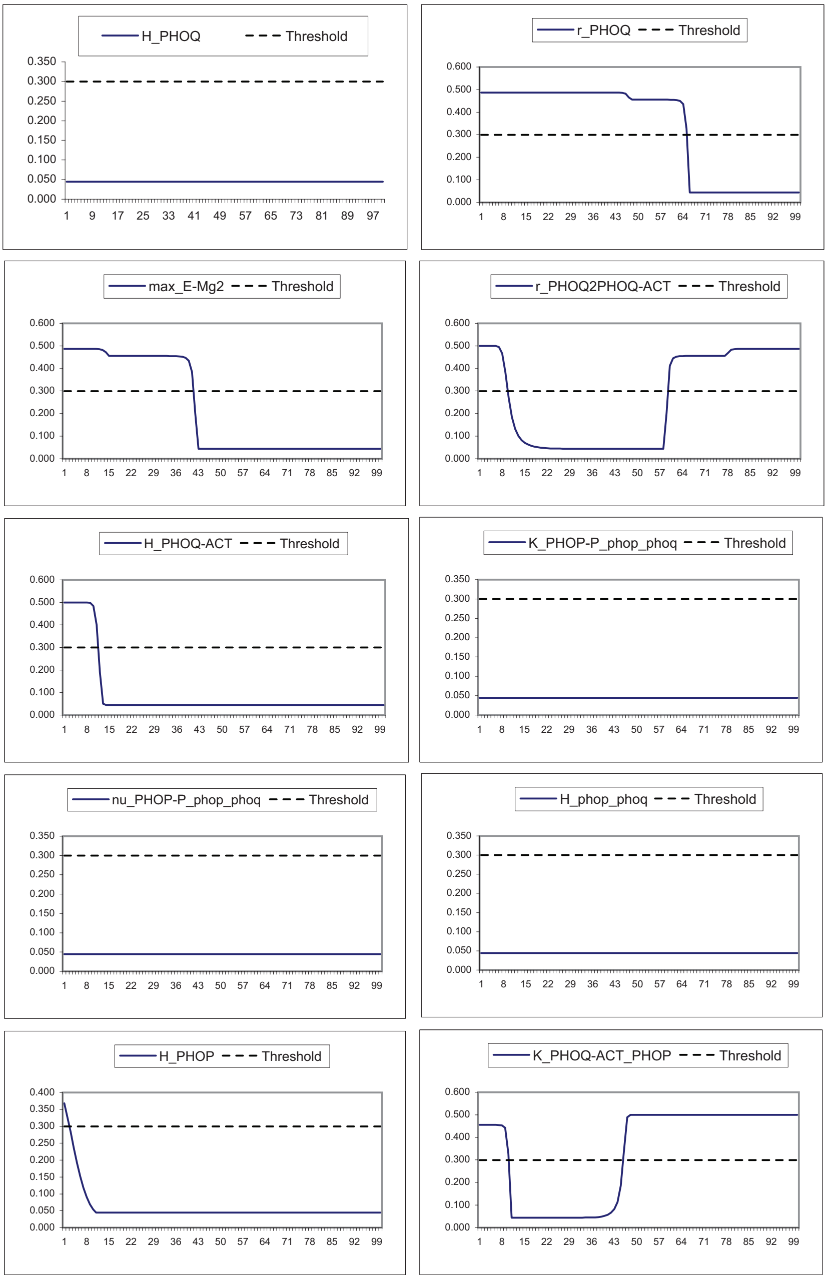


Figure 10.3 Robustness of parameters
Given a configuration, the model is evaluated by independently changing the value of each parameter in the entire range of its biological meaningful range of values (100 smaples). The model is valid if the obtained score is below a threshold value of 0.3 (indicated in red)

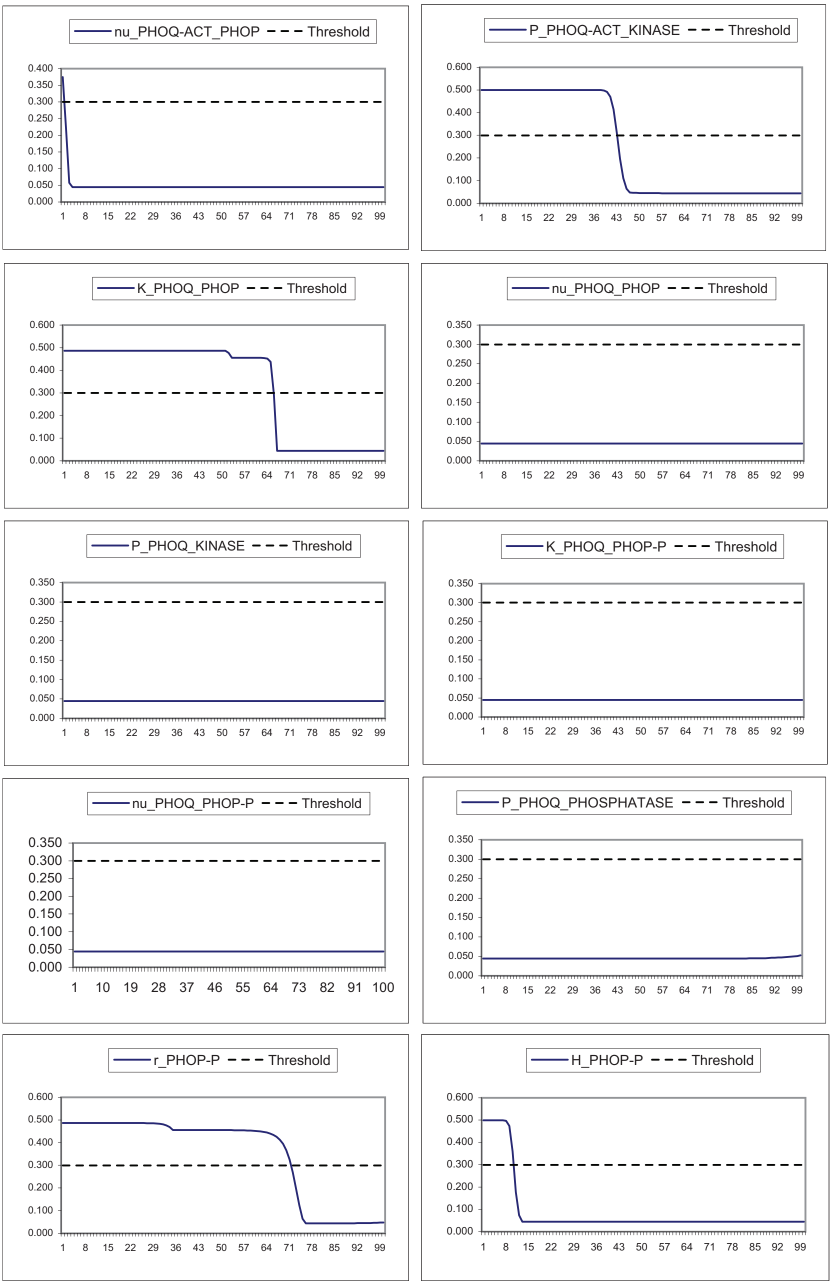


Figure 10.3 Robustness of parameters (Continued)

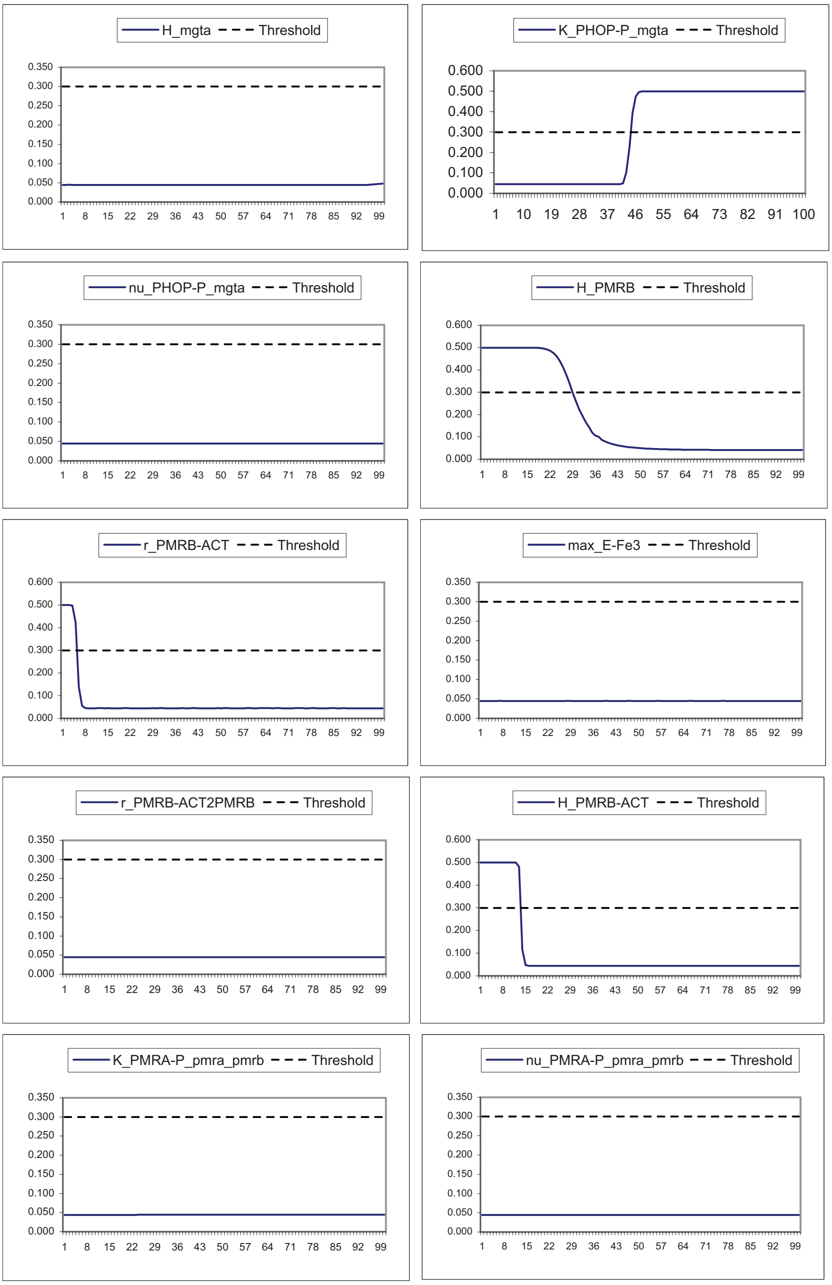


Figure 10.3 Robustness of parameters (Continued)

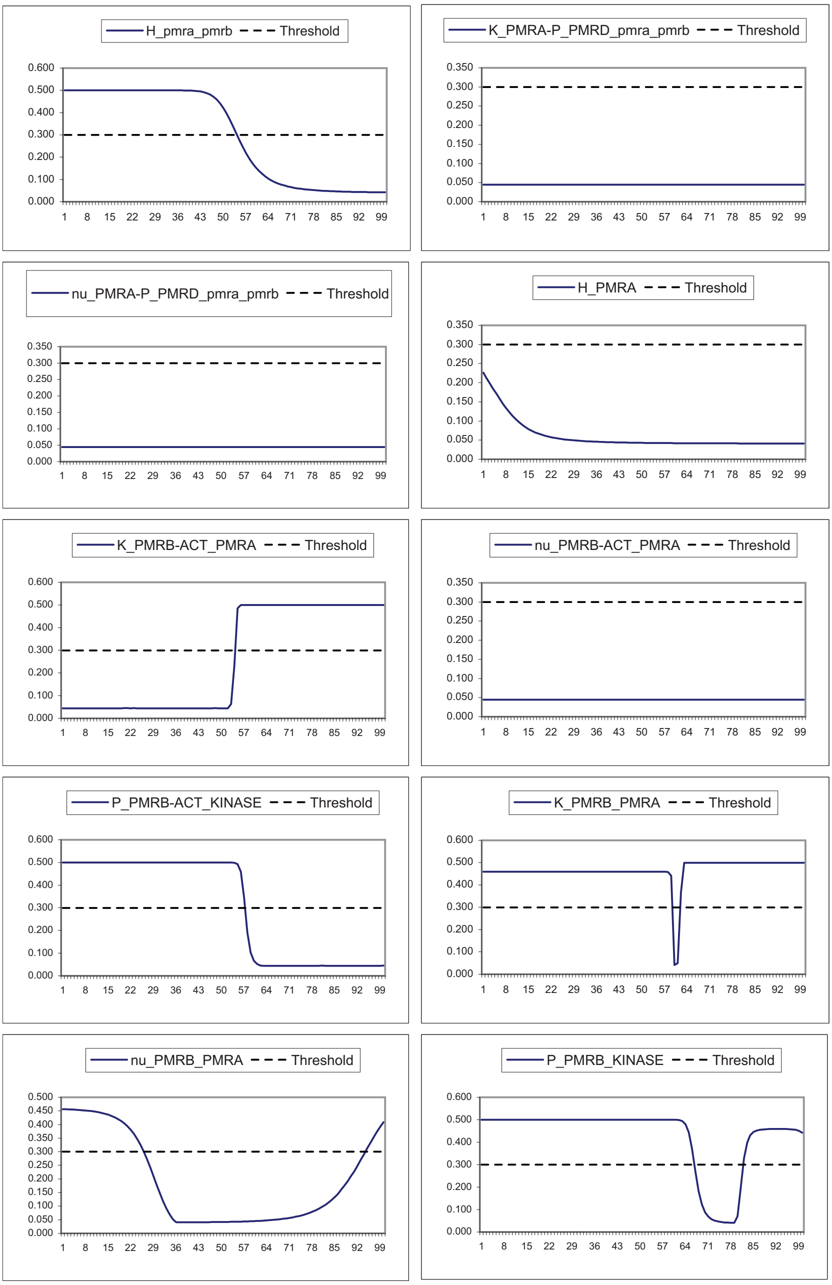


Figure 10.3 Robustness of parameters (Continued)

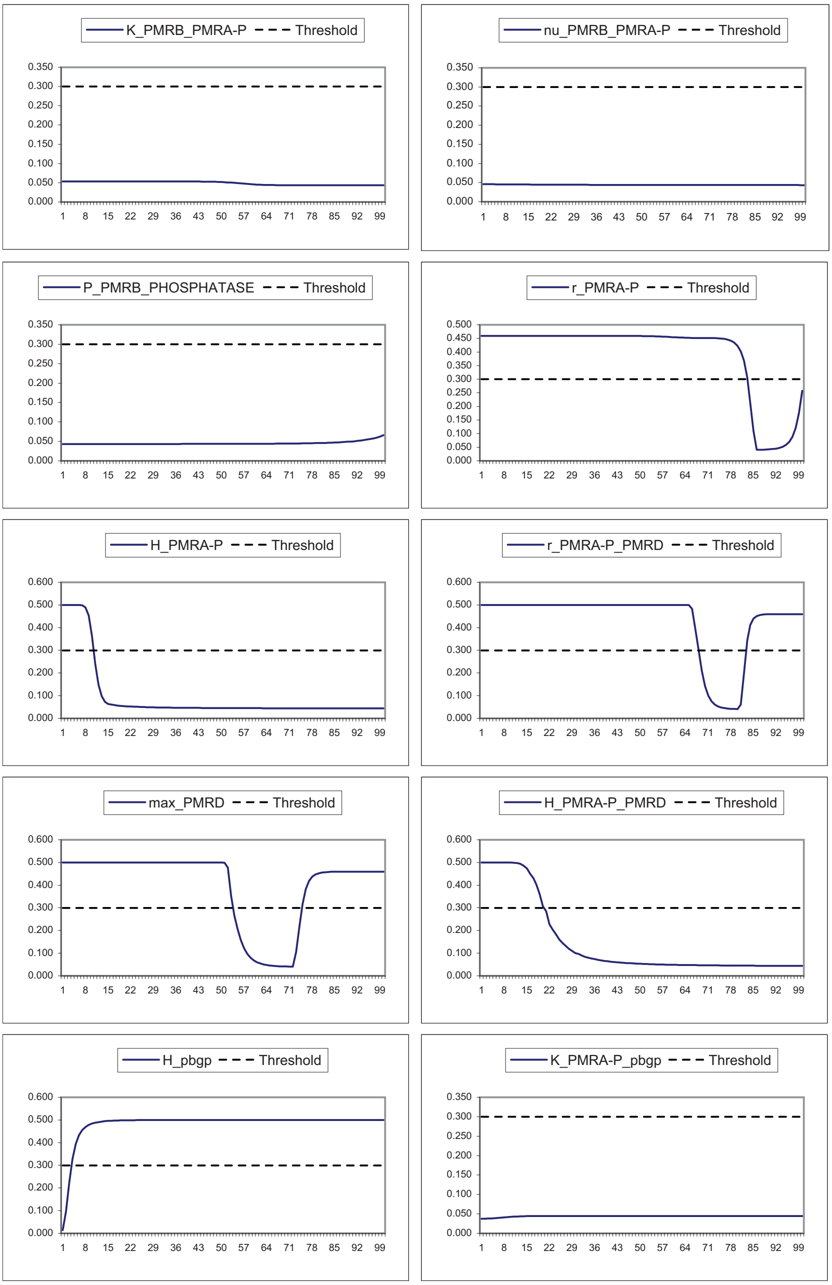


Figure 10.3 Robustness of parameters (Continued)

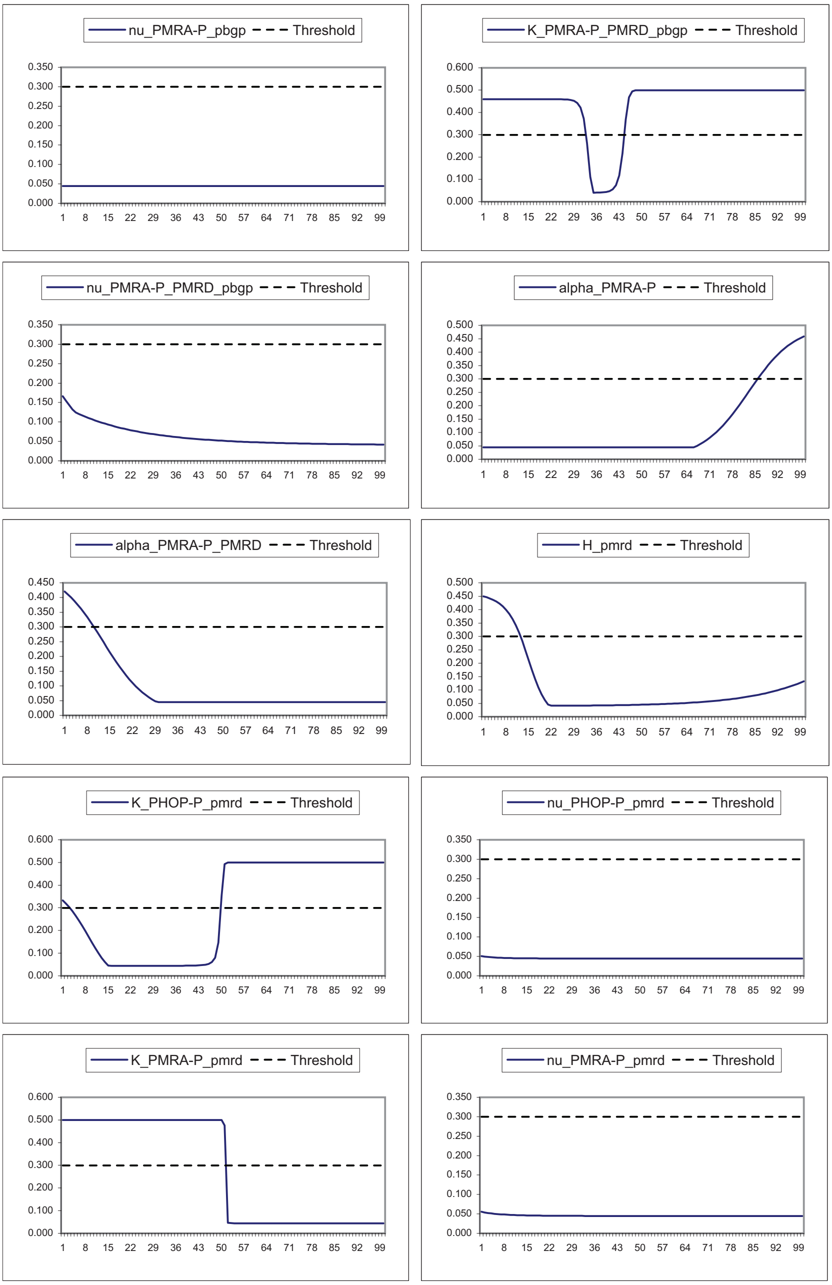


Figure 10.3 Robustness of parameters (Continued)

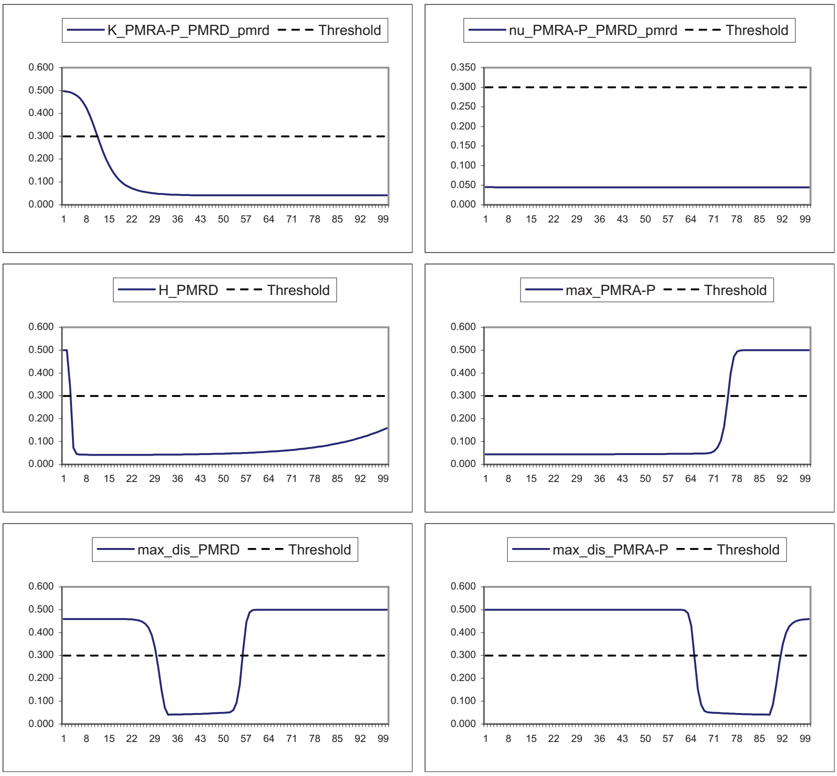
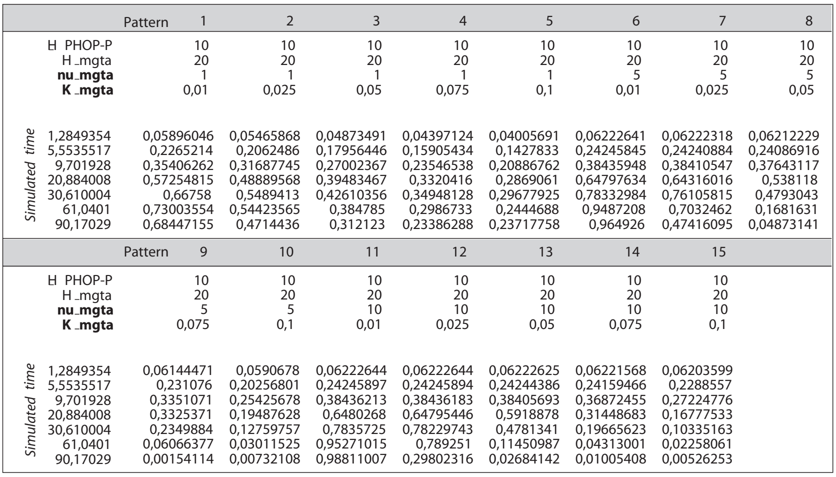


Figure 10.3 Robustness of parameters (Continued)

Table 10-2 Parameters that fulfill less than the 50% of their biological range

Parameter	Prom.	Prop.	Min	Max
H_pbgp	3,3	1	3	6
K_PMRA-P_PMRD_pbgp	24,7	0,7	12	95
r_PMRA-P_PMRD	25	0,5	14	42
K_PHOQ-ACT_PHOP	29	0,3	13	38
K_PHOP-P_pmrd	29,7	0,3	3	48
max_PMRD	32,8	0,6	19	57
r_PMRA-P	34,9	0,5	5	100
K_PMRB_PMRA	37,4	0,3	2	60
r_PHOP-P	43,3	0,1	17	100
r_PHOQ	43,4	0,1	21	69
K_PHOQ_PHOP	43,4	0	32	100
r_PHOQ2PHOQ-ACT	43,6	0	33	54
K_PMRA-P_pmrd	45,5	0,1	5	74
P_PHOQ-ACT_KINASE	46,4	0,1	13	71
P_PMRB_KINASE	47	0,4	15	100
K_PHOP-P_mgta	48,9	0	43	63

Table 10-3 Predicted patterns of behavior.



	Pattern	1	2	3	4	5	6	7	8
	H PHOP-P	10	10	10	10	10	10	10	10
	H _mgta	20	20	20	20	20	20	20	20
	nu_mgta	1	1	1	1	1	5	5	5
	K _mgta	0,01	0,025	0,05	0,075	0,1	0,01	0,025	0,05
Simulated time	1,2849354	0,05896046	0,05465868	0,04873491	0,04397124	0,04005691	0,06222641	0,06222318	0,06212229
	5,5535517	0,2265214	0,2062486	0,17956446	0,15905434	0,1427833	0,24245845	0,24240884	0,24086916
	9,701928	0,35406262	0,31687745	0,27002367	0,23546538	0,20886762	0,38435948	0,38410547	0,37643117
	20,884008	0,57254815	0,48889568	0,39483467	0,3320416	0,2869061	0,64797634	0,64316016	0,538118
	30,610004	0,66758	0,5489413	0,42610356	0,34948128	0,29677925	0,78332984	0,76105815	0,4793043
	61,0401	0,73003554	0,54423565	0,384785	0,2986733	0,2444688	0,9487208	0,7032462	0,1681631
	90,17029	0,68447155	0,4714436	0,312123	0,23386288	0,23717758	0,964926	0,47416095	0,04873141

	Pattern	9	10	11	12	13	14	15
	H PHOP-P	10	10	10	10	10	10	10
	H _mgta	20	20	20	20	20	20	20
	nu_mgta	5	5	10	10	10	10	10
	K _mgta	0,075	0,1	0,01	0,025	0,05	0,075	0,1
Simulated time	1,2849354	0,06144471	0,0590678	0,06222644	0,06222644	0,06222625	0,06221568	0,06203599
	5,5535517	0,231076	0,20256801	0,24245897	0,24245894	0,24244386	0,24159466	0,2288557
	9,701928	0,3351071	0,25425678	0,38436213	0,38436183	0,38405693	0,36872455	0,27224776
	20,884008	0,3325371	0,19487628	0,6480268	0,64795446	0,5918878	0,31448683	0,16777533
	30,610004	0,2349884	0,12759757	0,7835725	0,78229743	0,4781341	0,19665623	0,10335163
	61,0401	0,06066377	0,03011525	0,95271015	0,789251	0,11450987	0,04313001	0,02258061
	90,17029	0,00154114	0,00732108	0,98811007	0,29802316	0,02684142	0,01005408	0,00526253

Distinct patterns of expression are obtain by scanning the parameters of a feasible solution. The predictions are obtained by choosing the configuration for the final refined model that has the least phosphorylated Phop half-life value (H PHOP_P) and independently scanning the half-maximal activation of the generic gene *mgtA* (K_mgta) and its Hill coefficient (nu_mgta).

Table 10-4 Predicted patterns of behavior. Correlation of predictions to GFP experiments.

Gene		*phoP*	*mgtA*	*rstA*	*pmrD*	*slyB*	*mig-14*
Predicted patterns	1	0,61928709	0,85794978	0,33981874	0,82900028	0,26492335	0,48246059
	2	0,74576937	0,90652632	0,50200732	0,90152895	0,42424263	0,62833589
	3	0,84385231	0,91513845	0,64837599	0,93632341	0,58115667	0,75165981
	4	0,88615311	0,89874129	0,72579435	0,93636486	0,67187368	0,8116634
	5	0,81397157	0,84270697	0,65716106	0,88230393	0,64110106	0,74696069
	6	0,48457839	0,78407633	0,18265797	0,73523995	0,11935587	0,33487934
	7	0,85897756	0,96137959	0,6485592	0,9710322	0,49942263	0,75927577
	8	0,84352487	0,49976237	**0,94611432**	0,62670515	**0,92220393**	**0,92202348**
	9	0,55070375	0,11792405	0,74954543	0,24777596	0,9178142	0,65077755
	10	0,28659218	-0,1507399	0,54579546	-0,0273447	0,79958289	0,39458262
	11	0,46897421	0,77224203	0,1670765	0,72251153	0,10739606	0,31967558
	12	**0,9130103**	**0,97569855**	0,71356089	**0,98141336**	0,53376224	0,80470675
	13	0,80105453	0,42983635	0,91562486	0,56325458	0,91377424	0,89447241
	14	0,42572242	-0,0079116	0,63967864	0,11352873	0,88694265	0,52465998
	15	0,12272233	-0,2913141	0,40322016	-0,1797741	0,69813201	0,22577177
Best correlation		**12**	**12**	**8**	**12**	**8**	**8**

Gene		*mgtC*	*pagP*	*pagK*	*pagC*	*pcgL*
Predicted patterns	1	0,27716907	0,55674403	0,5991426	0,18018035	0,32097919
	2	0,74576937	0,67109489	0,72241444	0,33232432	0,47099036
	3	0,55281042	0,7628738	0,82205318	0,48287298	0,60867876
	4	0,62534764	0,8035574	0,86710168	0,56966587	0,68207881
	5	0,60067196	0,78255531	0,83482741	0,54888883	0,64997028
	6	0,14282156	0,4369099	0,47037	0,04220185	0,17710809
	7	0,50921551	0,74791195	0,80498073	0,41359311	0,58302558
	8	0,89572365	**0,835517**	**0,8686341**	0,90023069	**0,9324809**
	9	0,73294236	0,53028742	0,59107194	0,85041257	0,73332059
	10	0,51280449	0,25338073	0,32540817	0,68762304	0,50931899
	11	0,13228195	0,42638344	0,45821103	0,03199406	0,16522582
	12	0,50909639	0,71588885	0,79386547	0,42468576	0,58693351
	13	**0,91940832**	0,83067834	0,84277191	**0,91707216**	0,93035611
	14	0,64766201	0,4095504	0,4853395	0,81048726	0,63651178
	15	0,34921301	0,07199035	0,15531988	0,55652528	0,34925742
Best correlation		**13**	**8**	**8**	**13**	**8**

.

The correlation between each prediction and the experimentally obtained results is evaluates and selected the one with best score (highlighted in bold). Three predictions best represent the GFP curves: pattern 12 predicts genes with early rise time and high level of expression; pattern 13 correlates to genes that exhibit late rise time and low level of expression; and pattern 8 recovers genes with an intermediate kinetic behavior

Bibliography

(1997). Handbook of Evolutionary Computation. Bristol, UK, IOP Publishing Ltd.

Abel, K. M. (2004). "Foetal origins of schizophrenia: testable hypotheses of genetic and environmental influences." Br J Psychiatry **184**: 383-5.

Agrawal, R., H. Mannila, et al. (1996). Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press**:** 307-328.

Agrawal, R. and J. C. Shafer (1996). "Parallel mining of association rules." Ieee Transactions on Knowledge and Data Engineering **8**(6): 962-969.

Akil, M., J. N. Pierri, et al. (1999). "Lamina-specific alterations in the dopamine innervation of the prefrontal cortex in schizophrenic subjects." Am J Psychiatry **156**(10): 1580-9.

Alberts, B., A. Johnson, et al. (2003). Biolog\'ia molecular de la c\'elula. Cuarta Edici\'on, Omega.

Alcalá, R., O. Cordón, et al. (2002). Insurance Market Risk Modeling with Hierarchical Fuzzy Rule Based Systems. Proceedings of the 4th International Conference on Enterprise Information systems (ICEIS).

Alizadeh, A. A., M. B. Eisen, et al. (200). "Distinct types of diffuse large B-cell." Nature **403**(503): 511.

Alm, E., K. Huang, et al. (2006). "The evolution of two-component systems in bacteria reveals different strategies for niche adaptation." PLoS Comput Biol **2**(11): e143.

Alon, U. (2007). An introduction to System Biology. London, CRC Press, Taylor & Francis Group.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Attwood, T. K. and D. J. Parry-Smith (2002). Introducci\'on a la Bioinform\'atica, Prentice Hall.

Azevedo, R. B., R. Lohaus, et al. (2005). "The simplicity of metazoan cell lineages." Nature **433**(7022): 152-6.

Babuska, R. (1998). Fuzzy Modeling for Control. Norwell, MA, USA, Kluwer Academic Publishers.

Badner, J. A. and E. S. Gershon (2002). "Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia." Mol Psychiatry **7**(4): 405-11.

Bailey , T. L. and C. Elkan (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, Menlo Park, California, AAAI Press.

Bailey, T. L. and C. Elkan (1995). "The value of prior knowledge in discovering motifs with MEME." Proc Int Conf Intell Syst Mol Biol **3**: 21-9.

Bailey, T. L., N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res **34**(Web Server issue): W369-73.

Baldi, P., Y. Chauvin, et al. (1998). "Computational applications of DNA structural scales." Proc Int Conf Intell Syst Mol Biol **6**: 35-42.

Bar-Joseph, Z., G. K. Gerber, et al. (2003). "Computational discovery of gene modules and regulatory networks." Nat Biotechnol **21**(11): 1337-42.

Barash, Y., Elidan, G., Friedman, N., Kaplan, T. (2003). Modeling Dependencies in Protein-DNA Binding Sites. RECOMB'03.

Bardosey, A. and L. Ducjstein (1995). "Fuzzy rule-based modeling with application to geophysical, biological and engineering systems." CRC Press.

Barnard, A., A. Wolfe, et al. (2004). "Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes." Curr Opin Microbiol **7**(2): 102-8.

Batchelor, E. and M. Goulian (2003). "Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system." Proc Natl Acad Sci U S A **100**(2): 691-6.

Bauer, E. and R. Kohavi (1999). "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants." Machine Learning **36**(1-2): 105-139.

Beer, M. A. and S. Tavazoie (2004). "Predicting gene expression from sequence." Cell **117**(2): 185-98.

Ben-Dor, A., L. Bruhn, et al. (2000). "Tissue classification with gene expression profiles." Computer Biol. **7**(559): 583.

Benitez-Bellon, E., G. Moreno-Hagelsieb, et al. (2002). "Evaluation of thresholds for the detection of binding sites for regulatory proteins in Escherichia coli K12 DNA." Genome Biol **3**(3): RESEARCH0013.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society, Series B (Methodological) **57**(1): 289-300.

Bennage, W. A. and A. K. Dhingra (1995). "Single and Multiobjective Structural Optimization in Discrete-Continuous Variables Using Simulated Annealing." International Journal for Numerical Methods in Engineering **38**: 2753-2773.

Benson, D., I. Karsch-Mizrachi, et al. (2005). "GenBank." Nucleic Acids Research **34**: 16-20.

Berenji, H. R. and P. Khedkar (1992). "Learning and tuning fuzzy logic controllers through reinforcements." IEEE Trans Neural Netw **3**(5): 724-40.

Berg, D., B. Jabs, et al. (2001). "Echogenicity of substantia nigra determined by transcranial ultrasound correlates with severity of parkinsonian symptoms induced by neuroleptic therapy." Biol Psychiatry **50**(6): 463-7.

Berg, D., C. Siefker, et al. (2001). "Echogenicity of the substantia nigra in Parkinson's disease and its relation to clinical findings." J Neurol **248**(8): 684-9.

Berg, J., J. Tymoczko, et al. (2003). Bioqu\'imica. Quinta Edici\'on, Reverte.

Bezdek, J. (1998). "Fuzzy clustering." Handbook of Fuzzy Computation: f6.1:1-f6.6:19.

Bezdek, J. C. (1998). Pattern Analysis. Handbook of Fuzzy Computation. W. Pedrycz, P. P. Bonissone and E. H. Ruspini. Bristol, Institute of Physics**:** F6.1.1-F6.6.20.

Bezdek, J. C., S. K. Pal, et al. (1992). Fuzzy models for pattern recognition : methods that search for structures in data. New York, IEEE Press.

Borgelt, C. a. K., R. (2002). Induction of Association Rules: Apriori Implementation. 15th Conference on Computational Statistics (Compstat), Berlin, Germany, Physica Verlag.

Breier, A., T. P. Su, et al. (1997). "Schizophrenia is associated with elevated amphetamine-induced synaptic dopamine concentrations: evidence from a novel positron emission tomography method." Proc Natl Acad Sci U S A **94**(6): 2569-74.

Brenner, S. (2000). "Genomics. The end of the beginning." Science **287**(5461): 2173-4.

Browning, D. F. and S. J. Busby (2004). "The regulation of bacterial transcription initiation." Nat Rev Microbiol **2**(1): 57-65.

Brukner, I., R. Sanchez, et al. (1995). "Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides." EMBO J **14**(8): 1812-8.

Buck, M. J. and J. D. Lieb (2004). "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." Genomics **83**(3): 349-60.

Caligiuri, M. P., J. B. Lohr, et al. (1993). "Parkinsonism in neuroleptic-naive schizophrenic patients." Am J Psychiatry **150**(9): 1343-8.

Calvano, S. E., W. Xiao, et al. (2005). "A network-based analysis of systemic inflammation in humans." Nature.

Cannon, T. D., M. O. Huttunen, et al. (2000). "The inheritance of neuropsychological dysfunction in twins discordant for schizophrenia." Am J Hum Genet **67**(2): 369-82.

Carter, C. J. and C. J. Pycock (1980). "Behavioural and biochemical effects of dopamine and noradrenaline depletion within the medial prefrontal cortex of the rat." Brain Res **192**(1): 163-76.

Cavicchio, D. J. (1970). Adaptive search using simulated evolution, University of Michigan.

Chakos, M. H., D. I. Mayerhoff, et al. (1992). "Incidence and correlates of acute extrapyramidal symptoms in first episode of schizophrenia." Psychopharmacol Bull **28**(1): 81-6.

Charnes, A. and W. W. Cooper (1961). Management Models and Industrial Applications of Linear Programming Vol. 1, John Wiley & Sons, New York.

Cheeseman, P. and R. W. Oldford (1994). Selecting models from data : artificial intelligence and statistics IV. New York, Springer-Verlag.

Cho, R. J., M. J. Campbell, et al. "A genome-wide transcriptional analysis of the mitotic cell cycle."

Coello-Coello, C., D. V. Veldhuizen, et al. (2002). Evolutionary Algorithms for Solving Multi-Objective Problems, Kluwer.

Collado-Vides, J., B. Magasanik, et al. (1991). "Control site location and transcriptional regulation in Escherichia coli." Microbiol Rev **55**(3): 371-94.

Collins, F. S., M. Morgan, et al. (2003). "The Human Genome Project: Lessons from Large-Scale Biology." Science **300**(5617): 286-290.

Conlon, E. M., X. S. Liu, et al. (2003). "Integrating regulatory motif discovery and genome-wide expression analysis." Proc Natl Acad Sci U S A **100**(6): 3339-44.

Consortium, E. (2002). Elvira: An Environment for Probabilistic Graphical Models. 1st European Workshop on Probabilistic Graphical Models.

Cook, D., L. Holder, et al. (2001). "Structural Mining of Molecular Biology Data." IEEE Engineering in Medicine and Biology, special issue on Advances in Genomics **4**(20): 67-74.

Cooper, G. F. and E. Herskovits (1992). "A Bayesian Method for the Induction of Probabilistic Networks from Data." Machine Learning **9**(4): 309-347.

Cordon, O., H. F., et al. (1997). "Applicability of the fuzzy operators in the design of fuzzy logic controllers." Fuzzy Sets and Systems(86): 15-41.

Cordon, O., F. Herrera, et al. (2002). "Linguistic modeling by hierarchical systems of linguistic rules." Ieee Transactions on Fuzzy Systems **10**(1): 2-20.

Cotik, V., R. Romero-Zaliz, et al. (2005). "A Hybrid Promoter Anaysis Methodology for Prokaryotic Genomes." Special issue on ``Bioinformatics'', Fuzzy Sets and Systems **152**(1): 83-102.

Cotik, V., R. R. Zaliz, et al. (2005). "A hybrid promoter analysis methodology for prokaryotic genomes." Fuzzy Sets and Systems **152**(1): 83-102.

Cover, T. M. and J. A. Thomas (2006). Entropy, Relative Entropy and Mutual Information. Elements of Information Theory.

Cox, R. S., 3rd, M. G. Surette, et al. (2007). "Programming gene expression with combinatorial promoters." Mol Syst Biol **3**: 145.

Craddock, N., M. C. O'Donovan, et al. (2005). "The genetics of schizophrenia and bipolar disorder: dissecting psychosis." J Med Genet **42**(3): 193-204.

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-90.

D. C. Grainger, T. W. Overton, et al. (2004). "Genomic studies with Escherichia coli MelR protein: applications of chromatin immunoprecipitation and microarrays." J Bacteriol, **186**: 6938-43.

Darwin, C. (1859). On The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London, John Murray.

de Campos, L. M. a. P., J.M. (2001). "Stochastic local search algorithms for learning belief networks: searching in the space of orderings." Lecture Notes in Artificial Intelligence **2143**: 228-239.

de Erausquin, G. A. (2004). "Transactivation of cell death signals by glutamate transmission in dopaminergic neurons." Crit Rev Neurobiol **16**(1-2): 107-19.

De Jong, K. (1975). An analysis of the behaviour of a class of genetic adaptive systems, University of Michigan.

Deb, K. (2001). Multi-objective optimization using evolutionary algorithms. Chichester ; New York, John Wiley & Sons.

Der, G. and B. S. Everitt (1996). A handbook of statistical analyses using SAS, CHAPMAN-HALL.

Driankov, A. L., D. Phil, et al. (1993). "An introduction to fuzzy control." Springer-Verlag.

Durbin, R. (1998). Biological sequence analysis : probabilistic models of proteins and nucleic acids. Cambridge, Cambridge Univ. Press.

Egan, M. F., T. E. Goldberg, et al. (2001). "Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia." Proc Natl Acad Sci U S A **98**(12): 6917-22.

Elemento, O., N. Slonim, et al. (2007). "A universal framework for regulatory element discovery across all genomes and data types." Mol Cell **28**(2): 337-50.

Everitt, B. and G. Der (1996). A handbook of statistical analysis using SAS. London, Chapman & Hall.

Falkenauer, E. (1998). Genetic Algorithms and Grouping Problems. New York, John Wiley & Sons.

Flint, J. (2003). "Analysis of quantitative trait loci that influence animal behavior." J Neurobiol **54**(1): 46-77.

Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-61.

Freedman, R., A. Olincy, et al. (2003). "The genetics of sensory gating deficits in schizophrenia." Curr Psychiatry Rep **5**(2): 155-61.

Gasch, A. P. and M. B. Eisen (2002). "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering." Genome Biol **3**(11): RESEARCH0059.

Gasperoni, T. L., J. Ekelund, et al. (2003). "Genetic linkage and association between chromosome 1q and working memory function in schizophrenia." Am J Med Genet B Neuropsychiatr Genet **116B**(1): 8-16.

Gass, S. I. and T. L. Saaty (1955). "The computational algorithm for the parametric objective function." Naval Research Logistics Quarterly **2**: 39.

Gertz, J., L. Riles, et al. (2005). "Discovery, validation, and genetic dissection of transcription factor binding sites by comparative and functional genomics." Genome Res **15**(8): 1145-52.

Goldberg, D. and J. J. Richardson (1987). "Genetic algorithms with sharing for multimodal function optimization." Proceedings Second International Conference on Genetic Algorithm: 41-49.

Goldberg, T. E., M. F. Egan, et al. (2003). "Executive subprocesses in working memory: relationship to catechol-O-methyltransferase Val158Met genotype and schizophrenia." Arch Gen Psychiatry **60**(9): 889-96.

Goni, J. R., A. Perez, et al. (2007). "Determining promoter location based on DNA structure first-principles calculations." Genome Biol **8**(12): R263.

Gottesman, II and J. Shields (1973). "Genetic theorizing and schizophrenia." Br J Psychiatry **122**(566): 15-30.

Groisman, E. A. (2001). "The pleiotropic two-component regulatory system PhoP-PhoQ." J Bacteriol **183**(6): 1835-42.

Groisman, E. A. and C. Mouslim (2006). "Sensing by bacterial regulatory systems in host and non-host environments." Nat Rev Microbiol **4**(9): 705-9.

Guet, C. C., M. B. Elowitz, et al. (2002). "Combinatorial synthesis of genetic networks." Science **296**(5572): 1466-70.

Gutierrez-Rios, R. M., D. A. Rosenblueth, et al. (2003). "Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles." Genome Res **13**(11): 2435-43.

Halpern, A. L. and W. J. Bruno (1998). "Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies." Mol Biol Evol **15**(7): 910-7.

Hans, A. E. (1988). "Multicriteria optimization for highly accurate systems." Multicriteria Optimization in Engineering and Sciences **19**: 309-352.

Harrison, P. J. and D. R. Weinberger (2005). "Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence." Mol Psychiatry **10**(1): 40-68; image 5.

Hasty, J., D. McMillen, et al. (2002). "Engineered gene circuits." Nature **420**(6912): 224-30.

Haugen, S. P., M. B. Berkmen, et al. (2006). "rRNA promoter regulation by nonoptimal binding of sigma region 1.2: an additional recognition element for RNA polymerase." Cell **125**(6): 1069-82.

Heckerman, D., D. Geiger, et al. (1995). "Learning Bayesian Networks - the Combination of Knowledge and Statistical-Data." Machine Learning **20**(3): 197-243.

Hellendoorn, H. and C. Thomas (1993). "Deffuzification in fuzzy controllers." Journal of Intelligent Fuzzy Systems **1**: 109–123.

Hering, J. A., P. R. Innocent, et al. (2004). "Beyond average protein secondary structure content prediction using FTIR spectroscopy." Appl Bioinformatics **3**(1): 9-20.

Herrera, F. and M. Lozano (2005). "Editorial Real coded genetic algorithms." Soft Computing **9**(4): 223-224.

Herrera, F., M. Lozano, et al. (1995). "Tuning fuzzy logic controllers by genetic algorithms." International Journal of Approximate Reasoning **12**(3): 299-315.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics **15**(7-8): 563-77.

Hoch, J. A. (2000). "Two-component and phosphorelay signal transduction." Curr Opin Microbiol **3**(2): 165-70.

Hocking, R. R. (1976). "The analysis and selection of variables in linear regression." Biometrics(32): 1-49.

Horan, W. P., D. L. Braff, et al. (2008). "Verbal working memory impairments in individuals with schizophrenia and their first-degree relatives: findings from the Consortium on the Genetics of Schizophrenia." Schizophr Res **103**(1-3): 218-28.

Hughes, J. D., P. W. Estep, et al. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae." J Mol Biol **296**(5): 1205-14.

Ishihama, A. (1993). "Protein-protein communication within the transcription apparatus." J Bacteriol **175**(9): 2483-9.

Jacob, F. (1998). Of Flies, Mice, and Men, Harvard University Press.

Janky, R. and J. van Helden (2007). "Discovery of conserved motifs in promoters of orthologous genes in prokaryotes." Methods Mol Biol **395**: 293-308.

Kaasinen, V., E. Nurmi, et al. (2001). "Personality traits and brain dopaminergic function in Parkinson's disease." Proc Natl Acad Sci U S A **98**(23): 13272-7.

Kærn, M. (2003). Regulatory dynamics in engineered gene networks. 4th International Systems Biology Conference. Washington University, St. Louis.

Kanz, C., P. Aldebert, et al. (2005). "The EMBL Nucleotide Sequence Database." Nucleics Acids Research **33**(suppl_1): D29-33.

Kato, A. and E. A. Groisman (2004). "Connecting two-component regulatory systems by a protein that protects a response regulator from dephosphorylation by its cognate sensor." Genes Dev **18**(18): 2302-13.

Kato, A., T. Latifi, et al. (2003). "Closing the loop: the PmrA/PmrB two-component system negatively controls expression of its posttranscriptional activator PmrD." Proc Natl Acad Sci U S A **100**(8): 4706-11.

Kato, A., H. Tanabe, et al. (1999). "Molecular characterization of the PhoP-PhoQ two-component system in Escherichia coli K-12: identification of extracellular Mg2+-responsive promoters." J Bacteriol **181**(17): 5516-20.

Kimura, S., T. Kawasaki, et al. (2004). "OBIYagns: a grid-based biochemical simulator

with a parameter estimator." Bioinformatics **22**(10): 1646–1648.

Klir, G. J. and T. A. Folger (1988). Fuzzy sets, uncertainty, and information. London, Prentice Hall International.

Klosgen, W. (1996). Explora: A Multipattern and Multistrategy Discovery Assistant. Advances in Knowledge Discovery and Data Mining, MIT Press**:** 249-271.

Kohavi, R. and G. H. John (1997). "Wrappers for feature subset selection." Artificial Intelligence **97**(1-2): 273-324.

Kolomeets, N. S. and N. A. Uranova (1999). "Synaptic contacts in schizophrenia: studies using immunocytochemical identification of dopaminergic neurons." Neurosci Behav Physiol **29**(2): 217-21.

Kox, L. F., M. M. Wosten, et al. (2000). "A small protein that mediates the activation of a two-component system by another two-component system." Embo J **19**(8): 1861-72.

Kursawe, F. (1991). "A Variant of Evolution Strategies for Vector Optimization." PPSN I: Proceedings of the 1st Workshop on Parallel Problem Solving from Nature: 193-197.

Lee, D. and H.-S. Seung (1999). "Learning the parts of objects by non-negative matrix factorization." Nature **401**(6755): 788-791.

Lee, S. and H. Wang (1992). "Modified Simulated Annealing for Multiple Objective Engineering Design Optimization." Journal of Intelligent Manufacturing **3**: 101-108.

Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.

Lejona, S., A. Aguirre, et al. (2003). "Molecular characterization of the Mg2+-responsive PhoP-PhoQ regulon in Salmonella enterica." J Bacteriol **185**(21): 6287-94.

Lencer, R., P. Trillenberg, et al. (2004). "Smooth pursuit deficits in schizophrenia, affective disorder and obsessive-compulsive disorder." Psychol Med **34**(3): 451-60.

Leung, T. H., A. Hoffmann, et al. (2004). "One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers." Cell **118**(4): 453-64.

Li, C. and W. Hung Wong (2001). "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application." Genome Biol **2**(8): RESEARCH0032.

Li, C. and W. H. Wong (2001). "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection." Proc Natl Acad Sci U S A **98**(1): 31-6.

Li, H., V. Rhodius, et al. (2002). "Identification of the binding sites of regulatory proteins in bacterial genomes." Proc Natl Acad Sci U S A **99**(18): 11772-7.

Liu, M., M. Tolstorukov, et al. (2004). "A mutant spacer sequence between -35 and -10 elements makes the Plac promoter hyperactive and cAMP receptor protein-independent." Proc Natl Acad Sci U S A **101**(18): 6911-6.

Lockhart, D. J. and E. A. Winzeler (2000). "Genomics, gene expression and DNA arrays." Nature **405**(6788): 827-836.

Mamdani, E. H. (1974). "Applications of fuzzy algorithm for control a simple dynamic plant." Proceeding of the IEEE **12**: 1585-1577.

Mangan, S. and U. Alon (2003). "Structure and function of the feed-forward loop network motif." Proc Natl Acad Sci U S A **100**(21): 11980-5.

Manson McGuire, A. and G. M. Church (2000). "Predicting regulons and their cis-regulatory motifs by comparative genomics." Nucleic Acids Res **28**(22): 4523-30.

Martinez-Antonio, A. and J. Collado-Vides (2003). "Identifying global regulators in transcriptional regulatory networks in bacteria." Curr Opin Microbiol **6**(5): 482-9.

Marton, M. J., J. L. D. H. A. Bennett, et al. (1998). "Drug target validation and identification of secondary drug target effects using DNA microarrays." Nat Med **4**(11): 1235-1236.

Masuda, N. and G. M. Church (2003). "Regulatory network of acid resistance genes in Escherichia coli." Mol Microbiol **48**(3): 699-712.

McAdams, H. H. and A. Arkin (1999). "It's a noisy business! Genetic regulation at the nanomolar scale." Trends Genet **15**(2): 65-9.

McClellan, J. M., E. Susser, et al. (2007). "Schizophrenia: a common disease caused by multiple rare alleles." Br J Psychiatry **190**: 194-9.

McCue, L., W. Thompson, et al. (2001). "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes." Nucleic Acids Res **29**(3): 774-82.

Meir, E., E. M. Munro, et al. (2002). "Ingeneue: a versatile tool for reconstituting genetic networks, with examples from the segment polarity network." J Exp Zool **294**(3): 216-51.

Meir, E., G. von Dassow, et al. (2002). "Robustness, flexibility, and the role of lateral inhibition in the neurogenic network." Curr Biol **12**(10): 778-86.

Mejia-Roa, E., P. Carmona-Saez, et al. (2008). "bioNMF: a web-based tool for nonnegative matrix factorization in biology." Nucleic Acids Res **36**(Web Server issue): W523-8.

Michalewicz, Z. (1994). Genetic algorithms + data structures = evolution programs (2nd, extended ed.), Springer-Verlag New York, Inc.

Michalewicz, Z. and D. B. Fogel (2000). How to solve it: modern heuristics. New York, NY, USA, Springer-Verlag New York, Inc.

Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: simple building blocks of complex networks." Science **298**(5594): 824-7.

Minagawa, S., H. Ogasawara, et al. (2003). "Identification and molecular characterization of the Mg2+ stimulon of Escherichia coli." J Bacteriol **185**(13): 3696-702.

Mitchell, T. (1997). Machine Learning. New York, McGraw-Hill.

Mitchell, T. M. (1997). Machine learning. New York, McGraw-Hill.

Mjolsness, E., T. Mann, et al. (1999). From Coexpression to Coregulation: An Approach to Inferring Transcriptional Regulation Among Gene Classes from Large-Scale Expression Data, BEACON eSpace at Jet Propulsion Laboratory

Moses, A. M., D. Y. Chiang, et al. (2003). "Position specific variation in the rate of evolution in transcription factor binding sites." BMC Evol Biol **3**: 19.

Moses, A. M., D. A. Pollard, et al. (2006). "Large-scale turnover of functional transcription factor binding sites in Drosophila." PLoS Comput Biol **2**(10): e130.

Nadon, R. and J. Shoemaker (2002). "Statistical issues with microarrays: processing and analysis." Trends Genet **18**(5): 265-71.

Oshima, T., H. Aiba, et al. (2002). "Transcriptome analysis of all two-component regulatory system mutants of Escherichia coli K-12." Mol Microbiol **46**(1): 281-91.

Osman, I. (1995). "An introduction to Meta-heuristics." Operational Research Tutorial Papers Series, Annal Conference OR37 - Canterbury.

Osyczka, A. (1978). "An Approach to Multicriterion Optimization Problems for Engineering Design." Computer Methods in Applied Mechanics and Engineering **15**: 309-333.

Paunio, T., A. Tuulio-Henriksson, et al. (2004). "Search for cognitive trait components of schizophrenia reveals a locus for verbal learning and memory on 4q and for visual working memory on 2q." Hum Mol Genet **13**(16): 1693-702.

Pedersen, A. G., L. J. Jensen, et al. (2000). "A DNA structural atlas for Escherichia coli." J Mol Biol **299**(4): 907-30.

Pedrycz, W., P. Bonissone, et al. (1998). Handbook of fuzzy computation, Institute of Physics.

Peralta, V., M. J. Cuesta, et al. (2000). "Differentiating primary from secondary negative symptoms in schizophrenia: a study of neuroleptic-naive patients before and after treatment." Am J Psychiatry **157**(9): 1461-6.

Pharoah, P., A. Antonio, et al. (2008). "Polygenes, risk prediction, and targeted prevention of breast cancer." New England Journal of Medicine **26**(358): 2796-803.

Pond, S. L., S. D. Frost, et al. (2005). "HyPhy: hypothesis testing using phylogenies." Bioinformatics **21**(5): 676-9.

Pritsker, M., Y. C. Liu, et al. (2004). "Whole-genome discovery of transcription factor binding sites by network-level conservation." Genome Res **14**(1): 99-108.

Ptashne, M. and A. Gann (2002). Genes and signals. New York, Cold Spring Harbor Laboratory Press.

Purcell, S., B. Neale, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-75.

Qin, Z. S., L. A. McCue, et al. (2003). "Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites." Nat Biotechnol **21**(4): 435-9.

Quinlan, J. R. (1993). C4.5 : programs for machine learning. San Mateo, Calif., Morgan Kaufmann Publishers.

Rissanen, J. (1989). Stochastic complexity in statistical inquiry. Singapore, World Scientific.

Robison, K., A. M. McGuire, et al. (1998). "A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome." J Mol Biol **284**(2): 241-54.

Romero Zaliz, R., I. Zwir, et al. (2004). Generalized analysis of promoters: a method for DNA sequence description. Applications of Multi-Objective Evolutionary Algorithms. C. a. L. Coello Coello, G. Singapore, World Scientific**:** 427-450.

Ronen, M., R. Rosenberg, et al. (2002). "Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics." Proc Natl Acad Sci U S A **99**(16): 10555-60.

Rosania, G. R., Y. T. Chang, et al. (2000). "Myoseverin, a microtubule-binding molecule with novel cellular effects." Nat Biotechnol **18**(3): 261-262.

Rosner, B. (1986). Fundamentals of biostatistics. Boston, Mass., Duxbury Press.

Rubio-Escudero, C., O. Harari, et al. (2007). Modeling Genetic Networks: Comparison of Static and Dynamic Models. Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics, Valencia, Spain, Springer.

Ruspini, E. and I. Zwir (2001). "Automated Generation of Qualitative Representations of Complex Object by Hybrid Soft-computingMethods." Pattern Recognition: From Classical to Modern Approaches: 453-474.

Ruspini, E. H. and I. Zwir (1999). Automated Qualitative Description of Measurements. Proceedings of the 16th IEEE Instrumentation and Measurement Technology Conf., Venice, Italy.

Ruspini, E. H. and I. Zwir (2002). Automated generation of qualitative representations of complex objects by hybrid soft-computing methods. Pattern recognition : from classical to modern approaches. S. K. Pal and A. Pal. New Jersey., World Scientific**:** 454-474.

Salgado, H., S. Gama-Castro, et al. (2004). "RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12." Nucleic Acids Res **32**(Database issue): D303-6.

Salgado, H., A. Santos-Zavaleta, et al. (2001). "RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12." Nucleic Acids Res **29**(1): 72-4.

Saporta, G. (1996). "Probabilits, analyse des donnes et statistiques." Technip.

Schwarz, G. (1978). "Estimating the dimension of a model." Annals of Statistics **6**: 461-464.

Shi, Y., T. Latifi, et al. (2004). "Transcriptional control of the antimicrobial peptide resistance ugtL gene by the Salmonella PhoP and SlyA regulatory proteins." J Biol Chem **279**(37): 38618-25.

Shin, D. and E. A. Groisman (2005). "Signal-dependent Binding of the Response Regulators PhoP and PmrA to Their Target Promoters in Vivo." J Biol Chem **280**(6): 4089-94.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Stat Appl Genet Mol Biol **3**: Article3.

Spirtes, P., C. Glymour, et al. (1991). "From Probability to Causality." Philosophical Studies **64**(1): 1-36.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Sugeno, M. and T. Yasukama (1993). "A Fuzzy-logic-based Approach to Qualitative Modeling." IEEE Transactions on Fuzzy Systems **1**(1): 7-31.

Tavazoie, S., J. Hughes, et al. (1999). "Systematic determination of genetic network architecture." Nature Genet. **22**(3): 281-285.

Tavazoie, S., J. D. Hughes, et al. (1999). "Systematic determination of genetic network architecture." Nat Genet **22**(3): 281-5.

Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol **23**(1): 137-44.

Tucker, D. L., N. Tucker, et al. (2002). "Gene expression profiling of the pH response in Escherichia coli." J Bacteriol **184**(23): 6551-8.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.

van Someren, E. P., L. F. Wessels, et al. (2002). "Genetic network modeling." Pharmacogenomics **3**(4): 507-25.

von Dassow, G., E. Meir, et al. (2000). "The segment polarity network is a robust developmental module." Nature **406**(6792): 188-92.

Wade, J. T., K. Struhl, et al. (2007). "Genomic analysis of protein-DNA interactions in bacteria: insights into transcription and chromosome organization." Mol Microbiol **65**(1): 21-6.

Wahde, M., J. A. Hertz, et al. (2001). Reverse Engineering of Sparsely Connected Genetic Regulatory Networks Proceedings of the 2nd Workshop of Biochemical Pathways and Genetic Networks.

Wilkins, C. L. (2000). "Data mining with Spotfire Pro 4.0." Analytical Chemistry **72**(15): 550a-550a.

Winfield, M. D. and E. A. Groisman (2004). "Phenotypic differences between Salmonella and Escherichia coli resulting from the disparate regulation of homologous genes." Proc Natl Acad Sci U S A **101**(49): 17162-7.

Wolff, A. L. and G. A. O'Driscoll (1999). "Motor deficits and schizophrenia: the evidence from neuroleptic-naive patients and populations at risk." J Psychiatry Neurosci **24**(4): 304-14.

Wray, G. A., M. W. Hahn, et al. (2003). "The evolution of transcriptional regulation in eukaryotes." Mol Biol Evol **20**(9): 1377-419.

X. Tu, T. Latifi, et al. (2006). "The PhoP/PhoQ two-component system stabilizes the alternative sigma factor RpoS in Salmonella enterica." Proc. Natl. Acad. Sci. **(in press)**.

Yamamoto, K., H. Ogasawara, et al. (2002). "Novel mode of transcription regulation of divergently overlapping promoters by PhoP, the regulator of two-component system sensing external magnesium availability." Mol Microbiol **45**(2): 423-38.

Yeung, K. Y. and W. L. Ruzzo (2001). "Principal component analysis for clustering gene expression data." Bioinformatics **17**(9): 763-74.

Zadeh, L. (1973). "Outline of a new approach to the analysis of complex systems and decision processes." IEEE Transactions on Systems, Man, and Cybernetics(3): 28-44.

Zadeh, L. A. (1965). "Fuzzy Sets." Information Control **8**: 338-353.

Zadeh, L. A. (1975). "The concept of a linguistic variable and its application to approximate reasoning." Information Sciences **8**: 119-249.

Zadeh, L. A. (1992). Knowledge Representation in Fuzzy Logic. An Introduction to Fuzzy Logic Applications in Intelligent Systems. Boston, Kluwer**:** 1-25.

Zhang, N. L. (2004). "Hierarchical Latent Class Models for Cluster Analysis." Journal of Machine Learning Research **5**: 697-723.

Zheng, D., C. Constantinidou, et al. (2004). "Identification of the CRP regulon using in vitro and in vivo transcriptional profiling." Nucleic Acids Res **32**(19): 5874-93.

Zwir, I., O. Harari, et al. (2007). "Gene promoter scan methodology for identifying and classifying coregulated promoters." Methods Enzymol **422**: 361-85.

Zwir, I., H. Huang, et al. (2005). "Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation." Bioinformatics **21**(22): 4073-4083.

Zwir, I., D. Shin, et al. (2005). "Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica." Proc Natl Acad Sci U S A **102**(8): 2862-7.

Zwir, I., P. Traverso, et al. (2003). Semantic-oriented analysis of regulation: the PhoP regulon as a model network. Proceedings of the 3rd International Conference on Systems Biology (ICSB), St. Louis, USA.