

Jens Warfsmann

**A nonlinear systemic approach to genome analysis**

## Doctoral Thesis

Directed by Dr. Hilario Ramírez Rodrigo  
This work has been realized within the doctoral program  
*Biochemistry and Molecular Biology (113.89.1)*



University of Granada, November 2015

Editor: Universidad de Granada. Tesis Doctorales  
Autor: Jens Warfsmann  
ISBN: 978-84-9125-816-2  
URI: <http://hdl.handle.net/10481/43537>



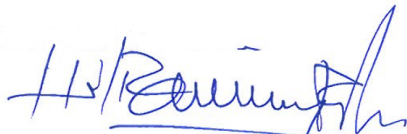
COMPROMISO DE RESPETO DE LOS DERECHOS DE AUTOR

El doctorando Jens Warfsmann y el director de la tesis Hilario Ramírez Rodrigo , garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección del director de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada 15.11.2015

Director de la Tesis

Fdo.:



HILARIO RAMÍREZ  
RODRIGO

Hilario Ramírez Rodrigo

Doctorando

Fdo.:



Jens Warfsmann



*"[...] los sistemas biológicos son producto de la evolución. Si las matemáticas son el arte de lo perfecto y la física es el arte de lo óptimo, la biología no es más que el arte de lo satisfactorio: cualquier cosa sirve, siempre que funcione. [...]"*

— Sidney Brenner, *El País*, 12-01-1999



---

# Contents

<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xv</b>
<b>1 Resumen en castellano</b>	<b>1</b>
<b>2 Introduction</b>	<b>15</b>
2.1 Complex systems . . . . .	15
2.1.1 Features of complex systems . . . . .	17
Nonlinearity . . . . .	17
Feedback . . . . .	18
Spontaneous order . . . . .	18
Robustness and lack of central control . . . . .	19
Emergence . . . . .	19
2.1.2 Mathematical background . . . . .	19
2.2 Cancer . . . . .	22
2.2.1 Clonal evolution in cancer . . . . .	22
2.2.2 Epigenetics in cancer . . . . .	23
2.3 Guanine-cytosine content . . . . .	26
2.3.1 GC-content in prokaryotes . . . . .	27



2.3.2	GC-content in mitochondria . . . . .	28
2.4	Support Vector Machines . . . . .	29
2.4.1	Linearly separable binary classification . . . . .	30
2.4.2	Using soft hyperplanes for binary classification with not-linearly separable data . . . . .	32
2.4.3	Dimensional scaling-up with kernels to perform non- linear classification . . . . .	33
<b>3</b>	<b>Objectives</b>	<b>37</b>
<b>4</b>	<b>Methods</b>	<b>39</b>
4.1	Recurrence analysis . . . . .	39
4.1.1	Phase space reconstruction . . . . .	39
4.1.2	Recurrence plots . . . . .	42
4.1.3	Recurrence quantitative analysis and lacunarity . . . . .	42
4.2	Classification . . . . .	44
4.2.1	Support vector machines . . . . .	44
	Linearly separable binary classification . . . . .	44
	Non-linearly separable binary classification . . . . .	47
4.3	Evaluation . . . . .	48
<b>5</b>	<b>Results</b>	<b>49</b>
5.1	Work flow . . . . .	49
5.2	Marker-less cancer classification . . . . .	56
5.2.1	Data acquisition and preparation . . . . .	56
	The original DNA methylation data . . . . .	56
	Preparation of the data files and working directories . . . . .	58
	DNA methylation density time series . . . . .	60
5.2.2	Time delayed reconstruction . . . . .	63
5.2.3	Systemic characterization of epigenetic changes in dif- ferent cancer types: Binary classification of tumor and normal cells . . . . .	67
5.2.4	Comparison of tumor versus normal classification on CpG sites with cancer specific differential methyla- tion and systemic features . . . . .	77

5.2.5	Tumor versus normal classification on randomly selected non cancer specific DNA-methylation sites. . .	78
5.2.6	Tumor versus normal classification on global $\beta$ -value sums . . . . .	79
5.3	Functional and gene position distances in the tomato genome	85
5.4	Systemic identification of taxa . . . . .	94
5.4.1	Data acquisition and preparation . . . . .	95
5.4.2	Pan . . . . .	95
5.4.3	Caniformia . . . . .	96
5.5	Anti joint recurrence plot . . . . .	104
<b>6</b>	<b>Discussion</b>	<b>105</b>
6.1	On cancer . . . . .	108
6.2	On tomato . . . . .	115
6.3	On chimpazees and Caniformia . . . . .	116
<b>7</b>	<b>Conclusions</b>	<b>119</b>
	<b>Bibliography</b>	<b>123</b>
	<b>Online resources</b>	<b>147</b>
	<b>Acknowledgments</b>	<b>149</b>



---

## List of Figures

2.1	Research fields associated with complexity science. . . . .	16
2.2	Robust reactions of the system: to stay or to change. . . . .	21
2.3	Global depiction of epigenomic alterations during oncogenesis. . . . .	24
2.4	Nucleotide composition continuum of completely sequenced mitochondrial and plastid DNA sequences. . . . .	29
2.5	Support vector machines: Finding hyperplanes. . . . .	31
2.6	Support vector machines: Search for the maximum margin hyperplane. . . . .	32
2.7	Support vector machines: Finding the “convex hull”. . . . .	33
2.8	Support vector machines: Mapping into higher dimension spaces. . . . .	34
5.1	Work flow of a nonlinear systemic approach for genome analysis. . . . .	50
5.2	Influence of the threshold $\varepsilon$ on the RQA measures shown for the standard Rössler system. . . . .	52
5.3	Standard Rössler attractor for $a = 0.2$ ; $b = 0.2$ ; $c = 5.7$ . . . . .	53
5.4	Time series build from the y-coordinates of a standard Rössler system and its reconstruction. . . . .	54
5.5	Distance contour plot and recurrence plot ( $\varepsilon = 0.22$ ) for the reconstruction shown in 5.4. . . . .	55
5.6	Visualization of the directed acyclic graph describing the first entry from a magetab document. . . . .	59
5.7	Structure of an analysis working directory. . . . .	61

5.8	DNA methylation density time series from a normal head and neck sample. . . . .	63
5.9	Graphical representation of reconstruction parameters. . . . .	64
5.10	Multi scatter plot of a reconstructed phase space. . . . .	65
5.11	Projection of a reconstructed phase space to the first three dimensions. . . . .	66
5.12	a) Distance contour plot and b) recurrence plot ( $\varepsilon = 0.126$ ) for the reconstruction shown in 5.10. . . . .	68
5.13	Recurrence plots of phase spaces representing a) colon and rectal adenocarcinoma, b) normal colon tissue and c) a difference plot (AJRP) of both. . . . .	69
5.14	Cluster analysis of RQA measures and lacunarity. . . . .	70
5.14	Cluster analysis of RQA measures and lacunarity. Continued . . .	71
5.14	Cluster analysis of RQA measures and lacunarity. Continued . . .	72
5.15	Pairwise comparison of RQA measures and lacunarity. . . . .	73
5.16	ROCs and AUCs used to evaluate the binary classifications of complete data sets. . . . .	76
5.17	Performance of a marker-based and a marker-less head and neck cancer versus normal tissue SVM binary classification. . . . .	78
5.18	ROC curves and AUCs for head and neck tumor versus normal tissue SVM binary classification based on DNA-methylation data from TCGA. . . . .	79
5.19	Boxplots for $\beta$ -value sums . . . . .	80
5.20	Boxplots of $\beta$ -value sums including CpG sites located only in BLOCKS. . . . .	81
5.21	Boxplots of $\beta$ -value sums including CpG sites located only in regions belonging to cancer related gene symbols, BLOCKS and cDMR. . . . .	82
5.22	Boxplots of $\beta$ -value sums including CpG sites located only in regions belonging to cancer related gene symbols and cDMR. . .	83
5.23	Boxplots of $\beta$ -value sums including CpG sites located only in cDMR. . . . .	84
5.24	The ITAG annotation pipeline <sup>1</sup> and the activity diagram of the PCB gene ontology annotation protocol. . . . .	86

5.25 A casual drawing of mated distance patterns of the tomato genome.	93
5.26 Casual drawing of plausible and not plausible Chimpanzee mi- grations . . . . .	97



---

## List of Tables

5.1	The head of a TCGA 3rd level HumanMethylation450 data file.	58
5.2	Description of the bar code meta data. . . . .	62
5.3	Statistical summary of all delays calculated for the marker-less cancer classification. . . . .	66
5.4	Statistical summary of all embedding dimensions calculated for the marker-less cancer classification. . . . .	66
5.5	Sample number distribution for the different data sets used in the tumor versus normal tissue classification on systemic features and complete CpG sets. . . . .	74
5.6	Areas under the curve (AUC) in % for marker-less classifications.	74
5.7	Areas under the curve (AUC) in % for marker-less classifications. Continued . . . . .	74
5.8	Wilcoxon rank sum test with continuity correction on $\beta$ value sums for complete and reduced DNA methylation data sets of different cancer types. . . . .	81
5.9	Areas under the curve (AUCs) in % for the $\beta$ -value sum classifications. . . . .	84
5.10	Functional distances as a function of inter gene distances of the + strand for the tomato chromosomes 1 – 12. . . . .	87
5.10	Continued . . . . .	88
5.10	Continued . . . . .	89



5.11 Gene base position distance time series, their reconstructed phase space distance plots and the corresponding recurrences plots. . .	90
5.11 Continued . . . . .	91
5.11 Continued . . . . .	92
5.12 Sample number distribution for Mt genomes of the genus Pan. .	96
5.13 Graphs from the recurrence analysis of time series based on the CG content of Pan Mt-genome sequences. . . . .	98
5.14 ROCs and AUCs for systemic classifications on CG content of Mt-genome sequences from the genus Pan. . . . .	100
5.15 Sample number distribution for Mt genomes of the suborder Caniformia. . . . .	101
5.16 Graphs from the recurrence analysis of time series based on the CG content of Caniformia Mt-genome sequences. . . . .	102

---

## List of Tables

5.1	The head of a TCGA 3rd level HumanMethylation450 data file.	58
5.2	Description of the bar code meta data. . . . .	62
5.3	Statistical summary of all delays calculated for the marker-less cancer classification. . . . .	66
5.4	Statistical summary of all embedding dimensions calculated for the marker-less cancer classification. . . . .	66
5.5	Sample number distribution for the different data sets used in the tumor versus normal tissue classification on systemic features and complete CpG sets. . . . .	74
5.6	Areas under the curve (AUC) in % for marker-less classifications.	74
5.7	Areas under the curve (AUC) in % for marker-less classifications. Continued . . . . .	74
5.8	Wilcoxon rank sum test with continuity correction on $\beta$ value sums for complete and reduced DNA methylation data sets of different cancer types. . . . .	81
5.9	Areas under the curve (AUCs) in % for the $\beta$ -value sum classifications. . . . .	84
5.10	Functional distances as a function of inter gene distances of the + strand for the tomato chromosomes 1 – 12. . . . .	87
5.10	Continued . . . . .	88
5.10	Continued . . . . .	89

5.11	Gene base position distance time series, their reconstructed phase space distance plots and the corresponding recurrences plots. . .	90
5.11	Continued . . . . .	91
5.11	Continued . . . . .	92
5.12	Sample number distribution for Mt genomes of the genus Pan. .	96
5.13	Graphs from the recurrence analysis of time series based on the CG content of Pan Mt-genome sequences. . . . .	98
5.14	ROCs and AUCs for systemic classifications on CG content of Mt-genome sequences from the genus Pan. . . . .	100
5.15	Sample number distribution for Mt genomes of the suborder Caniformia. . . . .	101
5.16	Graphs from the recurrence analysis of time series based on the CG content of Caniformia Mt-genome sequences. . . . .	102

*To my marvelous family.*



## Resumen en castellano

La investigación llevada a cabo ha tenido como objetivo fundamental la implementación, puesta a punto y análisis preliminar de un procedimiento original para la discriminación sistémica, de alta sensibilidad, de genomas o epigenomas representativos de condiciones tipo.

Los resultados obtenidos, presentados en esta memoria, avalan el enfoque sistémico propuesto, que se basa en dos conjeturas fundamentales: por una parte la consideración de que las dinámicas adaptativas a las que se ven sometidos los Genomas pueden ser analizadas desde la óptica de los Sistemas Complejos Adaptativos y, en particular, desde el marco conceptual y metodológico de las teorías que sobre Complejidad y Caos determinista han venido desarrollándose durante los últimos veinte años y que se han aplicado con éxito en otros campos. Por otra parte, la presunción de que las secuencias de ADN pueden ser conceptualizadas como series temporales multivariantes no lineales y ser tratadas como tales a nivel de modelos formales. Ello es posible porque conceptualmente una serie temporal es en esencia una colección ordenada de valores observacionales relativos a una de las variables de estado del sistema. No es la temporalidad en sentido estricto sino la ordinalidad de los datos lo que determina su dimensión "temporal", de tal modo que para todo valor  $v_i$  de la variable considerada puede establecerse de manera unívoca el valor precedente  $v_{i-1}$  y el subsecuente  $v_{i+1}$ .

Aunque no conocemos la existencia, por el momento, de ningún planteamiento teórico riguroso al respecto, ninguno de estos dos presupuestos son extraños al campo del Análisis del DNA<sup>2,3</sup>. La aplicación de métodos de análisis de Time

Series a secuencias de DNA viene ya, de hecho, empleándose de forma puntual desde finales de los ochenta<sup>4-7</sup>. Por otra parte, la representación de genomas, interactomas, proteomas y otros en términos de Redes Complejas Adaptativas (una estrategia frecuente en análisis de Sistemas Complejos) es ya habitual en numerosos contextos de la moderna Biología de Sistemas<sup>7-10,3</sup>. Se trata este último de un enfoque singularmente potente, ya que desde los trabajos pioneros de de Erdős y Rennyi<sup>11,12</sup>, el marco formal y metodológico del análisis de redes aleatorias está muy bien establecido y es fácil extrapolar las conclusiones obtenidas de los modelos de red a los Sistemas Complejos originales, más elusivos al análisis.

Este tipo de enfoque puede además revelar datos no solo acerca la estructura del Sistema Complejo original al que representa sino también acerca de sus propiedades dinámicas. Es el caso, por ejemplo, del notable trabajo de Albert Barabasi y otros acerca de la naturaleza autosimilar de la estructura de muchas de las redes de sistemas biológicos, de su particular resistencia a perturbaciones aleatorias o de su tolerancia a los errores<sup>8</sup>. Bastante menos frecuentes son, con alguna excepción de ámbito limitado<sup>13</sup>, los modelos dinámicos no lineales en el campo de la Genómica, en comparación, sobre todo, con la amplia variedad de problemas de física, química ingeniería y biomedicina donde este tipo de metodología se ha aplicado con éxito.

Barabasi señala que en tanto que la emergencia de las redes biológicas complejas es el resultado de dinámicas de autoorganización gobernadas por leyes simples de carácter genérico, comparten características prominentes que las definen<sup>9</sup>. Entre ellas están la estructura de escala libre y la organización jerárquica de los módulos funcionales. Desde esta perspectiva debe admitirse que junto al carácter intrínsecamente estocástico de muchos de los procesos dinámicos en biología, sus atractores serán con frecuencia atractores extraños (en el sentido de Prigogine) que incluyan en mayor o menor medida una componentes de caos determinista capaz de ser reconocida en la estructura de las redes complejas que los representan. Y en efecto, el carácter autosimilar de la distribución de nodos en muchas de estas redes -que siguen una ley potencial- confirmaría en efecto ese carácter determinista.

A nivel de mutación, los cambios observados en el genoma de las células tumorales son de carácter puntual y afectan a elementos clave del proceso

tumorigénico. Con frecuencia son cambios acumulativos que se producen en un conjunto muy específico de genes y cuya "lógica" puede ser establecida de modo directo porque afectan a elementos directamente implicados en los procesos de diferenciación y proliferación celular, supresión de tumores, organización del material nuclear, vulnerabilidad frente a determinados agentes o en procesos de comunicación y adherencia celular. Este conjunto singular, y relativamente reducido, de oncogenes que sufren alteraciones altamente correlacionables con el proceso de malignización celular y carcinogénesis ha permitido la definición de marcadores tumorales con un valor diagnóstico reconocido. El análisis de estos marcadores permite, en efecto, apoyar el diagnóstico de determinados tipos de cánceres y estimar su agresividad y, en definitiva, el pronóstico de la enfermedad.

Que el proceso de malignización celular va igualmente acompañado de cambios específicos en la metilación del DNA y que estos cambios son determinantes para la implantación y el desarrollo del tumor son hechos reconocidos desde hace tiempo Feinberg and Vogelstein<sup>14</sup>. No es extraño que los primeros esfuerzos fuesen orientados a la identificación de alteraciones específicas en la metilación del DNA equiparables a los marcadores tumorales anteriormente descubiertos. Sorprendentemente, y a la luz sobre todo de los numerosos estudios llevados a cabo a partir de análisis de metilación a escala genómica de DNA humano, hoy sabemos que los cambios epigenéticos que acompañan al proceso de carcinogénesis son sustancialmente diferentes. Pese a que se han identificado modificaciones específicas en la metilación de las islas CpG de promotores de determinados factores de transcripción y otros elementos directamente implicados en la implantación del tumor, lo que apuntaría hacia la existencia de unos "marcadores tumorales de metilación", estadísticamente correlacionables y potencialmente equivalentes a los marcadores mutacionales, lo cierto es que la metilación del DNA en células tumorales es cualitativa y radicalmente diferente a la de sus correspondientes homólogos celulares sanos.

De hecho las modificaciones epigenéticas observadas en las células tumorales afectan a extensas regiones del genoma y presentan además una característica adicional inesperada: las modificaciones no son únicas sino que muestran una manifiesta heterogeneidad entre los diferentes clones tumorales. Es por ello que numerosos autores hablan de una "desregulación epigenética" que acompañaría



al proceso de tumorigénesis. Más que como un proceso de desregulación, algunos autores han vinculado estos cambios a un hipotético proceso potencial de adaptación Darwiniana de los diferentes clones mediante ajuste epigenético de su reguloma<sup>15,16</sup>. Según esta perspectiva, la metilación diferencial de los diferentes clones tumorales incrementaría las posibilidades de implantación del tumor mediante un proceso intraevolutivo de selección natural de aquellos metilomas que mejor favoreciesen su desarrollo.

Los datos actuales son todavía insuficientes para poder asegurar que tal tipo de proceso tenga entidad real y mucho menos para, en caso positivo, estimar su relevancia potencial o su universalidad en el mecanismo de tumorigénesis. Entre otras cosas porque en la actualidad no disponemos aún de un modelo coherente y bien definido del papel de la regulación Epigenética del genoma ni ontogénica ni filogénicamente hablando.

Circunstancialmente, y a falta de pruebas directas, la hipotética existencia de tal tipo de mecanismos vendría avalada por dos características: *a*) Las modificaciones epigenéticas serían variables y heterogéneas y *b*) una vez desencadenado el proceso por uno o varios mecanismos (aún desconocidos), las metilaciones diferenciales deberían ser numerosas (a escala genómica) y seguir patrones sistémicos. Ambas características coinciden con los datos observados.

Tal dispositivo epigenético de optimización adaptativa, en caso de existir, conferiría al biosistema una ventaja evolutiva muy relevante: constituiría un auténtico dispositivo de plasticidad adaptativa de carácter reversible y rápido. Un sistema de optimización mediante aprendizaje (al modo de un algoritmo de "machine learning"), que *a*) debería probablemente estar sujeto a un control preciso; *b*) tendría naturaleza sistémica, de modo que las diferentes configuraciones alcanzadas podrían ser consideradas como atractores alternativos o propiedades emergentes y *c*) no podría ser identificado/interpretado fácilmente mediante las técnicas estadísticas convencionales. Para su análisis se requerirían métodos no lineales propios de la dinámica de sistemas complejos adaptativos.

El análisis de recurrencia de los atractores desplegados mediante embedding de series temporales de densidad de metilación de DNA, de densidad de pares GC y de otras posibles variables genómicas relevantes, mediante aplicación de los teoremas de Taken-Ruelle<sup>17-19</sup> y Poincaré<sup>20</sup> se nos planteó claramente, en este contexto, como una de las posibilidades a investigar y por ello buena parte

de este trabajo se ha centrado en el desarrollo de un protocolo metodológico completo que permitiese trasladar su probado potencial al campo de la genómica estructural, la filogenómica y la epigenómica de metilación de ADN.

De manera global, los resultados obtenidos indican que este planteamiento no solo es posible sino que nos ha permitido obtener descripciones altamente compactas de secuencias de ADN que retienen muchas de las características estructurales esenciales de los sistemas originales, hasta el punto de poder ser discriminados eficientemente mediante métodos de inteligencia artificial basados en algoritmos de aprendizaje automático ("machine learning"), en nuestro caso del tipo de vectores soporte ("support vector machines").

Quedan pendientes, sin duda, cuestiones relevantes, como la posibilidad de aplicar el teorema de ergodicidad<sup>21</sup> a las series temporales de DNA, sobre las que se mantiene todavía un encendido debate teórico acerca del ámbito de aplicación de los modelos empleados<sup>22,10,23</sup>, pero que, de cualquier forma, no nos ha impedido en nuestro caso llevar a cabo predicciones precisas en el terreno de la Epigenética del cáncer o la relación filogenética de comunidades de chimpancés en el África Central. También nos ha permitido acercarnos a la estructura del genoma del tomate desde una perspectiva sistémica que abre nuevas perspectivas sobre su anotación, en la que también se ha participado.

Así por ejemplo, mediante el protocolo de análisis cuantitativo de recurrencia implementado en este trabajo ha sido posible representar fragmentos de DNA de 240 megabases (cromosoma I humano completo) en términos de solo siete valores escalares. Cuando la representación corresponde a la secuencia de metilación de muestras procedentes de células sanas o cancerosas, este único vector 7-dimensional permite discriminar los patrones epigenéticos con porcentajes de acierto superiores, en ocasiones, al 98%. Cuando la serie temporal procede de distancias intergénicas, es posible representar el Genoma completo del tomate en términos de 12 vectores 7-dimensionales (84 valores), que deberían ser potencialmente suficientes para abordar una gran variedad de problemas, a medida que se vayan disponiendo de datos suficientes en el futuro.

La implementación del protocolo propuesto ha requerido abordar previamente una serie de aspectos metodológicos fundamentales. La teoría establece que para desplegar el atractor del sistema en el espacio n-dimensional adecuado, las coordenadas de los hiperpuntos se construyen a partir de la serie temporal

inicial, tomando valores sucesivos desfasados en un cierto desplazamiento ("delay") que debe ser previamente estimado. Además la propia dimensionalidad del espacio de fases debe ser también determinada de forma adecuada. La estimación adecuada de ambos parámetros no es sencilla: el algoritmo de "falso vecino más proximo" ("false nearest neighbor" o FFN) empleado para estimar la dimensión del despliegue del atractor en el espacio de fases ("embedding") depende de la elección del umbral y, por su parte, la estimación del desplazamiento requiere fijar estrategias adecuadas para la detección de mínimos. Por estas razones y porque interesaba además comprobar que el software desarrollado por nosotros se comportaba de la manera adecuada, fue necesario emplear un modelo de referencia conocido – el atractor de Roessler en nuestro caso – que se empleó como banco de pruebas de nuestro protocolo experimental. Las pruebas llevadas a cabo con este modelo nos permitieron poner a punto el método y confirmar su efectividad para desplegar el atractor a partir de series temporales de una de sus variables, verificándose que, como predice el Teorema de Takens, el atractor reconstruido retiene las características topológicas del original.

Una vez establecida de forma preliminar su validez, la herramienta desarrollada se aplicó al estudio de la deriva Epigenética que, en términos de metilación de DNA, acompaña al proceso de carcinogénesis. El estudio se hizo, además, con el objetivo de establecer la importancia del enfoque sistémico, en el sentido indicado anteriormente, como enfoque capaz de desvelar aspectos sistémicos, difícilmente identificables con las aproximaciones convencionales. Para ello se implementó un procedimiento de análisis cuantitativo de recurrencia (RQA) de los diagramas de recurrencia obtenidos a partir de los mapas de distancia de los puntos del atractor desplegado a partir de las series temporales iniciales (ver Material y Métodos). Los parámetros de RQA obtenidos fueron empleados, en la mayoría de los casos, para efectuar clasificaciones binarias mediante un algoritmo de aprendizaje automático basado en vectores soporte (SVM).

Estos resultados confirman que la compresión de la secuencia de metilación de un cromosoma humano completo en un único vector RQA 7-dimensional retiene la información necesaria para identificar de forma muy efectiva la deriva epigenética que acompaña al proceso de malignización celular, al menos en los tipos estudiados. En realidad, el protocolo es lo suficientemente sensible

como para que el simple examen visual de las proyecciones bidimensionales de los vectores de RQA nos permita ya establecer diferencias manifiestas entre células normales y células tumorales, con independencia del cáncer de que se trate. Resulta llamativo constatar la dispersión de los patrones de metilación en células tumorales cuando se comparan con los correspondientes valores de células sanas señalada anteriormente.

Como ya se ha mencionado, desde los trabajos pioneros de Prigogine<sup>24-26</sup> y otros, sabemos que los sistemas complejos adaptativos (CAS) se comportan dinámicamente como sistemas no lineales cuyos atractores finales comportan a menudo "escenarios" de estabilidad singularmente complicados ("atractores extraños"). Aunque normalmente no es posible saber cómo son en realidad estos atractores, la reconstrucción a partir de series temporales permitiría, como es nuestro caso, disponer de modelos topológicamente equivalentes, que aún retienen una información valiosa sobre el sistema original. Según lo dicho anteriormente, desde esta perspectiva se podría considerar que la metilación diferencial de los clones tumorales que aparecen durante la carcinogénesis representan, de hecho, "soluciones adaptativas" del metabolismo tumoral (cuya deriva sería aquí considerada como la dinámica de un sistema CAS) y, por tanto, como configuraciones estables de atractores sistémicos no lineales, susceptibles de ser analizados mediante nuestro protocolo experimental.

Si esta premisa fuese correcta, la deriva epigenética ("desregulación" para algunos autores) no podría ser satisfactoriamente explicada en términos de "marcadores" epigenéticos. En otras palabras, los cambios no obedecerían necesariamente a modificaciones específicas de la metilación de posiciones concretas (como sí tiende a suceder en el caso de los cambios mutacionales). Por el contrario, la "deriva metilacional" sería una consecuencia de los mecanismos intraevolutivos potenciales que operarían durante la carcinogénesis. Mecanismos que, por otra parte, no implican en absoluto que tengan que descartarse la existencia de posiciones específicas cuyos cambios en el estado de metilación son estadísticamente correlacionables con el proceso de carcinogénesis. Por el contrario, la deriva epigenética asociada al proceso de implantación del tumor sería compatible con la existencia de ciertas posiciones esenciales para alcanzar el "nicho" adaptativo (cuyos cambios en el estado de metilación fuesen por tanto invariantes) y pese a ello seguir siendo un proceso esencialmente

sistémico, difícilmente caracterizable en su totalidad por la sola presencia de estas invariencias. En otras palabras, es concebible que todas las "soluciones adaptativas" compartan ciertas invariencias pero respondan a una dinámica propia de sistemas CAS.

Aunque al comienzo de este trabajo no disponíamos aún de pruebas concluyentes acerca de la existencia de una dinámica sistémica de este tipo, dos argumentos diferentes apoyaban dicha posibilidad: por una parte el entorno ambiental del tumor durante su fase de implantación es manifiestamente hostil, por lo que las células tumorales se encuentran inicialmente bastante lejos de su óptimo adaptativo. Por otra parte, la posible implantación de mecanismos intraevolutivos de tipo Darwiniano estaría facilitada en el tumor emergente por su intrínsecamente rápida velocidad de crecimiento. En caso de ocurrir, un mecanismo de sistémico de este tipo sería, además, difícil de detectar mediante procedimientos convencionales.

Para profundizar en torno a esta cuestión fundamental se diseñaron estrategias diferentes para valorar la eficacia de las predicciones basadas en marcadores tumorales (basadas en metilación diferencial de sitios CpG) y las predicciones basadas en criterios sistémicos (markerless). En un primer grupo de experimentos se compiló una lista de sitios CpG metilados diferencialmente en células sanas y tumorales. Por otra parte se identificaron dos tipos de motivos con significación Epigenética que han sido descritos en la literatura con anterioridad. En su conjunto los tres tipos de elementos constituían más del 50% del cromosoma I. A continuación se crearon series temporales en las que se eliminaron sistemáticamente estos elementos relacionados con cáncer del material genético de partida, de modo que las time series solo contenían el material restante. Globalmente, los resultados obtenidos demostraron que la fracción de cromosoma restante retiene aún la firma epigenética que permite discriminar las células normales de las cancerosas. Una posible interpretación de los resultados obtenidos es que, en efecto, se confirma que la firma epigenética que caracteriza la malignización celular "está en el todo y en la parte". Se trataría en otras palabras de una característica emergente de carácter significativamente sistémico.

Ciertamente, otra posible interpretación es que no todos los elementos significativos en el proceso han sido identificados en la actualidad. Aun cuando

esta posibilidad es, por razones obvias, muy difícil de descartar por completo, resulta a nuestro juicio bastante más difícil de justificar. Suponiendo que la pérdida de capacidad discriminativa del clasificador empleado sea una medida del peso relativo que los supuestos elementos desconocidos tendrían sobre el total, estaríamos hablando de que aún faltarían por identificar entre un 20% Y un 40% de elementos relevantes en el proceso de carcinogénesis. Y que estos hipotéticos elementos desconocidos serían además comunes a la práctica totalidad de los 11 cánceres estudiados.

La validez atribuida en la literatura por otros autores a los marcadores de metilación se vió claramente confirmada cuando se compararon las predicciones realizadas con nuestro protocolo experimental y las llevadas a cabo con diferentes sets de símbolos relacionados con cáncer, mediante entrenamiento directo (sin reconstitución de series temporales en el espacio de fases ni RQA) del mismo algoritmo de aprendizaje, en condiciones comparables. Los resultados obtenidos para cáncer de cabeza y cuello mostraron performances muy elevadas prácticamente idénticas en ambos casos. Sorprendentemente, bastó un número relativamente reducido de marcadores para lograr un AUC de 97.4% frente a 98.6% en la predicción sin marcadores.

Por ello y para recabar más datos acerca del posible carácter sistémico de la deriva epigenética asociada a la carcinogénesis se llevó a cabo un tercer tipo de experimentos en los que las predicciones se realizaron previo entrenamiento del algoritmo de aprendizaje con muestras aleatorias de sitios de metilación no relacionados con cáncer (de tamaño reducido, comparable al número de marcadores empleados anteriormente). Aunque el estudio llevado a cabo es aún preliminar, los resultados obtenidos, sorprendentemente, continuaron siendo buenos (AUC mayores que 90%) incluso limitando el tamaño de las muestras a solo 18 pseudomarcadores. La validez de estos resultados fue confirmada empleando controles compuestos por muestras idénticas aunque con las posiciones aleatorizadas.

En su conjunto, nuestros datos sugieren que las diferencias en el patrón epigenético de las células cancerosas respecto a las normales se debe no solo a la existencia de unos marcadores de metilación bien definidos, cuya correlación con el proceso de malignización celular, ya establecida en la literatura, ha sido confirmada por nuestras observaciones con un procedimiento alternativo basado

en un algoritmo de SVM, sino que también tienen un carácter sistémico que se potencialmente se extiende a toda la secuencia de DNA y que puede ponerse de manifiesto con un número muy reducido de posiciones potencialmente metilables.

Asumiendo que las diferencias de valores beta se extienden a toda la secuencia, pareció interesante investigar si la simple estimación total del grado de metilación (suma total de valores beta) sería suficiente para discriminar entre células controles y tumorales, pese a que no tenemos constancia de ningún estudio previo que apoye tal suposición. Por ello se diseñaron una serie de experimentos destinados a comparar las distribuciones de sumas beta en muestras de células tumorales frente a sanas, bajo diferentes condiciones, empleando como criterio de discriminación los valores de  $p$  obtenidos mediante el test no paramétrico de Wilcoxon para suma de rangos con corrección.

Hemos encontrado que cuando se incluyen en el análisis todos los sitios CpG nuestros resultados indican por una parte que los valores de sumas beta son bastante parecidos entre células normales y cancerosas, por lo que la discriminación es generalmente difícil tanto si los resultados se estiman en términos de promedios de sumas beta, como si se interpreta en términos de  $p$  de test de Wilcoxon o se emplea un algoritmo de clasificación basado en SVM. Por otra parte, cuando se incluyen todos los sitios CpG se constata que cada tumor ofrece una respuesta diferente: en tanto que en el caso del carcinoma hepatocelular (LIHC) la discriminación es relativamente buena con cualquiera de los estimadores empleados, el carcinoma tiroideo papilar (THCA) no puede ser discriminado en ningún caso a partir de los valores de suma beta.

Más interesante fue el comportamiento heterogéneo de los diferentes tumores en relación con la posibilidad de ser discriminados en términos de suma beta, cuando las muestras se restringen a los elementos relacionados con la deriva Epigenética del cáncer. Así por ejemplo, los símbolos de genes relacionados al cancer (CRGS), compilados ad hoc en este estudio, permiten una discriminación relativamente buena en el caso de cáncer de mama (BRCA), colon and adenocarcinoma rectal (COAD), LIHC y en menor medida adenocarcinoma de pulmón (LUAD) y carcinoma renal papilar (KIRP), en tanto que el carcinoma pulmonar de células escamosas (LUSC), THCA y carcinoma endometrial (UCEC) se muestran refractarios con esta muestra.

Cuando se consideran las sumas beta de los dominios hipometilados BLOCKS, la capacidad de discriminación entre células normales y tumorales aumentó en casi todos los casos, aunque fue insuficiente para discriminar dos de ellos: PRAD y THCA.

Finalmente, cuando las muestras corresponden a los dominios hipermetilados cDMR, todos los tumores pueden ser discriminados en términos de suma beta con una elevada significación estadística, con la excepción de THCA. Si las muestras se construyen a partir de regiones hiper- e hipometiladas los resultados obtenidos son, como cabía suponer, notablemente peores, ya que los valores globales de suma beta se componen de elementos positivos y negativos que se cancelan mutuamente.

Teniendo en cuenta la heterogeneidad de muestras y procedimientos, los resultados obtenidos mostraron una coherencia bastante razonable. En su conjunto, estos resultados dibujan un escenario en el que se confirma claramente que la deriva Epigenética que acompaña al proceso de carcinogénesis en todos los tumores estudiados gravita sobre las regiones hipermetiladas cDMR, que resultan fundamentales para la caracterización de las células tumorales en todos los casos excepto en THCA y extensas regiones hipometiladas, cuya aportación al perfil epigenético de los diferentes cánceres es variable y podría estar ausente en dos de ellos, PRAD y THCA. Claramente los "símbolos" o "motivos" (CRGS) resultan más específicos y tendrían poco peso en la definición del perfil epigenético de LUSC, THCA y UCEC.

En tanto que LIHC puede ser prácticamente discriminado en cualquiera de las condiciones del ensayo y THCA es difícilmente discriminable en casi todas las condiciones, ambos tipos de tumor podrían representar los dos extremos en relación a su carácter sistémico. Así, el perfil epigenético de LIHC afectaría a una gran parte del cromosoma, indicando que la deriva es en este caso muy acusada o que tiene una importante componente sistémica. THCA representaría el otro extremo: los cambios epigenéticos que acompañan el proceso de génesis tumoral serían mínimos en este caso, lo que lo alejaría del modelo sistémico, y tendrían un carácter bastante más específico o no sufrirían una deriva significativa, lo que, en cualquier caso, indicaría que las células malignizadas se encontrarían mucho más cerca de su nicho adaptativo óptimo, ya desde el principio. Quedaría por tanto justificado el hecho de que THCA es también el único



de los tumores estudiados frente al que nuestro análisis sistémico obtuvo malos resultados.

Otro conjunto de experimentos incluidos en esta memoria estuvieron destinados a la posible aplicación de nuestro protocolo experimental al análisis estructural de genomas. El estudio se llevó a cabo sobre el genoma completo del tomate, dado que una parte del desarrollo de esta memoria ha sido realizado por mi dentro del "Plant Computational Biology"-group (PCB) del Max-Planck-Institute for Plant Breeding Research, como parte del International Tomato Annotation Group (ITAG) y del Tomato Genome Consortium, contribuyendo al GO anotación funcional del genoma del tomate (*Solanum lycopersicum*). A partir de los 19662 genes anotados (57% del total de genes codificantes) y de la secuencia completa del genoma, se intentó analizar la posible existencia de correlaciones significativas entre distancia física intergénica y distancia funcional a partir de las tres ontologías de genes (GO): procesos biológicos (BP), función molecular (MF) and componente celular (CC). El segundo objetivo de este experimento era el de obtener los doce vectores RQA que representan en nuestro modelo sistémico al genoma completo del tomate, en términos de distancias intergénicas.

Nuestros resultados apuntan a que no existe una correlación obvia entre distancia física y distancia funcional a partir de ninguna de las GO empleadas. Las tendencias de los perfiles obtenidos se explican en todos los casos a partir de las distribuciones observadas entre las distancias intergénicas. Con los datos actuales no es posible, de todas formas, completar este análisis con el nivel resolutivo que sería necesario para llevar a cabo RQA a partir de las distancias funcionales. En tanto que la metodología está disponible, la disponibilidad de nuevos datos permitirá avanzar en esta dirección en el futuro.

En cuanto a los resultados de RQA de los doce cromosomas del tomate, nuestros datos revelan RPs muy diferentes a los obtenidos a partir del epigenoma del cromosoma I humano. Algunos de los valores encontrados indicarían que se trata de un sistema con una baja predictibilidad, contrariamente a lo que sugerían los RPs procedentes de genoma humano. Puesto que el alcance de nuestros datos es aún muy limitado, también en este caso será preciso caracterizar un mayor número de genomas para poder interpretar correctamente el significado de estas diferencias.

El último bloque de experimentos incluidos en este estudio tuvo como principal motivación explorar la utilidad potencial de la metodología propuesta para analizar procesos de divergencia adaptativa entre organismos próximos. En este caso, nuestro estudio se centró en la discriminación de genomas mitocondriales porque el número de organismos secuenciados es notablemente mayor que el de genomas nucleares: actualmente 8753 genomas completos. Otra diferencia básica con los experimentos anteriores es que en este caso el análisis se realizó sobre series temporales de densidad de pares CGs. Los resultados de estos experimentos nos han permitido, en el primer caso, predecir tres posibles migraciones de tres subespecies del chimpancé común y en el segundo caso, clasificar perfectamente cinco especies de la superfamilia/suborden Caniformia.



# Introduction

## 2.1 Complex systems

According to the reductionist hypothesis any matter and process, for instance, – me – thinking of and writing this text, storing it on the hard disk or printing it and of course – you – reading it, is in the end controlled by the same fundamental laws acting on some elementary particles. But, from the bottom up, predicting the next word I am going to write based on observations of some elementary particles located in my brain would break down already on quantum mechanic level. In 1972, the later Nobel laureate, Anderson<sup>27</sup> was one of the first motivating for complex systems science. He mentioned that if we order science disciplines hierarchically (elementary particle physics → many body physics → chemistry → molecular biology → cell biology → ... → psychology → social sciences), at each level of complexity new properties appear and new own laws and concepts to describe them are necessary.

Now, more then 40 years later, "Complex system" is still one of those terms used in the scientific and philosophical literature which lack a precise definition. Nevertheless, a minimal consensus, to which probably most researchers would agree may be stated as follows:

**Definition 1.** *A complex system is build of interacting parts or agents that show emergent behavior which can not be trivially deduced from the behavior of the individual agents<sup>28</sup>.*

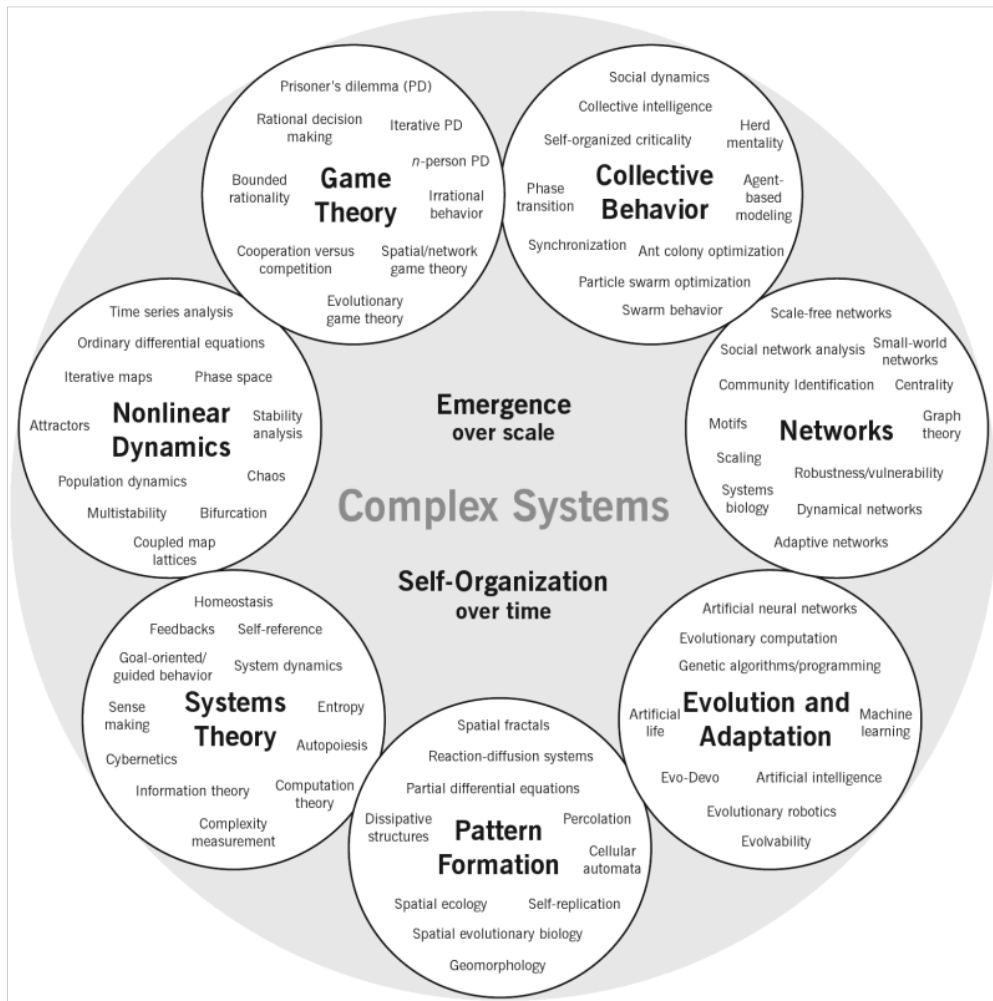


Figure 2.1: Research fields associated with complexity science. Created by Hiroki Sayama, D.Sc., Collective Dynamics of Complex Systems (CoCo) Research Group at Binghamton University, State University of New York, 26 November 2010, Creative Commons Attribution-ShareAlike 3.0 Unported License (CC-BY-SA 3.0), [link to the license](#), cf. <sup>30</sup>.

There are many examples of complex systems including for instance ecosystems, financial markets, the brain, the immune system, insect colonies, flocking or schooling behavior in birds or fish, human societies, the cosmos itself<sup>28</sup> and cancer<sup>15,29</sup>. Complex systems are also subject of numerous research fields (Figure 2.1).

### 2.1.1 Features of complex systems

To define complex systems it is fundamental to figure out whether complexity is a single natural phenomenon found to be present in various different systems, which could be the subject of a single theory, or is it domain-specific, and no common features and laws can be identified<sup>31</sup>. Ladyman et al.<sup>31</sup> reviewed various attempts to characterize complex systems and they compiled some features widely associated with this topic.

#### Nonlinearity

Homogeneity and superposition are together necessary and sufficient conditions for linear systems:

**Definition 2.** *A system is called linear if and only if it possesses both homogeneity and superposition properties. That is, if  $x_1(t) \rightarrow y_1(t)$  and  $x_2(t) \rightarrow y_2(t)$  and for any real numbers  $k_1$  and  $k_2$ , the relationship*

$$\{k_1x_1(t) + k_2x_2(t)\} \rightarrow \{k_1y_1(t) + k_2y_2(t)\} \quad (2.1)$$

*is true, then the system is linear<sup>32</sup> ...*

...otherwise, it is nonlinear. Complex systems are often said to be nonlinear. Rickles et al.<sup>33</sup>, for instance, state: "A necessary condition, owing to nonlinearity, of both chaos and complexity is sensitivity to initial conditions." In contrast, others<sup>31,34</sup> argue that there are examples of complex systems subject to game-theoretic and quantum dynamics which obey linear dynamics and therefore nonlinearity is not a necessary condition. It is also not a sufficient condition because there are nonlinear systems which are not complex (e.g. a single chaotic pendulum). Unfortunately, the authors do not provide concrete examples neither citations which make it difficult to follow their arguments. Moreover, Xiao-Feng and Yuan-Ping<sup>35</sup> state that the linearity of quantum mechanics (which is mathematically described by linear quantum dynamics, cf<sup>36</sup>) limits their application and it can not be used to study complex systems. Whether nonlinearity is a necessary condition for complex systems or not, these contradictions do not restrict to argue that nonlinearity is at least important for some complex systems.

## Feedback

Feedback is a necessary condition for complex systems<sup>31</sup>. Simplified, feedback means that the output of some process becomes an input to another. In complex systems feedback occurs between levels of organization in the way that lower level interactions between agents generate some pattern in a higher level which back-react causing new patterns on the lower level, and so on. Feedback can be either positive or negative. A positive feedback enhances or amplifies an effect or variable, a negative one reduces it<sup>33</sup>.

An example for an effective and complex feedback-control mechanism are tumor cells that turn on the expression of the multi-drug-resistance 1 (MDR1) gene. This gene encodes the P-glycoprotein, an ATP-dependent efflux pump, which exports drugs out of the cells and thereby giving rise to multi drug resistance<sup>29,37</sup>.

Feedback is not a sufficient condition for complexity. Cruise controls, as used in modern cars, consist basically of four components: a sensor which measures the vehicle speed, a cruise control unit, the throttle valve and a throttle position sensor. In fact, the control unit receives feedback from the sensors and operates appropriate on the throttle valve, but the system is not complex.

## Spontaneous order

For complex systems it is a necessary condition that they exhibit "some kind of spontaneous order"<sup>31</sup>. I use quotation marks here, because it's a citation, but also to point out that the notion of order is not necessary clear. However, according to the authors, it should be related to symmetry, organization, periodicity, determinism and pattern. Whatsoever order exactly is, pure randomness and total order is incompatible with complexity. Spontaneous order implies that disorder is also a necessary condition for complex systems. If not, from where does spontaneous order emerge?

The concept of spontaneous order seems quite strange and is not always comprehensible for everyone. In political and economical discourses you may even find derogative comments like this one, where Damon Linker, a senior correspondent at TheWeek.com, abused Friedrich August von Hayek (a pioneer of the concept of spontaneous order): "[...] the idea of spontaneous order

might be the silliest and most harmful of all [...]”<sup>38,39</sup>.

### Robustness and lack of central control

Flocks of birds like greylag geese (*Anser anser*) or common starlings (*Sturnus vulgaris*) are beautiful and fascinating examples for complex systems and their robustness. Robustness is a central and also necessary condition for complexity. The order, in our example the typical flock form, which emerges spontaneously from the interactions and feedback of neighbored birds is stable under perturbations like gusts of wind, erratic motions or random elimination of some birds<sup>31</sup>. Furthermore, the lack of central control is fundamental to the robustness of complex systems. It is not possible to knock out leading birds and break down the flock formation, simply because they do not exist. Lack of central control is another necessary condition for complexity.

### Emergence

In the first paragraphs of this introduction it has been already mentioned that emergence is a necessary condition of complexity. Emergence is a quiet mystic notion. It seems, that one characteristic of emergence is a 'downwards causation', where the emerged properties have an effect on the lower levels of the system. Emergence is either purely epistemological; or it is ontological, in which case we could not understand it<sup>31</sup>. Although, emergence is difficult to grasp – its there. Just think about the following words from Jochen Fromm<sup>40</sup>:

- one water molecule is not fluid
- one gold atom is not metallic
- one neuron is not conscious
- one amino acid is not alive

## 2.1.2 Mathematical background

As shown in figure 2.1 there are many ways to make complex systems mathematically accessible. In live science the most known approach is probably



through networks<sup>41,42</sup>. In this work we have focused on nonlinear dynamics, whose mathematical background is given in detail in chapter 4 "Methods".

Just one word more before closing this section. The term "attractor" is used from time to time in this work. Figure 2.2 gives an illustrative explanation of its meaning.<sup>43</sup>

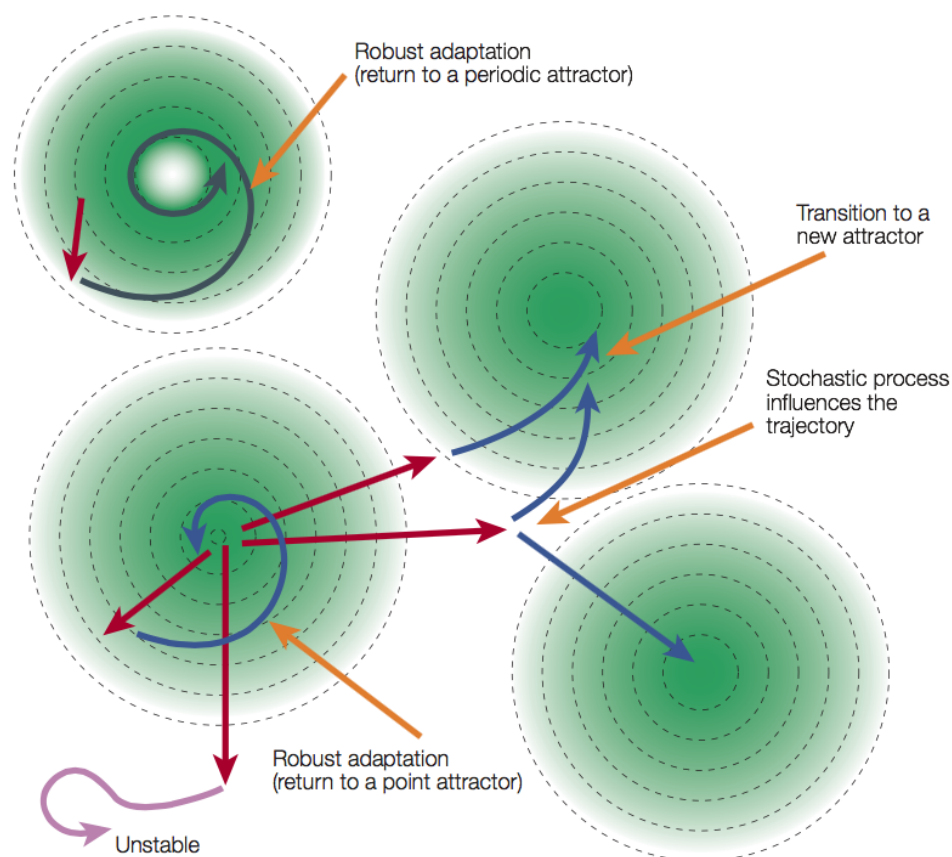


Figure 2.2: Robust reactions of the system: to stay or to change. The state of a system can be shown as a point in the state space. In this case, the state space is simplified into two dimensions. Perturbations forcefully move the point representing the system's state. The state of the system might return to its original attractor by adapting to perturbations, often using a negative feedback loop. Bacterial chemotaxis is an example. There are basins of attractions in the state space within which the state of the system moves back to that attractor. If the boundary is exceeded, the system might move into an unstable region or move to other attractors. Positive feedback can either move the system's state away from the current attractor, or push the system towards a new state. The cell cycle involves a combination of positive and negative feedbacks that facilitate transition between two attractors (G1 and S/G2/M) creating a bistable system. Often, stochastic processes affect transition between attractors, as seen in  $\lambda$ -phage fate decision, but maintenance of a new state has to be robust against minor perturbations. Reprinted by permission from Macmillan Publishers Ltd: *Nature Reviews Genetics* 5, 826-837, copyright (November 2004)

## 2.2 Cancer

### 2.2.1 Clonal evolution in cancer

The first impulsions to the hypothesis that neoplasms develop as a clone from a single cell of origin and that the neoplastic progression appears to be driven by sub-clonal selection have arisen in the fifties of the last century. More than 20 years later, in 1976, Nowell<sup>16</sup> established the evolutionary theory of cancer. Within this concept cancer clones are interpreted as asexual unicellular organisms, the clonal selection is equivalent to the Darwinian natural selection and the affected tissue becomes the ecosystem where a kind of micro-evolution takes place. In addition, cancer has been validated as a complex adaptive system<sup>44-46</sup>.

In terms of clonal dynamics, cancer cell doubling time (approximately 1-2 days) is much faster than tumor doubling time (60-200 days) which means that most tumor cells die in the competition for space and resources<sup>47,48</sup>. But, on the one hand, it means also that "survivors" are each time better adapted and more resistant. Cancer treatment, which is a sort of artificially induced extreme stress, does not inhibit the evolutionary process, but "provides a selective pressure for the proliferation of variant cells that resist the treatment"<sup>44</sup>. If a treatment is not 100% successful it can, (in my opinion) by analogy with antibiotic resistance, result "in cells with improved fitness and malignant potential"<sup>44</sup>. As the tissue ecosystems are open systems, the selective pressure applies in a similar way to the exposure of genotoxines, such as cigarette carcinogens or ultraviolet light, infection, hormone or inflammatory levels and other stress factors. On the other hand, mathematical modeling<sup>49</sup> has shown that more robust and malignant phenotypes of cancer clones are less likely in stable or homogeneous micro-environments.

Cancer stem cells exhibit a mandatory trait of self-renewal. Apart from that, any phenotypic feature that allows the cells to continue to survive and proliferate within its micro-environment can lead to almost infinite evolutionary trajectories which may end in extreme sub-clonal (epi)genetic heterogeneity and unique genomic profiles<sup>50-53</sup>.

## 2.2.2 Epigenetics in cancer

Classical genetics and genomics alone can neither explain embryogenesis nor phenotypical differences in monozygotic twins or cloned animals, but this gap of knowledge can be filled by epigenetics<sup>54</sup>. Epigenetics refers to mechanisms that initiate, in response to environmental stimuli (in a wider sense), and maintain heritable patterns of gene function and regulation without changing the genomic sequence. The underlying epigenetical mechanisms which can rise to different phenotypes are one or a combination of the following: DNA methylation, post-translational modifications of histone proteins, chromatin remodeling, and non-coding RNAs (for instance miRNA)<sup>55</sup>. Epigenetics has also an important impact in disease development, especially cancer. Figure 2.3 depicts briefly epigenomic aberrations during oncogenesis. In this work we focus on DNA methylation which is the only known epigenetic modification of the DNA<sup>56</sup> and is found in approximately 70-80% of CpG dinucleotides in adult mammalian somatic cells<sup>57,58</sup>. Non-CpG methylation is prevalent in embryonic stem cells<sup>59,60,57</sup> and has also been observed in neural development<sup>61</sup>.

**Definition 3.** *DNA methylation is the addition of a methyl group to DNA at the 5-carbon of the cytosine pyrimidine ring.*

DNA Methylation plays a key role in the control of gene expression and genomic stability in cells. It regulates important biological processes, such as embryonic development, genomic imprinting, X chromosome inactivation and carcinogenesis<sup>56,62</sup>.

In contrast to normal cells, cancer cells suffer drastic aberrations in DNA methylation which can be of either directions hyper- or hypomethylation. It has been widely accepted that hypermethylations are often observed in promoter CpG islands (CGIs) and also in non-promotor CGI shores whereas hypomethylation occurs mainly genome-wide in gene-poor areas, repetitive elements, retrotransposons and introns and can lead to genomic instability<sup>55,54,63</sup>.

Newer investigations based on bisulfite sequencing, however, have revealed additional cancer-specific differentially methylated regions (cDMRs) with increased stochastic variation in CpGs mainly far from islands and shores and large blocks of contiguous hypomethylation affecting more than half of the genome. Due to their results, the authors suggest in future efforts to DNA

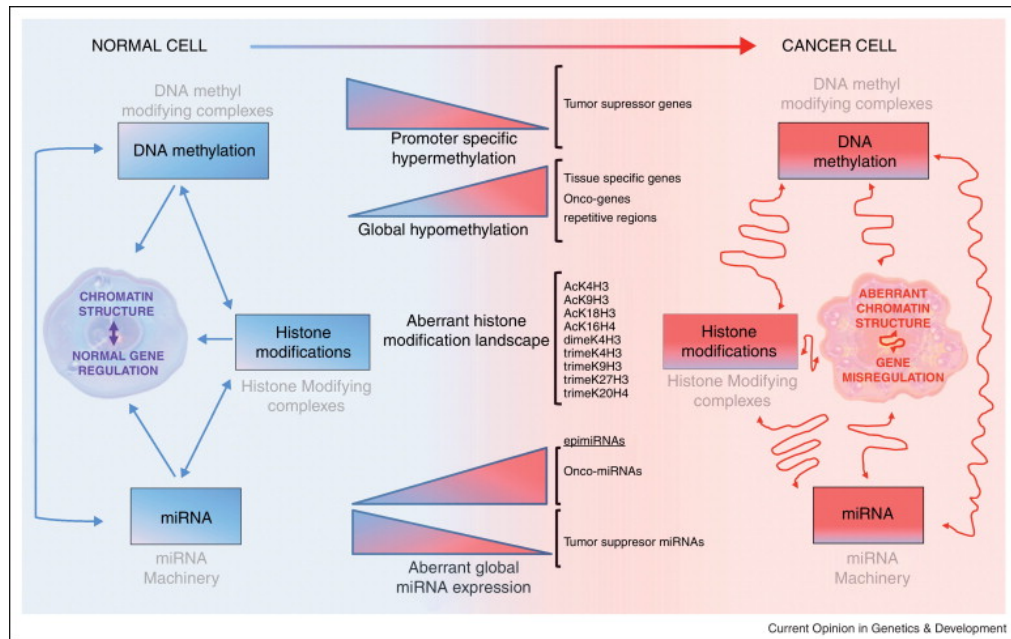


Figure 2.3: Global depiction of epigenomic alterations during oncogenesis. In conjunction with accumulation of genetic lesions, there is an aberrant pattern for the different epigenetic effectors: DNA methylation, histone modifications, and miRNAs. In normal cells, the interplay between the epigenetic factors and the chromatin structure leads to a tuned gene regulation. However, in cancer cells tumor suppressor genes promoters become hypermethylated and with an altered global pattern of histone modifications resulting in aberrant gene silencing. Moreover, global hypomethylation leads to chromosome instability and fragility. Epigenetic changes, including DNA methylation and histone modifications are responsible for abnormal mRNA and miRNA expression producing altered activation of oncogenes and silencing of tumor suppressor genes. Reprinted by permission from Macmillan Publishers Ltd: *Current Opinion in Genetics & Development* Volume 22, Issue 1, 50-55, copyright (February 2012)<sup>55</sup>

methylation for cancer screening, not only to focus on cancer specific profiles, but more at defining the cancer-specific DNA methylation "as the departure from a narrowly defined normal profile"<sup>64</sup>.

Much has been written and speculated about the underlying mechanisms which could explain the high variability of cancer DNA signatures, but – to my knowledge – no satisfactory standard model has been approved yet. In acute myeloid leukemia as an example<sup>65</sup>, mutations in epigenetic regulator proteins, including isocitrate dehydrogenases (IDH1/2), methylcytosine dioxygenases (TET2) and DNA methyltransferases (DNMT3A), have been described<sup>66–68</sup> and may be interpreted as drivers of aberrant DNA methylations. Some functional studies have also linked these proteins to distinct DNA methylation phenotypes<sup>69–73</sup>.

Another substantial question is whether and how can cancer development be linked to aging. Recent research has shown the relation between aging and DNA methylation alterations<sup>74–77</sup>. Consistent, yet tissue-specific changes in DNA methylation have been reported to come along with age<sup>78,79</sup> and result into a diverging epigenomic landscape.

This phenomenon is often referred to as 'epigenetic drift'<sup>65</sup>. Coherent explanation for the 'epigenetic drift', whether based on the assumption that it results from stochastic events or environmental factors, remain to be disputed<sup>78,80</sup>. For both hypotheses arguments can be found. On the one hand, it has been shown in mice that environmental factors can indeed lead to epigenetic deregulation<sup>80–82</sup>. On the other hand, a notion of increasing entropy of DNA methylation along with age has been proposed<sup>78,79</sup>.

A study on Illumina Infinium HumanMethylation450 BeadChip © based DNA methylation pattern across whole blood samples from 656 individuals (19 to 101 years old) has revealed more than 70000 CpG sites associated with the age-depend 'epigenetic drift'. These sites have been used to create a linear model which was able to predict the age of an individual with 96% accuracy<sup>78</sup>. Different tissues and gender have been fitted to this model. Considering this data it seems evident that the 'epigenetic drift' develop genome-wide and across tissues. Characteristic for the 'epigenetic drift' is that the high variability in DNA methylation or increase in methylation entropy can be found between individuals as well as neighboring CpG sites. These results are fairly similar to

the stochastic variation reported by Hansen et al.<sup>64</sup> (see two paragraphs above) and allow to conclude that cancerous tissues may be considered as prematurely aged<sup>65</sup>.

## 2.3 Guanine-cytosine content

The base composition, in particular the guanine-cytosine content (GC content), is a fundamental genomic property and a standard measure in genome projects. In absence of whole genome DNA sequences the base composition can be estimated, for instance, biochemically by DNA temperature melting analysis (TMA). This method takes into account that the double hydrogen bond connecting adenine with thymine is weaker than the triple hydrogen bond which binds guanine and cytosine. Thus, the difference in GC content results in different melting temperatures<sup>83</sup>. Once the melting temperature is known, the GC content can be calculated as follows:  $\%GC = 2.44(T_m - 81.5 - 16.6 \log[Na^+])$ , where  $\%GC$  is the GC content,  $T_m$  is the melting temperature and  $[Na^+]$  is the concentration of sodium ions<sup>84</sup>. Easier, faster and cheaper is the measurements of the GC content via flow cytometry (FCM). This method is based on synchronous measurements of a sample and a control, each with two different dyes. One dye measures the total DNA content and the other, base specific one, is used to calculate the portion of adenine and thymine, for instance. Once a portion is known, the others can be calculated easily:  $\%C = \%G$ ;  $\%A = \%T$ ;  $\%G + \%C = 100 - (\%A + \%T)$ , where  $\%G$ ,  $\%C$ ,  $\%A$ ,  $\%T$  are the portions of guanine, cytosine, adenine and thymine respectively Šmarda et al.<sup>85</sup>. If the complete genome sequence is available, then the GC content is simply calculated by the formula  $\%GC = (G + C)/(A + T + G + C)$ , where  $\%GC$  is the GC-content and  $A, T, G$  and  $C$  are the counts of adenine, thymine, guanine and cytosine, respectively. For complete and high quality genome sequences this is the most accurate method to obtain the GC-content.

The GC-content in genome sequences of prokaryotes, eukaryotes and organelles shows a wide spectra and is often highly variable. Processes which might in combination influence the base composition of a genome are, inter alia, mutation, recombination, random genetic drift and selection<sup>86,87</sup>. It's evident to guess that this high variability might be connected to some function

and it is therefore not surprising that already in the fifties of the last century its relation to phylogeny has been predicted<sup>88</sup>.

### 2.3.1 GC-content in prokaryotes

Mann and Chen<sup>89</sup> state in their review that the GC-content in bacteria genomes range between 20% and 75%. Recently, even values between 17% and 75% have been reported<sup>90</sup>. This high variability of the GC-content is discussed to form part of a response to environmental adaptation, whereat two trends can be observed. On the one hand, the GC-content is correlated to environmental niches and lifestyle. On the other hand, bacteria are able to apply incongruence in GC-content to delineate horizontally transferred genetic elements<sup>89</sup>. Two major processes have arisen from long debates to explain the extreme variability of the GC-content in bacteria. The mutual hypothesis propose that the GC-content is driven by genome-specific mutational biases, whereas the selectionist hypothesis based on selective processes in different organisms. The latter implies also codon position specific GC-content variation pattern<sup>90</sup>. This hypothesis will be resumed below. Free-living bacteria tend to higher GC-content and larger genome sizes, as a result to more complex and highly alterable environments (also reviewed by Bentley and Parkhill<sup>91</sup>). Whereas, parasites and endosymbionts occupying poor or limited environments show smaller genome sizes and higher AT-content, which may have been induced by translesion repair mechanism, phage insertion or cytosine degradation. The synthesis of GTP and CTP requires more energy, therefore the mutational bias towards a higher AT-content may confer a selective advantage. Higher AT-content impact also the length of coding sequences<sup>92</sup> because the probability towards stop codons increases and consequently the length of the coding sequences decreases.

Three major processes, transduction, transformation and conjugation, contribute to lateral DNA transfer of free-living organisms and impact their genomic GC-content<sup>93,94</sup>. Generalized transduction is an integration of non-specific DNA fragments from a donor organism via a bacteriophage. In contrast, specialized transduction means the incorporation of phage specific genetic material. Transformation is the process where the incorporation of exogenous DNA into the cell from its environment is directly taken up through the cell membrane. Con-



jugation implies direct contact between cells where the transmission of genetic material is typically assist by plasmids.

The integrated external DNA may show abrupt differential GC-content respectively to the background base composition of the host genome. This characteristic has been used as a indicator of non-self genomic content to detect pathogenicity islands (PAIs) in a host genome<sup>95</sup>. But GC-content is not sufficient to identify PAIs accurately<sup>96</sup>. With the age of the insert this effect obfuscate.

A relationship between GC-content and optimal growth temperature has been debated<sup>97-99</sup> and has not been sufficiently verified. In contrast, an increased GC-content for aerobes has been shown by Naya et al.<sup>100</sup>.

### 2.3.2 GC-content in mitochondria

Analyzing the base composition of complete mitochondrial and plastid DNA sequences shows a strong bias towards a high adenine and thymine content<sup>101,102</sup>. This phenomenon remains poorly understood. However, recently it was also shown that GC-rich organelles exist. Figure 2.4 shows nicely the wide spectra of base compositions in organelles. "The origins of AT richness within mtDNAs and ptDNAs are thought to reflect the endosymbiotic history of these genomes, their location within the cell, the unique population-genetic features that define organelles, and selection for metabolic and translational efficiency"<sup>103</sup>. Reasons that explain the persisting bias towards AT-richness might be that organelle DNAs inhabit a highly mutagenic environment where high concentrations of reactive oxygen species benefits GC  $\rightarrow$  AT mutations<sup>104</sup>.

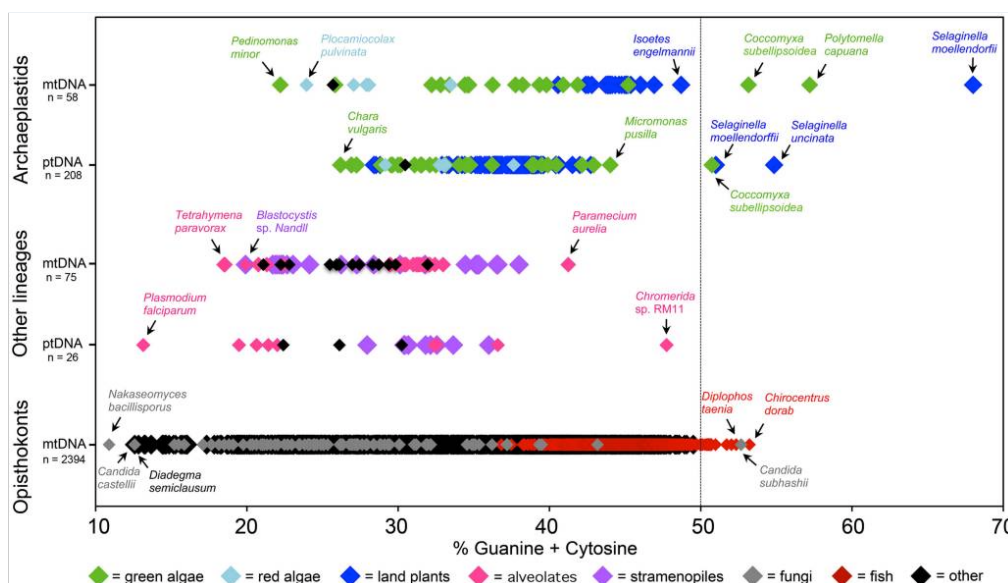


Figure 2.4: Nucleotide composition continuum of completely sequenced mitochondrial DNA (mtDNA) and plastid DNA (ptDNA) sequences. Most of the complete organelle genome sequences deposited in GenBank have a GC content below 50%, with the exception of those from certain green algae, lycophytes, fish, and fungi. The number of genome sequences (n) within each group is shown beside the y-axis. Mitochondrial and plastid genome sequences were downloaded from GenBank on January 1, 2012. This figure is taken from *Smith*<sup>103</sup> respecting the CC-BY-SA 3.0 license.

## 2.4 Support Vector Machines

Support vector machines (SVM) are outstanding machine learning algorithms that have been successfully applied to a wide variety of problems in very different fields. Together with neural networks, they are referred, sometimes, like “black box” algorithms, indicating that their performances are founded on rather hidden relationships among the income data that have to be discovered along the training process and that have no obvious relationships with the final model outcomes.

SVM have recently gained considerable popularity mostly due to its very success in mining, classification, regression and other complex prospective analysis. A second reason of its favor among massive data analyzers is that many good libraries in different programming environment have implemented excellent (and

fast) algorithms that make reasonably easy working with SVM models, despite that procedures to identify the support vectors relies on fairly technical vector geometry handling and some other tricky math difficult to interpret for non specialists. Some good introductions to SVM can be found in <sup>105–107</sup>.

Conceptually, SVM algorithms are nonlinear generalizations of the generalized portrait algorithm developed in Russia in the sixties <sup>108,109</sup>. In this context, they come from the so called statistical learning theory (SLT), later developed by Vapnik and others <sup>110–112</sup>. In one paragraph, SLT theory tries to characterize properties of learning machines which enable them to generalize models built by training with known data, to unseen data.

Traditionally, SVM have been used in binary classification problems. They can be used, however, in almost any numeric machine learning scenery, including regression and a variety of prediction, mining and pattern recognition models. They have been widely used in bioinformatics and biomedicine as well as in many other areas of experimental sciences and engineering. Classical works where SVM have largely demonstrated their potential includes, for example, languages and text categorization and analysis, speech recognition, industrial failure analysis, security breaches, intelligent search algorithms, terrorism prevention or earthquakes prediction.

Essentially, a support vector machines algorithm works on an  $n$ -dimensional space where points are the features vectors of each sample data (“training” data). An SVM can be imagined as a surface model that defines a boundary between two sets of data that share a binary condition. The target of an SVM is finding hyperplanes which leads to fairly homogeneous partitions of data on either side on the base of that condition (see figure 2.5).

### 2.4.1 Linearly separable binary classification

When performing binary classification and at least one hyperplane exists able to completely separate all the elements on the basis of the binary condition, we refer the sample as to be linearly separable. The main goal of the algorithm, in this straightforward case, is to find the optimum hyperplane among all the matching possibilities. To achieve that, SVM algorithms search for the maximum margin hyperplane (MMH) that creates the greatest separation between

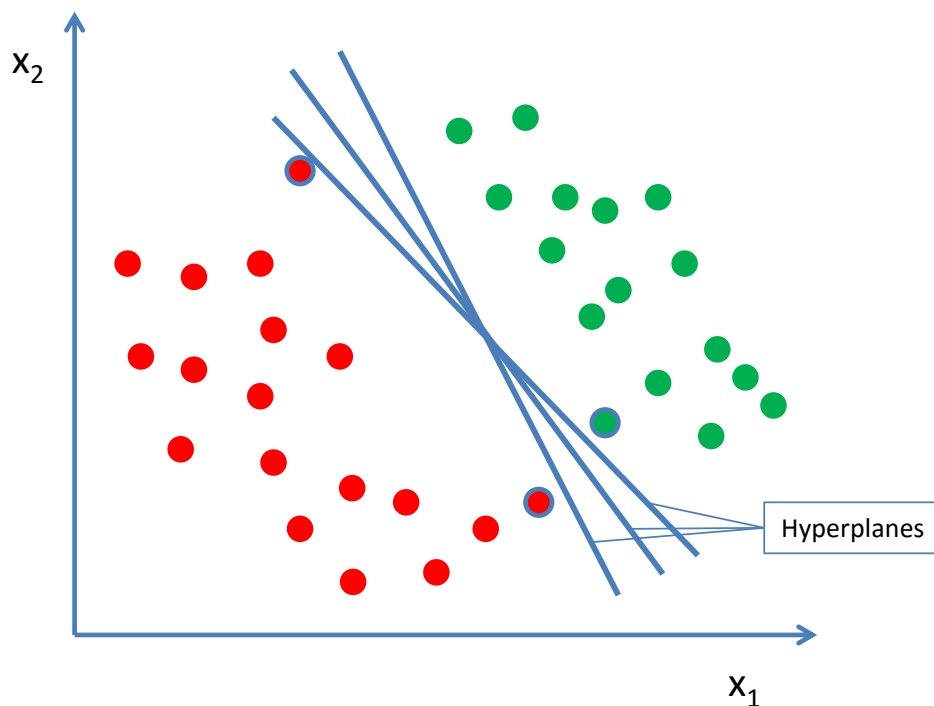


Figure 2.5: Support vector machines: Finding hyperplanes.

the two classes. (see figure 2.5, where three possible planes have been drawn, but where it can be supposed that the ones leading to the greatest separation will generalize the best to future data). The support vectors (shown in figure 2.6 as dashed lines) are those related to the MMH closest points from each condition. Each class (condition) may have one or more support vector and, reciprocally, support vectors can be used to calculate the maximum margin hyperplane in a very efficient way, even if the number of features is extremely large.

One way of calculating the MMH is by previous finding of what is called the “convex hull” of each binary condition (defined as the closed polyline drawn from the outer boundaries of the two groups of data points. The MMH can be, then, calculated as the perpendicular bisector of the shortest line between the two convex hulls (see figure 2.7). This is usually a complex calculation that usually has to be carried out by sophisticated quadratic optimization algorithms (See details in Method). This step can be processor intensive. However, there

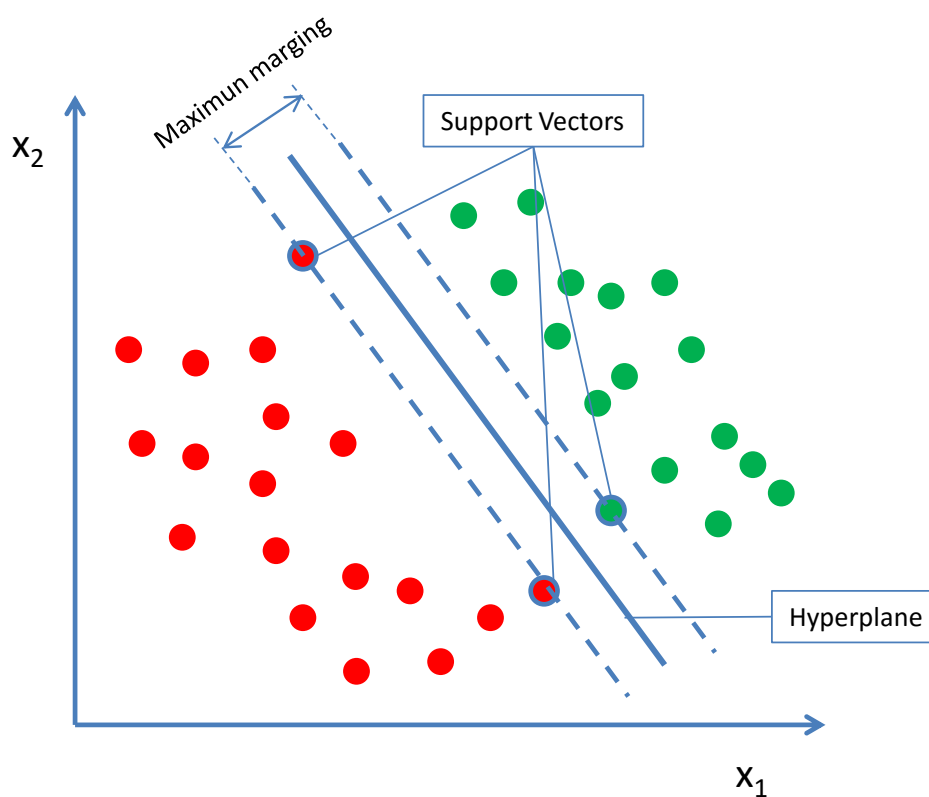


Figure 2.6: Support vector machines: Search for the maximum margin hyperplane.

are a number of very efficient algorithms able to quickly achieve the solution even starting with very large training datasets.

### 2.4.2 Using soft hyperplanes for binary classification with not-linearly separable data

Frequently, the full splitting of binary classes in two groups linearly divided by a hyperplane is not possible as some data points can fall on the wrong side of margins. To challenge this more than usual contingency, SVM algorithms can define “slack variables”, able of creating soft- margin hyperplanes in which some point are allowed to fall on the incorrect side of the margin with a “cost” that can be conveniently adjusted by the model. All the point that violates the

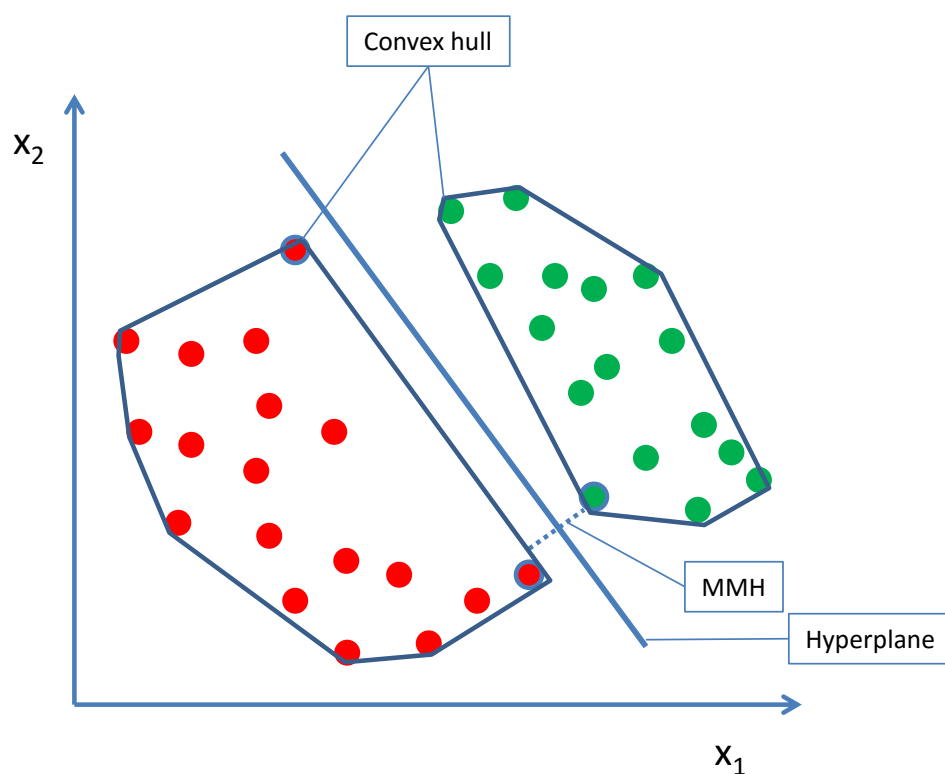


Figure 2.7: Support vector machines: Finding the “convex hull”.

constraints are penalized with this cost value, so that rather than finding the maximum margin, the algorithm will try to minimize an overall cost function (further details can be found in Method). High costs will restrict tolerance to achieve full separation, determining harder boundaries. On the opposite, low values of the cost parameter will give priority to a wider, soft overall margin. To achieve the appropriate generalization capacity, it will be usually important to ensure the proper balance between these two priorities when training the SVM model.

### 2.4.3 Dimensional scaling-up with kernels to perform non-linear classification

Using a slack variable to define soft hyperplanes is not the only way to approach the problem of non-linearity. A remarkable capacity of SVM algorithms

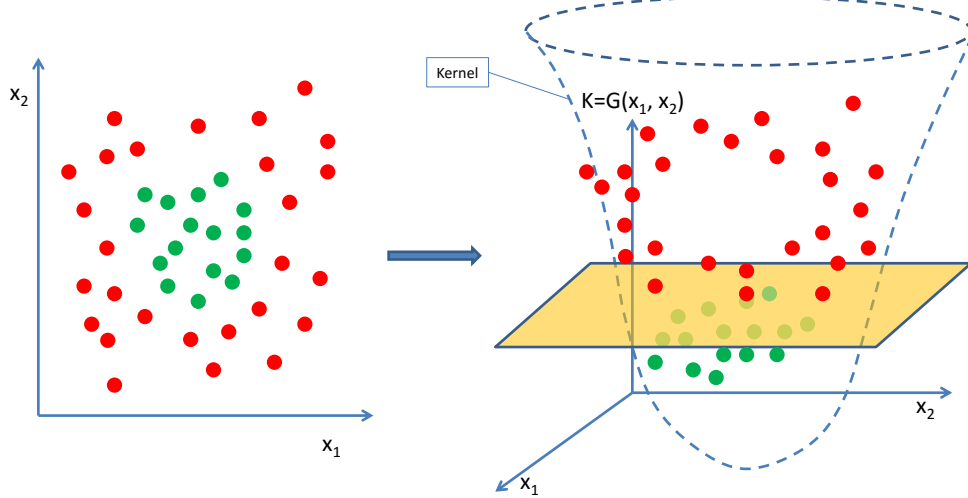


Figure 2.8: Support vector machines: Mapping into higher dimension spaces.

is their ability to map the problem into a higher dimension space using the so called “kernel functions”. Upgrading dimensionality is actually an important option to linearize non-linear classifications. This concept can be visualized in figure 2.8, where some gaussian function can map into a three-dimensional space a non-linearly classifiable bi-dimensional sample, on the basis of a “color” condition (green versus red in the picture). It is fairly intuitive that while there is not possible to linearly separate points by color in the 2D space, the gaussian transformed 3D points can be neatly split by the plane (see figure 4). In general terms, by using a variety of non-linear kernels, SVMs can add additional dimensions to the data to generate upgraded distributions susceptible of linear splitting in the way of the example. Conceptually, the kernel mapping adds new features to the model by defining new mathematical relationships among the initial, measured characteristics. In this way, SVM models become extremely powerful classifiers that can learn concepts that were not explicitly described by the original (observed) measures.

There are a number of kernel functions that have been widely used and are implemented in most libraries (i.e.: linear, polynomial, sigmoid, gaussian, etc). Unfortunately, there are no objective nor reliable criteria for choosing a particular kernel in a given learning problem. Kernel performance will heavily

depend on the concept to be learned, the training-set size and composition and the relationships among the features. In other words, the choice of kernel is arbitrary in many cases and frequently a trial and error protocol, including the evaluation of candidate models on a validation dataset, is advised.





## Objectives

The main objectives in this thesis are:

1. To develop a protocol or work flow which details the steps necessary to apply recurrence based nonlinear systemic analysis on complete genomes. This implies, in a first phase, data acquisition and preprocessing, generation of time series, phase space reconstruction via embedding as proposed by Taken<sup>19</sup>, obtaining characteristics recurrence plots and their recurrence quantitative analysis.
2. To develop and implement a software which operates this work flow and is apt for high-throughput analyses and reproducible research. The aim is that the software takes as input the genomic data, for instance DNA methylation data, and performs automatically the analyses and evaluations.
3. to characterize nuclear and mitochondrial genomes in terms of recurrences fingerprints.
4. To study the utility of this patterns with the aim to learn about the dynamical structure, its relation to function and other properties of genomes in different setting of genetic adaptation.
5. In the case that the objective mentioned above have an positive outcome, we want to extend the work flow setting up predictive or diagnostic tools based on machine learning, specially support vector machines, and explore

possible applications to problems emerging from adaptation to stress, in a wider sense, including diseases like cancer or species diversification. The latter would be an alignment-free approach.

## Methods

### 4.1 Recurrence analysis

#### 4.1.1 Phase space reconstruction

A time series describes the changes of a single system variable (e.g. DNA Methylation Density) over the time (which is any ordered sequence). In contrast, a phase space (or state space) represents all possible states of a system. We can describe a state at time  $t$  by its state variables  $x_1(t), x_2(t), \dots, x_d(t)$  that form a vector in a  $d$  dimensional space (please note: variable and parameter names taken from the references may have changed in this text to keep a consistent terminology). The trajectory of the vector over the time shows the temporal evolution of the system<sup>113</sup> – the so called phase orbit<sup>114</sup>.

Natural systems such as genomes are multivariate and there models usually don't cover the complete set of state variables. Reasons to this might be a) a leak of knowledge; b) experimenters are technically not able to measure all variables; or c) the data volume is not viable. Nevertheless, it might be possible to reconstruct an image of the phase space from a single signal using time delays<sup>115,19</sup> if we can map the time series  $Y = \{y_n\}$  by  $y_n = f(x_n)$ . The sequences of the reconstruction vectors would be

$$\{(y_n, y_{n+\tau}, \dots, y_{n+(m-1)\tau}) \in \mathbb{R}^m\}_n^{N-(m-1)} \quad (4.1)$$

where  $m$  is the dimension of the reconstructed phase space,  $\tau$  is the time delay,  $N$  is the number of measures and ideally the condition  $m > 2d$  is fulfilled (see

also Scheluter<sup>116</sup> or Casdagli et al.<sup>117</sup>)

To get a time delayed reconstruction from a time series an adequate time delay  $\tau$  has to be estimated to capture the dynamics of the system. To take non-linear correlations into account Fraser and Swinney<sup>118</sup> proposed time delayed mutual information to determine the delay  $\tau$ . Mutual information measures the general dependency of two variables. Its definition as a function of  $\tau$  is:

$$S = - \sum_{ij} p_{ij}(\tau) \ln \frac{p_{ij}(\tau)}{p_i p_j}, \quad (4.2)$$

where  $p_i$  and  $p_j$  are the probabilities to find a value in the  $i$ -th or  $j$ -th position of a time series, and  $p_{ij}(\tau)$  is the joint probability to observe the values for  $i$  and  $j$  at distance  $\tau$ <sup>119,120</sup>. We use the function *mutual* from the R-package *timeseriesChaos*<sup>121</sup> and choose the delay which produces the first local minimum of mutual information for the reconstruction.

The second parameter we have to choose to reconstruct from the time series is the embedding dimension  $m$ . With an adequate dimension the univariate projection (the time series) can be unfolded to a multivariate state space that restores the topology of the original system. Takens<sup>19</sup> proofed that the attractor is unfolded and all self-crossings of the orbit disappear when when  $m > 2d$ , where  $d$  is the dimension of the original system.

Actually, in our experiments we can not verify whether the conditions of Takens' theorem are fulfilled or not because the dimension of the original system and also the appropriate mapping function are unknown to us. Nevertheless,  $m > 2d$  is a sufficient condition and successful embeddings of unknown systems have been described in the literature. A classical example is the phase space reconstruction from a time series of measured time intervals between successive drops falling from a dripping tap<sup>122,123</sup>. More recently, it has been shown that the reconstruction of apparent random gamma-ray bursts time profiles reveal the existence of a well-defined strange attractor<sup>124</sup> or, to have also a life science example, Beninca et al.<sup>125</sup> have used embedding of time series to calculate the Lyapunov exponents and demonstrate chaos in a plankton food web.

A frequently used way to estimate the minimum embedding dimension is the method of *false nearest neighbors* proposed by Kennel et al.<sup>126</sup>. We have implemented this method in our R-package *bract* following the original

idea: A projection may place points into neighborhood which in the orbit of the unfolded attractor are far a way from each other. If we increment the dimension and observe that the number of false nearest neighbors drop to zero, we have embedded the attractor. The square of the Euclidean distance between the point  $y(n)$  and its  $r$ th nearest neighbour  $y^{(r)}(n)$  in  $d$  dimensions is

$$R_d^2(n, r) = \sum_{k=0}^{d-1} [x(n + k\tau) - x^{(r)}(n + k\tau)]^2. \quad (4.3)$$

To increment the dimension we add a  $(d+1)$ th coordinate to each of the vectors  $y(n)$ . The new coordinate is just  $x(n + d\tau)$ . The distance between the same points in the new dimension is now

$$R_{d+1}^2(n, r) = R_d^2(n, r) + [x(n + d\tau) - x^{(r)}(n + d\tau)]^2. \quad (4.4)$$

If we notice that, when passing to the next higher dimension, the distance between  $y(n)$  and  $y^{(r)}(n)$  increase much, then the embedding has still errors. An increase of

$$\left[ \frac{R_{d+1}^2(n, r) - R_d^2(n, r)}{R_d^2(n, r)} \right]^{1/2} = \frac{|x(n + \tau d) - x^{(r)}(n + \tau d)|}{R_d(n, r)} > R_{tol}, \quad (4.5)$$

reveals false neighbors. As threshold we used the authors suggestion  $R_{tol} = 10$ .

To exclude neighbors which are nearest but distant Kennel et al.<sup>126</sup> propose a second criterion:

$$\frac{R_{d+1}(n)}{R_A} > A_{tol} \quad (4.6)$$

where  $A_{tol}$  is a threshold which we set to two and

$$R_A^2 = \frac{1}{N} \sum_{n=1}^N [x(n) - \bar{x}]^2 \quad (4.7)$$

and

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x(n) \quad (4.8)$$

In our experiments, we used, if possible, the dimension where for the first time the percentage of false nearest neighbors reaches 0%.

### 4.1.2 Recurrence plots

Recurrences are a fundamental but undervalued characteristic of chromosomes. To study recurrences in dynamical systems Eckmann et al.<sup>17</sup> introduced a method which reduced the multidimensional trajectory of the attractor to a two-dimensional recurrence plot (RP). The RP can be described by the matrix

$$R_{i,j}(\varepsilon) = \Theta(\varepsilon - \|\vec{x}_i - \vec{x}_j\|), \quad i, j = 1, \dots, N, \quad (4.9)$$

where  $\vec{x}_i$  is a point in the embedding at time  $i$ ,  $N$  the number of points,  $\varepsilon$  an arbitrary threshold,  $\|\cdot\|$  is a norm and  $\Theta(\cdot)$  is the Heaviside function<sup>127,128</sup>

$$\Theta(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} \quad (4.10)$$

### 4.1.3 Recurrence quantitative analysis and lacunarity

To quantify the recurrence plots we calculate for each of them the lacunarity<sup>129</sup> and the recurrence quantification analysis (RQA) variables recurrence rate  $RR$ , determinism  $DET$ , entropy  $ENTR$ , ratio  $RATIO$ , laminarity  $LAM$  and trapping time  $TT$ <sup>130,128</sup>. We calculate the recurrence rate ( $RR$ ) or density of recurrent points by the formula

$$RR(\varepsilon) = \frac{1}{N^2} \sum_{i,j=1}^N R_{i,j}(\varepsilon) \quad (4.11)$$

$DET$ ,  $ENTR$  and  $RATIO$  are measures based on the frequencies of diagonal line lengths. Its histogram  $P(\varepsilon, l)$  is

$$P(\varepsilon, l) = \sum_{i,j=1}^N (1 - R_{i-1,j-1}(\varepsilon))(1 - R_{i+l,j+l}(\varepsilon)) \prod_{k=0}^{l-1} R_{i+k,j+k}(\varepsilon) \quad (4.12)$$

The predictability or determinism of the system increases with the amount of diagonal lines in the recurrence plot. We measure it by

$$DET = \frac{\sum_{l=l_{min}}^N lP(\varepsilon, l)}{\sum_{l=1}^N lP(\varepsilon, l)}, \quad (4.13)$$

where  $l_{min}$  is the minimum length of a diagonal line. We use  $l_{min} = 7$ . To measure the complexity of the system we calculate the Shannon information entropy of all diagonal lines lengths distributed over the histogram. It is based on the probability  $p(l) = P(l)/N_l$  to find a diagonal line of length  $l$  in the recurrence plot.

$$ENTR = - \sum_{l=l_{min}}^N p(l) \ln p(l) \quad (4.14)$$

The RQA measure ratio is just

$$RATIO = \frac{DET}{RR} \quad (4.15)$$

and may reveal transitions in the dynamics.

$LAM$  and  $TT$  are measures based on the frequencies of vertical lines lengths  $v$ . The histogram is

$$P(\varepsilon, v) = \sum_{i,j=1}^N (1 - R_{i-1,j-1}(\varepsilon))(1 - R_{i+l,j+l}(\varepsilon)) \prod_{k=0}^{v-1} R_{i+k,j+k}(\varepsilon) \quad (4.16)$$

In analogy to the determinism we can calculate the laminarity

$$LAM = \frac{\sum_{v=v_{min}}^N vP(v)}{\sum_{v=1}^N vP(v)}, \quad (4.17)$$

where  $v_{min}$  is the minimal length of a vertical line. We use  $v_{min} = 7$ . The time where the state of the system is trapped we compute by the average vertical line length

$$TT = \frac{\sum_{v=v_{min}}^N vP(v)}{\sum_{v=v_{min}}^N P(v)} \quad (4.18)$$

We recommend to read the review written by Marwan et al.<sup>128</sup> to get further details about the theory and application of the recurrence quantification analysis.

To calculate the lacurarity we used the gliding box algorithm re-described by Plotnick et al.<sup>131</sup>. It requires a  $r \times r$  box that moves column- and row-wise over the entire RP matrix to count the box mass  $S$  (number of 1's) for each box. This gives us the frequency distribution of box masses  $n(S, r)$ . The number of total boxes is  $N(r) = (M - r + 1)^2$ , where  $M$  is the size of the matrix (e.g. number of columns). The probability distribution is given by  $Q(S, r) = n(S, r)/N(r)$ .



Now we can determine the first and the second moment of the distribution which are  $Z^{(1)} = \sum S Q(S, r)$  and  $Z^{(2)} = \sum S^2 Q(S, r)$  and finally the lacunarity  $\Lambda(r) = Z^{(2)} / (Z^{(1)})^2$ .

## 4.2 Classification

We used the RQA measures RR, DET, ENTR, RATIO, LAM and TT and the lacunarity for binary classifications of tumor and normal cells, subspecies of chimpanzee and species belonging to the superfamily Caniformia. To avoid that very sparse or dense recurrence plots bias the classification we accepted only samples which meet the condition  $0.05 \leq RR \leq 0.2$ . From the filtered data entries for each class, cases and controls, we selected 80% for model training and the rest to perform the tests. We scaled the training and test data with the R function *scale*. To train the support vector machine we used the function *svm* from the R-package *e1071*<sup>132</sup>. We applied, except for the parameter *scale*, the default configuration for classifications: *scale=FALSE*, *type=C-classification*, *kernel=radial*, *gamma=1/(data dimension)*, *cost=1*, *tolerance=0.001*, *epsilon=0.1*, *fitted=TRUE*, *seed=1*, *probability=TRUE*. The tests or predictions based on the obtained models we performed with the function *predict.svm* from the same package. If necessary, we use the function *tune* to optimize classification parameters.

### 4.2.1 Support vector machines

The following lines, inspired by a tutorial from Tristan Fletcher<sup>133</sup>, provide the mathematical background to support vector machines.

#### Linearly separable binary classification

Let's consider that we have a sample with  $L$  training points  $x_i$ , each of them having  $N$  features (dimension  $N$ ) and can be binary classify in terms of a binary condition  $y$  (see figure 2.5, where that condition would be the points "color"), whose possible values are:  $+1$  or  $-1$ . That is, suppose a set:  $\{x_i, y_i\}$  where  $i = 1 \dots L, y_i \in \{-1, 1\}, x \in \mathbb{R}^D$

Let's also presume that our sample is linearly separable in terms of  $y$ . That means that we could separate both classes with an hyperplane that, in case of dimension 2 (as in figures 2.5 or 2.6), would be a line. The equation of this general hyperplane would be:  $w \cdot x + b = 0$  where  $w$  is normal to the hyperplane,  $\frac{b}{\|w\|}$  is the perpendicular distance from the hyperplane to the origin and  $\frac{2}{\|w\|}$  is the distance between these two planes.

Basically, the problem is to find the values of  $w$  and  $b$  so that our training set can be described by:

$$x_i \cdot w + b \geq +1 \quad \text{if } y_i = +1 \quad (4.19)$$

$$x_i \cdot w + b \leq -1 \quad \text{if } y_i = -1 \quad (4.20)$$

We can combine both equations to give:

$$y_i(x_i \cdot w + b) - 1 \geq 0, \forall i \quad (4.21)$$

As  $\frac{2}{\|w\|}$  is the distance between planes, each margin is given by  $\frac{1}{\|w\|}$ , a quantity usually named "SVM margin". The idea here is, consequently, maximize this margin or, in other words, minimize the Euclidean norm  $\|w\|$ . To make easier the handling of this problem, it is convenient to reformulate the goal as minimizing  $\frac{1}{2}\|w\|^2$ , which is equivalent and make it possible to perform quadratic programming (QP) optimization, a very efficient way to achieve the solution. In summary, we need to find:

$$\min \frac{1}{2}\|w\|^2 \quad \text{such that} \quad y_i(x_i \cdot w + b) - 1 \geq 0, \forall i \quad (4.22)$$

To manage the constrains in this minimization, we need to allocate them Lagrange multipliers  $\alpha$ , where  $\alpha_i \geq 0, \forall i$ :

$$L_p \equiv \frac{1}{2}\|w\|^2 - \alpha[y_i(x_i \cdot w + b) - 1], \forall i \quad (4.23)$$

$$\equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^L \alpha_i [y_i(x_i \cdot w + b) - 1] \quad (4.24)$$

$$\equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^L \alpha_i y_i(x_i \cdot w + b) + \sum_{i=1}^L \alpha_i \quad (4.25)$$

We must find the values of  $w$  and  $b$  which minimize, and the  $\alpha$  that maximize the expression. This can be done by differentiation of  $L_p$  with respect to  $w$  and

$b$  and making these derivatives equal to zero:

$$\frac{\delta L_p}{\delta w} = 0 \rightarrow w = \sum_{i=1}^L \alpha_i y_i x_i \quad (4.26)$$

$$\frac{\delta L_p}{\delta b} = 0 \rightarrow \sum_{i=1}^L \alpha_i y_i = 0 \quad (4.27)$$

Now we can substitute these equivalences into  $L_p$  expression:

$$L_D \equiv \sum_{i=1}^L \alpha_i - \frac{1}{2} \sum_{i,j=1}^L \alpha_i \alpha_j y_i y_j x_i \cdot x_j \quad s.t. \quad \alpha_i \geq 0, \forall i, \sum_{i=1}^L \alpha_i y_i = 0 \quad (4.28)$$

$$\equiv \sum_{i=1}^L \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i H_{i,j} \alpha_j \quad s.t. \quad H_{i,j} \equiv y_i y_j x_i \cdot x_j \quad (4.29)$$

$$\equiv \sum_{i=1}^L \alpha_i - \frac{1}{2} \alpha^T H \alpha \quad s.t. \quad \alpha_i \geq 0, \forall i, \sum_{i=1}^L \alpha_i y_i = 0 \quad (4.30)$$

A new expression, known as Dual Form of the Primary  $L_p$  that only requires dot products of each input vector  $x_i$  to be calculated (something being important when using Kernels).

Rather than minimizing  $L_p$ , we have now to maximize  $L_D$ :

$$\max_{\alpha} \left[ \sum_{i=1}^L \alpha_i - \frac{1}{2} \alpha^T H \alpha \right] \quad s.t. \quad \alpha_i \geq 0, \forall i, \sum_{i=1}^L \alpha_i y_i = 0 \quad (4.31)$$

Using a QP solver algorithm we can calculate and then  $w$  as  $\sum_{i=1}^L \alpha_i y_i x_i$ . This last relationship can also be used to calculate  $b$ , having in mind that any point being a support vector has to verify  $y_i(x_i \cdot w + b) = 1$  or, substituting by the above expression:

$$y_s \left( \sum_{m \in S} \alpha_m y_m x_m \cdot x_s + b \right) = 1 \quad (4.32)$$

We can now multiply both sides by  $y_s$  ( $s$  is the set of index of support vectors):

$$y_s^2 \left( \sum_{m \in S} \alpha_m y_m x_m \cdot x_s + b \right) = y_s \quad (4.33)$$

An having in mind that the only values of  $y_s$  can be  $\pm 1$ ,  $b$  can be expressed as:

$$b = y_s - \sum_{m \in s} \alpha_m y_m x_m \cdot x_s \quad (4.34)$$

Rather than using only one arbitrary support vector we can make the average of all of them. So, finally:

$$b = \frac{1}{N_s} \sum_{s \in S} (y_s - \sum_{m \in S} \alpha_m y_m x_m \cdot x_s) \quad (4.35)$$

### Non-linearly separable binary classification

When data point cannot be linearly separated we can use two different strategies: soften hyperplane margins or using a so called *kernel trick*. Softening margins can be achieved by introducing a positive slack variable  $\xi_i$ ,  $i = 1 \dots L$ :

$$x_i \cdot w + b \geq +1 - \xi_i, \quad \xi_i \geq 0, \forall i \quad \text{if } y_i = +1 \quad (4.36)$$

$$x_i \cdot w + b \leq -1 + \xi_i, \quad \xi_i \geq 0, \forall i \quad \text{if } y_i = -1 \quad (4.37)$$

That, by combining, gives:

$$y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0, \quad \xi_i \geq 0, \forall i \quad (4.38)$$

In this case, those points falling in the wrong side of the hyperplane are penalized proportionally to the distance. To do that, the objective function has to be reformulated by introducing a parameter  $C$  that weighs the trade-off between the slack variable and the size of margins:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^L \xi_i \quad \text{such that } y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0, \forall i \quad (4.39)$$

As above, we can reformulate this objective function as a Lagrangian that has to be minimize with respect to  $w$ ,  $b$  and and maximize with respect to  $\alpha$ . The resulting expression for  $L_D$  has the same form than before, however, it can be shown that has to be  $\alpha \leq C$ . So the target will be:

$$\max_{\alpha} \left[ \sum_{i=1}^L \alpha_i - \frac{1}{2} \alpha^T H \alpha \right] \quad \text{s.t.} \quad 0 \leq \alpha_i \leq C, \forall i, \quad \sum_{i=1}^L \alpha_i y_i = 0 \quad (4.40)$$

And parameters can be calculated in the same way.

The alternative approach is to use Kernel functions. Kernels are, in general, functions of the form:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (4.41)$$

There are many different possibilities. Probably the most commonly used kernel functions (included in all the most popular libraries) are the linear, polynomial, sigmoidal and gaussian:

$$\text{LINEAR} \quad K(x_i, x_j) = x_i \cdot x_j \quad (4.42)$$

$$\text{POLYNOMIAL} \quad K(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad (4.43)$$

$$\text{SIGMOID} \quad K(x_i, x_j) = \tanh(Kx_i \cdot x_j - \delta) \quad (4.44)$$

$$\text{GAUSSIAN} \quad K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\alpha^2}} \quad (4.45)$$

### 4.3 Evaluation

With the known labels and the predictions as input we have drawn receiver operating characteristic curves (ROC, sensitivity versus specificity) and have determined the areas under curve (AUC) for all performed classifications using the R-package `pROC`<sup>134</sup>.

Besides, we have carried out cluster analysis on the AUCs using the Bioconductor R-package `ComplexHeatmap`<sup>135</sup>. We have clustered the RQA measures and lacunarity in the same way for each cancer type separately and also taking all together.

To see the distributions of RQA measure pairs for cases and controls of each cancer type we have produced multi scatter plots using the function `spLOM` from the R-package `lattice` on the filtered RQA and lacunarity data.

## Results

In this chapter we demonstrate that viewing genomes as adaptive complex systems opens a innovative macroscopic perspective to genome analysis with new possibilities alongside classical statistics.

### 5.1 Work flow

Curiously, just in very few occasions recurrence plot based analysis have been applied to genomes, limited to study structural correlations in a human DNA fragment as well as in the yeast and *Caenorhabditis elegans* genomes<sup>13,136</sup>. Although recurrence quantification analysis has a history which goes back to the early nineties there is, to our knowledge, no further application to genomic problems neither a protocol or a software, which can be used without expert knowledge and strong mathematical background, for biological or medical research in this field available. In this work we have developed a new protocol and its accompanying software (`bract`) providing a nonlinear systemic approach to genome analysis. The work flow is robust and adaptable enabling us to perform a variety of experiments. The figure 5.1 shows an activity diagram of the work flow. Its has four mayor components: 1) Data acquisition and preparation; 2) recurrence analysis; 3) classification by support vector machines and 4) evaluation. Naturally, the data acquisition and preparation is the most variable part of the work flow and will therefore be described separately for each experiment in the corresponding sections. In contrast, the steps to process the recurrence analysis are invariant, but include, of course, minor configuration changes

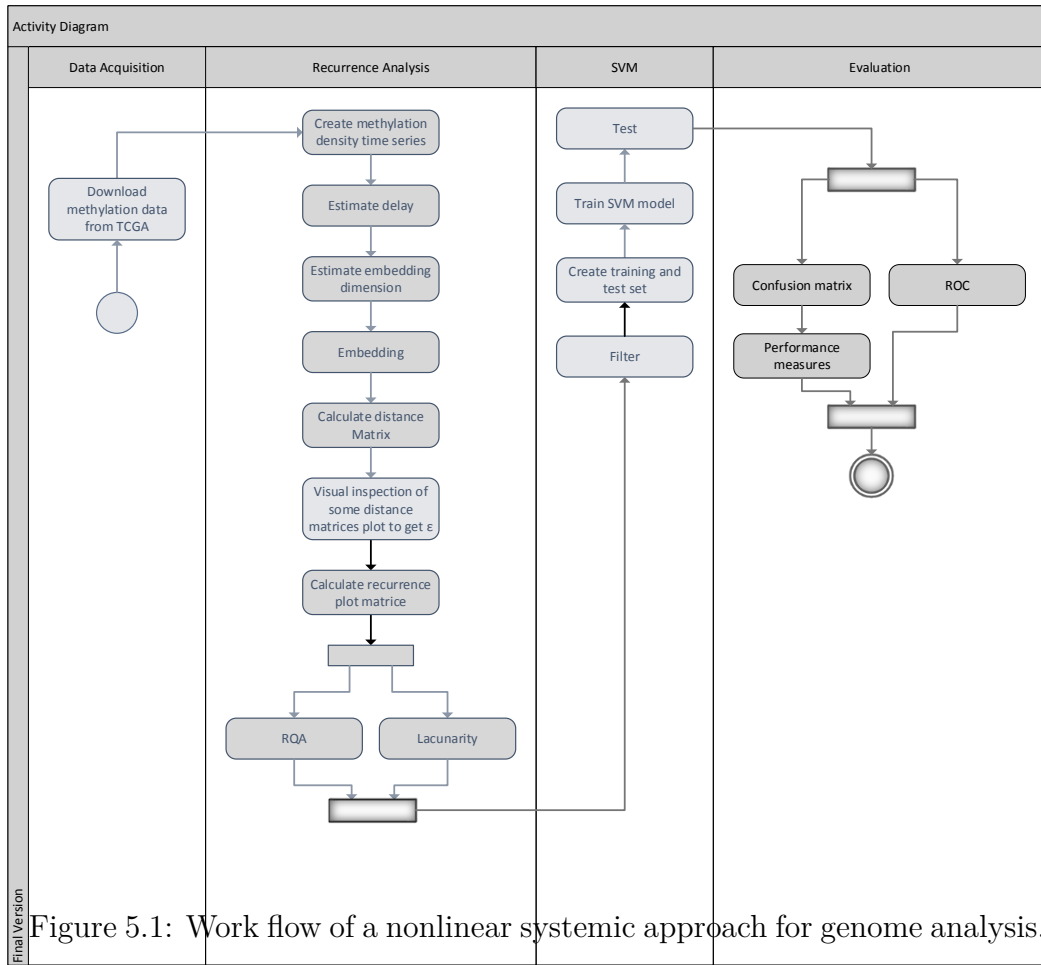


Figure 5.1: Work flow of a nonlinear systemic approach for genome analysis.

**Figure 1:** Diagram of the main activities of the method.

which adapt the work flow to the experimental problems. Although, some software packages to perform the complete recurrence quantification analysis are available (see <http://www.recurrence-plot.tk/programmes.php>), non was apt for our purposes. Either the projects were out dated, the methods were only accessible through GUIs and could not be used in a batch computing system, the software could not be extended easily or was not open source. These have been arguments enough to start the development from the scratch in R – having a future integration into bioconductor<sup>137,138</sup> in mind.

Basically, each step of this part of the work flow is extensively discussed in the literature (see methods), but even so we have been confronted with some pitfalls. Specially, we found it difficult to estimate the embedding dimension. Most, if not all, software related to this field, among them the R-package tseriesChaos<sup>121</sup> and TISEAN<sup>119</sup>, include a function to estimate the embedding

dimension referring to the method of false nearest neighbors (FNN) proposed by Kennel et al.<sup>126</sup>. But, these software functions are variants of the Kennel-method and need additional parameters. The most critical one was a threshold which restricts the radius of nearest neighbors to be include in the statistics. We have not been able to set this parameter with reasonable values. All intents failed. Plotting FNN as a function of embedding dimension we got fluctuating curves with first local minima often close to 40% of false nearest neighbors and reaching 0% FNN at very high dimension or never. We reimplemented the FNN method in our software package `bract` as described in 40. Finally we got useful graphs (see figure 5.9b) and have been able to estimate the embedding dimension choosing the dimension where for the first time the percentage of false nearest neighbors reaches 0% or alternatively 1%. Another, similar problem was to set the threshold radius,  $\varepsilon$ , necessary to get the recurrence plots from the distance matrices. No hint from the literature was applicable to our data sets. We estimated an approximate value for  $\varepsilon$  empirically, supported by contour distance plots and graphs showing  $\varepsilon$  as a function of the recurrence quantification analysis (RQA) measures as shown here for the Rössler attractor in figure 5.2. Finally, based on that, we defined a factor  $a$  which is part of the equation to calculated the threshold radius:

$$\varepsilon = a\sigma_D \quad (5.1)$$

where  $\sigma_D$  is the standard deviation of the euclidean distance matrix of the reconstructed phase space. This has to be done only ones for each kind of experiment (Rössler, systemic cancer classification, systemic identification of taxa, etc.). After setting  $a$  in equation 5.1 the `bract` pipeline runs each steps of the work flow automatically.

To test if our time series reconstructions and RQAs are reasonable, we applied our software to the well studied Rössler attractor (see figure 5.3). The Rössler attractor is defined by three nonlinear differential equations formulated as follows:

$$\dot{X} = -(Y + Z) \quad (5.2)$$

$$\dot{Y} = X + aY \quad (5.3)$$

$$\dot{Z} = b + (X - c)Z \quad (5.4)$$



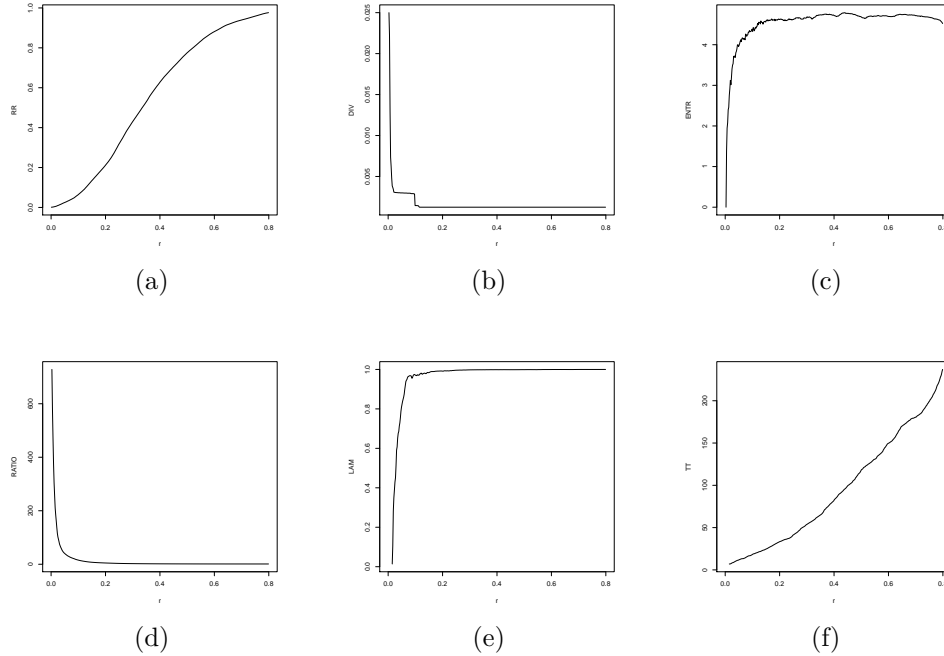


Figure 5.2: Influence of the threshold  $\varepsilon$  on the RQA measures a) RR, b) DIV, c) ENTR, d) RATIO, e) LAM and f) TT shown for the standard Rössler system.

We created a time series from the equation  $\dot{Y} = X + aY$  and submitted it to our pipeline. The initial conditions, estimations for  $\tau$  and  $m$  and resulting graphs we show in figure 5.4. Comparing the reconstructed phase space with the original attractor one recognizes their eye-catching similarity (note that the reconstruction is rotated). For the reconstructed phase space we created a recurrence plot with  $\varepsilon = 0.22$ . The obtained recurrence plot (figure 5.5) is very similar to others described in the literature (for instance here Marwan<sup>139</sup>). The recurrence quantification analysis of this RP also results in plausible values –  $RR = 0.14$ ,  $DET = 0.991$  and  $RATIO = 7.1$  – where the high values of DET and RATIO indicate a high predictability of the Rössler attractor. In summary – the new bract pipeline which processes our work flow passed the tests and is ready to make a nonlinear systemic approach applicable for genome analysis.

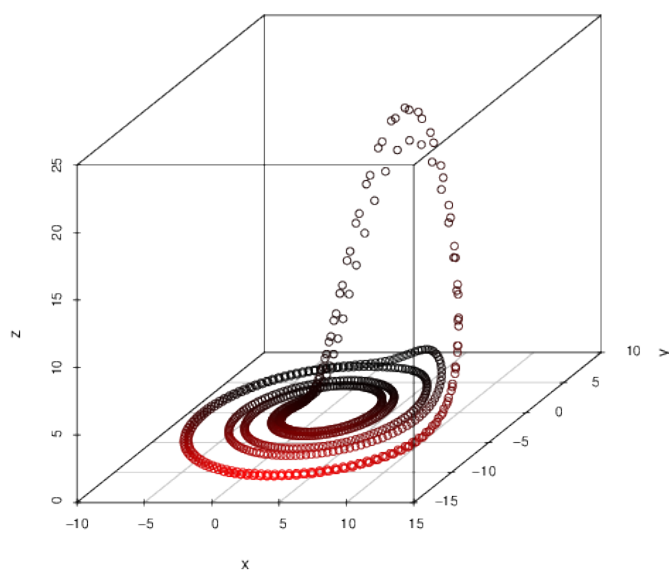
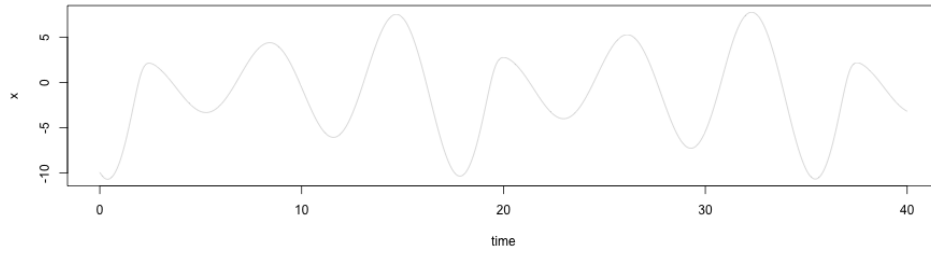
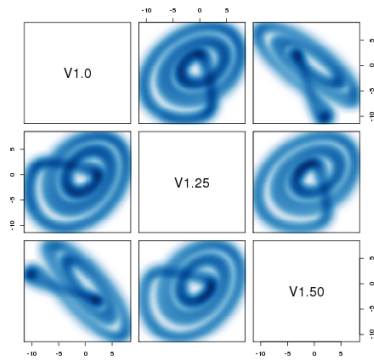


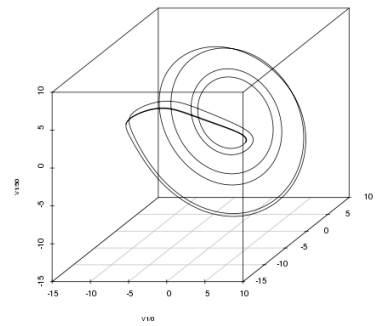
Figure 5.3: Standard Rössler attractor for  $a = 0.2$ ;  $b = 0.2$ ;  $c = 5.7$ . The initial conditions have been  $x_0 = -1.894$ ;  $y_0 = -9.92$ ;  $z_0 = 0.025$ . The time measure started at 0.0 and ended at 40 incrementing 0.05 between the observations.



(a)



(b)



(c)

Figure 5.4: a) Time series build from the y-coordinates of a standard Rössler system. The time measure started at time  $t = 0.0$  and ended at time  $t = 40.0$ . The time between observations was  $\Delta = 0.05$ . The initial conditions were  $x = -1.894$ ,  $y = -9.92$  and  $z = 0.025$ . Further we set  $a = 0.2$ ,  $b = 0.2$  and  $c = 5.7$ . b) Multi-scatter plot and c) 3d plot of its phase space reconstruction ( $\tau = 25$  and  $m = 3$ ). To get  $\tau$  we used the first minimum of the average mutual information index and  $m$  was obtained when the percentage of false nearest neighbors had reached less than 1%.

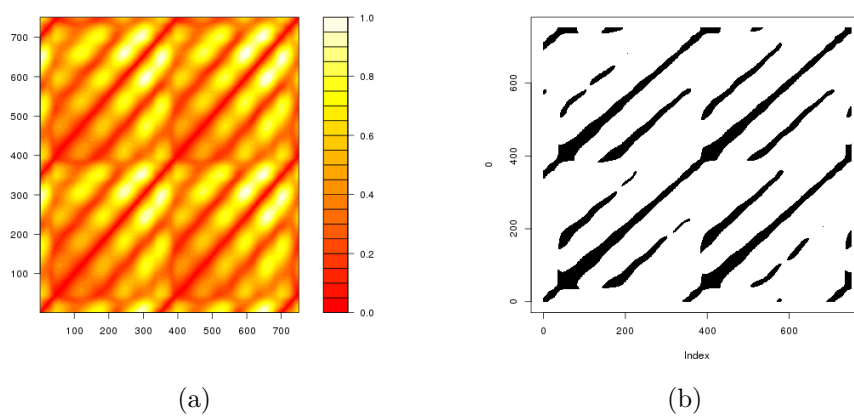


Figure 5.5: Distance contour plot and recurrence plot ( $\varepsilon = 0.22$ ) for the reconstruction shown in 5.4.

## 5.2 Marker-less cancer classification

### 5.2.1 Data acquisition and preparation

#### The original DNA methylation data

We obtained the data samples for our experiments from *The Cancer Genome Atlas (TCGA)*<sup>140</sup>. We used the data access matrix<sup>141</sup> to download public available, from Illumina Infinium HumanMethylation450 BeadChip<sup>140</sup> derived, level 3, tumor- and normal-matched data from all batches for different diseases. We selected the following diseases which are not restricted under the term of the TCGA's publication guidelines and having a sample set size  $s_{tumor} > 30 \wedge s_{normal} > 30$ : Breast cancer (BRCA), colon and rectal adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), clear cell kidney carcinoma (KIRC), papillary kidney carcinoma (KIRP), prostate adenocarcinoma (PRAD), papillary thyroid carcinoma (THCA) and uterine corpus endometrial carcinoma (UCEC).

TCGA compiles the requested data to a uncompressed or, if desired, compressed tarball. The size of uncompressed tumor DNA methylation data packages is quite huge and can easily pass ten gigabytes. The size of normal tissue DNA methylation data packages is notably lower, about one gigabyte.

The accompanying meta-data describes the relationship between important entities of the experiment. This information is basically provided by two tab-delimited MAGE-TAB documents with standard format<sup>142,143</sup>. One is the **I**nvestigation **D**escription **F**ormat (IDF), which contains information about the submitter contact details, the experiment and protocols. It links to an other file in **S**ample and **D**ata **R**elationship **F**ormat (SDRF), which describes exactly what its format name says. Moreover, the SDRF is a textual description of a directed acyclic graph (DAG). Detailed guidelines on the creation of SDRF files for TCGA can be read on the National Cancer Institute (NCI) wiki<sup>144</sup>. On the same page it is recommended to visualize the SDRF with the script *expt\_check.pl* from the EBI's *TAB2MAGE* perl package<sup>145</sup>. But *TAB2MAGE* depends on the outdated version 2.7.0 of the XML parser Xerces-C++<sup>146</sup> and its installation requires a disproportionate effort. We chose the perl module

`Bio::MAGETAB` instead<sup>147</sup>. The figure 5.6 shows a fraction of the DAG build from the IDF file

```
jhu-usc.edu_LIHC.HumanMethylation450.1.11.0.idf.txt
```

and its corresponding SDRF file

```
jhu-usc.edu_LIHC.HumanMethylation450.1.11.0.sdrf.txt
```

By the way, we found that the protocol name entry of the IDF file was incomplete. We corrected the file appending

```
jhu-usc.edu:image_acquisition:HumanMethylation450:01,  
↪ jhu-usc.edu:feature_extraction:HumanMethylation450:01 and  
↪ jhu-usc.edu:within_bioassay_data_set_function:  
↪ HumanMethylation450:01
```

to the corresponding line. The complete, huge, DAG can be visualized executing the following command with a command-line interpreter (shell):

```
read_magetab.pl -x -r -g  
↪ "jhu-usc.edu_LIHC.HumanMethylation450.1.11.0.png" -i  
↪ "jhu-usc.edu_LIHC.HumanMethylation450.1.11.0.idf.txt"
```

The level three DNA methylation data, namely the  $\beta$ -values, are stored in tab-delimited ASCII text files inside the directory

```
.../DNA_Methylation/JHU_USC__HumanMethylation450/Level_3
```

of the tarball. There is one data file for each sample containing information about the references to the composite element, calculated  $\beta$ -values, gene symbols, chromosome names and genomic coordinates (according to the human genome version hg18). A TCGA, level three, HumanMethylation450 data file has 485579 rows whereof 485577 contain data. Its size is about 21Mb. Table 5.1 shows part of the head of the data file

```
jhu-usc.edu_LIHC.HumanMethylation450.1.lvl-3.  
↪ TCGA-BC-A10Q-01A-11D-A132-05.txt
```

How we use each column of the data file is described in the corresponding sections below.

Table 5.1: The head of a TCGA 3rd level HumanMethylation450 data file. Shown are the rows from number two to six.

Composite Element REF	Beta_value	Gene_Symbol	Chromosome	Genomic_Coordinate
cg00000029	0.72753994338749	RBL2	16	53468112
cg00000108	NA	C3orf35	3	37459206
cg00000109	NA	FNDC3B	3	171916037
cg00000165	0.786574582300652		1	91194674
⋮		⋮	⋮	⋮

### Preparation of the data files and working directories

We have seen already that the size of the original data files is large. To decrease CPU-usages and upload-time to the high performance clusters (HPC) we extract the data for the chromosome of interest from the original data. This is done using a script from our R-package bract in the following way:

```
extractChrS.R --chromosome="1"
↪ "([:print:])*HumanMethylation450([[:print:]]*)\\.txt" %$"
↪ %@TODO: put $ after txt when methods is finished
```

The size of a single file containing chromosome 1 data is now reduced to 2Mb.

The names of the data files have the following pattern:

```
<domain for a TCGA center>_<disease study>.<vendor-specific
↪ technology platform>.<batch>.<data level>.<barcode>.txt
```

During our analysis the name of the DNA methylation data file serve as a stem for resulting files and will be extended by some more tags. To reduce the length of the file names, we eliminate redundant or (for us) irrelevant parts applying simple bash commands inside the data directory:

```
for file in *.txt ; do mv $file
↪ ${file//HumanMethylation450./} ; done
for file in *.txt ; do mv $file ${file//lv1-3./} ; done
for file in *.txt ; do mv $file ${file//jhu-usc.edu_/} ; done
```

The primary identifier of a sample is its universally unique identifier (UUID)<sup>148</sup>. It is a 128 bit number represented by 32 lowercase hexadecimal digits looking like this: 506c8b0c-c7cf-4f9b-8289-73106563c9f9. I mention the UUID here

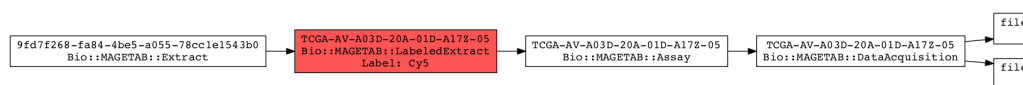


Figure 5.6: Visualization of the directed acyclic graph describing the first entry from the magetab document `jhu-usc.edu_LIHC.HumanMethylation450.1.11.0`

just for completeness, it's currently not directly relevant to the data we use or produce, but it may be useful to get further information on the samples in future investigations. The bar code<sup>149</sup> is a human readable identifier for each sample. It is important to know that a bar code can change if the associated meta data changes. The bar code, in the file names allows to distinguish tumor from normal data in a cumbersome way (see Tab.5.2). To identify normal tissue data files at first sight we append the tag NRML to the disease type. This is also done by a simple replace statement:

```
for file in *.txt ; do mv $file ${file//LIHC/LIHCNRML} ; done
```

For example, the tumor tissue data file

```
jhu-usc.edu_LIHC.HumanMethylation450.1.lvl-3.  
↔ TCGA-BC-A10Q-01A-11D-A132-05.txt
```

has been renamed to

```
LIHC.1.TCGA-BC-A10Q-01A-11D-A132-05.txt
```

and the normal tissue data file

```
jhu-usc.edu_LIHC.HumanMethylation450.4.lvl-3.  
↔ TCGA-AV-A03D-20A-01D-A17Z-05.txt
```

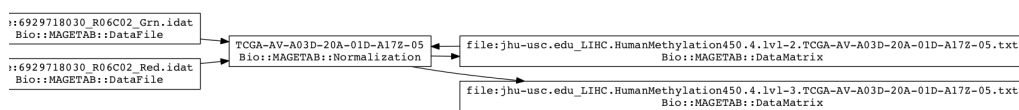
has been shortened to

```
LIHCNRML.4.TCGA-AV-A03D-20A-01D-A17Z-05.txt
```

remaining the most important information.

To guarantee that the results of the quite complex and cpu-intensive experiments on marker-less cancer classification are reproducible, we created a





directory structure as shown in figure 5.7 with all necessary input data and configurations inside (electronic supplement). To launch the processes which perform the time series generation and recurrence analysis to a HPC batch system, the script `run_svm_ch32rqa.sh` has to be executed. The SVM classifications and their evaluation starts running the script `run_svm_eval.sh`. Both has to be done for each cancer type separately.

### DNA methylation density time series

For each sample we construct a series of consecutive chromosomal fragments applying a sliding window which moves from the first base position up to the genomic coordinate of the last data record of the respective chromosome. Within each window we sum the beta values and divide by the window size. The size and the overlap of the windows depend on the chromosome sequence size, the wanted time series resolution and computational power. We use for the analysis of chromosome 1 a window size of 1Mb and move the window by 100000 positions.

During the execution of the experiments presented in this section we have created a total of 61604 time series. Figure 5.8 shows a typical example of a DNA methylation density time series. Characteristic for this kind of time series is the dominant peak at the beginning, a length of roughly 2500 time units and a large horizontal line in the middle.

The noticeably, more or less centric line at zero methylation density level is technically a gap caused by the centromere. But time series gaps are not allowed in the analyses, therefore missing values had to be replaced with a numeric value. We have chosen zero because it is intuitively associated with *nothing*. This produced a systematic bias, because these zeros can be confused with zero methylation density. Though no harm has been done as this bias is present in the same way in all time series it has no major impact on the classification performance.

It is important to note that the values on the abscissa represent the ordered

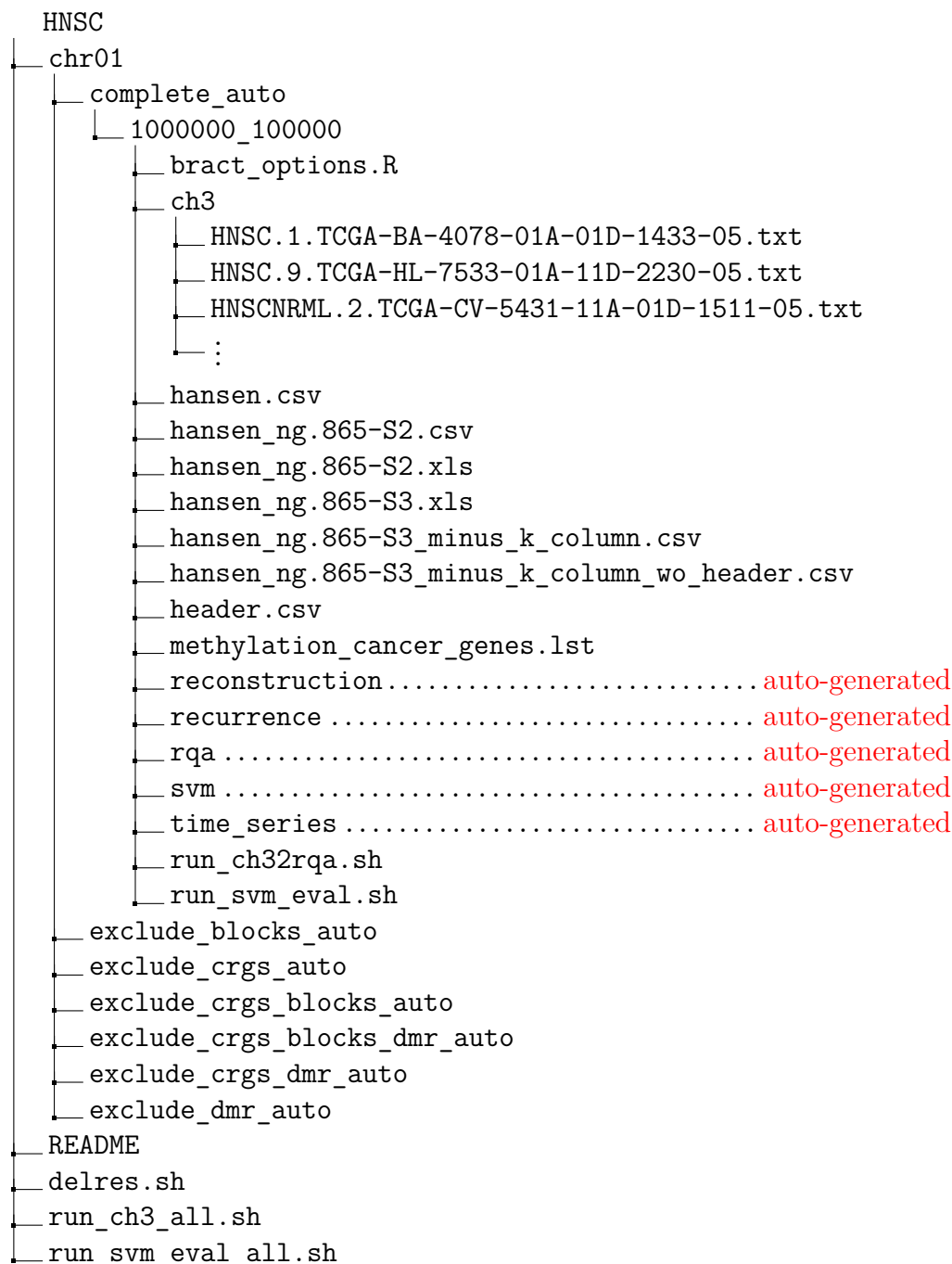


Figure 5.7: Structure of an analysis working directory.

Label	Identifier for	Value	Value description	Possible values
Project TSS	Project name Tissue source site	TCGA BC	TCGA project Liver hepatocellular carcinoma from UNC	TCGA See Code Tables Report <sup>150</sup> → select "Tissue Source Site"
Participant Sample	Study participant Sample type	A10Q 01	– A solid tumor	Any alpha-numeric value Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report <sup>150</sup> for a complete list of sample codes → select "Sample Type"
Vial	Order of sample in a sequence of samples	A	The first vial	A to Z
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	11	The 11th portion of the sample	01-99
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report <sup>150</sup>   select "Portion Analyte"
Plate	Order of plate in a sequence of 96-well plates	A132	The A132nd plate	4-digit alphanumeric value
Center	Sequencing or characterization center that will receive the aliquot for analysis	05	Johns Hopkins / University of Southern California, GCC	See Code Tables Report <sup>150</sup>   select "Center"

Table 5.2: Description of the bar code meta data taken from the bar code page of the NCI wiki<sup>149</sup> and adapted to an example used in this section.

sequence of the window numbers, not the base position on the chromosome. To get the starting base positions of the window on the chromosome one has to add the observation time interval in the following manner:

$$p = 1 + (w - 1)dt \quad (5.5)$$

where  $p$  is a starting position of the window on the chromosome,  $w$  is a window number taken from the abscissa shown in figure 5.8 and  $dt$  is the observation time interval or in other words, how many positions the window has been moved in the density time series generation process.

Another important issue is to remember that the ordinates are a kind of maximum-value-scaled relative densities. Therefore, the y-values shown in figure 5.8 do not allow conclusions to be drawn for the average  $\beta$ -value of the respective windows.

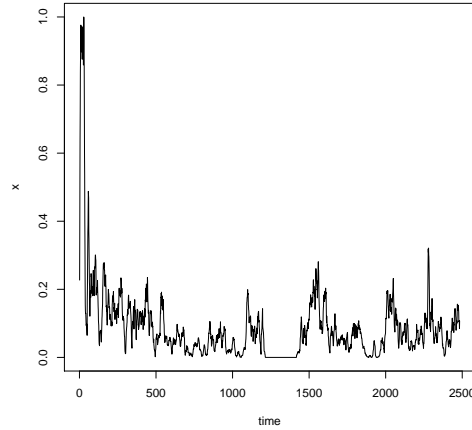


Figure 5.8: DNA methylation density time series from a normal head and neck sample (TCGA-CV-5431-11A-01D-1511-05). The settings used to create this time series have been: window size  $w = 1000000$ , Observation time interval (means by how many positions the window has been moved)  $dt = 100000$ .

### 5.2.2 Time delayed reconstruction

To perform a time delayed reconstruction two parameters, the delay and the embedding dimension, have to be estimated.

To get the delays we have calculated the intrinsic average mutual information index (AMI) considering time lags in a range from 1 to 200 for all 61604 time series. The example in figure 5.9a shows that the dependency of two variables decreases exponentially and then stabilize having an AMI between 0.2 and 0.3. No prominent minimum can be identified. We estimate the delays following the recommendation about the frequently used criteria proposed by Fraser and Swinney<sup>118</sup> taking the time lag which produces the first local minimum of mutual information. In the example the chosen delay is 18.

Considering all delays calculated for the marker-less cancer classification, the values range between 7.00 and 42.00. The mean of the delays is 20.43 and the median is 19.00 (see also table 5.3).

In addition, to get the embedding dimensions used in the time delayed reconstructions, we have calculated, with a custom function from our bract

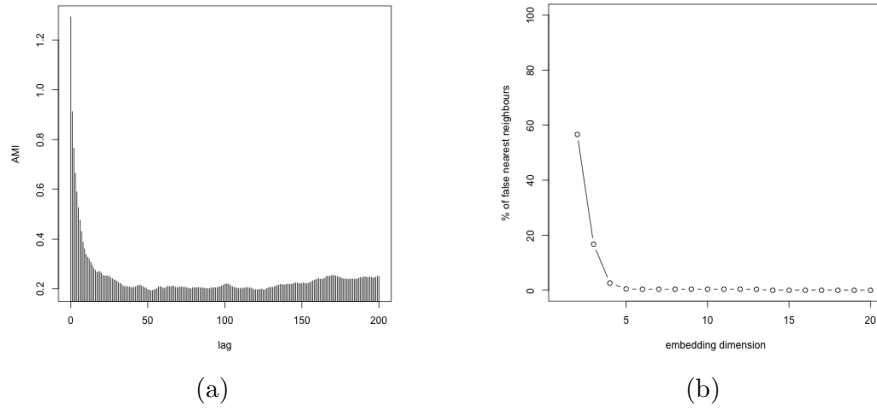


Figure 5.9: Graphical representation of a) the average mutual information for different time lags and b) the percentage of false nearest neighbours for different dimensions calculated for the time series show in figure 5.8.

package, the percentage of false nearest neighbors for each time series following strictly the method described by Kennel et al.<sup>126</sup> examining embedding dimensions in a range from 1 to 20. The example in figure 5.9b shows that this curve also decreases exponentially, but than stabilize having 0% false nearest neighbors. The dimension chosen for the embedding is the first one reaching 0% false nearest neighbors, in this case 14.

Considering all embedding dimensions calculated for the marker-less cancer classification, the values range between 5.0 and 20.0. The mean of the embedding dimensions is 15.5 and the median is 16.0 (see also table 5.4).

We have already seen that the embedding dimensions used to reconstruct the phase spaces in our experiments on marker-less cancer classification are mostly very high. To see whether the phase spaces are regular or even strange attractors are present, we visualize them using multi-scatter plots showing all possible pairs of 2-dimensional projections (see figure 5.10). We also have projected the phase spaces into their first three dimensions and plot them as shown in figure 5.11.

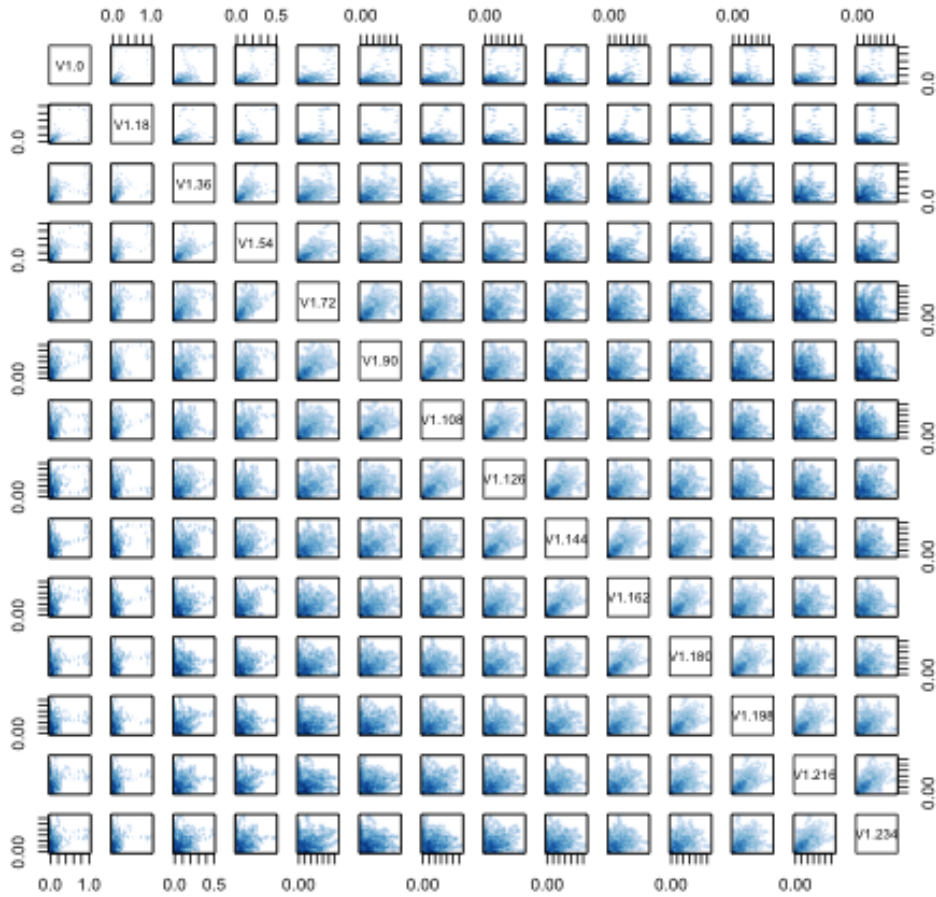


Figure 5.10: Multi scatter plot of a reconstructed phase space for a DNA methylation density time series from a normal head and neck sample (TCGA-CV-5431-11A-01D-1511-05). For this reconstruction a delay  $\tau = 18$  and an embedding dimension  $m = 14$  have been used.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7.00	18.00	19.00	20.43	23.00	42.00

Table 5.3: Statistical summary of all delays calculated for the marker-less cancer classification.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5.0	13.0	16.0	15.5	19.0	20.0

Table 5.4: Statistical summary of all embedding dimensions calculated for the marker-less cancer classification.

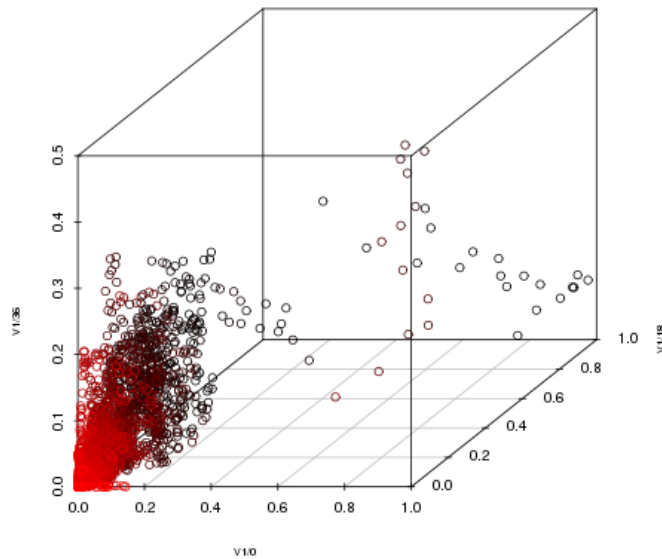


Figure 5.11: Projection of the reconstructed phase space to the first three dimensions for a DNA methylation density time series from a normal head and neck sample (TCGA-CV-5431-11A-01D-1511-05). For this reconstruction a delay  $\tau = 18$  and an embedding dimension  $m = 14$  have been used.

### 5.2.3 Systemic characterization of epigenetic changes in different cancer types: Binary classification of tumor and normal cells

Adaptive Complex Systems (ACS) with no freedom restricts tend to behave dynamically as nonlinear systems exhibiting complex stability landscapes (“strange attractors”) that can be analyzed by a variety of well known approaches<sup>151</sup>. From this perspective, we have considered clonal *methylomas* as “adaptive solutions” of the tumor metabolism drift (here considered as a case of ACS dynamics) and, thus, as a stable setting of the complex, nonlinear system attractor, susceptible of systemic characterization by an appropriate procedural work flow.

In our methodological strategy (summarized in figure 5.1), chromosome-1-DNA-methylation density data, obtained from Illumina Infinium HumanMethylation450 BeadChip ©<sup>152</sup>, were used to create density time series and submitted to an embedding procedure, as proposed by Takens<sup>19</sup>. This embedding procedure is able to provide the reconstruction of topologically equivalent images of their phase space by unfold the system attractor with the fitting delay and embedding dimension (see details in Methods, page 39). We have used this protocol/theorem to reconstruct phase spaces for TCGA samples of 11 different cancer types (with a total sample number of  $n_{cases} = 4363$  and  $n_{controls} = 652$ ): Breast cancer (BRCA), colon and rectal adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), clear cell kidney carcinoma (KIRC), papillary kidney carcinoma (KIRP), prostate adenocarcinoma (PRAD), papillary thyroid carcinoma (THCA) and uterine corpus endometrial carcinoma (UCEC). We have determined delay and embedding (by mutual information<sup>118</sup> and false nearest neighbor methods, respectively<sup>126</sup>) and found that optimum delay ranges between 7 and 42 while embedding dimension typically spans from values of 5 to 20 (table 5.3 and table 5.4).

Recurrence plots (RP) (e.g figure 5.12 (b)) built from the distance-maps (e.g figure 5.12 (a)) of these unfolded phase space images showed tumor specific, consistent pattern that reveal an underlying stochastic structure ostensibly compatible with the existence of deterministic components. In all cases, tumor



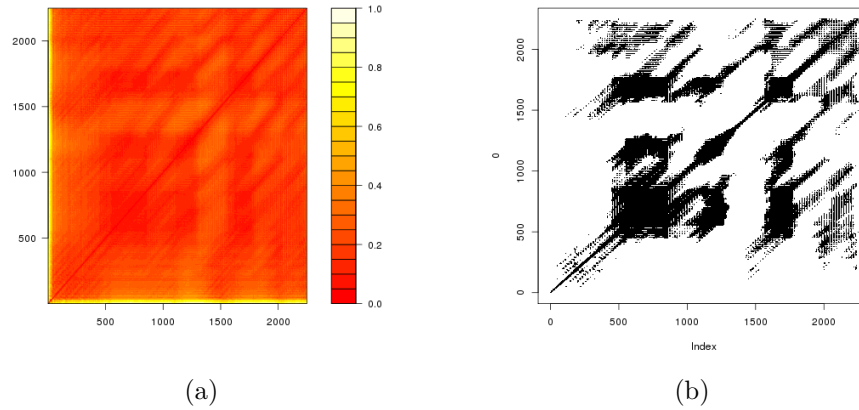


Figure 5.12: a) Distance contour plot and b) recurrence plot ( $\varepsilon = 0.126$ ) for the reconstruction shown in 5.10.

and normal tissues samples showed roughly similar recurrence patterns, although minor differences with no obvious trends can be visualized by differential plots (see figure 5.13 (c)), indicating that the obtained RPs were potentially sensible to epigenetic changes during carcinogenesis.

To explore if embedded RPs of methylation density time series produced can be used to characterize complex and adaptive dynamics of tumor epigenetics, recurrence quantification analysis (RQA)<sup>153,154,128,155–157</sup> was carried out on these images, by measuring six RP standard parameters: recurrence rate (RR), determinism (DET), entropy (ENTR), ratio (RATIO), laminarity (LAM) and trapping time (TT) as well as the lacunarity (LAC) of each sample<sup>131</sup> (see definitions and additional information in Methods). As a representative example, in figure 5.15, the scatter plots show all pair projection of these parameters in six different cancer types.

Heat maps shown in figures 5.14 evidenced again that RQA parameters from tumors are fairly more heterogeneous than those of controls, giving rise to broader clustered dendrograms. This feature can be more clearly seen in the scatter plots of figure 5.15.

We have seen in all cases a markedly greater dispersion of RQA values in tumor tissues when compared to normal (control) tissues (figures 5.15 (a) – (f)). This result not only confirm that methylation changes during carcino-

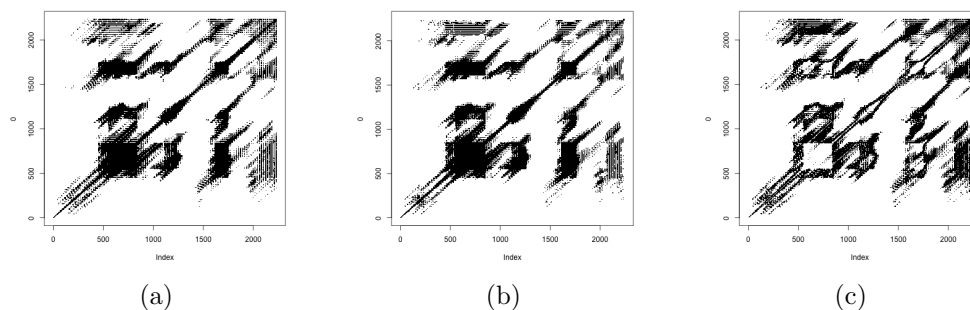


Figure 5.13: Recurrence plots of phase spaces representing a) colon and rectal adenocarcinoma, b) normal colon tissue and c) a difference plot (AJRP) of both.

genesis also increment heterogeneity of epigenetic patterns in tumor tissues, as previously reported<sup>64</sup>, but also indicates that our RQA protocol is able to capture this differences. In some cases, like in endometrial samples, normal tissues tend to show one only cluster in all projections (figure 5.15 (f)). More frequently, however, normal tissues display more than one definite spot (figures 5.15 (e)). Current data situation make impossible to decide whether this tendency to exhibit multi-cluster patterns is due to the cellular heterogeneity of samples (organ biopsies) or due to the existence of alternative, stable epigenetic landscapes. In tumor tissues, the observed dispersion is mainly diffuse or, at least, not as definite as in normal tissues, occupying wider regions of the phase space projections and showing in almost all cases centroids neatly different. Spots pattern are strikingly more similar when tumors affect the same organ. That is the case in the lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) (figures 5.15 (a) and (b)) and in clear cell kidney carcinoma (KIRC) and papillary kidney carcinoma (KIRP) (figures 5.15 (c) and (d)). These features are consistent with the epigenetic specificity that would be presumably in an adaptive scenario with different tumors and tissues.

We used these seven-dimensional RQA vectors as the starting point for training a machine learning algorithm based on support vectors and built binary classification models for each of the studied cancer types. Table 5.5 summarize number distribution of samples for the different data sets. For all data sets the number of tumor samples is notably larger than the number of normal samples.

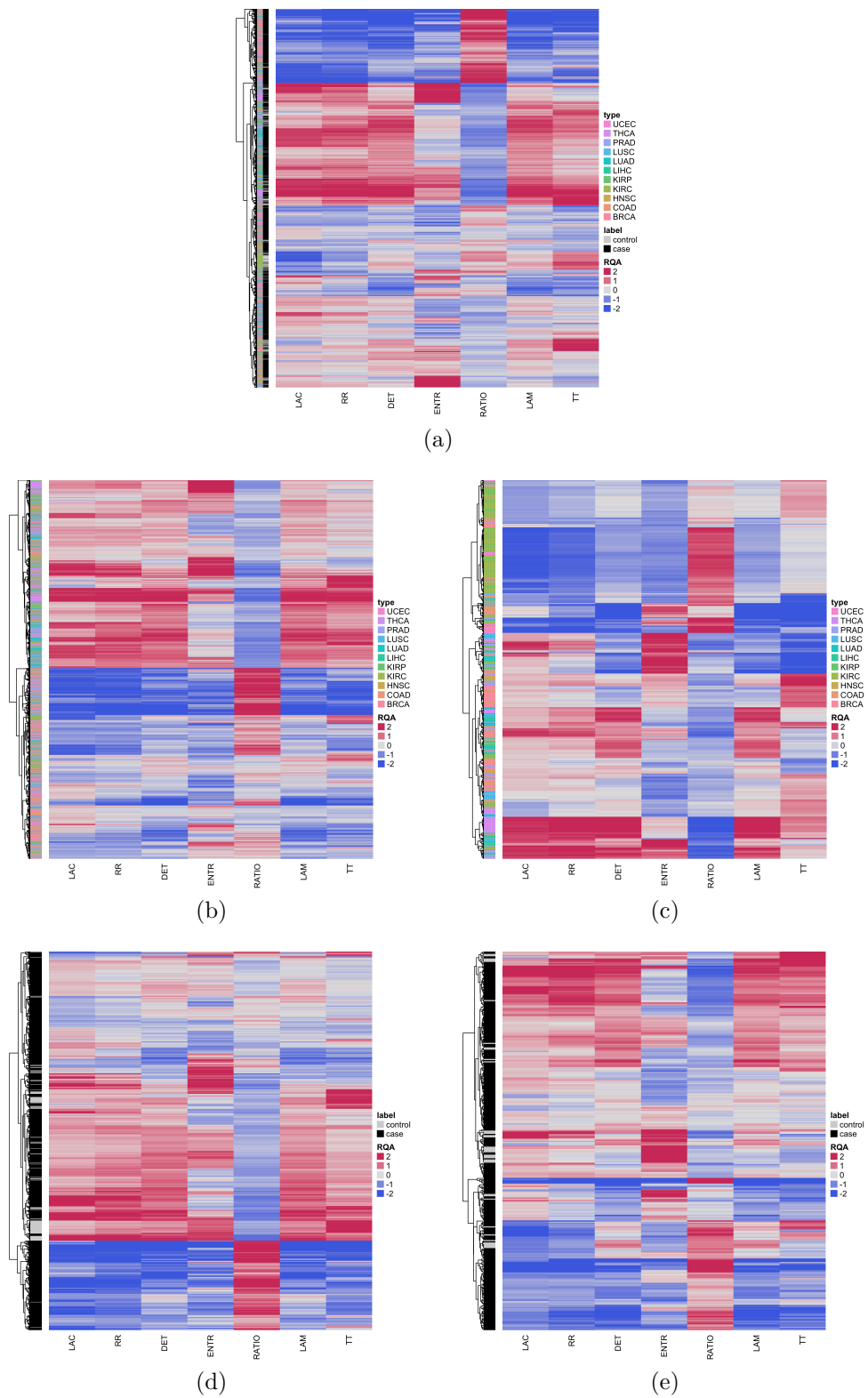


Figure 5.14: Cluster analysis of RQA measures and lacunarity showing a) all cases and controls for all cancer types, b) cases for all tumor types, c) controls for all cancer types and cases and controls for each tumor d) BRCA, e) COAD, f) HNSC, g) KIRC, h) KIRP, i) LIHC, j) LUAD, k) LUSC, l) PRAD, m) THCA and n) UCEC.

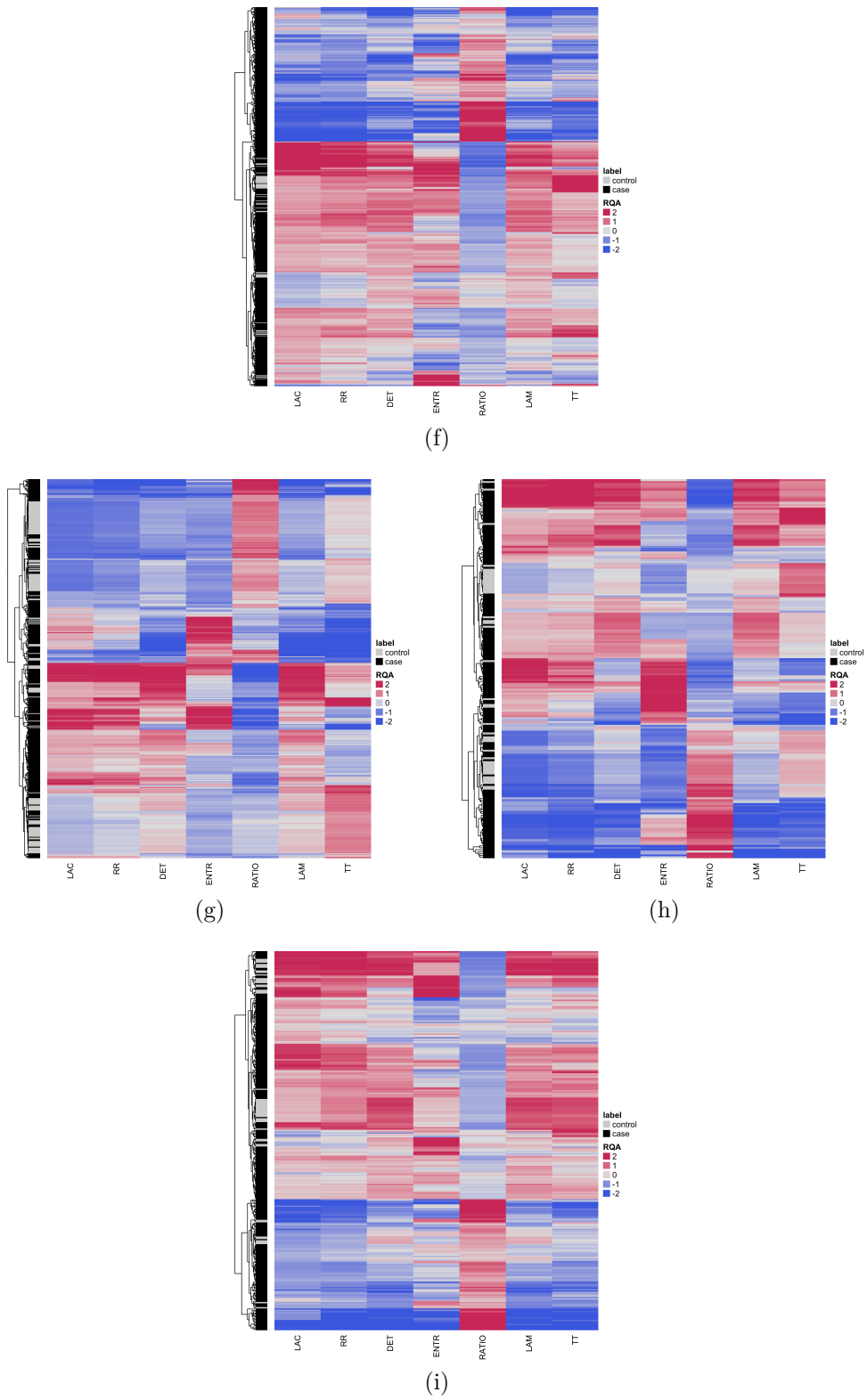


Figure 5.14: Cluster analysis of RQA measures and lacunarity. Continued

...

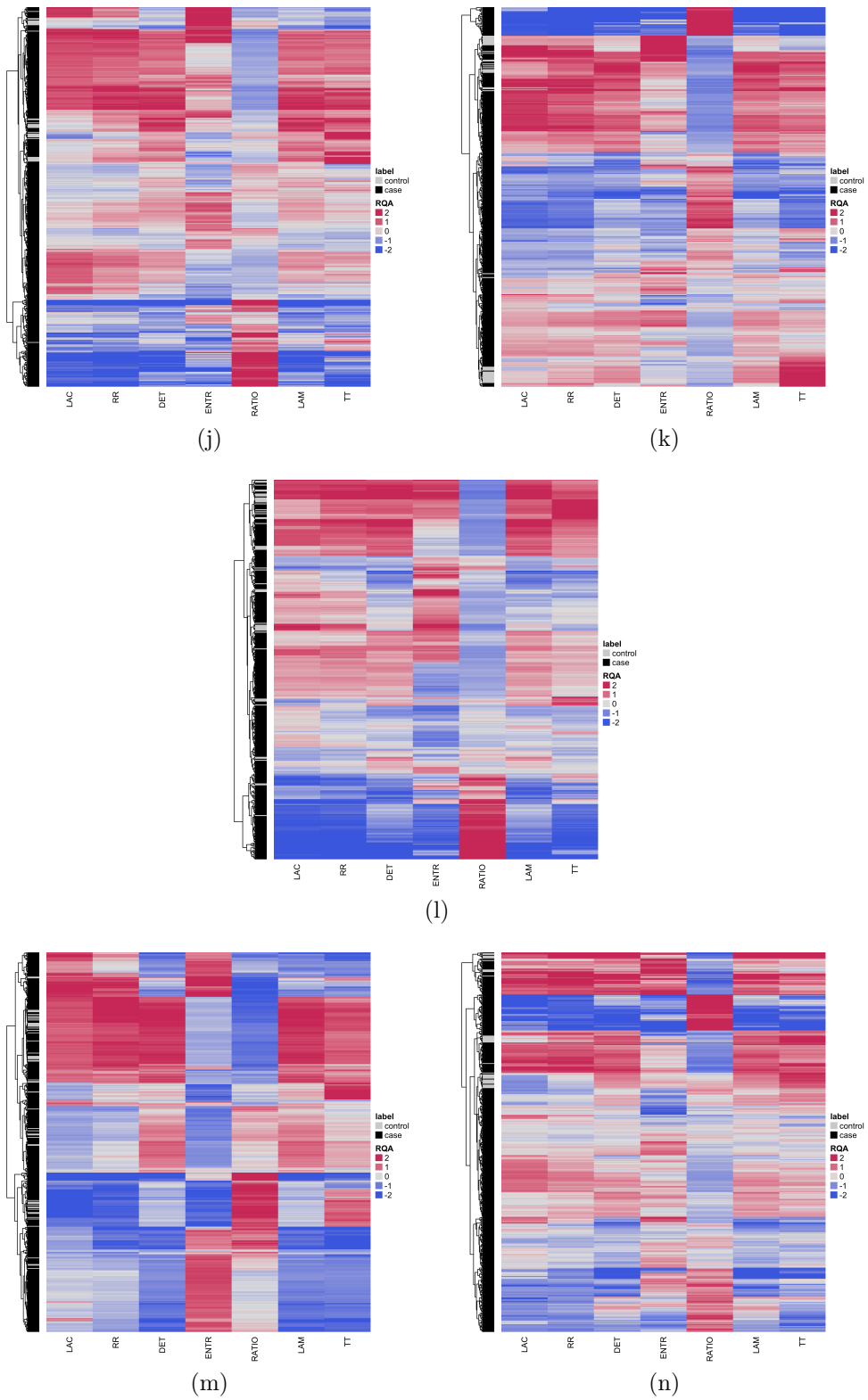
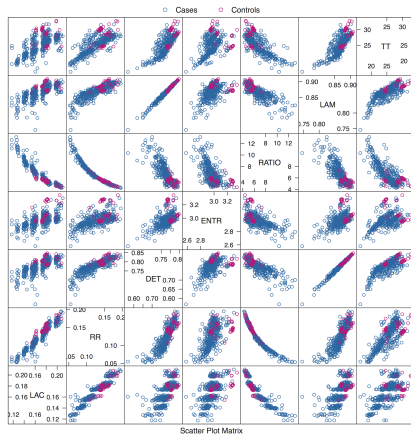
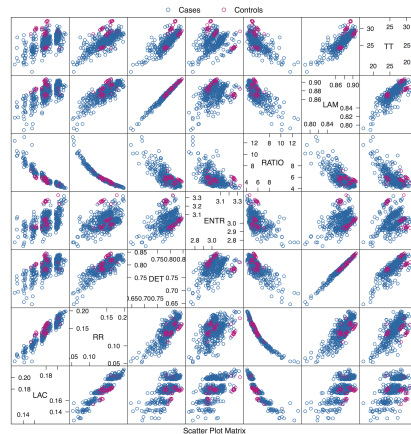


Figure 5.14: Cluster analysis of RQA measures and lacunarity. Continued

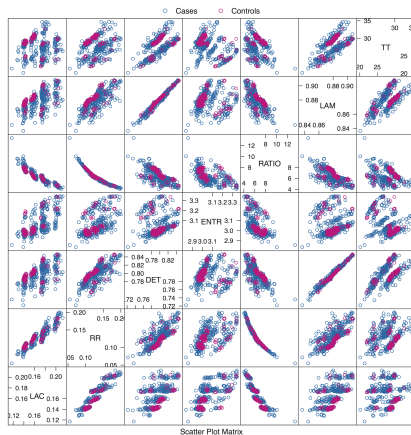
...



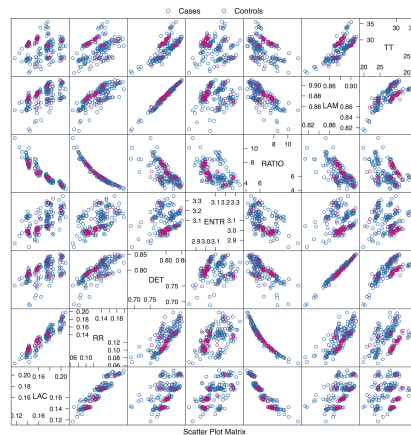
(a)



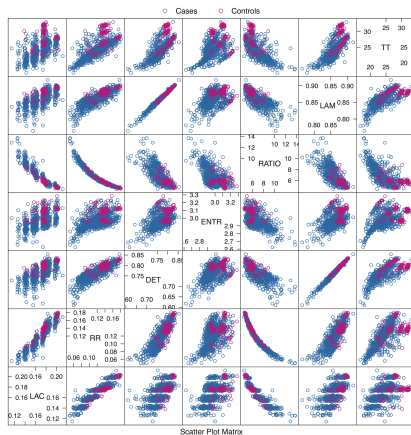
(b)



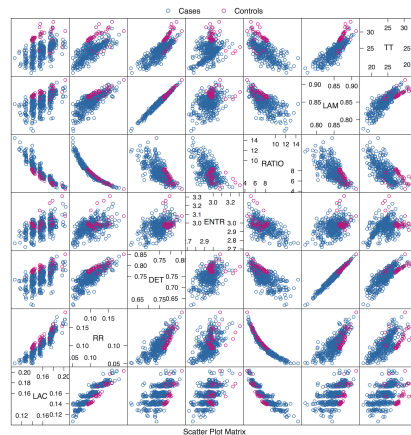
(c)



(d)



(e)



(f)

Figure 5.15: Pairwise comparison of RQA measures and lacunarity showing cases and controls from the tumor types a) LUSC, b) LUAD, c) KIRC, d) KIRP, e) BRCA and f) UCEC.

Table 5.5: Sample number distribution for the different data sets used in the tumor versus normal tissue classification on systemic features and complete CpG sets.

	brca	coad	hnscc	kirc	kirp	lihc	luad	lusc	prad	thca	ucec
tumor matched	725	291	530	301	156	257	465	359	340	515	424
normal matched	96	38	50	160	45	50	32	42	49	56	34
TM training	566	232	420	240	123	196	361	284	268	331	326
NM training	76	30	39	128	36	40	25	33	39	44	27
TM testing	142	58	106	60	31	49	91	72	68	83	82
NM testing	20	8	10	32	9	10	7	9	10	11	7

Table 5.6: Areas under the curve (AUC) in % for marker-less classifications on complete and reduced DNA methylation data sets of different cancer types versus their corresponding normal samples. BLOCKS, CRGS, cDMR and their combinations have been eliminated from the original data sets. From BLOCKS and cDMR we have considered those CpG-sites which match the HumanMethylation450 chip and their evidence threshold (Watson strand, cancer) is  $E > 0.5$ .

Cancer type	All CpG sites	BLOCKS	CRGS	CRGS & BLOCKS
BRCA	90.1	95.3	97.3	95.6
COAD	87.9	97.0	92.2	97.0
HNSC	96.4	83.2	94.5	91.0
KIRC	95.5	91.3	88.6	88.4
KIRP	92.1	99.6	91.3	80.8
LIHC	84.5	95.9	97.8	97.4
LUAD	98.1	96.3	98.1	99.3
LUSC	99.7	98.4	100.0	100.0
PRAD	89.9	85.4	92.5	88.6
THCA	74.6	88.8	93.4	77.1
UCEC	98.6	98.7	100.0	92.2

Table 5.7: Areas under the curve (AUC) in % for marker-less classifications. Continued . . .

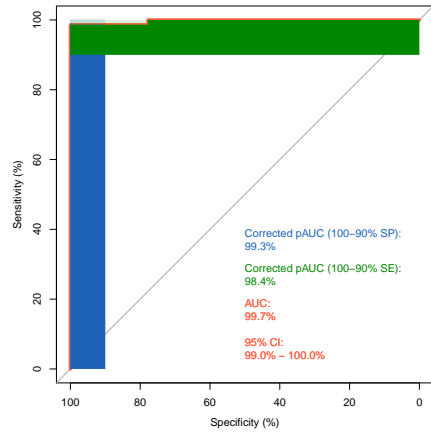
Cancer type	CRGS & BLOCKS & cDMR	CRGS & cDMR	cDMR	
BRCA		83.4	94.9	95.6
COAD		94.7	90.0	76.3
HNSC		81.0	90.6	90.7
KIRC		86.4	90.2	92.2
KIRP		80.0	91.6	78.5
LIHC		93.8	88.5	87.3
LUAD		98.3	93.9	98.5
LUSC		89.7	93.0	79.9
PRAD		87.3	86.6	79.7
THCA		85.9	79.5	65.5
UCEC		89.5	98.0	96.9

When using all CpG site methylation density time series from chromosome I, we could successfully train and predict tumor and normal tissues of the embedded RQA in all cases. Every tumor except one (THCA, 74.6%), showed performances that were, in terms of AUCs, on or above 84.5% and most of them (seven cases from eleven) were 90% or better (Table 5.6, column "All CpG sites" and figure 5.16).

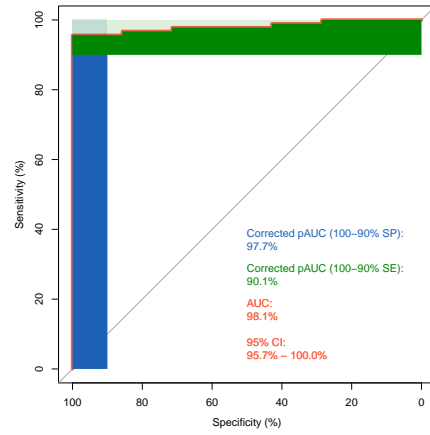
If tumor specific alterations of DNA-methylation were systemic then we should still be able to classify the samples with similar performance though we exclude cancer related CpG sites from the original data sets. To test this assumption we first compiled CpG sites that have been previously associated and documented with cancer related gene symbols (CRGS)<sup>158-165</sup> (electronic supplement). Again, chromosome I was used in this study as it holds the highest number of such sites. After applying the same work flow on the reduced data sets (excluding 6386 CpG sites or 16,65%) we obtained indeed mainly comparable results.

We have extended this rule out analysis to other two types of specific cancer related CpG regions, known to undergo well defined methylation changes during carcinogenesis: BLOCKS (large, up to several Mb, blocks of hypomethylation<sup>64</sup>) and cDMRs (regions of cancer specific DNA methylation variation<sup>64</sup>). In summary, the exclusion of one single or a combination of these sections (CRGS, BLOCKS or cDMR), our SVM-based classification approach still worked properly, preserving (or even improving) in most of cases its predictive performance (Table 5.6 and 5.7 and electronic supplement). This resilience cannot be explained by classical statistical correlations.

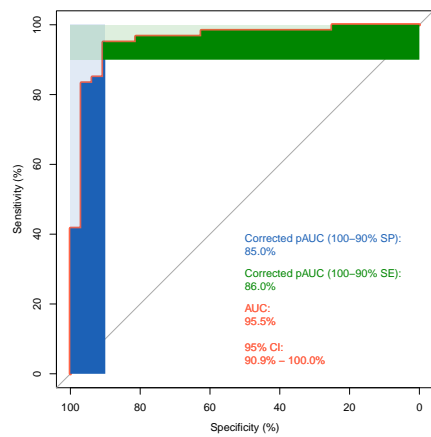




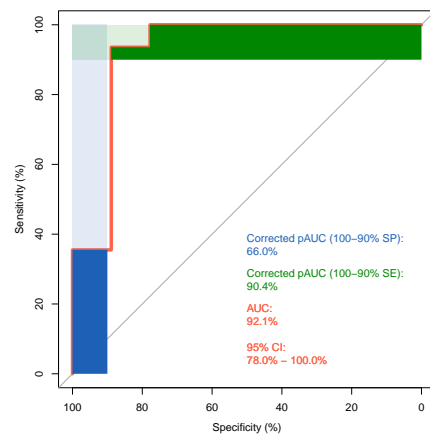
(a)



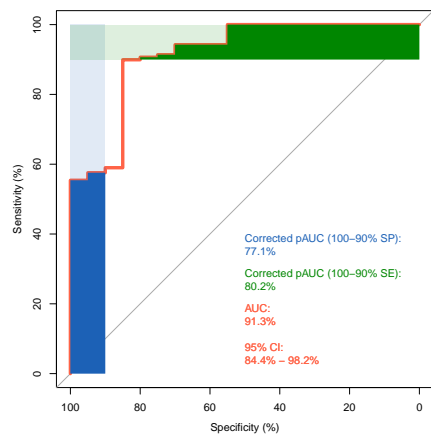
(b)



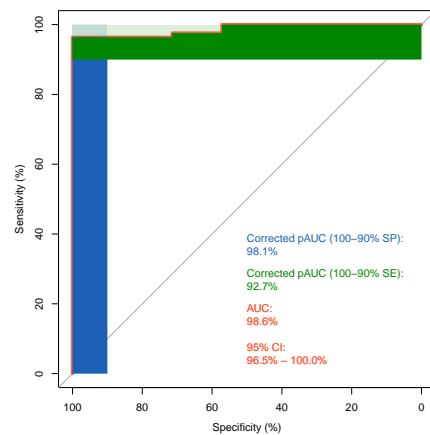
(c)



(d)



(e)



(f)

Figure 5.16: ROCs and AUCs used to evaluate the binary classifications of complete data sets for a) LUSC, b) LUAD, c) KIRC, d) KIRP, e) BRCA and f) UCEC

### 5.2.4 Comparison of tumor versus normal classification on CpG sites with cancer specific differential methylation and systemic features

To see if there is any difference between the performance of marker (based on differential DNA-methylated CpG sites) and marker-less (systemic) predictions we compared both methods each other on the classification of head and neck tumor and normal tissues. For that purpose we first had to ensure that both kind of classifications run under comparable conditions. Therefore, for all results shown here we have reused the same HNSC samples mentioned above – more precisely  $n_{cases} = 403$  and  $n_{controls} = 40$  for training and  $n_{cases} = 101$  and  $n_{controls} = 10$  for testing. We chose head and neck cancers (HNSC) to compare both methods because on the one hand we needed many samples to train our supervised method and on the other hand was a large number of markers an advantage to test various combinations.

We classified the original TCGA data with a support vector machine on CpG sites with head and neck cancer specific differential DNA-methylation taken from the supplementary table 7 published by Fernandez et al.<sup>161</sup>. We selected only those cancer specific differential DNA-methylation CpG sites which are also present in the Illumina Infinium HumanMethylation450 BeadChip © - the data type we have used for systemic classifications - and having an average  $\beta$ -value assigned ( $n = 54$ ). To observe roughly how the number of CpG sites bias the performance, we run classifications with different data sets ( $m_8, m_{15}, \dots, m_{23}, m_{54}$ ) varying the number of hypermethylated CpG sites ( $m$ ) and using always all (eight) differential hypomethylated CpG sites. Therefore  $m_{23}$  is a marker data set with eight hypomethylated and ,from the list, the first 15 mappable sites having an average  $\beta$  value assigned in the TCGA data set. The binary classifications performed, as expected, with  $AUCs > 94\%$ , which means simplified, the markers archive their function.

It turns out that the classification with 18 "markers" gives a slightly better result, having an  $AUC_{m18} = 97.4\%$ , than the others which range between  $AUC_{m8} = 94.5\%$  and  $AUC_{m17} = 96.9\%$  (electronic supplement). Details of the best result and its comparison to the marker-less performance we show in figure 5.17. In summary, the performance of the marker-less classification

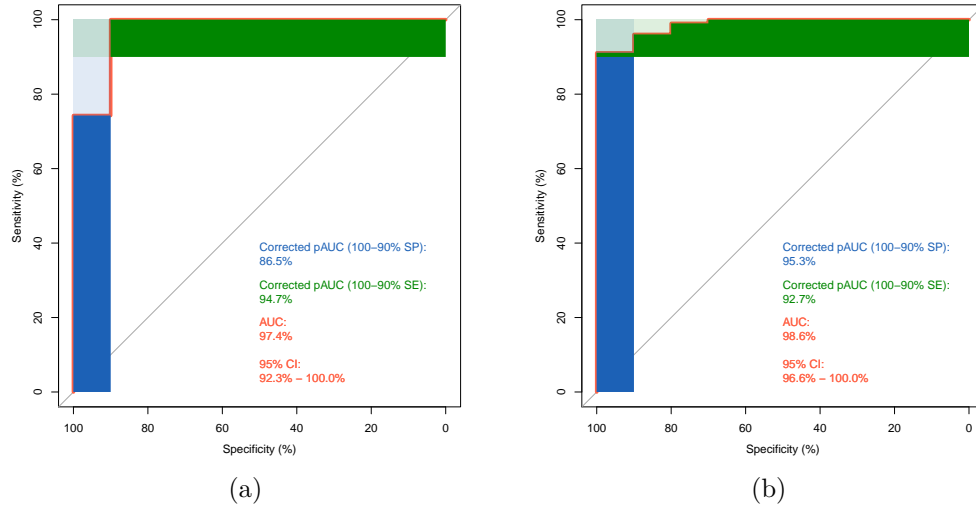


Figure 5.17: Performance of a marker-based (a) and a marker-less (b) head and neck cancer ( $n_{cases} = 106$ ) versus normal ( $n_{controls} = 10$ ) tissue SVM binary classification.

(AUC: 98.8%) is comparable or even slightly better than the best one obtained for marker-based classifications (AUC: 97.4%).

Finally, we tested also if there is a difference from which part of the list we take the marker CpG sites. This was not the case because reverting the order of the CpG sites in the list, we obtain similar results (electronic supplement). This time, we have obtained the best performance using the combination of nine hypo- and eight hypo-methylated CpG sites ( $AUC_{mr17} = 97.4\%$ ).

### 5.2.5 Tumor versus normal classification on randomly selected non cancer specific DNA-methylation sites.

The omnipresence of variations in DNA-methylation typically observed in tumor cells compared to controls – specially the global hypomethylation – has give us reason to test if combinations of  $\beta$ -values from arbitrary non cancer specific CpG sites are suitable to classify both kind of tissues. From the 395885 measured CpG sites of the Illumina Infinium HumanMethylation450 BeadChip © data for

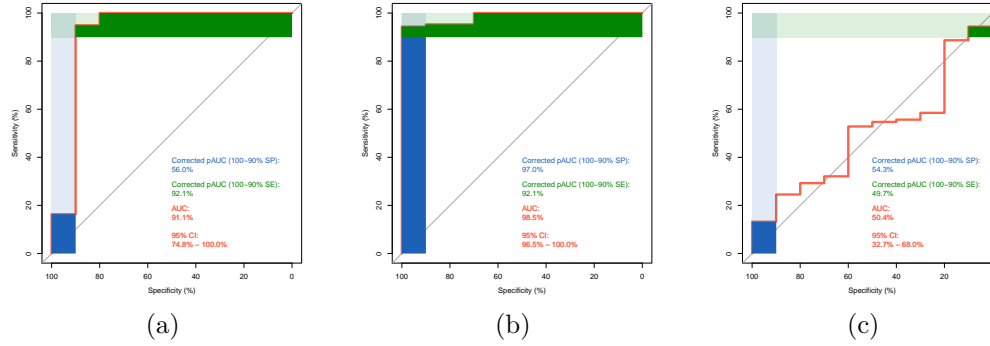


Figure 5.18: ROC curves and AUCs for head and neck tumor ( $n_{cases} = 98$ ) versus normal ( $n_{controls} = 10$ ) tissue SVM binary classification based on DNA-methylation data from TCGA. a) 100 non cancer specific CpG sites selected randomly from CpG sites of the Illumina Human-Methylation450 bead chip, b) 18 non cancer specific CpG sites selected randomly and c) the same 18 non cancer specific CpG sites, but trained on an randomly shuffled data set.

head and neck normal sample TCGA-CV-7263-11A-01D-2014-05 we exclude 71701 sites labeled with around 3210 different cancer related gene symbols. From the remaining 324184 supposed non-cancer specific CpG sites we select randomly 100 and 18 CpG sites, respectively, to use them as "pseudo-markers" in head and neck tumor versus normal tissue classifications.

We found that 18 pseudo markers ( $AUC_{pm18} = 98.5\%$ ) predict better than 100 ( $AUC_{pm100} = 91.1\%$ ) giving a predictive capability comparable to the cancer specific differential DNA-methylation CpG sites shown above (see figure 5.18).

### 5.2.6 Tumor versus normal classification on global $\beta$ -value sums

The total levels of DNA methylation decrease within repeat sequences, pericentromeric regions of chromosomes, some genes and most notably in megabase-large regions distributed over all chromosomes<sup>166,167</sup>. Our question was, if it would be possible to capture this phenomenon by a simple measure like the total  $\beta$ -value sums. We have compared distributions of  $\beta$ -value sums in tumor samples to distributions in normal samples under different conditions.

The box plots in figure 5.19 (a) show that the  $\beta$ -value sums, including all CpG sites, are for most samples relatively similar and difficult to discriminate.

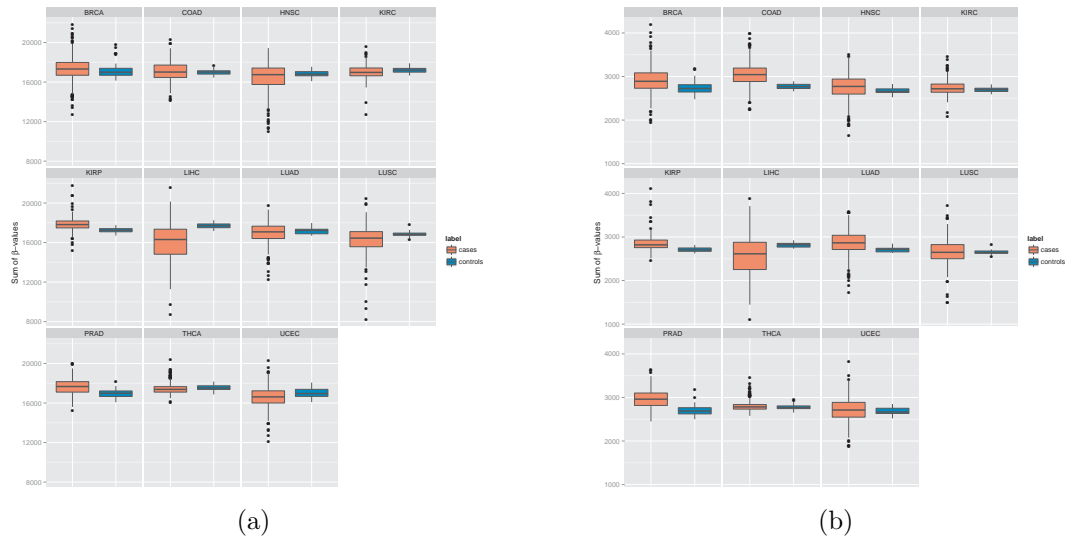


Figure 5.19: Boxplots of a)  $\beta$ -value sums including all CpG sites and b)  $\beta$ -value sums including CpG sites located only in regions belonging to cancer related gene symbols.

This is also supported by the p-values from the Wilcoxon rank sums test shown in the column *All CpG* of the table 5.8. However, tumor and normal samples for the experiments labeled with KIRP, LIHC and PRAD can be separated clearly.

If we concentrate our observations only to BLOCKS<sup>64</sup>, then the  $\beta$ -value sums are, as expected, sufficient to separate tumor and normal tissues in almost all cases. The box plots in figure 5.20 show clear separated medians and also the p-values calculated by the Wilcoxon rank sums test demonstrate the differences. An exception is again PRAD which is the only cancer type which can not be distinguished from normal tissues. The p-values range from  $9.07e^{-01}$  to  $2.58e^{-23}$  in the case that all CpG sites which match BLOCKS and the Illumina Infinium HumanMethylation450 BeadChip © are considered in the calculus. If we apply a evidence threshold  $E > 0.5$  for the BLOCKS, the p-values range from  $8.24e^{-01}$  to  $8.94e^{-24}$  (see also table 5.8).

Calculating  $\beta$ -value sums including only CpG sites located in regions belonging to cancer related gene symbols (CRGS) or CpG-sites situated a union of CRGS & BLOCKS & cDMR shows that in the half of the cases tumor tissues can be distinguished from normal ones (see figures 5.19 (b) and 5.21 and

Table 5.8: Wilcoxon rank sum test with continuity correction on  $\beta$  value sums for complete and reduced DNA methylation data sets of different cancer types versus their corresponding normal samples. BLOCKS, CRGS, cDMR and their combinations have been analyzed excluding other CpG-sites from the original data sets. From BLOCKS and cDMR we have considered those CpG-sites which match the HumanMethylation450 chip and their evidence threshold (Watson strand, cancer) is  $E > 0.5$ .

Type	All CpG	blocks	crags	crags & blocks & cdmr	crags & cdmr	cdmr
brca	1.89e-02	2.46e-12	5.79e-09	9.37e-01	3.79e-37	1.15e-36
coad	8.53e-01	3.24e-10	1.00e-10	1.32e-05	1.59e-18	2.80e-18
hnsk	4.54e-01	1.17e-11	1.14e-03	7.29e-01	6.21e-24	2.99e-23
kirc	1.32e-04	8.94e-24	4.91e-02	1.75e-01	7.24e-40	4.09e-36
kirp	1.53e-09	5.16e-06	2.40e-09	2.18e-07	9.18e-10	8.25e-12
lihc	2.52e-10	6.01e-16	6.43e-05	2.14e-07	5.69e-10	3.67e-08
luad	6.42e-01	2.10e-06	1.21e-05	1.66e-03	2.86e-16	8.70e-16
lusc	2.46e-03	4.98e-13	8.57e-01	1.64e-03	5.54e-20	1.01e-19
prad	3.04e-07	8.24e-01	1.56e-12	5.67e-06	1.04e-17	1.57e-17
thca	9.92e-03	2.34e-03	2.04e-01	2.46e-03	1.63e-06	1.90e-05
ucec	1.37e-02	9.64e-12	4.54e-01	6.61e-03	2.15e-15	1.22e-15

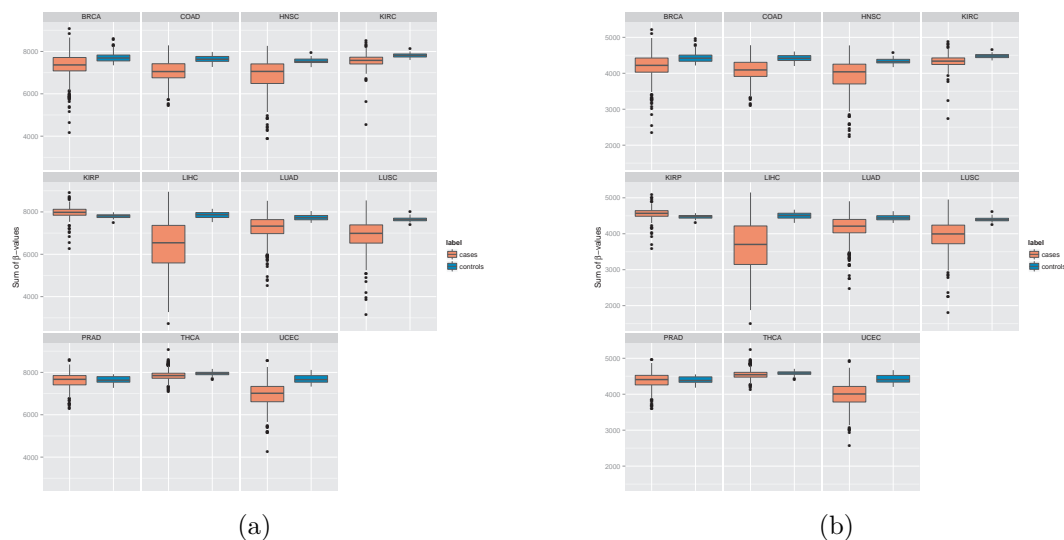


Figure 5.20: Boxplots of  $\beta$ -value sums including CpG sites located only in BLOCKS. In a) we have considered all CpG-sites which match the BLOCKS and the HumanMethylation450 chip and in b) we have selected only those from the BLOCKS having an evidence (Watson strand, cancer)  $E > 0.5$



Figure 5.21: Boxplots of  $\beta$ -value sums including CpG sites located only in regions belonging to cancer related gene symbols, BLOCKS and cDMR. In a) we have considered all CpG-sites which match the BLOCKS or cDMR and the Illumina Infinium HumanMethylation450 BeadChip © and in b) we have selected only those from the BLOCKS and cDMR having an evidence (Watson strand, cancer)  $E > 0.5$

table 5.8).

An outstanding differentiation between tumor and normal tissues we have obtained applying the  $\beta$ -value sums to CRGS & cDMR or cDMR only, which are known to be hypermethylate. The p-values calculated by the Wilcoxon rank sums test range from  $1.90e^{-05}$  to  $1.57e^{-41}$  on a union of values from both kind of experiments (see again table 5.8). The performance is also supported visually by clear separated medians shown in the box plots of the figures 5.22 and 5.23. These results are conform with the findings reported in Hansen et al. <sup>64</sup>.

The AUCs obtained by the evaluation of the tumor versus normal tissue classification on  $\beta$ -value sums are conform with our observations described above.  $\beta$ -value sums including all CpG sites, BLOCKS, CRGS, CRGS & BLOCKS and CRGS & BLOCKS & cDMR perform, with single exceptions, relatively poor having median AUCs around 70%. Here again the groups CRGS & cDMR and cDMR differ from the other results and show a markable performance having a median closed to 92%. Only the classification of THCA fails completely us-

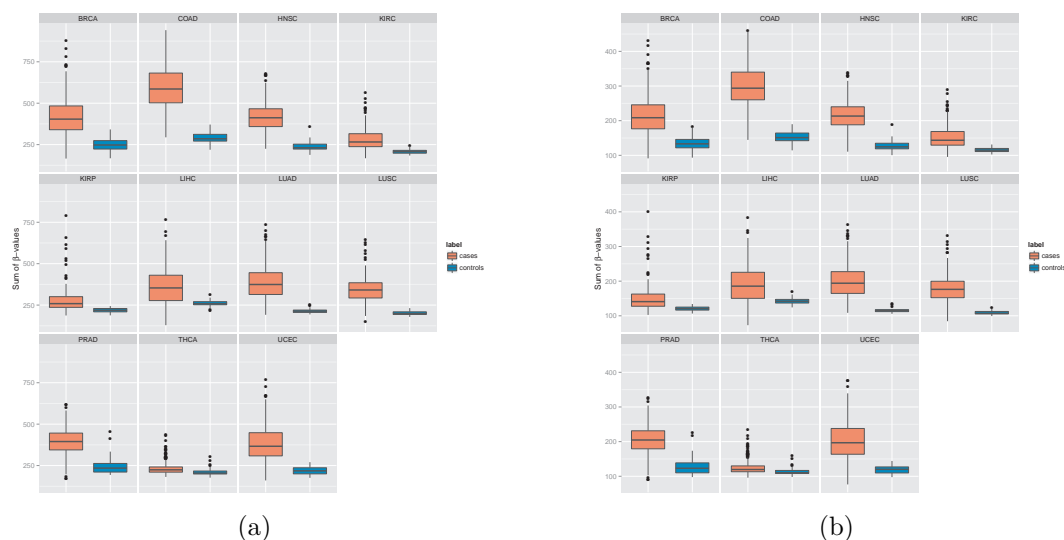


Figure 5.22: Boxplots of  $\beta$ -value sums including CpG sites located only in regions belonging to cancer related gene symbols and cDMR. In a) we have considered all CpG-sites which match the cDMR and the Illumina Infinium HumanMethylation450 BeadChip © and in b) we have selected only those from the cDMR having an evidence (Watson strand, cancer)  $E > 0.5$

ing support vector machines in all experiments based on  $\beta$ -value sums . All calculated AUCs based on  $\beta$ -value sum -based classifications and a statistical summary we show in table 5.9.

For us these experiments have been important, because they show that the effects or signals of hyper- and hypomethylation along a whole chromosome cancel each other out and can not be detected by a simple global  $\beta$ -value sum without prior knowledge on hyper- and hypomethylated regions.









Figure 5.23: Boxplots of  $\beta$ -value sums including CpG sites located only in cDMR. In a) we have considered all CpG-sites which match the cDMR and the Illumina Infinium HumanMethylation450 BeadChip © and in b) we have selected only those from the cDMR having an evidence (Watson strand, cancer)  $E > 0.5$

Table 5.9: Areas under the curve (AUCs) in % for the  $\beta$ -value sum classifications on various DNA methylation data sets of different cancer types versus their corresponding normal samples. BLOCKS, CRGS, cDMR and their combinations have been analyzed excluding other CpG-sites from the original data sets. From BLOCKS and cDMR we have considered those CpG-sites which match the HumanMethylation450 chip and their evidence threshold (Watson strand, cancer) is  $E > 0.5$ .

Cancer type	All CpG sites	blocks	crgs	crgs & blocks & cdmr	crgs & cdmr	cdmr
BRCA	69.48	74.34	54.1	77.28	92.07	92
COAD	77.97	80.3	81.14	82.84	97.46	96.61
HNSC	70.57	68.49	65.09	83.02	91.6	87.08
KIRC	72.34	81.2	64.5	64.14	97.44	96.36
KIRP	89.93	85.42	92.01	85.76	89.24	95.83
LIHC	91.54	80.96	93.27	95.96	58.85	77.88
LUAD	60.68	70.35	84.79	60.06	95.24	93.55
LUSC	56.17	96.91	65.28	66.82	98.46	98.46
PRAD	56.32	65.88	84.26	67.79	90.74	90.74
THCA	57.85	56.23	65.29	54.53	53.56	58.25
UCEC	75.63	93.11	64.37	81.68	100	98.99

## 5.3 Functional and gene position distances in the tomato genome

Lieberman-Aiden et al.<sup>168</sup> have shown that the local packing of chromatin is comparable with a fractal globule. They support this finding, for instance, showing a power law in measured and simulated contact probabilities as a function of genomic distances. In analogy, we have been interested to investigate if a similar behavior can be found observing protein functional distances as a function of gene base position distances in the tomato genome.

A prerequisite to investigate the relations between functional and gene position distances in tomato (*Solanum lycopersicum*) has been the availability of the chromosome sequences together with their structural and functional annotation. We, the "Plant Computational Biology"-group (PCB) at the Max-Planck-Institute for Plant Breeding Research, as part of the International Tomato Annotation Group (ITAG) and the Tomato Genome Consortium, have contributed the gene ontology (GO)<sup>169</sup> annotation and human readable descriptions<sup>170</sup> for protein sequences<sup>1</sup>. To perform the GO-term assignment we used, inter alia, the custom annotation platform  MANOS  <sup>1 2</sup> (electronic supplement) to launch the steps of our annotation protocol shown in figure 5.24 to our batch system. Interpro2go<sup>171,172</sup>, blast2go<sup>173</sup> and the in-house PhyloFUN<sup>174,175</sup>, an extension of the sifter pipeline<sup>176,177</sup>, have been used to map the GO annotation to the protein sequences. To avoid GO-term incongruities between the three methods,  MANOS  has been designed in the way that it synchronized and updated the outcomes of the integrated methods to the latest gene ontology version before the final results has been compiled and submitted to the ITAG repository. For the official ITAG2.3 tomato genome annotation release we have annotated 19662 or 57 % of 34727 tomato proteins with 39192 GO terms. 2108 GO terms of them are unique<sup>1</sup>. The gene models (including our GO annotation) of the ITAG 2.3 release are publicly available to download on the MIPS and SGN web sites<sup>178,179</sup>.

Taking the annotation, we have produces gene and functional distance pairs

---

<sup>1</sup>Developed specifically for this project by the author of this memory – Jens Warfmann.

<sup>2</sup>Logos designed and drawn by Alexander, Noelia and Irene

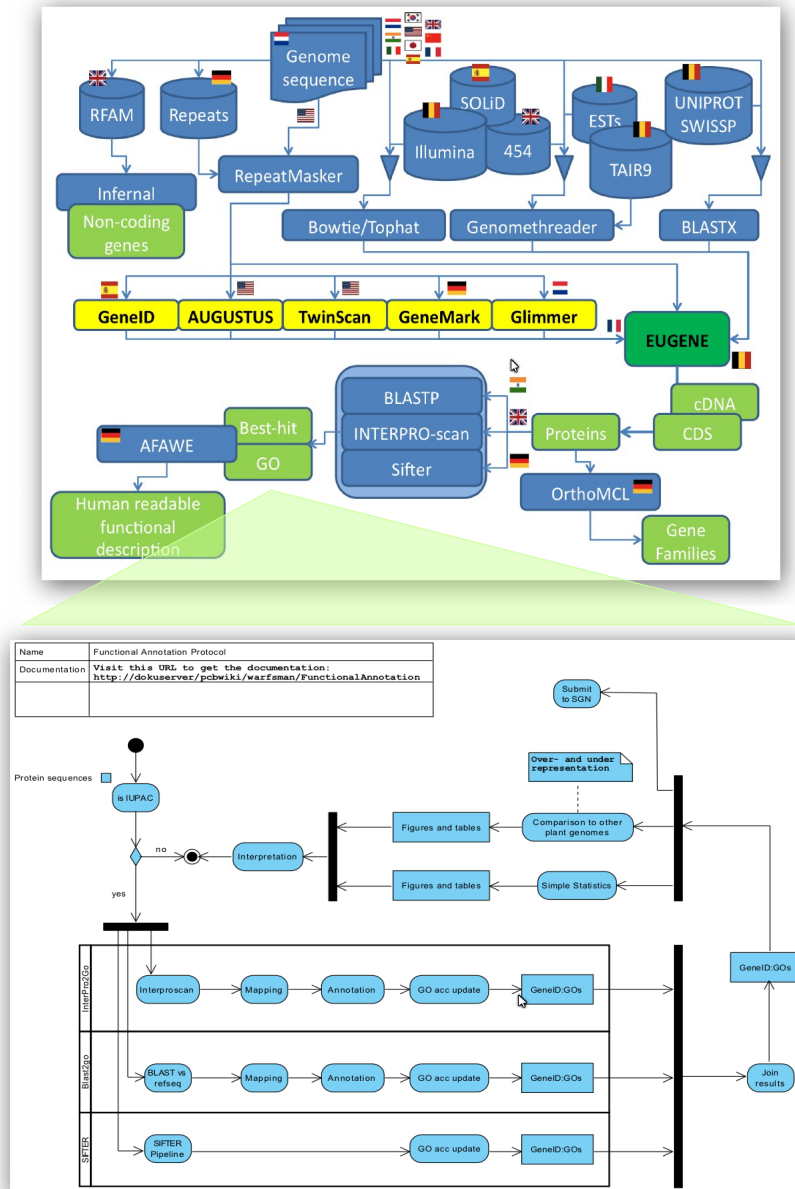


Figure 5.24: The ITAG annotation pipeline<sup>1</sup> and the activity diagram of the PCB gene ontology annotation protocol.

### 5.3. FUNCTIONAL AND GENE POSITION DISTANCES IN THE TOMATO GENOME

for all 12 tomato chromosomes with a custom script from our bract package (electronic supplement) as follows:

```
funseqdistS.R "ITAG2.3\_gene\_models.gff3"
```

In this experiment we have restricted our focus to genes located on the + strand only. We have calculated the distance of two genes as follows:  $D_n = p_{n+1}^{start} - p_n^{end}$ . In the case that the annotations overlap we apply the formula:  $D_n = p_{n+1}^{start} - p_n^{start}$ . In both equations D is the distance measured in base positions and p the position where the gene starts or ends. The functional distances we have calculated using the R-package GOsemSim<sup>180</sup> which combines information content (IC) and graph-based methods. It should be mentioned that the calculation of all versus all chromosomal functional distances (more than  $9e + 08$  combinations) are very cpu-intensive (and can take more then a week running on a state of art HPC-cluster). Therefore we have restricted the data sets selecting randomly 100000 samples of genes pairs from the tomato annotation.

Table 5.10: Here we show the functional distances as a function of inter gene distances of the + strand for the tomato chromosomes 1 – 12. From all possible gene pair combinations we toke here randomly 100000 samples.

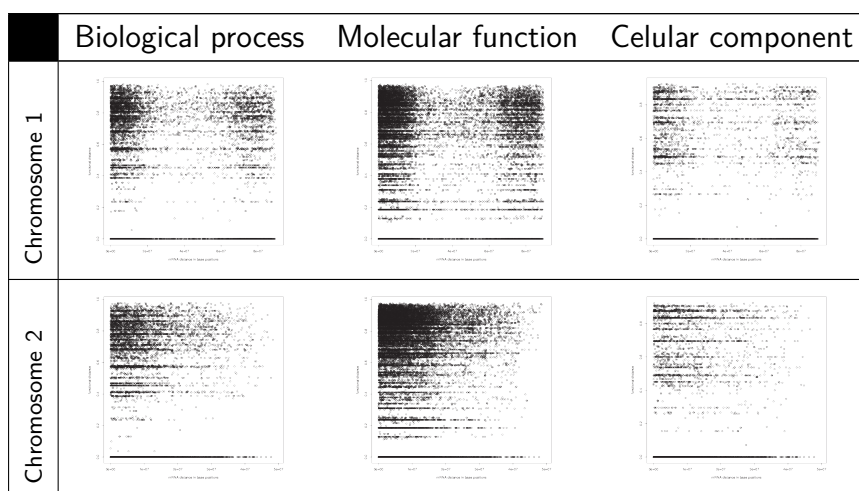


Table 5.10: Continued ...

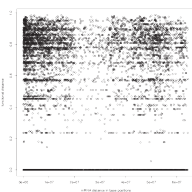
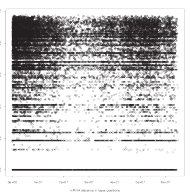
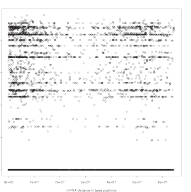
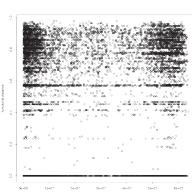
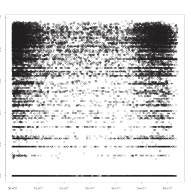

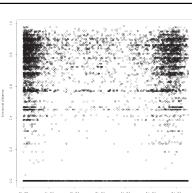
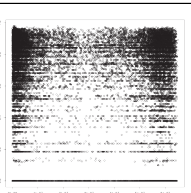
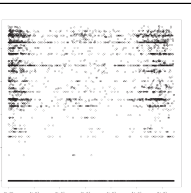
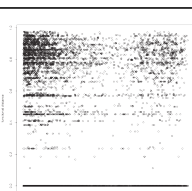
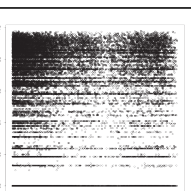
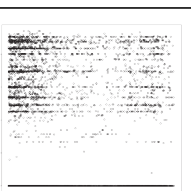
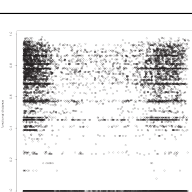
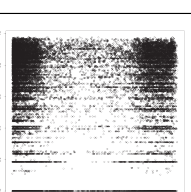
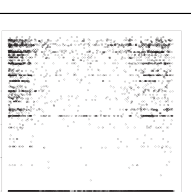
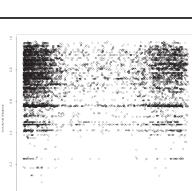
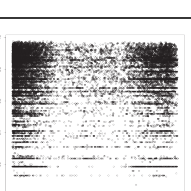
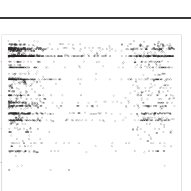
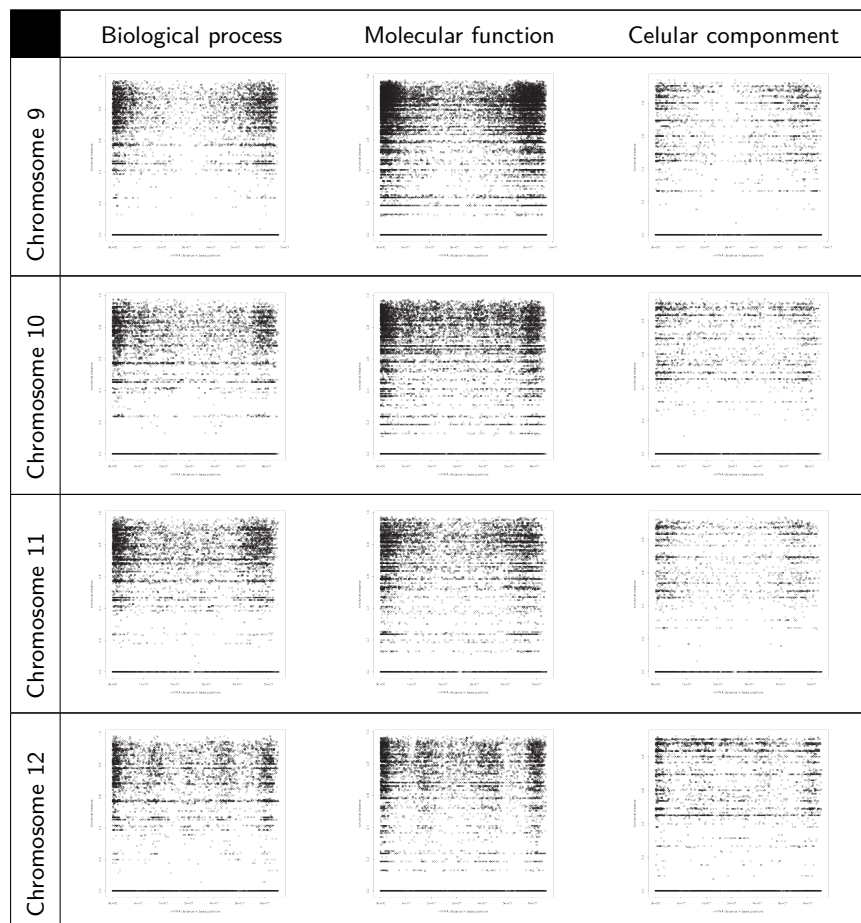
	Biological process	Molecular function	Cellular component
Chromosome 3			
Chromosome 4			
Chromosome 5			
Chromosome 6			
Chromosome 7			
Chromosome 8			

Table 5.10: Continued ...



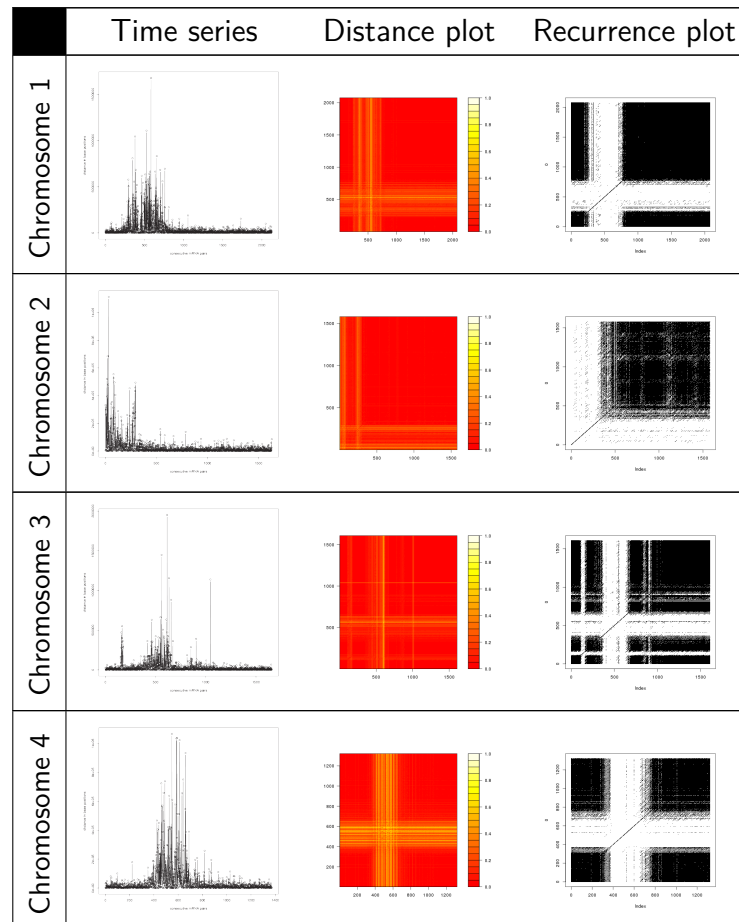
The results of this experiment are shown in table 5.10. The plots indicate that there is no obvious correlation between the functional and genomic distance. But some curious details can be observed. The most plots show that, considering the ontologies "biological process (BP)" and "cellular component (CC)", very short or large functional distances occur more frequently. However, in the ontology "molecular function (MF)" this trend is less pronounced. Many plots also leak medium inter gene distances.

Not clearly visible in the plots, but striking, if one inspects the underlying data matrix, is that it was frequently impossible to get the functional distance of gene pairs, because the assigned GO terms belong to different ontologies (BP, MF or CC). For instance, from 100000 measured chromosome 1 pairs we have obtained only 7661 ( $\sim 8\%$ ), 15587 ( $\sim 16\%$ ) and 4212 ( $\sim 4\%$ ) functional

distances bases on the BP, MF and CC ontology respectively (electronic supplement). This was also the reason why we could not create useful functional distance time series and perform the recurrence analysis to study their systemic properties. The obtained time series were too sparse, they do not even present long enough fragments with consecutive values.

To study dynamical aspects of the tomato (*Solanum lycopersicum*) genome we created time series based on consecutive gene base position distances for all 12 chromosomes and have analyzed recurrences with our bract pipeline. The pipeline is described in detail in the section [5.2 marker-less cancer classification](#).

Table 5.11: Gene base position distance time series (a), their reconstructed phase space distance plots (b) and the corresponding recurrences plots (c) for all twelve tomato chromosomes.



5.3. FUNCTIONAL AND GENE POSITION DISTANCES IN THE TOMATO GENOME

Table 5.11: Continued ...

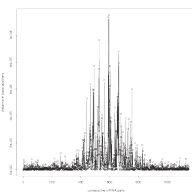
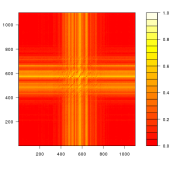
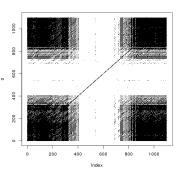
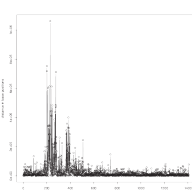
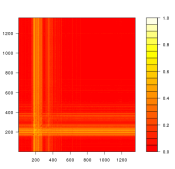
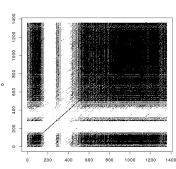
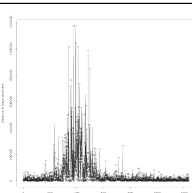
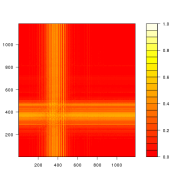
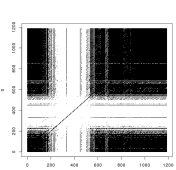
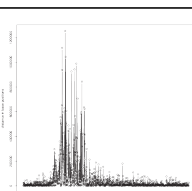
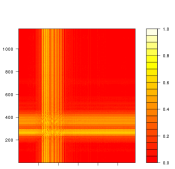
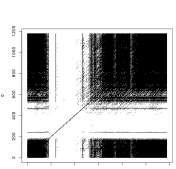
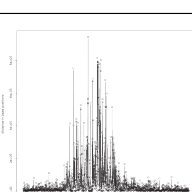
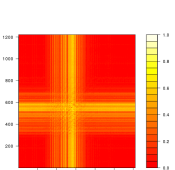
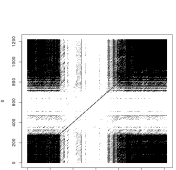
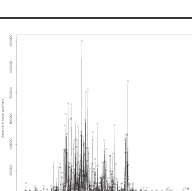
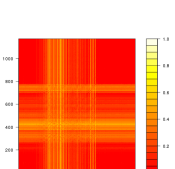
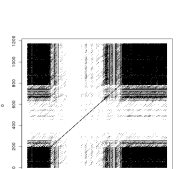
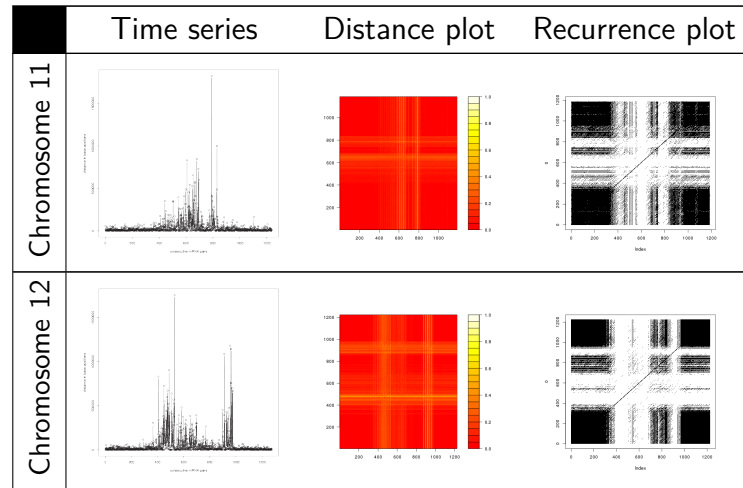
	Time series	Distance plot	Recurrence plot
Chromosome 5			
Chromosome 6			
Chromosome 7			
Chromosome 8			
Chromosome 9			
Chromosome 10			



Table 5.11: Continued ...



The plots in table 5.11 show that gene distances along the chromosomes are dominated mainly by two kind of blocks – short distances and longer ones. On the laterals we found mainly short distances blocks and between them a long distance block whose position varies. Some chromosomes, for instance chromosome 12, show a short-long-short-long-short pattern. In total, we found three different pattern (L, + and #) which in addition correlate with the main pattern classes found in the functional and gene position distance experiment (table 5.10). The mated pattern from both experiments we shown in the drawings of figure 5.25

DET (values between 0 and 1) and RATIO are both measures for the predictability of a system. As higher their values as better the predictability. From our recurrence quantification analysis of the tomato gene distance time series (electronic supplement) we have obtained DETs, which range between 0.0002895 and 0.2282000, and the RATIOS, ranging from 0.003456 to 1.264, that indicate a poor predictability of the system.

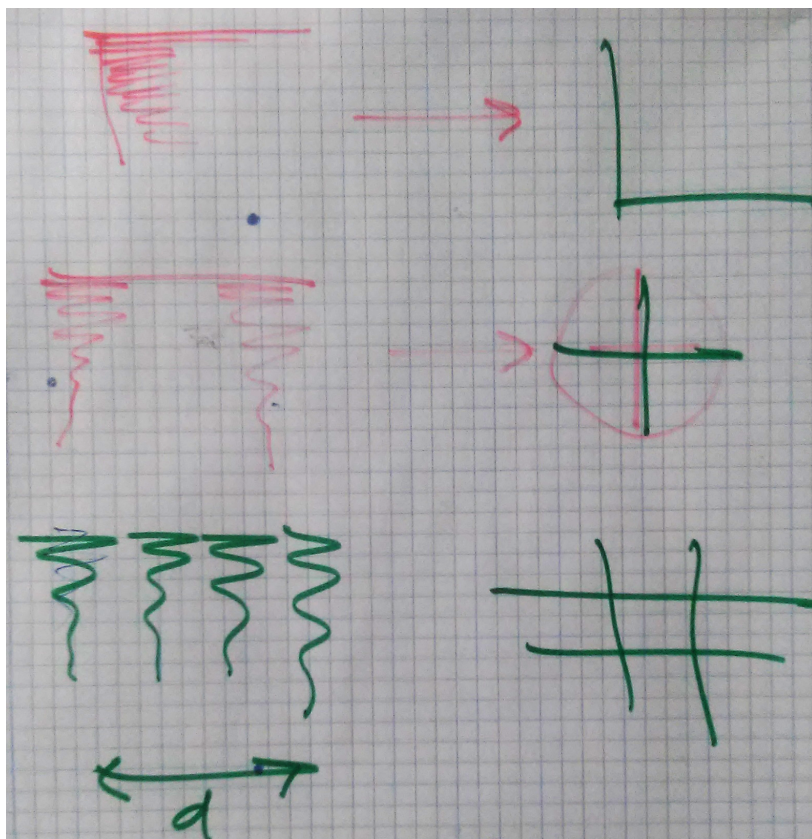


Figure 5.25: A casual drawing of mated distance patterns of the tomato genome. The main patterns found in the plots of protein functional distances as a function of intra-chromosomal gene base position distances are shown on the left side. The corresponding patterns from the recurrence analyses of reconstructed phase spaces based on gene distance time series are shown on the right side.

## 5.4 Systemic identification of taxa

As mentioned before, we consider genomes as adaptive complex systems which are solutions to selective pressure and contribute to the species diversification process. Under this assumption genomes should behave dynamically as non-linear systems and should manifest attractors which can be characterized and analyzed with our `bract` pipeline. Our principle motivation for the experiments in this section was to explore if under this assumption our proposed work flow is apt to analyze adaptive divergence processes between closely related organisms.

The principal problem to confront planning such an experiment is the poor data situation. There are not enough completely sequenced nuclear genomes available which allow us to train any model. However, the number of completely sequenced mitochondrial genomes is relatively high. On the ENA database<sup>181,182</sup> are currently 8753 (06.11.2015) mitochondrial genomes available. Moreover, the high variable mitochondrial DNA has been used, inter alia, for species bar-coding<sup>102</sup>, as a marker of molecular diversity<sup>183</sup> and phylogenetic reconstruction<sup>184</sup>. In genome papers, along graphical representation of chromosomes, is often displayed a track of GC content. The GC-content is found to be variable within different organisms and it is simple to calculate. These facts have motivated us to use the GC-content as the basis for the analyses shown here.

Our criteria to select the species to analyze were simple: We wanted one group of close and another of farer related species. Both groups should contain more than two species and for each of them have to be at least 15 different complete Mt genome sequences available. These criteria have reduced the large amount of mitochondrial data to a few completely sequenced Mt-genomes belonging to the genus *Pan* (chimpanzees) and the suborder *Caniformia* (or *Canoidea*, dog-like).

It should be mentioned also that we used for the taxa identification experiments our `bract` pipeline basically in the same way as in the marker-less cancer classification (see above in subsection 5.2). Both experiments differ only in the way the time series have been created and, with minor modifications, how the classifications have been evaluated. In this experiment we are confronted with the problem of multi-class classification. There are three solutions to this problem: directly using a pure multi-class algorithm or reducing the problem to multiple

binary classifications using a one versus rest or the one versus one strategy. As we are interested in evolutionary divergences between pairs of organisms we chose the one versus one strategy.

### 5.4.1 Data acquisition and preparation

To realize this experiment we have downloaded on October 06, 2015 all public available mitochondrial (Mt) genomes ( $n = 8753$ ) from the European Nucleotide Archive (ENA)<sup>181</sup> with the help of a script which is part of our R package bract:

```
download_ena.py -w organelle_genome -s mitochondria
```

This is done ones and allows us henceforth to work off-line. We have copied then only the wanted sequence files from our local mitochondrial genome sequence repository to the respective analysis working directories and have converted them from the EMBL to FASTA format. The criteria for the further species selection has been described already above.

For each Mt-genome we have generated a series of consecutive fragments applying a sliding window which moves from the first base position up to the last genomic coordinate of the DNA sequence. Within each window we sum the occurrences of both guanine and cytosine and divide by the window size. The size and the overlap of the windows depend on the sequence size, the wanted time series resolution and computational power. We use for the analysis of the Mt-genome sequences a window size of 1000 and move the window by 10 positions. This produce GC content time series with approximate 1500 values.

### 5.4.2 Pan

We have analyzed 124 mitochondrial genome sequences belonging to the genus Pan whose detailed distribution is shown in table 5.12. The number of available samples is notably lesser than in the cancer epigenomics experiment (see above in subsection 5.2). For each sequence we have create GC-content time series which have been analyzed by our bract pipeline. We excepted 1% false nearest neighbors to estimate the embedding dimension . We have used  $a = 1.2$  as factor in the formula (see equation 5.1) to get the cut off threshold  $\varepsilon$ . For each

Table 5.12: Sample number distribution for Mt genomes of the genus Pan.

	# used for training	# used for testing	total #
Pan paniscus	25	7	32
Pan troglodytes schweinfurthii	18	5	23
Pan troglodytes troglodytes	33	9	42
Pan troglodytes verus	16	5	21

species or subspecies, namely *Pan paniscus*, *Pan troglodytes schweinfurthii*, *Pan troglodytes troglodytes* and *Pan troglodytes verus* we have been able to reconstruct the time series, calculate the RQA measures and build a SVM model which could be used for classification. In figure 5.13 we show and compare the results of the most important pipeline steps. At first glance, images belonging to the same row are difficult to distinguish, except the recurrence plots. The time series look like semi-periodical serrated lines and exceed slightly 1500 time units. The graphs showing the average mutual information as a function of time delays decrease exponentially and stabilize, more or less, afterwards having AMIs below 0.5. the delays used for reconstruction range between  $\tau = 37$  and  $\tau = 39$ . The false nearest neighbors decrease fast from 20% to 0% giving embedding dimensions of  $m = 4$  and  $m = 5$

All classification which include *P. troglodytes troglodytes* failed or have performed poorly, all others – in contrast – show outstanding results (table 5.14). Hence, we can conclude that *Pan troglodytes troglodytes* and *Pan troglodytes schweinfurthii* as well as *Pan troglodytes troglodytes* and *Pan troglodytes verus* are among each other more closely related than *Pan troglodytes schweinfurthii* and *Pan troglodytes verus*. Under this assumption we propose some plausible emigration and diversification path which are shown in the casual figure 5.26.

### 5.4.3 Caniformia

From the suborder Caniformia we have analyzed 407 mitochondrial genome sequences. The sample number distribution for the species *Canis lupus familiaris*, *Martes pennanti*, *Urocyon cinereoargenteus*, *Ursus arctos* and *Ursus maritimus* is shown in table 5.15. The analyses of the Caniformia species have been performed under the same configuration as described for the genus Pan above.

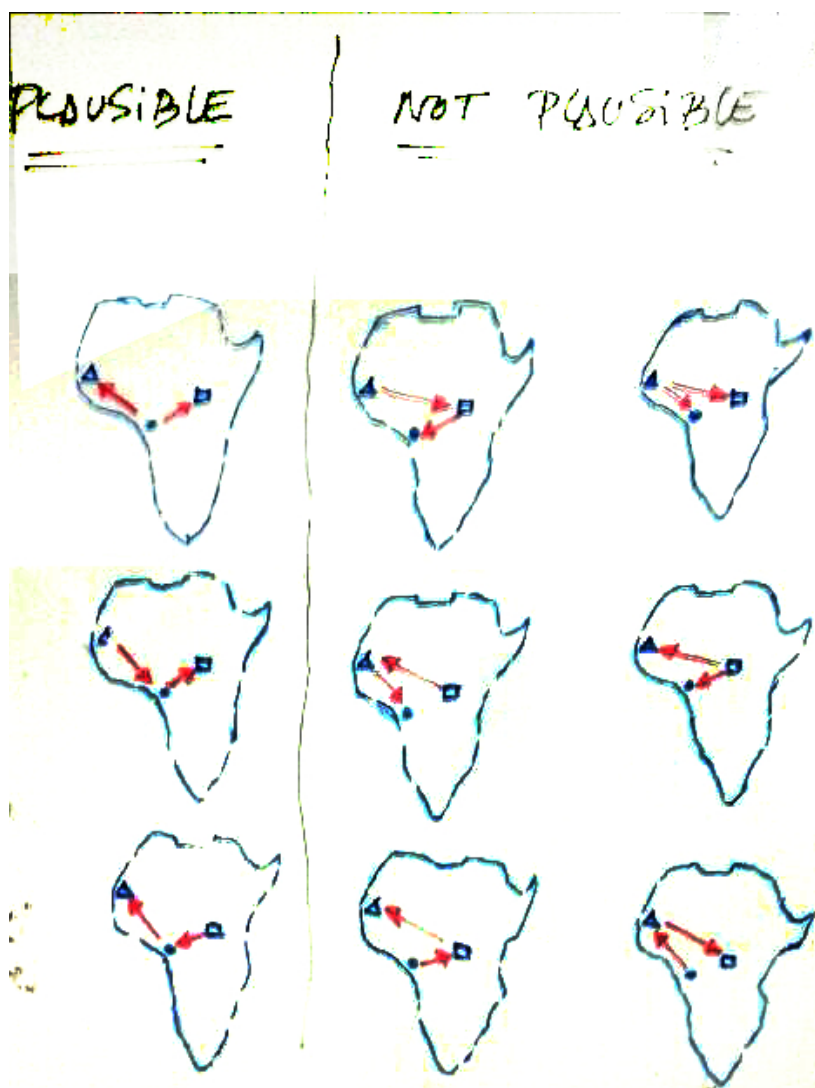


Figure 5.26: Casual drawing of plausible and not plausible Chimpanzee (*Pan troglodytes*) migrations and diversification paths. ▲ *Pan troglodytes verus*; ● *Pan troglodytes troglodytes*; ■ *Pan troglodytes schweinfurthii*.

The results for each step of the pipeline are similar to the ones for the genus Pan and shown in figure 5.16

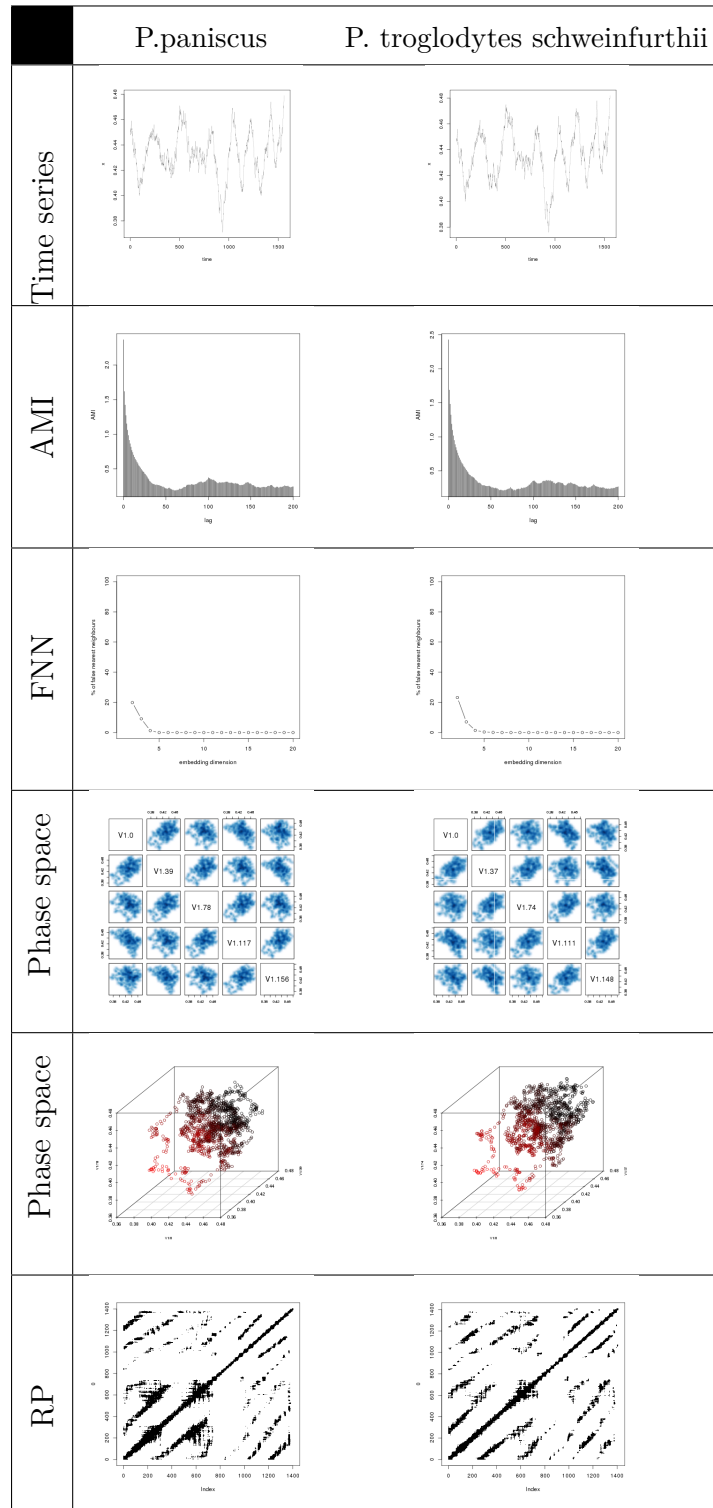
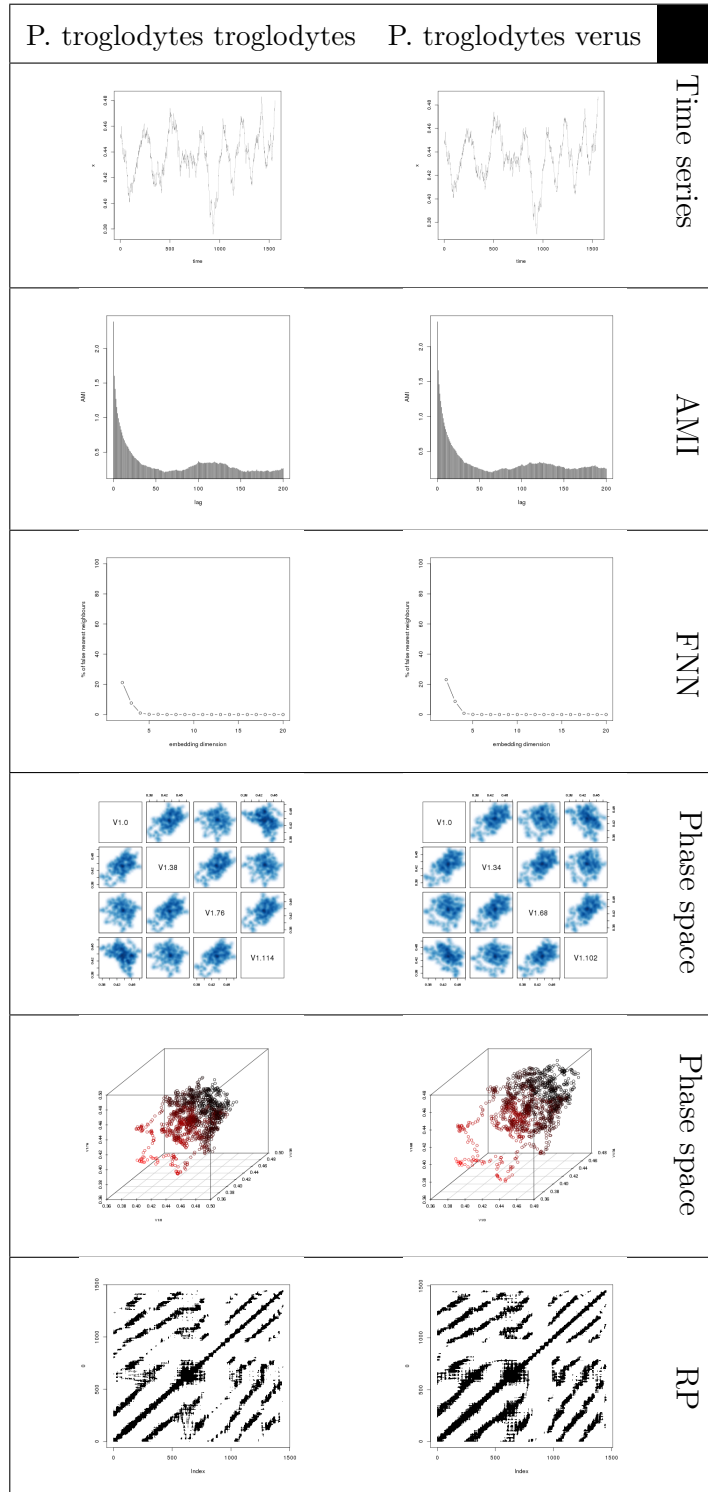


Table 5.13: Interim graphical results from the recurrence analysis of time series based on the CG content of Pan Mt-genome sequences.





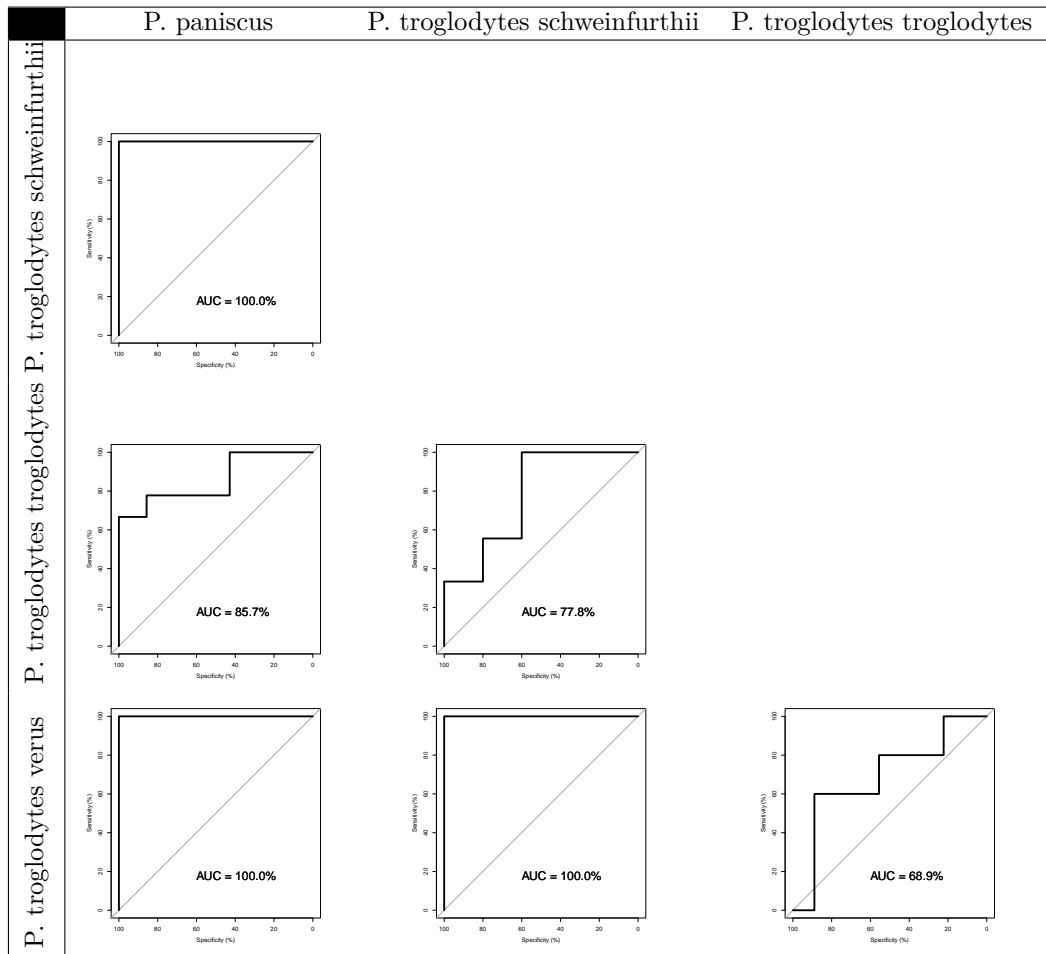


Table 5.14: ROCs and AUCs for systemic classifications on CG content of Mt-genome sequences from the genus Pan.

The results of the performance evaluation can be shortly described: All species of the suborder Caniformia can be classified correctly using an all versus all strategy. All AUCs are 100% (electronic supplement).

Table 5.15: Sample number distribution for Mt genomes of the suborder Caniformia.

Species	# used for training	# used for testing	total #
<i>Canis lupus familiaris</i>	105	27	132
<i>Martes pennanti</i>	12	4	16
<i>Urocyon cinereoargenteus</i>	17	5	22
<i>Ursus arctos</i>	53	14	67
<i>Ursus maritimus</i>	19	5	24

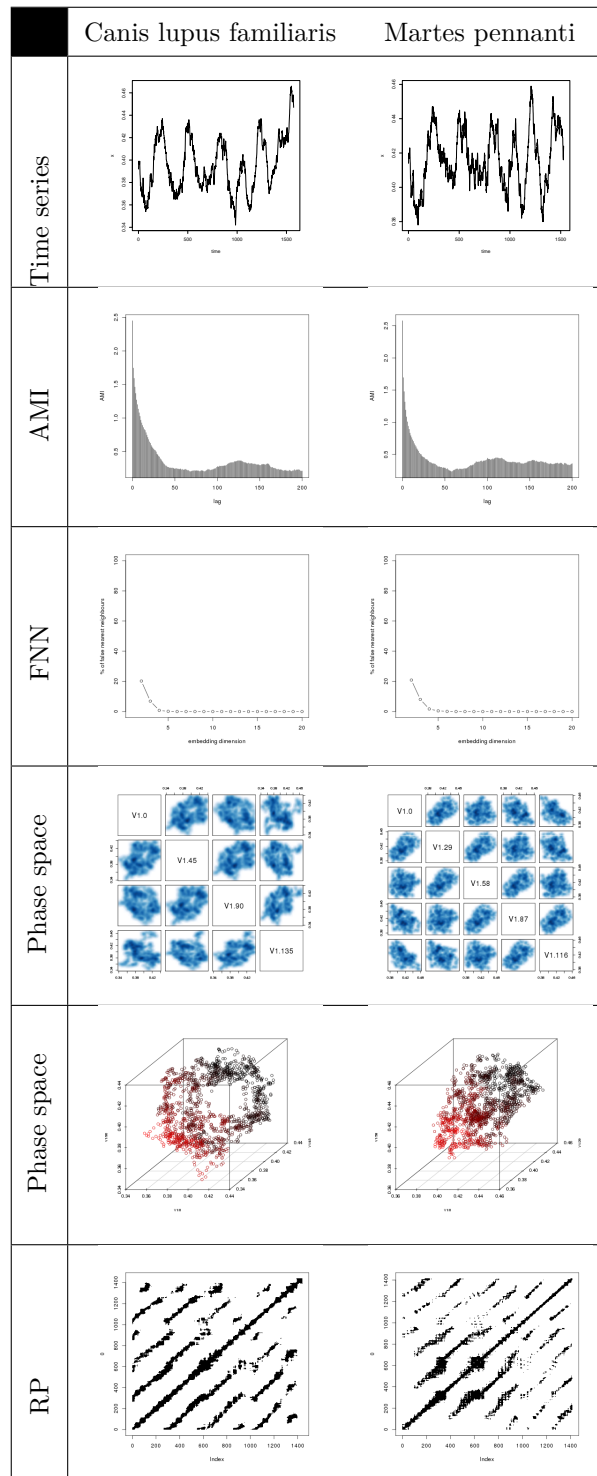
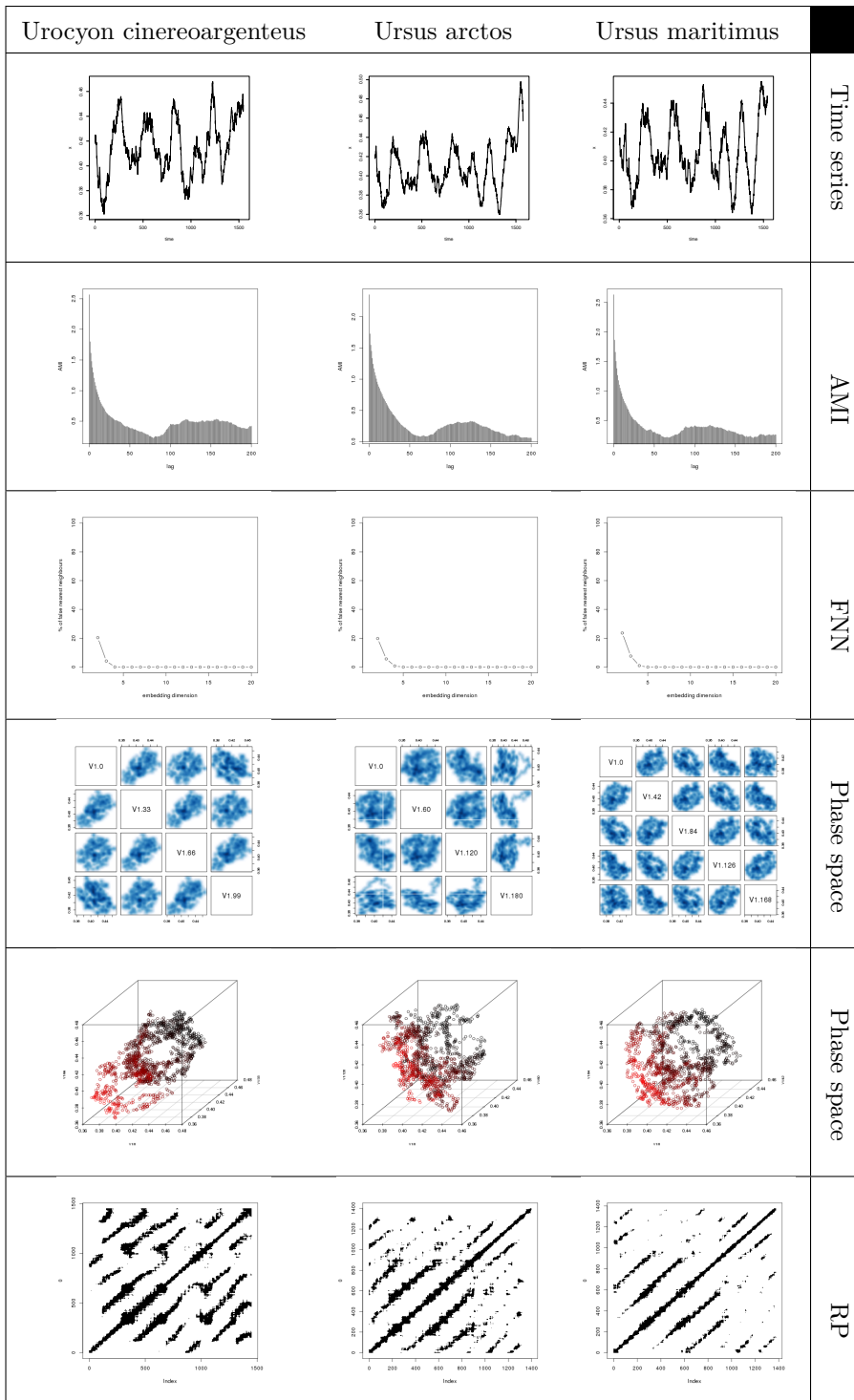


Table 5.16: Interim graphical results from the recurrence analysis of time series based on the CG content of Caniformia Mt-genome sequences.



## 5.5 Anti joint recurrence plot

To detect and visualize non synchronized regions in the comparison of two recurrence plots we formulate the new anti joined recurrence plot (AJRP) with the following equations:

$$R_{aj} = \Xi(R_1 - R_2) \quad (5.6)$$

where  $R_1$  and  $R_2$  are arbitrary recurrence plot matrices with identical dimensions and  $\Xi$  is

$$\Xi(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x \neq 0 \end{cases} \quad (5.7)$$

An example of its application is shown in figure [5.13](#).

## Discussion

While quantitative discrimination of gene-sized DNA fragments by standard alignment strategies and homology analysis can be easily achieved in a success and straightforward way, trying to scale up those approaches deeply further – as to chromosome wide or even genome wide sizes – is considerably very much difficult and, generally, much less meaningful<sup>185</sup>. Virtually all these alternatives are founded on local alignments or tree-guided alignment based algorithm pipelines<sup>186–189</sup>. They are computational costly and their efficiency drop rapidly with size and, overall, with evolutive distance. It has been recognized, however, that robust whole-genome alignment tools are going to be critical for the future of comparative genomics, despite there are a number of significant questions that haven't been properly addressed yet<sup>185</sup>.

In essence, our work cope this problem from a different perspective. The major aim of this research was, in fact, designing, implementing and preliminary characterizing of an original workflow able to make high-performance, systemic discrimination of genomes or epigenomes. This workflow can be flexibly applied to get binary classification of sequences without needing of previous alignment.

Globally, most of results included in this memory support the idea that genomes can be approached from a systemic perspective that we justify in two basic proposals: on the one hand, the assumption that genomes are subjected to adaptive dynamics that can be approached as complex adaptive systems (CAS). Consequently, they should be susceptible to be analyzed within the conceptual and methodological framework that complex theories and deterministic chaos theory have been developed and successfully applied during last twenty years

in many different fields<sup>151</sup>. On the other hand, the presumption that DNA sequences can be properly conceptualized as nonlinear time series – and be treated as such – at the level of formal models. It would be possible because, in essence, a time series is simply an ordered collection of observational data related to one of the state variables of the reference system. Thus, it is not temporality, *sensum strictum*, what determines its “temporal” dimension but the data ordinality, meaning that for each value  $v_i$  it is always possible to determine the precedent value  $v_{i-1}$  and the subsequent value  $v_{i+1}$ .

Although, to our knowledge, we do not have still any systematic formulation about it, none of these premises are completely aliens in the fields of DNA and genome analysis<sup>2,3</sup>. The using of time series analysis methodologies with DNA sequences are being used, punctually, since late eighties<sup>4-7</sup>. On the other hand, the mapping of genomes, proteomes, interactomes, and others biological systems objects in terms of complex adaptive networks (a frequently used approach in complex systems analysis) is already usual in numerous context of systems biology<sup>7-10,3</sup>. In fact, it has become a singularly powerful tool, due to its very well established framework, initially defined by the pioneer work of Erdos and Rennyi<sup>11,12</sup>.

When working with networks it is easy, actually, to extrapolate the obtained results to the original complex systems, largely more elusive. This kind of approach can provide information not only about the structure of the mapped complex system but also about its dynamic behavior and properties. It is the case, for example, of the outstanding work of Barabási and others about the self-similar nature of many biological system networks, their high resilience to random attacks or their high tolerance to errors<sup>8</sup>. With few exceptions of limited scope<sup>13</sup>, applying non linear dynamic models or analytical tools to the field of genomic or epigenomic are still rather unusual, particularly if compared to the very wide range of problems in physics, chemistry, biomedicine and engineering where these methodologies have been successfully applied.

Barabási and others have proposed that as topologies of biological complex networks are consequences of emergent properties determined by self-organization dynamics, ultimately ruled by simple laws of generic character, they should share a number of prominent features that define them<sup>9</sup>. Free-scale topologies and hierarchic organization of functional modules are two of

them. Since this perspective, it should be admitted that even considering the essentially stochastic nature of many dynamic processes in biology, their attractors must be frequently “strange attractors” (in the sense of Prigogine<sup>24</sup>). Consequently, they should have, in greater or less extension, a component of deterministic chaos that should be identified as structural features of the complex networks that map them. And, in fact, the self-similar topology of nodes distribution in many of these networks – following a power law – would confirm that deterministic character.

In this context, recurrence analysis of unfolded attractors, shown up by reconstructed time series based on different DNA features (methylation density, GC content, gene distance and other potentially relevant genomic variables) by applying the Taken - Ruelle<sup>17-19</sup> and Poincaré<sup>20</sup> theorems, seemed to us as a clear possibility to investigate. Thus, most of the work described in this memory was focused to develop a complete methodological protocol that would allow translating this theoretical framework, proved to have a high potential, to the field of structural genomics, phylogenomics and DNA methylation epigenomics.

In summary, our results indicate that this approach is not only possible but also can be effectively used to get highly compact descriptions of large DNA sequences (for instance chromosomes) that still retain many of the essential structural features of the original entities. Further, based on these descriptions (or pattern), models can be trained by machine learning algorithms, in our case support vector machines, and efficiently used to discriminate and classify unknown samples.

There are, undoubtedly, pending questions that might be relevant in the future, like the feasibility of applying the ergodicity theorem<sup>21</sup> to DNA time series, a controversial issue that still promotes vivid debates among theorists about the application scopes of the models and their meaning limits<sup>22,10,23</sup>. Anyway, it didn't prevent, in our case, reaching precise predictions in the field of cancer epigenetics or the genetic divergence among chimpanzee communities in Central Africa. It was also possible to do a preliminary systemic characterization of the tomato genome with regard to its functional annotation.

By means of our recurrence quantitative analysis protocol, we could analyze sequences up to 240 megabases (complete human chromosome I) and represent it by seven scalar values. When this analysis is made from DNA methylation



density time series coming from normal and tumor cells, our workflow enables us to discriminate their epigenetic patterns with performances that sometimes reached 98%. Starting from gene distances time series derived from the tomato genome we could map the complete genome in terms of 12 seven-dimensional vectors (84 scalar values) that can be used to challenge a wide variety of problems once we have the appropriate input data in the future.

A number of previous problems had to be afforded to get ready the implementation of the proposed workflow. As established by the theory, to unfold the reconstructed attractor in the appropriate  $n$ -dimensional space, coordinates of hyper-points must be defined from the starting time series, taking successive values with an appropriate delay that must be previously estimated. Moreover, dimensionality of the phase-space must be also determined in the appropriate way. None of these two parameters can be straightforwardly determined: the false nearest neighbor algorithm used to estimate the embedding dimension depends on the election of a bias value and, on the other hand, the estimate of a proper delay by computing the mutual information implies the adoption of adequate strategies for minimum detection.

For all these reasons and also to check the correct functioning of our software developed ad hoc for this purpose, it became necessary to use a well known reference model – the Rössler attractor in our case – as “workbench” to test the workflow. Preliminary experiments let us confirm that it was able to properly and efficiently unfold the Rössler attractor from time series built from one of its state variables and verify that the reconstructed attractor retained topological characteristics of the original.

Once established its validity, the RQA workflow was used to study the epigenetic drift that, in terms of DNA methylations, take place during the carcinogenesis process. This study was also aimed to establish the importance of the systemic approach, in the sense above mentioned, to get insights on systemic features that are difficult to capture using more conventional approaches.

## 6.1 On cancer

When referred to mutations, genomic changes encompassing the carcinogenesis process have been proved to be highly specific and particularly targeted to

key elements in tumor appearance and development. They progress as cumulative changes restricted to a very specific set of genes or regulatory elements whose implicit “logic” can be frequently established in a very clear manner as they are involved in either development, cell cycle regulation, cell proliferation, migration, differentiation, cells survival or apoptosis. Changes in this singular and relatively small set of genetic elements (proto-oncogenes) have proved to be strongly correlated, statistically, with cellular malignancy and carcinogenesis drive. Thus, despite their intrinsic value as potential research and/or therapeutic targets, they are usually referred as true cancer markers and have been largely recognized as predictive tools with increasing potential in cancer diagnosis. The analysis of some well defined tumor markers leads actually not only to determine the cancer type but also estimates tumor malignity and ultimately, illness prognosis.

The fact that carcinogenesis process is also linked to specific changes in methylation of DNA (firstly reported by Feinberg and Vogelstein<sup>14</sup>) and that these changes are indeed determinant in the tumor implant and development have been firmly establish along the last two decades. It is not surprising that many early efforts were intended to identify altered DNA methylation in located, specific positions or narrow regions confined to certain putative regulatory targets, in the same fashion as mutational tumor markers do. In fact, altered methylation of CpG islands within promoters of key transcription factors or genes involved in chromatin rearrangement, DNA methylation, cell proliferation and tumor repression have been reported as targets of these methylation changes and could have lead to support the existence of true “methylation markers”, statistically correlated with tumor cells and potentially equivalent to “classical” mutation markers.

Astoundingly, detailed methylation analysis, (especially after genome-wide Bisulfite sequencing methods became recently available) has revealed an entirely different genome methylation landscape, deeply altered in cancer cells and affecting up to 40% of the whole genome.

Suggestively, this cancer epigenetic drift has also revealed another essential characteristic: It cannot be considered as a well defined, unique shift from something like “normal epigenetic landscape” to something like “cancer epigenetic landscape”. In fact, alterations in DNA methylation across-cancer have

shown noticeable heterogeneity among different tumor clones (especially when compared to “normal” landscape). Some authors refer this situation as a tumor induced “epigenetic deregulation” where the observed changes in DNA methylation are essentially of stochastic nature, beyond the possibility of merely defining boundaries of generic hypomethylated and hypermethylated genomic regions<sup>64</sup>.

This point of view could implicitly lead to a model where stochasticity is particularly linked to randomness and, thus, make cancer epigenetic landscapes intrinsically unpredictable (at least, within the vast hypomethylated and hypermethylated entities above mentioned). This is of course a possibility that still would explain most of experimental observations and has been implicitly proposed by many previous studies analyzed by a variety of conventional statistic correlations.

To our mind there is a plausible alternative that would offer a suitable perspective of the reported potential Darwinian intra-evolutive mechanisms that might be operating during tumor development<sup>15,16</sup>. To this view, each clone would represent an adaptive “solution” to the effective process of carcinogenesis driving, whose final survival (and establishment) will be decided by a competitive selection process.

Taking these arguments, the existence of Darwinian intra-evolutive mechanisms of adaptation during tumor implant seems a rather plausible possibility. If that is the case, apparent stochasticity now should not arise from pure randomness but should have some kind of “structure”. In other words, changes in methylation of the different solutions (clones) should follow some particular patterns, better described from deterministic premises and hardly visible from conventional statistic correlation measures. We would be facing here, in other words, a deterministic chaos scenery.

To uncover those weak signals we have developed a model framework in which genome methylation patterns are considered time-dependent dynamical systems whose topological properties can be captured by embedding to a higher dimension<sup>128</sup> and properly described by subsequent recurrence quantitative analysis (RQA) by the six standard recurrence parameters described in methods (recurrence rate, determinism, entropy, ratio, laminarity and trapping time<sup>154,190</sup>) together with the additional measure of lacunarity<sup>129,131</sup>, to capture how gappy the structures are.

In total, we have analyzed data of eleven different cancer types (more than 4000 samples obtained from the Illumina Infinium HumanMethylation450 Bead-Chip © repository of The Cancer Genome Atlas Project). Taking chromosome I as the starting model we have analyzed more than fifty thousand CpG site methylation density time series. We have been able to successfully train binary classification models by support vector machines and used them to predict tumor and normal tissues on their embedded RQA in all cases. Except in one case (papillary thyroid carcinoma, 74.6%), all tumors showed good or very good performances that were, in terms of AUCs, on or above 84.5% and most of them (seven cases from eleven) were 90% or better.

These results suppose a strong evidence that compressing the methylation sequence of a complete human chromosome in a unique RQA 7-dimensional vector preserve the needed information to identify very effectively the epigenetic drift that take place during cellular malignancy, at least, in the studied models. In fact our protocol is sensible enough as to perceive significant differences between normal and tumor samples by simple visual inspection of 2-dimensional projections of the RQA vectors, independently of the type of tumor. It was rather evident that tumor cells show a markedly higher heterogeneity in their methylation patterns than the corresponding normal cells.

As mentioned above, from pioneer work of Prigogine<sup>24-26</sup>, complex adaptive systems can behave dynamically as nonlinear systems able to exhibit complex stability landscapes usually referred as “strange attractors”. Although it is not possible to know how these attractors look like, the ones that we have reconstructed here should be topologically related with them. From this perspective, it could be considered that during the carcinogenesis process, clonal methylomas, as mention above, would represent in fact “adaptive solutions” of the tumor metabolism drift (here considered as a case of CAS dynamics) and, thus, stable setting of the complex, nonlinear system attractor, susceptible of systemic characterization by our procedural workflow.

If this premise is correct, cancer epigenetic drift hardly could be satisfactorily explained in terms of epigenetic “markers”. In other words, the observed changes wouldn't be caused by specific changes of concrete positions (as actually happen in the case of mutational changes). By the contrary, methylation drift would be a consequence – as mentioned – of potential intraevolutive mechanisms

that would take place during carcinogenesis. It doesn't mean, of course, that the existence of specific positions or well defined locations whose methylation state correlates well with the carcinogenesis process have to be discarded. By the contrary, the cancer adaptive epigenetic drift could have some common invariances that could be essential to reach the new adaptive niche and still be an essentially systemic process that hardly could be explained or fully characterized, anyway, by these invariances. In other words, it would be conceivable that all adaptive "solutions" share some invariants but keep a CAS dynamic behavior.

While we did not have conclusive proofs of such a mechanism before our next set of experiments, there were two hypothetical reasons to support its feasibility. On the one hand, the environmental scenario in which the tumor has to grow is highly hostile and thus, tumor cells are initially quite a far from having a good adaptive fitness. On the other hand, these Darwinian intra-evolution-like mechanisms are presumably easy to implement for the emergent tumor due to its intrinsically high growing speed. Moreover, such a kind of mechanism, in case of happen, would probably be rather elusive to be detected by classical approaches.

To evaluate the performance of marker predictions (based on differential DNA-methylated CpG sites) and markerless prediction (based on systemic criteria) we carried out different types of experiments. In a first set of experiments, we compiled a list of differential DNA-methylated CpG sites (or Cancer Related Gene Symbols, CRGS). We also identified two types of regions with epigenetic significance previously described in the bibliography: large BLOCKS of hypomethylated regions and cDMR hypermethylated regions. As a whole, these three regions represented about 50% of the chromosome I. We derived different sets of time series in which systematically we eliminate one or more of these regions and proceeded to analyze them by RQA and performed support vector machine classifications.

In summary, this set of experiments showed that the remaining fraction of the chromosome still retained the epigenetic signature that permit to discriminate between normal and tumor cells. One plausible interpretation of these results is that such a epigenetic signature "is in the whole and in the part". It would be, in other words, an emergent characteristic of systemic nature.

Certainly, another possible interpretation is that not all the epigenetic sig-

nificant elements have been identified yet. Even when this option cannot be completely discarded, it hardly can be considered as the major cause of the obtained results. First of all, because assuming that the lost of the SVM discriminative capacity is a measure of the relative weight that hypothetical unknown elements would have on the entire response, we would be talking of about 20% to 40% of total cancer related (epigenetic) elements. In second term, these elements would be practically affecting all types of cancers and, finally, the following set of experiments were more coherent with a systemic behavior.

In fact, the weight attributed in the literature to the other methylation related regions was clearly confirmed when we compared predictions made by our approach and those obtained from different sets of cancer related elements through direct training (without embedding and RQA) of the same learning machine algorithm, in comparable conditions. Our results for the head and neck squamous cell carcinoma (HNSC) showed very high (and similar) performances in both cases. Surprisingly, it was enough a relatively reduced number of markers to get an AUC of 97.4% (against 98.6% reached with the markerless prediction).

To get further insights about the possible systemic character of the epigenetic drive associated to the carcinogenesis process, we planned a third kind of experiment in which predictions were made after training the machine learning algorithm with random, cancer unrelated methylation sites of similar sizes to the markers previously used. Although this is a preliminary study, our results were unexpectedly good ( $AUC > 90\%$ ), even limiting the number of pseudo-markers to only 18. The validity of these results was confirmed, nevertheless, by using controls made by shuffled samples (showing totally negative results).

Globally, our results suggested that epigenetic differences found between tumor cells and normal cells not only are due to well defined methylation markers, whose correlation with the cell malignancy has been previously established and now confirmed by our own experiments (founded in an alternative machine learning procedure). These differences are also due to a systemic component that potentially expand all the DNA sequence and whose presence can be made visible with only a few positions potentially methylables.

Assuming that differences in  $\beta$ -values expands along the sequence, it seemed interesting to investigate if a single estimation of the total degree of methylation

(total  $\beta$ -value sum ) would be enough to discriminate between normal and tumor cells. With that aim we planned a new set of experiments to compare total  $\beta$ -value sum distributions in tumor and control samples under different conditions and using three different estimation criteria: average of total  $\beta$ -value sums , p-values given by the Wilcoxon rank sum test with continuity correction and support vector machine based binary classification.

We have found that when including all CpG sites, total  $\beta$ -value sums are rather similar between normal and tumor cells and to discriminate them is generally difficult as if it is made on the base of averages, Wilcoxon based p-values or SVM binary classification. We have also observed a tumor dependent react: while liver hepatocellular carcinoma (LIHC) can be discriminated with any of the estimation criteria, papillary thyroid carcinoma (THCA) is refractory to classification on the basis of total  $\beta$ -value sums .

When samples are restricted to cancer related epigenetic elements (CRGS, BLOCKS and cDRM) the answers of different tumors is much more heterogeneous. CRGS, for example, lead to easily discriminate BRCA, COAD and LIHC and in less extension LUAD and KIRKP while LUSC, THCA and UCEC cannot be discriminated in most of cases.

When computing total  $\beta$ -value sums of hypomethylated BLOCKs regions, the discrimination capacity significantly increased in almost all cases, although it was not sufficient to classify PRAD and THCA cancers.

Finally, when samples were composed by hypermethylated cDMR domains, all tumors could be easily discriminated in terms of  $\beta$ -value sums with high statistic significance with the exception of THCA. If hyper- and hypomethylated domains are mixed, the discrimination capacity decrease notably, as expected, because total  $\beta$ -value sums in this case is composed by positive and negative terms that mutually cancel themselves.

Having in mind the heterogeneity of samples and procedures, our result showed reasonable coherence. In summary, they suggest that cancer epigenetic drive has a strong relationship with changes in cDMR hypomethylated regions in practically all cancer studied. This region is essential to characterize tumor cells in all cases except THCA. Large hypomethylated regions also contribute in a more variable proportion, depending of the type of cancer. Its contribution to PRAD and THCA is probably small. Finally, CRGS symbols are much more

specific and would have less weight in the epigenetic profile of LUSC, THCA and UCEC.

While liver hepatocellular carcinoma (LIHC) can be practically discriminated in any condition, THCA is difficult to classify, again, in most of conditions. They could represent extremes with regard their systemic character. LIHC epigenetic profile would potentially affect extensive regions of chromosome I, suggesting that in this case the epigenetic drift is very important (intense) and probably has an important systemic component. THCA, would represent the other extreme. Epigenetic changes in this case would be minimal or would affect other chromosomes. It is obvious that this study should be extended to the rest of chromosomes. In any case, a poor response in all the chromosomes would indicate that tumor cells are closer to their optimum adaptive niche from the very beginning and would be more distant from a systemic model. It seemed meaningful to us that THCA is also the only cancer that we couldn't predict properly.

In summary, we show that data driven or agnostic approaches which base exclusively on data intrinsic signals can solve problems even where knowledge based models are incomplete or not available. Nevertheless, we have to admit that the machine learning part of our analysis pipeline is limited to problems for which we have access to large data sets and as a consequence the answers to the following two important questions still remain open and are postponed to the near future: Can we classify pre or early neoplastic cells from normal ones? Are the non-normal systemic patterns only characteristic for tumors or in general for inflammations? Despite the data-poor situation we currently try to extent the application of our pipeline to other diseases (e.g. Alzheimer) and – motivated by preliminary experiments – also to the analysis of epigenomic stress response.

## 6.2 On tomato

Another different set of experiments included in this memory we focused to the potential application of our methodological tool to structural and functional genome analysis. In this case, the study was carried out on the complete tomato genome, as one part of this work was developed by me as a member of the del



"Plant Computational Biology"-group (PCB) at the Max-Planck-Institute for Plant Breeding Research, as part of the International Tomato Annotation Group (ITAG) and the Tomato Genome Consortium, where I contributed to the functional GO annotation of tomato genome. From 19662 annotated genes (57% of the total) and the complete genome sequence, we treated to analyze any possible correlation between gene physical distance and functional distance in terms of three standard ontologies: biological process (BP), molecular function (MF) and cellular component (CC). A second objective of these experiments was to obtain the twelve RQA vectors that represent the complete tomato genome, in terms of intergenic distances, in our systemic model.

Our results pointed out that there isn't any detectable correlation between physical distance and functional distance from any of the gene ontologies studied. Profiles can be actually explained in all cases from the particular distribution of intergenic distances calculated for each chromosome. With our current information, it is not possible for the moment to complete this functional analysis with the resolution level that would be necessary to get the RQA from functional distances. As the methodological workflow is ready and proved, the existence of new data would permit us going further in this direction.

With regard to embedding and RQA of tomato genome, we have analyzed them and found deep differences with respect to the ones that we got with the human epigenome (chromosome I). Some of the RQA parameters would suggest that tomato system has a very low predictability (opposite to our data from human genome). Also in this case it would be necessary characterize a significant number of genomes to be able of make a proper interpretation of the found differences.

### 6.3 On chimpazees and Caniformia

The last block of experiments included in this study had as major motivation to explore the potential of our systemic approach to analyze adaptive divergence among closely related genomes. Our study focused on classification of mitochondrial genomes because the number of sequenced organisms is notably higher than in the case of nuclear genomes: currently there are 8753 complete mitochondrial genomes available. In this case, our analysis was centered on CG

pairs density. We have carry out two sets of preliminary experiments focused on two animal groups well represented in the Mt-DNA database.

The first set was addressed to three chimpanzee (*Pan troglodytes*) subspecies: *schweinfurthii*, *troglodytes* and *verus*. RQ analysis of the embedded Mt-DNA have led us to get insights on the phylogenetic relationship of three subspecies of chimpanzee from Central Africa and confirm that our data are in good agreement with mainstream finding of previous studies.

It has been established<sup>191</sup> that Central and Western Africa populations of chimpanzees are divided into two geographically- and genetically-defined groups separated by the Sanaga river along East Nigeria and Cameroon. This first population split seemed to happen 250 kya ago and left *P.t. ellioti* and *P.t. verus* (in the west) in the north river side while *P.t. troglodytes* and *P. t. schweinfurthii* remained in the south riverside. Although there is not a definite consensus about the phylogenetic history of these populations<sup>192</sup>, it has been noticed that *P.t. troglodytes* and *P. t. schweinfurthii* are closer related between them than they are with *P.t. verus* and that, at least in part, divergences can be explained by an isolation-with-migration model<sup>191</sup>. Our results, by alone, would essentially confirm this relative distances and would reduce the number of possible migrations from 9 to only three (see figure 5.26).

Finally, the second set of experiments was devoted to test the potential efficiency of our workflow in the binary classification of Caniformia species. A total of 407 Mt genomes belonging to five different species of the suborder California were employed to carry out the study by recurrence analysis of time series based on the CG content of the Mt-genome sequences. RQA parameters were then used to train the SVM-based binary classification algorithm. Achieved performances were outstanding in all cases giving optimal values of both, sensibility and specificity (100% AUC). Even being a very preliminary result, these data pointed out the potential of the method and encourage us to perform further exploratory studies in different directions to find out new possibly applications of the proposed model.



## Conclusions

1. A new procedural workflow has been designed to perform genome analysis from a systemic perspective. The protocol has been pipelined and automated as open software (“bract”) accessible through the net (<https://github.com/bractproject/bract>). The tool has proved to be very efficient in discriminating closely related DNA sequences (genome sized) on the basis of their systemic features. Our results indicate that this approach can be effectively used to perform embedding and recurrence quantitative analysis of DNA time series to get highly compact descriptions of large DNA sequences by mapping them into one or a small set of 7-dimensional vectors that still retain many of the systemic features of the original entities. These compressed fingerprints can be used to efficiently train a support vectors-based, machine learning algorithm, able to do binary classification of the RQA parameters and to efficiently discriminate the original DNA on the base of those features.
2. This approach has been used to analyze the DNA methylation epigenetic drift that take place during carcinogenesis process in eleven different cancer types ( more than 4000 samples) obtained from the Cancer Genome Atlas project. Aside generating well defined hyper- and hypomethylated regions, as mentioned in the literature, our results have revealed that cancer epigenetic drift might also involve further along-the-genome, extensive changes that cannot be assimilated to any previously described entity and entailed a patent increase in heterogeneity. These epigenetic

changes, noticeably detected by our methodological approach, point out the existence of a systemic component that in some cases, like in the liver hepatocellular carcinoma, could be decisive and can be interpreted in terms of potential intra-evolutive (Darwinian) mechanisms where individual “clonal methylomas” would represent adaptive solutions of the demanding tumor metabolic requirements that have been presumably selected in a highly hostile environment. As shown by our results, different tumors could actually follow different adaptive epigenetic drifts or “cancer epigenetic landscapes” that can be also captured by the model.

3. Bract has been used to find potential non linear correlations between physic and functional gene distance, on the basis of the annotation of the tomato genome. While no apparent correlation could be detected, a mapping of the complete genome into twelve RQA vectors has been performed. Some of the RQA parameters would suggest that tomato system has a very low predictability (opposite to our data from human epigenomes). With our current information, it is not possible for the moment to complete this functional analysis with the resolution level that would be necessary to get the RQA from functional distances. The acquiring of new data about functional GO-annotation would be necessary to permit us going further in this direction.
4. Our protocol has proved to be useful to analyze adaptive divergence among closely related genomes. In the case of three major subspecies of Central and West Africa populations of chimpanzee, the systemic characterization of their mitochondrial genomes have permitted get insight about the possible phylogenetic relationships among these populations. Our results, by alone, would essentially confirm that *Pan troglodytes troglodytes* and *Pan troglodytes schweinfurthii* are closer related between them than they are with *Pan troglodytes verus* and that, at least in part, divergences can be explained by an isolation-with-migration model in what the suggested potential migrations, as directly deduced from the performed Mt-DNA analysis, match the (still no definite) consensus about the phylogenetic history of these populations.

In the context of systemic identification of taxa, bract shows to be extremely sensible in performing binary classification of Caniformia species. A total of 407 mitochondrial genomes belonging to five different species of the suborder Caniformia were employed to carry out a study by recurrence quantitative analysis of time series based on the CG content of the sequences. The achieved performances were outstanding in all cases, giving optimal values of both, sensibility and specificity (100% AUC). Even being a very preliminary result, these data pointed out the potential of the method and encourage us to perform further exploratory studies in different directions to find out new possibly applications of the proposed model.



---

## Bibliography

1. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–641, 2012.
2. Stuart Kauffman. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford University Press, 1996. ISBN 978-0-19-976185-2.
3. Jan Walleczek, editor. *Self-Organized Biological Dynamics and Nonlinear Control*. Cambridge University Press, 2000. ISBN 978-0-521-62436-7.
4. D.S. Stoffer, D.E. Tyler, A.J. McDougall, and G.A. Schachtel. Spectral analysis of DNA sequences. *Bull. Int. Stat. Inst.*, (Bk 1, 345-361; Bk 4, 63-69), 1993.
5. Serge Muyldermans and Andrew A. Travers. DNA Sequence Organization in Chromosomes. *Journal of Molecular Biology*, 235(3):855–870, 1994.
6. J. Maddox. Long-range correlations within DNA. *Nature*, 358(6382):103, July 1992. ISSN 0028-0836. doi: 10.1038/358103a0.
7. T. Subba Rao and Calyampudi Radhakrishna Rao. *Time Series Analysis: Methods and Applications*. Elsevier, 2012. ISBN 978-0-444-53858-1.
8. Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. Error and attack tolerance of complex networks : Article : *Nature*. *Nature*, 406(6794):378–382, 2000.



9. Albert-László Barabási, Zoltán N. Oltvai, and Stefan Wuchty. Characteristics of Biological Networks. In Eli Ben-Naim, Hans Frauenfelder, and Zoltan Toroczkai, editors, *Complex Networks*, number 650 in Lecture Notes in Physics, pages 443–457. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-22354-2 978-3-540-44485-5.
10. Andre S. Ribeiro and Stuart A. Kauffman. Noisy attractors and ergodic sets in models of gene regulatory networks. *Journal of Theoretical Biology*, 247(4):743–755, 2007.
11. P Erdős and A Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.
12. P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.
13. Clara Frontali and Elisabetta Pizzi. Similarity in oligonucleotide usage in introns and intergenic regions contributes to long-range correlation in the *Caenorhabditis elegans* genome. *Gene*, 232(1):87–95, 1999.
14. Andrew P. Feinberg and Bert Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895):89–92, 1983.
15. E. D. Schwab and K. J. Pienta. Cancer as a complex adaptive system. *Medical Hypotheses*, 47(3):235–241, 1996.
16. P. C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.
17. J.-P Eckmann, S. Oliffson Kamphorst, and D Ruelle. Recurrence Plots of Dynamical Systems. *Europhysics Letters (EPL)*, 4(9):973–977, 1987.
18. David Ruelle and Floris Takens. On the nature of turbulence. *Communications in Mathematical Physics*, 20(3):167–192, 1971.
19. Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, number 898 in Lecture Notes in Mathematics, pages 366–381.

- Springer Berlin Heidelberg, 1981. ISBN 978-3-540-11171-9 978-3-540-38945-3.
20. Poincaré, Henri. Sur le problème des trois corps et les équations de la dynamique. *Acta mathematica*, 13, 1890.
  21. L. Alaoglu and G. Birkhoff. General ergodic theorems. *Annals of Mathematics. Second Series*, 41:293–309, 1940.
  22. R. Serra, M. Villani, A. Barbieri, S. A. Kauffman, and A. Colacci. On the dynamics of random Boolean networks subject to noise: Attractors, ergodic sets and cell types. *Journal of Theoretical Biology*, 265(2):185–193, 2010.
  23. Michael G. Sadovskz and Ksenia A. Nikitina. Very Low Ergodicity of Real Genomes. *Journal of Siberian Federal University. Mathematics & Physics*, 7(4):530–532, 2014.
  24. G. Nicolis and Ilya Prigogine. *Exploring Complexity: An Introduction*. W.H. Freeman, 1989. ISBN 978-0-7167-1859-8.
  25. Ilya Prigogine and Isabelle Stengers. *Order Out of Chaos: Man's New Dialogue with Nature*. Flamingo, 1985. ISBN 978-0-00-654115-8.
  26. Ilya Prigogine. *From Being to Becoming: Time and Complexity in the Physical Sciences*. W. H. Freeman, 1980. ISBN 978-0-7167-1108-7.
  27. P. W. Anderson. More Is Different. *Science*, 177(4047):393–396, 1972.
  28. M. E. J. Newman. Resource Letter CS–1: Complex Systems. *American Journal of Physics*, 79(8):800–810, 2011.
  29. Hiroaki Kitano. Cancer as a robust system: implications for anticancer therapy. *Nature Reviews Cancer*, 4(3):227–235, 2004.
  30. Peter C. M. Molenaar and Richard M. Lerner. *Handbook of Developmental Systems Theory and Methodology*. Guilford Publications, 2013.

31. James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67, 2012.
32. A. N. Tripathi. *Linear Systems Analysis*. New Age International, 2007. ISBN 978-81-224-1164-5.
33. Dean Rickles, Penelope Hawe, and Alan Shiell. A simple guide to chaos and complexity. *Journal of Epidemiology and Community Health*, 61(11): 933–937, 2007.
34. R. S. MacKay. Nonlinearity in complexity science. *Nonlinearity*, 21(12): T273, 2008.
35. Pang Xiao-Feng and Feng Yuan-Ping. *Quantum Mechanics in Nonlinear Systems*. World Scientific, 2005. ISBN 978-981-4481-23-6.
36. Thomas F. Jordan. Why quantum dynamics is linear. *Journal of Physics: Conference Series*, 196(1):012010, 2009.
37. K. Nooter and H. Herweijer. Multidrug resistance (mdr) genes in human cancer. *British Journal of Cancer*, 63(5):663–669, 1991.
38. Damon Linker. Libertarianism’s terrible, horrible, no good, very bad idea, September 2014. URL <http://theweek.com/articles/443462/libertarianisms-terrible-horrible-no-good-bad-idea>.
39. W. W. In defence of spontaneous order. *The Economist*, 2014.
40. Jochen Fromm. *The emergence of complexity*.
41. Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.
42. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

43. Hiroaki Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11): 826–837, 2004.
44. Mel Greaves and Carlo C. Maley. Clonal evolution in cancer. *Nature*, 481 (7381):306–313, 2012.
45. Lauren M. F. Merlo, John W. Pepper, Brian J. Reid, and Carlo C. Maley. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6(12):924–935.
46. John W. Pepper, C. Scott Findlay, Rees Kassen, Sabrina L. Spencer, and Carlo C. Maley. SYNTHESIS: Cancer research meets evolutionary biology. *Evolutionary Applications*, 2(1):62–70, 2009.
47. Christoph A. Klein. Parallel progression of primary tumours and metastases. *Nature Reviews Cancer*, 9(4):302–312, 2009.
48. E. P. Malaise, N. Chavaudra, and M. Tubiana. The relationship between growth rate, labelling index and histological type of human solid tumours. *European Journal of Cancer (1965)*, 9(4):305–312, 1973.
49. Alexander R. A. Anderson, Alissa M. Weaver, Peter T. Cummings, and Vito Quaranta. Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. *Cell*, 127(5):905–915, 2006.
50. Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, 2011.
51. Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, 2013.
52. Andrea Sottoriva, Inmaculada Spiteri, Darryl Shibata, Christina Curtis, and Simon Tavaré. Single-Molecule Genomic Data Delineate Patient-Specific Tumor Profiles and Cancer Stem Cell Organization. *Cancer Research*, 73(1):41–49, 2013.

53. Shinichi Yachida, Siân Jones, Ivana Bozic, Tibor Antal, Rebecca Leary, Baojin Fu, Mihoko Kamiyama, Ralph H. Hruban, James R. Eshleman, Martin A. Nowak, Victor E. Velculescu, Kenneth W. Kinzler, Bert Vogelstein, and Christine A. Iacobuzio-Donahue. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319): 1114–1117, 2010.
54. Manel Esteller. Epigenetics in Cancer. *New England Journal of Medicine*, 358(11):1148–1159, 2008.
55. Juan Sandoval and Manel Esteller. Cancer epigenomics: beyond genomics. *Current Opinion in Genetics & Development*, 22(1):50–55, 2012.
56. Hoi-Hung Cheung, Tin-Lap Lee, Owen M. Rennert, and Wai-Yee Chan. DNA Methylation of Cancer Genome. *Birth defects research. Part C, Embryo today : reviews*, 87(4):335–350, 2009.
57. Ryan Lister, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A. Harvey Millar, James A. Thomson, Bing Ren, and Joseph R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.
58. Melanie R. Hassler and Gerda Egger. Epigenomics of cancer – emerging new concepts. *Biochimie*, 94(11):2219–2230, 2012.
59. Jonathan E. Dodge, Bernard H. Ramsahoye, Z. Galen Wo, Masaki Okano, and En Li. De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene*, 289(1–2):41–48, 2002.
60. Thomas R. Haines, David I. Rodenhiser, and Peter J. Ainsworth. Allele-Specific Non-CpG Methylation of the Nf1 Gene during Early Mouse Development. *Developmental Biology*, 240(2):585–598, 2001.
61. Ryan Lister, Eran A. Mukamel, Joseph R. Nery, Mark Urich, Clare A. Puddifoot, Nicholas D. Johnson, Jacinta Lucero, Yun Huang, Andrew J. Dwork, Matthew D. Schultz, Miao Yu, Julian Tonti-Filippini, Holger Heyn,

- Shijun Hu, Joseph C. Wu, Anjana Rao, Manel Esteller, Chuan He, Fate-meh G. Haghghi, Terrence J. Sejnowski, M. Margarita Behrens, and Joseph R. Ecker. Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science*, 341(6146):1237905, 2013.
62. María Berdasco and Manel Esteller. Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry. *Developmental Cell*, 19(5): 698–711, 2010.
63. Andrew P. Feinberg and Benjamin Tycko. The history of cancer epigenetics. *Nature Reviews Cancer*, 4(2):143–153, 2004.
64. Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyani, Benjamin Langmead, Oliver G. McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A. Irizarry, and Andrew P. Feinberg. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775, 2011.
65. T. Schoofs, W. E. Berdel, and C. Müller-Tidow. Origins of aberrant DNA methylation in acute myeloid. *Leukemia*, 28(1):1–14, 2014.
66. François Delhommeau, Sabrina Dupont, Véronique Della Valle, Chloé James, Severine Trannoy, Aline Massé, Olivier Kosmider, Jean-Pierre Le Couedic, Fabienne Robert, Antonio Alberdi, Yann Lécluse, Isabelle Plo, François J. Dreyfus, Christophe Marzac, Nicole Casadevall, Catherine Lacombe, Serge P. Romana, Philippe Dessen, Jean Soulier, Franck Viguié, Michaela Fontenay, William Vainchenker, and Olivier A. Bernard. Mutation in TET2 in Myeloid Cancers. *New England Journal of Medicine*, 360(22):2289–2301, 2009.
67. Maria E. Figueroa, Omar Abdel-Wahab, Chao Lu, Patrick S. Ward, Jay Patel, Alan Shih, Yushan Li, Neha Bhagwat, Aparna Vasanthakumar, Hugo F. Fernandez, Martin S. Tallman, Zhuoxin Sun, Kristy Wolniak, Justine K. Peeters, Wei Liu, Sung E. Choe, Valeria R. Fantin, Elisabeth Paietta, Bob Löwenberg, Jonathan D. Licht, Lucy A. Godley, Ruud Delwel, Peter J. M. Valk, Craig B. Thompson, Ross L. Levine, and Ari Melnick.

Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell*, 18(6):553–567, 2010.

68. Elaine R. Mardis, Li Ding, David J. Dooling, David E. Larson, Michael D. McLellan, Ken Chen, Daniel C. Koboldt, Robert S. Fulton, Kim D. DeLahanty, Sean D. McGrath, Lucinda A. Fulton, Devin P. Locke, Vincent J. Magrini, Rachel M. Abbott, Tammi L. Vickery, Jerry S. Reed, Jody S. Robinson, Todd Wylie, Scott M. Smith, Lynn Carmichael, James M. Eldred, Christopher C. Harris, Jason Walker, Joshua B. Peck, Feiyu Du, Adam F. Dukes, Gabriel E. Sanderson, Anthony M. Brummett, Eric Clark, Joshua F. McMichael, Rick J. Meyer, Jonathan K. Schindler, Craig S. Pohl, John W. Wallis, Xiaoqi Shi, Ling Lin, Heather Schmidt, Yuzhu Tang, Carrie Haipek, Madeline E. Wiechert, Jolynda V. Ivy, Joelle Kalicki, Glendoria Elliott, Rhonda E. Ries, Jacqueline E. Payton, Peter Westervelt, Michael H. Tomasson, Mark A. Watson, Jack Baty, Sharon Heath, William D. Shannon, Rakesh Nagarajan, Daniel C. Link, Matthew J. Walter, Timothy A. Graubert, John F. DiPersio, Richard K. Wilson, and Timothy J. Ley. Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *New England Journal of Medicine*, 361(11):1058–1066, 2009.
69. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*, 368(22):2059–2074, May 2013.
70. Masato Sasaki, Christiane B. Knobbe, Joshua C. Munger, Evan F. Lind, Dirk Brenner, Anne Brüstle, Isaac S. Harris, Roxanne Holmes, Andrew Wakeham, Jillian Haight, Annick You-Ten, Wanda Y. Li, Stefanie Schalm, Shinsan M. Su, Carl Virtanen, Guido Reifenberger, Pamela S. Ohashi, Dwayne L. Barber, Maria E. Figueroa, Ari Melnick, Juan-Carlos Zúñiga-Pflücker, and Tak W. Mak. IDH1(R132h) mutation increases murine haematopoietic progenitors and alters epigenetics. *Nature*, 488(7413): 656–659, 2012.

71. Lambert Busque, Jay P. Patel, Maria E. Figueroa, Aparna Vasanthakumar, Sylvie Provost, Zineb Hamilou, Luigina Mollica, Juan Li, Agnes Viale, Adriana Heguy, Maryam Hassimi, Nicholas Socci, Parva K. Bhatt, Mithat Gonen, Christopher E. Mason, Ari Melnick, Lucy A. Godley, Cameron W. Brennan, Omar Abdel-Wahab, and Ross L. Levine. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nature Genetics*, 44(11):1179–1181, 2012.
72. Xiao-Jing Yan, Jie Xu, Zhao-Hui Gu, Chun-Ming Pan, Gang Lu, Yang Shen, Jing-Yi Shi, Yong-Mei Zhu, Lin Tang, Xiao-Wei Zhang, Wen-Xue Liang, Jian-Qing Mi, Huai-Dong Song, Ke-Qin Li, Zhu Chen, and Sai-Juan Chen. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3a in acute monocytic leukemia. *Nature Genetics*, 43(4):309–315, 2011.
73. Myunggon Ko, Yun Huang, Anna M. Jankowska, Utz J. Pape, Mamta Tahiliani, Hozefa S. Bandukwala, Jungeun An, Edward D. Lamperti, Kian Peng Koh, Rebecca Ganetzky, X. Shirley Liu, L. Aravind, Suneet Agarwal, Jaroslaw P. Maciejewski, and Anjana Rao. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature*, 468(7325):839–843, 2010.
74. Marco P. Boks, Eske M. Derks, Daniel J. Weisenberger, Erik Strengman, Esther Janson, Iris E. Sommer, René S. Kahn, and Roel A. Ophoff. The Relationship of DNA Methylation with Age, Gender and Genotype in Twins and Healthy Controls. *PLoS ONE*, 4(8):e6767, 2009.
75. Brock C. Christensen, E. Andres Houseman, Carmen J. Marsit, Shichun Zheng, Margaret R. Wrensch, Joseph L. Wiemels, Heather H. Nelson, Margaret R. Karagas, James F. Padbury, Raphael Bueno, David J. Sugarbaker, Ru-Fang Yeh, John K. Wiencke, and Karl T. Kelsey. Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLoS Genet*, 5(8):e1000602, 2009.
76. Jordana T. Bell, Pei-Chien Tsai, Tsun-Po Yang, Ruth Pidsley, James Nisbet, Daniel Glass, Massimo Mangino, Guangju Zhai, Feng Zhang, Ana



- Valdes, So-Youn Shin, Emma L. Dempster, Robin M. Murray, Elin Grundberg, Asa K. Hedman, Alexandra Nica, Kerrin S. Small, Emmanouil T. Dermitzakis, Mark I. McCarthy, Jonathan Mill, Tim D. Spector, Panos Deloukas, and The MuTHER Consortium. Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population. *PLoS Genet*, 8(4):e1002629, 2012.
77. Reid S. Alisch, Benjamin G. Barwick, Pankaj Chopra, Leila K. Myrick, Glen A. Satten, Karen N. Conneely, and Stephen T. Warren. Age-associated DNA methylation in pediatric populations. *Genome Research*, 22(4):623–632, 2012.
78. Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Satta, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, 49(2):359–367, 2013.
79. Holger Heyn, Ning Li, Humberto J. Ferreira, Sebastian Moran, David G. Pisano, Antonio Gomez, Javier Diez, Jose V. Sanchez-Mut, Fernando Satten, F. Javier Carmona, Annibale A. Puca, Sergi Sayols, Miguel A. Pujana, Jordi Serra-Musach, Isabel Iglesias-Platas, Francesc Formiga, Agustin F. Fernandez, Mario F. Fraga, Simon C. Heath, Alfonso Valencia, Ivo G. Gut, Jun Wang, and Manel Esteller. Distinct DNA methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences*, 109(26):10522–10527, 2012.
80. Chris Murgatroyd, Alexandre V. Patchev, Yonghe Wu, Vincenzo Micale, Yvonne Bockmühl, Dieter Fischer, Florian Holsboer, Carsten T. Wotjak, Osborne F. X. Almeida, and Dietmar Spengler. Dynamic DNA methylation programs persistent adverse effects of early-life stress. *Nature Neuroscience*, 12(12):1559–1566, 2009.
81. Dengke K. Ma, Mi-Hyeon Jang, Junjie U. Guo, Yasuji Kitabatake, Min-lin Chang, Nattapol Pow-anpongkul, Richard A. Flavell, Binfeng Lu, Guo-li

- Ming, and Hongjun Song. Neuronal Activity–Induced Gadd45b Promotes Epigenetic DNA Demethylation and Adult Neurogenesis. *Science*, 323 (5917):1074–1077, 2009.
82. Victoria K. Cortessis, Duncan C. Thomas, A. Joan Levine, Carrie V. Breton, Thomas M. Mack, Kimberly D. Siegmund, Robert W. Haile, and Peter W. Laird. Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. *Human Genetics*, 131 (10):1565–1589, 2012.
83. Peter Yakovchuk, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*, 34(2):564–574, 2006.
84. J. Marmur and P. Doty. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *Journal of Molecular Biology*, 5(1):109–118, 1962.
85. Petr Šmarda, Petr Bureš, Jakub Šmerda, and Lucie Horová. Measurements of genomic GC content in plant genomes with flow cytometry: a test for reliability. *New Phytologist*, 193(2):513–521, 2012.
86. Michael Lynch. *The Origins of Genome Architecture*. Sinauer Associates Inc, Sunderland, Mass, 1 edition edition, March 2007. ISBN 978-0-87893-484-3.
87. Brian Charlesworth and Deborah Charlesworth. *Elements of Evolutionary Genetics*. Roberts & Company Publishers, Greenwood Village, Colo, February 2010. ISBN 978-0-9815194-2-5.
88. Ki Yong Lee, R Wahl, and E Barbu. Contenu en bases puriques et pyrimidiques des acides désoxyribonucléiques des bactéries. In *ANNALES DE L'INSTITUT PASTEUR*, volume 91, pages 212–224. MASSON EDITEUR 120 BLVD SAINT-GERMAIN, 75280 PARIS 06, FRANCE, 1956.

89. Scott Mann and Yi-Ping Phoebe Chen. Bacterial genomic G + C composition-eliciting environmental adaptation. *Genomics*, 95(1):7–15, 2010.
90. Luciano Brocchieri. The GC Content of Bacterial Genomes. *Journal of Phylogenetics & Evolutionary Biology*, 02(01), 2014.
91. Stephen D. Bentley and Julian Parkhill. Comparative Genomic Structure of Prokaryotes. *Annual Review of Genetics*, 38(1):771–791, 2004.
92. Eduardo P. C. Rocha and Antoine Danchin. Base composition bias might result from competition for metabolic resources. *Trends in Genetics*, 18(6):291–294, 2002.
93. Ravi Jain, Maria C. Rivera, and James A. Lake. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7):3801–3806, 1999.
94. Laura S. Frost, Raphael Leplae, Anne O. Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005.
95. Ren Zhang and Chun-Ting Zhang. A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics (Oxford, England)*, 20(5):612–622, 2004.
96. Sung Ho Yoon, Cheol-Goo Hur, Ho-Young Kang, Yeoun Hee Kim, Tae Kwang Oh, and Jihyun F Kim. A computational approach for identifying pathogenicity islands in prokaryotic genomes. *BMC Bioinformatics*, 6:184, 2005.
97. Héctor Musto, Hugo Naya, Alejandro Zavala, Héctor Romero, Fernando Alvarez-Valín, and Giorgio Bernardi. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Letters*, 573(1–3):73–77, 2004.

98. Huai-Chun Wang, Edward Susko, and Andrew J. Roger. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: Data quality and confounding factors. *Biochemical and Biophysical Research Communications*, 342(3):681–684, 2006.
99. Héctor Musto, Hugo Naya, Alejandro Zavala, Héctor Romero, Fernando Alvarez-Valén, and Giorgio Bernardi. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochemical and Biophysical Research Communications*, 347(1):1–3, 2006.
100. Hugo Naya, Héctor Romero, Alejandro Zavala, Beatriz Alvarez, and Héctor Musto. Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (GC%) in Prokaryotes. *Journal of Molecular Evolution*, 55(3):260–264, 2002.
101. Junko Kusumi and Hidenori Tachida. Compositional Properties of Green-Plant Plastid Genomes. *Journal of Molecular Evolution*, 60(4):417–425, 2005.
102. Xiang Jia Min and Donal A. Hickey. DNA Barcodes Provide a Quick Preview of Mitochondrial Genome Composition. *PLoS ONE*, 2(3):e325, 2007.
103. David Roy Smith. Updating Our View of Organelle Genome Nucleotide Landscape. *Frontiers in Genetics*, 3, 2012.
104. A. P. Martin. Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. *Molecular Biology and Evolution*, 12(6):1124–1131, 1995.
105. Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.
106. Kristin P. Bennett and Colin Campbell. Support Vector Machines: Hype or Hallelujah? *SIGKDD Explor. Newsl.*, 2(2):1–13, 2000.
107. Introduction. In *Support Vector Machines*, Information Science and Statistics. Springer New York, 2008. ISBN 978-0-387-77241-7.

108. V Vapnik and A Lerner. Pattern Recognition using Generalized Portrait Method. *Automation and Remote Control*, 24, 1963.
109. V Vapnik and A Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.
110. V. N. Vapnik and A. Ya. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, USSR, 1974.
111. *Estimation of Dependences Based on Empirical Data*. Information Science and Statistics. Springer New York, 2006. ISBN 978-0-387-30865-4. URL <http://link.springer.com/10.1007/0-387-34239-7>.
112. Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer New York, New York, NY, 2000. ISBN 978-1-4419-3160-3 978-1-4757-3264-1. URL <http://link.springer.com/10.1007/978-1-4757-3264-1>.
113. Martin Trauth, R. Gebbers, and N. Marwan. *MATLAB® Recipes for Earth Sciences*. Springer, 3rd ed. 2010 edition, 2010.
114. Boris P Bezruchko and Dmitry A Smirnov. *Extracting Knowledge From Time Series: An Introduction to Nonlinear Empirical Modeling*. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2010.
115. J. P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57(3):617–656, July 1985.
116. Schelter. *Handbook of Time Series Analysis - Recent Theoretical Developments and Applications*. Wiley-Blackwell, 2006.
117. Martin Casdagli, Stephen Eubank, J.Doyne Farmer, and John Gibson. State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, 51(1–3):52–98, 1991.
118. Andrew M. Fraser and Harry L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2): 1134–1140, February 1986.

119. Rainer Hegger, Holger Kantz, and Thomas Schreiber. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9:413, 1999.
120. Marc-Thorsten Hütt and Manuel Dehnert. *Methoden der Bioinformatik: Eine Einführung*. Springer-Verlag, 2006.
121. Antonio Fabio Di Narzo. *tseriesChaos: Analysis of nonlinear time series*, 2013. URL <http://CRAN.R-project.org/package=tseriesChaos>. R package version 0.1-13.
122. James P. Crutchfield, J. Doyne Farmer, Norman H. Packard, and Robert S. Shaw. Chaos. *Scientific American*, 255(6):38–49, 1986.
123. H. Broer, F. Takens, and B. Hasselblatt. *Handbook of Dynamical Systems*. Number Bd. 3 in Handbook of Dynamical Systems. Elsevier Science, 2010. ISBN 9780080932262.
124. G. Greco, R. Rosa, G. Beskin, S. Karpov, L. Romano, A. Guarnieri, C. Bartolini, and R. Bedogni. Evidence of Deterministic Components in the Apparent Randomness of GRBs: Clues of a Chaotic Dynamic. *Scientific Reports*, 1, 2011.
125. Elisa Beninca, Jef Huisman, Reinhard Heerkloss, Klaus D. Johnk, Pedro Branco, Egbert H. Van Nes, Marten Scheffer, and Stephen P. Ellner. Chaos in a long-term experiment with a plankton community. *Nature*, 451(7180):822–825, 2008.
126. Kennel, Brown, and Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review. A*, 45(6):3403–3411, 1992. PMID: 9907388.
127. Marco Thiel, M. Carmen Romano, and Jürgen Kurths. How much information is contained in a recurrence plot? *Physics Letters A*, 330(5): 343–349, 2004.
128. Norbert Marwan, M. Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, 2007.

129. C. Allain and M. Cloitre. Characterizing the lacunarity of random and deterministic fractal sets. *Physical Review A*, 44(6):3552–3558, 1991.
130. Joseph P. Zbilut, Nitza Thomasson, and Charles L. Webber. Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals. *Medical Engineering & Physics*, 24(1):53–60, 2002.
131. Roy E. Plotnick, Robert H. Gardner, and Robert V. O’Neill. Lacunarity indices as measures of landscape texture. *Landscape Ecology*, 8(3):201–211, 1993.
132. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang (libsvm C++-code), and Chih-Chen Lin (libsvm C++-code). e1071: Misc functions of the department of statistics (e1071), TU wien, 2014. URL <http://cran.r-project.org/web/packages/e1071/index.html>.
133. Tristan Fletcher. Support Vector Machines Explained, 2008. URL <http://www.tristanfletcher.co.uk/SVM%20Explained.pdf>.
134. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.
135. Zuguang Gu. *ComplexHeatmap: Making Complex Heatmaps*, 2015. URL <https://github.com/jokergoo/ComplexHeatmap>. R package version 1.0.0.
136. Zuo-Bing Wu. Recurrence plot analysis of DNA sequences. *Physics Letters A*, 332(3–4):250–255, 2004.
137. Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke

- Tierney, Jean Y. H. Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.
138. Wolfgang Huber, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D. Hansen, Rafael A. Irizarry, Michael Lawrence, Michael I. Love, James MacDonald, Valerie Obenchain, Andrzej K. Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K. Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015.
139. Norbert Marwan. How to avoid potential pitfalls in recurrence plot based data analysis. *International Journal of Bifurcation and Chaos*, 21(04): 1003–1017, 2011.
140. The cancer genome atlas home page. URL <http://cancergenome.nih.gov/>.
141. Data Matrix - Data Portal, . URL <https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm>.
142. Tim F. Rayner, Philippe Rocca-Serra, Paul T. Spellman, Helen C. Causton, Anna Farne, Ele Holloway, Rafael A. Irizarry, Junmin Liu, Donald S. Maier, Michael Miller, Kjell Petersen, John Quackenbush, Gavin Sherlock, Christian J. Stoeckert, Joseph White, Patricia L. Whetzel, Farrell Wymore, Helen Parkinson, Ugis Sarkans, Catherine A. Ball, and Alvis Brazma. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, 7(1):489, 2006.
143. FGED: MAGE-TAB. URL <http://fged.org/projects/mage-tab/>.
144. Sample and Data Relationship Format - TCGA - National Cancer Institute - Confluence Wiki, . URL <https://wiki.nci.nih.gov/x/9aFXAg>.
145. Tab2mage. URL <http://tab2mage.sourceforge.net/>.



146. Xerces-C++ XML Parser. URL <https://xerces.apache.org/xerces-c/>.
147. Tim F. Rayner. Bio-MAGETAB. URL <http://search.cpan.org/dist/Bio-MAGETAB/>.
148. X.667 : Information technology - Procedures for the operation of object identifier registration authorities: Generation of universally unique identifiers and their use in object identifiers. URL <http://www.itu.int/rec/T-REC-X.667-201210-I/en>.
149. TCGA barcode - TCGA - National Cancer Institute - Confluence Wiki, . URL <https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode>.
150. Code Tables Report - Data Portal, . URL <https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=Tissue%20Source%20Site>.
151. Murray Gell-Mann. What is complexity? Remarks on simplicity and complexity by the Nobel Prize-winning author of *The Quark and the Jaguar*. *Complexity*, 1(1):16–19, 1995.
152. Infinium HumanMethylation450 BeadChip - product\_info\_hm450.pdf. URL [http://www.illumina.com/content/dam/illumina-marketing/documents/products/product\\_information\\_sheets/product\\_info\\_hm450.pdf](http://www.illumina.com/content/dam/illumina-marketing/documents/products/product_information_sheets/product_info_hm450.pdf).
153. Joseph P. Zbilut and Charles L. Webber. Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*, 171(3):199–203, 1992.
154. C. L. Webber and J. P. Zbilut. Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, 76(2):965–973, 1994.
155. Charles L. Webber Jr., Norbert Marwan, Angelo Facchini, and Alessandro Giuliani. Simpler methods do it better: Success of Recurrence Quantification Analysis as a general purpose data analysis tool. *Physics Letters A*, 373(41):3753–3756, 2009.

156. Charles L. Webber, Jr and Joseph P. Zbilut. Recurrence Quantification Analysis of Nonlinear Dynamical Systems. In *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences Web Book*. .
157. Jr Charles L. Webber and Norbert Marwan. *Recurrence Quantification Analysis: Theory and Best Practices*. Jr., Norbert Marwan. Webber. Springer, 2014. ISBN 978-3-319-07155-8.
158. G. Andrés, N. Ashour, M. Sánchez-Chapado, S. Ropero, and J. C. Angulo. The study of DNA methylation in urological cancer: Present and future. *Actas Urológicas Españolas (English Edition)*, 37(6):368–375, 2013.
159. Dominique J. P. M. Stumpel, Pauline Schneider, Eddy H. J. van Roon, Judith M. Boer, Paola de Lorenzo, Maria G. Valsecchi, Renee X. de Menezes, Rob Pieters, and Ronald W. Stam. Specific promoter methylation identifies different subgroups of MLL-rearranged infant acute lymphoblastic leukemia, influences clinical outcome, and provides therapeutic options. *Blood*, 114(27):5490–5498, 2009.
160. Lien Van De Voorde, Reinhart Speeckaert, Dirk Van Gestel, Marc Bracke, Wilfried De Neve, Joris Delanghe, and Marijn Speeckaert. DNA methylation-based biomarkers in serum of patients with breast cancer. *Mutation Research/Reviews in Mutation Research*, 751(2):304–325, 2012.
161. Agustin F. Fernandez, Yassen Assenov, Jose Ignacio Martin-Subero, Balazs Balint, Reiner Siebert, Hiroaki Taniguchi, Hiroyuki Yamamoto, Manuel Hidalgo, Aik-Choon Tan, Oliver Galm, Isidre Ferrer, Montse Sanchez-Céspedes, Alberto Villanueva, Javier Carmona, Jose V. Sanchez-Mut, Maria Berdasco, Victor Moreno, Gabriel Capella, David Monk, Esteban Ballestar, Santiago Ropero, Ramon Martinez, Marta Sanchez-Carbayo, Felipe Prosper, Xabier Agirre, Mario F. Fraga, Osvaldo Grana, Luis Perez-Jurado, Jaume Mora, Susana Puig, Jaime Prat, Lina Badimon, Annibale A. Puca, Stephen J. Meltzer, Thomas Lengauer, John Bridgewater, Christoph Bock, and Manel Esteller. A DNA methylation fingerprint of 1628 human samples. *Genome Research*, 22(2):407–419, 2012.

162. Manel Esteller. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8(4):286–298, 2007.
163. Femke Simmer, Arie B. Brinkman, Yassen Assenov, Filomena Matarese, Anita Kaan, Lina Sabatino, Alberto Villanueva, Dori Huertas, Manel Esteller, Thomas Lengauer, Christoph Bock, Vittorio Colantuoni, Lucia Altucci, and Hendrik G. Stunnenberg. Comparative genome-wide DNA methylation analysis of colorectal tumor and matched normal tissues. *Epigenetics*, 7(12):1355–1367, 2012.
164. Jung H. Kim, Saravana M. Dhanasekaran, John R. Prensner, Xuhong Cao, Daniel Robinson, Shanker Kalyana-Sundaram, Christina Huang, Sunita Shankar, Xiaojun Jing, Matthew Iyer, Ming Hu, Lee Sam, Catherine Grasso, Christopher A. Maher, Nallasivam Palanisamy, Rohit Mehra, Hal D. Kominsky, Javed Siddiqui, Jindan Yu, Zhaohui S. Qin, and Arul M. Chinnaiyan. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome Research*, 21(7):1028–1041, 2011.
165. Maté Ongenaert, Leander Van Neste, Tim De Meyer, Gerben Menschaert, Sofie Bekaert, and Wim Van Criekinge. PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Research*, 36(Database issue):D842–846, 2008.
166. Stephen B. Baylin. 5 - Epigenetics and Cancer. In John Mendelsohn-Joe W. GrayPeter M. HowleyMark A. IsraelCraig B. Thompson, editor, *The Molecular Basis of Cancer (Fourth Edition)*, pages 67–78.e3. Content Repository Only!, Philadelphia, 2015. ISBN 978-1-4557-4066-6.
167. Benjamin P. Berman, Daniel J. Weisenberger, Joseph F. Aman, Toshinori Hinoue, Zachary Ramjan, Yaping Liu, Houtan Noushmehr, Christopher P. E. Lange, Cornelis M. van Dijk, Rob A. E. M. Tollenaar, David Van Den Berg, and Peter W. Laird. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature Genetics*, 44(1):40–46, 2012.
168. Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragozy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Pe-

- ter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, 2009.
169. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
170. groupschoof/AHRD, . URL <https://github.com/groupschoof/AHRD>.
171. interpro2go. URL <http://geneontology.org/external2go/interpro2go>.
172. Alex Mitchell, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, Amaia Sangrador-Vegas, Maxim Scheremetjew, Claudia Rato, Siew-Yit Yong, Alex Bateman, Marco Punta, Teresa K. Attwood, Christian J. A. Sigrist, Nicole Redaschi, Catherine Rivoire, Ioannis Xenarios, Daniel Kahn, Dominique Guyot, Peer Bork, Ivica Letunic, Julian Gough, Matt Oates, Daniel Haft, Hongzhan Huang, Darren A. Natale, Cathy H. Wu, Christine Orengo, Ian Sillitoe, Huaiyu Mi, Paul D. Thomas, and Robert D. Finn. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, 43(D1): D213–D221, 2015.
173. Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

174. groupschoof/PhyloFun, . URL <https://github.com/groupschoof/PhyloFun>.
175. Asis Hallab. *Protein Function Prediction using Phylogenomics, Domain Architecture Analysis, Data Integration, and Lexical Scoring*. Dissertation, Universität Bonn, Bonn, 2015.
176. Barbara E Engelhardt, Michael I Jordan, Kathryn E Muratore, and Steven E Brenner. Protein Molecular Function Prediction by Bayesian Phylogenomics. *PLoS Comput Biol*, 1(5):e45, 2005.
177. Anika Jöcker. *Automatic and manual functional annotation in a distributed web service environment*. Dissertation, Universität zu Köln, Köln, 2009.
178. MIPS ITAG2.3\_release, . URL [ftp://ftpmips.helmholtz-muenchen.de/plants/tomato/tomato\\_genome/ITAG\\_annotation/ITAG2.3\\_release/](ftp://ftpmips.helmholtz-muenchen.de/plants/tomato/tomato_genome/ITAG_annotation/ITAG2.3_release/).
179. SGN ITAG2.3\_release, . URL [ftp://ftp.solgenomics.net/genomes/Solanum\\_lycopersicum/annotation/ITAG2.3\\_release/](ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/annotation/ITAG2.3_release/).
180. Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7): 976–978, 2010. URL <http://bioinformatics.oxfordjournals.org/content/26/7/976>.
181. Rasko Leinonen, Ruth Akhtar, Ewan Birney, James Bonfield, Lawrence Bower, Matt Corbett, Ying Cheng, Fehmi Demiralp, Nadeem Faruque, Neil Goodgame, Richard Gibson, Gemma Hoad, Christopher Hunter, Mikyung Jang, Steven Leonard, Quan Lin, Rodrigo Lopez, Michael Maguire, Hamish McWilliam, Sheila Plaister, Rajesh Radhakrishnan, Siamak Sobhany, Guy Slater, Petra Ten Hoopen, Franck Valentin, Robert Vaughan, Vadim Zalunin, Daniel Zerbino, and Guy Cochrane. Improvements to services at the European Nucleotide Archive. *Nucleic Acids Research*, 38(suppl 1):D39–D45, 2010.

182. European Nucleotide Archive. URL <http://www.ebi.ac.uk/ena>.
183. N. Galtier, B. Nabholz, S. Glémin, and G. D. D Hurst. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology*, 18 (22):4541–4550, 2009.
184. Rute da Fonseca, Warren Johnson, Stephen O'Brien, Maria Ramos, and Agostinho Antunes. The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics*, 9(1):119, 2008.
185. Dent Earl, Ngan Nguyen, Glenn Hickey, Robert S. Harris, Stephen Fitzgerald, Kathryn Beal, Igor Seledtsov, Vladimir Molodtsov, Brian J. Raney, Hiram Clawson, Jaebum Kim, Carsten Kemena, Jia-Ming Chang, Ionas Erb, Alexander Poliakov, Minmei Hou, Javier Herrero, William James Kent, Victor Solovyev, Aaron E. Darling, Jian Ma, Cedric Notredame, Michael Brudno, Inna Dubchak, David Haussler, and Benedict Paten. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Research*, 24(12):2077–2089, 2014.
186. Mathieu Blanchette, W. James Kent, Cathy Riemer, Laura Elnitski, Arjan F.A. Smit, Krishna M. Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D. Green, David Haussler, and Webb Miller. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research*, 14(4):708–715, 2004.
187. Aaron E. Darling, Bob Mau, and Nicole T. Perna. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE*, 5(6):e11147, 2010.
188. Inna Dubchak, Alexander Poliakov, Andrey Kislyuk, and Michael Brudno. Multiple whole-genome alignments without a reference organism. *Genome Research*, 19(4):682–689, 2009.
189. Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, 21(9):1512–1528, 2011.

190. Charles L. Webber, Jr and Joseph P. Zbilut. Recurrence Quantification Analysis of Nonlinear Dynamical Systems. In *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences Web Book*. .
191. Matthew W. Mitchell, Sabrina Locatelli, Lora Ghobrial, Amy A. Pokempner, Paul R. Sesink Clee, Ekwoge E. Abwe, Aaron Nicholas, Louis Nkempi, Nicola M. Anthony, Bethan J. Morgan, Roger Fotso, Martine Peeters, Beatrice H. Hahn, and Mary Katherine Gonder. The population genetics of wild chimpanzees in Cameroon and Nigeria suggests a positive role for selection in the evolution of chimpanzee subspecies. *BMC Evolutionary Biology*, 15(1):1–15, 2015.
192. Rory Bowden, Tammie S. MacFie, Simon Myers, Garrett Hellenthal, Eric Nerrienet, Ronald E. Bontrop, Colin Freeman, Peter Donnelly, and Nicholas I. Mundy. Genomic Tools for Evolution and Conservation in the Chimpanzee: *Pan troglodytes ellioti* Is a Genetically Distinct Population. *PLoS Genet*, 8(3):e1002504, 2012.

---

## Online resources

The `bract` software package and electronic supplemental data, tables and graphs are available under the URL <https://github.com/bractproject/bract>





---

## Acknowledgments

I express my deepest gratitude to Hilario Ramírez Rodrigo for his excellent guidance, professional supervision and encouragement. I have enjoyed every minute we had for conversations and discussions – very special mixtures of science and culture – and I hope we will have many more in future.

Whenever I passed through the door into the Department of Biochemistry and Molecular Biology I have felt home and therefore I will miss that place. A very sincere thank you to José Antonio Lupiáñez Cara and his "crew".

I am also very thankful to my former group leader and colleagues from the "Plant Computational Biology" and "Crop Bioinformatics" groups. I appreciate their inspiration, friendly ears and the always great and pleasant atmosphere.

This work would not have been possible without the computational power and meaningful support from the Supercomputation Service of the University of Granada and the Crop Bioinformatics group of the University of Bonn. Thanks a lot to them!

To my parents and grandparents, I thank them for their education and unconditional love.

And, last but certainly not least, this work could not have been done without the enormous support by my family – Inma, Alexander, Noelia and Irene. They continuously warm my heart and soul. Words would never say how grateful I am to them.

FIN