UNIVERSIDAD DE GRANADA

ENGINEERING, EVOLUTION AND FOLDING OF ENZYMES WITH NATURAL AND NON NATURAL ACTIVITIES

(Programa de Doctorado en Química)

María Gloria Gámiz Arco

Departamento de Química Física

Facultad de Ciencias

Universidad de Granada

Noviembre 2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: María Gloria Gámiz Arco
ISBN: 978-84-1117-234-9
URI: http://hdl.handle.net/10481/72865

INDEX

LIST OF FIGURES

ABBREVIATIONS

3D	Three-dimensional
ASR	Ancestral sequence reconstruction
BLAST	Basic local alignment search tool
CAZy	Carbohydrate-active enzymes
CPk	*Candidatus Photodesmus katoptron*
DSC	Differential scanning calorimetry
GH1	Glycoside hydrolase family 1
GH5	Glycoside hydrolase family 58
GH10	Glycoside hydrolase family 10
GH13	Glycoside hydrolase family 1
GH17	Glycoside hydrolase family 17
GH18	Glycoside hydrolase family 18
GH25	Glycoside hydrolase family 25
GH26	Glycoside hydrolase family 26
GH39	Glycoside hydrolase family 39
GH51	Glycoside hydrolase family 51
GH53	Glycoside hydrolase family 53
HFSP	Human frontier science program
Km	Michaelis constant
LBCA	Last bacterial common ancestor
LPBCA	Last common ancestor of the cyanobacterial, deinococcus and thermus groups
N72	Node 72
TIM	Triosephosphate isomerase

ABSTRACT

Ancestral proteins refer to proteins from extinct organisms that no longer exit on Earth. However, phylogenetic and bioinformatic analyses of modern protein sequences can lead to plausible approximations to the sequences of their ancestors. The methodology is called ancestral sequence reconstruction and allows for further preparation of the ancestral proteins in the laboratory (resurrection) and their characterization. It has been shown that resurrected ancestral proteins are very powerful tools to address important issues in evolution, and also to provide solutions to practical problems in biotechnological/biomedical scenarios due to their remarkable properties, quite different, in many cases, from those of their modern counterparts. In this doctoral thesis, both aspects have been investigated as summarized below.

First, we have explored the relationship between *in vitro* and *in vivo* protein folding, in an evolutionary context. We have experimentally characterized the *in vitro* folding of a set of resurrected Precambrian and modern thioredoxins and found that, contrary to previous claims in the literature, the folding rates are not evolutionarily conserved. Thus, ancestral thioredoxins fold much faster in the test tube than their modern counterparts. Extensive mutational analyses have allowed us to identify mutation S74G as responsible for aggravating folding in *E. coli* thioredoxin. The evolutionary acceptance of this mutation is interpreted as an example of degradation of ancestral features at the molecular level. We propose that unassisted and efficient primordial folding was linked to fast folding encoded at the sequence/structure level. Once an efficient assistance machinery had emerged, mutations that impaired ancient sequence/structure determinants of folding efficiency could be accepted, since those determinants were no longer necessary. We conclude that *in vitro* and *in vivo* folding landscapes are disconnected and question the biologically relevance of the *in vitro* folding rate determinations, except as related to heterologous folding efficiency (see below).

In the second part of this thesis, we have addressed a pivotal and common problem in biotechnology, the inefficient heterologous expression of proteins. As a model system thioredoxin from *Candidatus Photodesmus katoptron*, an uncultured symbiotic bacteria of flashlight fish, has been used. Our results demonstrate its slow *in vitro* folding (it takes several hours to reach the native state) and inefficient expression in *E. coli*, leading mostly to insoluble protein. By using a few back-to-the-ancestral mutations at positions selected by computational modelling of the unassisted folding landscape we were able to rescue its inefficient expression.

Our results support that the folding of proteins in foreign hosts may be akin to some extent to unassisted folding due to the absence of coevolution of the recombinant protein with the natural chaperones of the new host. More generally, our results provide an approach based on sequence engineering to rescue inefficient heterologous expression with a minimal protein perturbation.

Finally, the Appendix is dedicated to briefly summarize our results on the use of ancestral proteins as powerful tools for engineering *de novo* enzymes with non natural activities. We have resurrected and exhaustively characterized an ancestral TIM-barrel protein belonging to family 1 glycosidase, which displays remarkable properties such as thermostability and enhanced conformational flexibility. Surprisingly, the ancestral glycosidase binds heme tightly and stoichiometrically, and its binding enhances catalysis. This finding is unexpected because none of the reported crystallographic structures of the ~1400 modern glycosidases shows a bound porphyrin. Moreover, these features reveals promising applications in custom catalysis and biosensor engineering.

RESUMEN

Las proteínas ancestrales son proteínas de organismos extintos, y aunque estas proteínas no existan en la actualidad, es posible derivar aproximaciones de sus secuencias a partir de las secuencias de sus descendientes modernos mediante la reconstrucción de secuencias ancestrales (ASR). De esta manera, las proteínas codificadas por secuencias ancestrales reconstruidas pueden ser obtenidas en el laboratorio (resucitadas) y caracterizadas experimentalmente. Las proteínas ancestrales son de gran utilidad para abordar problemas relevantes relacionados con la evolución, y también pueden tener importantes aplicaciones biomédicas y biotecnológicas debido a que presentan propiedades únicas, en muchos casos, bastante diferentes de las propiedades de sus homologas modernas. En esta tesis doctoral se ha abordado el uso de enzimas ancestrales en ambos aspectos, resumidos a continuación.

En primer lugar, hemos explorado la relación entre el plegamiento de proteínas *in vitro* e *in vivo*, en un contexto evolutivo. Hemos caracterizado experimentalmente el plegamiento *in vitro* de un conjunto de tiorredoxinas precámbricas y modernas y hemos descubierto que, al contrario de lo que ha sido afirmado recientemente en la literatura, la velocidad de plegamiento no ha sido conservada a lo largo de la evolución. Las tiorredoxinas ancestrales se pliegan mucho más rápido *in vitro* que sus homólogas modernas. Un extenso análisis mutacional nos ha permitido identificar que la mutación S74G es la responsable de hacer más lento el plegamiento en la tiorredoxina de *E. coli*. La aceptación evolutiva de esta mutación se interpreta como un ejemplo de degradación de rasgos ancestrales a nivel molecular. En nuestro trabajo, se propone que el plegamiento primordial, supuestamente no asistido y eficiente, está ligado a un plegamiento rápido que debe estar codificado a nivel de secuencia / estructura. Pero, una vez surgió la maquinaria de asistencia para el plegamiento, se aceptaron mutaciones que dañaban los determinantes de secuencia / estructura que codificaban para un plegamiento rápido *in vitro*, ya que esta característica dejó de ser necesaria. Por tanto, en este trabajo se concluye que los paisajes de plegamiento *in vitro* e *in vivo* están desconectados y se cuestiona la importancia biológica de la velocidad de plegamiento *in vitro*, excepto en los casos de expresión heteróloga (ver más abajo).

En la segunda parte de esta tesis, hemos abordado uno de los problemas más comunes en biotecnología, la ineficiente expresión heteróloga de proteínas. Como sistema modelo, se ha utilizado la tiorredoxina de *Candidatus Photodesmus katoptron*, una bacteria simbiótica no cultivable del pez linterna. Nuestros resultados demuestran que esta proteína presenta un

plegamiento *in vitro* muy lento (necesita varias horas para alcanzar su estado nativo) y su expresión es ineficaz en *E. coli*, lo que lleva principalmente a la obtención de proteína insoluble. Mediante el uso de mutaciones de vuelta al ancestro, en determinadas posiciones seleccionadas por un modelo computacional del paisaje de plegamiento, pudimos rescatar su expresión. Nuestros resultados apoyan que el plegamiento de proteínas en un huésped diferente, puede ser similar, en cierta medida, al plegamiento no asistido debido a la ausencia de coevolución de la proteína recombinante con las chaperonas naturales del nuevo huésped. De manera más general, nuestros resultados desarrollan un enfoque basado en la ingeniería de secuencias para rescatar la expresión heteróloga ineficaz con una perturbación mínima de la proteína.

Por último, encontramos el Apéndice, el cual está dedicado a resumir brevemente nuestros resultados sobre el uso de proteínas ancestrales como herramientas para el diseño de enzimas *de novo* con actividades no naturales. Hemos resucitado y caracterizado exhaustivamente una proteína ancestral con estructura de barril TIM perteneciente a la familia 1 de la glicosidasas, que presenta propiedades muy interesantes como una alta estabilidad o una gran flexibilidad conformacional. Además, esta glicosidasa une hemo con fuerza y de forma estequiométra, y su unión mejora la catálisis. Esto fue un descubrimiento sorprendente debido a que ninguna de las estructuras cristalográficas de las ~1400 glicosidasas modernas presenta una porfirina unida. Además, este hallazgo abre una puerta muy interesante, en el uso de esta enzima ancestral en ingeniería de catálisis personalizada y en su uso como biosensor.

1. INTRODUCTION

1.1. Resurrected ancestral proteins

Ancestral proteins refer to proteins from extinct organisms. However, these proteins no longer exit on Earth, analyses of modern protein sequences can be used to derive plausible approximations to their sequences. This methodology is called ancestral sequence reconstruction and allows for further preparation of the ancestral proteins in the laboratory (resurrection) and their characterization.

The possibility of arriving at plausible approximations to their sequences was first proposed by Nobel-prize-winner chemist Linus Pauling and molecular biologist Emile Zuckerkandl in their 1963 seminal paper[1]. These authors noted that by comparison of the sequences of homologous proteins in modern organisms, a reasonable estimate of the sequence of the protein in a common ancestor can be derived. However, their proposal remained essentially a theoretical possibility until the 1990s when advances in bioinformatics and the increasing availability of protein sequences facilitated the systematic reconstruction of ancestral sequences. Furthermore, advances in molecular biology techniques allowed for the actual preparation in the laboratory (the "resurrection") of the corresponding encoded proteins and subsequent experimental characterization.

The procedure of ancestral sequence reconstruction (ASR) typically comprises several steps including the alignment of a collection of homologous sequences from different modern organisms, the construction of a phylogenetic tree describing the evolutionary relationships between the modern descendants and their common ancestors, and finally, the expression and characterization of ancestral proteins for studying their functionality and biochemical properties[2]. The reader is referred to published literature (Liberles, 2008)[3] for details in the computational methodology and the different programs available for the required phylogenetic and bioinformatic analyses of modern protein sequences.

Early exercises of protein resurrections were reported, independently, by the Benner[4] and Wilson[5] groups back in 1990 and correspond to lysozymes and artiodactyl pancreatic ribonucleases, respectively. The relevance of these studies is unquestionable as they represent not only the realization of Pauling and Zuckerkandl´s idea but the beginning of the paleoenzymology. Of course, it is also possible to extract the DNA from fossils and convert this

information into protein products. However, this method has many obstacles[6] and, moreover, the age of the ancestral proteins obtained is very limited (the oldest extracted DNA has been dated to be 450-800 thousand years ago[7], right after the Cambrian explosion) in comparison with ancestral protein reconstruction, that may allow in some cases to study nodes close to the last common ancestor of life (LUCA).

During the last 30 years, a significant number of protein systems have been studied using ASR (the 2017 review by Gumulya and Gillam[8] reported around 50 different proteins), including detailed biophysical studies of Precambrian proteins from our research group[9,10]. All this accumulated experience gained has shown resurrected ancestral proteins to be very useful in two different contexts:

1.1.1. Resurrected ancestral proteins are very powerful tools to address important problems in evolution.

There are actually many examples reported in the literature where ancient proteins have been extensively used for understanding relevant issues in evolution and, also, learn more about the environment surrounding Precambrian life[11–14]. In this section, we are going to briefly go through some of these well-known examples to highlight the importance of the use of ancestral protein in evolutionary studies.

Thus, resurrected ancestral forms of the eukaryotic V-ATPase proton pump have been used to address the evolution of complexity in biomolecular machines (Finnigan et al., 2012)[15] and resurrected ancestral rhodopsins have been used to study the molecular origin and adaptive changes in the visual system of mammals (Bickelman et al., 2015)[16].

A remarkable feature of ancestral enzymes is that they often display an increased thermostability compared to their modern counterparts. Reconstructed elongation factors (EF-Tu)[13], nucleotide diphosphate kinases[11], thioredoxins[9] or β-lactamases[10] (studies on these two last proteins were carried out in our research group) have been shown to exhibit an increase in melting temperatures between 30 to 40 °C compared to their modern counterparts. This high stability provides essential information to understand the environment of their extinct hosts and it may reflect the adaptation to environmental changes over geological timescales when targeting billions-of-years-old Precambrian nodes[17,18].

The evolutionary history of uricases is also known thanks to the study of ancestral uricases by the Gaucher´s group[19]. Uricase is the enzyme responsible for the degradation of uric acid in most animals. Humans do not have uricases and can accumulate high levels of uric acid that produces diseases such as gout or kidney stones. The origin of high uric acid levels in humans is linked to the degradation of the uricase gene, due to a mutation that introduced a stop codon. It is surprising that evolution eliminated a useful gene without any evolutionary advantage. In 2014, Gaucher group[19] resurrected and characterized ancestral uricases. They found that ancestral uricases were accumulating mutations and progressively decreasing their activity over time before the completely inactivation of the gene. This reduction occurred at the end of the Oligocene, a period of time where the Earth was cooling and was difficult for our ancestors to find fruit. The loss of uricase may have provided a survival advantage because uric acid enhances the accumulation of fat from the metabolism of fructose, by upregulating some enzymes.

Another outstanding example was reported in 2015 by Benner and coworkers[20]. They actually brought back to life ancestral alcohol dehydrogenases across the hominid phylogeny and found a dramatically enhanced in alcohol degradation by one particular alcohol dehydrogenase (ADH4) dating back at about 10 million years ago. This result correlates well with the time our ancestors increased terrestrialization and, possibly had a greater access to fruit crops with a significant content in alcohol and, therefore, an increased ethanol quantity in the diet.

Insights into evolution of protein structures have been also gained by using a *vertical approach* based on the study of laboratory resurrections of proteins. Detailed structural characterization of various ancestral proteins such as precambrian thioredoxins[21] and beta-lactamases[22] (characterized in our research group), nucleoside disphosphate kinases[11], lysozymes[5], ligand-binding domain of a steroid receptor[23], ligand-binding domain of fish galectins[24], GFP-like proteins[25,26], lactate dehydrogenase[27] or malate dehydrogenase[27] have revealed a remarkable conservation of protein structure over several billion years, even though they display large sequence differences (up to ~50%) compared with their modern counterparts. In fact, extant structures of these proteins are considered as molecular fossils. These studies support the idea that evolution of the folds occurs in short periods of time followed by long periods where the structures do not change[21].

1.1.2. Resurrected ancestral proteins have relevant biomedical and biotechnological applications.

The use of ancestral proteins in biomedicine and biotechnology is actually a recent development[8,28] that relies on the fact that ancestral proteins were adapted to very different intra- and extra-cellular environments compared to their modern counterparts. As a result, resurrected ancestral proteins often display unique or extreme properties that, as shown below, may have great potential in biomedical and biotechnological applications. This is especially remarkable when targeting billions-of-years-old Precambrian nodes as the sequence differences compared to the extant proteins are expected to be greater and, therefore, their properties are likely to be quite different. Indeed, experimental and theoretical results in this regard have pointed out their **high stability**, **substrate and catalytic promiscuity**, **conformational flexibility** and **altered patterns of interaction with other partners *in vivo***[9,10,35–37,11,13,29–34]. All these are desired properties for molecular scaffolds in protein engineering as it is detailed below.

Increased stability seems to be a common outcome of ASR probably due to the thermophilic nature of ancient life. Thus, a large number of ancestral proteins such as lactamases[10], esterases and hydroxynitrile lyases[30], haloalkane dehalogenases[31], lignin-degrading enzymes[32], elongation factors[13], thioredoxins[9], nucleoside diphosphate kinases[11,33] or adenylate kinases[34] have displayed substantial stability enhancements. As an example, the denaturation temperature for the ancestral thioredoxin corresponding to the last common ancestor of the cyanobacterial and deinococcus and thermus groups (LPBCA) has been reported to be of ~123°C[35], this value is many degrees above than the denaturation temperature values of the modern human and *Escherichia coli* thioredoxins (~90°C in both cases)[35]. Interestingly enough, the reported increments in denaturation temperatures for resurrected proteins are typically much greater than those obtained by rational design or directed evolution approaches [38]. The molecular basis of the enhanced stability of ancient proteins might be related with previous results from our group that suggest conservation of site-specific amino acid preferences[39]. In any case, high stability is one of the most critical properties for biotechnological applications of enzymes for a number of reasons: i) low stability may limit practical applications of enzymes; ii) high stability is crucial for evolvability and iii) may improve pharmacokinetics of protein drugs. Actually, ancient uricases are being investigated for the treatment of hyperuricemia because of their convenient properties in terms of pharmacodynamics (enhanced *in vivo* stability) and low immunogenicity[19].

Another interesting example of the biomedical potential of laboratory resurrected proteins is given by coagulation factor VIII. Hemophilia A is a disorder caused by deficiency of functional coagulation factor VIII (FVIII), an essential component of blood coagulation pathways. In the early 1990s this disorder began to be treated using recombinant FVIII, and the hemophilia converted to a manageable disease. But the use of the recombinant FVIII have several limitations, such as poor biosynthetic efficiency, short shelf half-life, and potent immunogenicity. Doerin´s group[40] has used ancestral protein reconstruction to improve the pharmacological properties of FVIII and they obtain engineered coagulation factors VIII with improved activity, stability, biosynthesis potential, and reduced inhibition by anti-drug antibodies.

As mentioned before, promiscuity is another relevant feature commonly displayed by ancestral proteins. Enzyme promiscuity can be defined as the capability of an enzyme to catalyse a reaction other than the reaction for which it has been specialized[41]. Enzyme catalytic promiscuity indicates the ability of an enzyme to catalyze different chemical reactions through different reaction mechanisms whereas substrate promiscuity refers to the ability of enzymes to catalyze the same type of reaction on different substrates[29,42–49]. Although enzymes are very specific catalysts, it is possible that a protein can perform different tasks, although usually with a very low activity level[49,50]. This promiscuity in modern proteins is often considered as a vestige of the generalist nature of ancestral enzymes[43,51,52], as supported by a number or reported examples[30,43,51]. Indeed, enzyme promiscuity is a very desirable feature in protein engineering because promiscuous enzymes are very convenient starting points for the evolution of new catalytic functions or enhancing latent catalytic activities by applying directed evolution approaches[53].

It is widely accepted that promiscuity is related to conformational flexibility[54–56], another relevant feature commonly displayed by ancestral proteins. The native state of the protein should be understood as an ensemble of different conformations and dynamic interconversion between them, favored by specific mutations, may allow achieving different promiscuous activities.

One example of promiscuity associated to conformational flexibility was reported in 2012 by Ozkan´s group[36]. They studied the differences in conformational dynamics of the common ancestor of mineralocorticoid and glucocorticoid receptors (MR and GR). This ancestor (AncCR) was duplicated in the evolution and then diverged in its function to give rise to mineralocorticoid and glucocorticoid receptors. AncCR has been shown to have a promiscuous binding site that allows the binding to aldosterone and cortisol due its conformational flexibility.

Our group has also explored promiscuity in ancestral lactamases[10]. Modern and ancestral β-lactamases have basically the same 3D structure; ancestral β-lactamases degrade efficiently different lactam antibiotics such as benzylpenicillin, cefotaxime and ceftazidime, while the modern TEM-1 lactamase is specialized in the degradation of penicillin, and shows very low activity levels for the degradation of other antibiotics. Conformational dynamics studies show that these functional differences are due to the flexibility displayed by ancestral lactamases in the active site region that is required for the efficient binding of substrates of different sizes and shapes. On the other hand, modern TEM-1 lactamase shows a comparatively rigid active-site region, only efficient for the degradation of penicillin[57].

And finally, we highlight another relevant feature of ancestral proteins that has been investigated just very recently by our group: the *in vivo* altered pattern of interactions with other cellular components. Proteins *in vivo* interact with a large number of molecular partners and these interactions are essential for the majority of biological processes. It is reasonable to think that replacing a modern protein within a modern organism with one of its ancestors may affect many of these interactions. Given this situation, there are two plausible implications. On one hand, organismal fitness will be likely impared but, on the other hand, this replacement might be a strategy to evade pathogens. An illustrative example[37] from our research group will make this point clear. *E. coli* thioredoxin is an essential proviral factor for T7 bacteriophage, a virus that infects *E. coli*. For infection success, the virus must recruit bacterial thioredoxin to be part of its replisome where it specifically interacts with the viral gp5 polymerase. Viruses and their hosts co-evolve so modern viruses are adapted to recruit modern proviral factors. Delgado et al. [37] have replaced thioredoxin within *E. coli* for a number of its Precambrian representations showing acceptable levels of functionality in the cell. However, these ancestral proteins are not recruited by the virus for its replisome so infection cannot proceed. These results suggest a general strategy for engineering virus resistance which might have impact specifically in the engineering of plant virus resistance.

2. GENERAL APPROACH OF THIS DOCTORAL THESIS

The studies included in the present thesis are framed into the two different contexts, pointed out in the preceding section, where resurrected ancestral proteins have been revealed as powerful tools: i) to address important issues in evolution and ii) to provide solutions to practical problems in biotechnological/biomedical scenarios. Our general goal is, therefore, to use ancestral proteins to address, in an evolutionary context, important problems with potential practical implications.

Thus, our experimental studies can be divided into three different sections, described below, that pursue different objectives. Section 2.1. is related to context i) and Sections 2.2. and 2.3. are related to context ii):

2.1. Exploring the relation between protein folding in the test tube (*in vitro*) and protein folding within living organisms (*in vivo*)

Proteins are the most versatile macromolecules in living systems and play a major role in almost all biological processes[58,59]. In most cases, protein function is largely determined by the acquisition of a defined tridimensional structure. How a particular amino acid sequence dictates its biologically relevant structure is known as the protein folding problem[60]. Although significant progress has been made since 1962, when Anfinsen´s experiments[61] demonstrated that proteins can fold spontaneously *in vitro* without any assistance, the protein folding problem still remains a great challenge. In fact, it was listed as one of the 125 most important unresolved problems in science by Science, for its 125 anniversary, back in 2005.

Traditionally, early protein folding studies have been carried out in a test tube according to the general belief that the *in vitro* folding process may actually reflect the molecular events involved during the *in vivo* folding[62,63]. However, and thanks to the recent methodological and technical advances, we know now that folding *in vivo* is a very complex process, quite different than *in vitro.* In order to guarantee that proteins reach their functional tridimensional structures, modern organisms use a complex molecular machinery (including a diversity of chaperones and the ribosome itself) that guide and assist protein folding[64–68]. *In vivo*, the folding occurs co-

translationally, while the protein is being synthesized in the ribosome. Thus, the exit tunnel of the ribosome provides a confined space where the nascent chain begins to fold[69]. Next to the exit tunnel is bound the trigger factor, that is the first molecular chaperone that interacts with nascent chains emerging from the ribosome and guides folding[70]. Molecular chaperones maintain nascent polypeptide chains emerging from the ribosome in a non-aggregated state, shielding exposed hydrophobic amino acid residues that can give rise to aggregation in the highly crowded environment of the cytosol[70].

Having this in mind, it is not clear the evolutionary meaning of the *in vitro* folding studies and, furthermore, it is not clear the relationship between protein folding *in vivo* and *in vitro*. Indeed, the protein folding field needs a re-evaluation of some questions to effectively bridge the folding *in vitro* and *in vivo*. We have addressed this issue in an evolutionary context. Our working hypothesis is based on the idea that *in vitro* folding landscapes for modern proteins may represent moderately degraded versions of the landscapes for the unassisted folding of their ancestors. To experimentally test this hypothesis, systematic folding studies on modern and ancestral thioredoxins have been carried out.

Thioredoxin (Trx) seems to be an appropriate model system to this purpose. Thioredoxins are cytoplasmic oxidoreductases responsible for maintaining redox balance within cells. They are small globular protein of approximately 100 amino acid residues (12-kDa) with the conserved active site (CXXC)[71]. They are ubiquitous in all forms of life, from Archaea to mammals and are involved in a diversity of cellular processes[72–74]. Because of its abundance and relative ease of purification, thioredoxins have been frequently used as model system in numerous protein folding studies[75–79]**.** In addition, a number of ancestral thioredoxins, dating back to a time near the last common ancestor of life (LUCA), have been resurrected in our laboratory[9] and have been used to investigate relevant issues as the evolution of protein structures[21], the adaptation of proteins to climatic, ecological and physiological alterations[9], the design of hyperstable proteins[35] or engineering virus resistance[37]. In summary, our research group has a vast experience in working with modern and ancestral thioredoxins.

In this first part of the doctoral thesis, we have characterized the *in vitro* folding rates of ancestral and modern thioredoxins to study the changes in folding landscapes through evolution. Our results have allowed us to provide an evolutionary interpretation of the *in vitro* protein folding and to better understand the relationship between folding *in vitro* and *in vivo*.

2.2. Using ancestral proteins as a tool for improving heterologous expression

The second part of this thesis is focused on a practical aplication of ancestral reconstruction to a relevant problem in biotechnology: the inefficient expression of proteins of interest in surrogate hosts .

Heterologous expression is required for the production of many proteins of industrial and therapeutic interest. Since 1982, when the first heterologous human insulin produced in *E. coli* (by Gentech, Eli Lilly) was approved for clinical treatment by the FDA[80] a number of many other bioproducts have been overproduced in this expression system, thus becoming a corner stone in the biotech field. Indeed, *E. coli* is the most preferred host to express recombinant proteins due to its rapid growth, high yield of protein, cost-effectiveness, and easy scale-up process[81]. Moreover, heterologous expression is essential in metagenomics. Metagenomics is the study of genetic material recovered directly from environmental samples. It has revealed as a powerful and promising approach to discover novel proteins of industrial or biotechnological interest as it provides access to the enormous amount of genomes from inaccessible and/or uncultured microorganisms[82,83].

However, heterologous expression of proteins is highly problematic[84]. Functional expression of genes in a different host is often limited by various factors like inefficient transcription, translation or/and improper folding and assembly of the corresponding proteins caused by the lack of appropriate chaperones and cofactors[85]. All this results in common troubles such as low expression, formation of insoluble aggregates, degradation by host proteases and so on. Of course, different strategies to overcome these issues are available using, for instance, engineered hosts that have minimized the protease activity or to overexpress molecular chaperones for assisting the protein folding[81,86]. But, despite these advances, heterologous expression is still one of the major bottleneck in the biotech field, causing significant economic losses.

We aimed to investigate this pivotal issue, again in an evolutionary context. Our working hypothesis in this case is based on the idea that inefficient heterologous expression linked to low solubility might be likely due to the absence of co-evolution of the recombinant protein with the natural chaperones of the new host. Rermarkably enough, a number of recent publications have reported the increased yield obtained when overexpressing ancestral proteins as compared with their modern counterparts, including phosphate-binding protein[87], periplasmic binding protein[88]. serum paraoxonase[89], coagulation factor VIII[40], titin[90], haloalkane dehalogenases[31], cytidine and adenine base editors[91], diterpene cyclase[92], rubisco[93],

endoglucanases[94], L-amino acid oxidases[95], laccases[96], front-end Δ-6-desaturases[97] and fatty acid photo-decarboxylases[98]. These results, in our opinion, might be reflecting the fact that ancestral folding was unassisted (according to our studies in Section 5.1.) and that clues for efficient heterologous expression might be encoded in the respective ancestral sequences (note that the preparation of ancestral proteins in modern hosts is necessarily heterologous).

To experimentally test our hypothesis we have used the thioredoxin from *Candidatus Photodesmus katoptron (CPk)*, an obligate symbiont, lacking the neccesary genes for amino acid synthesis and for metabolism of energy sources like glucose[99], with a high evolutionary rate of its genome[99]. Actually, *CPk* thioredoxin is an excellent model system for a number of reasons: (i) it is a protein from an unculturable bacteria, so its preparation mimics the same scenario of metagenomic studies; (ii) according to our experimental results, it displays a very slow *in vitro* refolding rate, spending around 8 hours to reach the native state; (iii) its heterologous expression in *E. coli* at 37 ºC leads mostly to insoluble protein (≈ 90% of insoluble protein) and (iv) it has radically opposed folding features compared to ancestral thioredoxins. Thus, ancestral thioredoxins have an efficient folding *in vitro* and an efficient heterologous expression, so a direct comparison can be established with the *CPk* thioredoxin and relevant conclusions can be extracted.

In this part of the thesis we aimed to ascertain the molecular determinants of the ancestral folding efficiency and explore whether recue of inefficient heterologous folding can be engineered on the basis of a few selected back-to-ancestor mutations. In this regard, we have collaborated with Professor Athi N. Naganathan (Indian Institute of Technology Madras, Chennai, India). His computational predictions, based on a statistical-mechanical model of the folding landscape to determine the regions of protein that are likely to be unfolded in aggregation-prone intermediate states, have been crucial to target specific regions in *CPk* thioredoxin responsible for the impared folding. As a result, a number of single-mutant variants and chimeras have been extensively characterized in terms of *in vitro* kinetics and *in vivo* solubility measurements. Our results have shown that it is possible to rescue inefficient heterologous expression with only 1-2 back-to-ancestor mutations.

2.3. Ancestral glycosidases as molecular scaffolds for protein engineering

The last part of this doctoral thesis is also focused on a practical application of ancestral proteins to a relevant problem in biotechnology: engineering *de novo* activities.

Current technological applications of proteins are limited to their natural activities (or closely related activities). Indeed, modern proteins are the result of a long-term evolution experiment that has been running for over 4 billion-years resulting in highly specialized catalysts. The development of general procedures to obtain tailored-made enzymes capable of efficiently catalyze non-natural chemical reactions still remains as one of the most relevant problems one can face in protein engineering. Thus, the availability of *de novo* enzymes would provide sustainable alternatives to many industrial processes and would have a huge impact in biotechnology, biomedicine and industry.

We hypothesize that the rather limited success of most previous attempts to generate efficient *de novo* enzymes is due to: i) the use of modern proteins as scaffolds, highly specialized to perform very specific tasks contrary to their ancestors, likely promiscuous generalists; and ii) the limitations of the conventional directed evolution experimental approaches that either explore only a small fraction of protein sequence space or cannot select for high activity levels[100]. New methodologies, such as RNA display, have facilitated the generation of large libraries of protein variants that explore a larger sequence space. In fact, Seelig and Szostak[101] were able to find a *de novo* RNA ligase in a library of approximately 10^{13} variants prepared from a non-catalytic scaffold, suggesting that for the design of new activities is not essential a low level *de novo* activity. However, the levels of catalysis they obtained were very modest. The problem here was that the ultra-high throughput screening approaches could not be easily combined with an appropriate selection for high catalytic activity.

Accordingly, we propose the use of resurrected ancestral proteins that, as detailed in the Introduction, display properties (enhanced stability, catalytic promiscuity and conformational flexibility) that contribute to high evolvability and are also very convenient in molecular scaffolds for protein engineering. In fact, a previous study from our group has provided experimental evidence of the use of ancestral proteins for the generation of new activities. Ancestral β-lactamases were resurrected in our laboratory and their biophysical features were exhaustively studied. They displayed both, hyperstability and promiscuity linked to enhanced conformational flexibility[10]. The capability of ancestral lactamases to generate new activities was probed and promising results were obtained. With only a mutation in the

ancestral lactamase, a new and non-natural activity, the Kemp elimination activity, was generated[22].

To experimentally test our hypothesis we have selected as a model system a TIM barrel protein. The TIM barrel fold is the most common protein fold (it actually represents about 10% of known domain structures) and it is found in many different enzyme families, catalyzing completely unrelated reactions[102]. Finding the proper family for ancestral reconstruction was actually a major task. To this end an extensive study of the TIM-barrel superfamilies in the CATH database[103] has been carried out. The TIM-barrel superfamily that best fit our requirements was the glycosidase superfamily.

Glycosidases are well-known and extensively characterized TIM-barrel proteins. They are responsible for the hydrolysis of glycosidic linkages in a wide diversity of molecules[104]. The glycosidic bond is one of the strongest bonds in nature[105] and glycosidases can break them (with an estimated half-life of about 5 million years) in a few milliseconds[106]. They are present in almost all living organisms where they play diverse and different roles[107]. They have been extensively studied, because of its interest in industrial and biotechnological applications[108].

Several families of glycosidases were reconstructed in collaboration with Dr. Eric Gaucher (School of Biology, Georgia State University). They were resurrected in our laboratory and initial results (high yield protein expression and high stability) led us to select an ancestral glycosidase, called N72 from family 1 glycosidases (GH1), for exhaustive biochemical and biophysical characterization.

The ancestral glycosidase selected is the phylogenetic node corresponding to the ancestor of modern bacterial and eukaryotic enzymes. Our results support the unusual properties of this ancestral glycosidase, including enhanced conformational flexibility, optimum activity temperature within the range of family 1 glycosidases from thermophilic organisms and, surprisingly enough, specific heme binding with a concomitant allosteric modulation of catalysis. All these features open new and exciting opportunities for using this molecular scaffold for custom catalysis.

This study is actually part of a collaborative project led by Dr. Sanchez-Ruiz (Universidad de Granada) and funded by the prestigious Human Frontier Science Program (HFSP). The other research groups involved are led by: Dr. Gaucher (Georgia State University), Dr. Kamerlin (Uppsala University) and Dr. Seelig (University of Minnesota).

The results have been published in Nature Communications and are included in the Appendix of the present doctoral thesis as my personal contribution has been relevant (note my co-first authorship). However it is NOT part of the compendium of publications that the present thesis conforms to.

3. OBJECTIVES

The main objectives of this doctoral thesis are:

1. Provide an evolutionary interpretation of the relationship between the *in vitro* and *in vivo* protein folding. This objective includes two specific goals:
    a. An exhaustive experimental characterization of the *in vitro* folding rates of modern and Precambrian thioredoxins dating back up to ~4 billion years ago.
    b. An extensive mutational analysis on extant and ancestral thioredoxins to ascertain the molecular determinants for ancestral folding efficiency.
2. Propose a general strategy for the rescue of inefficient heterologous expression based on a few back-to-the-ancestor mutations. For achieving this final objective is necessary to address two specific goals:
    a. Investigate the relationship between efficiency of heterologous expression and *in vitro* folding rate. To this end, again a set of modern and ancestral thioredoxins has been used as a model system.
    b. Computational calculations to guide the search for particular regions of the protein likely to be unfolded in aggregation-prone intermediate states.
    c. Prepare in the laboratory different variants/chimeras of the thioredoxins of interest to experimentally test the theoretical predictions on the regions/positions relevant for efficient heterologous expression
3. Searching for promising scaffolds for the generation of *de novo* enzymes with non-natural activities using ancestral TIM-barrel proteins. This objective includes as specific goals:
    a. Exhaustively study of TIM barrel families in CATH database to select an adequate family for the reconstruction of ancestral proteins.
    b. Reconstruction and resurrection of ancestral glycosidases.
    c. Biophysical and biochemical characterization of ancestral glycosidases with unexpected properties relevant for protein engineering.

4. RESULTS

In this chapter, we present the main results obtained in this doctoral thesis. They are classified in different sections according to the proposed objectives.

OBJETIVE 1

Publication 1

Candel, A.M., Romero-Romero, M.L., Gamiz-Arco, G., Ibarra-Molero, B., and Sanchez-Ruiz, J.M. (2017). Fast folding and slow unfolding of a resurrected Precambrian protein. Proc. Natl. Acad. Sci. 114, E4122–E4123. doi: 10.1073/pnas.1703227114

Impact Factor (2017): 9.504

Category name /Quartile in Category: Multidisciplinary Sciences / Q1

Publication 2

Gamiz-Arco, G., Risso, V.A., Candel, A.M., Inglés-Prieto, A., Romero-Romero, M.L., Gaucher, E.A., Gavira, J.A., Ibarra-Molero, B., and Sanchez-Ruiz, J.M. (2019). Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding. Biochem. J. 476, 3631–3647. doi: 10.1042/BCJ20190739

Impact Factor (2019): 4.097

Category name /Quartile in Category: Biochemistry & Molecular Biology / Q2

OBJETIVE 2

Publication 3

Gamiz-Arco, G., Risso, V.A., Gaucher, E.A., Gavira, J.A., Naganathan, A.N., Ibarra-Molero, B., and Sanchez-Ruiz, J.M. (2021). Combining Ancestral Reconstruction with Folding-Landscape Simulations to Engineer Heterologous Protein Expression. J. Mol. Biol. 433, 167321. doi: 10.1016/j.jmb.2021.167321

Impact Factor (2020): 5.469

Category name /Quartile in Category: Biochemistry & Molecular Biology / Q2

PUBLICATION 1

Fast folding and slow unfolding of a resurrected Precambrian protein

Fast folding and slow unfolding of a resurrected Precambrian protein

Adela M. Candel[a], M. Luisa Romero-Romero[a,1], Gloria Gamiz-Arco[a], Beatriz Ibarra-Molero[a], and Jose M. Sanchez-Ruiz[a,2]

Tzul et al. (1) report different unfolding rates and similar folding rates for a number of thioredoxins. The authors interpret this result as evidence of the principle of minimal frustration. Their study includes several resurrected Precambrian thioredoxins that we have previously prepared and characterized (2–5).

We agree that the principle of minimal frustration is essential to understand protein evolution. However, approximate folding-rate invariance is easily explained without invoking this principle. Thioredoxin kinetic stability relies on a transition state that is substantially unstructured (5, 6). Therefore, mutations that changed unfolding rates to tune kinetic stability during evolution likely had much less effect on folding rates, as implied by the well-known principles of ϕ-value analysis (7).

Moreover, our experimental results are not consistent with folding-rate invariance being a general feature of thioredoxins. Fig. 1 shows folding–unfolding rates for the modern *Escherichia coli* thioredoxin and a resurrected Precambrian thioredoxin. The unfolding of the ancestral protein is ∼three orders-of-magnitude slower than the unfolding of the modern protein, indicating enhanced kinetic stability. However, in clear disagreement with the results reported in Tzul et al. (1) for the very same proteins, we find the following two features. (*i*) There are deviations from linearity (rollovers) in the folding branches at low denaturant concentrations. "Rollovers" are commonly attributed to proline isomerization and the presence of significantly populated intermediate states (8). Such kinetic complexities reflect ruggedness of the folding landscape and have been previously reported to occur for thioredoxins (9). (*ii*) There is a faster folding rate for the ancestral protein. The ancestral versus modern folding-rate enhancement extrapolates to ∼two orders-of-magnitude at low denaturant concentrations and it is linked to a higher slope of the folding branch (the folding *m* value). This finding suggests a role for the residual structure of the unfolded state, because the unfolding slope is unchanged. Substantial folding-rate enhancement is also observed in the rollover region. The intriguing possibility arises that the higher folding rate for the ancestral protein is an adaptation to inefficient Precambrian folding chaperones.

Why are these two features not apparent in the data of Tzul et al. (1)? We obtained our kinetic data at pH 7 using guanidinium hydrochloride, whereas pH 2 was used in Tzul et al. Such an acidic pH is destabilizing, thus allowing a weaker denaturant (urea) to be used and bringing the unfolding rates to the stopped-flow (milliseconds) time scale. However, destabilizing conditions may buffer the experimental consequences of ruggedness in folding landscapes. In fact, it is well known that kinetic intermediates are more readily observed under strongly stabilizing conditions (8). Furthermore, extensive protonation of residues at acidic pH may distort the folding landscapes in a manner that is not physiologically and evolutionary relevant. Note that comparatively few organisms are acidophiles and that many acidophiles actually pump out protons to keep a neutral intracellular medium.

We conclude that the experiments at pH 2 reported by Tzul et al. (1) miss important features of ancestral thioredoxin folding and do not provide any clear evidence for the principle of minimal frustration in the evolution of folding landscapes.

Acknowledgments

Work in the authors' laboratory is supported by European Fund of Local Development Funds and Grants BIO2015-66426-R from Ministry of Economy and Competitiveness/European Fund of Local Development (to J.M.S.-R.) and P09-CVI-5073 from the "Junta de Andalucía" (to B.I.-M.).

[a]Departamento de Quimica Fisica, Facultad de Ciencias, Universidad de Granada, Granada 18071, Spain

Author contributions: J.M.S.-R. designed research; A.M.C. and G.G.-A. performed research; M.L.R.-R. and B.I.-M. contributed new reagents/analytic tools; A.M.C. and B.I.-M. analyzed data; and J.M.S.-R. wrote the paper.

The authors declare no conflict of interest.

[1]Present address: Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel.

[2]To whom correspondence should be addressed. Email: sanchezr@ugr.es.

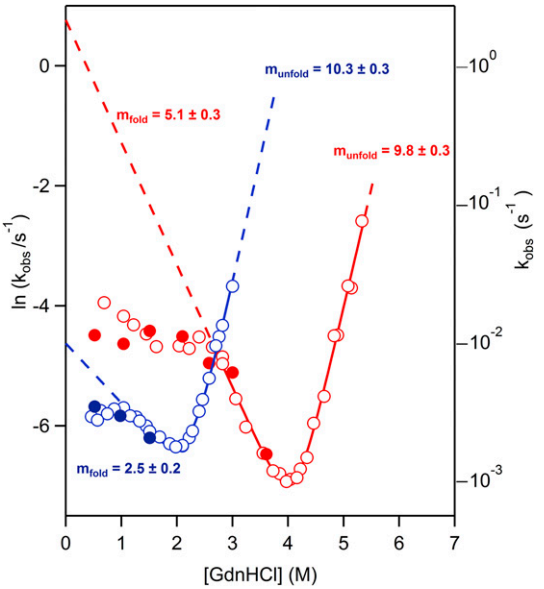


Fig. 1. Folding–unfolding rates for the modern *E. coli* thioredoxin (blue) and the resurrected thioredoxin corresponding to the last common ancestor of the cyanobacterial, *Deinococcus* and *Thermus* groups (LPBCA thioredoxin; red). The proteins were purified, as we have previously described (2–5). Rate-constant values were determined at pH 7 (Hepes buffer 50 mM) from the time-dependence of the protein fluorescence (open datapoints), following protocols we have previously described and used for thioredoxins (5, 6). We also followed kinetics using interrupted refolding with double-jump unfolding assays (closed datapoints), a methodology that we have previously described (10) and that provides a direct determination of amount of native protein. The protein concentrations used in interrupted refolding experiments are ∼10-fold higher than those used in fluorescence kinetic experiments. The agreement between the results obtained with the two methodologies therefore supports that our data are not distorted by protein association. Multiexponential kinetic profiles were observed at denaturant concentrations corresponding to the rollover regions and only data corresponding to the slower phase (i.e., the phase leading to the native state) are shown in these cases. Monoexponential kinetic profiles were observed at denaturant concentrations outside the rollover regions. Rate constants derived from these monoexponential profiles were fitted using a two-state kinetic model. These fits are shown with continuous lines, whereas dashed lines represent their extrapolation outside the experimental range of monoexponential kinetics. Despite the uncertainties involved in such extrapolations, it is clear that the ancestral versus modern rate enhancement for the folding from the unfolded state approaches ∼two orders-of-magnitude under native conditions. Values of the kinetic denaturant *m* values derived from the fits are given in kJ·mol^{-1}·M^{-1} alongside the corresponding folding and unfolding branches. GdnHCl, guanidinium hydrochloride.

1 Tzul FO, Vasilchuck D, Makhatadze GI (2017) Evidence for the principle of minimal frustration in the evolution of protein folding landscapes. *Proc Natl Acad Sci USA* 114:E1627–E1632.
2 Perez-Jimenez R, et al. (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol* 18:592–596.
3 Ingles-Prieto A, et al. (2013) Conservation of protein structure over four billion years. *Structure* 21:1690–1697.
4 Risso VA, et al. (2015) Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol Biol Evol* 32:440–455.
5 Romero-Romero ML, et al. (2016) Selection for protein kinetic stability connects denaturation temperatures to organismal temperatures and provides clues to Archaean life. *PLoS One* 11:e0156657.
6 Godoy-Ruiz R, et al. (2006) Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments. *J Mol Biol* 362:966–978.
7 Fersht AR, Matouschek A, Serrano L (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* 224:771–782.
8 Matouschek A, Kellis JT, Jr, Serrano L, Bycroft M, Fersht AR (1990) Transient folding intermediates characterized by protein engineering. *Nature* 346:440–445.
9 Georgescu RE, Li JH, Goldberg ME, Tasayco ML, Chaffotte AF (1998) Proline isomerization-independent accumulation of an early intermediate and heterogeneity of the folding pathways of a mixed α/β protein, *Escherichia coli* thioredoxin. *Biochemistry* 37:10286–10297.
10 Ibarra-Molero B, Sanchez-Ruiz JM (1997) Are there equilibrium intermediate states in the urea-induced unfolding of hen egg-white lysozyme? *Biochemistry* 36:9616–9624.

PUBLICATION 2

Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

Research Article

Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding

Gloria Gamiz-Arco[1], Valeria A. Risso[1], Adela M. Candel[1,*], Alvaro Inglés-Prieto[1,†], Maria L. Romero-Romero[1,‡], Eric A. Gaucher[2], Jose A. Gavira[3], Beatriz Ibarra-Molero[1] and Jose M. Sanchez-Ruiz[1]

[1]Departamento de Quimica Fisica, Facultad de Ciencias, Unidad de Excelencia de Química Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain; [2]Department of Biology, Georgia State University, Atlanta, GA 30303, U.S.A; [3]Laboratorio de Estudios Cristalográficos, Instituto Andaluz de Ciencias de la Tierra, CSIC-University of Granada, Avenida de las Palmeras 4, 18100 Armilla, Granada, Spain

Correspondence: Beatriz Ibarra-Molero (beatriz@ugr.es) or Jose M. Sanchez-Ruiz (sanchezr@ugr.es)

Evolution involves not only adaptation, but also the degradation of superfluous features. Many examples of degradation at the morphological level are known (vestigial organs, for instance). However, the impact of degradation on molecular evolution has been rarely addressed. Thioredoxins serve as general oxidoreductases in all cells. Here, we report extensive mutational analyses on the folding of modern and resurrected ancestral bacterial thioredoxins. Contrary to claims from recent literature, *in vitro* folding rates in the thioredoxin family are not evolutionarily conserved, but span at least a ~100-fold range. Furthermore, modern thioredoxin folding is often substantially slower than ancestral thioredoxin folding. Unassisted folding, as probed *in vitro*, thus emerges as an ancestral vestigial feature that underwent degradation, plausibly upon the evolutionary emergence of efficient cellular folding assistance. More generally, our results provide evidence that degradation of ancestral features shapes, not only morphological evolution, but also the evolution of individual proteins.

*Present address: Eurofins Villapharma Research. Parque Tecnológico de Fuente Álamo. Crta. El Estrecho – Lobosillo km 2.5 30320 Fuente Álamo, Murcia, Spain
†Present address: Research Center for Molecular Medicine of the Austrian Academy of Sciences. Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria
‡Present address: Max Planck Institute for Molecular Cell Biology and Genetics, Dresden 01307, Germany

Received: 4 October 2019
Revised: 19 November 2019
Accepted: 21 November 2019

Accepted Manuscript online: 21 November 2019
Version of Record published: 10 December 2019

Introduction

Instances of so-called imperfect (poor or suboptimal) 'design' have been extensively studied in records of evolutionary history, and have served as evidence that living organisms, rather than being designed, are the products of complex evolutionary forces and histories [1,2]. Glaringly questionable 'design', such as the recurrent laryngeal nerve in mammals, thus suggests that evolutionary tinkering with previously functional features can limit the possible outcomes of new functions in the future. Still, many examples of imperfect morphological 'design' are simply related to the evolutionary degradation of ancestral features that are no longer useful. Examples are abundant and include human's limited capability to move the ears (linked to the degradation of barely used muscles) as well as the presence of a tailbone, vestigial leg bones in whales and vestigial wings in flightless birds. Evolutionary degradation is primarily a consequence of the inability of natural selection to purge mutations that impair a feature, once the feature has ceased to be useful (i.e. once it has ceased to confer a functional selective advantage). Darwin did realize that 'rudiments' are evidence of descent from ancestral forms and discussed many examples in the first chapter of *The Descent of Man* (1871). More recently, the discovery that genomes include large numbers of pseudogenes has provided a clear example of evolutionary degradation at the molecular level. Thus, reversing the mutations that originally led to the silencing of a given gene does not typically restore the function of the encoded protein [3] because, after a gene is silenced, it is no longer subject to purifying natural selection and quickly accumulates other degrading mutations.

PORTLAND PRESS

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

Plausibly, the evolutionary degradation of useless ancestral features is widespread, not only in morphological evolution, but also during the course of molecular evolution. Other than pseudogenes [3], however, molecular examples appear to have been rarely discussed in the literature, if at all. We argue here that evolutionary analysis of protein folding processes may provide clear examples of evolutionary degradation at the molecular level. This is so because folding *in vivo* within modern organisms [4–9] is protected and assisted by a complex folding-assistance machinery, including chaperones and the chaperone functionality of the ribosome, while, on the other hand, folding studies in the test tube (*in vitro* folding) probe unassisted folding. Since evolution has no foresight, however, folding assistance cannot have arisen before protein folding itself, inasmuch as the components of the folding-assistance machinery are proteins themselves that need to be folded to be functional. It follows that the most ancient proteins could plausibly fold with little or no assistance. Therefore, unassisted folding may have been relevant at a primordial stage, prior to (or concomitantly with) the emergence of folding assistance.

We propose, therefore, that the *in vitro* folding process for modern proteins of ancient evolutionary origin may bear signatures of evolutionary degradation. Thioredoxins, general oxidoreductases that display a wide substrate scope and that are involved in a diversity of cellular processes [10,11], should provide an excellent model system to explore this possibility. They are present in all known cells (eukaryotes, bacteria and archaea) and it is thus plausible that they existed at a very early stage, even perhaps preceding the emergence of an efficient folding assistance. Indeed, thioredoxins can fold without assistance in the test tube. It has been known for many years [12], however, that thioredoxins have a 'folding problem' related to the presence of a proline residue in *cis* conformation at position 76 (we use *Escherichia coli* thioredoxin numbering throughout). *Cis*-prolines in native protein structures create folding kinetic bottlenecks [13–15], since isomerization is slow, the *trans* conformer is favoured in unfolded polypeptide chains and may become trapped in intermediate states in the folding landscape thus further slowing down folding. For thioredoxins, mutational escape from the problem is not possible, since position 76 is close to the catalytic disulfide bridge and the presence of a proline at that position is required for a fully functional active-site conformation [12]. Pro76 is thus strictly conserved in thioredoxins.

Here, we first study the folding *in vitro* of *E. coli* thioredoxin and two of its resurrected Precambrian ancestors. An extensive mutational analysis allows us to explain the slower folding of the modern protein in terms of a single amino acid replacement that aggravated the folding problem created by the *cis*-proline at the active site. Furthermore, the identified replacement points to a region of the thioredoxin molecule where mutations can be reasonably expected to impact the folding rate. Experimental analysis of a set of modern bacterial thioredoxins selected to represent natural sequence diversity in this region shows that, contrary to what it has been claimed in recent literature [16,17], *in vitro* folding rates are not evolutionarily conserved. In fact, *in vitro* folding for some of the studied modern thioredoxins occurs in the approximately hour time scale and is between 1 and 2 orders of magnitude slower than both the inferred ancestral folding and the folding of other modern thioredoxins. These results suggest an interpretation of *in vitro* folding as a degraded version of primordial unassisted folding. More generally, our results provide evidence that degradation shapes evolution not only at the morphological level but also at the level of individual enzymes.

Materials and methods

Protein expression and purification

E. coli, LBCA and LPBCA thioredoxins as its variants studied in this work were prepared without purification tags following procedures we have previously described in detail [18–20]. Proteins representing bacterial thioredoxins (Figure 6) were prepared with His-tags using affinity chromatography. Mutations were introduced using the QuikChange Lighting Site-Directed Mutagenesis kit (Agilent Technologies) and checked by DNA sequencing.

Protein solutions were prepared by exhaustive dialysis at 4°C against 50 mM Hepes (pH 7). Protein concentrations were determined spectrophotometrically using known values for the extinction coefficients. Solutions of guanidine in 50 mM Hepes (pH 7) were prepared as previously described [18–20]. Prior to use, urea was purified by ion-exchange chromatography as previously described [21]. Guanidine and urea concentrations were determined by refractometry.

Activity determinations

Most reported activity determinations are based on the insulin turbidimetric assay [22], as described previously [18]. Briefly, thioredoxin catalysis of the reduction in insulin by DTT is determined by following the aggregation of

the β-chain of insulin. An aliquot of a thioredoxin solution is added to 0.5 mg/ml of bovine pancreatic insulin and 1 mM DTT at pH 6.5 and the rate is calculated from the slope of a plot of absorbance versus time at the inflexion point (see Supplementary Figure S1 for illustrative examples). Values given (Figure 3c and Supplementary Table S1) are the average of at least three independent measurements.

For some selected variants, we also assayed thioredoxin activity with thioredoxin reductase coupled to the reduction in Ellman's reagent (5,5′-dithiobis(2-nitrobenzoic acid) or DTNB) at pH 8, essentially as described by Slaby and Holmgren [23]. Concentrations used were 0.02 μM for the reductase, 0.5 mM DTNB and 0.25 mM NADPH. Thioredoxin concentrations were typically in the range 0.15–0.20 μM. Values reported are the average of at least three independent determinations.

Unfolding and folding kinetics studied by steady-state fluorescence measurements

Kinetic data for non-mutated *E. coli* thioredoxin and LPBCA thioredoxin given in Figures 3, 4 and Supplementary Figure S10 are taken from [20]. All other kinetic data shown were obtained in this work. All experiments were performed at 25°C. Folding–unfolding kinetics were studied using procedures we have previously described in detail [20,24]. Briefly, we measured the time-dependence of the fluorescence emission at 350 nm with excitation at 276 nm, after suitable guanidine- or urea-concentration jumps. For experiments in guanidine solutions, we typically used 20-fold dilution from ∼4 to 5 M guanidine or from zero guanidine concentration for experiments carried at denaturant concentrations approximately above or below the denaturation midpoint. For experiments in urea solutions, we typically used 20-fold dilution from ∼10 M urea or from zero urea concentration for experiments carried at denaturant concentrations approximately above or below the denaturation midpoint. The ancestral LBCA and LPBCA thioredoxins are highly stable and are not fully denatured despite high concentrations of urea. Folding rates for these proteins in urea solutions (Figure 6) were obtained by first denaturing them in concentrated guanidine followed by a high dilution into urea solutions, in such a way that the final guanidine concentration was very low (∼0.1 M). Typically, the protein concentration in the fluorescence kinetic experiments was on the order of 0.05 mg/ml.

Unfolding kinetics could be adequately fitted with a single exponential equation from which the rate constant could be easily calculated (see Supplementary Figure S2 for representative examples). Many folding profiles could also be well described by a single exponential within the time range of the manual mixing experiments. However, two exponential terms were required to achieve good fits in many other cases (see Supplementary Figure S3 for representative examples).

Finally, the long-time, equilibrium fluorescence values derived from the analyses of kinetic profiles were used to assess the thermodynamic stability of the modern thioredoxins studied in this work. Profiles of equilibrium fluorescence intensity versus urea concentration were fitted assuming a linear dependence of the unfolding free energy with denaturant concentration [25] within the narrow transition range and using linear pre- and post-transition baselines. Values of the urea midpoint concentration ($C_{1/2}$, the urea concentration at which the unfolding free energy is zero) and the denaturant-concentration dependence of the unfolding free energy ($m = -\mathrm{d}\Delta G/\mathrm{d}[\mathrm{urea}]$) were obtained from the fits. These values are collected in Supplementary Table S4, where additional details of the fitting process are provided.

Using double-jump unfolding assays to determine the relevant kinetic phase of the major folding channel

Unlike unfolding, which often occurs in a single kinetic phase, protein folding is typically a complex process involving several parallel kinetic channels leading to the native state, as well as the transient population of intermediate states in many of these channels [26]. *In vitro* folding of thioredoxin is certainly known to conform to this scenario [27]. In this work, the complexity of *in vitro* thioredoxin folding is revealed by the multi-exponential folding profiles found in some cases (Supplementary Figure S3) and by clear rollovers in the folding branches of all the Chevron plots reported (Figures 3, 4, 6 and Supplementary Figure S10). In general, the folding rate of any given protein (i.e. the rate that defines the time scale of the protein folding process) could be defined in terms of the main slow phase of an experimental folding kinetic profile obtained using a suitable physical property. Still, it is absolutely essential to ascertain that this phase does indeed reflect the major kinetic channel that leads to the native protein. It would be conceivable, for instance, that most of the protein arrived to the native state in a slower phase that does not bring about a significant change in the

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

physical property being measured (steady-state fluorescence, in our case) and which is, therefore, not detected. Also, it would be conceivable that most of the protein arrives to the native state in a fast phase and that the slower phase detected in the kinetic folding profiles reflects a minor structural re-arrangement of the native ensemble or, alternatively, the folding of a small fraction of the protein from a kinetically trapped intermediate state. Thus, even if a single exponential phase is detected by the physical property used, there is the possibility that folding actually occurred during the dead time of the kinetic experiment. Furthermore, a very slow phase of small amplitude could just reflect instrumental drift. These and other interpretation uncertainties plagued the *in vitro* protein folding field since its beginnings. However, pioneers of the field found reliable ways around these problems on the basis of carefully designed 'jump assays' in which protein samples are extracted at certain times and transferred to solutions of selected composition for experimental assessment (see, for instance [13,14]). Here, we have specifically used double-jump unfolding assays, a methodology that aims at providing a direct determination of the amount of native protein [28,29]. The rationale behind this approach is that the unfolding of the native state of a protein is much slower than the unfolding of non-native or intermediate states. The amount of native state in a protein solution can then be determined from the unfolding kinetics followed in the appropriate time scale after transfer to denaturing conditions. Obviously, unfolding assays exploit the high activation free energy barrier for unfolding to determine the amount of native protein, i.e. they exploit the free energy barrier that confers kinetic stability to the native protein [30,31]. They are, therefore, particularly appropriate for this work because following folding kinetics using double-jump unfolding assays does define the time scale required for the development of kinetic stabilization. That is, they define the time span in which the unassisted folding chain is susceptible to undesirable interactions and alterations, which is a parameter of direct evolutionary significance.

For most the thioredoxin variants studied here, we have followed the folding kinetics under selected conditions by carrying out unfolding assays at different times after transfer of a denatured protein to native conditions, in such a way that folding kinetic profiles of the amount of native state versus time are obtained. In a typical experiment (see Supplementary Figure S4 for a representative example), we used a concentrated solution of unfolded protein in ∼4 M guanidine (*E. coli* thioredoxin and its variants) or ∼5 M guanidine (LBCA thioredoxin, LPBCA thioredoxin and their variants) and we started the folding process by a suitable dilution (within the 2–10-fold range) into a low-concentration guanidine solution to reach a final protein concentration on the order of 1 mg/ml. At given times, aliquots were extracted and transferred (20-fold dilution) to ∼3 M guanidine for *E. coli* thioredoxin and its variants or to ∼5 M guanidine for the ancestral thioredoxins and their variants, and the unfolding kinetics were determined by fluorescence. The fraction of native state (X_N) versus time (t) profile for folding at low denaturant concentration is easily obtained from the amplitude of the unfolding kinetic phase using a suitable control experiment (Supplementary Figure S4). Supplementary Figure S5 shows several representative examples of X_N versus t profiles that illustrate the strong effect of the S/G exchange at position 74 on the folding rate. In all cases, these profiles could be well described by single exponentials, with initial time and long-time values close to zero and unity. This indicates that these profiles probe the relevant kinetic phase of the major folding channel. Of course, it cannot be ruled out that, in several cases, small amounts of protein reach the native state through faster or slower channels, since the initial time and long-time values of the profiles actually differ somewhat from zero and unity (see also Figure 6b).

Comparison of the folding profiles from double-jump assays (X_N versus t) with those obtained using steady-state fluorescence revealed three different scenarios: (1) fluorescence profiles could be well fitted by a single exponential and the rate constant derived from such fits agreed with the value obtained from the X_N versus t profiles (see Supplementary Figure S6 for an illustrative example). (2) Two exponentials were required to fit the fluorescence folding profiles and it was the rate constant from the faster, larger amplitude phase that agreed with the value obtained from the X_N versus t profiles (see Supplementary Figure S7 for an illustrative example). (3) Two exponentials of roughly similar amplitude were required to fit the fluorescence folding profiles and it was the rate constant from the slower phase that agreed with the value obtained from the X_N versus t profiles (see Supplementary Figure S8 for an illustrative example).

Finally, it is important to note that following folding kinetics through double-jump unfolding assays is considerably time consuming. Therefore, our approach has been to carry out extensive folding kinetic studies on the basis of steady-state fluorescence measurements and to determine only a limited number of double-jump X_N versus t profiles in order to identify in the fluorescence profiles the relevant kinetic phase of the major folding channel. For the sake of simplicity and clarity, folding branches of the chevron plots given in Figures 3–5 and Supplementary Figure S10 show only the rate constants for such relevant kinetic phase and do

not differentiate between data derived from fluorescence profiles and data derived from double-jump X_N versus t profiles. However, in Supplementary Figure S9 we provide Chevron plots that include the comparison between rate constants values derived from fluorescence profiles and from double-jump X_N versus t profiles. Also, Figure 6b shows profiles of folding followed by double-jump unfolding assays for several modern bacterial thioredoxins.

Results and discussion

Modern versus ancestral thioredoxin folding

We first compared the folding of modern *E. coli* thioredoxin with that of two of its resurrected Precambrian ancestors (Figure 1a): the thioredoxins encoded by the reconstructed sequences for the last bacterial common ancestor (LBCA thioredoxin) and the last common ancestor of the cyanobacterial, deinococcus and thermus groups (LPBCA thioredoxin). These two phylogenetic nodes correspond to organisms that existed ∼4 and 2.5 billion years ago, respectively [18,32,33]. We have previously characterized LPBCA and LBCA thioredoxins, as well as several other resurrected Precambrian thioredoxins, in detail [18–20,33–35]. They are properly folded, highly stable, active enzymes that share essentially an identical three dimensional (3D)-structure with *E. coli* thioredoxin (Figure 2), despite their low sequence identity with the modern protein (for sequences and structures, see Figure 1 in [33]). The structures of the three proteins under study bear a proline residue at position 76 in *cis* conformation that is strictly conserved in thioredoxins.

Figure 3 shows chevron plots of rate constant versus denaturant concentration for *E. coli* thioredoxin and the ancestral LPBCA and LBCA thioredoxins. These plots include folding and unfolding branches. We have used guanidine, a strong denaturant, for the experiments in Figure 3 in order to achieve denaturation of the highly stable ancestral thioredoxins. However, urea, a weaker denaturant, is used in other experiments reported in this work but we show that the choice of denaturant does not affect our conclusions. We specifically define folding rates in terms of the relevant kinetic phase of the major folding channel as identified by double-jump unfolding assays (see Materials and methods for details).

As it is clear from Figure 3a, the *in vitro* folding of the ancestral thioredoxins is substantially faster than the folding of their modern *E. coli* counterpart. Rate constants reported in this figure have been determined at 25° C for both the modern and ancestral proteins, while high environmental temperatures (∼ 65–80°C) have been suggested for the Archaean [36]. Experimental determinations of the temperature dependence of the rate constants (Supplementary Figure S11) show that consideration of the different living temperatures of the modern and ancient hosts does not eliminate the folding rate difference. On the contrary, the folding of the ancestral LPBCA and LBCA thioredoxins at the proposed Archaean temperatures is actually around two orders of magnitude faster than the folding of *E. coli* thioredoxin at the optimal living temperature (37°C) of *E. coli* (Supplementary Figure S11).

For the three proteins studied, the folding rate is determined by the presence of a *cis*-proline at position 76, as shown by the observation that the folding rate is considerably increased by mutating proline 76 to alanine (compare Figure 3a,b). Of course, replacing proline at position 76 impairs activity [12], as shown by assays based on the aggregation of insulin (Figure 3c and Supplementary Table S1) as well as assays based on the interaction of thioredoxin with thioredoxin reductase (Supplementary Table S2). Interestingly, in the latter assays, the activity of variants of the most ancient LBCA thioredoxin appear depressed (Supplementary Table S2), possibly reflecting the consequences of coevolution between thioredoxin and thioredoxin reductase [37]. Still, the overall picture is that replacing P76 impairs activity, which explains why proline is conserved at position 76 in thioredoxins. Another critical point to note here is that the folding rates are similar for the three proteins when alanine is present at position 76 and diverge when proline is present at the position. Therefore, the slower folding of *E. coli* thioredoxin is attributed to mutational differences with the ancestral proteins that aggravate the kinetic bottleneck created by proline 76. Our efforts to identify such degrading mutations are described in the next section.

Mutational basis for the slow folding of *E. coli* thioredoxin as compared with the ancestral LPBCA and LBCA thioredoxins

The sequence of *E. coli* thioredoxin and the ancestral thioredoxins studied here differ at ∼40–50 amino acid positions for a protein of ∼110 residues [18,33] and, in principle, many different mutations could be responsible for the slower folding of the modern protein. Still, sequence differences in the neighbourhood of pro76

Biochemical Journal (2019) 476 3631–3647
https://doi.org/10.1042/BCJ20190739

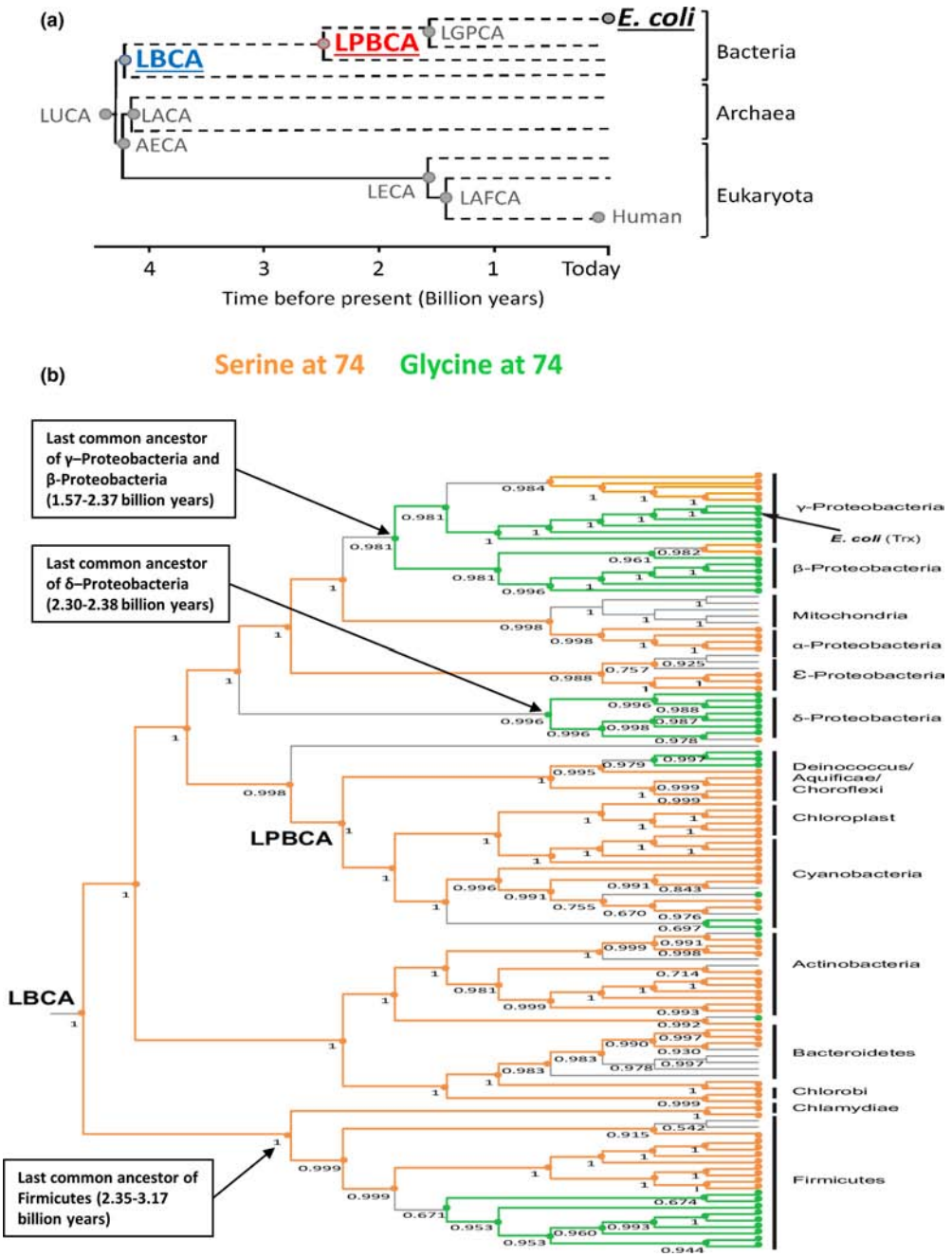


Figure 1. Thioredoxin phylogenetic tree used for ancestral sequence reconstruction [18].

(**a**) Schematic representation of the phylogenetic tree showing geological time. The nodes studied in this work are underlined and are defined in the text. See [18] for the definition of other nodes. (**b**) Bacterial section of the tree showing the evolutionary history of the amino acid (serine or glycine) present at position 74. The ages provided for some key nodes are taken from the *Timetree of Life* [32]. Numbers alongside the nodes stand for the posterior probability of the most likely residue (orange circles for serine and green circles for glycine). Most 74S residues in the modern thioredoxins are linked to the serine residue in the last common bacterial ancestor (LBCA node) through evolutionary trajectories that display serine conservation; these trajectories are highlighted in orange. Most 74G residues in modern thioredoxins can be linked to previous S74G replacements through trajectories that display glycine conservation; these trajectories are highlighted in green. The overall pattern suggests entrenchment related to coevolution with the many macromolecular partners of thioredoxin [10,11] (see Supplementary discussion in Supplementary material for details).

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

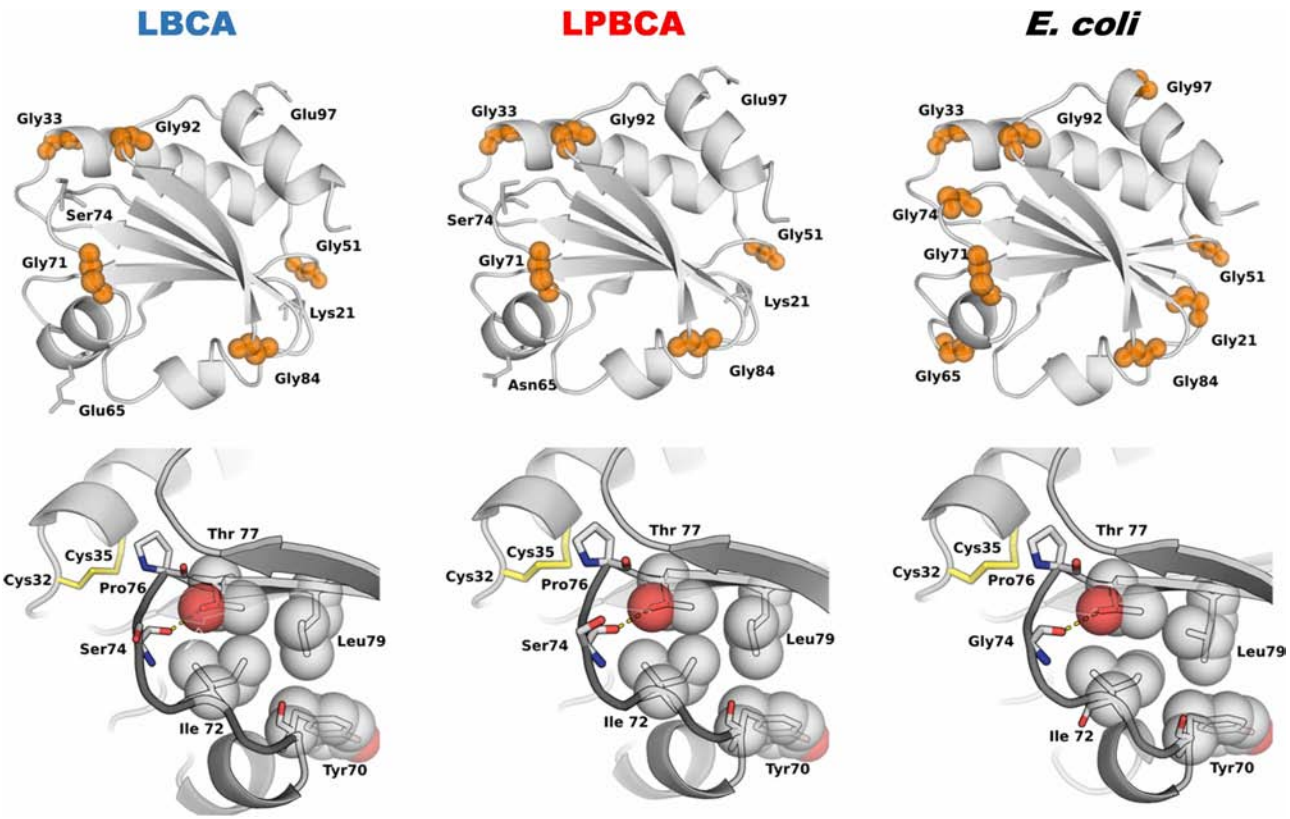


Figure 2. 3D-structures of modern and ancestral thioredoxins.

3D-structures of *E. coli* thioredoxin (modern, PDB code 2TRX), and LPBCA thioredoxin (ancestral, ∼2.5 billion years, PDB code 2YJ7) and LBCA thioredoxin (ancestral, ∼4 billion years, PDB code 4BA7) are shown. Positions with glycine residues in the modern protein are labelled. Note that the modern and ancestral thioredoxins share the thioredoxin fold, despite low sequence identity [33]. Blow-ups of the region including the *cis*-pro76 are shown below the full structures. The active-site disulfide bridge, the Gly/Ser residues at position 74 and residues that presumably stabilize the loop that includes Pro76 are highlighted.

should provide obvious candidates and one such difference stands out (Figure 2): serine is the residue at position 74 in LBCA and LPBCA thioredoxins, while glycine is the residue at position 74 in *E. coli* thioredoxin. Mutational analyses show that the S/G exchange at position 74 indeed accounts for most of the observed folding rate difference between the modern and the ancestral proteins (Figure 4a). Replacement of the ancestral residue at position 74 (serine) with glycine thus slows down folding in the ancestral LBCA and LPBCA thioredoxins, while the back-to-the-ancestor G74S in *E. coli* thioredoxin increases the rate of folding.

Many experiments support the robustness of our identification of S74G as a folding-degrading mutation. First, the folding rate enhancement obtained upon the reverse, back-to-the-ancestor G74S replacement in *E. coli* thioredoxin is reproduced when the active-site disulfide has been reduced and also when using urea, instead of guanidine, as denaturant (Figure 4b). In addition, *E. coli* thioredoxin actually has four additional glycine residues (at positions 21, 65, 74 and 97) with respect to the ancestral LPBCA and LBCA thioredoxins (Figure 2). However, extensive mutational studies (Supplementary Figure S10) indicate that it is only the glycine/ancestral-state replacement at position 74 that affects the folding rate.

A mutation (S74G) that aggravated the folding problem created by the active site *cis*-proline occurs in the line of descent that led to *E. coli* thioredoxin

As discussed above, the effect of the G/S exchange on the thioredoxin folding rate is experimentally robust. However, while the fact that glycine is the modern residue at position 74 (the residue present in *E. coli*

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

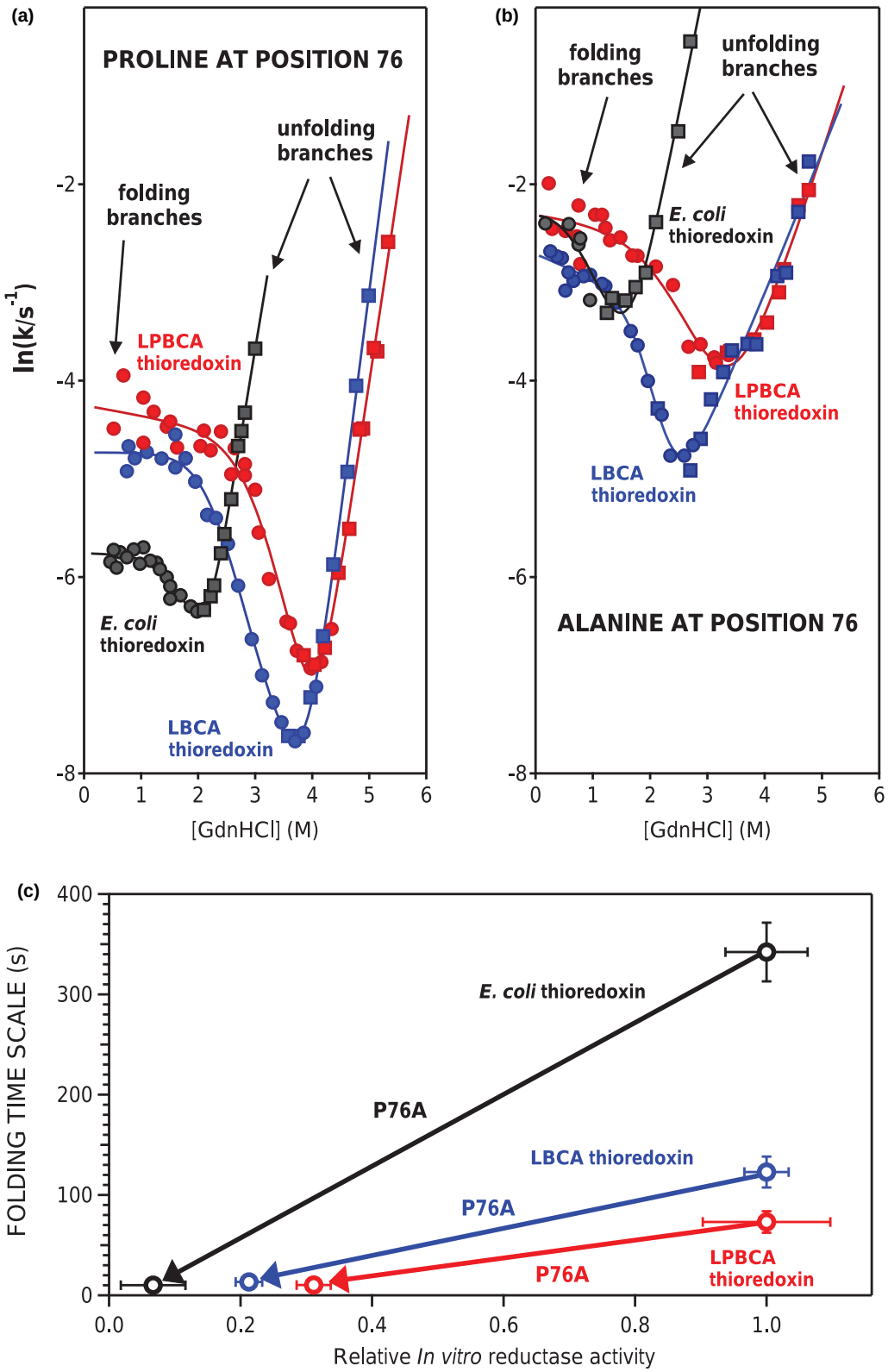


Figure 3. Folding–unfolding rates for *E. coli* thioredoxin and two resurrected Precambrian thioredoxins (see Figure 1). Part 1 of 2

(**a**) Chevron plots (including folding and unfolding branches) of the logarithm of the rate constant versus guanidine concentration for the 'wild-type' proteins that display the conserved proline at position 76. Circles and squares refer to the data

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

Figure 3. Folding–unfolding rates for *E. coli* thioredoxin and two resurrected Precambrian thioredoxins (see Figure 1). Part 2 of 2

obtained in the folding and unfolding directions, respectively. (**b**) same as (**a**) except for the P76A variants of the three proteins. (**c**) Plot of the time scale for folding versus *in vitro* reductase activity for the modern *E. coli* thioredoxin and the ancestral LPBCA and LBCA thioredoxins. Values of the folding time scale are calculated as the inverse of the folding rate constant extrapolated to zero denaturant concentration (see Materials and methods for details). Data for the 'wild-type' forms and the P76A variants included in this plot are connected by arrows in order to highlight the effect of the mutation and the function-folding trade-off: eliminating the proline at position 76 accelerates folding, but impair function. Error bars are not shown when they are smaller than the data point.

thioredoxin) is an observable result, the identification of serine as the ancestral residue is a statistical inference. There could be some doubt, therefore, that the S74G replacement actually occurred in the line of descent that led to *E. coli* thioredoxin. This is of particular interest given discussions from the literature [38–41] that ancestral sequence reconstruction can potentially be biased from uncertainties in the process. In this case, however,

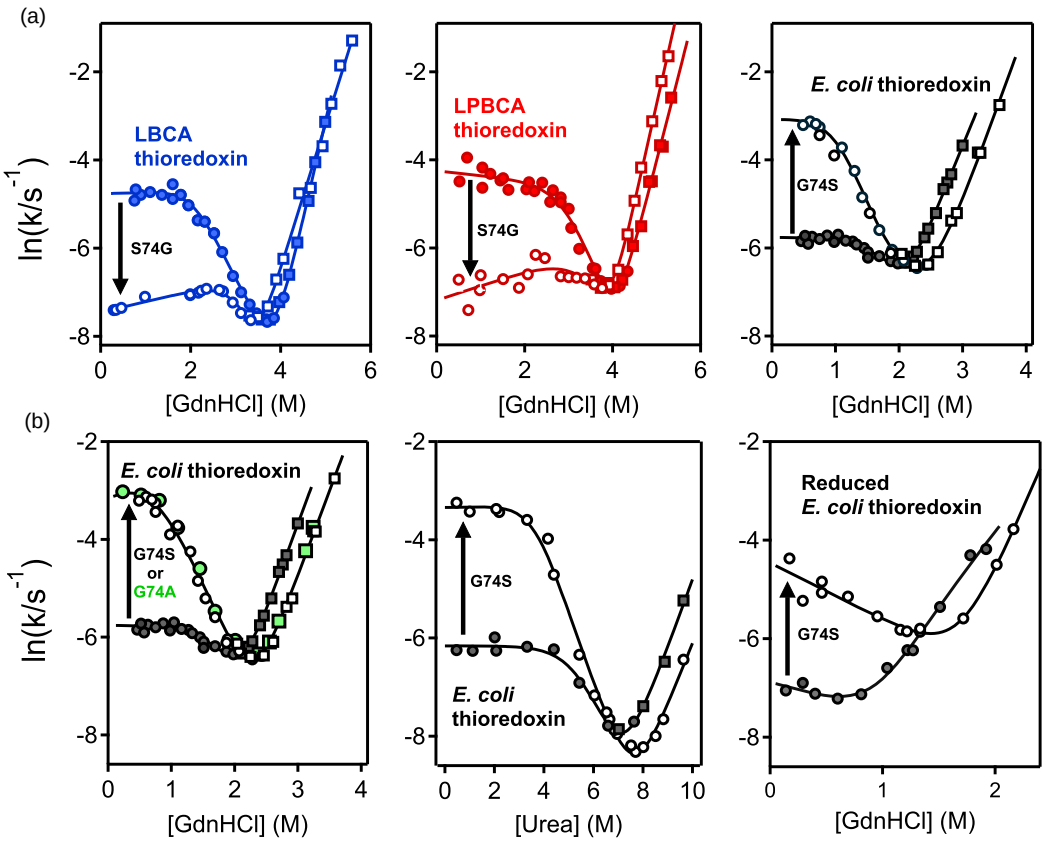


Figure 4. Effect of the G/S exchange at position 74 on thioredoxin folding rate.

Chevron plots of the logarithm of folding–unfolding rate versus guanidine concentration are shown for the 'wild-type' forms (closed data points) and variants (open data points). Circles and squares refer to the data obtained in the folding and unfolding directions, respectively. (**a**) Comparison between the modern *E. coli* thioredoxin and the ancestral LBCA and LPBCA thioredoxins. Note that glycine is the wild-type residue in *E. coli* thioredoxin, while serine is 'wild-type' in the ancestral thioredoxins. The effects of the mutation that replaces the wild-type are highlighted with an arrow. (**b**) Effect of the S74G mutation on *E. coli* thioredoxin folding and unfolding rates. Data obtained using urea and guanidine as denaturants are shown. In the latter case, data obtained using thioredoxin with a reduced active-site disulfide are also included. Note that the panel at the left also includes the effect of a G74A mutation (see text for details).

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

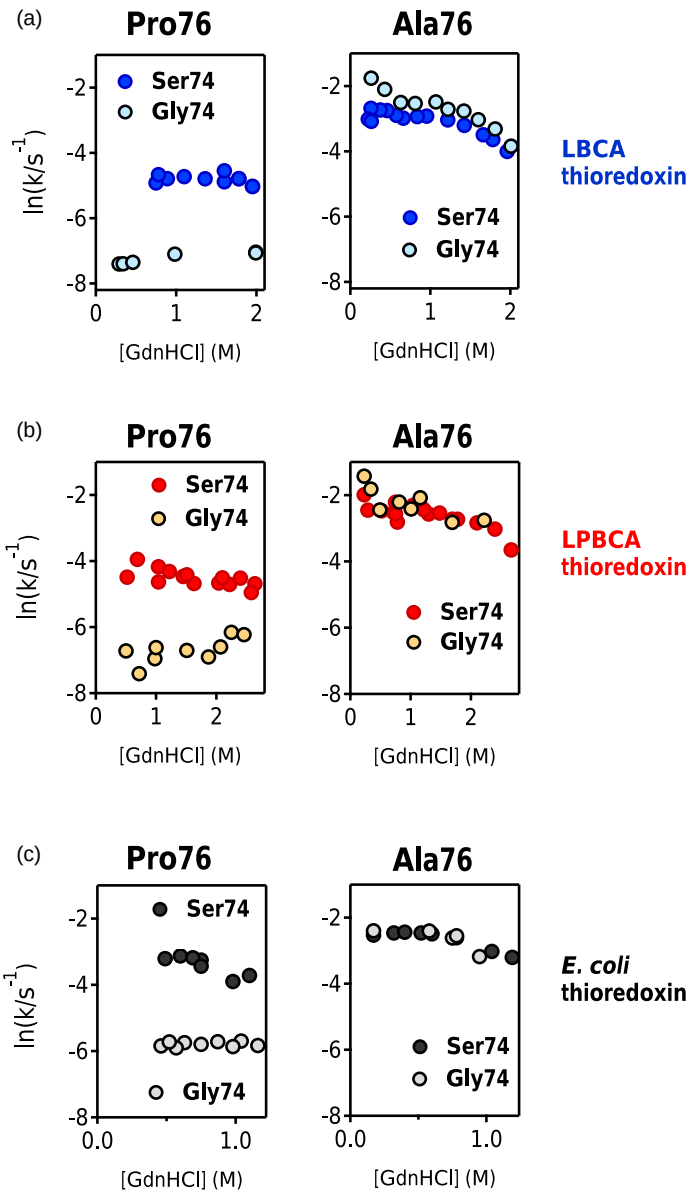


Figure 5. Double-mutant cycle analysis on the folding rate coupling between positions 74 and 76 in thioredoxins.
Folding rates for variants of: (**a**) ancestral LBCA thioredoxin; (**b**) ancestral LPBCA thioredoxin; (**c**) modern *E. coli* thioredoxin. Data for thioredoxin variants including P/A exchange at position 76, G/S exchange at position 74 and the combination of the two exchanges are shown. The plots shown are meant to make visually apparent the effect of the G74S mutation is substantial when proline is present at position 76 (plots at the left) but very low when alanine is present at position 76 (plots at the right). Note that glycine is the wild-type residue in *E. coli* thioredoxin, while serine is 'wild-type' in the ancestral thioredoxins. Proline is the wild-type residue at position 76 in all thioredoxins.

the identification of serine as the ancestral state at position 74 is quite robust. This follows first from the observation that serine is the consensus residue (i.e. the most frequent residue) at position 74 in modern thioredoxins (Figure 1b). Of course, discrepancies between consensus sequences and reconstructed ancestral sequences do exist and may have a phenotypic impact. Yet, as we have recently discussed [42], these discrepancies are typically restricted to positions at sites with a high sequence diversity and, consequently, high evolutionary rates. This does not appear to be the case for position 74 which is populated mainly by serine and glycine residues in modern bacterial thioredoxins (Figure 1b). Furthermore, the Bayesian posterior probabilities for the inferred

residues in both ancestors at position 74 is 100% (Figure 1b). We have previously shown that such sites rarely, if ever, are incorrectly inferred with such a high posterior probability [43].

Secondly, the S74G mutation decreases the folding rate of the ancestral LPBCA and LBCA thioredoxins by about one order of magnitude and the back-to-the-inferred-ancestor mutation G74S increases the folding rate of *E. coli* thioredoxin by about one order of magnitude (Figure 4 and Supplementary Figure S10). Therefore, the effect of the S/G exchange at position 74 on the folding rate is to a large extent independent of the background sequence (modern or ancestral). This implies that the mutational effect is reasonably robust against reconstruction uncertainties in other positions of the thioredoxin molecule.

Finally, the link between the effect of the S/G replacement at position 74 on folding kinetics and the *cis*-proline at position 76 is immediately revealed by a double-mutant cycle analysis of the coupling between positions 74 and 76 on the three thioredoxins studied (Figure 5). The S/G replacement at position 74 thus strongly affects the folding rate only when proline is at position 76 and not when pro76 has been replaced with alanine. Clearly, the mutation S74G did occur in the line of descent that led to *E. coli* thioredoxin and aggravated the folding problem created by the active site *cis*-proline.

Experimental study of a set of modern bacterial thioredoxins shows that folding rates are not evolutionarily conserved

The effect of the S/G exchange at position 74 on the thioredoxin folding rate is likely related to the fact that glycine residues have no side-chains which place little restriction on backbone dihedral angles and generate flexible links in polypeptide chains [44]. The 3D-structures of the modern and ancestral thioredoxins studied so far (Figure 2) reveal interactions that appear to stabilize the 70–79 segment in the conformation imposed by the *cis*-Pro76 (Figure 2, lower panel), namely hydrophobic contacts and a hydrogen bond between the backbone carbonyl of residue 74 (either G or S) and Thr77. Besides stabilizing the native protein structure with a *cis*-proline at position 76, these interactions should also favour local residual structures in the high energy regions of the folding landscape that favour the *cis* conformer, thus promoting correct folding. However, the flexible link generated by glycine residue at position 74 should allow many alternative conformations for the 70–79 segment to occur in the upper (high energy) regions of the folding landscape thus slowing down folding. This interpretation is strongly supported by the fact that replacing glycine at position 74 in *E. coli* thioredoxin with alanine brings about an increase in folding rate that is essentially identical with that produced by the G74S (left panel in Figure 4b), thus indicating folding rate enhancement is caused by the elimination of flexibility at position 74, either through the G74A mutation or through the G74S mutation.

The interpretation proposed above suggests that other amino acid replacements in the neighbourhood of position 74 could also impact the folding rate. In particular, a visual inspection of the 3D-structures (Figure 2) points to the residues at positions 70, 72, 74, 77 and 79 as being involved in interactions that could plausibly modulate the stability of the 70–79 in upper regions of the folding landscape. We, therefore, used these positions to guide the selection of a set of modern bacterial thioredoxins for experimental characterization. We performed a search in the NCBI Reference Sequence Database using the sequence of *E. coli* thioredoxin as query and considered the ∼5000 top hits. A substantial fraction of these sequences displayed differences with *E. coli* thioredoxin at positions 70, 72, 74, 77 and 79. We selected a small subset of modern thioredoxins to capture this sequence diversity in a meaningful way (Supplementary Figure S12). That is, for some of the proteins in the subset, most residues at positions 70, 72, 74, 77 and 79 are the same as in the ancestral LBCA or LPBCA thioredoxins (including the presence of serine at position 74). Yet, other proteins in the subset differ at several of the selected positions from the sequences of ancestral LBCA and LPBCA thioredoxins as well as from the sequence of *E. coli* thioredoxin. In all cases, the thioredoxin selected displayed the highest sequence identity with *E. coli* thioredoxin, given the amino acid residues present at positions 70, 72, 74, 77 and 79. The selected thioredoxins show similar activities in the insulin aggregation assay (Supplementary Table S3).

Folding rates for the 14 modern proteins in the subset determined using urea as denaturant (Figure 6a) span a ∼100-fold range, a result which is confirmed by double-jump unfolding experiments that directly probe the amount of native protein (Figure 6b, see Materials and methods for details). Of course, we cannot rule out that a substantial part of this observed folding rate variation is due to mutational changes outside the five positions we have used to guide the sequence selection. This, however, would not affect in the least the main implication of the data, namely that, contrary to what has been claimed in recent literature [16,17], *in vitro* thioredoxin

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

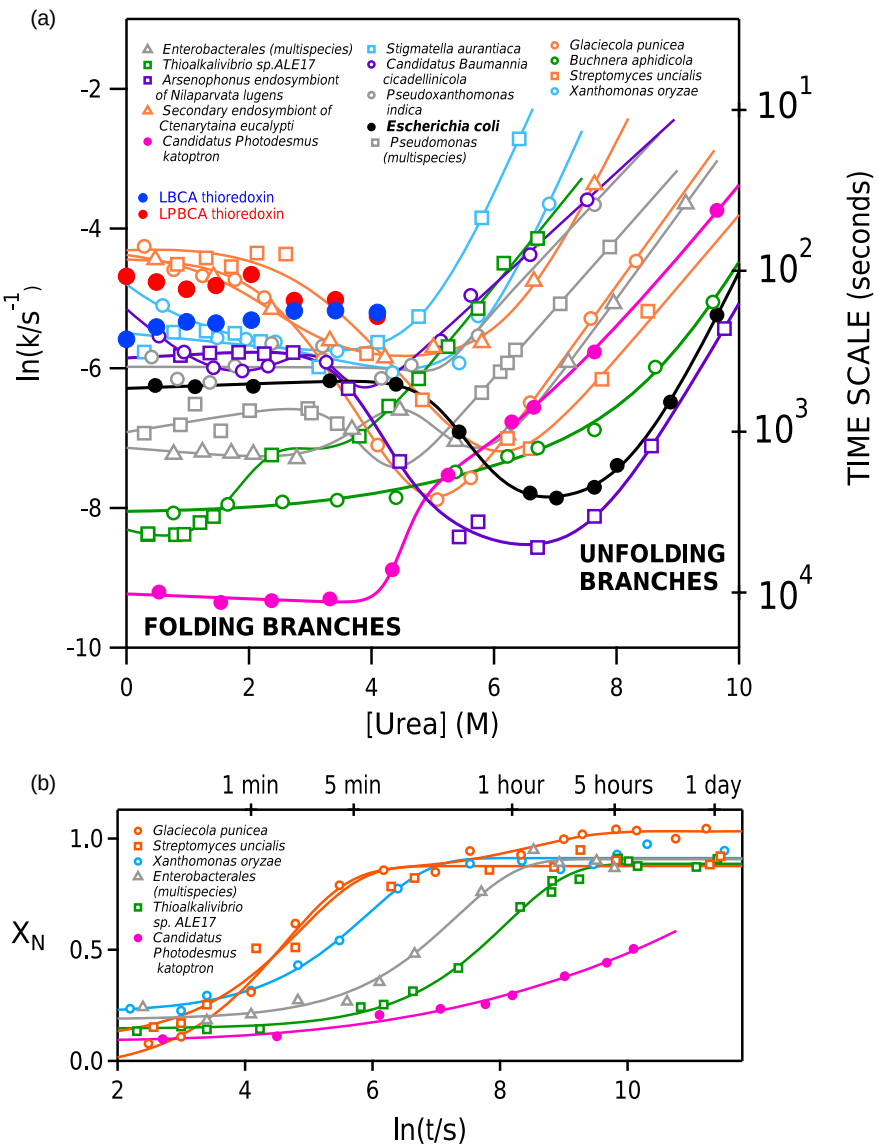


Figure 6. Folding rates are not conserved in the thioredoxin family.

(**a**) Chevron plots of folding–unfolding rate constant versus urea concentration for a set of 14 modern bacterial thioredoxins. Folding rates given correspond to the major slow phase of the fluorescence kinetic profiles (see Materials and methods for details). The time scale shown in the right axis is calculated as the inverse of the rate constant. For comparison, experimental folding data for LBCA and LPBCA thioredoxins are also included. Given the high stability of the ancestral proteins, their folding rates in urea solutions were obtained by denaturation in guanidine followed by high-dilution transfer to urea solutions, in such a way that the final guanidine concentration was low (~0.1 M). Note that the folding rate data in this plot span about two orders of magnitude. (**b**) Folding in 1 M urea as followed by double-jump unfolding assays, a methodology that allows a determination of the fraction of native protein. The profiles shown reveal the major folding channel (see Materials and methods for details) and are consistent with the range of folding times of the data shown in panel (**a**).

folding rates are not conserved, with some modern thioredoxins folding substantially faster and substantially slower than *E. coli* thioredoxin (closed black data points in Figure 6a).

We have also included in Figure 6a rate data for the ancestral LBCA and LPBCA thioredoxins in urea solutions. Clearly, while the ancestral proteins and some of the modern proteins studied fold in the approximately minute time scale, folding of some other modern thioredoxins occurs in the much slower approximately hour time scale.

Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

Folding on-rates do not correlate with stability in the thioredoxin family

Unfolding rates, as given by the unfolded branches of the chevron plots in Figure 6a, already provide a useful metric of kinetic stability [30,31]. To assess the thermodynamic stability of the modern proteins studied we fitted equilibrium profiles of fluorescence intensity versus urea concentration on the basis of the linear extrapolation model (see Material and methods for details). Fits were visually excellent (Figure 7a) and allowed us to derive values of a known metric of thermodynamic stability: the urea concentration at which unfolding free

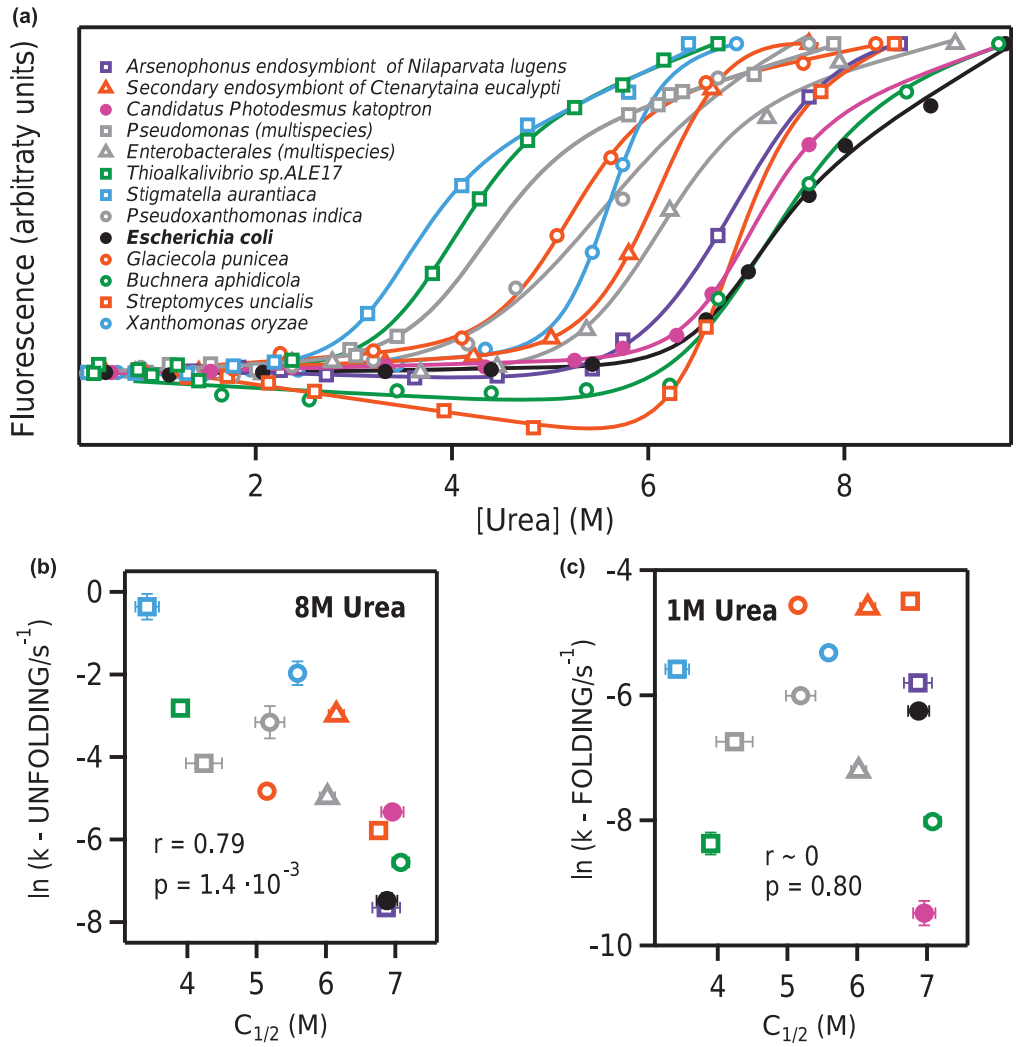


Figure 7. Folding on-rates do not correlate with stability in the thioredoxin family.

(**a**) Equilibrium profiles of fluorescence intensity versus urea concentration for modern thioredoxins. Lines represent the best fits of a model that assumes a linear dependence of the unfolding free energy with denaturant concentration within the narrow transition range and uses linear pre- and post-transition baselines (see Materials and methods for details and Supplementary Table S4 for the values obtained for the fitting parameters). (**b** and **c**) Plots of the logarithm of the unfolding rate constant (**b**) and folding rate constant (**c**) versus urea midpoint concentration (the urea concentration at which the unfolding free energy is zero). The values of the folding and unfolding rate constants are interpolated (for 1 M and 8 M urea concentration, respectively) from the corresponding branches of the Chevron plots. Associated errors are not shown when they are smaller than the size of the data points. Values of the correlation coefficient (*r*) and the probability that the correlation observed occurs by chance (*p*) are shown. Unfolding rates (kinetic stability) correlate with thermodynamic stability (**b**) but folding rates do not (**c**) (see the text for details).

Biochemical Journal (2019) 476 3631–3647
https://doi.org/10.1042/BCJ20190739

energy is zero. These $C_{1/2}$ values, as well as the m slopes that measure the urea-concentration dependence of the unfolding free energy, are given in Supplementary Table S4 for the modern thioredoxins.

Unfolding rates do show some correlation with $C_{1/2}$ values (Figure 7b). This association between kinetic and thermodynamic stabilities is to be expected from the fact that the transition state for thioredoxin unfolding is substantially unstructured, as we have previously noted and discussed [20,24]

Plotting the logarithm of folding rate versus $C_{1/2}$ produces a scattergram (Figure 7c). This lack of correlation, however, should not come as a surprise, as kinetic complexity and the presence of kinetic folding intermediates should disconnect folding rates from stability. Uncoupling of kinetics and thermodynamics was already advanced by Agard and co-workers in their seminal work on the kinetic stability of α-lytic proteases [45]. The thioredoxin family provides an example of a phenomenon, folding rate/stability uncoupling, that is likely to be widespread, given that the *in vitro* folding of many proteins, even small ones, is kinetically complex, as it has been known for many years [46].

Conclusions

It has been recently claimed that thioredoxin folding rates as determined *in vitro* are evolutionarily conserved [16,17]. This supposed conservation was furthermore taken as the first experimental evidence of a cornerstone of protein folding theory: the principle of minimal frustration. In very simple terms, the folding landscape was optimized (minimally frustrated) at a very early stage and remained so over billions of years of evolutionary history leading to folding rate conservation among modern proteins. As elaborated below in some detail, however, this proposal is inconsistent not only with well-known principles of evolutionary theory, but also with our current understanding of folding processes *in vivo*.

Proteins do not evolve in isolation because they are involved *in vivo* in a wide diversity of interactions [47]. In particular, protein folding within modern organisms relies on exceedingly complex intermolecular interactions that guide and assist the process [4–9]. Folding *in vivo* occurs co-translationally, local folding events may already take place within the ribosome exit tunnel and folding may be coupled to translation kinetics. Nascent chains are involved in many interactions as they emerge from the ribosomal tunnel, including interactions with the trigger factor, a protein that binds to exposed hydrophobic segments. The trigger factor is the first of many specialized molecules (folding chaperones) that, together with the modern ribosome, assist protein folding *in vivo*. As a result of the numerous intermolecular interactions involved, the conformational space explored by a folding chain within a modern organism *in vivo* may differ substantially from the conformational space that the folding chain explores *in vitro* [8]. Therefore, unassisted protein folding, as probed by *in vitro* studies, does not *necessarily* correlate with the biologically-relevant assisted folding that takes place within modern organisms.

Of course, it is conceivable that at a very early evolutionary stage, prior to the emergence of folding assistance, folding efficiency relied on fast folding that minimized the time scale the polypeptide chain spent in partially folded states which are susceptible to aggregation and other undesirable interactions. That is, fast unassisted folding, linked to a landscape with low (perhaps minimal) frustration, may have been required at a primordial stage. However, once folding assistance was available, mutations that impaired unassisted folding could be accepted. That is, as it is common in morphological evolution, a feature that it is no longer useful undergoes evolutionary degradation. However, while degradation at the morphological level may often be visually apparent, degradation of unassisted folding can only be revealed by *in vitro* folding experiments, since folding within modern cells is assisted.

Indeed, the *in vitro* experiments reported here are consistent with the evolutionary degradation of unassisted folding. We have unambiguously identified a mutation that substantially slows down *in vitro* thioredoxin folding and that was accepted in the line of descent that led to *E. coli* thioredoxin. Furthermore, we have shown that, while resurrected Precambrian thioredoxins and some modern thioredoxins fold *in vitro* in the approximately minute time scale, other modern thioredoxins approach the approximately hours time scale. Such variation in the folding rate, indicating different degrees of degradation, should not come as surprise. As Darwin already advanced in the first chapter of *The Descent of Man*, the variability of superfluous features that are not under natural selection should be a common observation in morphological studies. Indeed, such type of variability is often illustrated with the widely variable size of the human appendix [1] while here we have provided an example at the molecular level. It is also worth noting that folding in the approximately hours time scale is hardly of any biological significance and that assisted folding *in vivo* is likely to be much faster. Overall, it is clear that *in vitro* thioredoxin folding rates are not evolutionarily conserved. As we have previously

noted [20], recent claims to the contrary [16,17] are probably related to the use of destabilizing conditions which buffer the effect of landscape ruggedness on *in vitro* folding experiments, as it has been known for many years [48] and it is visually apparent in our data of Figure 6a. Note that the folding rates are roughly similar for most thioredoxins at 4 M urea, while they diverge as the denaturant concentration becomes lower and the solvent becomes less destabilizing. Furthermore, the very acidic pH (=2) employed in previous studies [16] brings about extensive protonation of residues and alterations of ionic interactions that are unlikely to be physiologically and evolutionary relevant [20].

Finally, beyond clearing up a relevant and consequential controversy in recent literature, this work has general implications of wide interest. It points to a simple evolutionary interpretation of *in vitro* protein folding as a degraded version of primordial unassisted folding and thus may contribute to clarify the much-debated issue of the relation between protein folding *in vivo* and protein folding *in vitro*. More generally, this work provides evidence that degradation shapes evolution not only at the morphological level, but also at the level of individual enzymes.

Abbreviations

DTNB, 5,5′-dithiobis(2-nitrobenzoic acid); LPBCA, last common ancestor of the cyanobacterial, deinococcus and thermus groups; LBCA, last bacterial common ancestor.

Authors Contributions

G.G.-A. purified the several variants of modern and ancestral thioredoxins; she also performed and analyzed the experiments aimed at determining their folding–unfolding kinetics. V.A.R. performed bioinformatics analyses and provided essential input for the evolutionary interpretation of the data. A.I.-P. performed preliminary experiments that pointed to a crucial role of glycine residues. A.M.C. and M.L.R.-R. performed preliminary experiments that pointed to the enhanced folding and unfolding rates for the ancestral thioredoxins. E.A.G. assessed the robustness of the reconstructed ancestral states and provided essential input for the evolutionary interpretation of the data. J.A.G. provided essential input for the structural analysis of the mutational effects on folding rates. B.I.-M. designed and supervised the folding–unfolding kinetic experiments. B.I.-M. and J.M.S.-R. directed the project. J.M.S.-R. provided the general evolutionary interpretation and wrote the paper. All authors read the manuscript and provided useful comments.

Funding

This research was supported by FEDER Funds, grant BIO2015-66426-R from the Spanish Ministry of Economy and Competitiveness (J.M.S.-R.), grant RGP0041/2017 from the Human Frontier Science Program (J.M.S.-R. and E.A.G.) and National Institutes of Health 1R01AR069137 (E.A.G.), Department of Defence MURI W911NF-16-1-0372 (E.A.G.).

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

References

1 Coyne, J.A. (2009) *Why Evolution is True, Ch. 3*, Oxford Univ. Press, New York, NY
2 Dawkins, R. (2009) *The Greatest Show on Earth: the Evidence for Evolution, Ch. 11*, Bantam Press, London, U.K.
3 Kratzer, J.T., Lanaspa, M.A., Murphy, M.N., Cicerchi, C., Graves, C.L., Tipton, P.A. et al. (2014) Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 3763–3768 https://doi.org/10.1073/pnas.1320393111
4 Kaiser, C.M., Goldman, D.H., Chodera, J.D., Tinoco, I. and Bustamante, C. (2011) The ribosome modulates nascent protein folding. *Science* **334**, 1723–1727 https://doi.org/10.1126/science.1209740
5 Oh, E., Becker, A.H., Sandikci, A., Huber, D., Chaba, R., Gloge, F. et al. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* **147**, 1295–1308 https://doi.org/10.1016/j.cell.2011.10.044
6 Zhang, G. and Ignatova, Z. (2011) Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *Curr. Opin. Struct. Biol.* **21**, 25–31 https://doi.org/10.1016/j.sbi.2010.10.008
7 Kim, Y.E., Hipp, M.S., Bracher, A., Hayer-Hartl, M. and Hartl, F.U. (2013) Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.* **82**, 323–355 https://doi.org/10.1146/annurev-biochem-060208-092442
8 Balchin, D., Hayer-Hartl, M. and Hartl, F.U. (2016) In vivo aspects of protein folding and quality control. *Science* **353**, aac4354 https://doi.org/10.1126/science.aac4354
9 Thommen, M., Holtkamp, W. and Rodnina, M.V. (2017) Co-traslational protein folding: progress and methods. *Curr. Opin. Struct. Biol.* **42**, 83–89 https://doi.org/10.1016/j.sbi.2016.11.020

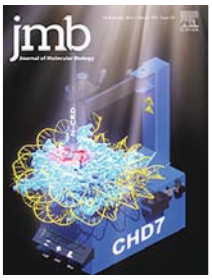
Biochemical Journal (2019) **476** 3631–3647
https://doi.org/10.1042/BCJ20190739

10 Holmgren, A. (1985) Thioredoxin. *Annu. Rev. Biochem.* **54**, 237–271 https://doi.org/10.1146/annurev.bi.54.070185.001321
11 Kumar, J.K., Tabor, S. and Richardson, C.C. (2004) Proteomic analysis of thioredoxin-targeted proteins in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 3759–3764 https://doi.org/10.1073/pnas.0308701101
12 Kelley, R.F. and Richards, F.M. (1987) Replacement of proline-76 with alanine eliminates the slowest kinetic phase in thioredoxin folding. *Biochemistry* **26**, 6765–6774 https://doi.org/10.1021/bi00395a028
13 Brandts, J.F., Halvorson, H.R. and Brennan, M. (1975) Consideration of the possibility that the slow step on protein denaturation reactions is due to cis–trans isomerism of proline residues. *Biochemistry* **14**, 4953–4963 https://doi.org/10.1021/bi00693a026
14 Schmid, F.X. and Baldwin, R.L. (1978) Acid catalysis of the formation of the slow-folding species of RNase A: evidence that the reaction is proline isomerization. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4764–4768 https://doi.org/10.1073/pnas.75.10.4764
15 Schmidpeter, P.A. and Schmid, F.X. (2015) Prolyl isomerization and its catalysis in protein folding and protein function. *J. Mol. Biol.* **427**, 1609–1631 https://doi.org/10.1016/j.jmb.2015.01.023
16 Tzul, F.O., Vasilchuk, D. and Makhatadze, G.I. (2017) Evidence for the principle of minimal frustration in the evolution of protein folding landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E1627–E1632 https://doi.org/10.1073/pnas.1613892114
17 Tzul, F.O., Vasilchuk, D. and Makhatadze, G.I. (2017) Evidence for the evolutionary conservation of folding kinetics in the thioredoxin protein family. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4124 https://doi.org/10.1073/pnas.1704669114
18 Perez-Jimenez, R., Inglés-Prieto, A., Zhao, Z.M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P. et al. (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* **18**, 592–596 https://doi.org/10.1038/nsmb.2020
19 Romero-Romero, M.L., Risso, V.A., Martinez-Rodriguez, S., Ibarra-Molero, B. and Sanchez-Ruiz, J.M. (2016) Engineering ancestral protein hyperstability. *Biochem. J.* **473**, 3611–3620 https://doi.org/10.1042/BCJ20160532
20 Candel, A.M., Romero-Romero, M.L., Gamiz-Arco, G., Ibarra-Molero, B. and Sanchez-Ruiz, J.M. (2017) Fast folding and slow unfolding of a resurrected Precambrian protein. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4122–E4123 https://doi.org/10.1073/pnas.1703227114
21 Acevedo, O., Guzman-Casado, M., Garcia-Mira, M.M., Ibarra-Molero, B. and Sanchez-Ruiz, J.M. (2002) pH corrections in chemical denaturant solutions. *Anal. Biochem.* **306**, 158–161 https://doi.org/10.1006/abio.2002.5668
22 Holmgren, A. (1979) Thioredoxin catalyzes the reduction of insulin disulfides by dithiothreitol and dihydrolipoamide. *J. Biol. Chem.* **254**, 9627–9632 PMID: 385588
23 Slaby, I. and Holmgren, A. (1975) Reconstitution of *Escherichia coli* thioredoxin from complementing peptide fragments obtained by cleavage at methionine-37 or arginine-73. *J. Biol. Chem.* **250**, 1340–1347 PMID: 803502
24 Godoy-Ruiz, R., Ariza, F., Rodriguez-Larrea, D., Perez-Jimenez, R., Ibarra-Molero, B. and Sanchez-Ruiz, J.M. (2006) Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments. *J. Mol. Biol.* **362**, 966–978 https://doi.org/10.1016/j.jmb.2006.07.065
25 Greene, R.F. and Pace, C.N. (1974) Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, α-chymotrypsin, and β-lactoglobulin. *J. Biol. Chem.* **249**, 5388–5393 PMID: 4416801
26 Radford, S.E., Dobson, C.M. and Evans, P.A. (1992) The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* **358**, 302–307 https://doi.org/10.1038/358302a0
27 Georgescu, R.E., Li, J.H., Goldberg, M.E., Tasayco, M.L. and Chaffotte, A.F. (1998) Proline isomerization-independent accumulation of an early intermediate and heterogeneity of the folding pathways of a mixed α/β protein, *Escherichia coli* thioredoxin. *Biochemistry* **37**, 10286–10297 https://doi.org/10.1021/bi9805083
28 Mücke, M. and Schmid, F.X. (1994) A kinetic method to evaluate the two-state character of solvent-induced protein denaturation. *Biochemistry* **33**, 12930–12935 https://doi.org/10.1021/bi00209a025
29 Ibarra-Molero, B. and Sanchez-Ruiz, J.M. (1997) Are there equilibrium intermediates in the urea-induced unfolding of hen-egg white lysozyme? *Biochemistry* **36**, 9616–9624 https://doi.org/10.1021/bi9703305
30 Sanchez-Ruiz, J.M. (2010) Protein kinetic stability. *Biophys. Chem.* **148**, 1–15 https://doi.org/10.1016/j.bpc.2010.02.004
31 Colon, W., Church, J., Sen, J., Thibeault, J., Trasatti, H. and Xia, K. (2017) Biological roles of protein kinetic stability. *Biochemistry* **56**, 6179–6186. https://doi.org/10.1021/acs.biochem.7b00942
32 Hedges, S.B. and Kumar, S. (eds.) 2009. *The Timetree of Life*, Oxford Univ. Press, New York, NY
33 Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J.M., Gaucher, E.A. et al. (2013) Conservation of protein structure over four billion years. *Structure* **21**, 1690–1697 https://doi.org/10.1016/j.str.2013.06.020
34 Risso, V.A., Manssour-Triedo, F., Delgado-Delgado, A., Arco, R., Barroso-delJesus, A., Ingles-Prieto, A. et al. (2015) Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol. Biol. Evol.* **32**, 440–455 https://doi.org/10.1093/molbev/msu312
35 Delgado, A., Arco, R., Ibarra-Molero, B. and Sanchez-Ruiz, J.M. (2017) Using resurrected ancestral proviral proteins to engineer virus resistance. *Cell Rep.* **19**, 1247–1256 https://doi.org/10.1016/j.celrep.2017.04.037
36 Garcia, A.K., Schopf, J.W., Yokobory, S., Akanuma, A. and Yamagishi, A. (2017) Reconstructed ancestral enzymes suggest long-term cooling of Earth's photic zone since the Archean. *Proc. Natl. Acad. Sci. U.S.A* **114**, 4619–4624 https://doi.org/10.1073/pnas.1702729114
37 Napolitano, S., Reber, R.J., Rubini, M. and Glockshuber, R. (2019) Functional analyses of ancestral thioredoxins provide insight into their evolutionary history. *J. Biol. Chem.* **294**, 14105–14118 https://doi.org/10.1074/jbc.RA119.009718
38 Williams, P.D., Pollock, D.D., Blackburne, B.P. and Goldstein, R.A. (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput. Biol.* **2**, e69 https://doi.org/10.1371/journal.pcbi.0020069
39 Groussin, M., Hobbs, J.K., Szöllősi, G.J., Gribaldo, S., Arcus, V.L. and Gouy, M. (2015) Toward more accurate ancestral protein genotype-phenotype reconstructions with the use of species tree-aware gene trees. *Mol. Biol. Evol.* **32**, 13–22 https://doi.org/10.1093/molbev/msu305
40 Eick, G.N., Bridgham, J.T., Anderson, D.P., Harms, M.J. and Thornton, J.W. (2017) Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Mol. Biol. Evol.* **34**, 247–261 https://doi.org/10.1093/molbev/msw223
41 Vialle, R.A., Tamuri, A.U. and Goldman, N. (2018) Alignment modulates ancestral sequence reconstruction accuracy. *Mol. Biol. Evol.* **35**, 1783–1797 https://doi.org/10.1093/molbev/msy055

42 Risso, V.A., Gavira, J.A., Gaucher, E.A. and Sanchez-Ruiz, J.M. (2014) Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins* **82**, 887–896 https://doi.org/10.1002/prot.24575
43 Randall, R.N., Radford, C.E., Roof, K.A., Natarajan, D.K. and Gaucher, E.A. (2016) An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat. Commun.* **7**, 12847 https://doi.org/10.1038/ncomms12847
44 Scott, K.A., Alonso, D.O., Sato, S., Fersht, A.R. and Daggett, V. (2007) Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2661–2666 https://doi.org/10.1073/pnas.0611182104
45 Jaswal, S.S., Sohl, J.L., Davis, J.H. and Agard, D.A. (2002) Energetic landscape of α-lytic protease optimizes longevity through kinetic stability. *Nature* **415**, 343–346 https://doi.org/10.1038/415343a
46 Kim, P.S. and Baldwin, R.L. (1982) Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* **51**, 459–489 https://doi.org/10.1146/annurev.bi.51.070182.002331
47 Gun, D. and Gruebele, M. (2019) Weak chemical interactions that drive protein evolution: crowding, sticking, and quinary structure in folding and function. *Chem. Rev.* **119**, 10691–10717 https://doi.org/10.1021/acs.chemrev.8b00753
48 Matouscheck, A., Kellis, J.T., Serrano, L., Bycroft, M. and Fersht, A.R. (1990) Transient folding intermediates characterized by protein engineering. *Nature* **346**, 440–445 https://doi.org/10.1038/346440a0
49 Anderson, D.W., McKeown, A.N. and Thornton, J.W. (2015) Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife* **4**, e07864 https://doi.org/10.7554/eLife.07864

PUBLICATION 3

Combining ancestral reconstruction with folding-landscape simulations to engineer heterologous protein expression



Combining Ancestral Reconstruction with Folding-Landscape Simulations to Engineer Heterologous Protein Expression

Gloria Gamiz-Arco [1†], Valeria A. Risso [1†], Eric A. Gaucher [2], Jose A. Gavira [3], Athi N. Naganathan [4*], Beatriz Ibarra-Molero [1*] and Jose M. Sanchez-Ruiz [1*]

*1 - **Departamento de Quimica Fisica,** Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain*
*2 - **Department of Biology,** Georgia State University, Atlanta, GA 30303, USA*
*3 - **Laboratorio de Estudios Cristalograficos,** Instituto Andaluz de Ciencias de la Tierra, CSIC, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, Avenida de las Palmeras 4, Armilla, Granada 18100, Spain*
*4 - **Department of Biotechnology,** Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India*

***Correspondence to Athi N. Naganathan, Beatriz Ibarra-Molero and Jose M. Sanchez-Ruiz:** athi@iitm.ac.in (A.N. Naganathan), beatriz@ugr.es (B. Ibarra-Molero), sanchezr@ugr.es (J.M. Sanchez-Ruiz) @jmsanchezruiz (J.M. Sanchez-Ruiz), @AthiNaganathan (A.N. Naganathan), @Gavirius (J.A. Gavira)*
https://doi.org/10.1016/j.jmb.2021.167321
Edited by Daniel Otzen

Abstract

Obligate symbionts typically exhibit high evolutionary rates. Consequently, their proteins may differ considerably from their modern and ancestral homologs in terms of both sequence and properties, thus providing excellent models to study protein evolution. Also, obligate symbionts are challenging to culture in the lab and proteins from uncultured organisms must be produced in heterologous hosts using recombinant DNA technology. Obligate symbionts thus replicate a fundamental scenario of metagenomics studies aimed at the functional characterization and biotechnological exploitation of proteins from the bacteria in soil. Here, we use the thioredoxin from *Candidatus Photodesmus katoptron*, an uncultured symbiont of flashlight fish, to explore evolutionary and engineering aspects of protein folding in heterologous hosts. The symbiont protein is a standard thioredoxin in terms of 3D-structure, stability and redox activity. However, its folding outside the original host is severely impaired, as shown by a very slow refolding *in vitro* and an inefficient expression in *E. coli* that leads mostly to insoluble protein. By contrast, resurrected Precambrian thioredoxins express efficiently in *E. coli*, plausibly reflecting an ancient adaptation to unassisted folding. We have used a statistical-mechanical model of the folding landscape to guide back-to-ancestor engineering of the symbiont protein. Remarkably, we find that the efficiency of heterologous expression correlates with the *in vitro* (*i.e.,* unassisted) folding rate and that the ancestral expression efficiency can be achieved with only 1–2 back-to-ancestor replacements. These results demonstrate a minimal-perturbation, sequence-engineering approach to rescue inefficient heterologous expression which may potentially be useful in metagenomics efforts targeting recent adaptations.

Introduction

Many proteins of industrial and therapeutic interest are produced in heterologous hosts using recombinant DNA technology.[1–3] Moreover, heterologous expression is unavoidable in metagenomics studies aimed at the functional characterization and biotechnological exploitation of proteins from organisms that cannot be cultured in the lab,[4,5] *i.e.,* the majority of the bacteria in soil.[6] Unfortunately, over-expression of a foreign gene poses a serious challenge to an organism and may lead to non-functional species, such as misfolded protein, proteolyzed protein or, more typically, insoluble protein aggregates.[7] These problems may be alleviated by using engineered hosts that have been modified for instance to minimize protease activity or to over-express molecular chaperones that assist correct folding.[2,8] Despite advances in host engineering, heterologous expression of functional proteins remains a major biotechnological bottleneck. For instance, about half of the proteins targeted in structural genomics initiatives could not be purified.[9]

Ancestral sequence reconstruction (ASR) uses phylogenetic analyses and sequences of modern protein homologs to compute statistically plausible approximations to the corresponding ancestral sequences.[10,11] During the last ~ 25 years, proteins encoded by reconstructed sequences ("resurrected" ancestral proteins) have been widely used as tools to address important problems in molecular evolution.[12–16] They have also been found to provide new possibilities for protein biomedical applications and protein engineering.[17–21] Ancestral proteins may considerably differ from their modern counterparts in terms of sequence and their experimental preparation necessarily involves heterologous expression, as the ancient original hosts are not available. Therefore, the fact that many resurrected ancestral proteins have been purified and studied experimentally emerges as remarkable in itself, even after acknowledging the obvious publication bias in favour of positive experimental outcomes.

Moreover, a substantial number of studies have actually reported improved heterologous expression of ancestral proteins as compared with their modern counterparts. Examples include: phosphate-binding protein,[22] periplasmic binding protein,[23] serum paraoxonase,[24] coagulation factor VIII,[25] titin,[26] haloalkane dehalogenases,[27] cytidine and adenine base editors,[28] diterpene cyclase,[29] rubisco,[30] endoglucanases,[31] L-amino acid oxidases,[32] laccases,[33] front-end Δ6-desaturases[34] and fatty acid photo-decarboxylases.[35] On a related note, recent studies on ancestral proteins have noted an improved capability to yield crystals suitable for X-ray structural determination.[36,37]

Regardless of the specific mechanisms responsible for efficient ancestral folding in modern organisms, it is clear that ancestral reconstruction may provide a basis for the sequence engineering of efficient heterologous expression. Yet, reconstructed ancestral sequences typically display extensive differences with respect to the corresponding modern sequences while, in many cases, researchers will be interested in minimally modifying the targeted modern sequence, in such a way that the properties of the encoded protein are barely altered. Minimal sequence perturbation would be particularly desirable when targeting recent adaptations, as, for instance, in metagenomics efforts at contaminated sites aimed at obtaining pollutant-degrading enzymes.[38] Overall, it is of interest to determine whether rescue of inefficient heterologous folding can be engineered on the basis of a few selected back-to-ancestor mutations.

Obligate symbionts typically display high evolutionary rates[39,40] and their proteins may be expected to differ considerably from their modern and ancestral homologs in terms of both sequence and properties. Consequently, the outcome and implications of evolutionary processes may become apparent even in small symbiont proteins that are amenable to detailed biomolecular characterization. Thioredoxins are small proteins (about 110 amino acid residues) that function as general redox catalysts in all known cells.[41] Here we use the thioredoxin from *Candidatus Photodesmus katoptron*, an uncultured symbiont of flashlight fish, as a simple model to explore evolutionary and engineering aspects of protein folding in heterologous hosts.

Flashlight fish (*Anomalopidae*) use light from sub-ocular bioluminescent organs to communicate, hunt prey and disorient predators.[42] Light is produced by luminous bacteria of the *Vibrionaceae* family of *Proteobacteria.* These bacteria are obligate symbionts and have not been cultured in the lab, thus replicating a fundamental scenario of metagenomics studies. Still, their genomes can be sequenced, since the sub-ocular organs of *Anomalopidae* fish harbour large numbers of *Vibrionaceae* in the absence of other bacteria.[43] *Candidatus Photodesmus katoptron*, the luminous bacterium of the *Anolamops katoptron* fish, shows extensive genome reduction and it is highly evolutionary divergent.[44–46] As expected from the high evolutionary rate of its original host, the sequence of the thioredoxin from *Candidatus Photodesmus katoptron* (*CPk* thioredoxin from now on) differs substantially from all known sequences of thioredoxins from other species. Despite this observation, it is similar to other modern thioredoxins (in particular, to its *E. coli* homolog) in terms of 3D-structure, stability and redox activity. However, as described below, its folding outside the original host appears severely impaired.

CPk thioredoxin displays a very slow refolding *in* vitro, reaching the native state in the time scale of hours[47] and its expression in *E. coli* at 37 °C leads mostly to insoluble protein. By contrast, resurrected

Precambrian thioredoxins have been extensively studied[47–51] and have been found to fold fast *in vitro* and efficiently in *E. coli* despite their huge sequence differences with modern thioredoxins in general and *E. coli* thioredoxin in particular.

Albeit inefficient, the heterologous folding of *CPk* thioredoxin in *E. coli* is not fully impaired and leads to about 20% of soluble protein at 37 °C, a yield that can be increased by carrying out the expression at lower temperatures to decrease protein aggregation.[52] This is a crucial feature that allows us to interrogate the folding properties of *CPk* thioredoxin variants. That is, the effect of sequence modifications on heterologous folding efficiency at 37 °C can be determined and correlated with the biomolecular properties of the corresponding variants of *CPk* thioredoxin, since these variants can actually be prepared in the lab.

We have used computational modelling of the folding landscape to guide back-to-the-ancestor engineering of *CPk* thioredoxin. Specifically, we have used a recently-developed,[53] block version of the Wako-Saitô-Muñoz-Eaton statistical-mechanical model of the folding landscape[54–57] to determine regions of the symbiont thioredoxin that are likely to be unfolded in aggregation-prone intermediate states[58] and we have performed back-to-ancestor sequence-engineering targeted to those regions. The ancestral protein we have used as reference is LPBCA thioredoxin, a putative Precambrian thioredoxin that we have previously characterized in detail[47–51] and that folds efficiently in *E. coli*, despite having only 58% sequence identity with *E. coli* thioredoxin.

Our current study includes several modern/ancestral chimeras, as well as a many single-mutant, back-to-ancestor variants, and allows us to generate several conclusions of general interest:

Folding in the heterologous host is likely akin to unassisted folding. This is supported by (i) the success of the approach used, which involves computational modelling of the unassisted folding landscape, (ii) the fact that the efficiency of heterologous expression correlates with the *in vitro* folding rate, (iii) the very limited rescue of inefficient heterologous expression by chaperone over-expression. Consequently, it appears plausible that ancestral folding efficiency reflects an adaptation to ancient unassisted folding.[47]

Stabilization does improve heterologous folding efficiency, but this cannot be explained by global stabilization alone. Rather, it is linked to specific stabilizing mutations at crucial positions in late-folding regions.

Although the sequences of the ancestral LPBCA thioredoxin and the modern *CPk* thioredoxin differ at 60 positions, the ancestral folding efficiency can be re-enacted in the symbiont thioredoxin with only 1–2 back-to-ancestor mutations. This result provides proof of concept for a minimal-perturbation, sequence-engineering approach to rescue inefficient heterologous folding with potential application in metagenomics.

Results and discussion

Sequences and biomolecular properties of the modern and ancestral thioredoxins studied in this work. Our current study utilizes the thioredoxin from the symbiont *Candidatus Photodesmus katoptron* (*CPk* thioredoxin), a modern *E. coli* homolog and a resurrected ancestral thioredoxin corresponding to the last common ancestor of the cyanobacterial, *Deinococcus* and *Thermus* groups, a Precambrian phylogenetic node dated at $\sim$ 2.5 billion years ago. This LPBCA thioredoxin, as well as other resurrected Precambrian thioredoxins, has been previously characterized in detail.[47–51]

CPk thioredoxin is highly divergent at the sequence level. A BLAST search in the non-redundant protein sequences (nr) database using as query the sequence of *CPk* thioredoxin yields a *vibrio* protein with only 75% identity as the closest hit. Sequence identity of this thioredoxin from *Candidatus Photodesmus katoptron* (belonging to the *Vibrionaceae* family of *Proteobacteria*) with the thioredoxin from *E. coli* (belonging to the *Enterobacteriaceae* family of *Proteobacteria*) is even lower: 69%. The ancestral LPBCA thioredoxin displays even lower sequence identity to both modern proteins: 57% and 45% with the thioredoxins from *E. coli* and *Candidatus Photodesmus katoptron*, respectively.

Despite the extensive sequence differences (Figure 1a), the three proteins share the thioredoxin fold (Figure 1b). The 3D-structures of *E. coli* thioredoxin and LPBCA thioredoxin have been previously reported.[49,59] The determination of the X-ray structure for *CPk* thioredoxin has been addressed in this work. However, despite numerous attempts using different crystallization conditions and approaches, we failed to obtain crystals of diffraction quality for the wild-type *CPk* thioredoxin (see Methods for details). The structure shown in Figure 1b for *CPk* thioredoxin actually corresponds to an engineered version of the protein in which a short loop (70–77) of the symbiont protein has been replaced by the corresponding loop in the ancestral LPBCA thioredoxin (see details below), a replacement which involves only 4 mutational changes. This variant of *CPk* thioredoxin did produce crystals suitable for diffraction and led to its structural model at 2.85 Å resolution.

The agreement between the three structures shown in Figure 1b is consistent with previous structural work that supported conservation of thioredoxin structure over the span of life on Earth.[49] *In vitro* redox activity, as determined by the insulin aggregation assay and by the assay with thioredoxin reductase coupled to DTNB, is also similar for the three thioredoxins (Figure S1). The two

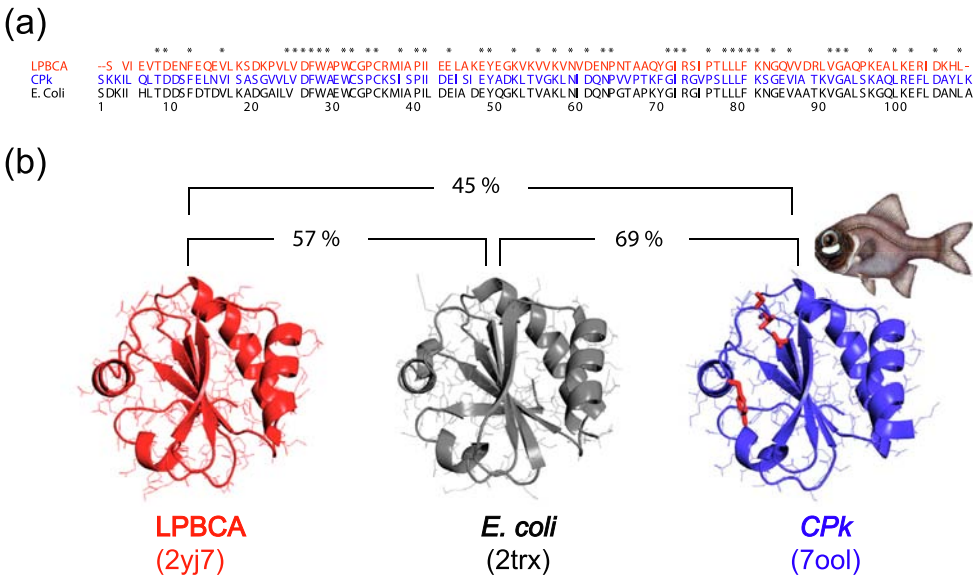


Figure 1. Sequences and structures of modern and ancestral thioredoxins. **a** Alignment of sequences from modern thioredoxins from *E. coli* and *Candidatus Photodesmus katoptron* (*CPk*), and a resurrected ancestral thioredoxin corresponding to the last common ancestor of the cyanobacterial, *Deinococcus* and *Thermus* groups (LPBCA thioredoxin). Positions with identical residues in the three sequences are labelled with asterisks. **b** 3D-structures for the three thioredoxins studied. The experimental structure of *CPk* thioredoxin actually corresponds to a variant with four back-to-ancestor replacements (highlighted in red; see main text for details). The PDB identifiers are shown, as well as the sequence identity percentages between the three proteins. *Candidatus Photodesmus katoptron* is a symbiont of flashlight fish. An illustration of a flashlight fish is shown here, alongside the *CPk* thioredoxin structure, as well as in the graphical abstract. Illustration used by courtesy of Encyclopædia Britannica, Inc., copyright 2011; used with permission.

modern thioredoxins display similar high stability as shown by midpoint of about 8 M for urea denaturation experiments[47] and by denaturation temperatures above 80 °C (see details below). The ancestral LPBCA thioredoxin is a hyperstable protein that cannot be denatured by urea at room temperature and that has a denaturation temperature of about 123 °C.[48,60]

In vitro **folding behaviour of the modern and ancestral thioredoxins studied in this work.** *In vitro* folding of thioredoxins has been known for many years to be a kinetically complex process involving intermediate states and parallel channels to arrive at the native state.[61,62] Such complexities reflect ruggedness of the folding landscape and are revealed by multi-exponential folding kinetics and rollovers in the folding branches of Chevron plots (*i.e.,* plots of folding-unfolding rate constant versus denaturant concentration). Figure 2a shows Chevron plots for the three thioredoxins studied here.[47] To identify the kinetic phase that leads to the native state, *i.e.,* the kinetic phase that defines the time-scale of refolding, we used double-jump unfolding assays,[47,63,64] which allow for a direct determination of the amount of native protein during the folding experiments. These assays (see Methods for details) are a specific instance of the well-known "jump assays" that were developed by pioneers of the *in vitro* protein folding field to resolve the kinetic complexities of folding processes.[65,66] While the interpretation of the time dependence of a protein physical property may not be straightforward, double-jump unfolding assays lead to a profile for the fraction at native state *versus* time and reveal immediately the time scale in which the native state is reached upon refolding *in vitro*. Such profiles are given in Figs 2a and b for the three thioredoxins studied. It is clear that folding *in vitro* of *CPk* thioredoxin is substantially slower than the folding of *E. coli* thioredoxin and LPBCA thioredoxin, as discussed above. Furthermore, the *in vitro* folding of *CPk* thioredoxin is also inefficient, as shown by the fact that substantial amounts of protein fail to reach the native state (Figure 2b). The effect is more pronounced the higher the total concentration of protein, suggesting that *in vitro* folding inefficiency is linked to protein aggregation (which is, in fact, observed visually). The rate constants derived using double-jump unfolding assays (closed symbols in the plots of Figure 2a) indicate that the native state is reached mostly in the slow kinetic phase detected by fluorescence.

While the focus of this work lies on folding rates and folding efficiency, it is also of interest to comment briefly on the unfolding rates since these are related with kinetic stability, an important protein property. *E. coli* thioredoxin is known to be a kinetically stable protein. This is clearly shown

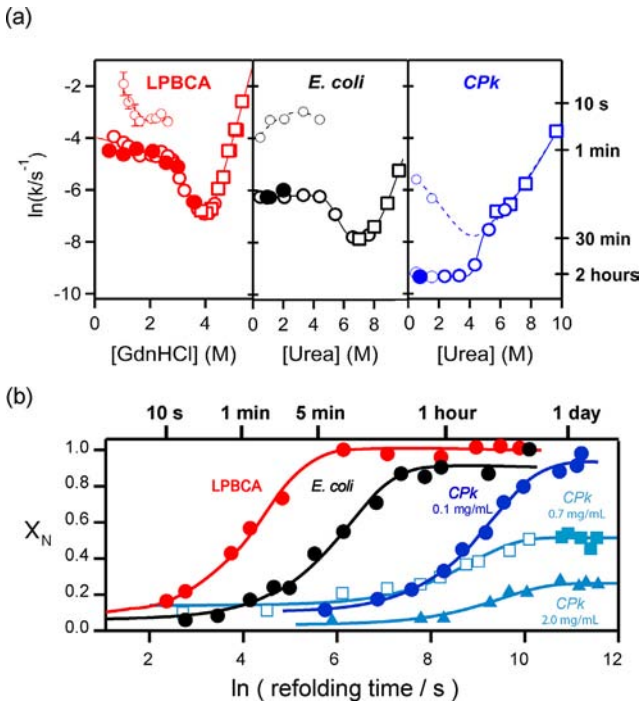


Figure 2. *In vitro* folding of modern and ancestral thioredoxins. **a** Chevron plots of folding-unfolding rate constant at pH 7 and 25 °C *versus* denaturant concentration for the three thioredoxins studied. Urea is used as denaturant for *E. coli* and *CPk* thioredoxins. LPBCA thioredoxin, however, is highly stable and cannot be denatured by urea at 25 °C. The chevron plot using the stronger denaturant guanidinium hydrochloride is shown for this protein. Still, the folding rates for LPBCA thioredoxin at low denaturant concentration obtained with urea and guanidine are in good agreement.[47] Circles and squares refer to experiments performed in the folding and unfolding directions, respectively. Error bars are standard errors derived from fits to the experimental profiles and are not shown when they are smaller than the size of the data point. Data are taken from Gamiz-Arco et al.[47] and were derived from fluorescence kinetic profiles. These profiles are often multiphasic in the folding direction, with the slow folding phase leading to the native state, as shown by the agreement with the folding rates (closed symbols) derived using double-jump unfolding assays. Lines shown are meant to guide eye. **b** Profiles of fraction of native state *vs.* time obtained by using double-jump unfolding assays. The refolding time is the time the protein is allowed to refold after the first jump, i.e. the time elapsed between the start of the folding process and its interruption (see Methods for details). Experiments were performed at pH 7, 25 °C in the presence of 1 M urea (see Methods for details). Protein concentration was 0.1 mg/mL, except for the profiles for *CPk* thioredoxin at 0.7 mg/mL and 2 mg/mL. The open symbols in the profile for *CPk* thioredoxin at 0.7 mg/mL are taken from Gamiz-Arco et al.[47]. The lines represent the best fits of a single exponential. Note that an exponential has a sigmoidal shape in a plot versus ln t. In both **a** and **b**, typical values of the lifetime (calculated as the inverse of the first-order rate constant) are indicated to highlight the large differences in folding time-scale between the three proteins.

by the extrapolation to zero denaturant concentration of the unfolding branch in the Chevron plot which leads to a very low unfolding rate constant and an unfolding time scale on the order of a few months.[67] The unfolding of LPBCA thioredoxin is about three orders of magnitude slower than the unfolding of *E. coli* thioredoxin,[51] indicating a much-enhanced kinetic stabilization for the ancestral protein. By contrast, the unfolding of the symbiont *CPk* thioredoxin is somewhat faster than the unfolding of *E. coli* thioredoxin (Figure 2a). Still, an extrapolation to zero denaturant concentration of the unfolding branch of the Chevron plot yields an unfolding time scale on the order of days. Therefore, although the symbiont thioredoxin is less kinetically stable than *E. coli* thioredoxin, it seems to retain some substantial level of kinetic stability. The differences in kinetic stability between the three thioredoxins may reflect, at least in part, the different living temperatures of the host organisms, as we have previously discussed.[68]

Efficiency of the heterologous expression in *E. coli* of modern and ancestral thioredoxins. We determined the efficiency of expression in *E. coli* at 37 °C as the ratio of soluble protein to total protein determined after overexpression for 3 h, as recommended by standard protocols (see Methods for details).

As was to be expected, expression of *E. coli* thioredoxin in *E. coli* is highly efficient, leading to

essentially 100% soluble protein. Remarkably, the expression of the ancestral LPBCA thioredoxin is also highly efficient, despite the extensive sequence differences with *E. coli* thioredoxin (only 57% sequence identity). On the other hand, expression of *CPk* thioredoxin in *E. coli* only leads to about 20% soluble protein (Figure 3).

Chaperone over-expression is a common host-engineering strategy to improve heterologous protein expression.[2,8] We attempted to rescue the inefficient folding of the *CPk* thioredoxin in *E. coli* by complementation with plasmids containing genes of the following *E. coli* chaperones: trigger factor, groES, groEL, dnaK, dnaJ. We used five separate combinations of these chaperones, as shown in Figure 3a. Only very moderate enhancements in folding efficiency were observed (Figure 3).

A plausible evolutionary narrative. The experiments described above (Figures 2 and 3) reveal a striking disparity between the folding behaviour of the *CPk* thioredoxin and LPBCA thioredoxin outside their original hosts. The ancestral protein folds fast *in vitro* and its expression in *E. coli* is efficient. By contrast, expression of the symbiont thioredoxin in *E. coli* produces a substantial amount of insoluble protein and refolding experiments show that it reaches the native state *in vitro* in the time scale of hours. This slow and inefficient folding can hardly be assumed to correspond to the situation *in vivo* in the original host. Note that the synthesis of a $\sim$ 100 residue protein by bacterial ribosomes takes about 5 seconds.[58] Obviously, it is difficult to understand that a small protein, which can in principle fold fast, has been selected during evolution to fold in its original host in a time scale $\sim$ 3 orders of magnitude above the time required for synthesis in the ribosome. A much more likely scenario is that folding of *CPk* thioredoxin in its original host is fast, plausibly allowing for co-translational folding, and efficient. This fast/efficient *in vivo* folding would obviously be the result of the interaction of the protein with the cellular folding assistance machinery, mainly the ribosome itself[69,70] and the ribosome-binding chaperones (the trigger factor), which would guide and assist co-translational folding, although a role for downstream chaperones and the specific environment in the symbiont (pH, redox, etc.) cannot be ruled out.

In fact, it is not unreasonable that the folding of thioredoxins in at least some modern organisms be assisted. A commonly accepted view is that folding is often inefficient and, therefore, it may require assistance, for proteins larger than 100 residues.[58] According to this, thioredoxins, with about 110 residues, would be a borderline case. However, thioredoxins have a serious folding problem that was already noted by Fred Richards many years ago[61]: the proline residue at position 76, which is essential for activity and strictly conserved, is in cis conformation, which generates a well-known kind of folding kinetic bottleneck. In fact, folding of *E. coli* thioredoxin in the test tube is known to be a complex process.[62] Overall, it appears plausible that thioredoxin folding *in vivo* may benefit from assistance and there can be little doubt that this is in fact the case with the symbiont thioredoxin we study in this work, since its unassisted folding outside its original host is seriously impaired. Such assistance, however, would not be available for *CPk* thioredoxin in the heterologous *E. coli* host, since co-evolution in the original host would lead to its adaptation to the folding assistance machinery of the symbiont. Note that, not only the symbiont thioredoxin, but also most of the symbiont chaperones and ribosomal proteins are highly divergent at the sequence level (see Tables S2 and S3). Therefore, co-evolution of interacting symbiont proteins is a likely scenario.

As we have recently noted,[47] since evolution has no foresight, folding assistance cannot have arisen before protein folding itself and, therefore, the most ancient proteins could plausibly fold with little or no assistance. That is, ancestral folding efficiency plausibly relied on fast unassisted folding that limits the transient population of aggregation-prone partially-unfolded states. Consequently, efficient folding of resurrected ancestral proteins in modern organisms may plausibly reflect an ancient adaptation to unassisted folding. Thioredoxins could in fact provide a clear example of this ancient unassisted-folding scenario, because there are thioredoxins in the three domains of life and their emergence can, therefore, be traced back to the last universal common ancestor, LUCA. Certainly, it is not known when efficient folding assistance emerged. Still, it is clear that efficient unassisted folding would no longer be a useful feature after the evolutionary emergence of cellular folding-assistance, which would thus allow the evolutionary acceptance of mutations that impair the ancestral feature. Such degradation of unassisted folding would be of no consequence for folding in the original host, but will lead to inefficient expression in a heterologous host, where folding assistance would not be available due to co-evolution in the original host. Furthermore, it is not clear how long it took for the ancient trait (efficient unassisted folding) to degrade after cellular folding assistance was available. One possibility is that substantial degradation of ancient unassisted-folding of small proteins, such as thioredoxin, may have only occurred to a substantial extent in organisms with a high evolutionary rate, such as the obligate symbiont we consider in this work.

To summarize, it appears plausible that the folding of the most ancient thioredoxins was unassisted, that the folding of *CPk* thioredoxin in its original host is assisted and that, as a result of co-evolution in the original host, efficient assistance is not available for the symbiont thioredoxin in the heterologous *E. coli* host.

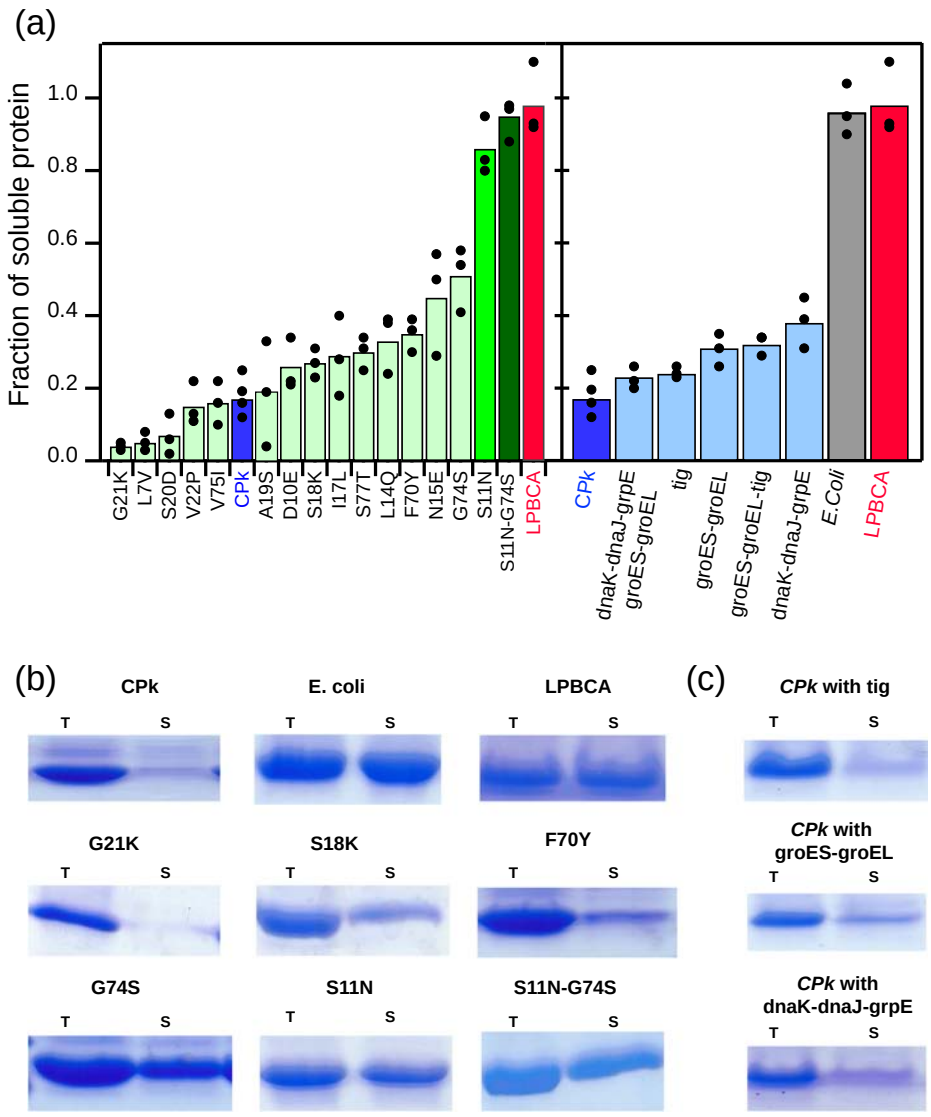


Figure 3. Expression in *E. coli* of modern and ancestral thioredoxins. **a** Fraction of soluble protein obtained upon expression of thioredoxin proteins in *E. coli* at 37 °C. The box at the left displays data for LPBCA thioredoxin, *CPk* thioredoxin, variants of the latter with single mutations and one double-mutant variant (S11N-G74S). The box at the right displays data for *E. coli* thioredoxin, LPBCA thioredoxin, *CPk* thioredoxin and data for the later with over-expression of various chaperone teams. Bars represent the average of several independent determinations and the individual values are also shown. **b** Illustrative examples of the experimental determination of the fraction of soluble protein for LPBCA thioredoxin, *E. coli* thioredoxin, *CPk* thioredoxin and variants of the latter. **c** Representative examples of the experimental determination of the fraction of soluble protein for *CPk* thioredoxin with over-expression of chaperone teams. In both **b** and **c**, T and S represent "total" and "soluble", respectively. For illustration, only sections of the SDS-PAGE gels with the thioredoxin bands are shown. Complete gels are shown Figures S10-S13.

Evolutionary narratives are necessarily speculative. Still, the narrative we propose has the merit of immediately suggesting an approach to the sequence-engineering of heterologous expression. That is, suitable back-to-ancestor mutations could lead to a more efficient heterologous expression. Furthermore, since folding in the heterologous host is, at least to some extent, unassisted, computational modelling of the unassisted folding-landscape may be used to guide back-to-ancestor engineering for efficient heterologous expression. Computational modelling of the folding landscape for thioredoxins is described in the next section.

Computational modelling of the folding landscape for modern and ancestral thioredoxins. Here we use a recently developed version[53] of the Wako-Saitô-Muñoz-Eaton (WSME) statistical model of protein folding[54–57] to assess the main features of the folding landscape of modern and ancestral thioredoxins (Figure 4).

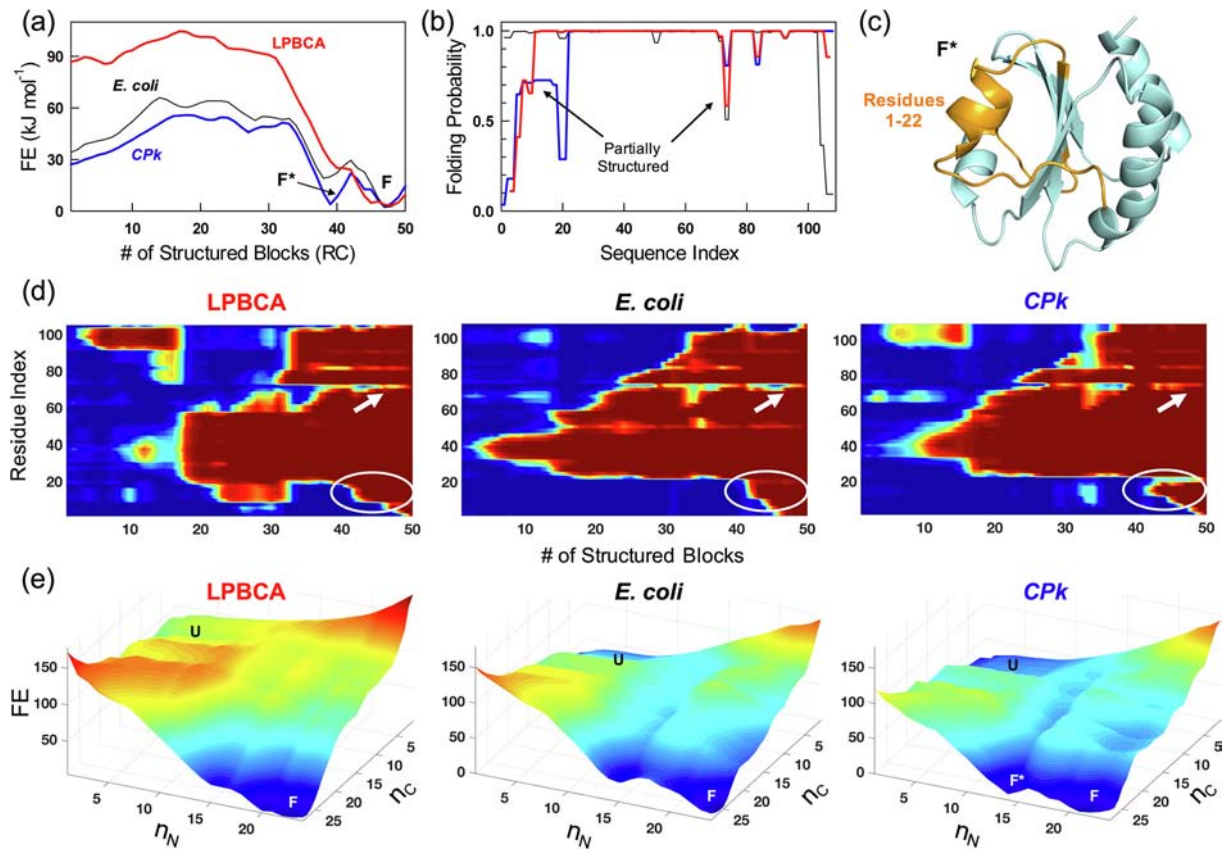


Figure 4. Statistical mechanical modelling of the folding landscape for modern and ancestral thioredoxins. A block version of the Wako-Saitô-Muñoz-Eaton model was used in all the calculations shown here. **a** Profiles of free energy versus number of structured blocks at 37 °C for the modern thioredoxins from *E. coli* and *Candidatus Photodesmus katoptron* (*CPk*), and the ancestral LPBCA thioredoxin. Note that a partially-unfolded intermediate (F* arrow), clearly differentiated from the fully folded protein (F), is distinctly observable only in *CPk* thioredoxin. **b** Residue folding probabilities as a function of sequence index at 37 °C following the colour code in panel **a**. **c** The predicted structure of F* with the partially structured residues 1–22 highlighted in orange. **d** Folding probability, coloured in the spectral scale from blue (0) to red (1), as a function of a plausible reaction coordinate, the number of structured blocks, for the three thioredoxin studied. The N-terminal region that folds the last is highlighted by white ovals while the 70–77 region is highlighted by an arrow. **e** Free energy landscapes (z-axis in $kJ \cdot mol^{-1}$) as function of n_N and n_C, the number of structured blocks in the N- and C-terminal half of the protein, respectively, for the three thioredoxins studied here.

As mentioned above, thioredoxin folding has been known from many years to be a complex process that involves intermediate states and parallel channels to arrive to the native state,[61,62] reflecting a rugged folding landscape. We do not aim here at reproducing these kinetic complexities in detail, but mostly at identifying regions of the thioredoxin molecule that are likely to be unfolded in intermediate states of the folding landscape. The rationale behind this approach is that such unfolded or partially-structured regions may be involved in aggregation and other undesirable interactions[58] and are, therefore, obvious targets for engineering efforts aimed at rescuing inefficient folding. This notion is consistent with early studies on the directed evolution of proteins for enhanced heterologous expression which demonstrated improvements in the stability and solubility of intermediates.[71]

We employ the block version of WSME model (the bWSME model), which considers 2–3 consecutive residues as blocks to reduce the protein phase space and thus rapidly calculate free energy profiles (see Methods). In the block description, the model still considers > 490,000 microstates compared to the residue-level version that would involve considering the contribution to the partition function from > 9.9 million microstates. We first reproduce the apparent experimental equilibrium stability differences to calibrate the model. The resulting one-dimensional

folding free energy profiles as a function of the number of structured blocks, the reaction coordinate (RC), are quite similar for the three proteins, but with one major difference – *CPk* thioredoxin populates a partially structured intermediate (F*) on the folding side of the main barrier with a population of 27% (Figure 4a). However, neither of LPBCA or *E. coli* thioredoxins populate F* significantly (<0.1%). It is important to note that this difference is not a consequence of the larger stability of *CPk* thioredoxin as *E. coli* thioredoxin does not populate F* despite exhibiting similar stability (black in Figure 4a). Moreover, WSME model calculations reveal that the shape of the free-energy profiles is conserved under iso-stability conditions (of 25 kJ mol^{-1}; Figure S2). These observations indicate that the intermediate F* is intrinsic to the *CPk* thioredoxin conformational landscape and is independent of the overall thermodynamic stability.

To identify the regions of *CPk* thioredoxin that are partially structured, we computed residue folding probabilities that quantify the extent to which every residue is structured at a given thermodynamic condition. At 37 °C, the N-terminal region of *CPk* thioredoxin (residues 1–22) is partially structured when compared to its ancestral counterpart (Figure 4b). Differences in folding probabilities are also evident in two other regions: residues 70–77 that harbours a critical cis-proline[47,61] and to a lesser extent in the residue stretch 83–84. We further calculated the probability of every region of the protein to be structured as a function of the reaction coordinate (Figure 4d), and find that the N-terminal region of the protein folds the last, thus revealing the identity of F* (Fig 4c and white ovals in Figure 4d). It can also be seen that the residue-stretch 70–77 exhibits equilibrium fluctuations during the folding for both the proteins (arrows in Figure 4d). The two-dimensional free energy landscape (Figure 4e), constructed by accumulating partial partition functions involving combinations of a given number of residues structured at N- and C-terminal halves of the protein, highlights that F* is the most populated state in *CPk* thioredoxin apart from multiple partially structured states that likely contribute to the slow folding of *CPk* thioredoxin (Figure 4e) compared to LPBCA thioredoxin.

There is a remarkable congruence between the computational predictions and the experimental folding kinetic data for *CPk* thioredoxin. Thus, a rate for reaching the native state much slower than expected from the typical shape of a chevron plot (compare continuous and discontinuous lines in the plot for *CPk* thioredoxin in Figure 2a) is the pattern expected from the accumulation of an intermediate. Furthermore, the structure predicted for F* (Figure 4c) is supported by results obtained with modern/ancestral chimeras described in the next section.

Efficiency of heterologous expression in *E. coli* of modern/ancestral thioredoxin chimeras. On the basis of the folding-landscape computations described in the preceding section, we selected two regions as targets for the engineering of heterologous folding: the N-terminal 1–22 fragment, which includes a short α-helix and large stretches of non-regular structure, and the 70–77 loop which includes a cis-proline residue that has been known for many years to be critical for thioredoxin folding.[47,61] These two regions are predicted to fold comparatively late and to remain unfolded during most of the protein folding process (Figure 4d). Plausibly, therefore, they may be involved in intermolecular processes that lead to insoluble protein. Since the heterologous expression of the ancestral LPBCA thioredoxin is efficient, we have prepared chimeras in which these regions are replaced by the corresponding ancestral sequences (see Figure 5a). These replacements involve 17 mutational changes in the case of the 1–22 fragment and 4 mutational changes in the case of the 70–77 loop. We designate the chimeras as *CPk*-[1-22] thioredoxin and *CPk*-[70-77] thioredoxin. We have also studied the "double chimera" in which both regions are simultaneously replaced by the corresponding ancestral sequences: *CPk*-[1-22]-[70-77] thioredoxin. The three chimeras show substantially improved heterologous expression in *E. coli* with *CPk*-[1-22]-[70-77] thioredoxin approaching 100% of soluble protein (Figure 5c and d).

In addition to the two regions referred to in the preceding paragraph, we have also selected for experimental analysis the 33–52 region which matches the longest α-helix in the thioredoxin molecule. This region is selected to provide an obvious control experiment, since it is predicted to fold early (Figure 4d) and, according to our working hypothesis, we do not expect its replacement with the corresponding ancestral sequence to improve the efficiency of heterologous folding. The experimental results on *CPk*-[33–52] thioredoxin conform to this expectation (Figure 5c and d).

Efficiency of heterologous expression of single-mutant variants of the symbiont thioredoxin. As described in the preceding section, efficient heterologous expression of the symbiont thioredoxin is achieved through replacement of the 1–22 and 70–77 regions with the corresponding ancestral sequences in LPBCA thioredoxin. To explore the individual mutational contributions to the rescue, we have prepared 15 back-to-ancestor, single-mutant variants. These include the four back-to-ancestor mutations in the 70–77 region and 11 back-to-ancestor mutations in the 1–22 region. We have excluded from this analysis the initial 1–6 N-terminal segment, since it appears to be unstructured in the native structures. For all the 15 single-mutant variants

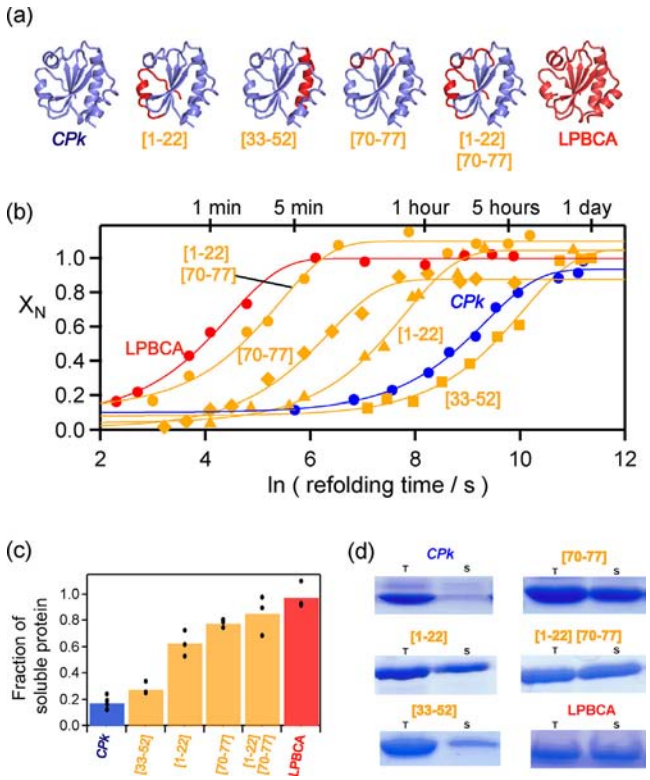


Figure 5. *In vitro* folding and expression in *E. coli* of modern/ancestral chimeras. **a** Definition and structural description of the studied modern/ancestral chimeras. The thioredoxin backbone is coloured to indicate the origin of the sequence: modern *CPk* thioredoxin (blue) or ancestral LPBCA thioredoxin (red). **b** Profiles of fraction of native state *versus* time for the *in vitro* folding of *CPk* thioredoxin, LPBCA thioredoxin and the several modern/ancestral chimeras defined in **a**. The values of the fraction of native state are derived from double-jump unfolding assays (see Methods for details). The plot is labelled with characteristic time values to highlight the wide range of folding times for the proteins studied. The lines represent the best fits of a single exponential. **c** Fraction of soluble protein obtained upon expression in *E. coli* at 37 °C of *CPk* thioredoxin, LPBCA thioredoxin and the four chimeras defined in **a**. Bars represent the average of several independent determinations and the individual values are also shown. **d** Representative examples of the experimental determination of the fraction of soluble protein for LPBCA thioredoxin, *CPk* thioredoxin and the modern/ancestral chimeras defined in **a**. Complete gels are shown Figures S10-S13.

we have determined the heterologous expression efficiency in *E. coli* (Figure3a). The most remarkable result of these studies is that a single mutation at the 1–22 segment, S11N, rescues most of the inefficient heterologous folding. Combining this mutation with the best-rescuing mutation in the 70–77 loop leads to a double mutant variant of the symbiont thioredoxin, S11N/G74S, that approaches 100% soluble protein.

Inefficient heterologous expression of CPk thioredoxin is rescued by back-to-the-most-ancient-ancestor mutations. It is important to note, first of all, that the inefficient heterologous folding of the symbiont *CPk* thioredoxin in *E. coli* is rescued by mutations that are back-to-ancestor, but not back-to-*E. coli.* The reason is obviously that the residues at position 11 and 74 in *E. coli* thioredoxin and *CPk* thioredoxin are identical: S and G, respectively[48] (see also Figure 1 in Ingles-Prieto et al.[49]). Furthermore, the N at position 11 and S at position 74 are very likely the residues present in the most ancient thioredoxins, as indicated by the posterior probabilities in the thioredoxins corresponding to the last common ancestor of bacteria (LBCA) and the last archaeal-eukaryotic common ancestor (AECA). The values are 0.999 (N at position 11, LBCA), 1.000 (S at position 74, LBCA), 0.991 (N at position 11, AECA) and 0.990 (S at position 74, AECA). As we have previously shown, sites with such high probability are very rarely, if ever, incorrectly predicted.[72] Our reason for providing the posterior probabilities at LBCA and AECA is that these nodes are immediately below the last universal common ancestor (LUCA) in the phylogeny used for ancestral reconstruction. The sequences of proteins in LUCA cannot be reconstructed by standard methodologies because an outgroup to determine LUCA in the phylogenetic tree is not available. Still, since N and S are with very high probability the residues at positions 11 and 74 in the nodes immediately below LUCA, it can be reasonably inferred that N and S were also the residues

at positions 11 and 74 in LUCA thioredoxin. Overall, even if ancestral reconstruction is unavoidable uncertain to some extent, there is little doubt about the identity of the amino acids in the most ancient thioredoxins at the crucial positions 11 and 74.

Correlation between heterologous folding efficiency and the *in vitro* folding rate. The modern-ancestral chimeras, the single-mutant variants and the double S11N/G74S variant studied here are all active and show levels of *in vitro* redox activity similar to the modern *CPk* thioredoxin and *E. coli* thioredoxin, as well as the ancestral LPBCA thioredoxin (Figure S1). They differ substantially, however, in terms of *in vitro* refolding rate. We have used double-jump unfolding assays to determine the rate constant for the last stage of *in vitro* refolding process, *i.e.,* the stage than leads to the native protein and defines the time scale of folding. These experiments reveal a very large (∼400-fold) range of folding rates (Figures 2b, 5b and 6) with the time scale in which these proteins reach the native state *in vitro* varying between a few minutes and many hours. Remarkably, there is a good correlation between the efficiency of heterologous folding and the folding rate *in vitro* (Figure 6a), supporting that efficient heterologous folding is achieved through the reduction of the time partially-unfolded states are significantly populated during the folding process.

Relation between heterologous folding efficiency and protein stability. We have used the denaturation temperature, as determined by differential scanning calorimetry, as a simple metric for the stability of the modern/ancestral chimeras, the single-mutant variants and the double S11N/G74S variant. For some selected variants, we have also performed urea denaturation studies. All the variants studied incorporate back-to-ancestor modifications. The ancestral sequence used as reference is that of LPBCA thioredoxin, a hyperstable protein with a very high denaturation temperature.[48,60] As anticipated, therefore, the back-to-ancestor modifications produce stability enhancements with respect to the *CPk* thioredoxin background in most cases. Furthermore, there appears to be a reasonable correlation between heterologous folding efficiency and stability, as described by the denaturation temperature values (Figure 6b).

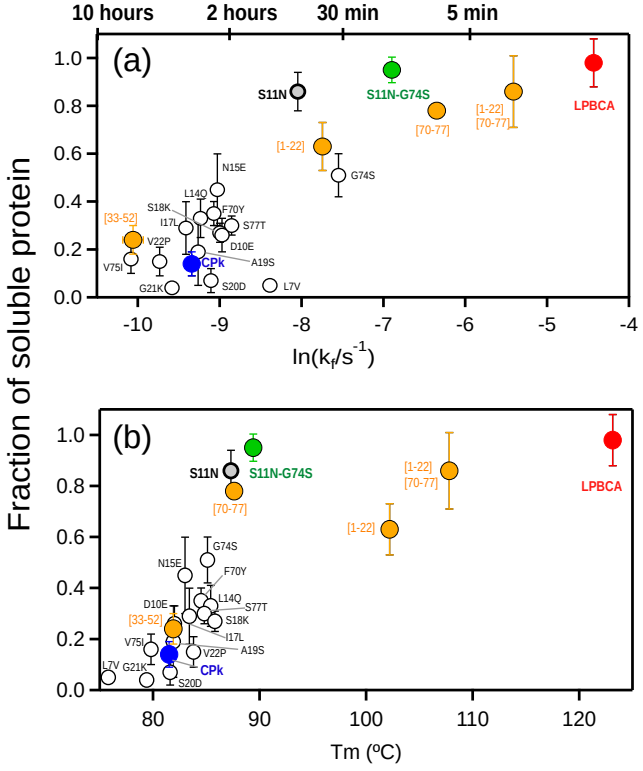


Figure 6. Correlations of the efficiency of heterologous expression with *in vitro* folding rate and protein stability. **a** Plot of fraction of soluble protein obtained in the expression in *E. coli versus* the logarithm of the *in vitro* folding rate constant including *CPk* thioredoxin, LPBCA thioredoxin, several variants of *CPk* thioredoxin and several modern/ancestral chimeras (Figure 5). Typical values of the lifetime are indicated to highlight the wide range of folding timescales. **b** Plot of fraction of soluble protein obtained in expression in *E. coli versus* denaturation temperature values derived from differential scanning calorimetry experiments. The proteins included here are the same as those included in the plot of panel **a**.

It is important to note, however, that global stability alone cannot explain the rescue of inefficient heterologous folding by back-to-ancestor modifications. This is more clearly seen when comparing the data of *CPk*-[1–22] thioredoxin with those for the variant of *CPk* thioredoxin with the single S11N mutation (Figure 7). Both scanning calorimetry data and chemical denaturation profiles indicate that replacement of the 1–22 segment in *CPk* thioredoxin with the corresponding LPBCA ancestral sequence brings about a very large stabilization, while the single S11N mutation produces a much more moderate stability enhancement (Figure 7a and b). Yet, heterologous folding of the single S11N variant is even more efficient than that of *CPk*-[1–22] thioredoxin (Figure 7c). Obviously, much of the stabilization brought about by the replacement of the 1–22 segment in *CPk* thioredoxin with its ancestral counterpart has little effect on heterologous folding efficiency.

The pattern described above for the heterologous folding efficiency is replicated by the *in vitro* folding rates. That is, the folding rate for the S11N variant of *CPk* thioredoxin is similar to the folding rate of the *CPk*-[1–22] thioredoxin (Figure 6a), despite of the much higher stability of the chimera as probed by the denaturation temperature value (Figure 6b). This result is not surprising, since stability enhancement does not necessarily lead to a faster folding rate. In fact, no significant correlation between stability and folding rate was found in an experimental analysis of a set 13 modern thioredoxins from different species.[47] As another example, the design of a completely symmetric β-trefoil led to a protein displaying both very high stability and very slow folding.[73,74] More generally, the uncoupling between thermodynamics and kinetics linked to the presence of populated intermediates is a known phenomenon, already noted in the seminal work of Agard and coworkers on α-lytic proteases.[75]

Overall, our results are consistent with the notion that protein stabilization may improve heterologous expression,[76] but support that the rescuing effect of stabilization is linked to specific mutations in regions of the protein that are likely unfolded in aggregation-prone intermediate states and that have a strong effect on the unassisted folding rate.

Finally, a somewhat intriguing result of our extensive experimental study on the stability of *CPk* thioredoxin variants deserves some attention. While most of the variants show the expected stability enhancement, this is not the case with *CPk*-[33–52] thioredoxin. This modern/ancestral chimera involves 10 back-to-ancestor amino acid replacements in the longest α-helix of the thioredoxin structure, which is expected to fold early according to our statistical-mechanical calculations (Figure 4). *CPk*-[33–51] thioredoxin displays a stability similar to that of the *CPk* thioredoxin background as shown by both the denaturation temperature values and the chemical denaturation profiles (Figure S3). One interesting possibility is that the ancestral stabilization is an adaptation to the need of efficient unassisted folding and high kinetic stability in an ancient environment. Consequently, it is not implemented in regions of the protein that are folded in aggregation-prone intermediates of the folding landscape and in the transition state that determines kinetic stability. Another possibility is that early folding of the long helix is crucial, mutations that impair its stability are not therefore accepted during evolution and, consequently, modern thioredoxins preserve the ancestral stability of the long helix. These interpretations are obviously speculative at this stage and will be explored in future work.

Structural basis of the rescue of inefficient heterologous folding. Inefficient heterologous expression of *CPk* thioredoxin in *E. coli* is rescued to a substantial extent by a single S11N mutation. Position 11 is part of a type IV turn involving residues 8–11 (Figure 8a). Turns are known to be crucial for protein folding in general, as they allow the polypeptide chain to fold onto itself and generate interactions that pertain to the native structure.[77] Presence of an asparagine residue at position 11 promotes the turn conformation because this residue can form stabilizing hydrogen bonds with the threonine at position 8 (Figure 8b). On the other hand, a serine at position 11 is not predicted to form stabilizing hydrogen bonds with the threonine at position 8 (Figure 8b) and consequently facilitates alternative conformations for the 8–11 segment in the high energy regions of the folding landscape. That is, the back to the ancestor mutation at position 11 should favour the local native conformation with respect local unfolded conformations, thus decreasing the time during which the corresponding partially unfolded states are significantly populated in the course of folding.

A similar explanation can be deduced for the effect of the G74S mutation included in the S11N/G74S variant that approaches 100% soluble protein in heterologous expression. As we have previously noted,[47] effects on folding of the G/S exchange at position 74 in thioredoxins are very likely related the fact that glycine has no side chain and places little restriction in local backbone conformation. The flexible link generated by the presence of a glycine residue will allow many different conformations in the high energy region of the folding landscape. This is particularly relevant for the 70–79 loop, since it also includes the proline residue at position 76, which is in the rare cis-conformer in the native structure.[61] Presence of a glycine residue at position 74 thus enables many conformations for the 70–79 loop that are not consistent with the native cis conformation for Pro76.

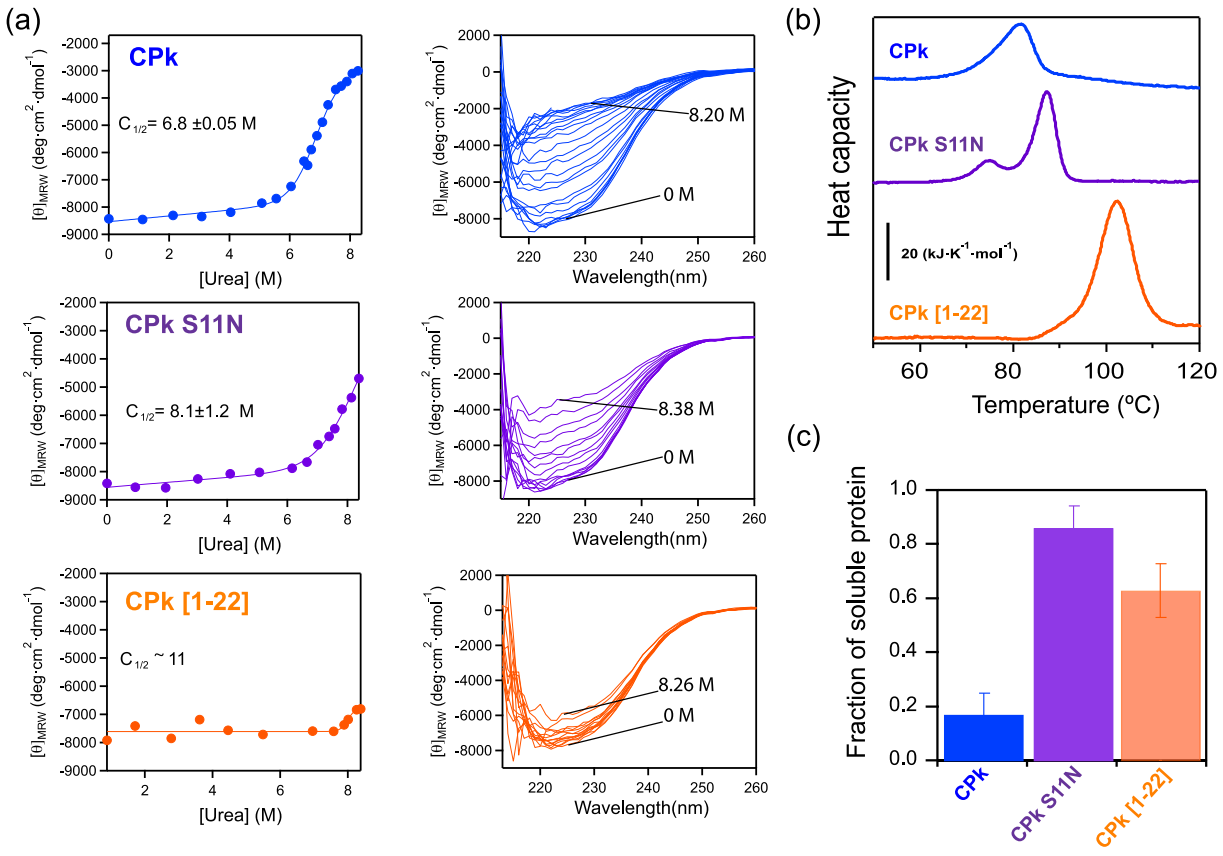


Figure 7. Relation between stabilization and rescue of inefficient heterologous expression. **a** and **b** Stability of the symbiont *CPk* thioredoxin, its S11N variant and the chimera in which the 1–22 segment has been replaced by the corresponding ancestral LPBCA sequence (*i.e., CPk*-[1–22] thioredoxin). In the experiments shown in **a** stability is probed by urea-induced denaturation followed by circular dichroism. Both, the original spectra at several urea concentrations and the profiles of ellipticity at 222 nm *versus* denaturant concentration are shown. The continuous lines represent the best fits of a two-state model (see Methods) and mid-point urea concentrations derived from the fits are shown. In **b** stability is studied by differential scanning calorimetry (see Methods and Figure S15 for details). While the mutation S11N has a significant but moderate stabilizing effect, the chimera is highly stable as revealed by a high denaturation temperature and resistance to denaturation at 25 °C by high urea concentrations. **c** Fraction of soluble protein obtained upon expression in *E. coli* at 37 °C of *CPk* thioredoxin, its S11N variant and the *CPk*-[1–22] thioredoxin chimera. Despite its much-enhanced stability, the chimera is less efficient at rescuing heterologous expression than the S11N variant.

Concluding remarks

Our results support that the folding of proteins in heterologous hosts may be akin to some extent to unassisted folding. For the specific protein system studied here, this is supported by (i) the success of the approach used to rescue inefficient heterologous expression, which involved computational modelling of the unassisted folding landscape, (ii) the fact that the efficiency of heterologous expression correlates with the *in vitro* folding rate, (iii) the very moderate rescue of inefficient heterologous expression by chaperone over-expression.

Unassisted folding in heterologous hosts may conceivably result from the overexpression of the protein exceeding the capacity of the folding-assistance machinery. This is a reasonable scenario, in particular since co-evolution may have led to the adaptation of the protein to the assistance machinery of its original host. Consequently, the fact that resurrected ancestral proteins often show improved heterologous expression as compared with their modern counterparts[22–35] plausibly reflects an ancient adaptation to unassisted folding. Efficient unassisted folding would no longer be a useful feature after the evolutionary emergence of cellular folding-assistance, thus allowing the evolutionary acceptance of mutations that impair the ancestral feature. Reversal of such mutations could then lead to a more efficient heterologous expression.

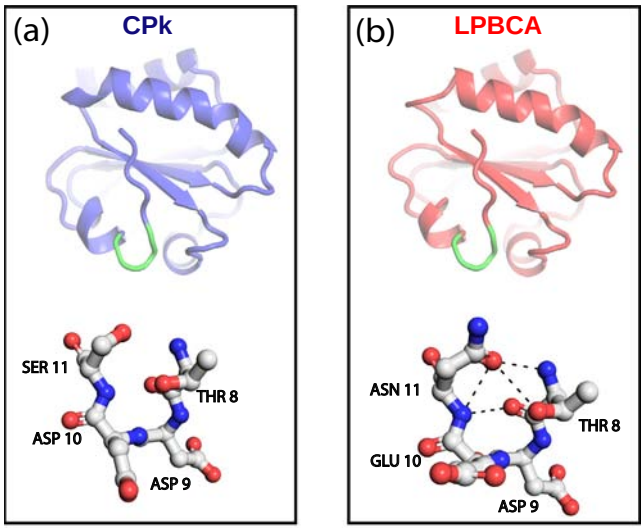


Figure 8. 3D-structures of the modern *CPk* thioredoxin and the ancestral LPBCA thioredoxin highlighting a type IV turn including position 11. Highlights of the turn show the hydrogen bonds involving the residue in position 11 with other residues in the turn as predicted by WHAT IF[94] with hydrogen-bond network optimization.[95] Hydrogen bonds are only predicted for the ancestral residue, which is therefore expected to stabilize the native turn conformation and disfavour non-native conformations.

Our results support that a few selected back-to-ancestor mutations can re-enact the folding efficiency of the resurrected ancestral proteins and point, therefore, to a minimal-perturbation, sequence-engineering approach to resolve inefficient heterologous expression. Some details of the practical application of the approach are noted below.

Prediction of back-to-ancestor mutations should be feasible for most protein systems, given the availability of large sequence databases and various software packages for the several steps of ancestral sequence reconstruction (for a recent account, see ref. [21]). Certainly, experimental screening for expression of all possible variants with single back-to-ancestor mutations may not be practical, in particular for large proteins. However, our results support that a limited screening of modern/ancestral chimeras may provide variants with enhanced expression. Furthermore, screening of individual mutations can be focused to protein regions that are expected to be unfolded in aggregation-prone intermediates populated during the folding process. Our results support that such regions can be predicted as the late-folding regions in folding landscape computations. Our version of the Wako-Saitô-Muñoz-Eaton statistical-mechanical model requires only a homology structure model as starting point and, most importantly, employs a block description to drastically reduce the number of microstates in the computation, thus allowing for a fast prediction even for large proteins. It is also worth noting that, for the protein system studied here, convincing molecular explanations can be put forward for the effect of the S11N and G74S mutations. This suggest the additional possibility of using rational design to determine the specific back-to-ancestor mutations that rescue inefficient heterologous folding.

Overall, our results open up the possibility of rescuing inefficient heterologous expression linked to low solubility by introducing a comparatively small number of back-to-ancestor mutations targeted to specific protein regions. In this way, the sequence would be minimally altered, and, likely, the properties of the encoded protein would be barely altered. This approach could be particularly useful in metagenomics studies that depend on sequence-based screening. Such studies identify enzymes having potential new activities of interest (pollutant or plastic degradation, for instance) based on the sequence similarity with known enzymes. In this scenario, achieving efficient expression with minimal sequence alteration will effectively contribute to the characterization of the new activity in the laboratory.

Methods

Expression and purification of thioredoxin variants. We followed procedures we have previously described in several publications[47–49,60] with small modifications. Briefly, genes encoding *Candidatus Photodesmus katoptron* thioredoxin (*CPk*) and the *CPk* chimeras (*CPk*[1–22] and *CPk*[33-52]) were synthesized with a His-tag at

the C-terminal and codon optimized for expression in *E. coli* cells. Mutations required for the single-mutant *CPk* variants (L7V, D10E, S11N, L14Q, N15E, I17L, S18K, A19S, S20D, G21K, V22P, F70Y, G74S, V75I, S77T), the double-mutant S11N/G74S variant and the chimera involving the loop[70–77] were introduced using the QuikChange Lighting Site-Directed Mutagenesis kit (Agilent Technologies) and the sequences were confirmed by DNA sequencing. Genes were cloned into pET24b(+) plasmid (GenScript Biotech) and transformed into *E. coli* BL21 (DE3) cells (Agilent). Protein expression was induced by 1 mM IPTG and cells were incubated overnight at 25 °C in LB medium and. Cell pellets were sonicated and His-tagged proteins were purified using affinity chromatography (HisGraviTrap column from GE Healthcare).

E. coli thioredoxin, LPBCA thioredoxin and the G74S variant of the later used in the experiments reported in this work were prepared following procedures similar to those described above, except that His-tags were not used. Therefore, these proteins were purified[49] by ion-exchange chromatography (Fractogel EMD DEAE column) followed by gel filtration chromatography on HiLoad Superdex 75 column. Our previous studies[47] indicate that the presence of a His-tag has a very small effect on the folding kinetic features of thioredoxins. Also, the presence of a His-tag does not have a significant effect on the efficiency of heterologous expression for *CPk* thioredoxin. The purification procedure based on ion-exchange chromatography and gel filtration was also used to prepare the non-His-tagged *CPk* thioredoxin used for crystallization (see further below).

Folding-unfolding experiments reported in this work were performed with thioredoxin solutions in 50 mM Hepes, pH 7. These solutions were prepared either by dialysis against the buffer at 4 °C or by passage through PD-10 desalting columns (GE Healthcare). Protein concentrations were measured spectrophotometrically using known values for the extinction coefficient. Guanidine and urea solutions in 50 mM HEPES, pH 7 were initially prepared by weight, but their concentrations were subsequently determined from refraction index measurements[78,79] using an Atago R500 hand refractometer. Urea solutions were purified by passage of the stock solution through an AG501-X8(D) ion-exchange resin (Bio-rad) before use.[79]

Double-jump unfolding assays. Double-jump unfolding assays, provide an estimate of the amount of native state in a protein solution.[47,63,64] They are based on the fact that the unfolding of the native state is much slower than the unfolding of intermediate (non-native and partially-unfolded) states. Therefore, the amount of native state in a protein solution can be estimated from the amplitude of the native-state unfolding kinetics observed upon transferring an aliquot to denaturing conditions, since intermediate states will unfold in a much shorter time-scale. Double-jump unfolding assays can be used to follow *in vitro* protein refolding kinetics (by performing the assays at several times during refolding), which provides an immediate assessment of the folding time-scale, *i.e.,* the time scale in which the folding polypeptide chain reaches the native state. This is particularly convenient when folding is a complex process involving multi-exponential kinetics and parallel kinetic channels, as is the case with thioredoxins.[61,62] We have recently described and discussed in some detail the use of double-jump unfolding assays to follow thioredoxin refolding kinetics.[47] Briefly, folding is initiated by transferring protein unfolded in a solution with a high denaturant concentration to a solution with a low denaturant concentration (first jump). At different times (t_1) after the first jump, aliquots are transferred to denaturing conditions (second jump) and the unfolding kinetics are followed by measuring fluorescence as a function of time (t_2). Plots of fluorescence intensity *versus* t_2 time after the second jump are well described by single exponentials, with lifetimes corresponding to the unfolding rate constant at the high denaturant concentration used (Figure S4). The amplitude of the exponential, however, varies reflecting the amount of native protein at the time (t_1) at which refolding was interrupted (representative examples are given in Figures S5-S8). After normalization with a suitable control amplitude, the amplitudes lead to a profile of fraction of native protein *versus* refolding time (t_1) (Figs. 2, 5 and S9). Specific details of the procedure used are discussed below.

In most cases, proteins were denatured in urea concentration within the range 7.5–9 M and, after fluorescence determinations had indicated that the unfolding process was essentially complete, the folding kinetics was initiated by dilution into typically 1 M urea, although some experiments at other final urea concentrations were performed, as shown in panel a of Figure 2). In most cases, the protein concentration in the folding kinetics experiments (i.e. after the first jump) was on the order 0.1 mg/mL, although additional experiments at higher protein concentrations were carried out with *CPk* thioredoxin (see Figure 2b). Typically, the second jumps involved a 1:15 dilution and the unfolding kinetics were determined by following the protein fluorescence at 350 nm as a function of time. The exact composition of the denaturing solution is immaterial for the result of the experiment, as long as the same composition is used in all the unfolding profiles corresponding to given folding experiment. Both, high urea concentrations (within the range 8–9.5 M) and high guanidine concentrations (within the range 3–5 M) were used. Of course, it is important that the denaturant concentration used does indeed unfolds the protein variant studied. In order to select denaturation concentrations that fulfill this

criterion, we determined the unfolding branches of Chevron plots for all the thioredoxins studied here (Figure S4). A few of the thioredoxins studied here are highly stable and they cannot be denatured by urea at 25 °C, not even using the highest urea concentrations experimentally available. In these cases, the initial unfolding step was performed in concentrated guanidine (within the range 3–4.5 M) and the dilution into native conditions was designed to ensure a low guanidine concentration in the folding kinetics experiment (within the range 0.1–0.3 M). Figures S5–S8 show several representative examples of the experiments we have described. Folding kinetic profiles for all the proteins studied here are shown in Figures 2b, 5b and S9.

In all cases, we carried out control experiments in which the native protein (at the same concentration used in the folding kinetic determinations) was transferred to the denaturing solution and the unfolding kinetics was followed by fluorescence at 350 nm. The amount of native protein at each time is then calculated as the ratio of the amplitude of the unfolding kinetics determined from the aliquot extracted at that time to the amplitude of the unfolding kinetics for the control. The resulting profiles of fraction of native state (X_N) versus time conformed to a single exponential. Note, however, that the kinetic profiles shown in Figures 2b, 5b and S9 use logarithm of time in the x-axis to highlight differences in folding time scale and that a single X_N *vs.* t exponential appears as sigmoidal in a plot of X_N *vs.* lnt.

Data of fraction of native protein *vs.* time were fitted with the following equation: $X_N = X_\infty+(X_0-X_\infty)\exp(-kt)$, where k is the first-order rate constant, and X_0 and X_∞ are short-time and long-time limiting values of the fraction of native protein. Note that X_0 and X_∞ do not necessarily equal zero and unity, respectively. Small differences with the control may certainly cause X_∞ to depart somewhat from unity. More importantly, thioredoxin folding is a complex process involving parallel kinetic channels and intermediate states.[61,62] A value of X_0 significantly higher than zero might reflect a that a fraction of the molecules reaches the native state in a shorter time scale than that probed by our experiments (*i.e.,* a fast folding kinetic channel). Likewise, a value of X_∞ significantly smaller than unity that a fraction of the molecules reaches the native state in a longer time scale than that probed in our experiments (*i.e.,* a slow folding kinetic channel). In practice, however, the values of X_0 and X_∞ determined from the fitting of the equation to our experimental folding profiles are reasonably close to 0 and 1 in essentially all cases. This implies that our experiments do identify the kinetic phase leading to the native state in the major folding channel. Therefore, we used as a metric of the folding time-scale the life-time calculated from the rate constant value derived from the fittings (*i.e.,* 1/k).

Protein solubility measurements. Solubility of overexpressed thioredoxins variants in *E. coli* BL21(DE3) strain was checked based on SDS-PAGE, following standard protocols. Briefly, at least 3 independent clones of each thioredoxin variant were grown up to an optical density of 0.6 and induced with 1 mM IPTG for 3 hours at 37 °C. A 90 mL aliquot of the final culture was centrifuged at 4000 rpm, 10 min at 4 °C and the collected pellet was re-suspended in 6 mL of lysis buffer containing 20 mM Tris, pH 7.5, 50 mM NaCl and a protease inhibitor tablet (Roche cOmplete™). After sonication, two aliquots were taken. One aliquot was subjected to SDS-PAGE to estimate the total amount of protein. Other aliquot was centrifuged (15000 rpm for 10 min at 4 °C) and the supernatant was subjected to SDS-PAGE to provide the amount of soluble protein. The SDS-PAGE gels obtained in this work are shown in Figures S10-S13.

Quantification of total and soluble thioredoxin fractions was carried out by SDS-PAGE on 15% Tris-glycine SDS-polyacrylamide gels and using ImageJ software (https://imagej.nih.gov/ij/) for image analysis of the thioredoxin bands stained by Coomassie dye. Illustrative densitometry profiles are given in Figure S14. At least, three independent measurements were performed for each protein variant. The average value, the standard deviation and the individual values are given in Figures 3 and 5.

In addition, we attempted to rescue the inefficient heterologous folding of *CPk* thioredoxin by co-overexpressing *E. coli* chaperones. Five plasmids designed to express the following "chaperone teams" were purchased from TAKARA Bio Inc: pG-KJE8 (expressing dnaK-dnaJ-grpE-groES-groEL), pGro7 (expressing groES-groEL), pKJE7 (expressing dnaK-dnaJ-grpE), pG-Tf2 (expressing groES-groEL-tig) and pTf16 (expressing the trigger factor). Chaperone plasmids were transformed into BL21(DE3) chemical competent cells containing the different thioredoxin plasmids.

Activity measurements. Activity of thioredoxin proteins was measured using the insulin turbidimetric assay[80] (Holmgren, 1979) as we have previously described.[47] Briefly, in this assay, disulfides reduction by dithiothreitol (DTT) catalysed by thioredoxin causes insulin aggregation, which is followed spectrophotometrically at 650 nm. The reaction mixture contains 0.1 M phosphate buffer pH 6.5, 2 mM EDTA, 0.5 mg/mL of bovine pancreatic insulin and a final thioredoxin concentration of 1.5 μM. The reaction is initiated by addition of DTT to a 1 mM final concentration. Activity values for each variant reported were obtained from the maximum value of plots of dA_{650nm}/dt *versus* time. A total of 3 independent measurements were carried out for each thioredoxin variant. The resulting

average values and the corresponding standard deviations are reported in Figure S1.

In addition, for some variants, thioredoxin activity was also assayed with thioredoxin reductase coupled to the reduction of DTNB[81] as we have previously described.[47] Final conditions in the cuvette were: 0.05 M Tris-HCl, 2 mM EDTA pH 8, 0.05 mg/mL BSA, 0.5 mM DTNB, 0.25 mM NADPH and 0.15 μM *CPk* variants. Reaction was started by addition of thioredoxin reductase to a final concentration of 0.02 μM and monitored spectrophotometrically. A total of three independent measurements were performed for each thioredoxin variant. The resulting average values and the corresponding standard deviations are reported in Figure S1.

Urea-induced equilibrium denaturation monitored by CD and fluorescence measurements. The urea-induced equilibrium denaturation of *CPk* thioredoxin, its S11N single mutant and *CPk*[1–22] chimera was studied by using far-UV circular dichroism measurements at 25 °C. Protein concentration was ~ 0.8 mg/mL in a 1 mm cuvette. The urea dependence of ellipticity at 222 nm could be adequately fitted by a two-state model that assumes a linear dependence of the unfolding free energy with denaturant concentration within the transition region and linear pre- and post-transition baselines, as previously described.[47] Values of the midpoint urea concentration ($C_{1/2}$) and the slope of the urea-dependence of the unfolding free energy (m) derived from these fits are given in Figure 7.

Denaturation Temperature measured by Differential Scanning Calorimetry. The denaturation temperatures of wild *CPk* thioredoxin and single- and double-mutant variants (L7V, D10E, S11N, L14Q, N15E, I17L, S18K, A19S, S20D, G21K, V22P, F70Y, G74S, V75I, S77T, S11N/G74S) and modern/ancestral chimeras (*CPk* [1-22], *CPk* [70-77], *CPk* [1-22]/[70-77], *CPk* [33-52]) were determined by differential scanning calorimetry as the temperature for maximum of the calorimetric transition. Experimental DSC thermograms are shown in Figure S15. Note that, in a few cases, two transitions were reproducibly seen in the thermograms, perhaps reflecting a decreased unfolding cooperativity upon mutation. In these cases, the maximum of the major transition was used as a metric of thermal stability. The experiments were performed with a MicroCal Auto-PEAQ DSC calorimeter (Malvern), at pH 7 in 50 mM HEPES buffer. Typically, protein concentration was within the 0.4–0.6 mg/mL range and scan rate was 240 K/h. Standard protocols well established in our laboratory for thioredoxins were followed.[48,82] In all cases, protein solutions for the calorimetric experiments were prepared by exhaustive dialysis and the buffer from the last dialysis step was used in the reference cell of the calorimeter. Calorimetric cells were kept under excess pressure to prevent degassing during the scan. Several buffer–buffer baselines were recorded to ensure proper calorimeter equilibration prior to the protein run.

Statistical-mechanical modeling of the folding landscape. The WSME model approach, its variants and parameterization are described elsewhere in detail.[57] Briefly, the model is Gō-like in its energetics requiring a starting structure and coarse-grains the polymer at the residue level. Thus, every residue is assumed to sample two sets of conformations – folded (represented as binary variable *1*) and unfolded (*0*) – contributing to a set of 2^N microstates for a *N*-residue protein. In the variant of the model used in the current work, we make two approximations to the original version.[54,56] First, we consider microstates restricted to only single-stretches of folded residues (single sequence approximation, SSA), two-stretches of folded residues (double sequence approximation, DSA) with no interaction across the island, and DSA with interaction across the structured islands if they interact in the native structure and even if the intervening residues are unfolded.[83] Second, we further reduce the accessible phase space by considering 2–3 consecutive residues to fold and unfold together (*i.e.* as a single block or unit). This model has been validated against the residue-level approximations and enables rapid predictions.[53] In addition, we include contributions from van der Waals interactions, all-to-all electrostatics,[57] simplified solvation terms and sequence and structure specific conformational entropy. At the time the simulations were performed, the 3D structure of *CPk* thioredoxin was not available. Therefore, homology modeling was employed to generate a model using the Robetta server[84] and the *E. coli* thioredoxin as a template. Still, the alpha carbon RMSD between the homology model and the experimental structure is 0.524 Å.

All simulations were performed at 37 °C, pH 7.0 and 0.1 M ionic strength conditions with the following parameters: atomic-level interaction energy (ξ) of −70 J mol^{-1} for every heavy atom van der Waals interaction identified with a 6 Å spherical cut-off and excluding the nearest neighbours for LPBCA and *CPk* thioredoxin (−76 J mol^{-1} for *E. coli* thioredoxin), change in heat capacity (ΔC_p^{cont}) of −0.36 J mol^{-1} K^{-1} per native contact, entropic penalty for fixing non-proline, non-glycine and residues in well-determined secondary structures (ΔS^{conf}) as −16.5 J mol^{-1} K^{-1} per residue, entropic penalty for glycine and coil residues as –22.56 J mol^{-1} K^{-1} per residue (accounting for the excess disorder in these regions,[85] and 0 J mol^{-1} K^{-n} per residue for proline residues given the limited backbone flexibility of prolines. Free energy profiles and surfaces as a function of the reaction coordinate, the number of structured blocks, are generated by accumulating partial partition functions corresponding to specific number of folded units.

Residue folding probabilities are calculated by summing up the probabilities of states in which the residue of interest is structured.

Crystallization and structural determination. Freshly purified *CPk*[70–77] thioredoxin was concentrated to 25 mg/ml prior setting the crystallization screening. Initial crystallization trials were carried out using the hanging-drop vapor diffusion method. Drops were prepared by mixing 1 μL of protein solution with the reservoir in a 1:1 ratio, and equilibrated against 500 μL of each precipitant cocktail of the HR-I & PEG/Ion™ crystallization screening (Hampton Research). Crystallization trials were kept at 293 k in an incubator. After one-week crystalline material was observed in conditions #33, #34, #43 and #11 of the HR-I kit and #13 of the PEG/Ion screening kit. Crystals were fished from the drop and transferred to the cryo-protectant solution prepared with the mother liquid supplemented with 15% (v/v) glycerol and subsequently flash-cooled in liquid nitrogen and stored until data collection.

Crystals were diffracted at ID23-1 beam-line of the European Synchrotron Radiation Facility (ESRF), Grenoble, France. The best diffracting crystals were obtained in condition #43 of the HR-I. The diffraction data were indexed and integrated using XDS[86] and scaled with SCALA from the CCP4 suite.[87] Thioredoxin crystals belonged to the $I2_13$ space group with only two monomers in the asymmetric unit and therefore with a usually high water content, almost 74%, as determined from Matthews´ coefficient, 4.71.[88] The molecular replacement solution was found using Molrep[89] and the coordinates of the PDB ID. 2TRX, chain A, locating the two monomers in the asymmetric unit. Refinement was done with phenix.refine[90] including manual building and water inspection with Coot[91] and using the Titration-Libration-Screw (TLS)[92] grouping since the initial steps. Model quality was checked using MolProbity[93] implemented within the Phenix suite.[90] Refinement statistics and quality indicators of the final model are summarized in Table S1. Coordinates and structure factors have been deposited at the PDB with accession code 7OOL.

Hydrogen bond analysis was done using WHAT IF[94] with hydrogen-bond network optimization.[95]

Acknowledgements

This work was supported by Human Frontier Science Program Grant RGP0041/2017 (J.M.S.-R. and E.A.G.), National Science Foundation Award #2032315 (E.A.G.), National Institutes of Health Award #R01AR069137 (E.A.G.), Department of Defense MURI Award #W911NF-16-1-0372 (E.A.G.), Spanish Ministry of Science and Innovation/FEDER Funds Grants RTI-2018-097142-B-100 (J.M.S.-R.) and BIO2016-74875-P (J.A.G.) and the Science, Engineering and Research Board (SERB, India) Grant MTR/2019/000392 (A.N.N.). We are grateful to the European Synchrotron Radiation Facility (ESRF), Grenoble, France, for the provision of time and the staff at ID23-1 beamline for assistance during data collection.

Author contributions

J.M.S.R. designed the research. G.G.-A. purified the modern/ancestral chimeras and the thioredoxin variants; she also performed and analysed the experiments aimed at determining their folding kinetics and biomolecular properties. V.A.R. performed experiments addressed at determining the efficiency of heterologous expression and provided essential input for the molecular interpretation of mutational effects on expression efficiency. E.A.G. provided essential input for the evolutionary interpretation of the data. J.A.G. determined the X-ray structure of the symbiont protein and provided essential input regarding its interpretation and implications. A.N.N. performed the computational simulations of the folding landscape for thioredoxins and provided essential input regarding their engineering implications. B.I.M. and J.M.S.-R. directed the project. J.M.S.-R. wrote the first draft of the manuscript to which V.A.R., J.A.G., A.N.N. and B.I.M. added crucial paragraphs and sections. All authors discussed the manuscript, suggested modifications and improvements, and contributed to the final version.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2021.167321.

Received 17 August 2021;
Accepted 17 October 2021;
Available online 21 October 2021

Keywords:
ancestral sequence reconstruction;
computational modelling of protein folding landscapes;
heterologous protein expression;
proteins from uncultured organisms;
obligate symbionts

† These authors contributed equally.

References

1. Walsh, G., (2010). Post-translational modifications of protein biopharmaceuticals. *Drug Discovery Today* **15**, 773–780.
2. Baeshen, M.N., Al-Hejin, A.M., Bora, R.S., Ahmed, M.M. M., Ramadan, H.A.I., Saini, K.S., Baeshen, N.A., Redwan, E.M., (2015). Production of biopharmaceuticals in *E. coli*: current scenario and future perspectives. *J. Microbiol. Biotechnol.* **25**, 953–962.
3. Tripathi, N.K., Shrivastava, A., (2019). Recent developments in bioprocessing of recombinant proteins: expression hosts and process development. *Front. Bioeng. Biotechnol.* **7**, 420.
4. Uchiyama, T., Miyazaki, K., (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* **20**, 616–622.
5. Calderon, D., Peña, L., Suarez, A., Villamil, C., Ramirez-Rojas, A., Anzola, J.M., Garcia-Betancur, J.C., Cepeda, M. L., Uribe, D., Del Portillo, P., Mongui, A., (2019). Recovery and functional validation of hidden soil enzymes in metagenomic libraries. *MicrobiolOpen* **8**, e572
6. Daniel, R., (2005). The metagenomics of soil. *Nat. Rev. Microbiol.* **3**, 470–478.
7. Baneyx, F., Mujacic, M., (2004). Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.* **22**, 1399–1408.
8. Selas Castiñeiras, T., Williams, S.G., Hitchcock, A.G., Smith, D.S., (2018). *E. coli* strain engineering for the production of advanced biopharmaceutical products. *FEMS Microbiol. Lett.* **365**, 15.
9. I. Acebrón, L. Plaza-Vinuesa, B. de las Rivas, R. Muñoz, J. Cumella, F. Sánchez-Sancho, J.M. Mancheño, Structural basis of the substrate specificity and instability in solution of a glycosidase from Lactobacillus plantarum. Biochim. Biophys. Acta Proteins Proteom. 1865 (2017) 1227-1236.
10. Pauling, L., Zuckerkandl, E., (1963). Chemical paleogenetics. Molecular “restoration studies” of extinct forms of life. *Acta Chem. Scan.* **17S**, 9–16.
11. Liberles, D.A., (2007). Ancestral Sequence Reconstruction. Oxford University Press, Oxford.
12. Benner, S.A., Sassi, S.O., Gaucher, E.A., (2007). Molecular paleoscience: systems biology from the past. *Adv. Enzymol. Relat. Areas Mol. Biol.* **75**, 1–132.
13. Hochberg, G.K.A., Thornton, J.W., (2017). Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* **46**, 247–269.
14. Gumulya, Y., Gillam, E.M., (2017). Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the “retro” approach to protein engineering. *Biochem. J.* **474**, 1–19.
15. S.D. Copley, Setting the stage for evolution of a new enzyme. Curr. Opin. Struct. Biol. 69 (2021) 41-49.
16. Selberg, A.G.A., Gaucher, E.A., Liberles, D.A., (2021). Ancestral sequence reconstruction: from chemical paleogenetics to maximum likelihood algorithms and beyond. *J. Mol. Evol.* **89**, 157–164.
17. Cole, M.F., Gaucher, E.A., (2011). Exploring models of molecular evolution to efficiently direct protein engineering. *J. Mol. Evol.* **72**, 193–203.
18. Risso, V.A., Gavira, J.A., Mejia-Carmona, D.F., Gaucher, E.A., Sanchez-Ruiz, J.M., (2013). Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.* **135**, 2899–2902.
19. Risso, V.A., Sanchez-Ruiz, J.M., Ozkan, S.B., (2018). Biotechnological and protein engineering implications of ancestral protein resurrection. *Curr. Opin. Struct. Biol.* **51**, 106–115.
20. Trudeau, D.L., Tawfik, D.S., (2019). Protein engineers turned evolutionist – the quest for the optimal starting point. *Curr. Opin. Struct. Biol.* **60**, 46–52.
21. M.A. Spence, J.A. Kaczmarski, J.W. Saunders, C.J. Jackson, Ancestral sequence reconstruction for protein engineers. Curr. Opin. Struct. Biol. 69 (2021) 131-141.
22. Gonzalez, D., Hiblot, J., Darbinian, N., Miller, J.C., Gotthard, G., Shohreh, A., Chabriere, E., Elias, M., (2014). Ancestral mutations as a tool for solubilizing proteins: the case of a hydrophobic phosphate-binding protein. *FEBS Open Bio* **4**, 121–127.
23. Withfield, J.H., Zhang, W.H., Herde, M.K., Clifton, B.E., Radziejewski, J., Janovjak, H., Henneberger, C., Jackson, C.J., (2015). Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci.* **24**, 1412–1422.
24. Trudeau, D.L., Kaltenbach, M., Tawfik, D.S., (2016). On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.* **33**, 2633–2641.
25. Zakas, P.M., Brown, H.C., Knight, K., Meeks, S.L., Gaucher, E.A., Doering, C.B., (2017). Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat. Biotechnol.* **35**, 35–37.
26. Manteca, A., Schönfelder, J., Alonso-Caballero, A., Fertin, M.J., Barruetabeña, N., Faria, B.F., Herrero-Galán, E., Alegre-Cebollada, J., De Sancho, D., Perez-Jimenez, R., (2017). Mechanochemical evolution of the giant muscle protein titin as inferred from resurrected proteins. *Nat. Struct. Mol. Biol.* **24**, 652–657.
27. Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R., Damborsky, J., (2017). Ancestral haloalkane dehalogenases show robustness and unique substrate specificity. *ChemBioChem* **18**, 1448–1456.
28. Koblan, L.W., Doman, J.L., Wilson, C., Levy, J.M., Tay, T., Newby, G.A., Maianti, J.P., Raguram, A., Liu, D.R., (2018). Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846.
29. Hendrikse, N.M., Charpentier, G., Nordling, E., Syrén, P.-O., (2018). Ancestral diterpene cyclases show increased thermostability and substrate acceptance. *FEBS J.* **285**, 4660–4673.
30. Gomez-Fernandez, B.J., Garcia-Ruiz, E., Martin-Diaz, J., Gomez de Santos, P., Santos-Moriano, P., Plou, F.J., Ballesteros, A., Garcia, M., Rodriguez, M., Risso, V.A., Sanchez-Ruiz, J.M., Whitney, S.M., Alcalde, M., (2018). Directed *–in vitro-* evoliution of Precambrian and extant Rubiscos. *Sci. Rep.* **8**, 5532.
31. Barruetabeña, N., Alonso-Lerma, B., Galera-Prat, A., Joudeh, N., Barandiaran, L., Aldazabal, L., Arbulu, M., Alcvalde, M., De Sancho, D., Gavira, J.A., Carrion-Vazquez, M., Perez-Jimenez, R., (2019). Resurrection of efficient Precambrian endoglucanases for lignocellulosic biomass hydrolysis. *Commun. Chem.* **2**, 76.
32. Nakano, S., Minamino, Y., Hasebe, F., Ito, S., (2019). Deracemization and stereoinversion to aromatic D-amino acid derivatives with ancestral L-amino acid oxidase. *ACS Catal.* **9**, 10152–10158.

33. Gomez-Fernandez, B.J., Risso, V.A., Rueda, A., Sanchez-Ruiz, J.M., Alcalde, M., (2020). Ancestral resurrection and directed evolution of fungal mesozoic laccases. *Appl. Environ. Microbiol.* **86**, e0078–e120.
34. Li, D., Damry, A.M., Petrie, J.R., Vanhercke, T., Singh, S. P., Jackson, C.J., (2020). Consensus mutagenesis and ancestral reconstruction provide insight into the substrate specificity and evolution of the from-end Δ6-desaturase family. *Biochemistry* **59**, 1398–1409.
35. Sun, Y., Calderini, E., Kourist, R., (2021). A reconstructed common ancestor of the fatty acid photo-decarboxylase clade shows photo-decarboxylation activity and increased thermostability. *ChemBioChem* **22**, 1833–1840.
36. Nicoll, C.R., Bailleul, G., Fiorentini, F., Mascotti, M.L., Fraaije, M.W., Matevi, A., (2020). Ancestral sequence reconstruction unveils the structural basis of function in mammalian FMOs. *Nat. Struct. Mol. Biol.* **27**, 14–24.
37. Schriever, K., Saez-Mendez, P., Rudraraju, R.S., Hendrikse, N.M., Hudson, E.P., Biundo, A., Schnell, R., Syrén, P.O., (2021). Engineering of ancestors as a tool to elucidate structure, mechanism, and specificity of extant terpene cyclase. *J. Am. Chem. Soc.* **143**, 3794–3807.
38. Ufarté, L., Laville, É., Duquesne, S., Potocki-Veronese, G., (2015). Metagenomics for the discovery of pollutant degrading enzymes. *Biotechnol. Adv.* **33**, 1845–1854.
39. Moran, N.A., (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. U. S.A.* **93**, 2873–2878.
40. Woolfit, M., Bromhan, L., (2003). Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol. Biol. Evol.* **20**, 1545–1555.
41. Holmgren, A., (1985). Thioredoxin. *Annu. Rev. Biochem.* **54**, 237–271.
42. Morin, J.G., Harrington, A., Nealson, K., Krieger, N., Baldwin, T.O., Hastings, J.W., (1975). Light for all reasons: versatility in the behavioural repertoire of the flashlight fish. *Science* **190**, 74–76.
43. Hendry, T.A., de Wet, J.R., Dougan, K.E., Dunlap, P.V., (2016). Genome evolution of the obligate but environmentally active luminous symbionts of flashlight fish. *Genome Biol. Evol.* **8**, 2203–2213.
44. Hendry, T.A., Dunlap, P.V., (2011). The uncultured luminous symbiont of *Anomalops katoptron* (Beryciformes: Anomalopidae) represents a new bacterial genus. *Mol. Phylogenet. Evol.* **61**, 834–843.
45. Hendry, T.A., Dunlap, P.V., (2014). Phylogenetic divergence between the obligate luminous symbionts of flashlight fish demonstrates specificity of bacteria to host genera. *Environ. Microbiol. Rep.* **6**, 331–338.
46. Hendry, T.A., de Wet, J.R., Dunlap, P.V., (2014). Genomic signatures of obligate host dependence in the luminous bacterial symbiont of a vertebrate. *Environ. Microbiol.* **16**, 2611–2622.
47. G. Gamiz-Arco, V.A. Risso, A.M. Candel, A. Inglés-Prieto, M.L. Romero-Romero, E.A. Gaucher, J.A. Gavira, B. Ibarra-Molero, J.M. Sanchez-Ruiz, Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted folding. Biochem. J. 476 (2019) 3631-3647.
48. Perez-Jimenez, R., Ingles-Prieto, A., Zhao, Z.M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Manyes, S., Kappock, T.J., Tanokura, M., Holmgren, A., Sanchez-Ruiz, J.M., Gaucher, E.A., Fernandez, J.M., (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* **18**, 592–596.
49. Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J.M., Gaucher, E.A., Sanchez-Ruiz, J.M., Gavira, J.A., (2013). Conservation of protein structure over four billion years. *Structure* **21**, 1690–1697.
50. V.A. Risso, F. Manssour-Triedo, A. Delgado-Delgado, R. Arco, Barroso-delJesus, A. Ingles-Prieto, R. Godoy-Ruiz, J.A. Gavira, E.A. Gaucher, B. Ibarra.Molero, J.M. Sanchez-Ruiz, Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. Mol. Biol. Evol. 32 (2015) 440-455
51. Candel, A.M., Romero-Romero, M.L., Gamiz-Arco, G., Ibarra-Molero, B., Sanchez-Ruiz, J.M., (2017). Fast folding and slow unfolding of a resurrected Precambrian protein. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4122–E4123.
52. Rosano, G.L., Ceccarelli, E., (2014). Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* **5**, 172.
53. Gopi, S., Aranganathan, A., Naganathan, A.N., (2019). Thermodynamics and folding landscapes of large proteins from a statistical mechanical model. *Curr. Res. Struct. Biol.* **1**, 6–12.
54. Wako, H., Saitô, N., (1978). Statistical mechanical theory of the protein conformation. 2. Folding pathway for protein. *J. Phys. Soc. Jpn.* **44**, 1939–1945.
55. Muñoz, V., Thompson, P.A., Hofrichter, J., Eaton, W.A., (1997). Folding dynamics and mechanism of β-hairpin formation. *Nature* **390**, 196–199.
56. Muñoz, V., Eaton, W.A., (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 11311–11316.
57. Naganathan, A.N., (2012). Predictions from an Ising-like statistical mechanical model on the dynamic and thermodynamic effects of protein surface electrostatics. *J. Chem. Theory Comput.* **8**, 4646–4656.
58. Balchin, D., Hayer-Hartl, M., Hartl, F.U., (2016). In vivo aspects of protein folding and quality control. *Science* **353**, aac4354.
59. Katti, S.K., LeMaster, D.M., Eklund, H., (1990). Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.* **212**, 167–184.
60. Romero-Romero, M.L., Risso, V.A., Martinez-Rodriguez, S., Ibarra-Molero, B., Sanchez-Ruiz, J.M., (2016). Engineering ancestral protein hyperstability. *Biochem. J.* **473**, 3611–3620.
61. Kelley, R.F., Richards, F.M., (1987). Replacement of proline-76 with alanine eliminates the slowest kinetic phase in thioredoxin folding. *Biochemistry* **26**, 6765–6774.
62. Georgescu, R.E., Li, J.H., Goldberg, M.E., Tasayco, M.L., Chaffotte, A.F., (1998). Proline isomerization-independent accumulation of an early intermediate and heterogeneity of the folding pathway of mixed α/β protein, *Escherichia coli* thioredoxin. *Biochemistry* **37**, 10286–10297.
63. Mücke, M., Schmidt, F.X., (1994). A kinetic method to evaluate the two-state character of solvent-induced protein denaturation. *Biochemistry* **33**, 12930–12935.
64. Ibarra-Molero, B., Sanchez-Ruiz, J.M., (1997). Are there equilibrium intermediates in the urea-induced unfolding of hen-egg-white lysozyme. *Biochemistry* **36**, 9616–9624.

65. Brandts, J.F., Halvorson, H.R., Brennan, M., (1975). Consideration of the possibility that the slow step on protein denaturation reactions is due to cis-trans isomerism of proline residues. *Biochemistry* **14**, 4953–4963.
66. Schmid, F.X., Baldwin, R.L., (1978). Acid catalysis of the formation of the slow-folding species of RNase A: evidence that the reaction if proline isomerization. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4764–4768.
67. Godoy-Ruiz, R., Ariza, F., Rodriguez-Larrea, D., Perez-Jimenez, R., Ibarra-Molero, B., Sanchez-Ruiz, J.M., (2006). Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments. *J. Mol. Biol.* **362**, 966–978.
68. Romero-Romero, M.L., Risso, V.A., Martinez-Rodriguez, S., Gaucher, E.A., Ibarra-Molero, B., Sanchez-Ruiz, J.M., (2016). Selection for protein kinetic stability connects denaturation temperaturas to organismal temperaturas and provides clues to Archaean life. *PLoS ONE* **11**, e0156657
69. Kaiser, C.M., Goldman, D.H., Chodera, J.D., Tinoco, I., Bustamante, C., (2011). The ribosome modulates nascent protein folding. *Science* **334**, 1723–1727.
70. Samelson, A.J., Bolin, E., Costello, S.M., Sharma, A.K., O'Brien, E.P., Marqusee, S., (2018). Kinetic and structural comparison of a protein's cotranslational folding and refolding pathways. *Sci. Adv.* **4**, eaas9098.
71. Roodveldt, C., Aharoni, A., Tawfik, D.S., (2005). Directed evolution of proteins for heterologous expression and stability. *Curr. Opin. Struct. Biol.* **15**, 50–56.
72. Randall, R.N., Radford, C.E., Roof, K.A., Natarajan, D.K., Gaucher, E.A., (2016). An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat. Commun.* **7**, 12847.
73. Broom, A., Doxey, A.C., Lobsanov, Y.D., Berthin, L.G., Rose, D.R., Howell, P.L., McConkey, B.J., Meiering, E.M., (2012). Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric globular protein. *Structure* **20**, 161–171.
74. Broom, A., Ma, M., Xia, K., Rafalia, H., Trainor, K., Colón, W., Gosavi, S., Meiering, E.M., (2015). Designed protein reveals structural determinants of extreme kinetic stability. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14605–14610.
75. Jaswal, S.S., Sohl, J.L., Davis, J.H., Agard, D.A., (2012). Energetic landscape of α-lytic protease optimizes longevity through kinetic stability. *Nature* **415**, 343–346.
76. Goldenzweig, A., Goldsmith, M., Hill, S.E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, R.L., Aharoni, A., Silman, I., Sussman, J.L., Tawfik, D.S., Fleishman, S.J., (2016). Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell* **63**, 337–346.
77. Marcelino, A.M., Gierasch, L.M., (2008). Roles of β-turns in protein folding: from peptide models to protein engineering. *Biopolymers* **89**, 380–391.
78. Garcia-Mira, M.M., Sanchez-Ruiz, J.M., (2001). pH corrections and protein ionization in wáter/guanidinium chloride. *Biophys. J.* **81**, 3489–3502.
79. Acevedo, O., Guzman-Casado, M., Garcia-Mira, M.M., Ibarra-Molero, B., Sanchez-Ruiz, J.M., (2002). pH corrections in chemical denaturant solutions. *Anal. Biochem.* **306**, 158–161.
80. Holmgren, A., (1979). Thioredoxin catalyzes the reduction of insulin disulfides by dithiothreitol and dihydrolipoamide. *J. Biol. Chem.* **254**, 9627–9632.
81. Slaby, I., Holmgren, A., (1975). Reconstitution of *E. coli* thioredoxin from complementing peptide fragments obtained by cleavage at methionine-37 or arginine-73. *J. Biol. Chem.* **250**, 1340–1347.
82. Georgescu, R.E., Garcia-Mira, M.M., Tasayco, M.L., Sanchez-Ruiz, J.M., (2001). Heat capacity analysis of oxidized *Escherichia coli* thioredoxin fragments (1–73, 74–108) and their noncovalent complex. Evidence for the burial of apolar surface in protein unfolded states. *Eur. J. Biochem.* **268**, 1477–1485.
83. Kubelka, J., Henry, E.R., Cellmer, T., Hofrichter, J., Eaton, W.A., (2008). Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 18655–18662.
84. Song, Y., DiMaio, F., Wang, R.Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., Baker, D., (2013). High-resolution comparative modelling with RosettaCM. *Structure* **21**, 1735–1742.
85. Rajasekaran, N., Gopi, S., Narayan, A., Naganathan, A.N., (2016). Quantifying protein disorder through measures of excess conformational entropy. *J. Phys. Chem. B* **120**, 4341–4350.
86. W. Kabsch, XDS. Acta Crystallogr. Sct. D Biol. Crystallogr. 66 (2010) 125-132.
87. Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R. J., Vagin, A., Wilson, K.S., (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 235–242.
88. Kantardjieff, K.A., Rupp, B., (2003). Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **12**, 1865–1871.
89. Vagin, A., Teplyakov, A., (2010). Molecular replacement with MOLREP. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 22–25.
90. Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G. J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., Richardson, J. S., Terwilliger, T.C., Zwart, P.H., (2010). PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221.
91. Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, K., (2010). Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 486–501.
92. Painter, J., Merritt, E.A., (2006). TLSMD web server for the generation of multi-group TLS models. *J. Appl. Crystallogr.* **39**, 109–111.

93. Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C., (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 12–21.
94. Vriend, G., (1990). WHAT IF: A molecular modelling and drug design program. *J. Mol. Graph.* **8**, 52–56.
95. Hooft, R.W.W., Sander, C., Vriend, G., (1996). Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins* **26**, 363–376.

5. DISCUSSION

Here we discuss the main results obtained in this thesis that support the use of resurrected ancestral proteins as molecular tools for testing hypothesis about relevant problems in evolution and also as valuable scaffolds for biotechnological and biomedical applications. In particular, in Section 5.1 we ask about the relationship between protein folding *in vitro* and *in vivo*, in an evolutionary context. Results on this topic have been published in the journals PNAS (Candel, A.M., Romero-Romero, M.L., Gamiz-Arco, G., Ibarra-Molero, B., and Sanchez-Ruiz, J.M. (2017). Fast folding and slow unfolding of a resurrected Precambrian protein. Proc. Natl. Acad. Sci. U. S. A. 114, E4122–E4123) and Biochemical Journal (Gamiz-Arco, G., Risso, V.A., Candel, A.M., Inglés-Prieto, A., Romero-Romero, M.L., Gaucher, E.A., Gavira, J.A., Ibarra-Molero, B., and Sanchez-Ruiz, J.M. (2019). Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding. Biochem. J. 476, 3631–3647) and correspond to Publication 1 and 2, respectively. In Section 5.2 a practical biotechnological application of ancestral reconstruction to engineer efficient heterologous protein expression is proposed. The results have been recently published in the Journal of Molecular Biology (Gamiz-Arco G, Risso VA, Gaucher EA, Gavira JA, Naganathan AN, Ibarra-Molero B, Sanchez-Ruiz JM. Combining ancestral reconstruction with folding-landscape simulations to engineer heterologous protein expression. J Mol Biol. 2021 Oct 20:167321. doi: 10.1016/j.jmb.2021.167321) and correspond to Publication 3.

At this point it is convenient to mention that the model system used for these studies is thioredoxin for the following reasons: thioredoxins are small cytoplasmic proteins of approximately 100 amino acid residues[71] that can be easily purified. They are present in all known cells (eukaryotes, bacteria and archaea) and are involved in a diversity of cellular processes[72,73]. They have been extensively studied in bibliography and our group has an extensive experience in working with thioredoxins[109,110,119,120,111–118]. Even more, they are an excellent model system for our evolutionary studies as we have previously resurrected and characterized a number of Precambrian thioredoxins dating back between ~1.4 and ~4 billion years[9,21,35,37]. Ancestral thioredoxins display considerable sequence differences with respect to their modern counterparts (up to ~50%) and also different properties. We have shown that ancestral thioredoxins display enhanced stability compared to their modern mesophilic counterparts[9,18], higher levels of redox activity in acid conditions[9] and, interestingly, the 3D-structure of these proteins has been conserved over billions of years. Indeed, ancestral and

modern thioredoxins display the same fold and only small structural changes have occurred throughout evolution[21].

5.1. Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding

5.1.1. Comparison of the *in vitro* folding of two ancestral thioredoxins and modern *E. coli* thioredoxin

The starting point of our study is the experimental determination of *in vitro* folding rates of modern *E. coli* thioredoxin and the representations of the last bacterial common ancestor (LBCA) and the last common ancestor of the cyanobacterial, deinococcus and thermus groups (LPBCA) thioredoxins, dating back ~4 and ~2.5 billion years ago, respectively. The corresponding chevron plots representing rate constants *versus* guanidinium concentration are displayed in Figure 1a.

From these results we can extract some relevant conclusions:

i. The unfolding of the ancestral proteins is approximately three orders of magnitude slower than the unfolding of the modern thioredoxin suggesting an increased kinetic stability. The higher activation energy barrier separating the native and transition states is consistent with an evolutionary narrative. Ancestral proteins lived in an aggressive environment and enhanced kinetic stability would have potentially acted as a protection mechanism for the biologically-relevant native state[121,122].

ii. An important feature to highlight is the significant deviations from linearity (rollovers) at low denaturant concentrations in the folding branches of the chevron plots. Rollovers indicate complex kinetics with transiently populated intermediates[122–124]. In this case, the rollover may be due to proline isomerization, a kinetic complexity that have been previously reported for thioredoxins[75]. Isomerization is a slow process where kinetically-trapped intermediates with the wrong proline conformer might be populated in the folding process[125–127]. Thioredoxins have a proline in cis conformation at position 76 in their native conformation[128] and its isomerization has been reported to be rate-limiting in the *in vitro* refolding. Interestingly, *E.coli* thioredoxin has four trans prolines (residues

34, 40, 64 and 68) whereas the ancestral LPBCA and LBCA thioredoxins have six trans prolines (residues 22, 30, 34, 64, 40 and 95), that can complicate their folding kinetics.

iii. The most intriguing result comes from the faster ancestral folding. The folding rates for the ancestral thioredoxins, extrapolated to 0 M denaturant, are nearly two orders of magnitude faster than the folding of their modern counterpart, indicating that ancestral proteins have a more efficient folding *in vitro.* Furthermore, we have also experimentally tested that the mutation Pro76Ala in our modern and ancestral thioredoxins has two inmediate consequences: on one hand, folding significantly accelerates becoming similar for the three proteins (Figure 1b) and, on the other hand, reductase activity is completely abolished (Figure 1c) in all cases. The structural interpretation seems to be straightforward: removing Pro76 eliminates the well-known kinetic bottleneck created by its cis conformation[128] and therefore increases folding rate, but also hampers function since Pro76, strictly conserved in all thioredoxins, is crucial for the active-site structure. However, the story cannot be that simple, since the folding of LPBCA and LBCA thioredoxins is faster than that of *E. coli* thioredoxin, even when proline is at position 76. Clearly, something else might be aggravating its folding problem.

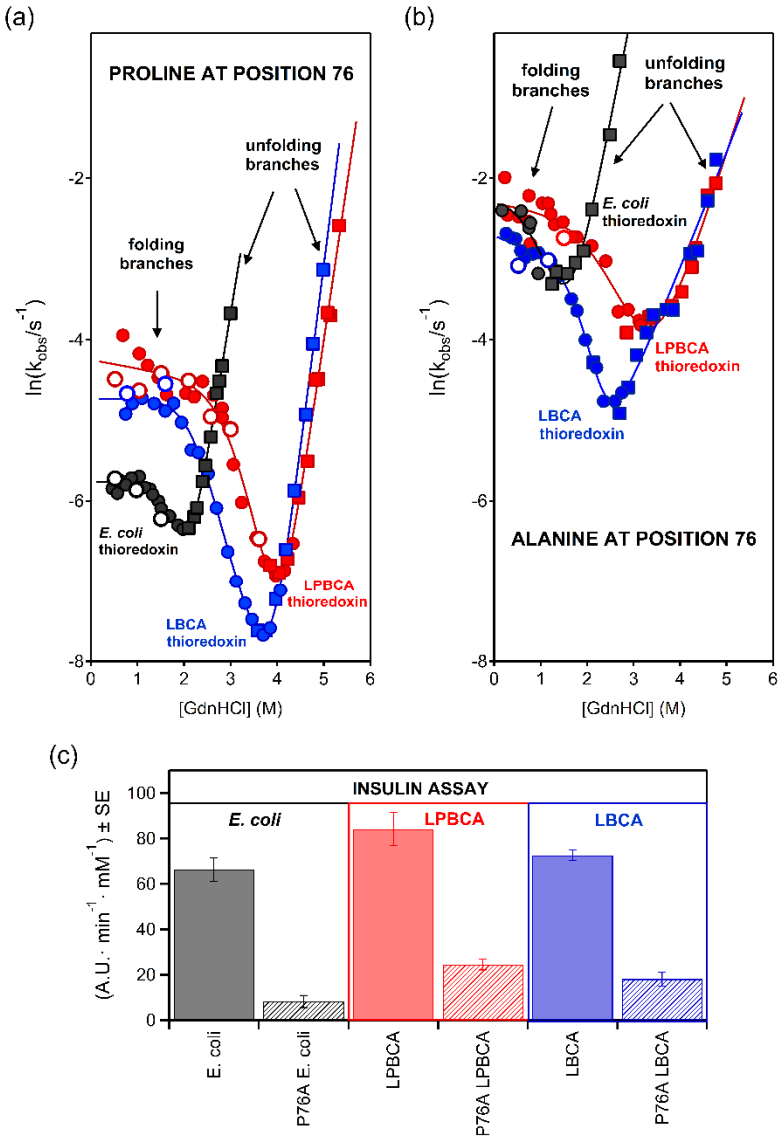


Figure 1. Effect of the amino acid changes at position 76 on *in vitro* thioredoxin folding *and* activity for the modern *E. coli* thioredoxin and the ancestral resurrected LPBCA and LBCA thioredoxins. (a) Chevron plots for the wild-type thioredoxins studied with proline at position 76. Circles refer to the data obtained in the folding direction and squares refer to the data obtained in the unfolding direction. (b) Same as (a) but the plot represent the thioredoxin variants studied with the Pro76Ala mutation (b) Plot of the *in vitro* reductase activity measured by the insulin aggregation assay[129] for the wild-type thioredoxins and their Pro76Ala variants. (Reprinted and modified with permission of Portland Press from Gamiz-Arco et al., 2019)[130].

5.1.2. Identification of the residue responsible for the slow folding of *E. coli* thioredoxin. Interpretation in an evolutionary context.

Next we aimed to investigate the molecular determinants for the efficient folding in ancestral thioredoxins. *E. coli* thioredoxin sequence differs at about 40-50 amino acid positions from the sequences of LBCA and LPBCA (Figure 2) so an exhaustive search, by mutational analysis, does not seem to be a good strategy to follow.

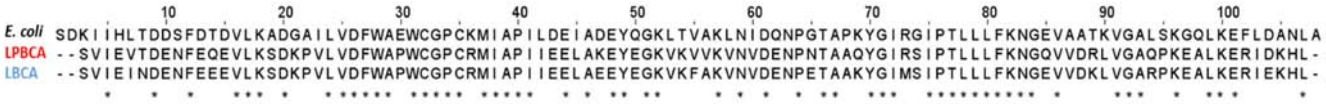


Figure 2. Sequences of ancestral and modern *E.coli* thioredoxins. Alignment of sequences from *E. coli* thioredoxin, and the resurrected ancestral thioredoxins corresponding to the last common ancestor of the cyanobacterial, Deinococcus and Thermus groups (LPBCA thioredoxin) and LBCA (last bacterial common ancestor). Positions with identical residues in the three sequences are marked with asterisks.

In principle, positions with glycine are likely good candidates as its lack of sidechain confers more conformational flexibility to the polypeptide backbone and therefore, can complicate the folding process[131]. Ancestral thioredoxins display glycines at position 33, 51, 71, 84 and 92. Interestingly, *E. coli* thioredoxin has 4 additional glycine residues at positions 21, 65, 74 and 97 (Figure 2). Mutational analyses show that the Ser/Gly exchange at position 74 is the responsible of the folding rate difference between the modern and the ancestral proteins (Figure 3).

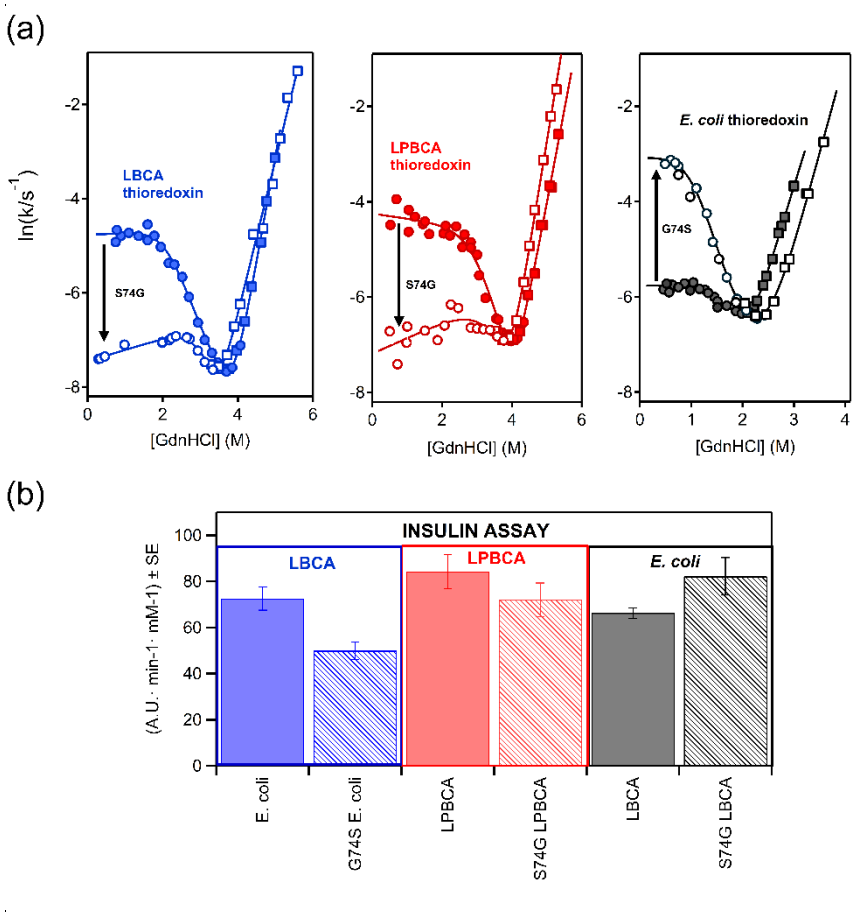


Figure 3. Effect of the amino acid changes at position 74 on *in vitro* thioredoxin folding and activity for the modern *E. coli* thioredoxin and the ancestral resurrected LPBCA and LBCA thioredoxins. (a) Chevron plots for the *E.coli* thioredoxin and the two ancestral LBCA an LPBCA thioredoxins and their respective variants with an exchange Ser/Gly at position 74. Circles refer to the data obtained in the folding direction and squares refer to the data obtained in the unfolding direction. (b) *In vitro* enzyme activity followed by the insulin aggregation assay[129]. These results show that reductase activity is not affected by Ser/Gly replacement at position 74. (Reprinted and modified with permission of Portland Press from Gamiz-Arco et al., 2019)[130].

Interestingly, serine is the residue at position 74 in LBCA and LPBCA thioredoxins while glycine is the residue at position 74 in *E. coli* thioredoxin. We have experimentally found that back-to-the-ancestor replacement (Gly74Ser) in the modern protein increases its folding rate to the level of the ancestral proteins and, conversely, the Ser74Gly ancestral variants slow down folding (Figure 3).

Position 74, spatially located close to the cis Pro76, seems to be the key position. Thus, Ser74Gly mutation has revealed as a folding-degrading feature. Moreover, results obtained from double-mutant cycle (Figure 4) analysis support that the degrading effect of the Ser74Gly mutation on the folding of modern thioredoxin is linked to the kinetic bottleneck created by cis Pro76.

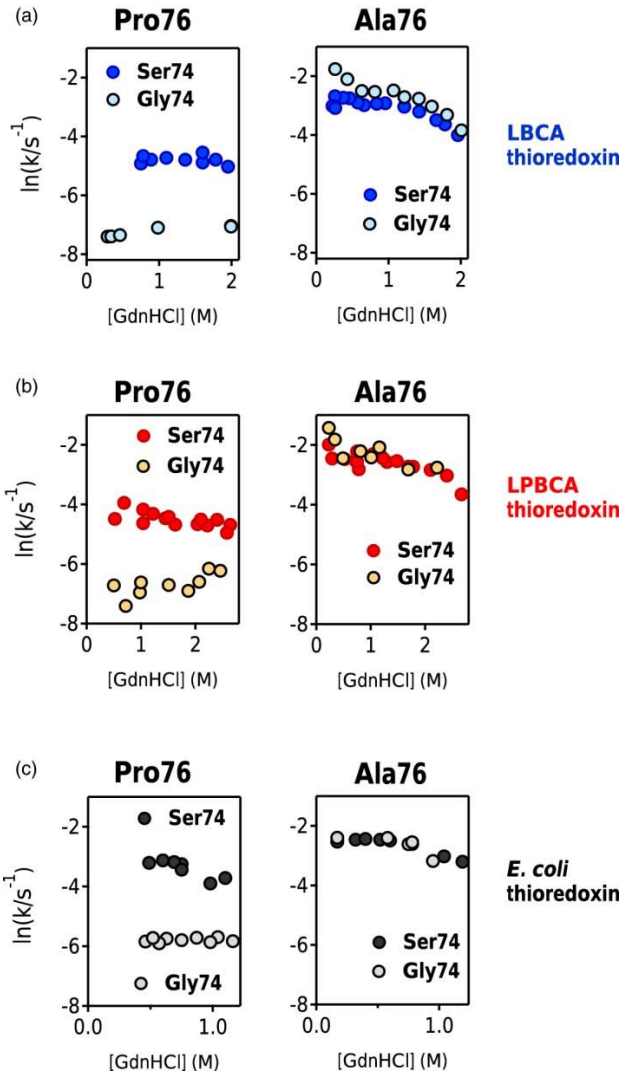


Figure 4. Double-mutant cycle analysis that shows the effects of the amino acid at position 74 and 76 on the folding rate of thioredoxins. Folding rates for (a) ancestral LBCA thioredoxin; (b) ancestral LPBCA thioredoxin; (c) modern *E. coli* thioredoxin. These data include Pro/Ala exchange at position 76, Gly/Ser exchange at position 74 and the combination of the two exchanges. These results confirm that the effect of the Gly74Ser mutation is only produced when proline is present at position 76 (left) and it is insignificant when alanine is present at position 76 (right). (Reprinted with permission of Portland Press from Gamiz-Arco et al., 2019)[130].

We rationalize the structural basis for this effect as follows. To stabilize the biologically relevant cis conformation of Pro76 it seems to be critical evolutionary conserved interactions established within the 70-79 loop, namely, hydrophobic interactions and a hydrogen bond between the backbone carbonyl amino acid at position 74 (Gly or Ser) and Thr77, as reveals by the 3D structures of the three proteins (Figure 5). It is reasonable to think that these interactions may also favour local residual structures that may promote correct folding. Conversely, the flexible link generated by Gly74, as compared to Ser, should allow conformational diversity in the 70-79 loop slowing down folding. This statement is supported by our kinetic data on Gly74Ala *E. coli* thioredoxin. Thus, alanine at position 74 produces an increase in folding rate similar to Gly74Ser mutation.

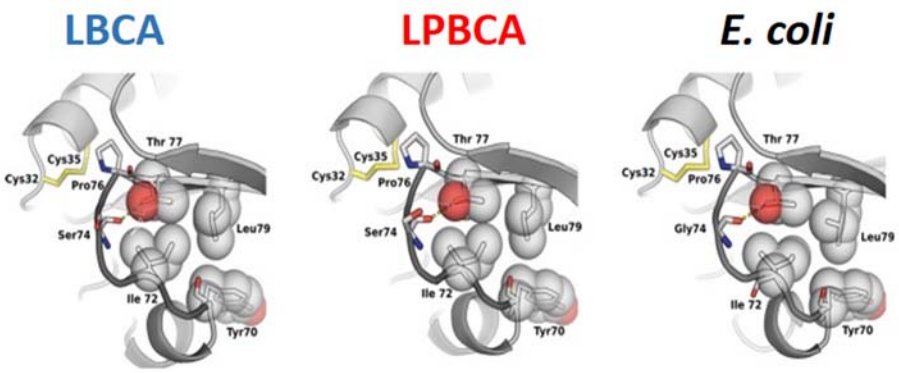


Figure 5. Blow-ups of the region including the 70-79 loop in *E. coli* (PDB code 2TRX), LPBCA (PDB code 2YJ7) and LBCA (PDB code 4BA7) thioredoxins. Important positions for the stability of the loop and the active site disulfide bridge are labelled. (Reprinted with permission of Portland Press from Gamiz-Arco et al., 2019)[130].

The evolutionary acceptance of this mutation cannot be understood without invoking the protecting role of the cellular folding-assistance machinery. The following plausible evolutionary narrative for the evolution of protein folding thus arises. It is reasonable to think that protein folding at a very early stage was unassisted and unprotected. Primordial folding efficiency was then linked to fast folding encoded at the sequence/structure level. Once an efficient assistance machinery had emerged, mutations that impaired ancient sequence/structure determinants of folding efficiency could be accepted, since those determinants were no longer necessary. *In vitro* folding landscapes for some modern proteins may then represent moderately degraded versions of the landscapes from the unassisted folding of their ancestors.

The Ser74Gly mutation in the line of descent to *E. coli* thioredoxin appears as a clear example of this scenario. Residue occupancy at position 74 in the thioredoxin phylogenetic tree used for ancestral sequence reconstruction (Figure 6) seems to be consistent with the interpretation described above. A rapid inspection reveals that amino acids different than Ser or Gly barely appears at position 74. Interestingly, serine is present in the majority of ancestral thioredoxins and in all the oldest phylogenetic nodes. On the other hand, in modern thioredoxins both, glycine and serine at this position, are found in a similar percentage. The first ancestral thioredoxin that accepted a glycine at position 74 was the "Last common ancestor of γ-proteobacteria and β-proteobacteria" about 2 billion years ago. Apart from it, there are more recent nodes that have accepted glycine at position 74 and this glycine has been conserved in their trajectory towards modern descendants. Whenever a glycine is accepted in this position,

it is maintained, indicating that although it is not the best choice in terms of folding, evolution should have selected it for a certain reason. As mentioned before, this position is located in a very important region of the protein, close to the active site, and glycine might be involved in the interaction surface with other molecular partners, being to some extent necessary.

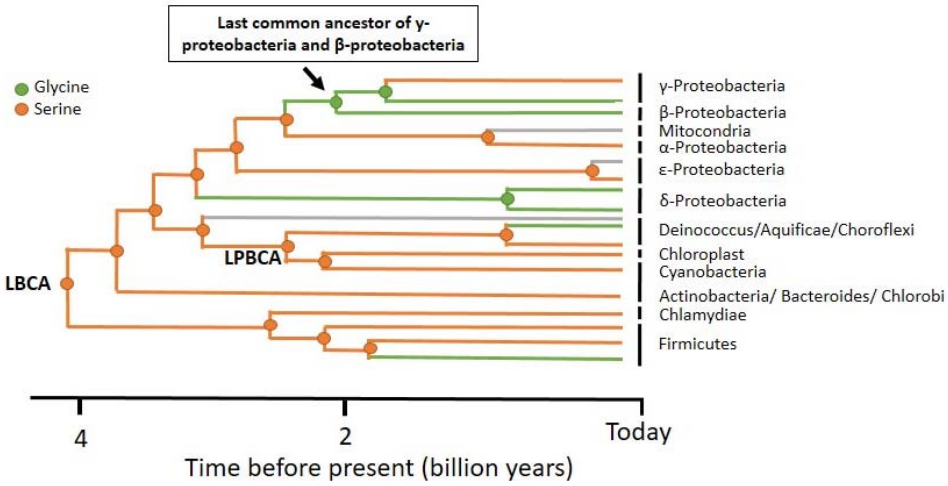


Figure 6. Schematic representation of the bacterial section of the thioredoxin phylogenetic tree used in ancestral sequence reconstruction[9] showing the evolutionary history of the amino acid at position 74. Phylogenetic tree showing geological time of the different nodes. The ancestral thioredoxins studied in this thesis are labeled. The nodes present a different colour depending of the amino acid at position 74 (green: glycine, orange: serine and grey: other).

In summary, results in this section support our hypothesis that *in vitro* folding landscapes for some modern proteins may then represent moderately degraded versions of the landscapes from the unassisted folding of their ancestors. Then, it is expected that different degrees of degradation might have taken place within modern proteins. To explore this idea we have extended out kinetic studies to a set of modern bacterial thioredoxins. Results are described in the following section.

5.1.3. Non-conservation of folding rates in modern thioredoxins

Our previous results on the Ser74Gly replacement in *E. coli* thioredoxin as degrading mutation opens up the possibility that other positions in the neighborhood (70-79 loop, that includes the active site cis Pro76) might have impact in the folding rate. To explore this idea, a set of 14 modern bacterial thioredoxins (including *E.coli* thioredoxin) has been selected in a NCBI

Reference Sequence Database search, representing natural sequence diversity in positions within the 70-79 loop, potentially important for its stability.

The experimental characterization of their *in vitro* folding (Figure 7), using fluorescence kinetics and double jumps experiments, has confirmed that it is not a conserved feature. The range of folding rates for modern thioredoxins is wide (at least a ~100-fold range). We propose that these huge differences have not an important impact *in vivo* since the folding assistance machinery assists protein folding in the cellular environment.

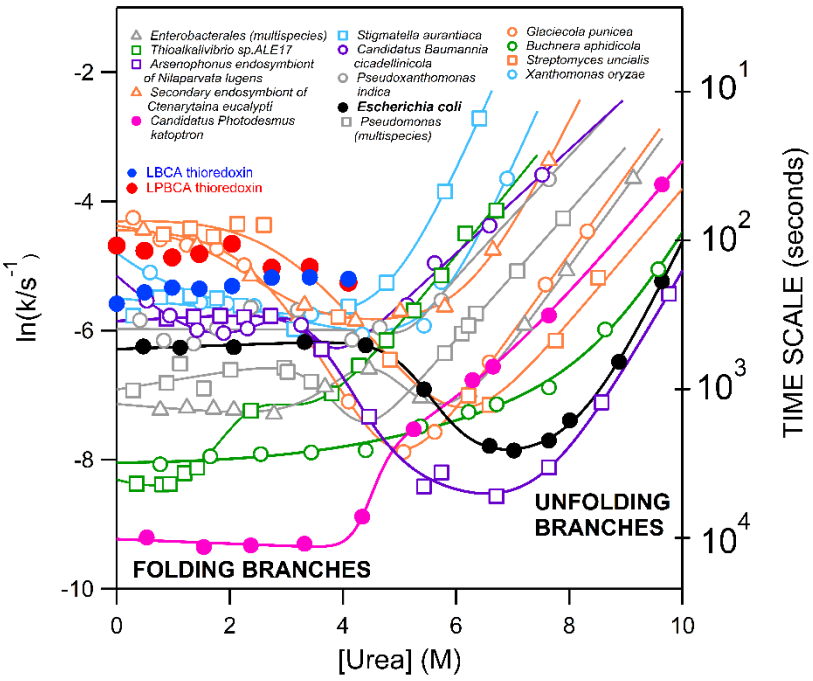


Figure 7. Chevron plots of 14 modern thioredoxins. Plots of folding–unfolding rate constants *versus* urea concentration for the modern thioredoxins selected. Folding rates given correspond to the phase of the fluorescence kinetic that leads to the native state. Folding data for LBCA and LPBCA thioredoxins are also included. (Reprinted with permission of Portland Press from Gamiz-Arco et al., 2019)[130].

Overall, in this section we have confirmed that *in vitro* thioredoxin folding rates have not been conserved through the evolution, contrary to claims from recent literature[132]. In fact, while some modern thioredoxins have an efficient folding similar to the ancestral thioredoxins, in the minute timescale, other modern thioredoxins need some hours to achieve their folded state.

Our study has suggested that the folding landscape was optimized at a very early stage, but billions of years of evolution have degraded the folding rate in modern proteins, due mainly to the assistance provided by the ribosome and the chaperones inside the cells[62,66,133–135]. Once

folding assistance was available, mutations that slow down the folding in vitro could be accepted. In conclusion, all the results obtained in this first part of the doctoral thesis provide a simple evolutionary interpretation of *in vitro* protein folding and help to better understand the relation between protein folding *in vitro* and *in vivo*.

5.2. Combining ancestral reconstruction with folding-landscape simulations to engineer heterologous protein expression

The second part of this thesis is focused on developing a methodology, useful in practical application scenarios, which involves the improvement of inefficient heterologous protein expression.

Successful heterologous expression of proteins is one of the major challenges in biotechnology nowadays. *E. coli* is the most preferred host for protein production due to its rapid growth, easy manipulation, low-cost production and the ability to express proteins at very high levels[136,137]. Nevertheless, obtaining substantial amount of a certain protein in its functional native form is not often an easy task. Oftentimes, non-functional protein such as misfolded, proteolyzed or aggregated protein[138] is the final result. That is a serious limitation in the production of recombinant proteins and a critical problem in the development of biotechnology.

Of course, there are a number of approaches that can be used to improve the production of recombinant proteins such as changing the expression vector, the culture conditions, minimizing protease activity, using fusion tags or supplementing the host with chaperones that assist the proper protein folding[81,86,136]. Despite these possible solutions that sometimes result useful, the reality is that efficient heterologous expression is still a major problem in biotechnology.

As mentioned, heterologous expression is essential in the production of proteins in the industrial, pharmaceutical and research field, but also it is pivotal in metagenomics, a growing field, where the uncovering of novel genomes necessarily goes through heterologous expression of encoded proteins.

We believe that inefficient heterologous expression linked to low solubility is likely due to the absence of coevolution of the recombinant protein with the natural chaperones of the new host. As shown in Section 5.1.1. and 5.1.3., modern *in vitro* folding rates are slower than

the folding rates for the resurrected ancestral thioredoxins. Our results suggest that unassisted primordial folding was efficient because it was fast and the expected slow *in vitro* folding of modern proteins is linked to the ancient mutations that impaired ancestral determinants for folding efficiency and that the acceptance of these mutations is linked to the emergence of the cellular folding-assistance machinery.

Therefore, we do have a suitable model system to investigate a general approach to rescue inefficient heterologous folding in an evolutionary context: on one hand ancestral thioredoxins, extensively studied in our laboratory, are necessarily expressed in foreign hosts and, as previously shown in the literature, many other ancestral proteins display an increased yield when expressed in *E. coli*[31,40,95–98,87–94]. On the other hand, a set of modern thioredoxins, spanning a wide range of folding rates, is also available in our laboratory (Section 5.1.3.). In particular, we have focused on thioredoxin from *Candidatus Photodesmus katoptron* (*CPk*), an obligate symbiont of flashlight fish (thus mimicking a typical scenario of metagenomic studies), revealed as the worst *in vitro* folder within our modern thioredoxins (Figure 7). As detailed below, *in vitro* folding kinetic studies and *in vivo* efficiency studies (soluble *versus* insoluble protein) on a number of chimeras and thioredoxin variants have been performed aiming to decipher suitable back-to-the-ancestral mutations that could lead to a more efficient heterologous expression.

5.2.1. *In vitro* folding and expression efficiency in *E. coli* of modern and ancestral thioredoxins

For this study, we have selected the modern thioredoxin from the *Candidatus Photodesmus katoptron* (*CPk* thioredoxin) as target protein to improve its inefficient heterologous expression. As controls of efficient expression, *E. coli* and the ancestral LPBCA thioredoxins have been used.

These proteins are very highly divergent at the sequence level. Sequence identity of *Candidatus Photodesmus katoptron* thioredoxin and *E. coli* thioredoxin is 69%. The ancestral LPBCA thioredoxin displays also a low sequence identity with modern proteins: 57% and 45% with *E. coli* and *Candidatus Photodesmus katoptron* thioredoxins, respectively. Despite the extensive sequence differences, the three proteins are similar in terms of 3D-structure and redox activity (Figure 8)

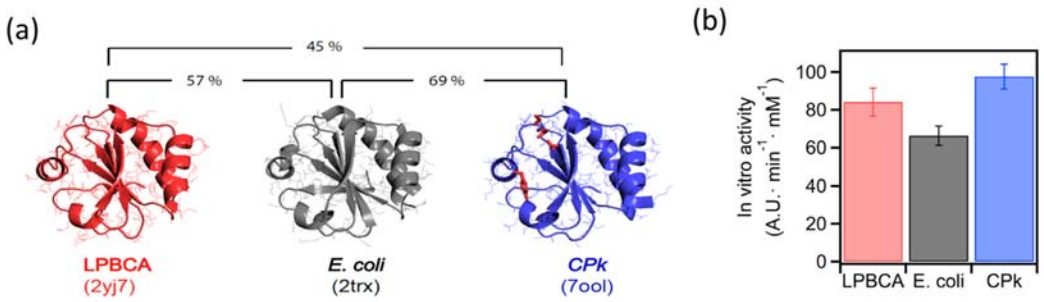


Figure 8. Structure and redox activity of modern and ancestral thioredoxins studied in this section. (a) 3D-structures of the ancestral thioredoxin corresponding to the last common ancestor of the cyanobacterial, *Deinococcus* and *Thermus* groups (LPBCA thioredoxin), modern thioredoxins from *E. coli* and *Candidatus Photodesmus katoptron* (*CPk*). The structure determined for *CPk* thioredoxin actually corresponds to a variant with four back-to-LPBCA mutations (Phe70Tyr, Gly74Ser, Val75Ile and Ser77Thr), due to the impossibility of obtaining crystals of diffraction quality for the wild-type *CPk* thioredoxin. In the figure is also shown the PDB identifiers for each thioredoxin and the sequence identity percentages between the proteins. (b) *In vitro* activities of the thioredoxins of panel (a) using the insulin turbidimetric assay129. The values are the average of 3 independent measurements and the error bars represent the standard deviations. (Reprinted and modified with permission of Elsevier from Gamiz-Arco et al., 2021b)[139]

In Section 5.1., we have previously studied the folding kinetics of these proteins and now, we also studied their efficiency of expression in *E. coli* (Figure 9). *CPk* thioredoxin displays a very slow refolding *in vitro*, reaching the native state in the time scale of hours and its expression in *E. coli* at 37 °C leads mostly to insoluble protein. Moreover, double-jump unfolding assays show that the *in vitro* folding of *CPk* thioredoxin leads to protein aggregation (Figure 10). The effect is more pronounced the higher the total concentration of protein, suggesting that *in vitro* folding inefficiency is linked to protein aggregation.

By contrast, resurrected LPBCA thioredoxin folds fast *in vitro* and expresses efficiently in *E. coli* (Figure 9). As expected, *E.coli* thioredoxin is purified in high yield and in a soluble and functional form, but its *in vitro* folding rate is in between the ancestral and the CPk thioredoxins.

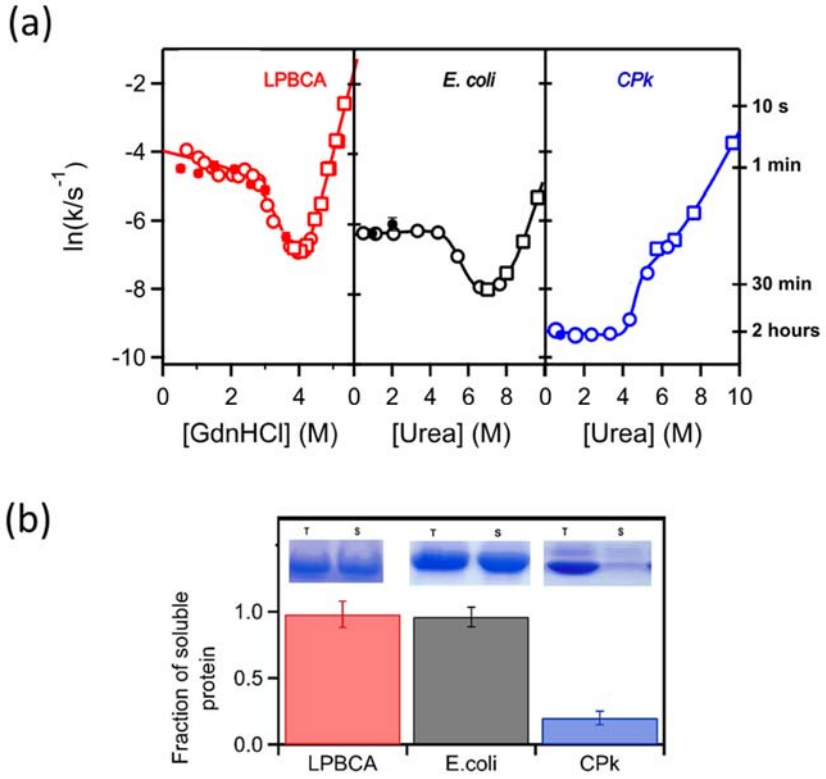


Figure 9. Comparison of the efficiency of *in vitro* and *in vivo* folding of the modern *E. coli* and *CPk* thioredoxins and the ancestral LPBCA thioredoxin. (a) Chevron plots of folding-unfolding rate constants derived from fluorescence kinetic profiles *versus* denaturant concentration. Circles refer to folding experiments and squares refer to unfolding experiments. The plot only includes the folding phase which leads to the native state, as indicate the double-jump unfolding assays (closed symbols). (b) Fraction of soluble protein obtained upon expression of thioredoxins in *E. coli* at 37 °C. The graph includes the sections of the SDS-PAGE gels with the thioredoxin bands used to determinate the fraction of soluble protein. T and S represent "total" and "soluble", respectively. (Reprinted and modified with permission of Elsevier from Gamiz-Arco et al., 2021b)[139].

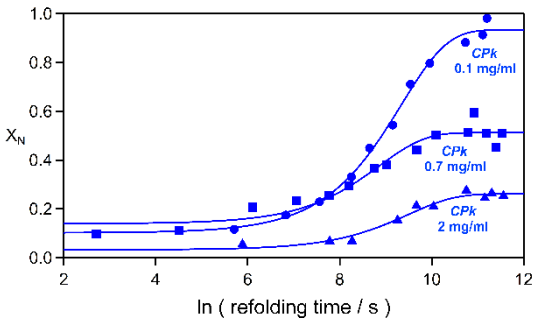


Figure 10. Double-jump unfolding assays of modern *Candidatus Photodesmus katoptron* (*CPk*) thioredoxin at different protein concentrations. This figure shows the profiles of fraction of native state *versus* refolding time obtained by double-jump unfolding assays. The refolding time is the time the protein is allowed to refold after its denaturation. The lines represent the fits to a single exponential. These results confirm the aggregation of *CPk* thioredoxins in its folding at high protein concentrations. (Reprinted and modified with permission of Elsevier from Gamiz-Arco et al., 2021b)[139].

The results of this section suggest a clear relationship between the efficiency of heterologous expression and *in vitro* folding rate.

Our first attempt to overcome the expression problem of *CPk* thioredoxin was to simultaneously overproduce the functionally cooperating chaperone network of *E. coli*[81,86]. This strategy has been used succesfully to alleviate heterologous expression problems in several cases in *E.coli*[140–148]. However, our results only show a negligible improvement in the soluble fraction of CPk thioredoxin (Figure 11).

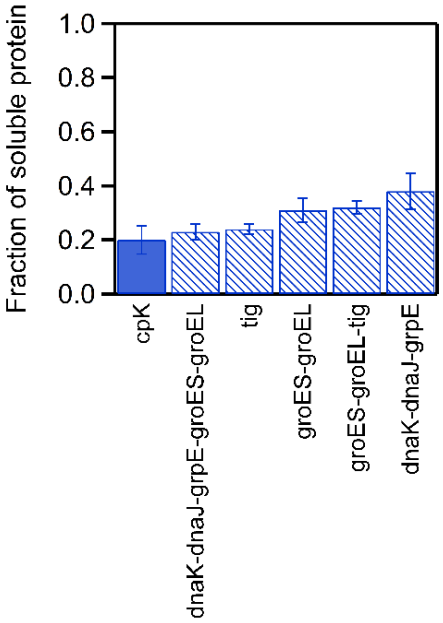


Figure 11. Expression of *CPk* thioredoxin in *E. coli* with over-expression of different chaperones. Representation of the fraction of soluble protein obtained in *E. coli* at 37 °C. The chaperones used were pG-KJE8 (expressing dnaK-dnaJ-grpE-groES-groEL), pGro7 (expressing groES-groEL), pKJE7 (expressing dnaK-dnaJ-grpE), pG-Tf2 (expressing groES-groEL-tig) and pTf16 (expressing the trigger factor). (Reprinted and modified with permission of Elsevier from Gamiz-Arco et al., 2021b)[139]

These results can be explained taking into account coevolution between thioredoxin and the folding assistance machinery in its natural host. Indeed, the folding assistance machinery of *E.coli* is very different from that of *CPk* (sequence identity of chaperones and ribosomal proteins of *CPk* and *E. coli* differs from a 41% to a 96,6%). Therefore, it appears plausible that the assistance machinery of *E.coli* is not able to assist *CPk* thioredoxin in its folding.

5.2.2. Computational predictions of the folding landscapes for modern and ancestral thioredoxins

Following our rationale that inefficient recombinant expression of proteins is linked to inefficient folding in foreign hosts that lack their natural chaperones, theoretical predictions based on the Wako-Saitô-Muñoz-Eaton statistical-mechanical model[149–152] for the folding

landscapes of the three proteins of interests have been carried out aiming to identify specific regions prone to be unfolded and that may favour aggregation processes. As shown below, this has turned out to be an excellent tool for guiding the back-to-the-ancestor engineering. Calculations were made by our collaborator Dr. Athi N. Naganathan at the Indian Institute of Technology Madras, Chennai, India.

A direct comparison between folding landscapes corresponding to LPBCA and *CPk* thioredoxins is very informative as both proteins are necessarily expressed in a heterologous host and, therefore, their folding is not assisted, at least to some extent, by the folding assistance machinery of *E. coli* .Two interesting regions were identified by Dr. Naganathan to be partially structured in *CPk* thioredoxin compared to its ancestral counterpart (Figure 12a). The N-terminal region of *CPk* thioredoxin (residues 1-22) and the region of the critical cis-proline (residues 70-77), that we have already studied in Section 5.1.2. These regions may be the responsible for an inefficient heterologous expression of *CPk*. Moreover, Dr. Athi N. Naganathan has also calculated the probability of every region of the protein to be structured as a function of the reaction coordinate (Figure 12b) and has found that the N-terminal region of the protein is the last region of the protein to fold. These two regions were targeted for an exhaustive mutational analysis to test our hypothesis.

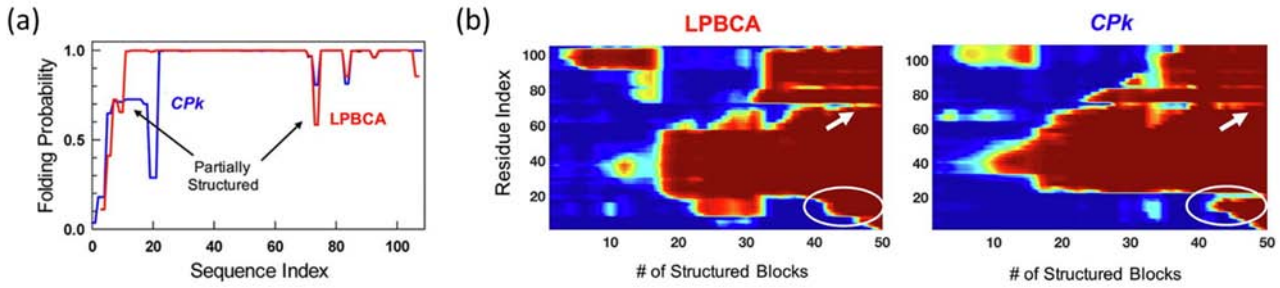


Figure 12. Computational modelling of the folding landscape for modern and ancestral thioredoxins for the modern thioredoxin from *Candidatus Photodesmus katoptron* (*CPk*) and ancestral LPBCA thioredoxin. A block version of the Wako-Saitô-Muñoz-Eaton statistical model was used. (a) The probability of finding residue folded as a function of residue index. (b) Folding probability, coloured from blue (0) to red (1), as a function of the number of structured residues predicted by the model. The N-terminal and the 70–77 regions are highlighted by white ovals and an arrow, respectively. (Reprinted and modified with permission of Elsevier from Gamiz-Arco et al., 2021b)[139].

5.2.3. Heterologous expression efficiency in E. *coli* of *CPk* thioredoxin variants and chimeras

To experimentally test the effect of the computational predictions on the *in vitro* folding and the efficiency of the heterologous expression, we prepared three chimeras and several single mutants, in which the potential important regions/positions have been replaced by the corresponding sequence of LPBCA. The chimera called *CPk*-[1-22] thioredoxin involves 17 mutations in the 1-22 fragment, and the chimera called *CPk*-[70-77] thioredoxin includes 4 mutations in the 70-77 loop. We have also prepared a chimera (*CPk*-[1-22]-[70-77] thioredoxin) in which both regions are replaced by the LPBCA sequence. The single variants involve individual mutations in the 1-22 fragment (Leu7Val, Asp10Glu, Ser11Asn, Leu14Gln, Asn15Gln, Ile17Leu, Ser18Lys, Ala19Ser, Ser20Asp, Gly21Lys, Val22Pro) and in the 70-77 loop (Phe70Tyr, Gly74Ser, Val75Ile, Ser77Thr).

The results of this extensive mutational study indicate that the important positions responsible for the inefficient folding and heterologous expression of *CPk t*hioredoxin are positions 11 and 74. The most remarkable effect comes from Ser11Asn mutation that surprisingly improves protein expression (showing more than 80% of soluble protein). The second most important effect is attributed to replacement at the already discussed position 74, widely studied by us in Section 5.1.2. Thus, the heterologous expression of Gly74Ser *CPk* thioredoxin results in 60 % of soluble protein. And finally, the double mutant is expressed totally soluble in *E. coli*, without affecting the functional properties of the protein.

The structural basis for these effects seems to be clear. We refer to the reader to Section 5.1.2. for the Gly74Ser replacement. The impact of mutation at position 11 is quite similar. The residue at this position is part of a type IV β-turn (Figure 13). β-turns are crucial for protein folding because they impose restrictions and generate interactions that may help the folding[153]. The substitution of a serine at position 11 by an asparagine promotes the β-turn formation because this residue can form hydrogen bonds with the threonine at position 8 (Figure 13). This reduces the formation of different alternative conformations for the 8-11 segment during the folding and, consequently, the time that partially unfolded states requires to find the correct spatial positioning of residues that leads to the native structure of the β-turn.

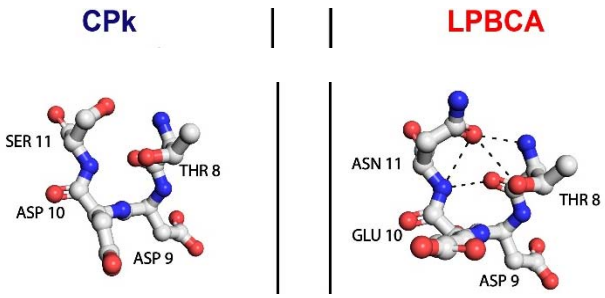


Figure 13. Representation of the β-turn type IV in the modern *CPk* thioredoxin and the ancestral LPBCA thioredoxin. This turn is formed by the residues at position 8-11 of the thioredoxins. The hydrogen bonds predicted by WHAT IF[154] that involve the residue in position 11 are highlighted. (Reprinted and modified with permission of Elsevier from Gamiz-Arco et al., 2021b)[139]

5.2.4. Relationship between heterologous expression, *in vitro* folding rate and stability

Finally, the extensive characterization of the wild-type modern/ancestral thioredoxins, modern/ancestral chimeras and single and double mutants, carried out in Sections 5.1. and 5.2, in terms of *in vitro* folding rates, heterologous folding efficiency and stability, provide a substantial amount of experimental information that allows us to extract relevant conclusions.

In first place, a plot of fraction of soluble protein obtained in *E. coli versus* the logarithm of the *in vitro* folding rate constant shows a good correlation between both properties (Figure 13) suggesting that an efficient heterologous expression is achieved when the time partially-unfolded states become significantly populated in the folding process is reduced. Actually, there is a linear correlation (Figure 14b) for thioredoxins that have a slow *in vitro* folding rate. These proteins are prone to aggregate and have an inefficient heterologous expression. Conversely, proteins with a fast *in vitro* folding (small lifetime values in the time range of some minutes) have an efficient heterologous expression in *E.coli,* leading essentially to 100% soluble protein (Figure 14a).

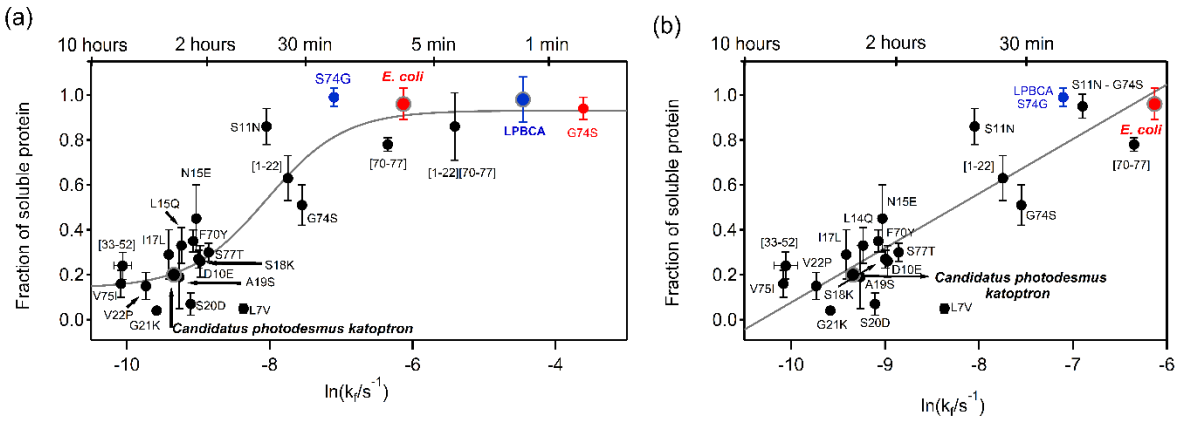


Figure 14. Correlation of the efficiency of *in vivo* heterologous expression with *in vitro* folding rate. (a) Plot of fraction of soluble protein obtained in *E. coli* at 37 °C *versus* the logarithm of the *in vitro* folding rate constant for *CPk*, *E. coli*, LPBCA thioredoxins and their variants (b) Blow-up of panel (a) clearly shows the linear correlation between the efficiency of heterologous expression and *in vitro* folding rate for all the proteins that have an inefficient folding *in vitro*.

In second place, the role of stability as a determinant of heterologous expression has been also explored. Values for denaturation temperature of modern/ancestral proteins, chimeras, the single-mutant variants and the double Ser11Asn/Gly74Ser variant obtained by DSC experiments have been taken as a metric for stability. It is important to note that the chimeras, single and double mutants contain back-to-LPBCA mutations. LPBCA is a hyperstable protein that displays a much higher denaturation temperature that *CPk* thioredoxin (background protein in most cases), so the majority of these mutations are expected to produce stability enhancements.

A plot of fraction of soluble protein obtained in *E. coli versus* stability reveals a reasonable correlation between both properties (Figure 15) for the vast majority of the proteins. However, it is important to highlight that stability cannot solely explain the rescue of inefficient heterologous expression. Thus, there are some highly stable variants that do not display an efficient heterologous expression as is the case of the *CPk*-[1-22] chimera. Moreover, there are other variants that do not require a really high stability to display an efficient heterologous folding, as in the case of Ser11Asn *CPk* or the double mutant Ser11Asn/Gly74Ser *CPk*. Both variants have similar stability as the wild-type protein *CPk*, but their heterologous expression is much more efficient. A direct comparison between *CPk* [1-22] and the Ser11Asn mutant leaves no doubts that the stability cannot be the only reason to explain heterologous folding efficiency.

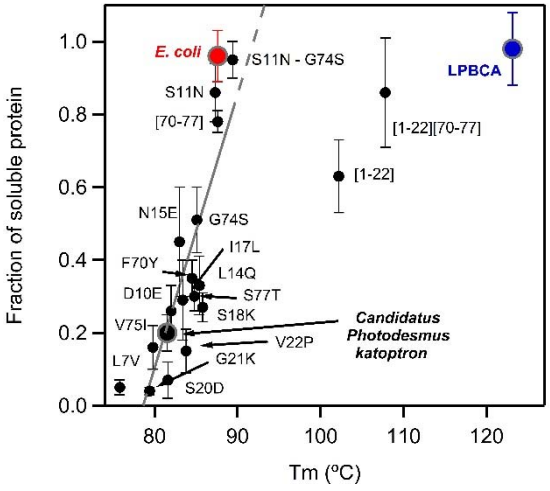


Figure 15. Correlation between heterologous folding efficiency and stability. Plot of fraction of soluble protein expressed in *E. coli* at 37 °C *versus* denaturation temperature values for LPBCA, *E. coli* and *CPk* thioredoxins and several *CPk* variants and chimeras.

And finally, the role of stability as a determinant of folding rates has been also explored. Actually, this has been a subject of discussion for a long time in the bibliography. A number of studies reported are consistent with the idea that stability can play an important role in the folding rates of protein with the same topology[155,156].

Our data corresponding to the wild-type thioredoxins and their variants also show a reasonable correlation between folding rate and stability (Figure 16). The observed pattern is very similar to that described above for the heterologous efficiency and the stability. Similarly, there are some outliers as is the case of the *CPk*-[1-22] thioredoxin, that despite being one of the most stable proteins, its folding rate is slow.

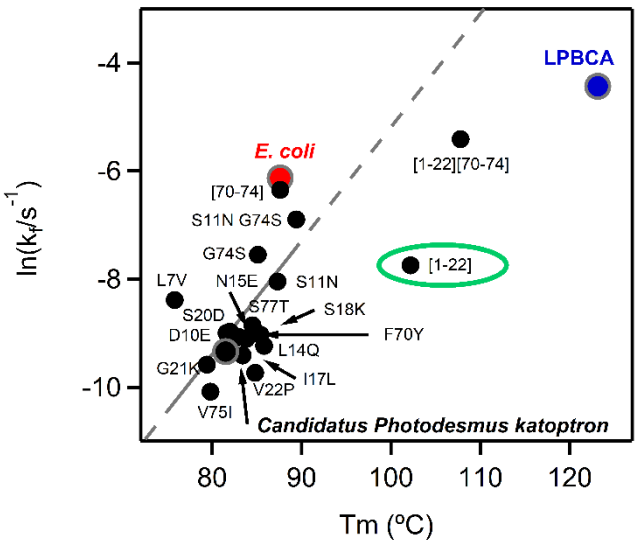


Figure 16. Correlation of *in vitro* folding and stability. Plot of the logarithm of refolding rate constants obtained by double-jump assays at 1M urea *versus* denaturation temperature values for LPBCA, *E. coli* and *CPk* thioredoxins and several *CPk* variants and chimeras, obtained by DSC experiments. The *CPk*-[1-22] thioredoxin is marked in the graph to highlight that clearly does not follow the correlation pattern between *in vitro* folding and stability.

On the other hand, we also have stability data for the modern thioredoxins (Figure 17) studied in section 5.1.3. In this occasion, values for urea midpoint concentration ($C_{1/2}$, the urea concentration at which the unfolding free energy is zero), extracted from the analyses of kinetic profiles *versus* urea concentration at equilibrium conditions, have been used as a metric for stability. Figure 17 clearly shows that no correlation between *in vitro* folding rates and $C_{1/2}$ is apparent. Another interesting example that supports this results is the reconstruction of a globular protein formed by several β-trefoil subdomains, that exhibits extremely high thermal stability[157] (unfolds at a high temperature of ~94°C) but shows a remarkably slow folding kinetics[158] (its folding half-life is on the order of 1 hour).

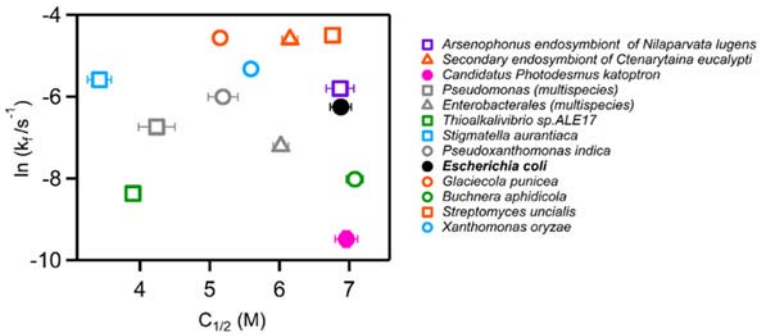


Figure 17. Correlation between *in vitro* folding and stability for modern thioredoxins. Plot of folding rate constant *versus* urea midpoint concentration for 13 modern thioredoxins (see section 5.1.3.). The values of the folding and unfolding rate constants were obtained interpolating the corresponding branches of the Chevron plots to 1 M urea. Stability was measured from fluorescence equilibrium profiles *versus* urea concentration where the $C_{1/2}$ values (the urea concentration at which unfolding free energy is zero) were calculated. (Reprinted with permission of Portland Press from Gamiz-Arco et al., 2019)[130].

Overall, we can conclude that exits a good correlation between efficiency of heterologous folding and *in vitro* folding rate, indicating that a reduction of the time the protein spends in partially unfolded states during the folding process is essential for an efficient heterologous expression. Moreover, stability play a role in folding rates and heterologous expression and an enhanced stability can sometimes produce an improvement in heterologous expression, but only when changes imply regions of the protein that are likely unfolded in intermediate states prone to aggregation.

In fact, previous attemps to improve heterologous expression by improving stability have been performed. In particular, in 2016, Fleishman´s group designed a novel structure and sequence-based methodology[159] (PROSS, http://pross.weizmann.ac.il) that can dramatically improve the protein stability and heterologous expression levels[159–168], using an automated algorithm based on Rosetta modeling and phylogenetic sequence information. Their strategy for improving heterologous expression is based on increasing the stability of proteins. Indeed, using this methodology heterologous expression can be improved, but often are required a large number of mutations. For example, in their seminal paper[159], they designed an acetylcholinesterase variant with 51 mutations that improved dramatically its expression in E. coli (~2,000-fold) and its stability (20°C higher). Our approach is really different because our hypothesis is based on the relationship between heterologous expression and in vitro refolding, and the rescue of inefficient heterologous is based on reducing the time proteins spend in partially-unfolded states prone to aggregation in the folding process. Moreover, our methodology achieve the expression efficiency with only 1–2 replacements, so protein changes are minimal compared with the high number of mutations used by PROSS. Our approach can be

especially useful in metagenomics, where is crucial maintain the unique properties of enzymes isolated from uncultured microbes.

6. CONCLUSIONS

The following main conclusions can be extracted from the studies presented as part of this doctoral thesis:

1. Ancestral thioredoxins fold faster and more efficiently *in vitro* than their modern counterparts.

2. Based on an extensive mutational analysis, we have unambiguously identified mutation Ser74Gly as responsible for aggravating folding in *E. coli* thioredoxin and described in detailed the molecular basis for its effect on its folding rate. The evolutionary acceptance of this mutation is rationalized invoking the protecting role of the cellular folding-assistance machinery. We propose that at a primordial stage, protein folding was unassisted and unprotected. Folding efficiency was then linked to fast folding encoded at the sequence/structure level. Once an efficient assistance machinery had emerged, mutations that impaired ancient sequence/structure determinants of folding efficiency could be accepted, since those determinants were no longer necessary. Ser/Gly replacement reveals as an example of degradation of ancestral features at the molecular level.

3. Contrary to previous claims in the literature, *in vitro* folding rates are not evolutionarily conserved in the thioredoxin family, indicating different degrees of degradation within modern proteins. Overall, our results support that *in vitro* protein folding is, to some extent, disconnected from *in vivo* protein folding. *In vitro* folding landscapes for some modern proteins may then represent moderately degraded versions of the landscapes from the unassisted folding of their ancestors.

4. We have rescued the inefficient heterologous expression of thioredoxin from *Candidatus Photodesmus katoptron*, an obligate symbiont of flashlight fish, using a few back-to-the-ancestral mutations at positions selected by computational modelling of the unassisted folding landscape. Our results support that the folding of heterologous proteins in foreign hosts may be akin to some extent to unassisted folding due to the

absence of coevolution of the recombinant protein with the natural chaperones of the new host.

5. We proposed a successful methodology to rescue inefficient heterologous expression with a minimal perturbation in the protein.

6. A promising scaffold for the design of proteins with new catalytic activities has been obtained and characterized (See Appendix). We have resurrected an ancestral glycosidase with relevant properties (enhanced stability and conformational flexibility) that may contribute to enzyme evolvability. Remarkably, we have shown that the ancestral glycosidase binds heme tightly and stoichiometrically at a specific site. Furthermore, we have demonstrated that heme binding allosterically enhances catalysis. Overall, our results support the potential of the ancestral scaffold for custom catalysis and biosensor engineering.

6. CONCLUSIONES

De los estudios presentados en esta tesis doctoral se pueden extraer las siguientes conclusiones:

1. Las tiorredoxinas ancestrales presentan un plegamiento más rápido y eficiente *in vitro* que sus homólogas modernas.

2. Un extenso análisis mutacional nos ha permitido identificar que una única mutación ancestral, Gly74Ser, es la responsable del plegamiento lento *in vitro* de la tiorredoxina de *E. coli* y se ha descrito en detalle la base molecular de su efecto en la velocidad de plegamiento. Esta mutación fue aceptada en la evolución cuando apareció la maquinaria de asistencia al plegamiento en las células. De esta manera, nosotros proponemos que, en una etapa primordial, las proteínas se plegaban en el medio celular sin asistencia ni protección. Esta eficacia en el plegamiento estaba ligada a un plegamiento rápido codificado al nivel de la secuencia/estructura. Una vez que apareció la maquinaria de asistencia para el plegamiento de las proteínas, éstas fueron aceptando mutaciones que dañaban los determinantes de la eficiencia de plegamiento a nivel de secuencia/estructura. La mutación Ser/Gly es un claro ejemplo de esto, de la degradación de una característica ancestral a nivel molecular en la evolución.

3. Al contrario de lo que ha sido recientemente afirmado en la literatura, las tasas de plegamiento *in vitro* no se conservan evolutivamente en la familia de las tiorredoxinas, lo que indica diferentes grados de degradación dentro de las proteínas modernas. En general, nuestros resultados apoyan que el plegamiento de proteínas *in vitro* está, hasta cierto punto, desconectado del plegamiento *in vivo*. Los paisajes de plegamiento *in vitro* para algunas proteínas modernas pueden representar versiones moderadamente degradadas de los paisajes de plegamiento sin asistencia de sus antepasados.

4. Hemos rescatado la expresión heteróloga ineficiente de la tiorredoxina de *Candidatus Photodesmus katoptron*, un simbionte obligado del pez linterna, mediante mutaciones de vuelta al ancestro en posiciones clave. Estas posiciones han sido seleccionadas mediante el modelado computacional del paisaje de plegamiento no asistido. Nuestros resultados apoyan que el plegamiento de proteínas en huéspedes heterólogos puede ser similar, en cierto modo, al plegamiento no asistido, debido a la ausencia de

coevolución de la proteína recombinante con las chaperonas naturales del nuevo huésped.

5. Desarrollo de una metodología cuyo objetivo es el rescate de la expresión heteróloga ineficaz, alterando de forma mínima la proteína.

6. Resurrección y caracterización de una glicosidasa ancestral con el objetivo de encontrar un buen punto de partida para la generación de proteínas con nuevas funcionalidades. Nuestra glicosidasa ancestral presenta excelentes propiedades (estabilidad mejorada, promiscuidad catalítica y flexibilidad conformacional) que pueden contribuir a su evolvabilidad. El hallazgo más sorprendente fue que la glicosidasa ancestral era capaz de unir hemo, de forma firme y estequiométrica, en un sitio específico. Además, se ha comprobado que la unión alostérica de hemo mejora la catálisis. En general, nuestros resultados apoyan el uso de esta glicosidasa ancestral como buen punto de partida para la generación de enzimas con nuevas actividades enzimáticas y en la ingeniería de biosensores.

7. PERSONAL CONTRIBUTION

FAST FOLDING AND SLOW UNFOLDING OF A RESURRECTED PRECAMBRIAN PROTEIN

- Contribution in unfolding and folding kinetics assays
- Contribution in double-jump unfolding assay

NON-CONSERVATION OF FOLDING RATES IN THE THIOREDOXIN FAMILY REVEALS DEGRADATION OF ANCESTRAL UNASSISTED-FOLDING

- Site-direct mutagenesis
- Proteins expression and purification
- Activity measurements
- Thermal Stability Studies
- Unfolding and folding kinetics
- Double-jump unfolding assays

COMBINING ANCESTRAL RECONSTRUCTION WITH FOLDING-LANDSCAPE SIMULATIONS TO ENGINEER HETEROLOGOUS PROTEIN EXPRESSION

- Site-direct mutagenesis
- Proteins expression and purification
- Activity measurements
- Chemical and thermal denaturation studies
- Unfolding and folding kinetics
- Double-jump unfolding assays

HEME-BINDING ENABLES ALLOSTERIC MODULATION IN AN ANCIENT TIM-BARREL GLYCOSIDASE (see Appendix)

- Exhaustive study of TIM-barrel families in CATH database for ancestral sequence reconstruction
- Proteins expression and purification
- Determination of the secondary structure and correct folding by circular dichroism

- Determination of association state through gel filtration chromatography
- Thermal denaturation studies
- Profiles of optimum activity temperature
- Database search for optimum temperature values for family-1 glycosidases
- Michaelis Menten kinetics of modern and ancestral glycosidases without heme of β-glucosidase and β-galactosidase activity
- Activity measurements using p-Nitrophenyl linked substrates of ancestral glycosidase wihout heme

APPENDIX

Ancestral glycosidases as molecular scaffolds for designing new catalytic properties

A publication entitled “Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase” (Gamiz-Arco, G., Gutierrez-Rus, L.I., Risso V.A., Ibarra-Molero, B., Hoshino, Y., Petrović, D., Justicia, J., Cuerva, J.M., Romero-Rivera, A., Seelig, B., Gavira, J.A., Kamerlin, S.C.L., Gaucher, E.A., Sanchez-Ruiz, J.M. Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase. Nat Commun **12**, 380 (2021). doi:10.1038/s41467-020-20630-1) is included in this appendix. Please, note that it is NOT part of the compendium of publications that the present thesis conforms to.

In this appendix, we discuss part of the results obtained in an ambitious and collaborative project funded by the prestigious Human Frontier Science Program (HFSP). The research groups involved are led by: Dr. Sanchez-Ruiz (IP; Universidad de Granada), Dr. Gaucher (Georgia State University), Dr. Kamerlin (Uppsala University) and Dr. Seelig (University of Minnesota). The results of this collaboration have been recently published in Nature Communicatons " (Gamiz-Arco, G., Gutierrez-Rus, L.I., Risso V.A., Ibarra-Molero, B., Hoshino, Y., Petrović, D., Justicia, J., Cuerva, J.M., Romero-Rivera, A., Seelig, B., Gavira, J.A., Kamerlin, S.C.L., Gaucher, E.A., Sanchez-Ruiz, J.M. Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase. Nat Commun **12**, 380 (2021). doi:10.1038/s41467-020-20630-1) and the corresponding publication is attached to this appendix.

The ultimate goal of this project is to generate enzymes with new activities using ancestral proteins as molecular scaffolds. As we have mentioned in the Introduction, ancestral proteins display relevant properties (enhanced stability, catalytic promiscuity and conformational flexibility)[9,10,35,36,11,13,29–34], that may contribute to enzyme evolvability. Such broad properties allow them to better search through functional landscapes in a more efficient manner and that is why ancestral proteins seem to be very convenient candidates as starting points for protein engineering[28,169]. Results included in the Nature Communications paper represent the first step of this enterprise, namely, obtaining what it seems to be an excellent scaffold for generating new functionalities. In addition, its complete biophysical characterization has been carried out.

Selecting the appropriate model system

Selecting the appropriate model system for our studies was one of the most important task, for obvious reasons. We selected the TIM-barrel fold because it is the most common protein fold, accounting for about 10% of known protein structures, and it can efficiently catalyze a wide variety of reactions[170] thanks to its versatile "catalitic face". The TIM-barrel fold (Figure 18) consists of central eight-stranded parallel β-sheets surrounded by eight α-helices. The TIM barrel fold has a "catalytic face" and a "stability face". The "catalytic face" contains the active site and it is formed by residues at the C-terminal ends of the β-strands and the loops that connect the β-strands with the subsequent α-helices. The "stability face" is actually the opposite face of the barrel and it is formed by the loops at the N-terminal portions of the β-strands. This face is important for conformational stability[171].

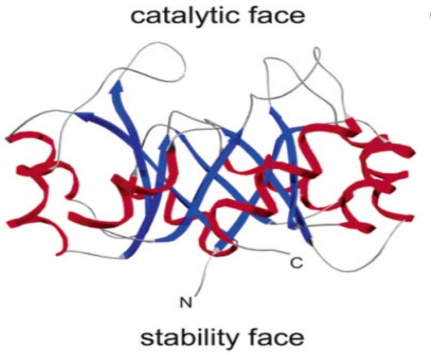


Figure 18. Schematic description of the TIM barrel fold. TIM barrel fold is formed by eight-stranded parallel β-sheets (blue) surrounded by eight α-helices (red). (Reprinted, with permission of American Chemical Society, from Sterner and Höcker, 2005)[171].

TIM-barrel may serve as an ideal scaffold for the generation of proteins having novel functions, although modern TIM-barrels proteins are fine-tuned by natural selection and highly specialized to perform specific reactions. By contrast, ancient proteins, as we have indicated throughout the present thesis, serve as enzymatic generalists. To select the family of TIM-barrel proteins for ancestral sequence reconstruction, we have exhaustively examined the CATH database[103]. CATH classifies proteins by their domains, based on the sequence, structural and functional information[172,173] and groups protein domains into superfamilies when there is sufficient evidence that they share a common ancestor. At the time of our study (June 2017), TIM-barrel proteins were classified in CATH in 29 superfamilies. Most TIM-barrel superfamilies are divided into several functional families according to the reactions they catalyze. We have studied each of these 29 superfamilies and their functional families, using very precise selection criteria: we selected only families with a considerable number of representatives (sequences and 3D structures), monomeric proteins and without the requirement of metals or cofactors. Moreover, the sequences of our selected families had to cover a wide range of taxa in order to have a complete phylogenetic tree that targeted very ancient phylogenetic nodes. The superfamily that contained the largest number of functional families best fitted to our requirements was the glycosidase superfamily. Within this superfamily, we found 9 functional families that were good candidates for ancestral sequence reconstruction. In each functional family, we found a representative structure that we could classify in CAZy database (Carbohydrate-Active enZYmes Database; http://www.cazy.org)[107]. CAZy is a classification of all the glycosidases based on the amino acid sequence and folding similarities. Our 9 functional families corresponded to 6 CAZy families (GH5, GH10, GH17, GH18, GH25 and GH53), that were selected for reconstruction. Studying in detail CAZy, we noted that CATH database did not cover all the TIM-barrel glycosidase families found in CAZy. Our previous selection could be incomplete. Therefore, we studied CAZy in detail and we found additional families, which also

fulfilled the previous requirements for ancestral glycosidase reconstruction, such as GH1, GH5, GH13, GH26, GH39 and GH51.

Glycosidases are involved in numerous biological and biotechnological processes. They are responsible for the hydrolysis of glycosidic linkages in a wide diversity of molecules[104]. Glycosidic bonds are one of the strongest bonds in nature[105] and glycosidases can accelerate it by factors approaching 10^{17}-fold[174]. GH are present in almost all living organisms where they play diverse and different roles[107]. Their families are categorized into two classes: retaining or inverting, depending if there is a change in the anomeric oxygen configuration during the hydrolysis of the glycosidic bond[175]. These two mechanisms of glycosidases were postulated by Koshland in 1953[176].

Ancestral sequence reconstruction of glycosidases was performed by our collaborator, Dr. Eric Gaucher (Georgia State University, Atlanta, USA) who resurrected the sequences corresponding to a total of 18 nodes of the selected families. We did several attempts to resurrect all these ancestral glycosidases in the laboratory, but nodes from family 17 and 18 were impossible to over-express in *E. coli* cells. Once the ancestral glycosidases were resurrected, their stability was measured by differential scanning calorimetry. Results showed that ancestral glycosidases from node 72 and 73, both belonging to glycosidases family 1, were the most thermostable proteins. Moreover, we studied another ancestral glycosidase from family 1, the node 125. This node displayed a lower stability than the previously mentioned proteins (its denaturation temperature was ~ 10 degrees below). Finally, we selected the node 72 (from now on "ancestral glycosidase") for an exhaustive biochemical and biophysical characterization because of its thermostability and its high yield production in the laboratory. The other two nodes were difficult to purify as they had tendency to aggregate over time.

The sequence of our ancestral glycosidase differed considerably from the sequences of modern glycosidases. In fact, the closest hit in a BLAST (Basic Local Alignment Search Tool) search had only a 62% sequence identity with our ancestral glycosidase. We expected that these huge sequence differences may be translated in different and surprising properties in the ancestral scaffold.

Our ancestral glycosidase belongs to family 1 glycosidase, one of the largest and taxonomically diverse families of glycosidases. Members of glycoside hydrolase family 1 (GH1) are widely distributed in all three domains of life. We exhaustively studied this family in the CAZy database, which currently includes 162 protein members for the GH1 family. Their most common enzymatic activities are β-glucosidases and β-galactosidases. However, other

commonly found activities are 6-phospho-β-glucosidase and 6-phospho-β-galactosidase, β-mannosidase, β-D-fucosidase and β-glucuronidase[177].

As a control for our study, we selected four modern bacterial family 1 glycosidases that are descendants from N72 and also listed in the CAZy database. Thus, we selected β-glucosidase from *Halothermothrix orenii,* β-glucosidase from *Marinomonas sp.* (strain MWYL1), β-glucosidase from *Saccharophagus degradans* (strain 2-40 T), and β-glucosidase from *Thermotoga maritime*. These modern GH1 glycosidases have been widely studied in bibliography in terms of stability and association state.

Biophysical characterization of ancestral glycosidase

An exhaustive biophysical and biochemical characterization of the ancestral glycosidase has been carried out in terms of relevant features for protein engineering, as described below:

- **Stability**

The stability of the ancestral glycosidase was determined by both, differential scanning calorimetry experiments and measurements of optimum activity temperature. The enzyme has a denaturation temperature of about 72 °C and an optimum temperature of 65 °C (Figure 18).

To put our results in context, an exhaustive bibliographic search for reported optimum temperatures on family 1 glycosidases has been performed. We found a linear relationship between the optimum activity temperature of the enzyme and the environmental temperature of their host organisms, showing that the optimum temperature of the enzyme is an indicative of the living temperature of the organisms (Figure 20). In addition, this representation reflects that our ancestral glycosidase has an optimum temperature within the range of optimum temperatures typical for thermophilic organisms. In particular, we have directly compared the stability of ancestral glycosidase and the modern glycosidase from the thermophilic *Halothermothrix orenii*, and our results proved that both stabilities are similar (Figure 19).

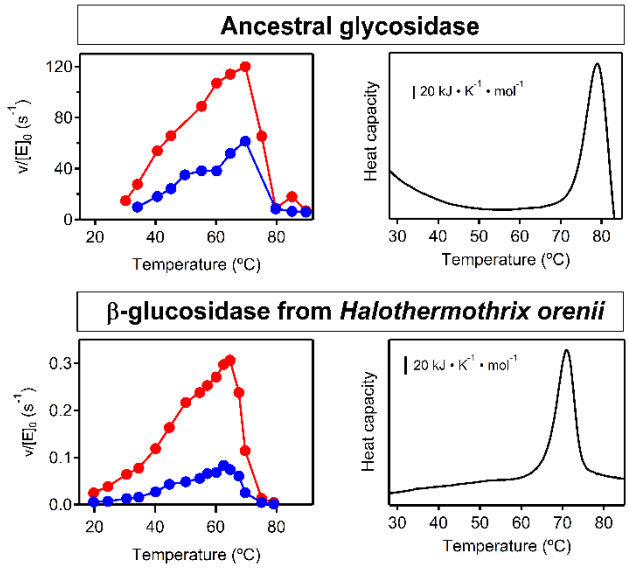


Figure 19. Stability of the ancestral glycosidase and the modern glycosidase from *Halothermothrix orenii*. Experimental determinations of optimum temperature (left panels) using two different substrates 4-nitrophenyl-β-D-glucopyranoside (red) and 4-nitrophenyl-β-D-galactopyranoside (blue). The right panels show differential scanning calorimetry profiles. (Reprinted and modified with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178].

Our results indicate that the living temperature of our glycosidase´s ancestral host was about 52 °C. This supports the fact that ancestral proteins display higher stability than their modern mesophilic counterparts, previously mentioned in the Introduction. This high stability provides essential information to understand the environment of their extinct microorganisms and may reflect the warmer temperatures of the Earth in the past [17,18].

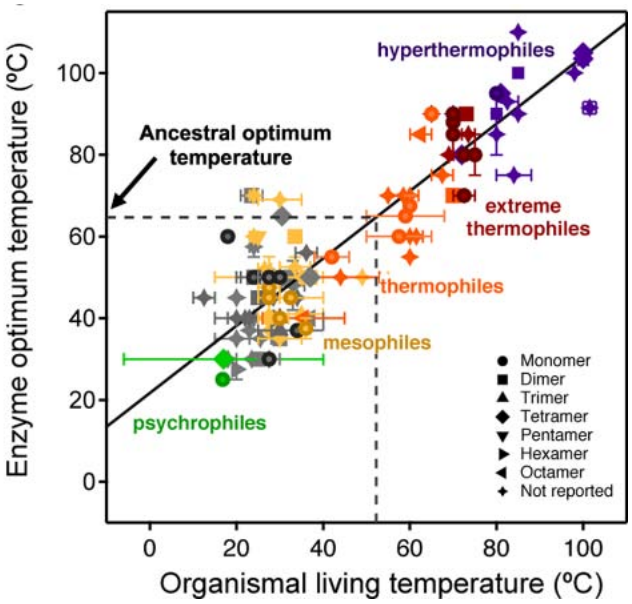


Figure 20. Optimum activity temperature for modern family 1 glycosidases. This plot shows enzyme optimum temperature *versus* living temperature of the host organism. Horizontal and vertical bars represent ranges of organismal living temperatures and enzyme optimum temperatures provided in the literature. Organisms were classified as hyperthermophiles, extreme thermophiles, thermophiles, mesophiles, psychrophiles as is published in literature. (Reprinted with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178]

- **Conformational flexibility**

Ancestral glycosidase displays large regions with greatly enhanced conformational flexibility, contrary to what is observed in the modern glycosidase from *Halothermothrix orenii*. This enhanced flexibility was demonstrated by both experimental and computational approaches:

Proteolysis

Ancestral glycosidase was highly susceptible to proteolysis using thermolysin, whereas the modern glycosidase remained essentially intact even when high concentrations of thermolysin were used. This pronounced difference is likely due to an enhanced conformational flexibility in ancestral protein and a subsequent exposure of the thermolysin cleavage sites.

3D structure

The 3D structure of ancestral glycosidase (PDB ID: 6Z1H) has been elucidated thanks to our collaborator Dr. José Antonio Gavira (Instituto Andaluz de Ciencias de la Tierra (CSIC-IACT), Granada, Spain). Remarkably, analysis of ancestral structure reveals two large missing regions (Figure 21), between residues 234 – 282 and 307- 331, in the electronic density map from X-ray crystallography. Probably, these missing regions are very flexible and are continuously changing within a diversity of protein conformations.

In addition, a direct comparison of B- factors values (Figure 20) corresponding to both structures indicate high flexibility in two alpha helices and several loops in the ancestral glycosidase. On the other hand, the barrel core shows low conformational flexibility in both structures, as expected.

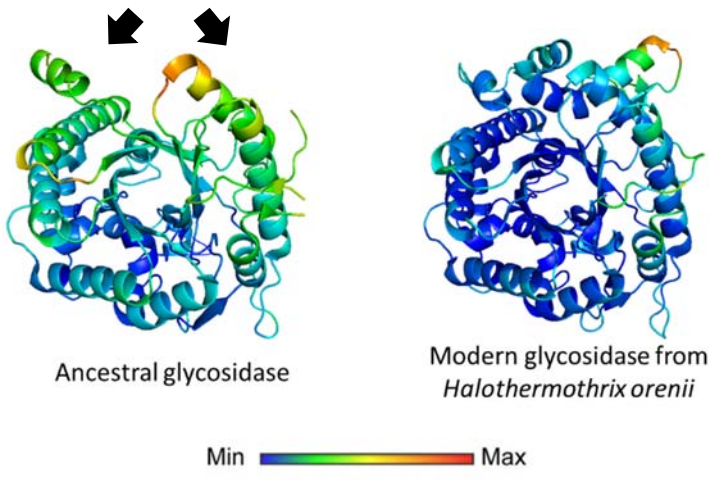


Figure 21. 3D structures of the ancestral glycosidase and modern glycosidase from *Halothermothrix orenii.* It is important to note the missing sections (marked with arrows) in the electronic density of the ancestral glycosidase with regard to the modern protein. Both proteins are coloured according to normalized B-factor.

Molecular dynamics simulations

Moreover, the enhanced flexibility of the ancestral protein was confirmed through molecular dynamics (MD) simulations performed by Dr. Lynn Kamerlin (Uppsala University, Sweden). MD simulations indicated enhanced flexibility in the region 227–334 in the ancestral glycosidase, compared with the modern glycosidase from *Halothermothrix orenii*. These experiments are in concordance with X-ray crystallography because the missing regions in the X-ray structure are included in the flexible regions shown by molecular dinamics simulations.

- **Catalysis and Promiscuity**

We have determined the β-glucosidase and β-galactosidase activities for the ancestral enzyme and the four modern control glycosidases (Figure 22).

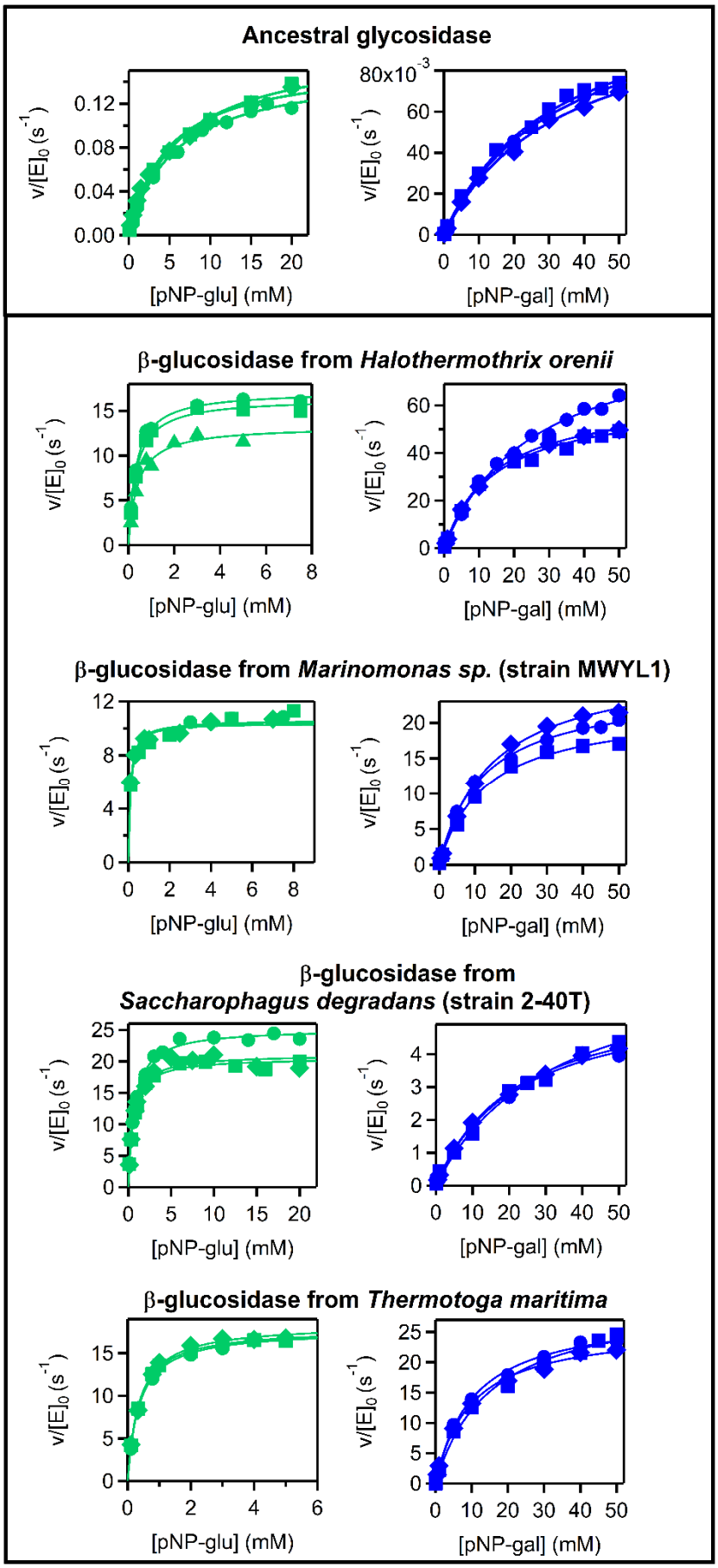


Figure 22. Michaelis plots for ancestral and modern glycosidases. Rate *versus* substrate concentration for the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl- β -D-galactopyranoside catalyzed by the ancestral glycosidase and the modern glycosidase from *Halothermothrix orenii*, *Marinomonas sp.* (strain MWYL1), *Saccharophagus degradans* and *Thermotoga maritime*. The different data points correspond to 3 experimental replicates The lines are the best fits to the Michaelis-Menten equation. (Reprinted and modified with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178].

Two interesting conclusions can be extracted:

- The catalytic efficiency of ancestral enzyme is lower, with a turnover number about two orders of magnitude below the values for the modern glycosidases (Figure 22). This result can be related to the higher conformational flexibility displayed by the ancestral glycosidase that implies a diversity of conformations of which only a few are active.

  The ancestral glycosidase does not display substrate specificity for β-glucosidase and β-galactosidase activities. Our ancestral glycosidase shows similar Km's values for the β-glucopyranoside and the β-galactopyranoside substrates. This fact is interesting because the family 1 glycosidases often reflects specialization in the substrate affinity, showing substantially higher Km values for the β-galactosidase activity than for the β-glucosidase activity[177]. We can clearly see this effect in the modern enzymes we have studied (Figure 22). This result reflects substrate promiscuity of the ancestral enzyme, as seen before in a number of other ancestral proteins[10,30]. A possible argument is that the ancestral glycosidase was specialized for a different substrate or for a different reaction.

It is worth mentioning, special efforts have been made in order to find promiscuity, even in very low levels, in our ancestral glycosidase. For this purpose, we tested substrate promiscuity of ancestral glycosidase against a wide range of substrates for which modern descendants are specialized (Figure 23) such as 4-nitrophenyl-β-D-fucopyranoside, 4-nitrophenyl-β-D-lactopyranoside, 4-nitrophenyl-β-D-xylopyranoside and 4-nitrophenyl-β-Dmannopyranoside. In all cases, we found that the levels of catalysis of the ancestral protein were reduced compared to the modern proteins.

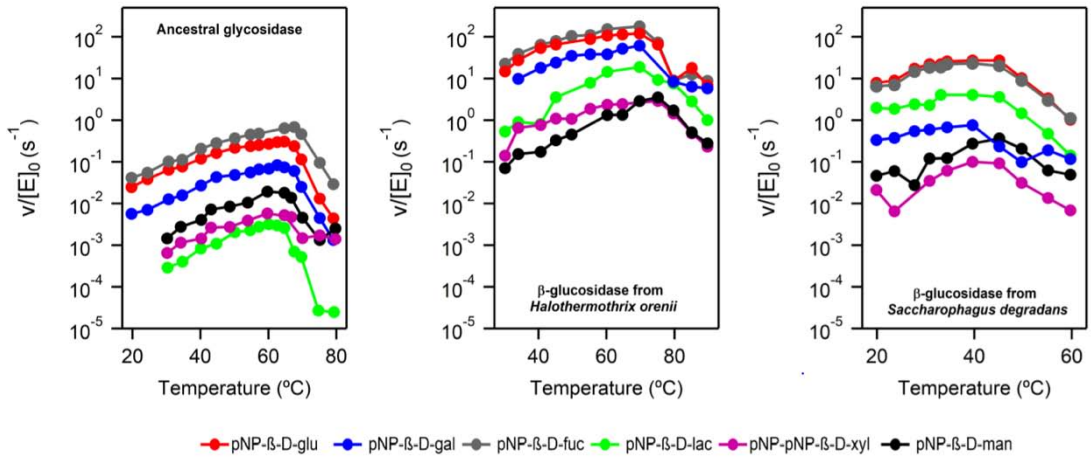


Figure 23. Profiles of activity versus temperature for the ancestral glycosidase (left) and two modern glycosidases. The activity of ancestral glycosidase (left) and the modern β-glucosidases from Halothermothrix orenii (middle) and Saccharophagus degradans (right) was measured using the following substrates: 4-nitrophenyl-β-D-glucopyranoside, 4-nitrophenyl-β-D-galactopyranoside, 4-nitrophenyl-β-D-fucopyranoside, 4-nitrophenyl-β-D-lactopyranoside, 4-nitrophenyl-β-D-xylopyranoside and 4-nitrophenyl-β-Dmannopyranoside. (Reprinted and modified with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178].

Finally, we also checked the catalytic activity against different substrates including derivatives of disaccharide (maltose, cellobiose), several substrates with an α anomeric carbon and 4-nitrophenyl-β-D-glucopyranoside6-phosphate to test for the 6-phosphate-β-glucosidase activity with no success. Actually, we have not found any substrate with a catalysis level substantially higher than in the modern glycosidases.

- **Heme binding: an unexpected property**

During the process of protein purification, we noticed that the preparations of purified ancestral glycosidase showed a light-reddish color. The absorption spectrum of the ancestral glycosidase showed an absorption peak around 400 nm corresponding to the Soret band (Figure 24), indicating that the glycosidase can bind a heme group[179]. The ratio between heme and protein was very low (≈ 0.02). This small heme:protein ratio suggests that all the previously described experiments were heme-free.

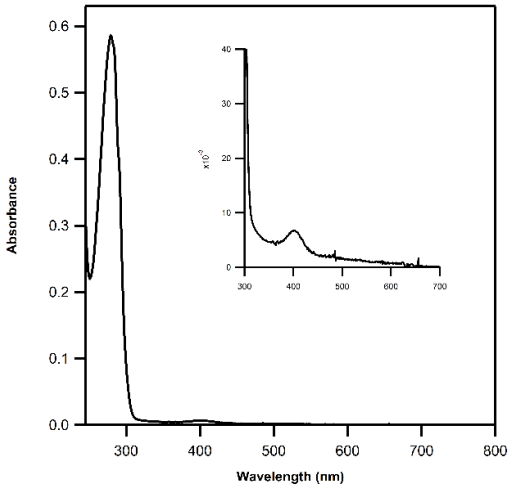


Figure 24. Heme binding to the ancestral glycosidase. Example of a UV–VIS spectra for a common preparation of the ancestral glycosidase. The plot shows the protein absorption band at about 280 nm and the Soret band at about 400 nm due to the presence of heme. (Reprinted and modified with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178].

The confirmation of the capability of the ancestral enzyme to bind heme, specifically, the binding of one heme per protein molecule, was shown by *in vitro* experiments, mass spectrometry and X-ray crystallography. Remarkably, the 3D structure of the ancestral glycosidase with heme has revealed that heme binding rigidifies the ancestral protein (Figure 25), so there are fewer missing regions in the 3D structure, in comparison to the structure of protein without heme.

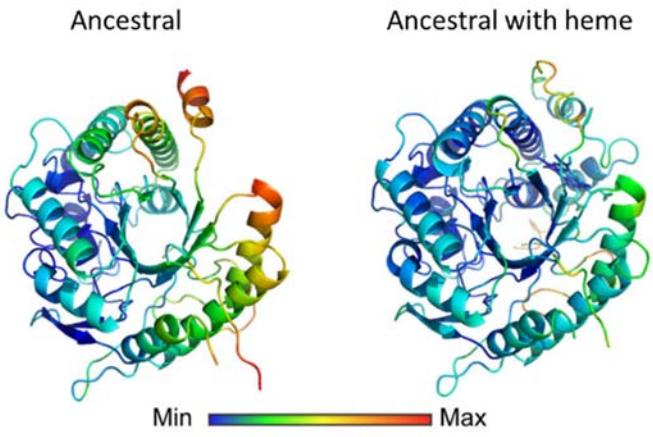


Figure 25. Comparison between the ancestral structure determined by X-ray crystallography in the absence (PDB ID: 6Z1H) and presence (PDB ID: 6Z1M) of bound heme. It is important to note that the missing sections in the electronic density are present mostly in the ancestral protein without heme. Both structures are coloured according to normalized B-factor values. (Reprinted and with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178].

Interestingly, the heme has an effect in enzymatic activity (Figure 26). Activity determinations with the heme-saturated ancestral enzyme show that heme increases the activity of the ancestral protein by ~3 fold.

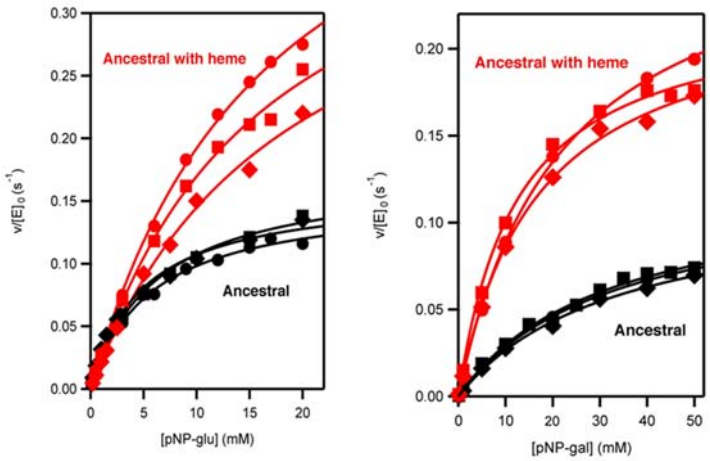


Figure 26. Effect of heme binding on catalytic activity of the ancestral glycosidase. Michaelis-Menten plots of rate versus substrate concentration for β-D-glucosidase activity (left) and β-galactosidase activity (right) catalyzed by the ancestral glycosidase with (red) and without heme bound (black). (Reprinted with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178].

A closer inspection of the 3D structure revealed that the heme is located near the enzyme active site but does not have direct access to it (Figure 27), so it is unlikely that the heme binding has a direct effect on catalysis. Therefore, the increase in activity observed upon heme binding is due to the restriction of movement of several of the residues implied in the catalysis or in substrate binding. Although these positions are not substantially altered, sub-Å changes in the protein may be responsible for the changes in the activity.

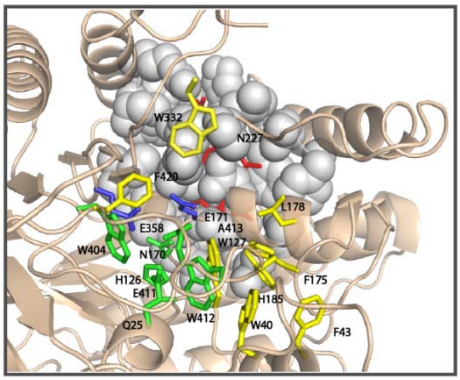


Figure 27. Active site of the ancestral glycosidase. In the image is highlighted the catalytic carboxylic acid residues (blue) and the residues involved in binding of the substrate (yellow and green). The residues that block the connection between the heme and the active site are marked with grey spheres. (Reprinted with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178].

The ability to bind heme in ancestral glycosidase was a completely surprising observation. To the best of our knowledge, no modern glycosidase binds heme or any porphyrin. At the time of our study (April 2020), none of the ~5500 reported crystallographic structures of ~1400 modern glycosidases reported in CAZy binds heme or a porphyrin.

We checked if heme binding to the ancestral glycosidase was an accidental issue promoted by its high conformational flexibility or, by contrast, was a functional ancestral feature. If this is the case, we may expect that some modern glycosidases and more recent ancestors of this family show some capability to bind heme, supporting to some extent the principle of the degradation of useless characteristics through the evolution (Section 5.1.)[180]. The results proved that modern glycosidases can bind heme but in an inefficient way (Figure 26), whereas the more recient nodes showed a heme-binding capability intermediate between modern and ancestral N72 glycosidase (Figure 28).

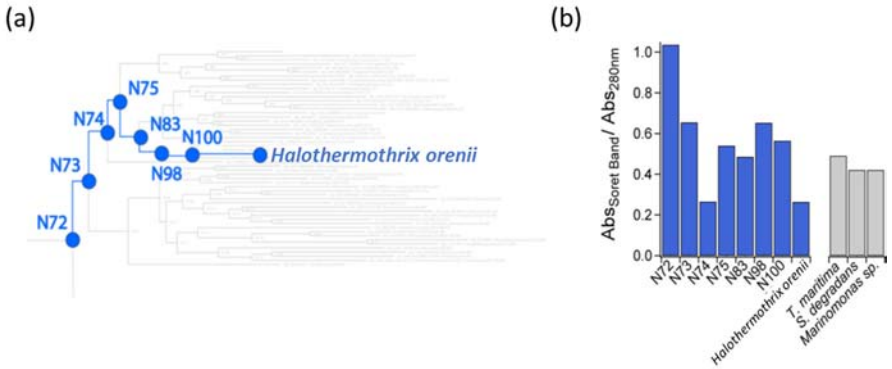


Figure 28. Degradation of heme binding in the family 1 glycosidases through evolution. (a) Section of the phylogenetic tree used for the reconstruction of family 1 glycosidases. The nodes in the evolutionary

trajectory from node 72 to *Halothermothrix orenii* are highlighted in blue. (b) Ratio of absorbance between the maximum of the heme Soret band and the absorbance at 280 nm (the maximum of the protein aromatic absorption band) for ancestral and modern family 1 glycosidases. (Reprinted and modified with permission of Springer Nature from Gamiz-Arco et al., 2021a)[178].

Overall, results described in this study provide fundamental information to understand the evolution of family 1 glycosidases and, by extrapolation, the evolution of the TIM barrel fold. Ancestral glycosidases bind specifically heme, although this feature has been degraded through evolution because it was no longer useful. Moreover, the features of the ancestral glycosidase (high stability, lower levels of activity, lack of specialization for substrates and high conformational flexibility) seem to be consistent with an ancestral protein that existed in an early stage in the evolution of family 1 glycosidases. Knowing all this information, we propose three different alternatives to the evolution of the TIM-barrel fold:

(i) A number of authors conclude that the TIM barrel fold have evolved from previously existing partial barrels through recombination and duplication [181182183184]. Our results seem to be consistent with this proposal because the resulting protein after fragment fusion is expected to be flexible, one of the principal characteristics of our ancestral glycosidase.

(ii) Another theory about the evolution of TIM-barrel fold proposes that the glycosidase scaffold was formed by a fusion event that implied a heme-binding domain. In this scenario, heme had a functional role in the heme-containing domain, which was no longer useful when the domain was fused with the larger glycosidase scaffold.

(iii) Heme is present in the ancestral glycosidase because heme is a cofactor that facilitated the origin of family 1 glycosidases by selecting them from a random pool of amino acid sequences[185].

In addition to evolutionary implications, our results have potential impact in biotechnological applications of enzymes. We have achieved our initial goal proposed for this project: the resurrection of an ancestral protein that display convenient properties (enhanced stability, catalytic promiscuity and conformational flexibility) for its use as molecular scaffold in protein engineering. Moreover, we have found a surprising feature in this ancestral protein that

opens up new possibilities due to the catalytic versatility of porphyrins and their use in protein engineering for the catalysis of non-natural reactions[186].

Moreover, the allosteric modulation of catalysis due heme binding suggests potential applications of the ancestral scaffold in engineering of biosensors. The heme binding site could be redesigned to bind a substance of interest that could tentatively produce large catalytic changes in glycosidase activity and thus be easily detected.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20630-1 OPEN

Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase

Gloria Gamiz-Arco[1,8], Luis I. Gutierrez-Rus[1,8], Valeria A. Risso[1], Beatriz Ibarra-Molero[1], Yosuke Hoshino[2], Dušan Petrović[3,7], Jose Justicia[4], Juan Manuel Cuerva[4], Adrian Romero-Rivera[3], Burckhard Seelig[5], Jose A. Gavira[6], Shina C. L. Kamerlin[3✉], Eric A. Gaucher[2✉] & Jose M. Sanchez-Ruiz[1✉]

Glycosidases are phylogenetically widely distributed enzymes that are crucial for the cleavage of glycosidic bonds. Here, we present the exceptional properties of a putative ancestor of bacterial and eukaryotic family-1 glycosidases. The ancestral protein shares the TIM-barrel fold with its modern descendants but displays large regions with greatly enhanced conformational flexibility. Yet, the barrel core remains comparatively rigid and the ancestral glycosidase activity is stable, with an optimum temperature within the experimental range for thermophilic family-1 glycosidases. None of the ~5500 reported crystallographic structures of ~1400 modern glycosidases show a bound porphyrin. Remarkably, the ancestral glycosidase binds heme tightly and stoichiometrically at a well-defined buried site. Heme binding rigidifies this TIM-barrel and allosterically enhances catalysis. Our work demonstrates the capability of ancestral protein reconstructions to reveal valuable but unexpected biomolecular features when sampling distant sequence space. The potential of the ancestral glycosidase as a scaffold for custom catalysis and biosensor engineering is discussed.

[1] Departamento de Quimica Fisica. Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain. [2] Department of Biology, Georgia State University, Atlanta, GA 30303, USA. [3] Science for Life Laboratory, Department of Chemistry-BMC, Uppsala University, BMC Box 576, S-751 23 Uppsala, Sweden. [4] Departamento de Quimica Organica. Facultad de Ciencias, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, 18071 Granada, Spain. [5] Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, Minnesota, United States of America, & BioTechnology Institute, University of Minnesota, St. Paul, MN, USA. [6] Laboratorio de Estudios Cristalograficos, Instituto Andaluz de Ciencias de la Tierra, CSIC, Unidad de Excelencia de Quimica Aplicada a Biomedicina y Medioambiente (UEQ), Universidad de Granada, Avenida de las Palmeras 4, Granada 18100 Armilla, Spain. [7] Present address: Hit Discovery, Discovery Sciences, Biopharmaceutical R&D, AstraZeneca 431 50 Gothenburg, Sweden. [8] These authors contributed equally: Gloria Gamiz-Arco, Luis I. Gutierrez-Rus. ✉email: lynn.kamerlin@kemi.uu.se; egaucher@gsu.edu; sanchezr@ugr.es

Pauling and Zuckerkandl proposed in 1963 that the sequences of modern protein homologs could be used to reconstruct the sequences of their ancestors[1]. While this was mostly only a theoretical possibility in the mid-twentieth century, ancestral sequence reconstruction has become a standard procedure in the twenty-first century, due to advances in bioinformatics and phylogenetics, together with the availability of increasingly large sequence databases. Indeed, in the last ~20 years, proteins encoded by reconstructed ancestral sequences ("resurrected" ancestral proteins, in the common jargon of the field) have been extensively used as tools to address important problems in molecular evolution[2,3]. In addition, a new and important implication of sequence reconstruction is currently emerging linked to the realization that resurrected ancestral proteins may display properties that are desirable in scaffolds for enzyme engineering[4–6]. For instance, high stability and substrate/catalytic promiscuity have been described in a number of ancestral resurrection studies[5,7]. These two features are known contributors to protein evolvability[8,9], which points to the potential of resurrected ancestral proteins as scaffolds for the engineering of new functionalities[4,10].

More generally, reconstruction studies that target ancient phylogenetic nodes typically predict extensive sequence differences with respect to their modern proteins. Consequently, proteins encoded by the reconstructed sequences may potentially display altered or unusual properties. Regardless of the possible evolutionary implications, it is of interest, therefore, to investigate which properties of putative ancestral proteins may differ from those of their modern counterparts and to explore whether and how these ancestral properties may lead to new possibilities in biotechnological applications. Here, we apply ancestral sequence reconstruction to a family of well known and extensively characterized enzymes. Furthermore, these enzymes display 3D-structures based on the highly common and widely studied TIM-barrel fold, a fold which is both ubiquitous and highly evolvable[11–13]. Yet, we find upon ancestral resurrection a diversity of unusual and unexpected biomolecular properties that suggest new engineering possibilities that go beyond the typical applications of protein family being characterized.

Glycosidases catalyze the hydrolysis of glycosidic bonds in a wide diversity of molecules[14]. The process typically follows a Koshland mechanism based on two catalytic carboxylic acid residues and, with very few exceptions, does not involve cofactors. Glycosidic bonds are very stable and have an extremely low rate of spontaneous hydrolysis[15]. Glycosidases accelerate their hydrolysis up to ~17 orders of magnitude, being some of the most proficient enzymes functionally characterized[16]. Glycosidases are phylogenetically widely distributed enzymes. It has been estimated, for instance, that about 3% of the human genome encodes glycosidases[17]. They have been extensively studied, partly because of their many biotechnological applications[14]. Detailed information about glycosidases is collected in the public CAZy database (Carbohydrate-Active enZYmes Database; http://www.cazy.org)[18] and the connected CAZypedia resource (http://www.cazypedia.org/)[19]. At the time of our study, glycosidases are classified into 167 families on the basis of sequence similarity. Since perturbations of protein structure during evolution typically occur more slowly than sequences change[20], it is not surprising that the overall protein fold is conserved within each family. Forty eight of the currently described glycosidase families display a fold consistent with the TIM barrel architecture. Often, common ancestry between different TIM-barrel families cannot be unambiguously demonstrated[12]. Therefore, the TIM-barrel may be considered as a "superfold" in the sense of Orengo et al.[21], and simply sharing this fold does not necessarily imply evolutionary relatedness.

Here, we study family 1 glycosidases, which are of the classical TIM-barrel fold. Family 1 glycosidases (GH1) commonly function as β-glucosidases and β-galactosidases, although other activities are also found in the family[22]. GH1 enzymes are present in the three domains of life and have been traced back to LUCA[23]. We focus on a putative ancestor of modern bacterial and eukaryotic enzymes and find a number of unusual properties that clearly differentiate the ancestor from the properties of its modern descendants. The ancestral glycosidase thus displays much-enhanced conformational flexibility in large regions of its structure. This flexibility, however, does not compromise stability as shown by the ancestral optimum activity temperature which is within the typical range for family 1 glycosidases from thermophilic organisms. Unexpectedly, the ancestral glycosidase binds heme tightly at a well-defined site in the structure with concomitant allosteric increase in enzyme activity. Neither metalloporphyrin binding nor allosteric modulation appears to have been reported for any modern glycosidases, despite the fact that these enzymes have been extensively characterized. Overall, this work demonstrates the potential of ancestral reconstruction as a tool to explore sequence space to generate combinations of properties that are unusual or unexpected compared to the repertoire from modern proteins.

Results

Ancestral sequence reconstruction. Ancestral sequence reconstruction (ASR) was performed based on a phylogenetic analysis of family 1 glycosidase (GH1) protein sequences (see the Methods for details). GH1 protein homologs are widely distributed in all three domains of life and representative sequences were collected from each domain, including characterized GH1 sequences obtained from CAZy as well as homologous sequences contained in GenBank. The phylogeny of GH1 homologs consists of four major clades (Fig. 1a and Fig. S1). One clade is composed mainly of archaea and bacteria from the recently proposed Candidate Phyla Radiation (CPR)[24], while the other three clades include bacteria and eukaryotes. The archaeal/CPR clade largely contains uncharacterized proteins and was thus excluded from further analysis. In the bacterial/eukaryotic clades, eukaryotic homologs form a monophyletic clade within bacterial homologs. For our current study, the common ancestors of bacterial and eukaryotic homologs are selected for ASR analysis (N72, N73, and N125) because many homologs have been characterized and there is substrate diversity between the enzymes in the different clades.

Selection of an ancestral glycosidase for experimental characterization. We prepared, synthesized, and purified the proteins encoded by the most probabilistic sequences at nodes N72, N73, and N125. While the three proteins were active and stable, we found that those corresponding to N73 and N125 had a tendency to aggregate over time. We therefore selected the resurrected protein from node 72 for exhaustive biochemical and biophysical characterization. For the sake of simplicity, we will subsequently refer to this protein as the ancestral glycosidase in the current study.

It is important to note that the sequence of the ancestral glycosidase differs considerably from the sequences of modern proteins. The set of modern sequences used as a basis for ancestral reconstruction span a range of sequence identity (26–59%) with the ancestral glycosidase (Table S1). Also, using the ancestral sequence as the query of a BLAST (Basic Local Alignment Search Tool) search in several databases (non-redundant protein sequences, UniProtKB/Swiss-Prot, Protein Data Bank, Metagenomic proteins) yields a closest hit with only a 62% sequence identity to the ancestral glycosidase. These sequence differences translate into unexpected biomolecular properties.

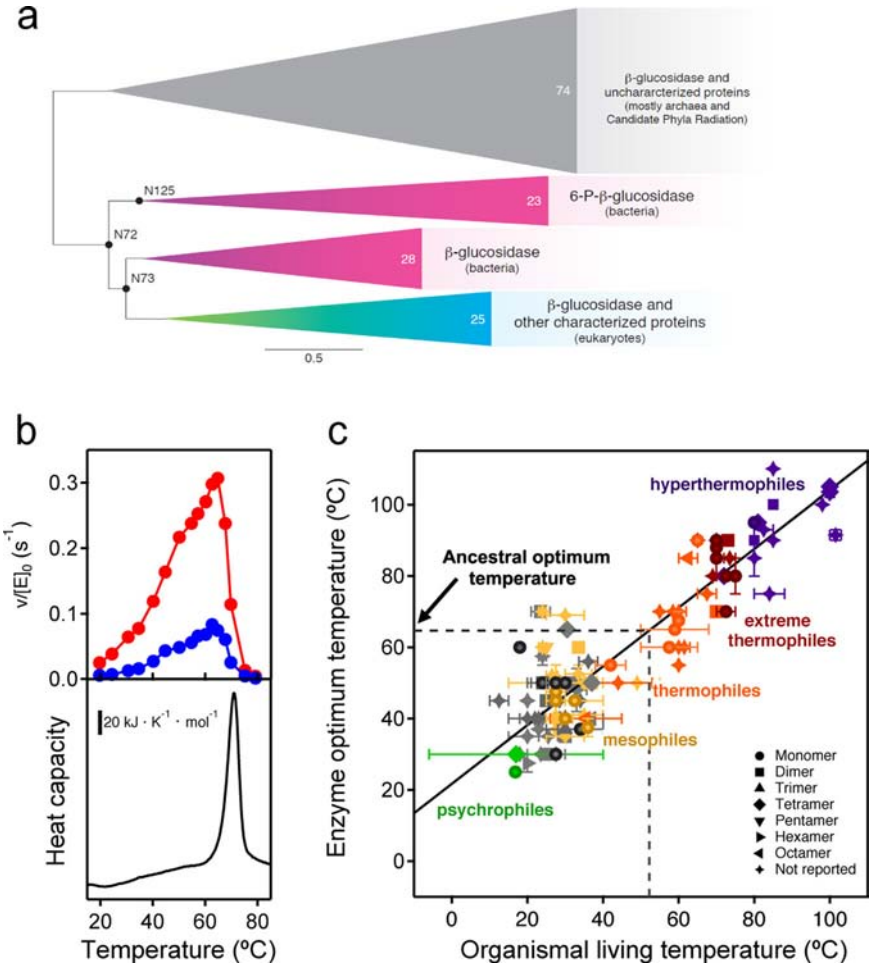


Fig. 1 Ancestral sequence reconstruction of family 1 glycosidases (GH1) and assessment of ancestral stability. a Bayesian phylogenetic tree of GH1 protein sequences using 150 representative sequences. Triangles correspond to four major well-supported clades (see supplemental Fig. S1 for nodal support) with common functions indicated. Numbers inside the triangles correspond to the number of sequences in each clade. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. Reconstructed ancestral sequences were inferred at the labeled nodes and the protein at node 72 was exhaustively characterized. **b** Determination of the optimum temperature for the ancestral glycosidase (upper panel) using two different substrates 4-nitrophenyl-β-D-glucopyranoside (red) and 4-nitrophenyl-β-D-galactopyranoside (blue). $v/[E]_0$ stands for the rate over the total enzyme concentration. The lower panel shows a differential scanning calorimetry profile for the ancestral glycosidase. Clearly, the activity drop observed at high temperature (upper panel) corresponds to the denaturation of the protein, as seen in the lower panel. **c** Plot of enzyme optimum temperature versus living temperature of the host organism for modern family 1 glycosidases. Data (Supplementary Dataset 1) are derived from literature searches, as described in Methods. Horizontal and vertical bars are not error bars, but represent ranges of organismal living temperatures and enzyme optimum temperatures when provided in the literature. Color code denotes the organisms that published literature describes as hyperthermophiles, extreme thermophiles, thermophiles, mesophiles, psychrophiles; gray color is used for organisms that have not been thus classified (plants that live at moderate temperatures in most cases). The line is a linear-squares fit ($T_{OPT} = 21.68 + 0.824T_{LIVING}$). Correlation coefficient is 0.89 and $p \sim 8.8 \times 10^{-45}$ (probability that the correlation results from chance). An environmental temperature of about 52 °C can be estimated from the optimum temperature of the ancestral glycosidase.

Stability. As it is customary in the glycosidase field, we assessed the stability of the ancestral glycosidase using profiles of activity versus temperature determined by incubation assays[25]. These profiles typically reveal a well-defined optimum activity temperature (Fig. 1b) as a result of the concurrence of two effects. At low temperatures, the expected Arrhenius-like increase of activity with temperature is observed. At high temperatures, protein denaturation occurs and causes a sharp decrease in activity. For the ancestral glycosidase, this interpretation is supported by differential scanning calorimetry data (lower panel in Fig. 1b) which show a denaturation transition that spans the temperature range in which the activity drops sharply.

The profiles of activity versus temperature (Fig. 1b) show a sharp maximum at 65 °C for the optimum activity temperature of the ancestral glycosidase. In order to ascertain the implications of this value for the evaluation of the ancestral stability, we have searched the literature on family 1 glycosidases for reported optimum temperature values (see Methods for details and Supplementary Dataset 1 for the results of the search). The values for ~130 modern enzymes show a good correlation with the environmental temperature of their respective host organisms (Fig. 1c). Therefore, the enzyme optimum temperature is an appropriate reflection of stability in an environmental context for this protein family. The optimum temperature value for ancestral glycosidase is within the experimental range of optimum activity temperatures for family 1 glycosidases from thermophilic organisms and it is consistent with an ancestral environmental temperature of about 52 °C (Fig. 1c).

Conformational flexibility. Remarkably, despite its "thermophilic" stability, large regions in the structure of the ancestral

glycosidase are flexible and/or unstructured, as demonstrated by both experiment and computation (Fig. 2).

Proteolysis is known to provide a suitable probe of conformational diversity and the protein energy landscape[26], since most cleavable sites are not exposed in folded compact protein states. The ancestral glycosidase is highly susceptible to proteolysis and degradation is already apparent after only a few minutes incubation at a low concentration of thermolysin (0.01 mg/mL,

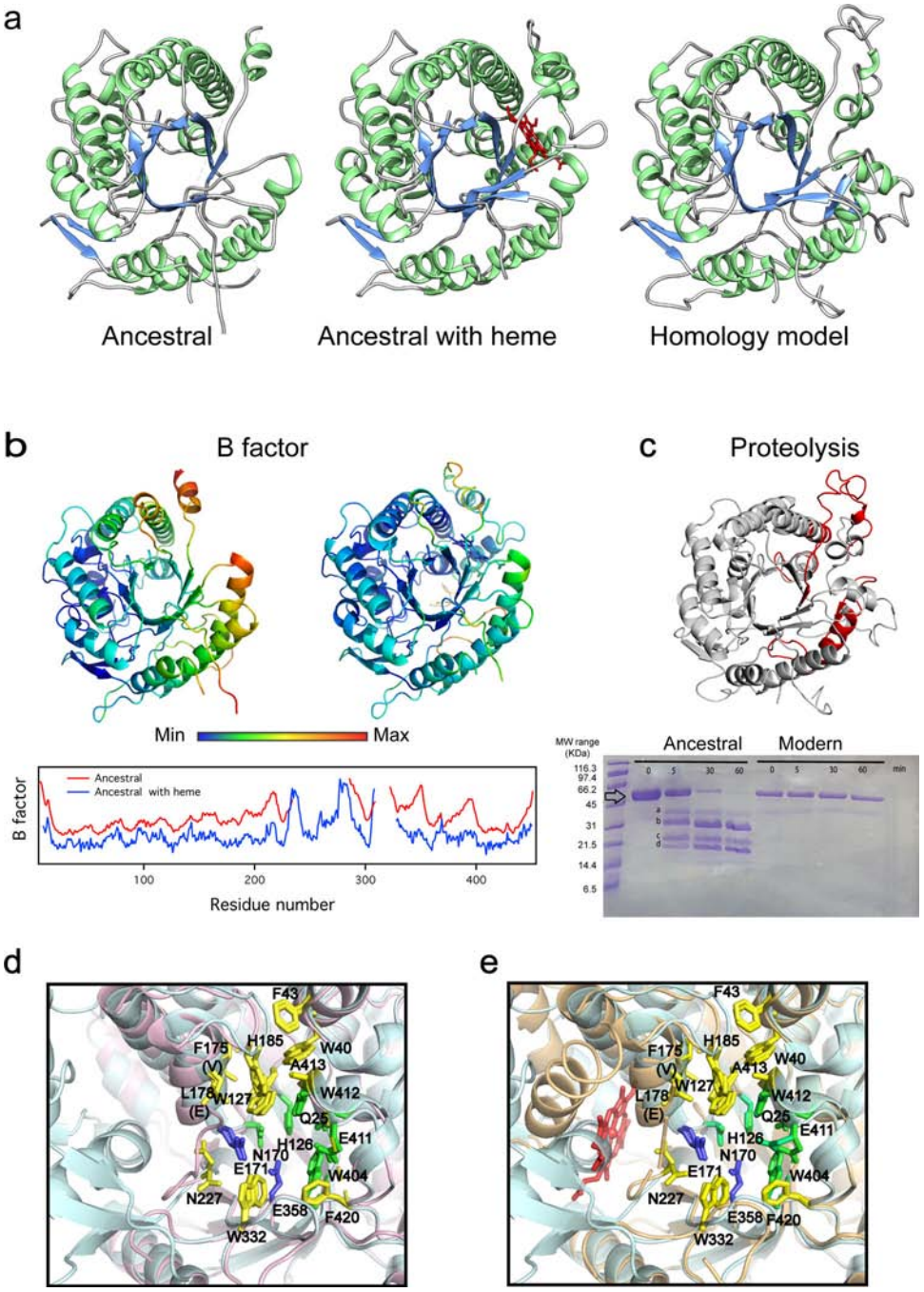


Fig. 2 3D Structure of the ancestral glycosidase as determined by X-ray crystallography. a Comparison between the ancestral structure determined in the absence (left) and presence (middle) of bound heme (red) and a homology model constructed as described in Supporting Information. The visual comparison reveals the missing sections in the electronic density of the ancestral protein, mostly in the protein without heme bound. **b** 3D structure of the ancestral protein without and with heme bound color-labeled according to normalized B-factor value and profiles of normalized B-factor versus residue number for the ancestral protein without (red) and with (blue) bound heme. Values are not shown for the sections that are missing in the experimental structures. **c** Proteolysis experiments with the ancestral glycosidase and the modern glycosidase from *Halothermothrix orenii*. The major fragments are labeled *a*, *b*, *c* and *d*. Molecular weights (MW) are shown for the markers used. Five independent experiments were performed with similar results. Mass spectrometry of the fragments predicts cleavage points within the red labeled sections in the shown structure. **d** Superposition of the structure of the ancestral glycosidase with that of the modern glycosidase from *Halothermothrix orenii* showing the critical active-site residues. **e** Superposition of the structures of the ancestral glycosidase without and with heme bound showing the critical active-site residues. In both (**d**) and (**e**), the highlighted active-site residues include the catalytic carboxylic acid residues (blue) and the residues involved in binding of the glycone (yellow) and aglycone (green) parts of the substrate.

Fig. 2c, and Fig. S2). Conversely, the modern glycosidase from the thermophilic *Halothermothrix orenii* remains essentially unaffected after several hours with the same concentration of the protease (Fig. 2c) or even with a ten times larger protease concentration. These two glycosidases, modern thermophilic and putative ancestral, are monomeric, as determined from gel filtration chromatography and analytical ultracentrifugation (Figs. S3 and S4) and display similar values for the optimum activity temperature (70 °C and 65 °C, respectively: Fig. 1b and S5). Therefore, their disparate susceptibilities to proteolysis can hardly be linked to differences in overall stability, but rather to enhanced conformational flexibility in the ancestral enzyme that exposes cleavable sites.

Furthermore, there is a large region missing in the electronic density map of the ancestral protein from X-ray crystallography (Fig. 2a), while the rest of the model agrees with a homology model based on modern glycosidase structures (see Supplementary Methods for details). At the achieved resolution of 2.5 Å, it should be possible to trace the course of a polypeptide chain in space, provided that such course is well defined. Therefore, the missing regions very likely correspond to regions of high flexibility. In addition, flexibility is also suggested by the B-factor values in regions that are present in the ancestral structure (Fig. 2b).

Lastly, molecular dynamics (MD) simulations (Fig. 3) also indicate enhanced flexibility in specific regions as shown by cumulative 15 µs simulations of the substrate-free forms of the ancestral glycosidase (both with and without heme: see below) as well as the modern glycosidase from *Halothermothrix orenii* (PDB ID: 4PTV)[27] [https://www.rcsb.org/structure/4PTV]. Both ancestral and modern proteins have the same sequence length, and similar protein folds with a root mean square deviation (RMSD) difference of only 0.7 Å between the structures. However, our molecular dynamics simulations indicate a clear difference in flexibility in the region spanning residues 227–334, which is highly disordered in the ancestral glycosidase but ordered and rigid in the modern glycosidase, with root mean square fluctuation (RMSF) values of <2 Å (Fig. 3). We also analyzed the interactions formed between residues 227–334 and the rest of the protein by counting the total intramolecular hydrogen bonds formed along the MD simulations. We observe that on average, the modern glycosidase forms 115 ± 11 hydrogen bonding interactions during our simulations, whereas the ancestral glycosidase forms either 101 ± 12/101 ± 13 hydrogen bonding interactions (in the presence of heme and absence of heme, respectively). This suggests that the higher number of intramolecular hydrogen bonds formed between residues 227–334 and the protein can contribute to the reduced conformational flexibility observed in the case of the modern glycosidase, compared to the ancestral glycosidase.

It is important to note that there is a clear structural congruence between the results of the experimental and computational studies described above. That is, the missing regions in the X-ray structure (Fig. 2a) match the high-flexibility regions in the molecular dynamics simulations (Fig. 3) and include the proteolysis cleavage sites determined by mass spectrometry (Fig. 2c). Overall, regions encompassing two alpha helices and several loops appear to be highly flexible or even unstructured in the ancestral glycosidase. The barrel core, however, remains structured and shows comparatively low conformational flexibility, which may explain the high thermal stability of the protein (see Discussion).

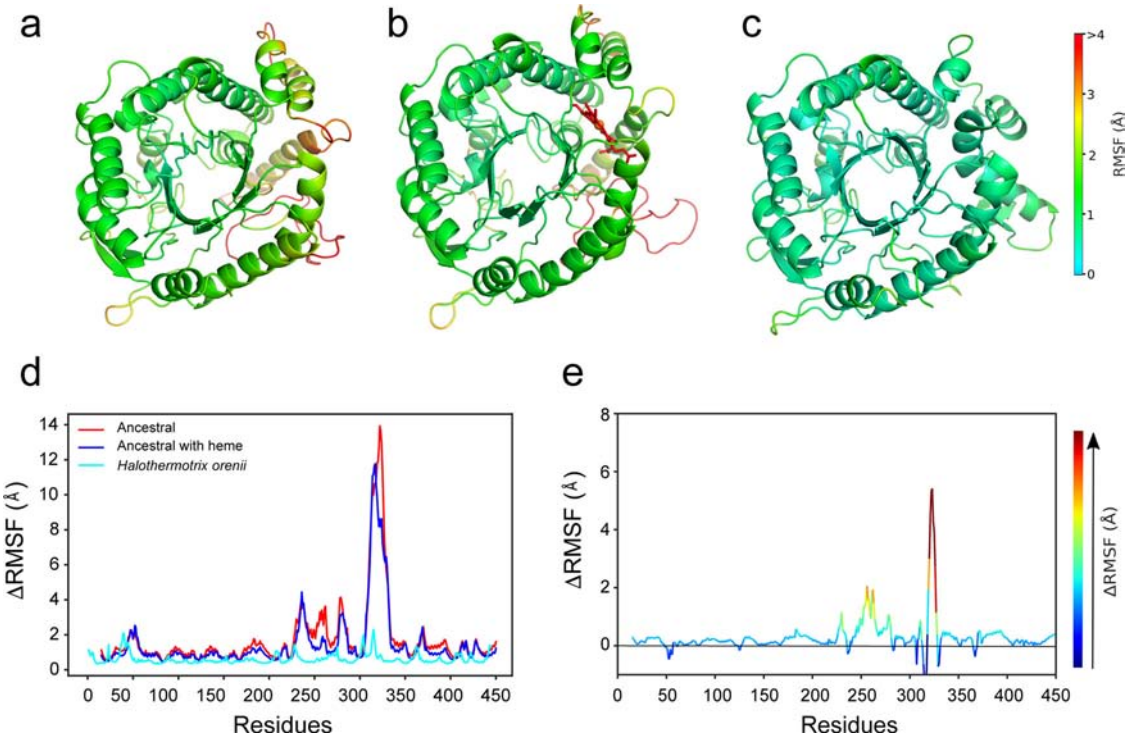


Fig. 3 Molecular dynamics simulations. Representative snapshots from molecular dynamics simulations of ancestral and modern glycosidases, showing the ancestral glycosidase both (**a**) without and (**b**) in complex with heme, as well as (**c**) the corresponding modern protein from *Halothermothrix orenii*. Structures were extracted from our simulations based on the average structure obtained in the most populated cluster using the hierarchical agglomerative algorithm implemented in CPPtraj[64]. All protein structures are colored by calculated root mean square fluctuations (RMSF) over the course of simulations of each system (see the color bar). Shown are also (**d**) absolute and (**e**) relative RMSF (Å) for each system, in the latter case showing the RMSF of the ancestral glycosidase without heme relative to the heme bound structure. Note the difference in the color bars between panels (**a**–**c**), which describes absolute RMSF per system, and panel (**e**), which describes relative RMSF. The numerical scale of the color bar on panel (**e**) corresponds to the *y*-axis of this panel.

Catalysis. We determined the Michaelis–Menten parameters for the ancestral enzyme with the substrates typically used to test the standard β-glucosidase and β-galactosidase activities of family 1 glycosidases (4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside) (Fig. 4 and Tables S2 and S3). We also compared the results with the catalytic parameters for four modern family 1 glycosidases, specifically those from *Halothermothrix orenii*, *Marinomonas sp.* (strain MWYL1), *Saccharophagus degradans* (strain 2-40 T), and *Thermotoga maritima* (Figs. S6–S9 and Tables S2 and S3). Modern glycosidases are highly proficient enzymes accelerating the rate of glycoside bond hydrolysis up to about 17 orders of magnitude[16]. The ancestral enzyme appears to be less efficient and shows a turnover number about two orders of magnitude below the values for the modern glycosidases studied here (Fig. 4). It is important to note, nevertheless, that the turnover number for the ancestral enzyme is still ~13 orders of magnitude higher than the first-order rate constant for the uncatalyzed hydrolysis of β-glucopyranosides, as determined by Wolfenden through Arrhenius extrapolation from high-temperature rates[15]. The catalytic carboxylic acid residues as well as the residues known to be responsible for the interaction with the glycone moiety of the substrate[28] are present in the ancestral enzyme and appear in the static X-ray structure in a configuration similar to that observed in the modern proteins

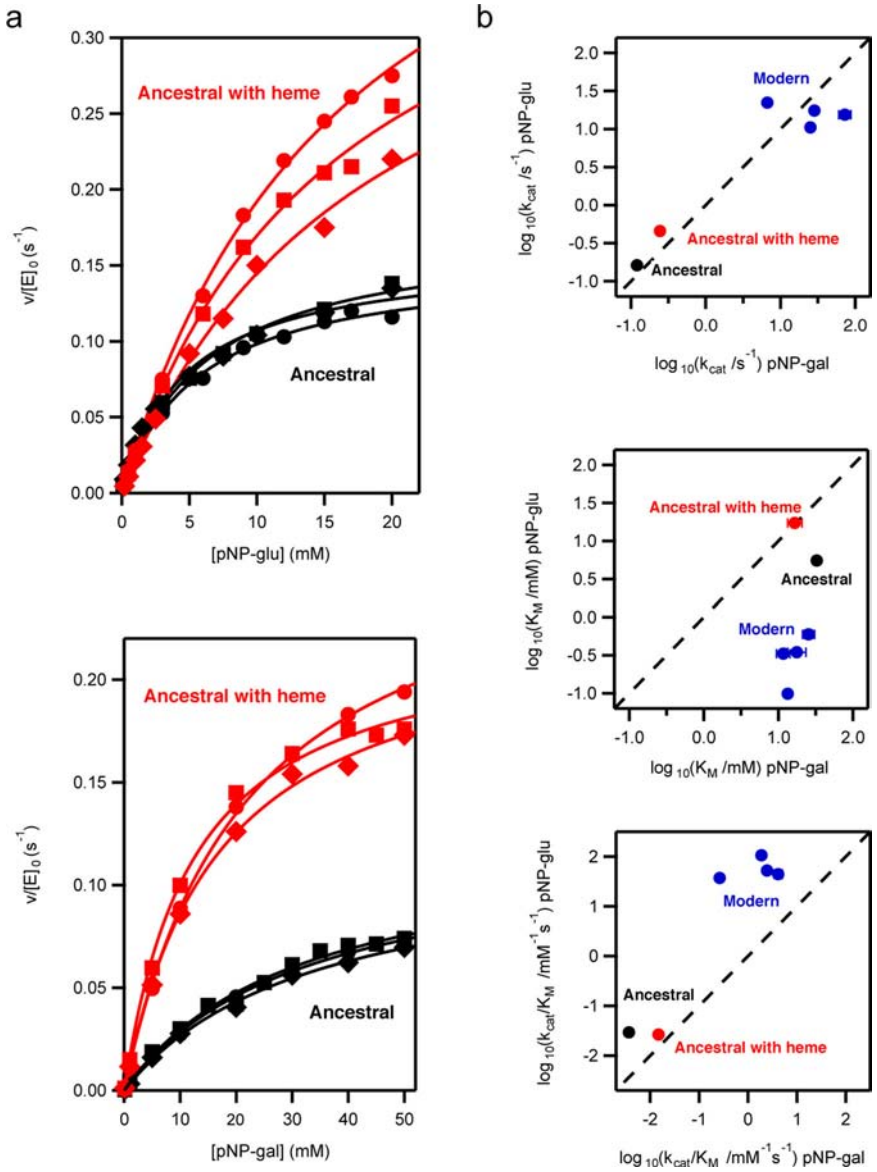


Fig. 4 Ancestral versus modern catalysis by family 1 glycosidases. a Michaelis plots of rate versus substrate concentration at pH 7 and 25 °C for hydrolysis of 4-nitrophenyl-β-D-glucopyranoside (upper panel) and 4-nitrophenyl-β-D-galactopyranoside (lower panel) catalyzed by the ancestral glycosidase with and without heme bound. $v/[E]_0$ stands for the rate over the total enzyme concentration. The lines are the best fits of the Michaelis–Menten equation. The different symbols (diamond, square, circle) refer to the triplicate experiments (involving two different protein preparations) performed for each protein/substrate combination. Michaelis plots for the four modern proteins studied in this work can be found in Figs. S6–S9. The values for the catalytic parameters derived from these fits are collected in Tables S2 and S3. **b** Logarithm of the Michaelis-Menten catalytic parameters for a glucopyranoside substrate versus a galactopyranoside substrate. pNP-glu and pNP-gal stand, respectively, for 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside. k_{cat}, K_M, and k_{cat}/K_M stand for the turnover number, the Michaelis constant and the catalytic efficiency. The values shown are averages of the values derived from the triplicates and the associated errors are the corresponding standard deviations. Note that, in most cases, the associated errors are smaller than the size of the data points.

(Fig. 2d). There are a few differences in the identity of the residues responsible for the binding of the aglycone moiety of the substrate[28], but these differences occur in positions that are variable in modern family 1 glycosidases (Fig. S10 and Table S4). Overall, the comparatively low activity of the ancestral protein is likely linked to its conformational flexibility. That is, the protein in solution is sampling a diversity of conformations of which only a few are active towards the common substrates. From an evolutionary point of view, the comparatively low ancestral activity may reflect an early stage in the evolution of family 1 glycosidases before selection favored greater turnover (see "Discussion").

Also, it is interesting to note that, although both β-glucosidase and β-galactosidase activities are typically described for family 1 glycosidases, these enzymes are commonly specialized as β-glucosidases[22]. This specialization does not occur, however, at the level of the turnover number, which is typically similar for both kinds of substrates. Instead, specialization occurs at the level of the substrate affinity, as reflected in lower values of the Michaelis constant (K_M) for β-glucopyranoside substrates as compared to β-galactopyranoside substrates[22]. This pattern is indeed observed in the modern enzymes we have studied (Fig. 4), which are described in the literature as β-glucosidases. On the other hand, this kind of specialization is not observed in the ancestral glycosidase, which shows similar K_M's for the β-glucopyranoside and the β-galactopyranoside substrates. This lack of specialization may again reflect an early stage in the evolution of family 1 glycosidases, an interpretation which would seem generally consistent with the fact that resurrected ancestral proteins often display promiscuity[5,7,9,29]. On the other hand, it can be argued that the ancestral glycosidase was specialized for a different kind of substrate. To explore this possibility, we determined catalytic rates for a wide range of glycosidase substrates. These studies are briefly described below:

(1) Using the same methodology employed with *4*-nitrophenyl-β-D-glucopyranoside and *4*-nitrophenyl-β-D-galactopyranoside (Fig. 1), we determined profiles of catalytic rate versus temperature for the ancestral glycosidase and the modern glycosidases from *Halothermothrix orenii* and *Saccharophagus degradans* using as substrates *4*-nitrophenyl-β-D-fucopyranoside, *4*-nitrophenyl-β-D-lactopyranoside, 4-nitrophenyl-β-D-xylopyranoside and 4-nitrophenyl-β-D-mannopyranoside. In all cases (Fig. S11), we found the levels of catalysis of the ancestral protein to be reduced in comparison with the modern proteins. We also found that the levels of catalysis for the β-glucopyranoside and β-fucopyranoside substrates were similar, but this pattern is also observed with the modern proteins. (2) We carried out single activity determinations at 25 °C for the ancestral glycosidase with a wider range of substrates, including derivatives of disaccharides (maltose, cellobiose) and several substrates with an α anomeric carbon (Table S5). However, we did not find any substrate with a catalysis level substantially higher than that of those previously determined for 4-nitrophenyl-β-D-glucopyranoside and *4*-nitrophenyl-β-D-galactopyranoside and, in many cases (in particular with the α substrates), no substantial activity was detected. (3) Since some of the proteins that descended from the N72 node are 6-phosphate-β-glucosidases (Fig. 1A and S1), we tested the activity of our ancestral glycosidase against *4*-nitrophenyl-β-D-glucopyranoside-6-phosphate (Fig. S12). We found the catalytic efficiency to be ~40 fold smaller than that determined with the corresponding non-phosphorylated substrate. (4) Glycosidases are typically described[14] as being very promiscuous for the aglycone moiety of the substrate (the part of the substrate that is replaced with *p*-nitrophenyl in the substrates commonly used to assay glycosidase activity) while they are more specialized for the glycone moiety of the substrate. However, the flexibility in certain regions of the ancestral structure could perhaps favor the hydrolysis of substrates with larger aglycone moieties. To explore this hypothesis, we tested four synthetic substrates with aglycone moieties larger than the usual *p*-nitrophenyl group (Fig. S13). We revealed that ancestral levels of catalysis are substantially reduced with respect to those obtained for the modern glycosidase from *Halothermothrix orenii*, used here as comparison.

Heme binding and allosteric modulation. Overall, it appears reasonable that our resurrected ancestral enzyme reflects an early stage in the evolution of family 1 glycosidases, perhaps following a fragment fusion event (see Discussion), at which catalysis was not yet optimized and substrate specialization had not yet evolved. The presence of a large unstructured and/or flexible regions in the ancestral structure could perhaps reflect the absence of a small molecule that binds within that region. While these proposals are speculative, the experimental results described in detail below, show that the ancestral glycosidase does bind heme tightly and stoichiometrically at a site in the flexible regions. This was a completely unexpected observation given the large number of modern glycosidases that have been characterized in the absence of any porphyrin rings.

We curiously noticed that most preparations of the ancestral glycosidase showed a light-reddish color after elution from an affinity column (Fig. 5). UV–Vis spectra revealed the pattern of bands expected for a heme group[30], including the Soret band at about 400 nm and, in some cases, even the weaker α and β bands (i.e., the Q bands) in the 500–600 nm region (Fig. 6). From the intensity of the Soret band, a very low heme:protein ratio of about 0.02 was estimated for standard enzyme preparations, indicating that all the experiments described above were performed with essentially heme-free protein. However, the amount of bound heme in protein preparations was substantially enhanced by including hemin in the culture medium (heme with iron in the +3 oxidation state) or 5-aminolevulinic acid, the metabolic precursor of heme (Fig. 5). Heme:protein ratios of about 0.10 and 0.18, respectively, were then obtained (Fig. 6a). These results suggest that the ancestral glycosidase does have the capability to bind heme, but also that, as is commonly the case with modern heme-binding proteins[31], the limited amount of heme available in the expression host, combined with the high protein overexpression levels used, leads to low heme:protein ratios. The capability of the ancestral enzyme to bind heme was first shown by the in vitro experiments described next, and then confirmed by mass spectrometry and X-ray crystallography as subsequently described.

Heme has a tendency to associate in aqueous solution at neutral pH, a process that is reflected in a time-dependent decrease in the intensity of the Soret band, which becomes "flatter" upon the formation of dimers and higher associations[32]. However, the process is reversed upon addition of the essentially heme-free ancestral glycosidase (Fig. 6b), indicating that the protein binds heme and shifts the association equilibria towards the monomeric state. Remarkably, heme binding is also reflected in a several-fold increase in enzymatic activity which occurs on the seconds time scale when the heme and enzyme concentration are at a ~micromolar concentration (Fig. 6c). Determination of activity after suitable incubation times for different heme:protein ratios in solution yielded a plot with an abrupt change of slope at a stoichiometric ratio of about 1:1 (Fig. 6d). These experiments were carried out with ~micromolar heme and protein concentrations, indicating therefore a tight, sub-micromolar binding. This interaction was confirmed by microscale thermophoresis experiments that yielded an estimate of 547±110 nM for the heme dissociation constant (see "Methods" and Fig. S14 for details). Indeed, in agreement with this tight binding, increases in activity upon heme addition to a protein solution were observed (Fig. 6c) even with concentrations of ~50 nanomolar.

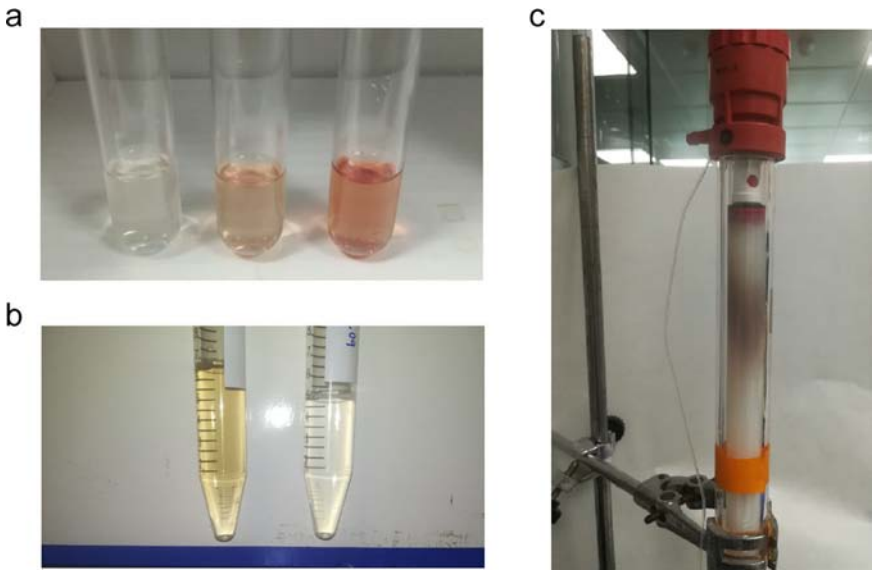


Fig. 5 Heme binding to the ancestral glycosidase is visually apparent. a The ancestral protein was prepared by Ni-NTA affinity chromatography. The pictures show the samples eluted from the columns for three different preparations that differed by the addition of 20 μM hemin (middle) and 0.4 mM 5-aminolevulinic acid (right) to the culture medium. Neither hemin nor 5-aminolevulinic acid had been added to the culture medium in the preparation on the left. Protein concentrations in these samples were ~10 mg/mL. **b** An ancestral glycosidase sample with a low amount of bound heme (right) was incubated with an excess of heme. A PD10 column and FPLC (fast protein liquid chromatrography) were then used to remove the unbound heme. The resulting protein preparation is shown on the left. Protein concentration is ~0.5 mg/mL. **c** The position of the ancestral protein with bound heme in a FPLC column is revealed by a reddish-brown band.

The 1:1 stoichiometry of heme/protein was confirmed by experiments in which the protein was incubated with an excess of heme and free heme was removed through exclusion chromatography (2 passages through PD10 columns). The protein was then quantified by the bicinchoninic acid method[33] with the Pierce™ BCA Protein Assay Kit while the amount of heme was determined using the pyridine hemochrome spectrum[34] after transfer to concentrated sodium hydroxide (see Methods for details). This resulted in a heme/protein stoichiometry of 1.03±0.03 from five independent assays.

The experiments described above allowed us to set up a procedure for the preparation of the ancestral protein saturated with heme and to use this preparation for activity determinations and crystallization experiments. The procedure (see Methods for details) involved in vitro reconstitution using hemin but did not include any chemical system capable of performing a reduction. It is therefore safe to assume that our heme-bound ancestral glycosidase contains iron in the +3 oxidation state. Activity determinations with the heme-saturated ancestral enzyme corroborated that heme binding increases activity by ~3 fold (see Michaelis plots in Fig. 4). Both mass spectrometry (Fig. S15) and X-ray crystallography confirmed the presence of one heme per protein molecule (Fig. 7, S16 and S17), which is located at the same site, with the same orientation and involved largely in the same molecular interactions in the three protein molecules (A, B, C) observed in the crystallographic unit cell. Besides interactions with several hydrophobic residues, the bound heme interacts (Fig. 7a) with Tyr264 of α-helix 8 (as the axial ligand), Tyr350 of α-helix 13, Arg345 of β-strand B and, directly via a water molecule, with Lys 261 of β-strand B, although this latter interaction is only observed in chain A. The bound heme shows B-factor values similar to those of the surrounding residues (Fig. 7b), it is well-packed and 95% buried (Fig. 7c). Indeed, the accessible surface area of the bound heme is only 43 $Å^2$ compared to the ~800 $Å^2$ accessible surface area for a free heme[35]. The interactions of the bound heme in the ancestral glycosidase are overall similar to those described for modern b-type heme proteins[35–37]. As observed in modern heme-binding proteins, the ancestral heme-binding pocket is enriched in hydrophobic and aromatic residues and propionate anchoring is achieved through interactions with arginine, tyrosine and lysine residues. Certainly, tyrosine, the axial ligand in the ancestral glycosidase, is not the most common axial ligand in modern heme proteins, but it is found in several cases, including catalases (see, Protein Data Bank (PDB) ID 1QWL for the 3D structure of the catalase from *Helicobacter pylori*). Interestingly, the amino acid residues that interact with the heme in the ancestral glycosidase are somewhat conserved, and are indeed the consensus residue from the set of modern glycosidases used as the starting point for ancestral reconstruction (Table S6). The fraction of each consensus residues in the modern protein is, however, less than unity and the sequences of modern glycosidases in the set differ from the ancestral sequence at many of the positions involved in heme interactions in the ancestral protein (Table S7).

Heme binding clearly rigidifies the ancestral protein, as shown by fewer missing regions in the electronic density map, in contrast to the structure of the heme-free protein (see Fig. 2a–c). This is also confirmed by molecular dynamics (MD) simulations of the ancestral glycosidase both with and without heme bound (Fig. 3 and S18). Figure S18 shows the backbone RMSD (root mean square deviation) over ten individual 500 ns MD simulations per system, and, from this data, it can be seen that while the RMSD is fairly stable in the case of the modern protein, the ancestral glycosidases (both with and without heme bound) are initially quite far from their equilibrium structures, due to the high flexibility of the missing regions of the protein which require substantial equilibration. In addition, we note that while the overall average RMSD for the ancestral protein with heme bound is slightly lower than for the ancestral protein without heme (Fig. S18), the standard deviation is higher. This is due to the greater flexibility of the reconstructed missing loop (see the "Methods" section), which allows it to sample a larger span of conformations depending on whether the loop is interacting with the bound heme or not (we observe both scenarios in our simulations of the heme-bound ancestral glycosidase).

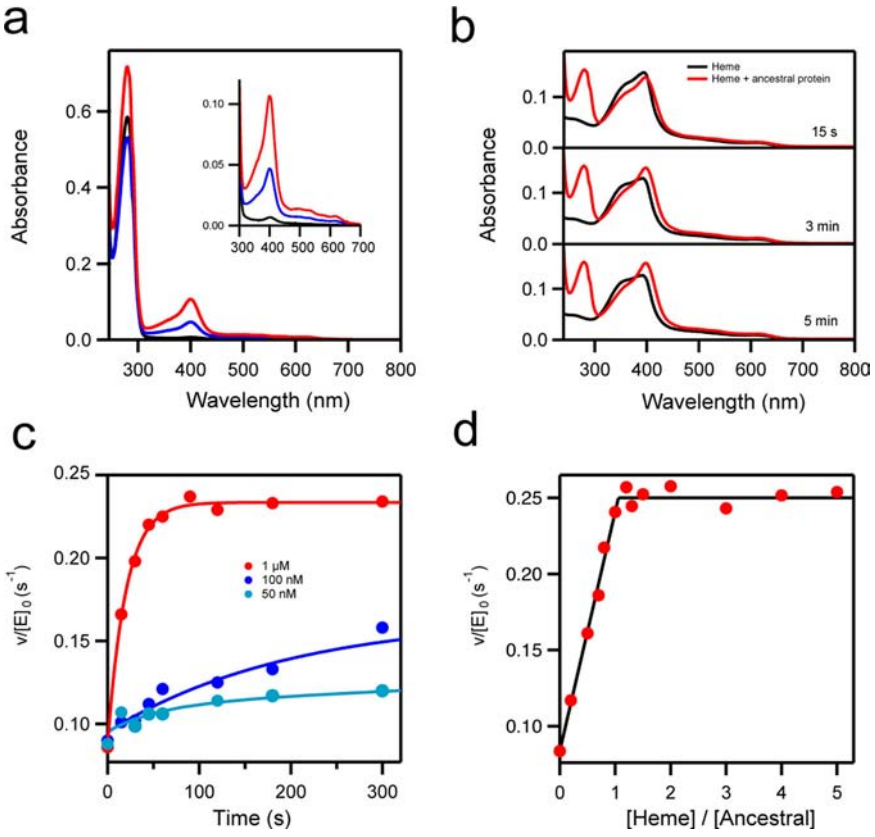


Fig. 6 Heme binding to the ancestral glycosidase. a UV–VIS spectra for preparations of the ancestral glycosidase showing the protein absorption band at about 280 nm and the absorption bands due to the heme (the Soret band at about 400 nm and the Q bands at higher wavelengths). Black color is used for the protein obtained using the original purification procedure without the addition of hemin or hemin precursor. Blue and red are used to refer, respectively, to preparations in which hemin and 5-aminolevulinic acid (the metabolic precursor of heme) were added to the culture medium. **b** Binding of heme to the ancestral glycosidase in vitro as followed by changes in VIS spectrum. Spectra of a heme solution 1 μM in the absence (black) or presence (red) of a similar concentration of ancestral protein. The "flat" Soret band of free heme is linked to its self-association in solution, while the bound heme is monomeric and produces a sharper Soret band. **c** Kinetics of binding of heme to the ancestral glycosidase as followed by the increase in enzyme activity (rate of hydrolysis of *4*-nitrophenyl-β-glucopyranoside; see Methods for details). In the three experiments shown a heme to protein molar ratio of 1.2 was used. The protein concentration in each experiment is shown. Note that activity increase is detected even with concentrations of 50 nM, indicating that binding is strong. The lines are meant to guide the eye and thus have no quantitative purpose. (**d**) Plot of enzyme activity versus [heme]/[protein] ratio in solution for a protein concentration of 1 μM. Activity was determined after a 5 min incubation and the plot supports a 1:1 binding stoichiometry. $v/[E]_0$ stands for the rate over the total enzyme concentration.

In contrast, in the absence of the heme, the loop is always in a flexible open conformation leading to a higher overall RMSD but a lower standard deviation as a narrower range of conformations are sampled in our simulations. As neither the loop nor the heme has access to the active site (Fig. S17), these differences are unlikely to have a direct effect on catalysis.

The higher flexibility of the ancestral protein without heme bound can also be seen from comparing RMSF (root mean square fluctuation) values across the protein. That is, the MD simulations performed without the heme bound show that most of the protein has higher flexibility (Fig. 3e, with ΔRMSF values greater than 0 in most of the sequence), particularly in the regions where the B-factors also indicate high flexibility (Fig. 2b). This is noteworthy, as the only difference in starting structure between the two sets of simulations is the presence or absence of the heme; the starting structures are otherwise identical. The MD simulations show that removing the heme from the heme-bound structure has a clear effect on the flexibility of the whole enzyme, increasing it relative to the heme bound structure (Fig. 3e), again also indicated by the B-factors (Fig. 2b). There are two regions where this difference is particularly pronounced. The first spans residues 25–265, which is located where the heme Fe(III) atom forms an interaction with the Tyr264 side chain as an axial ligand. Removing the heme removes this interaction, thus inducing greater flexibility in this region. The second region with increased flexibility spans 319–327, where again we observe that removing the heme increases the flexibility of this region.

Lastly, we note that the heme is located near the enzyme active site (at about 8 Å from the catalytic glutamate at position 171) but does not have direct access to this site as revealed in the structure (Fig. S17). Therefore, the increase in activity observed upon heme binding is an allosteric effect likely linked to dynamics (see "Discussion") since heme binding does not substantially alter the position/conformation of the catalytic carboxylic acids nor the residues involved in substrate binding according to the static X-ray structures (Fig. 2e). In fact, examining backbone RMSF values of key catalytic residues (Fig. S19) indicates that the flexibility of several of these residues is reduced upon moving from the ancestral glycosides without heme, to adding the heme, to the modern glycosidase, in a clear decreasing trend. We note that the observed effects are subtle and sub-Å; however, there are several experimental studies that suggest that sub-Å changes in dynamics can be catalytically important[38–40].

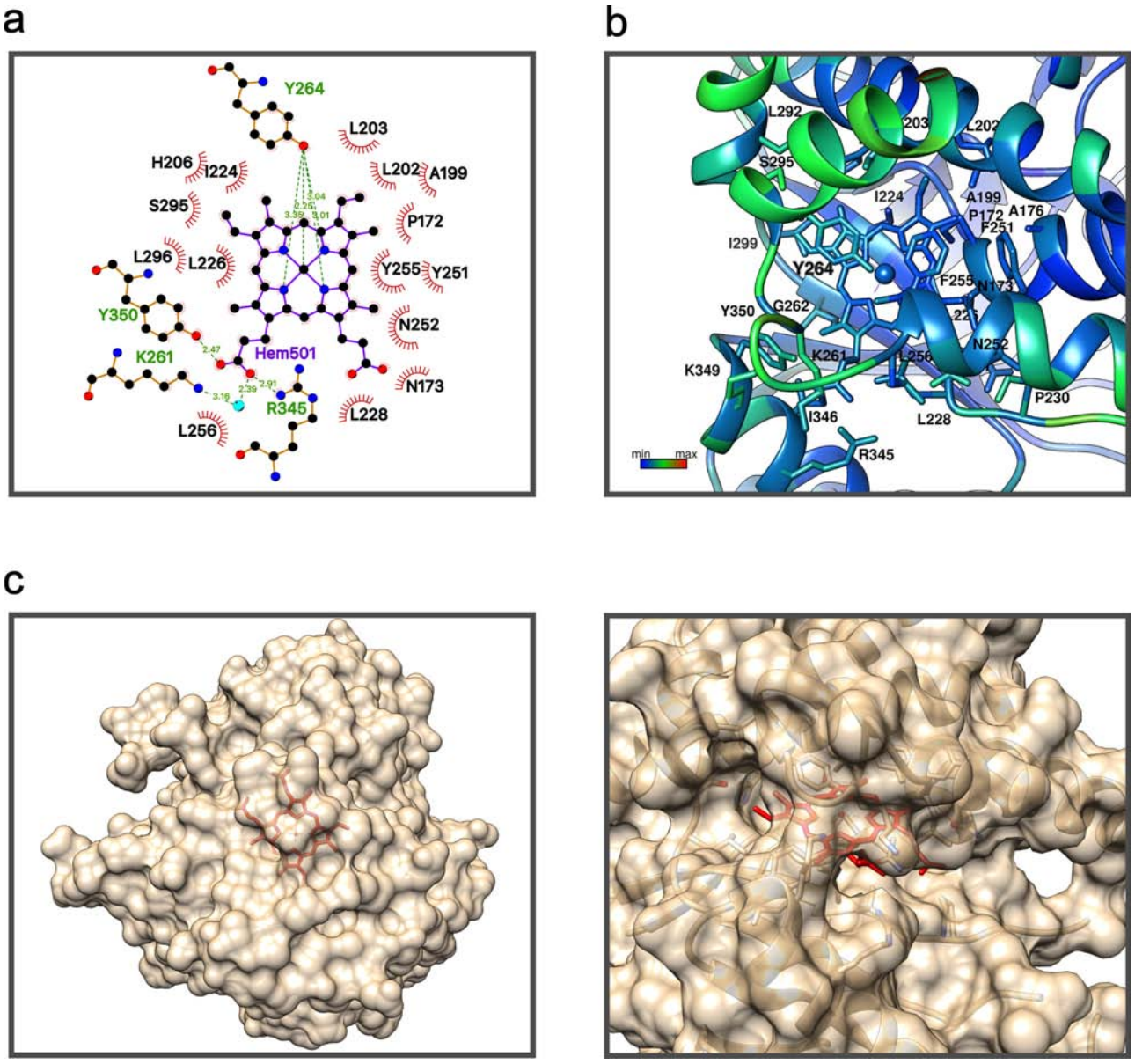


Fig. 7 Local molecular environment of the bound heme in the ancestral protein. a Schematic representation of the heme molecule and the neighbor residues in the 3D structure. **b** Heme group and residues directly interacting with the heme colored by B value. **c** Van der Waals surface of the ancestral protein shown in translucent brown, so that it becomes visually apparent that the heme (shown in red) is mostly buried.

Discussion

The TIM-barrel is the most common enzyme fold, accounting for ~10% of known enzyme structures and providing a scaffold for an enormous diversity of biomolecular functions[11–13]. It is composed of eight parallel (β/α) units linked by hydrogen bonds forming a cylindrical core ("the barrel") with secondary structure elements connected by loops. The high capability of the fold to accommodate a wide diversity of different natural functions is likely linked to its modular architecture, with the barrel (and the αβ loops) providing stability and allowing a substantial degree of flexibility, variability, and, therefore, evolvability for the βα loops. That is, the barrel provides a stable platform that can accommodate loops of different sequences and conformations at the so-called catalytic face.

Remarkably, the differences in conformational flexibility between different parts of the molecule appear to be even more pronounced in our ancestral TIM-barrel glycosidase. Stability is still guaranteed by a rigid barrel core, but flexibility is greatly enhanced and extends to large parts of the structure, as shown by a combination of computational and experimental results. Conformational flexibility implies that the protein in solution is sampling a diversity of conformations. On the one hand, this may prevent the enzyme from reaching the highest levels of catalysis for a given natural reaction since the protein ensemble may not be shifted towards the most active conformations. Indeed, while modern glycosidases approach catalysis levels up to 17 orders of magnitude above the rate of spontaneous glycoside bond hydrolysis[16], the ancestral glycosidase displays turnover numbers about two orders of magnitude below the modern glycosidases studied here (Fig. 4 and Tables S2 and S3). On the other hand, flexibility is key to the emergence of new functions and contributes to evolvability, since minor conformations that catalyze alternative reactions may be enriched by subsequent evolution[41–44]. Therefore, the ancestral TIM-barrel described here holds promise as a scaffold for the generation of de novo catalysts, an important and largely unsolved problem in

enzyme engineering. We have recently shown[45] that completely new enzyme functions can be generated through a single mutation that generates both a cavity and a catalytic residue, provided that conformational flexibility around the mutation site allows for substrate and transition-state binding[10,43,44]. The combination of a rigid core that provides stability with high flexibility in specific regions makes the ancestral protein studied here an excellent scaffold to develop this minimalist approach to de novo catalysis (work in progress).

Catalytic features of the ancestral glycosidase, such as diminished activity levels and lack of specialization for glucopyranoside substrates, would seem consistent with an early stage in the evolution of family 1 glycosidases. It has been proposed that TIM-barrel proteins originated through fusions of smaller fragments[46]. The high conformational flexibility in some regions of the ancestral glycosidase structure would then also seem consistent with an early evolutionary stage, since fragment fusion is not expected to immediately lead to efficient packing and conformational rigidity in all parts of the generated structure. On the other hand, the capability of the reconstructed ancestral glycosidase to bind heme tightly and stoichiometrically at a well-defined site is rather surprising. None of the ~5500 X-ray structures for the ~1400 glycosidases currently reported in CAZy shows a porphyrin ring. It is certainly possible that heme binding to the ancestral glycosidase is simply an accidental byproduct of the high conformational flexibility at certain regions of the structure, although the tightness of the binding and the specificity of the molecular interactions involved argue against this possibility. In any case, this is an issue that can be investigated by studying modern glycosidases. If heme binding is a functional ancestral feature (a product of selection), we may expect that at least some modern glycosidases show some inefficient, vestigial capability to bind heme, in keeping with the general principle that features that become less functional undergo evolutionary degradation[47,48]. No mention of heme binding to modern family 1 glycosidases can be found in the CAZypedia resource[22], but, of course, there is no reason why researchers in the glycosidase community should have tested heme-binding capabilities. As such, we have done so in this work for the four modern enzymes we already characterized in terms of catalysis (Fig. 4), i.e., the modern family 1 glycosidases from *Halothermothrix orenii*, *Thermotoga maritima*, *Marinomonas sp.* (strain MWL1), and *Saccharophagus degradans* (strain 2-40). When 5-aminolevulinic acid, the metabolic precursor of heme, was added to the culture medium, the four modern proteins were isolated with an appreciable amount of bound heme, although their heme-binding capability is clearly much reduced when compared with the ancestral glycosidase (Fig. 8 and Fig. S20). We furthermore carried out the same type of experiment with proteins corresponding to reconstructions at five nodes in the line of descent that leads from the ancestral glycosidase at node 72 to the modern glycosidase from *Halothermothrix orenii* (Fig. 8). We also found appreciable, but typically much lower amounts of bound heme, as compared with the "older" ancestral protein at node 72 (Fig. 8 and Fig. S21). Finally, we purified the heme-free forms of the proteins at these five ancestral nodes and studied their heme-binding capability in vitro. These proteins have glycosidase activity levels intermediate between those of the ancestral glycosidase at node 72 and the modern glycosidase from *Halothermothrix orenii* (Fig. S22) and unambiguously display in vitro heme-binding capability at micromolar concentrations, as shown by the presence of the Soret band in UV–VIS spectra (Fig. S23) and further confirmed by mass spectrometry (Figs. S24–S26). Evolutionary degradation of ancestral heme binding, however, is clearly revealed by analyses of elution profiles from gel filtration chromatography in terms of protein concentration, heme concentration, and glycosidase activity (Fig. 8). Thus, while heme binding to the most ancient studied node (our ancestral glycosidase at node 72) produces active monomers to a large extent, a trend towards a decreased amount of heme-bound monomers and appearance of higher association states upon heme binding is observed in the evolutionary line leading to the modern glycosidase from *Halothermothrix orenii*. One interesting possibility is that heme binding to the monomers of the less ancient proteins brings about conformational changes that trigger protein association.

It emerges that heme binding to the ancestral glycosidase at node 72 is not an oddity or an artifact of reconstruction. In contrast, it appears probable that heme binding to ancient family 1 glycosidases did specifically occur, and that it also underwent degradation at an early evolutionary stage to lead to a rudimentary capability with substantial variability, as it is commonly observed with vestigial features that are not subject to selection[47]. Overall, this suggests a complex evolutionary history for this family of enzymes involving perhaps a fortuitous (i.e., contingent) early fusion event with a heme-containing domain. In this scenario, heme had a functional role in the isolated heme-containing domain, which was no longer required when the domain was fused with the larger glycosidase scaffold, thus enabling the subsequent degradation of the heme-binding capability. In order to find some evidence for this hypothesis, we used the Dali sever[49] to search in the Protein Data Bank for structural alignments of the alpha helices involved in heme binding from our ancestral glycosidase. However, we did not find any convincing match, as the best obtained structural alignments had RMSD values of ~4 Å and Z scores of 2 or higher (see Fig. S27 for further details). It is possible that the structure of the ancestral heme-containing domain was distorted upon fusion and subsequent evolution, and is therefore difficult to identify in searches of modern protein structures. Another possibility is that there was never a fusion event with a heme-containing domain and that heme was already present even at the most ancient stages in the origins of family 1 glycosidases. This would be consistent with an interpretation of cofactors as molecular fossils that facilitated the primitive emergence of proteins by selecting them from a random pool of polypeptides[50].

It is important to note at this stage that relevant protein-engineering implications are independent of any evolutionary interpretations. In this context, heme-binding to the ancestral glycosidases, regardless of its evolutionary origin and implications, opens up new engineering possibilities. This is so because directed laboratory evolution can be used to enhance or modify any functionality, provided that a certain level of functionality is used to start the process. The capability to "seed" levels of new functionalities may become a critical bottleneck in protein-engineering projects. Our work uncovers a heme-binding capability and a possibility of allosteric regulation that were previously unknown in glycosidase enzymes. Potential practical implications are briefly discussed below.

Metalloporphyrins are essential parts of many natural enzymes involved in redox and rearrangement catalysis and can be engineered for the catalysis of non-natural reactions[51]. Remarkably, however, the combination of the highly evolvable TIM-barrel scaffold and the catalytically versatile metalloporphyrins is exceedingly rare among known modern proteins. A porphyrin ring is found in only 13 out of the 7637 PDB entries that are assigned the TIM-barrel fold according to the CATH classification[52]. These 13 entries correspond to just two proteins. One of them, uroporphyrinogen decarboxylase, is an enzyme involved in heme biosynthesis, while in the other identified case, flavocytochrome B2, the bound heme is far from the active site. By contrast, heme appears at about 8 Å from the catalytic Glu171 in our ancestral glycosidase. While its connection with the

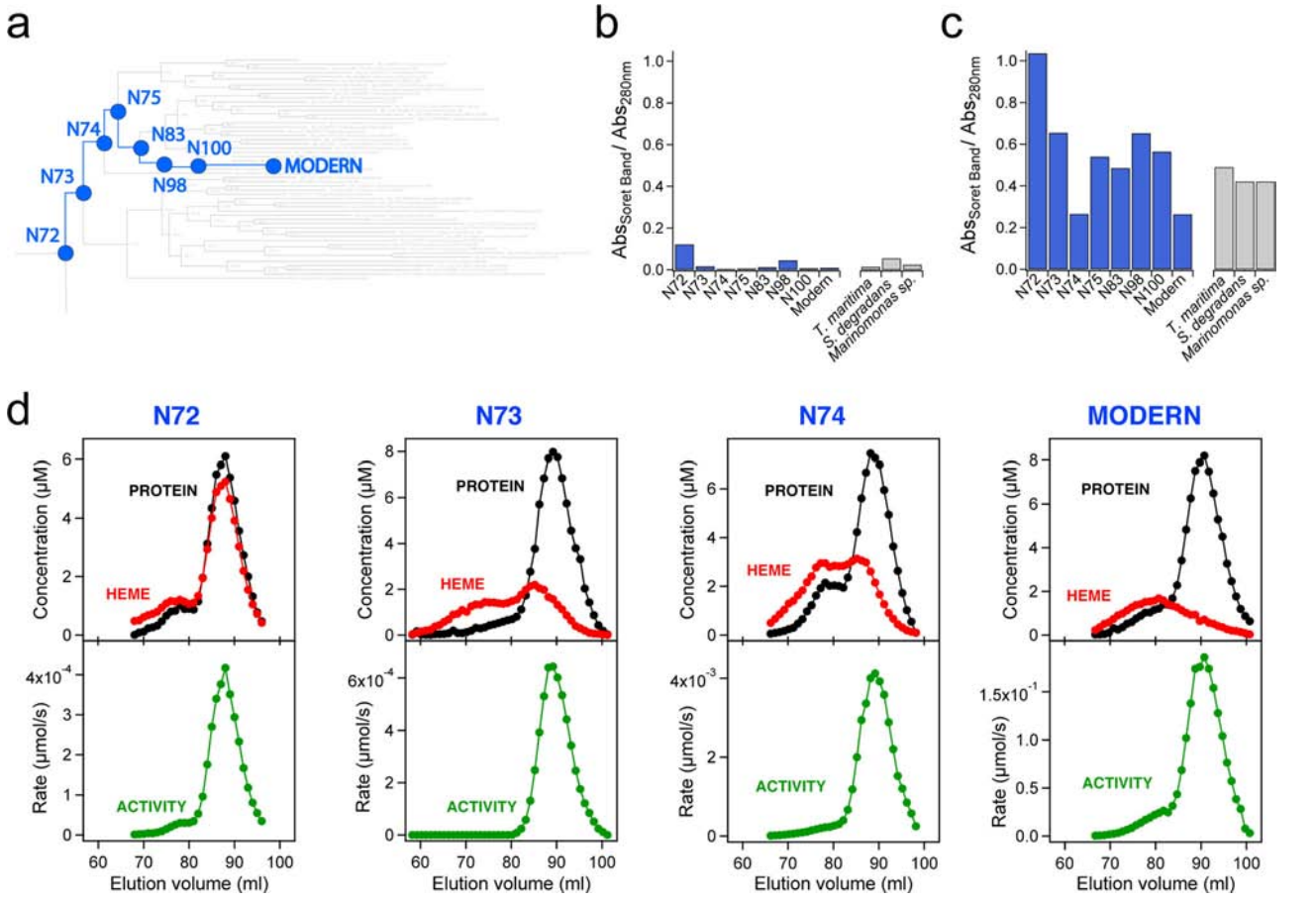


Fig. 8 Evolutionary degradation of ancestral heme binding. a Section of the phylogenetic tree used for the Bayesian analysis of family 1 glycosidases. The nodes in the evolutionary trajectory from node 72 to *Halothermothrix orenii* (labeled "MODERN") are highlighted and labeled. See Fig. S1 for a complete and detailed version of the phylogenetic tree. **b**, **c** Ratio of absorbance (Abs) at the maximum of the heme Soret band to the absorbance at the maximum of the protein aromatic absorption band for ancestral (see panel A) and modern family 1 glycosidases. Data in (**b**) correspond to protein preparations in which 0.4 mM 5-aminolevulinic acid (the metabolic precursor of heme) was added to the culture medium and the protein was purified by Ni-NTA affinity chromatography and further passage through a PD10 column (see Figs. S20 and S21 for further detail). For the data in (**c**), heme-free protein samples at ~5 μM were incubated for 1 h at pH 7 with a 5-fold excess of heme and free heme was removed through size exclusion chromatography (2 passages through PD10 columns) before recording the UV–VIS spectra (Fig. S23). **d** Profiles of protein concentration (bicinchoninic acid method[33]), heme concentration (pyridine hemochrome spectrum[3]) and glycosidase activity (with 0.25 mM 4-nitrophenyl-β-D-glucopyranoside) upon elution from gel filtration chromatography (HiLoad 16/600 Superdex 200 pg GE Healthcare). For these experiments, heme was gradually added to ~30 μM samples of ancestral (N72, N73, and N74) and modern (*Halothermothrix orenii*) glycosidases up to a ~5-fold excess and, after several hours, free heme was removed using PD10 columns. The elution volume for the main protein concentration peak is consistent with the monomeric association state (Fig. S3).

natural active site is blocked by several side chains in the determined structure (Fig. S17), it would appear feasible to use protein engineering to establish a conduit. As a simple illustration of this possibility, mutating Pro172, Asn173, Ile224, Leu226, Asn227, and Pro 272 to alanine in silico (Fig. S17 and Table S8) increase the accessible surface area of heme from barely 43 $Å^2$ to about 300 $Å^2$ and exposes the side of the heme facing the active site. The possibility thus arises that the engineering of metalloporphyrin catalysis, through rational design and/or laboratory evolution, would benefit from the evolutionary possibilities afforded by the flexible βα loops at a TIM-barrel catalytic face.

More immediate engineering possibilities arise from the allosteric modulation of catalysis in the ancestral protein, a phenomenon that, to our knowledge, has never been reported for modern glycosidases. Heme binding rigidifies the ancestral glycosidase and causes a several-fold activity enhancement. Heme is not expected to catalyze glycoside hydrolysis and, in any case, the bound heme does not have access to the active site in the experimentally determined ancestral structure. The activity enhancement upon heme binding is therefore an allosteric effect likely linked to dynamics, as it is the case with other allosteric effects reported in the literature[53]. Regardless of whether this feature is truly ancestral or just a byproduct of the enhanced conformational flexibility of the putative ancestral glycosidase, it is clear that it can provide a basis for biosensor engineering. For instance, computational design and laboratory directed evolution could be used to repurpose the heme-binding site for the binding of a targeted substance of interest and to achieve a large concomitant change in glycosidase activity. The development of this application should be facilitated by the availability of a wide diversity of synthetic chemical probes for the sensitive detection of glycosidase activity[17]. In total, we anticipate that unusual combinations of protein features will generate new possibilities in protein biotechnology and engineering.

Methods

Ancestral sequence reconstruction. Characterized GH1 protein sequences were retrieved from the Carbohydrate-active enzyme database (CAZy)[18], including β-glucosidase (accession numbers: ACI19973.1 and AAL80566.1), 6-P-β-glucosidase (AIY91871.1), β-mannosidase (AAL81332.1), and myrosinase (AAK32833.1). These characterized protein sequences were utilized as seeds to identify additional homologous sequences. GH1 homologs were retrieved for all three domains of life from GenBank (http://www.ncbi.nlm.nih.gov/) using BLASTp, with the cutoff threshold of $<1 \times 10^{-5}$. Sequences with the minimum length of 300 amino acids were included in the dataset. Taxonomically redundant sequences were excluded. A total number of 150 sequences were collected for further analysis.

Sequences were aligned using T-Coffee. Initial non-bootstrapped phylogenetic trees were constructed using RAxML (ver. 8.2.11) to identify major clusters and to eliminate spurious sequences[54]. The RAxML analysis was performed with the hill-climbing mode using the gamma substitution model. These initial trees were used as starting point for more thorough Bayesian analysis. MrBayes (ver. 3.2.6) was conducted using the WAG amino acid replacement model with a gamma distribution and invariable sites model for at least 1,000,000 generations, with sampling at intervals of 100 generations, and two runs with four chains per run in order to monitor convergence[55]. Twenty-five percent of sampled points were discarded as burn-in. The tree topology was broadly identical between the RAxML and MrBayes analyses.

Ancestral sequences were reconstructed using FastML (ver. 3.1) with the WAG amino acid replacement model with a gamma distribution for variable replacement rates across sites[56].

Database searches. We searched the CAZy database in order to ascertain the presence of porphyrin rings in reported glycosidase structures. We systematically went through all 167 glycoside hydrolase families of the database in March 2020. For each family, we checked the structure section and we individually examined all the links provided to the protein data bank. Overall, we examined 5565 PDB files corresponding to 1435 different glycosidase enzymes. We did not find a single example of a reported structure with a bound porphyrin ring.

We also used the CAZy database as a starting point of an extensive literature search for optimum temperature values of family 1 glycosidases. We examined the section of characterized enzymes for family 1 glycosidases, the references included in the corresponding GenBank links, as well as the publications that cite those references in a Google Scholar search. Several hundred published articles were examined for experimental activity versus temperature profiles and reported values of the optimum temperature. We found such data for 126 different family 1 glycosidases. In many cases, the oligomerization state of the enzymes was also provided in the original references. The environmental temperatures (optimum growth temperatures) of the corresponding host organisms could be found in most cases in the "Bergey's Manual of Systematic Bacteriology" although, in some cases, literature searches were performed to find the optimum temperatures. Most organisms were classified as hyperthermophilic, extreme thermophilic, thermophilic, mesophilic or psychrophilic in Bergey's manual or the relevant literature references. We have used this classification to color code Fig. 1c, since it leads to clear and intuitive data clusters. All the information related to the values of the optimum temperature for activity and the organismal living temperature is collected in Supplementary Dataset 1.

In order to find examples of proteins with the TIM-barrel fold and a bound porphyrin ring in the reported structures, we checked (March-2020) all entries in the protein data bank that are classified as TIM-barrels in the CATH database. We examined a total of 7637 PDB files and found a porphyrin ring in only 13 of them. These 13 structures correspond to two proteins: flavocytochrome B2, a multi-domain protein in which the porphyrin ring is located in the non-TIM-barrel domain, and uroporphyrinogen decarboxylase, which is an enzyme involved in heme biosynthesis.

Protein expression and purification. The different proteins studied in this work were purified following standard procedures. Briefly, genes for the His-tagged proteins in a pET24b(+) vector with kanamycin resistance were cloned into *E. coli* BL21 (DE3) cells, and the proteins were purified by Ni-NTA affinity chromatography in HEPES buffer. The His tag was placed at the C-terminus, i.e., at a position that is well removed from the catalytic face of the barrel, the regions of enhanced conformational flexibility in the ancestral protein and the heme-binding site. Since the ancestral protein is susceptible to proteolysis, we included protease inhibitors in all steps of the purification (cOmplete® EDTA-free Protease Inhibitor Cocktail from Roche, ref. 11873580001). Protein solutions were prepared by exhaustive dialysis against the desired buffer (typically 50 mM HEPES pH 7) or by passage through PD10 columns. Protein purity was assessed by gel electrophoresis (Fig. S28). Proteins were properly folded, as judged by circular dichroism spectra (Fig. S29) [Far-UV CD spectra from 250 to 210 nm were recorded for extant and ancestral glycosidases, at 25 °C, using a Jasco J-715 spectropolarimeter equipped with a PTC-348WI. Buffer conditions were 50 mM HEPES, pH 7.0, protein concentration was within 0.2–0.6 mg/mL range and a 1 mm pathlength cuvette was used. An average of 30 scans was performed in each case. Blank subtraction was always carried out prior to mean residue ellipticity calculation, $[\Theta]_{MRW}$].

Analytical ultracentrifugation. Samples of the ancestral glycosidase in HEPES 50 mM, NaCl 150 mM, pH 7.0 were used. The assays were performed at 48,000 rpm (185,463 xg) in an XL-I analytical ultracentrifuge (Beckman-Coulter Inc.) equipped with both UV–VIS absorbance and Raleigh interference detection systems, using an An-50Ti rotor. Sedimentation profiles were recorded simultaneously by Raleigh interference and absorbance at 280 nm. Differential sedimentation coefficient distributions were calculated by least-squares boundary modeling of sedimentation velocity data using the continuous distribution c(s) Lamm equation model as implemented by SEDFIT 16.1c[57]. These experimental values were corrected to standard conditions using the program SEDNTERP[58] (version 20120111 Beta) to obtain the corresponding standard values (s20,w).

Sedimentation equilibrium assays (SE) for GH1-N72 were carried out at speeds ranging from 8,000 to 11,000 rpm (5,152 xg to 9,740 xg) and at 280 nm, using the same experimental conditions and instrument as in the SV experiments. A last high-speed run (48,000 rpm, 185,463 xg) was done to deplete protein from the meniscus region to obtain the corresponding baseline offsets. Weight-average buoyant MW of GH1-N72 were obtained by fitting a single-species model to the experimental data using the HeteroAnalysis 1.1.60 program[59] once corrected for temperature and solvent composition with the program SEDNTERP[58] (version 20120111 Beta).

Preparation of the ancestral protein with bound heme and determination of the heme to protein ratio. Stock solutions of hemin (heme with iron in the +3 oxidation state) were prepared daily in 1.4 M sodium hydroxide. Prior to use, the stock solution was diluted (typically 1:100) into HEPES buffer 50 mM, pH 7 and this solution was immediately used.

The ancestral protein with bound heme was prepared by incubating the protein with a 5-fold excess of heme for about one hour, followed by passage through a PD10 column and a Superdex-200 column to eliminate the non-bound heme. The heme to protein ratio in the resulting protein samples could be roughly estimated from the absorbance of the Soret band and the protein band at 280 nm in UV–VIS spectra. This procedure is not exact because the Soret band may depend on the interactions of the bound heme and, also, heme can show some absorption at 280 nm. For more accurate characterization protein concentration was determined by the bicinchoninic acid method[33] with the Pierce™ BCA Protein Assay Kit (ThermoFisher Scientific). A method based on pyridine hemochrome spectra[34] was used to determine the amount of heme. Briefly, 25 µl of 0.1 M potassium ferricyanide were added to a mixture of 2 mL or pyridine, 2 mL 0.1 M NaOH and 2 mL water. This solution was mixed with the protein solution in a 1:1 volume ratio and an excess of sodium dithionate was added. Lastly, the amount of heme was calculated from the absorbance of the pyridine hemochrome at 556 nm after correction for the absorbance of a blank. Using this approach, a heme to protein stoichiometry of 1.03±0.03 was determined from 5 independent measurements.

UPLC mass spectrometry. Ultra performance liquid chromatography (UPLC) was performed using a Waters Acquity H Class UPLC connected to a mass spectrometry Waters Synapt G2 Triwave® system. A 2.1 × 100 mm Protein BEH C4 column of 300 Å pore size and 2.1 µm particle size at a flow of 0.2 mL/min was used for chromatography. The mobile phase was a mixture of 0.1% formic acid–water (A) with 0.1% formic acid–acetonitrile (B) and the elution gradient were as follows: 0–10.33 min, 98–55% A; 10.33–20 min, 55–30% A; 20–21.57 min, 30% A; 21.57–23,33 min, 30–2% A; 23.33–30 min, stay 2% A. Mass spectrometry conditions were as follows: the ionization source of ESI was operated inion mode of positive (ESI +) and 2.2 kV of capillary voltage. Temperature of desolvation was 400 °C, and ion source was 100 °C. Desolvent and cone gas (nitrogen) flow velocity were 600 L/h.

Microscale thermophoresis quantification of heme-protein interaction. The motion of molecules in microscopic temperature gradients (microscale thermophoresis) is sensitive to changes in properties induced by a binding event and can be used to quantify a diversity of intermolecular interactions[60]. For these experiments, we used a His-tagged protein labeled with a fluorescent probe using the His-tag labeling kit from NanoTemper technology. A 200 nanomolar protein solution was titrated at 25 °C with heme concentrations ranging from 55 nM to 2 µM. We did not use higher heme concentration to minimize the possibility of heme association, a process that would decrease the concentration of the monomeric heme that is competent for binding. The experiments were performed with Monolith NT.115 pico from NanoTemper technology. The data were acquired with MO. Control software, version 1.6 (NanoTemper Technologies GmbH). The binding curve and affinity were modeled and analyzed in the MO.Control software, version 1.6. Three replicate experiments were performed to yield an average value for the heme dissociation constant of 547 ± 110 nanomolar. Relevant plots and validation reports for the three experiments are shown in Fig. S14.

Activity determinations. Glucosidase and galactosidase activities were tested following the absorbance of p-nitrophenol at 405 nm upon the hydrolysis of 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside[27]. Rates were calculated from the initial absorbance vs. time slope and the known extinction coefficient of p-nitrophenol at pH 7. Experiments at different substrate

concentrations were carried out to arrive at Michaelis plots for the ancestral and several modern glycosidases studied in this work. For the rate determination at a wide range of substrate concentrations we used a protocol designed to minimize any changes in buffer composition that could distort the profiles. Thus, for a rate measurement at a given substrate concentration, an enzyme solution in HEPES buffer at pH 7 was mixed with an equal volume of substrate dissolved in pure water. To minimize pH changes, the initial enzyme solution was prepared in 200 mM buffer to yield a final buffer concentration of 100 mM. We confirmed that the pH changes upon mixing were negligible. See legends to Figs. S6–S9 for details on data analysis.

As it is common in the literature, values of the optimum activity temperatures were determined from the profiles of activity versus temperature derived from measurements performed after several-minute incubations at each temperature[25]. Briefly, the protein was incubated at the desired temperature with 1 mM substrate in HEPES buffer 50 mM pH 7 and, after 10 min, the reaction was stopped by adding sodium carbonate to a concentration of 0.5 M. The amount of substrate hydrolyzed was determined from the absorbance of p-nitrophenol at 405 nm. We confirmed that the 10-min incubation only hydrolyzed a fraction of the substrate present and, therefore, that the amount of substrate hydrolyzed after a 10-min. incubation is a suitable metric of enzyme activity. Profiles of activity versus temperature were determined using both 4-nitrophenyl-β-D-glucopyranoside and 4-nitrophenyl-β-D-galactopyranoside. The profiles for the ancestral glycosidase are shown in Fig. 1b and those for the modern glycosidase from *Halothermothrix orenii* are given in Fig. S5. In all cases, the profiles show a sharp maximum from which an unambiguous determination of the optimum temperature is possible. Note also that there is good agreement between the optimum temperature values derived using the two different substrates. Differential scanning calorimetry experiments were performed as we have previously described in detail[9].

Glycosidase substrates were obtained from commercial sources, except *4*-nitrophenyl-β-D-glucopyranoside-6-phosphate, which was prepared by us on the basis of a published procedure[61]. The chemical identity of the prepared compound was confirmed by mass spectrometry and nuclear magnetic resonance.

Proteolysis experiments. For proteolysis experiments, the ancestral glycosidase and the modern glycosidase from *Halothermothrix orenii* at a concentration of 1 mg/mL were incubated at 25 °C with thermolysin for different times in HEPES buffer 50 mM pH 7 containing 10 mM calcium chloride. Stock solutions of thermolysin were prepared fresh in the same solvent at a concentration of 1 mg/mL and were diluted 1:10 when added to the protein solution. The reaction was stopped by the addition of EDTA to a final concentration of 12.5 mM and aliquots were loaded into 15% (w/v) SDS-PAGE gels for electrophoresis. In some experiments, fragments separated by electrophoresis were extracted, desalted and subjected to LC-MS/MS analysis for mass determination. Fragment masses were determined by MALDI and their sequences were investigated using peptide mapping fingerprinting and MALDI-TOF/TOF (Fig. S2). This allowed us to locate approximately the cleavage sites, as shown in Fig. 2c.

Crystallization and structure determination. The ancestral glycosidase, dissolved in 150 mM NaCl, 50 mM HEPES pH 7.0, was concentrated to 35 mg/mL and to 70 mg/mL for the vapor-diffusion (VD) and counter-diffusion crystallization experiments, respectively. We checked by SDS electrophoresis that the concentrated protein used for crystallization was not proteolyzed. Hanging-drops VD experiments were prepared by mixing 1 µL of protein solution with the reservoir, in a 1:1 ratio, and equilibrated against 500 µL of each precipitant cocktail HR-I (Crystal Screen 1, Hampton Research). Capillary counter-diffusion experiments were set up in capillaries of 0.3 mm inner diameter using the CSK-24, AS-49 and PEG448-49 screening kits[62]. A similar procedure was followed for the crystallization of the ancestral glycosidase-heme complex, using two fixed concentrations at 75 and 30 mg/ml for the counter-diffusion and VD experiments. Experiments were performed at 293 K.

Crystals of the ancestral glycosidase were obtained only in condition #41 of HR-I, whilst the GH1N72-Heme complex crystallized in conditions #6 and #9 of HR-I and PPP8 of the mix of PEG counter-diffusion screen. Crystals were extracted either from the capillary or fished directly from the drop and subsequently cryoprotected by equilibration with 15 % (v/v) glycerol prepared in the mother liquid, flash-cooled in liquid nitrogen and stored until data collection. Crystals were diffracted at the XALOC beamline of the Spanish synchrotron light radiation source (ALBA, Barcelona). Indexed data were scaled and reduced using the CCP4 program suite[63].

Initial data sets were obtained for the ancestral glycosidase crystals diffracting the X-ray to 2.5 Å. The clean (without water, ligands, etc.) 3D model of the β-glucosidase from *Thermotoga maritima* (PDB ID. 2J78) [https://www.rcsb.org/structure/2J78] was used as search model for molecular replacement[63]. Two monomers were found in the asymmetric unit as expected from the Matthews coefficient for the P2(1) space group. Refinement, including Titration-Libration-Screw (TLS) parametrization, water pick, and model validation was carried out with PHENIX suite[64]. Unidentifiable amino acids in the highly disordered region have been assigned as poly-UKN chains C and D corresponding to the 18 and 14 (poly-Alanine) of chains A and B, respectively.

Crystals of the ancestral glycosidase-heme complex belong to the same space group than the ancestral glycosidase but were not isomorphous. The determined unit cell was bigger accommodating three polypeptide chains in the asymmetric as determined from the Matthews coefficient. A similar protocol was followed to place the three monomers in the unit cell by molecular replacement and to refine the structure. After a first refinement round the presence of one protoporphyrin ring in each polypeptide chain was determined. It was also clear that disordered regions of the ancestral glycosidase model were visible in the heme complex model.

The summary of data collection, refinement statistics, and quality indicators are collected in Table S9. The coordinates and the experimental structure factors have been deposited in the Protein Data Bank with ID 6Z1M [https://www.rcsb.org/structure/6Z1M] and 6Z1H [https://www.rcsb.org/structure/6Z1H] for the ancestral glycosidase with and without bound heme, respectively.

Figures displaying 3D-structures have been prepared using PyMOL (The Pymol Molecular Graphics System, Schrödinger, LLC). The 2D-interaction diagram of Fig. 7a was prepared using LigPlot$^+$ (https://www.ebi.ac.uk/thornton-srv/software/LIGPLOT/).

Molecular dynamics simulations. Molecular simulations were performed on both ancestral glycosidases (this work, PDB ID: 6Z1M [https://www.rcsb.org/structure/6Z1M], 6Z1H [https://www.rcsb.org/structure/6Z1H]) and the modern glycosidase from *Halothermothrix orenii* (PDB ID: 4PTV [https://www.rcsb.org/structure/4PTV]). The structure of the heme-free ancestral glycosidase was obtained by manually deleting the heme coordinates from the corresponding heme-bound crystal structure. The missing regions of the ancestral glycosidases were reconstructed using MODELLER. Histidine protonation states were selected based on empirical pK_a estimates performed using PROPKA 3.1 and visual inspection. All other residues were placed in their standard protonation states at physiological pH. The heme group was described using a bonded model, creating a bond between the Tyr264 side chain and the Fe(III) atom of the heme (Fig. S30). We used MCPB.py as implemented in AMBER19[65] to obtain the necessary parameters for creating the bonding pattern between the Fe(III) atom and the 4 nitrogen atoms of the heme and the tyrosine side chain oxygen. The resulting structure was then optimized, and frequency calculations were performed at the ωB97X-D/6-31 G* level of theory followed by the Seminario method[66] to obtain the force constants from the Hessian of the frequency calculation. This functional is a long-range (LC) corrected hybrid functional (see Chai and Head-Gordon[67] and references cited therein), which describes short-range interactions using an exchange functional, and long-range interactions using 100% Hartree-Fock exchange. ωB97X-D further improves on the concept of LC functionals, by systematic optimization of the functional, including the inclusion of an extra parameter that allows for an adjustable fraction of short-range exchange. In addition, this functional incorporates an empirical dispersion correction, following the DFT-D scheme[68]. We chose this functional for our parameterization as it yielded optimized structures of the heme complex that were similar to that observed in the heme complex, which is important in order to be able to maintain the heme in the binding pocket in a planar (non-distorted) conformation in our bonded model. This was further corroborated in our MD simulations, where the heme maintained a planar conformation without any unphysical distortions of dihedral angles. We note that only the heme and the Tyr264 side chain were considered as QM atoms in our model, as this is the region that was necessary to parameterize in our bonded model, following Li and Merz[69].

Partial charges were obtained for the heme and for the Tyr264 side chain at the HF/6-31 G* level of theory, using the restrained electrostatic potential (RESP) approach, following the MCPB.py protocol, and performing the calculations using Gaussian 09 Rev. E.01. Periodic boundary conditions (PBC) were used, and all systems were solvated in a truncated octahedral box filled with TIP3P water molecules[70], with 10 Å from the solute to the box edges in all directions. The truncated octahedron can fill space without leaving any gaps, and since our protein has a globular shape, a truncated octahedral box is the most suitable box shape to reduce the number of water molecules necessary to fill the box, which saves substantial computational time. with a distance of 10 Å from the solute to the surface of the box. The box was then filled with TIP3P water molecules. Na^+ and Cl^- counter ions were added to the system to neutralize each enzyme. The protein was described using the AMBER ff14SB force field[71], and the heme was described using the General AMBER Force Field (GAFF)[72].

Following system preparation, the LEaP module of AMBER19 was used to generate the topology and coordinate files for the MD simulations, which were performed using the CUDA version of the PMEMD module of the AMBER19 simulation package. The solvated system was first subjected to a 5000 step steepest descent minimization, followed by a 5000 step conjugate gradient minimization with positional restraints on all heavy atoms of the solute, using a 5 kcal mol^{-1} Å^{-2} harmonic potential. The minimized system was then heated up to 300 K using the Berendsen thermostat[73], with a time constant of 1 ps for the coupling, and 5 kcal mol^{-1} Å^{-2} positional restraints (again a harmonic potential) applied during the heating process. The positional restraints were then gradually decreased to 1 kcal mol^{-1} Å^{-2} over five 500 ps steps of NPT equilibration, using the Berendsen thermostat and barostat to keep the system at 300 K and 1 atm. For the production runs, each system was subjected to either 500 ns of sampling in an NPT ensemble at constant temperature (300 K) and constant pressure (1 atm), controlled by the Langevin thermostat, with a collision

frequency of 2.0 ps^{-1}, and the Berendsen barostat with a coupling constant of 1.0 ps. A 2 fs time step was used for all simulations, and snapshots were saved from the simulation every 5 ps. The SHAKE algorithm[74] was applied to constrain all bonds involving hydrogen atoms. A 10 Å cutoff was applied to all nonbonded interactions, with the electrostatic interactions being treated with the particle mesh Ewald (PME) approach[75]. 10 independent simulations were performed for each starting structure during 500 ns (for RMSD convergence, see Fig. S18). All the subsequent analyses were performed with the CPPTRAJ toolkit from Ambertools19[65]. Parameters used to describe the heme, input files, snapshots from our simulations, and simulation trajectories (with water molecules and ions removed to save file size) are available for download from Zenodo (https://zenodo.org) at https://doi.org/10.5281/zenodo.3857791.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data included in the figures and supporting the findings of this study are available from the corresponding authors upon reasonable request. Atomic coordinates and the experimental structure factors have been deposited in the Protein Data Bank (https://www.rcsb.org) with ID 6Z1M and 6Z1H for the ancestral glycosidase with and without bound heme, respectively. Parameters used to describe the heme, input files, snapshots from our molecular dynamics simulations, and simulation trajectories (with water molecules and ions removed to save file size) are available for download from Zenodo (https://zenodo.org) at https://doi.org/10.5281/zenodo.3857791. Source data are provided with this paper.

Received: 3 June 2020; Accepted: 11 December 2020;
Published online: 15 January 2021

References

1. Pauling, L. & Zuckerkandl, E. Chemical paleogenetics. Molecular "restoration studies" of extinct forms of life. *Acta Chem. Scan.* **17S**, 9–16 (1963).
2. Hochberg, G. K. A. & Thornton, J. W. Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).
3. Gumulya, Y. & Gillam, E. M. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering. *Biochem. J.* **474**, 1–19 (2017).
4. Cole, M. F. & Gaucher, E. A. Exploiting models of molecular evolution to efficiently direct protein engineering. *J. Mol. Evol.* **72**, 193–203 (2011).
5. Risso, V. A., Sanchez-Ruiz, J. M. & Ozkan, S. B. Biotechnological and protein engineering implications of ancestral protein resurrection. *Curr. Opin. Struct. Biol.* **51**, 106–115 (2018).
6. Trudeau, D. L. & Tawfik, D. S. Protein engineers turned evolutionists—the quest for the optimal starting point. *Curr. Opin. Biotechnol.* **60**, 46–52 (2019).
7. Siddiq, M. A., Hochberg, G. K. & Thornton, J. W. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* **47**, 113–122 (2017).
8. Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167 (2009).
9. Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.* **135**, 2899–2902 (2013).
10. Gardner, J. M., Biler, M., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. L. Manipulating conformational dynamics to repurpose ancient proteins for modern catalytic functions. *ACS Catal.* **10**, 4863–4870 (2020).
11. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).
12. Nagano, N., Orengo, C. A. & Thornton, J. M. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765 (2002).
13. Goldman, A. D., Beatty, J. T. & Landweber, L. F. The TIM barrel architecture facilitated the early evolution of protein-mediated metabolism. *J. Mol. Evol.* **82**, 17–26 (2016).
14. Grunwald, P. *Biocatalysis: Biochemical Fundamentals and Applications* 2nd edn. (World Scientific, New York, 2017).
15. Wolfenden, R., Lu, X. & Young, G. Spontaneous hydrolysis of glycosides. *J. Am. Chem. Soc.* **120**, 6814–6815 (1998).
16. Zechel, D. L. & Withers, S. G. Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res.* **33**, 11–18 (2000).
17. Burke, H. M., Gunnlaugsson, T. & Scanian, E. M. Recent advances in the development of synthetic chemical probes for glycosidase enzymes. *Chem. Commun.* **51**, 10576–10588 (2015).
18. Lombard, V. et al. The Carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
19. CAZypedia Consortium. Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes. *Glycobiology* **28**, 3–8 (2018).
20. Ingles-Prieto, A. et al. Conservation of protein over four billion years. *Structure* **21**, 1–8 (2013).
21. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372**, 631–634 (1994).
22. Withers, S. Glycoside hydrolase family 1. CAZypedia, available at http://www.cazypedia.org/. Accessed 19 April 2020.
23. Weiss, C. W. et al. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
24. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
25. Wood, T. M. & Bhat, K. M. Methods for measuring cellulase activities. *Methods Enzymol.* **160**, 87–112 (1988).
26. Park, C., Zhou, S., Gilmore, J. & Marqusee, S. Energetics-based protein profiling on a proteomic scale: identification of proteins resistant to proteolysis. *J. Mol. Biol.* **368**, 1426–1437 (2007).
27. Hassan, N. et al. Biochemical and structural characterization of a thermostable β-glucosidase from *Halothermothrix orenii* for galacto-oligosaccharide synthesis. *Appl. Microbiol. Biotechnol.* **99**, 1731–1744 (2015).
28. Marana, S. R. Molecular basis of substrate specificity in family 1 glycoside hydrolases. *IUBMB Life* **58**, 63–73 (2006).
29. Devamani, T. et al. Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J. Am. Chem. Soc.* **138**, 1046–1056 (2016).
30. Vanderkooi, G. & Stotz, E. Reductive alteration of heme α hemochromes. *J. Biol. Chem.* **240**, 3418–3424 (1965).
31. Fiege, K., Querebillo, C. J., Hildebrandt, P. & Frankenberg-Dinkel, N. Improved method for the incorporation of heme cofactors into recombinant proteins using Escherichia coli Nissle 1917. *Biochemistry* **57**, 2747–2755 (2018).
32. Inada, Y. & Shibata, K. The Soret band of monomeric hematin and its changes on polymerization. *Biochem. Biophys. Res. Commun.* **9**, 323–327 (1962).
33. Smith, P. K. et al. Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150**, 76–85 (1985).
34. Berry, E. A. & Trumpower, B. L. Simultaneous determination of hemes a, b, and c from pyridine hemochrome spectra. *Anal. Biochem.* **161**, 1–15 (1987).
35. Smith, L. J., Kahramam, A. & Thornton, J. M. Heme proteins—diversity in structural characteristics, function, and folding. *Proteins* **78**, 2349–2368 (2010).
36. Schneider, S., Marles-Wright, J., Sharp, K. H. & Paoli, M. Diversity and conservation of interactions for binding heme in b-type heme proteins. *Nat. Prod. Rep.* **24**, 621–630 (2007).
37. Li, T., Bonkovsky, H. L. & Guo, J. Structural analysis of heme proteins: implications for design and prediction. *BMC Struct. Biol.* **11**, 13 (2011).
38. Wiita, A. P. et al. Probeing the chemistry of thioredoxin catalysis with force. *Nature* **450**, 124–127 (2007).
39. Sigala, P. A. et al. Testing geometrical discrimination within an enzyme active site: constrained hydrogen bonding in the ketostteroid isomerase oxyanion hole. *J. Am. Chem. Soc.* **130**, 13696–13798 (2008).
40. Lüdtke, S. et al. Sub-ångström-resolution crystallography reveals physical distortions that enhance reactivity of a covalent enzymatic intermediate. *Nat. Chem.* **5**, 762–767 (2013).
41. James, L. C. & Tawfik, D. S. Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368 (2003).
42. Bershtein, S. & Tawfik, D. S. Advances in laboratory evolution of enzymes. *Curr. Opin. Chem. Biol.* **12**, 151–158 (2008).
43. Petrović, D., Risso, V. A., Kamerlin, S. C. L. & Sanchez-Ruiz, J. M. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* **15**, 20180330 (2018).
44. Pabis, A., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. L. Cooperativity and flexibility in enzyme evolution. *Curr. Opin. Struct. Biol.* **48**, 83–92 (2018).
45. Risso, V. A. et al. *De novo* active sites for resurrected Precambrian enzymes. *Nat. Commun.* **8**, 16113 (2017).
46. Höcker, B., Beismann-Driemeyer, S., Heltwer, S., Lustig, A. & Sterner, R. Dissection of a $(\beta\alpha)_8$-barrel enzyme into two folded halves. *Nat. Struct. Biol.* **8**, 32–36 (2001).
47. Gamiz-Arco, G. et al. Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding. *Biochem. J.* **476**, 3631–3647 (2019).
48. Randall, R. N., Radford, C. E., Roof, K. A., Natarajan, D. K. & Gaucher, E. A. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat. Commun.* **7**, 12847 (2016).
49. Holm, L. DALI and the persistence of protein shape. *Protein Sci.* **29**, 128–140 (2020).
50. Chu, X.-Y. & Zhang, H.-Y. Cofactors as molecular fossils to trace the origin and evolution of proteins. *ChemBioChem* https://doi.org/10.1002/cbic.202000027 (2020).

51. Chen, K. & Arnold, F. H. Engineering new catalytic activities in enzymes. *Nat. Catal.* **3**, 203–213 (2020).
52. Dawson, N. L. et al. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acid Res.* **45**, D289–D295 (2017).
53. Naganathan, A. N. Modulation of allosteric coupling by mutations: from protein dynamics and packing to altered native ensembles and function. *Curr. Opin. Struct. Biol.* **54**, 1–9 (2019).
54. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
55. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
56. Ashkenazy, H. et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* **40**, W580–W584 (2012).
57. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**, 1606–1619 (2000).
58. Laue, T. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. in *Analytical Ultracentrifugation in Biochemistry and Polymer Science* (eds Harding, S. E., Rowe, A. J. & Horton, J. C.) 90–125 (Royal Society of Chemistry, Cambridge, 1992).
59. Cole, J. L. Analysis of heterogeneous interactions. *Methods Enzymol.* **384**, 212–232 (2004).
60. Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P. & Duhr, S. Molecular interactions studies using microscale thermophoresis. *Assay. Drug Dev. Technol.* **9**, 342–353 (2011).
61. Acebrón, I. et al. Structural basis of the substrate specificity and instability in solution of a glycosidase from *Lactobacillus plantarum*. *BBA—Proteins Proteom.* **1865**, 1227–1236 (2017).
62. González-Ramírez, L. A. et al. Efficient screening methodology for protein crystallization based on the counter-diffusion technique. *Cryst. Growth Des.* **17**, 6780–6786 (2017).
63. Collaborative, C. P. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **D 50**, 760–763 (1994).
64. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr.* **D 66**, 213–221 (2010).
65. Case, D. A. et al. *AMBER 2019*. (University of California, San Francisco, 2019).
66. Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.* **60**, 1271–1277 (1996).
67. Chai, J.-D. & Head-Godon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
68. Sato, T., Tsuneda, T. & Hirao, K. Long-range corrected density functional study on weakly bound systems: Balanced descriptions of various types of molecular interactions. *J. Chem. Phys.* **126**, 234114 (2007).
69. Li, P. & Merz, K. M. MCPB.py: a python based metal center parameter builder. *J. Chem. Inf. Model* **56**, 599–604 (2016).
70. Jorgensen, W. L., Chandrasenkar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
71. Maier, J. A. et al. ff14SB: improving the accuracy of protein side chain and backbone parameters From ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
72. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
73. Beredsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
74. Ryckaert, J. P., Cicotti, G. & Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
75. Darden, T., York, D. & Pedersein, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

Acknowledgements

This work was supported by Human Frontier Science Program Grant RGP0041 (J.M.S.-R., E.A.G., B.S., and S.C.L.K.), NIH grant R01AR069137 (E.A.G.), Department of Defense grant MURI W911NF-16-1-0372 (E.A.G.), the Swedish Research Council (2019-03499) (S.C.L.K.), the Knut and Alice Wallenberg Foundation (2018.0140 and 2019.0431) (S.C.L.K.), Spanish Ministry of Economy and Competitiveness/FEDER Funds Grants BIO2015-66426-R (J.M.S.-R.) RTI2018-097142-B-100 (J.M.S.-R.) and BIO2016-74875-P (J.A.G.). The simulations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX partially funded by the Swedish Research Council through grant agreement no. 2016-07213. We acknowledge the Spanish Synchrotron Radiation Facility (ALBA, Barcelona) for the provision of synchrotron radiation facilities and the staff at XALOC beamline for their invaluable support. We are also grateful to Victoria Longobardo Polanco (Proteomic Unit, Institute of Parasitology and Biomedicine "López-Neyra") for help with mass spectrometry experiments and data analyses and to Juan Román Luque Ortega (Molecular Interactions Facility, Centro de Investigaciones Biológicas Margarita Salas) for help with ultracentrifugation experiments and data analyses.

Author contributions

B.S., S.C.L.K., E.A.G., and J.M.S.-R. designed the research. G.G.-A. and L.I.G.-R. prepared the protein variants and designed, performed and analyzed experiments addressed at determining their catalytic and biophysical features, under the supervision of V.A.-R. and B.I.-M., who also provided essential input regarding the interpretation of these properties. V.A.R. was in charge of mass spectrometry, ultracentrifugation, and thermophoresis experiments. Y.H. carried out ancestral sequence reconstruction under the supervision of E.A.G., who also provided essential input for the interpretation of the results in an evolutionary context. D.P. performed homology modeling under the supervision of S.C.L.K. Organic synthesis was performed by J.J. and J.M.C. who provided essential input regarding the properties of the synthesized compound. A.R.-R. carried out MD simulations under the supervision of S.C.L.K., and they provided the general interpretation and implications of the simulations. L.I.G.-R. and V.A.R. carried out protein crystallization. J.A.G. determined the X-ray structures and provided essential input regarding their interpretation and implications. J.M.S.-R. wrote the first draft of the manuscript to which B.S., S.C.L.K., and E.A.G. added crucial paragraphs and sections. All authors discussed the manuscript, suggested modifications and improvements, and contributed to the final version.

Funding

Open Access funding provided by Uppsala University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-20630-1.

Correspondence and requests for materials should be addressed to S.C.L.K., E.A.G. or J.M.S.-R.

Peer review information *Nature Communications* thanks Vickery Arcus, John Mitchell, Matilda Newton, and other, anonymous, reviewers for their contributions to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Pauling, L. & Zuckerkandl, E. Chemical paleogenetics. Molecular “restoration studies” of extinct forms of life. *Acta Chem. Scand.* **17**, S9–S16 (1963).

2. Garcia, A. K. & Kaçar, B. How to resurrect ancestral proteins as proxies for ancient biogeochemistry. *Free Radic. Biol. Med.* **140**, 260–269 (2019).

3. Liberles, D. A. Ancestral Sequence Reconstruction. *Ancestral Seq. Reconstr.* 1–272 (2008). doi:10.1093/ACPROF:OSO/9780199299188.001.0001

4. Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P. & Benner, S. A. The ribonuclease from an extinct bovid ruminant. *FEBS Lett.* **262**, 104–106 (1990).

5. Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. & Wilson, A. C. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* **345**, 86–89 (1990).

6. Zaucha, J. & Heddle, J. Resurrecting the Dead (Molecules). *Comput. Struct. Biotechnol. J.* **15**, 351–358 (2017).

7. Willerslev, E. *et al.* Ancient Biomolecules from Deep Ice Cores Reveal a Forested Southern Greenland. *Science* **317**, 111 (2007).

8. Gumulya, Y. & Gillam, E. M. J. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the ‘retro’ approach to protein engineering. *Biochem. J.* **474**, 1–19 (2017).

9. Perez-Jimenez, R. *et al.* Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* **18**, 592–6 (2011).

10. Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A. & Sanchez-Ruiz, J. M. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian β-Lactamases. *J. Am. Chem. Soc.* **135**, 2899–2902 (2013).

11. Akanuma, S. *et al.* Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl. Acad. Sci.* **110**, 11067–11072 (2013).

12. Risso, V. A., Gavira, J. A. & Sanchez-Ruiz, J. M. Thermostable and promiscuous Precambrian proteins. *Environ. Microbiol.* **16**, 1485–1489 (2014).

13. Gaucher, E. A., Govindarajan, S. & Ganesh, O. K. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704–707 (2008).

14. Alcalde, M. When directed evolution met ancestral enzyme resurrection. *Microb. Biotechnol.* **10**, 22–24 (2017).

15. Finnigan, G., Hanson-Smith, V., Stevens, T. & Thornton, J. Evolution of increased complexity in a molecular machine. *Nature* **481**, 360–364 (2012).

16. Bickelmann, C. *et al.* The molecular origin and evolution of dim-light vision in mammals. *Evolution* **69**, 2995–3003 (2015).

17. Knauth, L. P. Temperature and salinity history of the Precambrian ocean: implications for the course of microbial evolution. *Geobiol. Object. Concepts, Perspect.* 53–69 (2005). doi:10.1016/B978-0-444-52019-7.50007-3

18. Romero-Romero, M. L. *et al.* Selection for Protein Kinetic Stability Connects Denaturation Temperatures to Organismal Temperatures and Provides Clues to Archaean Life. *PLoS One* **11**, e0156657 (2016).

19. Kratzer, J. T. *et al.* Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc. Natl. Acad. Sci.* **111**, 3763–3768 (2014).

20. Carrigan, M. A. *et al.* Hominids adapted to metabolize ethanol long before human-directed fermentation. *Proc. Natl. Acad. Sci.* **112**, 458–463 (2015).

21. Ingles-Prieto, A. *et al.* Conservation of protein structure over four billion years. *Structure* **21**, 1690 (2013).

22. Risso, V. A. *et al.* De novo active sites for resurrected Precambrian enzymes. *Nat. Commun. 2017 81* **8**, 1–13 (2017).

23. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science (80-.).* **317**, 1544–1548 (2007).

24. Konno, A., Kitagawa, A., Watanabe, M., Ogawa, T. & Shirai, T. Tracing Protein Evolution through Ancestral Structures of Fish Galectin. *Structure* **19**, 711–721 (2011).

25. Kim, H. *et al.* A Hinge Migration Mechanism Unlocks the Evolution of Green-to-Red Photoconversion in GFP-like Proteins. *Structure* **23**, 34–43 (2015).

26. Kim, H. *et al.* Acid–Base Catalysis and Crystal Structures of a Least Evolved Ancestral GFP-like Protein Undergoing Green-to-Red Photoconversion. *Biochemistry* **52**, 8048–8059 (2013).

27. Boucher, J. I., Jacobowitz, J. R., Beckett, B. C., Classen, S. & Theobald, D. L. An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *Elife* **3**, e02304 (2014).

28. Risso, V. A., Sanchez-Ruiz, J. M. & Ozkan, S. B. Biotechnological and protein-engineering implications of ancestral protein resurrection. *Curr. Opin. Struct. Biol.* **51**, 106–115 (2018).

29. Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167 (2009).

30. Devamani, T. *et al.* Catalytic Promiscuity of Ancestral Esterases and Hydroxynitrile Lyases. *J. Am. Chem. Soc.* **138**, 1046–1056 (2016).

31. Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R. & Damborsky, J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *ChemBioChem* **18**, 1448–1456 (2017).

32. Ayuso-Fernández, I., Martínez, A. T. & Ruiz-Dueñas, F. J. Experimental recreation of the evolution of lignin-degrading enzymes from the Jurassic to date. *Biotechnol. Biofuels* **10**, 1–13 (2017).

33. Garcia, A. K., Schopf, J. W., Yokobori, S., Akanuma, S. & Yamagishi, A. Reconstructed ancestral enzymes suggest long-term cooling of Earth's photic zone since the Archean. *Proc. Natl. Acad. Sci.* **114**, 4619–4624 (2017).

34. Nguyen, V. *et al.* Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science (80-.).* **355**, 289–294 (2017).

35. Romero-Romero, M., Risso, V., Martinez-Rodriguez, S., Ibarra-Molero, B. & Sanchez-Ruiz, J. Engineering ancestral protein hyperstability. *Biochem. J.* **473**, 3611–3620 (2016).

36. Glembo, T. J., Farrell, D. W., Gerek, Z. N., Thorpe, M. F. & Ozkan, S. B. Collective Dynamics Differentiates Functional Divergence in Protein Evolution. *PLOS Comput. Biol.* **8**, e1002428 (2012).

37. Delgado, A., Arco, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. Using Resurrected Ancestral Proviral Proteins to Engineer Virus Resistance. *Cell Rep.* **19**, 1247–1256 (2017).

38. Wijma, H. J., Floor, R. J. & Janssen, D. B. Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.* **23**, 588–594 (2013).

39. Risso, V. A. *et al.* Mutational Studies on Resurrected Ancestral Proteins Reveal Conservation of Site-Specific Amino Acid Preferences throughout Evolutionary History. *Mol. Biol. Evol.* **32**, 440–455 (2015).

40. Zakas, P. M. *et al.* Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat. Biotechnol.* **35**, 35–37 (2016).

41. Gupta, R. D. Recent advances in enzyme promiscuity. *Sustain. Chem. Process.* **4**, 1–7 (2016).

42. Ding, W., Ji, X., Li, Y. & Zhang, Q. Catalytic Promiscuity of the Radical S-adenosyl-L-methionine Enzyme NosL. *Front. Chem.* **4**, (2016).

43. O'Brien, P. J. & Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, R91–R105 (1999).

44. Copley, S. D. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.* **7**, 265–272 (2003).

45. Bornscheuer, U. T. & Kazlauskas, R. J. Catalytic Promiscuity in Biocatalysis: Using Old Enzymes to Form New Bonds and Follow New Pathways. *Angew. Chemie Int. Ed.* **43**, 6032–6040 (2004).

46. Khersonsky, O., Roodveldt, C. & Tawfik, D. S. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.* **10**, 498–508 (2006).

47. Humble, M. S. & Berglund, P. Biocatalytic Promiscuity. *European J. Org. Chem.* **2011**, 3391–3401 (2011).

48. Pandya, C., Farelli, J. D., Dunaway-Mariano, D. & Allen, K. N. Enzyme Promiscuity: Engine of Evolutionary Innovation. *J. Biol. Chem.* **289**, 30229–30236 (2014).

49. Copley, S. D. An evolutionary biochemist's perspective on promiscuity. *Trends Biochem. Sci.* **40**, 72–78 (2015).

50. Copley, S. D. Shining a light on enzyme promiscuity. *Curr. Opin. Struct. Biol.* **47**, 167–175 (2017).

51. Jensen, R. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).

52. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).

53. Kazlauskas, R. J. Enhancing catalytic promiscuity for biocatalysis. *Curr. Opin. Chem. Biol.* **9**, 195–201 (2005).

54. Tokuriki, N. & Tawfik, D. S. Protein Dynamism and Evolvability. *Science (80-.).* **324**, 203–207 (2009).

55. Pabis, A., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. Cooperativity and flexibility in enzyme evolution. *Curr. Opin. Struct. Biol.* **48**, 83–92 (2018).

56. James, L. C. & Tawfik, D. S. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368 (2003).

57. Zou, T., Risso, V. A., Gavira, J. A., Sanchez-Ruiz, J. M. & Ozkan, S. B. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Mol. Biol. Evol.* **32**, 132–143 (2015).

58. Berg, J. M., Tymoczko, J. L. & Stryer, L. Protein Structure and Function. in *Biochemistry* (W H Freeman, 2002).

59. Groß, A., Hashimoto, C., Sticht, H. & Eichler, J. Synthetic Peptides as Protein Mimics. *Front. Bioeng. Biotechnol.* **3**, 211 (2016).

60. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The Protein Folding Problem. *Annu. Rev. Biophys.* **37**, 289 (2008).

61. EPSTEIN, C. J. & ANFINSEN, C. B. The reversible reduction of disulfide bonds in trypsin and ribonuclease coupled to carboxymethyl cellulose - PubMed. *J Biol Chem* **237**, 2175–2179 (1962).

62. Balchin, D., Hayer-Hartl, M. & Hartl, F. U. In vivo aspects of protein folding and quality control. *Science (80-.).* **353**, aac4354 (2016).

63. Yon, J. M. & Betton, J. M. Protein folding in vitro and in the cellular environment. *Biol. Cell* **71**, 17–23 (1991).

64. Hartl, F. U. & Hayer-Hartl, M. Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.* **16**, 574–581 (2009).

65. Cabrita, L., Dobson, C. & Christodoulou, J. Protein folding on the ribosome. *Curr. Opin. Struct. Biol.* **20**, 33–45 (2010).

66. Kaiser, C., Goldman, D., Chodera, J., Tinoco, I. & Bustamante, C. The ribosome modulates nascent protein folding. *Science* **334**, 1723–1727 (2011).

67. Thommen, M., Holtkamp, W. & Rodnina, M. V. Co-translational protein folding: progress and methods. *Curr. Opin. Struct. Biol.* **42**, 83–89 (2017).

68. Samelson, A. J. *et al.* Kinetic and structural comparison of a protein's cotranslational folding and refolding pathways. *Sci. Adv.* **4**, eaas9098 (2018).

69. Liutkute, M., Samatova, E. & Rodnina, M. V. Cotranslational Folding of Proteins on the Ribosome. *Biomol. 2020, Vol. 10, Page 97* **10**, 97 (2020).

70. Lakshmipathy, S. K., Gupta, R., Pinkert, S., Etchells, S. A. & Hartl, F. U. Versatility of Trigger Factor Interactions with Ribosome-Nascent Chain Complexes. *J. Biol. Chem.* **285**, 27911 (2010).

71. Gleason, F. & Holmgren, A. Thioredoxin and related proteins in procaryotes. *FEMS Microbiol. Rev.* **4**, 271–297 (1988).

72. Arne, H. Thioredoxin. *Annu. Rev. Biochem.* **54**, 237–271 (1985).

73. Kumar, J., Tabor, S. & Richardson, C. Proteomic analysis of thioredoxin-targeted proteins in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3759–3764 (2004).

74. Modi, T., Huihui, J., Ghosh, K. & Ozkan, S. B. Ancient thioredoxins evolved to modern-day stability–function requirement by altering native state ensemble. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170184 (2018).

75. Georgescu, R., Li, J., Goldberg, M., Tasayco, M. & Chaffotte, A. Proline isomerization-independent accumulation of an early intermediate and heterogeneity of the folding pathways of a mixed alpha/beta protein, Escherichia coli thioredoxin. *Biochemistry* **37**, 10286–10297 (1998).

76. Santos, J. *et al.* Structural Selection of a Native Fold by Peptide Recognition. Insights into the Thioredoxin Folding Mechanism. *Biochemistry* **48**, 595–607 (2009).

77. Mancusso R, Cruz E, Cataldi M, Mendoza C, Fuchs J, Wang H, Yang X, T. M. Reversal of Negative Charges on the Surface of Escherichia coli Thioredoxin: Pockets versus Protrusions. *Biochemistry* **43**, 3835–3843 (2004).

78. de Lamotte-Guéry, F. *et al.* Structural and functional roles of a conserved proline residue in the alpha2 helix of Escherichia coli thioredoxin. *Protein Eng. Des. Sel.* **10**, 1425–1432 (1997).

79. Gleason, F. Mutation of conserved residues in Escherichia coli thioredoxin: effects on stability and function. *Protein Sci.* **1**, 609–616 (1992).

80. Johnson, I. S. Human Insulin from Recombinant DNA Technology. *Science (80-.).* **219**, 632–637 (1983).

81. Baeshen, M. *et al.* Production of Biopharmaceuticals in E. coli: Current Scenario and Future Perspectives. *J. Microbiol. Biotechnol.* **25**, 953–962 (2015).

82. Coughlan, L. M., Cotter, P. D., Hill, C. & Alvarez-Ordóñez, A. Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. *Front. Microbiol.* **6**, 672 (2015).

83. Kamble, P. & Vavilala, S. L. Discovering novel enzymes from marine ecosystems: a metagenomic approach. *Bot. Mar.* **61**, 161–175 (2018).

84. Uchiyama, T. & Miyazaki, K. Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* **20**, 616–622 (2009).

85. Katzke, N., Knapp, A., Loeschcke, A., Drepper, T. & Jaeger, K. Novel Tools for the Functional Expression of Metagenomic DNA. *Methods Mol. Biol.* **1539**, 159–196 (2017).

86. Selas Castiñeiras, T., Williams, S., Hitchcock, A. & Smith, D. E. coli strain engineering for the production of advanced biopharmaceutical products. *FEMS Microbiol. Lett.* **365**, (2018).

87. Gonzalez, D. *et al.* Ancestral mutations as a tool for solubilizing proteins: The case of a hydrophobic phosphate-binding protein. *FEBS Open Bio* **4**, 121 (2014).

88. Whitfield, J. H. *et al.* Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci.* **24**, 1412–1422 (2015).

89. Trudeau, D., Kaltenbach, M. & Tawfik, D. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol. Biol. Evol.* **33**, 2633–2641 (2016).

90. Manteca, A. *et al.* Mechanochemical evolution of the giant muscle protein titin as inferred from resurrected proteins. *Nat. Struct. Mol. Biol.* **24**, 652–657 (2017).

91. Koblan, L. *et al.* Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–848 (2018).

92. Hendrikse, N. M., Charpentier, G., Nordling, E. & Syrén, P. O. Ancestral diterpene cyclases show increased thermostability and substrate acceptance. *FEBS J.* **285**, 4660–4673 (2018).

93. Gomez-Fernandez, B., Garcia-Ruiz, E., Martin-Diaz, Gomez e Santos, P., Santos-Moriano, P., Plou, J. M., Ballesteros, A., Garcia, M., Rodriguez, M., Risso, V. A., Sanchez-Ruiz, J. M., Whitney, S. M., Alcalde, M. Directed -in vitro- evolution of Precambrian and extant Rubiscos. *Sci. Rep.* **8**, 5532 (2018).

94. Barruetabeña, N. *et al.* Resurrection of efficient Precambrian endoglucanases for lignocellulosic biomass hydrolysis. *Commun. Chem.* **2**, (2019).

95. Nakano, S., Minamino, Y., Hasebe, F. & Ito, S. Deracemization and Stereoinversion to Aromatic d-Amino Acid Derivatives with Ancestral l-Amino Acid Oxidase. *ACS Catal.* **9**, 10152–10158 (2019).

96. Gomez-Fernandez, B. J., Risso, V. A., Rueda, A., Sanchez-Ruiz, J. M. & Alcalde, M. Ancestral resurrection and directed evolution of fungal mesozoic laccases. *Appl. Environ. Microbiol.* **86**, e00778-20 (2020).

97. Li, D. *et al.* Consensus Mutagenesis and Ancestral Reconstruction Provide Insight into the Substrate Specificity and Evolution of the Front-End Δ6-Desaturase Family. *Biochemistry* **59**, 1398–1409 (2020).

98. Sun, Y., Calderini, E. & Kourist, R. A Reconstructed Common Ancestor of the Fatty Acid Photo-decarboxylase Clade Shows Photo-decarboxylation Activity and Increased Thermostability. *Chembiochem* **22**, 1833–1840 (2021).

99. Hendry, T. A., De Wet, J. R. & Dunlap, P. V. Genomic signatures of obligate host dependence in the luminous bacterial symbiont of a vertebrate. *Env. Microbiol* **16**, 2611–22 (2014).

100. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nat. 2013 5037476* **503**, 418–421 (2013).

101. Seelig, B. & Szostak, J. W. Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nat. 2007 4487155* **448**, 828–831 (2007).

102. Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164 (1999).

103. Sillitoe, I. *et al.* CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **43**, D376–D381 (2015).

104. Herscovics, A. Glycosidases of the Asparagine-linked Oligosaccharide Processing Pathway. *Glycobiology* **4**, 113–125 (1994).

105. Raich, L., Nin-Hill, A., Ardèvol, A. & Rovira, C. Enzymatic Cleavage of Glycosidic Bonds: Strategies on How to Set Up and Control a QM/MM Metadynamics Simulation. *Methods Enzymol.* **577**, 159–183 (2016).

106. Wolfenden, R. & Snider, M. J. The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* **34**, 938–945 (2001).

107. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009).

108. Grunwald, P. *Biocatalysis: Biochemical Fundamentals And Applications (Second Edition)*. (2017).

109. Perez-Jimenez, R. *et al.* Diversity of chemical mechanisms in thioredoxin catalysis revealed by single-molecule force spectroscopy. *Nat. Struct. Mol. Biol. 2009 168* **16**, 890–896 (2009).

110. Georgescu, R. E., Garcia-Mira, M. M., Tasayco, M. L. & Sanchez-Ruiz, J. M. Heat capacity analysis of oxidized Escherichia coli thioredoxin fragments (1–73, 74–108) and their noncovalent complex. *Eur. J. Biochem.* **268**, 1477–1485 (2001).

111. Perez-Jimenez, R., Godoy-Ruiz, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. The effect of charge-introduction mutations on E. coli thioredoxin stability. *Biophys. Chem.* **115**, 105–107 (2005).

112. Perez-Jimenez, R., Godoy-Ruiz, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. The Efficiency of Different Salts to Screen Charge Interactions in Proteins: A Hofmeister Effect? *Biophys. J.* **86**, 2414–2429 (2004).

113. Godoy-Ruiz, R., Perez-Jimenez, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. Relation Between Protein Stability, Evolution and Structure, as Probed by Carboxylic Acid Mutations. *J. Mol. Biol.* **336**, 313–318 (2004).

114. Pey, A. L., Rodriguez-Larrea, D., Gavira, J. A., Garcia-Moreno, B. & Sanchez-Ruiz, J. M. Modulation of Buried Ionizable Groups in Proteins with Engineered Surface Charge. doi:10.1021/ja909298v

115. Rodriguez-Larrea, D. *et al.* Role of conservative mutations in protein multi-property adaptation. *Biochem. J.* **429**, 243–249 (2010).

116. Romero-Romero, M. L., Inglés-Prieto, A., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. Highly Anomalous Energetics of Protein Cold Denaturation Linked to Folding-Unfolding Kinetics. *PLoS One* **6**, e23050 (2011).

117. Garcia-Seisdedos, H., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. How many ionizable groups can sit on a protein hydrophobic core? *Proteins Struct. Funct. Bioinforma.* **80**, 1–7 (2012).

118. Godoy-Ruiz, R., Perez-Jimenez, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. A Stability Pattern of Protein Hydrophobic Mutations that Reflects Evolutionary Structural Optimization. *Biophys. J.* **89**, 3320–3331 (2005).

119. Godoy-Ruiz, R. *et al.* Natural Selection for Kinetic Stability Is a Likely Origin of Correlations between Mutational Effects on Protein Energetics and Frequencies of Amino Acid Occurrences in Sequence Alignments. *J. Mol. Biol.* **362**, 966–978 (2006).

120. Pey, A. L. *et al.* Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins Struct. Funct. Bioinforma.* **71**, 165–174 (2008).

121. Sanchez-Ruiz, J. M. Protein kinetic stability. *Biophys. Chem.* **148**, 1–15 (2010).

122. Tsytlonok, M. & Itzhaki, L. S. The how's and why's of protein folding intermediates.

Arch. Biochem. Biophys. **531**, 14–23 (2013).

123. Matouschek, A., Kellis, J. T., Serrano, L., Rycroft, M. & Ersht, A. R. F. Transient folding intermediates characterized by protein engineering Kinetics detects folding intermediate. (1990).

124. Jackson, S. E. & Fersht, A. R. Folding of Chymotrypsin Inhibitor 2. 1. Evidence for a Two-State Transition. *Biochemistry* **30**, 10428–10435 (1991).

125. Brandts, J. F., Halvorson, H. R. & Brennan, M. Consideration of the possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues. *Biochemistry* **14**, 4953–4963 (2002).

126. Schmid, F. X. & Baldwin, R. L. Acid catalysis of the formation of the slow-folding species of RNase A: Evidence that the reaction is proline isomerization. *Proc. Natl. Acad. Sci.* **75**, 4764–4768 (1978).

127. Schmidpeter, P. A. M. & Schmid, F. X. Prolyl Isomerization and Its Catalysis in Protein Folding and Protein Function. *J. Mol. Biol.* **427**, 1609–1631 (2015).

128. Kelley, R. F. & Richards, F. M. Replacement of proline-76 with alanine eliminates the slowest kinetic phase in thioredoxin folding. *Biochemistry* **26**, 6765–6774 (2002).

129. Holmgren, A. Thioredoxin catalyzes the reduction of insulin disulfides by dithiothreitol and dihydrolipoamide. *J Biol Chem* **254**, 9627–32 (1979).

130. Gamiz-Arco, G. *et al.* Non-conservation of folding rates in the thioredoxin family reveals degradation of ancestral unassisted-folding. *Biochem. J.* **476**, 3631–3647 (2019).

131. Scott, K., Alonso, D., Sato, S., Fersht, A. & Daggett, V. Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2661–2666 (2007).

132. Tzul, F. O., Vasilchuk, D. & Makhatadze, G. I. Evidence for the principle of minimal frustration in the evolution of protein folding landscapes. *Proc. Natl. Acad. Sci.* **114**, E1627–E1632 (2017).

133. Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M. & Hartl, F. U. Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annu Rev Biochem* **82**, 323–355 (2013).

134. Oh, E. *et al.* Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* **147**, 1295–1308 (2011).

135. Zhang, G. & Ignatova, Z. Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *Curr. Opin. Struct. Biol.* **21**, 25–31 (2011).

136. Kaur, J., Kumar, A. & Kaur, J. Strategies for optimization of heterologous protein expression in E. coli: Roadblocks and reinforcements. *Int. J. Biol. Macromol.* **106**, 803–822 (2018).

137. Shariati, F. S., Keramati, M., Valizadeh, V., Cohan, R. A. & Norouzian, D. Comparison of E. coli based self-inducible expression systems containing different human heat shock proteins. *Sci. Rep.* **11**, 1–10 (2021).

138. Baneyx, F. & Mujacic, M. Recombinant protein folding and misfolding in Escherichia coli. *Nat. Biotechnol. 2004 2211* **22**, 1399–1408 (2004).

139. Gamiz-Arco, G. *et al.* Combining Ancestral Reconstruction with Folding-Landscape Simulations to Engineer Heterologous Protein Expression. *J. Mol. Biol.* **433**, 167321

(2021).

140. Nishihara, K., Kanemori, M., Kitagawa, M., Yanagi, H. & Yura, T. Chaperone Coexpression Plasmids: Differential and Synergistic Roles of DnaK-DnaJ-GrpE and GroEL-GroES in Assisting Folding of an Allergen of Japanese Cedar Pollen, Cryj2, in Escherichia coli. *Appl. Environ. Microbiol.* **64**, 1694 (1998).

141. Folwarczna, J., Moravec, T., Plchova, H., Hoffmeisterova, H. & Cerovska, N. Efficient expression of Human papillomavirus 16 E7 oncoprotein fused to C-terminus of Tobacco mosaic virus (TMV) coat protein using molecular chaperones in Escherichia coli. *Protein Expr. Purif.* **85**, 152–157 (2012).

142. Marco, A. de. Protocol for preparing proteins with improved solubility by co-expressing with molecular chaperones in Escherichia coli. *Nat. Protoc.* **2**, 2632–2639 (2007).

143. Jhamb, K. & Sahoo, D. Production of soluble recombinant proteins in Escherichia coli: effects of process conditions and chaperone co-expression on cell growth and production of xylanase. *Bioresour. Technol.* **123**, 135–143 (2012).

144. Voulgaridou, G.-P., Mantso, T., Chlichlia, K., Panayiotidis, M. I. & Pappa, A. Efficient E. coli Expression Strategies for Production of Soluble Human Crystallin ALDH3A1. *PLoS One* **8**, e56582 (2013).

145. Yan, X., Hu, S., Guan, Y. & Yao, S. Coexpression of chaperonin GroEL/GroES markedly enhanced soluble and functional expression of recombinant human interferon-gamma in Escherichia coli. *Appl. Microbiol. Biotechnol.* **93**, 1065–1074 (2012).

146. Nausch, H. *et al.* Recombinant Production of Human Interleukin 6 in Escherichia coli. *PLoS One* **8**, e54933 (2013).

147. Levy, R., Weiss, R., Chen, G., Iverson, B. & Georgiou, G. Production of correctly folded Fab antibody fragment in the cytoplasm of Escherichia coli trxB gor mutants via the coexpression of molecular chaperones. *Protein Expr. Purif.* **23**, 338–347 (2001).

148. Ronez, F., Desroche, N., Arbault, P. & Guzzo, J. Co-expression of the small heat shock protein, Lo18, with β-glucosidase in Escherichia coli improves solubilization and reveals various associations with overproduced heterologous protein, GroEL/ES. *Biotechnol. Lett.* **34**, 935–939 (2012).

149. Muñoz, V. & Eaton, W. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11311–11316 (1999).

150. Muñoz, V., Thompson, P., Hofrichter, J. & Eaton, W. Folding dynamics and mechanism of beta-hairpin formation. *Nature* **390**, 196–199 (1997).

151. Naganathan, A. N. Predictions from an Ising-like Statistical Mechanical Model on the Dynamic and Thermodynamic Effects of Protein Surface Electrostatics. *J. Chem. Theory Comput.* **8**, 4646–4656 (2012).

152. Wako, H. & Saitô, N. Statistical Mechanical Theory of the Protein Conformation. II. Folding Pathway for Protein. *J. Phys. Soc. japan* **44**, 1939–1945 (1978).

153. Marcelino, A. M. C. & Gierasch, L. M. Roles of β-turns in protein folding: From peptide models to protein engineering. *Biopolymers* **89**, 380–391 (2008).

154. Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52–56 (1990).

155. Dinner, A. R. & Karplus, M. The roles of stability and contact order in determining protein folding rates. *Nat. Struct. Biol. 2001 81* **8**, 21–22 (2001).

156. Sato, S., Xiang, S. & Raleigh, D. P. On the Relationship Between Protein Stability and Folding Kinetics: A Comparative Study of the N-terminal Domains of RNase HI, E. coli and Bacillus stearothermophilus L9. *J. Mol. Biol.* **312**, 569–577 (2001).

157. Broom, A. *et al.* Modular Evolution and the Origins of Symmetry: Reconstruction of a Three-Fold Symmetric Globular Protein. *Structure* **20**, 161–171 (2012).

158. Broom, A. *et al.* Designed protein reveals structural determinants of extreme kinetic stability. *Proc. Natl. Acad. Sci.* **112**, 14605–14610 (2015).

159. Goldenzweig, A. *et al.* Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 337–346 (2016).

160. Georgoulia, P. S., Bjelic, S. & Friedman, R. Deciphering the molecular mechanism of FLT3 resistance mutations. *FEBS J.* **287**, 3200–3220 (2020).

161. Brazzolotto, X., Igert, A., Guillon, V., Santoni, G. & Nachon, F. Bacterial Expression of Human Butyrylcholinesterase as a Tool for Nerve Agent Bioscavengers Development. *Mol. 2017, Vol. 22, Page 1828* **22**, 1828 (2017).

162. Campeotto, I. *et al.* One-step design of a stable variant of the malaria invasion protein RH5 for use as a vaccine immunogen. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 998–1002 (2017).

163. Goldsmith, M. *et al.* Overcoming an optimization plateau in the directed evolution of highly efficient nerve agent bioscavengers. *Protein Eng. Des. Sel.* **30**, 333–345 (2017).

164. Lambert, A. R., Hallinan, J. P., Werther, R., Głów, D. & Stoddard, B. L. Optimization of Protein Thermostability and Exploitation of Recognition Behavior to Engineer Altered Protein-DNA Recognition. *Structure* **28**, 760-775.e8 (2020).

165. Malladi, S. K. *et al.* One-step sequence and structure-guided optimization of HIV-1 envelope gp140. *Curr. Res. Struct. Biol.* **2**, 45–55 (2020).

166. Trudeau, D. L. *et al.* Design and in vitro realization of carbon-conserving photorespiration. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11455–E11464 (2018).

167. Tullman, J., Christensen, M., Kelman, Z. & Marino, J. P. A ClpS-based N-terminal amino acid binding reagent with improved thermostability and selectivity. *Biochem. Eng. J.* **154**, 107438 (2020).

168. Warszawski, S. *et al.* Design of a basigin-mimicking inhibitor targeting the malaria invasion protein RH5. *Proteins* **88**, 187–195 (2020).

169. Cole, M. & Gaucher, E. Exploiting models of molecular evolution to efficiently direct protein engineering. *J. Mol. Evol.* **72**, 193–203 (2011).

170. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).

171. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the (βα) 8 - Barrel Enzyme Fold . *Chem. Rev.* **105**, 4038–4055 (2005).

172. Das, S. *et al.* Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics* **31**, 3460–3467 (2015).

173. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilles and domain superfolds. *Nat. 1994 3726507* **372**, 631–634 (1994).

174. Zechel, D. L. & Withers, S. G. Glycosidase Mechanisms: Anatomy of a Finely Tuned Catalyst. (2000). doi:10.1021/AR970172

175. Davies, G. & Henrissat, B. Structures and mechanisms of glycosyl hydrolases. *Structure* **3**, 853–859 (1995).

176. KOSHLAND, D. E. STEREOCHEMISTRY AND THE MECHANISM OF ENZYMATIC REACTIONS. *Biol. Rev.* **28**, 416–436 (1953).

177. Withers, S. Glycoside Hydrolase Family 1. CAZypedia. Available at: https://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_1. (Accessed: 19th October 2021)

178. Gamiz-Arco, G. *et al.* Heme-binding enables allosteric modulation in an ancient TIM-barrel glycosidase. *Nat. Commun.* **12**, 380 (2021).

179. VANDERKOOI, G. & STOTZ, E. Reductive Alteration of Heme a Hemochromes. *J. Biol. Chem.* **240**, 3418–3424 (1965).

180. Randall, R., Radford, C., Roof, K., Natarajan, D. & Gaucher, E. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat. Commun.* **7**, 12847 (2016).

181. Straus, D. & Gilbert, W. Genetic engineering in the Precambrian: structure of the chicken triosephosphate isomerase gene. *Mol. Cell. Biol.* **5**, 3497–3506 (1985).

182. Dietmar Lang, Ralf Thoma, Martina Henn-Sax, Reinhard Sterner, M. W. Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science* **289**, 1546–1550 (2000).

183. M, H.-S., B, H., M, W. & R, S. Divergent evolution of (betaalpha)8-barrel enzymes. *Biol. Chem.* **382**, 1315–1320 (2001).

184. Richter, M. *et al.* Computational and experimental evidence for the evolution of a (beta alpha)8-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. *J. Mol. Biol.* **398**, 763–773 (2010).

185. Chu, X.-Y. & Zhang, H.-Y. Cofactors as Molecular Fossils To Trace the Origin and Evolution of Proteins. *ChemBioChem* **21**, 3161–3168 (2020).

186. Chen, K. & Arnold, F. H. Engineering new catalytic activities in enzymes. *Nat. Catal. 2020 33* **3**, 203–213 (2020).