Evolutionary modes in protein observable space: the case of thioredoxins

Running Title: Protein evolutionary modes

Sara Del Galdo, *Dept. of Chemical Science and Technology, University of Roma Tor Vergata, via della Ricerca Scientifica, 00133 Rome, Italy*

Josephine Alba, *Dept. of Chemistry, Sapienza University of Rome, P.le A. Moro, 5, 00185, Rome, Italy*

Andrea Amadei*, *Dept. of Chemical Science and Technology, University of Roma Tor Vergata, via della Ricerca Scientifica, 00133 Rome, Italy*

Marco D'Abramo*, *Dept. of Chemistry, Sapienza University of Rome, P.le A. Moro, 5, 00185, Rome, Italy*

Contact Information: Marco D'Abramo, email: marco.dabramo@uniroma1.it

Andrea Amadei, email: andrea.amadei@uniroma2.it

Abstract

In this article we investigated the structural and dynamical evolutionary behaviour of a set of 10 thioredoxin proteins as formed by three extant forms and seven resurrected ones in laboratory. Starting from the crystallographic structures, we performed all-atoms molecular dynamics simulations and compare the trajectories in terms of structural and dynamical properties. Interestingly, the structural properties related to the protein density (i.e. the number of residues divided by the excluded molecular volume) well describe the protein evolutionary behaviour. Our results also suggest that the changes in sequence as occurred during the evolution have affected the protein essential motions, allowing us to discriminate between ancient and extant proteins in terms of their dynamical behaviour. Such results are yet more evident when the bacterial, archeal and eukaryotic thioredoxins are separately analysed.

Keywords

Protein evolution, protein dynamics, thioredoxin, essential motions, molecular dynamics

Introduction

The knowledge of the causes which have driven the evolution of proteins – in terms of sequence and structure from their first appearance on the earth – represents a very intriguing challenge. In fact, the possibility to describe the factors which have ultimately determined the protein molecular evolution could pave the way to rationalize the adaption of different forms of life to external changes as well as to a better understanding of the proteins as molecular machines. However, dealing with protein evolution requires information on extinct molecules which have disappeared along the history. Starting from the seminal observation that "most or all apparently heterologous gene derive ultimately from a common gene ancestor" (Pauling et al. 1963), different strategies have been proposed to reconstruct, with a certain degree of accuracy, putative sequences of proteins that no longer exist. These Ancestral Sequence Reconstruction (ASR) methods (Fitch 1971; Chang and Donoghue 2000; Hall 2006; Benner et al. 2007; Liberles 2008; Arenas et al. 2017) combine multiple alignments of extant protein sequences, phylogenetic analysis and the probability of amino acid substitution to infer a putative ancient protein sequence.

Differently from "horizontal" approaches, where the process determining the protein structure and function evolution is inferred from sequence comparison of extant proteins, thus losing the evolutionary aspects, phylogenetic approaches give the possibility to estimate protein sequences along the history. That is, a phylogenetic tree obtained from the analysis of extant protein family members belonging to the three different domains of life might be combined with a multiple sequence alignment and a substitution model of evolution to provide a statistical inference on the ancestral sequence at any internal node of the tree (Arenas 2015). By such an approach, many different ancestral proteins were resurrected (Chang et al. 2002; Benner et

al. 2007; Ortlund et al. 2007; Harms and Thornton 2010; Perez-Jimenez et al. 2011; Hart et al. 2014), and some of them crystallized (Ortlund et al. 2007; Ingles-Prieto et al. 2013). From an evolutionary point of view, such data constitute an invaluable source of knowledge which can be used to shed light on the key factors regulating the evolution of the protein along the history. It is worth to mention here that ASR methods are affected by errors because of the limits of the substitution model used. Therefore, new strategies are continuously proposed to improve the ASR accuracy and those considering structural constraints represent one of the more promising approaches (Arenas et al. 2017). However, the evaluation of the ASR accuracy in the case of the resurrected proteins studied in this work is beyond the scope of this paper.

To the best of our knowledge, two recent papers on thioredoxins represent the most complete study on resurrected proteins, where a set of seven Precambrian thioredoxin enzymes were resurrected, crystallized and tested to measure their enzymatic activity (Perez-Jimenez et al. 2011; Ingles-Prieto et al. 2013). These enzymes – ubiquitous in all living organisms (Holmgren et al. 1975) – were probably present in primitive life forms, as suggested by the archetypical active site (CXXC) and the conserved fold. It is worth noting that this structural dataset is actually the largest present in literature (Perez-Jimenez et al. 2011; Ingles-Prieto et al. 2013). In those papers (Perez-Jimenez et al. 2011; Ingles-Prieto et al. 2013), the melting temperature (T_m) was suggested as a possible evolutionary observable, which left genetic footprints on ancestral organisms (Boussau et al. 2008).

Considering the well-established link between protein function, structure and dynamics and the availability of this set of thioredoxin protein structures mentioned above, several questions immediately arise: how do the changes in protein sequences due to the selective pressure influence the protein motions? To what extent? Do such differences depend on the organism life domain?

Inspired by such questions, we report here our investigation on the structural and dynamical behaviour (as provided by molecular dynamics simulations) of this set of thioredoxin proteins (Fig. 1) with the aim to provide a link between protein evolution and protein structure and dynamics.

Materials and Methods

The structures of the 10 thioredoxin (Trx) proteins were taken from the Protein Data Bank: 1ERU, extant eukaryota (Human); 2TRX, extant bacteria (Ecoli); 2E0Q, extant archea (Archea); 4BA7, last bacterial common ancestor (LBCA); 2YJ7, last common ancestor of the cyanobacterial, deinococcus and thermus groups (LPBCA); 2YN1, last common ancestor of γ-proteobacteria (LGPCA); 3ZIV, archaea/eukaryota common ancestor (AECA); 2YNX, last archaeal common ancestor (LACA); 2YOI, last eukaryotic common ancestor (LECA); 2YPM, last common ancestor of fungi and animals (LAFCA). The reconstruction of the ancestral Trx enzymes as well the associated divergence times (Fig. 1) have been taken from the literature (Ingles-Prieto et al. 2013). Briefly, in that paper the sequences of ancient thioredoxins have been estimated by phylogenetic tree encompassing more than 200 diverse Trxs sequences from the three domains of life. The associated phylogenetic tree has been used to estimate the divergence dates to nodes in the tree using multiple fossil calibrations (Yang et al. 2006; Rannala et al. 2007) on the hypothesis that root of the tree lies between bacteria and the common ancestor of archaea and eukaryotes.

In this work, we used the reconstruction of the ancestral Trx enzymes previously published (Perez-Jimenez et al. 2011; Ingles-Prieto et al. 2013) (i.e., the inferred sequences of the ancient thioredoxins, the corresponding X-ray structures and the estimated divergence dates) to describe the behavior of different structural and dynamical properties of this protein set along the history as explained in the following subsections.

Note that all the definitions of ancestors depend on the accuracy of the reconstructed tree and the ancestral state reconstruction algorithms used.

Molecular Dynamics simulations

The Molecular Dynamics (MD) simulations were performed using the Gromacs software package (Hess et al. 2008) and the amber99sb-ildn force field. The SPC model (Berendsen et at. 1993) was used to mimic the water and sodium ions were added to neutralize the total thioredoxin charge. All the ten thioredoxins were simulated with periodic boundary conditions in the isothermal-isochoric ensemble (NVT), using an integration step of 2 fs and keeping the temperature constant (298 K) by the velocity rescaling thermostat (Bussi et al. 2007). The bonds were constrained using the LINCS algorithm (Hess et al. 1997) and for short range interactions a cut-off radius of 1.1 nm was employed. To compute long range interactions the particle mesh Ewald method (Darden et al. 1997) was used with grid search and cut-off radii of 1.1 nm. We calibrated the density of the boxes containing the water-protein solutions in order to obtain an identical pressure, within the noise (~ 10 bar), to the one provided by an MD simulation of a pure SPC box with a density corresponding to the liquid water experimental density at 298 K (we used as reference density 33.32 molecules per nm^3) according to the procedure described in our recent work (Del Galdo et al. 2015). We performed for all the systems a productive MD simulation lasting 100 ns. Essential dynamics analysis (Amadei et al. 1993) was applied to each single trajectory and to a combination of them in order to highlight the phylogenetic-based differences in protein essential motions. The overlap (s) between the covariance matrices (A, B) is defined by (Hess 2002):

$$s(A, B) = 1 - d(A, B)/(tr\,A + tr\,B)^{1/2}$$

where *tr* is the trace of the covariance matrix and d(A,B) is the difference between the covariance matrices A and B as defined by Hess (2002).

Structural analysis

The protein mean excluded volume, i.e. the mean volume enclosed by the solvent-accessible surface, has been estimated by averaging (over the productive MD simulation) the protein excluded volume along MD and using a probe radius of 0.14 nm according to the method reported in literature (Eisenhaber et al. 1995)□. The protein partial molecular volume was computed by the method reported in our previous work (Del Galdo et al. 2015)□, which is based on the evaluation of the mean protein excluded volume, the mean volume of the protein hydration shell and the hydration shell SPC density increment with respect to the reference SPC density (bulk density).

The number of hydrogen bonds and the residues with secondary structures have been calculated by Gromacs tools. The solvent (polar, apolar and total) accessible and excluded surface areas were calculated by finding solvent-exposed vertices of intersecting atoms (Fraczkiewicz et al. 1998)□. The number of proline residues as well as the B-factors have been directly extracted by the corresponding pdb files.

Principal component analysis of the structural observables

For these structure-related properties a principal component analysis has been performed. The melting temperature was added to this set, because it has been suggested as a possible evolutionary observable, as indicated by its increase along the history (Perez-Jimenez et al. 2011)□. Due to the different magnitude and physical meaning of such observables, the covariance matrix was built using the adimensional rescaled shifts with respect to their averages for all the observables. That is, for the melting temperature and density-related properties

$$\Delta T'_m = (T_m - \langle T_m \rangle) / \sigma_{Tm}$$

$\Delta\rho' = (\rho - \langle\rho\rangle) / \sigma_\rho$

$\Delta v' = (v - \langle v\rangle) / \sigma_v$

with the angle brackets indicating the mean over time and σ the square root of the variance.

Results

To investigate the evolutionary behaviour of the 10 thioredoxin proteins, we analysed the corresponding molecular dynamics trajectories using both fluctuation-related properties (protein principal motions and entropy estimates) and structure-related observables (protein density, molecular volume, hydrogen bond contents, amount of secondary structures and solvent accessible surface area).

It is worth noting that although our set is composed by proteins belonging to the same class, these proteins span a quite large sequence identity interval (between 0.25 and 0.92, see Table 1).

Structural behaviour

To address the evolutionary behaviour of thioredoxin observables and possibly uncover their correlation, we considered 12 different properties: the experimental melting temperature, the solvent excluded surface area, the polar, apolar and total solvent accessible surface area, the number of proline residues, the B-factors, the number of hydrogen bonds, the fraction of residues having secondary structure, the residue density within the protein excluded volume (ρ) and the partial molecular volume (v).

We excluded the relative density increment of the solvent density within the protein hydration shell with respect to the solvent bulk density, because it is nearly constant all along the evolutionary time (Fig. S1 in S.I.). This set of 11 observables was analyzed by means of

principal component analysis, thus providing a set of eigenvectors and associated eigenvalues describing the evolutionary behaviour within such an 11-dimensional space.

The spectrum of eigenvalues (Fig. S2 in S.I.) obtained from the diagonalisation of the covariance matrix shows that the first eigenvector accounts for ~ 50 % of the total fluctuations, the second for ~ 30 % and the third for ~ 10 %. The first eigenvector has the (nearly identical) major component values for the T_m, the residue density (ρ) and the partial molecular volume (ν). Due to the limited sample size formed by 10 proteins, which is further reduced in the case of separate analysis for the archeal (3 structures), eukaryotic (4 structures), and bacterial (4 structures) subsets, we decided to restrict our analysis on these three observables. Such a choice is largely justified by the principal component analysis performed on the global observable space defined by the 12 observables previously described, which shows that the major component of the first eigenvector (explaining about the 50% of the total fluctuations) is dominated by the T_m, the residue density (ρ) and the partial molecular volume (ν).

Therefore, these three observables not only represent the best choice for the description of the evolutionary behaviour, but also include the melting temperature and the density related properties which are simple, physically-sound observables. We recently observed that the protein density well correlates with the protein optimal growth temperature (Amadei et al. 2017)□. We would like to stress here that there are not direct thermodynamic relations connecting the difference in the heat capacity and/or the melting temperature – experimentally found to be correlated with thioredoxin evolution – with the protein density and/or the partial molecular volume.

In figure 2, we show the evolutionary trend of these three observables for archeal, eukaryiotic and bacterial thioredoxins.

From this figure, it is evident that the T_m, the residue density (ρ) and the partial molecular volume (v) significantly change along the evolutionary time (the sole exception being the T_m of the archeal thioredoxins, which decreases of only 3 K along the history).

In figure 3 we report the evolutionary trajectory for the sample of proteins considered within such a re-scaled 3-D observable space. Very interestingly, all the points are not spread over the planes, but they are rather well aligned along the diagonal, indicating a significant correlation among these properties.

In order to identify a possible single *generalised* observable able to describe the evolutionary trend, we performed the diagonalization of the $\Delta T'_m$, $\Delta\rho'$, $\Delta v'$ covariance matrix, providing by means of its eigenvectors, the relevant modes and corresponding observables within such a space (Amadei et al. 1993).

By using such eigenvectors (v_1, v_2, v_3) as new basis set of the observable space, we can express the evolutionary trajectory in terms of the three corresponding *generalised* observables (g_1, g_2, g_3, each corresponding to a specific linear combination of the original properties and defined by the projection of the observable trajectory on the eigenvectors). For eukaryotes, archea and bacteria the largest covariance matrix eigenvalue (corresponding to the v_1 eigenvector) provides 90%, 99% and 94% of the total square fluctuations, respectively, clearly indicating that the essential information on the evolutionary trend of thioredoxins can be obtained by the g_1 evolutionary behaviour (i.e. filtering out the small fluctuations along v_2 and v_3). It is worth to note that the v_1 eigenvectors have nearly identical component absolute values (~0.58) with the $\Delta\rho'$ component sign opposite to the others, thus indicating that the corresponding *generalized* observable is given by a virtually homogenous mixing of the original properties ($\Delta T'_m$, $\Delta\rho'$, $\Delta v'$), which are characterized by anti-correlation between $\Delta\rho'$ vs $\Delta T'_m$,$\Delta v'$.

As shown in figure 3 (lower right panel) the evolutionary trajectories along v_1, show a remarkable variation around 2 Gyears ago, well matching the archaean-proterozoic era

transition believed to correspond to a significant decrease of the global earth temperature (Lowe et al. 2004).

Therefore, our results suggest that global temperature decrease not only induced the T_m decrease as reported in literature (Perez-Jimenez et al. 2011), but possibly provided a significant residue density increase (partial molecular volume decrease), at least for thioredoxins. Interestingly, the partial molecular volume decrease is essentially due to the protein volume decrease as the hydration shell water density is virtually constant for all the investigated thioredoxins (see Fig. S1 for the eukaryotic thioredoxins). Finally, it is worth to mention that our results pinpoint the evolutionary trend of these three observables as occurring in thioredoxins; in different proteins, as bacterial ribonuclease H1 (RNH) proteins where the T_m is poorly correlated with evolution (Hart et al. 2014), it has been found that the low heat capacity of unfolding, due to the presence of residual structure in the unfolded state, is the major determinant for the RNHs difference in thermostability. Unfortunately, in that paper (Hart et al. 2014) only one structure of a resurrected RNH was made available, preventing the possibility to estimate the behavior of the observables used in the Trxs data set along the history.

Dynamical behavior

Principal motions

The essential motions describing the overall thioredoxin dynamics were calculated by principal component analysis performed on a single trajectory obtained by concatenating the ten thioredoxin trajectories of the C-alpha atoms.

The first eigenvector describes the 60% of the total variance, indicating that a remarkable amount of the protein motion of the thioredoxins is concentrated along this direction. The components of the first eigenvector (Fig. S3) show that the first essential motion is mainly

described by the N-terminal region, the α-helix 1, 2 and 3 and the β-sheet 1, 4 and 5, (see Fig. 1 for the secondary structure assignment).

The analysis of the sampled structures on the essential subspace as characterized by their projection on the first two eigenvectors shows three main regions, each corresponding to a different branch (Fig. S4). That is, the region of the subspace explored is characteristic of the life-domain. It is worth noting that the first eigenvector discriminates between archeal/bacterial vs eukaryotic thioredoxins, whereas the second eigenvector is able to discriminate between archeal and bacterial proteins, too.

To shed light on this different behaviour among eukaryotic, archeal and bacterial branches, the essential motions as described by the single thioredoxin trajectories were compared by pairs. The degree of similarity between the essential motions as well as the sequence identities are reported in the Table 1 for all the couples.

These values clearly show that the essential motions within the bacterial branch are well conserved, being the overlap between 0.65 (*E. coli* vs LGPCA) and 0.48 (*E. coli* and LPBCA). The comparison between the overlap of the essential motions and the sequence identities also points out that consecutive (along the history) thioredoxins share high sequence identities and high motion overlaps only within the same branch. In fact, LACA and LECA although sharing a high sequence identity (0.56) shows a relatively low motion overlaps (0.209), thus indicating that evolutionary steps have affected the protein essential motions.

Interestingly, also the AECA and LACA show a quite large overlap of the essential motions when compared to the bacterial thioredoxins, whereas the three remaining thioredoxins belonging to the eukaryotic branch (LECA, LAFCA and Human) show a very limited overlap

with respect to the other thioredoxins, i.e. the maximum overlap is 0.49 between LAFCA and Human which share the 85% of sequence identity.

All these results indicate that the archeal thioredoxins are rather similar in terms of sequence identity (Fig. S5) and protein motions to their ancestor as well as to the bacterial thioredoxins. The eukaryotic branch thioredoxins, on the other hand, are more distant in both sequence identity and essential motions with respect to the other thioredoxins.

Entropy

To obtain further insights in the thioredoxin protein behaviour, the estimation of their configurational fluctuation entropies has been performed using the covariance matrix of atomic coordinates as indicated by Schlitter (1993)□. In all the branches, the entropy shows a remarkable decrease along the history (Fig. 4), with a pronounced variation around 2 Gyears ago corresponding to the archaean-proterozoic era transition, believed to correspond to a significant decrease of the global earth temperature (Lowe et al. 2004)□. Interestingly, such a steep variation is also observed – in the same time interval – for the T_m, the partial molecular volume, the residue density and the associated *generalised* observable derived in the previous section.

From these results, it is tempting to suggest the hypothesis that the evolutionary behaviour of thioredoxins could be essentially entropically driven, with unfolded states of ancient and extant proteins similar in free energy and the ancient folded states (characterized by high melting temperatures) entropically stabilized by the residue density decrease.

After the Archean period, proteins were allowed to enhance local interactions – for example to optimize their biological functions – with a consequent entropy loss and melting temperature decrease, both permitted by the lower earth temperature.

Conclusions

In this work we calculated several global protein observables to characterize the structural/molecular evolution of thioredoxins.

In addition to the melting temperature, which was already found to change along the evolution, we calculated several other properties routinely used in large-scale classification of proteins as well as a new set of properties related to the protein density.

Our data point out that the melting temperature, the protein density and the partial molecular volume are the major components of the main evolutionary motion in thioredoxins. Therefore, by means of principal component analysis, we studied the evolutionary behavior within such a 3-dimensional space, which highlights an interesting trend along the history. In fact, the supposed global earth temperature decrease, dated in the archaean-proterozoic era transition, well matches the remarkable variation of the evolutionary trajectory along the main essential evolutionary mode.

The analysis of the molecular dynamics trajectories points out that the dynamical behavior of the thioredoxins is – as expected – driven by the structural changes induced by the protein sequence variations (upon evolution). However, the evolutionary step corresponding to the archaean-proterozoic era transition determines remarkable changes of the protein entropy.

Our study shows that structural-molecular evolution of thioredoxin proteins can be well described by a set of generalized protein observables, and that, among several properties, those related to the protein density are some of the most representative. The possibility to extend our

study to different protein families – using experimental and/or homology modeling approaches – could provide additional information on the relationship between protein structure and protein evolution, thus representing a very interesting research theme to investigate in the next future.

Acknowledgments

This work was partially funded by Sapienza, University of Rome (Progetto di Ateneo 2017).

The authors gratefully acknowledge NVIDIA and CINECA for computational support.

References

Amadei A, Del Galdo S, D'Abramo M (2017) Density discriminates between thermophilic and mesophilic proteins. J Biomol Struct Dyn 1–9. doi:10.1080/07391102.2017.1385537

Amadei A, Linssen AB, Berendsen HJ (1993) Essential dynamics of proteins. Proteins 17:412–25. doi: 10.1002/prot.340170408

Arenas M (2015) Trends in substitution models of molecular evolution. Front Genet. doi: 10.3389/fgene.2015.00319

Arenas M, Weber CC, Liberles DA, Bastolla U (2017) ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability. Syst Biol. doi: 10.1093/sysbio/syw121

Benner S, Sassi S, Gaucher E (2007) Molecular paleoscience: systems biology from the past. In: Toone E by EJ (ed) Adv Enzymol Relat Areas of Molecular Biology, Volume 75: Protein Evolution. pp 1–132

Berendsen HJC, Grigera JR, Straatsma TP (1993) The missing term in effective pair potentials. J Chem Phys 98:10089–10092

Boussau B, Blanquart S, Necsulea A, et al (2008) Parallel adaptations to high temperatures in the Archaean eon. Nature. doi: 10.1038/nature07393

Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126:014101. doi: 10.1063/1.2408420

Chang BS, Jonsson K, Kazmi MA, et al (2002) Recreating a functional ancestral archosaur visual pigment. Mol Biol Evol 19:1483–1489. doi: 10.1093/oxfordjournals.molbev.a004211

Chang BSW, Donoghue MJ (2000) Recreating ancestral proteins. Trends Ecol. Evol. 15:109–114

Darden T, Tork D, Pedersen L (1997) Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. J Comput Chem 18:1463–1472

Del Galdo S, Marracino P, D'Abramo M, Amadei A (2015) In silico characterization of protein partial molecular volumes and hydration shells. Phys Chem Chem Phys 17:31270–31277. doi: 10.1039/C5CP05891K

Eisenhaber F, Lijnzaad P, Argos P, et al (1995) The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J Comput Chem 16:273–284. doi: 10.1002/jcc.540160303

Fitch WM (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. Syst Biol 20:406–416. doi: 10.1093/sysbio/20.4.406

Fraczkiewicz R, Braun W (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J Comput Chem 19:319–333. doi: 10.1002/(SICI)1096-987X(199802)19:3<319::AID-JCC6>3.0.CO;2-W

Hall BG (2006) Simple and accurate estimation of ancestral protein sequences. Proc Natl Acad Sci U S A 103:5431–6. doi: 10.1073/pnas.0508991103

Harms MJ, Thornton JW (2010) Analyzing protein structure and function using ancestral gene reconstruction. Curr Opin Struct Biol 20:360–6. doi: 10.1016/j.sbi.2010.03.005

Hart KM, Harms MJ, Schmidt BH, et al (2014) Thermodynamic system drift in protein evolution. PLoS Biol 12:e1001994. doi: 10.1371/journal.pbio.1001994

Hess B (2002) Convergence of sampling in protein simulations. Phys Rev E 65:031910. doi: 10.1103/PhysRevE.65.031910

Hess B, Bekker H, Berendsen HJC, Fraaije J (1997) LINCS: A Linear Constant Solver for Molecular Simulations. J Comput Chem 18:1463–1472

Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput 4:435–447. doi: 10.1021/ct700301q

Holmgren A, Soderberg BO, Eklund H, Branden CI (1975) Three-dimensional structure of Escherichia coli thioredoxin-S2 to 2.8 A resolution. Proc Natl Acad Sci 72:2305–2309. doi: 10.1073/pnas.72.6.2305

Ingles-Prieto A, Ibarra-Molero B, Delgado-Delgado A, et al (2013) Conservation of protein structure over four billion years. Structure 21:1690–1697. doi: 10.1016/j.str.2013.06.020

Liberles DA (2008) Ancestral Sequence Reconstruction. Oxford University Press

Lowe DR, Tice MM (2004) Geologic evidence for Archean atmospheric and climatic evolution: Fluctuating levels of CO2, CH4, and O2 with an overriding tectonic control. Geology 32:493–496. doi: 10.1130/G20342.1

Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW (2007) Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. Science (80-) 317:1544–1548. doi: 10.1126/science.1142819

Pauling L, Zuckerkandl E (1963) Chemical paleogenetics: Molecular ‘restoration studies’ of extinct forms of life. Acta chem. scand 17:S9–S16

Perez-Jimenez R, Inglés-Prieto A, Zhao Z-M, et al (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat Struct Mol Biol 18:592–596. doi: 10.1038/nsmb.2020

Rannala B, Yang Z (2007) Inferring speciation tunes under an episodic molecular clock. Syst Biol 56:453–466. doi: 10.1080/10635150701420643

Schlitter J (1993) Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chem Phys Lett 215:617–621. doi: 10.1016/0009-2614(93)89366-P

Yang Z, Rannala B (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23:212–226. doi: 10.1093/molbev/msj024

Table 1. The Matrix of the sequence identity (elements below the diagonal) and overlap between essential motions as obtained from the overlap of the covariance matrices (elements above the diagonal)

Protein		LBCA	LPBCA	LGPCA	Ecoli	AECA	LACA	LECA	LAFCA	Human	Archea
	pdb id.	**4ba7**	**2yj7**	**2yn1**	**2trx**	**3ziv**	**2ynx**	**2yoi**	**2ypm**	**1eru**	**2e0q**
LBCA	**4ba7**	1	0.653	0.618	0.546	0.682	0.406	0.238	0.264	0.169	0.401
LPBCA	**2yj7**	0.87	1	0.544	0.476	0.614	0.463	0.216	0.224	0.157	0.390
LGPCA	**2yn1**	0.67	0.68	1	0.655	0.476	0.283	0.206	0.223	0.159	0.299
Ecoli	**2trx**	0.56	0.57	0.83	1	0.449	0.261	0.201	0.258	0.154	0.243
AECA	**3ziv**	0.85	0.76	0.60	0.53	1	0.469	0.272	0.277	0.181	0.426
LACA	**2ynx**	0.77	0.72	0.56	0.51	0.92	1	0.209	0.208	0.148	0.228
LECA	**2yoi**	0.57	0.53	0.44	0.37	0.59	0.56	1	0.490	0.390	0.201
LAFCA	**2ypm**	0.51	0.48	0.42	0.36	0.54	0.51	0.85	1	0.323	0.187
Human	**1eru**	0.36	0.34	0.30	0.25	0.38	0.35	0.55	0.58	1	0.154
Archea	**2e0q**	0.51	0.46	0.41	0.40	0.61	0.53	0.48	0.46	0.38	1

Figure Captions

Fig. 1 Sequence alignment for the ten thioredoxins studied in this work and the associated secondary structure (upper panel). Thioredoxin protein names and associated geological time (lower left table). Lower right panel: ribbon representation of the bacterial extant thioredoxin (PDB id. 2trx)

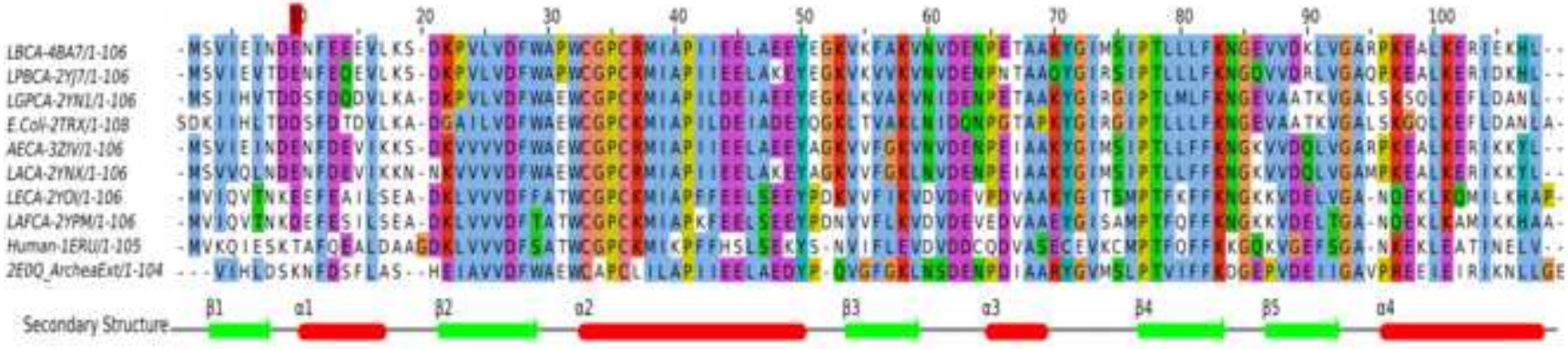
Fig. 2 Time evolution of the melting temperature (T_m upper left panel), the partial molecular volume (lower left panel) and the residue density within the protein excluded volume (ρ, upper right panel). Black, red and green circles refer to proteins associated to the eukaryotic, bacterial and archeal branch, respectively. The dashed lines serve as a guide for the eye. Note that the melting temperatures of the two oldest resurrected thioredoxins belonging to the archeal-eukaryotic and bacterial branches coincide

Fig. 3 Plot of the adimensional rescaled shifts ($\Delta T'_m$, $\Delta\rho'$, $\Delta\nu'$) for the ten thioredoxin proteins (upper plots and lower left plot) and the projection of the protein observables ($\Delta T'_m$, $\Delta\rho'$, $\Delta\nu'$) on the first eigenvector along the time (lower right plot; the dashed lines serve as a guide for the eye). Black, green and red circles refer to eukaryotic, archeal and bacteria thioredoxins, respectively

Fig. 4 Changes in the thioredoxin entropy with respect to the corresponding common ancestors (AECA and LBCA). Black, green and red points indicate eukaryotic, archeal and bacterial thioredoxins, respectively. The dashed lines serve as a guide for the eye

Figure 1

Click here to access/download;Figure;Fig1.tif



Name	Trx corresponding to	PDB id.	Time before present (billion years)
LBCA	last bacterial common ancestor	4ba7	-4.2
LPBCA	last common ancestor of the cyanobacterial, deinococcus and thermus groups	2yj7	-2.4
LGPCA	last common ancestor of g-proteobacteria	2yn1	-1.6
Ecoli	Ecoli	2trx	0
AECA	archaea/eukaryota common ancestor	3ziv	-4.2
LACA	last archaeal common ancestor	2ynx	-4.1
LECA	last eukaryotic common ancestor	2yoi	-1.6
LAFCA	last common ancestor of fungi and animals	2ypm	-1.4
Human	Human	1eru	0
Archea	Archea	2e0q	0

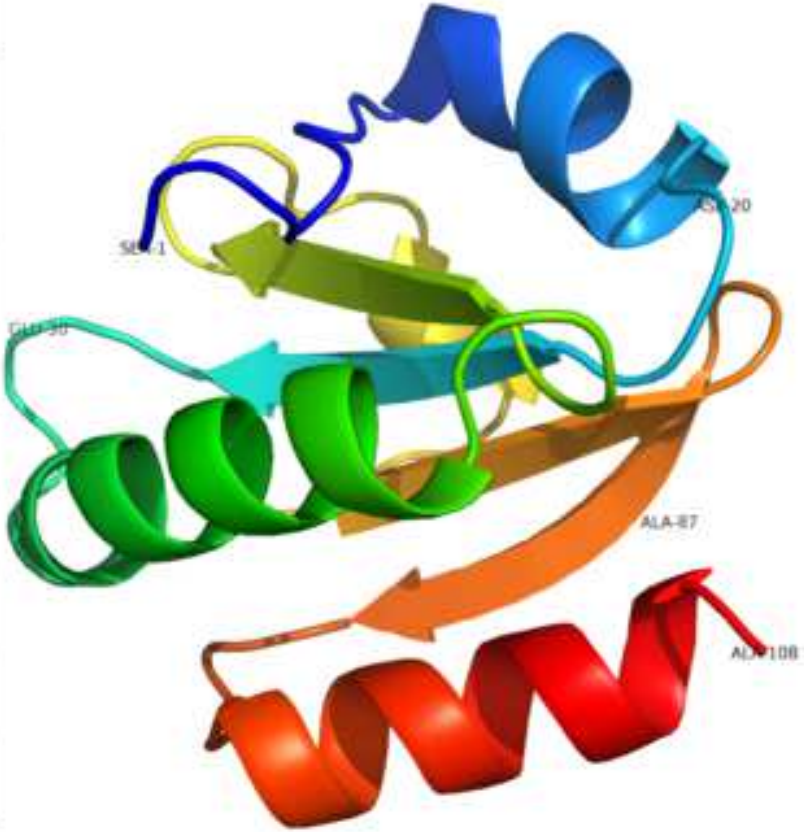
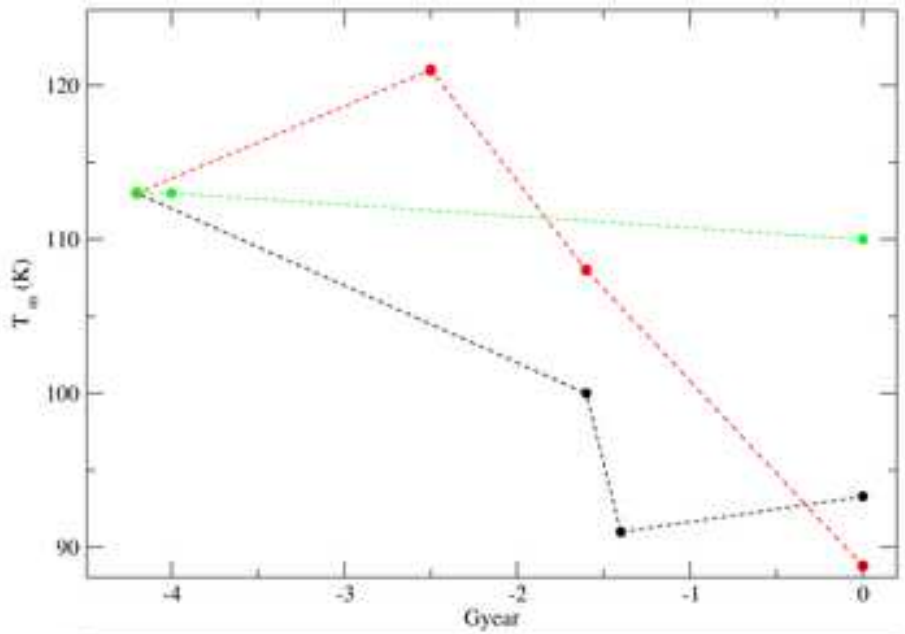
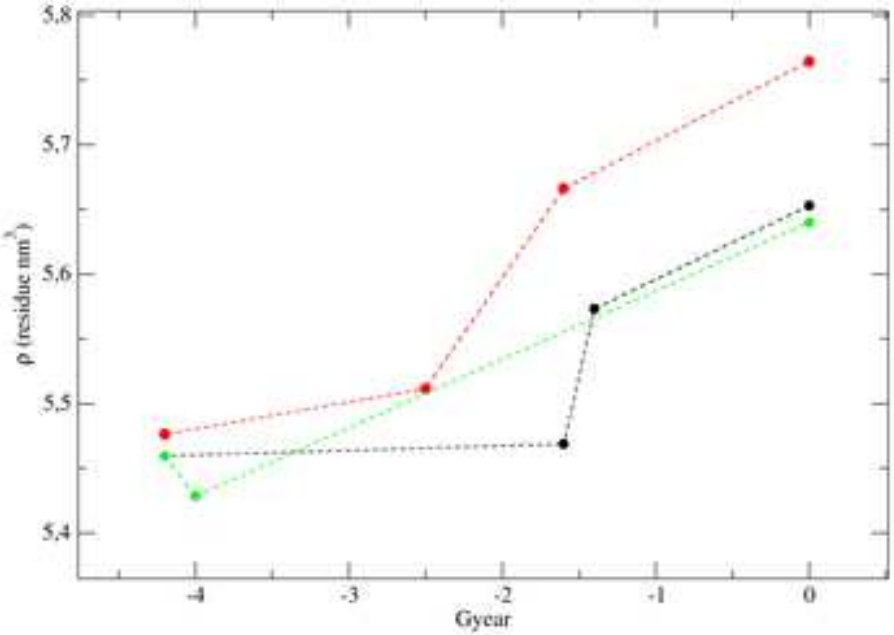


Figure 2

Click here to access/download;Figure;Fig2.tif





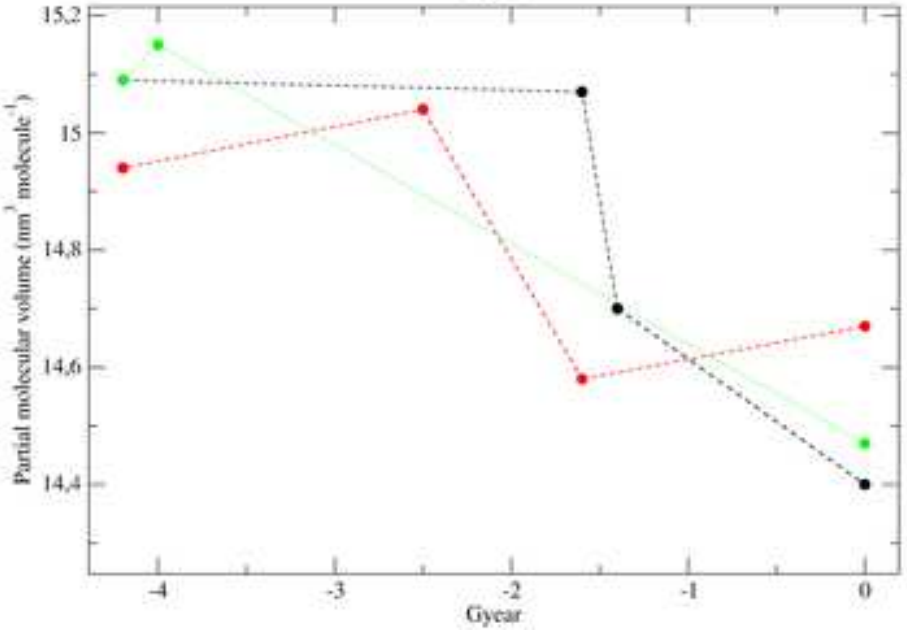


Figure 3

Click here to access/download;Figure;Fig3.tif

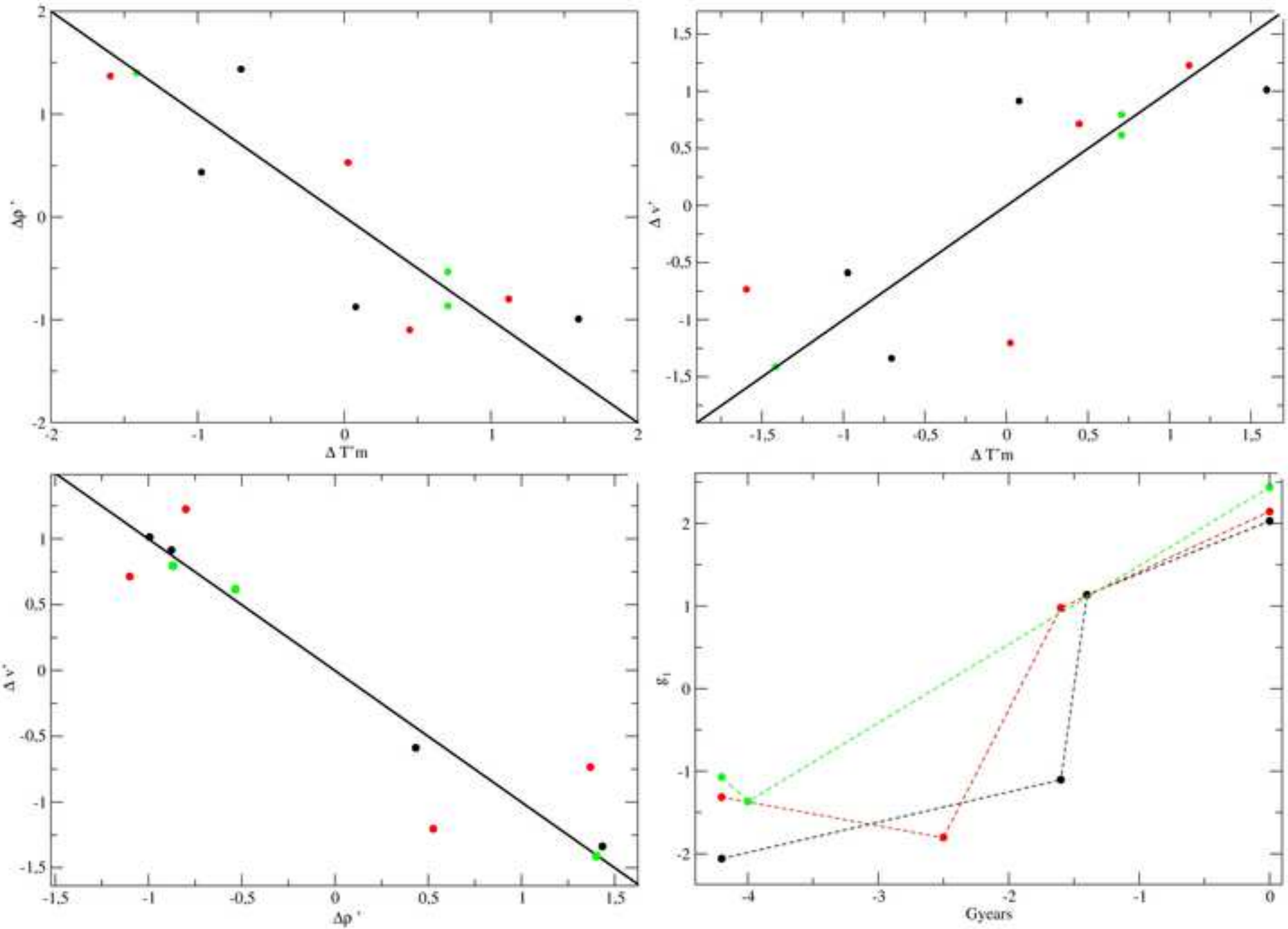


Figure 4

Click here to access/download;Figure;Fig4.tif