PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Research

Cite this article: Modi T, Huihui J, Ghosh K, Ozkan SB. 2018 Ancient thioredoxins evolved to modern-day stability–function requirement by altering native state ensemble. *Phil. Trans. R. Soc. B* **373**: 20170184.
http://dx.doi.org/10.1098/rstb.2017.0184

Accepted: 20 February 2018

One contribution of 17 to a discussion meeting issue 'Allostery and molecular machines'.

Subject Areas:
biophysics, evolution, computational biology, structural biology

Keywords:
evolution, thioredoxin, native state ensemble, dynamic flexibility index, molecular dynamics, protein conformational dynamics

Author for correspondence:
S. Banu Ozkan
e-mail: banu.ozkan@asu.edu

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4050779.

THE ROYAL SOCIETY PUBLISHING

Ancient thioredoxins evolved to modern-day stability–function requirement by altering native state ensemble

Tushar Modi[1], Jonathan Huihui[2], Kingshuk Ghosh[2] and S. Banu Ozkan[1]

[1]Department of Physics and Center for Biological Physics, Arizona State University, Tempe, AZ 85281, USA
[2]Department of Physics and Astronomy, University of Denver, Denver, CO 80209, USA

SBO, 0000-0002-9351-3758

Thioredoxins (THRXs)—small globular proteins that reduce other proteins—are ubiquitous in all forms of life, from Archaea to mammals. Although ancestral thioredoxins share sequential and structural similarity with the modern-day (extant) homologues, they exhibit significantly different functional activity and stability. We investigate this puzzle by comparative studies of their (ancient and modern-day THRXs') native state ensemble, as quantified by the dynamic flexibility index (DFI), a metric for the relative resilience of an amino acid to perturbations in the rest of the protein. Clustering proteins using DFI profiles strongly resemble an alternative classification scheme based on their activity and stability. The DFI profiles of the extant proteins are substantially different around the α3, α4 helices and catalytic regions. Likewise, allosteric coupling of the active site with the rest of the protein is different between ancient and extant THRXs, possibly explaining the decreased catalytic activity at low pH with evolution. At a global level, we note that the population of low-flexibility (called hinges) and high-flexibility sites increases with evolution. The heterogeneity (quantified by the variance) in DFI distribution increases with the decrease in the melting temperature typically associated with the evolution of ancient proteins to their modern-day counterparts.

This article is part of a discussion meeting issue 'Allostery and molecular machines'.

1. Introduction

Modern proteins have evolved through small changes from ancient times. Much of this information is encoded in protein classes from different species in the three kingdoms of life (bacteria, Archaea and Eukarya). With the advances in phylogenetics and DNA-synthesis techniques, the various ancient genes, including those from the last common ancestors of bacteria, bilaterian animals and vertebrates, have been resurrected in the laboratories. These studies have provided crucial insights on the environmental adaptations and the evolution of functions [1–8]: (i) ancestral proteins are more robust, showing high thermal and chemical stability [9–13] and, more interestingly, (ii) protein structures are conserved more than protein sequences throughout the molecular evolution [12–15]. Thus, the current challenge in molecular evolution is to understand the molecular mechanism of how nature alters the function and biophysical properties through amino acid substitutions, while conserving the three-dimensional structure.

In parallel to the advancements in phylogenetics, recent biophysical studies of proteins have shown that all positions are dynamically linked to each other within a network of interactions, where the strength of each link varies across the protein. This network of interactions leads to intrinsic fluctuations—encoded in the structure and sequence—that govern protein function [4,16–32]. The obsolete view of the single native structure has long been replaced by 'an ensemble of substates' that accurately represent the native state [25]. To shed light into the mechanism of evolution and study how evolution shapes the native ensemble, we have developed a method called dynamic flexibility index (DFI) analysis

[33]. DFI *is a position-specific measure that quantifies* the resilience of a given position (amino acid) to the perturbations occurring at various parts of the protein using linear response theory. Hence, it mimics the multidimensional response when the protein's conformational space is probed upon interaction with small molecules or other cellular constituents. The DFI is related to a given site's relative contribution to the conformational entropy of the protein. Because it is a position-specific metric, it also allows us to quantify the change in flexibility per position throughout evolution. The DFI identifies flexible and rigid positions within the three-dimensional interaction network of the protein structure. The low DFI sites are rigid sites (i.e. hinge sites) and are robust to perturbations owing to their interaction network within the three-dimensional structure. However, they efficiently transfer perturbations to the rest of the protein chain, similar to joints in a skeleton. Thus, they play a critical role in conformational dynamics, and usually correspond to functionally critical, conserved sites in a protein [33]. On the other hand, high DFI sites are flexible, thus mutations/substitutions on these highly flexible sites are more acceptable. Our earlier work of DFI analysis on over 100 human proteins has shown that there is a strong positive correlation between DFI and evolutionary rates. This demonstrates that rigid sites are more conserved while highly evolving sites correspond to sites with high flexibility [33]. DFI analysis of evolution of different protein families including GFP proteins [34], β-lactamase inhibitors [12] and nuclear receptors [35] has shown that alteration of conformational dynamics through allosteric regulations leads to functional changes. Furthermore, our site-specific dynamics-based metric, dynamic coupling index (DCI), reveals that enzymatic function is regulated by dynamically coupled residues, which form an allosteric communication network with the active sites. Evolution uses substitutions on these sites to regulate dynamics of the active sites and binding interfaces [34,36].

Here we analysed the evolution of thioredoxin (THRX). THRXs are versatile, small globular protein molecules comprising about 108 amino acid residues. They belong to a class of oxidoreductase enzymes present in all living organisms from Archaebacteria to humans. These are labelled as the ultimate *moonlighting* proteins with functions as ubiquitous as being a reducing agent for other proteins in biological reactions [37,38].

THRXs from Archaea to humans share about 27–69% sequence similarities and a common three-dimensional fold with a central β-sheet core surrounded by four α-helices [13,39]. The structure of the reduced and oxidized states of THRX has been studied extensively over the last few decades. Their structures contain a highly conserved functional site comprising of two neighbouring redox-active cysteines, Cys-Gly-Pro-Cys (CGPC) [40,41]. The cysteines participate in various redox reactions. The protein is maintained in its reduced state with its functional site having two thiols with cysteines using another class of reducing agents called thioredoxin reductases (ThxRs) like nicotinamide adenine dinucleotide phosphate, flavin adenine dinucleotide (FAD), etc. [38,40,42]. The reduced state of THRX facilitates the reduction of the target protein while turning itself into an oxidized state with a disulfide bond between its two functional cysteines. Although the oxidized state and the reduced state of the protein are very similar to each other in their fold (with most of the differences localized around the disulfide active site), chemical denaturation [43] and thermal [44,45] unfolding experiments suggest that THRXs are more stable in their oxidized state. The mutagenesis analyses of THRXs have shown that the C-terminal α-helix, α4, is known to play a critical role in the folding kinetics of the protein [46]. A shorter helix, α3, has been identified as important for the thermal stability of the fold, having several stabilizing mutations [47]. The central core with β-strand β5 acts like a bridge for interaction between the two α-helices [47] (figure 1).

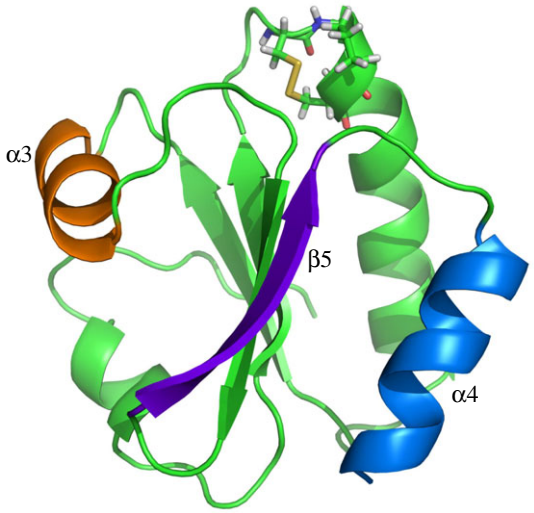

Figure 1. Cartoon representation of AECA THRX. The active sites, CGPC, are shown in stick representation; the β-strand, β5, is shown in purple; and the α-helices, α3 in orange and α4 in blue.

Despite a significant difference in the sequences between ancestral and extant enzymes, the ancestral THRXs, which existed almost 4 billion years ago, display the canonical THRX fold with only minimal structural changes [3,39,48]. However, during 4 billion years of evolution, the stability of THRXs has decreased to adapt to the cooling temperatures of the Earth (in mesophilic lineage). Moreover, the catalytic rates of ancestral THRXs at low pH are different from those of extant ones.

In summary, THRXs achieved adaptation to a cooler and less acidic Earth by altering their stability and changing their catalytic rates while maintaining the same three-dimensional fold. To provide insights into how variation in sequence alters stability and function while conserving the three-dimensional structure, we explore how the native state ensemble has evolved throughout evolution of THRX. To this aim, we performed 1-μs-long molecular dynamics simulations for all ancestral and extant THRXs, and obtained their DFI profiles. Comparison of the DFI profiles between ancestral and extant THRXs on each branch shows a common pattern. The α3 helix, which contributes the most to stability, exhibits enhanced flexibility in modern THRXs, correlating with the decrease in stability observed in modern enzymes. The increased flexibility of α3 was also associated with increased rigidity in the α4 helix along with conserved rigidity of the β-sheet core. This may have helped to maintain the three-dimensional fold through the evolution of THRXs. The clustering of the DFI profiles of all nine THRXs and mapping that onto a two-dimensional landscape of experimentally measured stability and catalytic rates suggest that the native ensemble of THRXs holds the clue to adaptation at a cooler ambient temperature and lower acidic environment.

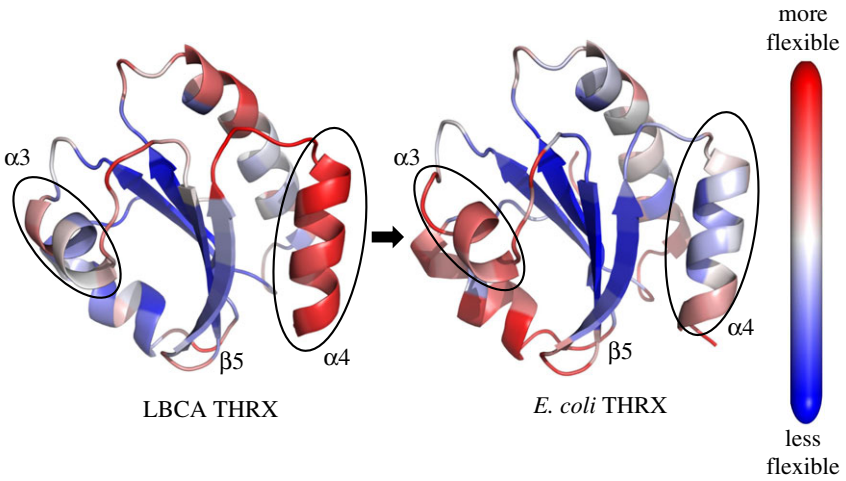

Figure 2. Cartoon representations of ancestral and extant THRXs from bacterial thioredoxins LBCA and *E. coli* THRX, respectively, colour coded with their DFI profiles. Red sites are the most flexible and blue sites are the least flexible (i.e. rigid). We observe a shift in the flexibility profiles of the helices α3 and α4 through evolution owing to changes in thermal stability. By contrast, the β-sheet core remains rigid throughout evolution, highlighting its importance in mediating the interaction between α3 and α4 helices necessary for the THRX fold.

2. Results and discussion

(a) The change in dynamic flexibility index profiles provides insight on decrease in melting temperatures during evolution of THRX

We note a marked difference between the flexibility profiles of the ancestral and extant THRXs, particularly the flexibilities of helices α3 and α4. Through mutagenesis analysis it has been shown that disruption of α3 impacts the overall stability of the THRXs [47]. Interestingly, we observed the enhancement in the flexibility of α3 in *Escherichia coli* (figure 2).

DFI profile comparison between LBCA and *E. coli* (figure 2) shows flexibility–rigidity compensation between the α3 helix and the α4 helix. Specifically, we note that the α3 helix in *E. coli* has substantially higher DFI values when compared with LBCA, while the opposite is seen in the α4 helix. However, the core region of the β-sheet does not show any noticeable alteration in flexibility between *E. coli* and LBCA apart from a slight enhancement in rigidity of the β5 β-strand.

These observations can be further correlated with the measured differences in stability due to differences in amino acids in these regions. After performing a sequence alignment between *E. coli*, LBCA and LPBCA, we note that the ancestral protein LPBCA (close homologue of LBCA) has a critical mutation, P68A (*E. coli* position 68 versus the LPBCA aligned position), in the α3 helix region when compared with *E. coli*. This substitution is typically stabilizing [47]. In the context of two other critical mutations, G74S and K90 L, seen in LPBCA, it has been hypothesized that the α3 helix may tilt towards arginine (R) 89 (another mutation in LPBCA), allowing for a strong charge–dipole interaction [47]. However, *E. coli* THRX does not allow such favourable interactions owing to the different amino acids in these positions. Particularly, 89 is threonine (T) instead of arginine (R) in *E. coli*. Thus, the loss of this favourable interaction may explain the lowered stability of *E. coli* THRX and the enhanced flexibility of the α3 helix as evident from their DFI profiles.

We tested this hypothesis in the context of LBCA. Similar to LPBCA, we note that LBCA also has all the three critical mutations, i.e. P68A, G74S and K90 L. Furthermore, after alignment, we see that position 89 has a positively charged amino acid, lysine, in LBCA.

Motivated by these similarities between LBCA and LPBCA, we further analysed the relative orientation between the α3 helix dipole and the side chain of lysine (K) (at the equivalent of position 89) in LBCA. We observed that the relative orientation between the α3 helix dipole and K89 (as seen in sequence alignment) is more rigid in LBCA (electronic supplementary material, figure SI-1), whereas T, at position 89 in *E. coli*, has a fluctuating orientation with the α3 helix (electronic supplementary material, figure SI-1). This difference indicates that a favourable interaction between the charge and the dipole (of the α3 helix) may be operative in LBCA—similar to LPBCA—and responsible for the higher stability of LBCA and also the rigidity of its α3 helix when compared with *E. coli* THRX.

Turning our attention to the α4 helix, guided by mutagenesis experiments [47], we again noted a set of key mutations between *E. coli* and LPBCA responsible for folding stability differences between the two proteins. For example, S95P and Q98A—both occurring near the end of the α4 helix in LPBCA—cause a reduction in stability. It has been hypothesized that both the serine (S) and glutamine (Q) in *E. coli* may stabilize the loop connecting the α4 helix and the β5 β-strand by possibly using dipole–dipole interactions. However, upon mutating to alanine (A) (for Q98A) and proline (P) (for S95P) in LPBCA, these dipolar interactions are lost, causing a decrease in stability and consequently high flexibility. A second set of mutations, L94Q and F102R (also present in the α4 helix region), in LPBCA have been further implicated to lowering the melting temperature by possibly destabilizing the hydrophobic network between the α4 helix and the β5 β-strand [47]. All these mutations are also present in LBCA, as determined by the sequence alignment, apart from L94R, present in LBCA, instead of L94Q in LPBCA. This suggests that the same physical principles, i.e. loss of charge–dipole or dipole–dipole interaction and/or disruption of the hydrophobic network, may be responsible for the higher flexibility of the α4 helix in LBCA compared with *E. coli*, as observed by DFI analysis.

Similar changes, particularly the compensation for the change in DFI profiles of the α3 and α4 helices, have also been

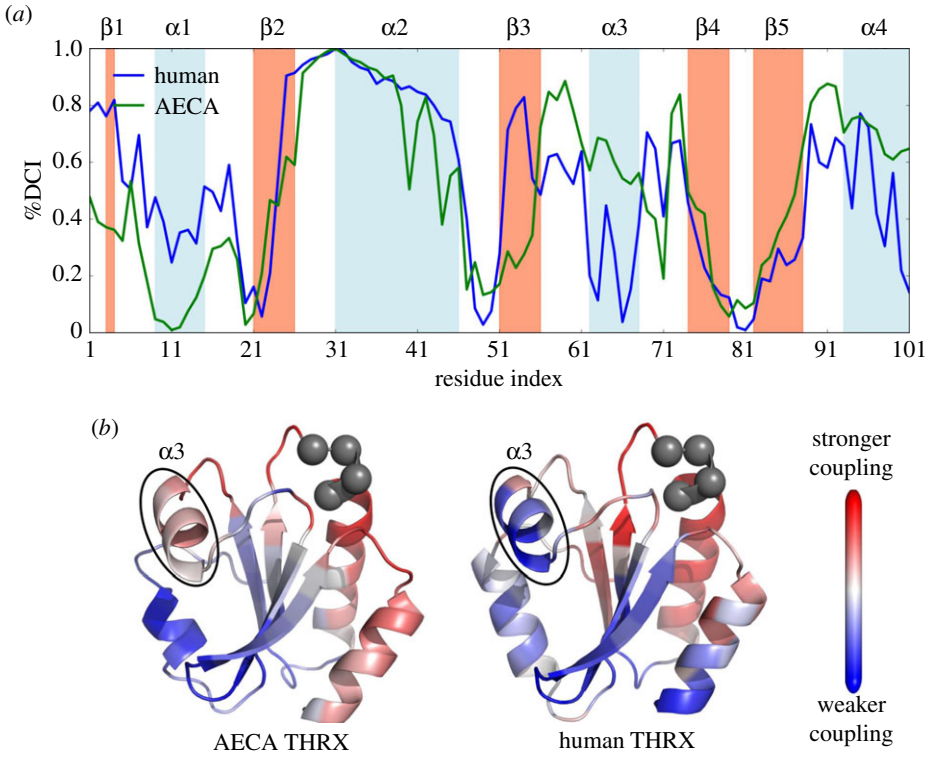

Figure 3. (*a*) Comparison of the coupling of catalytic sites (CGPC) in ancestral THRX (AECA) with extant human THRX using a percentile ranking of the dynamic coupling index (%DCI). We observe a striking difference in the couplings of the α3 region in the respective proteins with their corresponding active sites, suggesting their role in altering catalytic rates. This difference can also be visualized on the cartoon representations (*b*) of AECA and human THRXs where red represents sites highly coupled to the active sites, CGPC (grey spheres) and blue as sites with no significant coupling.

observed in the evolutionary branch of human THRXs when the DFI profiles of AECA THRX and human THRX were compared [47] (electronic supplementary material, figure SI-2). Overall, based on the mutagenesis analysis [47] and DFI comparison between ancestral and extant THRXs, a plausible mechanism of how THRX has evolved to lower stability is suggested. While the increased flexibility of the α3 helical region achieved a decrease in stability to adapt to cooler ambient temperatures, the rigidity conservation of the core and increased rigidity in helix α4 ensured conservation of the canonical THRX fold throughout evolution. The proposed mechanism should be further verified with experimental analysis in the future.

(b) Dynamic coupling of the active site alters throughout evolution

Comparison between the catalytic rates of the last ancestral THRX and their extant variant on each branch in phylogeny shows that the kinetic rates of disulfide bond reduction have decreased during the evolution at pH 5 [3]. Particularly, in the human branch, there is an approximately six-fold decrease in the kinetic rates between its first ancestor, AECA, and the modern-day Human THRX. Our earlier work on protein evolution shows that nature uses distal sites that are dynamically coupled with the active sites to control the active site's dynamics [49–51]. Thus, we analysed how the DCI (see Material and methods for the definition) of the active site changed during THRX evolution (figure 3*a*).

Interestingly, we observed that the dynamic coupling of the α3 helix with the active site has decreased drastically between the ancestral and extant THRXs in the human branch (figure 3*a*,*b*). We noted the same difference between the ancestral and extant THRXs in the *E. coli* branch (electronic supplementary material, figure SI-3). Previous studies have also shown that the α3 helix is part of the substrate-binding region and the change in dynamics of this binding region plays a critical role in the catalytic activity of the protein [52]. Thus, the decrease in the allosteric dynamic coupling of the α3 helix region with the catalytic site may be linked with the lowered activity observed in modern-day THRXs. However, careful mutagenesis experiment is needed to conclusively prove or disprove this hypothesis.

(c) The variance in dynamic flexibility index profile distribution correlates with change in melting temperature

DFI analysis of an ancestral versus extant THRXs in the *E. coli* and human branches provides insight on how the 'fine-tuning' of flexibility profiles of some functionally important structural features is used during evolution. Comparing the distribution of flexibility of various residue positions in ancestral THRXs with the extant ones can additionally reveal information about the change in their native landscape through evolution. The distribution of the flexibility of residues in each THRX protein is obtained by binning the residues according to their DFI scores (figure 4).

Interestingly, the distribution of the DFI profile for the 4-Gyr-old THRX (LBCA) differs from that of extant THRX (*E. coli*) (figure 4*a*,*b*). The high flexibility tail region of the distribution gets more populated in modern THRXs. Likewise, the probability density at low DFI has also increased in THRX of modern organisms, representing a gain of high-flexibility and low-flexibility regions during evolution. This behaviour was observed not only in the bacterial but also

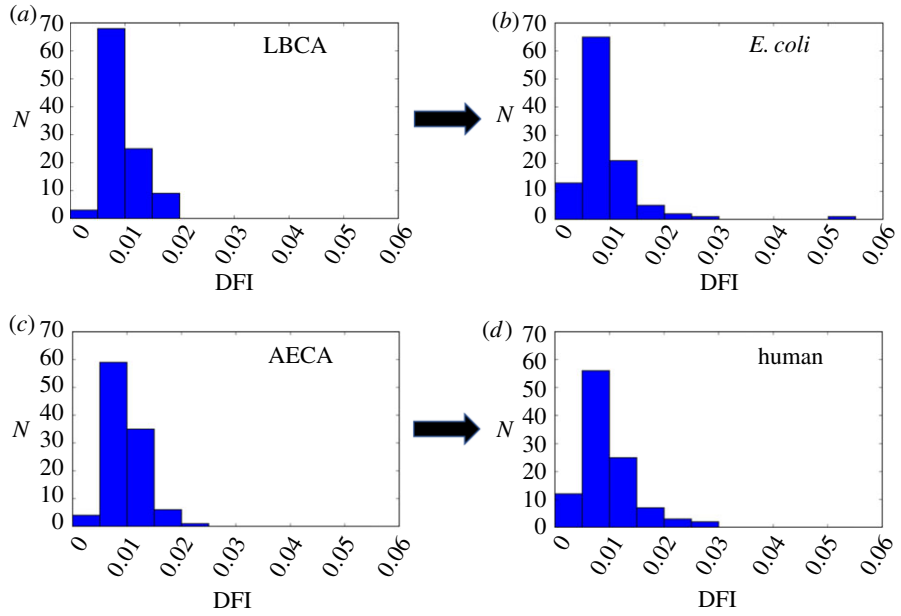

Figure 4. The distributions of DFI profiles in the ancestral and extant THRX proteins belonging to the ancestor of the bacterial branch, LBCA (*a*), evolving to *E. coli* (*b*), and the ancestor of the human branch, AECA (*c*), evolving to extant human THRX (*d*). We observe that through evolution the residues populate low-flexibility and high-flexibility regions in the distribution, making it wider. The fraction of residues populating high-flexibility (DFI greater than 0.02) regions increased in *E. coli* from 0% to approximately 4% and in humans from 1% to 4%. (Online version in colour.)

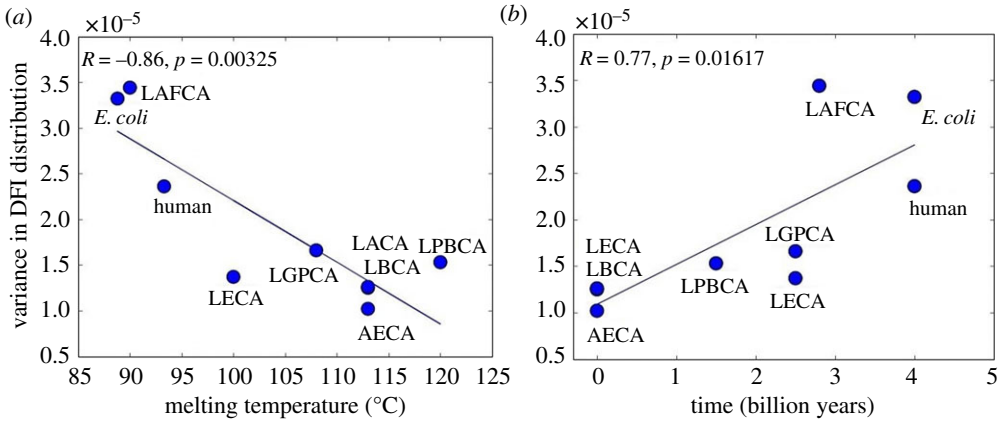

Figure 5. (*a*) The variance in distribution of the flexibility profiles in ancestral and modern-day THRX proteins is observed to correlate strongly with their melting temperatures [8], $R = -0.86$, indicating the implication of flexibility profiles in their thermostability. (*b*) The variance in DFI distributions is observed to increase with evolution [8] and is strongly correlated with the time of evolution of the proteins, $R = 0.77$. (Online version in colour.)

in the human branch of THRX's evolutionary tree (figure 4*c*,*d*). The redistribution of flexible sites in THRX structures helps proteins 'fine-tune' their activity in accordance with the functional requirement.

This characteristic pattern of increasing 'width' of the distributions with evolution is further supported by the high correlation between the variance of the DFI distributions of THRX proteins and their corresponding melting temperatures (correlation, $R = -0.86$ and $p = 3.2 \times 10^{-3}$) (figure 5*a*). In addition, because the decreasing melting temperature correlates with the evolution time, we also note significant correlation between the evolution time of THRX proteins and their variance in DFI distributions (correlation, $R = 0.77$ and $p = 1.6 \times 10^{-2}$) (figure 5*b*).

Overall, this result is in agreement with our previous results [12,34] that evolution shapes the conformational landscape of the native state. We observed two major changes in DFI distributions consistent in all branches. First, the DFI profiles change as the sequences evolve. Second, the probability distributions at the beginning (ancestral sequences) are more compact, having a higher probability towards mid- to high DFI values. This type of distribution ensures evolvability through mutation of various flexible positions in the protein because flexible sites typically correspond to highly evolving residues [53–55]. On the other hand, as we get closer to the modern enzymes, the distribution widens. There is an increase in the probability of the low DFI range along with a longer tail of high DFI values. In other words, a well-distributed set of very rigid and very flexible sites could be an evolutionary mechanism to adapt to low temperature and/or adjust to functional need.

(d) Dynamic flexibility index captures the functional evolution in thioredoxin

To test how the DFI profiles of nine THRXs capture the change in function throughout evolution, we clustered their DFI profiles using PCA analysis (see Material and methods). The

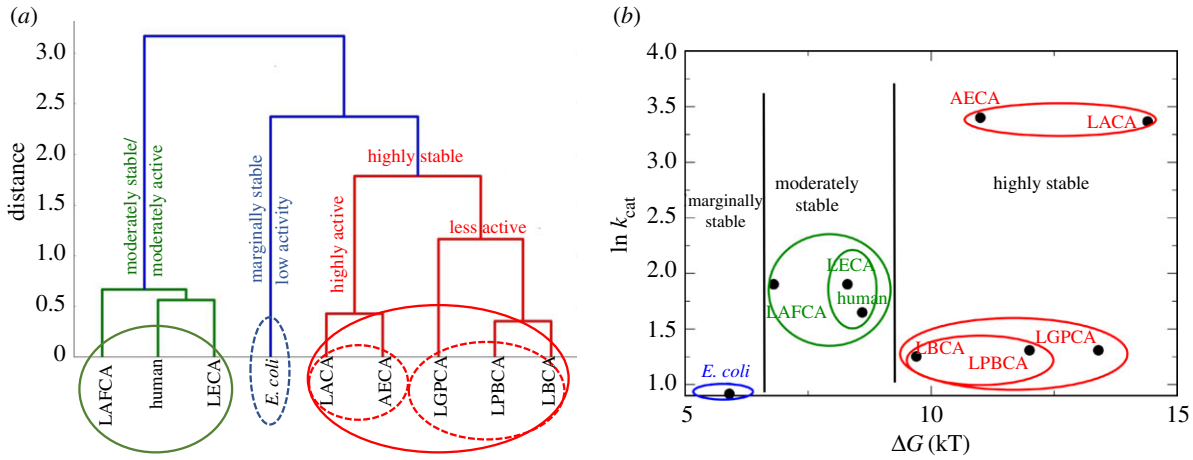

Figure 6. Clustering of ancestral and extant THRXs based on their DFI profiles (*a*) and based on their experimental rate constants of disulfide bond reduction [56] and change in free energy of folding $\Delta G_{folding}$ [56] (*b*). The two clustering criteria give similar results as they segregate the nine THRX proteins based on their activities and thermal stability.

THRXs that clustered together in accordance with the similarities in their flexibility profiles (figure 6*a*) also exhibit similar rate constants for disulfide bond reduction obtained from single-molecule experiments [3]. For example, ancestral THRXs LACA and AECA, belonging to the Archaea branch, have very high rate constants. Consistently, DFI clustering arranges them in the same group. On the other hand, LBCA THRX, from the bacterial branch, which evolved around the same period as AECA (around 4 billion years ago), is in a different cluster with LPBCA and LGPCA. LBCA, LPBCA and LGPCA all have much lower rates for disulfide bond reduction than those of LACA and AECA. Interestingly, human, LAFCA and LECA THRX, all grouped together in a common cluster, share similar kinetic rates of disulfide bond reduction. We also note *E. coli*, having the lowest reduction rate of all other THRXs, is clustered in a separate group (figure 6*a*).

While the above argument clearly explains subclasses, the clustering also raises a question about merging different subgroups to bigger groups. For example, how do we rationalize that the LACA–AECA dyad is grouped in the same class (red) with the triad LGPCA, LPBCA–LBCA in spite of the two subclasses having widely different catalytic rates? Interestingly, we note LACA, AECA, LGPCA, LPBCA and LBCA are all highly stable compared with the triad LAFCA, human and LECA. Thus, DFI-based classification can be best understood by both catalytic rates (k_{cat}, $M^{-1}\,s^{-1}$) [3] and free energy of folding [56]. It is interesting to note that although DFI solely uses the native state ensemble, it can successfully sculpt the landscape in these two coordinates (figure 6*b*). Based on stability and catalytic activity, THRXs can be grouped into three broad categories: (i) highly stable agents (AECA and LACA with high activity; LBCA, LPBCA and LGPCA with low activity), (ii) moderately stable agents (human, LAFCA and LECA) with moderate catalytic activity and (iii) marginally stable, low catalytic activity (*E. coli*) (figure 6*a*,*b*).

In summary, clustering solely based on DFI profiles successfully captures the clustering based on catalytic rates and stability. This suggests that evolution shaped the native state ensemble of THRXs to adapt and function at cooler temperatures and lower acidic ambient conditions. It is also in agreement with our previous analysis on protein evolution [12,50], highlighting that evolution exploits native state conformational dynamics to alter function.

3. Conclusion

Despite the significant structural similarities between ancestral and extant THRXs, they evolved towards lower stability and kinetic turnover rates. To gain insights into the underlying molecular mechanism, we explored how changes in the native state ensemble might have impacted the evolution of THRX proteins. To do this, we compared the difference in their DFI profiles. The enhanced flexibility of the α3 helix in the extant proteins, compared with their ancestral counterparts, is compensated by the decrease in flexibility of the α4 helix and may be responsible for lowering their stability to adapt to cooler ambient temperatures, while keeping the fold conserved. We further noted that the dynamic coupling of important positions with the catalytic site has changed during evolution. Particularly, the decrease in allosteric dynamic coupling of the α3 helical region—critical in substrate binding—with the catalytic site in extant THRXs may be associated with the decrease in catalytic activity at lower acidic conditions.

Comparison of the distribution of the flexibility of residues between ancestral and extant proteins revealed that the population density of high- and low-flexibility sites increases as they evolve. These common features observed in evolution suggest a '*fine-tuning*' of their native ensemble to adjust to ambient conditions in accordance with the evolution in their function. The high correlation between the variance of flexibility distribution of proteins and their melting temperature quantitatively supports this hypothesis.

In addition, clustering these proteins based on their flexibility profiles closely matches grouping using their kinetic rates of disulfide bond reduction and thermal stability. These observations, in agreement with our previous results, highlight that nature uses native state conformational dynamics to adapt and evolve.

4. Material and methods

(a) Molecular dynamics simulation protocol

All starting structures were taken from the Protein Data Bank using the respective accession numbers for the proteins LBCA (4BA7), AECA (3ZIV), LAFCA (2YPM), LECA (2YOI), LACA (2YNX), LGPCA (2YN1), LPBCA (2YJ7), *E. coli* (2TRX) and human (1ERU) THRX [57–59]. Next the H++ web server was used to predict the protonation state of the histidine side chains [60–62]. The

refined structure was then loaded into TLEAP using the ff14SB force field [63]. The disulfide bond was kept in the oxidized state. Protein hydrogens were then added and a 9.0 Å cubic box of TIP3P surrounding water atoms was added, followed by neutralizing ions [64,65]. Systems were then energy-minimized using the SANDER module of AMBER 14 [66–68]. The first cycle of minimization reduced the energy and steric clashes of the solvent with the protein restricted using harmonic restraints. The second cycle of minimization was then performed without the harmonic restraints, so the entire solution could adjust to the local minimum.

Heating, density equilibration and production were all then run using the GPU-accelerated PMEMD module of AMBER 14 [68]. These simulations were performed with periodic boundary conditions and the bond lengths of all covalent hydrogen bonds were constrained using SHAKE [67]. Direct-sum, non-bonded interactions were cut off at 9.0 Å, and long-range electrostatics was calculated using the particle mesh Ewald method [69–71]. The heating phase ran over 100 ps from 0 to 300 K. The density of the system was then allowed to equilibrate over 5 ns at constant temperature and pressure. All production simulations were run using the Langevin thermostat and Berendsen barostat to keep temperature and pressure, respectively, constant. A timestep of 2 fs was used and structural conformations were saved every 10 ps. All simulations were allowed to progress to 1 µs of total simulation time, deemed the minimal required simulation time for convergence based on our earlier study [72].

(b) Dynamic flexibility index

DFI measures the amount of resilience a residue experiences in response to perturbations in the rest of the protein. The dynamic response profile of the protein is explored using the perturbation response scanning (PRS) technique [73,74]. The original approach is based on the elastic network model (ENM) in which the nodes represent Cα atoms [54,75] and the pairwise potential between each atom is given by the potential of a harmonic spring. A small perturbation in the form of random Brownian kick is applied sequentially to each Cα atom in the elastic network. As a first-order approximation, this perturbation mimics the forces exerted by an approaching protein or a ligand in a crowded cellular environment. The perturbations on a single residue result in a cascade of perturbations to all other atoms in the network, inducing a global response. The fluctuation response profile of the positions upon perturbation of a single residue is obtained using linear response theory and given by the equation

$$\Delta \boldsymbol{R}_{3N\times 1} = \mathbf{H}^{-1}_{3N\times 3N}\, \mathbf{F}_{3N\times 1}, \tag{5.1}$$

where $\mathbf{H}$ is the Hessian matrix, a $3N \times 3N$ matrix composed of the second-order derivatives of the harmonic potential with respect to the components of the position vectors for the chain of length N, giving the position covariance of the residue pairs in equilibrium, $\mathbf{F}$ is the external unit force vector applied at N residues in the protein and $\Delta \boldsymbol{R}$ is the response of the force. The force is applied in all the directions at each residue and the magnitude of the response profile is averaged to give an isotropic measure of response.

However, a disadvantage of the ENM-based PRS model is that the coarse-grained network makes it insensitive to changes arising from the biochemical properties of amino acids. Therefore, in order to compare the family of THRXs, having similar back-bone structures, we replace the inverse of the Hessian with the covariance matrix obtained from molecular dynamic simulations (discussed in the previous section).

$$\Delta \boldsymbol{R}_{3N\times 1} = \boldsymbol{G}_{3N\times 3N}\mathbf{F}_{3N\times 1}. \tag{5.2}$$

Here, $\boldsymbol{G}$ is the covariance matrix containing the dynamic properties of the system. The covariance matrix contains the data for long-range interactions, solvation effects and biochemical specificities of all types of interactions. The DFI is calculated by computing the fluctuation response of all the residues in a protein (i.e. ΔR_i^j, $i = 1,..N$) by applying a unit force (isotropic) as perturbation at a specific position (i.e. site j) using equation (4.2). We repeat this single-residue perturbation for all residues (i.e. $1 \le j \le N$) in the protein. Consequently, the perturbation matrix, A is constructed,

$$\boldsymbol{A}_{N\times N} = \begin{bmatrix} |\Delta \boldsymbol{R}^1|_1 & \cdots & |\Delta \boldsymbol{R}^N|_1 \\ \vdots & \ddots & \vdots \\ |\Delta \boldsymbol{R}^1|_N & \cdots & |\Delta \boldsymbol{R}^N|_N \end{bmatrix}, \tag{5.3}$$

where $|\Delta \boldsymbol{R}^j|_i = \sqrt{\langle(\Delta \boldsymbol{R})\rangle^2}$ is the magnitude of fluctuation response at site 'i' due to the perturbations at site 'j'. A sum of a given row of perturbation matrix gives the net average displacement of the residue from its equilibrium position when all the residues are perturbed by an isotropic unit force one at a time. The DFI score of a position 'i' is defined as the net response of that position normalized with the net displacement of the whole protein when all the residues are perturbed, i.e.

$$\mathrm{DFI}_i = \frac{\sum_{j=1}^{N} |\Delta \boldsymbol{R}^j|_i}{\sum_{i=1}^{N}\sum_{j=1}^{N} |\Delta \boldsymbol{R}^j|_i}. \tag{5.4}$$

Thus, a higher DFI score of a residue position 'i' implies a more flexible site and a low score implies a rigid site with lower response to perturbations in the protein.

Recently, we have extended this method to identify dynamic coupling between any given residue and functionally important residues by introducing a new metric called the DCI. The DCI metric can identify sites that are distal to functional sites but impact active-site dynamics through dynamic allosteric coupling [49,50,76]. This type of allosteric coupling is important; sites with strong dynamic allosteric coupling to functionally critical residues (dynamic allosteric residue coupling (DARC) spots), regardless of separation distance, also contribute to the function. Thus, a mutation at such a site can disrupt the allosteric dynamic coupling or regulation, leading to functional degradation. As defined, the DCI is the ratio of the sum of the mean square fluctuation response of the residue 'i' upon functional site perturbations (i.e. catalytic residues) to the response of residue 'i' upon perturbations on all residues. DCI enables us to identify DARC spot residues, which are more sensitive to perturbations exerted on residues critical for function. This index can be used to identify the residues involved in allosteric regulation. It is expressed as

$$\mathrm{DCI}_i = \frac{\sum_{j}^{N_{\mathrm{functional}}} |\Delta \boldsymbol{R}^j|_i / N_{\mathrm{functional}}}{\sum_{j=1}^{N} |\Delta \boldsymbol{R}^j|_i / N}, \tag{5.5}$$

where $|\Delta \boldsymbol{R}^j|_i$ is the response fluctuation profile of residue 'i' upon perturbation of residue 'j'. The numerator is the average mean square fluctuation response obtained over the perturbation of the functionally critical residues $N_{\mathrm{functional}}$, and the denominator is the average mean square fluctuation response over all residues. Similar to DFI, the DCI profiles can also be computed using the covariance matrix obtained from molecular dynamics simulations or the inverse Hessian of the ENM. In this study, we used the covariance matrices.

(c) Clustering the DFI profiles of thioredoxin proteins

We clustered the DFI profiles of different extant and ancestral THRXs by comparing their percentile rankings. To compare the flexibility profiles, the proteins are aligned according to their multiple sequence alignment and are concatenated into a data matrix $\boldsymbol{X}$. Singular value decomposition (SVD) is a statistical procedure to factorize the data into the orthonormal basis, which represents the vector space containing data. It is similar to principal component analysis which could be used to understand the structure of data or to increase the signal-to-noise ratio in data by eliminating the redundant dimensions and mapping it on a lower-dimensional space. Clustering by SVD acts as an effective noise filter by isolating the highest variances among data points in the top principal

vectors. Consequently, the remaining insignificant singular vectors can be omitted from the reconstruction.

The DFI profiles of all proteins are merged into a matrix $\boldsymbol{X}$, of dimensions ($\boldsymbol{m} \times \boldsymbol{n}$). Here $\boldsymbol{m}$ is the number of datasets (proteins) we are clustering together, each having $\boldsymbol{n}$ number of attributes (number of residues). On performing SVD, $\boldsymbol{X}$ is decomposed as follows:

$$[X]_{m\times n} = [U]_{m\times m}[\Sigma]_{m\times n}[V]_{n\times n} \tag{5.6}$$

Here, $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices with orthonormal columns and are called left singular vectors and right singular vectors, respectively, and $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements known as the singular values of $\boldsymbol{X}$.

The singular values of $\boldsymbol{X}$, by convention, are arranged in a decreasing order of their magnitude; $\sigma = \{\sigma_i\}$ represent the variances in the corresponding left and right singular vectors. The set of highest singular values representing the largest variance in the orthonormal singular vectors can be interpreted to show the characteristics in the data $\boldsymbol{X}$ and the right singular vectors create the orthonormal basis which spans the vector space representing the data. The left singular vectors contain weights indicating the significance of each attribute in the dataset as $w_i = \Sigma_{k=1}^{r}\sigma_k|u_{ik}|$. Using these features of the decomposed singular vectors, we can create another matrix, $\boldsymbol{X}^*$ using only the highest 'r' singular values which can mimic the basic characteristics of the original dataset. Thus, $\boldsymbol{X}^*$ can be represented as

$$[X^*]_{m\times r} = [V^*]_{m\times r}[\Sigma^*]_{r\times r}. \tag{5.7}$$

Here, $\boldsymbol{\Sigma}^*$ contains only largest $\boldsymbol{r}$ singular values and $\boldsymbol{V}^*$ contains the corresponding right singular vectors. The data are now clustered hierarchically based on the pairwise distance between different proteins in the reconstructed DFI data with reduced dimensions.

For a pair of datasets (or between flexibility profiles of any two proteins) j_1 and j_2, the distance between them in the original set of data was given by

$$d_{12} = \sqrt{\sum_{i=1}^{n}\left(X_i^{j1} - X_i^{j2}\right)^2}, \tag{5.8}$$

which in reduced dimensions can be calculated as

$$d_{12} = \sqrt{\sum_{i=1}^{r}\left(X_i^{*j1} - X_i^{*j2}\right)^2}. \tag{5.9}$$

These pairwise distances are used as the parameters for clustering the flexibility profiles of THRX. We calculated the multiple sequence alignment of ancestral and modern THRXs. The DFI profiles are aligned with respect to LBCA, taking into consideration the gaps in sequences of other THRX proteins, and the data are clubbed into a dataset matrix $\boldsymbol{X}$. The three largest singular values are used for reconstruction of data and clustering. The pairwise distance between each protein using the equation above is used for clustering them hierarchically.

A bottom-up approach is used for the hierarchical clustering, where initially each protein is assigned its own cluster and then, in successive iteration, closest clusters are merged together into a common cluster. In this approach, the distance between clusters is defined by the average pairwise distance between their components (average linkage clustering [77]). In the end, the clusters are represented hierarchically using a dendrogram, where the vertical axis denotes the Euclidean distance between various clusters and among their sub-clusters.

Data accessibility. This article has no additional data.

Authors' contributions. S.B.O. and K.G. conceived and designed the study. T.M. and J.H. performed and analysed simulations, and generated the results. T.M., J.H., K.G. and S.B.O. wrote the manuscript.

Competing interests. We declare we have no competing interests.

Funding. Support from NSF-MCB Award 1715591 and Scialog Fellow Award by RCSA and the Gordon & Betty Moore Foundation is gratefully acknowledged by S.B.O. K.G. acknowledges support from National Science Foundation (NSF) (award number 1149992) and Research Corporation for Science Advancement.

Acknowledgement. We thank Lucas Sawle for the help with the all-atom molecular dynamics simulation.

References

1. Carroll SM, Bridgham JT, Thornton JW. 2008 Evolution of hormone signaling in elasmobranchs by exploitation of promiscuous receptors. *Mol. Biol. Evol.* **25**, 2643–2652. (doi:10.1093/molbev/msn204)
2. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. 2007 Crystal structure of an ancient protein. *Science* **317**, 1544–1548. (doi:10.1126/science.1142819)
3. Perez-Jimenez R *et al.* 2011 Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.* **18**, nsmb.2020. (doi:10.1038/nsmb.2020)
4. Wilson C *et al.* 2015 Kinase dynamics. Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **347**, 882–886. (doi:10.1126/science.aaa1823)
5. Bar-Rogovsky H, Hugenmatter A, Tawfik DS. 2013 The evolutionary origins of detoxifying enzymes: the mammalian serum paraoxonases (PONs) relate to bacterial homoserine lactonases. *J. Biol. Chem.* **288**, 23 914–23 927. (doi:10.1074/jbc.M112.427922)
6. Smith SD, Wang S, Rausher MD. 2013 Functional evolution of an anthocyanin pathway enzyme during a flower color transition. *Mol. Biol. Evol.* **30**, 602–612. (doi:10.1093/molbev/mss255)
7. Boucher JI, Jacobowitz JR, Beckett BC, Classen S, Theobald DL. 2014 An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *eLife* **3**, e02304. (doi:10.7554/eLife.02304)
8. Ingles-Prieto A, Ibarra-Molero B, Delgado-Delgado A, Perez-Jimenez R, Fernandez JM, Gaucher EA, Sanchez-Ruiz JM, Gavira JA. 2013 Conservation of protein structure over four billion years. *Structure 1993.* **21**, 1690–1697. (doi:10.1016/j.str.2013.06.020)
9. Trudeau DL, Kaltenbach M, Tawfik DS. 2016 On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.* **33**, 2633–2641. (doi:10.1093/molbev/msw138)
10. Akanuma S, Nakajima Y, Yokobori S, Kimura M, Nemoto N, Mase T, Miyazono K, Tanokura M, Yamagishi A. 2013 Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl Acad. Sci. USA* **110**, 11 067–11 072. (doi:10.1073/pnas.1308215110)
11. Hart KM, Harms MJ, Schmidt BH, Elya C, Thornton JW, Marqusee S. 2014 Thermodynamic system drift in protein evolution. *PLoS Biol.* **12**, e1001994. (doi:10.1371/journal.pbio.1001994)
12. Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. 2015 Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. *Mol. Biol. Evol.* **32**, 132–143. (doi:10.1093/molbev/msu281)
13. 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. (doi:10.1093/molbev/msu312)
14. Bridgham JT, Ortlund EA, Thornton JW. 2009 An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**, 515–519. (doi:10.1038/nature08249)
15. Choi I-G, Kim S-H. 2006 Evolution of protein structural classes and protein sequence families. *Proc. Natl Acad. Sci. USA* **103**, 14 056–14 061. (doi:10.1073/pnas.0606239103)
16. McLeish TCB, Cann MJ, Rodgers TL. 2015 Dynamic transmission of protein allostery without structural change: spatial pathways or global modes? *Biophys. J.* **109**, 1240–1250. (doi:10.1016/j.bpj.2015.08.009)

17. Bahar I, Lezon TR, Yang L-W, Eyal E. 2010 Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* **39**, 23–42. (doi:10.1146/annurev.biophys.093008.131258)
18. Bahar I, Chennubhotla C, Tobi D. 2007 Intrinsic enzyme dynamics in the unbound state and relation to allosteric regulation. *Curr. Opin. Struct. Biol.* **17**, 633–640. (doi:10.1016/j.sbi.2007.09.011)
19. Chennubhotla C, Yang Z, Bahar I. 2008 Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. *Mol. Biosyst.* **4**, 287–292. (doi:10.1039/b717819k)
20. Tobi D, Bahar I. 2005 Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl Acad. Sci. USA* **102**, 18 908–18 913. (doi:10.1073/pnas.0507603102)
21. Boehr DD, Dyson HJ, Wright PE. 2006 An NMR perspective on enzyme dynamics. *Chem. Rev.* **106**, 3055–3079. (doi:10.1021/cr050312q)
22. Mazal H, Aviram H, Riven I, Haran G. 2018 Effect of ligand binding on a protein with a complex folding landscape. *Phys. Chem. Chem. Phys.* **20**, 3054–3062.
23. Kar G, Keskin O, Gursoy A, Nussinov R. 2010 Allostery and population shift in drug discovery. *Curr. Opin. Pharmacol.* **10**, 715–722. (doi:10.1016/j.coph.2010.09.002)
24. Tsai C-J, Nussinov R. 2014 A unified view of 'how allostery works'. *PLoS Comput. Biol.* **10**, 1–12.
25. Tokuriki N, Tawfik DS. 2009 Protein dynamism and evolvability. *Science* **324**, 203–207. (doi:10.1126/science.1169375)
26. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D. 2007 A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **450**, 913–916. (doi:10.1038/nature06407)
27. Zheng W, Brooks BR, Thirumalai D. 2006 Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl Acad. Sci. USA* **103**, 7664–7669. (doi:10.1073/pnas.0510426103)
28. Dima RI, Thirumalai D. 2006 Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Sci. Publ. Protein Soc.* **15**, 258–268. (doi:10.1110/ps.051767306)
29. Liu T, Whitten ST, Hilser VJ. 2007 Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble. *Proc. Natl Acad. Sci. USA* **104**, 4347–4352. (doi:10.1073/pnas.0607132104)
30. Gruber R, Horovitz A. 2016 Allosteric mechanisms in chaperonin machines. *Chem. Rev.* **116**, 6588–6606. (doi:10.1021/acs.chemrev.5b00556)
31. Buchenberg S, Sittel F, Stock G. 2017 Time-resolved observation of protein allosteric communication. *Proc. Natl Acad. Sci. USA* **114**, E6804–E6811. (doi:10.1073/pnas.1707694114)
32. Sawle L, Huihui J, Ghosh K. 2017 All-atom simulations reveal protein charge decoration in the folded and unfolded ensemble is key in thermophilic adaptation. *J. Chem. Theory Comput.* **13**, 5065–5075. (doi:10.1021/acs.jctc.7b00545)
33. Nevin Gerek Z, Kumar S, Banu Ozkan S. 2013 Structural dynamics flexibility informs function and evolution at a proteome scale. *Evol. Appl.* **6**, 423–433. (doi:10.1111/eva.12052)
34. Kim H *et al.* 2015 A hinge migration mechanism unlocks the evolution of green-to-red photoconversion in GFP-like proteins. *Structure 1993.* **23**, 34–43.
35. Glembo TJ, Farrell DW, Gerek ZN, Thorpe MF, Ozkan SB. 2012 Collective dynamics differentiates functional divergence in protein evolution. *PLoS Comput. Biol.* **8**, e1002428. (doi:10.1371/journal.pcbi.1002428)
36. Kumar A, Butler BM, Kumar S, Ozkan SB. 2015 Integration of structural dynamics and molecular evolution via protein interaction networks: a new era in genomic medicine. *Curr. Opin. Struct. Biol.* **35**(Suppl. C), 135–142. (doi:10.1016/j.sbi.2015.11.002)
37. Holmgren A. 1985 Thioredoxin. *Annu. Rev. Biochem.* **54**, 237–271. (doi:10.1146/annurev.bi.54.070185.001321)
38. Mustacich D, Powis G. 2000 Thioredoxin reductase. *Biochem. J.* **346**, 1–8. (doi:10.1042/bj3460001)
39. Romero-Romero ML, Risso VA, Martinez-Rodriguez S, Ibarra-Molero B, Sanchez-Ruiz JM. 2016 Engineering ancestral protein hyperstability. *Biochem. J.* **473**, 3611–3620. (doi:10.1042/BCJ20160532)
40. Eklund H, Gleason FK, Holmgren A. 1991 Structural and functional relations among thioredoxins of different species. *Proteins* **11**, 13–28. (doi:10.1002/prot.340110103)
41. Weichsel A, Gasdaska JR, Powis G, Montfort WR. 1996 Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer. *Structure* **4**, 735–751. (doi:10.1016/S0969-2126(96)00079-2)
42. Arnér ESJ, Holmgren A. 2000 Physiological functions of thioredoxin and thioredoxin reductase. *Eur. J. Biochem.* **267**, 6102–6109. (doi:10.1046/j.1432-1327.2000.01701.x)
43. Chakrabarti A, Srivastava S, Swaminathan CP, Surolia A, Varadarajan R. 1999 Thermodynamics of replacing an α-helical Pro residue in the p40S mutant of *Escherichia coli* thioredoxin. *Protein Sci.* **8**, 2455–2459. (doi:10.1110/ps.8.11.2455)
44. Ladbury JE, Kishore N, Hellinga HW, Wynn R, Sturtevant JM. 1994 Thermodynamic effects of reduction of the active-site disulfide of *Escherichia coli* thioredoxin explored by differential scanning calorimetry. *Biochemistry* **33**, 3688–3692. (doi:10.1021/bi00178a027)
45. Godoy-Ruiz R, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM. 2004 Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations. *J. Mol. Biol.* **336**, 313–318. (doi:10.1016/j.jmb.2003.12.048)
46. Vazquez DS, Sánchez IE, Garrote A, Sica MP, Santos J. 2015 The *E. coli* thioredoxin folding mechanism: the key role of the C-terminal helix. *Biochim. Biophys. Acta* **1854**, 127–137. (doi:10.1016/j.bbapap.2014.11.004)
47. Cabrera ÁC, Sánchez-Murcia AP, Gago F. 2017 Making sense of the past: hyperstability of ancestral thioredoxins explained by free energy simulations. *Phys. Chem. Chem. Phys.* **19**, 23 239–23 246. (doi:10.1039/C7CP03659K)
48. Risso VA, Gavira JA, Gaucher EA, Sanchez-Ruiz JM. 2014 Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins* **82**, 887–896. (doi:10.1002/prot.24575)
49. Larrimore KE *et al.* 2017 Plant-expressed cocaine hydrolase variants of butyrylcholinesterase exhibit altered allosteric effects of cholinesterase activity and increased inhibitor sensitivity. *Sci. Rep.* **7**, 10419. (doi:10.1038/s41598-017-10571-z)
50. Kumar A, Glembo TJ, Ozkan SB. 2015 The role of conformational dynamics and allostery in the disease development of human ferritin. *Biophys. J.* **109**, 1273–1281. (doi:10.1016/j.bpj.2015.06.060)
51. Butler BM, Kumar A, Gerek ZN, Sanderford M, Kumar S, Ozkan SB. 2000 Dynamic coupling of residues in proteins explains enigmatic pathogenic missense variation in humans. *Rev*.
52. Perez-Jimenez R *et al.* 2009 Diversity of chemical mechanisms in thioredoxin catalysis revealed by single-molecule force spectroscopy. *Nat. Struct. Mol. Biol.* **16**, 890. (doi:10.1038/nsmb.1627)
53. Gerek ZN, Ozkan SB. 2010 A flexible docking scheme to explore the binding selectivity of PDZ domains. *Protein Sci.* **19**, 914–928.
54. Gerek ZN, Ozkan SB. 2011 Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. *PLoS Comput. Biol.* **7**, e1002154. (doi:10.1371/journal.pcbi.1002154)
55. Gerek ZN, Keskin O, Ozkan SB. 2009 Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior. *Proteins* **77**, 796–811. (doi:10.1002/prot.22492)
56. Tzul FO, Vasilchuk 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. (doi:10.1073/pnas.1613892114)
57. Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. 2013 Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.* **135**, 2899–2902. (doi:10.1021/ja311630a)
58. Katti SK, LeMaster DM, Eklund H. 1990 Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.* **212**, 167–184. (doi:10.1016/0022-2836(90)90313-B)
59. Capitani G, Markovic-Housley Z, Jansonius JN, Val GD, Morris M, Schürmann P, Biochimie L. 1998 Crystal structures of thioredoxins f and m from spinach chloroplasts. In *Photosynthesis: mechanisms and effects* (ed. G Garab), pp. 1939–1942. Dordrecht, the Netherlands: Springer.

60. Anandakrishnan R, Aguilar B, Onufriev AV. 2012 H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **40**, W537–W541. (doi:10.1093/nar/gks375)
61. Myers J, Grothaus G, Narayanan S, Onufriev A. 2006 A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins* **63**, 928–938. (doi:10.1002/prot.20922)
62. Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A. 2005 H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **33**, W368–W371. (doi:10.1093/nar/gki464)
63. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. 2015 ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713. (doi:10.1021/acs.jctc.5b00255)
64. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. 1983 Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935. (doi:10.1063/1.445869)
65. Neria E, Fischer S, Karplus M. 1996 Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **105**, 1902–1921. (doi:10.1063/1.472061)
66. Case D *et al.* 2014 AMBER 14. University of California, San Francisco.
67. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, DeBolt S, Ferguson D, Seibel G, Kollman P. 1995 AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **91**, 1–41. (doi:10.1016/0010-4655(95)00041-D)
68. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. 2013 Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888. (doi:10.1021/ct400314y)
69. Darden T, York D, Pedersen L. 1993 Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10 089–10 092. (doi:10.1063/1.464397)
70. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. 1995 A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593. (doi:10.1063/1.470117)
71. Hockney RW, Eastwood JW. 1988 *Computer simulation using particles*. Bristol, PA: Taylor & Francis, Inc.
72. Sawle L, Ghosh K. 2016 Convergence of molecular dynamics simulation of protein native states: feasibility versus self-consistency dilemma. *J. Chem. Theory Comput.* **12**, 861–869. (doi:10.1021/acs.jctc.5b00999)
73. Bolia A, Gerek ZN, Keskin O, Banu Ozkan S, Dev KK. 2012 The binding affinities of proteins interacting with the PDZ domain of PICK1. *Proteins* **80**, 1393–1408. (doi:10.1002/prot.24034)
74. Bolia A, Ozkan SB. 2016 Adaptive BP-dock: an induced fit docking approach for full receptor flexibility. *J. Chem. Inf. Model.* **56**, 734–746. (doi:10.1021/acs.jcim.5b00587)
75. Atilgan C, Gerek ZN, Ozkan SB, Atilgan AR. 2010 Manipulation of conformational change in proteins by single-residue perturbations. *Biophys. J.* **99**, 933–943. (doi:10.1016/j.bpj.2010.05.020)
76. Li Z, Bolia A, Maxwell JD, Bobkov AA, Ghirlanda G, Ozkan SB, Margulis CJ. 2015 A rigid hinge region is necessary for high-affinity binding of dimannose to cyanovirin and associated constructs. *Biochemistry* **54**, 6951–6960. (doi:10.1021/acs.biochem.5b00635)
77. Day WHE, Edelsbrunner H. 1984 Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classif.* **1**, 7–24. (doi:10.1007/BF01890115)