Proteome-scale Analysis of Vertebrate Protein Thermoadaptation Modulated by Dynamic Allostery and Protein Solvation

Zhen-lu Li [1*] and Matthias Buck [1,2*]

[1]Department of Physiology and Biophysics, Case Western Reserve University, School of Medicine, 10900 Euclid Avenue, Cleveland, Ohio 44106, U. S. A. [2]Department of Pharmacology; Department of Neurosciences, Case Western Reserve University, School of Medicine, 10900 Euclid Avenue, Cleveland, Ohio 44106, U. S. A.

E-mail corresponding authors: zhenlu.li@case.edu, matthias.buck@case.edu

Subtitle: Trends in the selection of amino acids in the temperature adaptation of vertebrate organisms

Abstract

Despite large differences in behaviors and living conditions, vertebrate organisms share the great majority of proteins often with subtle differences in amino acid sequence. By comparing a set of substantially homologous proteins between model vertebrate organisms at a sub-proteome level, we discover a pattern of amino acid conservation and a shift in amino acid use, noticeably with an apparent distinction between homeotherms (warm-blooded species) and poikilotherms (cold-blooded species). Importantly, we establish a connection between the thermoadaptation of protein sequences manifest in the evolved proteins and two of their physical features: a change in their proteins dynamics and in their solvation. For poikilotherms such as frog and fish, the lower body temperature is expected to increase the association of proteins due to a decrease in protein dynamics and correspondingly lower entropy penalty on binding. In order to prevent overly-sticky protein association at low temperatures, we find that poikilotherms enhance the solvation of their proteins by favoring polar amino acids on their protein's surface. This study unveils a general mechanism behind amino acid choices that constitute part of the thermoadaptation of vertebrate organisms at the molecular level.

Introduction

The diversity of living organisms is a result of evolution of genes and biological macromolecules. In general, todays organisms share a large number of homologous proteins of similar sequences with small differences. Protein sequences evolved in a way to optimize their proteins function, allowing the organisms to adapt to their different environments. The role of site specific mutations has been demonstrated in multiple proteins, as an organism adapts to extreme cold (e.g. L-Lactate dehydrogenase A chain, LDHA),[1] oxygen availability (Hemoglobin),[2] light (Rhodopsin),[3,4] high altitude (p53),[5] or the scarcity of food (Melanocortin 4 receptor).[6] In spite of sharing a large number of near identical proteins between the different organisms, vertebrates have an obvious and wide ranging difference in body temperature, most noticeably between warm-blooded and cold-blooded animals (homeotherms and poikilotherms, respectively). The activity of proteins is typically highly sensitive to temperatures. In a human being, a small increase of body temperature by 2-5℃ can seriously damage cells, tissues and organs. Thus, it is remarkable that homologous proteins could adapt and function at a broad range of temperatures. Mechanisms of protein thermoadaptation related to protein flexibility, stability and dynamics have been recognized in several proteins studied, e.g. by comparing a protein from a normal and in a thermophilic setting. [1,7,8] However, different strategies may be employed- some of which are not so obvious. Importantly, it is largely unknown whether a general principle for thermoadaptation may exist, at a proteome-wide level of scale over a broad range of organisms.

The function of proteins is rooted in their properties as dynamic macromolecules, exhibited by regional conformational fluctuations, e.g. mainchain θ, ψ changes in active site loops or changes to a different sidechain rotamer, and, in some cases by global configurational transitions such as changes in the orientation or relative distances of protein domains. However, there are also thermal fluctuation of the amino acid constituent atoms around their average positions in the protein's structure.[9] The process of protein recognition or association could in principle be entirely driven by changes in local protein thermal fluctuation without altering the average

structure of ground state of a protein---a process commonly termed as entropy driven dynamic allostery.[10] In order to adjust to the environmental temperature that alters protein dynamics and function, an organism would have accumulated certain types of amino acid changes (mutations), possibly at preferred sites, as part of a process of adaptation to any change in their body temperature. In addition, proteins like other molecules are hydrated in the aqueous solution that exists in- or outside the cell. The adaptation of a protein surface, but also in some cases of interior positions, to water represents another important factor influencing protein recognition and activity.[11,12] The "life on the edge of solubility" hypothesis posits that the expression of a protein in the cellular environment occurs very close to its solubility limit under most physiological solution.[13] Therefore, even small site mutations altering protein solvation may, in some cases dramatically influence the total protein solubility (e.g. sickle cell Hemoglobin). Overall, the dynamics and solubility of a protein are both crucial for its activity and for the well-being of the organism.[14-21] Thus, an evolutionary selection for a required optimal protein function- that is fitness, is likely modulated by the protein dynamics and solvation. In this study, we reveal a role for changes in protein dynamics and solvation in the thermoadaptation of proteins by analyzing sub-proteomes of multiple vertebrate organisms.

Results and Discussion

Conservation and Occurrence Shift of 20 Amino Acids Over Sub-proteomes of Model Vertebrate Organisms

We compared the sequence of homologous proteins between human (Homo sapiens) and several organisms typically chosen as representative models or reference in studies (Fig. 1a): Compared with human proteins are those of chimpanzee (Pan troglodytes), mouse (Mus musculus), chicken (Gallus gallus), African clawed frog (Xenopus laevis), and zebrafish (Danio rerio) over a selected 3607, 14046, 1921, 1565, and 2353 protein pairs respectively (see also Table S1). For each protein- a homolog pair between human and another organism- we compare the difference in the content for each one of the 20 amino acids (scaling by protein lengths) to give the value of occurrence shift $\Delta C_i(AA)$ (see Methods). For example, 143 amino acids constitute the Hemoglobin subunit alpha (HSA) in zebrafish, of which 18 are alanine. The homologous human protein has 21 alanines in 142 amino acids. Thus, the difference in the content of alanine ΔC(Ala; HSA) in Hemoglobin subunit alpha between zebrafish (12.59%) and human (14.79%) is -2.20 % [ΔC(Ala; HSA)=C(Ala; HSA; zebrafish)-C(Ala; HSA; human)]. In the same way, the differences were calculated for the 2353 paired proteins between zebrafish and human. The % difference for Ala is collected into 9 bins to give an overview of all protein pairs and then plotted as a histogram (Fig. 1b). Rather than using the number of protein pairs in each bin as the y-axis we can divide by their number to give the % of proteins in that bin. Statistically, of the 2353 proteins, 220 proteins (9%) have only subtle alternations in the sequence content of alanine within $\pm$0.125%. For 823 proteins (35%), zebrafish has less alanine than the human protein with a difference in the absolute content of $\leq$ -1%. Similarly, there are 444 proteins (19%) in zebrafish which have more alanine $\geq$ 1% compared to the human protein homolog. Therefore, there is an obvious asymmetry or shift in amino acid use of alanine over a considerable proportion of the proteome between zebrafish and human proteins. In the same way, the results for all the other amino acids between the above-mentioned model organisms are analyzed and summarized in Fig. 1c (also Fig. S1).

In general, the extent of variation in amino acid utilization rely on the closeness of a model organism to the human sequences. For example, amino acid sequences between chimpanzee and human proteins are highly identical with only very few changes. Mouse and chicken proteins have moderate changes compared to the human homologs. In contrast, for the sequences of African clawed frog and of fish, there are many amino acid changes relative to the human proteins. Our protein sequence comparison yields four major observations: i) For all the model organisms examined, the occurrence of aromatic amino acids (tryptophan, phenylalanine, tyrosine), of histidine as well as cysteine are mostly conserved. For a majority of the homologous protein pairs the difference in the occurrence of these five amino acids is within $\pm$0.125%. Therefore, they are considered to be the most stable amino acids in protein adaptation/evolution. Remarkably, tryptophan, phenylalanine and histidine are essential amino acids that cannot be synthesized de novo by humans. Especially, the biosynthetic enzymes and the pathway for generating tryptophan is highly conserved and likely evolved only once in the early evolution of cells.[22,23] The aromatic amino acids, as well as cysteine (for cell exterior proteins forming disulfide bonds) play a key role for protein stability, with Trp and Phe usually located in protein interiors, while

His and Tyr are typically found at the surface regions for specific protein-protein interactions sites. His and Cys, are also often highly conserved in chelating cofactors or ions. Thus, there factors are related to the conservation of these amino acids for the structural- and/or catalytic functions of the proteins.

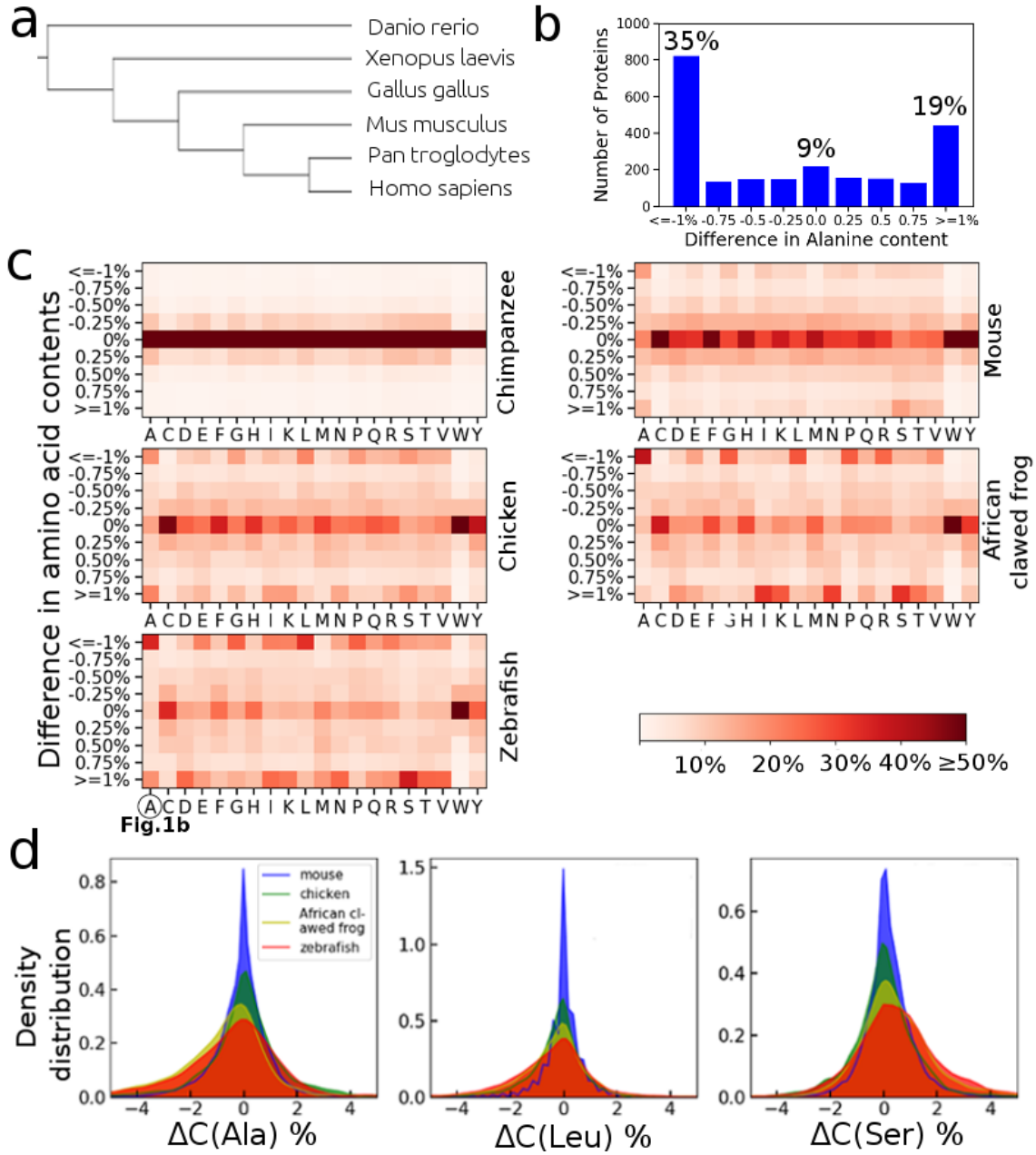


Figure 1: Conservation and change of the utilization of the 20 amino acids, comparing paired sub-proteomes of model vertebrate organisms. *(a) Phylogenetic tree of representative organisms considered in this study: Human (Homo sapiens, chimpanzee (Pan troglodytes), mouse (Mus musculus), chicken (Gallus gallus), African clawed frog (Xenopus laevis), and zebrafish (Danio rerio). (b) The distribution of ΔC(Ala) for 2353 homologous proteins between zebrafish and human. The difference in the occurrence of alanine of an individual protein pair, ΔCi(Ala) is put into one of 9 bins from -1% to 1% with an increment of 0.25%. Differences < -1% or >1% are added to the lowest and highest bin, respectively. This is done for all protein pairs, leading to the histogram plotted for Ala, as shown. The distributions of ΔC for the other amino acids and other model organisms are given over the protein pairs in (c). The y-axis value is the extent of the difference in absolute amino acid content (bin-ranges, same as x-axis in b). The x-axis indicates the 20 amino acids. [A(la) is circled for zebrafish as this column corresponds to Fig. 1b]. The color bar indicates the fraction (%) of proteins in the comparison that have a content difference in the bins (see the unscaled y-axis in b and fractions given in figure). (d) Density distribution of ΔC for alanine, leucine and serine between a model organism and human (data for chimpanzee is not shown for better visualization). Note that except for the extremes, the binned range is -1 to +1%, showing that a significant number of human/zebrafish compared proteins are in the extreme bins in the figure panels b-d).*

The remaining 15 amino acids are not as well conserved as the above mentioned 5 amino acids. For example, the content in glutamine varies, however, there is not a significant shift in amino acid utilization in either direction. By contrast, we see a shift for all the other 14 amino acids, although to different extents. ii) There is a trend of exchange of aspartic acid with glutamic acid; and exchange of lysine with arginine between human and other organisms, as already known many years ago from the tendency to usually tolerate/find advantages in substitutions that preserve amino acid similarity.[24] However, rather noticeably, the proteins of African clawed frog and zebrafish have less positively charged amino acids but more negatively charged amino acids than the homologous human proteins (Fig.1c and Fig. S2). Supposing a charge of +1 for lysine and arginine and -1 for aspartic and glutamic acid, on average the African clawed frog has a reduced positive charge by -0.14 units per protein and more negative charge by -0.43 per protein in comparison to the human protein homologs. For zebrafish, the values are -0.28 (less basic amino acids) and -0.66 (more acidic amino acids). At the extreme it is known that most proteins in bacteria have a lower isoelectric point, than human proteins – allowing the easy purification of human proteins when they are expressed recombinantly in the prokaryote. The reasons for these shifts are not fully understood but may be related to issues with the long-term stability of polypeptide chains around acidic residues. In contrast to the observed difference of -0.57 and -0.94 in net charge per protein for African clawed frog and zebrafish, the difference in average net charge between chimpanzee, mouse, chicken and human is +0.16, -0.11 and +0.32 charge units per protein respectively (Fig. S2).

Remarkably, we notice an obvious difference between warm-blooded animals (human, chimpanzee mouse and chicken), and cold-blooded animals (African clawed frog and zebrafish) with respect to the extent of variation in polar/non-polar amino acids. iii) A great number of human proteins have a higher content of alanine, glycine, leucine and proline, but a lower content of isoleucine and valine (Fig. 1c, d). Such changes may be related to the dynamics of the polypeptide chain, with glycine, alanine and in a sense proline, providing greater flexibility whereas isoleucine and valine reduce flexibility due to β-branched sidechains. iv) There is an increase in the relative proportion (again in terms of occurrence in the population of paired proteins) of polar amino acids (serine, threonine, asparagine) as well as methionine in African clawed frog and zebrafish (Fig. 1c, d) compared to the warm-blooded organisms. A possible rationale, increased solubility, for this trend is investigated below.

In general, the conservation as well as the shift in the occurrence of amino acid across homologous proteins reflect the complexity of the organism and the demands that its behavior and ultimately environment places on it. The latter comprises multiple factors, for example, the availability of food and nutrition, the metabolic level of an organism, and the difficulty of the biosynthesis of an amino acid as well as the stability of nucleotides in DNA/RNA. Thus, we would not expect that a single mechanism can explain all the shifts in amino acid occurrence, as summarized above in points i to iv. However, we notice a distinct difference between warm-blood and cold-blood organisms considering the utilization of multiple polar/non-polar amino acids. Next, we examine whether there is a correlation of amino acid changes with the obvious difference in body temperatures between ward-blooded and cold-blooded organisms.

Protein Association Modulated by Quenching/Tempering of Protein Dynamics

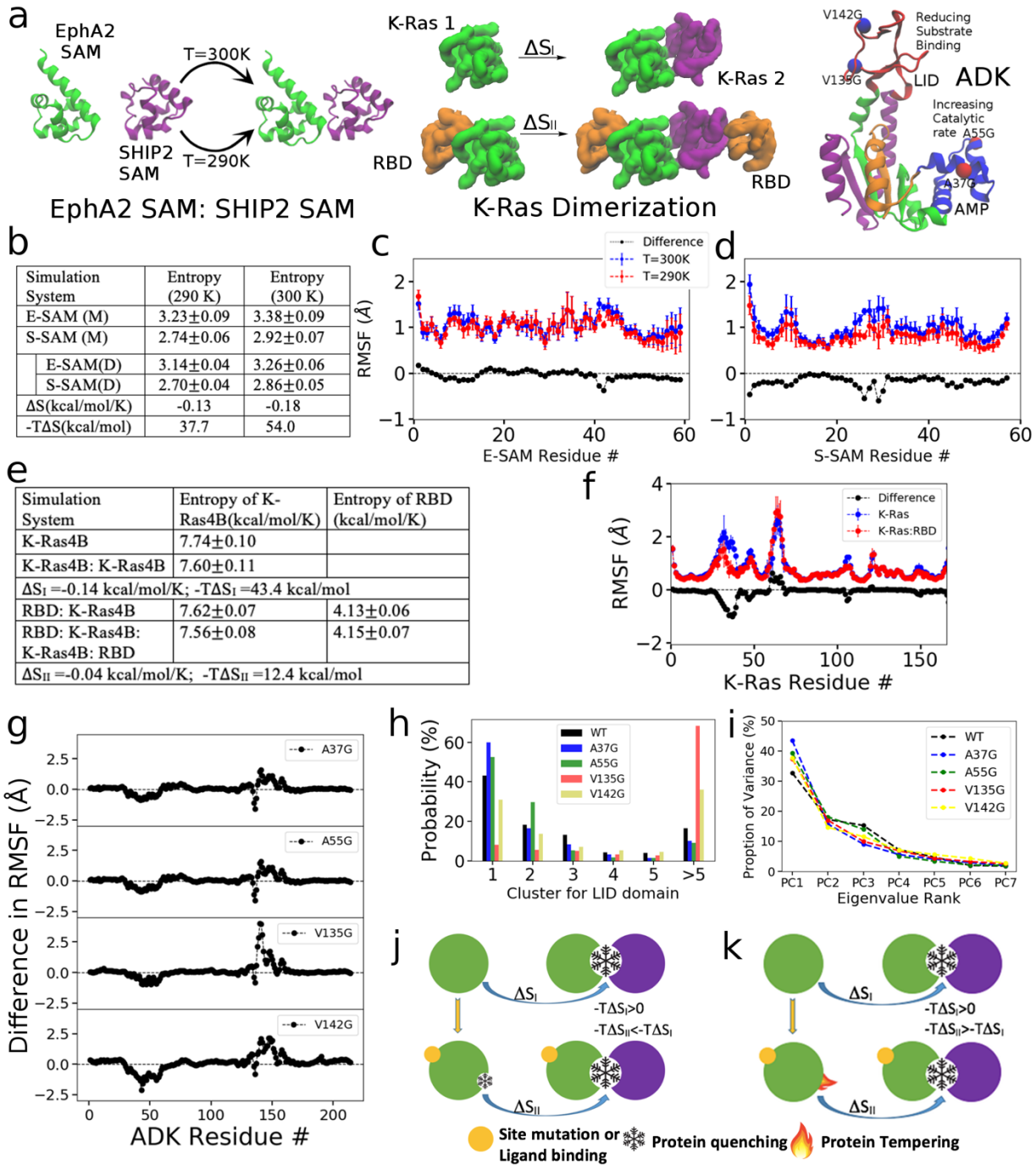
A cold-blooded organism, for example the African clawed frog, has a wide tolerance for changes in body temperature ranging from 10 to 30°C; for zebrafish, a range of body temperature between 16°C and 30°C is suitable for it to swim.[25] Both are lower than the temperature of humans (median 37°C), other mammals (median 38.1°C) and birds (median 40°C). The apparent drop in body temperature for cold-blooded organisms would attenuate, or "quench" many protein fluctuations, typically leading to enhanced protein association. As an example, we compare the variation of internal protein entropy for the EphA2 SAM: SHIP2 SAM association at two different temperatures (300 K and 290 K; Fig. 2a left). All atom molecular dynamics simulations were run for this system at the two temperatures, separately with 4 repeats, totaling 12 μs (Table S2a). We calculated the entropy of each SAM domain, individually and in complex from the trajectories as described in the Methods section. At T=300 K, the penalty due to the entropy decrease on SAM domain association, $-T\Delta S$, is 54.0 kcal/mol. The entropy penalty at the lower temperature of 290 K is 37.7 kcal/mol (Fig. 2b). This is because the decrease of the temperature already decreases some of the local fluctuation of residues (reduced entropy in the unbound state) for both E-SAM and S-SAM (Fig. 2c, d and Fig. S3). Therefore, at the lower temperature, the reduced protein flexibility goes hand in hand with a diminished requirement for a change in internal dynamics upon binding and results in a reduced entropy penalty for protein association. This lesser entropy penalty at low

temperature is typically counteracted by an increase in protein-water interactions – thus increasing the enthalpy penalties for removing waters from the protein-protein interacting surfaces.[26,27] Globally, the equilibrium of proteins in bound and unbound states depends on the free energy difference between the two states ($e^{-\Delta G/kT}$). Therefore, a drop of temperature would overall favor the protein bound state as long as $\Delta G<0$. Decreased temperature is also expected to slow down the interaction kinetics between and within proteins, which is an unwanted effect for enzyme reactions and cellular signaling processes.

We speculate, that the cold-blooded organisms may select amino acids that disrupt protein-protein interfaces (PPI) to resist an over-sticky protein association; however, the evolution rate of PPI is 3-orders of level slower than the rate of evolution of an average protein sequence.[28,29] Therefore, except by directly disrupting the PPI, adaptation may select mutations (outside of the PPI) that either elevate protein internal dynamics or increase the affinity of protein for water --- both would offset the entropy loss caused by the drop of temperature. Next, we further investigate the principle of protein association modulated by quenching or tempering of protein internal dynamics, before we return to the issue of solvation.

The role of amino acid residue mutations/ligand binding, both in regions away from a proteins' functional site on protein dynamic allostery, association and function have been well demonstrated for several proteins.[30-34] Recently Mattos and colleagues discovered a process of K-Ras4B dimerization stimulated by C-Raf Ras Binding Domain (herein RBD).[35] Here, similarly to the SAM domain association, above, we compare the entropy change. We consider two processes for the monomer-homodimer transition: K-Ras4B homodimerization by itself (Process I, ΔS_I) and K-Ras4B homodimer formation following K-Ras binding of the effector protein, RBD (Process II, ΔS_{II}) (Fig. 2a middle). We carried out molecular dynamics simulations on each set of systems with at least 4 repeats, giving a total simulation time of 13 μs (Table S2b). Given that the protein-protein interaction interfaces in K-Ras4B dimerization are the same, ideally, the enthalpy contribution in the free energy change associated with Process I and Process II are canceled. Therefore, the entropy difference ($\Delta\Delta S = \Delta S_{II}-\Delta S_I$) between the two processes reflects at least a dominant part of the free energy difference. As shown in Fig. 2e, the formation of K-Ras dimer is at the cost of conformational entropy (ΔS_I= -0.14 kcal/mol/K, $-T\Delta S_I$= 43.4 kcal/mol); however, with effector (Raf RBD) binding, the entropy penalty of K-Ras4B is lessened to -0.06 kJ/mol/K. Plus a small gain of entropy of 0.02 kcal/mol/K for the RBD upon binding, results in a net loss of entropy for Process II of -0.04 kcal/mol/K ($-T\Delta S_{II}$ = 12.4 kcal/mol). Fig. 2f shows that residues at the K-Ras dimerization interface have already gone through a quenching in their dynamics during RBD binding, therefore there is a diminished need for a further rigidification in the transformation from the K-Ras: RBD 1:1 complex to a 2:2 complex upon dimerization via K-Ras (Fig. S4), and consequently a reduced entropy penalty.

Considering the opposite: an increase, also known as tempering, of protein internal dynamics caused by ligand binding/site mutations is predicted to weaken protein-protein association. Inspired by an experimental study by Hilser and colleagues,[30] we carried out a set of simulations (in duplicate) totaling 5 μs for wild Adenosine kinase (ADK) and for four ADK mutants in their inactive state (Fig. 2a right, and Table S2c). The mutation of V135G or V142G in the LID domain increases the mainchain fluctuations (RMSF) of the protein (Fig. 2g and Fig. S5a), and considerably increases the conformational flexibility of the LID domain (Fig. 2h). The increased dynamics of the LID domain weakens the LID domain's substrate binding as reflected by an increase of Km in kinetic experiments.[30] Both set of mutations cause non-local perturbations. This is achieved by connecting residues between the LID and AMP domains via optimal allosteric pathways in the protein (Fig. S5b-c). The mutation of Ala37 to Gly or Ala55 to Gly in the AMP domain have a lesser influence on the protein's flexibility (Fig. 2g and Fig. S5d). These mutations have less influence on substrate binding, but largely increase the catalytic rates.[30] The collective mode analysis indicates that the proportion of soft modes, PC1 and PC2, are increased, in particularly for mutations Ala37 to Gly and Ala55 to Gly (Fig. 2i and Fig. S5e). PC1 reports on relative motions that approach the center of the AMP and LID domains, similar to those that appear in a catalytic process; while PC2 indicates crisscross motions (Fig. S5f). These hinge motions and relevant collective motions likely speed up the transition rates in a catalytic reaction especially for A37G and A55G mutations as observed experimentally.[30]



Simulation System	Entropy (290 K)	Entropy (300 K)
E-SAM (M)	3.23±0.09	3.38±0.09
S-SAM (M)	2.74±0.06	2.92±0.07
E-SAM(D)	3.14±0.04	3.26±0.06
S-SAM(D)	2.70±0.04	2.86±0.05
ΔS(kcal/mol/K)	-0.13	-0.18
-TΔS(kcal/mol)	37.7	54.0

Simulation System	Entropy of K-Ras4B(kcal/mol/K)	Entropy of RBD (kcal/mol/K)
K-Ras4B	7.74±0.10	
K-Ras4B: K-Ras4B	7.60±0.11	
ΔS_I =-0.14 kcal/mol/K; $-T\Delta S_I$ =43.4 kcal/mol		
RBD: K-Ras4B	7.62±0.07	4.13±0.06
RBD: K-Ras4B: K-Ras4B: RBD	7.56±0.08	4.15±0.07
ΔS_{II} =-0.04 kcal/mol/K; $-T\Delta S_{II}$ =12.4 kcal/mol		

Figure 2: Protein association is modulated by quenching or tempering protein internal dynamics. *(a) Dynamic allostery systems considered in this study plotted as mainchain ribbon or space filling. (Left) EphA2 SAM (green): SHIP2 SAM (purple) heterodimer formation at two different temperatures (300 K and 290 K); (Middle) The monomer-homodimer transition of K-Ras4B (green and purple) by itself (Process I) and K-Ras4B: RBD (yellow) as a 2:2 complex (Process II); (right) ADK structures in the open state. Segments N-terminus: 1-29; AMP: 30-73; Hinge: 74-121; LID: 122-159; C-terminus: 160-214 are shown in different colors. (b) Entropy change in the E-SAM: S-SAM heterodimer formation at 300 K and 290 K. (c, d) Quenching of protein internal dynamics for E-SAM and S-ASM due to the decrease in temperature. (e) Entropy change in a monomer-homodimer transition of K-Ras4B by itself (Process I) and K-Ras4B: RBD as a complex (process II). (f) Quenching of protein internal dynamics of K-Ras4B caused by RBD association. (g) Protein conformational flexibility for wild type ADK and ADK mutants. Plot of the difference in RMSF between wild type ADK and ADK mutants. (h) Conformational clustering for ADK LID domain. (i) Principle Component Analysis (PCA) of ADK and ADK mutants: the proportion of variance versus eigenvalue rank. (j, k) A schematic model of quenching-stimulated and tempering-repressed protein association.*

Entropy and Solvation Energy Over Sub-proteomes of Human and Model Organisms

Fig. 2j and 2k summarize the mechanism of protein association modulated by protein internal dynamics - a quenching or tempering mechanism: one diminishes, while the other one increases the entropy penalty due to internal protein dynamics in protein association. Based on this mechanism, in cold-blooded organisms, amino acid mutations that elevate protein dynamics should offset the loss of entropy caused by a drop of body temperature.[30] Here, we perform a sequence-based analysis of the polypeptide chain entropy and solvation free energy for the sets of proteins examined in Fig. 1. These calculations are based on available experimental values of amino acid backbone entropy and solvation free energy (Fig. 3 and Fig. S6).[36,37] Surprisingly, we find --- the African clawed frog and zebrafish have an overall lower backbone entropy in comparison to warm-blooded species - chicken, mouse, chimpanzee and human (Fig. 3a). Therefore, the frog and fish did not accumulate residues that largely increase protein dynamics. This is in accord with the amino acid constituents, as cold-blooded organisms have a relatively high abundance of Ile and Val (lowest backbone entropy) and low content of Ala and Gly (highest backbone entropy).

Importantly, the fish and frog proteins have an increase in affinity of the protein toward water (Fig. 3a). The scattering of the entropy/enthalpy values for 2353 homologous proteins between zebrafish and human indicates a general trend that more zebrafish proteins have an increased solvation in combination with a decreased protein entropy (Fig. 3b). The solvation increase is in accord with the increased occurrence of polar amino acids (Ser, Thr, Asn) in cold-blooded organisms. This increase of polar amino acids is a result of their tendency to replace Ala and Gly (Fig. 3c). The increase in the occurrence of methionine also contributes to increased solvation, as methionine frequently replaces leucine (Fig. 3c). The latter is more hydrophobic than the methionine. Though not included in the analysis, other contributors to the increased solvation in zebrafish proteins should include the greater number of charged (acidic) amino acids, as zebrafish has more of them (point ii, above).

So far we have made the connection between the shift in amino acid content and the thermoadaptation principle between cold-blooded and warm-blooded organisms. The sequence-based analysis gives us entropy/enthalpy information over a large number of proteins. However, we have excluded proline in the estimation of entropy, and four additional charged amino acids (Lys, Arg, Asp, Glu) in the calculation of solvation free energy due to the unavailability of experimental data.[36,37] In order to gain further evidence at the structure level, of the 2353 homologous proteins between human and zebrafish, we focused on 166 proteins with available and relatively stable structures and performed molecular dynamics simulations for both the human and homologous zebrafish proteins (Fig. 3 and Table S3). We firstly look at mutation sites in the 166 proteins: 96.4% of the identified site mutations are found at exactly the water-interacting protein surface (0-4 Å to water); 3.6% are at the interior of protein (>4 Å) (Fig. 3d). The fact that a larger number of mutations occur at the protein surface compared to its interior is well known to be associated with the larger average effect of the latter on protein solvation and themodynamic stability. We observed overall consistent results with those of the sequence-based analysis of protein solvation energy. In the collection of 166 proteins, there is not an obvious difference in backbone fluctuations (RMSF of Cα atoms) between zebrafish and human proteins (Fig. 3e). However, the zebrafish proteins have enhanced Poisson Boltzmann electrostatic interactions with the solvent as well as a larger solvent accessible surface area on average (Fig. 3e). The protein's enhanced affinity for water is partially caused by the enhanced enthalpy for hydrogen bond formation between protein and water (Fig .3e).

Together, both our sequence-based and structure-based energy analyses support the finding that the zebrafish proteins have an overall enhanced degree of solvation. The solvation increase would lead to a bigger loss of protein solvation energy in protein association. This is helpful for preventing an over-sticky association caused by low temperature during enzyme catalysis or cellular signaling processes.

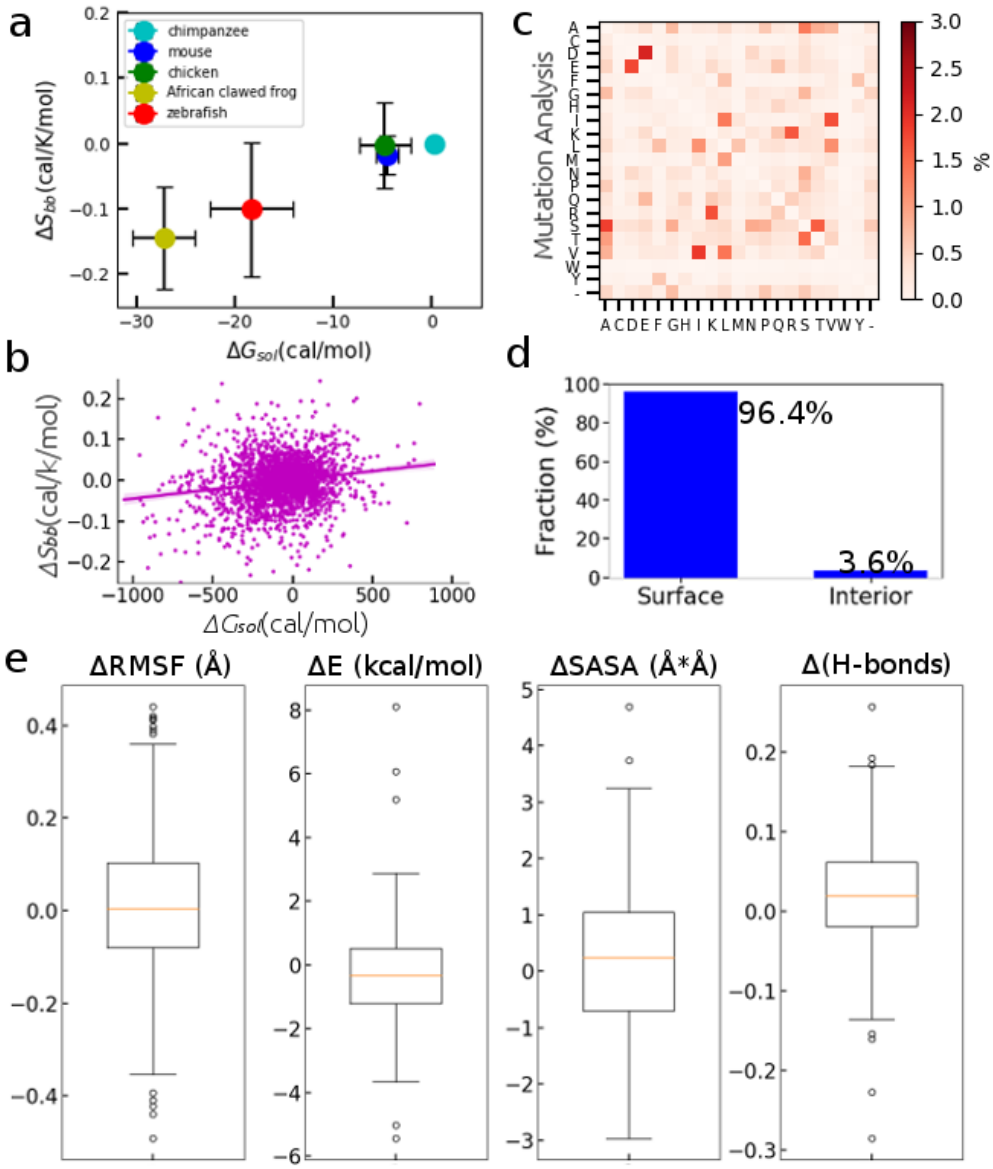


Figure 3: Energetics and entropy analysis over a sub-proteome between human and model organisms. *(a) Estimation of polypeptide chain entropy and solvation free energy for homologous proteins. The average values and standard errors of the sequence based difference in entropy and solvation free energy between protein homologous pairs from a model organism and humans are plotted. The entropy and solvation energy is averaged over the sequence length of a protein. (b) The fits and scatter of ΔS and ΔG for 2353 homologous proteins between zebrafish and human. (c) Amino acid mutation analysis for 2353 homologous proteins between zebrafish and human. (d) Relative distribution of positions of site mutations at the protein surface and interior based on the structure of 166 paired proteins between zebrafish and human. (e) Energetic analysis for the simulations of 166 paired proteins between zebrafish and human. Boxplots of ΔRMSF (Root mean squared fluctuation), ΔE (Poisson Boltzmann electrostatic interaction), ΔSASA (Solvent available surface area), and ΔH-bonds (number of hydrogen bonds between solvent and protein). All values in e) are scaled by the sequence length, i.e. are the average change per amino acid.*

Elevated Protein Dynamics and Protein Solvation in L-Lactate dehydrogenase A chain (LDHA) for Antarctic Fish in Extreme Cold

Lastly, we present a case study of protein L-Lactate dehydrogenase A chain (LDHA) of Antarctic fishes (scientific name, notothenioid). In this situation, the protein needs to adapt to an extreme cold of -2 to 0 ℃. The protein was found to have a decreased substrate binding affinity but an increased rate of catalysis at near-zero temperature.[1] Looking at the sequence, the LDHA of Antarctic fishes have an overall consistent occurrence of amino acids, as the sub-proteome level results for zebrafish, i.e., the higher abundance of Ser, Thr, Met, Val, and lower content of Ala and Leu (Fig. 4a). Noticeably, the Antarctic fishes have more Gly residues, opposite

from the general direction of the shift of Gly occurance comparing other model organisms (Fig. 1). We compare the dynamics of a monomeric LDHA of three different species with available protein structures: Mackerel icefish (a notothenioid fish), living in waters of -2 to 0 °C, Common carp living in normal water (ranging from 3 to 35 ℃) and lastly Human (37 ℃) (Fig. S7a). Molecular simulations for LDHA molecule of human, common carp and Mackerel icefish were performed in triplicate for 500 ns at three different temperatures - 310K (37℃), 290K (17℃) and 273K (0℃). The LDHA protein is composed of several different sub-domains or lobes: briefly, the N-terminal-, catalytic and the C-terminal lobes (Fig. 4b) and three regions encompassing the active site as clefts 1, 2 and 3. As expected, the Mackerel icefish and Common carp proteins have increased solvent accessible surface area (Fig. 4c), which is consistent with the general trend (Fig. 3). However, the Mackerel icefish also has a very high level of protein flexibility especially in the cleft 2 region (Fig. 4d and Fig. S7b). The clustering analysis over each of protein trajectories shows an increased population of conformations of LDHA away from the native structure for the Mackerel icefish (Fig. 4e), indicating that the LDHA of Mackerel icefish explores a broader conformation space, especially at lower temperature (17 and 0 ℃; Fig. 4h and Fig. S8a). This is in accord with the sequence difference that exists largely in cleft 2, where the Common carp have one more glycine than the Human protein, while the Mackerel icefish have added two more glycines than the Common carp near the same location (Fig. S8b). These glycines are responsible for the gain of flexibility in the Markerel icefish protein. Glycine has no side chain and is able to go through dihedral angles changes not allowed by the other residues. Mutations of other residues to glycine have been suggested to contribute to cold adaptation for enzymes such as Adenosine kinase and alkaline phosphatase.[30,38] In summary, as a general trend, the proteins from cold-blooded organisms have increased solvation by the occurrence of a larger number of polar amino acids on their surface. However, in order to adapt to extreme environmental conditions, certainly the proteins of an organism, here, -the Antarctic fish, can make use of additional mechanisms for optimal adaptation.

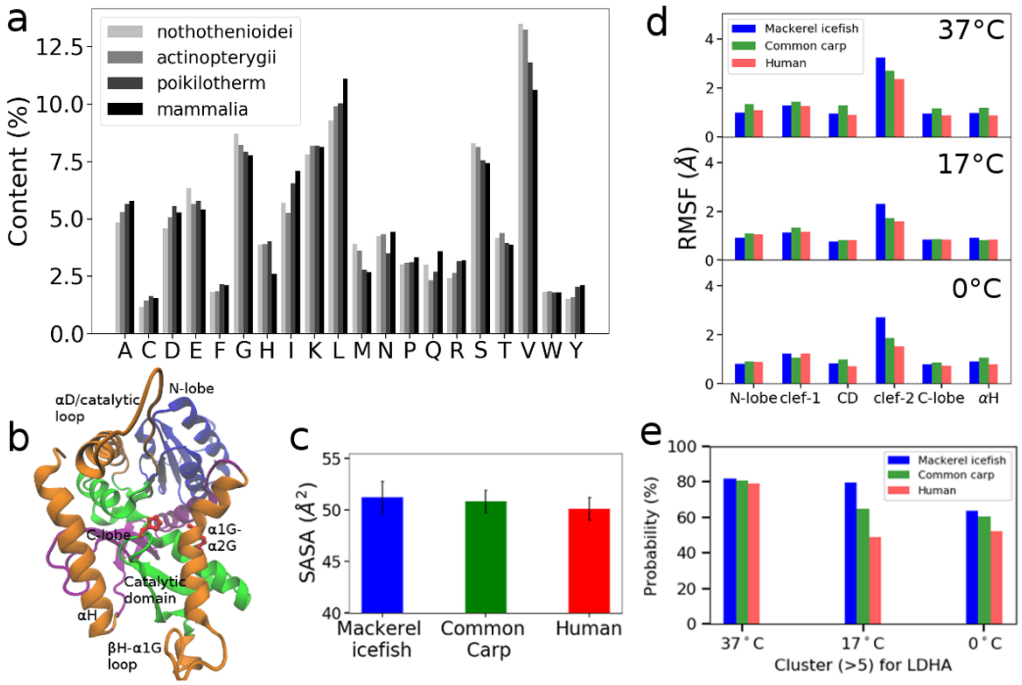


Figure 4: Cold adaptation of L-Lactate dehydrogenase A chain (LDHA) of Antarctic fishes in extreme cold*. (a) Comparing the occurrence of different amino acids for Antarctic fishes (nothothenioidei), other fishes (actinopterygii), cold-blooded organisms excluding fishes (poikilotherm) and mammalia. (b) LDHA (Mackerel icefish) structures and domain partitioning. N-lobe: 1-94 (blue); Catalytic subdomain: 129-210 (green); C-lobe: 245 to 307 (purple). Three secondary structural elements encompassing the active site are marked in green: αD/catalytic loop (95-128); βH –α1G loop and α1G-α2G (211-244); and αH (308 to 331). His 192 and Arg167 are space-filled to indicate the location of the active site. The starting residues of 1-19 are excluded for the analysis due to high flexibility. (c) Surface accessible surface area of LDHA of three different organisms (Mackerel icefish, Common carp and Human). (d) RMSF (average by amino acid residues in select regions) of LDHA of three different organisms at three different temperatures (37, 17 and 0 ℃). The values are averaged over residues in each of the six subdomains. (h) Cluster analysis of LDHA from three different organisms at different simulation temperatures.*

Perspective and Summary

Understanding the evolution of proteins at the molecular level broadens our insight into the origin of species and the biological adaptation to their environment. The principle of natural evolution also gives insights into directed evolution in laboratory strategies for the optimization of functional enzymes and macromolecules to address challenges in health management and green chemistry. Instances of protein molecular adaptation have been recognized for several proteins, which have been studied, across all three domains of life - animal,[1-6] plants as well as microorganisms. For microorganisms noticeably examples include the adaptation of thermobacteria and recently of the bacterial CRISPER-Cas system,[39,40] as well as the mutational tolerance of virus proteins at fever-like temperature.[41] For plants, the impact of global warming on proteins, nutrient and water availability for crops are attracting attention of scientists worldwide.[42] Increasingly it is recognized that properties such a dynamic/allosteric protein network pathways and protein solubility are involved in the evolution of diseases such as cancers and neurological disorders.[43,44] In addition to the role of mutations in protein-coding genes in their adaptation, the regulation of protein expression levels by mutations in non-coding regions reflects another important way for an increased fitness in changed environments.[45]

While instances of individual protein adaptation have been reported, it is unclear whether there is a general principle. In this study, we investigated the conservation and shifts in the occurrence of amino acids over pairwise homologous sequences for sub-proteomes between human and multiple vertebrate organisms. We identified highly conservative amino acids as tryptophan, phenylalanine, tyrosine, histidine and cysteine. Also, we discovered a distinction between proteins from warm-blooded and cold-blooded species. Specifically, we saw an increase in the occurrence of polar amino acid (Ser, Asn, Thr) as well as Ile, Val, Met in fish and frog proteins; and a change in the occurrence of Leu, Ala, Gly, and Pro when comparing proteins from lower vertebrate species, such as fishes, with homolog proteins from humans, mammals and birds. We then associated these changes with a modulation of protein association by changes in protein internal dynamics, a quenching or tempering mechanism---the former diminishes while the latter mechanism increases the entropy penalty typically seen in protein association. In order to counteract an undesirable increased tendency towards protein association caused by low body temperature, fish and frog proteins considerably increase their affinity to water by having a greater number of polar amino acids, methionine and acidic amino acids. By contrast, human proteins have a high-level protein backbone flexibility but also higher thermodynamic stability, which are likely optimal for the high metabolic levels but also longer life-span of humans. For Antarctic fish, the L-Lactate dehydrogenase A chain (LDHA) protein has both an elevated level of protein dynamics and solvation, in order to adapt to the extreme cold of the Antarctic ocean. To end this paper, we would emphasize that it is unlikely that a common mechanism can explain all the conservation and the changes in amino acid occurrence as discovered in this study. More studies from a wider perspective of physicists, biologists and naturalists will be needed to give deeper insights into the underlying mechanisms. To "get the ball rolling", we propose a principle of amino acid occurrence change related to vertebrate thermoadaptation which is mediated by the role of amino acids in determining protein dynamic allostery and protein solvation.

Acknowledgements

We thank for Dr. Carla Mattos for providing the structure of the dimeric Ras protein and information on C-Raf RBD-stimulus Ras dimerization. This work is supported by a NIH R01 grant from the National Eye Institute R01EY029169 and previous grants from NIGMS (R01GM073071 and R01GM092851) to the Buck lab. The simulations were run at the Ohio Supercomputer Center (OSC), as well as local computing resource in the core facility for Advanced Research Computing at Case Western Reserve University. Anton Computer time was provided by the Pittsburgh Supercomputing Center (PSC) through Grant R01GM116961 from the National Institutes of Health.

Author Contribution

Z.L. and M.B designed the studies, interpreted the data and wrote the manuscript. Z.L. performed the sequence/evolution analysis and computational modeling.

Competing Interests: The authors declare no competing interests.

Methods

Comparative sequence analysis over sub-proteomes of human and model organisms: Protein sequences between Human (Homo sapiens) and multiple model vertebrate organisms --- chimpanzee (Pan troglodytes), mouse (Mus musculus), chicken (Gallus gallus), African clawed frog (Xenopus laevis), and zebrafish (Danio rerio) were compared. All sequences of proteins of a species (including both "reviewed" and "un-reviewed" items) were downloaded from Uniprot database. Based on the Uniprot database, the total number of unique protein-coding genes of human is 20621 (gene counts). For chimpanzee, mouse, chicken, African clawed frog and zebrafish, the number is 23051, 21986, 18117, 43236, 25703 respectively (See Table S1). However, for the African clawed frog and zebrafish, they have much more individual-species specific genes than mammals or chicken. Therefore, the gene count of an individual zebrafish or African clawed frog should be much lesser.[46] With the downloaded sequences, we firstly analyzed the homologous proteins between human and a model organism by matching the name of individual protein species. We identified 4136, 15674, 2681, 2016, 3383 homologous proteins between human and chimpanzee, mouse, chicken, African clawed frog, zebrafish respectively. It should be realized that these numbers are not the real total number of homologous proteins between human and a model organism. Many of the protein sequences of a model organism were not well reviewed on Uniprot websever and have not been linked to a specific protein species, for example for those proteins marked and temporally named as "uncharacterized protein". Only the number of homologous proteins between mouse and human should be close to the real number, as mouse has widely available reviewed protein sequences due to the extensive use of mice in the laboratory. Sequences of homologous proteins were pair-wise compared between the human and its homolog in each of the model organisms. We excluded the homologous proteins with a difference in sequence length >5%. In this way, we focused on homologous proteins with a high similarity in their sequence length. As a result of this filter, we analyzed 3607, 14046, 1921, 1565, 2353 homologous protein pairs between human and chimpanzee, mouse, chicken, African clawed frog, and zebrafish (see Table S1). For each protein, we compare the difference in absolute content for each one of the 20 amino acids (AA) in a protein pair/dividing by its length (in case the length is different, by the average of the two). Then the results from all other protein pairs are calculated similarly and collected in 9 different ranges (or bins) from -1% to 1% with an increment of 0.25%. Differences < -1% or >1% are added to the lowest or highest bin, respectively. The distribution of $\Delta C(AA)$ that arises is then over all the protein pairs (Fig. S1). The y—axis is the number of proteins in each bin for a given amino acid comparison, but could also be rescaled by dividing by the number of proteins, giving a % scale, making an easy comparison possible between amino acids and between the species comparisons. This % scale is the color bar plotted in Fig. 1c.

Protein association modulated by quenching/tempering of protein dynamics: We performed molecular dynamics simulation for monomeric SAM and dimeric SAM at two different temperatures of 300 K and 290 K. Each simulation system was performed for 500 ns in quadruplicate. We simulated monomeric K-Ras4B and dimeric K-Ras4B, as well as K-Ras4B in complex with C-Raf RBD and a 2:2 RBD-K-Ras4B: K-Ras4B-RBD complex in solution. Each simulation was performed for 500 ns at 310 K with four to eight runs starting with different velocity assignments, in order to obtain standard errors. For Adenosine kinase (ADK), we carried out simulations in duplicate for wild ADK and four ADK mutants in their inactive open state. Each simulation was performed for 500 ns at 310 K. Trajectories were saved at every 10 ps. The root mean squared fluctuation (RMSF), clustering, principal mode analysis and dynamics network analysis were analyzed with standard scripts in VMD. The conformational entropy of the proteins was calculated with the quasi-harmonic approximation method developed by Andricioaei and Karplus.[47] The quasi-harmonic approximation method estimates the upper limit of protein conformational entropy and is implemented in Wordom[48]. Typically, the method overestimates the entropy value by 2-fold or more.[49] Since the entropy penalty was compared in a symmetric thermodynamic cycle (Process I and Process II), this only affects the absolute value but not the qualitative trends. Non-hydrogen atoms were selected and the entropy was calculated over trajectory intervals 100-200ns, 200-300ns, 300-400ns, 400-500ns and averaged.

Sequence-based energetic and entropy analysis: The average entropy of an amino acid in a polypeptide chain was obtained from Ref. 36; and the solvation free energy of an isolated amino acid side chain analogue was obtained from Ref. 37. The polypeptide chain entropy of a protein was estimated by adding up the entropy of the constituent amino acids. Proline was not included in the calculation of backbone entropy due to the unavailability of a value. The entropy was further averaged over the length of protein sequence (S_{bb} entropy per amino acid). The apparent entropy difference of a specific human and model organism protein pair was calculated as ΔS_{bb}= S_{bb}(model organism) – S_{bb}(human). The solvation free energy of protein was estimated by adding up by the solvation free energy of the isolated amino acids. This estimates the upper limit of a protein solvation free energy. The difference in protein solvation free energy between a protein pair was calculated as ΔG_{sol}= G_{sol}(model organism) – G_{sol}(human). Charged residues were not included in the calculation of solvation free energy due to the unavailability of experimental data for these amino acids. The distributions of ΔS ($\Delta S = \{\Delta S_1, \Delta S_2, \ldots \Delta S_N\}$) and ΔG ($\Delta G=\{\Delta G_1, \Delta G_2, \ldots \Delta G_N\}$) over the N homologous proteins between human and model organisms, as well the average value and standard variation of ΔS and ΔG are in plotted Fig. 3 and Fig. S6.

Molecular modeling of 166 homologous proteins between human and zebrafish: We analyzed 2353 human proteins and as many of their homologs we could find in zebrafish. Sequence alignment between zebrafish and human proteins were carried out with bioPython with Clustalw module. Of 2353 proteins, 644 have available PDB structures for the human proteins. Selectively, we choose structures that have a resolution better than 5 Å, cover at least 80% protein sequence, and have overall stable structures without too flexible regions such as long loops. This reduced the set to 166 soluble proteins. A full list of proteins studied is given in Table S3. Modeller was used to build the initial structure of each protein, adding missing residues and short loops.[50] For human proteins, N-terminal and C-terminal residues were added automatically if the missing segments are less than 5 residues long in a PDB. If there were more than 5 residues missing, the N- or the C-terminal residue of a protein was kept the same as in a PDB structure. With the high sequence similarity, the structure of the homologous zebrafish protein was constructed based on the corresponding human protein PDB with Modeller. The modeled protein systems were subjected to standard molecular dynamics simulations each of 100 ns. Disulfide bonds were patched with module DISU in the CHARMM force field. The results were analyzed by comparing the properties of the protein pairs in a manner similar to the sequence-based energetic analysis. Specifically, ΔRMSF (Root mean squared fluctuation), ΔE (Poisson Boltzmann electrostatic interaction), ΔSASA (Solvent available surface area), and ΔH-bonds (number of hydrogen bonds between solvent and protein) were calculated. For instance, the RMSF of a zebrafish protein was averaged over the sequence length, and was then compared to the value of a human protein: $\Delta RMSF_i$= $RMSF_i$ (zebrafish) – $RMSF_i$(human). The distribution of ΔRMSF ($\Delta RMSF = \{\Delta RMSF_1, \Delta RMSF_2, \ldots \Delta RMSF_{166}\}$) for the 166 simulated proteins were plotted with a boxplot in Fig. 3.

Case study of L-Lactate dehydrogenase A chain (LDHA): In the sequence analysis of LDHA, 17 Antarctic fishes were considered for nothothenioidei (Antarctic spiny plunderfish, Patagonian toothfish, Antarctic cod, Humped rockcod, Charcot's dragonfish, Yellowbelly rockcod, Gaudy notothen, Mackerel icefish, Emerald rockcod, Blackfin icefish, Ocellated icefish, Rockcod, Bald rockcod, Patagonian blennie, Maori cod, Black southern cod, Antarctic eelpout); 10 fishes (excluding the Antarctic fish) were used for actinopterygii (Pacific barracuda, Pelican barracuda, Lucas barracuda, Zebrafish, Killifish, Common carp, Long-jawed mudsucker, Shortjaw mudsucker, Blackeye goby, Spiny dogfish); Apart from the fish, 8 species were used for cold-blooded organisms (African claweded frog, Axolotl, Eastern fence lizard, Ball python, Florida scrub lizard, Chinese soft-shelled turtle, Red-eared slider turtle, American alligator) and 10 species for mammals (Gray short-tailed opossum, Rabbit, Pig, Wild yak, Bovine, Crab-eating macaque, Sumatran orangutan, Chimpanzee, Mouse, Rat, Human). Our standard simulation procedure was applied to the LDHA protein. PDB entry 4ajp, 1v6a and 2v65 were used for human, common carp and Mackerel icefish protein respectively. Each simulation system was performed in triplicate for 500 ns at three different temperatures - 310K (37°C), 290K (17°C) and 273K (0°C).

General simulation parameters for all modeling: All proteins were solvated by TIP3P water with 150 mM NaCl; except on Anton (which uses its own code) all simulations were performed with the NAMD/2.12 package.[51] For simulations of 100 ns in length, they were performed on the GPU cluster at Case Western Reverse

University or at the Ohio supercomputer center (OSC). A time step of 2 fs was employed. The SHAKE algorithm was applied for all covalent bonds to hydrogen. A Langevin thermostat of 310 K and a semi-isotropic Langevin scheme at 1 bar was used. Simulations on K-Ras4B, the SAM domains, ADK and LDHA. were performed on the Anton 2 supercomputer specialized for MD simulation.[52] The CHAMRM36m force field was used in all simulations.[53] The van der Waals (vdW) potential was cut off at 12 Å and smoothly shifted to zero between 10 and 12 Å. The long-range electrostatic interactions were calculated with the Particle-Mesh Ewald (PME) method.

Reference

1. Fields, P. A., and G. N. Somero. "Hot Spots in Cold Adaptation: Localized Increases in Conformational Flexibility in Lactate Dehydrogenase A4 Orthologs of Antarctic Notothenioid Fishes." *Proceedings of the National Academy of Sciences* 95, no. 19 (1998): 11476-1481. doi:10.1073/pnas.95.19.11476.
2. Projecto-Garcia, J., C. Natarajan, H. Moriyama, R. E. Weber, A. Fago, Z. A. Cheviron, R. Dudley, J. A. Mcguire, C. C. Witt, and J. F. Storz. "Repeated Elevational Transitions in Hemoglobin Function during the Evolution of Andean Hummingbirds." *Proceedings of the National Academy of Sciences* 110, no. 51 (2013): 20669-0674. doi:10.1073/pnas.1315456110.
3. Hill, Jason, Erik D. Enbody, Mats E. Pettersson, C. Grace Sprehn, Dorte Bekkevold, Arild Folkvord, Linda Laikre, Gunnar Kleinau, Patrick Scheerer, and Leif Andersson. "Recurrent Convergent Evolution at Amino Acid Residue 261 in Fish Rhodopsin." *Proceedings of the National Academy of Sciences* 116, no. 37 (2019): 18473-8478. doi:10.1073/pnas.1908332116.
4. Liu, Yang, Hai Chi, Longfei Li, Stephen J. Rossiter, and Shuyi Zhang. "Molecular Data Support an Early Shift to an Intermediate-Light Niche in the Evolution of Mammals." *Molecular Biology and Evolution* 35, no. 5 (2018): 1130-134. doi:10.1093/molbev/msy019.
5. Zhao, Y., J.-L. Ren, M.-Y. Wang, S.-T. Zhang, Y. Liu, M. Li, Y.-B. Cao, H.-Y. Zu, X.-C. Chen, C.-I. Wu, E. Nevo, X.-Q. Chen, and J.-Z. Du. "Codon 104 Variation of P53 Gene Provides Adaptive Apoptotic Responses to Extreme Environments in Mammals of the Tibet Plateau." *Proceedings of the National Academy of Sciences* 110, no. 51 (2013): 20639-0644. doi:10.1073/pnas.1320369110.
6. Aspiras, Ariel C., Nicolas Rohner, Brian Martineau, Richard L. Borowsky, and Clifford J. Tabin. "Melanocortin 4 Receptor Mutations Contribute to the Adaptation of Cavefish to Nutrient-poor Conditions." *Proceedings of the National Academy of Sciences* 112, no. 31 (2015): 9668-673. doi:10.1073/pnas.1510802112.
7. Nguyen, Vy, Christopher Wilson, Marc Hoemberger, John B. Stiller, Roman V. Agafonov, Steffen Kutter, Justin English, Douglas L. Theobald, and Dorothee Kern. "Evolutionary Drivers of Thermoadaptation in Enzyme Catalysis." *Science* 355, no. 6322 (2016): 289-94. doi:10.1126/science.aah3717.
8. Taylor, T. J. and Vaisman, I. I. Discrimination of thermophilic and mesophilic proteins. *BMC Structural Biology* **10,** (2010): S5.
9.Henzler-Wildman, Katherine A., Ming Lei, Vu Thai, S. Jordan Kerns, Martin Karplus, and Dorothee Kern. "A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis." *Nature* 450, no. 7171 (2007): 913-16. doi:10.1038/nature06407.
10. Cooper, A., and D. T. F. Dryden. "Allostery without Conformational Change." *European Biophysics Journal* 11, no. 2 (1984): 103-09. doi:10.1007/bf00276625.
11. Bellissent-Funel, Marie-Claire, Ali Hassanali, Martina Havenith, Richard Henchman, Peter Pohl, Fabio Sterpone, David Van Der Spoel, Yao Xu, and Angel E. Garcia. "Water Determines the Structure and Dynamics of Proteins." *Chemical Reviews* 116, no. 13 (2016): 7673-697. doi:10.1021/acs.chemrev.5b00664.
12. Guin, Drishti, and Martin Gruebele. "Weak Chemical Interactions That Drive Protein Evolution: Crowding, Sticking, and Quinary Structure in Folding and Function." *Chemical Reviews* 119, no. 18 (2019): 10691-0717. doi:10.1021/acs.chemrev.8b00753.
13. Vecchi, Giulia, Pietro Sormanni, Benedetta Mannini, Andrea Vandelli, Gian Gaetano Tartaglia, Christopher M. Dobson, F. Ulrich Hartl, and Michele Vendruscolo. "Proteome-wide Observation of the Phenomenon of Life on the Edge of Solubility." *Proceedings of the National Academy of Sciences* 117, no. 2 (2019): 1015-020. doi:10.1073/pnas.1910444117.
14. Bigman, Lavi S., and Yaakov Levy. "Proteins: Molecules Defined by Their Trade-offs." *Current Opinion in Structural Biology* 60 (2020): 50-56. doi:10.1016/j.sbi.2019.11.005.

15. Jr, Richard N. Mclaughlin, Frank J. Poelwijk, Arjun Raman, Walraj S. Gosal, and Rama Ranganathan. "The Spatial Architecture of Protein Function and Adaptation." *Nature* 491, no. 7422 (2012): 138-42. doi:10.1038/nature11500.
16. Raman, Arjun S., K. Ian White, and Rama Ranganathan. "Origins of Allostery and Evolvability in Proteins: A Case Study." *Cell* 166, no. 2 (2016): 468-80. doi:10.1016/j.cell.2016.05.047.
17. Buck, Matthias, Jonathan Boyd, Christina Redfield, Donald A. Mackenzie, David J. Jeenes, David B. Archer, and Christopher M. Dobson. "Structural Determinants of Protein Dynamics: Analysis of 15N NMR Relaxation Measurements for Main-Chain and Side-Chain Nuclei of Hen Egg White Lysozyme." *Biochemistry* 34, no. 12 (1995): 4041-055. doi:10.1021/bi00012a023.
18. Lisi, George P., and J. Patrick Loria. "Allostery in Enzyme Catalysis." *Current Opinion in Structural Biology* 47 (2017): 123-30. doi:10.1016/j.sbi.2017.08.002.
19. Zhang, Yan, Pemra Doruker, Burak Kaynak, She Zhang, James Krieger, Hongchun Li, and Ivet Bahar. "Intrinsic Dynamics Is Evolutionarily Optimized to Enable Allosteric Behavior." *Current Opinion in Structural Biology* 62 (2020): 14-21. doi:10.1016/j.sbi.2019.11.002.
20. Weinkam, Patrick, Yao Chi Chen, Jaume Pons, and Andrej Sali. "Impact of Mutations on the Allosteric Conformational Equilibrium." *Journal of Molecular Biology* 425, no. 3 (2013): 647-61. doi:10.1016/j.jmb.2012.11.041.
21. Williams, G. J., A. S. Nelson, and A. Berry. "Directed Evolution of Enzymes for Biocatalysis and the Life Sciences." *Cellular and Molecular Life Sciences* 61, no. 24 (2004): 3034-046. doi:10.1007/s00018-004-4234-5.
22. Merino, Enrique, Roy A. Jensen, and Charles Yanofsky. "Evolution of Bacterial Trp Operons and Their Regulation." *Current Opinion in Microbiology* 11, no. 2 (2008): 78-86. doi:10.1016/j.mib.2008.02.005.
23. Gutiérrez-Preciado, A., Romero, H. and Peimbert, M. An Evolutionary Perspective on Amino Acids. *Nature Education* **3**, no. 9 (2010): 29.
24. Henikoff, S., and J. G. Henikoff. "Amino Acid Substitution Matrices from Protein Blocks." *Proceedings of the National Academy of Sciences* 89, no. 22 (1992): 10915-0919. doi:10.1073/pnas.89.22.10915.
25. Wakamatsu, Yuma, Kazutoyo Ogino, and Hiromi Hirata. "Swimming Capability of Zebrafish Is Governed by Water Temperature, Caudal Fin Length and Genetic Background." *Scientific Reports* 9, no. 1 (2019). doi:10.1038/s41598-019-52592-w.
26. Hota, Prasanta K., and Matthias Buck. "Thermodynamic Characterization of Two Homologous Protein Complexes: Associations of the Semaphorin Receptor Plexin-B1 RhoGTPase Binding Domain with Rnd1 and Active Rac1." *Protein Science* 18, no. 5 (2009): 1060-071. doi:10.1002/pro.116.
27. Li, Zhen-Lu, and Matthias Buck. "Modified Potential Functions Result in Enhanced Predictions of a Protein Complex by All-Atom Molecular Dynamics Simulations, Confirming a Stepwise Association Process for Native Protein–Protein Interactions." *Journal of Chemical Theory and Computation* 15, no. 8 (2019): 4318-331. doi:10.1021/acs.jctc.9b00195.
28. Qian, W., X. He, E. Chan, H. Xu, and J. Zhang. "Measuring the Evolutionary Rate of Protein-protein Interaction." *Proceedings of the National Academy of Sciences* 108, no. 21 (2011): 8725-730. doi:10.1073/pnas.1104695108.
29. Zhang, Jianzhi, and Jian-Rong Yang. "Determinants of the Rate of Protein Sequence Evolution." *Nature Reviews Genetics* 16, no. 7 (2015): 409-20. doi:10.1038/nrg3950.
30. Saavedra, Harry G., James O. Wrabl, Jeremy A. Anderson, Jing Li, and Vincent J. Hilser. "Dynamic Allostery Can Drive Cold Adaptation in Enzymes." *Nature* 558, no. 7709 (2018): 324-28. doi:10.1038/s41586-018-0183-2.
31. Barman, Arghya, and Donald Hamelberg. "Coupled Dynamics and Entropic Contribution to the Allosteric Mechanism of Pin1." *The Journal of Physical Chemistry B* 120, no. 33 (2016): 8405-415. doi:10.1021/acs.jpcb.6b02123.
32. Kaplan, Mohammed, Siddarth Narasimhan, Cecilia De Heus, Deni Mance, Sander Van Doorn, Klaartje Houben, Dušan Popov-Čeleketić, Reinier Damman, Eugene A. Katrukha, Purvi Jain, Willie J.c. Geerts, Albert J.r. Heck, Gert E. Folkers, Lukas C. Kapitein, Simone Lemeer, Paul M.p. Van Bergen En Henegouwen, and Marc Baldus. "EGFR Dynamics Change during Activation in Native Membranes as Revealed by NMR." *Cell* 167, no. 5 (2016). doi:10.1016/j.cell.2016.10.038.

33. Popovych, Nataliya, Shangjin Sun, Richard H. Ebright, and Charalampos G. Kalodimos. "Dynamically Driven Protein Allostery." *Nature Structural & Molecular Biology* 13, no. 9 (2006): 831-38. doi:10.1038/nsmb1132.
34. Tzeng, Shiou-Ru, and Charalampos G. Kalodimos. "Dynamic Activation of an Allosteric Regulatory Protein." *Nature* 462, no. 7271 (2009): 368-72. doi:10.1038/nature08560.
35. Packer, Morgan, Jillian A. Parker, Jean K. Chung, Zhenlu Li, Young Kwang Lee, Trinity Cookis, Hugo Guterres, Steven Alvarez, Md Amin Hossain, Daniel P. Donnelly, Jeffrey N. Agar, Lee Makowski, Matthias Buck, Jay T. Groves, and Carla Mattos. "Raf Promotes Dimerization of the Ras G-domain with Increased Allosteric Connections." 2020. doi:10.1101/2020.07.15.205070.
36. D'aquino, J. Alejandro, Javier Gómez, Vincent J. Hilser, Kon Ho Lee, L. Mario Amzel, and Ernesto Freire. "The Magnitude of the Backbone Conformational Entropy Change in Protein Folding." *Proteins: Structure, Function, and Genetics* 25, no. 2 (1996): 143-56. doi:10.1002/(sici)1097-0134(199606)25:23.0.co;2-j.
37. Wolfenden, R., L. Andersson, P. M. Cullis, and C. C. B. Southgate. "Affinities of Amino Acid Side Chains for Solvent Water." *Biochemistry* 20, no. 4 (1981): 849-55. doi:10.1021/bi00507a030.
38. Mavromatis, Konstantinos, Iason Tsigos, Maria Tzanodaskalaki, Michael Kokkinidis, and Vassilis Bouriotis. "Exploring the Role of a Glycine Cluster in Cold Adaptation of an Alkaline Phosphatase." *European Journal of Biochemistry* 269, no. 9 (2002): 2330-335. doi:10.1046/j.1432-1033.2002.02895.x.
39. Sawle, Lucas, and Kingshuk Ghosh. "How Do Thermophilic Proteins and Proteomes Withstand High Temperature?" *Biophysical Journal* 101, no. 1 (2011): 217-27. doi:10.1016/j.bpj.2011.05.059.
40. Sternberg, Samuel H., Hagen Richter, Emmanuelle Charpentier, and Udi Qimron. "Adaptation in CRISPR-Cas Systems." *Molecular Cell* 61, no. 6 (2016): 797-808. doi:10.1016/j.molcel.2016.01.030.
41. Phillips, Angela M., Michael B. Doud, Luna O. Gonzalez, Vincent L. Butty, Yu-Shan Lin, Jesse D. Bloom, and Matthew D. Shoulders. "Enhanced ER Proteostasis and Temperature Differentially Impact the Mutational Tolerance of Influenza Hemagglutinin." *ELife* 7 (2018). doi:10.7554/elife.38795.
42. Asseng, S. et al. Climate change impact and adaption for wheat proteins. *Global Change Biology* **25**, no.1 (2019): 155-173 (2019).
43.Wang, Yu, Chenzhou Zhang, Nini Wang, Zhipeng Li, Rasmus Heller, Rong Liu, Yue Zhao, Jiangang Han, Xiangyu Pan, Zhuqing Zheng, Xueqin Dai, Ceshi Chen, Mingle Dou, Shujun Peng, Xianqing Chen, Jing Liu, Ming Li, Kun Wang, Chang Liu, Zeshan Lin, Lei Chen, Fei Hao, Wenbo Zhu, Chengchuang Song, Chen Zhao, Chengli Zheng, Jianming Wang, Shengwei Hu, Cunyuan Li, Hui Yang, Lin Jiang, Guangyu Li, Mingjun Liu, Tad S. Sonstegard, Guojie Zhang, Yu Jiang, Wen Wang, and Qiang Qiu. "Genetic Basis of Ruminant Headgear and Rapid Antler Regeneration." *Science* 364, no. 6446 (2019). doi:10.1126/science.aav6335.
44. Nelson, Peter T., Kari Stefansson, Jeffrey Gulcher, and Clifford B. Saper. "Molecular Evolution of Tau Protein: Implications for Alzheimer's Disease." *Journal of Neurochemistry* 67, no. 4 (1996): 1622-632. doi:10.1046/j.1471-4159.1996.67041622.x.
45. Loehlin, David W., Jesse R. Ames, Kathy Vaccaro, and Sean B. Carroll. "A Major Role for Noncoding Regulatory Mutations in the Evolution of Enzyme Activity." *Proceedings of the National Academy of Sciences* 116, no. 25 (2019): 12383-2389. doi:10.1073/pnas.1904071116
46. Howe K et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496,** (2013): 498-503.
47. Andricioaei, Ioan, and Martin Karplus. "On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations." *The Journal of Chemical Physics* 115, no. 14 (2001): 6289-292. doi:10.1063/1.1401821.
48. Seeber, Michele, Angelo Felline, Francesco Raimondi, Stefanie Muff, Ran Friedman, Francesco Rao, Amedeo Caflisch, and Francesca Fanelli. "Wordom: A User-friendly Program for the Analysis of Molecular Structures, Trajectories, and Free Energy Surfaces." *Journal of Computational Chemistry* 32, no. 6 (2010): 1183-194. doi:10.1002/jcc.21688.
49. Amaral, M., D. B. Kokh, J. Bomke, A. Wegener, H. P. Buchstaller, H. M. Eggenweiler, P. Matias, C. Sirrenberg, R. C. Wade, and M. Frech. "Protein Conformational Flexibility Modulates Kinetics and Thermodynamics of Drug Binding." *Nature Communications* 8, no. 1 (2017). doi:10.1038/s41467-017-02258-w.
50. Webb, Benjamin, and Andrej Sali. "Protein Structure Modeling with MODELLER." *Methods in Molecular Biology Protein Structure Prediction*, 2014, 1-15. doi:10.1007/978-1-4939-0366-5_1.

51. Phillips, James C., Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. "Scalable Molecular Dynamics with NAMD." *Journal of Computational Chemistry* 26, no. 16 (2005): 1781-802. doi:10.1002/jcc.20289.
52. Shaw, D. E. et al. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. SC14: International Conference for High Performance Computing, Networking, Storage and Analysis (2014). doi:10.1109/sc.2014.9.
53. Huang, Jing, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L De Groot, Helmut Grubmüller, and Alexander D. Mackerell. "CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins." *Nature Methods* 14, no. 1 (2016): 71-73. doi:10.1038/nmeth.4067.