OPEN ACCESS Freely available online

Collective Dynamics Differentiates Functional Divergence in Protein Evolution

Tyler J. Glembo[1], Daniel W. Farrell[2], Z. Nevin Gerek[1], M. F. Thorpe[1], S. Banu Ozkan[1]*

1 Center for Biological Physics, Department of Physics, Arizona State University, Tempe, Arizona, United States of America, **2** Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York, United States of America

Abstract

Protein evolution is most commonly studied by analyzing related protein sequences and generating ancestral sequences through Bayesian and Maximum Likelihood methods, and/or by resurrecting ancestral proteins in the lab and performing ligand binding studies to determine function. Structural and dynamic evolution have largely been left out of molecular evolution studies. Here we incorporate both structure and dynamics to elucidate the molecular principles behind the divergence in the evolutionary path of the steroid receptor proteins. We determine the likely structure of three evolutionarily diverged ancestral steroid receptor proteins using the Zipping and Assembly Method with FRODA (ZAMF). Our predictions are within ~2.7 Å all-atom RMSD of the respective crystal structures of the ancestral steroid receptors. Beyond static structure prediction, a particular feature of ZAMF is that it generates protein dynamics information. We investigate the differences in conformational dynamics of diverged proteins by obtaining the most collective motion through essential dynamics. Strikingly, our analysis shows that evolutionarily diverged proteins of the same family do not share the same dynamic subspace, while those sharing the same function are simultaneously clustered together and distant from those, that have functionally diverged. Dynamic analysis also enables those mutations that most affect dynamics to be identified. It correctly predicts all mutations (functional and permissive) necessary to evolve new function and ~60% of permissive mutations necessary to recover ancestral function.

Citation: Glembo TJ, Farrell DW, Gerek ZN, Thorpe MF, Ozkan SB (2012) Collective Dynamics Differentiates Functional Divergence in Protein Evolution. PLoS Comput Biol 8(3): e1002428. doi:10.1371/journal.pcbi.1002428

Editor: Ruth Nussinov, National Cancer Institute, United States of America and Tel Aviv University, Israel, United States of America

Received October 11, 2011; **Accepted** January 30, 2012; **Published** March 29, 2012

Funding: This research was supported in part by the National Science Foundation through TeraGrid resources provided by Ranger and the Fulton High Performance Computing Initiative at Arizona State University for computer time. SBO and ZNG acknowledge the support from 1U54GM094599. MFT and DWF thank NSF for support through grant DMS-0714953 The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Banu.Ozkan@asu.edu

Introduction

Proteins are effective and efficient machines that carry out a wide range of essential biochemical functions in the cell. Beyond being robust and efficient, the outstanding property of proteins is that they can evolve and they show a remarkable capacity to acquire new functions and structures. In fact, modern proteins have emerged from only a few common ancestors over millions to billions of years [1–3]. Moreover, the emergence of drug resistance and enzymes with the capacity to degrade new chemicals indicates the ongoing contemporary evolution of proteins [1–7]. Therefore, understanding the mechanism by which mutations lead to functional diversity is critical in many aspects from protein engineering to drug design and personalized medicine. Indeed, computational protein design through analysis of mutations has attained major breakthroughs, with profound biotechnological and biomedical implications: design of a new fold [8], design of new biocatalysts and biosensors [9–11], design of binding affinity [12,13], and design of proteins to bind non-biological cofactors [14]. Moreover, there are computational bioinformatics-based tools based on evolutionary information aspects to identify mutations leading to functional loss or disease [15–17].

From a phylogenetics perspective, horizontal and vertical approaches have been used to analyze the set of mutations that lead to changes in protein function throughout evolution [18]. The horizontal approach compares modern day proteins at the tips of the evolutionary tree. It identifies the amino acid residue differences within the functionally divergent members of a protein family based on primary sequence and structural analyses and then characterizes the functional role of these residues by swapping them between these family members through site-directed mutagenesis in the laboratory to check for loss of function [19–21]. Although the horizontal method gives insight into mutations critical to function, it often fails to identify permissive mutations necessary to switch function between family members. Protein function has evolved as mutations throughout history, i.e. *"vertically"*, in the ancestral protein lineages. Therefore, it is important to incorporate the historical background which contains both neutral and key function-switching mutations when examining function-altering mutations [18]. The vertical approach determines the likely ancestral sequences at nodes along the evolutionary tree and compares modern day proteins to their ancestors. Recent advances in molecular phylogenetic methods make it possible to obtain ancestral sequences by protein sequence alignments in a phylogenetic framework using Bayesian and Maximum Likelihood methods [22,23]. DNA molecules are synthesized coding for the most probable ancestral sequences and the protein expressed, allowing for experimental character-

Author Summary

Proteins are remarkable machines of the living systems that show diverse biochemical functions. Biochemical diversity has grown over time via molecular evolution. In order to understand how diversity arose, it is fundamental to understand how the earliest proteins evolved and served as templates for the present diverse proteome. The one sequence - one structure - one function paradigm is being extended to a new view: an ensemble of different conformations in equilibrium can evolve new function and the analysis of inherent structural dynamics is crucial to give a more complete understanding of protein evolution. Therefore, we aim to bring structural dynamics into protein evolution through our zipping and assembly method with FRODA. (ZAMF). We apply ZAMF to simultaneously obtain structures and structural dynamics of three ancestral sequences of steroid receptor proteins. By comparative dynamics analysis among the three ancestral steroid hormone receptors: (i) we show that changes in the structural dynamics indicates functional divergence and (ii) we identify all functionally critical and most of the permissive mutations necessary to evolve new function. Overall, all these findings suggest that conformational dynamics may play an important role where new functions evolve through novel molecular interactions.

ization of the ancient protein. The vertical approach has been used to gain insight into the underlying principles of protein function and evolution in several proteins including opsins [24,25], GFP-like protein [26,27], and others [28–32]. More recently, a vertical analysis of two ancestral nuclear receptors has been coupled with X-ray structure determination in successfully elucidating the switching of function between divergent members [33,34]. Such studies highlight the importance of including ancient protein structures into evolutionary studies.

Although coarse-grained and all-atom models have furthered our understanding of sequence/structure relationship in evolution, further study of the inherent structural dynamics is crucial to give a more complete understanding of protein evolution [35]. A small local structural change due to a single mutation can lead to a large difference in conformational dynamics, even at quite distant residues due to structural allostery [36–38]. Thus the one sequence-one structure-one function paradigm is being extended to a new view: an ensemble of different conformations in equilibrium that can evolve new function [1,39–41]. The importance of structural dynamics has been demonstrated by a recent experimental study which shows that mutations distant from a binding site can increase enzyme efficiency by changing the conformational dynamics [42]. The modulation of rigidity/flexibility of residues both near and distant from the active region(s) as related to promiscuous and specific binding has also been noted in tRNA synthetase complexes [43,44].

Here we have developed a method to predict structural and dynamic evolution of ancestral sequences by using a modified version of our protein structure prediction tool, Zipping and Assembly Method with FRODA (ZAMF) [45]. ZAMF combines two crucial features of ZAM [46], and FRODA [47,48] : i) FRODA is a constraint-based geometric simulation technique that speeds up the search for native like topologies by accounting only for geometric relationships between atoms instead of detailed energetics, ii) Molecular dynamics identifies the low free energy structures and further refines these structures toward the actual native conformation. Thus, it is a two-step multi-scale computational method that performs fast and extensive conformational sampling. As an outcome, we not only predict protein structures but also obtain detailed conformational dynamics of the predicted structures.

With modified ZAMF, we analyze the role of structural dynamics in the evolution of three ancestral steroid receptors (AncCR, AncGR1 and AncGR2), the ancestors of mineralocorticoid and glucocorticoid receptors (MR and GR). MR and GR arose by duplication of a single ancestor (AncCR) deep in the vertebrate lineage and then diverged function. MR is activated by aldosterone to control electrolyte homeostasis, kidney and colon function and other processes [33]. It is also activated by cortisol, albeit to a lesser extent [18]. On the other hand, GR regulates the stress response and is activated only by cortisol [33]. The structural comparison of human MR and GR (i.e. horizontal approach) suggested the two mutations (S106P and L111Q) to be critical in ligand specificity, however, swapping these residues between human MR and human GR yielded receptors with no binding activity [49]. Conversely, by resurrecting key ancestral proteins (AncCR, AncGR1 and AncGR2) in MR and GR evolution and determining the crystal structures, Thornton *et al.* were able to shed insight into how function diverges through time by using both functional and permissive (compensatory) mutations [33,34]. AncCR (main ancestor), ~470 million years old, is a promiscuous steroid receptor which is activated by aldosterone, cortisol, and deoxycortisol ligands. AncCR branched into the mineralocorticoid steroid receptors. AncGR1 (ancestor of sharks) is ~440 million years old with 25 mutations from AncCR and also promiscuously binds to and functions with aldosterone, cortisol, and deoxycortisol. AncGR1 later evolved into the Elasmobranch glucocorticoid receptor protein. AncGR2 (ancestor of humans and fish) is ~420 million years old with 36 mutations from AncGR1 and preferentially binds to cortisol alone. These two ancestral proteins, AncGR1 and AncGR2, which diverge functionally, have highly similar experimental structures that have <1 Å RMSD between them. Among 36 mutations between AncGR1 and AncGR2, two conserved mutations {S106P, L111Q} (i.e. group X) when introduced together are sufficient to increase cortisol specificity. However three more functionally critical conserved mutations {L29M, F98I, S212Δ} (i.e. group Y) are needed for the loss of aldosterone binding activity when they are introduced together with two other permissive (i.e. compensatory) mutations {N26T and Q105L} (i.e. group Z). Thus, making the X, Y, Z mutations in AncGR1 enables AncGR1 to function as AncGR2 (i.e. forward evolution) [34]. To make AncGR2 function as AncGR1 (backward evolution) the X, Y, Z mutations are insufficient and render the protein inactive. A fourth set of permissive mutations (W) is required to reverse function in addition to the X, Y, and Z, sets. The W mutation set is {H84Q, Y91C, A107Y, G114Q, L197M} [33]. A mutation between AncCR and AncGR1, Y27R, is also a necessary mutation to eventually alter function to cortisol specificity, though it was not experimentally considered as part of the X, Y, Z, or W mutation sets [34].

We ask here whether an analysis of the predicted 3-D structures and corresponding equilibrated dynamics can distinguish the functional divergence and function swapping mutations between AncCR, AncGR1, and AncGR2. By applying ZAMF, we obtain the 3-D structures within ~2.7 Å all-atom RMSD of the experimental structures. More importantly, when we analyze their structure-encoded dynamics, we observe that changes in the dynamics indicate functional divergence: that the most collective fluctuation profiles of AncCR and AncGR1 (i.e. the slowest mode) are much closer and distinctively separated from the functionally divergent AncGR2. Moreover, AncCR and AncGR1 have a more flexible binding pocket, suggesting the role of flexibility in their

promiscuous binding specificity. On the other hand, the mutations of AncGR2 lead to a rigid binding pocket, which suggests that as the binding becomes cortisol specific, evolution acts to shape the binding pocket toward a specific ligand. Finally, using their mean square fluctuation profiles and cross correlation maps to analyze the change in dynamics at each residue position enables us to distinguish critical mutations needed for swapping the function. Overall, all these findings suggest that conformational epistasis may play an important role where new functions evolve through novel molecular interactions and an analysis of detailed dynamics might provide insight into the mechanisms behind these novel interactions.

Results/Discussion

Structure Prediction and Identification of Function Altering Mutations through Structural Analysis

Many of the modern day homologs to ancestral proteins in the steroid receptor class of the nuclear receptor superfamily have high sequence similarity (~40–50%), and, as prediction accuracy scales with sequence similarity [50–52] our secondary structures for the ancestral sequences are sufficiently accurate to provide native-like structures [45]. Indeed, predicted secondary structures are all correct within one residue to the experimentally determined ancestral cortisol receptor protein [34]. Using these secondary structures as input to the assembly and refinement stages of ZAMF, we determine the 3D structure of the AncCR from its experimentally determined structure to 2.5 Å all atom RMSD (2.2 Å backbone), AncGR1 from its experimentally determined structure to 2.9 Å all atom RMSD (2.6 Å backbone) AncGR2 from its experimentally determined structure to 2.9 Å all atom RMSD (2.4 Å backbone) (Fig. 1 and Table S1). To test the accuracy of these predictions, we first compare the structural differences between the experimental structures. The experimental structures are very similar, with an RMSD of 1.49 Å between AncCR and AncGR1, 1.68 Å between AncCR and AndGR2, and 1.70 Å between AncGR1 and AncGR2. However alignment excludes the atoms of the mutational residues. We also ran a 4 ns REMD simulation of the experimentally determined AncCR and AncGR2 under the same conditions. The ensembles for AncCR and AncGR2 converges at ~2.5 Å backbone RMSD from their respective experimentally determined structures (Fig. S1). The 2.5 Å RMSD indicates that our predicted structures are as accurate as our force field permits. Closer analysis reveals that helix h9 in the predicted structure of AncGR2 is slightly less stable than in the experimental structure REMD simulations. However, both simulations show a high degree of flexibility in the loop region between helices h9 and h10 and ends of helices h9 and h10 at this loop region.

As these three proteins diverged in function and have >10% sequence mutation between each successive protein, we expect to see some differences in structure. Therefore, we first look at a mean square displacement (MSD) between the static structures of AncCR, AncGR1 and AncGR2. The MSD versus residue profile gives an indication of which residues are mutating, as mutated residues pack into stereochemically unique conformations (Fig. S2). Fig. S2 reveals conformational shifts in helices h7 and h10 and in the β-sheet region, b1. We attempt to determine which of the 36 mutated residues between AncGR1 and AncGR2 are critical for cortisol binding specificity through distinguishing residues having an MSD cutoff of >6 $Å^2$ between the AncGR1 and AncGR2 predicted structures. The residues identified from X, Y, Z and W sets are Y91C, Q105L, and S212Δ, with no false positives. The S212Δ and Q105L mutations are permissive mutations to shift function to cortisol specificity whereas Y91C is a permissive mutation necessary for "reverse evolution" i.e. to return binding promiscuity to AncGR2. Experimental work indicates that S212Δ removes a hydrogen bond and imparts greater mobility to the loop before the activation function (AF) helix, allowing it to hydrogen bond with helix h3, while Q105L indirectly restores a hydrogen

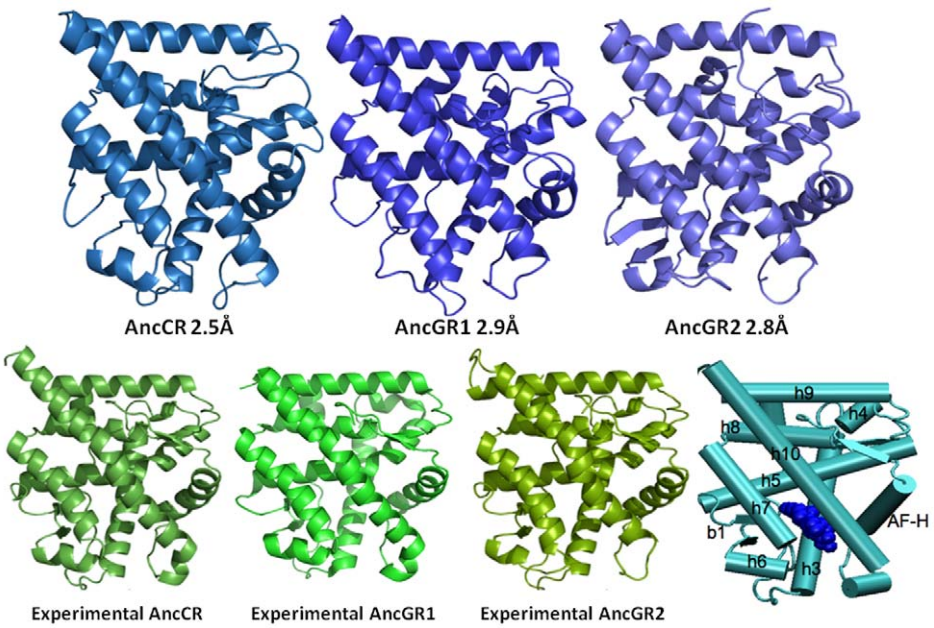

Figure 1. 3D structures of AncCR, AncGR1 and AncGR2. AncCR was within 2.5 Å all-atom RMSD from the experimentally determined AncCR. AncGR1 was within 2.9 Å all-atom RMSD from the experimentally determined crystal structure. AncGR2 was within 2.8 Å all-atom RMSD from the experimentally determined AncGR2. Included for reference is a cartoon figure with helices labeled for reference and the ligand is bound, represented in blue spheres.
doi:10.1371/journal.pcbi.1002428.g001

bond with the activation helix by allowing for tighter packing of helices h3 and h7 [34]. An analysis of hydrogen bonding patterns [53] shows the loss of the S212 hydrogen bond with V217 (in the loop before the AF helix) in the AncGR2 structure as compared to the AncCR/AncGR1 structures, agreeing with experimental results. Y91C is one of the W mutations required for reverse evolution of AncGR1 from AncGR2 and we find it forms a hydrogen bond with N86 in AncGR2 but does not in AncCR or AncGR1. Interestingly, none of these mutations occur in the binding pocket itself. Therefore, an MSD analysis is not sensitive enough to find functionally critical mutations in the binding pocket, and only finds a few of the necessary mutations to diverge function.

The Relationship with Functional Divergence and Structural Dynamics

We investigate the role of structural dynamics in functional divergence observed among the three ancestral steroid proteins. The extensive conformational sampling of our method enables us to capture the dynamics along with the most native-like structure (Fig. S4). We obtain the most collective modes of these three ancestral structures (i.e. slowest fluctuation profiles) through principal component analysis of our restraint-free trajectories (See Method). We then form an Mx3N matrix where the M columns are the eigenvectors weighted by their eigenvalues, with each M column being a 3 column super-element composed from the slowest modes of AncCR, AncGR1 and AncGR2 and N being the number of C-α atoms. We chose to analyze the top 10 slowest modes and therefore there are 30 columns. By performing a singular value decomposition on this matrix, we measure how the most collective motions of these three ancestral proteins are distributed in dynamic space. Interestingly, as shown in Fig. 2A, AncCR and AncGR1 are much closer and distinctively separated in dynamic space from the functionally divergent ancestor of the human glucocorticoid receptor, AncGR2. Clustering in dynamics space is significant because it shows that these structurally similar but functionally unique proteins differ in functionally governing dynamics, as observed in previous studies [42,54–56]. Moreover, previous studies indicate that functionally critical mutations alter modes that characterize biologically functional motion, while random sequence variations typically have non-statistically significant impact on those modes [57]. These findings indeed suggest that the governing functional dynamics is encoded within the structure and that only critical mutations lead to a shift in collective motion and therefore in binding selectivity as well [55,58].

Fig. 2B presents the color coded ribbon diagrams of these three ancestral proteins with respect to their functionally related collective fluctuation (obtained by PCA) profiles within a spectrum of red to blue, where rigid regions are denoted by blue/green and flexible regions are denoted with red/orange. Experimentally determined function altering mutations are highlighted in the sphere representation. Strikingly, residues in and near the functional site (i.e. binding site) are much more flexible for the two promiscuous enzymes (AncGR1 and AncCR) whereas the human ancestor AncGR2, which has affinity only to cortisol, has very rigid functional site residues. The new view of proteins states that, rather than a single structure with induced binding, proteins interconvert between bound and unbound conformations in the native ensemble. Thus, promiscuous binding proteins utilize greater flexibility to interconvert between a greater number of conformations in the native ensemble as compared to specific binding proteins. Therefore, our dynamic analysis agrees with the new view that while the promiscuous ancestors are more flexible

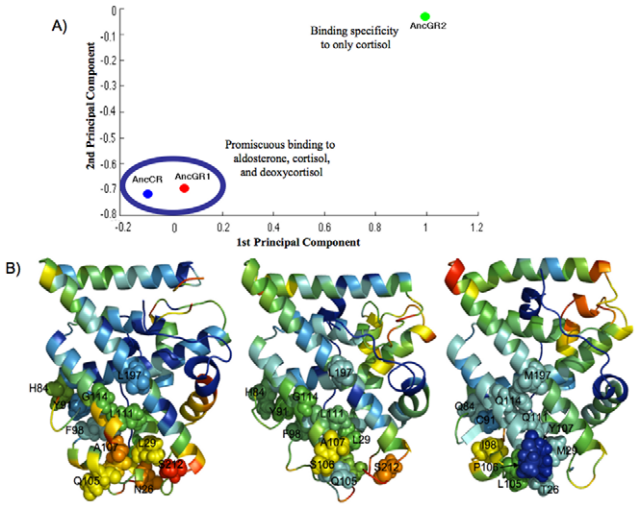

Figure 2. Plot and ribbon diagram of the dynamics of the three ancestral proteins characterized by slowest collective mode. (A) The first two principal components of AncCR, AncGR1 and AncGR2 plotted against each other. The principal components were found via a Singular Value Decomposition of the **G** matrix (See Methods). Higher order modes are mostly orthogonal or mixed and therefore not represented here. (B) 3D structures of AncCR, AncGR1 and AncGR2 colored by residue fluctuation. The critical mutations in AncCR and AncGR1 have greater flexibility and thus, higher binding promiscuity. AncGR2 has much lower flexibility in general amongst these residues and therefore more selective binding. The S212Δ mutation also rigidifies the lower loop at the bottom end of h10 by shortening the loop and removing degrees of freedom. This also alters the packing of h10 (the frontmost helix) and decreases flexibility.
doi:10.1371/journal.pcbi.1002428.g002

around the functional site, the functional site rigidifies as Nature biases towards binding only a single ligand with greater affinity [1].

Identification of Function Altering Mutations through Structural Dynamics

Upon confirmation that dynamics can indeed distinguish functional divergence, the next question is whether dynamics can indicate which residues in the protein are critical to diverging function. We investigate whether we can distinguish the mutations, including function altering and permissive (i.e. compensatory), that cause AncCR/GR1 to shift function to specifically bind cortisol as AncGR2 does, and also those that reverse the function of AncGR2 to promiscuously bind in the same way as AncCR/AncGR1.

To identify the critical residues for swapping function, we analyze how the fluctuation profile changes over these three successive ancestral proteins. Thus, using their most collective fluctuation profile (i.e the slowest mode obtained by PCA), we compute the net change in fluctuation from AncCR to AncGR1 and AncGR1 to AncGR2 and show them in a 2-D plot to distinguish the mutations that have a higher impact on the change in dynamics between AncGR2 and AncGR1 compared to those mutations affecting the change in dynamics between AncGR1 and AncCR (Fig. 3). The upper left region of the graph in Fig. 3 indicates mutations that most alter dynamics when comparing the function-altering mutation from AncGR1 (binding promiscuity) to AncGR2 (binding specificity to cortisol) whereas the lower right region of the plot indicates mutations that most alter dynamics when comparing AncCR and AncGR1, which do not diverge functionally. The central region of the graph (between the parallel cutoff lines) contains those mutations that do not alter the

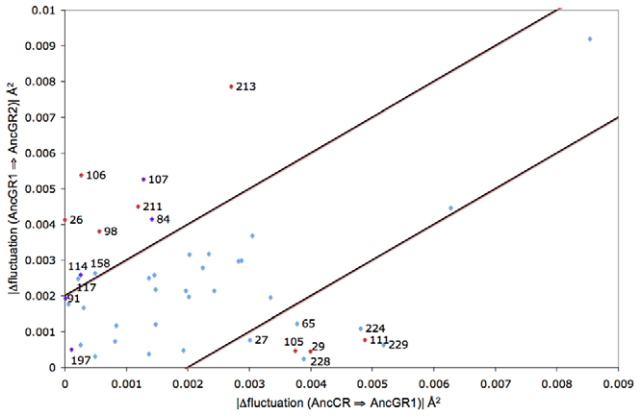

Figure 3. The change in fluctuation along the most collective mode between AncCR, AncGR1 and AncGR2. The X, Y, Z, and Y27R mutation groups necessary to alter function toward cortisol binding specificity are noted in red, and those permissive W mutations necessary to reverse function and recover promiscuous binding are noted in purple. A cutoff of ± 0.002 Å^2 is applied to differentiate mutations critical to altering dynamics as also used in Fig. 4. The upper left region of the graph indicates mutations that most alter dynamics when comparing the function-altering mutation from AncGR1 (binding promiscuity) to AncGR2 (binding specificity to cortisol) whereas the lower right region of the plot indicates mutations that most alter dynamics when comparing AncCR and AncGR1, which do not diverge functionally.
doi:10.1371/journal.pcbi.1002428.g003

dynamics in a significantly different manner between successive homologs. Interestingly, most of the function altering mutation sites such as 106, 212 (shown as 211 and 213 due to deletion) and most of the W mutations (mutations necessary for backward evolution, *e.g.* altering AncGR2 to become promiscuous) are in the upper left region. Permissive mutations 27, 29, 105, and the mutations in the activation function helix are in the lower right region of the plot. 111, a critical mutation for changing the specificity to cortisol only, is also in the lower right region. However, experimental analysis showed that the 111 mutation alone does not alter function in any appreciable manner. Thus, we propose it is only after permissive mutations alter the dynamics at site 111 can the necessary critical mutation at site 111 have a function altering effect. Additionally, certain mutations such as 214 and 173 both show large dynamic transitions. Mutation 214 is associated with the loop region that contains the critical mutation S212Δ, and it is in at the edge of a loop region. It undergoes transitions between being at the end of the h10 helix to being in the loop. The change in dynamics can be associated with the S212Δ mutation to identify the loop as a critical region. The 173 mutation is in a region that was not able to be crystallized in the experimental AncCR structure. Though the REMD simulations were determined to have converged, there is a possibility of some influence near site 173 due to the loop having to be built into the structure prior to REMD simulation. However, we expect that the shift in dynamics at mutation 173 may be correlated with movement of helix h10, and is therefore potentially significant.

We also obtain the net absolute change in the successive Δr^2 fluctuation profiles along the slowest mode using the formulation $\|\Delta\text{fluctuation}_{\text{AncCR-AncGR1}}|-|\Delta\text{fluctuation}_{\text{AncGR1-AncGR2}}\|$ for mutated residues based the alignment of AncCR and AncGR2 (Fig. 4A) and predict those residues with a net $|\Delta\Delta\text{fluctuation}|>0.002$ Å^2 to be critical. The forward mutations required to shift function to cortisol specificity are N26T, L29M, F98I, Q105L, S106P, L111Q, and S212Δ, and all of these are captured as critical as they are above the cutoff. The reverse mutations required to shift function from cortisol specific to promiscuous binding are H84Q, Y91C, A107Y, G114Q, and L197M. With the chosen cutoff, the identified permissive mutations are H84Q, A107Y, and G114Q, with Y91C only slightly below the cutoff. Interestingly, A107Y is the only W mutation that by itself partially recovered the promiscuous binding function [33] and it shows a high $|\Delta\Delta\text{fluctuation}|$ in our plot. We also find eight other mutated residues above the cutoff. Three of those are false positives I65L, Q117K and M158I. Each of these mutations occurred between AncCR and AncGR1, prior to a shift in function. Among mutations identified is Y27R, which is not explicitly in the X, Y, or Z set, yet it is highly conserved in the GR family and is an experimentally determined permissive mutation critical for GR function [34]. The three mutations at the activation function helix are also identified as critical. The other mutation above the cutoff is 211, which is correlated with S212Δ. Overall, our dynamic method identifies all mutations that are necessary for the evolution of GR function. We also distinguish three of the five mutations necessary for reversal of evolution (e.g. permissive mutations to AncGR2 which are necessary to recover the promiscuous binding of AncCR/AncGR1). Interestingly, many of the identified critical mutations such as N26T, H84Q, Y91C, F98I, Q105L, and S212Δ, are not interacting with the ligand, but rather are distant from the binding pocket (i.e. >5 Å from any atom in the ligand). Additionally, the high $|\Delta\Delta\text{fluctuation}|$ at the C-terminus is associated with the activation-function (AF) helix, which does not contain critical mutations but its dynamics is critical to function.

We also investigate the pairwise cross correlations of AncGR1 and AncGR2 (Fig. 4B). Interestingly, comparing the cross correlations reveals differences along the regions containing critical mutations. The cross-correlations between helix h5 (containing the critical mutation H84Q) and helix h7 (containing the critical mutations: Q105L, S106P, A107Y, L111Q, G114Q) become highly positively correlated in AncGR2 whereas there is no correlation in AncGR1. Analysis of hydrogen bonds [53] in predicted structures showed that additional hydrogen bonds are found between the β-sheet b1 and helices h5 and h7, indicating the observed increased correlation in AncGR2 is likely due to the repacking of helices h5 and h7 after mutation which incorporates/creates these new hydrogen bonds. Moreover, we also observe increased positive correlations between the AF-helix and helices h3 and h10 in AncGR2. These regions contain multiple permissive mutations (N26T, L29M, L197M, S212Δ) and thus, the change in correlations relate to the change in the stability of the AF helix caused by these permissive mutations necessary to alter function [34]. Furthermore, in Fig. 4C we compare the cross correlations of the most critical mutation for swapping the function to GR (X mutations) and the permissive mutations necessary to reverse the function to MR (W mutations) between AncGR1 and AncGR2. In AncGR2 these mutations are significantly more correlated than in AncGR1. This indeed suggests that W mutations play a critical role for GR function from the dynamics-perspective and therefore, they also need to be reversed along with the X, Y, Z mutation to recover the MR function.

To test the robustness of our method in other proteins we repeated our method for benign and disease associated mutations [59–61] in the human ferritin protein [62] (Fig. S5). We observe that, indeed, benign and disease associated mutations are individually clustered together while separated from each other in dynamics space.

In summary, by comparative dynamics analysis among the three ancestral steroid hormone receptors we identify all functionally critical and permissive mutations necessary to evolve

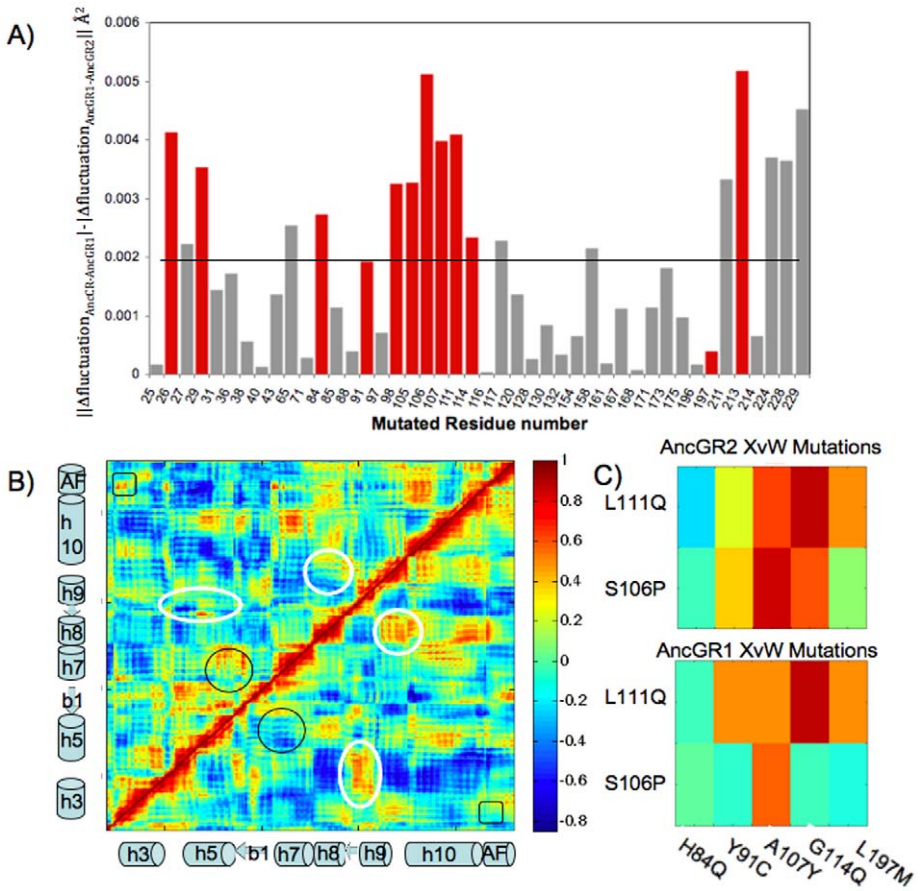

Figure 4. The change in net fluctuations and correlations of the mutated residues for successive evolution of MR to GR proteins. (A) The change in net fluctuation between successive ancestral proteins, AncCR, AncGR1 and AncGR2 for mutated residues. Those residues identified as critical to alter-function are noted in red. The activation-function (AF) helix contains mutations 224 and 229. A cutoff (solid line) results in all critical mutations identified except for Y91C and L197M. Y27R is noted as critical to function but sites 65, 117, and 158 are false positives. (B) The cross correlation map with AncGR2 on the upper left and AncGR1 on the lower right. Circled in black are changes in the cross correlation associated with critical residues near the binding pocket. Squared in black are the changes in cross correlation due to critical mutation N26T forming a hydrogen bond with the AF-helix. Circles in white are additional changes in cross correlation not associated with critical mutations. (C) The cross correlations between the X and W mutations. The correlation between X and W mutations is higher for AncGR2, whereas AncGR1 X functional mutations are uncorrelated, increasing the flexibility in the binding pocket and allowing for promiscuous binding.
doi:10.1371/journal.pcbi.1002428.g004

new function from the ancestral MR promiscuous binding proteins to the ancestral GR cortisol-specific binding proteins. We also identify 60% of the permissive mutations necessary to revert to ancestral function along with an additional functionally critical mutation. We observe significant loss of flexibility in key residues both near and distant from the binding pocket in the transition from promiscuous to specific binding. A loss in flexibilty agrees well with the new view of proteins being conformationally dynamic in which bound and unbound conformations are sampled within the native ensemble. Thus, proteins evolve not just through those mutations that alter function in the immediate sense, but also due to those mutations that are permissive and alter the dynamic space in which the protein exists, thereby giving the protein the potential to evolve new function.

Methods

Ancestral Protein Structure Prediction Based on Modern Homologs

We previously used the Zipping and Assembly Method with FRODA [ZAMF] [45–48,63] on a set of test proteins to predict the 3D structure from their 1D amino acid sequence. Here, we slightly modify ZAMF for the prediction of ancestral protein structures, particularly the three ancestral steroid receptor proteins, the corticoid receptor [AncCR], the glucocorticoid/corticoid receptor [AncGR1], and the glucocorticoid receptor [AncGR2] [33,34]. Since structure is more conserved than sequence [64–66], we incorporate structural data acquired from modern day homologues into our prediction method. The modified version of ZAMF as outlined in Fig. 5 includes several steps: (i) obtaining secondary structural motifs and common contacts based on modern homologs, (ii) generation of an unfolded ensemble, (iii) generation of compact-native like conformations using FRODA, and (iv) refinement by ZAMF. Overall, all these steps lead to an extensive search in conformational space, which comes with several advantages. First, we increased our prediction accuracy for native structures compared to the previous version of ZAMF. Second, we obtain converged dynamics trajectories through the refinement stage of ZAMF, which is used for dynamic evolution analysis of the ancient proteins. We summarize each step in our approach below.

I. Obtain secondary structural motifs and potential contact map of ancestral sequences. Usually, the first stage of ZAMF is to predict the secondary structural elements for

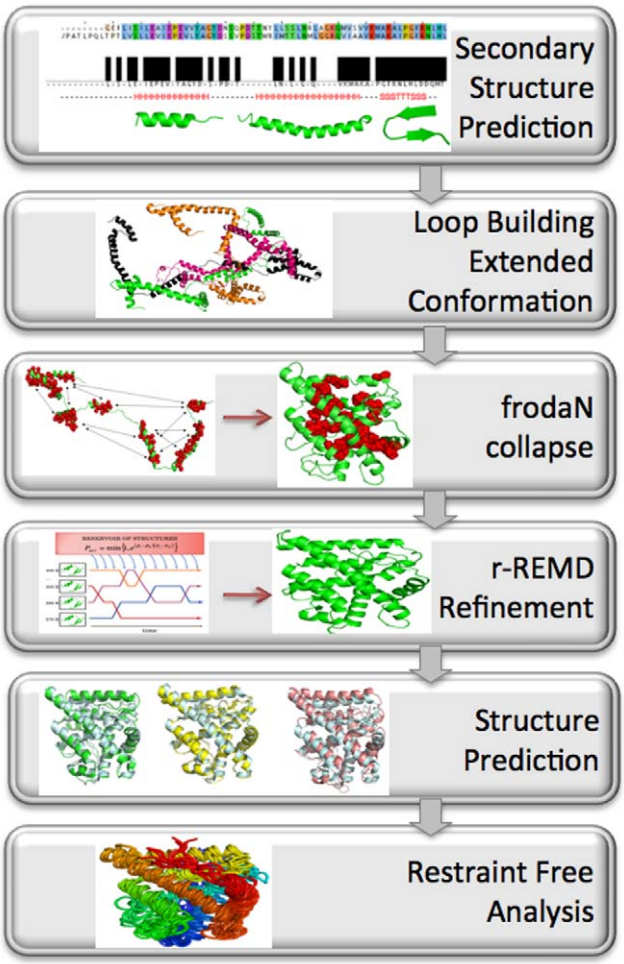

Figure 5. The secondary structure is predicted through multiple sequence alignment with modern day homologs. These secondary structural elements are then connected with loops in extended conformation to generate hundreds of conformations with high flexibility. Only a few are shown here. These structures all undergo a FRODA simulation which collapses them by adding attractive perturbations between all hydrophobic contact pairs (represented by arrows) into tightly packed structures with hydrophobic cores. A subset of hydrophobic residues are shown as spheres. After scoring, the collapsed structures they are ran in a restrained r-REMD simulation for 5 ns and then an unrestrained REMD simulation for 5 ns or until converged. The 3 ancestral structures are prediction to within 2.7 Å all atom RMSD of a similar experimentally determined structure. The final ensemble of restraint free generated structures are analyzed for dynamics using PCA.
doi:10.1371/journal.pcbi.1002428.g005

shortened sequences, i.e., 8mers, 12mers, 16mers etc of the protein using an *ab initio* approach. However, here we use the SSPRED online server [67] to confirm likely secondary structural elements by examining the secondary structure of modern day homologs such as mouse, human, and rat steroid receptor proteins [68–71] and aligning with the ancestral sequences. We choose the predicted secondary structural motifs such that they agree with the secondary structural motifs of modern day homologs at the regions with high sequence similarity. Furthermore, the information gleaned from the sequence alignment of the modern day homologs is also coupled with analysis of the 3D structure of the modern day homologs in order to generate a contact map for the target ancestral protein in question. For example, if segment of modern-day homolog 10–15 and 20–26 have identical residues with those of ancient sequence and there is a contact between 10 and 20, we use the contact 10, 20. In order to translate these contact maps between each other, we take into account insertions, deletions and differences in numbering from the sequence alignment. Finally, the consensus contacts across all maps (i.e. contacts overlap in all modern day homologs) are taken as the contact map for the ancestral proteins. The contact map includes both residue-residue distance contacts and also dihedral angle variations. This contact map is later used to couple with FRODA [47,72] during simulations that collapse the assembled secondary structural motifs into folded units.

II. Generation of unfolded assembled secondary structural motifs. The individual secondary structure elements are connected by building loops in extended conformation between secondary structures adjacent in sequence. We use a Monte Carlo technique in ZAMF [63] to build these loops and generate hundreds of unique conformations each with maximized radii of gyration, as shown in Fig. 4. Using many initial structures has the advantage of unbiasing the results from any individual initial structure.

III. Generation of collapsed folded conformations using geometric constraint-based FRODA. Each of these unique, "open" structures is then run in a FRODA simulation that enforces hydrophobic collapse through attractive perturbations between specific hydrophobic residue pairs in the previously mentioned contact map. No hydrophobic residues within loops are chosen and contacts within the same secondary structural motif are not considered a contact pair. During the simulation each of the residue-residue contacts are perturbed together if their separation distance in >7.0 Å. The run is prematurely ended if all the contacts from the contact list are found to be within a 7.0 Å cutoff distance at any time during the simulation. An additional hydrophobic collapse of all hydrophobic residues is done via a Monte Carlo accept/reject method with Boltzmann weighting between subsequent snapshots based on the difference of radius of gyration of hydrophobic residues. Other parameters of the FRODA simulation, such as momentum run-on between subsequent steps, remain the same as outlined in previous work [45].

The final collapsed structures from the FRODA simulations are then clustered into representative structures using a *k*-means clustering algorithm based on a 1.0 Å RMSD between atomic positions. These representative structures are scored and sorted based on both the radius of gyration of hydrophobic residues and also the number of hydrophobic contacts (<7.0 Å) (Fig. S6).

IV. Refinement and selection of the most native-like folded structure using ZAMF. We then move on to the refinement stage of ZAMF. The refinement stage involves a reservoir REMD (r-REMD) [73] step to both determine the most native conformation and also to further refine all conformations. The replicas and reservoir are filled with structures that are sorted according to the hydrophobic scoring function mentioned above. We then run multiple simulations where we narrow the conformational search space to avoid entrapment in local minima through residue-residue contact restraints based on the contact map of the ancestral protein. The local contacts are applied before the nonlocal ones to allow local refinement to occur before global refinement (tertiary structure). This approach is motivated by a hierarchical folding mechanism (search mechanism of ZAM). The restrained simulation is ran for 5 ns with replicas from 270K to 450K in the AMBER96 force field with generalized born implicit solvent model [74]. The residue-residue constraint is approximated to be at the center of mass of the residue and the

force constant is 0.5 kcal/(mol Å^2). After the restrained run, an unrestrained simulation with identical parameters is then run for at least 5 ns. After 5 ns, a convergence analysis is done, and if the protein is converged no further simulation is completed. If it is not, an additional 2 ns of simulation is run and convergence is checked. Continued 2 ns simulations are repeated until the protein has converged. The most dominant structure at the lowest replica is chosen as our prediction at the end of convergence. Our refinement protocol works well for ancestral sequences since their structure is close to modern day homologs whose structures are known. In other extreme cases where the starting initial model has lower resolution (i.e. 6–7 Å RMSD) from the original structure, our refinement protocol may fail and need additional alterations in order to reach to higher resolution structures.

Since we also generate an extensive amount of trajectory data, we use the unrestrained converged trajectories to analyze the dynamics of the ancestral structure as explained in detail below.

Principal Component Analysis for Identifying Functionally Important Dynamics

Convergence is critical and, as such, a sample window of 1 ns is slid along the trajectory at 0.5 ns intervals and Principal Component Analysis is done. The PCA is done by first aligning and centering each snapshot of the trajectory to remove the translations and rotations, generating a matrix $\mathbf{X_n}$ for each sampling window

$$\mathbf{X}_\mathrm{n} = \mathbf{x}_\mathrm{n} - \langle \mathbf{x}_\mathrm{n} \rangle \tag{1}$$

where $\mathbf{x_n}$ are 3N dimensional position vectors and the $< >$ denote a time average for a specific sampling window. Then, the covariance matrix of that sampling window, $\mathbf{C_{n,n}}$, is calculated by

$$\mathbf{C}_\mathrm{n,n} = \langle \langle \mathbf{X}_\mathrm{n} \rangle \langle \mathbf{X}_\mathrm{n} \rangle^\mathbf{T} \rangle \tag{2}$$

From the covariance matrix, the matrix of eigenvectors ($\mathbf{V_n}$) and the matrix of eigenvalues ($\mathbf{\Lambda_n}$) are

$$\mathbf{V}_\mathrm{n}^{-1} \mathbf{C}_\mathrm{n,n} \mathbf{V}_\mathrm{n} = \Lambda_\mathrm{n} \tag{3}$$

The eigenvectors and eigenvalues are sorted in order of decreasing eigenvalue and only the top 30 are kept as, once converged, any higher order (faster fluctuation/smaller positional deviations) are not relevant in determining biologically relevant large scale motion of the protein [75]. The reduced set of principal components is then

$$\mathbf{M}_\mathrm{n} = \mathbf{V}_\mathrm{n}^\mathbf{T} \mathbf{X}_\mathrm{n}^\mathbf{T} \tag{4}$$

The fluctuation profile along each mode is simply the Δr of each residue in that mode. By plotting these against each other, we confirm convergence when the Pearson correlation coefficient, P_{ij}, of the trajectory for sampling window i ($\mathbf{X}_\mathrm{i}$) and sampling window j ($\mathbf{X}_\mathrm{j}$) is >0.8

$$P_{i,j} = \frac{\mathbf{C}_\mathrm{i,j}}{\sigma_\mathrm{i} \sigma_\mathrm{j}} \tag{5}$$

σ_i and σ_j are the standard deviations of their trajectories. If the run has not converged it is continued until convergence is confirmed over a 3 ns window (Fig. S3). Using the Saguaro high performance computer at Arizona State University, a 250 residue protein with 40 temperature replicas (1 logical core per replica) finishes just under 300 ps/day. The most native like structures are assumed to be those that dominate the lowest temperature replica, while those in higher temperature replicas are dismissed.

After confirming convergence, in order to obtain the dynamics difference between the most collective motions (i.e. slowest frequency fluctuation profiles) of these three ancestral structures we apply the Singular Value Decomposition (SVD) technique to the matrix of dynamics profiles, $\mathbf{G}$ (i.e. the dynamics profile of each protein will be the column in the matrix, and each super-element, ik corresponds the X, Y, and Z fluctuations of the k^{th} residue in the sequence of protein i).

$$\mathbf{G}_\mathrm{n} = \left[\frac{\mathbf{V}_{\mathrm{n,\,Protein}i}}{\Lambda_{\mathrm{n,\,Protein}i}}, \frac{\mathbf{V}_{\mathrm{n,\,Protein}j}}{\Lambda_{\mathrm{n,\,Protein}j}}, \frac{\mathbf{V}_{\mathrm{n,\,Protein}k}}{\Lambda_{\mathrm{n,\,Protein}k}}, \ldots \right] \tag{6}$$

G matrix includes most collective modes of (i.e. global motion) individual proteins that we obtained separately from REMD trajectories. With construction of the G matrix our goal is to cluster the proteins with similar global motion. Since global dynamics (i.e. most spatially extensive collective mode) is most related to the function, proteins with similar global dynamics should cluster together and execute similar function. In order to do clustering we perform an SVD on G matrix

$$\mathbf{G}_\mathrm{n} = \mathbf{U}_\mathrm{n} \mathbf{S}_\mathrm{n} \mathbf{W}_\mathrm{n}{}^{-\mathbf{T}} \tag{7}$$

The first through mth values in each column of $\mathbf{W}$ can be plotted against each other to visualize the dynamic space occupied by each protein.

Supporting Information

Figure S1 The RMSD versus time plot for experimental structures of AncCR and AncGR2.
(PDF)

Figure S2 The Mean Square Displacement between our predicted structures for AncCR-AncGR1 (blue), AncCR-AncGR2 (green), and AncGR1-AncGR2 (red).
(PDF)

Figure S3 The plot of most collective mean square fluctuation of different sliding windows.
(PDF)

Figure S4 The dynamics of the experimental AncCR, AncGR1, and AncGR2 structures plotted in a reduced subspace.
(PDF)

Figure S5 Plot and ribbon diagram of the dynamics of the single mutation variant of human ferritin protein characterized by the slowest collective mode.
(PDF)

Figure S6 Radius of gyration of the hydrophobic residues versus the RMSD from the experimentally determined structure of AncCR for a single FRODA run.
(PDF)

Table S1 RMSD from Native Before and After REMD Simulation.
(PDF)

Author Contributions

Conceived and designed the experiments: TJG DWF MFT SBO. Performed the experiments: TJG. Analyzed the data: TJG ZNG SBO. Wrote the paper: TJG SBO.

References

1. James LC, Tawfik DS (2003) Conformational diversity and protein evolution - a 60-year-old hypothesis revisited. Trends Biochem Sci 28: 361–368.
2. Ma B, Shatsky M, Woflson HJ, Nussinov R (2009) Multiple diverse ligands binding at a single protein site: A matter of preexisting populations. Protein Sci 11: 184–197.
3. O'Brien PJ, Herschlag D (1999) Catalytic promiscuity and the evolution of new enzymatic activities. Chem Biol 6: R97–R105.
4. Jimenez R, Salazar G, Yin J, Joo T, Romesberg FE, et al. (2003) Protein dynamics and the immunological evolution of molecular recognition. Proc Nat Acad Sci U S A 101: 3803–3808.
5. Zimmerman J, Oakman EL, Thorpe IF, Shi X, Abbyad P, et al. (2006) Antibody Evolution Constains Conformational Heterogeneity by Tailoring Protein Dynamics. Proc Nat Acad Sci U S A 103: 13722–13727.
6. Radkiewicz JL, III CLB (2000) Protein Dynamics in Enzymatic Catalysis: Exploration of Dihydrofolate Reductase. J Am Chem Soc 122: 225–231.
7. Hespenheide BM, Rader AJ, Thorpe MF, Kuhn LA (2002) Identifying protein folding cores from the evolution of flexible regions during unfolding. J Mol Graphics Modell 21: 195–207.
8. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, et al. (2003) Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Science 21: 1364–1368.
9. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, et al. (2008) De novo computational design of retro-aldol enzymes. Science 319: 1387–1391.
10. Bolon DN, Mayo SL (2001) Enzyme-like proteins by computational design. Proc Nat Aca Sci U S A 98: 14274–14279.
11. Looger LL, Dwyer MA, Smith JJ, Hellinga HW (2003) Computational design of receptor and sensor proteins with novel functions. Nature 423: 185–190.
12. Joachimiak LA, Kortemme K, Stoddard BL, Baker D (2006) Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. J Mol Biol 361: 195–208.
13. Lazar GA, Dang W, Karki S, Vafa O, Peng JS, et al. (2006) Engineered antibody Fc variants with enhanced effector function. Proc Nat Aca Sci U S A 103: 4005–4010.
14. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, et al. (2005) Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. J Am Chem Soc 127: 1346–1347.
15. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.
16. Chen R, Davydov EV, Sirota M, Butte AJ (2010) Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. PLoS One 5: e13574.
17. Bromberg Y, Yachdav G, Rost B (2008) SNAP predicts effect of mutations on protein function. Bioinformatics 24: 2397–2398.
18. Harms MJ, Thornton JW (2010) Analyzing protein structure and function using ancestral gene reconstruction. Curr Opin Struct Biol 20: 360–366.
19. Donald JE, Shakhnovich (2009) SDR: a database of predicted specificity-determining residues in proteins. Nucleic Acids Res 37(suppl 1): D191–194.
20. Chakrabarti S, Lanczycki CJ (2007) Analysis and prediction of functionally important sites in proteins. Protein Sci 16: 4–13.
21. Buske FA, Their R, Gillam EM, Boden M (2009) In silico characterization of protein chimeras: relating sequence and function within the same fold. Proteins 77: 111–120.
22. Yang Z, Kumar S (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substition rates among sites. Mol Biol Evol 13: 650–659.
23. Liberies (2007) Ancestral Sequence reconstruction. USA: Oxford University Press.
24. Yokoyama S, Tada T, Zhang H, Britt L (2008) Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Nat Aca Sci U S A 105: 13480–13485.
25. Yokoyama S, Yang H, Starmer WT (2008) Molecular basis of spectral tuning in the red- and green-sensitive (M/L WS) pigments in vertebrates. Genetics 179: 2037–2043.
26. Field SF, Matz MV (2010) Retracing evolution of red flourescence in GFP-like proteins from Faviina corals. Mol Biol Evol 27: 225–233.
27. Ugalde JA, Chang BS, Matz MV (2004) Evolution of coral pigments recreated. Science 305: 1433.
28. Gaucher EA, Govindarajan S, Ganesh OK (2008) Paleotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451: 704–707.
29. Gaucher EA, Thomson JM, Burgan MF, Benner SA (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425: 285–288.
30. Kaiser SM, Malik HS, Emerman M (2007) Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. Science 316: 1756–1758.
31. Kuang D, Yao Y, Maclean D, Wang M, hampson DR, et al. (2006) Ancestral reconstruction of the ligand-binding pocket of Family C G protein-coupled receptors. Proc Nat Aca Sci U S A 103: 14050–14055.
32. Thomson JM, Gaucher EA, Burgan MF, Kee DWD, Li T, et al. (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. Nat Genet 37: 630–635.
33. Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461: 515–519.
34. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW (2007) Crystal Structure fo an Ancient Protein: Evolution by Conformational Epistasis. Science 317: 1544–1548.
35. Xia Y, Levitt M (2004) Simulating protein evolution in sequence and structure space. Curr Opin Struct Biol 14: 202–207.
36. Kar G, Keskin O, Gursoy A, Nussinov R (2010) Allostery and population shift in drug discovery. Curr Opin Pharmacol 10: 715–722.
37. Bakan A, Meireles LM, Bahar I (2011) ProDy: Protein Dynamics Inferred from Theory and Experiments. Bioinformatics 27: 1575–1577.
38. Liu Y, Gierasch LM, Bahar I (2010) Role of Hsp70 ATPase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with NEFs. PLoS Comp Biol 6: PMID: 20862304.
39. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307: 1113–1143.
40. Tokuriki N, Oldfield CJ, Uversky VN, Berezovsky IN, Tawfick DS (2009) Do viral proteins possess unique biophysical features. Trends Biochem Sci 34: 53–59.
41. Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 19: 596–604.
42. Jackson CJ, Foo J-L, Tokuriki N, Agriat L, Carr PD, et al. (2009) Conformational sampling, catalysis, and evolution of bacterial phosphotriesterase. Proc Nat Aca Sci U S A 106: 21631–21636.
43. Bhattacharyya M, Vishveshwara S (2011) Probing the Allosteric Mechanism in Pyrrolysyl-tRNA Synthetase Using Energy-Weighted Network Formalism. Biochemistry 50: 6225–6236.
44. Ghosh A, Vishveshwara S (2008) Variations in Clique and Community Patterns in Protein Structures during Allosteric Communication: Investigation of Dynamically Equilibrated Structures of Methionyl tRNA Synthetase Complexes. Biochemistry 47: 11398–11407.
45. Glembo TJ, Ozkan SB (2010) Union of Geometric Constraint-Based Simulations with Molecular Dynamics for Protein Structure Prediction. Biophys J 98: 1046–1054.
46. Ozkan SB, Wu GA, Chodera JD, Dill KA (2007) Protein folding by zipping and assembly. Proc Nat Aca Sci U S A 104: 11987–11992.
47. Wells S, Menor S, Hespenheide B, Thorpe MF (2005) Constrained geometric simulation of diffusive motion in proteins. Phys Biol 2: S127–S136.
48. Farrell DW, Speranskiy K, Thorpe MF (2010) Generating Stereochemically Acceptable Protein Pathways. Proteins 78: 2908–2921.
49. Li Y, Suino K, Daugherty J, Xu HE (2005) Structural and Biochemical Mechanisms for the Specificity of Hormone Binding and Coactivator Assembly by Mineralocorticoid Receptor. Mol Cell 19: 367–380.
50. Cozzetto D, Kryshtafovych A, Tramontano A (2009) Evaluation of CASP8 model quality predictions. Proteins 77: 157–166.
51. Jones TA, Kleywegt GJ (1999) Comparative Modeling: Assessment. Proteins 3: 30–46.
52. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A (2005) Assessment of Predictions Submitted for the CASP6 Comparative Modeling Category. Proteins 7: 27–45.
53. Peterson EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. J Comp Chem 13: 1605–1612.
54. Wintrode PL, Zhang D, Vaidehi N, Arnold FH, Goddard WAI (2003) Protein dynamics in a family of laboratory evolved thermophilic Enzymes. J Mol Biol 327: 745–757.
55. Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. Nature 11: 572–583.
56. McIntosh BE, Hogenesch JB, Bradfield CA (2010) Mammalian Per-Arnt-Sim proteins in environmental adaptation. Annu Rev Physiol 72: 625–645.
57. Zheng W, Brooks BR, Thirumalai D (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Nat Aca Sci U S A 103: 7664–7669.
58. Tawfik DS (2010) Messy biology and the origins of evolutionary innovations. Nat Chem Biol 6: 692–696.
59. Campagnoli MF, Pimazzoni R, Bosio S, Zecchina G, DeGobbi M, et al. (2002) Onset of cataract in early infancy associated with the 32G→C transtion in the iron responsive element of L-ferritin. Eur J Pediatr 161: 499–502.
60. Curtis ARJ, Fey C, Morris CM, Bindoff LA, Ince PG, et al. (2001) Mutation in the gene encoding ferritin light polypeptide causes dominant adult-onset basal ganglia disease. Nat Genet 28: 350–354.
61. Foglieni B, Ferrari F, Goldwurm S, Santambrogio P, Castiglioni E, et al. (2007) Analysis of ferritin genes in Parkinson's disease. Clin Chem Lab Med 45: 1450–1456.
62. Wang Z, Li C, ellenburg M, Soistman E, Ruble J, et al. (2006) Structure of human ferritin L chain. Acta Crystallogr D D62: 800–806.
63. Shell M, Ozkan S, Voelz V, G Wu KD (2009) Blind Test of Physics-Based Prediction of Protein Structures. Biophys J 96: 917–924.
64. Cygler M, Schrag JD, Sussman JL, Harel M, Silman I, et al. (1993) Relationship between sequence conservation and three-dimensional structure in a large family of esterases, lipases, and related proteins. Protein Sci 2: 366–382.

65. Redfern OC, Dessailly B, Orengo CA (2008) Exploring the structure and function paradigm. Curr Opin Struct Biol 18: 394–402.
66. Sadowski MI, Jones DT (2009) The sequence-structure relationship and protein funciton prediction. Curr Opin Struct Biol 19: 357–362.
67. Mehta PK, Heringa J, Argos P (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. Protein Sci 4: 2517–2525.
68. Biggadike K, Bledsoe RK, Coe DM, Cooper TW, House D, et al. (2009) Design and x-ray crystal structures of high-potency nonsteroidal glucocorticoid agonists exploiting a novel binding site on the receptor. Proc Nat Acad Sci U S A 106: 18114–18119.
69. Seitz T, Thoma R, Schoch GA, Stihle M, Benz J, et al. (2010) Enhancing stability and solubility of the glucocorticoid receptor ligand-binding domain by high-throughput library screening. J Mol Biol 403: 562–77.
70. Stehlin-Gaon C, Willmann D, Zeyer D, Sanglier S, Dorsselaer AV, et al. (2003) All-trans retinoic acid is a ligand for the orphan nuclear receptor ROR beta. Nat Struct Biol 10: 820–825.
71. Suino-Powell K, Xu Y, Zhang C, Tao YG, Tolbert WD, et al. (2008) Doubling the size of the glucocorticoid receptor ligand binding pocket by deacylcortivazol. Mol Cell Biol 28: 1915–1923.
72. Farrell DW, Speranskiy K, Thorpe MF (2010) Generating stereochemically acceptable protein pathways. Proteins 78: 2908–2921.
73. Roitberg A, Okur A, Simmerling C (2007) Coupling of Replica Exchange Simulations to a Non-Boltzmann Structure Reservoir. J Phys Chem B 111: 2415–2418.
74. Pearlman D, Case D, Caldwell J, Ross W, Cheatham T, III, et al. (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.
75. Amadei A, Linssen ABM, Berendson HJC (1993) Essential Dynamics of Proteins. Proteins 17: 412–425.