A Multiscale Hybrid Evolutionary Algorithm to Obtain Sample-based Representations of Multi-basin Protein Energy Landscapes

Rudy Clausen
Department of Computer Science,
George Mason University,
Fairfax, VA 22030
rclausen@gmu.edu

Amarda Shehu*
Department of Computer Science,
Department of Bioengineering
School of Systems Biology
George Mason University,
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT

The emerging picture of proteins as dynamic systems switching between structures to modulate function demands a comprehensive structural characterization only possible through an energy landscape treatment. Only sample-based representations of a protein energy landscape are viable in silico, and sampling-based exploration algorithms have to address the fundamental but challenging issue of balancing between exploration (broad view) and exploitation (going deep). We propose here a novel algorithm that achieves this balance by combining concepts from evolutionary computation and protein modeling research. The algorithm draws samples from a reduced space obtained via principal component analysis of known experimental structures. Samples are lifted from the reduced to an all-atom structure space where they are then mapped to nearby local minima in the all-atom energy landscape. From an algorithmic point of view, this paper makes several contributions, including the design of a local selection operator that is crucial to avoiding premature convergence. From an application point of view, this paper demonstrates the utility of the proposed evolutionary algorithm to advance understanding of multi-basin proteins. In particular, the proposed algorithm makes the first steps to answering the question of how sequence mutations affect function in proteins at the center of proteinopathies by providing the energy landscape as the intermediate explanatory link between protein sequence and function.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; J.3 [**Computer Applications**]: Life and Medical Sciences

General Terms

Algorithms

Keywords

protein energy landscape; multi-basin proteins; evolutionary algorithm; structurization; principal component analysis; multiscale modeling; decentralized selection

*Corresponding Author

BCB'14, September 20–23, 2014, Newport Beach, CA, USA.

ACM 978-1-4503-2894-4/14/09 ...$15.00.
http://dx.doi.org/10.1145/2649387.2649390.

1. INTRODUCTION

There is increasing evidence that proteins are dynamic molecules populating diverse stable and semi-stable structures to modulate their biological function and participate in complex chemical processes in the cell [3,16]. Modulation can be achieved through fast and small or slow and large structural fluctuations. Angstrom and sub-angstrom displacements are harnessed by enzymes into productive events [10, 39]. Other proteins switch between structures several angstroms away [1,18,21]. While the employment of structural diversity for functional diversity enriches the relationship between structure and function, it also challenges our understanding of how mutations affect function in proteinopathies.

The view that mutations cause loss of function by essentially removing the ability of a protein to assume an appropriate structure is rather simplistic. While loss of function or misfunction may be explained via loss of a crucial stable structure in many proteinopathies [36], the most complex human diseases, such as cancer, Amyotrophic lateral sclerosis (ALS), and others, involve proteins that switch between different structures in their natural state [11]. How do mutations cause dynamic proteins to misbehave? Revealing the relationship between sequence, structures, and loss, mis, or gain of function necessitates obtaining a comprehensive and detailed structural characterization that goes beyond the single-structure view of a protein [34].

Mapping out the different structural states a protein has at its disposal for biological activity is possible through an energy landscape treatment of structure modeling. Made possible by Dill and Wolynes, this treatment, founded upon statistical mechanics, provides a unifying framework for structure modulations in proteins [9,32]. By organizing structures into states associated with energy basins, the energy landscape provides a rationale for why certain structural states may be thermodynamically-favored over others for the pur-

pose of function and how this changes upon perturbations due to the presence of a ligand, cellular stress, changes in the environment, or mutations to the sequence [26].

Only sample-based representations of the protein energy landscape are viable in silico. Sampling-based search algorithms, while in principle able to obtain such representations, have to address the fundamental but challenging issue of how to spend computational resources when navigating a vast, high-dimensional, non-linear, and multimodal search space. The fundamental issue is how to balance between drilling down into minima (exploitation) while obtaining a broad view of the search space (exploration), so no stable or semi-stable states are missed due to premature convergence.

The literature on stochastic search or optimization algorithms for protein structure modeling is rich. While a review is beyond the scope of this paper, there is recently renewed interest in stochastic optimization under the umbrella of evolutionary computation for protein structure modeling [34]. Evolutionary algorithms (EAs), relying on the key idea of evolving a population of structures towards low-energy ones over generations, have been shown to be powerful for challenging problems, such as loop modeling and *de novo* structure prediction, when equipped with domain-specific expertise on representations of protein chains and state-of-the-art energy functions [19, 20, 27–31].

In this paper we propose a novel EA to explore the energy landscape of a protein and reveal energy basins. The algorithm strikes the right balance between exploration and exploitation by combining concepts from evolutionary computation and protein modeling research. The proposed EA draws samples from a reduced space obtained via principal component analysis (PCA) of known structures caught for a protein (whether in wildtype or variant form) in the wet laboratory. The driving hypothesis is that wet-lab wildtype and variant structures delineate the structure space of relevance and can be exploited to focus a search algorithm. Analysis of accumulated variance is conducted to determine the dimensionality and the axes of the search space (the principal components - PCs). Leveraging of wet-lab structures in this way allows reigning in computational complexity, as fewer dimensions (PCs) are explored in comparison to hundreds or thousands when using cartesian- or angular-based representations of protein chains. However, the algorithm implements multiscale modeling, as it lifts samples from the reduced to an all-atom structure space, where all-atom structures are then mapped to nearby local minima in the all-atom energy landscape.

The proposed EA contains several novel algorithmic components, including the leveraging of wet-lab structures to define a reduced structure space amenable to an informative grid-based structurization, a multiscale local improvement operator that switches between the reduced and all-atom search space to map samples into nearby local minima in the all-atom energy/fitness landscape, and a decentralized local selection operator exploiting the structurization. Analysis in this paper demonstrates that this operator, which essentially pitches newly-generated child samples to compete only with parent samples in neighboring cells of the structurization, is key to avoiding premature convergence. It is worth noting that the employment of a meaningful structurization is another novel algorithmic contribution in this paper, as typically arbitrary spatial structurizations are employed in spatially-structured EAs [7]. For the purpose of categorization, the EA proposed in this paper is a spatially-structured hybrid/memetic EA due its employment of structurization and a local improvement operator.

From an application point of view, this paper demonstrates the utility of the proposed EA to advance modeling and understanding of multi-basin proteins that exploit small or large structural displacements to carry out complex biological functions. In particular, the proposed EA makes the first steps towards answering the question of how sequence mutations affect function in proteins involved proteinopathies by providing the protein energy landscape as the intermediate explanatory link in the relationship between protein sequence and function.

We demonstrate the ability of the proposed EA to advance knowledge on the human Superoxide dismutase 1 (SOD1) enzyme, whose sequence mutations have been linked to familial ALS [5]. The energy landscape reconstructed by the algorithm for the wildtype and the dominant ALS-causing mutation in the US population provides a structural and energetic basis for the hypothesis that the mutation causes a toxic gain of function. Additional testing on multi-basin proteins exhibiting larger structural displacements (of several angstroms), such as HIV-1 Protease and Calmodulin (CaM), suggests the algorithm is scalable and can map known structural states onto the energy landscape, even revealing new ones. The results presented in this paper support the argument that the proposed algorithm extends the applicability of EAs to more challenging but also more powerful molecular modeling settings beyond *de novo* structure prediction that are of direct relevance to understanding disease.

The rest of this paper is organized as follows. The proposed EA is described in detail in section 2. Analysis of its performance and the energy landscapes it reveals on several proteins is provided in section 3. The paper concludes with a summary and directions of future work in section 4.

2. METHOD

The proposed EA follows the baseline framework in evolutionary computation, evolving a population of samples or individual over generations towards individuals of high fitness. An initial population is constructed through some mechanism to initialize the search. In each generation, parents are selected and subjected to reproductive operator(s) to produce offspring or child individuals. In a hybrid EA, each offspring is subjected to a local improvement operator that maps an offspring to a nearby local minimum prior to adding it to the population. As EAs mimick evolution, a selection operator implements natural selection, under which offspring compete with all or a subset of parents. The surviving individuals initialize the population for the next generation. Typically, as in the proposed EA, the process is repeated for a fixed number of generations.

A crucial ingredient of the proposed EA is its employment of a reduced representation of an individual to improve the computational efficiency of the reproductive operator. Specifically, an individual is a projection of a CA-trace of a protein structure onto a few PCs; the latter are obtained by conducting PCA prior to the execution of the algorithm over known wet-lab wildtype and variant structures of the protein of interest. This representation, detailed in section 2.1, is employed to define a reproductive operator that perturbs a parent in a randomly drawn vector in the PC space, resulting in a child. Details are provided in section 2.2.

The child is subjected to a local improvement operator, whose task is to improve the energetic profile of the child, essentially by moving it to a nearby structure residing in a local minimum of the energy landscape. Prior to that, the PC-based representation of the child needs to be lifted from the reduced space to an actual all-atom structure space, where the energy can be evaluated, and energetic minimization can be employed to implement the local improvement operator. The lifting mechanism employs multiple scales, essentially extracting first the CA trace from the PC-based representation of the generated child, then reconstructing the backbone from the CA trace, and then adding side chains onto the reconstructed backbone. The resulting all-atom (child) structure is then subjected to an energetic refinement protocol. Details are provided in section 2.3.

In a baseline EA, a centralized or global selection operator is employed, which pitches generated child individuals to compete with all parents for survival of the fittest. The fitness score in the proposed EA is the all-atom potential energy as evaluated through the Rosetta all-atom score12 function [17]. In the proposed EA, a decentralized local selection operator is employed instead to maintain structural diversity and avoid premature takeover of a population by a few lowest-energy structures. The key idea is to limit the pool of parents with which a child competes. The mechanism through which the proposed EA limits this pool is via structurization, as the PC map over which parent and child structures are projected readily provides information on which structures are neighboring and which ones are not. A two-dimensional structurization is employed here, essentially imposing a grid over the space of projections of parent and child structures over the top two PCs. Various neighborhood structures are investigated to define a neighborhood in the grid, so that a child can compete only with parents in its neighborhood. Details are provided in section 2.3.

Last, the construction of an initial population is key to proper initialization of an EA, and we detail it in section 2.5 before relating implementation details in section 2.6.

2.1 PC-based Representation

Wildtype and variant structures of a protein of interest are extracted from the Protein Data Bank (PDB) [2]. A consensus chain length is defined (possibly by excising few termini amino acids), structures whose chains miss internal amino acids are removed, and only variants with no more than a maximum number of mutations are considered. CA traces are extracted from the structures that pass the criteria, aligned to a reference trace and centered. The covariance matrix is subjected to Singular Value Decomposition. While PCA is generally not guaranteed to be effective, the EA only proceeds if at least 50% of the variance can be captured with the top two principal components (PCs). The EA directly searches in the low-dimensional PC map of m dimensions, ensuring that m PCs are sufficient to capture 90% of the variance in the original structure data. That is, each individual in the EA is an m-dimensional vector, with each element denoting the coordinate of the (CA trace) structure represented by that individual on the m axes of the PC map/space.

2.2 Reproductive Operator

The coordinates of an individual selected to serve as parent are perturbed to obtain a child as follows. A maximum step size $s_{\max}$ is defined. For each of the m coordinates of the parent, a step size s_i is sampled at uniform in $[-s_{\max}, +s_{\max}]$. This is then scaled according to the variance captured by the axes/PCs, as in: $s_{i,\text{scaled}} = s_i \cdot \frac{Var(PC_i)}{Var(PC_1)}$. Given that the PCs are ordered from the highest to the lowest variance, the idea is to carry out larger perturbations in the axes that capture more of the variance of the original structure data, thus preserving the scaling. Given the step size obtained this way for each coordinate of a parent individual, the corresponding coordinate $PC_{i,\text{child}}$ of its child is obtained as: $PC_{i,\text{child}} = PC_{i,\text{parent}} + s_{i,\text{scaled}}$. We note that each of the N parents in a population are subjected to this reproductive operator to obtain N offspring.

2.3 Local Improvement Operator

The reproductive operator operates in a reduced space, and the child traces it obtains may be invalid and need to be energetically improved. Multiscaling is used to do so. The first step is to rebuild the rest of the protein backbone from the CA trace of the child. We employ BBQ [12], one of the top backbone reconstruction protocols, to do so. Side chains are then added to the resulting backbone structure via the Rosetta *relax* protocol [17]. This protocol also allows conducting short energetic refinements of the all-atom structure while constraining motions of the backbone, which we employ here to improve a child structure and effectively map it to a local minimum of the all-atom (Rosetta score12) energy landscape. We note that various aspects onto whether the backbone reconstruction is accurate and whether the all-atom refinement step preserves the backbone have been studied by us before in the context of a randomized tree search for structure sampling [4] and deemed effective.

2.4 Local Selection Operator

To prevent premature convergence, we employ a local selection operator instead of a global one. After being improved, a child structure is not compared by its score12 value to all parents in a population, but only to a subset of parents deemed to be structurally similar to the child. Rather than use expensive structural comparison measures, we employ a simple yet informative structurization over the PC search space explored by the EA to determine neighbors. A grid is imposed on the top two PCs, effectively allowing us to map each EA individual in a cell of a two-dimensional (2d) structurization. The size of each cell grid is a parameter, and its value is controlled to allow separation of individuals while ensuring that not many cells are empty.

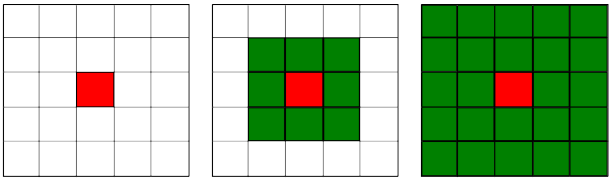

Figure 1: Illustration of some neighborhoods: the cell populated by the child is drawn in red; cells in green are additional neighboring cells when increasing the C parameter.

Neighborhoods can be defined over the structurization, through the use of a neighborhood size parameter, C. Three such neighborhoods are illustrated in Figure 1. The local se-

lection operator compares a child only to parents in a given Cx neighborhood, where x is a parameter. In section 3 we justify the selection of neighborhood size, as it has a direct effect on retainment of diversity over generations to avoid premature convergence. We note that in the event that no parents are in a given neighborhood, the child is compared to all parents in the population. Thus, for each child generated, either the child or a parent in its neighborhood are removed from the set of N parent and N child individuals. This reduces the population back to N individuals, which then serve as parents for the next generation. We note that the technique employed here to prevent premature convergence is also referred to as crowding in the evolutionary computation community [24].

2.5 Initial Population

The starting set S of CA traces subjected to PCA may be less than the size N of a population in the EA. Therefore, more structures need to be generated for the initial population. Moreover, these traces are extracted from structures of possibly different sequences (wildtype, variants) as opposed to one sequence of interest (whether that is the wildtype or a desired variant sequence). For this purpose, the initial population for a sequence of interest is obtained as follows. The CA traces in S are "threaded" onto the sequence of interest (this is the reason for a consensus length), projected onto the top m PCs, and then subjected to the multiscaling and local improvement operator. The latter two steps are repeated on randomly drawn individuals (the Rosetta *relax* protocol is a stochastic simulated annealing protocol, so different results are obtained) until the population reaches the desired size N and can now serve as the initial population for our EA.

2.6 Implementation Details

The algorithm is implemented in C/C++ and run on a 16 core red hat linux box with 3.2GhZ HT Xeon CPU and 8GB RAM. Run time ranges from 35 to 67 hours for protein chains ranging from 99 to 150 amino acids. Population size N=500 structures, and the algorithm is run for 100 generations (convergence in lowest fitness/energy values per generation was reached, data not shown). The rest of the parameter settings are listed in Table 1.

Table 1: Parameter values in our EA

System	$s_{\max}$	Cell Size	C
SOD1	2	2×2	49
HIV-1 Protease	1	1×1	49
CaM	10	10×10	25

3. RESULTS

We investigate here 3 proteins, SOD1, HIV-1 Protease, and CaM. We choose these systems due to their different sizes, from 99 to 150 amino acids, the availability of diverse structures in the PDB (from slightly over 1Å to 10Å away in structure space), as well as the richness of sequence mutations documented for SOD1 and HIV1-Protease. On each of these proteins we analyze various aspects of the energy landscape captured by the proposed EA, such as the location of known and novel structures, as well as implications for function in disease-involved variants.

3.1 Data Collection

Only X-ray structures were collected from the PDB for HIV-1 Protease. For SOD1 and CaM, NMR solution structures were allowed to further enrich the initial set of structures. The wildtype sequence of each of these proteins was obtained from the UniProt [22], and this sequence was used as reference to both define the sequence length and limit the number of mutations among available variants (and thus structures collected) to 3. Any structures with missing residues were rejected. These criteria allowed collecting 254 structures for HIV1-Protease, 697 structures for CaM, and 186 structures for SOD1. All SOD1 structures were subjected to the PCA. For HIV-1 Protease and CaM, a randomly-drawn subset was removed prior to running PCA, reserving these structures for an analysis on whether the EA can reproduce them. So, 54 of the 254 collected structures of HIV-1 Protease and 197 of the 697 collected structures of CaM were removed and reserved for this analysis.

3.2 PCA is Effective: Analysis of Variance

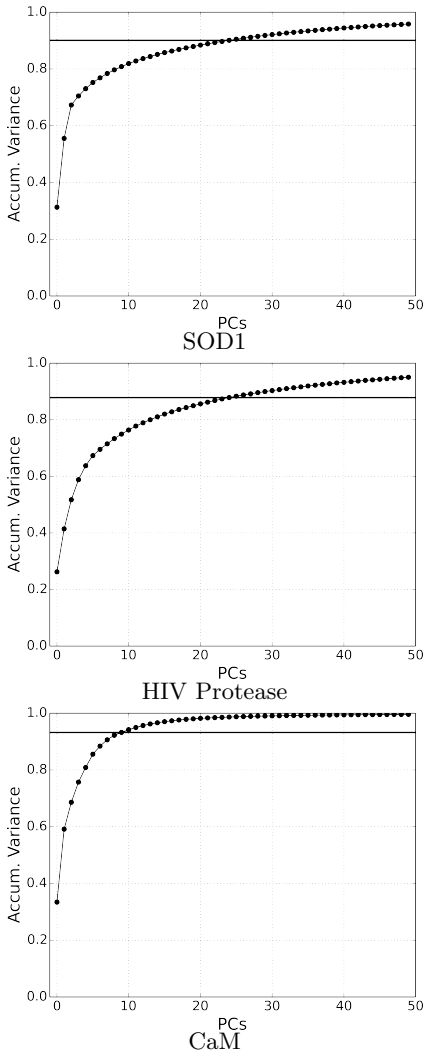

Figure 2: The accumulation of variance shows that PCA is effective for each of the proteins studied here. The horizontal line shows the 90% cutoff used to determine the top m PCs to be used as directions of search by the proposed EA.

PCA is an effective dimensionality reduction technique for each of the proteins considered here. Figure 2 draws the accumulation of variance and shows that the top two PCs capture close to 50% of the variance in the original structure data. This is important, as the first two PCs are used to define the structurization for the local selection operator. Moreover, the accumulation of variance analysis shown in Figure 2 is used to determine the number m of PCs for the reduced search space over which the proposed EA operates. A cumulative variance of 90% is reached at 25, 25, and 10 PCs for SOD1, HIV-1 Protease, and CaM, respectively.

3.3 Local Selection Retains Diversity

A detailed analysis has been conducted to determine the neighborhood size, C, for the local selection operator (data not shown). We summarize this analysis by demonstrating in Figure 3 that a local neighborhood C25 allows retaining more structural diversity longer than a global neighborhood, where a child is compared to the entire parent population. Figure 3 draws the average structural dissimilarity between any two structures in the population in a given generation and tracks this value across the generations. Structural dissimilarity is measured via Euclidean distance over the PC1 and PC2 coordinates of each individual in the population.

Figure 3 shows that structural diversity is lost very quickly when a global selection operator is used, indicating rapid convergence to a few minima in this setting. In contrast, a local selection operator with a C25 neighborhood preserves structural diversity longer, thus allowing the exploration of possibly diverse minima in a protein's energy landscape. This comparison provides the foundation for why we employ a local selection operator in the proposed EA.

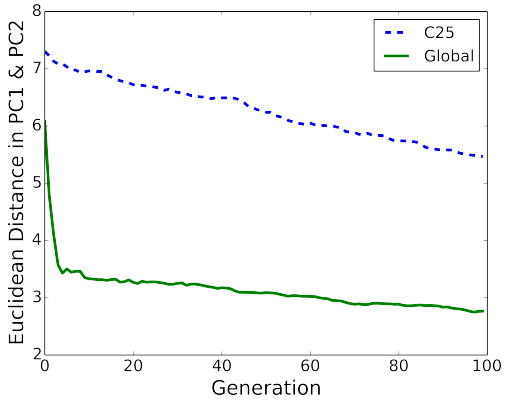

Figure 3: The average pairwise Euclidean distance over PC1 and PC2 coordinates of individuals in a generation is computed and tracked across the generations. Two settings are compared, one where a global selection operator is employed (green line), and one where a local selection operator with neighborhood structure C25 is employed (dashed blue line).

3.4 Analysis of EA-obtained Energy Landscape of Wildtype and Variant SOD1

Figure 4(a) shows all collected SOD1 structures superimposed on the top two PCs. The projections are color-coded based on the sequence variants they represent. The PC map in Figure 4(a) shows that PC1 separates the structures into two clusters. On the right one finds structures of mutant sequences, such as H46R and A4V . On the left, one finds those of mutant sequences, such as C111S, L38V, and I113T. Excluding the points labeled "Other" (mutations not tracked), the wildtype "WT" is the sequence for which structures are found in both clusters. This map suggests that there are two distinct structural states for SOD1. This is further supported by the results in Figure 4(b), which shows a bimodal distribution of pairwise lRMSD values between all collected structures (lRMSD stands for least Root-Mean-Squared-Deviation and is a popular yet imperfect structure dissimilarity measure [23]). These results are in full agreement with experimental studies, where SOD1 is shown to switch between an apo and holo structural state [37].

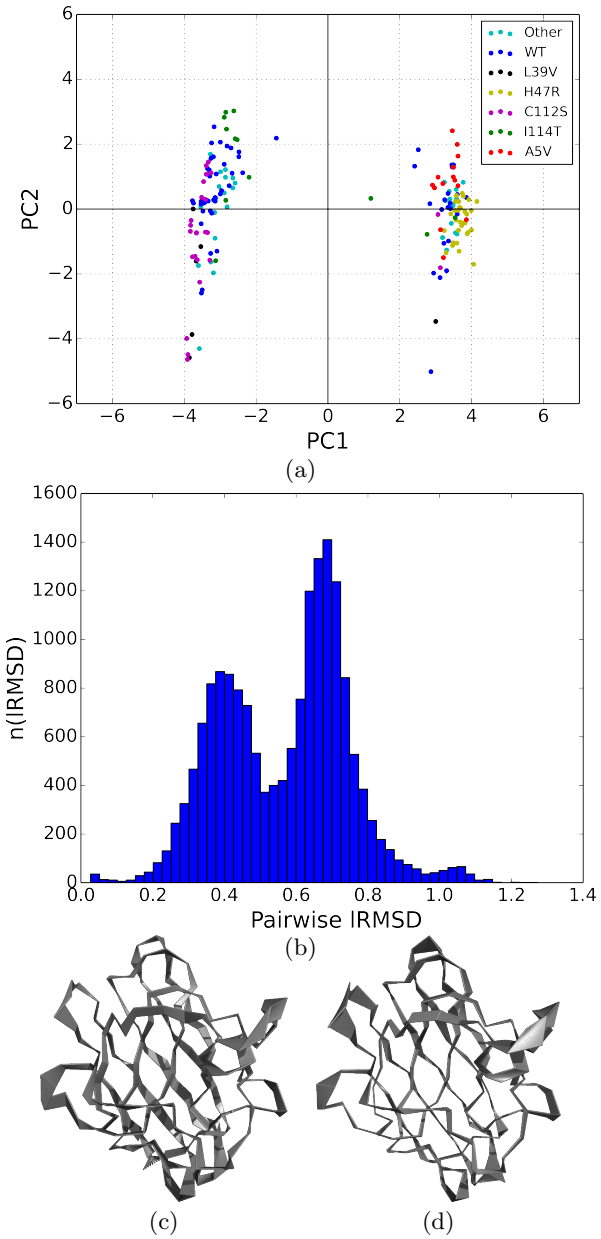

Figure 4: (a) The PCA map of SOD1 wet-lab structures shows two distinct clusters separated by PC1. (b) The distribution of pairwise lRMSDs among all structures is also bimodal. (c)-(d) Structural displacements along PC1 in (c) and along PC2 in (d) are illustrated on a selected structure.

The structural displacements that PC1 and PC2 capture on SOD1 are shown in Figure 4(c)-(d), respectively. Both PC1 and PC2 capture displacements in regions nearby the Cu and Zn-binding sites of SOD1 (center of structure) and loops, pointing to allosteric regulation of metal-binding in SOD1, in agreement with other studies [6].

The projection in Figure 4(a) of wet-lab structures for SOD1 in a reduced space suggests that only the wildtype has been captured to access both structural states in the wet laboratory. However, other variants may have access to more structures than what is documented in the PDB. A detailed energy landscape needs to be reconstructed. Therefore, EA is applied to the wildtype sequence and then separately to the US-dominant ALS-causing mutation, A5V, in SOD1. The energy landscapes are then compared, by visualizing the set of structures in the final generation of EA on each sequence, in Figure 5(a) for the wildtype sequence, and Figure 5(b) for the A5V variant. The structures are projected onto the top two PCs, and they are color-coded based on their energy values (Rosetta score12).

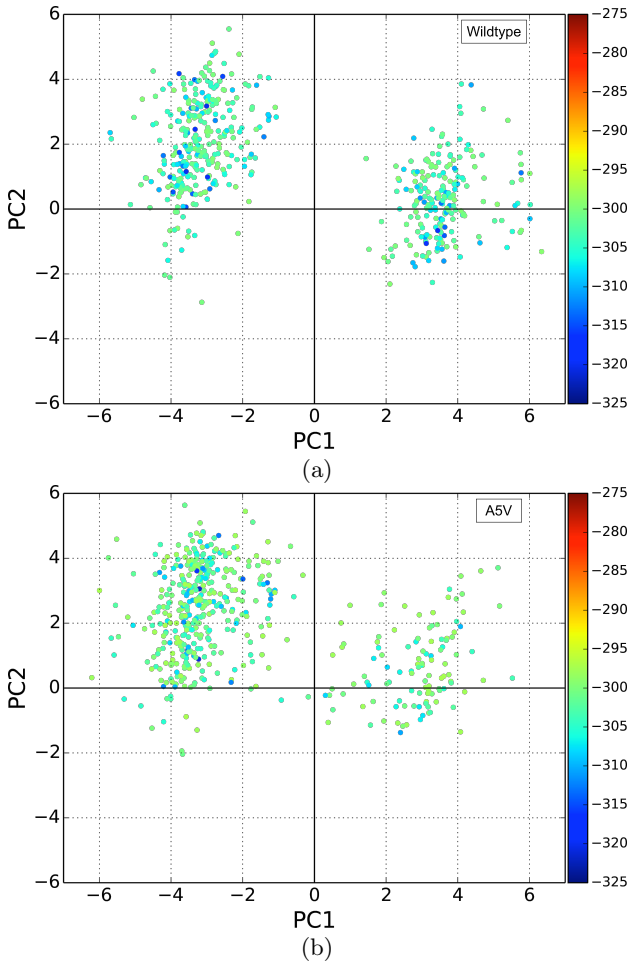

Figure 5: The last generation of computed structures by the EA is shown projected on the top two PCs for the SOD1 wildtype in (a) and the A5V mutant sequence in (b). Structures are color-coded by their Rosetta score 12 energy values.

At first glance, the landscapes of these two sequences seem very similar. There is, however, a significant difference. There are many more low-energy structures that seem to be available to the A5V variant as opposed to the wildtype. These structures are found along the 0 line in PC1. They seem to bridge between the two structural states for the A5V variant. In contrast, the EA cannot find low-energy structures to bridge between the two structural states in the wildtype. These results indicate that the A5V mutant sequence can switch more rapidly between two structural states through many more low-energy structures, essentially being more unstable than the wildtype and exhibiting a toxic gain of function. This conclusion seems to provide the structural basis for observations made in the wet laboratory, where the A5V mutant has been found to have higher tendency to engage in aggregation [8, 14, 33].

3.5 Analysis of EA-obtained Energy Landscape of HIV-1 Protease

Out of the 254 collected wet-lab structures for HIV-1 Protease, 54 drawn at random are withheld from the PCA. The structural displacements along each of the top two PCs obtained by PCA of the remaining 200 structures are illustrated on a selected structure and shown in Fig. 6(a)-(b). Displacement along PC1 corresponds to the vertical (open-close) motion of the top flaps that surround the active site. Displacement along PC2 corresponds to the orthogonal, horizontal movement of the flaps surrounding the active site. This is in agreement with other PCA-based analysis (of structures obtained via Molecular Dynamics) of the structural flexibility of HIV-1 Protease [38].

Figure 6(c) shows all 254 collected HIV-1 Protease structures superimposed on the top two PCs. The projections are color-coded: red to indicate the 200 structures subjected to PCA and green to indicate the 54 withheld. The distribution of structures in the reduced space in Figure 6(c) shows no distinct structural states. The distribution of pairwise CA lRMSDs between all collected structures of HIV1-Protease, shown in Figure 6(d), supports this, showing a unimodal distribution with a maximum pairwise lRMSD of 1.4Å. The distribution of TM-scores shown inside (TM-score is a measure that reflects localized structural changes [42] as opposed to lRMSD distributing changes on all atoms) shows that the dissimilarity is localized to specific regions on the chain. The pairwise TM-scores are all higher than 0.8 (a TM-score higher than 0.5 conveys high structural similarity [40]). This analysis suggests that HIV-1 Protease has a wide basin, with a wide range a structures thermodynamically-available to the wildtype sequence.

Application of the EA to the wildtype sequence to reconstruct its energy landscape supports the above observations. Figure 6(e) shows the set of structures in the final EA generation, projected on the top two PCs and color-coded by their energies. As can be observed from the distribution of these structures in the reduced space and their energies, the EA has not converged to any distinct structural states/minima in the landscape. Instead, a large collection of structures of comparable energies has been obtained. Given that HIV-1 Protease has a fast mutation rate and yet forms stable monomers (decoupled from its dimerization in the enzyme active state), these findings point to the conclusion that the landscape has a wide basin. It is worth noting that the modeling here is limited to the monomeric unit of the naturally-occurring dimer. However, even though 54 structures were withheld from the PCA, the EA has been able to capture the same structure space. Figure 6(f) shows the highest TM-

score between a withheld structure and any structure in the final generation of the EA. All these values are above 0.8, suggesting that many of the wet-lab structures are near the energy basin reconstructed by the EA.

3.6 Analysis of EA-obtained Energy Landscape of CaM

The final system, CaM, demonstrates the ability of our EA to reproduce landscapes for proteins with multiple structural states more than 13Å apart (pairwise lRMSD between structures with PDB ids 1CLL and 2F3Y is 13.44Å). As detailed in the accumulation of variance analysis above, due to these large concerted structural changes, only 10 PCs are needed to capture ≈90% of the variance.

The structural displacements along PC1 and PC2 are shown in Figure 7(a)-(b), respectively. Displacement along PC1 captures folding and unfolding motions of the intermediate α-helical domain, as well as opening and closing of the N- and C-terminal domains. Displacement along PC2 captures primarily motions of the N- and C-terminal domains. These findings are in agreement with other studies of CaM, which show that the structural variability is localized to folding and unfolding of the helical domain and concerted motions between the terminal domains [35].

Figure 7(c) summarizes the energy landscape reconstructed by the EA by projecting the structures computed in the final generation on the top two PCs. The structures are color-coded by their Rosetta score 12 energies. The 197 structures withheld from the PCA are also shown, drawn in gray. Figure 7(c) shows two distinct lowest-energy basins. Labeling by their PDB ids representative wet-lab structures that map to the location of these two basins in the PC map reveals that the two basins correspond to the closed, bound states of CaM. The deepest basin corresponds to the ligand-bound state (2F3Y), and the next deepest, though wider, basin corresponds to the protein-bound state (1NWD, in particular, is a structure found bound to the a dimer of glutamate decarboxylase C-termini [41]).

The EA-obtained landscape allows drawing several more conclusions about CaM. The superimposition of the withheld structures shows that the ligand-bound structures of CaM are actually shifted in the structure space by about 7Å. Figure 7(d) superimposes structures in this basin over the wet-lab structure under PDB id 2F3Y to show that the Rosetta energy function retains the overall topology of the ligand-bound state, and the shift is due to structural fluctuations in loops and termini. The structures in the next deepest basin are also shown, superimposed on the wet-lab structure with PDB id 1NWD in Figure 7(e). There is higher structural variability in this basin, but the overall topology is closed. Inspection of all collected wet-lab structures that map to the location of this second-deepest basin (data not shown) reveals that this basin captures all protein-bound structures of CaM, not just the one with PDB id 1NWD.

The landscape shows a third group of higher-energy structures not in a well-defined basin. These include the two structures that represent the calcium-bound and calcium-free (apo) states of CaM, labeled in Figure 7(c) by PDB ids 1CLL and 1CFD, respectively. There are low-energy structures bridging the transition from the calcium-binding and calcium-free states to the protein-bound (second-deepest basin) state in the landscape, but a transition to the ligand-bound (deepest basin) seems more difficult; the EA does not reveal low-energy structures to facilitate this transition. The location of the apo/calcium-free state (see label 1CFD) in the PCA map suggests that this state may bridge the transition from the calcium-bound (1CLL) to the ligand-bound (2F3Y) state. This is in agreement with other studies suggesting the calcium-bound state mediates the transition from the apo to the ligand-bound state [13]. Moreover, putting all the results obtained for CaM together, the transition from the calcium-bound to the protein-bound state probably occurs at a faster rate than that to the ligand-bound (most thermodynamically-stable) state due the presence of more low-energy structures to bridge the transition.

4. CONCLUSION

This paper has proposed a novel stochastic optimization algorithm to explore the structure space of intrinsically dynamic proteins that switch between structural states to modulate their biological activity. The algorithm is a hybrid, spatial EA with several novel algorithmic components, including leveraging of wet-lab structures to define a reduced search space, employment of multiscaling to map structures onto the underlying energy landscape of an amino-acid sequence of interest, and exploitation of PCA to define a structurization for a local selection operator.

The analysis presented in this paper indicates that the proposed EA is powerful and not prone to premature convergence. This characteristic allows it to reconstruct in detail energy landscapes of known multi-basin proteins. The algorithm is able to provide a link between sequence mutations and changes in function through the energy landscapes, as demonstrated on SOD1. It is scalable and able to explain relationships between known structural states of proteins, such as CaM, where structural dissimilarity between its functional states exceed 10Å. Work has also begun on employing Markov-based analysis of the generated structures to model the structural transitions and estimate transitions rates. This additional information is valuable at providing further detailed insight into the role of different structural states and how these states, and therefore, protein function is affected by sequence mutations [25].

The results presented here are encouraging and promise that the employment of stochastic optimization algorithms can provide detailed answers relevant to understanding the relationship between sequence, structure, and function in dynamic proteins, but several challenges and limitations remain. First, the algorithm is limited to proteins on which enough structures have been captured in the wet laboratory in order to define a credible reduced structure space. With more and more wet-lab efforts dedicated to disease-involved proteins, this is less likely to be a limitation in the future. The employment of PCA is a limitation, as other structure spaces may be nonlinear. Possibly fewer dimensions could underly the true structure space, and using PCA may add unnecessary dimensions. Nonlinear dimensionality reduction techniques that allow directly sampling in the reduced space, a key feature of the proposed EA, will be investigated. Finally, the results presented are conditionally dependent on the energy function employed, and they may be affected by biases in the particular function employed. It is possible, for instance, that the differences between the two deepest basins revealed by the proposed EA on CaM may be less striking when another energy function is employed. Future work will investigate the robustness of the results across dif-

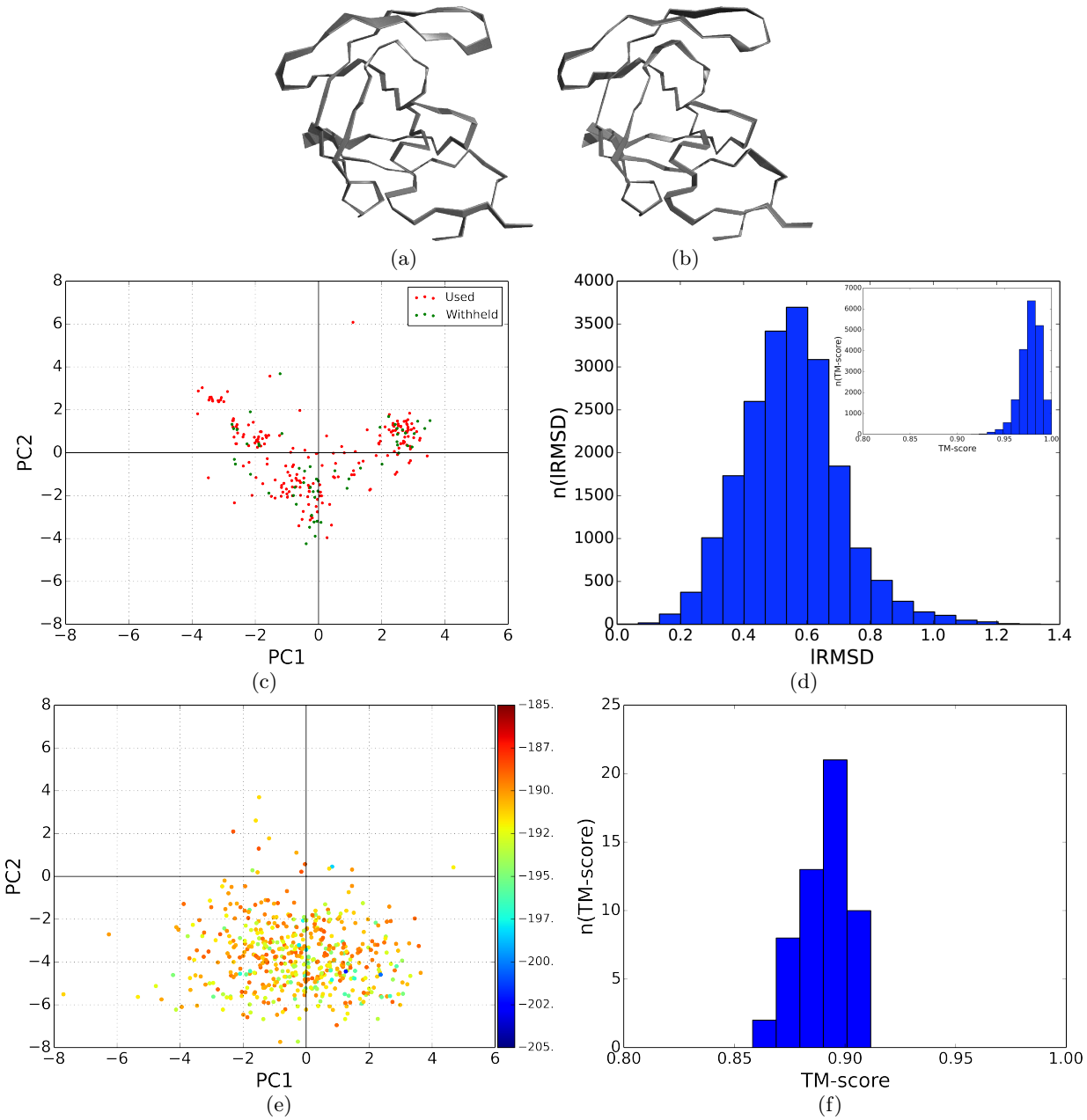

Figure 6: (a)-(b) Structural displacements along PC1 in (a) and along PC2 in (b) are illustrated on a selected structure for HIV-1 Protease. (c) HIV-1 protease wet-lab structures subjected to the PCA are color-coded in red, and those withheld are in green. (d) The distributions of pairwise lRMSDs and TM-scores among all collected structures are shown. (e) Structures in the last generation of the EA for HIV-1 Protease are projected on the top two PCs and color-coded by their energies. (f) Structures withheld from the PCA are compared to those in this final generation in terms of TM-score. The distribution of the maximum TM-score obtained for each withheld structure is shown.

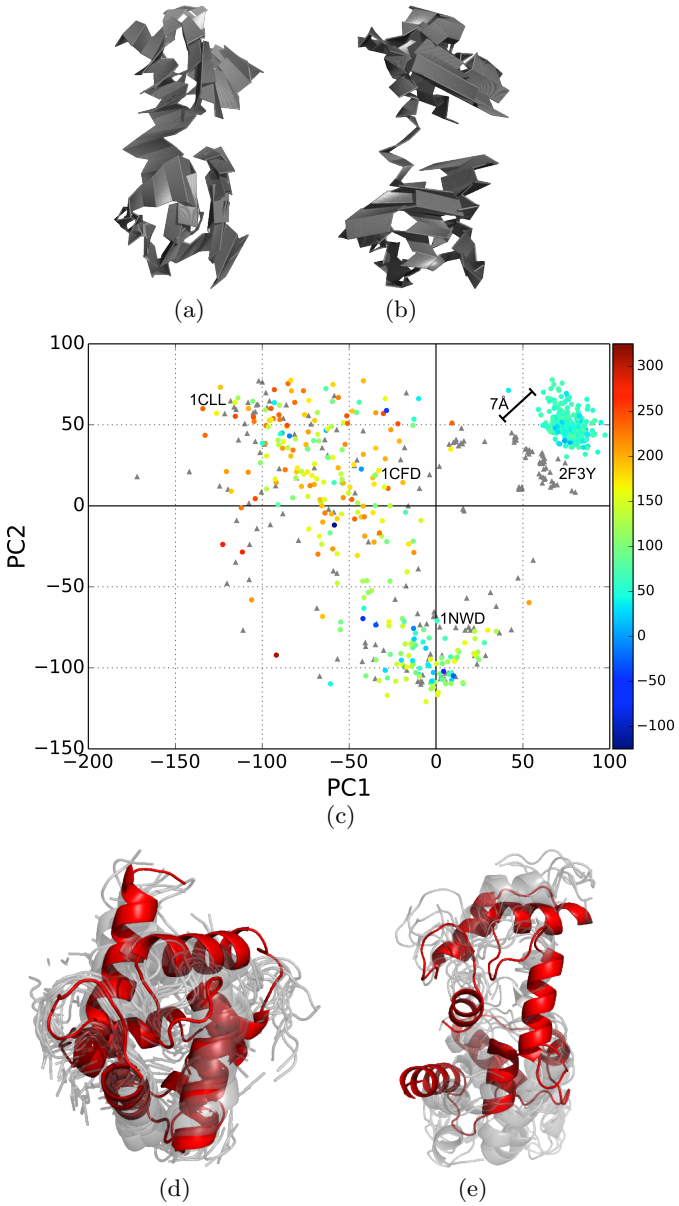

Figure 7: Structural displacements along PC1 in (a) and along PC2 in (b) are illustrated on a selected CaM structure (PDB id 1CFD). (c) The final population of structures generated by the EA for the wildtype sequence of CaM is shown projected on the top two PCs and color-coded by Rosetta score12 energy values. The 197 structures withheld from PCA are projected on the top two PCs and color-coded in gray. (d) Structures in the deepest basin revealed by the EA and corresponding to the closed ligand-bound state are superimposed (drawn in gray and transparent) over the representative closed ligand-bound structure of CaM (PDB id 2F3Y, drawn in opaque red). (e) Structures in the next deepest basin revealed by the EA and corresponding to another closed state of CaM (drawn in gray and transparent) are superimposed (drawn in gray and transparent) over a protein-bound state of CaM (PDB id 1NWD, drawn in opaque red). Structure rendering is done with VMD [15].

ferent protein energy functions.

Acknowledgment

This work is supported by NSF CCF No. 1016995, NSF IIS CAREER Award No. 1144106, and NSF CCF No. 1421001.

5. REFERENCES

[1] O. Beckstein, E. J. Denning, J. R. Perilla, and T. B. Woolf. Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open-closed transitions. *J. Mol. Biol.*, 394(1):160–176, 2009.

[2] H. M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, 10(12):980–980, 2003.

[3] D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem Biol*, 5(11):789–96, 2009.

[4] R. Clausen and A. Shehu. Exploring the structure space of wildtype ras guided by experimental data. In *ACM Conf on Bioinf and Comp Biol Workshops (BCBW)*, pages 757–764, Washington, D. C., September 2013.

[5] R. A. Conwit. Preventing familial ALS: a clinical trial may be feasible but is an efficacy trial warranted? *J Neurol Sci*, 251(1-2):1–2, 2006.

[6] A. Das and S. S. Plotkins. SOD1 exhibits allosteric frustration to facilitate metal binding affinity. *Proc. Natl. Acad. Sci. USA*, 110(10):3871–3876, 2013.

[7] K. A. De Jong. *Evolutionary Computation: A Unified Approach*. MIT Press, Cambridge, MA, 1st edition, 2006.

[8] M. DiDonato, L. Craig, M. Huff, M. Thayer, R. Cardoso, C. Kassmann, T. Lo, C. Bruns, E. Powers, J. Kelly, E. Getzoff, and J. Tainer. Als mutants of human superoxide dismutase form fibrous aggregates via framework destabilization. *J. Mol. Biol.*, 332(1):601–615, 2003.

[9] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4(1):10–19, 1997.

[10] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–121, 2005.

[11] A. Fernández-Medarde and E. Santos. Ras in cancer and developmental diseases. *Genes Cancer*, 2(3):344–358, 2011.

[12] D. Gront, S. Kmiecik, and A. Kolinski. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J. Comput. Chem.*, 28(29):1593–1597, 2007.

[13] J. Gsponer, J. Christodoulou, A. Cavalli, J. M. Bui, B. Richter, C. M. Dobson, and M. Vendruscolo. A coupled equilibrium shift mechanism in calmodulin-mediated signal transduction. *Structure*, 16(5):736–746, 2008.

[14] M. Hough, J. Grossmann, S. Antonyuk, R. Strange, P. Doucette, J. Rodriguez, L. Whitson, P. Hart, L. Hayward, J. Valentine, and S. Hasnain. Dimer

destabilization in superoxide dismutase may result in disease-causing properties: structures of motor neuron disease mutants. *Proc. Natl. Acad. Sci. USA*, 101(16):5976–5981, 2004.

[15] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics. *J. Mol. Graph. Model.*, 14(1):33–38, 1996. http://www.ks.uiuc.edu/Research/vmd/.

[16] K. Jenzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007.

[17] K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, and J. Meiler. Practically useful: What the rosetta protein modeling suite can do for you. *Biochemistry*, 49(14):2987–2998, 2010.

[18] D. Kern and E. R. Zuiderweg. The role of dynamics in allosteric regulation. *Curr. Opinion Struct. Biol.*, 13(6):748–757, 2003.

[19] Y. Li, I. Rata, and E. Jakobsson. Improving predicted protein loop structure ranking using a pareto-optimality consensus method. *BMC Struct Biol*, 10(22):1–14, 2010.

[20] Y. Li and A. Yaseen. Pareto-based optimal sampling method and its applications in protein structural conformation sampling. In *BOOKTITLE = AAAI Workshop,*, pages 32–37, Bellevue, Washington, July 2013.

[21] Q. Lu and J. Wang. Single molecule conformational dynamics of adenylate kinase: energy landscape, structural correlations, and transition state ensembles. *J. Am. Chem. Soc.*, 130(14):4772–4783, 2008.

[22] M. Magrane and the UniProt consortium. UniProt knowledgebase: a hub of integrated protein data. *Database*, 2011(bar009):1–13, 2011.

[23] A. D. McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr. A.*, 26(6):656–657, 1972.

[24] O. J. Mengshoel and D. E. Goldberg. The crowding approach to niching in genetic algorithms. *Evol Comput*, 16(3):315–354, 2008.

[25] K. Molloy, R. Clausen, and A. Shehu. On the stochastic roadmap to model functionally-related structural transitions in wildtype and variant proteins. In *Robotics: Science and Systems (RSS) Workshop*, pages 1–6, Berkeley, CA, 2014.

[26] K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 103(32):11844–11849, 2006.

[27] B. Olson and A. Shehu. Efficient basin hopping in the protein energy surface. In *IEEE Intl Conf on Bioinf and Biomed*, Philadelphia, PA, October 2012. 119-124.

[28] B. Olson and A. Shehu. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. *Proteome Sci*, 10(10):S5, 2012.

[29] B. Olson and A. Shehu. An evolutionary-inspired algorithm to guide stochastic search for near-native protein conformations with multiobjective analysis. In *AAAI Workshop*, pages 32–37, Bellevue, Washington, July 2013.

[30] B. Olson and A. Shehu. Multi-objective stochastic search for sampling local minima in the protein energy surface. In *ACM Conf on Bioinf and Comp Biol (BCB)*, pages 430–439, Washington, D. C., September 2013.

[31] B. Olson and A. Shehu. Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction. In *Intl Conf on Bioinf and Comp Biol (BICoB)*, Las Vegas, NV, 2014.

[32] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: the energy landscape perspective. *Annual Review of Physical Chemistry*, 48:545–600, 1997.

[33] T. Ratovitski, L. Corson, J. Strain, P. Wong, D. Cleveland, V. Culotta, and D. Borchelt. Variation in the biochemical/biophysical properties of mutant superoxide dismutase 1 enzymes and the rate of disease progression in familial amyotrophic lateral sclerosis kindreds. *Hum. Mol. Genet.*, 8(8):1451–1460, 1999.

[34] A. Shehu. Probabilistic search and optimization for protein energy landscapes. In S. Aluru and A. Singh, editors, *Handbook of Computational Molecular Biology*. Chapman & Hall/CRC Computer & Information Science Series, 2013.

[35] A. Shehu, L. E. Kavraki, and C. Clementi. Multiscale characterization of protein conformational ensembles. *Proteins: Struct. Funct. Bioinf.*, 76(4):837–851, 2009.

[36] C. Soto. Protein misfolding and neurodegeneration. *JAMA Neurology*, 65(2):184–189, 2008.

[37] R. W. Strange, S. Antonyuk, M. A. Hough, P. A. Doucette, J. A. Rodriguez, P. Hart, L. J. Hayward, J. S. Valentine, and S. Hasnain. The structure of holo and metal-deficient wild-type human cu, zn superoxide dismutase and its relevance to familial amyotrophic lateral sclerosis. *J. Mol. Biol.*, 328(4):877âĂŞ891, 2003.

[38] M. Teodoro and L. E. Kavraki. Understanding protein flexibility through dimensionality reduction. *J Comput Biol*, 10(3-4):617–634, 2003.

[39] M. Vendruscolo and C. M. Dobson. Dynamic visions of enzymatic reactions. *Science*, 313(5793):1586–1587, 2006.

[40] J. Xu and Y. Zhang. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[41] K. Yap, T. Yuan, H. Mal, T.K. AMD Vogel, and M. Ikura. Structural basis for simultaneous binding of two carboxy-terminal peptides of plant glutamate decarboxylase to calmodulin. *J. Mol. Biol.*, 328(1):193–204, 2003.

[42] Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.