A Data-driven Evolutionary Algorithm for Mapping Multi-basin Protein Energy Landscapes

RUDY CLAUSEN[1] and AMARDA SHEHU[1,2,3,*]

ABSTRACT

Evidence is emerging that many proteins involved in proteinopathies are dynamic molecules switching between stable and semi-stable structures to modulate their function. A detailed understanding of the relationship between structure and and function in such molecules demands a comprehensive characterization of their conformation space. Currently, only stochastic optimization methods are capable of exploring conformation spaces to obtain sample-based representations of associated energy surfaces. These methods have to address the fundamental but challenging issue of balancing computational resources between exploration (obtaining a broad view of the space) and exploitation (going deep in the energy surface). We propose a novel algorithm that strikes an effective balance by employing concepts from evolutionary computation. The algorithm leverages deposited crystal structures of wildtype and variant sequences of a protein to define a reduced, low-dimensional search space from where to rapidly draw samples. A multiscale technique maps samples to local minima of the all-atom energy surface of a protein under investigation. Several novel algorithmic strategies are employed to avoid premature convergence to particular minima and obtain a broad

[1]Department of Computer Science, George Mason University

[2]Department of Bioengineering, George Mason University

[3]School of Systems Biology, George Mason University

[*]Current address: Department of Computer Science, George Mason University, Fairfax, VA, USA 22030

view of a possibly multi-basin energy surface. Analysis of applications on different proteins demonstrates the broad utility of the algorithm to map multi-basin energy landscapes and advance modeling of multi-basin proteins. In particular, applications on wildtype and variant sequences of proteins involved in proteinopathies demonstrate that the algorithm makes an important first step towards understanding the impact of sequence mutations on misfunction by providing the energy landscape as the intermediate explanatory link between protein sequence and function.

Key words: Protein structure modeling, multi-basin energy landscape, evolutionary algorithm, dimensionality reduction, multiscale modeling.

1. INTRODUCTION

Increasingly, the accepted view of proteins is that of inherently dynamic molecules populating diverse thermodynamically-stable and semi-stable structures to modulate biological function and participate in various processes in the cell (Jenzler-Wildman and Kern, 2007; Boehr et al., 2009). Motions connecting functional structures of a protein can be fast and small or slow and large. For instance, fast motions in the angstrom or sub-angstrom range are often observed to be employed by enzymes (Eisenmesser et al., 2005; Vendruscolo and Dobson, 2006; Tousignant and Pelletier, 2004). Other motions allow proteins to switch between functional structures several angstroms away (Kern and Zuiderweg, 2003; Lu and Wang, 2008; Beckstein et al., 2009). The employment of structural diversity for function modulation is a phenomenon observed particularly in higher-order organisms to enrich the relationship between structure and function. However, this same enrichment challenges our ability to model and understand this relationship, particularly as it concerns elucidating the detailed role of protein sequence mutations in proteinopathies.

In dynamic proteins, it is much more difficult to understand how mutations result in misfunction or loss of function. In some proteinopathies, loss of protein function can be explained by loss of a crucial stable structure (Soto, 2003, 2008). However, proteins involved in some of the most complex human diseases, such as cancer, Amyotrophic lateral sclerosis (ALS), and others, switch between different functional structures in their wildtype (WT) form (Fernández-Medarde and Santos, 2011). How do variants cause misfunction? Answering this question requires obtaining a detailed structural characterization that goes beyond the single-structure view of a protein (Shehu, 2013).

Mapping out the menu of different functional structures that a protein has at its disposal for biological activity demands obtaining a comprehensive view of the conformation space and underlying energy surface. Computing the energy surface of a protein and then projecting it on few dimensions to visualize the energy landscape was introduced by Dill, Wolynes, and colleagues (Dill and Chan, 1997; Onuchic et al., 1997). By organizing structures via energetic states, the energy landscape provides a rationale for why certain structures may be thermodynamically-favored over others, and how this changes upon perturbations, such as presence of a ligand, cellular stress, environmental changes, or sequence mutations (Okazaki et al., 2006). The energy landscape view is, therefore, not only important to understand dynamic, multi-basin proteins, but it is also essential to elucidate the impact of mutations on function. Comparisons between energy landscapes reconstructed for WT and variant sequences of a protein may reveal the structural and energetic reasons behind misfunction in variants.

One of the challenges computational methods experience in obtaining a comprehensive view of the conformation space and underlying energy surface is that the conformation space is continuous, high-dimensional, and therefore not enumerable. As such, the conformation space can only be probed through a sample-based approach, where essentially a conformation is sampled at a time. The result is a sample-based representation of the protein energy surface that often comes at great computational cost. While a review of Molecular Dynamics (MD) and Monte Carlo (MC) methods for obtaining such representations is beyond the scope of this work, it is worth mentioning that all such methods are greatly affected by the structures used to initialize them. While many strategies exist to enhance their exploration capability beyond a small neighborhood around the initial structure in conformation space, the computational cost can well exceed a few weeks on a CPU (Adcock and McCammon, 2006).

Instead of the usual MD or MC methods, other stochastic optimization algorithms are devised specifically to address the issue of how to balance computational cost between obtaining a broad view of a vast, continuous, and high-dimensional conformation space while having the time to go deep down a non-linear and multi-basin (or multimodal) energy surface. This issue is also known as exploration versus exploitation, and addressing it is crucial to capturing many stable or semi-stable structural states of a protein and not converging prematurely to any particular state.

The exploration versus exploitation issue is the subject of much algorithmic research in stochastic optimization under the umbrella of evolutionary computation (EC) (De Jong, 2006). Evolutionary algorithms (EAs) originating in the EC community have been shown powerful for challenging problems, such as loop modeling, protein-ligand binding, and even *de novo* structure prediction (Shehu, 2013). While they are often designed to serve as black-box optimization tools for NP-hard problems, equipping EAs with domain-specific expertise, such as state-of-the-art protein representations and energy functions, has resulted in performance that rivals that of MC-based methods (Li et al., 2010; Olson and Shehu, 2012a,b; Li and Yaseen, 2013; Olson and Shehu, 2013, 2014).

Inspired by the recent performance of EAs for *de novo* structure prediction, in this paper we propose a novel EA to extend the *in silico* characterization of protein functional structures to multi-basin proteins. It is well-known that *de novo* structure prediction is a challenging problem even for small-to-medium size proteins with a single well-defined basin. Therefore, in this paper, the proposed EA is applied to a given protein sequence but exploits experimentally-available structures for the WT and other variant sequences of the protein under investigation. The key idea is that such structures, while reported on perhaps a different variant from the sequence under inves-

tigation, may serve as yet-to-be-discovered stable or semi-stable structures in the energy surface of the sequence under investigation. In particular, the proposed EA extracts from such structures information on the true dimensionality of the conformation space, its shape, and bounds.

We refer to the proposed algorithm as PCA-EA, as it uses a particular dimensionality reduction technique, Principal Component Analysis (PCA), to extract and so define a reduced search space from a collected set of experimentally-available structures for a protein under investigation. The input to PCA-EA is a particular protein sequence as well as a collection of structures found in the Protein Data Bank (PDB) (Berman et al., 2003) for that sequence and other variants of the protein. The output is an ensemble of conformations that are local minima in the all-atom energy surface of the input sequence.

The reduced search space allows PCA-EA to be computationally-efficient, as the algorithm draws samples from a space of few collective variables as opposed to hundreds or thousands of variables when using cartesian- or angular-based representations of protein chains. However, PCA-EA implements multiscale modeling, as it lifts drawn samples from the reduced space to the all-atom conformation space, where all-atom conformations are mapped to nearby local minima in the all-atom energy landscape. Other novel algorithmic components in PCA-EA allow it to delay convergence to a particular basin and instead explore the breadth of the energy surface.

From an application point of view, this paper demonstrates the utility of an EA to advance modeling and understanding of multi-basin proteins that exploit small or large structural displacements to carry out complex biological functions. Three proteins are selected that exhibit motions as small as 1.5Å and as large as 13Å. We demonstrate the ability of PCA-EA to advance knowledge on the human Superoxide dismutase 1 (SOD1) enzyme, whose sequence mutations have been linked to

familial ALS (Conwit, 2006). Additional testing on multi-basin proteins exhibiting larger structural displacements (of several angstroms), such as HIV-1 Protease and Calmodulin (CaM), suggests PCA-EA is scalable and can map known structural states onto the energy landscape, even revealing new ones.

The results presented here 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. In particular, PCA-EA makes the first steps towards answering the question of how sequence mutations affect function in proteins involved in proteinopathies by providing the protein energy landscape as the intermediate explanatory link in the relationship between protein sequence and function.

2. METHODS

As an EA, PCA-EA implements the key idea of evolving a population of samples or individual over generations towards individuals of high fitness. In each generation, a subset or all of the individuals are selected to serve as parents and subjected to reproductive operators. The resulting offspring either replace parents or compete with all or a subset of them for survival. Survival is based on a measure of fitness of each individual. Surviving individuals comprise the initial population for the next generation. Typically, this process is repeated for a fixed number of generations or until another stopping criterion is satisfied. The only population that is constructed through some other mechanism is the first population that initializes the algorithm.

Several important algorithmic components need to be defined. First, a mechanism is needed to construct the initial population. PCA-EA builds the initial population over a collection of experimentally-available structures. Second, a determination needs to be made on how to represent an individual. The choice of representation is directly related to the size and dimensionality of the search space. Rather than employing variables based on cartesian coordinates or dihedral angles of protein chains, PCA-EA employs collective variables that define a low-dimensional, reduced search space. Namely, collected experimental structures are subjected to a linear dimensionality reduction technique, PCA, to reveal the underlying axes of the search space. PCA-EA draws samples in this reduced space. Individuals are not structures but rather points in this reduced space revealed by the PCA. Third, once the representation and search space are defined, reproductive operators need to be specified to modify parents. PCA-EA makes use of asexual reproduction; that is, the operator modifies one parent in the reduced space to obtain an offspring. Fourth, a selection

mechanism is needed to obtain a set of individuals from the parents and their offspring to define the population for the next generation. The selection mechanism here is an overlapping one, where offspring compete with parents. Competition is based on fitness, and all-atom energy is used here to measure fitness. However, since individuals are not structures but points in the reduced space, a multiscale procedure is defined to map an individual to an all-atom conformation. Moreover, the procedure also improves the individual by mapping it to a conformation that is a nearby local minimum in the all-atom energy surface. Finally, once fitness values are obtained for parent and offspring, the selection mechanism ensures that offspring only compete with structurally-similar parents so as to preserve offspring longer and avoid take-over of the population by a few fittest parents; hence, retain structural diversity and avoid premature convergence. All these algorithmic components of PCA-EA are shown in a diagram in Fig. 1.

We now describe each algorithmic component, starting with the reduced representation in section 2.1, then the reproductive operator in section 2.2, the local improvement operator in section 2.3, the selection operator in section 2.4, and the specification of the initial population in section 2.5.

2.1. Representation of an Individual

Wildtype and variant structures of a protein of interest are extracted from the PDB. A consensus 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. Structures that pass these selection criteria are simplified to their CA traces, discarding all other backbone and side-chain atoms. The CA traces are aligned to a reference trace (arbitrarily selected to be the first trace in the set) through an optimal superimposition

procedure typically employed to calculate least root-mean-square-deviation (lRMSD) McLachlan (1972). The so-modified CA traces are then converted into atomic displacements by subtracting from them the average CA trace over the aligned set. The purpose for this data preparation is to capture internal structural fluctuations rather than differences due to rigid-body motions (translations and rotations in three-dimensional space). This data is stored in a matrix $A_{3k \cdot n}$, where k is the number of CA atoms (corresponding also to the number of amino acids in the protein sequence), and n is the number of structures collected.

A singular value decomposition of $1/\sqrt{n-1} \cdot A$ is conducted to obtain $1/\sqrt{n-1} \cdot A = U \cdot \Sigma \cdot V^T$. The procedure employed to conduct this decomposition is the *dgesvd* routine in lapack Anderson et al. (1990). The entire process described here, also referred to as PCA, essentially rotates the data to reveal principal axes in order of the variance they preserve (projecting data onto an axis reveals the variance captured by that axis). These axes, also referred to as principal components (PCs), are $3k$-dimensional vectors found in the columns of the U matrix. The Σ matrix contains in its diagonal the singular value σ_i for each corresponding PC_i. The *dgesvd* routine provides the PCs in the order of largest to smallest singular value. The singular values σ_i are square roots of eigenvalues e_i, which measure the variance of the data when projected onto PC_i. Analysis of the eigenvalues allows selecting a subset of the PCs in the eigenvalue-sorted ordering to capture a given accumulated variance. Namely, by sorting PCs from largest to smallest eigenvalue, one can estimate through $(\sum_{i=1,\ldots,i=j} e_i)/(\sum_{i=1,\ldots,n} e_i)$ the data variance that can be preserved if data are represented as j-dimensional vectors of entries that are projections over $\mathrm{PC}_1, \ldots, \mathrm{PC}_j$. Typically, PCA is considered effective if the top two PCs capture cumulatively more than 50% of the total variance. PCA-EA employs a cutoff of 90% cumulative variance to determine the set of m PCs that can be used as new search axes. That is, each individual in PCA-EA is an m-dimensional

point, with each element denoting the coordinate of the (CA trace) structure represented by that individual on the m axes of the PC map/space.

This reduced representation improves the computational efficiency of the reproductive operator, as typically m is much less than the number of variables that would have to be specified if cartesian coordinates or dihedral angles were to be used. While PCA is not guaranteed to be an effective reduction technique on all structure data, the reductions it provides are measurable over other variables that can be defined without specific insight onto a particular protein system at hand. While PCA-EA uses PCA to find general collective variables for any protein with functional structures in the PDB, other reduction techniques can be used. The algorithm can be used as a roadmap on how to integrate such information in an evolutionary algorithm.

2.2. Reproductive Operator

Each of the N parents in a population are subjected to the reproductive operator to obtain N offspring. This operator perturbs a parent in a randomly drawn vector in the PC space, resulting in an offspring. Specifically, the coordinates of an individual selected to serve as parent are perturbed to obtain an offspring as follows. A maximum step size $\lambda_{\max}$ is defined. For each of the m coordinates of the parent, a step size λ_i is sampled at uniform in $[-\lambda_{\max}, +\lambda_{\max}]$. This is then scaled according to the variance captured by the axes/PCs, as in: $\lambda_{i,\text{scaled}} = \lambda_i \cdot \frac{Var(PC_i)}{Var(PC_1)}$. 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 (the shape of the search space). Given the step size obtained this way for each coordinate of a parent individual, the corresponding coordinate $PC_{i,\text{offspring}}$ of its offspring is obtained as: $PC_{i,\text{offspring}} = PC_{i,\text{parent}} + \lambda_{i,\text{scaled}}$.

2.3. Fitness Evaluation and Local Improvement Operator

Once offspring are obtained, the objective of the local improvement operator is both to improve their energetic profile and evaluate their fitness. Since the reproductive operator operates in a reduced space, the offspring it obtains may correspond to an invalid, high-energy conformation. So, first offspring are mapped to conformations and then energetic refinement of these conformations is carried out. Since an offspring needs to first be mapped to a CA trace, then to a backbone, and then to an all-atom conformation, the mapping operates over various scales. First, the CA trace corresponding to an offspring o can be easily obtained as $o \cdot U^T + \langle \text{trace} \rangle$, where the $\langle \text{trace} \rangle$ vector contains the average CA trace (the latter is calculated as part of centering the data for PCA, as described in section 2.1).

Once the CA trace corresponding to an offspring is obtained, a backbone can be easily reconstructed. PCA-EA uses one of the top backbone reconstruction protocols, BBQ Gront et al. (2007). Side chains then need to be packed onto a backbone conformation. Various side-chain packing protocols exist but many use simplistic energy functions. One of the top current protocols that uses a state-of-the-art energy function is implemented as part of an energetic refinement in the *relax* procedure in the Rosetta structure prediction package Kaufmann et al. (2010). Because this package is open-source and written in C/C++, the *relax* procedure easily interfaces with PCA-EA. The procedure conducts a short Monte Carlo simulated annealing to obtain a local minimum conformation in the all-atom energy surface. Moreover, it allows restricting motions of the backbone, which we employ here in order to ensure that the resulting all-atom conformation corresponds to the offspring that was subjected to the local improvement operator. The result of this process is not only a local minimum all-atom conformation that is added to the PCA-EA recorded all-atom con-

formation ensemble Ω, but also a fitness value for the ofsspring, measured as the *score12* all-atom energy of its corresponding all-atom conformation.

It is worth noting that while the reduced search space is the same for all variant sequences of a protein under investigation, the local improvement operator associates a sequence-specific energy surface with obtained all-atom conformations. Hence, the Ω ensemble and associated *score12* energies are different from different application of PCA-EA on variant sequences of a protein. It is this feature that allows employing PCA-EA to compare energy surfaces of different variants of a protein.

2.4. Local Selection Operator

PCA-EA employs an overlapping evolutionary model, where offspring compete with parents for survival. However, instead of implementing a global/centralized selection operator, where offspring compete with all parents, PCA-EA employs a local/decentralized selection operator. The objective of this operator is to limit competition so as to increase the likelihood that suboptimal offspring will survive longer. In other words, the operator slows down take-over of a population by a few fittest individuals, thus delaying convergence in the interest of obtaining a broad view of the energy surface.

Competition is limited by allowing an offspring to compete only against structurally-similar parents. Instead of employing expensive structure comparison techniques, such as lRMSD, a coarse measure of structural similarity is estimated. Namely, all individuals, parent and offspring are projected onto the top two PCs. A grid is then laid over this map (also referred to as structurization of the search space), and cells of a given size are then defined over the grid. Individuals can be

considered structurally-similar or neighbors if they fall in the same cell or if they fall within a neighborhood of cells. An offspring is thus compared only to parents that fall in the same neighborhood. In the event that no parents are in a given neighborhood, the offspring is compared to all parents in the population.

Neighborhoods can be defined over the structurization through the use of a neighborhood size parameter, C. The local selection operator compares an offspring only to parents in a given Cx neighborhood, where x is a parameter. Part of our analysis in section 3 focuses on determining an effective value for this parameter in a trade-off between preserving structural diversity (exploration/breadth in search) and reaching regions of low energies (exploitation/depth in search). This particular approach that PCA-EA employs to prevent premature convergence is also referred to as the crowding approach to niching in EAs Mengshoel and Goldberg (2008).

2.5. Initial Population

CA traces of the collected experimentally-available structures are projected onto the top m PCs to yield individuals that constitute the initial population. However, the number of obtained individuals may be less than the desired size of a population. Therefore, more individuals need to be generated to populate the initial population. Additional CA traces are "threaded" onto the sequence of interest (this is the reason for a consensus length in the collection of structures from the PDB), projected onto the top m PCs, and then subjected to the 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 initial population reaches the desired size N. Analysis in section 3 investigates the effect of various population sizes to determine an effective one.

3. RESULTS

3.1. Experimental Setup

3.1.1. Systems of Study. We investigate here 3 proteins, SOD1, HIV-I Protease, and CaM. On SOD1 we investigate its WT sequence and three variants found in US and Asian populations. On HIV1-Protease and CaM we study the WT sequence. We choose these proteins due to their different sizes, from 99 to 150 amino acids and the availability of diverse structures in the PDB (from slightly over 1Å to 10Å away in structure space). On each of these proteins we analyze various aspects of the energy landscape reconstructed by PCA-EA, such as the location of known and novel structures, as well as implications for function in disease-involved variants.

3.1.2. Data Collection. Only X-ray structures are collected from the PDB for HIV-I Protease. For SOD1 and CaM, NMR solution structures are allowed to further enrich the collected set. The WT sequence of each of these proteins is obtained from the UniProt Magrane and the UniProt consortium (2011), and this sequence is used as reference to both define the sequence length and limit the number of mutations among available variants (and thus structures collected) to no more than 3. Any structures with missing internal amino acids are discarded. These criteria allow collecting 254 structures for HIV1-Protease, 697 structures for CaM, and 186 structures for SOD1. All SOD1 structures are subjected to the PCA. For HIV-I Protease and CaM, a randomly-drawn subset is removed prior to running PCA, reserving these structures for an analysis on whether PCA-EA could reproduce them in its computed ensemble. So, 54 of the 254 collected structures of HIV-I Protease and 197 of the 697 collected structures of CaM are removed and reserved for this

analysis.

3.1.3. Implementation Details. PCA-EA 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. Analysis below shows the effective population size, and neighborhood parameter C. The algorithm is run for 100 generations, but convergence is reached earlier on all systems. The parameter settings for each of the three protein systems studied here are listed in Table 1.

The rest of this section is organized as follows. In the first part of our analysis, we focus on various aspects of the algorithm, such as the effectiveness of the PCA, the impact of population size on performance, and the impact of the neighborhood size on retaining structural diversity and avoiding premature convergence. This analysis allows determining the effective values for the population size and neighborhood size parameters reported in Table 1. The second part of the analysis focuses on the results obtained by the algorithm on each of the three proteins selected here in their WT and disease-variant forms. A detailed investigation is conducted of features of energy landscapes and structural states obtained by PCA-EA.

3.2. Detailed Analysis of Parameter Value Selections in PCA-EA

3.2.1. Analysis of Variance to Determine Reduced Space. PCA is an effective dimensionality reduction technique for each of the proteins considered here. Fig.2 draws the accumulation of variance and shows that the top two PCs capture between 40% and 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. These two PCs are also used to project the PCA-EA ensemble on

two dimensions and visualize the energy landscape reconstructed by PCA on each system. The accumulation of variance analysis in Fig. 2 is also 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-I Protease, and CaM, respectively.

3.2.2. Analysis of Population Size in PCA-EA. A detailed analysis is conducted to determine an effective population size for each system. Typically, in EAs for structure modeling of small-to-medium size proteins, a population size in the hundreds is suggested Shehu (2013). Here we run the algorithm with three different population sizes, 300, 400, and 500. We measure two quantities to summarize the diversity (breadth) and energetic quality (depth) across generations. First, the Euclidean distance in the m-dimensional space of PCs is measured between any two individuals in a generation, and the average value is associated with a generation and plotted across generations. Second, the average *score12* value over all individuals in a generation is also recorded and plotted across generations.

The progression of this Euclidean-based measure of structural diversity is shown in Fig. 3(a1) for one of the systems here, SOD1, up to generation 50 (convergence is reached around generation 70 for this system). The progression of the energetic quality is shown in Fig. 3(b1) for the same system. Fig. 3(a1) shows that larger population sizes preserve structural diversity longer. This observation aligns with the expected behavior of EAs, where a larger population affords a broader view of the search space. Fig. 3(b1) also shows that larger population sizes reach lower-energy values in the energy surface. Thus, a larger population size provide better breadth/exploration and better depth/exploitation. Taken together, these results suggest that a population size of 500 is advantageous, and this analysis justifies our selection of population size of 500 for the rest of the

experiments and analysis in this paper.

3.2.3. Analysis of Neighborhood Size in Local Selection Operator. A detailed analysis has been conducted to determine the neighborhood size, C, for the local selection operator (keeping population size at 500). As before, the diversity and energetic quality of a generation are measured. The progression of these two quantities over generations is shown in Fig. 3(a2)-(b2) on one of the systems here that converges at generation 50. For comparison, in addition to the C9, C25, and C49 neighborhoods, a run of PCA-EA with a global selection operator is also analyzed (in this operator, an offspring competes with all parents, effectively having $C\infty$). Fig. 3(a2)-(b2) shows that the global selection operator results in rapid drop in diversity. Out of the different neighborhoods considered for the selection operator, in this particular system, either C25 or C49 are effective. It is worth noting that more rapid loss in diversity of C9 in the later generations is due to the lack of parents in particular neighborhoods. As PCA-EA starts converging, cells of the PC1-PC2 structurization become empty. In such cases, the local selection operator pitches an offspring against all parents. This analysis on the effect of the neighborhood size is conducted on each of the systems here (data not shown) to determine an effective value for the C parameter. The values yielded by this analysis on each of these systems are shown in Table 1.

3.3. Analysis of Applications of PCA-EA on Protein Systems

We now proceed with a detailed analysis of the results of PCA-EA on each of the protein systems considered here, starting with SOD1 WT and its variants.

3.3.1. Analysis of PCA-EA on WT and Variant SOD1. Fig. 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 Fig. 4(a) shows that PC1 separates the structures into two clusters. On the right one finds structures reported for the WT and variant sequences, such as H46R and A4V . On the left, one finds structures reported for the WT and variant sequences, such as C111S, L38V, G37R, and I113T. Excluding the points labeled "Other" (mutations not annotated), the WT and G37R are two sequences for which structures are found in both clusters. In particular, G37R seems to also occupy the middle of the plot.

The above observation on the organization of SOD1 functional structures is further supported by the results drawn in Fig. 4(b), which shows a bimodal distribution of pairwise CA lRMSDs between all collected structures. These results are in full agreement with experimental studies, where SOD1 is shown to switch between an apo and holo structural state Strange et al. (2003), which we refer to here as A and B from now on.

The PC1-PC2 projection of experimentally-available functional structures for SOD1 in Fig. 4(a) suggests that only the WT has been captured to access both structural states richly in the wet laboratory. However, other variants may have access to more structures than what is documented in the PDB. Energy landscapes need to be reconstructed. Therefore, PCA-EA is applied to the WT and then to three other variants. These variants include A4V (an alanine in position 4 in the WT is replaced with valine in this variant), which is the US-dominant ALS-causing variant, and two other variants, G37R and H46R, that are predominantly reported in Asia but are less understood with regards to pathogeneicity.

PCA-EA is applied to each of these four sequences to obtain four different conformation ensem-

bles. Conformations in each ensemble are analyzed in terms of their *score12* energy values. Only conformations with *score12* values no higher than −200 units are retained as functional conformations. This threshold is set by observing the variance of *score12* values of SOD1 experimentally-available structures that are threaded onto the WT sequence and are subjected to the *relax* procedure. The maximum obtained *score12* value is around −200 kcal/mol. Therefore, this value is considered the maximum at which a conformation can be determined functional/relevant.

An energy landscape is obtained with each subset of retained/functional conformations as follows. The landscape for each sequence is a projection of the energy surface (only of functional conformations) obtained by PCA-EA over the top two PCs for the purpose of visualization. This two-dimensional projection of the space of functional conformations is color-coded as follows. A grid is overlaid with cells of size 1. A cell is colored by the median *score12* value of the conformations that project to it. The bilinear interpolation in the *imshow* python utility is employed for this purpose. The color bar shows not the range of absolute *score12* energy values but instead the difference from the highest-energy value.

Fig. 5 shows the four energy landscapes thus constructed for each of the SOD1 sequences.

The Sod1 WT energy landscape shows two well-defined energy basins, labeled A and B (which correspond to the organization of experimentally-available structures in Fig. 4(a)), with a significant energy barrier in between. Some lower-energy regions break the barrier, suggesting that lower-energy paths can connect the two structural states A and B. The presence of these paths through the barrier, coupled with the slow gradient between the bottom of each of the basins and the energy barrier, may aid in a carefully-timed transition of SOD1 WT from one structural state to the other.

In comparison, the Sod1 A4V landscape has lost the well-defined energy basins observed in the WT but also has a lower-energy barrier between the two states. Compared to the WT, this barrier is not only lower in energy but also less substantial. This indicates that Sod1 A4V is able to switch more rapidly between the two structural states A and B through more low-energy structures, essentially being more unstable than the WT and so exhibiting a toxic gain of function. This conclusion seems to provide the structural basis for observations made in the wet laboratory, where the A4V variant has been found to have a higher tendency to engage in aggregation DiDonato et al. (2003); Hough et al. (2004); Ratovitski et al. (1999).

The two other variants, G37R and H46R, have rather similar landscapes; both show a rise of the A and B basins and a widening of the energy barrier. This makes it difficult to understand what the effects of these mutations are on the stability of SOD1. On the one had, the rise of the basins would cause the transition rate to increase. However, a denser and higher energy barrier would cause a decrease in the transition rate. Taken all together, the variants may have similar transition rates to the WT. This is in agreement with what is found in the current literature, where H46R is reported to maintain about 80% of its activity as in the WT (http://www.uniprot.org/uniprot/P00441). These two variants, while reported in the Asian population, are not found to cause ALS in the US population. Further research that goes beyond the single-chain analysis here may be needed to understand the possible cause of toxicity in these two variants in the Asian population.

3.3.2. Analysis of PCA-EA on HIV-I Protease. Out of the 254 structures collected from the PDB for HIV-I Protease, 54 drawn at random are withheld from the PCA. Regions on structure that are most affected and undergo large structural displacements along each of the top two PCs are illustrated on a selected structure and shown in Fig. 6(a)-(b). The displacements along PC1 affect

largely the same regions as the displacements along PC2. 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 (data not shown). This is in agreement with other PCA-based analysis in Teodoro and Kavraki (2003) (of structures obtained via MD) of the structural flexibility of HIV-I Protease.

Fig. 7(a) shows all 254 structures collected for HIV-I Protease from the PDB projected on the top two PCs. The projections are color-coded, with orange indicating the 200 structures subjected to PCA and blue indicating the 54 structures withheld from the PCA. The distribution of structures in the reduced PC-based space in Fig. 7(a) shows no distinct organization of structural states.

The distribution of pairwise CA lRMSDs between all collected structures of HIV-I Protease in Fig. 7(b) shows a unimodal distribution with a maximum pairwise CA lRMSD of 1.4Å. Taken together, these results suggest that HIV-I Protease has a wide basin, with a range of a structures that are thermodynamically-available to the WT sequence.

PCA-EA is applied to the HIV-I Protease WT sequence, and a subset of functional conformations is selected based on an energy threshold observed as the maximum energy value of experimentally-available functional structures after being subjected to the *relax* protocol. As described above for SOD1, an energy landscape can be associated with such conformations and visualized. The landscape is shown in Fig. 8. A broad region is associated with low energies, which reflects the fact that PCA-EA has obtained a wide range of functional conformations of comparable energies. Given that HIV-I 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 indeed a wide basin. It is worth noting that these observations are only relevant

for the monomeric unit of the naturally-occurring dimer, to which PCA-EA is limited. The PDB structures for HIV-I Protease are projected over the landscape. The structures withheld from the PCA are drawn as gray triangles, whereas those used by the PCA are drawn as black circles. The locations of these structures are on the landscape, including those not used by the PCA, which suggests PCA-EA captures the functional structures not used to define its reduced search space. Many of the functional structures deposited in the PDB are on the broad basin reported by PCA-EA for HIV-I Protease, but there are structures (including those documented for the WT in the PDB) in regions associated with higher energies by PCA-EA. This suggests that either PCA-EA has not fully explored these regions or that it has indeed found lower-energy functional structures in regions of the structure space not yet probed in the wet laboratory.

3.3.3. Analysis of PCA-EA on CaM. CaM demonstrates the ability of PCA-EA to reconstruct landscapes of proteins with multiple structural states more than 13Å apart (pairwise CA lRMSD between structures with PDB ids 1CLL and 2F3Y is 13.44Å). As shown 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 for CaM. The structural displacements along PC1 and PC2 are shown in Fig. 9(a)-(b), respectively. The largest displacements along PC1 include the long α-helix connecting the two structurally-similar domains in CaM; motions along PC1 capture folding and unfolding of this helix that brings the two N- and C-termini domains close to or far away from each-other. Motions along PC2 do not include the helix and capture primarily motions of the N- and C-terminal domains. These findings are in agreement with other studies, which show that the structural variability in CaM is localized to folding and unfolding of the connecting helix that regulates concerted motions between the N- and C-termini domains Shehu et al. (2009).

Fig. 10(a) shows all collected CaM structures superimposed on the top two PCs. The projections are color-coded, with blue indicating the structures subjected to PCA and green indicating the structures withheld from the PCA. The distribution of structures in the reduced space in Fig. 10(a) shows the existence of several distinct structural states. In particular, two regions of the space seem well-probed in the wet laboratory, those that correspond to the structural states captured under PDB id 2F3Y and 1NWD.

The distribution of pairwise CA lRMSDs between all collected structures of CaM, shown in Fig. 10(b), supports the existence of multiple well-defined structural states, showing a multimodal distribution with a maximum pairwise CA lRMSD of over 20Å. This analysis suggests that multiple basins are expected to be found in the CaM energy landscape.

Fig. 11(a) summarizes the CaM WT energy landscape reconstructed by PCA-EA by projecting the subset of functional conformations obtained by PCA-EA on the top two PCs. Projections of all PDB-collected structures used by the PCA and those withheld from it are shown on the landscape. The structures used by PCA-EA are drawn as black circles, whereas those withheld are drawn as gray triangles.

Fig. 11(a) shows a complex landscape with multiple low-energy regions. In particular, two broad basins are found. One, the deepest corresponds well to the ligand-bound state of CaM (PDB id 2F3Y). The other broad, but not as deep basin corresponds to the protein-bound state (the structure reported under PDB id 1NWD, which is found bound to the a dimer of glutamate decarboxylase C-termini Yap et al. (2003)). The CaM WT 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 Fig. 11(b1)-(b2) by PDB ids 1CLL and 1CFD,

respectively. This suggests a bias in Rosetta towards compact conformations.

The CaM WT landscape obtained by PCA-EA allows drawing several more conclusions. The superimposition of the withheld structures shows that the ligand-bound structures of CaM are actually shifted in the structure space by about 7Å. Fig. 11(b1) renders functional conformations found by PCA-EA that are in this basin and superimposes them over the wet-lab structure under PDB id 2F3Y. The conformations are of the same topology as the structure under PDB id 2F3Y, which confirms that the Rosetta energy landscape retains the overall topology of the ligand-bound state, and the shift is due to structural fluctuations in loops and termini. The functional conformations obtained by PCA-EA that are in the next broad basin are also shown, superimposed on the wet-lab structure with PDB id 1NWD in Fig. 11(b2). 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, including that with PDB id 1NWD.

4. DISCUSSION

This paper has proposed a novel stochastic optimization algorithm, PCA-EA, to explore the conformation space of dynamic proteins with complex energy surfaces. The algorithm reveals stable and semi-stable structural states of a given protein sequence by reconstruction of the energy landscape. Computational cost is controlled by leveraging information contained in experimentally-available structures of WT and variant forms of a protein. In particular, dimensionality reduction is employed to extract from such structures collective variables to define a reduced search space. The algorithm contains several novel components, including a local selection operator to avoid premature convergence to any particular region in the conformation space.

The analysis of applications of PCA-EA in this paper indicates that the algorithm is able to provide a link between sequence mutations and changes in function through the energy landscape, as demonstrated by comparison of the energy landscapes it reconstructs on the WT and three disease-involved variants of SOD1. PCA-EA is also scalable and able to explain relationships between known structural states of proteins, such as CaM, where experimentally-probed functional structures can be more than 10Å away from one another.

The results presented here are promising and suggest that further algorithmic research in EA-based exploration of protein conformation spaces is well warranted. Several directions of future work can be considered. An interesting direction concerns employing PCA-EA as a roadmap on how to integrate dimensionality reduction in search but considering different techniques that do not suffer the linearity limitation of PCA. Such techniques may reveal even lower-dimensional search spaces, but they must allow directly sampling in the reduced space. The latter is a key feature for any con-

formational search algorithm. Finally, as some of the results on CaM have suggested, there may be distinct biases in specific energy functions. It is possible, for instance, that the differences between the two deepest basins revealed by PCA-EA on CaM may be less striking when another energy function is employed, or that the other known states are indeed in basins of their own. Considering different energy functions is important, but this also increases computational demands, particularly when considering that some of the most popular physics-based force fields implemented as part of MD simulation software packages are not yet easily integrateable in conformational search algorithms written by researchers. It is expected that improvements in these packages to this end will facilitate comprehensive analysis and strengthen the applicability of powerful stochastic optimization methods for protein structure modeling.

ACKNOWLEDGMENTS

Funding for this work is provided in part by the National Science Foundation (Grant No. 1421001 and CAREER Award No. 1144106) and the Thomas F. and Kate Miller Jeffress Memorial Trust Award.

AUTHORS' CONTRIBUTIONS

R.C. and A.S conceived the algorithm proposed here, the experimental design, and analysis strategy, carried out data analysis. R. C. implemented the algorithm and performed production runs. A. S. wrote the article.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

References

Adcock, S. A. and McCammon, J. A. 2006. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106(5):1589–1615.

Anderson, E., Bai, Z., Dongarra, J., Greenbaum, A., McKenney, A., Du Croz, J., Hammerling, S., Demmel, J., Bischof, C., and Sorensen, D. 1990. Lapack: A portable linear algebra library for high-performance computers. In *Proceedings of the 1990 ACM/IEEE Conference on Super-*

computing, Supercomputing '90, pages 2–11, Los Alamitos, CA, USA. IEEE Computer Society Press.

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

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

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

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

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

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

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

Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A.,

Skalicky, J. J., Kay, L. E., and Kern, D. 2005. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–121.

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

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

Hough, M., Grossmann, J., Antonyuk, S., Strange, R., Doucette, P., Rodriguez, J., Whitson, L., Hart, P., Hayward, L., Valentine, J., and Hasnain, S. (2004. 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.

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

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

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

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

Li, Y. and Yaseen, A. 2013. Pareto-based optimal sampling method and its applications in protein

structural conformation sampling. In *AAAI Workshop*, pages 32–37, Bellevue, Washington. AAAI Press.

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

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

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

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

Okazaki, K., Koga, N., Takada, S., Onuchic, J. N., and Wolynes, P. G. 2006. 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.

Olson, B. and Shehu, A. 2012a. Efficient basin hopping in the protein energy surface. In *IEEE Intl Conf on Bioinf and Biomed*, pages 119–124, Philadelphia, PA. IEEE.

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

Olson, B. and Shehu, A. 2013. 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. ACM.

Olson, B. and Shehu, A. 2014. 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. ISCA.

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

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

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

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

Soto, C. 2003. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nat Rev Neurosci*, 4(1):49–60.

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

Strange, R. W., Antonyuk, S., Hough, M. A., Doucette, P. A., Rodriguez, J. A., Hart, P., Hayward,

L. J., Valentine, J. S., and Hasnain, S. 2003. 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):877891.

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

Tousignant, A. and Pelletier, J. N. 2004. Protein motions promote catalysis. *Chem. Biol.*, 11(8):1037–1042.

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

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

Table 1: PARAMETER VALUES in our EA

System	$\lambda_{\max}$	Cell Size	C	Pop Size
SOD1	2	2×2	49	500
HIV-1 Protease	1	1×1	49	500
CaM	10	10×10	25	500

1.

DIAGRAM OF ALGORITHM

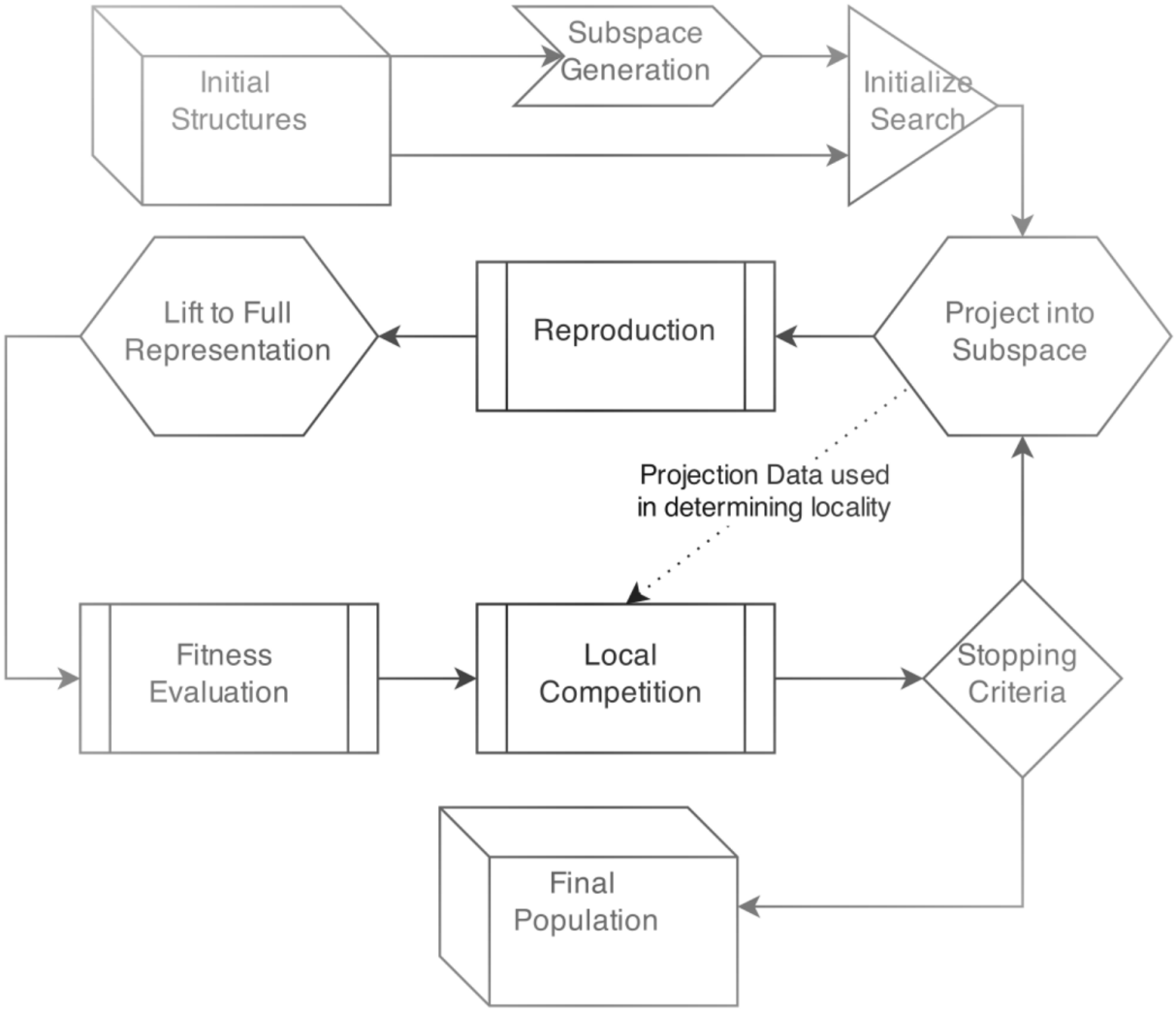


Fig. 1: Diagram shows all algorithmic components in PCA-EA.

2.

ACCUMULATION OF VARIANCE ANALYSIS

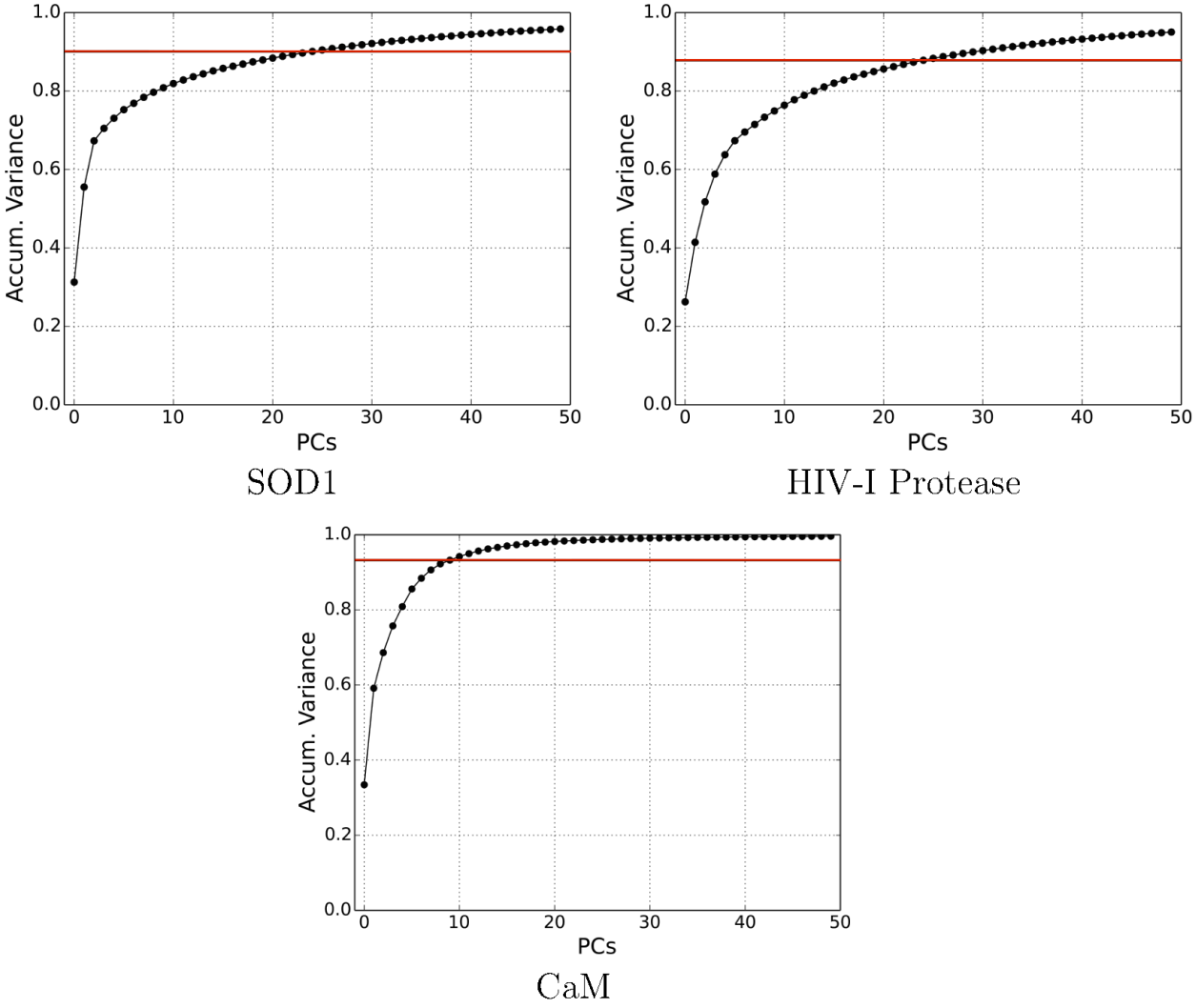


Fig. 2: The accumulation of variance is shown here. The red horizontal line shows the 90% cutoff.

3. POPULATION AND NEIGHBORHOOD SIZE ANALYSIS

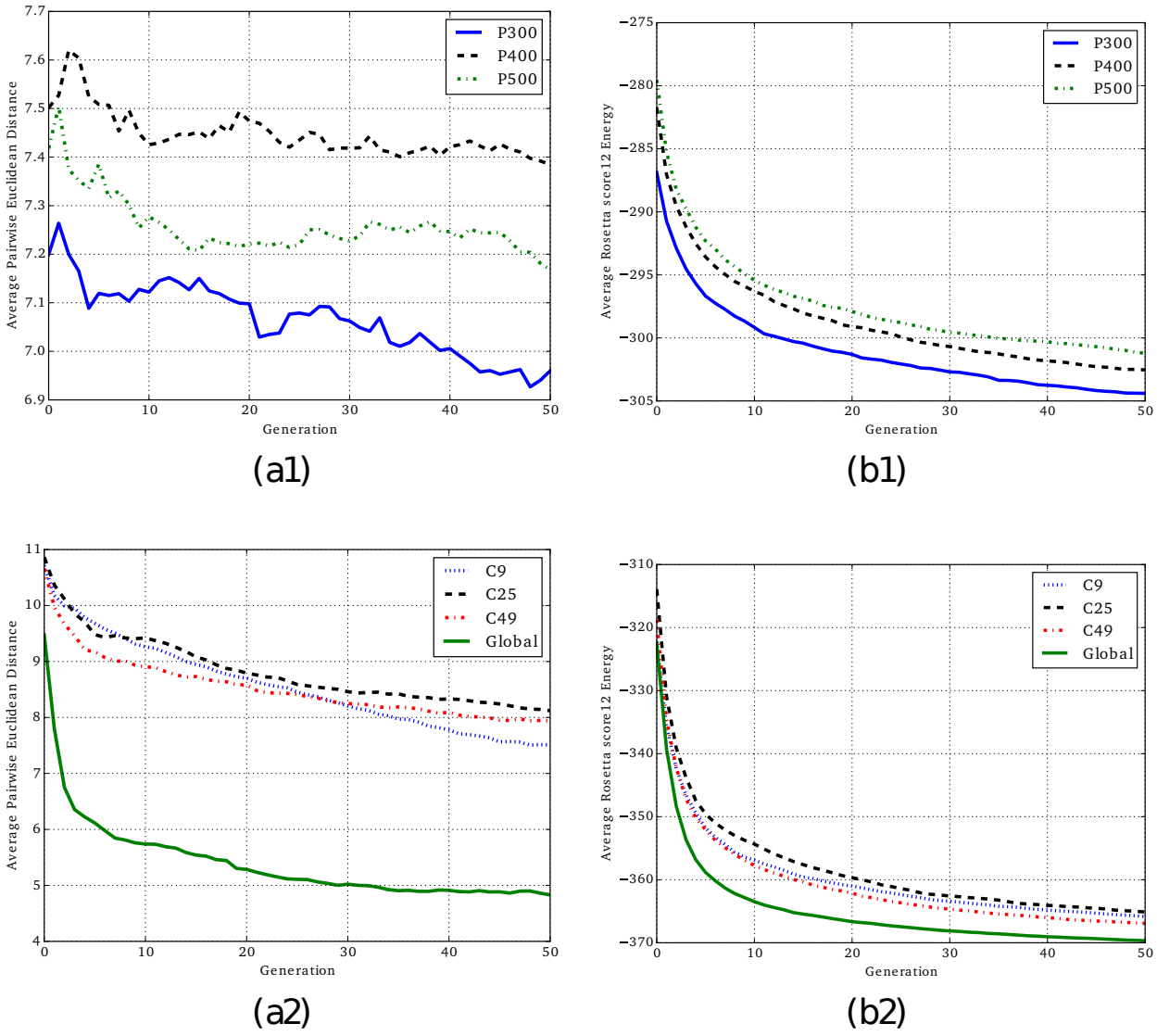
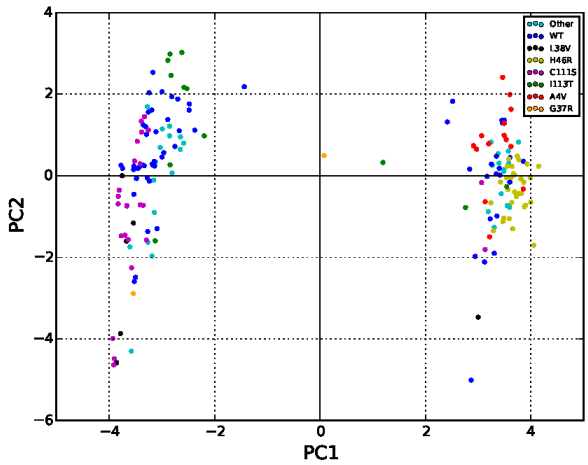


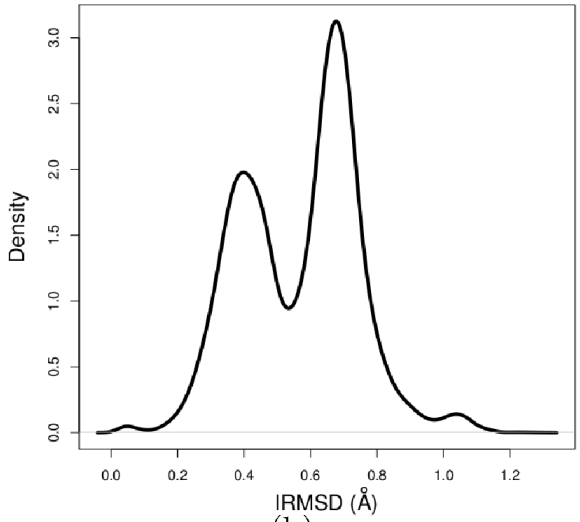
Fig. 3: Left panel: The average pairwise Euclidean distance in the m-dimensional PC space is computed over individuals in a generation and tracked across generations. Right panel: The average fitness is computed over individuals in a generation and tracked across generations. (a1)-(b1) Three settings are compared, varying the population size from 300 to 400 to 500. (a2)-(b2) Fixing population size to 500, neighborhood size is varied from C9, C25, C49, and C∞ (global).

4.

ANALYSIS OF SOD1 CRYSTAL STRUCTURES



(a)



(b)

Fig. 4: (a) A projection is shown of all structures collected from the PDB for HIV-I Protease on the top two PCs. Projections of structures available for WT and selected variants are drawn in different colors. The PC1-PC2 map of SOD1 experimentally-available structures shows two distinct clusters separated by PC1. (b) The distribution of pairwise CA lRMSDs among all structures collected from the PDB for SOD1 is bimodal. Kernel density estimation is employed to estimate the probability density function of pairwise lRMSD.

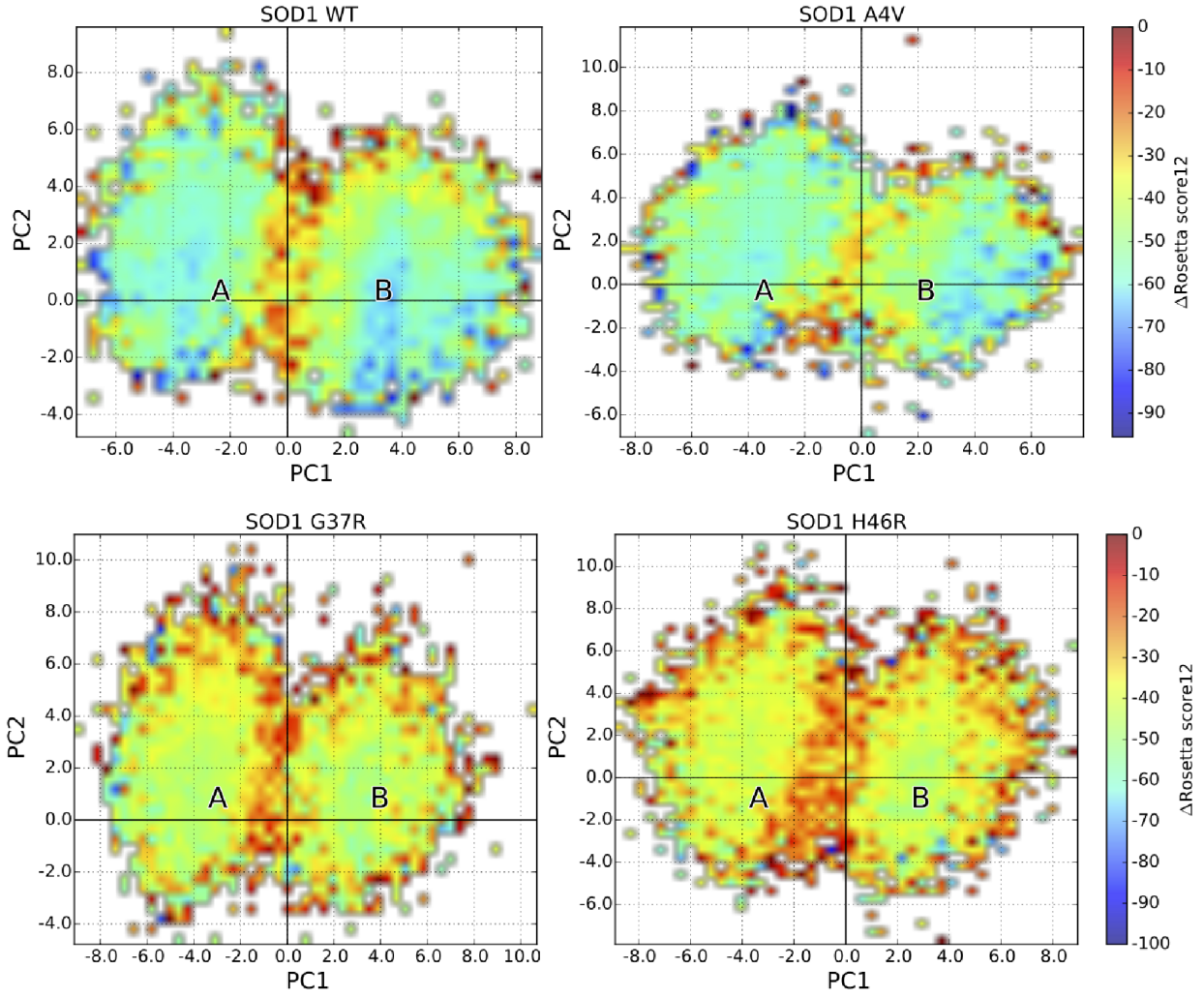


Fig. 5: Obtained energy landscapes are shown for WT, A4V, G37R, and H46R SOD1. The color bar shows not the range of absolute *score12* energy values but instead the difference from the highest-energy value.

6.

DISPLACEMENTS OF HIV-I ALONG PC1 AND PC2

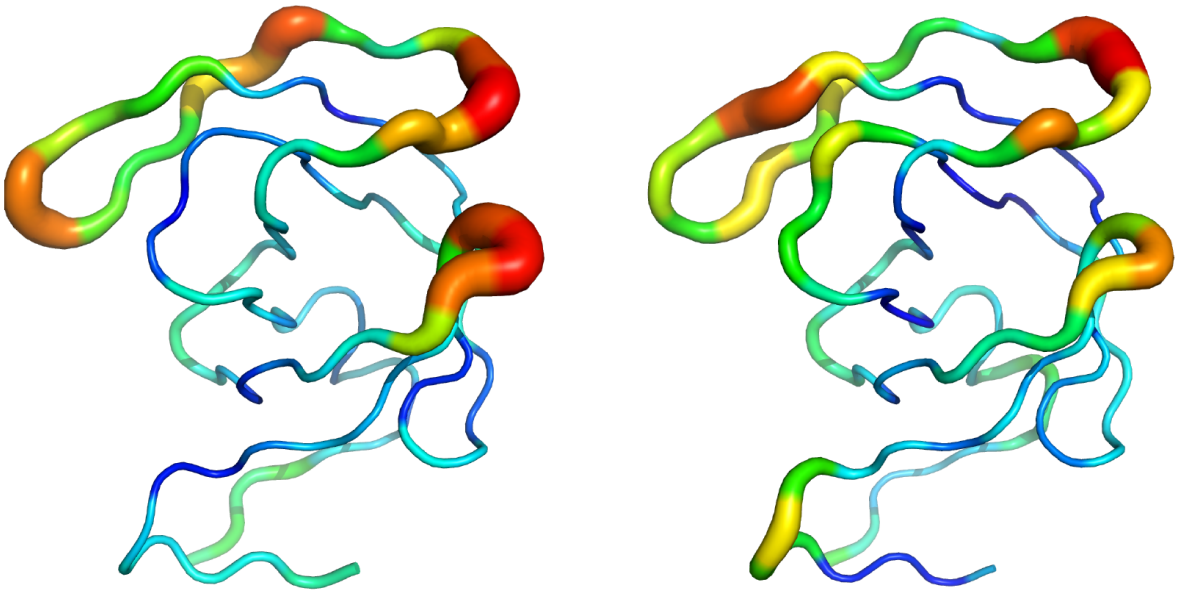
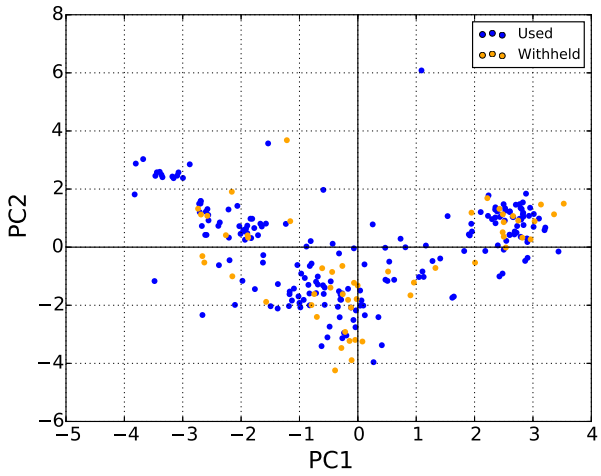


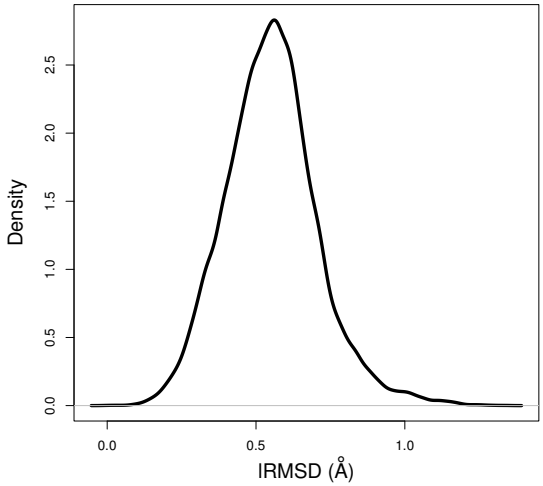
Fig. 6: (a)-(b) Regions undergoing displacements along PC1 in (a) and along PC2 in (b) are shown on a selected structure for HIV-I Protease. The coordinates in each PC are used to indicate displacements. A red-to-blue color scheme is used to indicate large-to-small displacements. Chain thickness is also used to indicate larger displacements.

7.

ANALYSIS OF HIV-I CRYSTAL STRUCTURES



(a)



(b)

Fig. 7: (a) A projection is shown of all structures collected from the PDB for HIV-I Protease on the top two PCs. Projections of structures used to obtain the PCs through PCA are drawn in blue. Projections of structures withheld from the PCA are drawn in orange. (b) The distribution of pairwise CA lRMSDs among all structures collected for HIV-I Protease from the PDB is unimodal. Kernel density estimation is employed to estimate the probability density function of pairwise lRMSD.

8.

HIV-I ENERGY LANDSCAPE

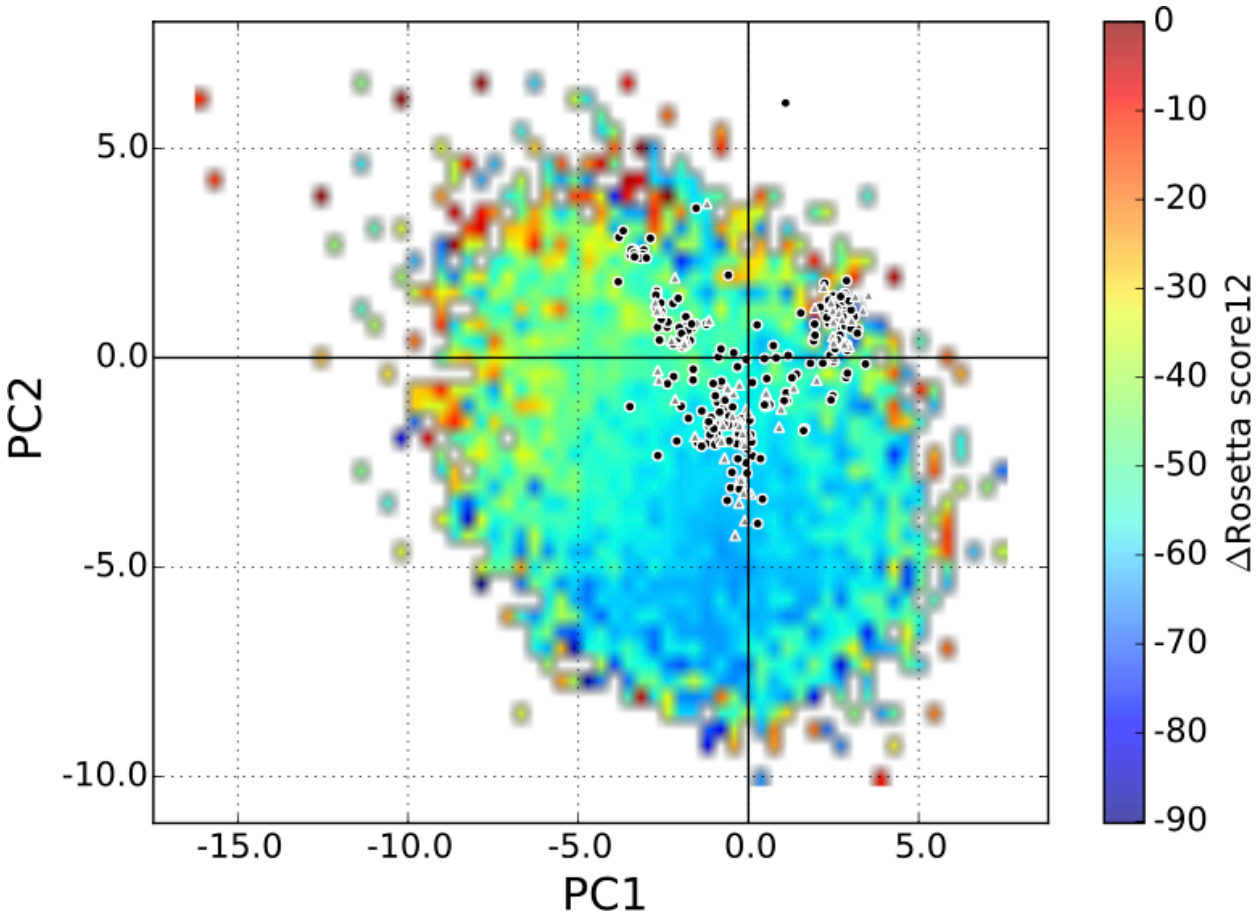
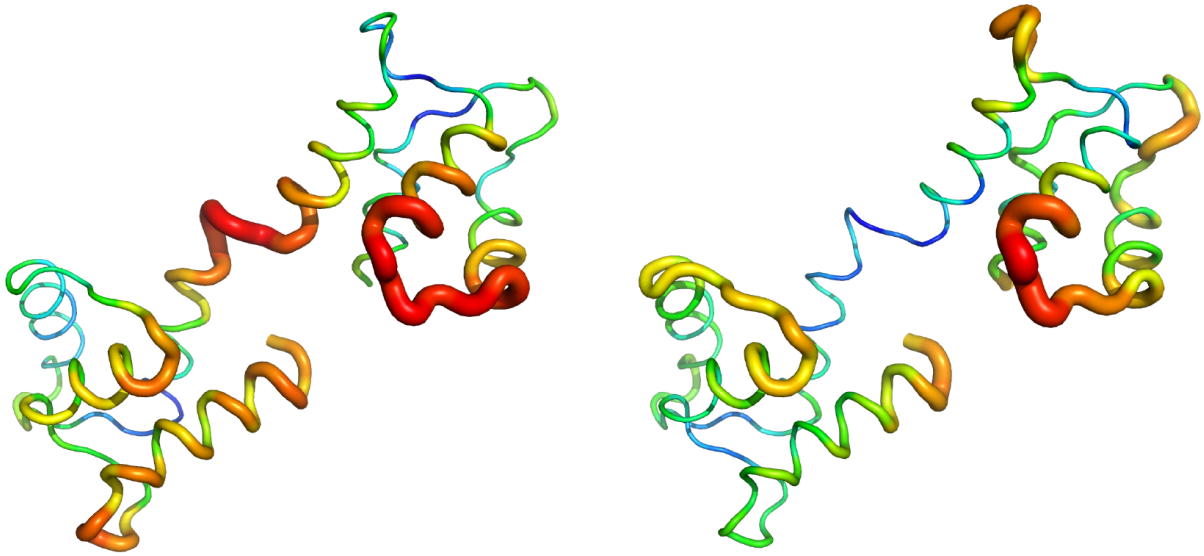


Fig. 8: Obtained energy landscape for HIV-I Protease WT. The color bar shows not the range of absolute *score12* energy values but instead the difference from the highest-energy value. The structures withheld from PCA are projected on the top two PCs and color-coded as gray triangles. The structures used by PCA are color-coded as black circles.

9.

DISPLACEMENTS OF CAM ALONG PC1 AND PC2

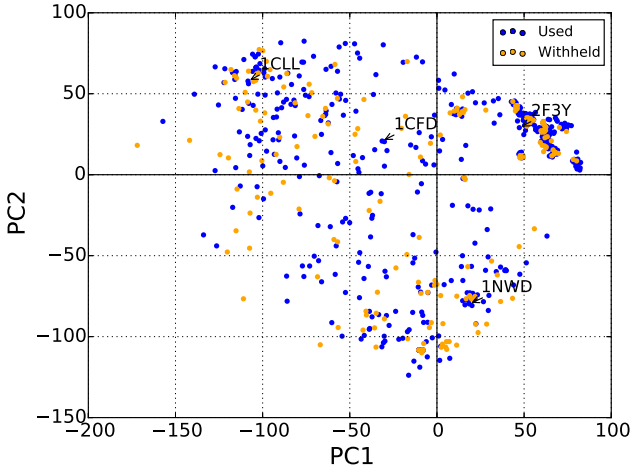


(a) (b)

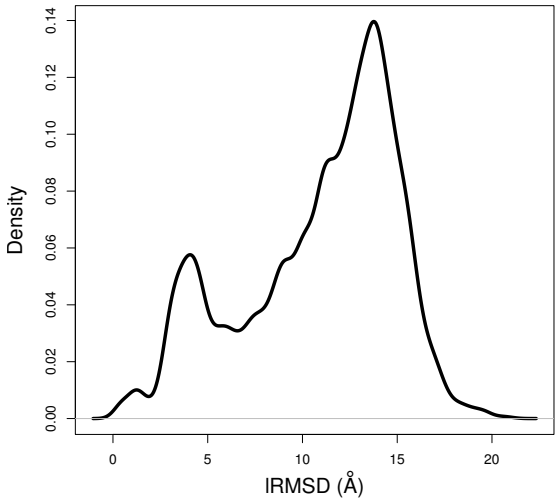
Fig. 9: (a)-(b) Regions undergoing structural displacements along PC1 in (a) and along PC2 in (b) are illustrated on a selected CaM structure (PDB id 1CFD). The coordinates in each PC are used to indicate displacements. A red-to-blue color scheme is used to indicate large-to-small displacements. Chain thickness is also used to indicate larger displacements.

10.

ANALYSIS OF CAM CRYSTAL STRUCTURES



(a)



(b)

Fig. 10: (a) A projection is shown of all structures collected from the PDB for CaM on the top two PCs. Projections of structures used to obtain the PCs through PCA are drawn in blue. Projections of structures withheld from the PCA are drawn in orange. Four well-studied functional states of CaM, represented by structures with PDB ids 1CLL, 1CFD, 2F3Y, and 1NWD are annotated on the PC1-PC2 map. (b) The distribution of pairwise CA lRMSDs between structures collected from the PDB for CaM is multimodal. Kernel density estimation is employed to estimate the probability density function of pairwise lRMSD.

11.

CAM ENERGY LANDSCAPE AND SELECTED STATES

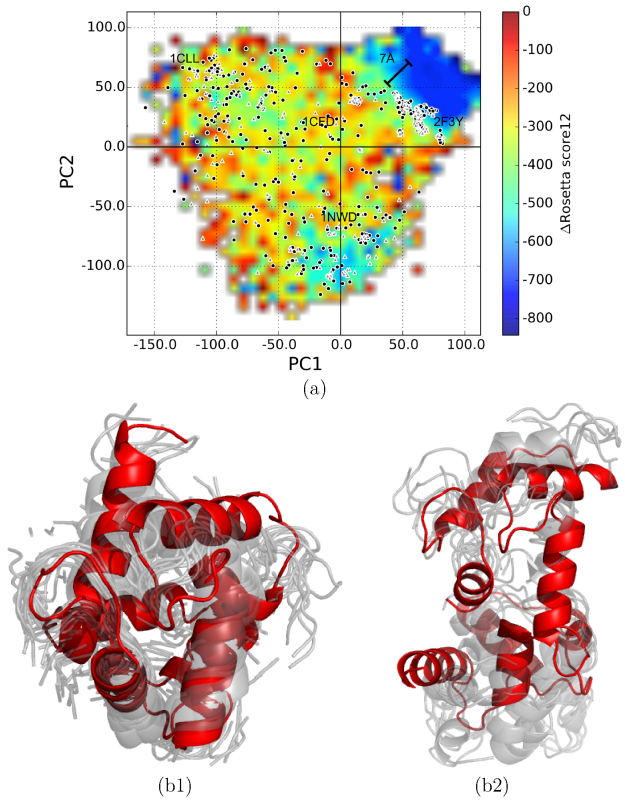


Fig. 11: (a) The population of functional conformations generated by PCA-EA for CaM WT is shown projected on the top two PCs and color-coded by Rosetta score12 energy values. The color bar shows the difference from the highest-energy value. The 197 structures withheld from PCA are projected on the top two PCs and color-coded as gray triangles. The structures used by PCA are color-coded as black circles. (b1) 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). (b2) 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).