From Optimization to Mapping: An Evolutionary Algorithm for Protein Energy Landscapes

Emmanuel Sapin, Kenneth A De Jong, and Amarda Shehu, *Members, IEEE*

Abstract—Stochastic search is often the only viable option to address complex optimization problems. Recently, evolutionary algorithms have been shown to handle challenging continuous optimization problems related to protein structure modeling. Building on recent work in our laboratories, we propose an evolutionary algorithm for efficiently mapping the multi-basin energy landscapes of dynamic proteins that switch between thermodynamically stable or semi-stable structural states to regulate their biological activity in the cell. The proposed algorithm balances computational resources between exploration and exploitation of the nonlinear, multimodal landscapes that characterize multi-state proteins via a novel combination of global and local search to generate a dynamically-updated, information-rich map of a protein's energy landscape. This new mapping-oriented EA is applied to several dynamic proteins and their disease-implicated variants to illustrate its ability to map complex energy landscapes in a computationally feasible manner. We further show that, given the availability of such maps, comparison between the maps of wildtype and variants of a protein allows for the formulation of a structural and thermodynamic basis for the impact of sequence mutations on dysfunction that may prove useful in guiding further wet-laboratory investigations of dysfunction and molecular interventions.

Index Terms—Protein structure modeling, energy landscape, basins, structural states, stochastic search, evolutionary algorithm, mapping, hall of fame, lineage- and neighborhood-aware improvement operator, reduced representation, collective variables.

1 Introduction

Stochastic search is often the only viable option to address complex optimization problems with vast variable spaces [1]. Evolutionary algorithms (EAs) have been shown capable of solving diverse optimization problems with nonlinear and multimodal fitness landscapes [2], ranging from the traveling salesman and scheduling [3], [4] to packing [5]. EAs have become popular in various sub-domains in computational structural biology, such as peptide and protein structure modeling [6], [7], [8], [9], [10], protein-ligand docking [11], protein-protein docking [12], [13], [14], and cryoEM-based registration for protein assembly structure determination [15].

Many EAs have been proposed in recent years to advance computational capabilities regarding the *de novo* protein structure prediction (PSP) problem, where the only *a priori* knowledge specific to the protein under investigation is its amino-acid sequence. PSP has a natural formulation under optimization, as spatial placements (structures) assumed by the (protein) amino-acid chain at equilibrium are minima of a potential energy function that sums up atomic interactions. While a comprehensive review of EAs for PSP is beyond the scope of this paper, the interested reader is referred to a recent review in [16].

At the core of PSP EAs is the assumption that the protein under investigation has a unique, structurally-homogeneous set or state of structures that can be discovered by treating PSP as an optimization problem of a potential energy function. However, many proteins are dynamic molecules switching between diverse structural states at equilibrium to regulate allosteric signaling, catalysis, and other processes central to the cell machinery [17]. There is now a demand for computational methods that can uncover a possibly rich set of structural states populated by a protein at equilibrium to understand the array of activities of a dynamic protein in the cell [18].

From a protein energy landscape perspective, this new task for computational methods shifts the focus from optimization to mapping: identifying the multiple local optima in the landscape, characterizing their neighborhoods (basins), thus expanding the view from single structures to thermodynamically-stable and semi-stable structural states, and visualizing the inter-basin energy barriers slowing down structure switching [19].

Given that protein structure spaces (and associated energy landscapes) are vast and high-dimensional, attempting to construct maps of protein energy landscapes in a *de novo* setting is computationally infeasible. As our recent efforts demonstrate, a structure-guided setting may prove useful. For many proteins of importance to human biology and disease, wet-laboratory studies have identified numerous structures for the wildtype (WT) and variant (mutated) sequences of a protein and documented them in publicly-accessible repositories, such as the Protein Data Bank (PDB) [20]. These structures encode, in their entirety, the function-constrained dynamics/motions of a protein's WT or variant sequences; such motions and the structures themselves can be leveraged to effectively build maps of protein energy landscapes with an EA.

In this paper we build on our previous work and propose an EA that uses a structure-guided approach to produce information-rich maps of multi-basin protein energy landscapes in a computationally-feasible manner. The EA

• *E. Sapin, A.Shehu, and K. A. De Jong are with the Department of Computer Science, George Mason University, Fairfax, VA, 22030. E-mail: amarda@gmu.edu*

makes use of a novel combination of global and local search to balance between exploration and exploitation. From now on, we refer to this mapping-oriented EA as PELMap-EA for Protein Energy Landscape Mapping EA. We illustrate its capabilities by applying it to several proteins and their disease-variants and showing that PELMap-EA produces comprehensive and detailed maps of the multi-basin energy landscapes associated with multi-state proteins. We further show that comparison between the PELMap-EA-constructed maps of WT and variants of a protein aids the formulation of a structural and thermodynamic basis for the impact of sequence mutations on dysfunction that may prove useful in guiding further wet-laboratory investigations of dysfunction and molecular interventions.

This paper is organized as follows. Section 2 provides background on protein energy landscapes and summarizes related work, focusing on evolutionary search techniques for structure spaces of dynamic proteins. PELMap-EA is described in Section 3. Analysis in Section 4 shows that PELMap-EA is effective at mapping and locating basins in complex energy landscapes. Section 5 concludes the paper with a discussion and future prospects.

2 Background and Related Work

The ability of a protein to switch between different structures at equilibrium [21] warrants characterizing the equilibrium structural dynamics of a protein as a means of exposing the range of activities of a protein in the cell [18]. At present, neither computational nor wet-laboratory methods can on their own span all the spatial and temporal scales involved in equilibrium protein dynamics [22].

Structure-switching proteins are expected to have energy landscapes rich in local optima with local neighborhoods (basins) of varying depths and breadths. As noted in Section 1, this setting may be more suitably addressed by mapping-oriented EAs leveraging novel algorithmic ingredients to enhance both exploration and exploitation. In our prior work on optimization-oriented EAs for multi-state proteins we have shown that known, wet-laboratory structures of a protein sequence under investigation and its variants can be leveraged to enhance exploration; the latter is accomplished by extracting from the structures, via statistical analysis, an effective reduced representation leading to a reduced variable space, as well as using the structures themselves to seed the initial population with possible local minima in the energy landscape.

An important realization in our laboratories has been that, even when the focus is on a particular protein sequence, structures caught in the wet laboratory for variants of this sequence (while stable for the variants) represent possible semi-stable states for the particular sequence under investigation (a semi-stable state is one in which a protein does not stay as long as in a stable one and is therefore harder to catch in the wet laboratory). This realization is based on the principle of conformational selection that now informs our understanding of the relationship between structural dynamics and function [23].

The principle of conformational selection states that perturbations, such as sequence mutations or presence of binding partners, do not change the structure space but rather the probabilities (related to energies) with which structures are assumed by a particular protein sequence or the uncomplexed versus the complexed protein [23]. This principle has been reaffirmed in studies of the structural dynamics of multi-state proteins, including the H-Ras enzyme studied in this paper. As summarized in [24], many proteins harbor an intrinsic susceptibility to sample multiple structural states, and perturbations do not directly induce structural changes but rather shift or bias pre-existing structures. Therefore, structures that are stable for variants may be higher-energy for the particular sequence under investigation, but including them in the initial population may provide an EA with new regions of the landscape that, through further exploitation, lead to previously-unknown basins.

Moreover, structures caught in wet laboratories for a specific protein sequence under investigation and other similar sequences additionally encode the essential dynamics (how groups of atoms move collectively) and can be leveraged via statistical analysis to extract modes of motion to serve as collective variables for analysis or search [25], [26], [27], [28], [29]; in fact, recent work shows that extending the analysis to the entire family to which a protein belongs provides useful information on dynamics; the particular protein sequence under investigation effectively selects modes of motion from the family dynamics [30], [31].

Based on the realization that structures deposited for sequence variants of a protein represent possible local minima in the structure space of the sequence under investigation, our laboratories have developed and analyzed in detail several optimization-oriented EAs for uncovering functionally-relevant stable and semi-stable structural states of an uncomplexed (unbound) protein. For instance, work in [32] shows that a Principal Component Analysis (PCA) of known, wet-laboratory structures of a protein sequence and its variants provides useful information about collective motions of atoms and the protein structure space probed in the wet laboratory. This information is integrated in well-known optimization-oriented evolutionary search strategies, such as CMA-ES [33], or novel generational EAs, such as PCA-EA in [34]. Significant analysis of various algorithmic components of PCA-EA in [35], [36] has lead, among other insights, to the design of an effective decentralized selection operator for delaying population take-over by the most fit individuals in an optimization setting [35].

An optimization setting that relies on analysis of all structures ever generated by an EA to uncover basins in the landscape may not be best suited for dynamic proteins. For this reason, in recent work [37] we investigate switching from an optimization setting to that of mapping a (multi-state) protein's multi-basin energy landscape. The goal is to obtain a *sample-based representation* of the energy landscape, which we refer to as a map. A useful map for our goal of elucidating stable and semi-stable structural states of a protein relevant for function modulation is one that uncovers as many broad and deep basins as can be afforded by a reasonable computational budget and an intelligent apportioning of this budget. In [37] we show that the evolutionary computation technique of a hall of fame is well-suited to serve as a sparse, resource-efficient, sample-based representation (map) of a protein's energy landscape.

The result of this series of work in our laboratories is

the firm belief that effective and efficient characterization of multi-basin protein energy landscapes requires a fundamental shift from optimization to mapping in a manner that effectively uses existing *a priori* structural information and interleaves global and local search. The result is PELMap-EA, which we now describe in detail.

3 METHODS

PELMap-EA is provided in pseudocode in Algorithm 1. The first shaded box in Algorithm 1 draws attention to lines 1-4, where wet-laboratory structures are subjected to PCA to extract the variable space (details are provided in Section 3.1). The next shaded box draws attention to lines 5-6, where the initial population is defined from the wet-laboratory structures and the defined variable space (details are provided in Section 3.2). The next three lines prepare a two-dimensional (2d) grid for the local selection operator, initialize the hall of fame, and declare a variable to track the status of an improved offspring.

PELMap-EA evolves a population of individuals in the variable space until a budget FMAX of fitness evaluations is exhausted, with $fcounter$ keeping count (lines 10-11). Inside the main loop, each parent in the population is selected to be subjected to the variation operator. The fitness of the parent is available, as well as two values specific to the novel improvement operator we propose and analyze here, $niters$, and fea. The first keeps track of the number of iterations spent improving individuals of a lineage, and the second records the fitness of the earliest ancestor. These two values, properly set in the initial population, are updated in lines 15-16 for the offspring yielded by the variation operator (line 14); the operator is detailed in Section 3.3.

The offspring is then subjected to an improvement operator, and the flag passed as input to PELMap-EA determines which operator is to be used. The local improvement operator, detailed in Section 3.4, returns the improved offspring, its fitness, and adds NrImprovItersMax to the counter of fitness evaluations (lines 17-20). Every improved offspring survives. This is not the case with the novel, lineage- and neighborhood-aware improvement operator, carried out in lines 21-22, and detailed in Section 3.5. The improvement may fail if the budget of iterations per lineage has been reached already in c, lines 23-24. Only one iteration is spent on an offspring, as well (line 25). The next shaded box takes effect only when the novel improvement operator is invoked. If the improved offspring survives, it is considered for inclusion in the hall of fame (lines 30-31), as detailed in Section 3.6. If not, then the lineage is terminated, and a new one (lines 28-29) is started with the parent replaced by an individual drawn at random (line 27).

A surviving offspring is added to the offspring set (line 32). The local selection operator compares offspring to parents to determine the subset of N individuals that survive in the next population (lines 33-34), as detailed in Section 3.7. PELMap-EA outputs the hall of fame, which is analyzed to compare the impact of the two improvement operators as well as to compare energy landscapes of WT and variants of a protein. We now proceed to relate details.

Algorithm 1 PELMAP-EA

Input: FMAX //budget of fitness/energy evaluations
N //Population Size
ProteinSequence //WT or variant
$\{T_1, \ldots, T_n\}$ //wet-lab CA traces for PCA
TargVar //for PCA
$\text{step}_{\max}$ //variation operator
NrImprovItersMax //for improvement
UseLocImprovOper //which operator
fitThreshold //for hall of fame update
distThreshold //for hall of fame update
$W_{\mathcal{G}}$ //grid cell width for selection operator
$Cx_{\mathcal{G}}$ //grid neighborhood for same operator

//extract collective vars from wet-lab CA traces
1: $\langle \text{trace} \rangle \leftarrow$ average trace over $\{T_1, \ldots, T_n\}$
2: **for** $1 \leq k \leq n$ **do** //convert to displacements
3: $T_k \leftarrow T_k - \langle trace \rangle$ //for centered covariance
4: $U, \Sigma, V^T, m \leftarrow$ PCAAndDim($\{T_1, \ldots, T_n\}$), TargVar

5: $i \leftarrow 0$ //population iterator
6: $\mathcal{P}_i \leftarrow$ InitOper(ProteinSequence, N, $\{T_1, \ldots, T_n\}$, $\langle trace \rangle$, $U_{[,1:m]}$)

7: $\mathcal{G} \leftarrow$ 2dGrid($U_{[,1]}$, $U_{[,2]}$, $W_{\mathcal{G}}$) //for selection operator
8: $\mathcal{H} \leftarrow \emptyset$ //initialize hall of fame
9: survives $\leftarrow 1$ //for offspring
10: $fcounter \leftarrow 0$ //counter of energy evaluations
11: **while** $fcounter <$ FMAX **do**
12: $\mathcal{C} \leftarrow \emptyset$ //set of offspring
13: **for** $\langle p, f(p), niters(p), fea(p) \rangle \in \mathcal{P}_i$ **do**
14: $c \leftarrow$ VarOper(p, $\Sigma_{1,\ldots,m}$, $\text{step}_{\max}$)
15: $niters(c) \leftarrow niters(p)$
16: $fea(c) \leftarrow fea(p)$
17: **if** UseLocImprovOper ≥ 1 **then**
18: $\langle c', f \rangle \leftarrow$ LocImprovOper(ProteinSequence, c, $\langle \text{trace}, \rangle$, $U_{[,1:m]}$, NrImprovItersMax)
19: survives $\leftarrow 1$
20: $fcounter \leftarrow fcounter +$ NrImprovItersMax
21: **else**
22: $\langle c', f, \text{survives} \rangle \leftarrow$ ImprovOper(ProteinSequence, c, $\langle \text{trace}, \rangle$, $U_{[,1:m]}$, NrImprovItersMax, $niters(c)$, $fea(c)$, $\mathcal{H}$)
23: **if** $\langle c', f \rangle$ == NIL **then**
24: CONTINUE //improvement may fail
25: $fcounter \leftarrow fcounter + 1$
26: **if** survives < 1 **then**
27: $p \leftarrow$ atRandom($U_{[,1:m]}$) //replace parent
28: $niters(p) \leftarrow 0$ //new lineage begins
29: $fea(p) \leftarrow f$ //from failed offspring
30: **else** //consider offspring for hall of fame
31: $\mathcal{H} \leftarrow$ UpdateHallOfFame($\mathcal{H}$, c', f, fitThreshold, distThreshold)
32: $\mathcal{C} \leftarrow \mathcal{C} \cup \{\langle c', f \rangle\}$ //add to offspring set
33: $i \leftarrow i + 1$
34: $\mathcal{P}_i \leftarrow$ LocSelectOper(P_{i-1}, $\mathcal{C}$, $\mathcal{G}$, $Cx_{\mathcal{G}}$)

Output: $\mathcal{H}$

3.1 From Expert Examples to Collective Variables

Details can be found in recent work [35], but we summarize the process in the interest of clarity. Briefly, the CA atoms (the CA atom is the main atom in each of the building blocks, the amino acids, that constitute a protein chain) are first extracted from each wet-laboratory structure, obtaining what are referred to as CA traces. The traces are aligned to the first trace, arbitrarily selected to be a reference, so atomic displacements can be obtained rather than differences due to rigid-body motions (in an abuse of terminology, in pseudocode in this paper we refer to the atomic displacements as traces). Differences of each displacement vector corresponding to a structure from an average displacement vector are then computed in preparation for a centered covariance matrix. The latter is subjected to the `dgesvd` routine in Lapack [38], obtaining the U matrix of PCs and the Σ matrix containing in its main diagonal the corresponding singular values (line 4 in Algorithm 1).

As described in [35], a target cumulative variance `TargetVar` is specified to determine the number m of principal components (PCs) needed to capture the specified cumulative variance among the atomic displacements. The m top, eigenvalue-sorted PCs then serve as collective variables and define the variable space (line 4 in Algorithm 1). A cumulative variance `TargetVar` of 90% results in low enough values of m that allow reconstructing physically-realistic protein traces (and recovering low-energy structures from them via energy minimizations). For each of the proteins studied here, $m < 25$, which is more than two orders of magnitude reduction from the number of Cartesian coordinates and more than an order of magnitude reduction from the number of dihedral angles. Once the m PCs are extracted, an individual in `PELMap-EA` is then a vector of coordinates in the m-dimensional space of PCs.

3.2 Initialization Operator

The initialization operator, whose pseudocode is shown in Algorithm 2, populates the initial population with N individuals, $n < N$ of which are obtained from the wet-laboratory structures (lines 2-8). It is important to note that the n wet-laboratory structures are threaded onto the sequence of interest (`ProteinSequence`); this is accomplished by first extracting the CA traces from the collected structures, and then by replacing the amino-acid sequence of an extracted CA trace with `ProteinSequence` (line 4). The resulting CA traces, now threaded onto the sequence of interest, are then subjected to the local improvement operator (line 5) in order to obtain all-atom structures that are low-energy in the Rosetta all-atom score12 energy surface [39] of the specific input `ProteinSequence` under investigation.

The rest of the $N-n$ individuals in the initial population are drawn at random, taking into account the boundaries of the m-dimensional embedding of the wet-laboratory traces (the boundaries are not shown as input in Algorithm 2 in the interest of a clear presentation of the pseudocode) and then improved (lines 9-14). The inclusion of individuals generated at random is justified in Section A.1 in Appendix A. An individual in the initial population begins a lineage, so fea is set to its fitness (lines 6 and 12) but the improvement (so as to obtain a low-energy structure for the sequence of interest) is not added to the budget, as the initialization operator is essentially a pre-processing step.

Algorithm 2 Initialization Operator

Input: `ProteinSequence`
N
$\{T_1, \ldots, T_n\}$
$\langle trace \rangle$
$U_{[,1:m]}$

1: $\mathcal{P} \leftarrow \emptyset$
2: $i \leftarrow 0$
3: **for** $i \leq n$ **and** $i \leq N$ **do**
4: $c_i \leftarrow T_i \cdot U_{[,1:m]}$ //project trace
5: $c_i^{'}, f_i \leftarrow$ LocImprovOper(`ProteinSequence`, c_i, $\langle$trace$\rangle$, $U_{[,1:m]}$, `NrImprovItersMax`)
6: $fea \leftarrow f_i$ //individual starts new lineage
7: $niters \leftarrow 0$ //no counting of its improvement
8: $\mathcal{P} \leftarrow \mathcal{P} \cup \{\langle c_i^{'}, f_i, niters, fs \rangle\}$
9: **for** $i < N$ **do**
10: $c_i \leftarrow$ atRandomIn($U_{[,1:m]}$)
11: $c_i^{'}, f_i \leftarrow$ LocImprovOper(c_i, $\langle$trace$\rangle$, $U_{[1,1:m]}$, `NrImprovItersMax`)
12: $fea \leftarrow f_i$ //individual starts new lineage
13: $niters \leftarrow 0$ //no counting of its improvement
14: $\mathcal{P} \leftarrow \mathcal{P} \cup \{\langle c_i^{'}, f_i, niters, fea \rangle\}$

Output: $\mathcal{P}$

3.3 Variation Operator

The variation operator is described in [34], [35]. In summary, it modifies a parent by a vector drawn in the variable space. The boundaries of the m-dimensional embedding of the wet-laboratory traces are not observed, as the ultimate goal is to generate new structures. The pseudocode and more details are provided in Section A.2 in Appendix A.

3.4 Local Improvement Operator

The local improvement operator, described in our recent work [34], [35] is summarized in pseudocode in Algorithm 3. A CA trace is recovered from an individual by adding its point-based representation in the m-dimensional PC space to the average trace/displacement vector (line 1); more details can be found in [35]. A top backbone reconstruction protocol, BBQ [40], is used to compute coordinates of missing backbone atoms from the CA atoms (line 2). Coordinates of missing side-chain atoms are then computed via the Rosetta *relax* protocol [39], which makes use of the provided protein sequence (line 3). The Rosetta package is open-source and written in C/C++, which allows easily interfacing with its *relax* protocol. The open source is a strong reason for choosing the Rosetta energy function and its *relax* protocol.

The CA trace of the all-atom structure resulting from the *relax* protocol is projected back onto the m PCs (lines 4-5) to extract from it the corresponding individual. The improved individual and its corresponding fitness, the all-atom Rosetta *score12* energy are returned to `PELMap-EA`.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2628745, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Algorithm 3 Local Improvement Operator.

Input: `ProteinSequence`
c
$\langle \text{trace} \rangle$
$U_{[,1:m]}$
`NrImprovItersMax`

1: $t \leftarrow c \cdot U^T_{[,1:m]} + \langle \text{trace} \rangle$ //recover CA trace
2: backboneStructure $\leftarrow$ BBQ(t) //recover backbone
3: $\langle s, f \rangle \leftarrow$ RosettaRelax(`ProteinSequence`, backboneStructure, `NrImprovItersMax`)
4: $t^{'} \leftarrow$ Trace(s) //corresponding trace
5: $c^{'} \leftarrow (t^{'} - \langle \text{trace} \rangle) \cdot \text{U}_{[,1:\text{m}]}$ //improved offspring

Output: $\langle c^{'}, f \rangle$

Here a lower-energy individual is considered more fit than a higher-energy, less fit individual. The replacement of the offspring with the result of the local improvement operator makes `PELMap-EA` a Lamarckian EA.

Since the Rosetta *relax* protocol is computing-intensive, the improvement of offspring is distributed on a multi-core architecture. A novel improvement operator is also introduced that determines whether an offspring is worth additional CPU cycles for further improvement. The operator makes use of a key property of minimization protocols for protein and peptide structures. Because protein energy functions are nonlinear and multimodal, they are typically locally optimized via MC-based techniques. Since these are not guaranteed to converge, they proceed in iterations, at each iteration measuring energetic improvement over the previous iteration, until the improvement fails to meet a minimum predefined value;the minimization is said to have converged. We exploit the iterative characteristic here to propose an effective and efficient improvement operator.

3.5 Lineage- and Neighborhood-Aware Improvement Operator

The new improvement operator is shown in pseudocode in Algorithm 4. Unlike the local improvement operator, which carries out `NrImprovItersMax` iterations of the MC-based minimization in the Rosetta *relax* protocol to improve an offspring, the lineage- and neighborhood-aware improvement operator spends only one iteration at a time on improving an offspring c (line 3 in Algorithm 4) until the maximum `NrImprovItersMax` has been reached on the lineage to which c belongs. If such a maximum has been reached, the operator returns NIL (lines 1-2).

The algorithm is provided with information on the lineage, $niters(c)$ (the number of improvement iterations spent on a lineage) and $fea(c)$ (the fitness of the earliest ancestor of c), as well as the hall of fame $\mathcal{H}$. Once an improved offspring $c^{'}$ is obtained, the number of iterations of the lineage is updated (line 4), and $c^{'}$ is made aware of the fitness of its earliest ancestor (line 5). These two pieces of information on the lineage of an improved offspring $c^{'}$, together with μ_f, the average fitness of neighbors of $c^{'}$ in the hall of fame (computed in line 10), are employed in line 11 to determine if $c^{'}$ survives (line 12). The specific formula employed combines the fitness of the earliest ancestor, as well, and the iterations budget. An upper bound of 1 on

Algorithm 4 Lineage- and Neighborhood-aware Improvement Operator

Input: `ProteinSequence`
c
$\langle \text{trace} \rangle$
$U_{[,1:m]}$
`NrImprovItersMax` //maximum nr. iterations
$niters(c)$ //nr. iterations over lineage
$fea(c)$ //fitness of earliest ancestor in //lineage
$\mathcal{H}$ //Hall of fame

1: **if** $niters(c) ==$ `NrImprovItersMax` **then**
2: RETURN NIL //lineage has spent budget
3: $c^{'}, f \leftarrow$ LocImprovIter(`ProteinSequence`, c, $\langle \text{trace} \rangle$, $U_{[,1:m]}$, 1)

//update lineage and make improved offspring aware
4: $niters(c^{'}) \leftarrow niters(c) + 1$
5: $fea(c^{'}) \leftarrow fea(c)$
6: `survives` $\leftarrow 0$
7: **if** $niters(c^{'}) == 1$ **then** //one iteration spent so far
8: `survives` $\leftarrow 1$
9: **else** //compare to neighbors
10: $\mu_f \leftarrow$ average fitness over neighbors in hall of fame
11: **if** $f < \frac{((5-niters(c^{'}))\times fea(c^{'})+\mu_f)}{(6-niters(c^{'}))}$ **then**
12: `survives` $\leftarrow 1$
13: **if** $niters(c^{'}) ==$ `NrImprovItersMax` **then** //maximum now reached
14: **if** $f < (0.9 \times \mu_f)$ **then** //but sufficiently fit
15: `survives`$\leftarrow 1$

Output: $\langle c^{'}, f,$ `survives` $\rangle$

the L1-norm between $c^{'}$ and neighboring individuals in $\mathcal{H}$ is used to obtain μ_f. If $c^{'}$ fails this strict test and the budget of iterations has been reached on it, it is compared only to the hall of fame. If none of these tests pass, the improved offspring does not survive, and lines 26-29 in Algorithm 1 take effect, where the lineage of this offspring is terminated, and a new lineage is initiated.

It is worth noting that the relationships in lines 11 and 13 have been devised and tested experimentally. Specifically, we choose not to compare an offspring to its direct parent but rather to its earliest ancestor in order to make the exploitation less greedy and give an offspring a chance to survive. The coupling of the improvement operator and the survival mechanism in an EA is novel and has applicability beyond the application domain on which `PELMap-EA` is tested in this paper.

3.6 Efficient Update of the Hall of Fame

The hall of fame is an effective evolutionary search technique to equip an EA with memory. We utilize the hall of fame to serve as a discrete, memory-efficient representation of an energy surface. In recent work, we have maintained all individuals ever generated by an EA [35] and then visualized the individuals via color-coded projections on the top two PCs, with colors indicating Rosetta *score12* values. This has allowed us to visualize energy basins [36], but it puts significant demands on memory and post-processing.

Due to the ruggedness of the protein energy surface and continuity of the protein structure space, hundreds of thousands of structures may need to be generated to capture a possibly large, diverse set of local energy minima.

Maintaining all individuals ever generated in memory is not practical. Instead, what is needed is a map of the energy surface whose resolution is tunable. We employ the hall of fame as a dynamically-updated, resolution-tunable map of the energy surface. As line 31 in Algorithm 1 shows, every surviving improved offspring is considered for addition to the hall of fame. The algorithm that updates the hall of fame is shown in pseudocode in Algorithm 5.

Algorithm 5 Update of Hall of Fame

Input: $\mathcal{H}$ //hall of fame
c //individual considered for inclusion
$f(c)$ //Rosetta *score12* fitness of individual
`fitThreshold`
`distThreshold`

1: **if** $f(c) \geq$ `fitThreshold` **then** //fails fitness test
2: RETURN
3: $c\_candidate\_for\_hof \leftarrow 1$ //flag
4: **for all** $\langle C, f(C)\rangle$ in the hall of fame **do**
5: **if** L1-norm(c, C) < `distThreshold` **then**
6: **if** $f(c) < f(C)$ **then** //c similar to C but fitter
7: $\mathcal{H} \leftarrow \mathcal{H} \setminus \{C\}$ //C removed from $\mathcal{H}$
8: **else**
9: $c\_candidate\_for\_hof \leftarrow 0$
10: **if** $c\_candidate\_for\_hof$ == 1 **then**
11: $\mathcal{H} \leftarrow \mathcal{H} \cup \{\langle c, f(c)\rangle\}$; //$c$ included $\mathcal{H}$

The decision to include an individual c in the hall of fame is made in two stages. In the first stage, the fitness $f(c)$ of the individual is inspected. If the fitness is not below a specified threshold (line 1), the individual is not considered for addition, as the focus is on updating the hall of fame with fit individuals.

If c passes the fitness test, c is flagged for possible inclusion (line 3) and is compared to neighboring individuals C already in the hall of fame (line 4). An individual C in the hall of fame is considered similar to c if their dissimilarity, measured via the fast L1-norm distance, (line 5), is below a specified threshold. If an individual C similar to c exists in the hall of fame, then their fitnesses are compared (line 6), and if the newly-considered individual c has better fitness, then the lesser-fit replica C is discarded from the hall of fame (line 7). The idea behind this decision is to update the hall of fame with individuals that may represent the same region in the variable space but allow further exploitation of a local minimum, providing thus an opportunity to update the map with deeper minima. If the individual c is similar to some individual C in the hall of fame but does not reside deeper in the local minimum containing C, there is no reason to update the hall of fame, and c is flagged as not a candidate for inclusion (line 9).

If the individual c is not similar to anyone in the hall of fame (and has already passed the fitness test), then it is flagged for inclusion (line 10), as it resides in a new region in the variable space not already represented in the hall of fame. The inclusion of an individual c in the hall of fame after all the tests pass is carried out in line 11. The result of the hall of fame update algorithm is that the hall of fame is a set of distinct local minima, separated by at least the defined distance threshold in the variable space.

3.7 Local Selection Operator

The selection operator selects N individuals off the combined pool of N parents and N offspring (line 22 in Algorithm 1). Details and pseudocode are provided in Section A.3 in Appendix A.

3.8 Parameter Values and Implementation Details

`PELMap-EA` is implemented in C/C++ and run until a total budget of `FMAX` $= 1,000,000$ Rosetta *score12* evaluations are exhausted. This effectively is the total number of iterations in the Rosetta *relax* protocol over all individuals. Many of the parameter values employed here are as in [37] (and are listed and described in Section A.5 in Appendix A), with the exception of the budget of evaluations allocated to the new improvement operator. As described above, `NrImprovItersMax` in the improvement operators is set to 5, but the new improvement operator exhausts this budget one iteration at a time. Details on parameter values employed for the Rosetta *relax* protocol are related in Section A.4 in Appendix A. For clarity, the presentation of `PELMap-EA` in Algorithm 1 is serial, but the local improvement of the offspring is distributed on a multi-core platform of 3.2GhZ HT Xeon CPUs with 9GB RAM. The experiments reported here are carried out on a 16-core platform, but, since the distribution is embarrassingly parallel, significant time savings can be obtained with more cores.

4 Results

The analysis focuses on evaluating the performance of the lineage- and neighborhood-aware improvement operator over the local improvement operator in `PELMap-EA`. We do so on three proteins of importance to human biology that have complex dynamics. On each test case, two sets of runs are performed, 5 runs of `PELMap-EA` with the local improvement operator, and 5 runs of `PELMap-EA` with the new, lineage- and neighborhood-aware improvement operator. The halls of fame obtained from each of the 5 runs of the algorithm without or with the new improvement operator are analyzed, and one is selected as representative to use for visualization. A hall of fame is visualized over PC1 and PC2, effectively showing only the first two coordinates of each individual and color-coding the PC1-PC2 coordinates based on the Rosetta *score12* energies of the corresponding all-atom structures. `PELMap-EA` with the new improvement operator is also applied to specific disease-implicated variants of H-Ras. Halls of fame of selected variants are compared to the hall of fame of the WT to understand the impact of sequence mutations on the H-Ras energy landscape. Important observations are drawn regarding the role of structure and energetics in H-Ras dysfunction.

This section is organized as follows. We first provide information on testing data sets. We then evaluate the novel improvement operator. Lastly, we draw differences between the H-Ras WT and selected variants.

4.1 Test Cases and Data Preparation

Performance is evaluated on the catalytic domain of H-Ras, to which we refer as H-Ras from now on, HIV-I Protease, and Calmodulin (CaM). H-Ras is a 166 amino-acid long enzyme that mediates signaling pathways central to cell proliferation, growth, and development. HIV-I Protease is an enzyme that assists the replication process of the HIV-I virus [41]. In its native form, the enzyme is a dimer, containing two identical chains, each 99 amino-acids long. Here we explore the structure space accessed by one, uncomplexed chain; per the conformation selection principle, the structure space of the uncomplexed chain is also populated in the complexed, dimeric form, though with possibly different population probabilities [42]. CaM, is a 148 amino-acid long enzyme that binds calcium and regulates over 100 target proteins, such as kinases, phosphodiesterases, calcium pumps, and motility proteins [43], [44], [45].

These proteins are well-studied in wet laboratories due to their central role in human biology and disease. Many wet-laboratory studies have caught these proteins assuming different structures depending on their molecular partner. For instance, it is known that H-Ras switches between two distinct structural states, deemed "on" (active) and "off" (inactive) to regulate its biological activity between GTP- and GDP-binding [27], [46]. These two states can be found in the PDB under entries 1qra [47] and 4q21 [46], respectively. The structures in these two entries are around 1.5Å away when comparing their CA atoms.

The ability of HIV-I Protease to undergo large-scale, concerted structural fluctuations has been observed in dry and wet laboratories [48]. Similarly, CaM, has been found to tune its biological function and recognize a diverse set of molecular partners due to its ability to assume structures more than 10Å away from one another in structure space when comparing the placements of CA atoms. These three proteins are classic examples of multi-state or multi-basin proteins, and the diversity of structures deposited for them in the PDB makes them ideal test cases for `PELMap-EA`.

Many multi-state proteins undergo sequence mutations that impact their ability to tune their biological function. H-Ras is one such protein, with sequence mutations that are implicated in a variety of human cancers [49], [50]. In addition to uncovering the thermodynamically stable and semi-stable structural states of multi-basin proteins, such as H-Ras, we also compare here the energy landscapes of WT and selected cancer-implicated variants to discover the role of structure and energetics in the relationship between protein sequence and function.

As described in section 3, PCA is applied to datasets collected for the three proteins, 43 structures for H-Ras, 254 structures for HIV-I Protease, and 697 structures for CaM. A cumulative variance of 90% is reached at $m = 10$, $m = 25$, and $m = 10$ PCs for H-Ras, HIV-I Protease, and CaM, respectively. The cumulative variance profiles are not shown here but have been presented in our recent work on analysis of the effectiveness of PCA for capturing functional motions of multi-basin proteins [35], [36]. Further details regarding the data collection protocol and the PDB identifiers of all collected structures for each protein studied here can be found in Sections B.1-3 in Appendix B.

4.2 Visual Rendering of the Hall of Fame

We show a hall of fame as a (gnuplot) point cloud. Each point corresponds to an individual in the hall of fame, positioned according to the individual's first two coordinates. The points are color-coded based on the Rosetta *score12* values of the individuals corresponding to them. This way of rendering the hall of fame effectively provides a 2d map of the *score12* all-atom energy landscape. For each of these proteins, the 2d map captures about 50% of the essential dynamics; that is, the cumulative variance of the top two PCs is about 50%. While gnuplot does provide interpolation options, no interpolation is carried out in the interest of providing an unbiased 2d map of the Rosetta *score12* all-atom energy landscape.

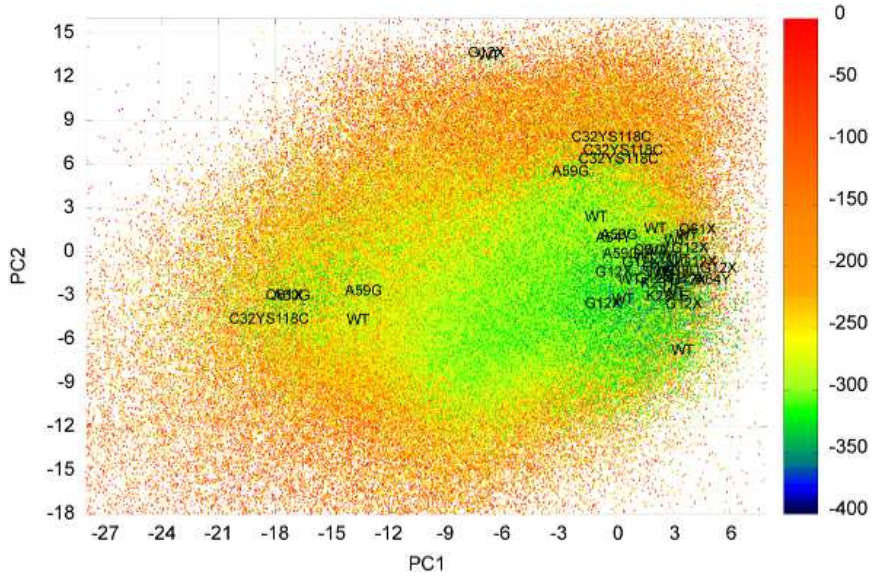
In addition, the color-coding scheme for the drawn maps is chosen so as to concentrate on low energies (specifically, red-to-yellow covers Rosetta *score12* values between 0 and -300 energy units, and yellow-to-blue covers energies between -300 to -400 units). The order in which the structures are plotted matters; a practical consequence of the exceptional ruggedness of protein energy surfaces is that similar structures can have markedly different energies. In particular, when the hall of fame is composed of hundreds of thousands of individuals corresponding to PC-projections of all-atom structures, it is very common to have structures with similar projections on the top two PCs; as a result, many points in the 2d projection of the hall of fame can be plotted on top of another; so, order is important for the high-energy structures not to occlude the view of low-energy, nearby structures. Since our goal here is to see whether low-energy structures are uncovered for a particular protein, the structures in the hall of fame are sorted based on their energies (from high to low energies). Plotting is then carried out in this order; the 2d projections of the structures with higher energies are plotted first, as it is acceptable for them to be covered by projections of lower-energy structures.

4.3 Impact of Lineage- and Neighborhood-aware Improvement Operator on the Hall of Fame

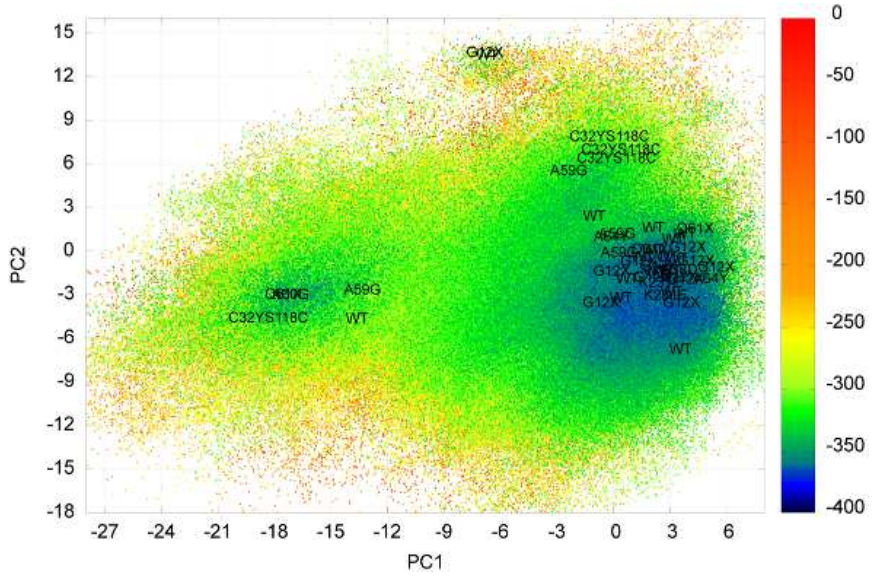
Fig. 1 juxtaposes the hall of fame obtained for H-Ras WT with the local improvement operator, shown in Fig. 1(a) to the hall of fame obtained by `PELMap-EA` with the lineage- and neighborhood-aware improvement operator, shown in Fig. 1(b). Each run of `PELMap-EA` uses the same budget of energy/fitness evaluations. On each plot, the X-ray structures are shown as well, by drawing their projections on the top two PCs. The annotations show whether the X-ray structures are reported for the WT or variants; in the case of variants, the specific mutation is denoted.

Comparison of the halls of fame in Fig. 1 shows that the lineage- and neighborhood-aware improvement operator is able to find more lower-energy structures. In particular, the new operator confers `PELMap-EA` with the ability to reproduce the "on" and "off" basins of H-Ras more faithfully. The off basin corresponds to the neighborhood of low-energy structures left of the PC1=-9 line, which also contains the projection of the "off" structure deposited in PDB entry 4q21; the "on" basin corresponds to the neighborhood of low-energy structures right of the PC1=-9 line, which also contains the projection of the "on" structure deposited in

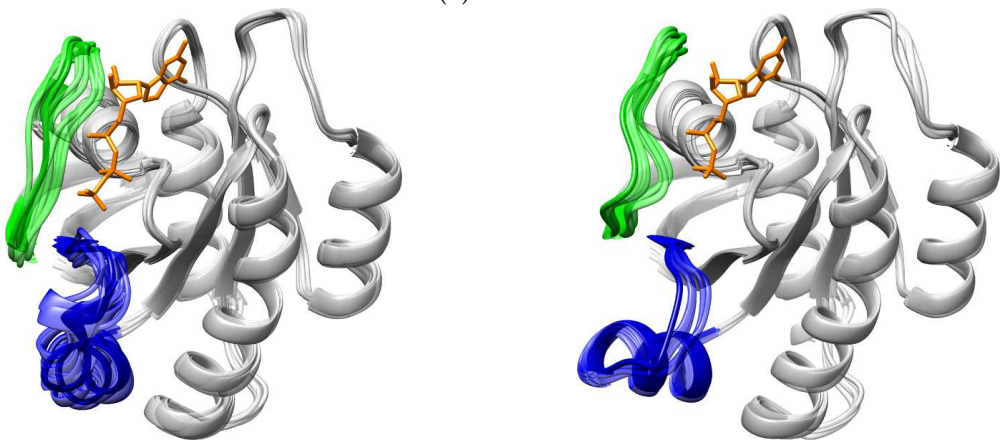
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2628745, IEEE/ACM Transactions on Computational Biology and Bioinformatics



(a)



(b)



(c)

Fig. 1. Halls of fame obtained for the H-Ras WT by PELMap-EA when using the local improvement operator, shown in (a), versus when using the lineage- and neighborhood improvement operator, shown in (b). Points are color-coded by Rosetta *score12* values of corresponding individuals in the hall of fame. Projections of the X-ray structures on the top two PCs are also shown. The annotations denote whether the X-ray structures correspond to WT or variants. In the case of variants, the mutations are listed. (c) shows 10 lowest-energy structures of the "off" basin (left) and "on" basin (right), drawn in transparent and superimposed over the known WT structures drawn in opaque (left: PDB id 4q21; right: PDB id 1qra). The SI and SII regions (residues $25-40$ and $57-75$, respectively) are drawn in green and blue, respectively. GDP (bound to "off" state) and GTP (bound to "on" state) are drawn in orange. Rendering is performed with chimera 10.1 [51].

PDB entry 1qra. The 10-lowest energy structures from each basin are selected and then drawn superimposed over the known structure corresponding to each basin in Fig. 1(c). Based on the structural similarity between structures selected from a basin and the wet-laboratory structure caught in that basin, PELMap−EA uncovers the correct "on" and "off" basins. Based on structural comparisons inside a basin and between basins (statistics are provided in Section E.3 in Appendix E), the intra-basin structural variations are roughly half of the inter-basin structural variations in H-Ras. The inter-basin structural variations are predominantly located on the SI and SII regions, as Fig. 1(c) shows; these regions undergo a concerted open-to-close motion as H-Ras transitions from the "on" to the "off" basin (as H-Ras binds GTP). The intra-basin structural variations are local and distributed over the H-Ras chain. More statistics regarding structural comparisons (and more structures) are related in Section E.3 in Appendix E.

In addition to uncovering the "on" and "off" basins, PELMap−EA finds other interesting basins possibly corresponding to semi-stable states that are difficult to catch in the wet laboratory. One such basin of interest, visible in the PC1-PC2 embedding in Fig. 1(b), emerges over the "on" basin, around the range $[4, 6]$ in PC2. Known wet-laboratory structures (under PDB entries 1lf0 and 6q21(D)) project in this region. The structure under PDB entry 1lf0 belongs to an H-Ras variant [52] and is structurally between the "on" and "off" known structures (PDB entries 4q21 and 1qra, respectively). PELMap−EA suggests these stable variant structures are low in energy and present possible semi-stable, short-lived states populated by H-Ras WT while transitioning between the stable "on" and "off" states. This finding further justifies employing stable structures of variants in the PCA analysis and the initial population.

The 2D embeddings shown Fig. 1 may hide interesting findings by essentially showing only about 54% of the H-Ras dynamics (PC1 and PC2 collectively capture 54% of the variance among known structures). It is only when considering 4d embeddings of the probed energy surface that interesting observations can be made regarding the ability of PELMap−EA to uncover new regions of the H-Ras energy landscape not yet probed in wet or dry laboratories.

To relate 4d embeddings, we make use of two-way conditioned quantile plots. The hall of fame is sorted along PC3 and PC4 and partitioned along the four quantiles of each PC; this results in 16 partitions, considering all PC3:Q[1-4] and PC4:Q[1-4] pairs. The individuals residing in some particular PC3:Q* and PC4:Q* partition are shown as before, by plotting their PC1-PC2 coordinates and color-coding them based on energy values. Appendix C shows all 16 plots. Fig. 2 shows 2 selected plots that demonstrate PELMap−EA's ability to uncover new regions along the [PC3:Q1, PC4:Q1] (left panel) and the [PC3:Q2, PC4:Q3] partition (right panel).

The "off" basin disappears in the [PC3:Q1, PC4:Q1] partition, whereas the "on" basin occupies a significant region of the space and reaches very deep in the energy landscape. A new low-energy region emerges at the top, and two structures project to it, one caught for the WT and the other for the G12V variant. This region presents a possibly new basin not well populated by structures probed in the wet laboratory due to its higher energy than the "on" basin;

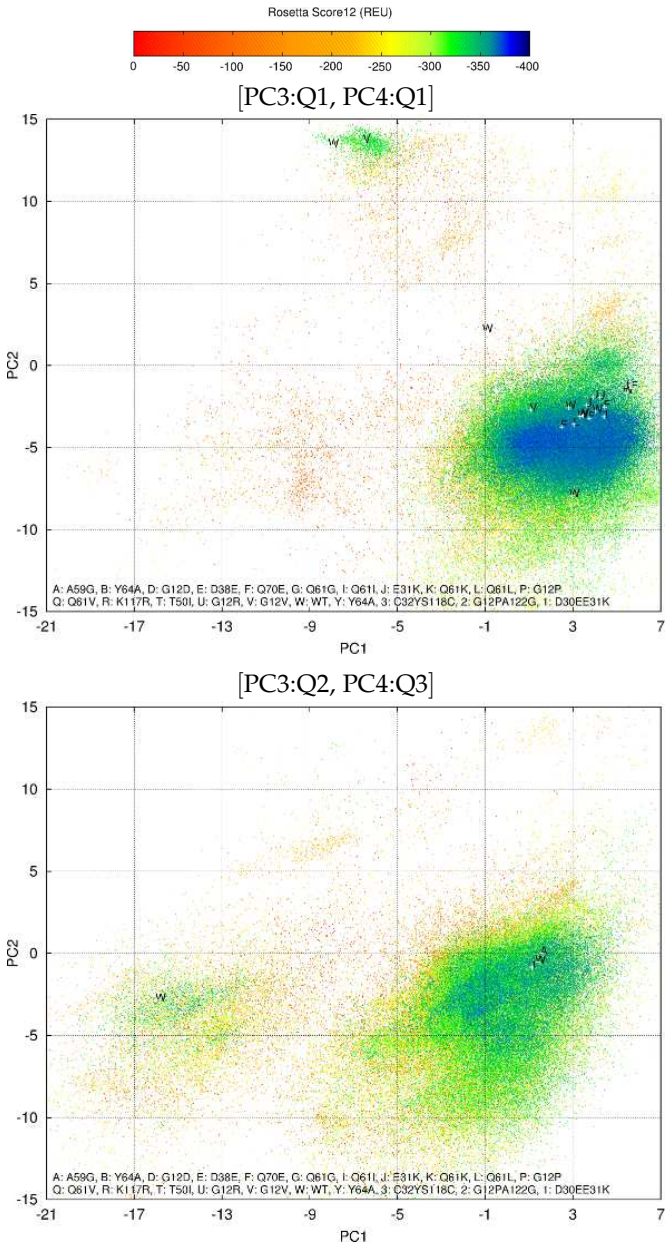


Fig. 2. Halls of fame obtained for the H-Ras WT by PELMap−EA with the new improvement operator are conditioned along quantiles of PC3 and PC4; Two such partitions are shown here, by projecting the structures mapping to a specific partition on the top two PCs and color-coding their projections by Rosetta *score12* energy values. Projections of X-ray structures that fall in a partition are also shown, indicating the sequence variant to which they belong.

however, this region may provide interesting insight into the structure switching mechanism in H-Ras. The [PC3:Q2, PC4:Q3] partition in Fig. 2 shows regions of the "on" basin not populated by any wet-laboratory structures, effectively uncovering a broader basin than would be indicated by the distribution of known wet-laboratory structures.

The two partitions shown (and the rest of the 16 possible ones related in Appendix C) allow reaching two conclusions. First, PELMap−EA with the new improvement operator uncovers new basins and new regions of known basins, thus supplementing wet-laboratory knowledge; Second, on H-Ras, the "on" basin is broader and deeper than the "off"

basin, and thus preferred at equilibrium by the uncomplexed H-Ras, thus poising H-Ras for preferential binding to GTP. These conclusions provide more detailed insight into the role of dynamics in function modulation in H-Ras. Once binding GTP, H-Ras cleaves the terminal phosphate of GTP, converting it to GDP. This conversion makes the "on" state less preferred, and prompts switching into its "off" state. This switching is very slow, and PELMap-EA provides a possible reason. The much deeper and broader "on" basin increases the time it takes to escape it via thermal fluctuations at equilibrium. Accessory proteins are needed to speed up the conversion and allow H-Ras to release GDP. The release prompts H-Ras to seek its most stable, uncomplexed state and promptly populate the "on" basin, starting the structure-switching mechanism anew.

Fig. 3 now compares the halls of fame obtained for HIV-I Protease and CaM without and with the novel improvement operator. Fig. 3(a)-(b) shows a broad, shallow basin for HIV-I Protease that is only captured by the novel improvement operator. Section E.1 in Appendix E shows two structures selected from the basin. The broad basin is in agreement with work in [41], which also shows a wide basin populated by many variants of HIV-I protease. The broad basin also provides a rationale for wet-laboratory observations that HIV-I protease has a fast mutation rate and yet forms stable monomers; in a broad basin, HIV-I protease can undergo mutations and yet populate diverse stable, functional structures. Some very low-energy structures are captured by the novel improvement operator. They correspond to projections in the $[0, 2]$ range on PC1 and $[-4, -2]$ range in PC2 and may be of interest to researchers investigating compound binding to HIV-I Protease.

Fig. 3(c)-(d) shows that the novel improvement operator is able to obtain many more lower-energy structures for CaM and obtain a more detailed map of CaM's energy landscape. The halls of fame for CaM have a characteristic hollow shape and point to an exceptionally complex landscape with many high-energy structures. In particular, the Rosetta *score12* all-atom energy function assigns high, unfavorable energies to many generated CaM structures; so the energy threshold `fitThreshold` in the hall of fame is set to 250 instead of 0 (the latter was sufficient for H-Ras and HIV-I protease). Many high-energy structures are found in the middle of the variable space. Section E.2 in Appendix E shows some of the lowest-energy selected structures and superimposes them over structures caught in wet laboratories, providing indication of what structural states correspond to the uncovered basins.

4.4 Comparison of H-Ras WT and Variant Landscapes

Having established the superiority of the novel, lineage- and neighborhood-aware improvement operator over the local improvement operator, we now apply PELMap-EA with the novel improvement operator to study landscapes of H-Ras variants. Specifically, we obtain halls of fame for several single- and double-mutant variants of H-Ras, such as G12V, G12D, G12S, Q61L, and C32YS118C of H-Ras (PELMap-EA is rerun in each instance, with the specific sequence under investigation as input). No significant differences are observed between the halls of fame obtained on G12D, G12S, C32YS118C and the WT, though in some variants shallower basins are obtained compared to the WT (data not shown). Differences are observed between the WT and two variants, Q61L and G12V, whose halls of fame are shown in Appendix D. Visual comparison of the halls of fame of these two variants suggests that the on/active basin is shallower and not as well populated in G12V as in Q61L. Moreover, comparison with the WT hall of fame drawn in Fig. 1(b) suggests that the "on" basin is much deeper and broader in the WT than in the variants. To quantify the visually-observed differences, density of state plots are drawn as histograms in Fig. 4. The comparison focuses on structures with energies no higher than -100 Rosetta energy units (REUs). Results for the WT are drawn in blue for reference, with the results for the two variants drawn in red.

Fig. 4 makes it clear that there are more structures in the halls of fame found for the WT than Q61L and G12V, and this contrast is particularly striking for Q61L. This is likely to be the result of PELMap-EA finding more often higher-energy structures for the variants than the WT, when the energy threshold is set to 0 for the inclusion test in the hall of fame. Comparison of the histograms also clarifies that more lower-energy structures are found for the WT than the two variants. This is particularly striking when comparing G12V to the WT; the number of structures with energies between -350 and -360 is much lower for G12V than the WT. Most of the lowest-energy structures that exist in the hall of fame obtained for the WT do not have any corresponding counterparts in the hall of fame obtained for G12V. This explains the shallower "on" basin in G12V as compared to the "on" basin in the WT. In summary, this analysis suggests that the mutations G12V and Q61L have a direct impact on the depth of the "on" basin. In general, the landscapes of the variants are elevated compared to the WT, and this is particularly the case for the "on" basin.

5 Conclusion

While there is merit in pursuing algorithmic treatments that operate in a *de novo* setting, the structures available for many proteins nowadays can be used to significantly improve the performance of computational methods. In the research reported here, the use of existing structural information has allowed us to formulate a powerful EA that can obtain comprehensive and detailed maps of complex protein energy landscapes. The hall of fame mechanism can be effectively employed to represent the protein energy landscape. The resolution of the representation can be controlled via the `distThreshold` parameter. This work has also contributed a novel improvement operator that is unique, to the best of our knowledge, in its direct interaction with the evolutionary history of an individual and the hall of fame, and in its indirect interaction with the selection operator. This lineage- and neighborhood-aware improvement operator is likely to be appealing to the broader evolutionary computation community.

The work presented here opens up several prospects for future research. One concerns pursuing nonlinear dimensionality reduction techniques to extract collective variables. Future work needs to address how to directly sample in a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2628745, IEEE/ACM Transactions on Computational Biology and Bioinformatics

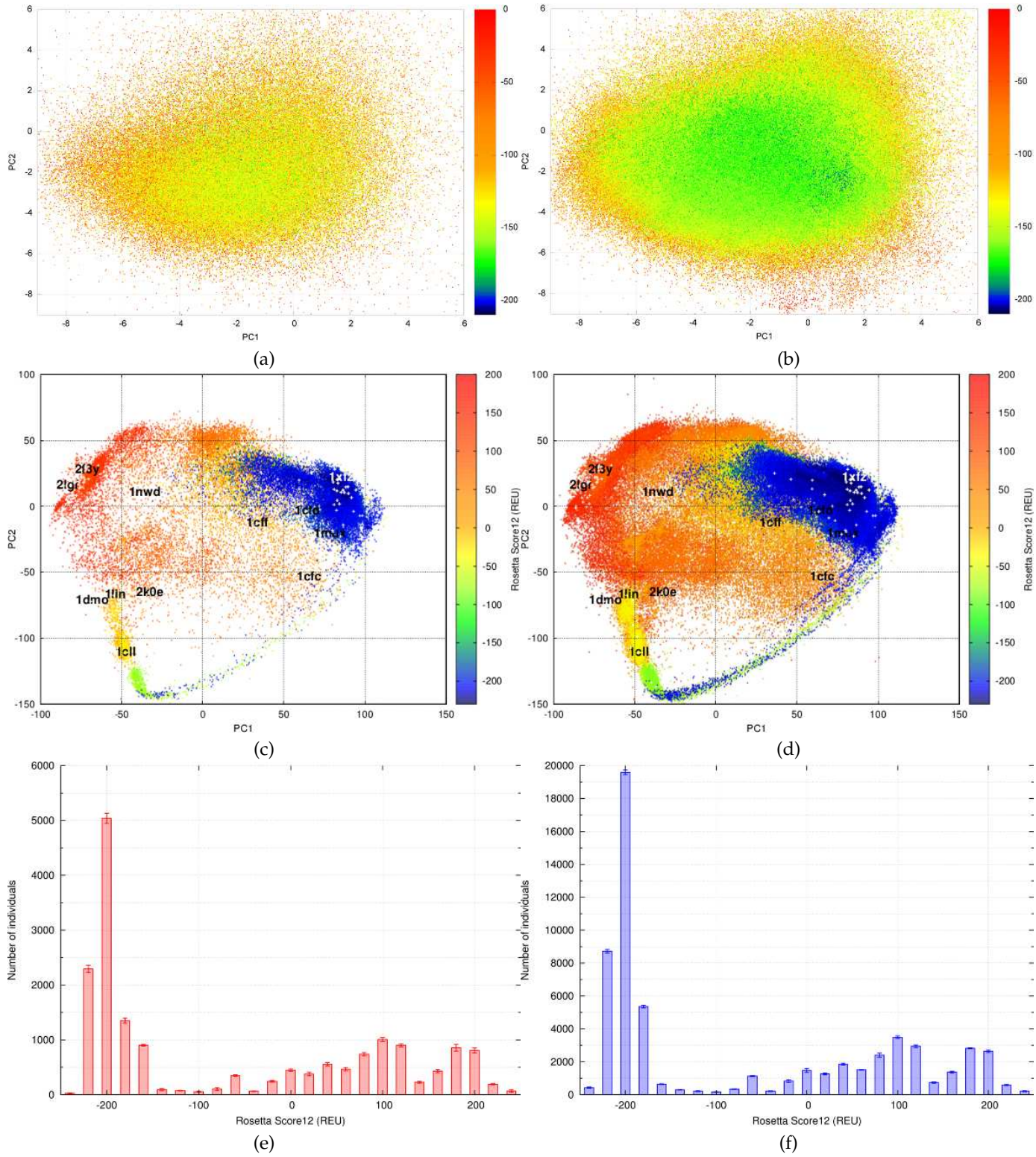
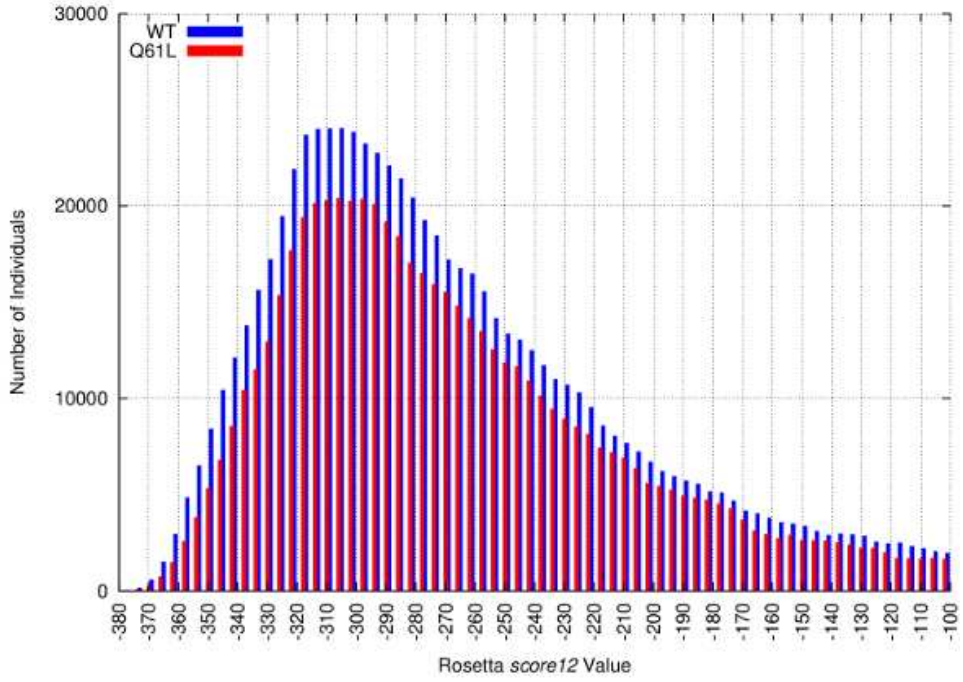
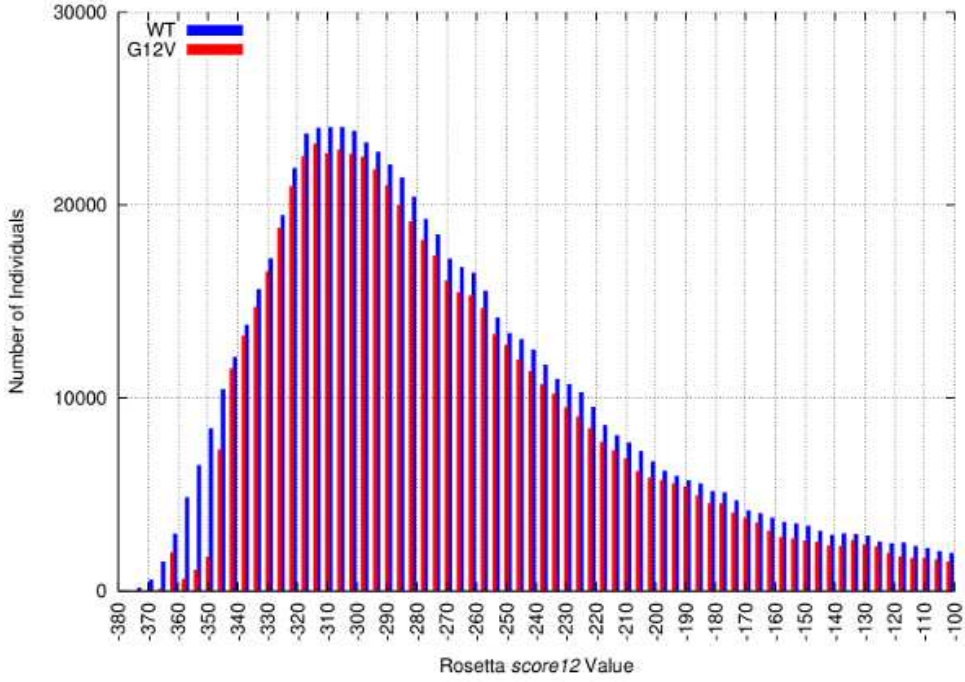


Fig. 3. Halls of fame obtained for HIV-I Protease by PELMap−EA when using the local improvement operator, shown in (a), versus when using the lineage- and neighborhood improvement operator, shown in (b). The juxtaposition of the local vs. the lineage- and neighborhood improvement operator is shown for CaM in (c) and (d), respectively. Points are color-coded by Rosetta *score12* values of corresponding individuals. (e) and (f) show the distribution of energies of individuals in the hall of fame without (e) and with the new improvement iterator (f) for CaM. The error bars show the magnitude of deviations over 5 runs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2628745, IEEE/ACM Transactions on Computational Biology and Bioinformatics



Q61L versus WT



G12V versus WT

Fig. 4. Density of state plots show the distribution of individuals in the halls of fame of Q61L and G12V variants (red) versus WT (blue).

non-linear variable space. Another direction concerns improving the reliability of predictions on the locations and depths of computed basins by employing various energy functions, given the presence of intrinsic errors in energy functions.

The proposed PELMap-EA has elucidated interesting findings regarding stable and semi-stable structural states of various proteins that can aid further studies. The results obtained for HIV-I protease can be used to anticipate drug resistance; projecting newly-found structures on the obtained landscape may reveal similarity with already-characterized structures and thus allow transferring information on stability, drug resistance, and possible effective inhibitors. The results obtained for CaM point to a complex energy landscape, with possibly more information to be mined for obtaining a mechanistic insight into this protein's rich set of structural changes. The results obtained for the different sequences of H-Ras, which in addition to reproducing wet-laboratory knowledge also point to the existence of possible semi-stable states yet to be discovered in wet laboratories, may aid the

quest for inhibitors to treat cancer. The area around the H-Ras "on" basin is very rugged and may be used in variants to alter the "on"-to-"off" switching routes and in turn the ability of H-Ras to regulate molecular recognition events. Structures in new, low-energy regions found by PELMap-EA may present interesting novel targets both to understand H-Ras regulation and to design possible novel molecular interventions.

Several studies have shown that the structural plasticity of proteins can be directly exploited for therapeutic benefit [24]; a prominent example is that of the cancer drug Gleevic/imatinib, which selectively binds and stabilizes an inactive form of Abl kinases. A comprehensive knowledge of the set of stable and semi-stable structural states of a protein may facilitate selective targeting of distinct inactive structures to alter structure switching [53]. By aiming to provide comprehensive and detailed maps of energy landscapes with reasonable computational budgets (5-15 days of CPU time), and by demonstrating that this budget allows obtaining maps of WT and variants of a protein that can then be directly compared, results obtained by PELMap-EA may be of use to molecular intervention studies. To this end, we make all structures and corresponding Rosetta *score12* energies in the halls of fame obtained by PELMap-EA available upon request.

ACKNOWLEDGMENT

This work is supported in part by NSF-CCF 1421001 and NSF-IIS 1144106. We thank Dr. Tatiana Maximova for assistance with rendering in chimera.

REFERENCES

[1] C. A. Floudas and P. M. Pardalos, *Encyclopedia of Optimization*. Norwell, MA: Kluwer Academic Publishers, 2001.

[2] J. Pinter, "Nonconvex optimization and its applications," in *Global Optimization: Scientific and Engineering Case Studies*, ser. Mathematics and Statistics, P. Panos, Ed. New York, NY: Springer Science and Business Media, 2006, vol. 85.

[3] F. Werner, "A survey of genetic algorithms for shop scheduling problems," in *Mathematics and Statistics*, ser. Heuristics: Theory and Applications, P. Siarry, Ed. Nova Science Publishers, 2013, pp. 161–222.

[4] D. Whitley, D. Hains, and A. E. Howe, "A hybrid genetic algorithm for the traveling salesman problem using generalized partition crossover," in *Parallel Problem Solving from Nature (PPSN)*, September 2010, pp. 566–575.

[5] A. Grosso, A. R. Jamali, M. Locatelli, and F. Schoen, "Solving the problem of packing equal and unequal circles in a circular container," *J. Global Optim.*, vol. 47, no. 1, pp. 63–81, 2010.

[6] D. Devaurs, K. Molloy, M. Vaisset, and A. Shehu, "Characterizing energy landscapes of peptides using a combination of stochastic algorithms," *IEEE Trans NanoBioScience*, vol. 14, no. 5, pp. 545–552, 2015.

[7] B. Olson and A. Shehu, "Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface," *Proteome Sci*, vol. 10, no. 10, p. S5, 2012.

[8] ——, "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.

[9] B. Olson, K. A. De Jong, and A. Shehu, "Off-lattice protein structure prediction with homologous crossover," in *Conf on Genetic and Evolutionary Computation (GECCO)*. New York, NY: ACM, 2013, pp. 287–294.

[10] 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)*, Washington, D. C., September 2013, pp. 430–439.

[11] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, "Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function," *J Comput Chem*, vol. 19, no. 14, pp. 1639–1662, 1998.

[12] J. Esquivel-Rodriguez, Y. D. Yang, and D. Kihara, "Multi-lzerd: Multiple protein docking for asymmetric complexes," *Proteins: Struct. Funct. Bioinf.*, vol. 80, no. 7, pp. 1818–1833, 2012.

[13] I. Hashmi and A. Shehu, "idDock+: Integrating machine learning in probabilistic search for protein-protein docking," *J Comp Biol*, vol. 22, no. 9, pp. 1–18, 2015.

[14] ——, "Hopdock: A probabilistic search algorithm for decoy sampling in protein-protein docking," *Proteome Sci*, vol. 11, no. Suppl 1, p. S6, 2013.

[15] M. Rusu and S. Birmanns, "Evolutionary tabu search strategies for the simultaneous registration of multiple atomic structures in cryo-EM reconstructions," *J Struct Biol*, vol. 170, no. 1, pp. 164–171, 2010.

[16] A. Shehu, "A review of evolutionary algorithms for computing functional conformations of protein molecules," in *Computer-Aided Drug Discovery*, ser. Springer Methods in Pharmacology and Toxicology, W. Zhang, Ed. Springer Verlag, 2015.

[17] K. Jenzler-Wildman and D. Kern, "Dynamic personalities of proteins," *Nature*, vol. 450, pp. 964–972, 2007.

[18] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu, "Principles and overview of sampling methods for modeling macromolecular structure and dynamics," *PLoS Comput Biol*, vol. 12, no. 4, p. e1004619, 2016.

[19] P. Stadler, "Fitness landscapes," *Appl Math & Comput*, vol. 117, pp. 187–207, 2002.

[20] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, pp. 980–980, 2003.

[21] D. D. Boehr, R. Nussinov, and P. E. Wright, "The role of dynamic conformational ensembles in biomolecular recognition," *Nature Chem Biol*, vol. 5, no. 11, pp. 789–96, 2009.

[22] D. Russel, K. Lasker, J. Phillips, D. Schneidman-Duhovny, J. A. Veláquez-Muriel, and A. Sali, "The structural dynamics of macromolecular processes," *Curr Opin Cell Biol*, vol. 21, no. 1, pp. 97–108, 2009.

[23] B. Ma, S. Kumar, C. Tsai, and R. Nussinov, "Folding funnels and binding mechanisms," *Protein Eng*, vol. 12, no. 9, pp. 713–720, 1999.

[24] B. J. Grant, A. A. Gorfe, and A. McCammon, "Large conformational changes in proteins: signaling and other functions," *Curr Opin Struct Biol*, vol. 20, no. 2, pp. 142–147, 2010.

[25] E. Babini, I. Bertini, F. Capozzi, C. Luchinat, A. Quattrone, and M. Turano, "Principal component analysis of the conformational freedom within the EF-hand superfamily," *J Proteome Res*, vol. 4, no. 6, pp. 1961–1971, 2005.

[26] A. A. Gorfe, B. J. Grant, and J. A. McCammon, "Mapping the nucleotide and isoform-dependent structural and dynamical features of Ras proteins," *Structure*, vol. 16, no. 6, pp. 885–896, 2008.

[27] B. J. Grant, A. A. Gorfe, and J. A. McCammon, "Ras conformational switching: Simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics," *PLoS. Comput. Biol.*, vol. 5, no. 3, p. e1000325, 2009.

[28] C. C. David and D. J. Jacobs, "Principal component analysis: A method for determining the essential dynamics of proteins," in *Protein Dynamics*, ser. Methods in Molecular Biology, 2013, vol. 1084, pp. 193–226.

[29] M. Gur, J. D. Madura, and I. Bahar, "Global transitions of proteins explored by a multiscale hybrid methodology: application to adenylate kinase," *Biophys J*, vol. 105, no. 7, pp. 1643–1652, 2013.

[30] Y. Liu and I. Bahar, "Sequence evolution correlates with structural dynamics," *Mol Biol Evol*, vol. 29, no. 9, pp. 2253–2263, 2014.

[31] L. Skjaerven, X. Q. Yao, G. Scarabelli, and B. J. Grant, "Integrating protein structural dynamics and evolutionary analysis with Bio3D," *BMC Bioinf*, vol. 15, no. 399, pp. 1–11, 2014.

[32] R. Clausen and A. Shehu, "Exploring the structure space of wild-type Ras guided by experimental data," in *ACM Conf on Bioinf and Comp Biol Workshops (BCBW)*, Washington, D. C., September 2013, pp. 757–764.

[33] R. Clausen, E. Sapin, K. A. De Jong, and A. Shehu, "Evolution strategies for exploring protein energy landscapes," in *Genet Evol Comput Conf (GECCO)*. New York, NY, USA: ACM, July 2015, pp. 217–224.

[34] R. Clausen and A. Shehu, "A multiscale hybrid evolutionary algorithm to obtain sample-based representations of multi-basin protein energy landscapes," in *ACM Conf on Bioinf and Comp Biol (BCB)*, Newport Beach, CA, September 2014, pp. 269–278.
[35] ——, "A data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes," *J Comp Biol*, vol. 22, no. 9, pp. 844–860, 2015.
[36] R. Clausen, B. Ma, R. Nussinov, and A. Shehu, "Mapping the conformation space of wildtype and mutant H-Ras with a memetic, cellular, and multiscale evolutionary algorithm," *PLoS Comput Biol*, vol. 11, no. 9, p. e1004470, 2015.
[37] E. Sapin, K. A. De Jong, and A. Shehu, "Evolutionary search strategies for efficient sample-based representations of multiple-basin protein energy landscapes," in *IEEE Intl Conf Bioinf and Biomed (BIBM)*, 2015, pp. 13–20.
[38] E. Anderson *et al.*, "LAPACK: A portable linear algebra library for high-performance computers," in *ACM/IEEE Conf Supercomputing*, Los Alamitos, CA, USA, 1990, pp. 2–11.
[39] 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*, vol. 49, no. 14, pp. 2987–2998, 2010.
[40] 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.*, vol. 28, no. 29, pp. 1593–1597, 2007.
[41] M. W. Chang, "Computational structure-based methods to anticipate hiv drug resistance evolution and accelerate inhibitor discovery," Ph.D. dissertation, University of California, San Diego, 2008.
[42] C. Tsai, B. Ma, and R. Nussinov, "Folding and binding cascades: shifts in energy landscapes," *Proc. Natl. Acad. Sci. USA*, vol. 96, no. 18, pp. 9970–9972, 1999.
[43] 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*, vol. 16, no. 5, pp. 736–746, 2008.
[44] 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.*, vol. 328, no. 1, pp. 193–204, 2003.
[45] B. W. Zhang, D. Jasnow, and D. M. Zuckermann, "Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin," *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 46, pp. 18 043–18 048, 2007.
[46] M. V. Milburn, L. Tong, A. M. deVos, A. Brünger, Z. Yamaizumi, S. Nishimura, and S. H. Kim, "Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic Ras proteins," *Science*, vol. 247, no. 4945, pp. 939–945, 1990.
[47] A. J. Scheidig, C. Burmester, and R. S. Goody, "The pre-hydrolysis state of p21(Ras) in complex with GTP: new insights into the role of water molecules in the GTP hydrolysis reaction of Ras-like proteins," *Structure*, vol. 7, no. 11, pp. 1311–1324, 1999.
[48] M. Teodoro, G. N. J. Phillips, and L. E. Kavraki, "Understanding protein flexibility through dimensionality reduction," *J Comput Biol*, vol. 10, no. 3–4, pp. 617–634, 2003.
[49] A. E. Karnoub and R. A. Weinberg, "Ras oncogenes: split personalities," *Nat Rev Mol Cell Biol*, vol. 9, no. 7, pp. 517–531, 2008.
[50] A. Fernández-Medarde and E. Santos, "Ras in cancer and developmental diseases," *Genes Cancer*, vol. 2, no. 3, pp. 344–358, 2011.
[51] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera–a visualization system for exploratory research and analysis," *J Comput Chem*, vol. 25, no. 13, pp. 1605–1612, 2004.
[52] B. E. Hall, D. Bar-Sagi, and N. Nassar, "The structural basis for the transition from Ras-GTP to Ras-GDP," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 19, pp. 12 138–12 142, 2002.
[53] B. J. Grant, S. Lukman, H. J. Hocker, J. Sayyah, J. H. Brown, J. A. McCammon, and A. A. Gorfe, "Novel allosteric sites on Ras for lead generation," *PLoS ONE*, vol. 6, no. 10, p. e25711, 2011.

Emmanuel Sapin Dr. Sapin is a postdoctoral fellow in the department of Computer Science at George Mason University. His research interests include computational structural biology, stochastic optimization methods, nature-inspired algorithms with a focus on evolutionary algorithms, and ant colony optimization in various domains, from emergence of complexity in discrete models to genome wide association studies and protein structure modeling. Sapin is a member of IEEE and ACM.

Kenneth A De Jong Dr. De Jong is a Professor of Computer Science and Associate Director of the Krasnow Institute at George Mason University. His research interests include genetic algorithms, evolutionary computation, machine learning, and adaptive systems. He is currently involved in research projects involving the development of new evolutionary algorithm (EA) theory, the use of EAs as heuristics for NP-hard problems, and the application of EAs to the problem of learning task programs in domains such as robotics, diagnostics, navigation and game playing. Support for these projects is provided by NSF, DARPA, ONR, and NRL. He is an active member of the Evolutionary Computation research community and has been involved in organizing many of the workshops and conferences in this area. He is the founding editor-in-chief of the journal Evolutionary Computation (MIT Press), and a member of the board of ACM SIGEVO. De Jong is a member of the IEEE and ACM.

Amarda Shehu Dr. Shehu is an Associate Professor in the Department of Computer Science at George Mason University. She holds affiliated appointments in the Department of Bioengineering and School of Systems Biology. Shehu's research contributions are in computational structural biology, biophysics, and bioinformatics with a focus on issues concerning the relationship between sequence, structure, dynamics, and function in biomolecules. Shehu's expertise is in tight coupling of robotics-inspired and evolutionary search techniques with domain-specific insight in biophysics for modeling equilibrium biomolecular structures and dynamics. Shehu's research is supported by various NSF programs, including Intelligent Information Systems, Computing Core Foundations, and Software Infrastructure. Shehu is the recipient of an NSF CAREER award and a member of the IEEE and ACM.