A Novel EA-based Memetic Approach for Efficiently Mapping Complex Fitness Landscapes

Emmanuel Sapin
Dept of Computer Science
George Mason University
Fairfax, VA 22030
esapin@gmu.edu

Kenneth De Jong
Dept of Computer Science
George Mason University
Fairfax, VA 22030
kdejong@gmu.edu

Amarda Shehu*
Dept of Computer Science
George Mason University
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT

Recent work in computational structural biology focuses on modeling intrinsically dynamic proteins important to human biology and health. The energy landscapes of these proteins are rich in minima that correspond to alternative structures with which a dynamic protein binds to molecular partners in the cell. On such landscapes, evolutionary algorithms that switch their objective from classic optimization to mapping are more informative of protein structure-function relationships. While techniques for mapping energy landscapes have been developed in computational chemistry and physics, protein landscapes are more difficult for mapping due to their high dimensionality and multimodality.

In this paper, we describe a memetic evolutionary algorithm that is capable of efficiently mapping complex landscapes. In conjunction with a hall of fame mechanism, the algorithm makes use of a novel, lineage- and neighborhood-aware local search procedure for better exploration and mapping of complex landscapes. We evaluate the algorithm on several benchmark problems and demonstrate the superiority of the novel local search mechanism. In addition, we illustrate its effectiveness in mapping the complex multimodal landscape of an intrinsically dynamic protein important to human health.

Keywords

mapping; fitness landscape; memetic evolutionary algorithm; protein modeling; energy landscape; computational structural biology.

1. INTRODUCTION

Increasingly, the focus of research in computational structural biology is on understanding the structure-to-function relationship in proteins of central importance to human biology and health. The emerging view of these proteins is that of intrinsically dynamic molecules in perpetual motion [17, 16], accessing and switching between different energetically-similar structures at equilibrium to bind to different molecular partners and thus regulate their complex biological activity in the cell [**?**, 2].

Intrinsically dynamic proteins with a rich array of thermodynamically stable and semi-stable structures with which they can participate in different cellular processes are challenging systems for computational modeling. On such systems, where little may be known about alternative structures that are functionally-relevant, the computational framework for structure modeling shifts from classic optimization to mapping; that is, the objective becomes to map the complex, multimodal energy landscape of a given protein in order to identify the minima in the landscape that correspond to functionally-relevant, stable or semi-stable structures of the protein at equilibrium.

Mapping energy landscapes of organic and inorganic multi-particle systems has been at the forefront of many theoretical developments in computational chemistry and physics [13, 21, 11, 20]. Mapping energy landscapes is now recognized to be key to understanding a wide range of molecular phenomena. While great progress has been made in mapping energy landscapes of atomic clusters [20], glasses [3], and even short peptides [12], mapping protein energy landscapes remains challenging. In systems such as glasses, atomic particles, and short peptides, the number of interacting atoms or particles does not exceed a few hundreds. On such systems, the application of evolutionary algorithms (EAs) can uncover local minima in the population of generated individuals. For instance, work in [21, 20, 12] employs a 1+1 memetic EA, which is effectively an iterated local search, better known as "basin hopping" in the life sciences literature.

EAs that rely mainly on exploitation and limit exploration to naive strategies (such as random restart) lose efficacy rapidly on landscapes of increasing modality. On these landscapes, sophisticated strategies are needed to balance between exploitation and exploration in order to avoid premature convergence. Various strategies have been proposed over the years to address adequate exploration and diversity maintenance, building on the pioneering efforts of Holland, De Jong, Goldberg, and Richardson [10, 4]. For instance, work in [5] proposes a genetic algorithm (GA) that makes use of a histogram-assisted fitness adjustment in order to prevent the GA from converging early to any particular optimum. In [15], a set of multi-population GA operators are proposed for complex fitness landscapes.

In addition to high modality, protein energy landscapes

*Corresponding Author

GECCO '16, July 20-24, 2016, Denver, CO, USA

DOI: http://dx.doi.org/10.1145/2908812.2908829

are also high-dimensional, as even a small-size protein can be composed of thousands of interacting atoms. Recent work proposes a series of increasingly sophisticated EA-based memetic algorithms (MAs) that are shown to be capable of uncovering competing structures corresponding to different minima in a protein energy landscape [7, 8, 9, 6]. Various strategies are shown key to success, such as employing domain-expert examples (experimentally-known structures of a protein) to construct the initial population and define the variable space, structurization of the variable space, decentralization of the selection operator, and others.

However, the majority of existing EAs designed to map complex fitness landscapes do not explicitly build a map of the landscape. Instead, they rely on quantitative or visual analysis of the individuals in the last population or all individuals ever generated to identify populated minima. Efforts to define landscape maps can be found in computer vision and computational structural biology. Two definitions of landscape maps have been operationalized in these fields. In [18], a tree structure is described for computer vision applications. Leaves in the tree are local minima, and internal nodes are barriers between adjacent minima. These tree-based maps are similar to the disconnectivity graphs proposed for biomolecular landscapes in computational structural biology [1]. Both are dependent on Monte Carlo or other dynamics simulation techniques to reveal local maxima or saddles that help a dynamic system navigate between two adjacent local minima in the landscape. Such techniques have limited exploration capability and are impractical in a random restart setting to obtain comprehensive maps of protein energy landscapes.

Recent work has proposed a different definition of a map that is not dependent on the above techniques and is additionally well suited for EAs, maps based on a "hall of fame" mechanism. In [19], an MA is presented that records distinct local minima in a hall of fame, effectively using the hall of fame as a map of the protein energy landscape. Halls of fame of several proteins are shown to contain known competing structures, thus reproducing biological knowledge and lending credibility to the constructed maps.

Inspired by latest work on mapping protein energy landscapes, we present here an EA-based MA for more effectively mapping complex fitness landscapes. The proposed algorithm builds on the latest published work [19], but additionally makes use of a novel, lineage- and neighborhood-aware local search operator. This operator and the hall of fame mechanism interact with each other in a way that allows the MA to more effectively apportion its computational resources towards exploitation of promising regions of the landscape while efficiently updating the global map of the landscape (the hall of fame).

In this paper we demonstrate that, in addition to mapping protein energy landscapes, the proposed EA-based MA is of broader, more general use for mapping landscapes of complex fitness functions. We illustrate this by first analyzing its performance on a set of increasingly-difficult black-box optimization problems. First, we select a problem where the fitness landscape contains a deep and broad basin of attraction from a list of problems proposed as benchmarks for black-box optimization in [14]. Based on this, we then construct two more problems with multimodal landscapes resembling protein energy landscapes. By varying the dimensionality of the variable space, we then obtain problem instances of increasing difficulty. We compare the performance of an MA with a naive local search procedure to that of our MA with its novel, lineage- and neighborhood-aware local search. We show the latter results in better, richer maps of complex fitness landscapes. In addition, we show that our MA also obtains better maps of complex protein energy landscapes and thus pushes forward the promise of EA-based MAs as general tools for mapping protein energy landscapes.

2. METHODS

In order to understand the novel local search procedure, we need to first describe the EA-based MA with a naive local search operator (to which we will refer as MA$^-$ from now on), and then relate details on the novel local search and the resulting MA (to which we will refer as MA$^+$). As mentioned in Section 1, the algorithmic skeleton of MA$^-$ has evolved from recent published work on protein energy landscapes [19]. Here we will summarize it and focus on describing implementations of the operators that allow its application for mapping multimodal fitness landscapes.

MA$^-$, shown in pseudocode in Algorithm 1, initializes the map (the hall of fame) to the empty set (line 1) and the running counter that keeps track of fitness function evaluations (line 2) so as to evaluate and compare MAs using a user-defined budget `FMAX` of fitness evaluations (line 5).

Algo. 1 MA$^-$

Require: `FMAX` //total computational budget
N Population Size
`NrImprovIters` budget for local search

1: $\mathcal{H} \leftarrow \emptyset$ //initialize hall of fame
2: $fcounter \leftarrow 0$ //counter of fitness evaluations
3: $i \leftarrow 0$ //generation
4: $\mathcal{P}_i \leftarrow$ InitOper(N)
5: **while** $fcounter <$ `FMAX` **do**
6: $\mathcal{C} \leftarrow \emptyset$ //set of offspring
7: **for** $\langle p, f \rangle \in \mathcal{P}_i$ **do**
8: $c \leftarrow$ VarOper(p) //generate offspring
9: $\langle c^{'}, f^{'} \rangle \leftarrow$ NaiveLocalSearch(c, `NrImprovIters`)
10: $fcounter \leftarrow fcounter +$ `NrImprovIters`
11: $\mathcal{C} \leftarrow \mathcal{C} \cup \{\langle c^{'}, f^{'} \rangle\}$ //add to offspring set
12: $\mathcal{H} \leftarrow$ UpdateHoF($\mathcal{H}$, $c^{'}$, $f^{'}$) //update hall of fame
13: $i \leftarrow i + 1$
14: $\mathcal{P}_i \leftarrow$ Replacement(P_{i-1},$\mathcal{C}$) //update population

Ensure: $\mathcal{H}$

The initial population is obtained via an initialization mechanism (line 4). For the benchmark problems studied here, coordinates for individuals are drawn uniformly at random from the given parameter ranges. On applications to proteins, different initialization mechanisms can be employed that take advantage of domain-specific knowledge. For example, work in [19] describes an effective initialization mechanism combines individuals drawn at random with those obtained from experts [19].

Once initialized, the population evolves over time as follows. Offspring are recorded in a set $\mathcal{C}$ initialized to the empty set (line 6). Each individual p in population P_i (line 7) is selected to obtain an offspring c via a variation operator (line 8). The variation operator for the benchmark problems studied here is a Gaussian perturbation operator,

which perturbs each coordinate of an individual by a value drawn from a zero-mean Gaussian distribution with a given variance.

The offspring is then subjected to a local search which seeks to improve the offspring. For this study a naive local search (line 9) chooses any of the coordinates of the offspring with equal probability and then applies a simple gradient descent on the chosen coordinate for a total of `NrImprovIters` iterations/cycles. The result of the local search is that the offspring is mapped to a nearby local minimum. Both the offspring c and its fitness f are replaced by the result $c^{'}$, $f^{'}$ of the local search.

The resulting, improved offspring $c^{'}$ is added to the offspring set (line 11), and is also considered for inclusion in the hall of fame (line 12). The algorithm that updates the hall of fame has been described in detail in [19], but we summarize it here for completeness. The offspring $c^{'}$ is compared to existing individuals in the hall of fame $\mathcal{H}$. Two individuals are considered to be similar (i.e., neighbors) if the Manhattan distance between them is below a problem-specific tunable threshold. If an individual i in $\mathcal{H}$ is similar to $c^{'}$, their fitnesses are compared. If $c^{'}$ has better fitness, i is removed from $\mathcal{H}$. The reason for this is that one wants to update the hall of fame with individuals that may reside in the same region in the variable space but allow further exploitation of a local minimum. If $c^{'}$ is not nearby any individual in $\mathcal{H}$, it represents a new region of the variable space, and $c^{'}$ is added to the hall of fame $\mathcal{H}$. More details can be found in [19] of this algorithm that runs in $\mathcal{O}(\textbf{cardinality}(\mathcal{H}))$.

Once N offspring are generated and each has been considered as a potential update the hall of fame, the offspring set $\mathcal{C}$ and the parent set P_i compete for survival (line 14). The replacement operator is also based on our prior work on proteins [8] and it proceeds as follows. Each offspring is compared to those parent individuals within a (tunable) Manhattan distance. The worst-fit parent in that neighborhood is replaced by the offspring. If the neighborhood is empty (there are no nearby individuals), the offspring replaces the worst-fit parent in the entire parent population. This replacement operator is effectively a decentralized or a local selection operator. Work in [9] shows that a decentralized operator retains diversity longer than global selection operators on proteins.

The MA$^-$ thus described has shown some promise on mapping protein energy landscapes [19], but here we present a modification that we demonstrate to be more effective memetic for complex multimodal fitness functions. In such cases, a fixed number of local improvement iterations does not allow apportioning computational resources to the most promising offspring, which is particularly important when improvement of offspring is the main consumer of the total computational budget. A smarter local search needs to know when to give up improving an offspring. It can only do so, however, if it has some view, even if limited, of the history of a particular region of the landscape, and of how the offspring it is asked to improve compares to nearby individuals. For this reason, we propose here a lineage- and neighborhood-aware local search that makes use of history via the concept of a lineage and makes use of a neighborhood by interacting with the hall of fame. The MA that integrates this novel local search, which we refer to as MA$^+$, is shown in pseudocode in Algorithm 2.

Algo. 2 MA$^+$

Require: `FMAX` //total computational budget
N Population Size
`NrImprovIters` budget for local search

1: $\mathcal{H} \leftarrow \emptyset$ //initialize hall of fame
2: $fcounter \leftarrow 0$ //counter of fitness evaluations
3: $i \leftarrow 0$ //generation
4: $\mathcal{P}_i \leftarrow$ InitOper(N)
5: **while** $fcounter <$ `FMAX` **do**
6: $\mathcal{C} \leftarrow \emptyset$ //set of offspring
7: **for** $\langle p, f, fea(p), niters(p)\rangle \in \mathcal{P}_i$ **do**
8: **if** $niters(p) \geq$ `NrImprovIters` **then**
9: CONTINUE //pick another parent
10: $c \leftarrow$ VarOper(p) //generate offspring
11: $niters(c) \leftarrow niters(p)$ //know your lineage
12: $fea(c) \leftarrow fea(p)$
13: $\langle c^{'}, f^{'}, continue\rangle \leftarrow$ AwareLocalSearch(c, $fea(c)$, $niters(c)$, $\mathcal{H}$)
14: $fcounter \leftarrow fcounter + 1$
15: **if** $continue < 1$ **then**
16: $p \leftarrow$ New individual //replace parent
17: $niters(p) \leftarrow 0$ //start new lineage
18: $fea(p) \leftarrow p$ //fitness not evaluated yet
19: **else** //consider offspring for hall of fame
20: $\mathcal{H} \leftarrow$ UpdateHoF($\mathcal{H}$, $c^{'}$, $f^{'}$)
21: $\mathcal{C} \leftarrow \mathcal{C} \cup \{\langle c^{'}, f\rangle\}$ //add to offspring set
22: $i \leftarrow i + 1$
23: $\mathcal{P}_i \leftarrow$ Replacement(P_{i-1},$\mathcal{C}$) //update population

Ensure: $\mathcal{H}$

Because of the novel local search, MA$^+$ needs to do some more bookkeeping over MA$^-$. For each individual p in a population, MA$^+$ needs to record not only its fitness f, but also the fitness of its earliest ancestor, $fea(p)$, and the number of improvement iterations already spent on its lineage, $niters(p)$ (line 7). When the initial population is constructed, $niters$ is set to 0, and the individual is its own earliest ancestor ($fea(p)$ is p). If an individual in a population belongs to a lineage that has exhausted the improvement budget for that lineage, the individual is not selected as a parent (lines 8-9), and MA$^+$. If the budget on a particular lineage has not been exhausted, the selected individual p is subjected to the variation operator (line 10). The resulting offspring c is made aware of its lineage; lines 11-12 show that $niters(c)$ is set to $niters(p)$, as the variation operator does not carry out any improvement, and $fea(c)$ is set to $fea(p)$.

The new local search operator, invoked in line 13 and described in pseudocode form in Algorithm 3 implements the following strategy: using only one improvement iteration together with lineage and neighborhood information, a determination is made as to whether this particular local search thread is worth continuing. If not (line15), its parent is replaced by an individual drawn at random in the variable/coordinate space (line 16), and a new lineage begins (lines 17-18). If the decision is to continue exploring that thread, the offspring is considered for an update to the hall of fame (line 20) and is added to the offspring set (line 21) for the eventual replacement operator (line 23).

These dynamically determined, context-sensitive thread-exploration decisions are the key to an improved use of a

fixed evaluation budget for simultaneous exploration and exploitation of complex multimodal landscapes.

As indicated in Algorithm 3, the lineage- and neighborhood -aware local search operator invokes the naive local search, with a budget of 1 iterations (line 2). Once an improved offspring $c^{'}$ is obtained, information on its lineage is updated (lines 3-4). 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 $\mathcal{H}$ (computed in line 9), are employed in line 10 to determine if this thread is worth further exploration (line 11).

Algo. 3 Aware Local Search

Require: c
$niters(c)$
$fea(c)$
$\mathcal{H}$

1: continue $\leftarrow 0$
2: $\langle c^{'}, f^{'} \rangle \leftarrow$ NaiveLocalSearch(c, 1)
3: $niters(c^{'}) \leftarrow niters(c) + 1$
4: $fea(c^{'}) \leftarrow fea(c)$
5: **if** $niters(c^{'}) == 1$ **then** //one iteration spent so far
6: `continue` $\leftarrow 1$
7: $fea(c^{'}) \leftarrow fea(c)$
8: **else** //compare improved offspring to neighbors in $\mathcal{H}$
9: $\mu_f \leftarrow$ average fitness over neighbors in $\mathcal{H}$
10: **if** $f^{'} < \frac{(fea(c^{'}) + niters(c^{'}) \times \mu_f)}{(1 + niters(c^{'}))}$ **then**
11: `continue` $\leftarrow 1$

Ensure: $\langle c^{'}, f^{'}$, `continue` $\rangle$

3. RESULTS

The key difference between MA$^-$ and MA$^+$ is that, based on context and lineage, MA$^+$ may terminate a local search thread before its budget is exhausted and initiate a new local search thread in its place. From an exploration/exploitation point of view, the idea is to dynamically reallocate exploitation resources to additional exploration. Since our focus is on mapping complex landscapes, our goal is to achieve a win-win situation in which the constructed maps exhibit both improved coverage (exploration) of the entire landscape while continuing to accurately identify the important local minima (exploitation).

Recall that both MA$^-$ and MA$^+$ construct landscape maps via an archival "hall of fame" mechanism in which entry is based on both fitness and geometric (neighborhood) criteria. If we run both MA$^-$ and MA$^+$ with the same overall search budgets and the same neighborhood criteria, maps that contain more sample points and lower average fitness values provide quantitative evidence for win-win situations in addition to the more qualitative visual evidence provided by plotting the maps produced.

In this section we summarize our initial results. We first analyze the behavior of MA$^-$ and MA$^+$ on 3 generic landscapes, carefully chosen to capture some of the features that make landscapes difficult to map. We follow that with an analysis of the ability of MA$^-$ and MA$^+$ to map the energy landscape of an important protein associated with a degenerative muscle disease in humans.

3.1 Generic Landscape Analysis

To evaluate the differences between MA$^-$ and MA$^+$ on mapping complex fitness landscapes, we began with three carefully chosen problems:

- A sphere with multiplicative noise: $f(x) = \sqrt{\sum_{i=1}^{n} x_i^2} \times noise(x)$
- The sum of two spheres with multiplicative noise: $f(x) = [\sqrt{\sum_{i=1}^{n}(x_i - 200)^2} + \sqrt{\sum_{i=1}^{n}(x_i + 200)^2}] \times noise(x)$.
- The product of two spheres with multiplicative noise: $f(x)\sqrt{\sum_{i=1}^{n}(x_i - 200)^2} \times \sqrt{\sum_{i=1}^{n}(x_i + 200)^2} \times noise(x)$.

where $noise(x)$ is a random number between 0 and 1 determined using a pseudorandom number generator seeded with x such that the noise is random but stays the same for all the evaluations of the same x.

The first function results in a landscape with one global minimum, while the landscapes for the other two contain two broad minima. The dimensionality n for each of these three problems was varied from $\{2, 5, 10, 20\}$, resulting in a total of 12 problem instances. The range of values for all of the coordinates x was fixed at $[-500, 500]$.

Both MA$^-$ and MA$^+$ were applied to each problem instance with a population size of 1000 and a total budget of $1,000,000$ fitness evaluations for each instance. The budget limit for each local search is set to 5.

Figures 1 and 2 summarize quantitative results obtained on all 12 problem instances. Figure 1 compares the number of hall-of-fame individuals obtained by MA$^+$ versus MA$^-$ on each problem instance. Figure 1 shows that MA$^+$ obtains more individuals in the hall of fame (i.e., better exploration) than MA$^-$ on each problem instance, even though both algorithms have the same budget of fitness evaluations.

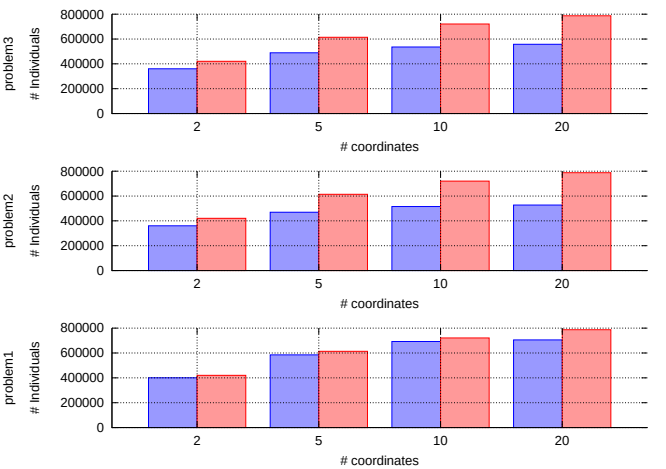


Figure 1: Number of individuals in the halls of fame obtained by MA$^-$ (light gray) and MA$^+$ (dark gray) on each of the 12 problem instances.

Figure 2 compares the average fitness of the $1,000$ lowest (best) individuals in the halls of fame obtained by MA$^+$ versus MA$^-$ on each problem instance. In each case, the averages for MA$^+$ are better/lower than MA$^-$. This provides us with initial quantitative evidence that, for a fixed

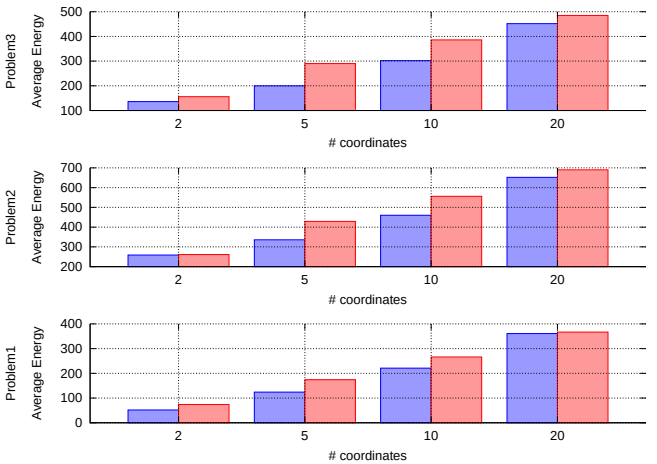


Figure 2: Average energies of the $1,000$ best/lowest-energy individuals in the halls of fame obtained by MA$^-$ (light gray) and MA$^+$ (dark gray) on each of the 12 problem instances.

budget, MA$^+$ improves both exploration and exploitation performace.

Figure 3 and Figure 4 provide some additional representative results of our comparative analysis of MA$^-$ and MA$^+$ on the generic landscapes.

Figure 3(a)-(b) shows typical results on the one-minimum landscape associated with the sphere with multiplicative noise. The halls of fame obtained by MA$^-$ and MA$^+$ are visualized by projecting individuals on the first two coordinates. The projections are color-coded by energies of the corresponding individuals in a blue-to-red color scheme (blue for low energy and red for high energy). A visual omparison of the projected halls of fame shows that MA$^+$ has made better use of its computational budget than MA$^-$: better low-energy coverage without sacrificing overall coverage. This result is made clearer in Figure 3(c), where the distributions of energy-binned individuals in the halls of fame are compared directly and shows that MA$^+$ obtains more individuals with low energy values.

Figure 4 shows typical results obtained on the two-minima landscape associated with the product of two spheres with multiplicative noise. Again we see a significant improvement in the ability of MA$^+$ to map the local minima without sacrificing overall coverage.

Figure 4(c) compares the energy distributions of the individuals in halls of fame generated MA$^-$ and MA$^+$ on the same problem but when $n = 20$. Figure 4(c) shows that on this challenging problem, MA$^+$ finds more individuals in *every* energy range/bin, reinforcing the claim of a win-win situation with respect to both exploration and exploitation.

3.2 Results on a Multi-state Protein

Recall that the application area motivating this research is the development of efficient and effective computational techniques for mapping the complex energy landscapes of proteins that can exist in more than one structural state, and can dynamically switch from one to another. The hypothesis is that these multiple states correspond to minimum energy basins and that knowing their location and the energy barriers between them will provide significantly to our understanding of these proteins.

As an initial assessment of this, we applied both MA$^-$ and MA$^+$ to the problem of mapping the energy landscape of superoxide dismutase (SOD1), an intrinsically dynamic protein implicated in amyotropic lateral sclerosis (ALS) , a muscle degenerative human disorder. As in earlier work [19], individuals here are represented as points in a variable space of 20 collective variables, which are extracted from a principal component analysis of experimentally-known structures of SOD1. The halls of fame produced by MA$^-$ and MA$^+$ are visualized by projecting individuals on the top two (collective) variables/coordinates. Figure 5 shows very clearly that the SOD1 landscape contains two distinct minima. The black dots in Figure 5 are projections of experimentally-known structures of SOD1. Their locations compare well with the location of the two minima found by MA$^-$ and MA$^+$, which suggests that both algorithms have managed to reproduce relevant features of the SOD1 energy landscape. Moreover, MA$^+$ has obtained a denser representation of both minima in its hall of fame. Both minima obtained by MA$^+$ are deeper and wider than those obtained by MA$^-$ given the same computational budget.

4. CONCLUSION

We have described a novel EA-based MA for mapping complex fitness landscapes. Inspired by recent work on mapping protein energy landscapes, the new MA employs an archival hall of fame as an explicit map of the fitness landscape and additionally makes use of a novel, lineage- and neighborhood-aware local search that interacts with both the hall of fame and with the EA survival mechanism. This coupling is specifically designed to provide a dynamic and context-sensitive mechanism for balancing the allocation of a fixed evaluation budget between exploration and exploitation.

The analysis of this new MA on three generic problems of growing difficulty show that the lineage- and neighborhood-aware local search equips the MA with a win-win mapping capability by simultaneously improving both exploration and exploitation. This suggests its usefulness for a broad range of landscape mapping problems from a variety of problem domains.

Since our primary focus is in computational structural biology, we intend to further investigate the proposed MA on mapping energy landscapes of intrinsically dynamic proteins and their disease variants. The preliminary application to SOD1 described in this paper suggests that this MA will perform better than the current state of the art.

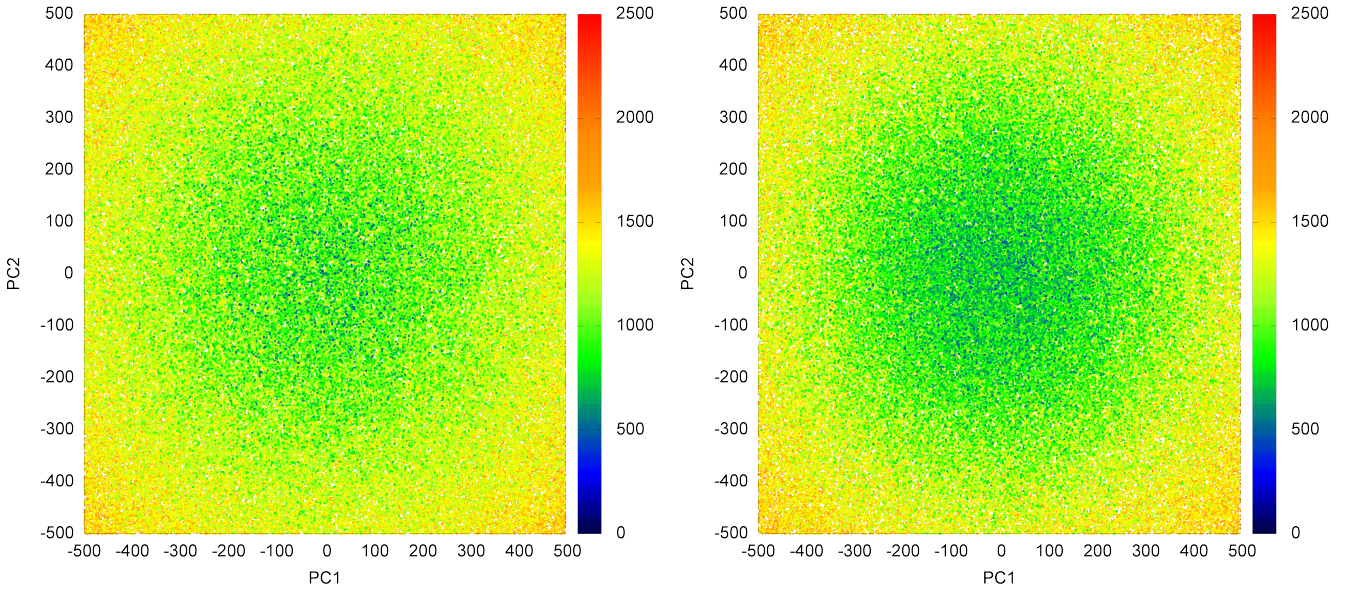
5. ACKNOWLEDGMENTS

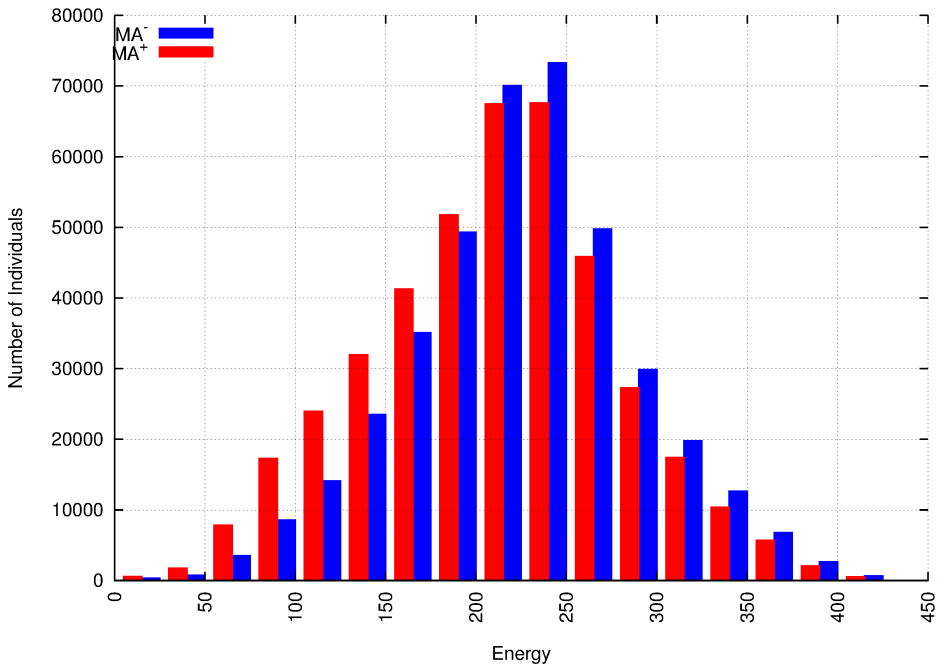
This work is supported in part by NSF CCF No. 1421001 and NSF IIS CAREER Award No. 1144106.

6. REFERENCES

[1] O. M. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J Chem Phys*, 106(4):1495âĂŞ151, 1997.

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





(c)

Figure 3: (a)-(b) Hall of fame maps obtained on the sphere with multiplicative moderate noise at $n = 10$ with $\mathbf{MA}^-$ in (a) and $\mathbf{MA}^+$ in (b). These maps are visualized by projecting individuals on the first two coordinates, color-coding by their energy values. (c) Individuals in the halls of fame obtained by $\mathbf{MA}^-$ and $\mathbf{MA}^+$ binned by their energies. Resulting histogram for $\mathbf{MA}^-$ (in light gray), is compared directly to that for $\mathbf{MA}^+$ (in dark gray).

[3] D. D. Boehr and P. E. Wright. How do proteins interact? 320(5882):1429–1430, 2008.

[4] P. Charbonneau, J. Kurchan, G. Parisi, P. Urbani, and F. Zamponi. Fractal free energy landscapes in structural glasses. *Nat Commun*, 5(4725):3725, 2013.

[5] W. Chen and K. Y. Szeto. Complex energy landscape mapping by histogram assisted genetic algorithm. In *Intl Conf Genet Algorithms (ICGA)*, pages 44–49, 1987.

[6] W. Chen and K. Y. Szeto. Complex energy landscape mapping by histogram assisted genetic algorithm. In *Genet Evol Comput Conf (GECCO)*, pages 673–680. ACM, 2010.

[7] R. Clausen, B. Ma, R. Nussinov, and A. Shehu.

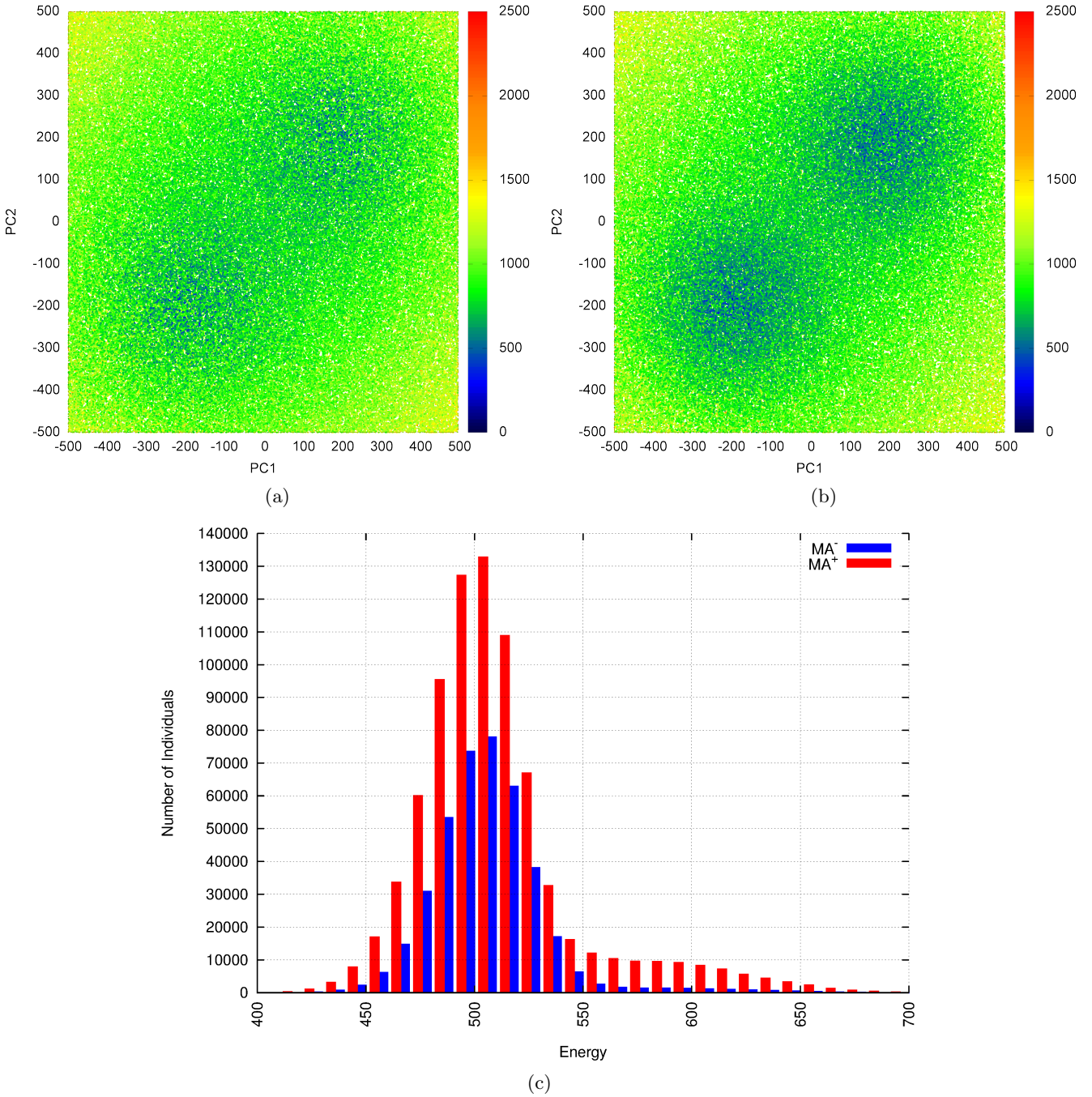
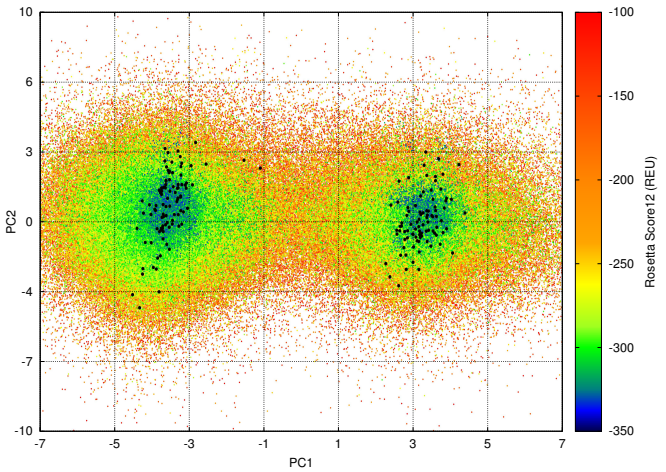


Figure 4: (a)-(b) Hall of fame maps obtained on the product of two spheres with multiplicative moderate noise at $n = 5$ with MA^- in (a) and MA^+ in (b). These halls of fame are visualized by projecting individuals on the first two coordinates, color-coding by their energy values. (c) Individuals in the halls of fame obtained by MA^- and MA^+ when $n = 20$ binned by their energies. Resulting histogram for MA^- (in light gray), is compared directly to that for MA^+ (in dark gray).

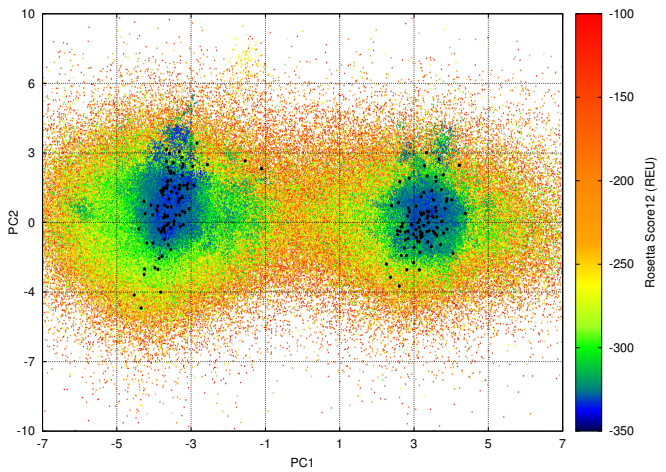
Mapping the conformation space of wildtype and mutant H-Ras with a memetic, cellular, and multiscale evolutionary algorithm. *PLoS Comput Biol*, 11(9):e1004470, 2015.

[8] R. Clausen, E. Sapin, K. A. De Jong, and A. Shehu. Mapping multiple minima in protein energy landscapes with evolutionary algorithms. In *Genet Evol Comput Conf (GECCO)*, pages 923–927, New York, NY, USA, July 2015. ACM.

[9] 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*



(a)



(b)

Figure 5: Halls of fame obtained for the SOD1 protein by MA^- in (a) and MA^+ in (b) are projected on the top two coordinates. Projections are color-coded by Rosetta `score12` energy values of corresponding individuals in the hall of fame. The black dots show projections of experimentally-known structures.

(BCB), pages 269–278, Newport Beach, CA, September 2014.

[10] R. Clausen and A. Shehu. A data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes. *J Comp Biol*, 22(9):844–860, 2015.

[11] K. A. De Jong. An analysis of the behavior of a class of genetic adaptive systems. Master's thesis, University of Michigan, 1975.

[12] P. G. Debenedetti and F. H. Stillinger. Supercooled liquids and the glass transition. *Nature*, 410(6825):259–267, 2001.

[13] D. Devaurs, K. Molloy, M. Vaisset, and A. Shehu. Characterizing energy landscapes of peptides using a combination of stochastic algorithms. *IEEE Trans NanoBioScience*, 14(5):545–552, 2015.

[14] J. P. K. Doye. The network topology of a potential energy landscape: A static scale-free network. *Phys Rev Lett*, 88(23):238701, 2002.

[15] S. Finck, N. Hansen, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2010: Experimental setup. Technical report, 2010.

[16] Y. B. Guo and K. Y. Szeto. Landscape mapping by multi-population genetic algorithm. In *Nature Inspired Cooperative Strategies for Optimization*, volume 236 of *Studies in Computational Intelligence*, chapter 14, pages 165–176. Springer, 2009.

[17] J. S. Hub and B. L. de Groot. Detection of functional modes in protein dynamics. *PLoS Comp Biol*, 5(8):e1000480, 2009.

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

[19] M. Pavlovskaia, K. Tu, and S. Zhu. Mapping the energy landscape of non-convex optimization problems. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 8932, pages 421–435. Springer, 2014.

[20] 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)*, pages 13–20, 2015.

[21] L. C. Smeeton, J. D. Farrell, M. T. Oakley, D. J. Wales, and R. L. Johnston. Structures and energy landscapes of hydrated sulfate clusters. *J Chem Theory Comput*, 11(5):2377âĂŞ2384, 2015.

[22] D. J. Wales, M. A. Miller, and T. R. Walsh. Archetypal energy landscapes. *Nature*, 394(6695):758–760, 1998.