Evolution Strategies for Exploring Protein Energy Landscapes

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

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

The focus on important diseases of our time has prompted many experimental labs to resolve and deposit functional structures of disease-causing or disease-participating proteins. At this point, many functional structures of wildtype and disease-involved variants of a protein exist in structural databases. The objective for computational approaches is to employ such information to discover features of the underlying energy landscape on which functional structures reside. Important questions about which subset of structures are most thermodynamically-stable remain unanswered. The challenge is how to transform an essentially discrete problem into one where continuous optimization is suitable and effective. In this paper, we present such a transformation, which allows adapting and applying evolution strategies to explore an underlying continuous variable space and locate the global optimum of a multimodal fitness landscape. The paper presents results on wildtype and mutant sequences of proteins implicated in human disorders, such as cancer and Amyotrophic lateral sclerosis. More generally, the paper offers a methodology for transforming a discrete problem into a continuous optimization one as a way to possibly address outstanding discrete problems in the evolutionary computation community.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences; I.6.3 [**Computing Methodologies**]: Simulation and Modeling—*Applications*

*Corresponding Author

GECCO '15, July 11 - 15, 2015, Madrid, Spain

DOI: http://dx.doi.org/10.1145/2739480.2754692

Keywords

evolution strategies; protein modeling; energy landscape.

1. INTRODUCTION

The focus on important diseases of our time has prompted many experimental labs to resolve and deposit functional structures of disease-involved proteins in structural databases such as the Protein Data Bank (PDB) [4]. Research in the wet laboratory has shown that proteins involved in some of the most complex human diseases, such as cancer, assume different functional structures and shift between them to modulate their biological function [11]. In disease, changes have occurred to the underlying chemical composition of a protein; namely, particular amino acids in the protein's amino-acid sequence have mutated from the healthy or wildtype (WT) version into a variant version. Variations in sequence have been related with human oncogenicity in proteins such as H-Ras and occurrence of familial Amyotrophic lateral sclerosis (ALS) in proteins such as SOD1.

The availability of structure data on disease-involved proteins opens up an interesting direction for experimentally-assisted or guided protein modeling research. Important questions remain unanswered that protein modeling can address. While wet-laboratory investigations can catch healthy and diseased-versions of a protein in different stable structures under physiological conditions, such investigations cannot provide information on the underlying energy landscape of the protein under investigation. While structures stable enough to be caught in the wet laboratory are associated with minima in the energy landscape, the location of such minima is unknown. In particular, the location of the global optimum is not known, as experimental techniques are more limited than computation in their ability to probe the structure space. Extracting information on the energy landscape and the location of the global optimum is important for two reasons. First, such information can help understand which of the experimentally-determined functional structures reside at the global optimum and are thus most stable. Second, the discovery of the global optimum can point to functional regions of the energy landscape that have yet to be probed in the wet laboratory. In both cases, precious insight

is obtained into the biological function of disease-involved proteins.

There is a growing body of protein modeling research in the evolutionary computation (EC) community. The main focus of evolutionary algorithms (EAs) has been on the *de novo* structure prediction problem (PSP), where no information is assumed to be available on the functional structure(s) of a given protein amino-acid sequence. These algorithms are limited to small proteins ($\leq$ 70 amino acids long) in their applicability [7, 20, 8, 13, 32, 6, 3, 14, 10] and are increasingly shown to be outperformed by domain-specific algorithms originating in the computational biology community with respect to their ability to discover functional structures [5, 31, 16, 34, 30, 23].

While progress has been made recently [24, 25, 26], EAs for PSP are limited in their utility for protein modeling research. Their assumption that only sequence information and no structural information is available on a protein is invalidated on proteins that play a central role in human diseases. Such proteins are also typically longer than the cases that EAs for PSP can currently handle. While current EAs can typically handle short protein chains $\leq$ 70 amino acids long, many interesting proteins have lengths of at least 100 amino acids. The variable space of these proteins is too high-dimensional for exploration. Exploiting any available information on where the relevant regions of this space are is beneficial, but the current challenge is how to do so. Thus, the challenge for EAs in this new subdomain in protein modeling research is how to make use of the functional structures found for a protein in its healthy and diseased forms in the wet laboratory.

The subfield of algorithms originating in the computational biology community for guided exploration of a protein's fitness landscape is also sparse. Some local search algorithms that are initiated from a given structure and explore the space in the vicinity of the initial structure exist [9, 33, 29]. Running times of these algorithms vary from a few hours to a few days on one CPU. In principle, these algorithms can be run in a restart mode, initiated from different structures, to explore the fitness landscape. However, their computational demands would also increase in this new setting, and their ability to populate regions of the space with no experimentally-available structures is questionable, given that they fundamentally control computational cost by limiting their exploration in a neighborhood around the initial structure.

For these reasons, in this paper we focus on how to make structure-guided exploration for protein modeling amenable to evolution strategies. In essence, one is provided with a discrete setting, where domain experts (wet-laboratory investigators) have revealed a set of fit individuals. One direction of research can elect to focus on the design of an effective representation and effective reproductive operators over the individuals.

In this paper, we propose a novel direction. We provide a transformation of the discrete setting into a continuous optimization one. In essence, the collection of experimentally-available functional structures for healthy and diseased versions of a protein are subjected to a dimensionality reduction technique. The technique reveals the underlying variable space, its axes, number of dimensions, shape of the space, and bounds. Once a continuous variable space is defined, evolution strategies for continuous stochastic optimization can then be readily employed. In particular, here we showcase the ability to apply the Covariance Matrix Adaptation Evolution Strategy (CMA-ES). Several adaptations have to be introduced to properly initialize the algorithm and accurately compute fitness. The reason we choose CMA-ES here is that CMA-ES has been shown effective for difficult non-linear non-convex black-box optimization problems in continuous domains with anywhere from 3 to 100 variables. CMA-ES has yielded competitive results for local and global optimization [18, 17, 2] and is now a well-established method with many different application domains [12, 27, 19].

The analysis in this paper focuses on understanding the performance of CMA-ES on this new problem. The analysis shows that CMA-ES is capable of rapidly exploring the structure space and converging to the global optimum on different test cases. In some test cases, the algorithm is able to reveal which functional structures caught in the wet-laboratory reside in the global optimum of the fitness landscape and are thus thermodynamically more stable. In others, the algorithm reveals new information on the location of the global optimum, thus suggesting further experimentation in the wet-laboratory to probe unknown functional regions of the structure space.

The contributions of this paper go beyond the specific application domain on structure-guided exploration for protein modeling research. More generally, what this paper introduces is a methodology for transforming a discrete problem into a continuous one. While here we show a proof-of-concept on a needed subfield in protein modeling research, the applicability of this methodology extends to to other problems in protein modeling, such as protein-ligand and protein-protein binding and modeling of protein dynamics in folding and more. Even more generally, ideas introduced here may allow making progress in addressing challenging discrete optimization problems regarding evolution of finite state machines, neural networks, and more. What would be typically thought of as discrete problems can in principle be converted to continuous optimization problems. An approach that captures the underlying structure of the variable space, which we realize here through a statistical analysis of examples provided from domain experts, may allow re-introducing powerful evolution strategies capable of handling challenging continuous optimization problems.

2. METHODS

We first describe the transformation from a discrete problem into a continuous optimization one in detail. Next we describe our adaptation of CMA-ES to make use of the proposed transformation.

2.1 Tranformation from a Discrete to Continuous Setting

The transformation from a discrete to a continuous setting makes use of a linear dimensionality reduction technique to expose the underlying variable space. In this particular application setting, we are provided with experimentally-available functional structures of a protein. These are extracted from the PDB in the form of PDB entries, and each PDB entry lists the x, y, z coordinates of all atoms that constitute the amino acids of a protein. Figure 1(a) draws the structure of a short protein chain of 21 amino acids to illustrate the atom composition of a protein. A protein consists of backbone or skeleton comprised of N, CA, C, and O

atoms. Each amino acid shares these atoms, and they repeat in a serial fashion. A bond connects the N of one amino acid to the O of the other amino acid, thus providing the amino-acid chain. The CA atom is the central carbon atom in each amino acid that connects to the side-chain group of atoms that are found in each amino acid. Effectively, side chains dangle off the backbone like beads in a necklace. Different structures can exist for the same protein chain, as the backbone is highly deformable. Some functional structures of two proteins that are used as test cases in this paper are shown in Figure 1. Only the backbone chain is shown in each of these functional structures, in blue, for ease of visibility.

The transformation we employ here extracts only the CA atoms of each structure obtained for a protein from the PDB. So, the x, y, z coordinates of the CA atoms of each of n functional structures collected for a protein from the PDB are deposited into a data matrix of n columns and $3m$ rows, where m is the number of CA atoms in the protein. The first column, corresponding arbitrarily to the first structure, is used as reference. All other columns are modified through a technique known as optimal fit [22] to remove differences due to rigid-body transformations. The resulting data matrix is then centered, removing the average x, y, and z coordinate of each CA atom from each entry, and then multiplied by $\frac{1}{\sqrt{n-1}}$. The matrix is then subjected to the lapack [1] *dgesvd* procedure, which provides a singular value decomposition of the matrix as a product $U \cdot \Sigma \cdot V^T$ (also known as a Principal Component Analysis – PCA [21]).

The columns of the U matrix, known as the principal components (PCs), are the new orthogonal axes over which the data has been projected. These axes are vectors of $3m$ entries, with entries $3 \cdot i$, $3 \cdot i + 1$, and $3 \cdot i + 2$ providing the x, y, z displacements for the CA of the i^{th} amino acid in the protein. The diagonal of the Σ matrix contains the singular values, which are the square roots of the eigenvalues corresponding to each of the PCs. The *dgesvd* routine provides the PCs in the order of largest to smallest singular value. Analysis of the corresponding eigenvalues, as described in section 3, allows selecting a subset of PCs to consider as variables to represent an individual.

2.2 Adaptation of CMA-ES

CMA-ES follows an iterative process, where individuals in a generation are sampled from a multivariate normal distribution in R^n and then fit individuals are selected to update the multivariate normal distribution and covariance matrix. Details on CMA-ES are available in [17].

We introduce two adaptations to CMA-ES. The first adaptation concerns the initialization of the multivariate normal distribution. Since the data matrix to which PCA is applied is centered, the distributions of projections of the experimentally-available structures onto the PCs have a mean of 0. Standard deviations for each of the normal distributions along each of the PCs/variables are set at twice the standard deviations of the projections of experimentally-available structures; the resulting distributions have been confirmed to cover all experimentally-available structures are covered.

The second adaptation concerns evaluating sampled individuals. The individuals in a CMA-ES generation are not actual protein structures. They are points in a multi-dimensional PC-space. However, we want to associate with each of them a fitness that is measured through a sophis-

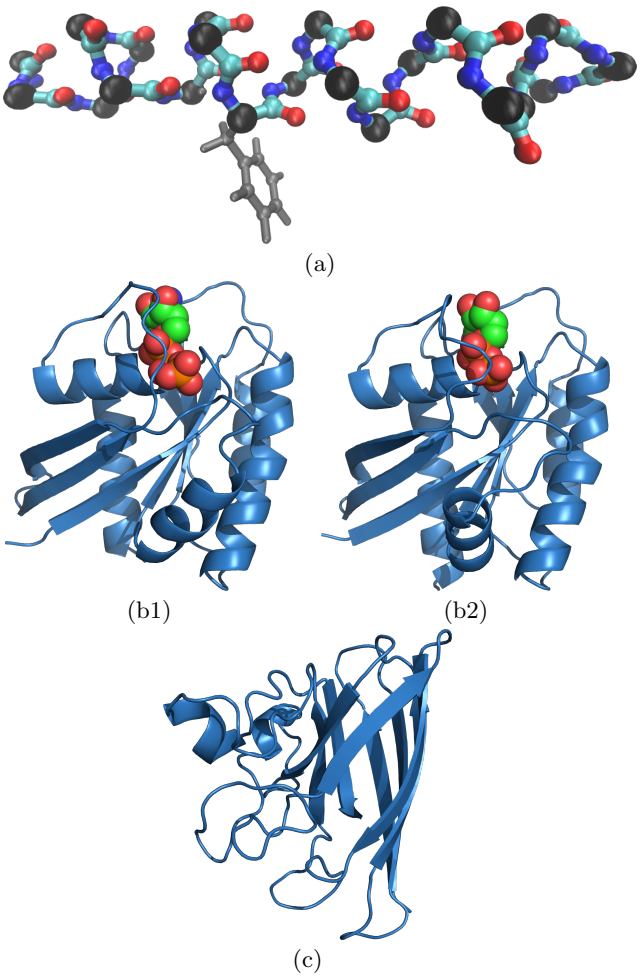

Figure 1: (a) illustrates a short protein chain of 21 amino acids. The main atoms, N, CA, C, and O, are shown for each amino acid. These atoms constitute the backbone chain of a protein and effectively provide the skeleton of a protein structure. Side-chain atoms for each amino acid dangle off the CA atom. The CA atom, which is the central atom in each amino acid, is drawn larger and in black. Side chain atoms are shown for only one amino acid here, drawn thinner and in gray, to illustrate their connetivity to the CA atom of an amino acid. (b)-(c) show two fit individuals (functional structures) for the healthy version of the H-Ras protein, corresponding to PDB entries (b1) 1qra and (b2) 4q21. Repectively, these represent the On (ligand-binding) and Off (non ligand-binding) states of H-Ras. (c) shows a fit individual for the healthy version of the SOD1 protein corresponding to PDB entry 1hl4. Structures are drawn in blue with Pymol [28]. For H-Ras, the ligand with which the protein interacts is drawn, as well, with ligand amino acids drawn as red and yellow spheres.

ticated all-atom energy function such as Rosetta *score12* (available in the open-source Rosetta structure prediction package [5]). The goal is to have the fitness landscape to correspond to the all-atom protein energy landscape, so that conclusions and observations made have biological significance. This requires recovering an actual protein structure with all-atom detail from an individual. The following technique is used. First, the x, y, z coordinates of all m CA atoms for an individual are recovered from an individual I represented through less than $3m$ variables through the operation $I \cdot U^T + A$, where the A vector contains the average x, y, z coordinates for each CA atom (this average vector was used to center the data matrix for PCA, as described above). Once the coordinates of CA atoms are recovered for an individual, a detailed all-atom structure can be then be computed. First, coordinates of the all backbone atoms can be filled in with statistics-based algorithms such as BBQ [15], which is one of the top backbone reconstruction protocols in protein modeling. Then, side-chain packing algorithms can be used to obtain optimal configurations of side-chain atoms of each amino acid given the backbone coordinates of the amino acids The *relax* protocol in the open-source Rosetta structure prediction package is used for this purpose. This protocol is easily integrated into CMA-ES, as it is open source.

2.3 Implementation Details

CMA-ES is run for 100 generations, and population size is set at 500. This takes $47-55$ hours on a CPU for CMA-ES. CMA-ES implemented in C/C++ and run on a 16 core red hat linux box with 3.2GhZ HT Xeon CPU and 8GB RAM. Analysis that looks into robustness of the algorithm uses 5 independent runs.

3. RESULTS

Test Cases: Performance is evaluated on 6 test cases. These include the WT and 3 variants of the H-Ras protein and the WT and one variant of the SOD1 protein. The variants of H-Ras are G12V, G12D, and Q61L; in G12V, for instance, the glycine amino acid at position 12 has been replaced with a valine amino acid in the variant. The variant considered for SOD1 is A4V, where the alanine amino acid at position 4 in the sequence has been replaced with a valine amino acid. These variants are linked to human diseases. In particular, the H-Ras variants are known to be oncogenic, whereas the SOD1 A4V variant is found in familial ALS.

Data Preparation: Experimentally-available structures of WT and variants of H-Ras and SOD1 are collected from the PDB. Variants with more than 3 amino-acid changes are discarded. PDB entries with missing internal amino acids are discarded, as well. This leaves 86 structures, each 166 amino acids long, for H-Ras and 186 structures, each 150 amino acids long, for SOD1. PCA is applied to each dataset, as described in section 2, to obtain the underlying variables for the WT and variants of H-Ras and SOD1.

Experimental Setup: The analysis focuses on the convergence of CMA-ES, as well as understanding where and how CMA-ES converges on the fitness landscape. The latter information is compared with the location of experimentally-available functional structures on the landscape. The performance of CMA-ES in terms of convergence

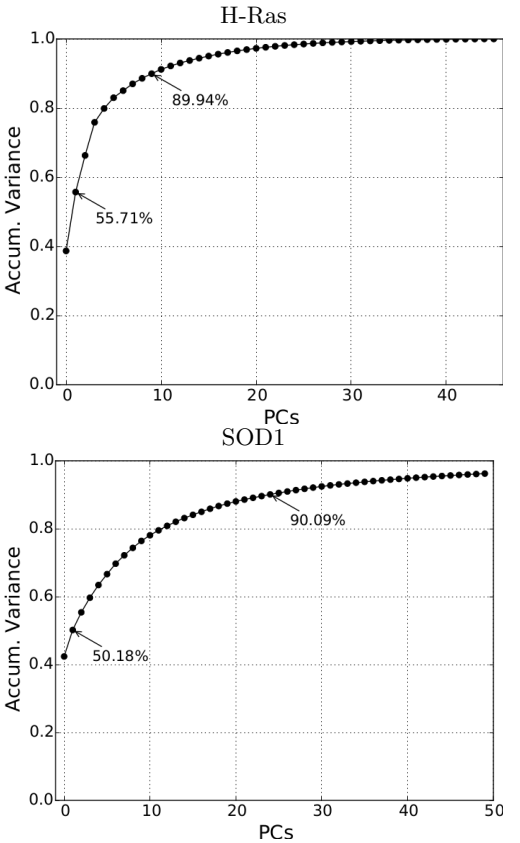

Figure 2: The accumulation of variance over the PCs is shown for H-Ras and SOD1. The accumulated variance at 2 PCs and 10 PCs is annotated.

3.1 PCA is Effective to Elucidate Underlying Variables

Figure 2 demonstrates that PCA is an effective dimensionality reduction tool for exposing the underlying variables for protein systems studied here. Figure 2 draws the accumulation of variance over the PCs as sorted by largest to smallest corresponding eigenvalue. Note that the eigenvalue obtained for a PC measures the variance of data projected on that PC. By sorting PCs from largest to smallest eigenvalue, one can estimate through $\frac{\sum_{i=1,\ldots,i=j} e_i}{\sum_{i=1,\ldots,N} e_i}$ the data variance that can be preserved if data are represented as j-dimensional vectors of entries that are projections over PC_1, ..., PC_j. In particular, Figure 2 shows that over 50% of the variance is preserved when employing only 2 PCs. In order to determine an optimal number of PCs that can be used as variables in the new PC-based representation, a variance threshold of 90% is used. This is reached at 10 PCs for H-Ras and 25 PCs for SOD1. That is, the number of variables in the representation used by the algorithms here is 10 for H-Ras WT and its variants and 25 for SOD1 WT and its variant.

3.2 CMA-ES Convergence

Two measurements are tracked across each of the 100 generations. The first concerns fitness improvement, measured as the difference in best fitness from generation i to generation $i+1$ (we note that low *score12* values are considered

high fitness). Figure 3 shows this analysis on two selected test cases, H-Ras WT and SOD1 WT.

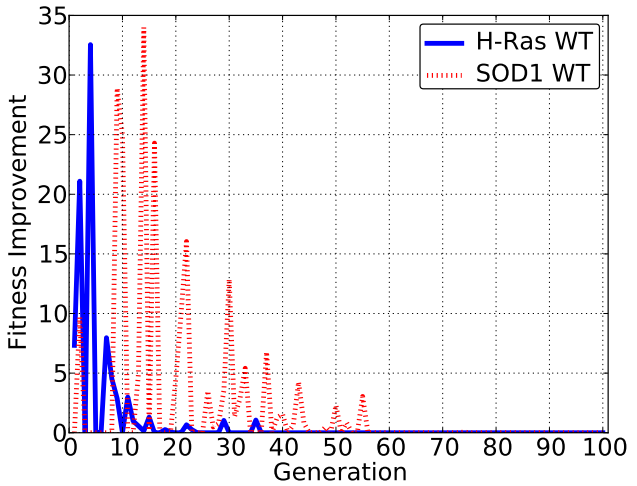

Figure 3: The fitness improvement from generation i to generation $i+1$ are tracked across the 100 generations.

The second measurement tracked across 100 generations concerns diversity, which is measured as the average Euclidean distance over all pairs of individuals in a population is tracked and plotted in Figure 4(a)-(b) for each of the 6 test cases. This analysis shows that CMA-ES converges quickly, within generation 20 in all H-Ras WT and variant test cases. For SOD1, where the variable space has twice the number of dimensions of that of H-Ras, convergence is reached at generation 40. These results are supported by the analysis on fitness improvement.

3.3 Detailed Analysis of CMA-ES

The convergence of CMA-ES is studied in greater detail. For two selected test cases, H-Ras WT and SOD1 WT, the two-dimensional gaussians over PC1 and PC2 are tracked over a few of the generations, starting with the initial generation and ending with the generation where CMA-ES is considered to have converged. Figure 5(a)-(b) draws these gaussians and individuals that constitute 2/3 of the population for each of the selected generations. The results show, as expected, narrowing of the distributions and movements of the means towards particular regions in the variable space.

To better illustrate where in the space CMA-ES converges, Figure 6(a)-(b) color-codes the gaussians by the average fitness of the corresponding population and draws only the locations of the experimentally-available structures in the PC1-PC2 projection of the variable space. These results show that the average fitness improves, as expected, and the generation where CMA-ES converges exposes which structures found in the wet laboratory constitute the global optimum of the fitness landscape. This discovery is important, as it reveals important information about landscape. In particular, for H-Ras WT, the experimentally-available structures, whose positions align well with the gaussian of the generation where CMA-ES converges, are considered to be the functional On state of H-Ras. In other words, CMA-ES has discovered that the On state of H-Ras WT is at the global optimum in the fitness landscape and is thus

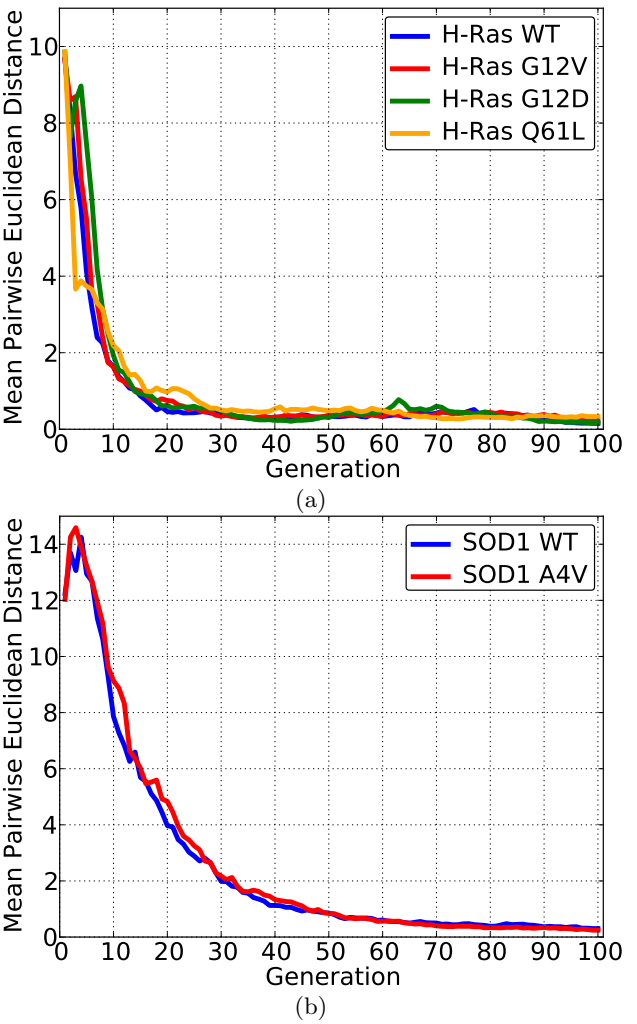

Figure 4: The average pairwise Euclidean distance is tracked across the 100 generations for H-Ras WT and its variants in (a) and SOD1 WT and its variant in (b).

more thermodynamically stable than other experimentally-determined structural states of H-Ras. On SOD1, the additional rendering of projections of the crystal structures shows that CMA-ES has converged to a distribution comprised of two distinct clusters of conformations. These largely cover the crystal structures known for SOD1 and further illustrates the ability of CMA-ES to capture these favorable regions of the SOD1 conformational space unpon convergence even when initiated from a wide gaussian. In particular, the results for SOD1 illustrate that there are two equally energetically-favorable regions of the conformational space, pointing to two possibly equally stable structural states for the SOD1 WT.

While not shown here, CMA-ES run on H-Ras variants converges to the same location as for the WT. This indicates that on both healthy and diseased versions of the H-Ras protein the global optimum is the same and contains the On state probed in the wet laboratory. This finding provides insight into what may be the reason for misfunction in the H-Ras variants. It is not that the variants lose

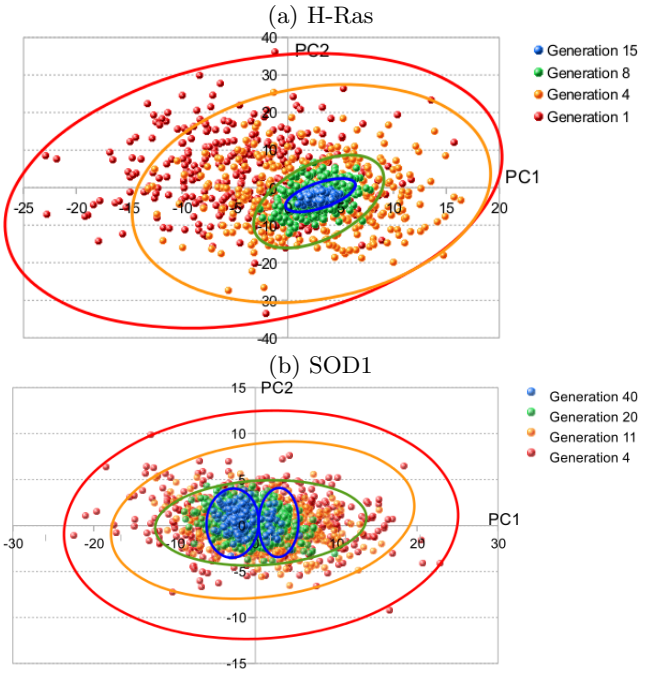

Figure 5: (a)-(b) illustrate the gaussians over PC1 and PC2 for H-Ras WT in (a) and SOD1 WT in (b) for 4 selected generations. The selected generations include the first and the one where CMA-ES is considered to have converged. Drawn points are 2/3 of the population in each selected generation.

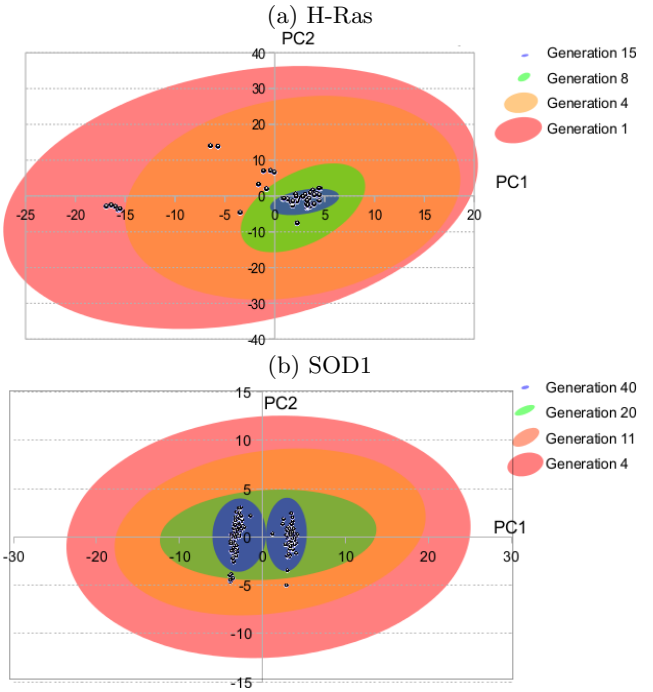

Figure 6: (a)-(b) Gaussians are now color-coded by the average fitness of corresponding population. A red-to-blue color scheme is used to indicate low-to-high fitness. The points drawn are projections of the X-ray structures on PC1 and PC2.

the ability to access the On state. Instead, possible barriers in the landscape may prevent them from switching from the Off to the On state at the same rate as the WT. This finding provides guidance for what further studies can investigate. On the other test cases, SOD1 WT and variant, CMA-ES reveals a global optimum that is different from the location of experimentally-available structures. This finding suggests CMA-ES proposes the existence of novel functional structures not yet captured in the wet laboratory.

In summary, this detailed analysis reveals that CMA-ES is effective and is able both to reproduce current knowledge, as in the case of H-Ras, and generate novel insight, as in the case of SOD1.

3.3.1 Restart Analysis of CMA-ES

Here the variability in when and where CMA-ES converges is studied in greater detail. In particular, two test cases are selected, H-Ras WT and SOD1 WT, and on each case CMA-ES is run 5 times. It is important to note that while the definition of the initial gaussian that initializes CMA-ES is the same on each independent run, the actual individuals sampled may be different, and thus the other gaussians of following populations can vary. Table 1 reports the mean and standard deviation of the mean of the gaussians over PC1 and PC2 in the generation where CMA-ES converges. The results in Table 1 demonstrate that CMA-ES is robust, as the standard deviations of the two tracked quantities are small.

4. CONCLUSION

This work has introduced a novel direction of research on evolution strategies for exploring protein energy landscapes. In particular, the exploitation of experimental data is proposed to guide EAs for protein modeling research. An interesting setup is proposed in this paper, where a discrete problem is transformed into a continuous optimization one. In particular, rather than pursue research on reproductive operators for three-dimensional protein structures, this work has instead focused on extracting information from a collection of experimentally-available structures on the underlying variables of the space, as well as the value ranges for these variables. Dimensionality reduction is used to extract such variables, which allows framing the problem as a continuous optimization one and employing both standard and novel evolutionary strategies for exploration.

One of the main observations in this paper is that this interesting exploitation of experimentally-available data allows the application and adaptation of powerful EAs such as CMA-ES. The results presented in this paper show that CMA-ES is effective, converges fast, and is able to reveal interesting information regarding the energy or fitness landscape of a protein under study. While the algorithm is essentially seeded only with information on what the initial shape and bounds of the variable space are, it manages to quickly find the global optimum of the landscape. Comparing the location of the optimum with the locations of the experimentally-available structures allows revealing which structures reside on the global optimum and which ones do not. In particular, for one of the proteins studied here, CMA-ES reveals that structures captured when the protein is in its On state are indeed of better fitness than other ones. This insight allows understanding that the On

Table 1: CMA-ES is run 5 times on two selected test cases, Ras WT and SOD1 WT. Where CMA-ES has reached convergence in terms of the means μ_{PC1} and μ_{PC2} of the normal distributions on PC1 and PC2 are tracked. Means and standard deviations of these two quantities over the 5 separate runs are reported here.

Test case	$\mu_{\mu_{PC1}}$ ($\sigma_{\mu_{PC1}}$)	$\mu_{\mu_{PC2}}$ ($\sigma_{\mu_{PC2}}$
H-Ras WT	3.32 (0.11)	-2.96 (0.14)
SOD1 WT	-2.45 (0.17)	6.39 (0.07)

state has higher thermodynamic stability than other structural states of H-Ras, thus revealing precious information on the function of H-Ras. For another test case, the algorithm reveals a global optimum that is further from the location of experimentally-available structures. This finding suggests further experimentation is needed in the wet laboratory to probe unknown functional regions of the structure space.

The work presented in this paper opens the way for several directions of further research. First, the interesting transformation here from a discrete problem into a continuous one may prompt more EC researchers to pursue novel evolution strategies for exploration and mapping of protein fitness landscapes. Other approaches beyond linear dimensionality reduction may be pursued.

The work presented here falls in the category of non black-box optimization algorithms that exploit existing knowledge about a problem. The exploitation of already-available information, as in the form of experimentally-available structures here, is an important direction to ground and guide optimization algorithms in a possibly very high-dimensional space. In particular, the combination of experiment and computation is an important direction that is bound to reveal more accurate biological insight onto the energy landscapes and thus function of protein molecules.

While this paper has focused on a protein modeling application, the setup introduced here to convert a discrete problem into a continuous optimization one may inspire progress in addressing challenging discrete optimization problems concerning evolution of finite state machines, neural networks, and more.

5. ACKNOWLEDGMENTS

This work is supported in part by NSF CCF No. 1421001, NSF IIS CAREER Award No. 1144106, and the Thomas F. and Kate Miller Jeffress Memorial Trust Award.

6. REFERENCES

[1] E. Anderson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammerling, J. Demmel, C. Bischof, and D. Sorensen. Lapack: A portable linear algebra library for high-performance computers. In *Proceedings of the 1990 ACM/IEEE Conference on Supercomputing*, Supercomputing '90, pages 2–11, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[2] A. Auger and N. Hansen. A restart cma evolution strategy with increasing population size. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 1769–1776. IEEE, 2005.

[3] D. Becerra, A. Sandoval, D. Restrepo-Montoya, and L. Nino. A parallel multi-objective ab initio approach for protein structure prediction. In *IEEE Intl Conf Bioinf and Biomed (BIBM)*, pages 137–141. IEEE, 2010.

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

[5] P. Bradley, K. M. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, Sep 2005.

[6] J. Calvo, J. Ortega, and M. Anguita. Comparison of parallel multi-objective approaches to protein structure prediction. In *J Supercomputing*, pages 253–260. CITIC UGR Univ Granada, Dept Comp Architecture & Comp Technol, Granada, Spain, 2011.

[7] C. Chira, D. Horvath, and D. Dumitrescu. An Evolutionary Model Based on Hill-Climbing Search Operators for Protein Structure Prediction. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 38–49, 2010.

[8] Cutello, V, G. Morelli, G. Nicosia, M. Pavone, and G. Scollo. On discrete models and immunological algorithms for protein structure prediction. *Natural Computing*, 10(1):91–102, 2011.

[9] B. L. de Groot, D. M. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins: Struct Funct Genet*, 29(2):240–251, 1997.

[10] R. Faccioli, I. N. da Silva, L. O. Bortot, and A. Delbem. A mono-objective evolutionary algorithm for Protein Structure Prediction in structural and energetic contexts. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 1–7. IEEE, 2012.

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

[12] C. Gagné, M. Beaulieu, J. NAND Parizeau, and S. Thibault. Human-competitive lens system design with evolution strategies. *Applied Soft Computing*, 8(4):1439–1452, 2008.

[13] M. Garza-Fabre, G. Toscano-Pulido, and E. Rodriguez-Tello. Locality-based multiobjectivization for the HP model of protein structure prediction. In *Intl Conf on Genet and Evol Comput (GECCO)*. 2012. ACM.

[14] M. M. Goldstein, E. E. Fredj, and R. B. R. Gerber. A new hybrid algorithm for finding the lowest minima of potential surfaces: approach and application to peptides. *J Comput Chem*, 32(9):1785–1800, July 2011.

[15] D. Gront, S. Kmiecik, and A. Kolinski. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from

alpha carbon coordinates. *J Comput Chem*, 28(29):1593–1597, 2007.

[16] J. Handl, J. Knowles, R. Vernon, D. Baker, and S. C. Lovell. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. *Proteins: Struct Funct Bioinf*, 80(2):490–504, 2012.

[17] N. Hansen and S. Kern. Evaluating the cma evolution strategy on multimodal test functions. In *Intl Conf on Parallel Problem Solving from Nature (PPSN)*, pages 282–291, 2004.

[18] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[19] O. Ibañez, L. Ballerini, O. Cordón, S. Damas, and J. Santamaria. An experimental study on the applicability of evolutionary algorithms to craniofacial superimposition in forensic identification. *Information Sciences*, 179(3):3998–4028, 2009.

[20] M. K. Islam, M. Chetty, and M. Murshed. Novel Local Improvement Techniques in Clustered Memetic Algorithm for Protein Structure Prediction. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 1003–1011. IEEE, Apr. 2011.

[21] D. G. Luenberger. *Introduction to Linear and Nonlinear Programming.* Addison-Wesley, 1973.

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

[23] K. Molloy, S. Saleh, and A. Shehu. Probabilistic search and energy guidance for biased decoy sampling in ab-initio protein structure prediction. *IEEE/ACM Trans Bioinf and Comp Biol*, 10(5):1162–1175, 2013.

[24] B. Olson, K. A. D. Jong, and A. Shehu. Off-lattice protein structure prediction with homologous crossover. In *Intl Conf on Genet and Evol Comput (GECCO)*, pages 287–294, New York, NY, 2013. ACM.

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

[26] J. Santos, P. Villot, and M. Dieguez. Emergent protein folding modeled with evolved neural cellular automata using the 3d hp model. *J of Comp Biol*, 21(11):823–845, 2014.

[27] J. Schaub, K. Mauch, and M. Reuss. Metabolic flux analysis in escherichia coli by integrating isotopic dynamic and isotopic stationary 13c labeling data. *Biotechnol Bioeng*, 99(5):1170–1185, 2008.

[28] L. Schrödinger. The PyMOL molecular graphics system, version 1.3r1, August 2010.

[29] A. Shehu, C. Clementi, and L. E. Kavraki. Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. *Proteins: Struct Funct Bioinf*, 65(1):164–179, 2006.

[30] A. Shehu and B. Olson. Guiding the search for native-like protein conformations with an ab-initio tree-based exploration. *Intl J Robot Res*, 29(8):1106–1127, 2010.

[31] A. Shmygelska and M. Levitt. Generalized ensemble methods for de novo structure prediction. *Proc Natl Acad Sci USA*, 106(5):94305–95126, 2009.

[32] A.-A. Tantar, N. Melab, and E.-G. Talbi. A grid-based genetic algorithm combined with an adaptive simulated annealing for protein structure prediction. *Soft Computing*, 12(12):1185–1198, 2008.

[33] S. A. Wells. Geometric simulation of flexible motion in proteins. *Methods Mol Biol*, 1084:173–192, 2014.

[34] D. Xu and Y. Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Struct Funct Bioinf*, 80(7):1715–1735, 2012.