An Evolutionary-inspired Probabilistic Search Algorithm to Structurally Characterize the Native State of a Novel Protein Sequence

Shehu lab www.cs.gmu.edu/~ashehu
{ssaleh2, amarda}@gmu.edu

Sameh Saleh[1] and Amarda Shehu[1,2,3]

[1]Department of Computer Science, [2]Department of Bioinf. & Comp Biol., [3]Department of Bioengineering

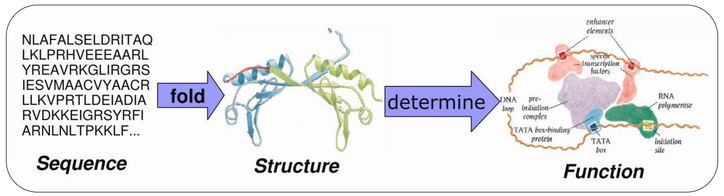
Abstract

Predicting the functionally-relevant three-dimensional structure of a protein from knowledge of its amino-acid sequence is a long-standing challenge of computational biology. Computationally, the task involves searching a high-dimensional space for low-energy conformations. Doing so naively is computationally intensive, but addressing the problem with efficient algorithms promises to broaden our understanding of the structure-function relationship in protein molecules.

We propose a probabilistic search algorithm that employs evolutionary search strategies and biophysics-inspired modeling of proteins. Specifically, protein conformations are treated as individuals to be evolved. Evolution is guided through a state-of-the-art physics-based energy function that measures the fitness of each individual and steers the surviving population towards fitter individuals. Fragment-based assembly, an established technique for protein structure prediction, is used to modify a parent and obtain a child conformation. Parent and children conformations are pooled together to be truncated back to the initial population size based on their energy fitness. The result is that evolved individuals are physically-realistic conformations that progressively occupy lower-energy regions in the underlying energy surface.

We demonstrate that the algorithm is efficient and yields native-like protein conformations for different sequences. Our results show that the search converges to low energy regions that lie near the known native structure. These results are promising and suggest that further research on the integration of evolutionary algorithms and memetic-based approaches will improve traversal of the protein energy surface and enhance sampling of native-like conformations.

Introduction



"[...] the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment." Anfinsen, C. B. Science 181, 1973

Experimental techniques that are devoted to resolving the native structure of a protein sequence cannot keep pace with the exponential explosion in the number of new protein sequences deposited to databases.

Determining the biologically-active structure of a protein sequence in-silico remains a central challenge in computational structural biology.

Exploring the protein conformational space in search of conformations that populate the protein native state is an NP-hard problem.

The protein conformational space is vast, continuous, and high-dimensional.

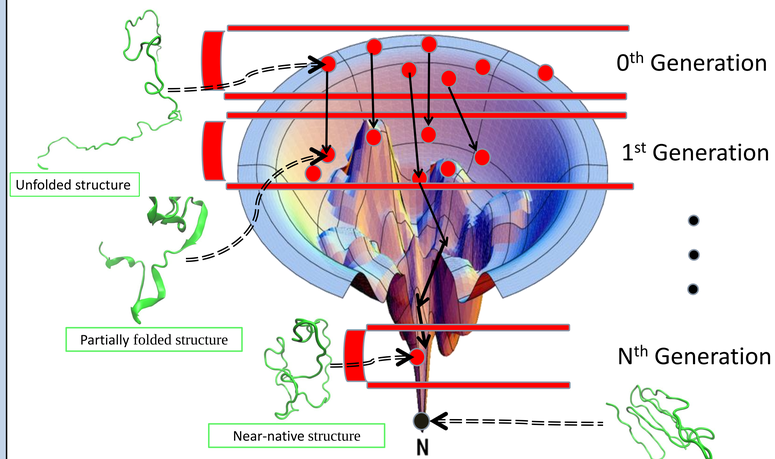
The protein energy surface is funnel-like, but rich in local minima of varying sizes.

The protein native state is associated with the basin of the protein energy surface.

We propose to revisit evolutionary search strategies and combine them with the state-of-the-art fragment assembly in computational biophysics in order to effective explore the protein conformational space and obtain lowest-energy conformations associated with the native state

Acknowledgments This work was partially supported by the Undergraduate Student Research Program to S.S., an NSF CCF Grant No.1144106 to A.S., and an NSF Career Award Grant No. 1016995 to A.S.

Methods



Each conformation represents a separate individual in the population.

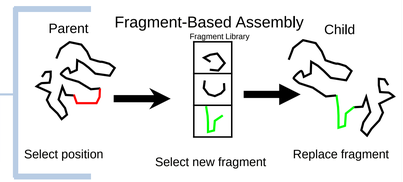
The initial population is instantiated from *n* number copies of the fully extended conformation of the protein, whereby each copy is altered with *n-2* fragment replacements to diversify the population. *N=100* was used.

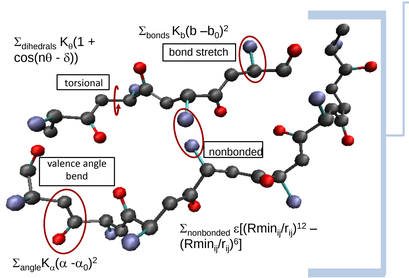
For each generation, from these *n* members of the population, *m* conformations are chosen to be copied and act as the children of the population. *M=100* and *M=75* were both tested. *M = 100* were used to arrive at the results.

Such a selection was made using fitness-proportional selection. The stochastic implementation precludes that the conformations with the higher fitness are more likely to have children.

These children copies of the parent are modified through asexual reproduction to produce a child conformation. A fragment configuration in the parent is replaced with a configuration selected from a pre-built fragment configuration library to produce the child.

The process is facilitated by fragment-based assembly, a state-of-the-art technique in computational biophysics that allows effectively obtaining realistic conformations.





The fitness of each individual is measured through a state-of-the-art energy function.

The children are integrated into the original parent population, which is then sorted in descending order based on the energy fitness function

Such a population is truncated to the original population size of *n* individuals and the process is repeated for *k* number of generations. *K=500* was used.

Unlike existing work that applies evolutionary search algorithms on toy protein models, the proposed method incorporate a state-of-the art energy function and latest biophysics techniques such as fragment-based assembly.

Results

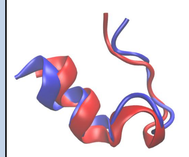
Four proteins were used to test our evolutionary search structure prediction method. The proteins used represent a variety of structure types with different folds and different length. The correctness of the algorithm is evaluated by calculating the root mean square deviation (RMSD) from the native structure. Since the evolutionary search uses the energy as the fitness of each conformation, the minimum energy helps assess the convergence of the evolutionary model as the number of generations approach infinity. Finally, the Q-values show the percentage of amino acid that are in correct placement as compared with the native.

In the graphs below, the relationship between decreasing RMSD and decreasing energy is shown as a scatterplot. Finally, the conformation of the lowest RMSD for each protein is visualized in **blue** and is superimposed on the native conformation which is displayed in **red**.

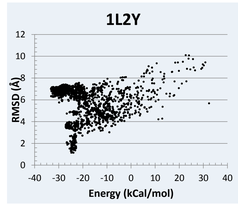
Systems of Study

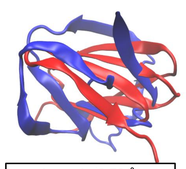
Name	Trp-Cage	Influenza A Virus NS1	Peptostreptococcus L	Trp RNA-binding
PDB ID	1l2y	1ail	1hz6A	1wapA
Fold	α	α	α/β	β
Size	20	70	67	68

PDB ID	1l2y	1ail	1hz6A	1wapA
Min RMSD (Å)	1.15	4.41	6.71	9.78
Min Energy (kCal/mol)	-33.26	-134.58	-105.59	-72.55
Max Q-score	74%	62%	47%	22%

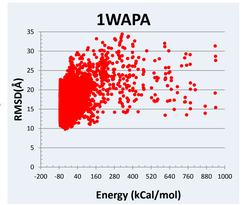


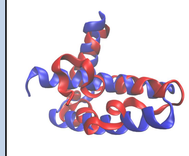
1l2y (1.15Å)



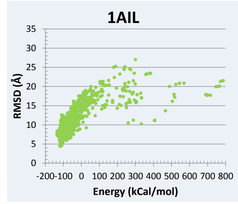


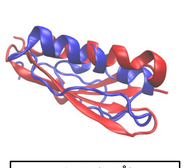
1-wapA (9.78 Å)



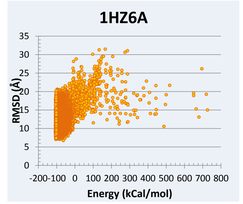


1ail (4.41 Å)





1hz6a (8.6Å)



Conclusions

Based on the analysis of the results gathered, the evolutionary search algorithm effectively predicts high-quality near-native conformations in two of the four proteins, *1ail* and *1l2y*. In the other two proteins, *1wapA* and *1hz6A*, the algorithm was able to iteratively approach the native structure over generations. By using a state-of-the-art fragment-based library, the fragment replacements are configured efficiently achieving structural diversity in the proteins, while maintaining physically realistic dimensions.

Future work will focus on achieving an optimization scheme that arrives at progressively decreasing local minima without quickly converging to the lowest energy. Possibilities of research incorporate a black-box approach that uses varying degrees of a Metropolis Monte Carlo optimization and a hill descent optimization.

Implementation details: The presented results were obtained by running the method on a 2.66GHz Opteron processor with 8GB of memory for 2 to 5 hours depending on protein length. The method was implemented in C/C++.