A new Distance Function for Protein Structures for the Decoy Selection Problem in *De-novo* Structure Prediction

Songyue Huang[1] and Amarda Shehu[1, 2, 3]

[1]Dept. of Computer Science, [2]Dept. of Bioengineering, [3]School of Systems Biology, George Mason University

[shuang7, amarda]@gmu.edu

ABSTRACT

We now have algorithms spewing out hundreds of thousands of three-dimensional structures computed for a given protein sequence in a few days on one CPU. These structures are known as decoys. The challenge is how to determine which of them is the native structure. This is a crucial part of the de novo structure prediction problem, where for a given protein sequence, we want to know what its native structure is as the first step to understanding anything about the function of that protein.

In this project I explore the utility of a novel distance function for comparison of protein structures in the process of decoy selection. The function allows highly efficient comparisons of structures, as it does not rely on structure superimpositions but instead on comparison of discretized mappings of protein backbone angles. Here I report on several characteristics of this function as well as its utility for decoy selection.

INTRODUCTION

My project addresses the so-called decoy selection problem in protein structure prediction. We now have algorithms spewing out hundreds of thousands of three-dimensional structures computed for a given protein sequence in a few days on one CPU. These structures are known as decoys. The challenge is how to determine which of them is the native structure. This is a crucial part of the de novo structure prediction problem, where for a given protein sequence, we want to know what its native structure is as the first step to understanding anything about the function of that protein [1].

This problem is open. It turns out that comparing by potential energies produces false positives [2], as functions that measure energy are noisy; in other words, the native structure may not have the lowest energy. Current algorithms ignore energy and instead groups structures together by structural similarity. They cluster structures, and typically the cluster with the most members is offered as containing the native structure [3].

There are two open questions in decoy selection: 1) how does one measure structural similarity? The predominant way is to compare by least root-mean-squared-deviation (lrmsd), but this is slow and requires optimally aligning two structures. With hundreds of thousands of structures, this comparison measure is very time-demanding. It is also not very descriptive [4]. Other, possibly cheaper and more accurate distance functions ought to be considered. 2) Once structures are clustered, which structure in which cluster should be predicted as representative of the native structure? A cluster may not be homogeneous; structures in it may be diverse, depending on the lrmsd threshold used for clustering. Should this process not include energy at all? A more rigorous scheme ought to be considered.

I explore the utility of a novel distance function for comparison of protein structures in the process of decoy selection. The function is inspired from work in [4], which transforms a protein's three-dimensional structure into a one-dimensional string by mapping each residue onto its corresponding basin.

METHODOLOGY

A three-dimensional protein structure is represented by dihedral angles. Angles are then mapped into letter codes, based on regions in the Ramachandran map observed to be populated by known native protein structures. In this way, a three-dimensional structure is represented as a linear string of 2n characters, each character denoting a bin/region in the Ramachandran map (2 angles per amino acid). The designation of regions in the Ramachandran map to allow this 1d mapping is shown below, as pr

Table I. *Basin Definitions*[a]

A: −62; −42	R: −68; −18	V: −93; 2	B: −120; 135	P: −64; 139
G: −93; 95	D: −134; 70	L: 51; 42	U: 82; −3	T: 55; −129
Y (Gly only) 77; −171				

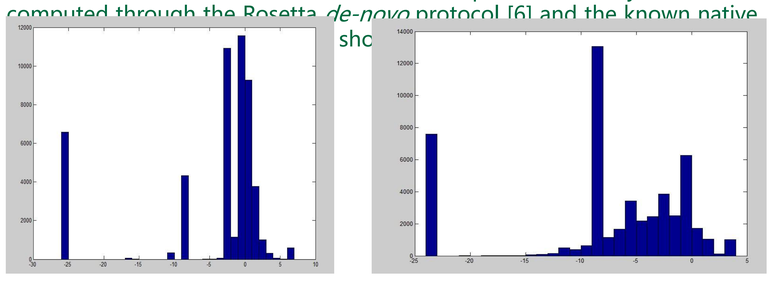
[a] Basin labels and ϕ,ψ-centroids in degrees. Basins were defined as $\phi\psi$-regions that extend $\pm 10^2$ beyond the centroid in each direction.

Any distance function can be defined to operate on such 1d representations of protein structures. The function designed here sums in-orde ne basin subs

Table VI. *The Basin Substitution Matrix Used in This Study*

	A	B	D	G	L	P	R	T	U	V	Y
A	0										
B	−3	1									
D	−1	0	1								
G	−2	1	1	1							
L	−2	1	1	1	3						
P	−2	1	1	1	1	1					
R	0	0	0	0	0	0	1				
T	−2	1	1	1	1	1	1	4			
U	−2	1	1	1	2	2	0	1	3		
V	−1	0	1	0	1	1	0	1	1	2	
Y	−2	1	1	1	2	1	0	1	2	1	1

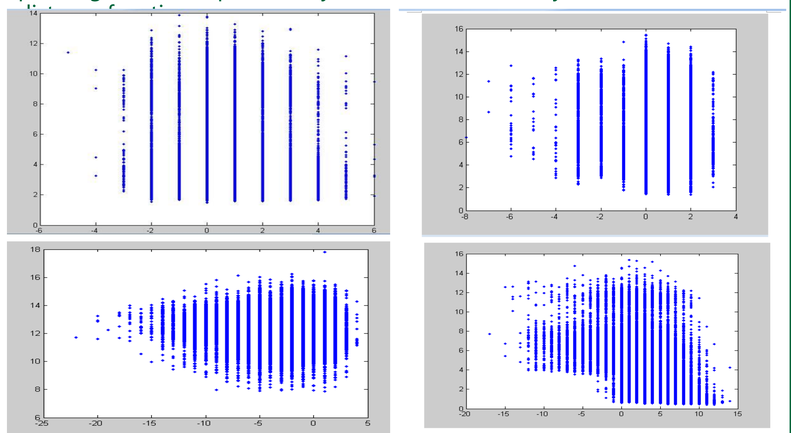
The distributions of distances between protein tertiary structures computed through the Rosetta *de-novo* protocol [6] and the known native sho



Two questions arise: (i) what is the relationship between this function and the well-known RMSD function [5]? (ii) how would this function perform in the context of clustering-based decoy selection?

RESULTS

The RMSD from the native structure is drawn on the Y axis, and the novel distance is shown on the X-axis. No correlation is observed, pointing to the complementary information offered by the novel



In the absence of correlation, no threshold can be determined for the novel distance function to establish similarity or dissimilarity of two structures. Hence, leader-based clustering algorithms are not amenable. Instead, we pursue a different algorithm, which represents a structure through its distances from a few selected landmark structures in a decoy set. Here are some results of this clustering algorithm. Weka is used [7].

```
=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation:     anglerelation
Instances:    50012
Attributes:   114
[list of attributes omitted]
Test mode:evaluate on training data

=== Model and evaluation on training set ===

EM
==

Number of clusters selected by cross validation: 1

                Cluster
Attribute            0
                   (1)
========================
FAI1
```

```
Relation:     distance
Instances:    50012
Attributes:   1
              distance
Test mode:evaluate on training data

=== Model and evaluation on training set ===

EM
==

Number of clusters selected by cross validation: 4

                Cluster
Attribute            0          1          2          3
                   (0.46)     (0.12)     (0.37)     (0.05)
==========================================================
distance
  mean             0.823     2.2707     0.4168         -2
  std. dev.       0.8502     1.0766     0.7252     0.0024


Time taken to build model (full training data) : 99.71 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      26195 ( 52%)
1       2823 (  6%)
2      18370 ( 37%)
3       2624 (  5%)

Log likelihood: -1.26675
```

ACKNOWLEDGEMENTS

This research is supported by a URSP Spring 2015 award to SH. Testing and experiments were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA (URL: http://orc.gmu.edu).

REFERENCES

[1] Alberts B, Johnson A, Lewis J, et al. Analyzing Protein Structure and Function. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002.
[2] Jiang Z, Qianqian Z, Yunyu S, et al. How well can we predict native contacts in proteins based on decoy structures and their energies? Proteins 2003, 52:598-608.
[3] Zhang Y and Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004, 25(6):865-871.
[4] Chellapa G D and Rose G D. Reducing the dimensionality of the protein-folding search problem. Protein Sci 2012, 21(8):1231-1240.
[5] Maiorov V N and Crippen G M. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol 1994, 235(2):625-634.
[6] Bradley P, Malmström L, Qian B. et al. Free modeling with Rosetta in CASP6. Proteins: Structure, Function, and Bioinformatics 2005, 61 (Suppl1): 128–134.
[7] Hall M, Frank E, Holmes G, et al. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009 11(1):10-18.