

# A Novel Computational Framework for Structural Classification of Proteins using Local Geometric Parameter Matching

Sumeet Dua, Ph.D.  
Data Mining Research Laboratory,  
Louisiana Tech University  
E-mail: sdua@latech.edu

Naveen Kandiraju  
Data Mining Research Laboratory,  
Louisiana Tech University  
E-mail: nka007@latech.edu

## Abstract

The objective of this study was to develop a novel and fast computational framework for classification of proteins using a series of secondary structure geometric parameter represented by an unexplored dihedral angle of a protein sequence. A dihedral angle is calculated between two planes represented by atom-tuplets  $[N(i), C(i), N(i+1)]$  and  $[C(i), N(i+1), C(i+1)]$ , of adjacent ( $i$  and  $i+1$ ) amino acids of a protein structure. The comparison of two such series of dihedral angles, each representing a different protein structure, is based on subsequence matching which not only gives the extent of match but also provides with the approximate demographic information of the match which then is used in classification of proteins. The technique is tested over 25 proteins belonging to 5 different families randomly selected from Alpha, Beta, Alpha and Beta (alpha/beta) and Multi-domain proteins (alpha and beta) classes. The classification rate is achieved with an accuracy of 88%.

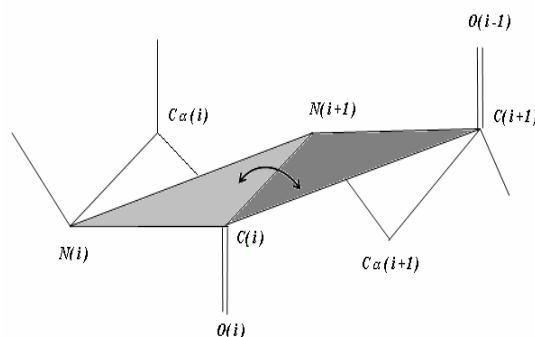
## 1. Introduction

The drastic increase observed in the size of protein structure databases and the rapid advances in Bioinformatics efforts in the post-genomic era, has necessitated the development of rapid protein structural comparison techniques. This need stems from the various advantages associated with identification of protein structural similarities, most significantly in aiding the prediction of an unknown protein function; a well studied problem that has caught the eye of many researchers in recent years. There are many structural comparison techniques in literature that use the geometric information of the amino acids [1], [2], [3]. We propose a novel computational framework that uses an unexplored dihedral angle as the geometric parameter in determining the similarity of two protein structures,

which is then employed to classify proteins in their respective families based upon the extent of similarity.

## 2. Methodology

Figure 1, describes the proposed NCNC dihedral angle made up of the Nitrogen and Carbon atoms of adjacent amino acids of a protein structure. Each protein structure can be represented as a series of above described dihedral angles, abstracting the problem of comparing two proteins structures to one of the comparison of two linear distributions.



**Figure 1. The proposed NCNC dihedral angle comprising of two N and two C atoms of adjacent amino acids in a protein sequence.**

The comparison of the thus obtained distributions is accomplished by a similarity-search mechanism that segments the distributions into overlapping subsequences followed by the discovery of the areas of relatively stationary harmonic behavior (called trails). These trails are then structured in a unique translational and scale invariant indexing schema to enable searching and reporting of local alignments. Table 1 gives the family wise distribution of the proteins the proposed method is tested on. Figure 2 and Figure 3 give the plots of dihedral distributions of 1ATT and 1AZX of Serpins family, and 1J9G and 1J8Q of Flavodoxin-related family respectively.

### 3. Results

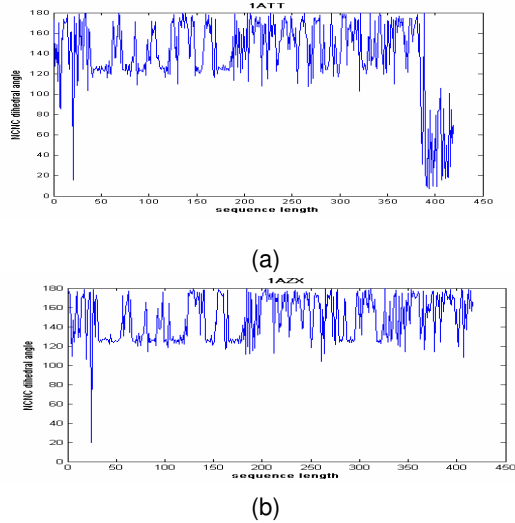


Figure 2. NCNC dihedral plots for proteins (a) 1ATT and (b) 1AZX of Serpins family.

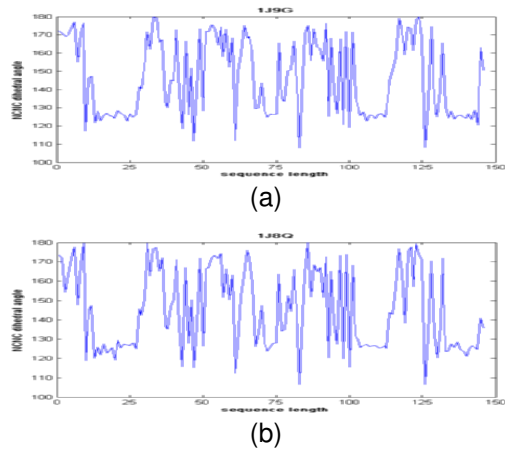


Figure 3. NCNC dihedral plots for proteins (a) 1J9G and (b) 1J8Q of the Flavodoxin-related family.

Degree of local similarity is calculated using our translational and scale invariant indexing schema, and the results represented with approximate positional information of the similitude match. Figure 4 demonstrates the regional information of local structural similarity matches between proteins 1ATT and 1AZX of the Serpins family and for proteins 1J9G and 1J8Q of the Flavodoxin-related family.

The experimental results demonstrate a cumulative true positive rate of 88% in classification, with a very low degree of false negatives.

Table 1. Proteins and their respective families

Family Name	Proteins selected from the family (PDB ids)
Serpins	1ATT, 1AZX, 2ACH, 2ANT, 7API
Flavodoxin-related	1C7E, IC7F, 1J9G, 1J8Q, 1J9E
Monodomain – cytochrome c	1B7V, 1K3G, IK3H, 1KIB, 1N9C
V set domains (antibody variable domain like)	1BJM, 2FB4, 2IG2, 3BJL, 4BJL
Subtilases	1SEL, 1OYV, 1SCJ, 1CSE, 2SEC

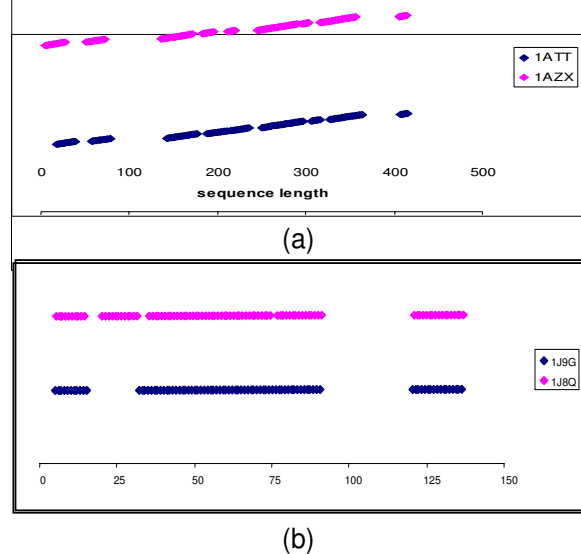


Figure 4. Approximate positional information of similarity between (a) 1ATT and 1AZX (b) 1J9G and 1J8Q

### 4. References

- [1] Jacek Leluk, Leszek Konieczny, and Irena Roterman, "Search for Structural similarity in proteins", Vol. 19, No. 1, *Bioinformatics* 2003, pp. 117-124.
- [2] S. Laiter, D. L. Hoffman, R. K. Singh, I. L. Vasiman, and A. Tropsha, "Pseudotorsional occo backbone angle as a single descriptor of protein secondary structure", *Protein Science*, 4(8), 1995, pp. 1633-1643.
- [3] C. Guda, E.D. Scheeff, P.E. Bourne, and I.N. Shindyalow, "A new algorithm for the alignment of multiple protein structures using Monte Carlo optimization", *Proceedings of Pacific Symposia on Biocomputing*. 2001.