

Geometric Invariant Theory applied to Protein Structure Classification

Pramod P. Wangikar¹, Ashish V. Tendulkar^{2,3}, Milind A. Sohoni³

¹Department of Chemical Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai 400 076 INDIA
Email: pramodw@iitb.ac.in Fax:+91-22-2572 6895

²Kanwal Rekhi School of Information Technology, Indian Institute of Technology, Bombay, Powai, Mumbai 400 076

³Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai 400 076

Keywords: Recurring patterns, graph theory, data mining, active site

Introduction: Proteins are conventionally classified based on similarity either of overall sequence or of geometric structure. Several methods are known for optimal pair-wise alignment of protein structures [8,9]. Much of the structure analysis hinges on the hypothesis that nearly all proteins have structural similarities and, in many cases share a common evolutionary origin [1]. Furthermore, substructures of small number of amino acids are known to be conserved across several proteins. The conserved amino acids are deemed to be important for function [2]. Methods have been developed to search for user-defined conformations, which are typically useful in searching for known active sites [2,3]. Thus, our first objective was to establish relationships between proteins based on recurring structural patterns, which could potentially be functional site patterns. In addition, we probe into the conserved structural patterns at the protein-protein interface.

Algorithm: We use recurring structural patterns of a small number of amino acids as keys to classification. We report a novel approach to the detection of recurring structural patterns via geometric invariants. These invariants help in deciding the super-imposability of candidate patterns without the computationally expensive step of actually constructing and verifying the superimposing transformation. A geometric invariant is a quantity, which is unchanged under a group of geometric transformations, in this case, the group of translations and rotations in 3-dimensional space. The simplest invariant is *content*, the labels of amino acids in the pattern. Examples of continuous invariants associated with a geometrical structure of amino acids, or *pattern*, are volumes, areas, lengths, etc. For our group of transformations, it has been shown that invariants suffice to decide superimposability of two structures [4,5]. Thus, we coordinatize a pattern by its evaluations on a fixed suite of N invariants. Using earlier work [3], we examine patterns from 3,400 non-redundant protein structures and the collection of patterns is subjected to the usual tools of data-mining. A cluster is a small region in this N-space, which has a large number of pattern-vectors, and thus is visited by a large number of proteins. Such a cluster corresponds to a recurring pattern.

Results: When this classification scheme is compared with the well accepted SCOP classification [6], we find that several of our clusters are visited by members of a single SCOP superfamily with the detected recurring pattern acting as a signature of that SCOP superfamily and not present in any other superfamily (Fig. 1 shows an example). In addition, we find several clusters visited by proteins from two to three SCOP superfamilies. This points to novel relationships between superfamilies and potentially common

functional sites / mechanisms of action across superfamilies. We find many of the pattern forming amino acid residues to be involved in functional / mechanistic role.

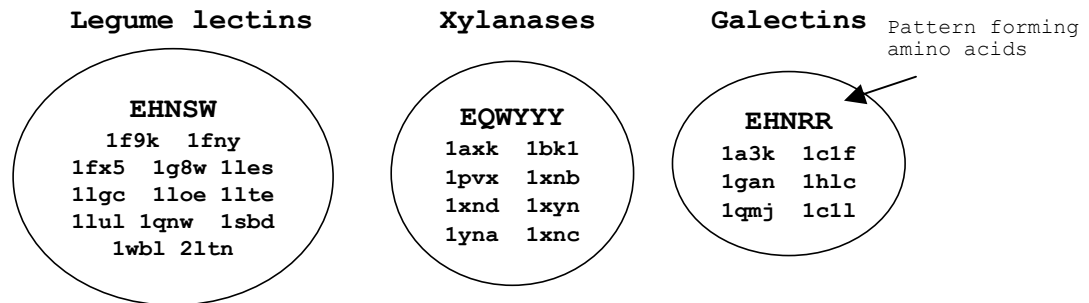


Fig. 1: Representative clusters: Functional site patterns of Concanavalin A like lectins / glucanases superfamily: Members of this superfamily get separated into three mutually exclusive clusters which correlate with the known functional categories. The pattern forming amino acids (one letter codes) of the structural pattern are shown with the pdb structures, which contain the recurring pattern

Conclusion: We apply our algorithm to classify the 3,400 non-redundant protein structures into 1,200 non-hierarchical clusters. This results in several signature patterns, which seem to decide the membership of a protein in a family. To exemplify, the important patterns include a “glutamate double bridge” of superoxide dismutase, the charge-charge interaction at the interface of the serine protease-inhibitor complex, and functional sites of lectins, xylanases, protein kinases, beta-lactamase, acid protease, transglycosidase, zinc protease, metalloproteases, cytochromes, cysteine protease, TIM beta/alpha barrel and several other protein families. The results provide a vast resource for the biologists for experimental validation of the proposed functional sites, and for the design of synthetic enzymes, inhibitors and drugs. This is the first report on application of the century old “Geometric Invariant Theory” for analysis of protein structures and opens new opportunities for geometric invariant-based extraction of relationships based on either local or global structural similarities.

References:

- [1] Murzin, A.G. (1992). Familiar strangers. *Nature*, **360**, 635.
- [2] Wallace, A.C., Laskowski, R.A. & Thornton, J.M. (1996). Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001-1013.
- [3] Wangikar, P.P., Tendulkar, A.V., Ramya, S., Mali, D.N. and Sarawagi, S. (2003). Functional Sites in Protein Families Uncovered via an Objective and Automated Graph Theoretic Approach. *J. Mol. Biol.* **326**, 955-978.
- [4] Hilbert, D. (1893) Uber die vollen Invariantensysteme. *Math. Ann.* **42**, 313-373
- [5] Weyl, H. (1939) *The Classical Groups, their Invariants and Representations* Princeton University Press, Princeton
- [6] Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-539.