

Protein classification and meta-organization. Methods for global organization of the protein universe.

ISMB 2002 tutorial proposal

Golan Yona
Department of Computer Science
Cornell University

April 5, 2002

1 Objectives

The focus of this tutorial is methods for protein classification and meta-organization. We will start with algorithms for sequence and structure comparison, and mathematical models of protein families. Then we will review databases of protein families (sequence and structure), and the underlying methods and algorithms used to create these databases. We will conclude with lessons we learned from previous and existing classification systems, as well as the new trends in this field, and the future problems that we will have to face as new data is continuously emerging from sequencing projects.

2 Instructor

Golan Yona is an Assistant Professor in the Department of Computer Science in Cornell university. Golan Yona holds a bachelor's degree in Physics and Mathematics and a Ph.D. in Computer Science from the Hebrew University of Jerusalem, Israel. He was a Burroughs-Wellcome postdoctoral fellow in Computational Molecular Biology at Stanford university. His research focuses on computational molecular biology, with an emphasis on large scale analysis of protein sequences and structures, exploring high-order organization within the protein space. He is the developer of the ProtoMap database of protein families and the BioSpace database of the protein models.

3 Teaching experience

Department of Computer Science, Cornell university

- Machine Learning (Spring 2001, Spring 2002)
- Problems and Perspectives in Computational Molecular Biology (Fall 2001, Spring 2002)

Institute of Computer Science, Hebrew University, Jerusalem, Israel

- Algorithms for molecular biology (Fall 1997)

4 Intended audience of the tutorial

Computer scientists, Statisticians, Mathematicians, Biologists. No special background is needed, though, obviously, general background in molecular biology is assumed.

5 Length

Half day

6 Outline

6.1 Background and significance

The ongoing sequencing efforts continue to discover the sequences of many new proteins, whose function is unknown. Currently, protein databases contain the sequences of more than 700,000 proteins, the majority of which has not been characterized yet. This explosion of biological data rules out directed research of all known protein sequences, and consequently, some domains of the protein space will remain untouched by experimentation in the near future. In this view, advanced and automatic tools to organize and analyze this data become a necessity.

The traditional way to analyze a new protein is to search sequence databases for similar, related proteins, using sequence comparison algorithms, such as BLAST [Altschul et al. 1997], FASTA [Pearson & Lipman 1988] and Smith-Waterman [Smith & Waterman 1981]. These algorithms can help identifying the biological function of new protein sequences, if significant similarity with known proteins is detected. In fact, to date, more sequences have been putatively characterized by database searches than by any other single technology. However, in many cases sequences have diverged to such an extent that sequence similarity is undetectable by current means of sequence comparison.

If structural information is available, then one can search for similar proteins in structure databases (PDB, SCOP, CATH, HSSP) using structure comparison algorithms such as [Holm & Sander 1997a, Shindyalov & Bourne 1998, Levitt & Gerstein 1998]. Structure is often conserved more than sequence, and detecting structural similarity may help infer function in cases where sequences have diverged greatly. However, structure information is available for only a small portion of the protein space. Moreover, structural similarities are even more elusive than sequence similarities and structure comparison algorithms can fail to detect all structural similarities. In some cases functional similarity does not even imply similar fold (e.g. the Cellulases family [Wilson & Irwin 1999]).

I will review the traditional ways of sequence and structure analysis, including iterative algorithms such as PSI-BLAST, the successes and the limitations.

6.2 Protein classification

Since the early days of genomic research molecular biologists have tried to make sense out of the accumulated information on protein sequences and structures by classifying proteins into families. Protein sequences are usually classified into sub-families, families and super-families. If structural information is available, they can be further classified into fold families and classes. Some of these relational classes are based on clear evolutionary relationships, such as families and super-families. The definition of other classes is less objective, and is subjective to our current perceptions of the protein fold mechanism and our knowledge of the protein structure repertoire.

Several mathematical representations were developed for protein families. Families are usually represented in the form of either consensus patterns and regular expressions, position-specific scoring matrices or profiles and HMMs. These forms/models differ in their mathematical complexity, as well as in their sensitivity/selectivity.

Comparing a new sequence with a database of protein families, using these representations is more effective than a standard database search in which each sequence is compared with only library sequence at a time. The models that are used to represent the families in those databases can capture subtle features of proteins that belong to those families and therefore are more sensitive in detecting remote homologies. Therefore, protein classification is extremely useful and important for function and structure analysis. This is especially important in large-scale annotation efforts, such as those that follow the large-scale sequencing projects.

I will emphasize the need and the importance of these systems, and review the terms related to protein classification, the concepts, and the different mathematical representations used for protein families.

6.3 Existing approaches

In general large-scale studies that considered all or many of the known protein sequences can be divided into 4 main categories: sequence-based studies, structure-based studies, studies that use alternative representations of proteins, and studies that use combinations of similarity measures.

6.3.1 Sequence-based studies

These studies are further divided into two categories: those focused on finding significant motifs, patterns and domains within protein sequences, and those which apply to complete proteins.

Motif and domain based analyses: Most of these studies yielded databases of protein motifs and domains. Among these are PROSITE [Bairoch 1991], Blocks [Henikoff et al. 1999], PRINTS [Attwood & Beck 1994], ProDom [Corpet et al. 1999], Pfam [Sonnhammer et al. 1997], and Domo [Gracy & Argos 1998]. These studies differ from each other in several aspects. Some are based on manual or semi-manual procedures (e.g. PROSITE, PRINTS), others are generated semi-automatically (Pfam) and the rest - fully automatically (e.g. ProDom, Blocks, Domo). Some focus on short motifs (PROSITE, PRINTS, Blocks) while others seek whole domains and try to define domain boundaries (Pfam, ProDom, Domo).

I will review each one of this databases, explaining the underlying algorithm, the overlap and the differences, and finally will compare all the databases.

Protein based analysis: The studies in this category are applied to whole protein sequences. Most of them draw directly on pairwise comparison [Gonnet et al. 1992, Harris et al. 1992], [Watanabe & Otsuka 1995, Koonin et al. 1996, Barker et al. 1996, Tatusov et al. 1997] [Krause & Vingron 1998, Yona et al. 1999]. All these works cluster the input database, using clustering algorithms. Selected studies will be described in more detail. I will explain the algorithms, the merits and the problems as a result of the application of the specific methodology.

6.3.2 Structure based studies

There are several publicly available classifications of protein architectures including SCOP [Murzin et al. 1995], CATH [Orengo et al. 1997] and FSSP/DALI [Holm & Sander 1997a]. Whereas SCOP is built by the careful manual curation of Dr. Alexei Murzin, both CATH and FSSP are built more or less automatically from structural alignments. CATH has a rather simple hierarchy with just four fold

classes and a few tens of architectures in each class. SCOP has a more complicated hierarchy with 7 fold classes, some containing over a hundred folds. The FSSP classification is automatic with a hierarchy built by the Z-score similarity of proteins in each branch of the tree. While the CATH and FSSP classifications use protein chains as the object of interest, SCOP breaks proteins into domains. I will review each of these studies, the underlying algorithms and the databases.

6.3.3 Alternative representations

Several studies employed alternative representations of protein sequences to the analysis of protein sequences and families. For example, based on dipeptide composition [van Heel et al. 1991, Ferran et al. 1994] or combination of compositional properties and other physical/chemical properties [Hobohm & Sander 1995]. In some cases these representations induced measures of similarity/dissimilarity between complete protein sequences, that were used to classify the sequences into clusters [van Heel et al. 1991], or to search for close relatives [Hobohm & Sander 1995]. In other cases, neural nets were applied to cluster the sequences based on the new representation [Wu et al. 1992, Ferran et al. 1994]. Other studies utilized information from multiple alignments to derive a new representation and study significant patterns in protein families [Han & Baker 1995, Casari et al. 1995].

I will review these studies, and emphasize the implications, and possible extensions of this general methodology.

6.3.4 Combined analyses

Several sequence-structure studies were carried out in the last few years (e.g. [Elofsson & Sonnhammer 1999, Han & Baker 1996, Rigoutsos et al. 1999]). I will review the conclusions of these studies, as well as other studies that actually combined several different metrics to cluster the protein space (e.g. [Yona & Levitt 2000b] structure-based metrics and sequence-based metrics)

6.4 Future directions

I will discuss the limitations of current systems, the need for a unified and exhaustive classification system, and future research problems in this field, from technical problems, to algorithmic and theoretical problems.

I will base the presentation on my personal research experience in this field for the last nine years, and our lessons from applying several different methodologies for the analysis of the protein space, while committing to extensive and balanced presentation of all related studies.

References

- [Altschul et al. 1997] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
- [Attwood & Beck 1994] Attwood, T. K. & Beck, M. E. (1994). PRINTS-a protein motif fingerprint database. *Protein Eng.* **7**, 841-848.
- [Bairoch 1991] Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **19**, 2241-2245.
- [Barker et al. 1996] Barker, W. C., Pfeiffer, F. & George, D. G. (1996). Superfamily classification in PIR-international protein sequence database. *Methods Enzymol.* **266**, 59-71.
- [Casari et al. 1995] Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171-178.

- [Corpet et al. 1999] Corpet, F., Gouzy, J., & Kahn, D. (1999). Recent improvements of the ProDom database of protein domain families. *Nucl. Acids Res.* **27**, 263-267.
- [Elofsson & Sonnhammer 1999] Elofsson, A. & Sonnhammer, E. L. (1999). A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics* **15:6**, 480-500.
- [Ferran et al. 1994] Ferran, E. A., Pflugfelder, B. & Ferrara P. (1994). Self-Organized Neural Maps of Human Protein Sequences. *Protein Sci.* **3**, 507-521.
- [Gonnet et al. 1992] Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443-1445.
- [Gracy & Argos 1998] Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. I. Integration of copositional similarity search, local similarity search and multiple sequence alignment. II. Delineation of domain boundaries from sequence similarity. *Bioinformatics* **14:2**, 164-187.
- [Han & Baker 1995] Han, K. F. & Baker, D. (1995). Recurring local sequence motifs in proteins. *J. Mol. Biol.* **251**, 176-187.
- [Han & Baker 1996] Han, K. F. & Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl. Acad. Sci. USA* **93**, 5814-5818.
- [Harris et al. 1992] Harris, N. L., Hunter, L. & States, D.J. (1992). Mega-classification: Discovering motifs in massive datastreams. In *Proc. of the 10th national conf. on AI*, 837-842, AAAI press/The MIT Press, Menlo park/Cambridge.
- [Henikoff et al. 1999] Henikoff, J. G., Henikoff, S. & Pietrokovski, S. (1999). New features of the Blocks Database servers. *Nucl. Acids Res.* **27**, 226-228.
- [Hobohm & Sander 1995] Hobohm, U. & Sander, C. (1995). A sequence property approach to searching protein database. *J. Mol. Biol.* **251**, 390-399.
- [Holm & Sander 1997a] Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.* **25**, 231-234.
- [Koonin et al. 1996] Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996). Protein sequence comparison at genome scale. *Methods Enzymol.* **266**, 295-321.
- [Krause & Vingron 1998] Krause, A. & Vingron, M. (1998). A set-theoretic approach to database searching and clustering. *Bioinformatics* **14:5**, 430-438.
- [Levitt & Gerstein 1998] Levitt, M & Gerstein, M. (1998). A Unified Statistical Framework for Sequence Comparison and Structure Comparison. *Proc. Natl. Acad. Sci. USA* **95**, 5913-5920.
- [Murzin et al. 1995] 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**, 536-540.
- [Orengo et al. 1997] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH-a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108.
- [Pearson & Lipman 1988] Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448.
- [Rigoutsos et al. 1999] Rigoutsos, I. Gao, Y., Floratos, A. & Parida, L. (1999). Building dictionaries of 1D and 3D motifs by mining the unaligned 1D sequences of 17 archaeal and bacterial genomes. In *the proceedings of ISMB 99*, 223-233.
- [Shindyalov & Bourne 1998] Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739-747.
- [Smith & Waterman 1981] Smith, T. F. & Waterman, M. S. (1981). Comparison of Biosequences. *Adv. App. Math.* **2**, 482-489 .
- [Sonnhammer et al. 1997] Sonnhammer, E. L., Eddy, S. R., Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-420.
- [Tatusov et al. 1997] Tatusov, R. L., Eugene, V. K. & David, J. L. (1997). A genomic perspective on protein families. *Science* **278**, 631-637.
- [van Heel et al. 1991] van Heel, M. (1991). A new family of powerful multivariate statistical sequence analysis techniques. *J. Mol. Biol.* **220**, 877-887.
- [Watanabe & Otsuka 1995] Watanabe, H. & Otsuka, J. (1995). A comprehensive representation of extensive similarity linkage between large numbers of proteins. *Comp. App. Biosci.* **11:2**, 159-166.

- [Wilson & Irwin 1999] Wilson, D. B. & Irwin, D. C. (1999). Genetics and properties of cellulases. *Adv. Biochem. Eng.* **65**, 2-21.
- [Wu et al. 1992] Wu, C., Whitson, G., Mclarty, J., Ermongkonchai A. & Chang, T. (1992). Protein classification artificial neural system. *Protein Sci.* **1**, 667-677.
- [Yona & Levitt 2000b] Yona, G. & Levitt, M. (2000). Towards a complete map of the protein space based on a unified sequence and structure analysis of all known proteins. *In the proceedings of ISMB 2000*, 395-406, AAAI press, Menlo Park.
- [Yona et al. 1999] Yona, G., Linial, N. & Linial, M. (1999). ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins*, **37**, 360-378.