

High Performance Computing for Protein Sequence Modeling

Raquel Melo Minardi¹, Karine Bastard²³⁴, François Artiguenave²³⁴⁵

¹Department of Computer Science, Federal University of Minas Gerais (UFMG)
Belo Horizonte, Brazil

²Genoscope, Institut de Génomique, CEA, Evry cedex

³UMR 8030, CNRS, Evry cedex

⁴University Evry Val d'Essonne, Evry cedex, France

⁵Medical Sciences Faculty, UNICAMP State University, Campinas, Brazil

raquelcm@dcc.ufmg.br, kbastard@genoscope.cns.fr,
francois@fcm.unicamp.br

Abstract. *Common bioinformatics approaches for protein function prediction are based on sequence classification and annotation transfer from known proteins to their closest homologous. These approaches are restricted to homogeneous superfamilies and are not able to predict new activities. Structural biology offers a new insight to overcome this problem by adding protein structure information. Using a 3D modeling approach, we developed a method to predict evolution of catalytic sites in superfamilies. We present results obtained during a computational Grand Challenge on the 350 Tflops CCRT French Supercomputing Facility that illustrate how high performance computing provide new perspectives for understanding protein evolution and function.*

Resumo. *As abordagens mais utilizadas para predição de função de proteínas são baseadas na classificação de sequências e na transferência de funções de proteínas conhecidas para seus homólogos mais próximos. Estas abordagens são restritas às super-famílias homogêneas e não são úteis na predição de novas atividades. A biologia estrutural oferece novos meios de superar esta limitação através da agregação da informação sobre as estruturas de proteínas. Neste trabalho, apresentamos os resultados do Grand Challenge, um desafio do supercomputador francês CCRT de 350TFlopss que ilustra as novas perspectivas que este tipo de tecnologia juntamente com as técnicas de e-Science nos fazem vislumbrar rumo ao entendimento das funções e evolução de proteínas.*

1. Introduction

With the increasing number of genomes being sequenced, a critical challenge concerns functional prediction of proteins encoded by these genes. Several methods are widely used, including sequence and gene context analysis. While these methods are efficient for closed homologous, they are limited to proteins sharing homologies with known proteins. To overcome such problems, structural information can be used, in combination with sequence data, in order to provide a more successful understanding of protein function on molecular basis. To gather as much as possible structural information, recent international initiatives are developing high throughput technologies for experimental resolution of protein structures. However, costs and technical difficulties still explain the huge gap between number of sequences and number of structures available. Homology Modeling is a reliable approach to bridge this gap as long as the target sequence has a minimum sequence similarity with at least one experimentally solved protein structure [Tramontano *et al.*, 2001].

To evaluate the potential target of structural approach for annotation, we examined PFAM database [Finn *et al.*, 2008] and estimated the size of data eligible to modeling in addition to sequence analysis (table 1). Looking at PFAM v. 23.0 reveals the high number of families without annotation, DUFs (Domain of Unknown Function) and UPFs (Uncharacterized Protein Family) representing 21% of the overall database. While the number of available structures represents only 10% of unknown families, about 200 families are potentially tractable by structural analysis. To complete this view, objectives of the PSI (Protein Structure Initiative) is to solve at least one protein structure for all PFAM families.

Table 1. PFAM (v. 23.0) families lacking annotation with structural information

Total PFAM entries	DUF	UPF	DUF/UPF with PDB entries
10340	2156	92	208

Depending on the accuracy of the structural information, one can reach different level of elucidation of the relationships between structure and function. At high level of resolution, biochemical reaction may be predicted using molecular modeling and substrate docking methods. At low-resolution level, fold assignment and 3D motif searching can support functional annotation. For example, conserved structural cavities in a protein family are an indicator of active sites. Residues in these cavities are subject to different selective pressures so that multiple alignments can reveal conserved profiles. Hidden Markov Models (HMMs) provide a coherent statistical theory for this analysis.

In this study, we applied a recent developed methodology to 83 DUFs families present in the Cloaca meta-genome data studied at Genoscope (1 million prokaryote sequence proteins from Evry Waste Water Anaerobic Plant). The modelling of 60,000 sequences (1,000 models for each sequence) were obtained during a Computational Grand Challenge proposed by the French Super-Computing Facility CCRT of the CEA (Commissariat à l'Énergie Atomique et aux Énergies Alternatives). The 60.10⁶ generated models are stored in a new structural database, integrating information on

sequence, structure and conserved pockets. Initial analysis of some families allowed the identification of new enzymatic functions and specificities.

2. Active Site Modeling and Clustering

In a recent work [Melo-Minardi *et al.*, 2010], we developed ASMC (Active Site Modeling and Clustering), a methodology for analysis of residues of protein cavities to detect determinant ones involved in catalytic or enzyme specificity. Briefly, the ASMC method is an unsupervised method for the classification of protein sequences based on structural information of protein pockets. ASMC combines homology modeling of family members, structural alignment of modeled active sites and a subsequent hierarchical conceptual classification. Comparison of profiles obtained from computed clusters allows the identification of residues correlated to subfamily function divergence, called specificity determining positions. Instead of using a global MSA, we use structural alignments of the predicted cavity residues. From these alignments, we are able to divide the protein families into groups of similar profiles using conceptual clustering [Fisher, 1987]. The analysis detects intra-family variations that can be responsible for function and/or specificity.

The different steps of the ASMC method are summarized in figure 1:

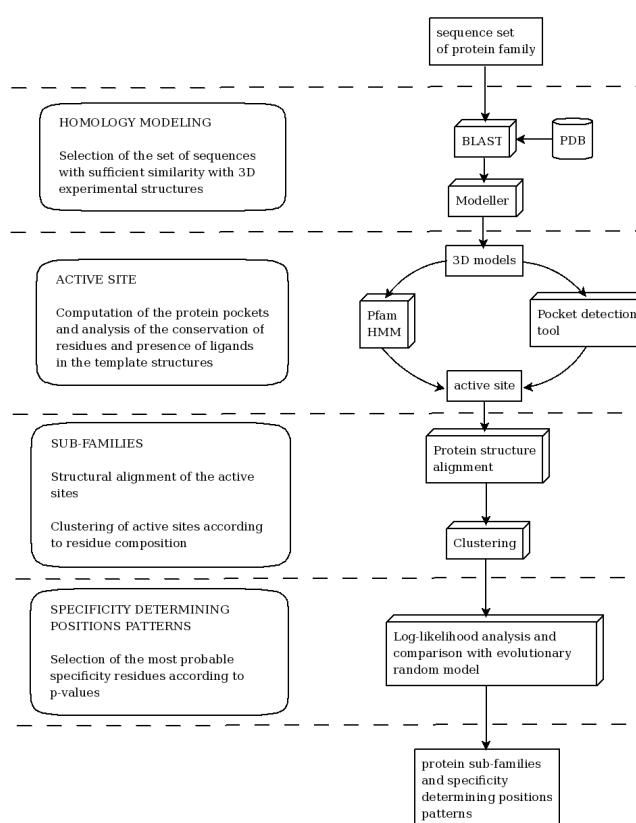


Figure 1. ASMC method diagram

3. Results

We scaled up the methodology to propose a systematic functional annotation of new families, focusing on poorly annotated ones. From 208 families identified in PFAM, we restricted our analysis to 83 families present in *in-house* data from the Evry wastewater Metagenomics project, which are of specific interest for the study of anaerobic prokaryote metabolism (main research project of Genoscope). For the 83 families, the 60.10^6 modelling jobs (Modeller v. 9.6) were run on the “Platine” supercomputing facility at CEA/CCRT. Platine is a cluster of Novascale server including 932 computing nodes and 26 administration or IO nodes. Each node is composed of 4 1.6 Ghz bi-cores Intel® Itanium. Each node has a 24 Gb memory. The Novascale servers, running linux, are interconnected by a Volaire network (InfiniBand DDR). The LSF batch system is provided by the Platform Computing company, and use resources of Slurm®. The jobs were dispatched on 4000 processors for a total of 280,000 CPU hours. The results were obtained in 70 hours. From the computed models, cavities were detected and conservation profiles were computed. Cavities’ key residues were identified and families were clustered into subfamilies. All results are available at (<http://bioinfo.speed.dcc.ufmg.br/3dbio/raquelcm/dufs/index.jsp?idioma=ingles>).

In order to illustrate perspectives opened by these results, we will present a detailed study on a family for which a new enzymatic activity has been characterized [Bellinzoni et al., 2011]. For this family, DUF849, data from structural modelling provided key insights for enzymatic mechanism elucidation and for analysis of evolution of the activity inside the protein superfamily. The main predicted clusters (figure 2) were tested for enzymatic activities and presented different response.

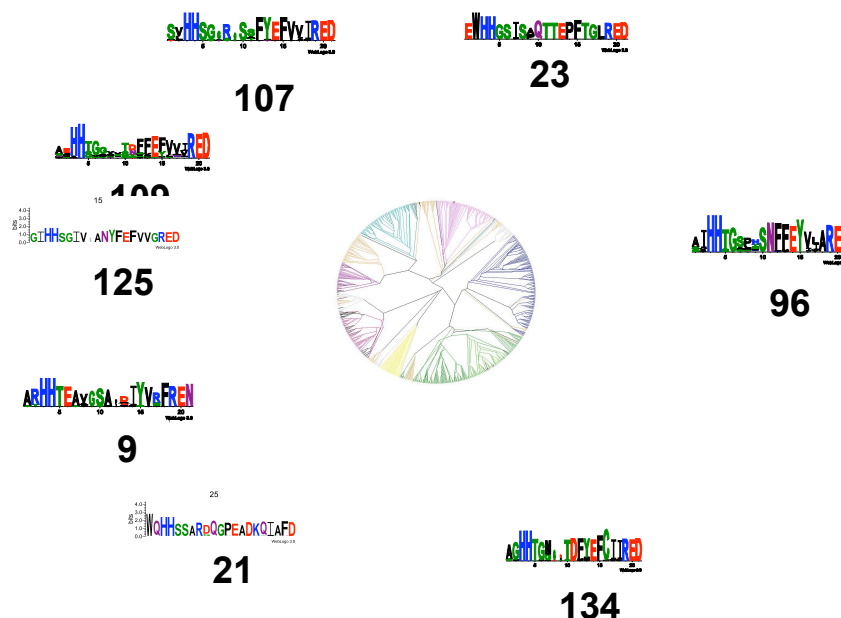


Figure 2. Mapping of catalytic site clusters and profiles on the evolutive tree of the DUF 849 family.

4. Conclusion

For the recent years, many efforts are provided to understand complex molecular and cellular systems including new computations analysis, modelling and simulation capabilities. This work presented an initiative to approximate two computational disciplines of the biology, bioinformatics and chemo-informatics. While bioinformatics focuses mainly to large scale and genomics sequence analysis, traditional chemoinformatics works focused on specific molecular and enzymatic aspects. Current development of informatics infrastructure and the development of supercomputing facilities open new perspectives which combine both statistical and data mining approaches to mathematical modelling approaches. Such approaches can bring new impulses to biology by providing *in silico* tools able to predict *de novo* biological functions.

References

- Bellinzoni, M., Bastard, K., Perret, A., Zaparucha, A., Perchat, N., Vergne, C., Wagner, T., Melo-Minardi, R., Artiguenave, F., Cohen, G., Weissenbach, J., Salanoubat, M., Alzari, P. (2011) 3-keto-5-amino-hexanoate cleavage enzyme: a common fold for an uncommon reaction, Submitted to PNAS.
- Eswar, N. et al. (2006) Comparative protein structure modeling using modeller. *Curr. Protoc. Bioinformatics*, Chapter 5, Unit 5.6. Eswar, N. et al. (2008) Protein structure modelling with Modeller. *Methods Mol. Biol.*, 426, 145–159.
- Finn R.D. et al. (2008) The Pfam protein families database. *Nucleic Acids Res.*, 36, D281–D288.
- Fisher, D. (1987) Knowledge acquisition via incremental conceptual clustering. *Mach. Learn.*, 2, 139–172.
- Guilloux, V.L. et al. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10.
- Holmes, G. et al. (1994) Weka: a machine learning workbench. In *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*.
- Madhusudhan, M. et al. (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.*, 22, 569–574.
- Melo-Minardi, R.C., Bastard, K. and Artiguenave, F. (2010) “Identification of subfamily-specific sites based on active sites modeling and clustering”. *Bioinformatics*, 26, 3075–3082.
- Pei, J. et al. (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, 22, 164–171.
- Shatsky, M. et al. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, 56, 143–156.
- Sol, A.D. et al. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, 326, 1289–1302.

- Sonnhammer,E. et al. (1997) Pfam: a comprehensive database of protein families based on seed alignments. *Proteins*, 28, 405–420.
- Tramontano,A. and Morea,V. (2003) Assessment of homology-based predictions in CASP5. *Proteins*, 53 (Suppl. 6), 652–368.
- Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, 18, 691–699.