Bioinformatic analysis of diverse protein superfamilies to design improved enzymes

Dmitry Suplatov, Nikolay Panin, Tatyana Shcherbakova, Ilyas Khaliullin, Vytas Švedas

Lomonosov Moscow State University, Belozersky Institute of Physicochemical Biology and Faculty of Bioengineering and Bioinformatics, Moscow, Russia, vytas@belozersky.msu.ru

Enzymes within a family usually share a common function but differ in more specific features and can be divided into subfamilies with different catalytic activity, substrate specificity, enantioselectivity, stability, etc. Evolution of proteins imposes constraints on sequence variation which can be studied by aligning sequences and structures of functionally diverse homologs. Bioinformatic analysis of resulting superimpositions of proteins within a superfamily can be used to decipher the natural mutation patterns and their implications for protein function and stability. Positions which are conserved in a column of a multiple alignment can define general properties of the entire superfamily (for example, have direct roles in enzyme catalytic machinery) but do not explain functional diversity. Another mutation pattern can be described as subfamily-dependent conservation – conserved within functional subfamilies but different between them. To describe these positions a term "subfamily-specific position(s)" or SSP(s) can be used to outline that distribution of amino acid types in a column is specific to functional subfamilies. Multiple methods have been developed to detect SSPs and highlight their potential functional role [1-4]. Majority of these studies, however, were limited to *in silico* research and so far only a few experimental evaluations have been performed.

We have recently developed a new method of bioinformatic analysis to identify function-related variable residues in protein structures that are responsible for functional divergence within superfamilies of homologous enzymes [5]. A new algorithm has been suggested to predict functional subfamilies, a novel scoring function of subfamily-dependent distribution of amino acids has been implemented, which takes into account sequence/structural information and physicochemical properties of amino acid side chains, and ranking is performed to select the most statistically significant hotspots automatically for further

evaluation. The Zebra method can be used as a tool to explore SSPs with different structural localization in order to understand their implication to structure-function relationship and protein function; interface is available on-line at http://biokinet.belozersky.msu.ru/zebra [6]. Subfamily-specific positions are preferentially associated with catalytic and allosteric sites in enzymes. Our results indicate that presence of the subfamily-specific positions is a very powerful factor for ranking of pockets and cavities in a protein structure by their functional significance. Method pocketZebra implements the power of bioinformatics and geometry-based structural approaches to identify and classify subfamily-specific binding sites in proteins by their functional importance, distinguish particular positions in the structure that determine selective accommodation of ligands, can be used to identify allosteric sites and to annotate proteins with unknown function [7]. Interface to pocketZebra is available on-line at http://biokinet.belozersky.msu.ru/pocketzebra.

The developed methodology has been applied to study structure-functional relationship in various enzyme superfamilies: α/β-hydrolases, Ntn-hydrolases, penicillin-binding proteins, etc. Large structure-based sequence alignments have been created for each superfamily. Remote evolutionary relatives were superimposed by structural comparison, while sequence-based alignments were assumed meaningful for closer homologs [8]. Systematic bioinformatic analysis of genomic and structural information corresponding to each selected superfamily of enzymes has been carried out to identify functionally important subfamily-specific positions. Mutations at subfamily-specific positions have been used as a tool to study their biological role and to design improved variants. Molecular modeling and *in silico* screening have been implemented to construct the corresponding *in silico* libraries of protein mutants and evaluate influence of selected residues on structural stability, binding and catalytic conversion of selected substrates. The most promising variants have been selected for experimental production and evaluation. It has been shown, that patterns of SSPs can be effectively used to design enzyme mutants with improved catalytic properties and to predict functional properties of enzymes. Substitutions at the subfamily-specific positions have led to catalytic promiscuity of a *Candida antarctica* lipase B [9]. Mutation of a subfamily-specific position has allowed to improve stability of *Escherichia coli* penicillin acylase in alkaline medium and resistance to inactivation by high concentrations of substrates [10, 11]. D-

aminopeptidase from *Ochrobactrum anthropi* with extended substrate specificity towards large substrates, which are not converted by the wild type enzyme, has been produced by mutations at the subfamily-specific positions.

We consider the subfamily-specific positions as an important tool to study structure-function relationship and regulation in large protein superfamilies, classify functionally important binding sites and annotate proteins with unknown function. From a practical perspective SSPs can be used as hotspots for directed evolution or rational design experiments to design novel efficient biocatalysts [12]. The role of the subfamily-specific positions in protein function and evolution should be further studied in more detail, first of all experimentally.

Financial support from RFBR (grant #14-08-00987). Lomonosov Moscow State University supercomputer clusters were used for resource-intensive computations.

1. G. Casari, C. Sander, A. Valencia (1995) A method to predict functional residues in proteins, *Nature Structural Biology*, **2:**171-178.

2. L. Mirny, M. Gelfand (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors, *J.Mol.Biol.,* **321:**7-20.

3. O. Kalinina, A. Mironov, M. Gelfand, A. Rakhmaninova (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families, *Protein Science*, **13**:443-456.

4. D. de Juan, F. Pazos, A. Valencia (2013) Emerging methods in protein co-evolution, *Nature Reviews Genetics*, **14(4)**:249-261.

5. D. Suplatov, D. Shalaeva, E. Kirilin, V. Arzhanik, V. Švedas (2014) Bioinformatic analysis of protein families for identification of variable amino acid residues responsible for functional diversity, *J.Biomol.Struct.Dyn.,* **32(1)**:75-87.

6. D. Suplatov, E. Kirilin, V. Takhaveev, V. Švedas (2014) Zebra: web-server for bioinformatic analysis of diverse protein families, *J.Biomol.Struct.Dyn.*, **32(11)**:1752-58.

7. D. Suplatov, E. Kirilin, M. Arbatsky, V. Takhaveev, V. Švedas (2014) pocketZebra: a web-server for automated selection and classification of subfamily-specific binding sites by bioinformatic analysis of diverse protein families. *Nucl. Acids Res.*, **42(W1)**:344-349.

8. D. Suplatov, V. Arzhanik, V. Švedas (2011). Comparative Bioinformatic Analysis of Active Site Structures in Evolutionarily Remote Homologues of α,β-Hydrolase Superfamily Enzymes, *Acta naturae*, **3(1)**:93-98.

9. D. Suplatov, W. Besenmatter, V. Švedas, A. Svendsen (2012) Bioinformatic analysis of alpha/beta-hydrolase fold enzymes reveals subfamily-specific positions responsible for discrimination of amidase and lipase activities, *Protein Eng.Des.Sel.*, **25(11)**:689-697.

10. D. Suplatov, N. Panin, E. Kirilin, T. Shcherbakova, P. Kudryavtsev, V. Švedas (2014) Computational design of a pH stable enzyme: understanding molecular mechanism of penicillin acylase's adaptation to alkaline conditions, *PLoS ONE*, **9(6)**:e100643.

11. T. Shcherbakova, N. Panin, D. Suplatov, I. Shapovalova, V. Švedas (2015). The βD484N mutant of penicillin acylase from Escherichia coli is more resistant to inactivation by substrates and can effectively perform peptide synthesis in aqueous medium, *J. Mol. Cat. B*, **112**:66-68.

12. D. Suplatov, V. Voevodin, V. Švedas (2015) Robust enzyme design: Bioinformatic tools for improved protein stability, *Biotechnol J.*, **10:**344–355