Determinants of protein function revealed by combinatorial entropy optimization

Boris Reva*, Yevgeniy Antipin*, Chris Sander *

*Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA, borisr@mskcc.org

Background. Genome projects are generating a rapidly increasing number of protein sequences, but our knowledge of functional details lags behind. Fortunately, functional constraints in evolution have created information-rich conservation patterns in protein families. If one can decode these patterns, one can derive detailed functional hypotheses. Here, we focus on decoding the patterns of specificity residues. Such residues are conserved in each protein subfamily, but differ between functionally diverse subfamilies.

Results. We present a new algorithm to solve the combinatorial complex problem of identifying specificity residues and, simultaneously, the corresponding optimal division into subfamilies. In our approach, called combinatorial entropy optimization (CEO), we optimize a conservation contrast function over different assignments of proteins to subfamilies. We validate the method by comparing sets of predicted specificity residues with sets of experimentally known functional residues, such as interaction residues observed in three-dimensional macromolecular complexes, and get good agreement between prediction and observation.

Conclusion. The method, at http://proteinkeys.org, takes a multiple sequence alignment as input and returns subfamilies and a set of specificity residues. The computed subfamilies may be used, e.g., to assign a likely function to new protein sequences or to choose maximally informative targets for structural genomics projects. The computed specificity residues may be used to design highly specific mutation experiments that test function with minimal side effects; to build sharper and more informative evolutionary trees that more accurately reflect functional relatedness; to predict interactions with proteins; and, to estimate the functional consequences of genetic variation [1]-[7].

References

[1] Reva B, Antipin Ye and Sander Ch.: Determinants of protein function revealed by combinatorial entropy optimization. Genome Biology 2007, 8:R232.

[2] Casari G, Sander C, Valencia A: A method to predict functional residues in proteins. Nat Struct Biol 1995, 2:171-178.

[3] Lichtarge O, Bourne HR, Cohen FE: An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996, 257:342-358.

[4] Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. Bioinformatics 2001, 17:1035-1046. PubMed Abstract — Publisher Full Text OpenURL

[5] Correlated mutation analyses on very large sequence families. Chembiochem 2002, 3:1010-1017.

[6] Switch-of-function mutants based on morphology classification of Ras superfamily small GTPases. Cell 2003, 113:315-328.

[7] Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics 2004, 83:970-979.