What Proteins Are Made From? Informational Way To Protein Alphabet

Alexander Gorban[*1], Mikhail Kudryashev[2], and Tatiana Popova[3]

[1]*Centre for Mathematical Modeling, University of Leicester, UK*
[2]*Krasnoyarsk State University, Krasnoyarsk, Russia*
[3]*Institute of Computational Modeling, Akademgorodok, Krasnoyarsk, Russia*

Keywords. Amino acid, protein, classification, relative entropy, alphabet reduction

What proteins are made from, as the working parts of the living cells protein machines? To answer this question, we need a technology to disassemble proteins onto elementary functional details and to prepare lumped description of such details. Our hypothesis is that informational approach to this problem is possible. We propose a way of hierarchical classification of amino acids that makes the primary structure of protein *maximally non-random.* In order to formalize this idea we follow [1,2] and analyse *frequency dictionaries* of short protein fragments, and *relative entropies* of such frequency dictionaries. The *entropic optimality principle* is formulated and applied for amino acids classifications for various databases of primary protein sequences. In contrary to the widespread *MaxEnt* approach (that is, of maximal disorder), we deal with the principle of maximal order. The following properties of amino acids binary informational classifications are studied 1) the *existence and uniqueness* of optimal classification for given frequency dictionary, 2) structure of the set of highly informative classifications in the vicinity of the optimal one, 3) *stability/instability* of the optimal classification with respect to variations of the frequency dictionary, 4) similarity between classifications constructed on the basis of 2-letter words frequencies and those constructed on the basis 3, 4 and 5-letter words frequencies. We compared informational binary classifications of amino acids with classifications obtained by other methods. Amino acids groupings mentioned in most of reviewed papers do have moderate similarity with two types of Hydrophobic/Polar classification while informational classifications is shifted to Charged/Uncharged property. Classification of [3] is the only one to be rather close to informational classifications. Detailed statistic data are published in preprint [2]. The binary informational classification gives us "the first letter of protein alphabet". Algorithms of *hierarchical information classification* are developed in order to find the following "letters". On each level of hierarchy we find the optimal classification that is independent (with given accuracy) of classifications obtained on the previous levels. The second level is surprisingly independent of all known "usual" amino acids classifications. Below the example of the first and the second classifications is presented for E-coli proteome (number of proteins is 5797 [4]):

1st classification: 1st class: A,C,D,E,F,G,I,K,L,M,N,P,S,T,V; 2nd class: H,Q,R,W,Y.
2nd classification 1st class: A,E,K,L,M,P,Q,R,T,V,W; 2nd class: C,D,F,G,H,I,N,S,Y.

Binary classification tree: Zero level (one class) {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}
First level (two classes) {A,C,D,E,F,G,I,K,L,M,N,P,S,T,V}, {H,Q,R,W,Y}
Third level (four classes) {A,E,K,L,M,P,T,V}, {C,D,F,G,I,N,S}, {Q,R,W}, {H,Y}.

The next step of the research program should be the informational classifications analysis of 2 and 3-symbol elements of primary structures presented as a sequence of 2 and 3-letters elements and so on.

References

1. Bugaenko N.N., Gorban A.N., Sapozhnikov A.N. Minimum entropy principle and classification of symbols for revealing statistical regularities in text. In: *Proc. of the first Int. Conf. on Bioinformatics of Genome Regulation and Structure*, 1998; Vol. 2: 280-282; http://www.bionet.nsc.ru/bgrs/author/gorban.html .
2. Gorban A.N., Kudryashev M., Popova T., On the Way to Protein Alphabet: Informational Classification of Amino Acids in Comparison to Other Classifications, e-print: http://arxiv.org/abs/q-bio.BM/0501019 .
3. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.* **2000**, 13(3), 149-52.
4. *Swiss Institute of Bioinformatics website*, Expert Protein Analysis System http://expasy.org/

[*] *University Road, Leicester, LE1 7RH, UK, ag153@le.ac.uk*