FUNCTIONAL PROFILE OF *FRANCISELLA TULARENSIS* PROTEINS

Lenka HERNYCHOVÁ, Aleš MACELA, Jiří STULÍK
Institute of Nuclear and Biological Medical Defence, Purkyně Military Medical Academy, Hradec Králové, Czech Republic

Introduction

In spite of the enormous quantity of information from genomic and proteomic analysis the data do not mirror precisely *in vivo* situation in respect to functional profile of cellular system studied. Generally, the function of molecular entities is strictly associated with their tertiary structure, subcellular localization inside individual cell compartments and basically on the consensual sequences, which compose the functional domains of protein molecule. Moreover, the metabolic and signaling cellular cascades rely on the protein-protein interactions, which are also out of the scope of, so called "expression proteomics". To overcome this gap, the construction of virtual proteome using the algorithms of bioinformatics and the systematic study of protein-protein interactions through the isolation of protein complexes called "cell-maps proteomics" precedes the creation of real bacterial proteome. Moreover, the proteomic study of isolated individual cellular compartments can bring more precise information of function-related protein expression.

Virtual proteome

Virtual proteome increases the preliminary value of possible interpretation of data obtained from genome and post-genome studies. Bioinformatics is one of the possible tools for the construction of virtual proteome that allows summarization of both 2-DE gel electrophoresis and mass spectrometry data in combination with genomic data and enable their "functional" interpretation including post-genomic properties (expression level of proteins, post-translation modifications of proteins, protein-protein interaction etc.).

The virtual proteome of *Francisella tularensis* microbe was created from the existing ORFeom by translation of tentative ORFs. Translated OFRs were subsequently evaluated by algorithms, which predict, on the bases of inserted amino acid sequences, the function of proteins (COG algorithm), their intracellular topology (PSORT), discrimination between soluble and membrane proteins (SOSUI), or prediction of protein to be secreted (SPScan, SignalP, Tmpred) [1].

Intracellular bacteria *Francisella tularensis* strain Schu4 was sequenced and nucleotide sequence data from genomic DNA has been partially assembled to provide 1·83 Mb of the genome sequence [2]. A preliminary analysis of the whole genome allows creating complete ORFeom. In total 1804 candidate ORFs were identified in the data set of which 1289 were thought to encode proteins (743 genes with known function, 133 genes coding for hypothetical proteins and 413 genes with no database match) [3]. The putative proteins encoded by the *F. tularensis* sequence data were assigned to one of 16 different categories, to enable a comparison with other bacteria whose genomes have previously been sequenced [3].

Prediction of secondary structure and topography of microbial proteins – Application of SOSUI and PSORT algorithms

A software system, SOSUI, was previously developed for discriminating between soluble and membrane proteins together with the prediction of transmembrane helices [4]. SOSUI system is a www-based tool with its Internet address http://www.tuat.ac.jp/~mitaku/sosui. The performance of the system was 99% for the discrimination between two types of proteins (membrane versus soluble protein) and 96% for the prediction of transmembrane helices [5]. This theoretical method does not depend on the sequence alignment but on the physicochemical properties of amino acid sequences. The average hydrophobicity of the most hydrophobic helices in membrane proteins had a characteristic relationship with the length of the protein [6]. Two prediction and two graphs are presented in the output page: (1) the type of protein, (2) the region of transmembrane helices when the protein is a membrane type, (3) a graph of the hydropathy plot, (4) helical wheel diagrams of all transmembrane helices. SOSUIsignal Beta Version calculates signal peptides from amino acid sequence of protein. The results from SOSUI algorithm for identified proteins are presented in tables 1 and 2. Other version of SOSUI program SOSUI (Batch) allows analysis of large amount of amino acid sequences of individual proteins. Preliminary analysis of *Francisella tularensis* ORFeom by SOSUI system predicted that 34% from putative proteins analyzed till now could be classified as a membrane proteins.

The topology of proteins listed in tables 1 and 2 was evaluated by PSORT algorithm. PSORT algorithm can be used for prediction of gram-positive and gram-negative bacterium, yeast, animal and plant proteins localization. Algorithm is displayed on Internet address: http://psort.nibb.ac.jp. The predicted candidate localization-sites for gram-negative bacterium are bacterial outer membrane, bacterial periplasmic space, bacterial inner membrane and bacterial cytoplasm. In gram-negative bacteria, most periplasmic and outer membrane proteins have a signal sequence in the N-terminus, which is cleaved off after the translocation to the cytoplasmic membrane. Some cytoplasmic membrane proteins have signal sequence un-cleavable, which remains on mature protein as transmembrane segment. PSORT first predicts the presence of signal sequences by McGeoch´s method [7] modified by Nakai and Kanehisa [8]. It considers the N-terminal basically charged region and the central hydrophobic region of signal sequences. A discriminant score is calculated from the three values: length of central hydrophobic region, peak value of central hydrophobic region, and net charge of N-terminal basically-charged region. A large positive discriminant score means a high possibility to possess a signal sequence whether it is cleaved off or not. Next, PSORT applies von Heijne´s method of signal sequence recognition [9]. It is a weight-matrix method and incorporates the information of consensus pattern around the cleavage sites and thus it can be used to detect uncleavable signal sequences. A large positive output means a high possibility that it has a cleavable signal sequence. The position of possible cleavage site is also reported. Next parameters that PSORT calculates are transmembrane segments and lipoproteins. Transmembrane segments exist in the cytoplasmic membrane proteins only. Thus, these segments can be regarded as the sorting signal into the cytoplasmic membrane. PSORT employs Klein et al.´s method to detect potential transmembrane segments [10]. The discrimination between the cytoplasmic membrane and the outer membrane is done as follows based on the experiment of Yamaguchi [11]. If a lipoprotein has a negatively charged residue at second or third position of the mature part, it is sorted to the inner membrane; otherwise, it is sorted to the outer membrane (see PSORT Users´ Manual at http://psort.nibb.ac.jp /helpwww.html).

Structure-function relationship of microbial proteins – Exploitation of COG program. The predicted functional profile of tentative proteins listed in tables 1 and 2 was obtained by applying of COGs algorithms. The database of Clusters of Orthologous Groups of proteins (COGs) has been incepted as phylogenetic classification of proteins from complete genomes [12]. COGs were delineated by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages to serve as platform for functional annotation of newly sequenced genomes and for studies on genome evolution, and post-genomic evaluation of protein analysis [13]. The COGs have been classified into 17 broad functional categories, including a class for which a general function prediction, usually that of biochemical activity, was feasible and a class of uncharacterized COGs. In strict sense, the COGNITOR program that accompanies the COGs database and assign new proteins, typically from newly sequenced genomes [14], was used for prediction of identified *Francisella tularensis* proteins function. COGs algorithm is a www-based tool with its Internet address http://www.ncbi.nlm.nih.gov/COG.

Conclusion

Computer algorithms evaluated data from genome and proteome analysis complete the mosaics of basic information ordered to virtual proteome. The complexity of information collected in virtual proteome offers the possibility to use the discriminative features for the creation of real proteome. The basic distinctness between virtual and real proteome can be seen in unambiguous expression of real protein species in *in vivo* system under strictly defined conditions. Moreover, the precise identification of protein species including posttranslational modification and evaluation of structure-function relationship using knockout gene and point mutation technologies should be the stipulate attribute of real proteome. The information created real proteome can backward influence the virtual proteome in respect to higher probability of protein properties prediction.

References

1. GOMEZ, M. – JOHNSON, S. – GENNARO, ML. Identification of secreted proteins of *Mycobacterium tuberculosis* by a bionformatic approach. *Infect. Immun.*, 2000, 68, p. 2323–2327.
2. KARLSSON, J., et al. Sequencing of the *Francisella tularensis* strain Schu 4 genome reveals the shikimate and purine metabolic pathways, targets for the construction of a rationally attenuated auxotrophic vaccine. *Microb. Compar. Genomics*, 2000, vol. 5, no. 1, p. 25–39.
3. PRIOR, RG., et al. Preliminary analysis and annotation of the partial genome sequence of *Francisella tularensis* strain Schu 4. *J. Applied Microbiol.*, 2001, 91, p. 614–620.
4. HIROKAWA, T. – BOON-CHIENG, S. – MITAKU, S. SOSUI: classification and secondary structure prediction

system for membrane proteins. *Bioinformatics*, 1998, 14, p. 378–379.
5. MITAKU, S., et al. Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system. *Biophys. Chem.*, 1999, 82, p. 165–171.
6. MITAKU, S. – HIROKAWA, T. Physicochemical factors for discriminatin between soluble and membrane proteins: hydrophobicity of helical segments and protein length. *Protein Eng.*, 1999, 12, p. 953–957.
7. McGEOCH, DJ. On the predictive recognition of signal peptide sequences. *Virus Res.*, 1985, 3, 271–286.
8. NAKAI, K. – KANEHISA, M.: Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins*, 1991, 11, p. 95–110.
9. Von HEIJNE, G. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.*, 1986, 14, p. 4683–4690.
10. KLEIN, P. – KANEHISA, M. – DeLISI, C. The detection and classification of membrane - spanning proteins. *Biochim. Biophys. Acta*, 1985, 815, p. 468–476.
11. YAMAGUCHI, K. – YU, F. – INOUYE, M. A single amino acid determinant of the membrane localization of lipoproteins in *E. coli*. *Cell*, 1988, 53, p. 423–432.
12. TATUSOV, RL. – KOONIN, EV. – LIPMAN, DJ. A genomic perspective on protein families. *Science*, 1997, 278, p. 631–637.
13. TATUSOV, RL., et al. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, 2000, 28, p. 33–36.
14. TATUSOV, RL., et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, 2001, 29, p. 22–28.

Acknowledgement: The presented study was fully supported by Ministry of Education, Youth and Sport, grant No. LN00A033.

Correspondence: Lenka Hernychová
Institute of Nuclear and Biological Medical Defence
Purkyně Military Medical Academy
Třebešská 1575
500 01 Hradec Králové
Czech Republic
e-mail: hernychova@pmfhk.cz

Received 16. 9. 2002