Discrimination of intracellular and extracellular proteins by single residue and residue-pair scores

Hiroshi Nakashima [1]
ab0011@jpnknzw1.bitnet

Ken Nishikawa [2]
nishikawa@peri.co.jp

[1] School of Allied Medical Professions,
Kanazawa University,
5-11-80 Kodatsuno, Kanazawa 920, Japan

[2] Protein Engineering Research Institute,
6-2-3 Furuedai, Suita, Osaka 565, Japan

Classification of proteins into groups is a first step to grasp the characteristics of sequences. There are many ways to classify proteins, e.g., in terms of purification procedure, component, function, structure and other criteria. Proteins are classified into "families" in the PIR database according to the degree of similarity in amino acid sequences. If classified proteins have correlation with the sequences, we might gain some insight into the general tendency. For example, membrane proteins have at least one stretch of hydrophobic residues in a sequence, so we could infer if a given protein to be a membrane protein or not by surveying a cluster of hydrophobic regions along the sequence.

Nishikawa et al. (1983) have reported that intracellular and extracellular proteins possess different amino acid compositions, and they are discernible from composition data alone. A similar distinction is observed for the cytoplasmic and extracellular domains of transmembrane proteins (Nakashima & Nishikawa, 1992). In this study, we re-examined the water soluble intracellular and extracellular proteins in terms of composition and frequencies of occurrence of amino acid pairs.

Proteins with signal peptides at the amino terminus were classified as extracellular and others were classified as intracellular. The signal peptide of an extracellular protein was excluded in the analysis. Membrane proteins were excluded from the analysis. We prepared two sets of sequence data, one was a training set to determine a parameter set of score and the other was a test set, and they were different from each other. Training set includes 894 proteins, containing 649 intracellular and 245 extracellular ones. Test set have 379 proteins, 225 intracellular and 154 extracellular proteins. The test set contains 128 proteins of known 3D structure.

[1]中島広志　金沢大学医療技術短期大学部, 〒 920 金沢市小立野 5-11-80
[2]西川　建　蛋白工学研究所, 〒 565 吹田市古江台 6-2-3

We defined single residue and residue-pair scores using composition and residue-pair frequencies, by which the type (intra- or extra-cellular) of a protein can be assigned from sequence data alone. According to the definition, a protein with a positive score is assigned as intracellular type and negative as extracellular one.

The single residue score of Met, Ile, Arg, His and Glu show a positive score implying that they prefer intracellular proteins and Cys, Trp, Asn, Ser and Tyr indicate a negative score implying that they prefer extracellular ones. The intracellular proteins are relatively rich in aliphatic (hydrophobic) as well as charged residues. Using the single residue score term, 78% of proteins in the test set were correctly identified. This is in accordance with previous work (Nishikawa et al., 1983), where the discrimination was done in the 20-dimensional composition space. As the residue-pair terms were added to the single residue term one by one starting from the nearest neighboring pair, the percentage of correctly identified proteins increased and the accuracy improved by 7% for intracellular and 9% for extracellular proteins. The percentage of proteins correctly identified by this method is 90% for the 894 training proteins and 86% for the 379 test proteins.

The reason why such difference of amino acid sequence exists between intracellular and extracellular proteins in not explained. One possible reason is the condition for extracellular proteins to be transported across the membrane lipid bilayer. Another possibility is the speed of protein folding might relate with the sequence. Nevertheless, this study shows that it is possible to infer a protein to be an intra- or a extra-cellular type.

This work is recently published (Nakashima & Nishikawa, 1994).

References

[1] Nakashima, H. & Nishikawa, K. (1992). *FEBS Lett.* 303, 141-146.

[2] Nakashima, H. & Nishikawa, K. (1994). *J. Mol. Biol.* 238, 54-61.

[3] Nishikawa, K., Kubota, Y. & Ooi, T. (1983). *J. Biochem.* 94, 997-1007.