Wanted: subcellular localization of proteins based on sequence

Frank Eisenhaber and Peer Bork

Which human proteins are destined for translocation to the extracellular space and, therefore, might be studied in more detail as easily accessible pharmaceutical targets? Which bacterial proteins are possible extracellular virulence factors? How many nuclear proteins are there in yeast? What is the proportion between extra- and intracellular proteins in bacteria or in single- or multicellular eukaryotes? Such questions are increasingly important as the attention shifts from sequencing to the interpretation of new genome data. But, surprisingly, they are very difficult to answer despite the availability of a large body of literature and database information.

Subcellular localization is a key functional characteristic of proteins. To cooperate towards the execution of a common physiological function (metabolic pathway, signal-transduction cascade, cytoskeleton, etc.), proteins must be localized in the same cellular compartment. Knowledge of subcellular localization helps the selection, from the set of genomically encoded sequences, of proteins worth being investigated more thoroughly. It also influences the design of experimental strategies for functional characterization. For example, it might be useful to check ATP binding for a cytosolic protein, whereas, in the case of extracellular proteins, proteolytic digestion assays appear much more appropriate.

Advanced database searches classify only 22% of SWISS-PROT

Today, the most successful method for functional assignment of new protein sequences is the examination of homologous proteins in sequence databases accompanied by pattern and motif searches[1]. Given the vast amount of information in DNA and protein sequence databases, the chance of finding an annotated homologue to a new protein sequence is already approaching 90% for bacteria and is increasing rapidly for eukaryotes. Cellular location can often be inferred (although sometimes only implicitly) from the annotated description of similar proteins[2,3]. However, protein function is hard to quantify for large sets of sequences. Even with the most advanced database search systems such as SRS[4] and relying on SWISS-PROT, currently the best-annotated protein identification resource[5], it is impossible to get exhaustive answers to the questions posed above. The major problem is the status of the sequence databank annotations written for reading by human experts in molecular and cell biology. Except for the keywords, the informative text is written in plain English, applying a large variety of terminology for the in-depth description of particular phenomena. For example, searches with 'intracell', 'cytoplasm', 'cytosol', 'extracell' and 'membran' only classify 22% of SWISS-PROT (release 34) into intracellular, extracellular and membrane-associated proteins. Thus, automatic assignment by homology for a primary screening of gene sequences is hampered by the detailed database annotation.

Sequences alone are not sufficient for a reliable classification

Alternatively, protein classification with respect to cellular localization can be attempted relying solely on amino acid sequence properties. Based on the small protein databases known in 1982, Nishikawa and Ooi[6] found that amino acid composition and cellular location are related. Although, as a trend, their finding still appears true today, the picture becomes blurred since many more proteins are available now[7,8]. Unfortunately, recent studies of the predictive power of amino acid compositional data for subcellular localization[9–11] were also restricted to small test sets. Other efforts concentrated on the analysis of targeting signals in protein sequences[12]. Their sequence patterns are not clear cut; therefore, the prediction accuracy is limited[13].

An intelligent annotation analyser might help

A feasible solution for classifying the whole database with respect to cellular location is to teach the computer to follow molecular and cell-biological rules and to extract conclusions from annotations not written for automatic evaluation (meta-analysis). The functional description of a protein, its catalytic or ligand-binding activity, among others, very often indicates the cellular compartment where the protein analysed is located. The relationship between annotational patterns and subcellular localizations can be encoded in a computer-readable rule for classification into intracellular, extracellular, membrane-related (transmembrane and lipid-anchored) and viral proteins.

We compiled a library of more than 1100 such rules based on a careful analysis of SWISS-PROT annotations. In the case of eukaryotes, intracellular location can be further detailed with respect to association with organelles. Some rules have a very general character, others are specific for a small family of proteins. For example, since energy-rich phosphate compounds are only available inside the cell, all proteins described with an ATPase activity, having a role in the regulation of nucleotide triphosphate concentration or requiring ATP- or GTP-binding, are assumed to have an intracellular portion. A selection of the most common problems encountered during rule formulation is presented in Box 1.

Our computer program, Meta-A(nnotator), can assign localization attributes to SWISS-PROT entries. Among the 59 021 proteins of the 34th release, 38 757 entries qualified as 'intracellular' based on automatic evaluation of the annotation, 17 131 entries have at least an extracellular portion, proteins described in 12 611 entries are membrane related and 7531 entries belong to viruses. In fact, we could derive useful location information for ~88% of all database entries. In 4868 cases where none of the cellular localizations was assigned, the protein was annotated as hypothetical. Only 2278 entries qualified as unknown, no evaluable annotation being found.

As an example, in the case of *Mycoplasma genitalium*, the only almost-complete genome in the 34th release of SWISS-PROT, Meta-A allows the evaluation of the relative proportions of intra- and extracellular proteins. This showed that 295 out of a total 425 proteins are (at least in part) intracellular, 99 are membrane related and 89 have an extracellular portion (112 are hypothetical or unknown). Only two proteins are evaluated as purely extracellular by Meta-A (MG040 and MG046); but at least the former is

The authors are at the European Molecular Biology Laboratory, Meyerhofstr. 1, Postfach 10.2209, D-69012 Heidelberg, Germany and at the Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Str. 10, 13122 Berlin-Buch, Germany.
E-mail: Eisenhaber@EMBL-Heidelberg.DE

 0962-8924/98/$19.00
PII: S0962-8924(98)01226-4

BOX 1 – EXAMPLES OF PITFALLS IN USING DATABASES FOR SEARCHING SUBCELLULAR LOCALIZATION OF PROTEINS

Hunting for information in sequence databases results in a variety of unexpected new observations and surprises. In part, they reflect the real difficulty in pressing increasingly better understood biological phenomena into an existing classification (e.g. the location of a protein can vary during its life cycle). But mistakes ranging from simple typos to dramatic factual errors[14] also happen since no human annotator is perfect.

• One entry – several contradictory locations

Since one SWISS-PROT entry might describe several versions of a given protein, an annotation-based automatic assignment of subcellular localization might result in assigning to several mutually exclusive cellular compartments. For example, the pig acyl-CoA ester carrier (P12026) is an intracellular protein, but a fragment is also annotated as an extracellular neuropeptide modulating GABA receptors and having antibacterial properties.

• One description – several possible meanings

The term periplasmic space can denote the layer between the plasmalemma and cell wall in bacteria (extracellular space) or the cytosol in centrolecithal eggs. In this case, as SWISS-PROT only uses 'periplasmic' in the former sense, it is therefore a strong indicator for extracellular localization.

• Conclusions based on incomplete or wrong annotations

The automatic annotation analyser can only find information that is contained in the entry and, since the information is scarce, it has to assume that it is correct. For example, the human CD6 precursor (P30203) is recognized only as extracellular from the SWISS-PROT entry but as having also a transmembrane part from the updated SWISS-NEW entry. The rules must also be stable against typographical errors (e.g. 'membane' or 'memrane'). Finally, annotation analysis is helpless if the annotation is completely wrong as in the case of the DAN-family proteins that are described as nuclear zinc-finger proteins[15] but, according to sequence-similarity studies (H. Hegyi and P. Bork, unpublished), appear to be cysteine-knot proteins, which are generally extracellular in localization.

likely to be lipid-anchored based on homology searches. Thus, from the sequence annotations, *M. genitalium* seems to have practically only membrane-based extracellular proteins.

The lists with the assignments of subcellular location for all SWISS-PROT entries, the rule libraries as well as future updates are available at the World Wide Web URL: `http://www.bork.embl-heidelberg.de/CELL_LOC/CELL_LOC.html`

In the future, an integration of the Meta-A assignments with SWISS-PROT (Ref. 5) and SRS (Ref. 4) is planned.

Acknowledgements

The authors thank T. Gibson and J. Schultz for helpful comments.

References

1 Bork, P. and Gibson, T. J. (1996) *Methods Enzymol.* 266, 162–184
2 Bork, P. *et al.* (1997) *FASEB J.* 11, 68–76
3 Yuan, Y. P. *et al.* (1997) *Cell* 88, 9–11
4 Etzold, T. and Argos, P. (1993) *Comput. Appl. Biosci.* 9, 59–64
5 Bairoch, A. and Apweiler, R. (1997) *Nucleic Acids Res.* 25, 31–36
6 Nishikawa, K. and Ooi, T. (1982) *J. Biochem.* 91, 1821–1824
7 Eisenhaber, F. *et al.* (1996) *Proteins* 25, 157–168
8 Eisenhaber, F., Frömmel, C. and Argos, P. (1996) *Proteins* 25, 169–179
9 Nakashima, H. and Nishikawa, K. (1994) *J. Mol. Biol.* 238, 54–61
10 Cedano, J. *et al.* (1997) *J. Mol. Biol.* 266, 594–600
11 Nakai, K. and Kanehisa, M. (1997) *Genomics* 14, 897–911
12 Claros, M. G., Brunak, S. and von Heijne, G. (1997) *Curr. Opin. Struct. Biol.* 7, 394–398
13 Nielsen, H. *et al.* (1997) *Protein Eng.* 10, 1–6
14 Bork, P. and Bairoch, A. (1996) *Trends Genet.* 12, 425–427
15 Ozaki, T. and Sakiyama, S. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 2593–2597