BIOINFORMATICS

Vol. 15 nos 7/8 1999
Pages 528–535

Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries

Frank Eisenhaber[1,2,*] and Peer Bork[1,2]

[1]Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Straße 10, 13122 Berlin-Buch and [2]European Molecular Biology Laboratory, Meyerhofstraße 1, Postfach 10.2209, D-69012 Heidelberg, Germany

Received on October 21, 1998; revised on December 10, 1998; accepted on February 23, 1999

Abstract

Motivation: *Computer-based selection of entries from sequence databases with respect to a related functional description, e.g. with respect to a common cellular localization or contributing to the same phenotypic function, is a difficult task. Automatic semantic analysis of annotations is not only hampered by incomplete functional assignments. A major problem is that annotations are written in a rich, non-formalized language and are meant for reading by a human expert. This person can extract from the text considerably more information than is immediately apparent due to his extended biological background knowledge and logical reasoning.*

Approach: *A technique of automated annotation evaluation based on a combination of lexical analysis and the usage of biological rule libraries has been developed. The proposed algorithm generates new functional descriptors from the annotation of a given entry using the semantic units of the annotation as prepositions for implications executed in accordance with the rule library.*

Results: *The prototype of a software system, the Meta_A(nnotator) program, is described and the results of its application to sequence attribute assignment and sequence selection problems, such as cellular localization and sequence domain annotation of SWISS-PROT entries, are presented. The current software version assigns useful subcellular localization qualifiers to ~88% of all SWISS-PROT entries. As shown by demonstrative examples, the combination of sequence and annotation analysis is a powerful approach for the detection of mutual annotation/sequence inconsistencies.*

Availability: *The software is available on request from Frank.Eisenhaber@embl-heidelberg.de. Results for the cellular localization assignment can be viewed at the URL http://www.bork.embl-heidelberg.de/CELL_LOC/CELL_LOC.html.*

Contact: *Frank.Eisenhaber@EMBL-Heidelberg.DE*

*To whom correspondence should be addressed:
European Molecular Biology Laboratory, Meyerhofstraße1, Postfact 10.2209, D–69012 Heidelberg, Germany.

Introduction

Wheelan and Boguski (1998) conclude their article with the heretical proposal to solve the problem of annotating the enormous amount of new genomic data just by eliminating most of the archived functional annotation, since, in their belief, the functional features can easily be recalculated with the existing sequence comparison and structure prediction programs. They are certainly right if the value added to the sequence consists simply of the inherited annotation of the best matching database sequence found in a BLAST search (Smith and Zhang, 1997; Bork and Koonin, 1998). On the other hand, Smith (1998) and, as we believe, an increasing majority of researchers, consider the nucleic acid and protein sequences only as raw material for biological research and their functional annotation as a non-trivial and, perhaps, an even more valuable complement of the sequence information which is composed of (i) experimental data with respect to biochemical functions, expression profiles, cellular and physiological impact, (ii) computed structural and functional predictions, (iii) literature references as well as (iv) human expert input.

Traditionally, the sequence part of a database entry is prepared for access by computer software and efficient programs for collecting families of similar sequences are available (Madden *et al*., 1996; Pearson, 1996; Altschul *et al*., 1997). In contrast, the retrieval of groups of genes with a related functional description in the database is a much more difficult task (Bork *et al*., 1998). The annotation section of an entry is mostly written in plain English, with a rich biological vocabulary that often varies in different areas of research. The texts are intended to be read by the specialized human researcher and are not well structured for computer-aided evaluations.

The most simple and the only generally available computer-based annotation analysers are keyword searching engines as implemented in SRS (Etzold *et al*., 1996) or ENTREZ (Schuler *et al*., 1996). Combined with pre-indexing of information units (keywords or short phrases) in databases, this approach allows fast retrieval of entries. Lexical analysis systems based on keyword analysis have found many applications in biomolecular information processing. The GENXREF system (Achard and Barillot, 1998; Achard and Dessen, 1998) compares the occurrence of keywords in entries of two different databases and, in cases of significant similarity, allows the automatic generation of inter-database links between entries supposed to be related by their functional annotation (database VIRGIL of links between GDB genes and Genbank sequences). Guigo (Guigo *et al*., 1991, 1993; Guigo and Smith, 1993) developed a tool for determining the most characteristic subset of keywords for the biological function of a protein family from their database annotation that can be inherited to uncharacterized members of the family. Andrade and Valencia (1995, 1998) addressed a similar question by analysing a set of MEDLINE abstracts.

The disadvantages of pure keyword searching approaches are 2-fold. First, the keyword context is lost and more complex semantic units cannot be retrieved. Second, functional annotation in the form of keywords creates the problem of categorized description of gene function at the molecular, cellular (e.g. organelle localization, involvement in metabolic pathways, signal transduction cascades, structural associates, and the like) and phenotypic levels with controlled vocabularies (Bork *et al*., 1998; Riley, 1998). A limited number of categories have been proposed and applied for the description of catalytic function (Overbeek *et al*., 1997), of protein functions in SWISS-PROT (Bairoch and Apweiler, 1998) and in FLYBASE (www.ebi.ac.uk:7081/docs/flydocs/flybase/controlled-vocabularies.txt, ftp.ebi.ac.uk/pub/databases/edgp/misc/ashburner/fly_function_tree). Such systems require not only an enormous discipline from database curators, but are also a constant source of problems and database inaccuracies as biological understanding improves. New categories will arise which are missing in the vocabulary, the meaning of other notions will shift with time, and the same keyword may become used with different meanings in older and newer database entries. Large-scale updates of database annotations are uncommon due to the huge effort required.

It should be noted that an expert reads an annotation in a greatly different manner compared with a keyword searching engine. Most importantly, he/she extracts considerably more information from the text than is actually written there due to an extended biological background knowledge and logical reasoning. In this work, we attempt to simulate this approach by combining lexical annotation analysis with the use of biological rule libraries. The prototype of a software system, the so-called Meta_A(nnotator) program, is described and the results of its application to sequence attribute assignment and sequence selection problems such as cellular localization and sequence domain annotation are presented.

Methods: The basic algorithm

Dissection of the database annotation into token-oriented semantic units

The information in sequence annotations within a database entry is organized in semantic units labelled by tokens. As an illustration, we consider the SWISS-PROT database (Bairoch and Apweiler, 1998) which is one of the best annotated protein sequence databases to date. For example, the 'DE' token is followed by the protein name, the 'KW' line contains keywords (words or small groups of words from a predefined list), the 'AC' precedes accession numbers, etc. Some tokens are subdivided by secondary tokens, e.g. the comments 'CC' (secondary tokens as '-!- PTM:', etc.) and the feature table 'FT' (secondary tokens as 'VARSPLIC', etc.). As a first step, the complete text associated with a given token which is possibly distributed over several lines has to be extracted from the annotation as a continuous string.

Application of token-specific biological rules

Each specific token is expected to be accompanied by a certain type of information. This is indeed generally the case, but not always followed. Who would await genetic information following the set of token and secondary token 'CC -!- SUBCELLULAR CLASSIFICATION:' as in the yeast fumarase description (FUMH_YEAST, P08417, information about nuclear localization of the gene is supplied, the protein itself is, of course, not a nuclear one)? Nevertheless, the variety of possible texts behind given tokens is relatively limited. Therefore, it is reasonable to conclude from the occurrence of some typical complex lexical pattern (to be described as a regular expression) in the text associated with a token that the protein in the entry has some property. The result of this deduction is called the primary attribute for the database entry. A biological rule of this type can be written in logical notation ($\wedge$ = ***and***, $\exists$ = ***exist***, $\rightarrow$ ***implication***) as

$$\textbf{\textit{token}}\ (\wedge\ \textbf{\textit{secondary token}})_{\textbf{\textit{optional}}} \wedge (\exists \textbf{\textit{pattern}}_i) \rightarrow \textbf{\textit{primary attribute}}_j \qquad (1)$$

associating a given lexical pattern ***i*** with the attribute ***j***. For example, the text unit 'CARTILAGE PROTEIN' in the text associated with token 'CC' and the secondary token '-!- FUNCTION:' allows one to infer that the protein is extracellularly located (=***primary attribute***$_j$).

Sometimes, it is even not necessary to analyse the text associated with a token since the existence of the token itself already implies the primary attribute. To illustrate, the token

'FT SIGNAL' implies the primary attribute 'secreted' as being valid for the given protein:

$$\textit{token}\ (\wedge\ \textit{secondary token})_{optional} \rightarrow \textit{primary attribute}_j \quad (2)$$

Deductive logics with primary attributes and generation of the final functional characteristics

After scanning the complete annotation, all possible primary attributes have been determined and are subjected to a second round of deduction logics. The biological rules applied here combine primary attributes with logical operators AND, OR and NOT for the determination of secondary attributes, e.g.:

$$\textit{primary attribute}_{i1} \wedge \textit{primary attribute}_{i2} \rightarrow \textit{secondary attribute}_j \quad (3)$$

As an illustration, a protein described as 'ANTIGEN' in the DE-line (= ***primary attribute***$_{i1}$) and as 'IMPLICATED IN.* PARASITE INVASION' ('.' implies any character and '*' any number of its repetitions in a regular expression) in the 'CC -!- FUNCTION:' description (= ***primary attribute***$_{i2}$) can also be considered as extracellularly located (= ***secondary attribute***$_j$). By the way, the characterization as antigen alone proved not sufficient to justify this conclusion (Godelaine *et al*., 1993), e.g. in the case of the cystic fibrosis antigen calgranulin A described in entry S108_HUMAN (P05109).

As another example, a protein of the respiratory chain is membrane-related in prokaryotes, but both membrane-related and mitochondrial in eukaryotes. In our implementation, both attributes are generated at the primary level and, at the secondary level, the qualifier 'mitochondrial' is removed for non-eukaryotic organisms.

Since the successive application of various implication rules might not be commutative, it reasonable to introduce some order. In our implementation, we start with all rules producing new attributes and only then apply rules that eliminate existing attributes. The remaining attributes are the newly generated functional characterizations for the sequence annotation studied.

Implementation: Meta_A(nnotator)

We have written a general computer program Meta_A(nnotator) in the C-language for the evaluation of human-readable annotations in token-oriented databases with exchangeable rule libraries. Our software prototype requires the following input which is specific for each application:

1. A description file of the database structure (e.g. of SWISS-PROT and SWISSNEW).
2. Files containing molecular and cell biological rules relating annotational patterns (described in the form of regular expressions) and functional attributes.
3. Files with logical combination rules for the analysis at the secondary level.

It should be emphasized that the computer program is a general tool for any token-oriented database but the ASCII-readable files (points 1–3) describing the database structure and the rule libraries are specific for each application. Whereas the computer program (~10 000 lines of C-code) was completed within 2 months, it took about another 2 months to produce the first version of the application-specific ASCII-readable files for the task of cellular localization or ~3 weeks for the domain assignment (see below). Thus, some work is required before the system can be used for a new application. The creation of the token-specific rule files relating lexical patterns with functional attributes (point 2) is the most time-consuming step in the preparation of the software system for a given application. Therefore, the following extremely simple notation of a rule was selected:

```
>ABCD    regular_expression
```

Here, '>' opens a line containing a rule. This sign is followed by all primary attributes encoded in the form of single letters, e.g. A, B, C and D. The remaining non-white part of the line is considered the lexical pattern written in the form of a regular expression.

At present, it takes Meta_A(nnotator) less than 45 min to generate cellular localization assignments for 74 019 entries of the SWISS-PROT Release 36 on an O2 SGI workstation when accessing the database over the local net, i.e. on average ~36 ms for a single entry.

Results

Sorting with respect to subcellular localization

The original impetus for this work was given by the seemingly simple task of sorting all proteins from SWISS-PROT with respect to cellular localization. It was decided to use homology relationships to infer the localization information of uncharacterized proteins (Bork *et al*., 1997; Yuan *et al*., 1997). Automatic, computer-aided selection methods are an important way to identify attractive target proteins among the haystack of new gene sequence data. One of the helpful decision criteria is the probable subcellular localization of the gene products (Eisenhaber and Bork, 1998). For example, extracellular proteins are good candidates in a search for virulence factors of pathogenic bacteria or for easily accessible entry points for pharmaceutical drugs, while proteins at other subcellular locations may, at the beginning, not be considered for such purpose.

However, a simple keyword analysis proved not sufficient to solve the sorting task. For example, only 20 283 out of the total 59 011 entries of Release 34 of SWISS-PROT entries have a commentary 'SUBCELLULAR LOCATIONS'. An SRS database request (Etzold *et al*., 1996) with the search

patterns 'cytoplasm', 'extracell' and 'membran' revealed just 4081 cytoplasmic, 1179 extracellular and 8792 membrane proteins, respectively, classifying a total of 22% of the database (some entries received several qualifiers).

Therefore, the Meta_A(nnotator) software and a localization rule library have been created. As final (secondary level) localization attributes, the following subcellular compartments were allowed: 'intracellular', 'membrane-related' (with transmembrane regions, with lipid anchors, or located nearby the membrane), 'extracellular', 'viral'; in the case of eukaryotic intracellular proteins, we also included 'nuclear', 'mitochondrial', 'chloroplast' and 'ER/Golgi'. To obtain at least some localization information for even tersely annotated entries, it was decided to allow any suitable, even indirect hint in the annotation to generate the attributes 'intracellular', 'membrane-related' and/or 'extracellular'. By contrast, we attempt to assign the remaining attributes as well as the qualifiers 'cytoplasmic' and 'transmembrane' in a restrictive manner with high reliability. As a rule, these localizations have been explicitly described in the annotation.

It also appeared practical to introduce the attribute 'HYPOTHETICAL' if the protein is described as hypothetical and no other useful localization assignment could be made. The program outputs 'UNKNOWN' for entries if none of the 11 attributes listed above is applicable.

In the newest version of the software package, 12 token-specific libraries of biological rules contain >1100 types of implications relating more or less complex lexical patterns with ~20 primary attributes. The database texts associated with protein names, taxonomy information, commentaries and the feature table are scanned. About 30 deduction rules for the secondary level of treatment determine the applicable subset among the allowed 12 final functional attributes.

Of course, we attempted to formulate rules that have a very general character. This was not always possible, some rules are specific for a small family of proteins. At the same time, it was attempted to avoid the formulation of rules valid for single entries only. For the sake of illustration, we describe a few conclusion types. (i) Generally, energy-rich phosphate compounds are available only inside the cell. Therefore, all proteins described with ATP/GTPase activity, having a role in the regulation of NTP concentration or requiring ATP or GTP binding, are assumed to be intracellular proteins. Similarly, proteins important for the biosynthesis of small organic compounds such as amino acids, nucleotides, and the like, are considered intracellular. (ii) Gene products with lipid anchors are generally membrane bound (except for myristylated proteins which are also observed in the cytosol), whereas GPI-anchored lipoproteins are thought to populate the extracellular surface of the plasmalemma, proteins with palmitate, farnesyl or geranylgeranyl modifications occupy positions on intracellular membranes or on the cytoplasmic side of the plasmalemma. A 'TYPE I MEMBRANE PROTEIN' was assumed as having intracellular, extracellular and membrane-related parts. The localization 'extracellular' is removed if the annotation reveals the involvement in an intracellular organelle. (iii) In SWISS-PROT, the extracellular localization is often described in a very detailed way. Simple cases are bone or cartilage proteins, a more specific version is 'IN THE AQUEOUS FLUID SURROUNDING OLFACTORY SENSORY DENDRITE'. Each such instance will generate a new rule.

The rules have been extensively checked against the database to avoid ambiguous or erroneous attribute assignments. For >4000 SWISS-PROT entries selected randomly or as affected by a given biological rule, the computer-generated assignments have been manually compared with the database annotation. The biological rule system has been refined to achieve only correct automatic assignments in all such cases (0% wrong assignments). This does not mean that a human expert produces only the computer-generated attributes, she or he might find more and more specific qualifiers based on the given annotation. A rule was finally accepted only if we could not find an entry with a computer-generated assignment that is in contradiction with the database annotation and/or with biological sense. Three aspects ease this task considerably: (i) many SWISS-PROT entries contain uniform text units in their annotation; (ii) each rule affects only a subset of entries, which is often small and can be analysed within reasonable amounts of time by a human expert; (iii) each entry is affected only by a few rules. Thus, changing one rule does not have dramatic effects on the whole system.

The Meta_A(nnotator) system has been in use in the laboratory for more than a year now; all discrepancies between software assignments and database annotations observed by group members resulted in subsequent rule corrections until the problem was removed. In each case, all entries affected by the rule change have been automatically subselected and manually checked. Thus, the rule system is being refined under the influence of constant expert advice. We also used sequence analysis with some cell compartment-specific sequence domain profiles to cross-check the annotation analysis independently on a large scale (see below).

We found a considerable number of entries for which automatic localization assignment is difficult (Eisenhaber and Bork, 1998). Since one SWISS-PROT entry may describe several versions of a given protein, an annotation-based automatic assignment of subcellular localization may result in a list of several, each other excluding cellular compartments. For example, the thymopoietin entries THPA_HUMAN (P42166) and THPB_HUMAN (P42167) contain data on different versions of the protein with alternatively intra- or extracellular localization. It is also difficult to make a computer-aided distinction between extracellular localization and the annotation of intracellular compartments such as endoplasmic or sarcoplasmic reticulum, Golgi apparatus, lyso-

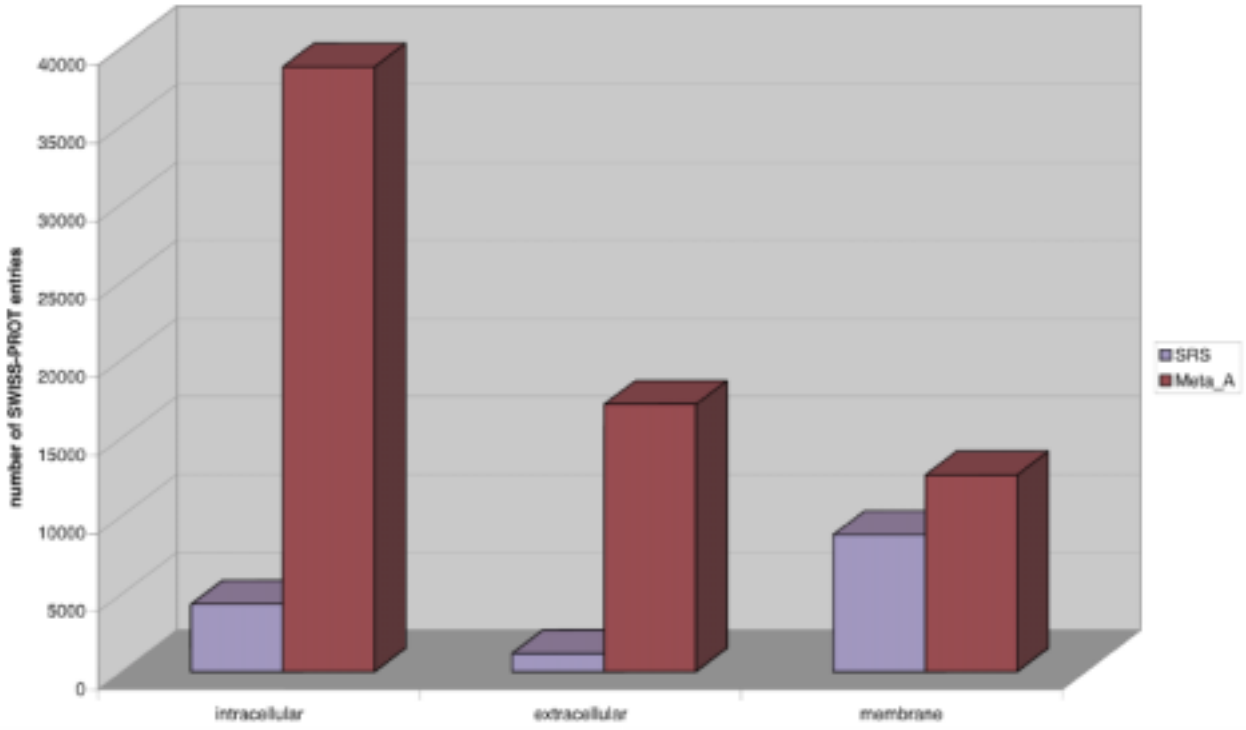

Fig. 1. The difference in the assignment efficiency of the categories 'intracellular', 'extracellular' and 'membrane-related' to SWISS-PROT entries (Release 34) with simple keyword searching engines (example SRS) and with a rule library-based approach [example Meta_A(nnotator)] is shown. In total, keyword searching assigns localization attributes to 22% of all entries, whereas rule libraries can achieve 88%. Clearly, membrane relationship is mostly described with a controlled vocabulary in SWISS-PROT and, therefore, rule libraries give less relative improvement in this case than for the other two categories.

some, vacuole, peroxisome, endosome or 'GRANULES INSIDE THE CELL WHICH MAY BE FINALLY SECRETED'. For some proteins, cellular localization changes during the life cycle. Thus, there are always entries for which the categorization is not fully adequate.

The different usage of biological terms in a given database may also create problems. Originally, we thought 'DNA-binding' indicated nuclear proteins in eukaryotic organisms, but it was found that some clearly extracellular proteins (e.g. the fibronectins of the FINC SWISS-PROT entries) also have DNA-binding activity under certain conditions. Whereas the keyword 'STORAGE PROTEIN' labels proteins localized extracellularly or in vacuoles and the endoplasmic reticulum throughout SWISS-PROT, the text pattern 'SEED STORAGE PROTEIN' is used purely for intracellular proteins.

In Table 1, we present the results of the annotation analysis of SWISS-PROT releases 34, 35 and 36 with respect to cellular localization. The results show clearly that the Meta_A(nnotator) system increases the total number of entries with useful assignments ~4-fold compared with simple keyword search systems (see also Figure 1). A manual check of a few hundred database entries qualified as unknown revealed that the overwhelming majority of those have such a terse annotation that even a human expert cannot decide about their cellular compartment (often no annotation at all, except for species taxonomy and authors). Thus, the rule system for localization approached some upper level; the addition of more new rules will hardly increase the assignment efficiency to any essential degree as long as no new entries with qualitatively new functionality (and, therefore, new lexical description) are included in the database.

At the same time, the rate of useful assignments decreases slowly with increasing release number (by 1% from Release 34 to Release 36). There are several reasons for this. First, proteins with completely new functional descriptions may enter the databases and these annotation types have not been included in the rule libraries. Fortunately, this does not happen very often. In such rare cases, periodic updates of the rule libraries will help. A second problem is much more serious. The fraction of extremely poorly annotated protein sequences is steadily increasing (e.g. 9.2% hypothetical proteins in Release 36 compared with 8.2% in Release 34). Whereas previously sequencing was the final step after a de-

tailed functional characterization of proteins, now genome projects produce large quantities of gene sequences not accompanied by publications of biochemical or molecular biological studies.

The corresponding rule libraries for cellular localization will receive support in the near future and will be regularly updated. Results for the localization assignment of SWISS-PROT entries and some META_A(nnotator) software information can be viewed at the URL http://www.bork.embl-heidelberg.de/CELL_LOC/CELL_LOC.html.

Table 1. Automatic analysis of annotations in the SWISS-PROT database with respect to cellular localization of proteins. Some entries have received several attributes, therefore the number of all attributes is larger than the total number of entries. The attributes 'HYPOTHETICAL' and 'UNKNOWN' are mutually exclusive and are also not compatible with any useful assignment

	Release 34	Release 35	Release 36
Total number of database entries	59 021	69 113	74 019
Intracellular	38 757	45 872	49 553
Membrane related	12 611	15 235	16 525
Extracellular	17 131	20 053	21 353
Cytoplasmic	7306	8516	9470
Transmembrane	8792	10 879	11 993
Mitochondrial	2917	3370	3860
Chloroplast	2772	2960	3133
Nuclear	6169	6600	7094
ER/Golgi	701	851	909
Viral	7531	7790	7908
Useful assignments	87.9 %	87.0 %	86.9 %
HYPOTHETICAL	4868	6361	6793
UNKNOWN	2278	2629	2921

Annotation of sequence domains and annotation/sequence inconsistencies

The SMART system (Schultz *et al.*, 1997) is both a collection of profiles for a large number of mobile protein domains in signalling and extracellular proteins, as well as a search tool for such domains in query sequences. To test the sensitivity and selectivity of the profiles, as well as the annotation quality in SWISS-PROT, it was decided to compare the possible domain annotations of proteins with the hits of the SMART tool in SWISS-PROT. A library of biological rules encoding possible types of domain descriptions based on protein names, commentaries, PROSITE (Bairoch *et al.*, 1997) links and feature table texts has been created, and is being constantly updated. At the moment, ~250 rules are used for analysing the annotation of ~80 types of sequence domains. In the near future, ~200 domain types will be covered. It must be said that this annotation analysis problem is much simpler than the localization analysis since domain description is more explicit and, therefore, the secondary deduction step does not need such a level of sophistication.

The results of this automatic database analysis are useful in several aspects. First, we found a few domains for which the SMART profiles did not recognize all SWISS-PROT representatives. For example, the profile for the GLA domain (hyaluronan-binding domain containing γ-carboxylate residues) did not find a whole subgroup of 27 sequences. This was a strong indication that such sequence families had to be reinvestigated and the corresponding profiles updated. In this way, there is an independent, automatic control mechanism for the SMART sequence analysis system. Second, non-annotated domain hits can be selected for further scientific analysis.

Since some domains are specific for certain types of cellular localization, the domain assignment by annotation, and especially by sequence analysis through the SMART tool, can also be used to cross-check the automatic cellular localization assignments since domain information was not used for the rule library applied in the cellular compartment recognition. Manual inspection of possible discrepancies allows correction of the rule libraries or the identification of entries with possible annotation or sequence inaccuracies.

Several interesting contradictions have been found. Typically, inconsistent use of terminology creates annotation inaccuracies. A 'growth factor' is per definition a polypeptide hormone regulating cell division and as such a mainly extracellular molecule (Lackie and Dow, 1995). The fibroblast growth factors (FGF) of human (FGFC_HUMAN, Q92912; FGFE_HUMAN, Q92915) and mouse (FGFB_MOUSE, P70378; FGFE_MOUSE, P70915) are annotated both as growth factors and as nuclear proteins. As the SMART tool revealed, all these proteins have the FGF domain which is a strong marker for extracellular localization. As another example, the glia maturation factors of human (GLMB_HUMAN, P17774) and rat (GLMB_RAT, Q63228) are also described as growth factors (pointer to extracellular localization), but, as sequence analysis showed, contain an actin depolymerization factor/cofilin-like (ADF) domain typical for proteins of the cytoskeleton.

Similarly, an oncogene is defined as a mutated and/or overexpressed version of a normal gene (= proto-oncogene) that in a dominant fashion can release the cell from normal restraints on growth, and thus alone, or in concert with other changes, converts a cell into a tumour cell. Proto-oncogenes are generally involved in signalling and regulation of cell growth, and as such are generally intracellular (Lackie and Dow, 1995). The NOV-proteins (NOV_CHICK, P28686; NOV_COTJA, P42642; NOV_HUMAN, P48745; NOV_MOUSE, Q64299) are all described as (proto-) oncogenes, but are mosaic proteins composed of four typically

extracellular sequence domains (Bork, 1993): IB (insulin growth factor-binding protein homologues), VWC (von Willebrand factor type C), TSP1 (thrombospondin type I repeat), CT (C-terminal module in matrix proteins). Either the understanding of a proto-oncogene has to be widened or the annotation is not fully adequate.

Sometimes, the contradiction points to a possible sequence inaccuracy. The hypothetical homeobox protein C02F12.5 of *Caenorhabditis elegans* (YL15_CAEEL, Q11101) is annotated as nuclear protein. It contains both a homeobox domain (in agreement with nuclear localization) as well as a Kunitz domain which is typical for extracellular proteins. Perhaps, there is an error in the DNA sequence assembly or in the gene annotation, resulting in fusion of two independent proteins.

Discussion

The development of a rule system-based automatic evaluator is an alternative to complete updates of sequence database annotations. The latter are very labour intensive and are scheduled with large time delays or may even never happen. Also, it cannot be expected that sequence database annotations will be constructed with a controlled vocabulary that contains keywords for all possible research questions. Therefore, the creation of libraries with biological rules might be the method of choice if researchers need the answer to complex database sorting and entry subselection problems in a reasonable time scale.

Moreover, rule systems may work well with a historically grown database and can create the appearance of a virtual annotation for the user that has never explicitly existed. In this case, the inertia of database development is a favourable circumstance since (i) rule libraries do need not to be heavily updated and (ii) different rules may be applied for various creation dates of database entries, i.e. it is possible to follow the shift in understanding of biological notions in this way. For example, the domain annotation of SMART Release 1.03 (Schultz *et al*., 1997) outputs automatically the cellular localization for all hits found in SWISS-PROT so that the user can conclude immediately about the typical cellular compartment for the corresponding sequence family.

It should be noted that an automatic annotation analyser, especially such a simple one as described here, will never be able to compete with a human expert in evaluating a given entry. In some cases, the automatically generated conclusion may be inaccurate and require refinement both because of the limited number of categories in the classification and the incompleteness or simplicity of biological rules. It should also be noted that most annotations are so scarce that the automatic annotation analyser has to assume that all information is correct, whereas a human expert with a more complex background and deduction scheme is often able to determine inconsistencies and to find the probably more correct answer. However, in the speed of evaluating thousands of entries, the automatic procedure is unbeatable and as such is a valuable tool for treating large amounts of data, the typical situation encountered in genome analyses.

Acknowledgement

The authors are grateful to Jörg Schultz for many suggestions that resulted in refinements of the biological rule libraries.

References

Achard,F. and Barillot,E. (1998) VIRGIL: a database of rich links between GDB and GenBank. *Nucleic Acids Res*., **26**, 100–101.

Achard,F. and Dessen,P. (1998) GenXref VI: automatic generation of links between two heterogeneous databases. *Bioinformatics*, **14**, 20–24.

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Andrade,M.A. and Valencia,A. (1995) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *ISMB*, **5**, 25–32.

Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

Bairoch,A. and Apweiler,R. (1998) The SWISS-PROT protein sequence databank and its supplement TrEMBL in 1998. *Nucleic Acids Res*., **26**, 38–42.

Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status and progress. *Nucleic Acids Res*., **25**, 217–221.

Bork,P. (1993) The modular architecture of a new family of growth regulators related to connective tissue growth factor. *FEBS Lett*., **327**, 125–130.

Bork,P. and Koonin,E.V. (1998) Predicting function from protein sequence: Where are the bottlenecks? *Nature Genet*., **13**, 313–318.

Bork,P., Hofmann,K., Bucher,P., Neuwald,A.F., Altschul,S.F. and Koonin,E.V. (1997) A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J*., **11**, 68–76.

Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol*., **283**, 707–725.

Eisenhaber,F. and Bork,P. (1998) Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol*., **8**, 169–170.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*., **266**, 114–128.

Godelaine,D., van Pel,A. and Beaufay,H. (1993) Presentation of mouse tum- P91A antigen from chimeric proteins with different subcellular localisations by class I molecules of the major histocompatibility complex. *Eur. J. Immunol*., **23**, 1727–1730.

Guigo,R. and Smith,T.F. (1993) Inferring correlation between database queries: Analysis of protein sequence patterns. *IEEE Trans. Patt. An. Mach. Learn*., **15**, 1030–1041.

Guigo,R., Johansson,A. and Smith,T.F. (1991) Automatic evaluation of protein sequence functional patterns. *Comput. Appl. Biosci*., **7**, 309–315.

Guigo,R., Vazquez,I., Rao,S. and Smith,T.F. (1993) A protein sequence database cross-field association system. *Proc. 26th Haw. Int. Conf. System Sci.: Biotech*., **1**, 822–833.

Lackie,J.M. and Dow,J.A.T. (1995) *The Dictionary of Cell Biology*. Academic Press, London.

Madden,T.L., Tatusov,R. and Zhang,J. (1996) Applications of network BLAST server. *Methods Enzymol*., **266**, 131–141.

Overbeek,R., Larsen,N., Smith,W., Maltsev,N. and Selkov,E. (1997) Representation of function: the next step. *Gene*, **191**, GC1–GC9.

Pearson,W.R. (1996) Effective protein sequence comparison. *Methods Enzymol*., **266**, 227–259.

Riley,M. (1998) Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol*., **8**, 388–392.

Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: Molecular biology database and retrieval system. *Methods Enzymol*., **266**, 141–162.

Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1997) SMART, a simple modular architecture research tool: Identification of signalling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.

Smith,T.F. (1998) Functional genomics—bioinformatics is ready for the challenge. *Trends Genet*., **14**, 291–293.

Smith,T.F. and Zhang,X. (1997) The challenges of the genomic sequence annotation or ‘The devil is in the details’. *Nature Biotechn*., **15**, 1222–1223.

Wheelan,S.J. and Boguski,M.S. (1998) Late-night thoughts on the sequence annotation problem. *Genome Res*., **8**, 168–169.

Yuan,Y.P., Schultz,J., Mlodzik,M. and Bork,P. (1997) Secreted fringe-like signaling molecules may be glycosyltransferases. *Cell*, **88**, 9–11.