Rolf Apweiler
heads the SWISS-PROT, TrEMBL and InterPro database activities at the EMBL Outstation – European Bioinformatics Institute.

Functional information in SWISS-PROT: The basis for large-scale characterisation of protein sequences

Rolf Apweiler

Date received (in revised form): 9th November 2000

Abstract

With the rapid growth of sequence databases, there is an increasing need for reliable functional characterisation and annotation of newly predicted proteins. To cope with such large data volumes, faster and more effective means of protein sequence characterisation and annotation are required. One promising approach is automatic large-scale functional characterisation and annotation, which is generated with limited human interaction. However, such an approach is heavily dependent on reliable data sources. The SWISS-PROT protein sequence database plays an essential role here owing to its high level of functional information.

Keywords: *bioinformatics, protein sequence, function annotation, database, automation*

INTRODUCTION

Although the first complete sequence of an organism was determined some 22 years ago, the 5-kilobase sequence of the bacterial virus phi-X174 achieved by Sanger in Cambridge,[1] it is only in the last few years that the technology of sequencing has developed to the stage that the sequencing of the complete genome of a living organism can be contemplated as a practical and routine possibility. A major breakthrough was the sequencing of the first complete eukaryote chromosome, chromosome III of *Saccharomyces cerevisiae*, in 1992 by an EU-funded consortium.[2] In 1995 the Institute of Genome Research (TIGR) group published the first complete sequence of a bacterial genome, that of *Haemophilus influenzae*.[3]

Since those dramatic events the complete sequences of more than 30 bacterial genomes have been published and at least 70 more are known to be nearing completion. Not only has the complete sequence of *S. cerevisiae* been achieved,[4] but so has that of the nematode worm *Caenorhabditis elegans*[5] and of the fruitfly *Drosophila melanogaster*,[6] and of the plant *Arabidopsis thaliana*, while the sequences of the yeast *Schizosaccharomyces pombe* and the sequences of several important protozoan parasites are well towards completion. In addition the complete genomes of many mitochondria and plastids have been determined. Large-scale sequencing of the genome of the laboratory mouse is well underway, in both the USA and Europe. The 'Holy Grail' of large-scale sequencing is, however, the determination of the sequence of the human genome, estimated at 3 billion base-pairs. The completion of the 'first draft' of this sequence was announced on 26th June 2000.

All these projects produce large amounts of sequence data, lacking experimental determination of the biological function. To cope with such large data volumes, faster and more effective means of creating functional annotation are required. One promising approach is automatic annotation, which is generated with limited human interaction.

Several solutions of automatic functional characterisation of unknown proteins are based on high-level sequence similarity searches against known proteins. Other methods collect the results of different prediction tools in a simple[7] or more elaborate[8] manner. However, the

Rolf Apweiler,
EMBL Outstation – European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton,
Cambridge CB10 1SD, UK

Tel: +44 (0) 1223 494435
Fax: +44 (0) 1223 494468
e-mail: apweiler@ebi.ac.uk

currently used solutions have several drawbacks, such as the following:

- Since many proteins are multifunctional, the assignment of a single function, which is still common in genome projects, results in loss of information and outright errors.
- Since the best hit in pair-wise sequence similarity searches is frequently a hypothetical protein or poorly annotated or has simply a different function, the propagation of wrong annotation is widespread.
- There is no coverage of position-specific annotation such as active sites.
- The annotation is not constantly updated and is thus quickly outdated.

pitfalls of automatic annotation

It is also important to emphasise that a single sentence describing some predicted properties of an unknown protein should not be regarded as full annotation, but rather more as an attempt to characterise a protein. Full annotation means the combination of extracting experimentally verified information from the literature with sequence analysis to add as much reliable and up-to-date information as possible about properties such as function(s) of the protein, domains and sites, catalytic activity, cofactors, regulation, induction, subcellular location, quaternary structure, diseases associated with deficiencies in the protein, the tissue specificity of a protein, developmental stages in which the protein is expressed, pathways and processes in which the protein may be involved, and similarities to other proteins.

THE ANNOTATION CONCEPT OF SWISS-PROT AND TrEMBL

The SWISS-PROT protein sequence database[9] strives to provide extensive annotation as defined above. However, owing to the increased data flow from genome projects to the sequence databases SWISS-PROT faced a number of challenges to its time- and labour-intensive way of database annotation. Maintaining the high quality of sequence and annotation in SWISS-PROT requires careful and detailed annotation of every entry with information retrieved from the scientific literature and from rigorous sequence analysis. This is the rate-limiting step in the production of SWISS-PROT. On one hand it is desirable to keep the high editorial standards of SWISS-PROT. But that means that there is a limit to how much the annotation procedures can be accelerated. On the other hand, it is also vital to make new sequences available as quickly as possible. To address this concern, the European Bioinformatics Institute (EBI) introduced in 1996 TrEMBL (translation of EMBL nucleotide sequence database). TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDS) in the EMBL database, except for CDS already included in SWISS-PROT.

To enhance the annotation of uncharacterised protein sequences in TrEMBL, the SWISS-PROT/TrEMBL group at the EBI developed a novel method for the prediction of functional information.[10] This method selects proteins in the SWISS-PROT protein sequence database, which belong to the same group of proteins as a given unannotated protein, extracts the annotation shared by all functionally characterised proteins of this group, and assigns this common annotation to the unannotated protein.

AUTOMATIC ANNOTATION OF TrEMBL

To implement this methodology for the automated large-scale functional annotation of proteins three major components are required. First of all, a reference database must serve as the source of annotation. SWISS-PROT is used as the reference database because of its highly reliable, well-annotated and standardised information.

automated large-scale functional annotation of TrEMBL protein sequence database records

Furthermore, a highly diagnostic protein family signature database must supply the means to assign proteins to groups. For this purpose PROSITE[11] was initially used, but since the beginning of 2001 InterPro[12] has been used to increase coverage and enhance reliability (reduction of false-positives and false-negatives). This will mean coverage will be increased and reliability (reduction of false-positives and false-negatives) enhanced. InterPro is a new integrated resource for protein families, domains and functional sites, developed initially as a means of rationalising the complementary efforts of the PROSITE, PRINTS,[13] Pfam[14] and ProDom[15] databases. By uniting these databases, it was possible to capitalise on their individual strengths, producing a single entity that is far greater than the sum of its parts. The use of InterPro allows the reliable classification of proteins into families and the recognition of the domain structure of multidomain proteins. InterPro can classify currently around 60 per cent of all known protein sequences and this information is incorporated into SWISS-PROT and TrEMBL in the form of database cross-references to InterPro and its member databases.

The final component needed for the implementation of our automated large-scale functional annotation methodology is a database (RuleBase) that stores and manages the annotation rules, their sources and their usage.

The actual flow of information during the automatic annotation can be divided into five steps:

- Use InterPro to extract the information necessary to assign proteins to groups ('conditions') and store the conditions in the RuleBase.
- Group the proteins in SWISS-PROT by the conditions.
- Extract from SWISS-PROT the common annotation shared by all functionally characterised proteins of each group and store this common annotation together with its conditions in the RuleBase. Now every rule consists of conditions and the annotation common to all proteins of this group characterised by these conditions.
- Group the unannotated TrEMBL entries by the conditions stored in the RuleBase.
- Add the common annotation to the unannotated TrEMBL entries. The predicted annotation will be flagged with evidence tags, which will allow users to recognise the predicted nature of the annotation as well as the original source of the inferred annotation.

As the reliability of the conditions is crucial to the reliability of the methodology, a multiple-step procedure is used to minimise false positive automatic annotation:

- The InterPro database used to extract conditions to assign proteins to groups integrates different computational techniques for the recognition of signatures diagnostic for different protein families or domains. All the different approaches integrated in InterPro (hidden Markov models (HMMs), Profiles, Fingerprints, Regular Expressions, etc.) have different strengths and weaknesses. The combination of the strengths of the different signature recognition methods, coupled with statistical and biological significance test, allows the various drawbacks of the individual methods to be overcome.
- An important condition in every rule is that the taxonomic classification of the unannotated protein sequences must be within the known taxonomic range of the experimentally characterised proteins. For instance, a match of an *a priori* prokaryotic signature against a human protein is regarded as violating

the conditions of the rule for this protein family. It is thus considered as false positive and filtered out.

- In cases where a protein family is characterised by more than one signature in InterPro, all signatures must be found in the unannotated protein sequence. For instance, bacterial rhodopsins have a signature for a conserved region in helix C and another signature for the retinal binding lysine. If an unannotated protein sequence matches the helix-C -pattern, but not the retinal-binding pattern, it will not be regarded as a bacterial rhodopsin.

The automation of functional annotation is of paramount importance to mine the avalanche of sequence data. Our approach for the second generation of automated annotation will hopefully overcome some limitations of the existing automatic annotation methods:

improving the quality of automated functional annotation

- By using only the annotation from a reliable reference database for our predictions, the propagation of wrong annotation, one of the big problems in functional annotation,[16] will be drastically reduced.

- By using the 'common annotation' of multiple entries, the implemented methodology will produce significantly fewer over-predictions than methods based on the best hit of a sequence similarity search.

- Using the 'common annotation' from a reliable reference database with standardised annotation and nomenclature allows the standardised annotation of uncharacterised proteins by avoiding the use of wrong nomenclature and of different descriptions for the same biological fact.

- Since the method will take both position-independent and position-specific common annotation available in the reference database into account, a much higher level of annotation will be achieved, including position-specific annotation such as active sites.

- The 'common annotation' approach can be used not only with protein families, but also with conditions aiming at a higher level in the protein family hierarchy. Only the annotation common to all members of this, for instance, super-family, will be copied over. The automatic annotation on a super-family level will obviously lead to more generic and limited annotation than on a family level.

- Our methodology is independent of the multidomain organisation of proteins. If a certain condition aims at a single domain that occurs with various other domains, it can be expected that only the annotation referring to this single domain will be found in all relevant characterised proteins. On the other hand, if the single domain always occurs with another domain, the information for the other domain will be picked up as well.

- The evidence tags will also allow the automatic update of the predicted annotation if the underlying conditions or the 'common annotation' in the RuleBase changes.

FUNCTIONAL INFORMATION IN SWISS-PROT

The functional annotation basis for the automatic annotation of TrEMBL is the functional information in the SWISS-PROT protein sequence database. Many other annotation approaches try to predict functions by comparative analysis with SWISS-PROT and other protein databases like TrEMBL, Genpept, etc. There are three main reasons for using only SWISS-PROT annotation in automatic approaches:

- The comprehensiveness of SWISS-

PROT. This may sound surprising, since SWISS-PROT contains currently (January 2001) only 92,000 proteins. Although these sequences represent – taking redundancy into account – only one-third of all known protein sequences, SWISS-PROT contains around 60 per cent of all proteins found in comprehensive protein sequence databases (like SWISS-PROT + TrEMBL or the protein entries in Entrez) with annotation of at least basic experimentally derived functional characterisation. The percentage was estimated from the number of papers (around 70,000) cited in SWISS-PROT records compared with the number of papers in all SWISS-PROT + TrEMBL or Entrez protein entries (around 110,000) together. The calculation was made by assuming that the proportion of papers reporting sequencing to papers reporting characterisation is the same in SWISS-PROT records as in TrEMBL records or in non-SWISS-PROT Entrez protein records. However, an inspection of citations from SWISS-PROT compared with citations from TrEMBL shows that SWISS-PROT contains a higher proportion of papers representing biochemical citation than do TrEMBL citations. This observation, together with the sequence redundancy in TrEMBL and the non-SWISS-PROT records of Entrez proteins, indicate that SWISS-PROT probably contains even more than 60 per cent of all annotated proteins with at least basic biochemical characterisation. Even more striking is the fact that more than 80 per cent of all functional annotation found in the comprehensive protein sequence database records (like SWISS-PROT + TrEMBL or protein entries in Entrez) is SWISS-PROT annotation.

functional information in SWISS-PROT

- Another important reason is the standardisation of annotation in SWISS-PROT. This unique feature of SWISS-PROT allows the extraction of 'common annotation' described above. Using the standardised SWISS-PROT annotation leads eventually also to a standardised annotation of TrEMBL.

- The last and maybe most important reason is the fact that SWISS-PROT distinguishes experimentally determined functions from those determined computationally. The following highlights how functional annotation is assigned in SWISS-PROT and how experimentally and computationally determined information is portrayed.

SWISS-PROT is, as already mentioned, a curated protein sequence data bank that strives to provide a minimal amount of redundancy, a high level of integration with other databases and a high level of annotation. Efforts are made to enter as much functional information as possible into the database. The annotation is mainly stored in the CC (Comment), FT (Feature Table), KW (Keyword) and DE (Description) lines. There are currently (January 2001) more than 500,000 CC lines, 440,000 FT lines and 120,000 DE lines in SWISS-PROT. The main sources of information are articles reporting sequencing and/or characterisation. When biochemical experiments have been undertaken to characterise a protein, this is added to the Reference Position (RP) line of the entry. This is part of the reference block and describes what has been determined in that publication. As an example, find below the RP line of an entry where the translated sequence, function and the phosphorylation of a protein have been determined:

```
RP   SEQUENCE FROM N.A., FUNCTION, AND
     PHOSPHORYLATION.
```

The SWISS-PROT format currently allows only one RP line of 75 characters (although this will change at a later date), and so we are not always able to list all experimentally determined

characteristics. In these cases 'CHARACTERIZATION' is added.

Now let us have a closer look at the functional information in DE, CC and FT lines. The DE (Description) line(s) lists all the names under which a particular protein is or has been known. The DE line gives also an indication about the characterisation of the protein. Here the SWISS-PROT entry with the accession number P29965 is used as an example:

The description (DE) lines in SWISS-PROT

```
ID   CD4L_HUMAN STANDARD; PRT; 261 AA.

AC   P29965;

DT   01-APR-1993 (Rel. 25, Created)

DT   01-APR-1993 (Rel. 25, Last sequence
     update)

DT   01-JUL-1999 (Rel. 38, Last annotation
     update)

DE   CD40 LIGAND (CD40-L) (TNF-RELATED
     ACTIVATION PROTEIN) (TRAP) (T CELL

DE   ANTIGEN GP39) (CD154 ANTIGEN).
```

Our example describes the protein as 'CD40 LIGAND'. That means that this protein has been experimentally characterised to be the 'CD40 LIGAND'. With the increasing amount of data coming from mega-sequencing projects, more and more proteins in SWISS-PROT will be found with no experimental characterisation. These proteins can be identified through their standardised labelling of the DE line.

All predicted protein sequences lacking any significant sequence similarity to characterised proteins are labelled as '*hypothetical proteins*'. The majority of these cases come from the genome-sequencing projects. Example:

```
DE   HYPOTHETICAL 33.8 KD PROTEIN C5H10.01
     IN CHROMOSOME I.
```

When a protein exhibits extensive sequence similarity to a characterised protein and/or has the same conserved regions then the label '*probable*' is used in the DE line. It is normally followed by the full name of a protein from the same family that it matches. Example:

```
DE   PROBABLE 5'-NUCLEOTIDASE PRECURSOR
     (EC 3.1.3.5).
```

The label '*putative*' is used in the DE line of proteins that exhibit limited sequence similarity to characterised proteins. These proteins often have a conserved site, eg ATP-binding site, but no other significant similarity to a characterised protein. It is most frequently used for sequences from genome projects. Example:

```
DE   PUTATIVE AMINO-ACID PERMEASE.
```

The assignment of the labels 'probable' and 'putative' is dependent primarily on the results of sequence similarity searches against SWISS-PROT and InterPro. It is important to point out here that no specific cut-off point is used to assign a protein as 'putative' or 'probable', i.e. it is not the case that < 50 per cent identity = putative and > 50 per cent = probable. An example illustrates why such assignments must involve curator judgments and cannot be based on specific cut-off points. Take the two *Drosophila* proteins Q9V5E3 (described as 'PROBABLE SERINE PROTEASE CG12133 (EC 3.4.21.-)') and Q9V4W6 (described as 'CG8586 PROTEIN'). A FastA search of both of these protein sequences against SWISS-PROT results in both cases in dozens of highly significant hits (with an *E*-value – the assessment of statistical significance based upon the extreme value distribution – of e-9 or lower) against known proteases. Also, both proteins show Pfam trypsin (AC number PF00089) and PRINTS chymotrypsin (AC number PR00722) signatures. So both proteins seem to belong to the chymotrypsin serine protease family. However, only Q9V5E3 can be a real serine protease, since only in this protein you can find both the serine (PROSITE AC number PS00135) and histidine (PROSITE AC number PS00134) active sites.

Now let us move on to the SWISS-PROT CC lines. In SWISS-PROT entry P29965 you can find the following CC lines:

the comment (CC) lines in SWISS-PROT

```
CC -!- FUNCTION: MEDIATES B-CELL PROLIFERATION IN THE ABSENCE OF CO-
CC     STIMULUS AS WELL AS IGE PRODUCTION IN THE PRESENCE OF IL-4.
CC     INVOLVED IN IMMUNOGLOBULIN CLASS SWITCHING.
CC -!- SUBUNIT: HOMOTRIMER.
CC -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN. ALSO EXISTS AS AN
CC     EXTRACELLULAR SOLUBLE FORM.
CC -!- TISSUE SPECIFICITY: SPECIFICALLY EXPRESSED ON ACTIVATED CD4+
CC     T-LYMPHOCYTES.
CC -!- DISEASE: DEFECTS IN CD40LG ARE THE CAUSE OF AN X-LINKED
CC     IMMUNODEFICIENCY WITH HYPER-IGM (HIGM1), AN IMMUNOGLOBULIN
CC     ISOTYPE
CC     SWITCH DEFECT CHARACTERIZED BY ELEVATED CONCENTRATIONS OF SERUM
CC     IGM AND DECREASED AMOUNTS OF ALL OTHER ISOTYPES. AFFECTED MALES
CC     PRESENT AT AN EARLY AGE (USUALLY WITHIN THE FIRST YEAR OF LIFE)
CC     RECURRENT BACTERIAL AND OPPORTUNISTIC INFECTIONS, INCLUDING
CC     PNEUMOCYSTIS CARINII PNEUMONIA AND INTRACTABLE DIARRHEA DUE TO
CC     CRYPTOSPORIDIUM INFECTION. DESPITE SUBSTITUTION TREATMENT WITH
CC     INTRAVENOUS IMMUNOGLOBULIN, THE OVERALL PROGNOSIS IS RATHER POOR,
CC     WITH A DEATH RATE OF ABOUT 10% BEFORE ADOLESCENCE.
CC -!- SIMILARITY: BELONGS TO THE TUMOR NECROSIS FACTOR FAMILY.
CC -!- DATABASE: NAME=CD40Lbase;
CC     NOTE=European CD40L defect database (mutation db);
CC     WWW="http://www.expasy.ch/cd40lbase/";
CC     FTP="ftp://ftp.expasy.ch/databases/cd40lbase".
CC -!- DATABASE: NAME=PROW; NOTE=CD guide CD154 entry;
CC     WWW="http://www.ncbi.nlm.nih.gov/prow/cd/cd154.htm".
```

the FeaTure Table (FT) lines in SWISS-PROT

The CC (Comments) lines contain various textual comments grouped under different topics. There are altogether 20 different topics. The current topics and their definitions are listed in Table 1.

The CC lines give, as the DE lines, an indication about the level of characterisation of a protein. In our example you can find experimentally verified information about the 'FUNCTION', the quaternary structure ('SUBUNIT'), the 'SUBCELLULAR LOCATION' and the 'TISSUE SPECIFICITY' of the protein. You also find a description of the 'DISEASE(s)' known to be associated with a deficiency of the protein, a description of the 'SIMILARITY' of the protein with other proteins, and a cross-reference to network 'DATABASE' resource(s) for this specific protein.

The labelling of a CC topic with '*by similarity*' indicates that these comments have been assigned because of similarity to an existing characterised entry. The label '*potential*' is in general used if there is no experimental proof for the information given in a CC topic for a protein, but typically members of the same protein family show the annotated characteristics. If comparative analysis reveals highly likely comments, then the label '*probable*' is used:

Further annotation is found in the FT (FeaTure) lines, which describe regions or sites of interest in the sequence:

Table 1: Current topics and definitions in SWISS-PROT

Topic	Description
ALTERNATIVE PRODUCTS	Description of the existence of related protein sequence(s) produced by alternative splicing of the same gene or by the use of alternative initiation codons
CATALYTIC ACTIVITY	Description of the reaction(s) catalysed by an enzyme
CAUTION	This topic warns you about possible errors and/or grounds for confusion
COFACTOR	Description of an enzyme cofactor
DATABASE	Description of a cross-reference to a network database/resource for a specific protein
DEVELOPMENTAL STAGE	Description of the developmental specific expression of a protein
DISEASE	Description of the disease(s) associated with a deficiency of a protein
DOMAIN	Description of the domain structure of a protein
ENZYME REGULATION	Description of an enzyme regulatory mechanism
FUNCTION	General description of the function(s) of a protein
INDUCTION	Description of the compound(s) which stimulate the synthesis of a protein
MASS SPECTROMETRY	Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods
MISCELLANEOUS	Any comment that does not belong to any of the other defined topics
PATHWAY	Description of the metabolic pathway(s) to which a protein is associated
POLYMORPHISM	Description of polymorphism(s)
PTM	Description of a post-translational modification
SIMILARITY	Description of the similarities (sequence or structural) of a protein with other proteins
SUBCELLULAR LOCATION	Description of the subcellular location of the mature protein
SUBUNIT	Description of the quaternary structure of a protein
TISSUE SPECIFICITY	Description of the tissue specificity of a protein

```
FT   DOMAIN 1 22 CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM 23 46 SIGNAL-ANCHOR (TYPE-II
     MEMBRANE PROTEIN).
FT   DOMAIN 47 261 EXTRACELLULAR
     (POTENTIAL).
FT   DISULFID 178 218 POTENTIAL.
FT   CARBOHYD 240 240 POTENTIAL.
FT   VARIANT 36 36 M -> R (IN H1GM1).
..   15 FT lines omitted
..
```

In general the feature table lists post-translational modifications, binding sites, active sites of an enzyme, the secondary structure, sequence conflicts and variations, signal sequences, transit peptides, pro-peptides, trans-membrane regions and other characteristics.

The feature table gives the user, as the CC and DE lines, an indication about the level of characterisation of a protein. In the example above only the variants are experimentally verified. Use of sequence similarity searches and prediction programs have derived the other features. For features not experimentally verified, SWISS-PROT uses the same labelling conventions as already described above for the CC lines. In our example it is known that this is glycosylated, disulphide bonds containing type II membrane protein, but the correct topology of the protein, the glycosylation site(s) and the disulfide bonds have not been experimentally confirmed. The label '*potential*' is used to indicate the predicted character of the information given in the features 'DOMAIN', 'DISULFID' and 'CARBOHYD'. Another label used to indicate that a feature has not been experimentally proven is '*by similarity*'. This label indicates that this feature has been assigned because of similarity to an existing characterised entry. Table 2 summarises the rules used to flag annotation added to entries where no, or only limited, experimental evidence is available.

CONCLUSION

The addition of functional information to SWISS-PROT and TrEMBL protein

Table 2: Flags describing the evidence level of SWISS-PROT annotation

Flag	Line	Explanation
Hypothetical protein	DE	Predicted proteins, which may not be translated.
Putative	DE	Some similarity over conserved domains/sites. Used to show a *tentative* classification.
Probable	DE	Strong similarity over conserved domains/sites. Used to show a *likely* classification.
	CC	Experiments performed but not entirely conclusive.
	FT	Experiments have shown the protein to be modified/processed but actual site has not been confirmed.
Potential	CC	Assignment by comparative analysis. Typically, members of the same family show the annotated characteristics.
	FT	Added to features resulting from prediction programs for signal sequences, trans-membrane regions, coiled-coils, etc.
By similarity	CC	Experimentally proven in member of the same family to which the new entry is believed to belong.
	FT	From alignment the domain/site exists and the sequence it is aligning against has the domain/site experimentally proven.

sequence entries is a time- and labour-intensive process. Great care is taken to ensure that the information added can be traced either to the relevant data source or back to the entry (or entries) reporting the experimentally determined characteristics. This effort is necessary as the exploitation of the sequence avalanche is heavily depending on reliable data sources as the basis for automatic large-scale functional characterisation and annotation by comparative analysis. The ongoing inclusion of additional functional information and a further improved labelling of the annotation in SWISS-PROT and TrEMBL with more advanced and rigorous evidence tagging will be of great importance for the development of new automatic annotation systems able to achieve a much higher level of accurate predicted annotation than today.

References

1. Sanger, F., Coulson, A. R., Friedmann, T. *et al.* (1978), 'The nucleotide sequence of bacteriophage phi-X174', *J. Mol. Biol.*, Vol. 125, pp. 225–246.
2. Oliver, S. G. *et al.* (1992), 'The complete DNA sequence of yeast chromosome III', *Nature*, Vol. 357, pp. 38–46.
3. Fleischmann, R. D. *et al.* (1995), 'Whole-genome random sequencing and assembly of Haemophilus influenzae Rd', *Science*, Vol. 269, pp. 496–512.
4. Goffeau, A. *et al.* (1997), 'The Yeast Genome Directory', *Nature*, Vol. 387 (suppl.), pp. 1–105.
5. The *C. elegans* Sequencing Consortium (1998), 'Genome sequence of the nematode *C. elegans*: A platform for investigating biology', *Science*, Vol. 282, pp. 2012–2018.
6. Adams, M. D. *et al.* (2000), 'The genome sequence of *Drosophila melanogaster*', *Science*, Vol. 287, pp. 2185–2195.
7. Frishman, D. and Mewes, H.-W. (1997), 'PEDANTic genome analysis', *Trends Genetics*, Vol. 13, pp. 415–416.
8. Scharf, M., Schneider, R., Casari, G. *et al.* (1994), 'GeneQuiz: a workbench for sequence analysis', In: Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D. Eds, 'Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 348–353.
9. Bairoch, A. and Apweiler, R. (2000), 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Res.*, Vol. 28, pp. 45–48.
10. Fleischmann, W., Moeller, S., Gateau, A. and Apweiler, R. (1999), 'A novel method for automatic and reliable functional annotation', *Bioinfomatics*, Vol. 15, pp. 228–233.
11. Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999), 'The PROSITE database, its status in 1999', *Nucleic Acids Res.*, Vol. 27, pp. 215–219.
12. Apweiler, R., Attwood, T. K., Bairoch, A.

et al. (2001), 'InterPro – An integrated documentation resource for protein families, domains and functional sites', *Nucleic Acids Res*., Vol. 29, pp. 37–40.

13. Attwood, T. K., Croning, M. D. R., Flower, D. R. *et al*. (2000), 'PRINTS-S: The database formerly known as PRINTS', *Nucleic Acids Res*., Vol. 28, pp. 225–227.
14. Bateman, A., Birney, E., Durbin, R. *et al*. (2000), 'The Pfam Protein Families Database', *Nucleic Acids Res*., Vol. 28, pp. 263–266.
15. Corpet, F., Servant, F., Gouzy, J. and Kahn, D. (2000), 'ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons', *Nucleic Acids Res*., Vol. 28, pp. 267–269.
16. Bork, P. and Koonin, E. V. (1998), 'Predicting functions from protein sequences – where are the bottlenecks?', *Nature Genetics*, Vol. 18, pp. 313–318.