Linking publication, gene and protein data

Paul Kersey and Rolf Apweiler

The computational reconstruction of biological systems, 'systems biology', is necessarily dependent on the existence of well-annotated data sets defining and describing the components of these systems, especially genes and the proteins they encode. Information about these components can be accessed either through structured bioinformatics databases, which store basic chemical and functional information abstracted from (or supplementing) the scientific literature, or through the literature itself, which is richer in content but essentially unstructured.

The systematic application of automated high-throughput molecular biology techniques has led to the generation of an immense quantity of data compared with that produced by the hypothesis-driven application of earlier technologies. For example, advances in DNA sequencing technology have made the study of complete genomes routine, and have been followed by similar progress in the study of transcripts, proteins and metabolites, although these generate far more complex datasets. In fact, the need for databases that store molecular biological data, and that allow analysis through computational algorithms, was apparent long before experimental techniques were as powerful — and as data generative — as they are today. The first attempts to catalogue information about proteins began in 1965, and by 1984 these had evolved into the Protein Information Resource[1], the Protein Data Bank[2] (which stores information about protein structures and was founded in 1971) and the EMBL data library[3] (the first database to store information about nucleic-acid sequences, founded in 1981). Today, these databases and their successors have been joined by numerous other resources, which store information on chemical entities, gene expression, molecular interactions and biochemical pathways. These databases often contain more raw data than could be practically published in a conventional hard-copy journal. However, the interpretation of these data is still dependent on inference drawn from hypothesis-driven experimentation, the details of which reside in free-text articles. The value of bioinformatics data is thus utterly dependent on the ability to make the correct links from the sequences to the scientific literature, and to extract the information that literature contains.

This article is a user's guide to linking gene, protein and publication data. It considers the principal primary resources in which such data can be found, as well as the difficulties and potential pitfalls when linking them, and some of the general approaches and specific tools that can be used to overcome these problems. We also discuss the semantic web, a recent technological framework proposed for the integration of distributed data that is attracting particular interest within the bioinformatics community. The need for new solutions is acute, as the quantity and variety of available data, including transcriptome, proteome, metabolome, molecular and genetic interaction, and image data continues to rise. However, this deluge of data is also an opportunity for development and has led to the growth of the discipline of systems biology, which aims to reconstruct entire biological systems through the modelling of their components. This requires that all components are clearly identified and well-described, which for a large system means, in practice, that descriptive annotation must be inferred, often automatically, from related

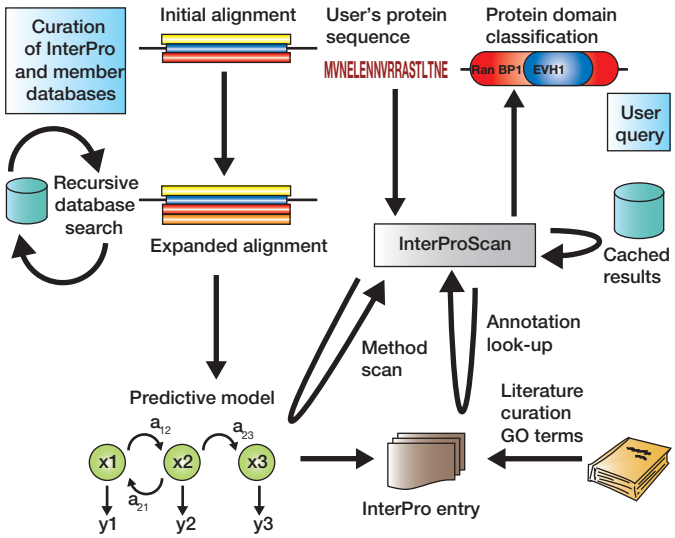

Figure 1 Protein analysis using InterProScan. Curators of member databases of InterPro identify proteins that share a known functional domain and manually identified sequences can be supplemented through iterative searching of the public databases until a comprehensive set is found. The sequences are aligned and a classifying function is built from the alignment, which is designed to recognise other proteins that posses the same domain. Many of these classifiers are built using hidden Markov models. Alternative classifying functions judged (by curators) to identify the same domain are grouped into a single entry in the InterPro database, and annotated with relevant information (expressed both in free text and in the terms of the GO controlled vocabulary). InterProScan is a programme that allows users to characterize a sequence by applying these classifying functions. Performance can be improved by precomputing the results for known sequences.

Paul Kersey and Rolf Apweiler are in the EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
e-mail: pkersey@ebi.ac.uk

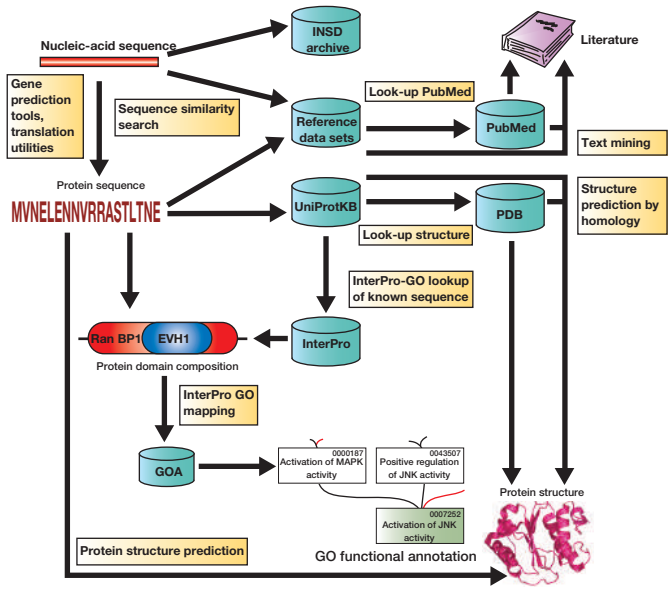

Figure 2 A schematic representation of a typical workflow for bioinformatics analysis. The figure shows a typical workflow for bioinformatics analysis, progressing from sequence to functional annotation, protein structure and literature. The complete analysis may be carried out entirely within a bioinformatics warehousing system (such as SRS), or as a sequence of separate operations performed in different environments. Starting with the sequence of a gene or protein, identical and/or similar sequences are identified in the public databases. The database records describing these sequences also contain general information about the sequence, curated links to structural information and relevant scientific literature. Protein sequence itself can also be directly analysed to predict its domain composition and structure. These may provide a more reliable indication of protein function than overall sequence similarity. It is often useful to express the results of these analyses in standard controlled vocabularies (such as GO), to allow the comparison and correspondence of information derived from different resources.

components in other systems. Therefore, the establishment and exploitation of reliable connections between the underlying source data is not only necessary for primary sequence analysis, but is also an essential part of the platform needed to support the computational, synthetic approaches to biology that are emerging in the post-genomic era.

PRIMARY BIOINFORMATICS DATABASES

Here, we discuss some of the main resources that function as primary repositories of bioinformatics data and the types of data that they contain, and introduce MEDLINE and PubMed, two databases that offer computerised access to scientific literature.

Nucleotide databases

The EMBL[3]–Genbank[4]–DDBJ[5] International Nucleotide Sequence Database (INSD) is the product of a collaborative effort to maintain a single global archive for nucleotide sequences. As most other bioinformatic data is dependent on inference from nucleotide sequence, it is arguably the single most important bioinformatics resource. Its comprehensive coverage is maintained through the editorial policy of most scientific journals — submitting authors have to deposit the sequence information that is associated with articles they publish. The joint database currently contains over 80 million sequences.

The information contained within the archive, however, is not uniformly well annotated. The contents of existing databases are used in most methods for gene prediction, and thus, once these databases have been updated, earlier predictions may differ from those that could now be made. In addition, individual submitters are free to choose what annotation they consider pertinent to add to a sequence. As a result, nothing can be inferred from the absence of information in database records — this particular information may simply have been unknown to, or ignored by, the submitter.

With the advent of whole genome sequencing, a new class of database has emerged for data from higher eukaryotic species, in which sequence and automatically generated annotation is made available in advance of not only experimental confirmation, but even the existence of definitive methods for gene prediction. The Ensembl[6] database of metazoan genome annotation is a good example of such a resource. The contents of this type of database are subject to frequent revision, especially when the sequence has only recently been deciphered, but they often provide more complete and up-to-date annotation than the traditional archives.

Protein databases

Most efforts at recording functional information have occurred in protein databases. Swiss-Prot, a subsection of the UniProt Knowledgebase[7] (UniProtKB) that is manually curated from the scientific literature, is widely considered to be the richest and most reliable store of functional information, and is frequently used to infer information about proteins not yet subjected to specific experimental study. The world-wide Protein Data Bank[2] functions as an archive for protein structures and information about posttranslational modifications can be found in the RESID database[8]

Other bioinformatics databases

Several hundred bioinformatics databases exist in total, with many resources focusing on particular taxonomic or methodological domains. For example, there are specific databases for many model organisms

BOX 1 DATA WAREHOUSES

A data warehouse is a database constructed to support efficient querying of the data it contains (in contrast to normalized databases designed to support data integrity, which are widely used to maintain primary resources). A feature of many data warehouses used in bioinformatics is that they provide generic query interfaces (for example, computer languages and graphical user interfaces) applicable to all the data they contain, thus enabling the addition of new data without the need for interface redesign. A single warehouse may be built from several different resources, but to allow the construction of queries that filter and/or extract information derived from more than one of these, the data must be fitted into a single model that captures the relationships between the different sources. This is often done by exploiting the cross-references that many of these sources contain. The centralization of data in a data warehouse is an alternative to the use of technologies designed to support distributed queries. Distributed approaches can avoid certain problems associated with data warehousing, such as the synchronization of updates, but queries accessing large quantities of data spread over many locations can be difficult to optimise.

BOX 2 PROBLEMS IN LINKING BIOINFORMATICS DATA TO LITERATURE

In theory, data warehousing technologies would seem to support the automatic generation of a single integrated data resource from independent, but cross-referenced, databases. However, there are certain inherent limitations to this approach. Accurate cross-references are vital for good results, but the volume of data now extant exceeds the capacity for manual curation of almost all bioinformatics databases — for example, about 220,000 records (7% of the total) in the UniProtKB are manually curated. Automatic methods for generating cross-references can be used instead, through the tracking of identifiers (as data is transferred between resources), or by a comparison of the properties (such as the sequence) of entities to establish their equivalence. However, these methods may not always produce the correct answers. Different databases may maintain different identifiers, and use different names, for the same biological object (synonymy and homonymy are particularly problematic when searching literature, in which references to genes and proteins occur in free text and do not necessarily follow well-defined standards). And the 'same' object may be assigned different properties in different resources.

Other problems may be encountered even when cross-references are well-maintained: first, the records in different databases may not be conceptually compatible. Numerous definitions exist for words like protein and gene, and their use is not necessarily standardised between resources. Second, different resources may differ in content, that is, the data they choose to include may not be the same. Third, the huge quantity of experimentally generated data, and the generally limited frequency at which submissions are updated, means that archive databases may contain much information that can be considered redundant or out-of-date. Fourth, the use of different standards and syntax for annotation is a further barrier to integration.

(such as bakers' yeast[9], worm[10], fruit fly[11] and mouse[12]), which are often the most detailed source of information about gene and protein function within their taxonomic scope. Other resources concentrate on particular families of (protein and functional RNA-encoding) genes.

Literature databases

The scientific literature itself can be accessed through computerized databases. The largest primary database containing biological literature is MEDLINE, which contains over 10 million citations. MEDLINE is frequently accessed as part of PubMed (www.ncbi.nih.gov/entrez/query/static/overview.html), a larger collection of literature databases. Many publishers are now making the full text of articles publicly available from their own archives over the World Wide Web, frequently using digital object identifiers (DOIs; www.doi.com), which provide a common mechanism for data access and many full-text articles are also freely available in the PubMed Central archive[13].

The simplest way that bioinformatics and literature databases can be linked is through the use of the cross-references, which many of these resources maintain to each other. The UniProtKB, for example, maintains cross-references to over 100 other resources (including MEDLINE and PubMed) in which more detailed information on particular aspects of a protein's function can be found. These cross-references are widely exploited by data warehouses (see Box 1).

PUBLIC RESOURCES FOR INTEGRATED BIOINFORMATICS

Many of the main bioinformatics databases are made available by a small number of institutions whose mission is to provide services to the scientific community (such as the European Bioinformatics Institute[14] (EBI) in the UK, and the National Centre for Biotechnology Information[15] (NCBI) in the USA). These resources provide an obvious first point of call for anyone attempting to integrate bioinformatics data.

The EBI maintains a large data warehouse that contains over 100 bioinformatics databases[16] using the SRS data warehousing system[17]. The NCBI Entrez server[18] offers similar functionality (see Box 2). However, there is a difference in the focus of the two systems: the interface to SRS is powerful but complex, allowing the construction of inter-database queries in a generic manner (for example, "get me the DNA sequences encoding proteins of known structure in these species"). Entrez offers a simpler interface, with less support for structured queries, but provides rapid retrieval of data of all types linked to a given search term (for example, "get me all the genes, proteins, genomes and literature that relate to ras1"). Both warehouses contain not only gene and protein databases, but also literature databases (for example, PubMed is incorporated into Entrez), allowing direct interlinking between these resources.

Ultimately, a warehouse's usefulness is dependent on the coherence of the data it contains (see Box 2). Approaches to improve the coherence among biological databases include the establishment of international collaborations to share data, the use of common controlled vocabularies (such as the Gene Ontology[19], GO, a hierarchical controlled vocabulary for describing the function of gene products) by different resources, and the wider development of standards for data representation (particularly in new areas of research, such as transcriptomics[20] and proteomics[21]), where the data complexity can be orders of magnitude greater than that of sequence data. Another approach is the production of well-annotated, non-redundant subsets of the complete data and the development of services based on these. The EBI's Integr8 project[22] provides reference data sets and analysis tools for species with completely deciphered genomes from which redundancy has been removed. Annotation in Integr8 is enhanced, updated and corrected (with respect to the original data submissions) through the transfer of data from actively curated resources and the performance of new computational analysis. The re-annotated genomes are available in EMBL-like format as Genome Reviews. The NCBI's RefSeq database[23] similarly provides comprehensive, non-redundant sequences for genomes, genes and proteins. The related resource Gene[24] provides a framework (a defined gene set) against which sequences from RefSeq and other resources are cross-referenced, and includes two-way links to literature in PubMed. Searching against resources such as these may provide clearer answers than those obtained when using the primary repositories, in which data redundancy, and non-standard or out-of-date annotation, may obscure the correct results.

Figure 3 From gene to protein to literature: a sample analysis. Beginning with a gene sequence (top left), a protein sequence (below gene sequence) is generated and can be analysed with InterProScan to detect sequences indicative of the presence of known protein domains and families within the sequence, generating a graphical overview (top right). The sequence is also compared with known sequences in the UniProtKB database using the BLAST algorithm and the output displays an alignment (middle) between the query sequence and its best matches. The best match for the sample query data is the RAS1 protein from *Saccharomyces cerevisiae*. The corresponding UniProtKB entry features curated links for this protein to known publications (lower left). Additional publications for this protein, retrieved using text-mining methods, can be found by querying iHOP (lower right).

SMALL-SCALE DATA MINING

Data mining refers to two phenomena: the analysis of large-scale data sets for the purpose of general inference (discussed in the next section) and the extraction of specific information that relates to some initial data of interest. If that initial data is a gene or protein name or ID, querying resources such as UniProtKB (for protein-centric queries), Integr8 (to obtain information in a genomic context) or SRS (for complex filtering over many resources) will usually identify a database record that contains much of the information that is already known about that entity. However, if the data consists of nucleotide or protein sequence, a comparison needs to be made to known sequences to identify similar or identical molecules that have already been annotated. Common algorithms used for this purpose include BLAST[25], FASTA[26] and Smith-Waterman[27], and are available through the websites and web services of the EBI, NCBI and other bioinformatics service providers.

Analysing sequence

High overall sequence similarity may be a good indicator of functional equivalence, but aspects of a protein's function may be inferred from the presence of particular domains, even if a protein's overall architecture has not been reporte or the complete function of closely matching sequences is not known. Several methods exist for domain identification in sequence, mostly using hidden Markov models[28]. InterPro[29] is a curated, integrative resource that combines methods for domain identification from 15 different member databases, in which redundant methods are merged and common annotation attached. InterProScan[30], the programme which applies these methods (online or as a local installation), is among the most powerful tools for characterizing unknown protein sequence. How InterPro entries are assembled and how a scan is applied to unknown sequence is illustrated in Fig. 1.

Predicting protein structure is more problematic, as the structures of most proteins remain unknown. Sequence similarity can be used to infer structure, providing the contribution of individual residues within the sequence to the structure is taken into account (and not just the overall degree of similarity). Tools such as SWISS-MODEL[31] generate a probable structure for a query sequence by identifying a similar sequence of known structures that can serve as a model. Certain structural features can also be predicted by direct sequence analysis. Programmes

that perform this type of analysis include TMHMM[32], which predicts trans-membrane helices, and GenThreader[33] (available through the PSI-PRED server[34]), a tool for predicting protein folds. However, prediction of higher-order structural features from sequence is more difficult than predicting secondary structure or domain composition, especially for proteins not related to those with known structures.

Literature mining

In Swiss-Prot, Entrez Gene and other well-curated databases, direct links exist from individual records to relevant publications. However, references to many papers relevant to a gene or protein are not likely to have been directly curated. Further literature can be found by directly searching literature databases with the name(s) associated with the protein, although care must be taken to avoid confusion caused by the existence of synonyms and homonyms. Abstracts in MEDLINE are labelled with terms from a controlled vocabulary (medical subject headings; MeSH[35]), and searching with relevant MeSH terms can be used to identify papers of interest. Unfortunately, most bioinformatics database records are not directly annotated with MeSH terms; instead, they are often annotated with terms from the GO vocabulary. GO is more tightly focused than MeSH — each contains approximately 20,000 terms, but MeSH is split into 16 principal subsections and covers geography and sociology in addition to biology, whereas GO only covers three specific aspects of biology. A resource to translate GO terms automatically into their equivalent MeSH terms is still in development.

Increasingly, natural language programming (NLP) techniques are being used to mine literature databases automatically, thus providing a more sophisticated way of identifying relevant publications. Programs using NLP can typically be rerun with greater frequency than a record can be revisited by a curator. These techniques are limited in their ability to accurately summarise the complete information contained in a single paper, but can effectively identify papers worthy of subsequent manual inspection. iHOP[36] is a web interface that provides access to papers that have been identified using natural language programming as relevant to genes from a number of sample genomes. This interface displays the specific text that was used to identify each paper as relevant, thereby allowing false positives to be quickly discarded.

An outline workflow that might be used in characterizing an unknown sequence is shown in Fig. 2 and a specific example is given in Fig. 3.

INFORMATION EXTRACTION FROM LARGE DATA SETS

Data mining for large data sets can be quite different to extracting data about individual sequences. The types of analysis performed are generally similar, but the development of automated procedures is usually essential as the data volume is increased. Many public bioinformatics servers place limits on the amount of data that can be requested by users in a single query, although the underlying software is often also available for local installation. Another route by which service providers are increasingly attempting to meet user demands for bulk data analysis is through the provision of programmatic access over the internet, with the widespread implementation of services meeting the Web Service standards proposed by the World Wide Web consortium (http://www.w3.org/2002/ws).

For scientists who have to analyse large quantities of data, but who lack programming expertise, the use of a workflow management tool may provide a solution. Taverna[37] is designed explicitly for use in the context of bioinformatics data, providing a graphical user interface for the assembly of multiple services, potentially running at diverse locations, into a single data processing pipeline.

Large data sets enable 'knowledge discovery' through the identification of patterns within the data. For example, individual pieces of protein interaction data may be combined to produce a local or global data interaction network. One such tool, the IntAct Hierarch view application[38], also allows such information to be overlaid with patterns of GO annotation, which in turn can enable the validation and interpretation of the results. Another example of large-scale data mining can be found on the UniProt website[7], where statistical patterns in curated data sets have been used to apply annotation to non-curated sequences. Statistical predictions can also be directly tested against the actual annotation in curated records, allowing each type of data to be validated against the other.

THE FUTURE OF BIOINFORMATICS DATA AND BIOINFORMATICS DATA SERVICES

The importance of automatic methods for linking gene, protein and literature data has grown as the number of known sequences has increased and the proportion of manually curated database records has fallen. As new, increasingly large-scale, experimental techniques continue to be developed, and the number of specialised databases holding data from their application grows in parallel, the ability to perform distributed

BOX 3 TOWARDS SYSTEMATIC RECONSTRUCTION OF WHOLE ORGANISM BIOLOGY

In addition to generic data warehousing models such as SRS and Entrez, an increasing range of more specialist data warehousing tools have been developed by groups specifically aiming to integrate data for particular species. The proliferation of different data models and user interfaces can be a problem for users. Therefore, an important development is the emergence of reusable frameworks (applicable to multiple species) such as GMOD39 (the generic model organism database project), GUS40 (the genomics unified schema), and BioCyc41 (a collection of 205 species-specific metabolic pathway databases). BioCyc is particularly interesting — for its content (a partial reconstruction of cellular metabolomes), but also for its mode of generation, whereby databases for most species are inferred computationally from a combination of genome sequence and a curated database of pathway information largely derived from well-studied species (such as *Escherichia coli*). As the number of biological systems, and the attributes of these systems, continues to increase, it is likely that only a decreasing portion of the generated data will be subject to individual experimental verification and subsequent curation into databases. Therefore, the approximate reconstruction of less-characterized systems by inference from well-characterized models will become an increasingly important factor in the interpretation of large-scale experiments.

queries over resources with separate physical locations is also growing in importance, as there is a limit to the size and complexity at which integrated data warehouses can be efficiently maintained. This ability is also necessary if local, and potentially private, data is to be integrated with information available from public resources. New bioinformatics warehousing systems, such as BioMart[42], have been designed with the explicit aim of supporting queries that can be executed across differently located resources, but ensuring the efficient execution of such queries is inherently difficult.

An alternative to accessing a single data warehouse is to retrieve data from multiple sources and merge this as needed. Projects such as MyGrid[43] are developing software to provide solutions to generic problems in analysing bioinformatics data distributed across many resources. However, it remains difficult to integrate data from different resources. For example, to write a programme to combine data from Entrez and the EBI SRS server requires domain-specific expertise to retrieve and parse the data from each source, and additional expertise to merge the data correctly. These are common problems for programmers attempting to mine information available on the internet, not just those working with biological data, and there is growing interest among bioinformaticians in adopting the emerging 'semantic web' technologies[44] designed to support distributed computing over a wider range of domains. The key concept of the semantic web is the publication of self-describing data, that is, data published together with the metadata that describes it. If such publication accords to standardised protocols, and if the descriptions themselves are machine-readable, standardised and shared across resources, then a programmer should be able retrieve and integrate distributed data by specifying a logical request to a semantically aware search engine without needing specific knowledge about the peculiarities of individual sources. This model is particularly attractive, not only due to the highly dispersed nature of much bioinformatics data, but also because the descriptive standards necessary to make it work have already been developed for particular sub-domains, such as microarray and molecular interaction data.

Table 1 Glossary

Controlled vocabulary	A limited range of terms used in bioinformatics databases to ensure standardised annotation (as opposed to free text). If terms are linked by defined relationships, a controlled vocabulary may constitute an ontology.
Data warehouse	A database constructed to support efficient querying of the data it contains, possibly constructed from many primary data sources.
Distributed queries	A query addressed, not to a single database, but to multiple data resources at potentially disparate locations on a network. Distributed queries work on distributed data, and are one application in the wider field of distributed computing.
Digital object identifier	A standard way for representing a piece of intellectual property expressed in a digital environment (for example, an online journal article).
GO	Gene ontology; a structured controlled vocabulary widely used for the annotation of gene products.
MeSH	Medical subject headings; a structured controlled vocabulary used in MEDLINE to summarise the contents of scientific articles.
Hidden markov models	HMMs; a statistical modelling technique that can be applied to alignments of sequences and used to identify other sequences with similar properties to those in the alignment.
Homonymy	The condition where two different entities share a common name (often true of genes).
Natural language programming	NLP; the discipline of computer science that attempts to understand the meaning of natural language. Widely used to automate processing of scientific literature, sometimes accurately.
-ome, -omics	Suffixes indicating the complete data (and the study of that data) of a particular type (indicated by the root) from particular biological systems (for example, genome/genomics, proteome/proteomics).
Parsing	The process of analysing the contents of a document (such as a flat-file database entry) in accordance with a set of rules defining its structure.
Ontology	In computer science, a model of a particular domain, amounting to a logical description of a set of related concepts.
Semantic web	A network of interlinked data available on the internet, capable of being understood by computer software; also used to refer to technologies associated with the development of such a network.
SRS	Sequence retrieval system; a data warehousing system widely used in bioinformatics, originally developed at the EBI and now maintained by BioWisdom Ltd.
Synonymy	The condition where a single entity has multiple names (often true of genes)
Web services	Programmatic interfaces available for remote computation, using the protocols associated with the world wide web; one potential component of the semantic web.

A significant challenge for the bioinformatics community is to continue to maintain and promote these standards, and to develop additional standards for data types such as small RNAs, where suitable representations do not yet exist, at a sufficient pace to match the ever-growing diversity and quantity of data. □

Note: Supplementary Information is available on the Nature Cell Biology website.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing financial interests.

1. Apweiler, R., Bairoch, A. & Wu, C. H. Protein sequence databases. *Curr. Opin. Chem. Biol.* **8,** 76–80 (2004).
2. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.* **10,** 980 (2003).
3. Cochrane, G. *et al.* EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res.* **34**, D10–D15 (2006).
4. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **34**, D16–D20 (2006).
5. Okubo, K., Sugawara, H., Gojobori, T. & Tateno, Y. DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res.* **34**, D6–9 (2006).
6. Birney, E. *et al.* Ensembl 2006. *Nucleic Acids Res.* **34**, D556–D561 (2006).
7. Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187–D191 (2006).
8. Garavelli, J. S. The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics* **4,** 1527–1533 (2004).
9. Christie, K. R. *et al. Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.* **32**, D311–D314 (2004).
10. Schwarz, E. M. *et al.* WormBase: better software, richer content. *Nucleic Acids Res.* **34**, D475–D478 (2006).
11. Grumbling, G. & Strelets, V. FlyBase: anatomical data, images and queries. *Nucleic Acids Res.* **34**, D484–D488 (2006).
12. Blake, J. A., Eppig, J. T., Bult, C. J., Kadin, J. A. & Richardson, J. E. The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res.* **34**, D562–D567 (2006).
13. Sequeira, E., McEntyre, J. & Lipman, D. PubMed Central decentralized. *Nature* **410,** 740 (2001).
14. Lopez, R., Duggan, K., Harte, N. & Kibria, A. Public services from the European Bioinformatics Institute. *Brief Bioinform.* **4,** 332–340 (2003).
15. Jenuth, J. P. The NCBI. Publicly available tools and resources on the Web. *Methods Mol. Biol.* **132,** 301–312 (2000).
16. Zdobnov, E. M., Lopez, R., Apweiler, R. & Etzold, T. The EBI SRS server-new features. *Bioinformatics* **18,** 1149–1150 (2002).
17. Etzold, T., Ulyanov, A. & Argos, P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.* **266,** 114–128 (1996).
18. Geer, R. C. & Sayers, E. W. Entrez: making use of its power. *Brief Bioinform.* **4,** 179–184 (2003).
19. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* **34**, D322–D326 (2006).
20. Whetzel, P. L. *et al.* The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics* **22,** 866–873 (2006).
21. Orchard, S. *et al.* Autumn 2005 Workshop of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) Geneva, September, 4–6, 2005. *Proteomics* **6,** 738–741 (2006).
22. Kersey, P. *et al.* Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res* **33,** D297–D302 (2005).
23. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
24. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54–D58 (2005).
25. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25 (2004).
26. Pearson, W. R. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.* **24,** 307–331 (1994).
27. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
28. Eddy, S. R. What is a hidden Markov model? *Nature Biotechnol.* **22,** 1315–1316 (2004).
29. Mulder, N. J. *et al.* InterPro, progress and status in 2005. *Nucleic Acids Res.* **33**, D201–D205 (2005).
30. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
31. Kopp, J. & Schwede, T. The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res.* **34**, D315–D318 (2006).
32. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–182 (1998).
33. McGuffin, L. J. & Jones, D. T. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19,** 874–881 (2003).
34. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16,** 404–405 (2000).
35. Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J. L. & Arluk, N. The MeSH translation maintenance system: structure, interface design, and implementation. *Medinfo* **11,** 67–69 (2004).
36. Hoffmann, R. & Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* **21**, ii252–ii258 (2005).
37. Oinn, T. *et al.* Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20,** 3045–3054 (2004).
38. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
39. Stein, L. D. *et al.* The generic genome browser: a building block for a model organism system database. *Genome Res.* **12,** 1599–1610 (2002).
40. Davidson, S. B. *et al.* K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. *IBM Systems Journal* **40,** 512–531 (2001).
41. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33,** 6083–6089 (2005).
42. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–3440 (2005).
43. Stevens, R. D., Robinson, A. J. & Goble, C. A. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* **19,** i302–i304 (2003).
44. Berners-Lee, T. & Hendler, J. Publishing on the semantic web. *Nature* **410,** 1023–1024 (2001).

Table S1 Some important bioinformatics resources and their applications

Institutions

Name	URL	Description
European Bioinformatics Institute (EBI)	www.ebi.ac.uk	European Centre for research and services in bioinformatics, part of EMBL (European Molecular Biology Laboratory)
National Center for Biotechnology Information	www.ncbi.nih.gov	Centre for bioinformatics research and services in the United States, part of the National Institutes of Health.

Primary Databases

Name	URL	Description
International Nucleotide Sequence Database Collaboration (INSDC)	www.insdc.org	Collaboration that maintains archival repository of nucleotide sequence data
EMBL Nucleotide Sequence Database	www.ebi.ac.uk/embl	European partner in the INSDC
Genbank	www.ncbi.nih.gov/Genbank	American partner in the INSDC
DNA Database of Japan (DDBJ)		Japanese partner in the INSDC
UniProt Knowledgebase (UniProtKB)	www.uniprot.org	Central access point for protein-related information maintained by the UniProt consortium
Swiss-Prot		Manually curated portion of the UniProtKB, commonly regarded as a gold standard for functional annotation
Gene	www.ncbi.nih.gov/query.fcgi?db=gene	Database of known genes maintained by NCBI. Provides reference system for NCBI annotation efforts.
RefSeq	www.ncbi.nih.gov/RefSeq	Database of reference DNA and protein sequences maintained by NCBI. Closely linked to Gene.
Ensembl	www.ensembl.org	Metazoan genome annotation database
Saccharomyces Genome Database (SGD)	www.yeastgenome.org	*Saccharomyces* genome database
Mouse Genome Informatics (MGI)	www.informatics.jax.org	Mouse genome database
FlyBase	flybase.bio.indiana.edu	*Drosophila* genome database
WormBase	www.wormbase.org	*Caenorhabditis* genome database
GOA	www.ebi.ac.uk/goa	Database of functional annotation, part-curated, part derived from InterPro analysis
IntAct	www.ebi.ac.uk	Database of protein interactions
BioCyc	www.biocyc.org	Collection of pathway and genome databases
World-wide Protein Data Bank (wwPDB)	www.wwpdb.org	Repository for protein structure information
ArrayExpress	www.ebi.ac.uk/arrayexpresss	European repository for microarray data
Gene Expression Omnibus (GEO)	www.ncbi.nlm.nih.gov/geo	American repository for microarray data
MEDLINE	www.nlm.nih.gov/pubs/factsheets/jsel.html	Database of information on biomedical literature

PubMed	www.ncbi.nih.gov/entrez/query/static/overview.html	Expanded version of MEDLINE, searchable through Entrez at NCBI
Sequence Analysis		
Name	**URL**	**Description**
Smith-Waterman (SW)	Programs running most sequence analysis algorithms are available from many bioinformatics service providers.	Dynamic programming algorithm for sequence analysis, widely considered good, but slow. Various hardware-optimised approximations to SW also exist.
Basic Local Alignment Search Tool (BLAST)		High speed algorithm for sequence comparison.
FAST-All (FASTA)		Another widely used sequence comparison algorithm
InterProScan	www.ebi.ac.uk/InterProScan	Domain analysis tool incorporating algorithms and annotations from 16 consortium members. Results for known sequences stored in InterPro database.
TMHMM	www.cbs.dtu.dk/services/TMHMM-2.0	Prediction of transmembrane helices in proteins
SWISS-MODEL	swissmodel.expasy.org	Automated protein modelling server
PSIPRED	bioinf.cs.ucl.ac.uk/psipred	Protein structure prediction server, offers access to the GenTHREADER program
Warehouses and web portals		
EBI SRS server	srs.ebi.ac.uk	Primary data warehouse/query engine maintained by EBI
NCBI Entrez server	www.ncbi.nih.gov/database/index.html	Primary data warehouse/query engine maintained by NCBI
Integr8	www.ebi.ac.uk/integr8	Provides structured overview of genes and genomes
iHOP	www.ihop-net.org/UniPub/iHOP	Portal for assessing automatically retrieved literature linked to genes and proteins
Generic database schemas		
GMOD	www.gmod.org	Schema maintained by several model organism databases
GUS	www.gusdb.org	Schema for functional genomics data
Integration tools		
BioMart	www.biomart.org	Data warehousing system designed for easy local install, interface generation and distributed querying
Taverna	taverna.sourceforge.net	Workflow manager for bioinformatics. Part of the MyGRID project.
MyGRID	mygrid.org.uk	Project developing a suite of tools for distributed computing in bioinformatics