Integrating Genomic and Proteomic Data: The Integr8 Project

Manuela Pruess, Paul Kersey, Rolf Apweiler

EMBL Outstation European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; mpr@ebi.ac.uk

Summary

Integr8 (http://www.ebi.ac.uk/integr8/) has been developed to provide an integration layer for the exploitation of genomic and proteomic data. High-quality databases from major bioinformatics centres in Europe are included, and some core data and the relationships of biological entities to each other and to entries in other databases are stored. Thus, a framework exists that allows for new kinds of data to be integrated, and an entity-centric view of complete genomes and proteomes is offered. Integr8 is an automatically populated database, providing different entry points to the data, depending on the user's entity of interest. The Proteome Analysis database for statistical analysis and the Genome Reviews for annotated genome information are the main developments within the Integr8 project. With the BioMart application, an interactive querying tool for performing customisable proteome analysis and data mining is offered. Future developments will especially focus on the Genome Reviews, including mapping not yet annotated protein sequences onto their corresponding genomes, generating new predictions for non-coding RNA genes, and generally extending the scope to lower metazoan organisms.

1 Introduction

Since the first bacterial genome (*H. influenzae*) has been sequenced in 1995 [1] and the first eukaryotic genome (*S. cerevisiae*) in 1996 [2], sequencing methods have evolved substantially, with most of the data coming now from the big sequencing projects. The number of completely sequenced genomes is increasing at a rapid speed, as is the amount of available proteomics data. All these different types of data present a huge challenge for data integration, because they are usually stored in different databases without overall data standards or universal links. To a certain amount data can be linked ('integrated') through manual, literature-based curation - but with the quantity of available data continuing to increase exponentially, automatic methods for data integration need to be developed. Integr8 [3] is such an approach, a joint project of major European molecular biology data providers. To ensure an optimal coverage of genomic and proteomic data known up to date, the databases included into the Integr8 project are, amongst others, the UniProt Protein Sequence Knowledgebase [4], the EMBL Nucleotide Sequence Database [5], the Ensembl Genome Browser [6], the Transcription Factor Database (TRANSFAC) [7], the Eukaryotic Promoter Database (EPD) [8], the European Macromolecular Structure Database (E-MSD) [9], the Homologous Vertebrate Genes Database (HOVERGEN) [10] and the RZPD Clone Database.

Integr8 is a species-centric portal on information about complete genomes and proteomes (Figure 1). It has been developed to combine and provide different forms of information, spanning sequences and functions derived from in silico comparisons or specific experiments. Additionally, literature is listed and statistical analysis and summary information for each species is available.

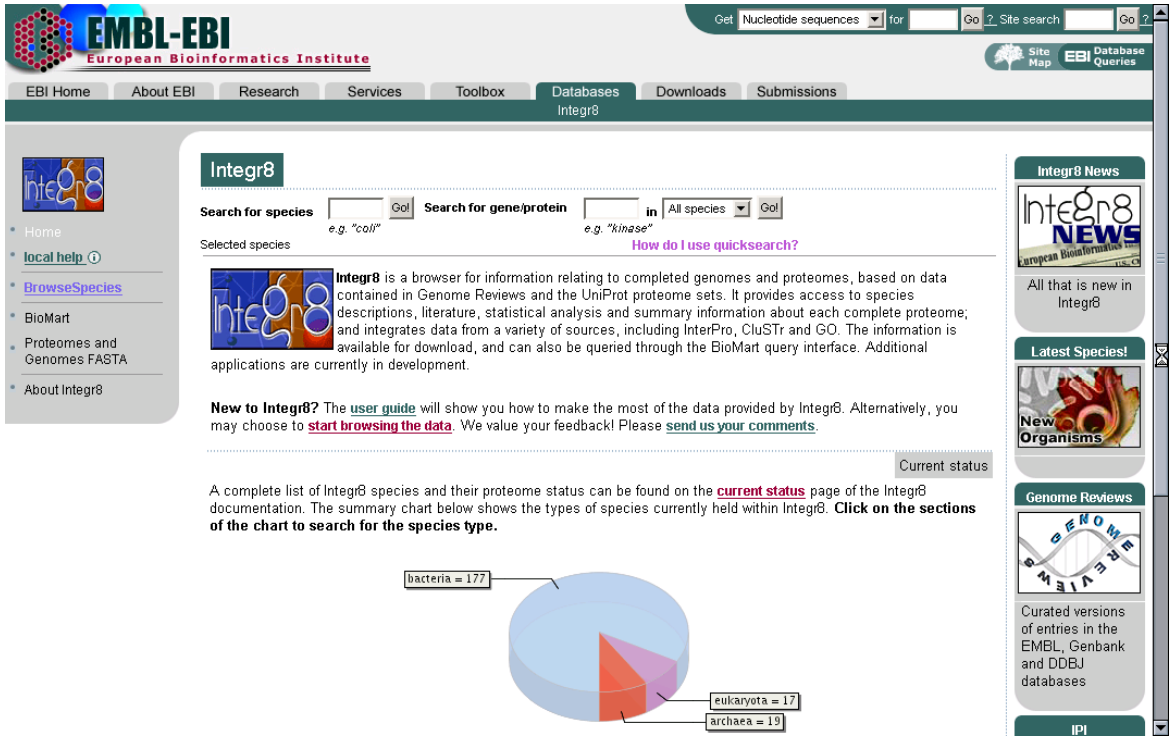

Figure 1: The Integr8 entrance page - users can search for species, genes or proteins, and get further information on whole proteomes and genomes.

In this paper we describe the features the genome and proteome analysis functions provide, the main sequence sources, the querying tool and some intended future developments.

2 Methods

2.1 Proteome Analysis

The Proteome Analysis Database [11] is a tool to analyse proteomes of completely sequenced organisms, comprising bacteria, archaea and eukaryota. The genomes can be translated into predicted protein coding sequences, and with the help of different resources comparative analyses can be performed. The database has been available as a single, already high integrative resource; now it is completely integrated into Integr8 and acts as its backbone.

To analyse proteins and proteomes, several protein information resources are used. One is the InterPro database [12] on protein families, domains and functional sites. InterPro is a powerful tool for describing the physical composition of a complete proteome and for inferring the biological consequences of this. Through Integr8, it is possible to discover the most common domains, families and functional sites present in each proteome, find the proteins with the largest number of different InterPro classifications, and determine the overall coverage of the proteome by InterPro methods (unclassified proteins are usually completely uncharacterised, and therefore are either novel or wrongly predicted). The complete set of all matches between InterPro methods and sequences in each proteome set is also available for download. In addition, precomputed comparisons are available for interesting combinations of species, allowing users to see the differential representation of InterPro matches between them. Additional comparisons can be customised by the user and generated interactively.

Another important tool for the statistical analysis of proteins is the CluSTr database [13], which offers an automatic classification of proteins into groups of related ones. The statistical

measures of protein similarity contained in the CluSTr database are used to support proteome analysis by the independent clustering of proteins in each complete proteome. The user is thus able to identify the largest and tightest clusters of proteins (grouped by similarity) in each species, identify singleton proteins (proteins without paralogues), and identify clusters with no or inconsistent InterPro annotation (which may indicate the presence of a common functional unit not yet detected by any of the InterPro methods). These results can be compared with the results of clustering proteins from all species via the CluSTr website. The absence of proteins from a particular species in a particular multi-species cluster, and the existence of loose clusters containing proteins only from a particular group of species, may give clues as to the functional differentiation of species.

With GO Slim, part of the Gene Ontology (GO) project [14] that describes genes and gene products according to molecular function, biological process and cellular component, a GO-based analysis of each complete proteome is also available. This provides a high-level overview of the functional bias of each proteome, summarising its composition in terms from the GO. To obtain comparable data for each complete proteome, the raw data in the GO annotation database (GOA) corresponding to each proteome set is mapped to a reduced set of high-level (i.e. less-specific) terms, known as GO Slim. Typical GO Slim terms in each of the 3 GO ontologies are 'transcription regulator activity' (function), 'cell cycle' (process), and 'external encapsulating structure' (cellular component). The overall coverage of each proteome in each of the 3 ontologies is provided, together with the percentage of each proteome made up by proteins to which each GO Slim term has been applied. The structure of GO allows a given lower level term may map to more than one GO Slim term (for example, a 'cellular physiological process' is both a 'cellular process' and a 'physiological process'), hence the cumulative proportion of proteins assigned to each high-level category may exceed 100%. In addition to this summary information, the complete set of GO terms assigned to each proteome is also available for download.

Integr8 also provides access to structural data about each proteome; the length distribution of each protein in the proteome and the cumulative amino acid usage are represented graphically. Proteins with known 3-D structures (contained in the Protein Data Bank [15]) from each proteome can be reviewed according to the SCOP superfamily classification [16]. Proteins with predicted 3-D structures inferred from sequence similarity are also available.

2.2 Genome Reviews

Gene sequences and the annotation that is provided by the original sequencing group are stored in EMBL/GenBank/DDBJ collaborating databases that exchange data on a daily basis [5, 17, 18]. EMBL/GenBank/DDBJ are archival databases and serve as repositories of the submitted data. The curation is limited to syntax checking and some very basic check on the integrity of the data. The inevitable consequence of the fact that the databases are archival is the inconsistency of the annotation. To overcome the problem that the archival entries in EMBL/GenBank/DDBJ can not be touched without the submitter's instructions, an additional database, the Genome Reviews, has been created in which the bulk of information is taken from the archival entries and the annotation in the newly created entries is improved, either by performing manual curation or by propagating the information from curated databases. Additional annotation is imported from data sources such as the UniProt knowledgebase, the GOA (GO Annotation) project, InterPro etc. In addition, annotations used inconsistently among the original submissions have been standardised, and deleted in cases where the coverage is low, making it easier to compare data across several genomes.

The Genome Reviews aim at providing comprehensively annotated genomic sequences of organisms with completely deciphered genomes. Enhanced versions of the original

EMBL/Genbank/DDBJ entries are presented in EMBL format, with many cross references, standardised gene/product names and gene/product synonyms, imported from the UniProt Knowledgebase, and systematic locus tag identifiers for nearly all genes, indicating the position of genes on the genome. Data imported from GOA has been used to annotate coding sequences with functional information. The Genome Reviews are used as the source of genome sequences in Integr8 (Figure 2).

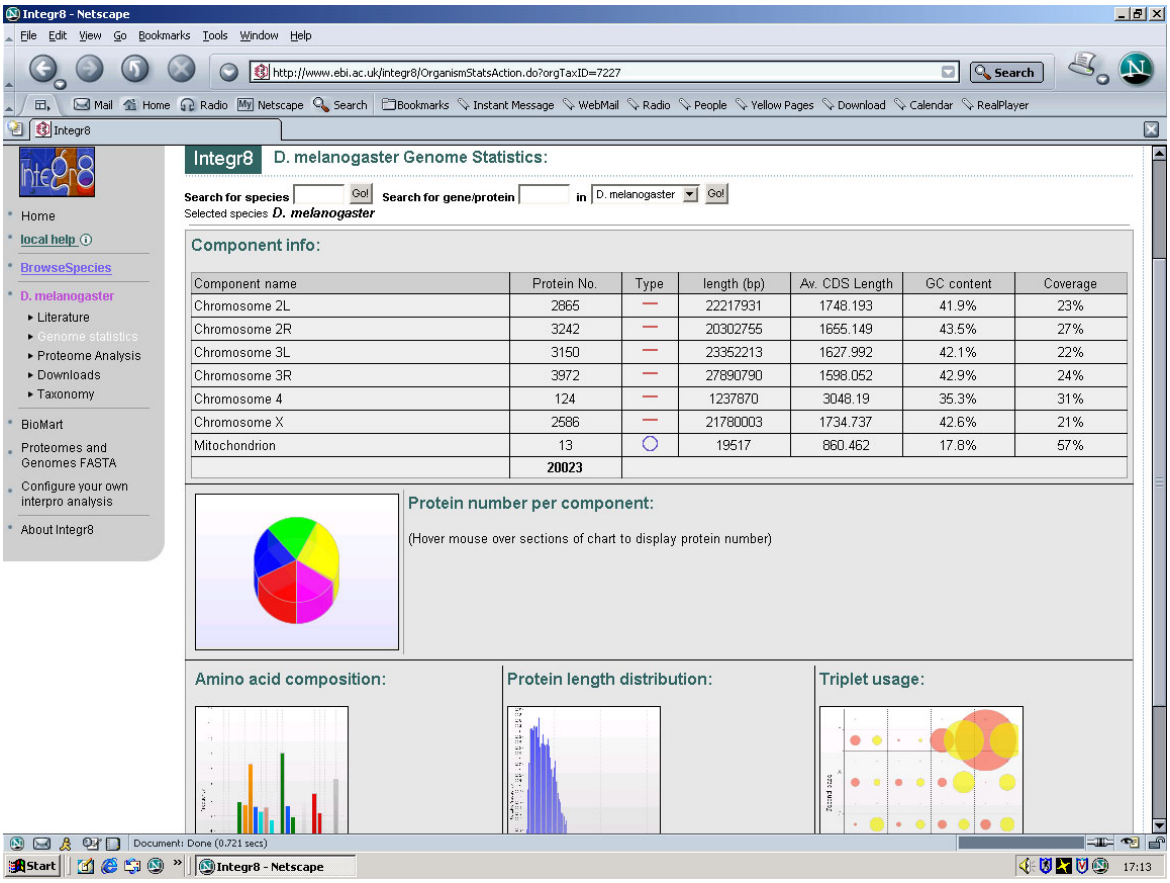

Figure 2: Genome statistics for Drosophila melanogaster. (Average CDS length, GC content and gene coverage are calculated by processing EMBL entries representing components of fully sequenced genomes.)

3 Results

3.1 The Integr8 data model

Integr8 stores information about biological objects or entities, like chromosomes, genes and polypeptides (mostly sequences or sequence features). Stable identifiers for the entities allow for the attachment of cross-references and annotation, and the relationships between them can be described. Links between genome, proteome and phenotype are presented. Web links to the source databases provide further information. Only core data is stored in the automatically populated database, because Integr8 does not attempt to replicate the full schemas of the source databases. A limited number of properties are stored for each entity, mainly name, sequence coordinates and evidence. It is possible to partially update Integr8 when a single source database changes, without the need of a complete recalculation of the database contents.

The Integr8 object model, described in detail in Kersey et al., 2003 [3], is designed to capture data from different molecular biology databases. UML (Universal Modelling Language) has been used to develop the model, and Integr8 is implemented as an n-tier system using a

modern object-oriented programming language (Java). An object-relational tool, OJB (ObJectRelationalBridge), has been used to specify the interface between the upper layers and an underlying relational database. The model also is able to represent richer genomic and proteomic information that is likely to become available in the future.

3.2 BioMart

In addition to the precomputed analyses, a web-based tool is available that allows users to configure their own analysis (according to InterPro classification) of any number of selected proteomes. The UniProt complete proteome sets (and corresponding data from the Genome Reviews) are also available through the BioMart server (http://www.ebi.ac.uk/biomart/). This is an integrated database and query system developed from the EnsMart system [19]. It allows users to generate their own statistical analysis of selected proteomes using combinatorial criteria such as InterPro, CluSTr or GO classification, UniProt annotation or gene location. BioMart also allows users to download customised data sets based on the selected criteria. Additionally, chained queries can be specified that allow the user to link between the proteome sets and other data sets contained within BioMart. The system is query-optimised, and it supports the flexible cross-querying of data from the different primary sources within and between Marts. A variety of interfaces exist including a web-based interface, stand-alone programs, and a Java library. BioMart allows the user to apply a number of filters to the complete data set (comprised of all proteomes), and then to select additionally what attributes of this data set are desired for download or display.

Initially applied only to Ensembl, BioMart is now extended to the UniProt Proteomes, the Macromolecular Structure Database [9], the Vertebrate Genome Annotation (VEGA) browser (http://vega.sanger.ac.uk/), and the Single Nucleotide Polymorphism database (dbSNP) [20]. Data from all major EBI databases, including the EMBL nucleotide sequence database, UniProt, InterPro, CluSTr, GO, MSD and ArrayExpress [21], are imported into the Mart data warehouse. The entire database can be downloaded and installed locally.

4 Discussion

Integr8, a collaboration of twelve of the main European providers of molecular biology data, is offering an overview of complete genomes and proteomes. Information from different sources is integrated, allowing comparisons between genomes and proteomes, links between different levels of data, and a consistent view of this variable data. Currently (as of November 2004), information and analysis of 213 different species is available in Integr8, spanning bacteria, archaea and eukaryota. The number of completely sequenced species is growing rapidly, and Integr8 provides with its object model the facilities to keep pace with these developments. With the BioMart browser, the user can do customised queries on various types of sequence types and annotations for numerous species. Pre-prepared data can be downloaded from the FTP site.

Especially the Genome Reviews database will be subject to a number of developments in the future. Currently, in Genome Reviews the problem that annotations describing DNA sequence features may be out of date or incorrect is addressed, but not the problem that the DNA sequence features themselves may be incorrect or absent. However, a substantial number of gene predictions may not encode real genes - for *Escherichia coli* for example it could be shown that it probably has only approximately 3,800 genes, in contrast to the claimed 4,300 genes, and a similar discrepancy seem to exist for almost all published genomes [22] -, and that other genes have not been described. Therefore methods are going to be developed to map protein sequences not annotated in the original EMBL genome entries (which may

represent corrected versions of originally annotated protein sequences, or novel protein sequences subsequently experimentally determined or predicted by alternative methods) onto their corresponding genomes. In most cases, it is possible to identify a putative sequence in the genomic DNA that encodes this protein, and the reason why the annotation was originally not made. In future releases of Genome Reviews, additional CDSs representing unannotated protein sequences obtained from trusted sources (including, but not limited to, the UniProt Knowledgebase) will be added to the Genome Reviews files, enabling the provision of a consistent view of the genome and proteome of each organism. The possibility of generating new predictions for non-coding RNA genes in a standardised fashion for all genomes is also under investigation. For all new features, sequence discrepancies will be annotated and the use of evidence tags will allow the source of new data to be clearly identified.

Currently, Genome Reviews files are available for all prokaryota, which account for over 80% of the annotated genomes in the public databases. It is planned to extend Genome Reviews to lower metazoan organisms in the near future, but not to higher metazoan species (in the case of the latter, incomplete or unfinished sequences or annotation are frequent problems, and their gene structure is typically more complex and less well determined; their genomes are interpreted in dedicated resources such as Ensembl). For the lower metazoan organisms, Genome Reviews will complement Ensembl.

5 Acknowledgement

Integr8 is supported by the European Commission (EC) TEMBLOR grant (no. QLRI-CT-2001-00015).

6 References

[1] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269:496-512, 1995.

[2] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin and S. G. Oliver. Life with 6000 genes. Science, 274:563-567, 1996.

[3] P. J. Kersey, L. Morris, H. Hermjakob and R. Apweiler. Integr8: Enhanced Inter-Operability of European Molecular Biology Databases. Methods of Information in Medicine, 42:154-160, 2003.

[4] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi and L. L. Yeh. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research, 32:D115-D119, 2004.

[5] T. Kulikova, P. Aldebert, N. Althorpe, W. Baker, K. Bates, P. Browne, A. Van Den Broek, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, M. Garcia-Pastor, N. Harte, C. Kanz, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, P. Stoehr, G. Stoesser, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu and R. Apweiler. The EMBL Nucleotide Sequence Database. Nucleic Acids Research, 32:D27-D30, 2004.

[6] E. Birney, T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, T. Down, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J.

Gilbert, M. Hammond, H.-R. Hotz, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, K. C. Woodwark, G. Cameron, R. Durbin, A. Cox, T. Hubbard and M. Clamp. An Overview of Ensembl. Genome Research 14:925-928, 2004.

[7] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Research, 31:374-378, 2003.

[8] C. D. Schmid, V. Praz, M. Delorenzi, R. Perier and P. Bucher. The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. Nucleic Acids Research, 32:D82-D85, 2004.

[9] A. Golovin, T. J. Oldfield, J. G. Tate, S. Velankar, G. J. Barton, H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Hussain, J. M. C. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, A. Pajon, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, J. Swaminathan, M. Tagari, S. Tromm, W. Vranken and K. Henrick. E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Research, 32:D211-D216, 2004.

[10] L. Duret, D. Mouchiroud and M. Gouy. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Research, 22:2360-2365, 1994.

[11] M. Pruess, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E. Kriventseva, V. Mittard, N. Mulder, I. Phan, F. Servant and R. Apweiler. The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. Nucleic Acids Research, 31:414-417, 2003.

[12] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P.Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. A. Sigrist, R. Vaughan and E. M. Zdobnov. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Research, 31:315-318, 2003.

[13] E. V. Kriventseva, F. Servant and R. Apweiler. Improvements to CluSTr: the database of Swiss-Prot+TrEMBL protein clusters. Nucleic Acids Research, 31:388-389, 2003.

[14] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig. et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 25:25-29, 2000.

[15] J. Westbrook, Z. Feng, L. Chen, H. Yang and H. M. Berman. The Protein Data Bank and structural genomics. Nucleic Acids Research, 31:489-491, 2003.

[16] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia and A. G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Research, 32:D226-D229, 2004.

[17] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler. GenBank: update. Nucleic Acids Research, 32:D23-D26, 2004.

[18] S. Miyazaki, H. Sugawara, K. Ikeo, T. Gojobori and Y. Tateno. DDBJ in the stream of various biological data. Nucleic Acids Research, 32:D31-D34, 2004.

[19] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox and E. Birney. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. Genome Research, 14:160-169, 2004.

[20] D. L. Wheeler, D. M. Church, R. Edgar, S. Federhen, W. Helmberg, T. L. Madden, J. U. PontiusG. D. Schuler, L. M. Schriml, E. Sequeira, T. O. Suzek, T. A. Tatusova and L. Wagner. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Research, 32:D35-D40, 2004.

[21] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra and S. A. Sansone. ArrayExpress - a public repository for microarray gene expression data at the EBI. Nucleic Acids Research, 31:68-71, 2003.

[22] M. Skovgaard, L. J. Jensen, S. Brunak, D. Ussery and A. Krogh. On the total number of genes and their length distribution in complete microbial genomes. Trends in Genetics, 17:425-428, 2001.