Nucleic Acids Research, 2005, Vol. 33, Database issue ***D297–D302***
doi:10.1093/nar/gki039

Integr8 and Genome Reviews: integrated views of complete genomes and proteomes

Paul Kersey*, Lawrence Bower, Lorna Morris, Alan Horne, Robert Petryszak, Carola Kanz, Alexander Kanapin, Ujjwal Das, Karine Michoud[1], Isabelle Phan[1], Alexandre Gattiker[1], Tamara Kulikova, Nadeem Faruque, Karyn Duggan, Peter Mclaren, Britt Reimholz[2], Laurent Duret[3], Simon Penel[3], Ingmar Reuter[4] and Rolf Apweiler

The EMBL Outstation–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [1]Swiss Institute of Bioinformatics, CMU, 1 rue Michel Servet, 1211 Geneva 4, Switzerland, [2]RZPD Deutsches Ressourcenzentrum für Genomforschung GmbH, Heubnerweg 6, 14059 Berlin, Germany, [3]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 Université Lyon 1, France and [4]BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbuttel, Germany

Received September 13, 2004; Revised September 17, 2004; Accepted September 27, 2004

ABSTRACT

Integr8 is a new web portal for exploring the biology of organisms with completely deciphered genomes. For over 190 species, Integr8 provides access to general information, recent publications, and a detailed statistical overview of the genome and proteome of the organism. The preparation of this analysis is supported through Genome Reviews, a new database of bacterial and archaeal DNA sequences in which annotation has been upgraded (compared to the original submission) through the integration of data from many sources, including the EMBL Nucleotide Sequence Database, the UniProt Knowledgebase, InterPro, CluSTr, GOA and HOGENOM. Integr8 also allows the users to customize their own interactive analysis, and to download both customized and prepared datasets for their own use. Integr8 is available at http://www.ebi.ac.uk/integr8.

INTRODUCTION

Since the advent of whole genome sequencing in the mid-1990s, the sequences of over 190 cellular organisms have been completely determined, annotated and deposited in the public repositories. The rate of deposition of such sequences is still increasing, with over 90 such genomes sequenced and made available since March 2003. The availability of these data has enabled the development of new ways to interpret information about individual genes and proteins in their biological context, and has underpinned the development of new experimental and theoretical fields such as transcriptomics, proteomics and systems biology. However, these new technologies have generated enormous quantities of data, meaning that the information needed to draw scientific conclusions is increasingly likely to be spread over many different primary resources, which do not necessarily maintain common identifiers for the data items they describe, or even agree on the definition of common terms (1). Coherently integrating such data, and offering access to it, has thus emerged as one of the most important challenges in bioinformatics. We have addressed these problems by releasing a new database, Genome Reviews, in which updated annotation is added to genomic sequence data; and a new web interface, Integr8, offering interactive access to data integrated from Genome Reviews and other resources, centred around organisms with completely sequenced genomes.

GENOME REVIEWS

Motivation

The International Nucleotide Sequence Database, a collaboration among the EMBL (2), GenBank (3) and DDBJ (4) nucleotide sequence databases, is the usual primary public repository for DNA and RNA sequence and annotation, including completed genome sequences. As repositories, these databases allow submitters to retain ownership of their own data and in consequence, annotation of different entries is often not standardized in format or update frequency (and as annotation for predicted genes is often inferred by similarity to other sequences, information can become out-of-date simply by the submission of additional entries to the databases).

*To whom correspondence should be addressed. Tel: +44 1223 494601; Fax: +44 1223 494468; Email: pkersey@ebi.ac.uk

Additionally, theoretical annotation inferred from a sequence (that is commonly present in database submissions) is usually not well integrated with data derived from laboratory experiments. These issues can only be addressed through active curation of database entries; but information introduced into well-annotated resources such as the UniProt Knowledgebase (5) cannot be incorporated into the archive genome sequence record except at the instigation of the submitter. Improved genome annotation has been made available by RefSeq (6). However, these data use their own identifiers and do not necessarily contain cross-references to databases such as EMBL or UniProt.

Therefore, we have launched Genome Reviews, to make genome sequence available with standardized, up-to-date annotation while maintaining cross-references to the primary submission (and to entries in other databases with cross-references to it). To ensure compatibility with existing tools, Genome Reviews is distributed in an extended version of the flat file format used by EMBL. The initial scope of Genome Reviews is prokaryotic genomes. Release 7 (made on August 4, 2004) contains files for 187 chromosomes and 105 plasmids, representing the complete genomes of 170 species.

Propagation of data to Genome Reviews

EMBL entries describe nucleotide sequences, features (annotated regions of sequence) and feature qualifiers (individual annotations attached to a feature). Additionally, there is also some annotation that is attached to the database entry itself as opposed to the sequence (e.g. the database accession number). The 'CDS' (CoDing Sequence) feature is used to identify subsequences within the overall DNA sequence that encode a protein sequence (proteins are the most widely annotated biological entity); the '/db_xref' qualifier indicates cross-references to entries in other databases. The '/protein_id' qualifier uniquely identifies each CDS annotated in the entry.

There are therefore three ways in which an entry in another database can be identified as referring to the same biological entity as a given EMBL (CDS) feature: if the EMBL feature cross-references that entry, if that entry cross-references the EMBL feature or if an entry in a third database cross-references both other entries. By tracking identifiers between databases, additional annotation belonging to a feature can be identified. The UniProt Knowledgebase (5), a well-annotated resource in which redundant submissions are merged, is a particularly useful database hub for retrieving annotation and cross-links to further resources.

To produce Genome Reviews, a particular preferred source is nominated for each type of annotation; and annotation of that type is imported from that source into Genome Reviews either as a supplement to or as a replacement for the annotation in the original submission. Where more than one resource may provide annotation of a certain type (e.g. gene names), redundant data are case-standardized and merged.

The annotation attached to other types of features (e.g. non-coding RNAs) has also been standardized, and redundant or rarely used features and feature qualifiers removed. In addition to the insertion/deletion of feature qualifiers associated with existing features in the original submission, new features have been added (for example, regions of DNA encoding the mature peptides produced after proteolytic cleavage of the primary translations) by mapping features annotated on protein sequences onto corresponding regions of DNA. CDSs identified by UniProt curators as false (i.e. unlikely to encode a real protein) have been removed.

Table 1. Incorporation of new data and data types in Genome Reviews (compared with parent entries in the EMBL nucleotide sequence database)

	Original EMBL entries	Genome Reviews entries
Number of feature types	30	11
Number of qualifier types	42	28
Number of feature qualifiers	4 649 864	6 783 847
Number of external databases cross-referenced	6	18
Number of 'mat_peptide' features	0	3825
Number of '/db_xref' qualifiers	631 881	2 527 269
Number of '/locus_tag' qualifiers	367 771	384 899
Number of evidence tags	0	5 474 235

The table shows how the total quantity of annotation has been increased (with some examples), while the number of feature and feature qualifier types has been reduced, with the remaining types used more consistently across all entries. Statistics were compiled by comparing Genome Reviews release 7.0 with EMBL release 79, incrementally updated to August 10, 2004.

Genome Reviews has been implemented using the Java programming language in conjunction with a relational database management system. An extension of the open source BioJava (7) EMBL parser has been used to prepare the files for distribution.

Content and format changes

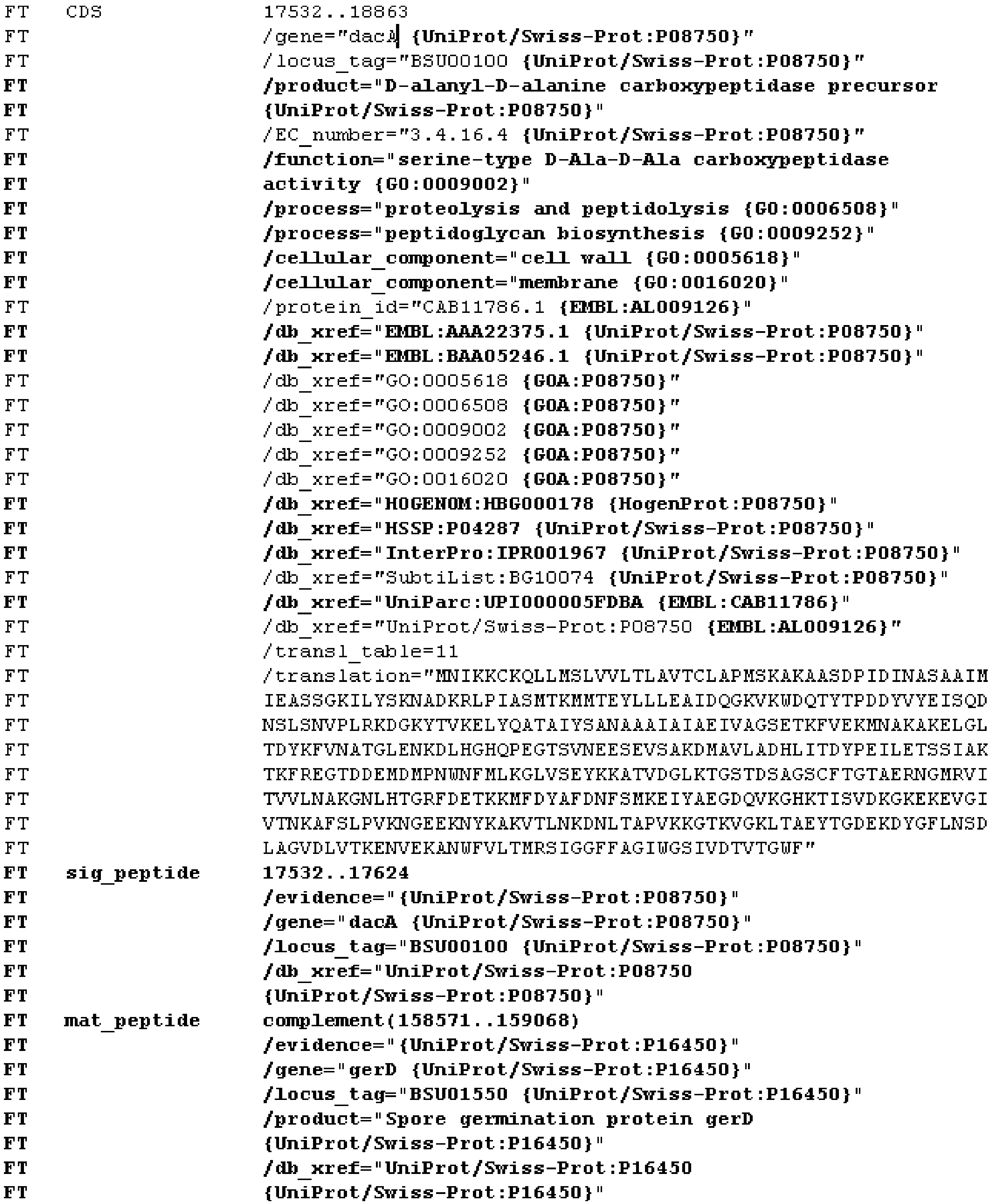
With the propagation of data from multiple sources into Genome Reviews, the EMBL flat file format has been extended to support the provision of clear evidence regarding the origins of each piece of data included in the file by the addition of evidence tags to feature qualifiers. Each evidence tag consists of the name of the database and where appropriate, the identifier of an entry within that database, from which data have been sourced. Additionally, new types of features and feature qualifiers have been introduced to describe imported data not previously present in EMBL entries. In spite of this, the number of different types of annotation has been reduced owing to the standardization of representation and the removal of redundant annotation. Some statistics indicating how Genome Reviews has been enhanced compared to the original submissions are given in Table 1. For example, the number of cross-references to other databases has been increased 4-fold and the number of cross-referenced databases 3-fold.

The effect of these changes on (as part of) an individual EMBL entry can be seen in Figure 1, which illustrates the introduction of new feature and feature qualifier types, new data (added using existing EMBL features and feature qualifiers) and the use of evidence tags.

INTEGR8

Aims and data sources

The Integr8 portal offers an overview of information about organisms with completely sequenced genomes; statistical analyses of their genomes and proteomes, individually



```
FT   CDS             17532..18863
FT                   /gene="dacA {UniProt/Swiss-Prot:P08750}"
FT                   /locus_tag="BSU00100 {UniProt/Swiss-Prot:P08750}"
FT                   /product="D-alanyl-D-alanine carboxypeptidase precursor
FT                   {UniProt/Swiss-Prot:P08750}"
FT                   /EC_number="3.4.16.4 {UniProt/Swiss-Prot:P08750}"
FT                   /function="serine-type D-Ala-D-Ala carboxypeptidase
FT                   activity {GO:0009002}"
FT                   /process="proteolysis and peptidolysis {GO:0006508}"
FT                   /process="peptidoglycan biosynthesis {GO:0009252}"
FT                   /cellular_component="cell wall {GO:0005618}"
FT                   /cellular_component="membrane {GO:0016020}"
FT                   /protein_id="CAB11786.1 {EMBL:AL009126}"
FT                   /db_xref="EMBL:AAA22375.1 {UniProt/Swiss-Prot:P08750}"
FT                   /db_xref="EMBL:BAA05246.1 {UniProt/Swiss-Prot:P08750}"
FT                   /db_xref="GO:0005618 {GOA:P08750}"
FT                   /db_xref="GO:0006508 {GOA:P08750}"
FT                   /db_xref="GO:0009002 {GOA:P08750}"
FT                   /db_xref="GO:0009252 {GOA:P08750}"
FT                   /db_xref="GO:0016020 {GOA:P08750}"
FT                   /db_xref="HOGENOM:HBG000178 {HogenProt:P08750}"
FT                   /db_xref="HSSP:P04287 {UniProt/Swiss-Prot:P08750}"
FT                   /db_xref="InterPro:IPR001967 {UniProt/Swiss-Prot:P08750}"
FT                   /db_xref="SubtiList:BG10074 {UniProt/Swiss-Prot:P08750}"
FT                   /db_xref="UniParc:UPI000005FDBA {EMBL:CAB11786}"
FT                   /db_xref="UniProt/Swiss-Prot:P08750 {EMBL:AL009126}"
FT                   /transl_table=11
FT                   /translation="MNIKKCKQLLMSLVVLTLAVTCLAPMSKAKAASDPIDINASAAIM
FT                   IEASSGKILYSKNADKRLPIASMTKMMTEYLLLEAIDQGKVKWDQTYTPDDYVYEISQD
FT                   NSLSNVPLRKDGKYTVKELYQATAIYSANAAAIAIAEIVAGSETKFVEKMNAKAKELGL
FT                   TDYKFVNATGLENKDLHGHQPEGTSVNEESEVSAKDMAVLADHLITDYPEILETSSIAK
FT                   TKFREGTDDEMDMPNWNFMLKGLVSEYKKATVDGLKTGSTDSAGSCFTGTAERNGMRVI
FT                   TVVLNAKGNLHTGRFDETKKMFDYAFDNFSMKEIYAEGDQVKGHKTISVDKGKEKEVGI
FT                   VTNKAFSLPVKNGEEKNYKAKVTLNKDNLTAPVKKGTKVGKLTAEYTGDEKDYGFLNSD
FT                   LAGVDLVTKENVEKANWFVLTMRSIGGFFAGIWGSIVDTVTGWF"
FT   sig_peptide     17532..17624
FT                   /evidence="{UniProt/Swiss-Prot:P08750}"
FT                   /gene="dacA {UniProt/Swiss-Prot:P08750}"
FT                   /locus_tag="BSU00100 {UniProt/Swiss-Prot:P08750}"
FT                   /db_xref="UniProt/Swiss-Prot:P08750
FT                   {UniProt/Swiss-Prot:P08750}"
FT   mat_peptide     complement(158571..159068)
FT                   /evidence="{UniProt/Swiss-Prot:P16450}"
FT                   /gene="gerD {UniProt/Swiss-Prot:P16450}"
FT                   /locus_tag="BSU01550 {UniProt/Swiss-Prot:P16450}"
FT                   /product="Spore germination protein gerD
FT                   {UniProt/Swiss-Prot:P16450}"
FT                   /db_xref="UniProt/Swiss-Prot:P16450
FT                   {UniProt/Swiss-Prot:P16450}"
```

Figure 1. Incorporation of new data into Genome Reviews. The figure shows a portion of Genome Reviews entry AL009126_GR from release 7.0. Data in boldface have been added to the corresponding portion of the original submission to the EMBL/GenBank/DDBJ nucleotide sequence databases.

and in comparison with each other, using various data from multiple resources; interactive interfaces allowing users to configure their own analyses; and access to the underlying data for download. Integr8 is built on three main data sources:

(i) Genome Reviews.
(ii) Non-redundant sets of UniProt entries representing each complete proteome. For prokaryotic organisms, these are constructed according to the HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes) specifications (8) used in the annotation of the UniProt Knowledgebase. Sets are also available for eukaryotic organisms, prepared by filtering the UniProt Knowledgebase using information from EMBL and model organism databases such as FlyBase (9). In some species (like human) where the level of multiple submissions is very high, additional entries are filtered out according to their sequence similarity.
(iii) IPI (the International Protein Index) (10) provides comprehensive protein sets for certain higher metazoan species, by combining data from the UniProt Knowledgebase, Ensembl (11) and RefSeq (6) in a non-redundant fashion.

For each proteome set, additional information has been integrated from other resources such as HAMAP (8), InterPro (12), CluSTr (13), the Gene Ontology Annotation database (GOA) (14) and others. A full list of resources available through Integr8 is provided in Table 2.

Gene and species search

A simple search form provides access to summary information relevant to each species. This information includes a description, a list of recent publications, a list of the components of its genome and information about the composition of these components (such as their length, average GC content and the length and codon usage in the CDSs they contain), represented textually or graphically as is appropriate. A typical page displaying such an analysis is shown in Figure 2.

The search facility can also be used to search for proteins belonging to the non-redundant proteome sets, and the genes that encode them.

Proteome Analysis

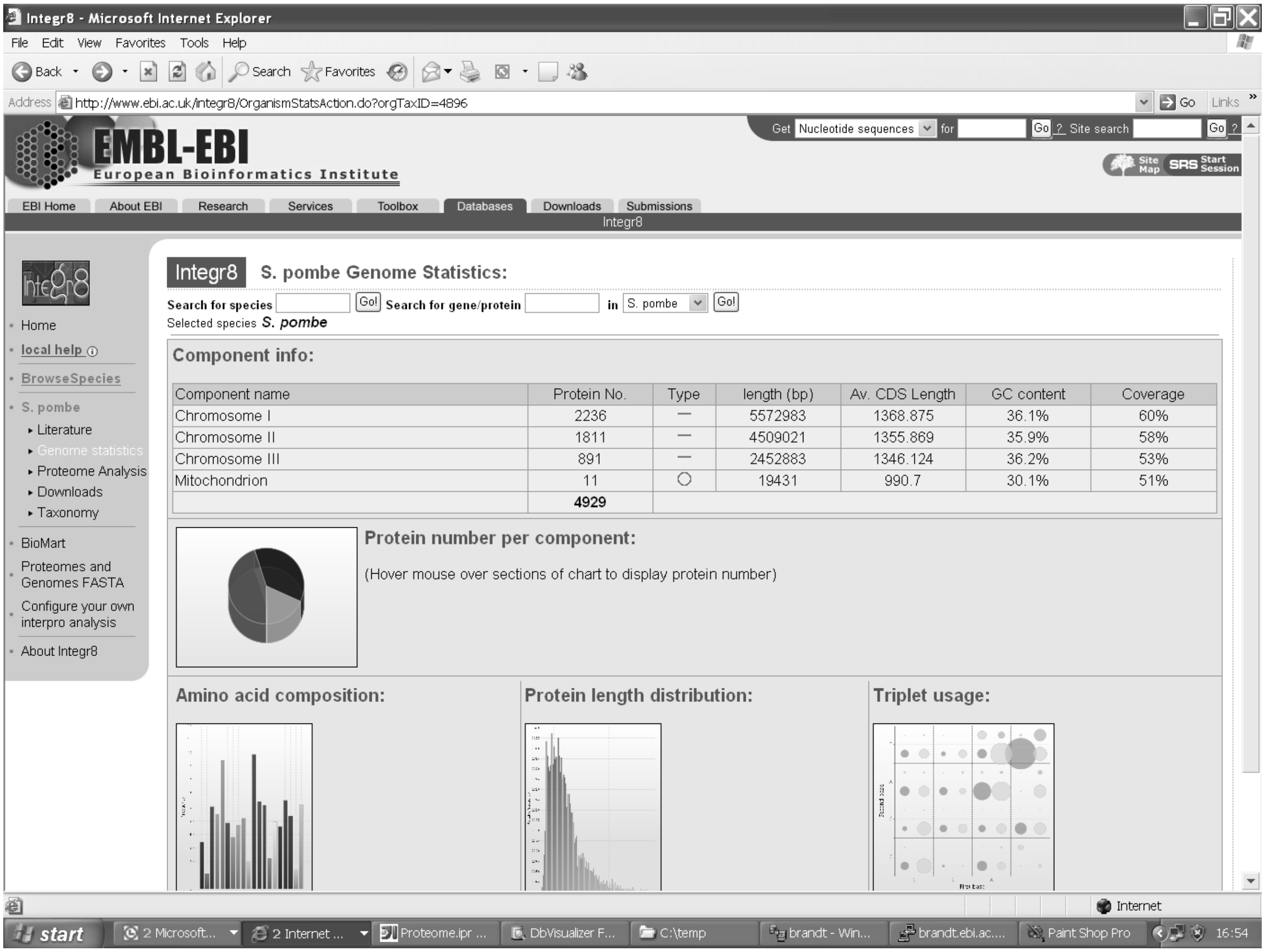
Integr8 has incorporated the Proteome Analysis Database (15), to provide information about the composition of complete proteomes. Individual proteins are classified according to InterPro (12), GO (16) and CluSTr (13), and an overview of the composition of each proteome constructed on these criteria is available. For example, for each proteome, users can identify the most common protein families and domains, proteins without close relatives, or clusters of related proteins unclassified by InterPro. GO classifications for each species are summarized using a reduced set of high-level terms (GO Slim), presenting an overview of proteome function even in species where more specific annotations might not be available. A major advance in the past year has been the doubling in the coverage of CluSTr, a database that categorizes proteins into a hierarchy of clusters based on overall sequence similarity. Individual hierarchies of clusters have now been prepared for each of 109 proteomes, enabling the relationship of all paralogous proteomes to be analysed in these species. Additional structural data are also available based on information derived from the Protein Data Bank (17) and HSSP (18).

Integr8 also offers comparative analysis, whereby the composition of multiple proteomes can be compared. A total of 160 comparative analyses (each featuring between two and four related species) have been pre-compiled; additional comparisons can be specified interactively.

Users can also configure their own comparative analysis in two ways. First, it is possible to configure a multi-species analysis based on InterPro classifications (19). Second, the data from Integr8 have been loaded into an additional search tool, BioMart, a development of the EnsMart data

Table 2. Resources integrated in Integr8

Database	Brief description of content/purpose	URL
CleanEx (26)	Gene expression data	http://www.cleanex.isb-sib.ch
CluSTr (13)	Clusters of proteins with similar sequences	http://www.ebi.ac.uk/clustr
EMBL nucleotide sequence database (2)	Nucleotide sequences and annotation	http://www.ebi.ac.uk/embl
Ensembl (11)	Predictions of gene structure and protein sequence	http://www.ensembl.org
Eukaryotic Promoter Database (27)	Promoters	http://www.epd.isb-sib.ch
Genome Reviews	Genome sequences with upgraded annotation	http://www.ebi.ac.uk/GenomeReviews
Gene Ontology (16)	Gene product classification hierarchy	http://www.geneontology.org
GOA (14)	GO annotations for proteins	http://www.ebi.ac.uk/GOA
HAMAP (8)	Bacterial gene families and annotation specifications	http://www.expasy.org/sprot/hamap
HOGENOM	Phylogenetically analysed protein clusters	http:/pbil.univ-lyon1.fr/databases/hogenom.html
HSSP (18)	Tertiary structure inferred from secondary structure	http://www.cmbi.kun.nl/gv/hssp
InterPro (12)	Protein domains, families and repeats	http://www.ebi.ac.uk/interpro
IPI (10)	Protein sequences	http://www.ebi.ac.uk/IPI
PDB (17)	Macromolecular structures	http://www.wwpdb.org
ReAlSplice	Splice sites and events	http://realsplice.bioinf.med.uni-goettingen.de
RefSeq (6)	Nucleotide and protein sequences and annotation	http://www.ncbi.nlm.gov/projects/RefSeq/
S/MARt Db	Scaffold/matrix attachment regions	http:/.smartdb.bioinf.med.uni-goettingen.de
UniProt Archive (5)	Protein sequences	http://www.ebi.ac.uk/uniparc
UniProt Knowledgebase (5)	Protein sequences and annotation	http://www.uniprot.org
RZPD Clone Database	DNA clones	http://www.rzpd.de
TRANSFAC (28)	Transcription factors	http://www.gene-regulation.com
UTRdb (29)	UTRs	http://bighost.area.ba.cnr.it/BIG/UTRHome


Integr8 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ebi.ac.uk/integr8/OrganismStatsAction.do?orgTaxID=4896
Go Links
Get Nucleotide sequences for Go ? Site search Go ?
EMBL-EBI
European Bioinformatics Institute
Site Map SRS Start Session
EBI Home About EBI Research Services Toolbox Databases Downloads Submissions
Integr8
Home
local help
BrowseSpecies
S. pombe
Literature
Genome statistics
Proteome Analysis
Downloads
Taxonomy
BioMart
Proteomes and Genomes FASTA
Configure your own interpro analysis
About Integr8
Search for species Go! Search for gene/protein in S. pombe Go!
Internet
start 2 Microsoft... 2 Internet ... Proteome.ipr ... DbVisualizer F... C:\temp brandt - Win... brandt.ebi.ac.... Paint Shop Pro 16:54

Integr8 **S. pombe Genome Statistics:**

Selected species *S. pombe*

Component info:

Component name	Protein No.	Type	length (bp)	Av. CDS Length	GC content	Coverage
Chromosome I	2236	—	5572983	1368.875	36.1%	60%
Chromosome II	1811	—	4509021	1355.869	35.9%	58%
Chromosome III	891	—	2452883	1346.124	36.2%	53%
Mitochondrion	11	○	19431	990.7	30.1%	51%
	4929					

Protein number per component:

(Hover mouse over sections of chart to display protein number)

Amino acid composition:

Protein length distribution:

Triplet usage:

Figure 2. Genome Statistics for the fission yeast *Schizosaccharomyces pombe*, as represented in the Integr8 browser.

warehousing system (20). BioMart provides the ability to search complete proteomes and genomes using combinatorial criteria, and to customize matching data for download. BioMart also supports interactive querying between the Integr8 data and other resources, such as ArrayExpress (21), Ensembl (11) and the European Macromolecular Structure Database (22).

Downloads

The following data are available for download from Integr8: (i) Genome Reviews files; (ii) UniProt complete proteome sets; (iii) IPI datasets; (iv) Files of InterPro matches for each proteome set; and (v) 'Chromosome tables', summary files mapping proteins represented in UniProt to their genomic locations. As noted above, users can additionally customize their own data for download through BioMart.

FUTURE DEVELOPMENTS

In Genome Reviews, we address the problem of annotations describing DNA sequence features being out of date or incorrect; but not the problem that the DNA sequence features themselves may be incorrect or absent. However, the use of techniques such as statistical (23) or proteomic (24) analysis has indicated that a substantial number of gene predictions may not encode real genes [in the case of one genome, the number of false CDSs has been estimated at 50% (25)], and that other genes have not been described. We are therefore developing methods 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 (by sequence similarity searching) a putative sequence in the genomic DNA that encodes this protein (and also to describe any difference between its translation and the actual protein sequence, thereby explaining 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. We are also investigating the possibility of generating new predictions for non-coding RNA genes in a standardized fashion for all genomes. 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.

Genome Reviews files are currently available for all prokaryota, which account for over 80% of the annotated genomes currently in the public databases. It is planned to extend Genome Reviews to lower metazoan organisms in the near future. In the case of higher metazoan species, different types of problem are typically encountered, such as incomplete or unfinished sequence or annotation (as opposed to the complete, but out-of-date, information often associated with data from simpler species); and gene structure is typically more complex and less well-determined. The best current interpretation of such genomes can be found in dedicated resources such as Ensembl (11). Genome Reviews will complement Ensembl, and will not extend to higher metazoan species.

AVAILABILITY

Updated versions of Genome Reviews and Integr8 are released on a bi-weekly schedule, in synchrony with releases of the UniProt Knowledgebase (5). Genome Reviews is available from its own website (http://www.ebi.ac.uk/GenomeReviews) or through the Integr8 site (http://www.ebi.ac.uk/integr8). Files can also be downloaded via the respective FTP sites (ftp://ftp.ebi.ac.uk/pub/databases/genome_reviews, ftp://ftp.ebi.ac.uk/pub/databases/integr8). BioMart is available at http://www.ebi.ac.uk/biomart.

ACKNOWLEDGEMENTS

This work has been funded by the award of grant number QLRI-CT-2001000015 from the European Union under the RTD program 'Quality of Life and Management of Living Resources'.

REFERENCES

1. Kersey,P.J., Morris,L., Hermjakob,H. and Apweiler,R. (2003) Integr8: enhanced inter-operability of European molecular biology databases. *Methods Inf. Med.*, **42**, 154–160.
2. Kulikova,T., Aldebert,P., Althorpe,N., Baker,W., Bates,K., Browne,P., van den Broek,A., Cochrane,G., Duggan,K., Eberhardt,R. *et al.* (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **32**, D27–D30.
3. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
4. Miyazaki,S., Sugawara,H., Ikeo,K., Gojobori,T. and Tateno,Y. (2004) DDBJ in the stream of various biological data. *Nucleic Acids Res.*, **32**, D31–D34.
5. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
6. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI reference sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
7. Mangalam,H. (2002) The Bio* toolkits—a brief overview. *Brief Bioinformatics*, **3**, 296–302.
8. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
9. FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.
10. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
11. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
12. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
13. Kriventseva,E.V., Servant,F. and Apweiler,R. (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Res.*, **31**, 388–389.
14. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
15. Pruess,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N., Phan,I., Servant,F. *et al.* (2003) The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. *Nucleic Acids Res.*, **31**, 414–417.
16. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
17. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.*, **10**, 980.
18. Dodge,C., Schneider,R. and Sander,C. (1998) The HSSP database of protein structure–sequence alignments and family profiles. *Nucleic Acids Res.*, **26**, 313–315.
19. Kanapin,A., Apweiler,R., Biswas,M., Fleischmann,W., Karavidopoulou,Y., Kersey,P., Kriventseva,E.V., Mittard,V., Mulder,N., Oinn,T. *et al.* (2002) Interactive InterPro-based comparisons of proteins in whole genomes. *Bioinformatics*, **18**, 374–375.
20. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
21. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
22. Boutselakis,H., Dimitropoulos,D., Fillon,J., Golovin,A., Henrick,K., Hussain,A., Ionides,J., John,M., Keller,P.A., Krissinel,E. *et al.* (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.*, **31**, 458–462.
23. Larsen,T.S. and Krogh,A. (2003) EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.
24. Jaffe,J.D., Stange-Thomann,N., Smith,C., DeCaprio,D., Fisher,S., Butler,J., Calvo,S., Elkins,T., FitzGerald,M.G., Hafez,N. *et al.* (2004) The complete genome and proteome of Mycoplasma mobile. *Genome Res.*, **14**, 1447–1461.
25. Skovgaard,M., Jensen,L.J., Brunak,S., Ussery,D. and Krogh,A. (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.*, **17**, 425–428.
26. Praz,V., Jagannathan,V. and Bucher,P. (2004) CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature. *Nucleic Acids Res.*, **32**, D542–D547.
27. Schmid,C.D., Praz,V., Delorenzi,M., Perier,R. and Bucher,P. (2004) The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension. *Nucleic Acids Res.*, **32**, D82–D85.
28. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
29. Pesole,G., Liuni,S., Grillo,G., Licciulli,F., Mignone,F., Gissi,C. and Saccone,C. (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.*, **30**, 335–340.