Full text provided by www.sciencedirect.com

Automated bacterial genome analysis and annotation

Paul Stothard[1] and David S Wishart[1,2]

More than 300 bacterial genome sequences are publicly available, and many more are scheduled to be completed and released in the near future. Converting this raw sequence information into a better understanding of the biology of bacteria involves the identification and annotation of genes, proteins and pathways. This processing is typically done using sequence annotation pipelines comprised of a variety of software modules and, in some cases, human experts. The reference databases, computational methods and knowledge that form the basis of these pipelines are constantly evolving, and thus there is a need to reprocess genome annotations on a regular basis. The combined challenge of revising existing annotations and extracting useful information from the flood of new genome sequences will necessitate more reliance on completely automated systems.

Addresses
[1] Departments of Biological Sciences & Computing Science, University of Alberta
[2] National Research Council, National Institute for Nanotechnology (NINT) Edmonton, Alberta, Canada T6G 2E8

Corresponding author: Wishart, David S (david.wishart@ualberta.ca)

Current Opinion in Microbiology 2006, **9**:505–510

This review comes from a themed issue on
Genomics
Edited by David W Ussery and Timothy D Read

Available online 22nd August 2006

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2006.08.002

Introduction

Bacterial genome sequences are exciting for a variety of reasons. Lurking within the strings of letters are the details of the proteins and pathways that enable bacteria to metabolize numerous compounds, inhabit a diverse range of environments, infect other organisms and share genetic material [1•]. Unraveling these details could potentially lead to the development of novel vaccines, the creation of useful antimicrobial compounds and the design of innovative strategies to modify bacteria for applications such as bioremediation [2,3•,4].

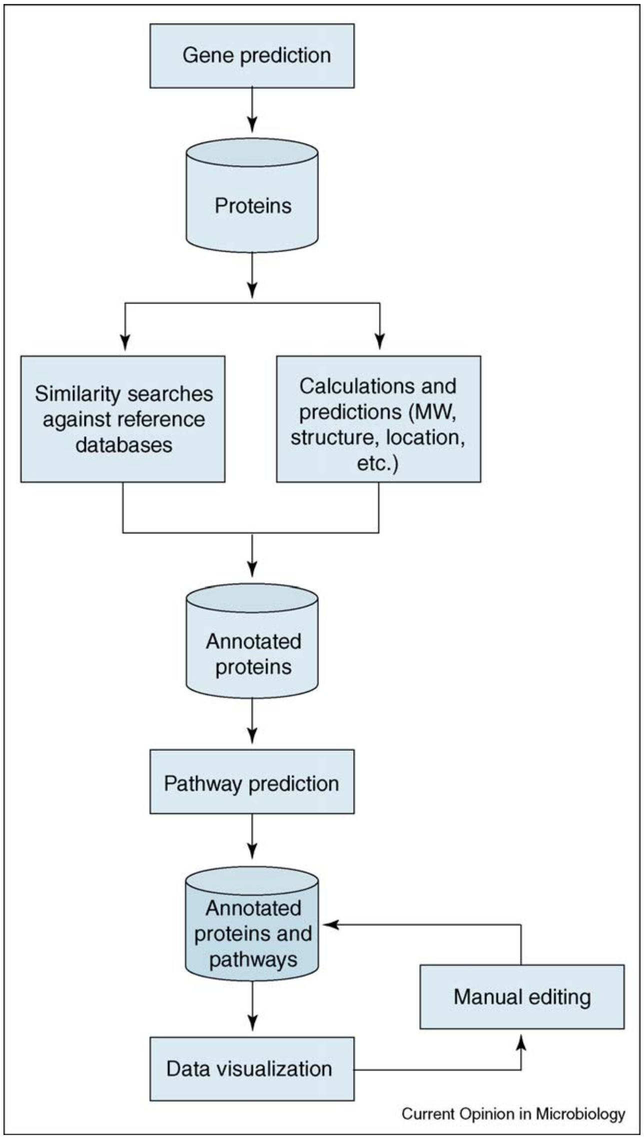
Usually the first step in interpretation of a genome is to use gene-prediction programs, which scan the sequence for regions that are likely to encode proteins or functional RNA products, depending on the particular program (Figure 1). The identified genes are then compared to databases of DNA or protein sequences in an attempt to identify related sequences. If hits of a certain similarity are identified, information about their function is transferred to the new sequence. In addition to general predictions of gene and protein function, annotation pipelines can add several other types of information to genes and genomes, such as protein chemical properties, protein structural properties, predicted operons, gene ontologies, evolutionary relationships and metabolic pathways. These annotations are obtained using sequence similarity searches, calculated from the predicted gene and protein sequences directly, or derived from comparisons of gene order between species. The final and most frequently overlooked step of the annotation process is to organize and present the results in a useful manner.

The degree to which this genome annotation procedure is automated varies. Some systems are completely automatic [5•], whereas others present problematic cases (e.g. genes for which there is a low similarity hit) to human curators, who must then decide on the appropriate action to take [6]. Most of the latest systems reach a compromise by generating the annotation set in an automated fashion and then allowing for manual revisions [7•,8,9]. Here we discuss several of the more recently developed software tools and websites that can be used to obtain bacterial genome annotations; these tools and websites are summarized in Table 1.

Annotation systems are constantly improving

When a genome is annotated, some of the predicted genes might not be similar to anything in the reference databases because they have diverged extensively from their relatives, they represent a novel uncharacterized sequence, or because they are random open reading frames that have been misidentified as genes. In cases where predictions are made, these could be incorrect because sequence similarity does not always imply functional similarity, or because the reference databases contain incorrect annotations. Even when a true orthologue is identified, there might be little in terms of functional information associated with it in the reference database. These challenges are continually being addressed through the development of more completely annotated databases, better gene prediction algorithms and more sensitive sequence comparison methods. Furthermore, new experimentally derived functional information is constantly being generated. For these reasons the most useful annotations are usually those derived most recently.

Figure 1


Gene prediction
Proteins
Similarity searches against reference databases
Calculations and predictions (MW, structure, location, etc.)
Annotated proteins
Pathway prediction
Annotated proteins and pathways
Manual editing
Data visualization
Current Opinion in Microbiology

A flowchart depicting the general procedure used to annotate bacterial genome sequences.

Choosing an annotation resource

For genomes that have already been released, annotations can be obtained from several online databases, maintained by dedicated research groups that continually reprocess bacterial genomes using custom annotation pipelines. In the case of newly sequenced genomes or sequence fragments, some of these same groups, as well as some others, offer web-based services for analyzing and annotating bacterial genomes. Alternatively, there are several complete genome annotation systems that can be downloaded and run locally. Each of these resources differs in terms of the particular annotation strategies used, what types of annotations are available, how the annotations are presented, and how much manual editing can be performed. Investigators looking for more control over the annotation process can also construct their own system using many freely available analysis modules and databases. Assembling an annotation pipeline from existing gene predictors, sequence comparison programs and databases requires the appropriate computing resources and bioinformatics expertise. Even the installation of a pre-built pipeline often requires some custom programming or, at the very least, a significant amount of software configuration or compiling. We explore each of these options in more detail below.

Databases of annotated genomes

When a genome sequence is deposited into the primary sequence databases (GenBank, EMBL and DDBJ) it usually includes a basic set of annotations in the form of predicted genes and protein functions. The quality and completeness of the annotation varies, depending on the particular programs and databases used by the sequence submitters. To a large extent this depends on when the sequence was deposited, as newer projects can take advantage of more complete databases and more advanced annotation software. To address these issues, several groups continually generate and update their own annotations for bacterial genomes. Some sites focus on the providing extremely detailed information for members of a particular species, such as *Escherichia coli* [10–12,13•]. These specialized resources often include extensive sets of experimentally determined annotations, literature citations and other manually added information not found elsewhere. For example, EcoCyc provides literature-derived annotations for *E. coli* [13•], and EcoGene uses evidence from a variety of sources to produce more accurate translation start sites [10]. A middle ground between these species-specific resources and the more comprehensive sites are those containing genomes from particular taxonomic groups [14,15•,16], or those containing genomes sequenced at specific sequencing centers [8,9]. Some of these sites also contain experimental data and many allow members of a particular research community to manually review and edit the genome annotations. Most bacterial genomes are not found in species-specific or family-specific databases, but rather are found in just over a half-dozen comprehensive microbial genome sites such as HAMAP [6], PUMA2 [7•], IMG [17], Entrez Genome [18••], PEDANT [19•], the Comprehensive Microbial Resource [20], MicrobesOnline [21] or BacMap [22]. The annotation systems used by these sites have the advantage of being tested on numerous genomes, and are often more refined than the systems built for one or a few genomes. The sheer quantity of information in these databases makes manual curation somewhat more difficult, although not infeasible: the HAMAP project relies on human curators to review problematic cases presented by the automatic annotation software [6]. In addition to providing a more consistent, comprehensive, and up-to-date set of annotations, all of these sites provide

Table 1

Software and databases that can be used as sources of bacterial genome annotations.

Name	Comments	URL	Refs
GenBank	Annotated collection of all publicly available DNA sequences (data is shared among GenBank, EMBL, and DDBJ)	http://www.ncbi.nlm.nih.gov/Genbank/	-
EMBL	Annotated collection of all publicly available DNA sequences (data is shared among GenBank, EMBL, and DDBJ)	http://www.ebi.ac.uk/embl/	-
DDBJ	Annotated collection of all publicly available DNA sequences (data is shared among GenBank, EMBL, and DDBJ)	http://www.ddbj.nig.ac.jp/	-
BASys	Annotation system that is fully automatic and web-accessible	http://wishart.biology.ualberta.ca/basys/	[5•]
HAMAP	Database of high quality annotated proteomes and protein families	http://ca.expasy.org/sprot/hamap/	[6]
PUMA2	Annotation system providing detailed metabolic pathway information	http://compbio.mcs.anl.gov/puma2/	[7•]
ASAP	Annotation system that deals primarily with genomes from enterobacteria	https://asap.ahabs.wisc.edu/asap/ASAP1.htm	[8]
MaGe	Annotation system that uses conservation of gene order to support predictions	http://www.genoscope.cns.fr/agc/mage/	[9]
EcoGene	Database containing an extensive set of annotations for *E. coli* K-12	http://ecogene.org/	[10]
CCDB	Database of *E. coli* K-12 annotations for use in cellular simulations.	http://redpoll.pharmacy.ualberta.ca/CCDB/	[11]
EcoCyc	Database of *E. coli* K-12 annotations including detailed metabolic pathways.	http://ecocyc.org/	[13•]
coliBase	Database of annotations and comparative genomics information for *E. coli* and its relatives	http://colibase.bham.ac.uk/	[14]
UCSC Archaeal Genome Browser	Database of archaeal genome annotations.	http://archaea.ucsc.edu/	[15•]
xBASE	Collection of databases providing annotations for a variety of bacterial species	http://xbase.bham.ac.uk/	[16]
IMG	Database of annotations for all publicly available bacterial genomes and several draft genomes	http://img.jgi.doe.gov/	[17]
Entrez Genome	Comprehensive collection of publicly available genome sequences and annotations	http://eutils.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome	[18••]
PEDANT	Database providing numerous annotations for all publicly available bacterial genomes	http://pedant.gsf.de/	[19•]
CMR	Database of bacterial genome sequences and annotations that includes numerous online analysis tools	http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi	[20]
MicrobesOnline	Database including annotations and useful comparative genomics tools and information	http://www.microbesonline.org/	[21]
BacMap	Database providing annotations and interactive graphical maps for all bacterial genomes	http://wishart.biology.ualberta.ca/BacMap/	[22]
TIGR Annotation Engine	Annotation service for new bacterial genome sequences	http://www.tigr.org/AnnotationEngine/	-
PSORTb	Software for the prediction of bacterial protein subcellular location	http://www.psort.org/psortb/	[23]
Proteome Analyst	Software for protein function and subcellular location prediction	http://www.cs.ualberta.ca/∼bioinfo/PA/	[24]
SABIA	Bacterial annotation system that can be downloaded and run locally	http://www.sabia.lncc.br/	[25]
MAGPIE	Genome annotation system that can be downloaded and run locally	http://magpie.ucalgary.ca/	[26]
GenDB	Bacterial annotation system that can be accessed online or run locally	http://www.cebitec.uni-bielefeld.de/groups/brf/software/gendb_info/	[27]
Artemis	Software for viewing and editing sequence annotations	http://www.sanger.ac.uk/Software/Artemis/	[28]
Bluejay	Software for viewing sequence annotations	http://bluejay.ucalgary.ca/	[29]
Glimmer	Gene prediction software for microbial genomes	http://www.tigr.org/∼salzberg/glimmer.html	[31]
BLAST	Software for sequence database searching and one of the most important components of sequence analysis systems.	http://www.ncbi.nlm.nih.gov/blast/	[32]
GeneMark	Gene prediction software	http://exon.gatech.edu/GeneMark/	[33]
tRNAscan-SE	Fast and accurate program for identifying tRNA genes in genomic sequences	http://selab.wustl.edu/	[34]
HMMER	Software for performing sensitive sequence database searches	http://hmmer.wustl.edu/	[35]
UniProt	High-quality database of protein sequences and annotations that is used by many sequence annotation systems as a source of functional predictions	http://www.pir.uniprot.org/	[36]
Pfam	Large database of protein families and domains that can be used to categorize new sequences.	http://www.sanger.ac.uk/Software/Pfam/	[37]
tmHMM	Software for predicting transmembrane regions in protein sequences	http://www.cbs.dtu.dk/services/TMHMM/	[38]
SignalP	Software for predicting the presence and location of signal peptide cleavage sites in protein sequences	http://www.cbs.dtu.dk/services/SignalP/	[39]
EMBOSS	Large collection of sequence analysis programs that can be incorporated into other bioinformatics applications	http://emboss.sourceforge.net/	[40]
Bioperl	Extensive and well-documented set of software modules for managing sequence-related data	http://www.bioperl.org/	[41]

specialized tools for exploring, searching and downloading information. Various comparative genomics tools and metrics can also be accessed. For example, Entrez Genome generates a graph for each bacterial genome, which plots the similarity between it and all other known bacterial genomes [18••]. Similarly, the Comprehensive Microbial Resource allows users to align entire bacterial genomes [20].

Web-based annotation pipelines

As a result of continued improvements in sequencing technology, more and more bacterial genomes are being sequenced by more laboratories than ever before. This is leading to a growing problem where sequence data are being generated by small teams of researchers, or smaller sequencing facilities that lack the computing resources and expertise needed to maintain or implement the software necessary for bacterial genome annotation. For these investigators the most convenient annotation tools will be those provided as online services. For example, the BASys server accepts a raw genomic sequence, and then uses more than 30 different programs to generate clickable genome maps and data for nearly 60 annotation fields corresponding to each gene [5•]. Another annotation pipeline is the TIGR Annotation Engine. Users can submit raw genome sequences by email for processing. Several values are returned for each gene predicted by the system, including common name, gene symbol, Enzyme Commission (EC) numbers, Gene Ontology (GO) terms, BLAST hits and paralogues. The results, which are stored in a relational database, can be manually edited. Another web-based system that supports automated and interactive annotation is PUMA2. Besides generating predicted protein functions, PUMA2 provides extensive metabolic pathway information [7•]. In addition to these complete analysis systems, there are several more specialized servers that can provide one or two annotations per gene. For example, PSORTb [23] can be used to predict the subcellular location of bacterial proteins, and Proteome Analyst [24] can generate subcellular location and GO predictions. These tools can accept entire proteomes, and can provide results more quickly than the complete annotation systems.

Annotation pipelines that can be run locally

The main drawbacks of submitting a sequence to a web server or an annotation service such as BASys, the TIGR Annotation Engine or PUMA2 are that the results might not be available for several days, and there is little flexibility in terms of which programs and databases are used during the analysis procedure. One solution to these issues is to install an annotation system locally. A local system can be accessed whenever needed, and in many cases can be adjusted to suit the particular needs of the project. The SABIA system, for example, can be installed and then controlled through a web interface. It can be used to perform sequence trace reading, assembly, gene prediction and function prediction [25]. A more extensive set of annotations can be obtained from the MAGPIE pipeline [26]: it uses more than fifteen public databases to characterize new sequences, and it provides graphical reports. Another standalone pipeline is the GenDB system [27]. It offers many of the features found in the server-based systems and has a modular design that makes it more suited to custom modifications. The installation of these systems is fairly involved because they all make use of several existing bioinformatics programs that must be obtained and installed separately. Somewhat distinct from these systems is the Artemis annotation tool [28]. It can be used to simultaneously view the results of multiple sequence analysis results in the context of a genome sequence. By allowing sequence zooming and scrolling, Artemis can serve as a useful tool for manual annotation review and editing. The annotations themselves must come from other programs, because analysis algorithms are not built into Artemis. This makes the program more flexible, but some users might have trouble importing analysis results. Another useful tool for exploring and evaluating genome annotations is BlueJay [29]; it allows data from BioMOBY-compliant services [30] or from MAGPIE [26] to be viewed graphically in an interactive environment.

Building an annotation pipeline

Most of the existing web-based and downloadable annotation systems are built using similar components. For example, BASys [5•], the TIGR Annotation Engine, and SABIA [25] all use the gene-prediction program Glimmer [31] to identify putative coding sequences. Similarly, all three use BLAST [32] to compare the predicted genes with a variety of sequence databases. Building a custom annotation pipeline, using these same components in combination with other annotation tools is an option for groups wanting as much control as possible over how sequences are analyzed. The process of assembling a custom annotation system involves either writing or downloading selected sequence analysis modules, connecting them so that the outputs from one can be passed on as the inputs to another, and storing the results. Module connection is usually accomplished using a scripting language such as Perl, and the data are usually stored in a relational database such as MySQL, or as a simple flat file. Almost all automated analysis systems make use of gene predictors such as Glimmer [31] or GeneMark [33] and the tRNA identification program tRNAscan-SE [34]. The predicted genes or proteins are then usually passed to BLAST [32] or HMMER [35] for comparison against some of the many freely available sequence and sequence family databases, such as UniProt [36] and Pfam [37]. Potential protein transmembrane regions and signal peptides can be identified using tmHMM [38] and SignalP [39], respectively. The EMBOSS package [40] offers several additional programs for sequence analysis, and the Bioperl library [41] contains many Perl modules that can be used for managing sequences and annotations, and for reading and writing sequence files. If the pipeline is designed carefully, new analysis programs can readily be incorporated as they become available.

Conclusions

Automated methods of genome annotation will become increasingly important in the future, as our capacity to generate genome sequences continues to increase. Several different annotation systems are available, each providing a distinct set of features. The most popular pipelines will probably be those that are web-accessible, although standalone and custom systems will continue to be useful for groups wanting more specific annotations in a shorter period of time. New methods of annotation are constantly being developed. For example, a better strategy for including the input of human experts in the annotation process has recently been presented [42], and a novel approach for identifying bacterial pseudogenes has been described [43]. These advances, along with more complete databases and new experimental knowledge, will no doubt lead to better annotation systems in the near future.

Update

Direct sampling of DNA from microbial communities, referred to as metagenomics, should greatly enhance our understanding of the diversity and structure of microbial ecosystems, as well as provide insight into the metabolic capabilities of bacteria that cannot be isolated or cultivated. A new software tool, IMG/M, has recently been developed for the analysis of metagenomics data [44••]. Based on the IMG system [17], IMG/M addresses some of the challenges associated with the analysis of sequences derived from a mixture of genomes. Several useful comparative tools are also provided. For example, two metagenomes can be compared in terms of the abundance of a gene family of interest.

Acknowledgements

The authors wish to acknowledge the financial support of The National Institute for Nanotechnology (NINT) and Genome Alberta (a division of Genome Canada) in preparing this review.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Fraser-Liggett CM: **Insights on biology and evolution from microbial genome sequencing**. *Genome Res* 2005, **15**:1603-1610.

The authors discuss the incredible diversity of bacterial genomes and the potential of metagenomics in understanding microbial communities.

2. Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, La T, Hampson DJ, Bellgard M, Wassenaar TM, Ussery DW: **Ten years of bacterial genome sequencing: comparative-genomics-based discoveries**. *Funct Integr Genomics* 2006, **6**:165-185.

3. • Raskin DM, Seshadri R, Pukatzki SU, Mekalanos JJ: **Bacterial genomics and pathogen evolution**. *Cell* 2006, **124**:703-714.

An in-depth examination of bacterial genome evolution and comparative genomics. It includes several detailed discussions of important discoveries that have been made using a combination of computational and experimental methods.

4. Barker JJ: **Antibacterial drug discovery and structure-based design**. *Drug Discov Today* 2006, **11**:391-404.

5. • Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS: **BASys: a web server for automated bacterial genome annotation**. *Nucleic Acids Res* 2005, **33**:W455-W459.

Currently BASys is the most accessible fully automatic annotation system. Users simply need to navigate to the BASys homepage and supply a genome sequence and an email address.

6. Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C *et al.*: **Automated annotation of microbial proteomes in SWISS-PROT**. *Comput Biol Chem* 2003, **27**:49-58.

7. • Maltsev N, Glass E, Sulakhe D, Rodriguez A, Syed MH, Bompada T, Zhang Y, D'Souza M: **PUMA2–grid-based high-throughput analysis of genomes and metabolic pathways**. *Nucleic Acids Res* 2006, **34**:D369-D372.

This study discusses a recently developed analysis system that provides numerous annotations for user-supplied bacterial genomes and for all publicly available bacterial genomes.

8. Glasner JD, Rusch M, Liss P, Plunkett G III, Cabot EL, Darling A, Anderson BD, Infield-Harm P, Gilson MC, Perna NT: **ASAP: a resource for annotating, curating, comparing, and disseminating genomic data**. *Nucleic Acids Res* 2006, **34**:D41-D45.

9. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Medigue C: **MaGe: a microbial genome annotation system supported by synteny results**. *Nucleic Acids Res* 2006, **34**:53-65.

10. Rudd KE: **EcoGene: a genome sequence database for *Escherichia coli* K-12**. *Nucleic Acids Res* 2000, **28**:60-64.

11. Sundararaj S, Guo A, Habibi-Nazhad B, Rouani M, Stothard P, Ellison M, Wishart DS: **The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate *in silico* modeling of *Escherichia coli***. *Nucleic Acids Res* 2004, **32**:D293-D295.

12. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T *et al.*: ***Escherichia coli* K-12: a cooperatively developed annotation snapshot–2005**. *Nucleic Acids Res* 2006, **34**:1-9.

13. • Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive database resource for *Escherichia coli***. *Nucleic Acids Res* 2005, **33**:D334-D337.

This illustrates the importance of manual curation efforts aimed at extracting knowledge from the literature.

14. Chaudhuri RR, Khan AM, Pallen MJ: **coliBASE: an online database for *Escherichia coli, Shigella* and *Salmonella* comparative genomics**. *Nucleic Acids Res* 2004, **32**:D296-D299.

15. • Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM: **The UCSC Archaeal Genome Browser**. *Nucleic Acids Res* 2006, **34**:D407-D410.

The authors describe an archaeal genome browser built using components of the UCSC Genome Browser. It includes many useful annotations along with excellent tools for exploring and extracting annotations of interest.

16. Chaudhuri RR, Pallen MJ: **xBASE, a collection of online databases for bacterial comparative genomics**. *Nucleic Acids Res* 2006, **34**:D335-D337.

17. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I *et al.*: **The integrated microbial genomes (IMG) system**. *Nucleic Acids Res* 2006, **34**:D344-D348.

18. •• Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S *et al.*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2006, **34**:D173-D180.

The Entrez Genome database is one of the most up-to-date and dependable resources for obtaining publicly available genome sequences and annotations.

19. • Riley ML, Schmidt T, Wagner C, Mewes HW, Frishman D: **The PEDANT genome database in 2005**. *Nucleic Acids Res* 2005, **33**:D308-D310.

A valuable source of annotations for publicly available genomes. More than 20 fields of annotation are supplied for each gene.

20. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O: **The comprehensive microbial resource**. *Nucleic Acids Res* 2001, **29**:123-125.

21. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP: **The MicrobesOnline web site for comparative genomics**. *Genome Res* 2005, **15**:1015-1022.

22. Stothard P, Van Domselaar G, Shrivastava S, Guo A, O'Neill B, Cruz J, Ellison M, Wishart DS: **BacMap: an interactive picture atlas of annotated bacterial genomes**. *Nucleic Acids Res* 2005, **34**:D317-D320.

23. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS: **PSORTb v.2.0. expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis**. *Bioinformatics* 2005, **21**:617-623.

24. Szafron D, Lu P, Greiner R, Wishart DS, Poulin B, Eisner R, Lu Z, Anvik J, Macdonell C, Fyshe A *et al.*: **Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations**. *Nucleic Acids Res* 2004, **32**:W365-W371.

25. Almeida LG, Paixao R, Souza RC, Costa GC, Barrientos FJ, Santos MT, Almeida DF, Vasconcelos AT: **A system for automated bacterial (genome) integrated annotation–SABIA**. *Bioinformatics* 2004, **20**:2832-2833.

26. Gaasterland T, Sensen CW: **MAGPIE: automated genome interpretation**. *Trends Genet* 1996, **12**:76-78.

27. Linke B, Rupp O, Giegerich R, Puhler A: **GenDB–an open source genome annotation system for prokaryote genomes**. *Nucleic Acids Res* 2003, **31**:2187-2195.

28. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation**. *Bioinformatics* 2000, **16**:944-945.

29. Turinsky AL, Ah-Seng AC, Gordon PM, Stromer JN, Taschuk ML, Xu EW, Sensen CW: **Bioinformatics visualization and integration with open standards: the Bluejay genomic browser**. *In Silico Biol* 2005, **5**:187-198.

30. Wilkinson MD, Links M: **BioMOBY: an open source biological web services proposal**. *Brief Bioinform* 2002, **3**:331-341.

31. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER**. *Nucleic Acids Res* 1999, **27**:4636-4641.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipmann DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-420.

33. Besemer J, Borodovsky M: **GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses**. *Nucleic Acids Res* 2005, **33**:W451-W454.

34. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic Acids Res* 1997, **25**:955-964.

35. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**:755-763.

36. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al.*: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2004, **32**:D115-D119.

37. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

38. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**:567-580.

39. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0**. *J Mol Biol* 2004, **340**:783-795.

40. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**:276-277.

41. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H *et al.*: **The Bioperl toolkit: Perl modules for the life sciences**. *Genome Res* 2002, **12**:1611-1618.

42. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes**. *Nucleic Acids Res* 2005, **33**:5691-5702.

43. Ochman H, Davalos LM: **The nature and dynamics of bacterial genomes**. *Science* 2006, **311**:1730-1733.

44. •• Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavrommatis K, Kunin V, Garcia Martin H *et al.*: **An experimental metagenome data management and analysis system**. *Bioinformatics* 2006, **22**:e359-e367.

This paper describes the IMG/M system for analyzing sequence data obtained from mixtures of microbial genomes (metagenomes). This system is one of the first to address metagenomics data, the analysis of which is inherently more difficult than the analysis of isolated genomes.