Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W455–W459***
doi:10.1093/nar/gki593

BASys: a web server for automated bacterial genome annotation

Gary H. Van Domselaar, Paul Stothard, Savita Shrivastava, Joseph A. Cruz, AnChi Guo, Xiaoli Dong, Paul Lu, Duane Szafron, Russ Greiner and David S. Wishart*

Departments of Biological Sciences and Computing Science, University of Alberta, Edmonton, AB, T6G 2E8, Canada

Received February 13, 2005; Revised April 14, 2005; Accepted May 6, 2005

ABSTRACT

BASys (Bacterial Annotation System) is a web server that supports automated, in-depth annotation of bacterial genomic (chromosomal and plasmid) sequences. It accepts raw DNA sequence data and an optional list of gene identification information and provides extensive textual annotation and hyperlinked image output. BASys uses >30 programs to determine ~60 annotation subfields for each gene, including gene/protein name, GO function, COG function, possible paralogues and orthologues, molecular weight, isoelectric point, operon structure, subcellular localization, signal peptides, transmembrane regions, secondary structure, 3D structure, reactions and pathways. The depth and detail of a BASys annotation matches or exceeds that found in a standard SwissProt entry. BASys also generates colorful, clickable and fully zoomable maps of each query chromosome to permit rapid navigation and detailed visual analysis of all resulting gene annotations. The textual annotations and images that are provided by BASys can be generated in ~24 h for an average bacterial chromosome (5 Mb). BASys annotations may be viewed and downloaded anonymously or through a password protected access system. The BASys server and databases can also be downloaded and run locally. BASys is accessible at http://wishart.biology.ualberta.ca/basys.

INTRODUCTION

Over the last decade, the rate at which bacterial genomes have been sequenced and made public has accelerated from roughly three per year to three per week. This surfeit of genomic sequence data should offer researchers an unprecedented opportunity to gain new insights into the inner workings of life; to develop novel vaccines, to discover new antimicrobial compounds or therapeutics; and to develop effective strategies for re-engineering bacteria for such applications as bioremediation. However, distilling useful knowledge from the enormous volume of data is a growing challenge for both bioinformaticians and microbiologists alike. To assist with the interpretation of genomic data, a number of automated genome annotation tools have been created, including GeneQuiz (1), PEDANT (2), Genotator (3), MAGPIE/BLUEJAY (4,5), GenDB (6) and the TIGR CMR (7). In order for these systems to perform at a high level of quality and throughput, these annotation systems are quite sophisticated that require a high degree of skill to implement and maintain, and also require considerable computing power to operate. The average microbiology laboratory, lacking these resources, has become reliant on centralized operations to provide them with their genome-scale annotations. This reliance creates a problem if the specific organisms they study have not been annotated, if the annotations they require are not available, or if the annotations are not current. Clearly, there is a need for the members of the scientific community to have the ability to generate their own detailed, up-to-date genome-scale annotations of the bacterial organisms they study.

An additional and critically important challenge is to present large volumes of data in an organized and intuitive way. Graphical genome maps are preferred for their ability to display genomic data in a format that is familiar and intuitive to microbiologists. Graphical maps, particularly those with zooming capability, also provide valuable information regarding the organization of open reading frames (ORFs), genes and operons. Most existing annotation systems provide genomic information in a text-only or a relatively unsophisticated graphical format with limited feature labeling and zooming. The ideal graphical system should provide multiple views of data at different scales, and should provide a convenient way to switch from graphical views to textual views and vice versa.

Here, we present a publicly available web server for bacterial genome annotation called BASys (Bacterial Annotation System). BASys accepts a bacterial genome (chromosome or

*To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 1071; Email: david.wishart@ualberta.ca

plasmid) file and optional gene identification file as input and then performs an exhaustive analysis of the genomic data, producing up to 60 separate annotations per gene. The annotated genomic data is returned as a colorful and navigable circular genome map with hyperlinks to detailed textual annotations for each gene, which include information regarding their source and validity. The complete set of maps and annotations is provided for download for off-line viewing using any modern web browser.

PROGRAM DESCRIPTION

BASys is composed of three parts: (i) a front-end web interface for submitting the raw genomic data, for scheduling the annotation and for monitoring or reporting the annotation progress; (ii) an annotation engine for analyzing the chromosome data and for generating the annotations; and (iii) a reporting system for rendering the various graphics, HTML and textual output produced by BASys. Each of these components is discussed in detail below.

Data submission and scheduling

BASys allows both anonymous and login-based access for the submission, monitoring and retrieval of genome annotations. For anonymous submissions, the user is emailed a secure URL for monitoring the progress of their annotations and for retrieving the annotations when they are complete. Alternatively, the user may register with BASys and then use a password-protected login to interact with the system. The login system allows users to submit and monitor multiple chromosomes and plasmid annotations, whereas the anonymous login is designed for single chromosome submissions. In either case, access to the submitted genomic data and results are restricted to the submitters, thus protecting possibly sensitive data from viewing by third parties.

BASys provides a web-based form for submitting the bacterial chromosome for annotation. The chromosome data must be uploaded as a FASTA-formatted file. Also required are the chromosome topology (circular or linear), the Gram-stain subtype, and a user-assigned string (i.e. name of chromosome) for monitoring annotation progress and for identifying the output files. If no additional information is supplied, BASys will attempt to predict the coding sequences within the chromosome using 'Glimmer', a popular *ab initio* microbial gene prediction program (8). Glimmer performs very well for genomes with GC content below ~60%. Above this value, Glimmer may generate a high number of false positive predictions and therefore should be used with caution. Other gene prediction programs, including extrinsic gene finders such as Critica (9) are currently being evaluated for inclusion in BASys. If gene positions are already known, or predicted in advance using an alternate gene finder, they can be supplied to BASys in a simple TAB-delimited format or as an NCBI '.ffn'-formatted FASTA file. NCBI '.ffn' files, which are available for many bacterial genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi), include the nucleotide coding sequences along with the location and direction along the chromosome, and thus can be used to 'correct' the coding regions from the chromosomal data containing frame shifts or other translational anomalies. Descriptions and links to examples of the various required and optional files are provided on the submission form for easy reference. Once the sequence file and optional input files have been submitted to BASys, they are added to a queue and scheduled for processing by the BASys annotation engine.

The computing power required to fully annotate a bacterial genome is unavoidably high. The BASys annotation engine running on a modern 32-bit Intel/Linux PC and annotating an average-sized bacterial genome of ~5 Mb or 3000 genes requires ~24 h to complete. To accommodate multiple users simultaneously performing long-running, resource-intensive genome annotations, we have implemented BASys as a distributed system operating in a clustered computing environment (Figure 1). The master node hosts the web server, and runs the queuing and scheduling system. Each slave node hosts a BASys annotation engine and genome report system. Upon receiving and validating the bacterial genomic data for annotation, the master node examines the current processing load for each slave, assigns the genome annotation job to the node with the highest availability, and then notifies the slave of the new job. The master node can also issue directives to suspend, resume, restart and remove the genome annotation jobs on the

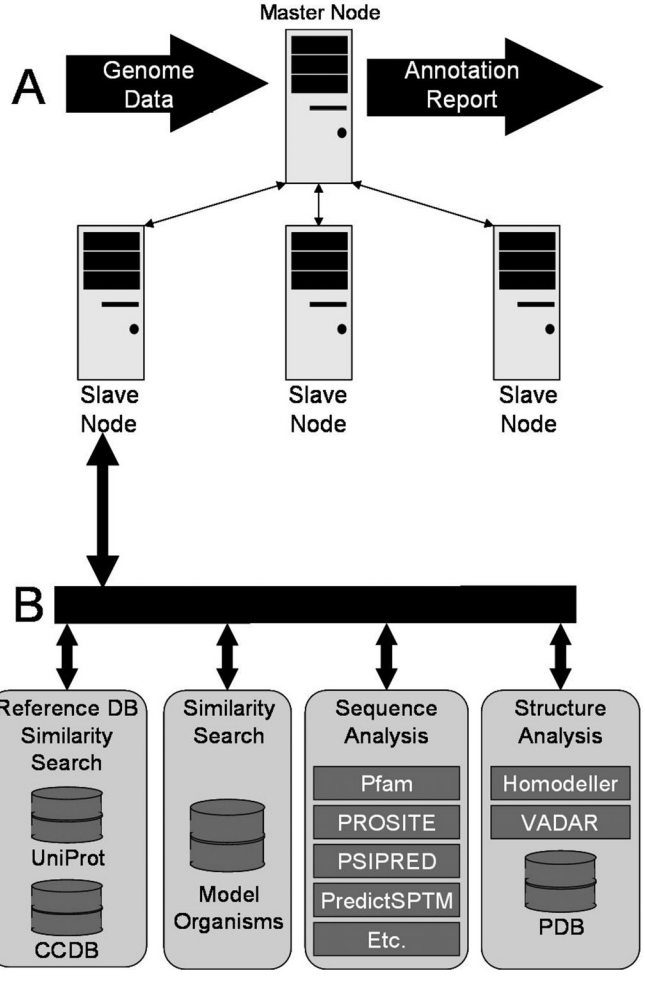

Figure 1. A schematic outline of the BASys architecture. (**A**) The web server, hosted on the master node, accepts genome data and schedules it for processing by the slave nodes. (**B**) Each slave node hosts a BASys annotation engine. The annotation pipeline combines similarity searching and sequence analysis to generate annotations.

slave nodes. Each slave node continually communicates its progress to the master node while generating the annotations and reports. Once complete, the master node notifies the submitter by email that the annotations are ready. The master and slave nodes use a MySQL client/server protocol to communicate directives and status, and the Apache web server/HTTP protocol to transfer the sequence data and reports.

The BASys annotation engine

The BASys annotation engine combines database comparison and computational sequence analysis in its processing pipeline (Figure 1). Translated coding sequences are initially compared using BLAST (10) to the extensively and expertly annotated reference databases UniProt (11) and the CyberCell comprehensive molecular database on *Escherichia coli* (12). These databases are valuable sources of high quality genomic and proteomic information including function, metabolic role, structural family and enzyme classification. Associated with each type of annotation is a similarity threshold. The similarity score between the query and database sequence is compared to the threshold value for each annotation type and qualifying annotations are transitively applied to the query sequence. For example, transferring feature information such as transmembrane domain locations requires a perfect match between query and database sequence, whereas the transfer of a general function assignment is more lenient, requiring a BLAST Expect-value of only 1×10^{-10} or better.

BASys attempts to fill the remaining annotations by subjecting the query sequence to a battery of additional similarity searches and sequence analyses. BLAST searches are conducted against a number of specialized databases including the protein sequences of model organisms (*Caenorhabditis elegans*, *Homo sapiens*, *Saccharomyces cerevisiae* and *Drosophila melanogaster*); a non-redundant database of bacterial protein sequences, the PDB database of 3D biological macromolecular structure data (13), and the COG database of orthologous groups of proteins (14). Various sequence analyses are also performed on the query sequence, including protein family analysis with Pfam (15), sequence motif analysis with PROSITE (16), signal peptide and transmembrane domain prediction with PredictSPTM (J. Cruz, unpublished data), and predicted secondary structure with PSIPRED (17). If the sequence has sufficient similarity to a sequence represented in the PDB database, then BASys may use HOMODELLER (X. Dong, unpublished data) to generate a homology model and subsequently perform a structural analysis using VADAR (18). Several additional annotations, such as protein molecular weight, isoelectric point and operon structure are calculated directly from the chromosomal, protein-coding nucleotide and translated protein sequence data. In all, a collection of ~60 distinct annotations is generated for each gene. A detailed summary of the annotation process for each field, along with its relevance and the program or database versions used, is available on the BASys web site (http://wishart.biology.ualberta.ca/basys/cgi/annotations.pl).

BASys is an ambitious annotation system. It draws on many different sources for its information and tries to assign as many annotations as possible to each gene. Although every precaution has been taken to provide the highest quality annotations possible, it is inevitable that some annotations will be incorrect, as is the case with all fully automated genome annotation systems. For this reason BASys provides, whenever possible, information on the source of the annotation, the evidence used to support the annotation and a rough indication of the quality of the annotation. For example, an annotation assigned from a similarity search would include the BLAST report as evidence, the database and version as the source, and a quality indicator such as 'marginal', 'strong' or 'clear'. This information is kept in a separate file to prevent cluttering up the main annotation report, but a hyperlink is provided to the file for easy inspection.

The BASys report generator

BASys provides its annotations in the form of a navigable circular genome map with hyperlinks to gene cards containing the annotations for each identified gene. A montage illustrating the various types of output produced by BASys is provided in Figure 2. For genome visualization and exploration, BASys relies heavily on the CGView application (19), a highly customizable circular genome rendering system that is well suited for use in bioinformatics pipelines. BASys passes information generated during the annotation process to CGView in the form of an XML document. CGView then renders this information as a series of hyperlinked PNG image files. The resulting maps show annotated genes and COG category classifications, when available. CGView renders the images at several different levels of resolution and maintains the circular view at every resolution level, making navigation quick and intuitive.

Each displayed gene in the genome map is labeled and hyperlinked to a gene card containing the corresponding annotations in a tabular format. Gene card fields derived from external sources are hyperlinked to those sources, and those derived from a similarity search are presented with a code identifying the source: [S] 100% sequence match to SwissProt, [H] Homology to a SwissProt entry or [C] Homology to a CCDB entry. Each gene card contains a hyperlink to an associated file containing more detailed descriptions of the source and quality of the annotations. Additionally, BASys annotations can be conveniently searched using server-side text and BLAST searching tools. Each search hit is hyperlinked to a corresponding BASys gene map and gene card. These text and sequence searching services are only available for annotations left on the BASys server. Downloaded (HTML or text) BASys files must be searched using client-side tools (such as standard word-processing software or local BLAST tools).

VALIDATION

Many of the algorithms and annotation strategies used in BASys were originally developed and validated during the construction of the CyberCell database (12), and the BacMap bacterial genome atlas (20). Several of the BASys annotations are calculated directly from gene and protein sequences, and hence are trivial to validate. Those derived from predictive programs or assigned transitively from a similarity search require somewhat deeper scrutiny. The protein functional assignment is generally considered to be the most important annotation, and most automated genome annotation programs are evaluated in terms of their ability to produce a correct functional assignment. However, objectively assessing the correctness of such assignments is a problem, as currently

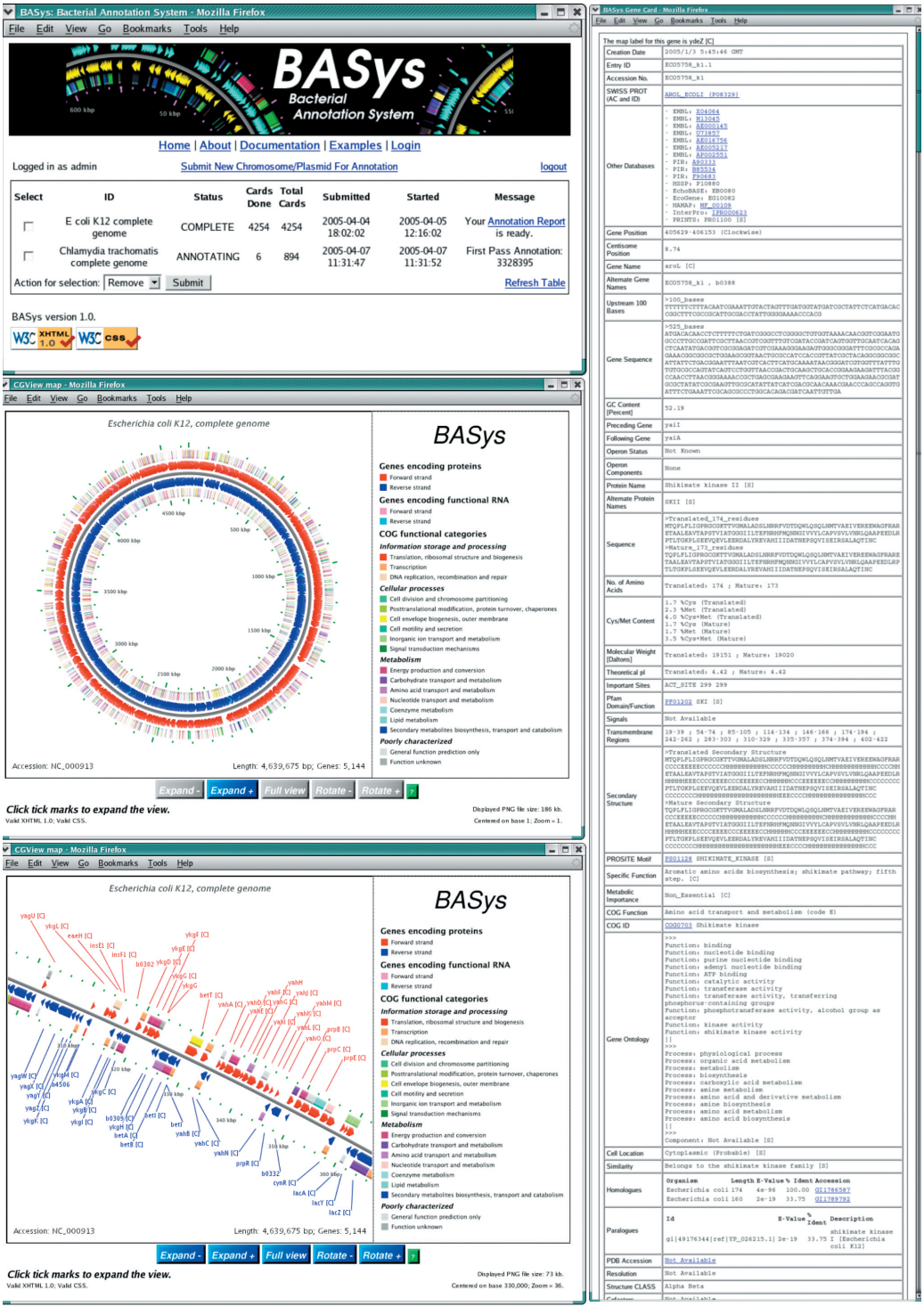

Figure 2. A screenshot montage of the BASys server showing the different kinds of graphical and textual output available. Shown are the annotation progress monitor for two genome annotation projects (*E.coli* and *C.trachomatis*) in various states of completion, a full-scale and expanded graphical genome map for *E.coli* and a sample annotation report.

there exists no single fully validated standard for comparing genome-scale annotations (21).

To assess the accuracy of our protein function assignments, we chose to use expertly annotated *Chlamydia trachomatis* proteins as a reference set (22). BASys and the complete and up-to-date non-redundant bacterial and SwissProt (Version 45) databases (with the *C.trachomatis* entries removed) were used to re-annotate the reference proteins. Matching functional assignments were obtained for 762 of the 894 entries (Table 1). BASys failed to assign a function to 33 entries having an assigned function in the reference set, corresponding to a false negative rate of 3.7%. BASys assigned 53 functions

Table 1. BASys annotation comparison with expert annotation

Category	Number
BASys assigment matches reference	492
No assignment by BASys or reference	270
Conflicting assignments	53
Assignment by BASys/unassigned by reference	46
Unassigned by BASys/assigned by reference	33

Protein function annotations were compared against an expertly annotated set of proteins from *C.trachomatis*. Additional analysis and discussion of the results are presented in the text.

that conflicted with those in the reference set, and 46 functions were assigned to proteins lacking function annotations in the reference set (see the following link: http://wishart.biology.ualberta.ca/basys/cgi/validation.pl for more details). Assuming these are overpredictions on the part of BASys, these values combine to give a false positive rate of 11%. BASys is not a conservative annotation system as it was specifically designed to attempt to annotate each gene as fully as possible. For this purpose, we feel that the observed false positive and false negative rates are acceptable. Overall, the prediction accuracy is very good and compares favorably with other automated systems (22). Because of the potential for error propagation, especially for automated annotation systems, some caution should be taken in using BASys annotations as a primary reference source. Certainly BASys results can be used as a first pass annotation that can serve as a useful starting point for later, downstream manual curation or correction. Furthermore, all annotations can be evaluated by examining the corresponding evidence file provided by BASys. Because users can download both HTML and text versions of the BASys annotations, it is relatively easy to use standard text editing software to manually correct or update any annotations at any time, without interfering with its graphic hyperlinks.

CONCLUSION

To summarize, BASys is a web server that permits high throughput, detailed and fully automated annotation of bacterial genomes. BASys consists of >30 separate and well-tested programs which are used to provide ~60 annotation fields for each gene. A fully navigable graphical map, which is hyperlinked to textual gene descriptions, is also generated to allow results to be easily browsed and evaluated. BASys' strengths are in its web accessibility, its depth and comprehensiveness of annotation and its user-friendly graphical interface. BASys does not (yet) handle partially assembled genomes nor does it offer the sophisticated ORF handling of stand-alone programs like GenDB or Magpie. The BASys web server is freely accessible at http://wishart.biology.ualberta.ca/basys.

ACKNOWLEDGEMENTS

Funding for this project was provided by the AICML, the Protein Engineering Network Centres of Excellence (PENCE Inc.) and Genome Prairie (a division of Genome Canada). Funding to pay the Open Access publication charges for this article was provided by Genome Canada.

Conflict of interest statement. None declared.

REFERENCES

1. Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. and Sander,C. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.
2. Frishman,D., Mokrejs,M., Kosykh,D., Karstenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Volz,A. *et al.* (2003) The PEDANT genome database. *Nucleic Acids Res.*, **31**, 207–211.
3. Harris,N.L. (1997) Genotator: a workbench for sequence annotation. *Genome Res.*, **7**, 754–762.
4. Gaasterland,T. and Sensen,C.W. (1996) MAGPIE: automated genome interpretation. *Trends Genet.*, **12**, 76–88.
5. Gordon,P. and Sensen,C. (1999) Bluejay: a browser for linear units in Java. *Proceedings of the 13th Annual International Symposium on High Performance Computing Systems and Applications*, Kingston, ON, Canada, pp. 183–194.
6. Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,J., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. and Puhler,A. (2003) GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.
7. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The comprehensive microbial resource. *Nucleic Acids Res.*, **29**, 123–125.
8. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
9. Bader,J.H. and Olsen,G.J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.*, **16**, 412–424.
10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipmann,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–420.
11. 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.
12. Sundararaj,S., Guo,A., Habibi-Nazhad,B., Rouani,M., Stothard,P., Ellison,M. and Wishart,D.S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate *in silico* modeling of *Escherichia coli*. *Nucleic Acids Res.*, **32**, D293–D295.
13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
14. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
15. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
16. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
17. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
18. Willard,L., Ranjan,A., Zhang,H., Monzavi,H., Boyko,R.F., Sykes,B.D. and Wishart,D.S. (2003) VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.*, **31**, 3316–3319.
19. Stothard,P. and Wishart,D.S. (2005) Circular genome visualization and exploration using CGView. *Bioinformatics*, **21**, 537–539.
20. Stothard,P., Van Domselaar,G., Shrivastava,S., Guo,A., O'Neill,B., Cruz,J.A., Ellison,M. and Wishart,D.S. (2005) BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucleic Acids Res.*, **33**, D317–D320.
21. Ouzounis,C.A. and Karp,P.D. (2002) The past, present and future of genome-wide re-annotation. *Genome Biol.*, **3**, Comment 1–6.
22. Iliopoulos,I., Tsoka,S., Andrade,M.A., Enright,A.J., Carroll,M., Poullet,P., Promponas,V., Liakopoulos,T., Palaios,G., Pasquier,C. *et al.* (2003) Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*, **19**, 717–726.