Nucleic Acids Research, 2007, Vol. 35, Database issue ***D533–D537***
doi:10.1093/nar/gkl823

SYSTOMONAS — an integrated database for systems biology analysis of *Pseudomonas*

Claudia Choi[1], Richard Münch[1], Stefan Leupold[1], Johannes Klein[1], Inga Siegel[1], Bernhard Thielen[2], Beatrice Benkert[1], Martin Kucklick[1], Max Schobert[1], Jens Barthelmes[2], Christian Ebeling[2], Isam Haddad[1], Maurice Scheer[1,4], Andreas Grote[1,3], Karsten Hiller[1], Boyke Bunk[1], Kerstin Schreiber[1], Ida Retter[1], Dietmar Schomburg[2] and Dieter Jahn[1,*]

[1]Institut für Mikrobiologie, Technische Universität Braunschweig, Spielmannstraße 7, D-38106 Braunschweig, Germany, [2]Institut für Biochemie, Universität zu Köln, Zülpicher Straße 47, D-50674 Köln, Germany, [3]Institut für Bioverfahrenstechnik, Technische Universität Braunschweig, Gaußstraße 17, D-38106 Braunschweig, Germany and [4] Fachbereich Informatik, Fachhochschule Wolfenbüttel, Am Exer 2, D-38302 Wolfenbüttel, Germany

Received August 11, 2006; Revised and Accepted October 5, 2006

ABSTRACT

To provide an integrated bioinformatics platform for a systems biology approach to the biology of pseudomonads in infection and biotechnology the database SYSTOMONAS (SYSTems biology of pseudOMONAS) was established. Besides our own experimental metabolome, proteome and transcriptome data, various additional predictions of cellular processes, such as gene-regulatory networks were stored. Reconstruction of metabolic networks in SYSTOMONAS was achieved via comparative genomics. Broad data integration is realized using SOAP interfaces for the well established databases BRENDA, KEGG and PRODORIC. Several tools for the analysis of stored data and for the visualization of the corresponding results are provided, enabling a quick understanding of metabolic pathways, genomic arrangements or promoter structures of interest. The focus of SYSTOMONAS is on pseudomonads and in particular *Pseudomonas aeruginosa*, an opportunistic human pathogen. With this database we would like to encourage the *Pseudomonas* community to elucidate cellular processes of interest using an integrated systems biology strategy. The database is accessible at http://www.systomonas.de.

MOTIVATION

Traditionally, metabolic and gene-regulatory networks were analysed separately. There are various tools for metabolic network reconstruction [e.g. (1–3)], and for the generation of gene-regulatory networks (4), i.e. the prediction of the regulation of certain genes by specific transcription factors. However, there still exist only a few tools combining both networks such as the Pathway Tools Omics Viewer (5). This poor connectivity between the two outlined approaches might be due to the fact that the required information is stored in different databases. Information on transcription factor binding sites can be found for example in RegulonDB (6) or PRODORIC (7), while metabolic reactions or pathways need to be retrieved from other database, such as BRENDA (8), BioCyc (5), KEGG (9), PseudoCyc (10), and UM-BDD (11). Combining knowledge from multiple disciplines and data resources will drive our understanding of cellular processes and lead to the prediction of the cellular behaviour in its entirety.

Consequently, we constructed the database SYSTOMONAS, which provides the basis for a systems biology approach. Here we focus on data integration for the biotechnologically and medically relevant group of bacteria, the pseudomonads.

CONTENT OF SYSTOMONAS

The complexity of a systems biology approach requires the focus on a certain well investigated organism. We have chosen the Gram negative proteobacterium *Pseudomonas aeruginosa*. This organism is a versatile soil bacterium and an important opportunistic pathogen causing persistent infection in immunocompromised patients (12). Our long term goal is the development of a dynamic model simulating the behaviour of *P.aeruginosa* during infection. The basis of our approach is SYSTOMONAS, a comprehensive database that includes data from all levels of analysis as microarray and proteomics data, metabolite measurements, sequence data, gene-regulatory networks and corresponding enzyme data.

*To whom correspondence should be addressed. Tel: +49 531 391 5801; Fax: +49 531 391 5854; Email: d.jahn@tu-bs.de

Table 1. Statistics of the metabolic network reconstruction for various *Pseudomonas* species in SYSTOMONAS

Organism	Proteins	Enzyme annotation In total	via KEGG	via PGDv2	via ENZYME	via BioCyc	Predicted
P.aeruginosa PAO1	5651	1509	1003	1017	393	493	241
P.aeruginosa PA14	6107	1442	−*	1139	—	—	303
P.fluorescens Pf-5	6137	1332	1067	—	—	—	265
P.fluorescens PfO-1	5736	1235	985	—	—	—	250
P.putida KT2440	5351	1168	897	—	23	—	268
P.syringae pv *phaseolicola*	5121	1118	938	—	—	—	180
P.syringae pv *syringae*	5089	1130	871	—	—	—	259
P.syringae pv tomato	5608	1100	851	—	5	—	249

Most enzyme information was provided by KEGG. PGD2, BioCyc, ENZYME also contributed to the functional annotation of enzymes. Further enzymes were annotated by comparative genomics (column 'Predicted'), which were missing in the databases mentioned afore. * strain absent in KEGG (version 13th April 2006).

Our database comprises information on eight different *Pseudomonas* species and strains, which genomes have been completely sequenced and functionally annotated. Besides the medically relevant *P.aeruginosa* the genera *Pseudomonas* contains various important plant pathogens and biotechnologically as well as ecologically interesting species.

Our initial focus was on metabolomics. However, all current information coming from genomics, transcriptomics and proteomics (13) is also stored in our database as data warehouse (14) or dynamically accessible via web services using SOAP interfaces (15), a platform-independent data transfer protocol (see section 'Database Techniques'). Besides other research groups and our own experimental results further data are retrieved from major general data resources. Major external sources of SYSTOMONAS are KEGG [Kyoto Encyclopedia of Genes and Genomes (9)], Pseudomonas Genome Database v2 [PGDv2 (16)], PRODORIC [PROcaryotIC Database Of gene-Regulation (4)], and BRENDA (8). KEGG provided metabolic reactions, compounds, glycans and pathways; PGDv2 and PRODORIC supplied protein, gene annotation, gene-regulatory and genome structure data. BRENDA supports kinetic and disease information. ENZYME (17) and BioCyc (5) provide further functional characterization of proteins.

Currently, SYSTOMONAS contains 10 034 proteins identified as enzymes, 195 transcription factor–gene relations, 14 250 measuring points of three independent metabolome experiments. Moreover, 11 exemplary protein spots from one proteome experiment were entered. Transcriptome data are provided by PRODORIC via SOAP (see section 'Database Techniques'). For *P.aeruginosa* PAO1 1509 unique proteins were annotated in SYSTOMONAS as enzymes. The 1509 annotated enzymes were retrieved from KEGG (1003), PGDv2 (1017), BioCyc (493), ENZYME (393) and from our own annotation (241). The corresponding annotation process is described in the following sections. By comparison, PseudoCyc contains 738 enzymes (version 9.6, http://v2.pseudomonas.com:1555/).

COMPARATIVE GENOMICS AND REGULATORY NETWORK PREDICTION

Comparing a *Pseudomonas* protein of interest with other well-characterized proteins may deliver useful insights into the evolution, distribution and species specific function. Therefore, we searched for all deduced proteins of the SYSTOMONAS database for orthologous proteins in other *Pseudomonas* species to obtain orthologous protein clusters. First, a restricted BLAST analysis (18) was performed on the protein sequences followed by a pairwise global alignment using the tool stretcher of the EMBOSS package (19). The homologous protein pairs can be obtained from the SYSTOMONAS protein table and visualized as multiple alignments, which are produced by MUSCLE [MUltiple Sequence Comparison by Log-Expectation (20)]. A more dynamic and flexible tool for the visualization of multiple alignments is provided by Jalview (21), which is also accessible from SYSTOMONAS. This tool does not only display multiple alignments but is also able to generate a phylogenetic tree for the protein group by different algorithms. The *E*-value of BLAST and the identity calculated by stretcher can be retrieved by activating the two multiple alignment tools.

Graphical maps of corresponding gene regions can be retrieved via a hyperlink to the BRENDA Genome Explorer, a tool within the BRENDA package (8). BRENDA Genome Explorer visualizes orthologous gene regions, which have a sequence identity of at least 50% in different organisms.

If the user is interested in the prediction of transcription factor binding sites and the deduction of corresponding regulons the tool Virtual Footprint (4) can be employed. We adapted this tool to SYSTOMONAS by limiting the analysis on *Pseudomonas* species.

METABOLIC NETWORK RECONSTRUCTION

Genes and proteins of eight *Pseudomonas* species and strains are carefully annotated by the *Pseudomonas* community and involved genome projects (16). PGDv2 is the resource for the continually updated *P.aeruginosa* PAO1 genome annotation. It also refers to genome annotation web sites of other *Pseudomonas* genome projects for the most up-to-date information. To reconstruct metabolic networks, we transferred known EC numbers of each *Pseudomonas* protein to its homologous partners. The enzyme designation of proteins was determined in three steps. First, the external databases KEGG (9), PGDv2 (16), ENZYME (17), BioCyc (5) provide EC numbers for the proteins. Second, homology analyses lead to putative orthologous protein pairs (see section 'Comparative Genomics'). Third, if the identity of the global alignment (stretcher) equals or exceeds 60%, all EC number connections to proteins were transferred to their orthologous

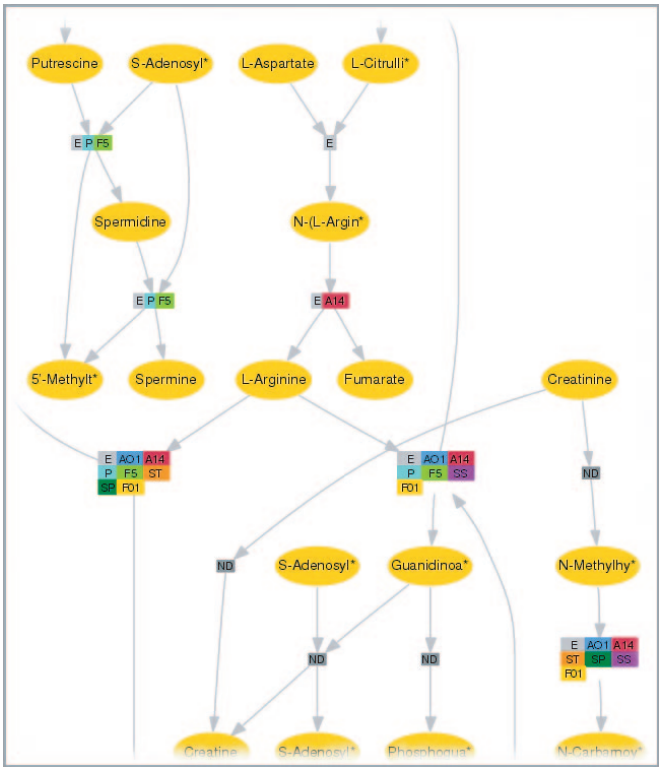

Figure 1. The visualization of metabolic pathways from KEGG in SYSTOMONAS is based on GraphViz using the dot layout. All known metabolic reactions are depicted here for the 'Urea cycle and metabolism of amino groups' pathway. Rectangles depict metabolic reactions, ellipses represent metabolites whose names are abbreviated with an asterisk * when the length exceeds 10 letters. Both types of nodes are clickable. Different colours for rectangles specify distinct *Pseudomonas* species, which catalyse the corresponding reaction. These pathways can be obtained from metabolic pathway entries. An abbreviation code for the species is provided with the visualization output (AO1 = *P.aeruginosa* PAO1, A14 = *P.aeruginosa* PA14, P = *P.putida* KT2440, Pf-5 = *P.fluorescens* F5, F01 = *P.fluorescens* PfO-1, ST = *P.syringae* pv tomato, SP = *P.syringae* pv *phaseolicola*, SS = *P.syringae* pv *syringae*)

protein partners. The EC number source on the website is indicated accordingly. EC numbers newly identified by our method are declared as 'predicted'. In order to improve the metabolic network reconstruction by applying another method, we also used the tool metaSHARK (1). This tool is able to identify potential enzyme-encoding genes in raw DNA sequences, which are not annotated yet. All newly detected EC numbers by metaSHARK are also indicated as 'predicted' in SYSTOMONAS. Table 1 lists parts of the corresponding database content.

Next, we imported KEGG-pathway data and enabled links to pathway maps via the SOAP interface provided by KEGG (9), in which *P.aeruginosa* enzymes are highlighted. A metabolic network including all involved enzymes, metabolic reactions and metabolites of all pseudomonads is delivered by our own adapted tool, which is based on GraphViz (22) and creates clickable image maps (Figure 1). The user immediately recognizes enzymatic reactions unique to one *Pseudomonas* species. All necessary information concerning this reaction is provided by mouse click.

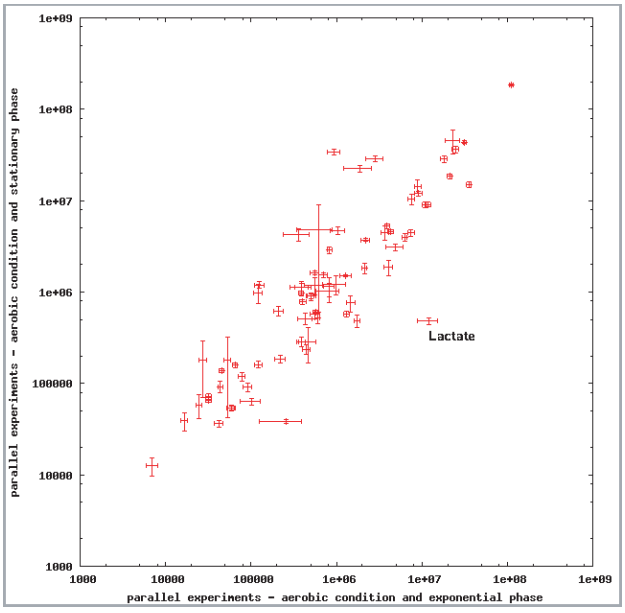

Figure 2. Semi-quantitative scatter plot for the comparison of metabolic profiles measured for *P.aeruginosa* PAO1 grown under aerobic conditions. Metabolites were analysed by GC/MS. Mean peak areas and standard deviations for the metabolites were calculated and plotted on a logarithmic scale using gnuplot (www.gnuplot.info). Metabolites measured from samples of exponentially growing cells under aerobic conditions are plotted along the *x*-axis against metabolites from samples of resting cells along the *y*-axis. The metabolite name for every data point is shown as tooltip while moving the mouse over the point (e.g. for the data point 'Lactate') and linked back to the corresponding database entry. If the metabolic profile during one experimental condition is similar to the condition compared, data points will arrange closely to the diagonal line.

METABOLOMICS DATA

The database structure of SYSTOMONAS is suitable for the simple storage of various types of experimental data. All currently available transcriptome and proteome data are deposited in the database. For a start, we included our own experimental data obtained for the *P.aeruginosa* strain PAO1 measured under different growth conditions. Our raw data analysed by GC/MS can be accessed via the query form 'omics data'. As an extra feature, the metabolomics data obtained for one specific growth condition can be plotted against another dataset using gnuplot (www.gnuplot.info). Data points are clickable and lead to the corresponding metabolite of the database (Figure 2). If the levels of specific metabolites differ significantly between different conditions, the measured values are found distinct to a fictive diagonal line. Experimental conditions and methods are indicated along with the raw data.

DATABASE TECHNIQUES

We have chosen the open source object-relational database management system PostgreSQL 8.0.3 (www.posgresql.org) for our database. This database is accessed with the scripting language PHP (www.php.net), which also allows the dynamic generation of the web interface. The web server is Apache 2.0 (http://httpd.apache.org).

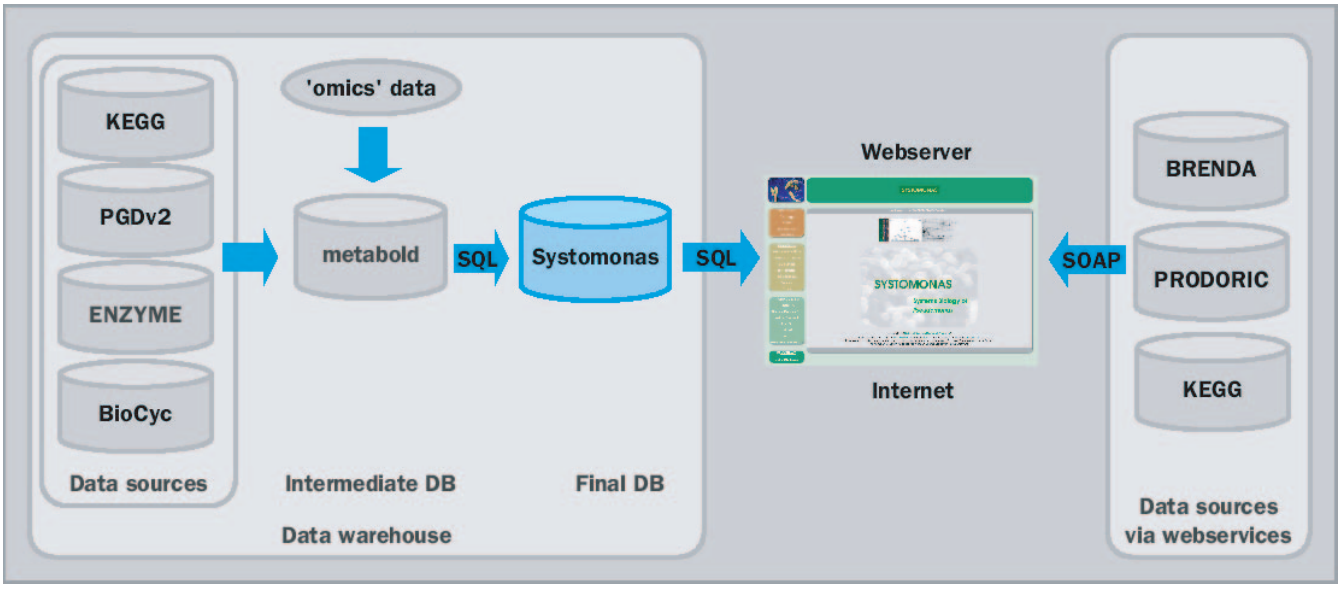

Figure 3. SYSTOMONAS architecture: combining the data warehouse concept and web services to provide a quick and dynamically updated data integration.

Table 2. External web services implemented on the SYSTOMONAS (S.) websites via SOAP

Database	Name	Function	S. website form
BRENDA www.brenda.uni-koeln.de/soap	getFunctionalData()	Kinetic data and corresponding references	EC
	getDisease()	Diseases and corresponding references	EC
PRODORIC www.prodoric.de/soap	getOperon()	Operon data and corresponding references	Gene
	getRegulatorsFromGene()	Transcription factors, DNA binding sites, and corresponding references	Interaction
	getProfile()	Experimental conditions for expression profile experiments	Transcriptomics
	getProfileParameter()	Expression profiles experiments and corresponding references	Transcriptomics
KEGG www.genome.jp/kegg/soap	soap_kegg_pathway()	Visualization of metabolic pathway maps	Pathway

These services complement a specific record of the indicated SYSTOMONAS website form by transferring the appropriate information from the given external database.

Data integration with SYSTOMONAS combines two different principle concepts, the data warehouse concept (14) and dynamic web services via SOAP. The advantage of a data warehouse is mainly its fast performance during the data retrieval. For this purpose, a major portion of the data of SYSTOMONAS, such as KEGG compounds, reactions or PRODORIC transcription factor—gene interactions is locally stored. The major advantage of SOAP is its up-to-dateness, since SOAP-transmitted information is corresponding to the most recent data of the consulted database. Several databases provide web services via SOAP, such as the major sequence databases (23–25). Several other databases, such as Atlas (26) are organized as data warehouses. The main data sources of SYSTOMONAS are stored and matched in an intermediate data container 'metabold', which supplies data to SYSTOMONAS (Figure 3). The web services via SOAP, which are used on the websites of SYSTOMONAS, are listed in Table 2. Whenever a webpage with these web services is accessed, the data is retrieved from the actual external database and amends the locally stored data of SYSTOMONAS. The API (application programming interface) is constructed by the SOAP extension of PHP.

AVAILABILITY

Currently, the data of SYSTOMONAS along with its visualization tools and web services can be accessed freely via a web-based user interface (http://www.systomonas.de). Additional information, such as kinetic data, operon structures or transcriptomics data are retrieved on-the-fly via web services from PRODORIC, BRENDA and KEGG (Table 2). We provide SBML (27) formatted files for downloading our metabolic and gene-regulatory networks along with a database copy at the SYSTOMONAS website.

ACKNOWLEDGEMENTS

The authors are much obliged to Frank Klawonn for advising in the statistical part of metabolomics comparison.

Many thanks go to Mathias Krull and Barbara Schulz, who proof-read this paper. This work was funded by the German Bundesministerium für Bildung und Forschung (BMBF) for the National Genome Research Network (NGFN2-EP, grant no. 0313398A), BMBF for the Bioinformatics Competence Center Intergenomics (Grant No. 031U110A/031U210A) and the Volkswagen Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Pinney,J.W., Shirley,M.W., McConkey,G.A. and Westhead,D.R. (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. *Nucleic Acids Res*., **33**, 1399–1409.
2. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18**, S225–S232.
3. Goesmann,A., Haubrock,M., Meyer,F., Kalinowski,J. and Giegerich,R. (2002) PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, **18**, 124–129.
4. Münch,R., Hiller,K., Grote,A., Scheer,M., Klein,J., Schobert,M. and Jahn,D. (2005) Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics*, **21**, 4187–4189.
5. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahrén,D., Tsoka,S., Darzentas,N., Kunin,V. and López-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*., **33**, 6083–6089.
6. Salgado,H., Gama-Castro,S., Peralta-Gil,M., Díaz-Peredo,E., Sánchez-Solano,F., Santos-Zavaleta,A., Martínez-Flores,I., Jiménez-Jacinto,V., Bonavides-Martínez,C., Segura-Salazar,J. *et al*. (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization and growth conditions. *Nucleic Acids Res*., **34**, D394–D397.
7. Münch,R., Hiller,K., Barg,H., Heldt,D., Linz,S., Wingender,E. and Jahn,D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res*., **31**, 266–269.
8. Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*., **32**, D431–D433.
9. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.
10. Romero,P. and Karp,P. (2003) PseudoCyc, a pathway-genome database for *Pseudomonas aeruginosa*. *J. Mol. Microbiol. Biotechnol*., **5**, 230–239.
11. Ellis,L.B.M., Roe,D. and Wackett,L.P. (2006) The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res*., **34**, D517–D521.
12. Lyczak,J.B., Cannon,C.L. and Pier,G.B. (2000) Establishment of *Pseudomonas aeruginosa* infection: lessons from a versatile opportunist. *Microbes Infect*., **2**, 1051–1060.
13. Schreiber,K., Bös,N., Eschbach,M., Jänsch,L., Wehland,J., Bjarnsholt,T., Givskov,M., Hentzer,M. and Schobert,M. (2006) Anaerobic survival of *Pseudomonas aeruginosa* by pyruvate fermentation requires an Usp-type stress protein. *J. Bacteriol*., **188**, 659–668.
14. Stein,L.D. (2003) Integrating biological databases. *Nature Rev. Genet*., **4**, 337–345.
15. Stein,L.D. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.
16. Winsor,G.L., Lo,R., Sui,S.J.H., Ung,K.S.E., Huang,S., Cheng,D., Ching,W.-K.H., Hancock,R.E.W. and Brinkman,F.S.L. (2005) *Pseudomonas aeruginosa* Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic Acids Res*., **33**, D338–D343.
17. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*., **28**, 304–305.
18. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*., **32**, W20–W25.
19. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
20. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
21. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
22. Gansner,E.R. and North,S.C. (2000) An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper*., **30**, 1203–1233.
23. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res*., **34**, D16–D20.
24. Okubo,K., Sugawara,H., Gojobori,T. and Tateno,Y. (2006) DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res*., **34**, D6–D9.
25. Cochrane,G., Aldebert,P., Althorpe,N., Andersson,M., Baker,W., Baldwin,A., Bates,K., Bhattacharyya,S., Browne,P., van den Broek,A. *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res*., **34**, D10–D15.
26. Shah,S.P., Huang,Y., Xu,T., Yuen,M.M.S., Ling,J. and Ouellette,B.F.F. (2005) Atlas–a data warehouse for integrative bioinformatics. *BMC Bioinformatics*, **6**, 34.
27. Finney,A. and Hucka,M. (2003) Systems biology markup language: level 2 and beyond. *Biochem. Soc. Trans*., **31**, 1472–1473.