BRIGEP—the BRIDGE-based genome–transcriptome–proteome browser

A. Goesmann*, B. Linke, D. Bartels, M. Dondrup, L. Krause, H. Neuweger, S. Oehm, T. Paczian, A. Wilke and F. Meyer

Bielefeld University, Center for Biotechnology (CeBiTec), D-33594 Bielefeld, Germany

Received February 14, 2005; Accepted March 16, 2005

ABSTRACT

The growing amount of information resulting from the increasing number of publicly available genomes and experimental results thereof necessitates the development of comprehensive systems for data processing and analysis. In this paper, we describe the current state and latest developments of our BRIGEP bioinformatics software system consisting of three web-based applications: GenDB, EMMA and ProDB. These applications facilitate the processing and analysis of bacterial genome, transcriptome and proteome data and are actively used by numerous international groups. We are currently in the process of extensively interconnecting these applications. BRIGEP was developed in the Bioinformatics Resource Facility of the Center for Biotechnology at Bielefeld University and is freely available. A demo project with sample data and access to all three tools is available at https://www.cebitec.uni-bielefeld.de/groups/brf/software/brigep/. Code bundles for these and other tools developed in our group are accessible on our FTP server at ftp.cebitec.uni-bielefeld.de/pub/software/.

INTRODUCTION

In the last few years, advances in high-throughput sequencing techniques have dramatically decreased time and costs needed for obtaining the DNA sequence of an organism. Currently, more than 1300 finished or ongoing genome projects are listed in the GOLD (1) database. For most organisms under investigation more and more experimental data are collected, e.g. by transcriptomics, proteomics and metabolomics experiments.

To process and analyze these large amounts of datasets, software packages for each of these areas have been developed in recent years [e.g. ARTEMIS (2) and ERGO (3) for genome annotation or BASE (4) and GECKO (5) for gene expression data analysis]. Approaches to link and integrate data originating from these different areas have also been found to be a useful means for collecting new knowledge in an easy and more intuitive way [e.g. PRIME (6) and HaloLex (www.halolex.mpg.de)]. In the Bioinformatics Resource Facility of the Center for Biotechnology, we have developed the following three applications: GenDB (7) for genome annotation, EMMA (8) for transcriptome analyses and ProDB (9) for proteome analyses. Each system exhibits a full-featured analysis software for the respective area, ranging from raw data processing to diverse and advanced functions for analyzing the processed data.

In this report, we briefly describe the newly developed web frontends available for each of these applications as well as first examples of their ongoing integration enabled by the BRIDGE integration layer (10). Using the resulting new web-based system BRIGEP, our collaborators from all over the world are able to process and analyze their data, as well as sharing it with other members of the community. As we have included a project management component into our software, a user or a community can also decide whether and when their data are made available to the public.

In the following, we illustrate the functionality of the three web interfaces and show examples of the ongoing integration of the data.

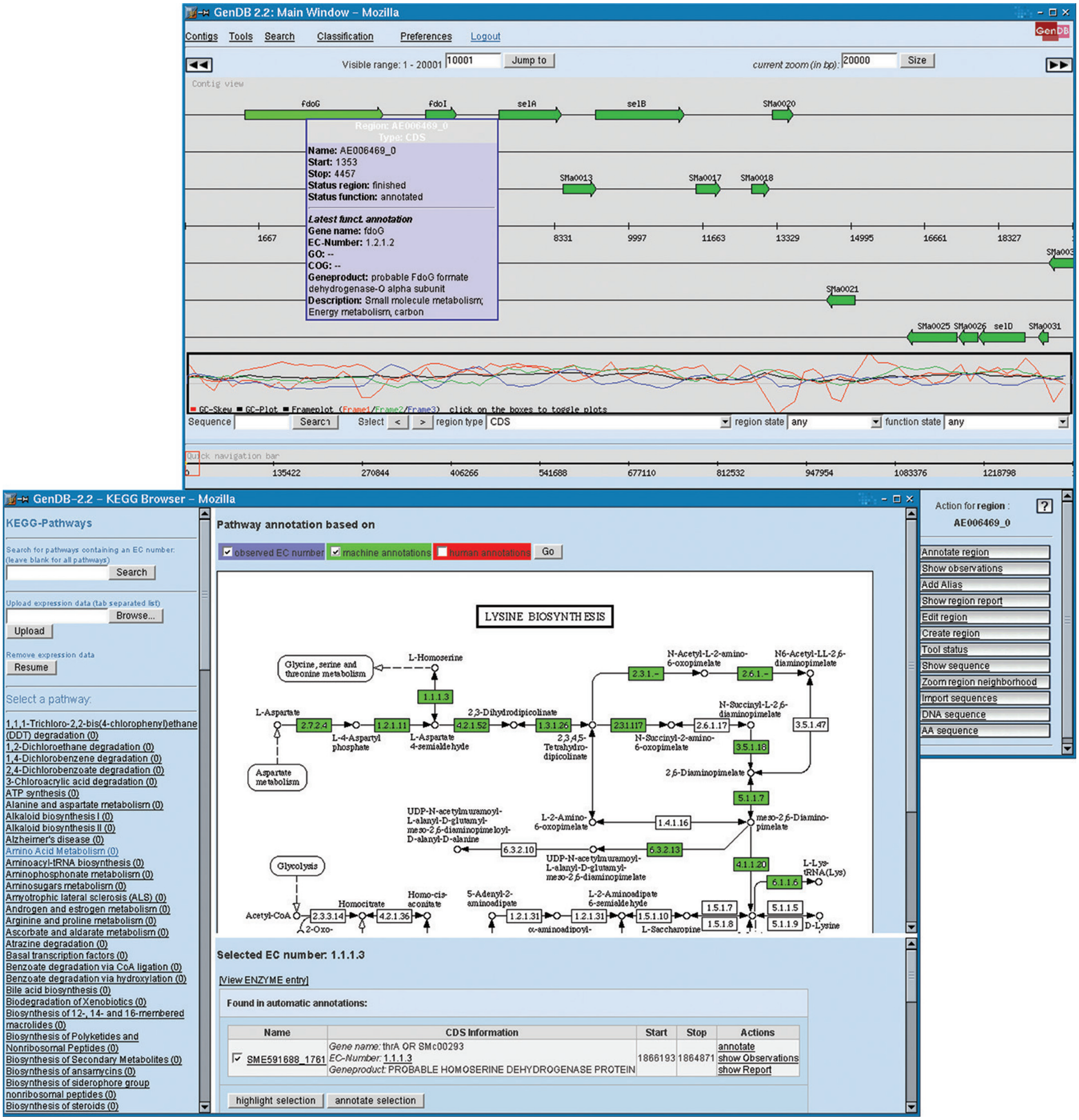
DESIGN AND IMPLEMENTATION

The BRIGEP system currently provides access to the three applications GenDB, EMMA and ProDB. The underlying data are stored in SQL databases. Each system has a three-tiered architecture based on an object-relational mapping provided by the in-house developed O2DBI (11) tool. The storage backend can be accessed via an application programmer's interface (API). Standard methods for retrieving, manipulating and deleting objects in a persistent manner are automatically generated, while additional methods for more complex tasks are added manually as extensions to each class.

*To whom correspondence should be addressed. Tel: +49 521 106 4827; Fax: +49 521 106 6419; Email: Alexander.Goesmann@CeBiTec.Uni-Bielefeld.DE

To be able to restrict data access to authorized users, data are organized in projects, e.g. a GenDB project is usually set up for the annotation of a single genome or a ProDB project for maintaining proteome data of a single organism. Projects and their members are managed using the General Project Management System, for which we have developed a web interface as well (not described here). The BRIDGE layer is used to interconnect data objects from different projects.

Based on the described API and the BRIDGE layer, which is provided in Perl, the web functionality is established using Perl CGI scripts. Via these scripts, authorized users (members of a project) can handle data input, manipulation, processing



GenDB 2.2: Main Window – Mozilla
Contigs Tools Search Classification Preferences Logout
GenDB
Visible range: 1 - 20001 10001 Jump to
current zoom (in bp): 20000 Size
Contig view
fdoG fdoI selA selB SMa0020
SMa0013 SMa0017 SMa0018
1667 8331 9997 11663 13329 14995 16661 18327
SMa0021
SMa0025 SMa0026 selD SMa0031
GC-Skew GC-Plot Frameplot (Frame1/Frame2/Frame3) click on the boxes to toggle plots
Sequence Search Select < > region type CDS region state any function state any
Quick navigation bar
0 135422 270844 406266 541688 677110 812532 947954 1083376 1218798
GenDB-2.2 – KEGG Browser – Mozilla
KEGG-Pathways
Search for pathways containing an EC number: (leave blank for all pathways)
Search
Upload expression data (tab separated list)
Browse... Upload
Remove expression data
Resume
Select a pathway:
1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation (0)
1,2-Dichloroethane degradation (0)
1,4-Dichlorobenzene degradation (0)
2,4-Dichlorobenzoate degradation (0)
3-Chloroacrylic acid degradation (0)
ATP synthesis (0)
Alanine and aspartate metabolism (0)
Alkaloid biosynthesis I (0)
Alkaloid biosynthesis II (0)
Alzheimer's disease (0)
Amino Acid Metabolism (0)
Aminoacyl-tRNA biosynthesis (0)
Aminophosphonate metabolism (0)
Aminosugars metabolism (0)
Amyotrophic lateral sclerosis (ALS) (0)
Androgen and estrogen metabolism (0)
Arginine and proline metabolism (0)
Ascorbate and aldarate metabolism (0)
Atrazine degradation (0)
Basal transcription factors (0)
Benzoate degradation via CoA ligation (0)
Benzoate degradation via hydroxylation (0)
Bile acid biosynthesis (0)
Biodegradation of Xenobiotics (0)
Biosynthesis of 12-, 14- and 16-membered macrolides (0)
Biosynthesis of Polyketides and Nonribosomal Peptides (0)
Biosynthesis of Secondary Metabolites (0)
Biosynthesis of ansamycins (0)
Biosynthesis of siderophore group nonribosomal peptides (0)
Biosynthesis of steroids (0)
Pathway annotation based on
☑ observed EC number ☑ machine annotations ☐ human annotations Go
LYSINE BIOSYNTHESIS
Glycine, serine and threonine metabolism
L-Homoserine
1.1.1.3
L-Aspartate 2.7.2.4 1.2.1.11 4.2.1.52 1.3.1.26 2.3.1.117 2.6.1.17 3.5.1.47
L-4-Aspartyl phosphate L-Aspartate 4-semialdehyde
2,3-Dihydrodipicolinate 2,3,4,5-Tetrahydrodipicolinate
N-Acetyl-L-2-amino-6-oxopimelate 2.3.1.- 2.6.1.- N6-Acetyl-LL-2,6-diaminopimelate
N-Succinyl-L-2,6-diaminopimelate N-Succinyl-2-amino-6-oxopimelate 3.5.1.18
2,6-Diaminopimelate 5.1.1.7
Aspartate metabolism
UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanyl-D-alanine
L-2-Amino-6-oxopimelate 1.4.1.16 meso-2,6-Diamino-pimelate
6.3.2.10 6.3.2.13 UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimelate
Glycolysis 4.1.1.20 L-Lys-tRNA(Lys) 6.1.1.6
Acetyl-CoA 2.3.3.14 2-Oxo- Homocitrate 4.2.1.36 Homo-cis-aconitate
5-Adenyl-2-aminoadipate 1.2.1.31 1.2.1.31 α-aminoadipoyl- L-2-Aminoadipate 6-semialdehyde 1.5.1.10 L-Saccharopine 1.5.1.7 1.5.1.8 5.1.1.5 5.1.1.9 D-Lysine
Action for region : AE006469_0
Annotate region
Show observations
Add Alias
Show region report
Edit region
Create region
Tool status
Show sequence
Zoom region neighborhood
Import sequences
DNA sequence
AA sequence
highlight selection annotate selection

Region: AE006469_0
Type: CDS

Name: AE006469_0
Start: 1353
Stop: 4457
Status region: finished
Status function: annotated

Latest funct. annotation
Gene name: fdoG
EC-Number: 1.2.1.2
GO: --
COG: --
Geneproduct: probable FdoG formate dehydrogenase-O alpha subunit
Description: Small molecule metabolism; Energy metabolism, carbon

Selected EC number: 1.1.1.3

[View ENZYME entry]

Found in automatic annotations:

	Name	CDS Information	Start	Stop	Actions
☑	SME591688_1761	*Gene name:* thrA OR SMc00293 *EC-Number:* 1.1.1.3 *Geneproduct:* PROBABLE HOMOSERINE DEHYDROGENASE PROTEIN	1866193	1864871	annotate show Observations show Report

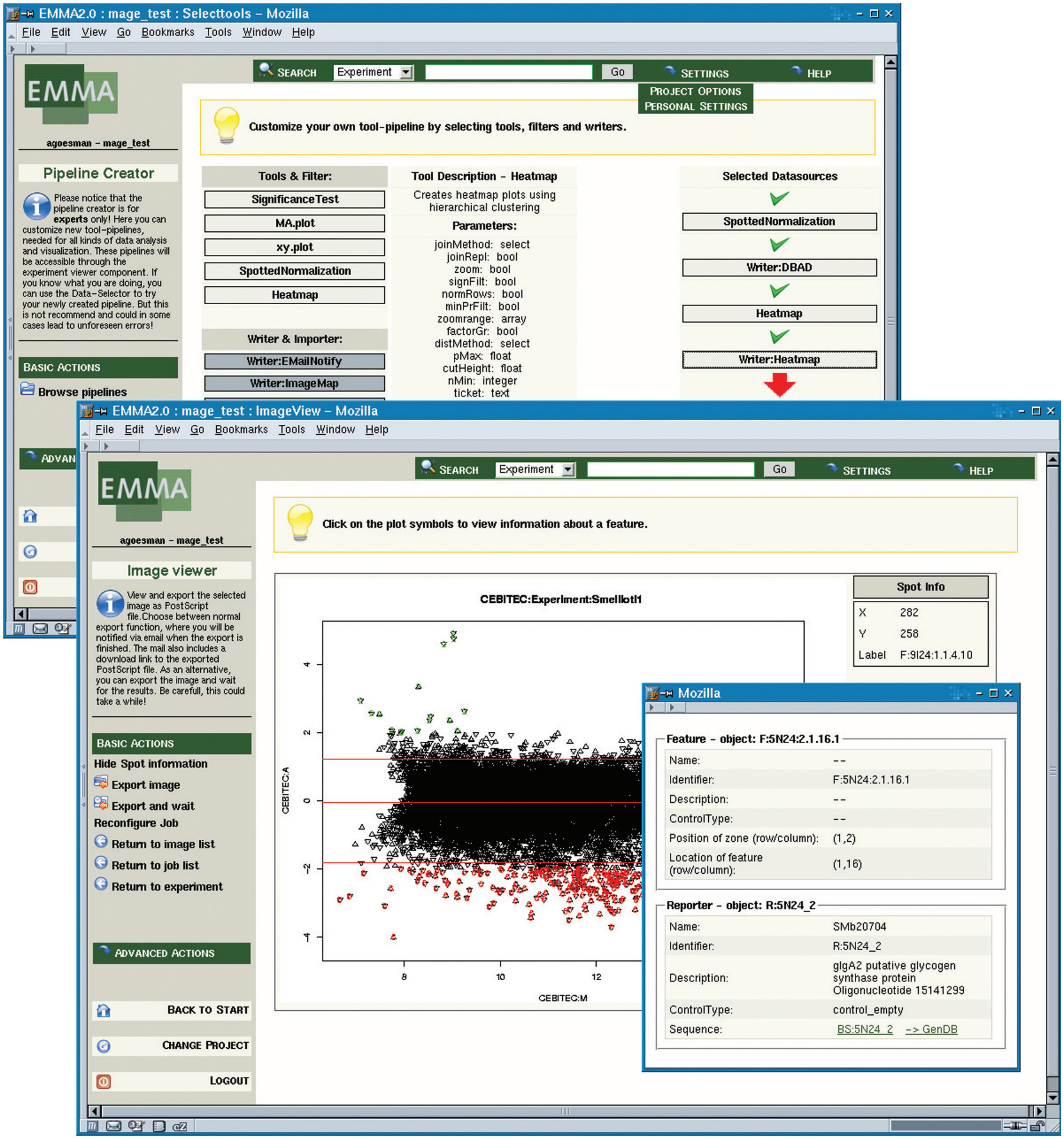
Figure 1. The newly implemented GenDB web interface provides a multitude of views for browsing a genome and for manipulating a genome annotation. The up-most screenshot shows the GenDB contig view that can be used for navigating from a region in the genome to a specific gene or region on a contig. An informal reconstruction of metabolic pathways can be visualized using the KEGG browser shown in the lower part of the screenshot: here, automatically annotated enzymes are highlighted in green.

and retrieval. Complex analyses like the automated annotation of a whole genome or the analyses of a large bundle of mass spectra (MS) can also be initiated and subsequently visualized via the web interface. Documentation for users as well as for programmers is provided in the form of a WIKI (http://moinmoin.wikiwikiweb.de/).

Details on the applications

In the following, we describe the three web applications in greater detail.

GenDB is an open source genome annotation system for prokaryotic genomes that has been in development for



agoesman – mage_test

Pipeline Creator

Please notice that the pipeline creator is for **experts** only! Here you can customize new tool–pipelines, needed for all kinds of data analysis and visualization. These pipelines will be accessible through the experiment viewer component. If you know what you are doing, you can use the Data–Selector to try your newly created pipeline. But this is not recommend and could in some cases lead to unforeseen errors!

Customize your own tool–pipeline by selecting tools, filters and writers.

Tools & Filter:	Tool Description – Heatmap	Selected Datasources
SignificanceTest	Creates heatmap plots using hierarchical clustering	✓
MA.plot	Parameters:	SpottedNormalization
xy.plot	joinMethod: select	✓
SpottedNormalization	joinRepl: bool	Writer:DBAD
Heatmap	zoom: bool	✓
	signFilt: bool	Heatmap
Writer & Importer:	normRows: bool	✓
Writer:EMailNotify	minPrFilt: bool	Writer:Heatmap
Writer:ImageMap	zoomrange: array	
	factorGr: bool	
	distMethod: select	
	pMax: float	
	cutHeight: float	
	nMin: integer	
	ticket: text	

agoesman – mage_test

Image viewer

View and export the selected image as PostScript file.Choose between normal export function, where you will be notified via email when the export is finished. The mail also includes a download link to the exported PostScript file. As an alternative, you can export the image and wait for the results. Be carefull, this could take a while!

Click on the plot symbols to view information about a feature.

Spot Info	
X	282
Y	258
Label	F:9I24:1.1.4.10

Feature – object: F:5N24:2.1.16.1

Name:	--
Identifier:	F:5N24:2.1.16.1
Description:	--
ControlType:	--
Position of zone (row/column):	(1,2)
Location of feature (row/column):	(1,16)

Reporter – object: R:5N24_2

Name:	SMb20704
Identifier:	R:5N24_2
Description:	glgA2 putative glycogen synthase protein Oligonucleotide 15141299
ControlType:	control_empty
Sequence:	BS:5N24_2 -> GenDB

Figure 2. Two examples of the user interface of the EMMA software. The up-most window displays the tool configuration wizard, which serves to define customized analysis pipelines. The building blocks of pipelines are functions or external programs known as plug-ins. The second screenshot depicts a scatterplot of the normalized log-expression versus the log-intensity where each gene is linked to its corresponding annotation.

more than five years. Given a genome sequence, the system integrates numerous tools to perform a gene prediction and a functional annotation of the genome.

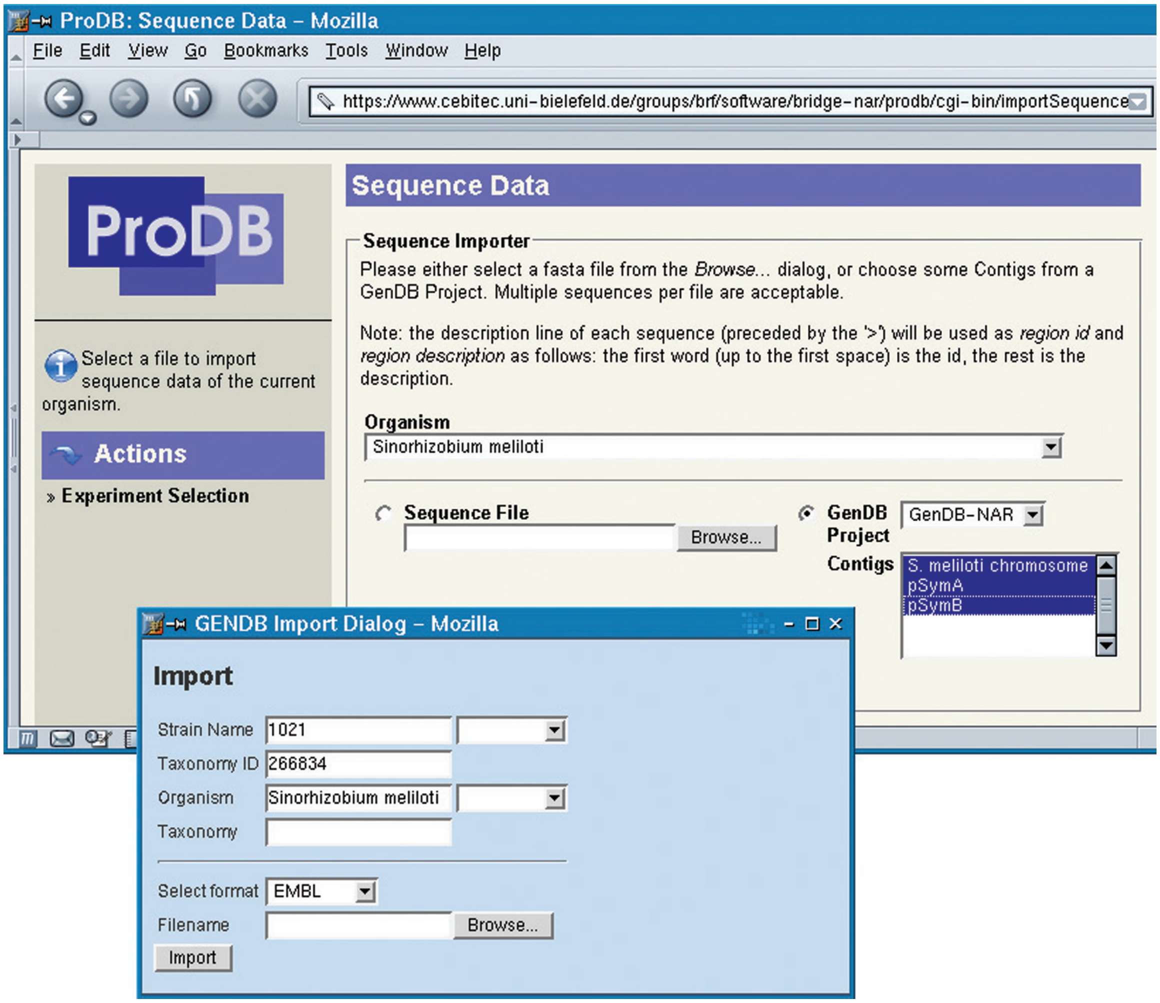
For the prediction of coding sequences (CDSs) we rely on an approach combining Glimmer (12) and Critica (13) [Reganor (14)]. For each CDS we perform an automatic function prediction (Metanor) using a combination of standard tools like BLAST (15), HMMer (16) and InterPro (17) as a basis for assigning a gene name, gene product, description, functional category, GO (18) numbers and other attributes. These automatic annotations can be curated and enriched manually via the web interface. In order to keep track of all automatic and manual annotations, GenDB stores a history of all annotations in the form of a list.

Among other views, the web interface provides a contig view (Figure 1) for easy navigation, a report on each CDS, a region editor for changing gene starts, a region creator for manually creating new genes and a virtual 2D gel. For navigating all genes according to their functional classification the system provides a KEGG (19) (Figure 1), COG (20) and GO browser. Import and export can be done for FASTA, EMBL and GenBank files. Currently, more than 25 genomes are being analyzed in various national and international cooperations using this web interface.

EMMA is a MAGE (21) compliant software platform for transcriptome data analysis including a LIMS component (ArrayLIMS). Data can be uploaded in standard formats and linked to the GenDB data. The system provides customizable pipelines for data processing and has a modular architecture that can easily be extended. Several visualization methods like scatter-plots or heat maps (Figure 2) are also available. EMMA features detailed reports about spots, genes and their corresponding measurements.

Data import and export are provided in a variety of formats. The complete MAGE-ML language is supported for array layouts, datasets and experimental descriptions. Datasets can be exported as tab-separated tables and in the binary format HDF5, which is also used for reliable storage of large quantification tables. EMMA supports the Array Description Format and MAGE-ML for defining array layouts. After creating the array layouts the contained sequences can be linked to GenDB automatically.

Fine grained access control is provided for every experiment, array and dataset on a user and group level. Upload and


ProDB: Sequence Data – Mozilla
File Edit View Go Bookmarks Tools Window Help
ProDB
Sequence Data
Actions
» Experiment Selection
Browse...
Import
GENDB Import Dialog – Mozilla
Browse...
Import

https://www.cebitec.uni-bielefeld.de/groups/brf/software/bridge-nar/prodb/cgi-bin/importSequence

Select a file to import sequence data of the current organism.

Sequence Importer

Please either select a fasta file from the *Browse...* dialog, or choose some Contigs from a GenDB Project. Multiple sequences per file are acceptable.

Note: the description line of each sequence (preceded by the '>') will be used as *region id* and *region description* as follows: the first word (up to the first space) is the id, the rest is the description.

Organism: Sinorhizobium meliloti

○ Sequence File

● GenDB Project: GenDB-NAR

Contigs: S. meliloti chromosome, pSymA, pSymB

Strain Name: 1021

Taxonomy ID: 266834

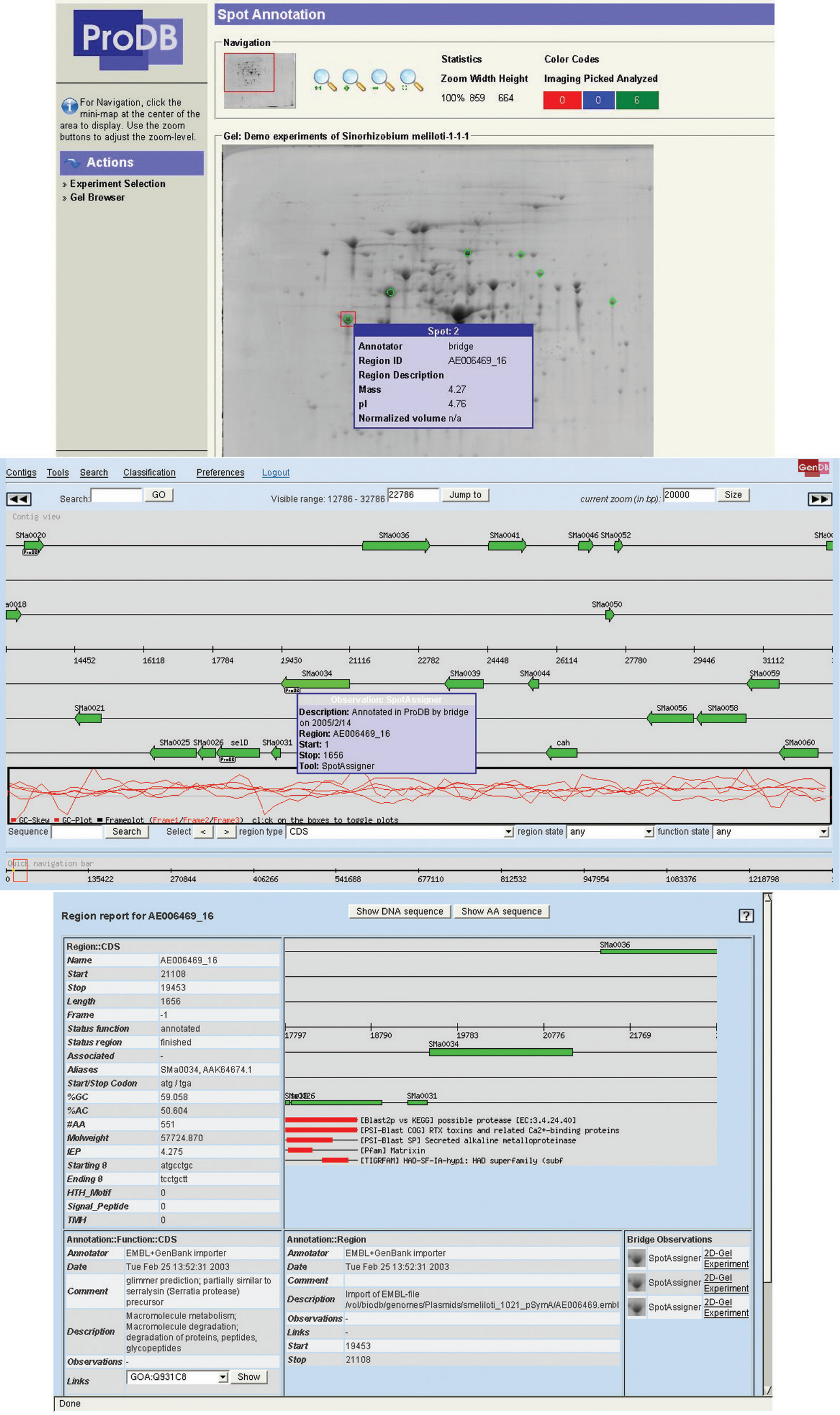
Organism: Sinorhizobium meliloti

Taxonomy:

Select format: EMBL

Filename:

Figure 3. Data import in GenDB and ProDB. In the GenDB Import Dialog the import for the *S.meliloti* sequence and annotations using an EMBL file is shown. The imported contigs can then be used for creating a protein database for the MS analyses in a ProDB project as shown in the upper screenshot.



For Navigation, click the mini-map at the center of the area to display. Use the zoom buttons to adjust the zoom-level.

Zoom	Width	Height
100%	859	664

Imaging	Picked	Analyzed
0	0	6

Gel: Demo experiments of Sinorhizobium meliloti-1-1-1

Spot: 2

Annotator	bridge
Region ID	AE006469_16
Region Description	
Mass	4.27
pI	4.76
Normalized volume	n/a

Search:

Visible range: 12786 - 32786 22786

current zoom (in bp): 20000

Observation: SpotAssigner

Description: Annotated in ProDB by bridge on 2005/2/14
Region: AE006469_16
Start: 1
Stop: 1656
Tool: SpotAssigner

Sequence: Search Select < > region type: CDS region state: any function state: any

Region report for AE006469_16

Region::CDS	
Name	AE006469_16
Start	21108
Stop	19453
Length	1656
Frame	-1
Status function	annotated
Status region	finished
Associated	-
Aliases	SMa0034, AAK64674.1
Start/Stop Codon	atg / tga
%GC	59.058
%AC	50.604
#AA	551
Molweight	57724.870
IEP	4.275
Starting 8	atgcctgc
Ending 8	tcctgctt
HTH_Motif	0
Signal_Peptide	0
TMH	0

```
[Blast2p vs KEGG] possible protease [EC:3.4.24.40]
[PSI-Blast COG] RTX toxins and related Ca2+-binding proteins
[PSI-Blast SP] Secreted alkaline metalloproteinase
[Pfam] Matrixin
[TIGRFAM] HAD-SF-IA-hyp1: HAD superfamily (subf
```

Annotation::Function::CDS	
Annotator	EMBL+GenBank importer
Date	Tue Feb 25 13:52:31 2003
Comment	glimmer prediction; partially similar to serralysin (Serratia protease) precursor
Description	Macromolecule metabolism; Macromolecule degradation; degradation of proteins, peptides, glycopeptides
Observations	-
Links	GOA:Q931C8 Show

Annotation::Region	
Annotator	EMBL+GenBank importer
Date	Tue Feb 25 13:52:31 2003
Comment	
Description	Import of EMBL-file /vol/biodb/genomes/Plasmids/smeliloti_1021_pSymA/AE006469.embl
Observations	-
Links	-
Start	19453
Stop	21108

Bridge Observations	
SpotAssigner	2D-Gel Experiment
SpotAssigner	2D-Gel Experiment
SpotAssigner	2D-Gel Experiment

Figure 4. Example for the ongoing integration of the BRIGEP applications. Spots in a 2D gel in ProDB are linked to CDS objects in GenDB. The spot marked with a red square in the up-most screenshot is highlighted in the GenDB contig view in the middle. The GenDB report for a selected region is enriched by images of corresponding proteins picked from a 2D gel in ProDB.

storage of experimental setups, RNA-extraction, hybridization conditions, scanned images and quantification data are handled by the included ArrayLIMS system.

ProDB is software for large-scale analysis of proteome data, including a LIMS component. ProDB stores experimental data, such as images of 2D gels or MS and allows automated data analysis and annotation of MS.

The system handles data from different mass spectrometer software (e.g. processed data from Bruker or Thermofinnigan) and will support the mzData standard from the PSI (22). Since ProDB stores MS together with numerous details about the experimental setup, the annotation is automatically linked to the corresponding spots on a gel.

The web interface provides data input and management of all experimental steps leading up to the MS data. We have implemented a common interface to different search engines like Mascot (23) or emowse (24) [contained in the EMBOSS package (25)]. In this interface, the user can define search sets consisting not only of one specific parameter for a MS search (e.g. peptide mass tolerance of 100 p.p.m.) but an interval for each parameter (e.g. peptide mass tolerance from 50 to 100 p.p.m. with steps of 25 p.p.m.). These sets of search parameters are used to analyze all selected MS. The results are then presented to the user for the annotation of the spectra.

Integrating the systems

Each of the described applications can be used stand-alone, but they can also be linked via the BRIDGE integration layer. Using this layer, data objects from different projects can be linked so that, for example, information originating from a GenDB project can be shown in the EMMA or ProDB web interface. We are in the process of tightly integrating the different applications to provide the user with the benefit of having all useful information present at each step of the analyses.

In this manner, we have linked the sequences spotted on an array to the corresponding GenDB genes. Thus, the user can jump directly from a spot in EMMA to the corresponding GenDB contig view, report or annotation dialog. Sequence data from GenDB can also be used to create the database for MS analyses. The results of the analyses can be linked to the corresponding sequence object stored within GenDB (Figure 3). Some examples of the ongoing integration can be found in the supplementary material.

To demonstrate the benefit of the integration a sample application is described as follows.

A sample application: enriching gene annotations with experimental evidence from 2D gel analyses. For this sample application we have imported the EMBL files of all replicons of the *Sinorhizobium meliloti* (26) genome into a GenDB project. Afterwards, we computed a new automatic functional annotation using the standard GenDB pipeline. The amino acid sequences of all CDSs and their updated gene annotations were then imported directly from GenDB into the ProDB demo project and installed as a searchable database for emowse.

Sample data from 2D gel analyses were imported into ProDB and several spots on a gel were identified from their peptide mass fingerprints. These spots were annotated within the ProDB system and by assigning the corresponding GenDB region the spot objects were linked directly to the GenDB CDS objects. At the same time, an observation was created within the GenDB system referring to the ProDB spot object, and these observations are listed in the GenDB region report showing a small image of the corresponding spot on a gel (Figure 4). Furthermore, the user can jump directly from the GenDB report to the ProDB 2D gel and experiment. Conversely, one can also navigate from an annotated spot to the GenDB contig view. Directly linking experimental data from 2D gel analyses is used in this example for creating enriched gene annotations with increased quality and reliability provided by experimental evidence.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank the users for their hints and suggestions to improve and extend the functionality and usability of our software. We are grateful to our student programmers Stefan Albaum, Thomas Bekel, Johannes T. Griebel, Sebastian Jünnemann, Tim Kahlke, Alexander Lenhardt, Daniel Paarmann, Oliver Rupp and Felix Tille. The work of A.G. and D.B. is funded by the 'Bundesministerium für Bildung und Forschung' (BMBF 0313105), B.L. and A.W. are funded by the 'Deutsche Forschungsgemeinschaft' (DFG PU28/25-3). M.D. acknowledges financial support by the 'International NRW Graduate School in Bioinformatics and Genome Research', L.K. and S.O. are supported by the 'DFG Graduiertenkolleg 635 Bioinformatik'. H.N. and T.P. are funded by the EU (GOCE-CT-2004-505403). Funding to pay the Open Access publication charges for this article was provided by Bielefeld University.

Conflict of interest statement. None declared.

REFERENCES

1. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res*., **29**, 126–127.
2. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.-A. and Barrell,B. (2000) Artemis: sequence visualisation and annotation. *Bioinformatics*, **16**, 945–994.
3. Overbeek,R., Larsen,N., Walunas,T., D Souza,M., Pusch,G., Selkov,E.J., Liolios,K., Joukov,V., Kaznadzey,D., Anderson,I. *et al*. (2003) The ERGO genome analysis and discovery system. *Nucleic Acids Res*., **31**, 164–171.
4. Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,A. and Peterson,C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol*, **3**, software 0003–1 0003.6.
5. Theilhaber,J., Ulyanov,A., Malanthara,A., Cole,J., Xu,D., Nahf,R., Heuer,M., Brockel,C. and Bushnell,S. (2004) GECKO: a complete large-scale gene expression analysis platform. *BMC Bioinformatics*, **5**, 195.
6. Dondrup,D., Goesmann,A., Bartels,D., Kalinowski,J., Krause,L., Linke,B., Rupp,O., Sczyrba,A., Pühler,A. and Meyer,F. (2003) EMMA: a platform for consistent storage and efficient analysis of microarray data. *J. Biotechnol*., **106**, 135–146.
7. Wilke,A., Rückert,C., Bartels,D., Dondrup,M., Goesmann,A., Hüser,A., Kespohl,A., Linke,B., Mahne,M., McHardy,A.C. *et al*. (2003)

Bioinformatics support for high-throughput proteomics. *J. Biotechnol.*, **106**, 147–156.

8. Facius,A., Englbrecht,C., Birzele,F., Groscurth,A., Schmid,B., Wanka,S. and Mewes,W. (2005) PRIME: a graphical interface for integrating genomic/proteomic databases. *Proteomics*, **5**, 76–80.
9. Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,C., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. and Pühler,A. (2003) GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.
10. Goesmann,A., Linke,B., Rupp,O., Krause,L., Bartels,D., Dondrup,D., McHardy,A.C., Wilke,A., Pühler,A. and Meyer,F. (2003) Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology. *J. Biotechnol.*, **106**, 157–167.
11. Linke,B. (2002) O2DBI II: ein Persistenzlayer für Perl-Objekte. Master's Thesis, Bielefeld University, Bielefeld, Germany.
12. 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.
13. Badger,H. and Olsen,G.J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.*, **16**, 512–524.
14. McHardy,A.C., Goesmann,A., Pühler,A. and Meyer,F. (2004) Development of joint application strategies for two microbial gene finders. *Bioinformatics*, **20**, 1622–1631.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
17. Apweiler,R., Attwood,T., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
18. The Gene Ontology Consortium (2000), Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
19. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
20. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
21. Spellman,P.T., Miller,M., Stewart,J., Troup,C., Sarkans,U., Chervitz,S., Bernhart,D., Sherlock,G., Ball,C., Lepage,M. *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*, **3**, research 0046.1–0046.9.
22. Orchard,S., Hermjakob,H. and Apweiler,R. (2003) The proteomics standard initiative. *Proteomics*, **3**, 1374–1376.
23. Perkins,D.N., Pappin,D.J., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
24. Eng,J.K., McCormack,A.L. and Yates,J.R.,III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.
25. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
26. Galibert,F., Finan,T., Long,S., Pühler,A., Abola,P., Ampe,F., Barloy-Hubler,F., Barnett,M., Becker,A., Boistard,P. *et al.* (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science*, **29**, 668–672.