Design and implementation of a database for genome annotation

Benoît de Hertogh, Leïla Lahlimi, Christophe Lambert, Jean-Jacques Letesson, Eric Depiereux

Title: Design and implementation of a database for *Brucella melitensis* genome annotation

Authors: Benoît De Hertogh, Leïla Lahlimi, Christophe Lambert, Jean-Jacques Letesson, Eric Depiereux

PII: S0378-1135(07)00433-6
DOI: doi:10.1016/j.vetmic.2007.09.010
Reference: VETMIC 3810

To appear in: *VETMIC*

Received date: 13-3-2006
Revised date: 15-1-2007
Accepted date: 13-9-2007

Please cite this article as: De Hertogh, B., Lahlimi, L., Lambert, C., Letesson, J.-J., Depiereux, E., Design and implementation of a database for *Brucella melitensis* genome annotation, *Veterinary Microbiology* (2007), doi:10.1016/j.vetmic.2007.09.010

Design and implementation of a database for *Brucella melitensis* genome annotation.

Benoît De Hertogh[1], Leïla Lahlimi[1], Christophe Lambert[2], Jean-Jacques Letesson[1], Eric Depiereux[1]*.

1. U.R.B.M., F.U.N.D.P., Rue de Bruxelles, 61, B-5000, Namur, Belgium.

2. BioXpr sa, Rue du séminaire, 22, B-5000, Namur, Belgium.

*To whom correspondence can be addressed:

Eric Depiereux,

Unité de Recherche en Biologie Moléculaire,

Facultés Universitaires Notre-Dame de la Paix,

Rue de Bruxelles, 61

B-5000 Namur,

Belgium

Tel: +32 81724415

Fax: +32 81724420

E-mail: **eric.depiereux@fundp.ac.be**

1 **Abstract**

2 The genome sequences of three *Brucella* biovars and of some species close to *Brucella sp.*

have become available, leading to new relationship analysis. Moreover, the automatic genome annotation of the pathogenic bacteria *Brucella melitensis* has been manually corrected by a consortium of experts, leading to 899 modifications of start sites predictions among the 3198 open reading frames (ORFs) examined. This new annotation, coupled with the results of automatic annotation tools of the complete genome sequences of the *B. melitensis* genome (including BLASTs to 9 genomes close to *Brucella*), provides numerous data sets related to predicted functions, biochemical properties and phylogenic comparisons.

To made these results avalaible, αPAGe, a functional auto-updatable database of the corrected sequence genome of *B. melitensis*, has been built, using the entity-relationship (ER) approach and a multi-purpose database structure. A friendly graphical user interface has been designed, and users can carry out different kinds of information by three levels of queries: (1) the basic search use the classical keywords or sequence identifiers; (2) the original advanced search engine allows to combine (by using logical operators) numerous criteria: (a) keywords (textual comparison) related to the pCDS's function, family domains and cellular localization; (b) physico-chemical characteristics (numerical comparison) such as isoelectric point or molecular weight and structural criteria such as the nucleic length or the number of transmembrane helix (TMH); (c) similarity scores with *E. coli* and 10 species phylogenetically close to *B. melitensis*; (3) complex queries can be performed by using a SQL field, which allows all queries respecting the database's structure.

The database is publicly available through a Web server at the following url:

http://www.fundp.ac.be/urbm/bioinfo/aPAGe

Keywords: *Brucella melitensis*, relational, database, annotation, db-main.

Introduction

Brucella melitensis

Brucellosis is a worldwide distributed bacterial disease, affecting reindeers in Alaska and Siberia, camels in the Middle East, cattles, pigs, goats, sheeps.... Sheeps and goats are the most susceptible. In human, the disease is known as Malta fever, undulating fever, mediterranean fever or Melitococcie. It causes abortion in females and orchitis in males. Humans catch this disease from contaminated meat or milk, aborted fetuses and slaughtering infected animals (Godfroid *et al.* 2005).

The six known species of *Brucella* (Proteobacteria; Alphaproteobacteria; Rhizobiales; Brucellaceae; *Brucella*), of which four are pathogenic for human (*B. melitensis, B. abortus, B. suis and B. canis*) are often regarded today as being biovars of one single species (for contradictory informations, see Chain *et al.*, 2005).

Table 1

The development of many chronic diseases involves interactions between environmental factors and genes that regulate important physiological processes. Several projects are trying to understand those processes, and to develop methods to predict toxicity and understand the genetic basis of differential susceptibility (Mattingly *et al.*, 2004). Genome sequencing projects led to new disciplines in biology, and promise a better comprehension of disease-associated genes (Miller and Kumar, 2001).

The availability of the genome sequence of three *Brucella* biovars and eight species phylogenetically close to *Brucella sp.* (Table 1) opens the way to genomic comparisons that may elucidate the molecular mechanisms of brucellosis, but also phylogenetic and evolutionary relationships, speciation and emergence of new gene families.

The correction consortium

The 3198 pCDS of the *Brucella melitensis* genome were sequenced and automatically annotated by Integrated Genomics Inc. (Delvecchio *et al.*, 2002). However, many errors were notified in predictions of the position of the start codon of these pCDS. The Research Unit in Molecular Biology (URBM) initiated a project of correction of the start position of all the genome sequences. This project was carried out in collaboration with four research teams, members of the European consortium of research COST845 (VLA Weybridge, U.K.; U. Navarra, Spain; U. Cantabria, Spain; Inserm U. 431, France) (Dricot, 2004).

Objectives

Through this paper, we had five main objectives: (1) to made available the Consortium corrections (2) and the results of the automatic annotation (3) in a polyvalent auto-updatable and easily upgradable (i.e. to others genomes) database (4) using a performing DataBase Management System (DBMS), in order to allow complex queries by the use of (5) performing and user friendly search tools.

Material and methods

The correction protocol

Each pCDS was manually checked by at least two scientific teams, using the following protocol:

1. Genome files of *Brucella melitensis* (NC_003318.gbk and NC_003317.gbk) had been downloaded from the National Center for Biotechnology Information (NCBI) ftp site. The free Artemis software was used to visualize annotations;
2. For each ORF, the predicted start was checked. The wrong starts were detected by carrying out the following steps:
    a. Control of the start codon. TTG is usually a wrong start site;

b. Detection of overlapping ORFs, which usually hide a wrong start codon;

c. Analyses of the corresponding protein sequence alignment against the Non-Redundant (NR) database from NCBI. If similar sequences in phylogenetically close organisms are longer or shorter, the start site is probably wrong;

3. A new start site was checked using Artemis software with, in increasing preferences corresponding to their observed frequency, start codons: ATG, GTG and TTG;

4. If several potential start positions were found, without operon or Ribosome Binding Site (RBS) (Salgado *et al.*, 2000) identified, the pCDS having the closest size to homologous sequences detected in NR was selected.

Otherwise, the longer pCDS with an ATG or GTG was chosen.

Functional annotation

Functional annotations of proteins translated from the ORFeome library are done *in silico* using the following programs:

1. BLASTP) against the NR protein database from GenBank and Swiss-Prot databases;

2. BLASTN against the genomes listed in Table 1;

3. hmmpfam against the 8183 proteins of the Protein Family database (Pfam 19.0);

4. Prediction of the cellular localizations by an updated version of PSORTII (Nakai *et al.*, 1999) for Gram-negative bacteria. PSORTII examines a given protein sequence for amino acid composition, similarity to proteins of known localization, presence of a signal peptide, transmembrane alpha-helices and motifs

corresponding to specific localizations. A probabilistic method integrates this analysis, returning a list of five possible localization sites with associated probability scores;

5. Prediction of the transmembrane segments are predicted by TMHMM v.2.0, the most reliable transmembrane prediction program (Möller, 2001);

6. Prediction of the secondary structure of each pCDS using the PSIPRED2 protein structure prediction server (McGuffin *et al.*, 2000);

7. Prediction of the three-dimensional structure by ESyPred3D (Lambert *et al.*, 2002), with alignment of the peptidic sequences against the Protein data bank (PDB) (Berman *et al.*, 2000). This program produces a multiple alignment of the query sequence with several sequences from the PDB, and builds a consensus of high reliability. This reliable alignment is subsequently used for the building of the homology model.

The DB-MAIN CASE tool

Because a genome annotation generates numerous data sets that cannot be easily managed with simple file management software, a powerful and complex knowledge-management software, DB-MAIN, had been used to create a database using the relational model in accordance with the ER model (Entity Relationship). DB-MAIN is a data-oriented Computer Aided Software Engineering (CASE) environment, designed to support most database engineering processes, including: (1) requirement analysis, conceptual design, normalisation, schema integration, logical design, physical design, schema's optimisation and code generation; (2) schema transformation, model transformation; (3) schema analysis, code analysis, data reverse engineering; (4) database migration, database evolution, database integration and federation, data wrapper design and generation; (5)

temporal database design, active database evaluation and generation.

In addition, the method-modelling component allows the user to define any of these sub-models (Englebert and Hainaut, 1999; http://www.db-main.be/).

The database design

The database design follows globally the "Merise" method. It starts with the extraction of the entities from the application domain (in this paper, the data sets provided by both manual and automatic annotation of the *B. melitensis* genome). These entities will constitute the database classes, which models the real-world organization and its important data elements and relationships.

Although there are many conceptual models drawn for biological databases, the conceptual schema remains specific to the application domains, which are various and depend on projects subject. The conceptual model is an extension of the Entity-Relationship model. It defines the logical relationships that link the database entities and the object attributes, in order to build the data structure. This step involves the description of the entire information content of the database.

The next step, the logical design, is the translation of the conceptual schema in logical schema. These schemas represent the same information, but the second one (logical) expresses the data through the construct of the DBMS. It shows how data are organised in a relational way: the database is represented as a group of related tables that can be managed by a relational database management system (RDBMS). Finally, the entity relationship model is translated into a database structure that can be afterwards easily expressed using structured query language (SQL).

Both conceptual and logical schemas are available on the database website.

The web interface

The data sets are stored in a MySQL (http://www.mysql.com/) relational database. The database can be queried through a web interface built using PHP Hypertext Preprocessor (PHP), a widely used general-purpose scripting language especially suited for web development and more specifically to be connected to a database (http://www.php.net/).

Machines

The annotation of the genome and the construction of the database were carried out on Silicon Graphics Octane duo and a cluster of PC (Table 2).

Results

Increasing of pCDS reliability

The correction project consortium has increased the database information reliability. Start positions have been corrected for 899 pCDS. The mean difference in position is 69 nucleotides or 23 amino acids. 565 pCDS have been shortened and 334 have been lengthened. The corrected pCDS are mentioned in the commentary field.

The αPAGe database

The sequence databases offer a great source of information for studies on biological variation, evolutionary patterns and protein family characterization. In order to find all pieces of information available, it is often necessary to search several databases or to click on several links. Facing the complexity to obtain information, each **α**PAGe entry includes the whole information about a sequence (functional and structural annotation, pCDS's properties, links to other biological databases...).

The identification and the functional characterization of genes, that may be involved in disease development or may be responsible for virulence, is a critical step. An advanced

search tool form–allows the combination of different criteria and may help to select a subset of pCDS for further analysis. Most databases propose advanced search tools based only on keywords searches. The **α**PAGe makes possible to perform searches combining several criteria, with the possibility to impose choice constraints (and/or) at each level (see below). The database is publicly available through a Web server at http://www.fundp.ac.be/urbm/bioinfo/aPAGe.

ORFeome library

Entire libraries composed of all protein-encoding Open Reading Frames (ORF) cloned in highly flexible vectors represent a new type of resource, which is needed to take full advantage of informations generated by the sequencing efforts and to address the new type of question and hypothesis generated in post-genomic area.

Thus, the complete genome sequence of *B. melitensis* had been used to generate a protein-coding ORF database. This ORFeome library contains 3091 Gateway entry clones, each one corresponding to a defined ORF. This strategy may help to validate the genome annotation and to create a resource to functionally characterize the proteome (Dricot *et al.*, 2004). The cloning state field displays the information available for this experiment such as the forward and reverse primers used for each cloned pCDS.

Moreover, the "Similarity (BLASTN) in close organisms" field allows a quick overview of the distribution of the homologous sequences in the genomes of 10 alpha-proteobacteria, phylogenetically close to each other, and *Escherichia coli K12*, as Gram-negative model.

In order to facilitate the databases mining, links are proposed to display the precomputed best hits against the following databases: GenBank, Protein Information Resource (PIR), Protein Data Bank (PDB), DNA Data Bank of Japan (dbj), the Protein Research Foundation (PRF) (www4.prf.or.jp/en/), the European Molecular Biology Laboratory (EMBL) and the Protein Family (Pfam) database.

Cellular localization and membrane topology

The cellular localization field contains three parts: the observed cellular localization (often unknown), the predicted cellular localization (PSORT II) and the compatibility between predicted and observed subcellular localization.

The membrane topology and the number of transmembrane helices (TMHs) are also precomputed (TMHMM).

Secondary and three-dimensional structure

The predicted secondary structures are coloured in the following way: helices, beta sheets, and coils are represented respectively in red, blue and grey. Transmembrane segments are written with underlined bold characters.

The 3D structure field reports the homology modelling procedure if a protein with a detectable similarity is found in the PDB. It includes two parts:

1. The "3D structure field", which proposes a link to download the 3D final model of the treated pCDS. The model may be displayed using an external software (i.e. the Swiss PDB viewer (http://www.expasy.org) or a browser plug in (Chime)).
2. The "Modelling characteristics" field summarizes information used during the modelling step: the template (protein with known 3D structure) used for the alignment, the model building, the template experimental method and its parameters (resolution and R-value), the percentage of sequence modelled and a link to find more information and coordinates of the template structure at the PDB web site. Models are updated each month or at each new PDB release.

pCDS properties

The last sets of predictions concerns pCDS start and end positions, GC content, theoretical

physico-chemical properties, that includes molecular weight and isoelectric point and nucleic and peptidic sequence in FASTA format.

Promoter and termination sites can easily be studied by displaying the sequence of a chosen number of nucleotides located upstream or downstream of the pCDS. This tool may be useful to identify the non-coding upstream and downstream of a pCDS. It also allows to define the pCDS extremities for the cloning experiments.

References and cross-references

The cross-reference field allows users to be redirected towards KEGG and translated EMBL (TrEMBL) public databases. In a reference field, users are invited to add references related to the pCDS treated.

User interface

The database can be browsed using a powerful graphical user interface. One may consider three levels of queries:

Basic search

The database can be browsed:

1. By the pCDS ID;
2. By keywords (text search). Various fields can be targeted: GenBank annotation, Swiss-Prot annotation, observed function, comments, predicted or observed localization, Pfam summary, templates used in the 3D modelling procedure or a key word related to the cloning state;
3. By the display of the complete list of ORFs;
4. By similarity (using BLAST). Peptidic or nucleic sequences can be blasted against the whole genomes of *B. melitensis*, *B. suis* and *B. abortus*;

5. By using regular expression to search for specific patterns (nucleic or peptidic).

Advanced search

An advanced search tool allows users to limit queries to some defined fields or to combine search terms with logical operators.

Following fields are available:

1. Keywords searches (text comparison) described in the basic search.
2. Quantitative data (numerical comparisons using "<" and ">" operators) including:
    a. Physico-chemical properties, such as isoelectric point or molecular weight;
    b. Structural properties such as:
        i. The number of trans-membrane helices (TMHs);
        ii. The target-template percentage identity (three dimensional structure prediction);
        iii. The percentage identity of the deduced sequence modelled using the ESyPred3D system.
    c. The pCDS properties such as:
        i. The percentage in GC;
        ii. The position of the first and the last nucleotide;
        iii. The nucleotidic and peptidic lengths.

It is also possible to add constraints (expected value range) related to similarities between the *B. melitensis* pCDS considered and the pCDS of at least one of the 11 species described in Table 1.

Figure 1

To allow combinations of quantitative data with the logical operators "and" and "or" and with the comparisons operators "<" and ">" is probably the most interesting particularity of the advanced search tool (Fig. 1). This unusual way to proceed permits to extract data in order to perform statistical analysis. For example, the study of the frequency distribution of the predicted pI shows a bimodal distribution (Fig. 2). It is well known that, in disruptive selection, selection pressures act against individuals in the middle of the trait distribution, resulting in a bimodal curve. Thus, this predicted pI bimodal distribution seems to suggest an important evolutionary selection pressure applied against *B. melitensis*, which unfavorise the *B. melitensis* ORFs around pI's of 8. However, for different authors, this multimodal distribution is an effect of allowed combinations of the charged amino acids, and not due to evolutionary causes (*i.e.* Nandy *et al*., 2005; Schwartz *et al*., 2001).
Similarly, data extraction for various statistical analysis, implying one or more fields, can easily be performed.

Figure 2

SQL search

For advanced users, data may be retrieved from the Web interface using SQL queries.

Conclusion and Perspectives

We developed, under DBMS, a relational updatable database dedicated to the pathogenic bacteria *Brucella melitensis*, from the sequenced genome manually corrected by a

specialists consortium and from the data generated *in silico* by several among the most powerful prediction programs.

Search tools are efficient, friendly, but also original. In the advanced search, users may combine, by using logical operators, several criteria related to the pCDS properties in order to select one or more specific pCDS: (1) keywords (textual comparison) related to the pCDS's function, family domains and cellular localization; (2) physico-chemical characteristics (numerical comparison) such as isoelectric point, molecular weight or structural criteria such as the nucleic length or the number of TMHs; (3) homology (or not) in 10 species phylogenetically close to *B. melitensis* (expected value range).

The structure of the database allows its extension to other genomes, especially those phylogenetically close to *Brucella melitensis*, but also to model organism like *Saccharomyces cerevisiae, Drosophila melanogaster* or *Homo sapiens*. The functional and structural annotation of these genomes could be incorporated automatically from nucleotidic sequences.

The extension of the advanced search to new fields, like TMHs or motives is also on the way. Such improvements will permit more complex queries.

New advanced search tool form that deal with all the data types and new links to biological databases for complementary information will also be added.

Finally, a special attention must be paid in order to make sure that the alteration is compatible with the existing database. The DB-MAIN CASE tool supports such meta-data evolution. The approach relies on a generic database model and on the transformational paradigm that states that database engineering processes can be modelled by schema transformation. Indeed, a transformation provides both structural and instance mappings that formally define how to modify database structures and contents.

Acknowledgements

The authors gratefully acknowledge the aMAZE team and BioXpr colleagues for their help for the database conceptual schema design. They also address a special thank the DB-MAIN team for their assistance and to Fabrice Berger and Xavier De Bolle for critical discussions and comments.

References

Alsmark, C.M., Frank, A.C., Karlberg, E.O., Legault, B.A., Ardell, D.H., Canback, B., Eriksson, A.S., Naslund, A.K., Handley, S.A., Huvet, M., Scola, B.L., Holmberg, M. and Andersson, S.G., 2004. The louse-borne human pathogen *Bartonella quintana* is a genomic derivative of the zoonotic agent *Bartonella henselae*. Proc. Natl. Acad. Sci. USA., 101, 9716-9721.

Barnett, M.J., Fisher, R.F., Jones, T., Komp, C., Abola, A.P., Barloy-Hubler, F., Bowser, L., Capela, D., Galibert, F., Gouzy, J., Gurjal, M., Hong, A., Huizar, L., Hyman, R.W., Kahn, D., Kahn, M.L., Kalman, S., Keating, D.H., Palm, C., Peck, M.C., Surzycki, R., Wells, D.H., Yeh, K.C., Davis, R.W., Federspiel, N.A. and Long, S.R., 2001. Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid. Proc. Natl. Acad. Sci. USA., 98, 9883-9888.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.M., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y., 1997. The complete genome sequence of *Escherichia coli* K-12. Science. 277, 1453-1474.

Chain PS, Comerci DJ, Tolmasky ME, Larimer FW, Malfatti SA, Vergez LM, Aguero F, Land ML, Ugalde RA, Garcia E., 2005. Whole-genome analyses of speciation events in pathogenic Brucellae. Infect. Immun. 73, 8353-61.

365 Delvecchio, V.G., Kapatral, V., Redkar, R.J., Patra, G., Mujer, C., Los, T., Ivanova, N.,
366 Anderson, I., Bhattacharyya, A., Lykidis, A., Reznik, G., Jablonski, L., Larsen, N.,
367 D'souza, M., Bernal, A., Mazur, M., Goltsman, E., Selkov, E., Elzer, P.H., Hagius, S.,
368 O'callaghan, D., Letesson, J.J., Haselkorn, R., Kyrpides, N. and Overbeek, R., 2002. The
369 genome sequence of the facultative intracellularpathogen *Brucella melitensis*. Proc. Natl.
370 Acad. Sci. USA. 99, 443-448.
371 Dricot, A., Rual, J.F., Lamesch, P., Bertin, N., Dupuy, D., Hao, T., Lambert, C., Hallez, R.,
372 Delroisse, J.M., Vandenhaute, J., Lopez-Goñi, I., Moriyon, I., Garcia-Lobo, J.M., Sangari,
373 F.J., Macmillan, A.P., Cutler, S.J., Whatmore, A.M., Bozak, S., Sequerra, R., Doucette-
374 Stamm, L., Vidal, M., Hill, D.E., Letesson, J.J. and Debolle, X., 2004. Generation of
375 *Brucella melitensis* ORFeome version 1.1. *Genome Research*. 14, 2201-2206.
376 Englebert, V. and Hainaut, J.L., 1999. DB-MAIN: A Next Generation Meta-CASE.
377 Journal of Information Systems. 24, 99-112.
378 Galibert F., Finan T.M., Long S.R., Puhler A., Abola P., Ampe F., Barloy-Hubler F.,
379 Barnett M.J., Becker A., Boistard P., Bothe G., Boutry M., Bowser L., Buhrmester J.,
380 Cadieu E., Capela D., Chain P., Cowie A., Davis R.W., Dreano S., Federspiel N.A., Fisher
381 R.F., Gloux S., Godrie T., Goffeau A., Golding B., Gouzy J., Gurjal M., Hernandez-Lucas
382 I., Hong A., Huizar L., Hyman R.W., Jones T., Kahn D., Kahn M.L., Kalman S., Keating
383 D.H., Kiss E., Komp C., Lelaure V., Masuy D., Palm C., Peck M.C., Pohl T.M., Portetelle
384 D., Purnelle B., Ramsperger U., Surzycki R., Thebault P., Vandenbol M., Vorholter F.J.,
385 Weidner S., Wells D.H., Wong K., Yeh K.C., Batut J., 2001. The composite genome of the
386 legume symbiont *Sinorhizobium meliloti*. Science. 293:668-72.
387 Godfroid J., Cloeckaert A., Liautard J.P., Kohler S., Fretin D., Walravens K., Garin-Bastuji
388 B., Letesson J.J., (2005) From the discovery of the Malta fever's agent to the discovery of a
389 marine mammal reservoir, brucellosis has continuously been a re-emerging zoonosis. Vet.
390 Res. 36:313-26.

391 Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B.
392 S., Cao, Y., Askenazi, M., Halling, C., Mullin, L., Houmiel, K., Gordon, J., Vaudin, M.,
393 Iartchouk, O., Epp, A., Liu, F., Wollam, C., Allinger, M., Doughty, D., Scott, C., Lappas,
394 C., Markelz, B., Flanagan, C., Crowell, C., Gurson, J., Lomo, C., Sear, C., Strub, G., Cielo,
395 C. and Slater, S., 2001. Genome sequence of the plant pathogen and biotechnology agent
396 *Agrobacterium tumefaciens* C58. Science. 294:2323-2328.
397 Halling S.M., Peterson-Burch B.D., Bricker B.J., Zuerner R.L., Qing Z., Li L.L., Kapur V.,
398 Alt D.P., Olsen S.C. (2005) Completion of the genome sequence of *Brucella abortus* and
399 comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*. J
400 Bacteriol. 187:2715-26.
401 Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A.,
402 Idesawa, K., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C.,
403 Kohara, M., Matsumoto, M., Matsuno, A., Mochizuki, Y., Nakayama, S., Nakazaki, N.,
404 Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M. and Tabata, S., 2000. Complete
405 genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA
406 Res. 7:331-338.
407 Lambert, C., Leonard, N., Xavier De Bolle and Depiereux, E., 2002. ESyPred3D:
408 Prediction of proteins 3D structures. Bioinformatics. 18:1250-1256.
409 Larimer, F.W., Chain, P., Hauser, L., Lamerdin, J., Malfatti, S., Do, L., Land, M.L.,
410 Pelletier, D.A., Beatty, J.T., Lang, A.S., Tabita, F.R., Gibson, J.L., Hanson, T.E., Bobst,
411 C., Torres, J.L., Peres, C., Harrison, F.H., Gibson, J. and Harwood, C.S. (2004) Complete
412 genome sequence of the metabolically versatile photosynthetic bacterium
413 *Rhodopseudomonas palustris*. Nat. Biotechnol. 22:55-61.
414 Mattingly, C.J., Colby, G.T., Rosenstein, M.C., Forrest, J.N. and Boyer, J.L., 2004.
415 Promoting comparative molecular studies in environmental health research: an overview of
416 the comparative toxicogenomics database (CTD). Pharmacogenomics J., 4, 5-8.

417 Mcguffin, L.J., Bryson, K. and Jones, D.T., 2000. The PSIPRED protein structure
418 prediction server. Bioinformatics. 16:404-405.
419 Miller, M. and Kumar, S., 2001. Understinding human disease mutations through the use
420 of interspecific genetic variation. Hum. Mol. Genet. 10:2319-2328.
421 Möller, S., Croning, M.D.R. and Apweiler, R., 2001. Evaluation of methods for the
422 prediction of membrane spaning regions. Bioinformatics. 17, 646-653.
423 Nandi S., Mehra N., Lynn A.M., Bhattacharya A., 2005. Comparison of theoretical
424 proteomes: Identification of COGs with conserved and variable pI within the multimodal
425 pI distribution. BMC Genomics. 6:116.
426 Nierman, W.C., Feldblyum, T.V., Laub, M.T., Paulsen, I.T., Nelson, K.E., Eisen, J.A.,
427 Heidelberg, J.F., Alley, M.R., Ohta, N., Maddock, J.R., Potocka, I., Nelson, W.C., Newton,
428 A., Stephens, C., Phadke, N.D., Ely, B., Deboy, R.T., Dodson, R;J., Durkin, A.S., Gwinn,
429 ML., Haft, D.H., Kolonay, J.F., Smit, J., Craven, M.B., Khouri, H., Shetty, J., Berry, K.,
430 Utterback, T., Tran, K., Wolf, A., Vamathevan, J., Ermolaeva, M., White, O., Salzberg,
431 S.L., Venter, J.C., Shapiro, L., Fraser, C.M. and Eisen, J., 2001. Complete genome
432 sequence of *Caulobacter crescentus*. Proc. Natl. Acad. Sci. USA.. 98, 4136-4141.
433
434 Paulsen, I.T., Seshadri, R., Nelson, K.E., Eisen, J.A., Heidelberg, J.F., Read, T.D., Dodson,
435 R.J., Umayam, L., Brinkac, L.M., Beanan, M.J., Daugherty, S.C., Deboy, R.T., Durkin,
436 A.S., Kolonay J.F., Madupu, R., Nelson, W.C., Ayodeji, B., Kraul, M., Shetty, J., Malek,
437 J., Aken, S.E.V., Riedmuller, S., Tettelin, H., Gill, S.R., White, O., Salzberg, S.L., Hoover,
438 D.L., Lindler, L.E., Halling, S.M., Boyle, S.M. and Fraser, C.M., 2002. The *Brucella suis*
439 genome reveals fundamental similarities between animal and plant pathogens and
440 symbionts. Proc Natl. Acad. Sci. USA.. 99, 13148-13153.
441 Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides, J., 2000. Operons in
442 Escherichia coli: genomic analyses and predictions. Proc. Natl. Acad. Sci., 97, 6652-6657.

443 Sanchez, D.O., Zandomeni, R.O., Cravero, S., Verdun, R.E., Pierrou, E., Faccio, P., Diaz,
444 G., Lanzavecchia, S., Aguero, F., Frasch, A.C., Andersson, S.G., Rossetti, O.L., Grau, O.
445 and Ugalde, R.A., 2001. Gene discovery through genomic sequencing of Brucella abortus.
446 Infect. Immun. 69, 865-868.
447 Schwartz, R., Ting, C.S., King, J. ,2001. Whole proteome pI values correlate with
448 subcellular localizations of proteins for organisms within the three domains of life.
449 Genome Res. 11:703-709.
450 Wood, D.W., Setubal, J.C., Kaul, R., Monks, D.E., Kitajima, J.P., Okura, V.K., Zhou, Y.,
451 Chen, L., Wood, G.E., Almeida, N.F., Woo, L., Chen, Y., Paulsen, I.T., Eisen, J.A., Karp,
452 P.D., Bovee, D., Chapman, P., Clendenning, J., Deatherage, G., Gillet, W., Grant, C.,
453 Kutyavin, T., Levy, R., Li, M.J., Mcclelland, E., Palmieri, A., Raymond, C., Rouse, G.,
454 Saenphimmachak, C., Wu, Z., Romero, P., Gordon, D., Zhang, S., Yoo, H., Tao, Y.,
455 Biddle, P., Jung, M., Krespan, W., Perry, M., Gordon-Kamm, B., Liao, L., Kim, S.,
456 Hendrick, C., Zhao, Z.Y., Dolan, M., Chumley, F., Tingey, S.V., Tomb, J.F., Gordon, M.
457 P., Olson, M.V. and Nester E.W., 2001. The genome of the natural genetic engineer
458 *Agrobacterium tumefaciens* C58. Science, 294, 2317-2323.

459 **Figure legends**

460

461 Fig. 1. Combinations of numerous fields and both logical and comparisons operators allow
462 to construct complex queries through a user-friendly interface.

463

464 . Fig. 2. The bimodal distribution of the predicted isoeletric point seems to suggest an
465 important evolutionary selection pressure to *B. melitensis*, but some authors invalid this
466 suggestion (see text) (each classes defines an half-unity of pI).

467

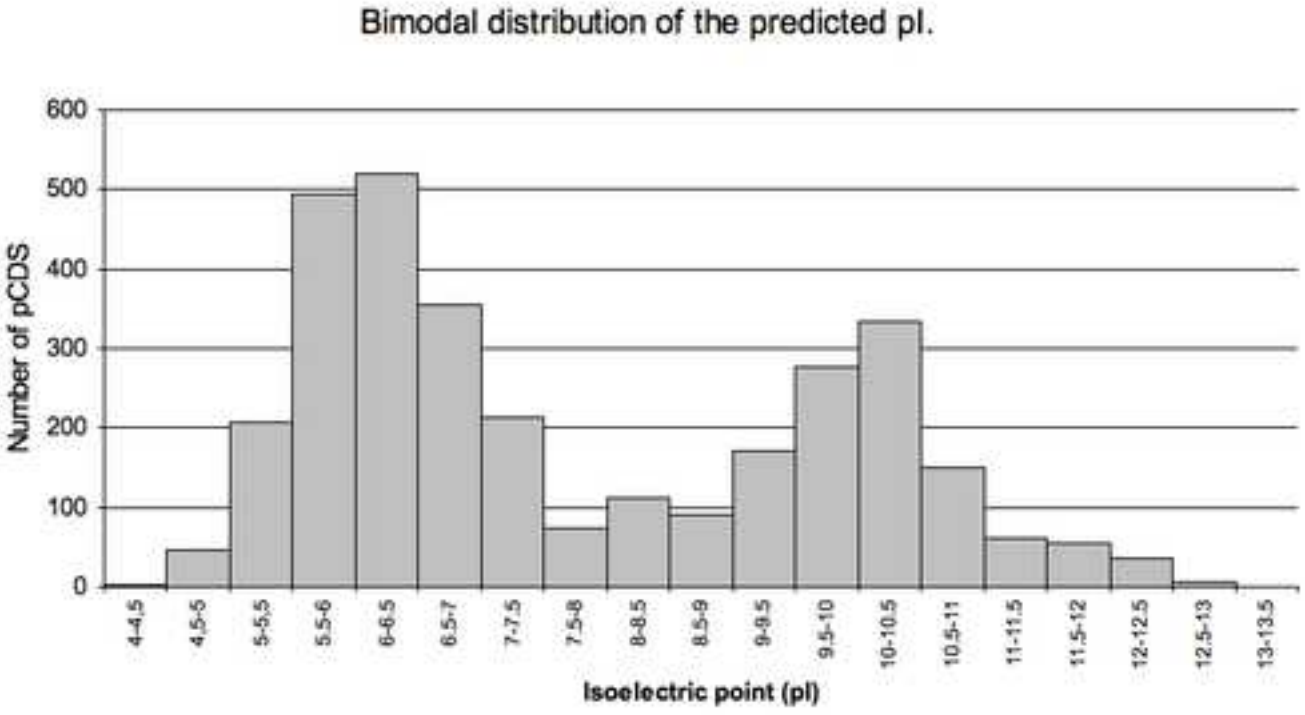

Table 1. Characteristics of the 10 species phylogenetically close to *Brucella sp.* present in the database.

Escherichia coli is used as a model Gram-negative bacteria (from EMBL-EBI and quoted authors).

Brucella biovars	Reference	Life cycle	Host	Genome	Orfs
Brucella abortus	Sanchez *et al.*, 2001; Halling *et al.*, 2005	pathogenic	human and livestock	2,1 MB and 1,2 Mb circular chromosomes	3,076
Brucella melitensis	Delvecchio *et al.*, 2002	pathogenic	human and livestock	2,1 MB and 1,2 Mb circular chromosomes	3,197
Brucella suis	Paulsen *et al.*, 2002	pathogenic	human and livestock	2,1 MB and 1,2 Mb circular chromosomes	3,256.
Close to Brucell*a*					
Agrobacterium tumefaciens C58	Wood *et al.*, 2001; Goodner *et al.*, 2001	pathogenic	numerous plants (crown, roots and stems)	2,8 MB circular chromosome; 0,54 Mb Plasmid pMLa; 0,2 Mb Plasmid pMLa	5,304
Bartonella henselae str. Houston-1,	Alsmark *et al.*, 2004	pathogenic	human and cat	1,9 Mb circular chromosome	1,464
Bartonella quintana str. Toulouse	Alsmark *et al.*, 2004	pathogenic	human specific	1,6 Mb circular chromosome	1,137
Caulobacter crescentus	Nierman *et al.*, 2001		dilute aquatic environment	1,2 Mb circular chromosome	3,718
Mesorhizobium loti	Kaneko *et al.*, 2000	symbiotic	nitrogen-fixing soil plant	7 Mb circular chromosomes; 0,4 Mb Plasmid pMLa; 0,2 Mb Plasmid pMLa	7,255
Sinorhizobium meliloti	Barnett *et al.*, 2001; Galibert *et al.*, 2001	symbiotic	nitrogen-fixing soil plant	3,6 Mb chromosome and two megaplasmids, pSyma (1,3 Mb) and pSymb (1,7 Mb)	6,148
Rhizobium leguminosarum	unpublished data	symbiotic	nitrogen-fixing soil plant	5 Mb circular chromosomes and six plasmid	-
Rhodopseudomonas palustris	Larimer *et al.*, 2004	metabolically versatile	soils and water	5,5 Mb circular chromosome; 8,427 bp plasmid pRPA	4,798
GRAM -					
Escherichia coli K12	Blattner *et al.*, 1997	pathogenic	human	4,6 Mb circular chromosome; 1 MB plasmid F	4,338

Table 2. Machines dedicated to the *B. melitensis* genome database.

	Processors	**CPU**	**RAM**	**OS**
Silicon Graphics Octane duo	2 x R10000 MIPS	225 Mhz	512 MB	IRIX 6.5
Priminfo Xeon Server	2 x 32 bits Intel Pentium 4	2.2 Ghz	2 GB	Red Hat Linux 7.3
6 nodes Priminfo Xeon	2 x 32 bits Intel Pentium 4	2.2 Ghz	2 GB	Red Hat Linux 7.3