Microbiology (2001), **147**, 2643–2649 Printed in Great Britain

FindTarget: software for subtractive genome analysis

Farid Chetouani, Philippe Glaser and Frank Kunst

Author for correspondence: Farid Chetouani. Tel: +33 1 45 68 87 48. Fax: +33 1 45 68 87 86.
e-mail: fchetou@pasteur.fr

Laboratoire Génomique des Microorganismes Pathogènes, Institut Pasteur, Département de Biologie Moléculaire, 25 rue du Dr Roux, 75724 Paris Cedex 15, France

***In silico* subtractive/differential genome analysis is a powerful approach for identifying genus- or species-specific genes, or groups of genes that are responsible for a unique phenotype. By this method, one searches for genes present in one group of bacteria and absent in another group. A software package has been developed, named FindTarget, that has a user-friendly web interface to facilitate differential genome analysis. The user chooses the genomes to compare, the similarity criteria and the thresholds to decide if a gene has a counterpart in another genome. The searches are based on BLASTP comparisons of proteomes. FindTarget also includes access to sequences, coloured multiple alignments, phylogenetic trees of conserved proteins and links to public annotated databases which provide a means for validation of the results. To illustrate this approach, a FindTarget search for genes putatively involved in the specificity of cell envelope synthesis of Gram-negative bacteria is presented. The results show that most of the identified genes are clearly involved in cell wall processes, underlining the power of such an approach in general and that of FindTarget in particular.**

Keywords: comparative genomics, bioinformatics, software, differential genome analysis

INTRODUCTION

The increasing number of complete bacterial genomes available in the public databases (for an overview see http://wit.integratedgenomics.com/GOLD/) offers new opportunities for understanding the relationship between genotype and phenotype using *in silico* genome comparisons. Differential genome analysis is an attempt to link genome content and phenotypic features according to the presence or absence of genes (Huynen *et al*., 1997). The method is based on the assumption that the genes responsible for a specific phenotype are conserved during evolution but lost in those genomes not showing that phenotype. Therefore, this method is used to search for those genes which are present in a group of genomes having a common phenotype, but which are absent in another group not showing this phenotype, as for instance the capacity to grow in the presence of an antibiotic or the ability to synthesize an outer membrane (a distinctive trait between Gram-negative and Gram-positive bacteria). This strategy may be a first step in the understanding of adaptive mechanisms of micro-organisms. For example, the comparison of the coding sequences of *Helicobacter pylori* with those of *Haemophilus influenzae* and *Escherichia coli* allowed Huynen and coworkers (1998) to detect 594 proteins specific for *Helicobacter pylori*, of which 398 had unknown functions, 123 corresponded to host interaction factors and the remaining 73 were species-specific. As the capacity to survive in the gastric environment is a specific property of *Helicobacter pylori* in comparison to *Haemophilus influenzae* and *E. coli*, the resulting list (73 proteins) contains candidate factors possibly required for survival in an acid gastric environment and thus also possible drug targets.

To date two complementary *in silico* methods have been developed allowing 'genome subtraction'. They are either based on computed clusters of homologous proteins or on pairwise protein comparisons. The first approach uses the following process to construct the protein families. First, all proteins of a sequence database, including those of complete genomes, are compared to each other with similarity search software, like BLASTP (Altschul *et al*., 1997) or FASTA (Pearson & Lipman, 1988). Then, the corresponding search outputs are processed according to default constraints to extract significant hits. Finally, the protein families are constructed using single transitive links: e.g. if proteins A and B are similar according to the constraints and

Table 1. Definition of the FindTarget input parameters and the corresponding values for the example given in the text

Parameter name	Definition	Value
Query genome	The studied genome	*Escherichia coli* K-12
Reference genome	A genome which shares similar genes with the query genome	*Campylobacter jejuni* NCTC11168 *Haemophilus influenzae* KW20 *Helicobacter pylori* 26695 *Neisseria meningitidis* Z2491
Selection criterion	The similarity criterion used to decide whether a gene from the query genome has a homologue in a reference genome	Expected $< 10^{-7}$
Match number	The minimum number of reference genome(s) containing a homologue of a query gene	4
Exclusion genome	A genome in which a query gene homologue should not be present	*Bacillus subtilis* 168 *Mycobacterium tuberculosis* H37Rv *Mycoplasma genitalium* G37
Exclusion criterion	The similarity criterion used to decide that a gene from the query genome does not contain a homologue in an exclusion genome	Expected $<10^{-6}$

proteins B and C are also similar then proteins A, B, C are stored in the same cluster. Software tools like CluSTr (Kriventseva *et al*., 2001), COG (Tatusov *et al*., 2001), HOBACGEN (Perriere *et al*., 2000), ProtoMap (Yona *et al*., 2000) or SYSTERS (Krause *et al*., 2000) provide access to such sets of homologous proteins, but only COG contains a tool, entitled 'Phylogenetic pattern search', which allows genome subtraction to select protein families. The second approach does not use fixed constraints. The user defines the similarity thresholds to decide whether a coding sequence is present or absent in a genome. The software Seebugs (Bruccoleri *et al*., 1998) belongs to this category and is based on protein sequence comparisons using the FASTA program.

To our knowledge, there are only two freely available resources providing a query engine for differential genome analysis: the reference website COG (http://www.ncbi.nlm.nih.gov/COG/) and the Seebugs software. The public database COG contains defined clusters of homologous genes for 34 of the 43 publicly available complete genomes (April 2001). If a user is interested in a specific cellular process or in genome data from a micro-organism not yet included in the COG database, the Seebugs software could be installed locally. However, as the authors of Seebugs admit in their documentation, installation is somewhat complex. Considering this situation, we have developed a software package with a user-friendly web interface for differential genome analysis which we have called FindTarget. The user chooses the similarity criteria to decide whether or not a gene has a counterpart in a set of selected genomes according to BLASTP comparisons between theoretical proteomes (predicted from the DNA sequences). Coloured multiple alignments and phylogenetic trees of conserved proteins are provided to help define relationships between gene products. For each selected gene a link to the corresponding entry in a public annotated genome database (if available) allows access to updated gene information. Any genome, even unfinished or 'private' genomes, can be added.

METHODS

Database format. The datasets managed by FindTarget are stored as text files. For each genome, the corresponding information is saved in distinct fields which contain the full species name, an abbreviated name of the organism, the sequences corresponding to the theoretical proteome and its release date, the position of each coding sequence on the chromosome (an optional field), the status of the genome sequence (complete or unfinished) and the website address of the genome database (if available). Using this format, it is also possible to include a 'virtual' proteome like a set of proteins corresponding to virulence factors.

If the database contains n genomes, $n \times (n-1)$ proteome BLASTP comparisons should theoretically be performed. To limit computation time, only the proteome versus proteome comparisons defined by the local FindTarget administrator are launched according to the research interest of the user. To reduce disk space usage, for each proteome versus proteome BLASTP comparison, only alignment properties with the best hits are saved. These are the protein name of the query and its best matching protein, the length of the query sequence and its best hit protein, the number of similar or identical amino acids for the best overlap region, the length of the best overlap region and its expected value and score, the number of amino acids found in all overlap regions for the query protein and for its best hit protein. Typically, for a comparison of two genomes encoding about 4000 proteins a parsed BLAST output file has a size of only 150 kilobytes. Several script utilities are provided for easy installation or update of the database. Due to its design, FindTarget is not restricted to BLASTP comparisons. It can also support other similarity search programs like FASTA (Pearson and Lipman, 1988) or PSI-BLAST (Altschul *et al*., 1997).

Differential genome analysis algorithm. During a FindTarget session, the user defines the input parameters to query the database. These parameters are presented in Table 1. To

Table 2. Selection/exclusion criteria definition

Criterion name	Definition
E value (database)	Number of alignments expected to have occurred by chance with a score equal to or higher than the score of the best overlap region between the query protein and its best match protein in a database (Altschul *et al*., 1997)
Score (database)	Bit score of the best overlap region between the query protein and its best match protein in a database
Percentage identity	$\frac{(\text{Number of identical aa in the best overlap region}) \times 100}{\text{Length of the best overlap region}}$
Percentage identity length min.	$\frac{(\text{Number of identical aa in the best overlap region}) \times 100}{\text{Length of the smaller protein (query or best match)}}$
Percentage positive	$\frac{(\text{Number of similar aa in the best overlap region}) \times 100}{\text{Length of the best overlap region}}$
Percentage positive length min.	$\frac{(\text{Number of similar aa in the best overlap region}) \times 100}{\text{Length of the smaller protein (query or best match)}}$
Percentage length min.	$\frac{\left[\begin{array}{l}\text{Length of the shortest protein (query or best}\\ \text{match) involved in all the overlap regions}\end{array}\right] \times 100}{\text{Length of the smaller protein (query or best match)}}$
log(*E* value) ratio*	log(*E* value) (reference genome)/log(*E* value) (exclusion genome)
Score ratio*	Score (reference genome)/Score (exclusion genome)

* The criterion is only an exclusion criterion.

increase flexibility, different selection and exclusion criteria are available (Table 2). The algorithm executed according to the chosen parameters is divided into two steps. First, the program selects all the proteins from the query genome which have a homologue in at least m reference genomes (m = match number) according to the selection criterion. A temporary list of query proteins is then generated. The next step is to reject from the temporary set all the query proteins that have a homologue in at least one exclusion genome according to the exclusion criterion. From this analysis, a final list of query proteins is retained. Typically, such an analysis takes 14 s on a Linux Pentium II 400 MHz computer with 128 megabytes of RAM.

Availability and installation procedure. The FindTarget source programs are written in Perl language (Wall *et al*., 1996), known among other advantages for its portability. Our software is currently available for Digital, Linux, Sun and Silicon Graphics Unix operating systems. The Unix-based FindTarget package, including on-line help, is available upon request. The address of the website for accessing the software is http://bioweb.pasteur.fr/seqanal/findtarget.

During software development every effort has been made to ensure installation is easy. Installation is a two-step process. First, on a Unix web server the following external software must be installed: BLAST (Altschul *et al*., 1997), BLAST 2 sequences (Tatusova & Madden, 1999), DisplayFam (Corpet *et al*., 1999), MultAlin (Corpet, 1988), Html4blast (http://bioweb.pasteur.fr/docs/softgen.html#html4blast) and Mview (Brown *et al*., 1998). Second, the configuration file of the FindTarget package has to be modified according to the local host machine (definitions of the filename directories, path to external softwares, email of the sofware administrator).

RESULTS AND DISCUSSION

Interface and functionalities

During a session, the user defines via her/his internet browser all the search parameters (Tables 1 and 2). The output lists all the selected proteins in the query genome with their best BLASTP match in each reference/exclusion genome(s) (Fig. 1). The query genes are then sorted by chromosomal position (if available) to identify potentially conserved gene clusters. For further inspection, it is possible to get a local (Tatusova & Madden, 1999) or global (Corpet, 1988) alignment between a translated query gene and its best hit in a reference/exclusion genome. For each gene name displayed, the user can also access the translated gene sequence. A link to the corresponding entry in a public annotated genome database, if available, is added to obtain detailed information about gene function (Fig. 1). Furthermore, local BLAST databases can be searched for sequence similarities.

Phylogenetic trees and coloured global multiple alignments between similar sequences, i.e. a query protein and all its best hits in the reference genome(s), are generated upon user request (Fig. 1). The alignments and the phylogenetic trees are very helpful to identify

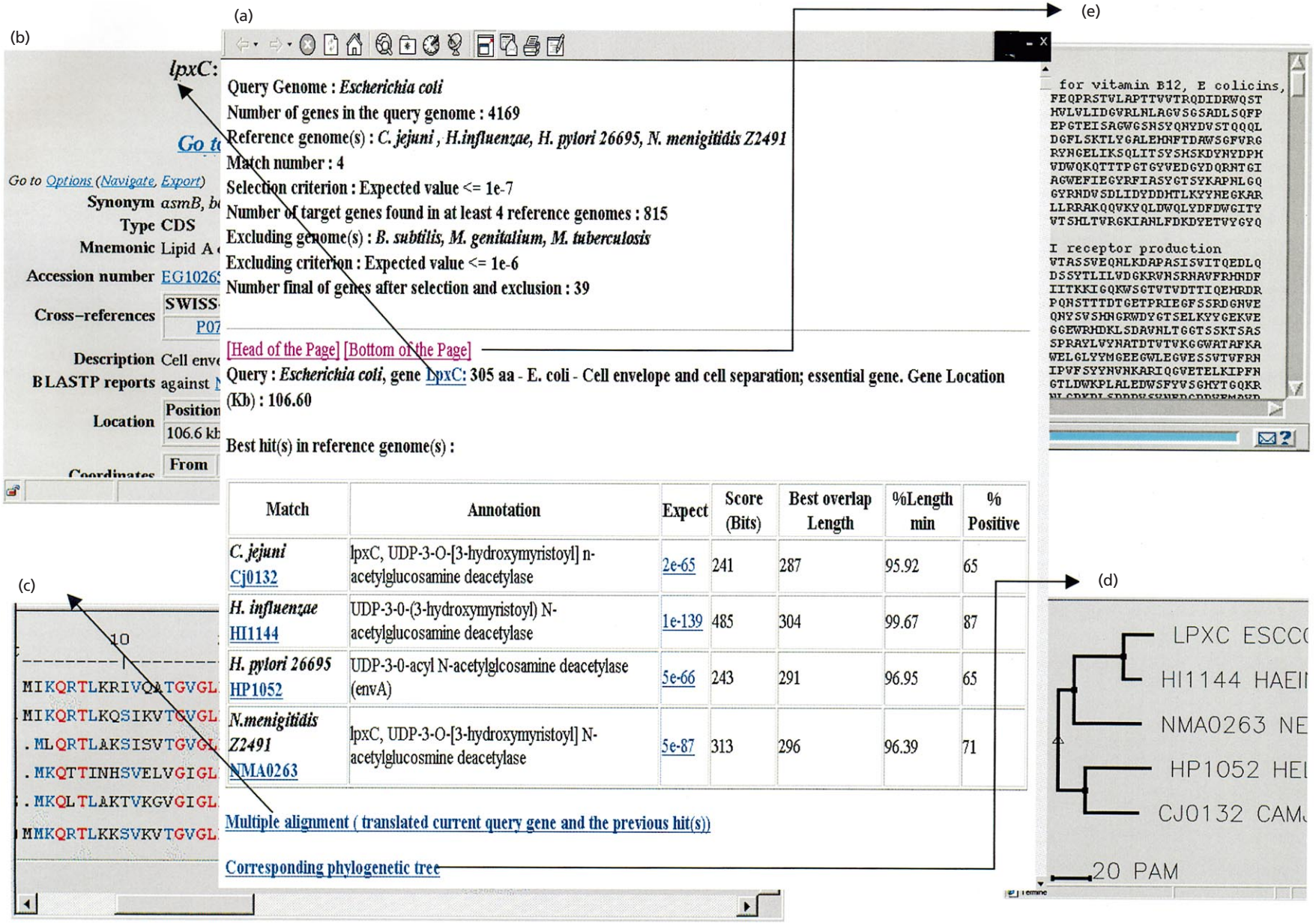

Fig. 1. FindTarget interface. According to the user parameters (see Table 1), the main output displays all the selected genes (a). For each gene, if available, a link to specialized genome databases is provided (b). By a simple click, multiple alignments and phylogenetic trees between conserved proteins are generated (c, d). Links to lists of selected gene lists and corresponding protein sequences are available (e).

Table 3. Results from a FindTarget session

The corresponding session parameters are given in Table 1. The selected *E. coli* gene names are followed by the corresponding annotation extracted from the SWISS-PROT (release 39.11) and COG (release of November 2000) databases.

Capsule formation
GutQ, YrbH: Predicted sugar phosphate aminotransferases
Lipid A biosynthesis
Ddg, HtrB: Lauroyl acyltransferase
LpxA: UDP-*N*-acetylglucosamine *O*-acyltransferase
LpxB: Lipid A disaccharide synthase
LpxC: UDP-3-*O*-acyl-*N*-acetylglucosamine acid deacetylase
Lipopolysaccharide biosynthesis
KdsB: CMP-2-keto-3-deoxyoctulosonic synthetase
KdtA: 3-Deoxy-D-manno-octulosonic-acid transferase
RfaC, RfaF: Lipopolysaccharide heptosyltransferase
Lipoprotein
RlpA: Rare lipoprotein A
Membrane phospholipid biosynthesis/turnover
DgkA: Diglyceride kinase
PgpA: Phospholipid phosphatase
Potential membrane proteins
YaeT: Probable outer-membrane protein
YhbX, YhjW, YijP, YjdB: Probable integral membrane protein
YbeQ: Probable periplasmic protein
Transport proteins and membrane receptors
BtuB: Vitamin B12 receptor precursor
CirA: Colicin I receptor precursor
ExbB, ExbD: biopolymer import/transport proteins
FucP: Fucose permease
TolQ: Uptake of group A colicins
YbhG: Multidrug resistance protein A
YcfW: Probable permease
YjdM: Uptake protein
YncD: Probable receptor
Others
Fbp: Fructose bisphosphatase
SlyD: *cis*/*trans*-Isomerase (filamentation if overexpressed)
SpeA: Arginine decarboxylase (spermidine biosynthesis)
SurE: Survival protein
YciA: Possible acyl CoA hydrolase
YdaO: Conserved hypothetical protein
YfhL: Probable ferredoxin
YgdP, YrbI: Function unknown

relationships between proteins. The global multiple alignments are important to ascertain that the aligned proteins share the same domain organization and not just a single domain. The program also permits the display of all the translated sequences or just the name of the selected genes for each genome (query/reference/exclusion) to 'copy and paste' them into a local file for further analyses (Fig. 1).

An application of FindTarget

To illustrate the utility of FindTarget, we used it to predict genes potentially implicated in Gram-negative membrane synthesis. The cell envelope is formed by the cytoplasmic membrane, the periplasm and the cell wall. In Gram-positive bacteria, the periplasm is defined as the volume directly surrounding the cytoplasmic membrane. In Gram-negative bacteria an additional outer membrane is present. In certain bacteria, a capsule (polysaccharide layer) may surround the cell envelope. The cytoplasmic membrane is a phospholipid bilayer containing membrane proteins. The cell wall consists of peptidoglycan (also called murein), a linear polysaccharide with peptide linkers. The Gram-negative outer membrane consists of a phospholipid bilayer, membrane proteins, lipoproteins and lipopolysaccharides (lipid A and O polysaccharide). The lipopolysaccharides are surface antigens anchored to the

outer membrane by a terminal lipid A core. The lipid A and O polysaccharides are unique to the outer membrane of Gram-negative bacteria (Neidhart *et al*., 1990).

The functions of the genes expected to be specific for Gram-negative bacteria are diverse. They may encode outer-membrane proteins or proteins involved in the interaction between the outer membrane and cytoplasmic membrane. They may also be involved in the synthesis and the degradation of membrane constituents. With the use of FindTarget, we searched for *E. coli* proteins having a homologue in a set of Gram-negative bacteria (*Campylobacter jejuni*, *Haemophilus influenzae*, *Helicobacter pylori* 26695, *Neisseria meningitidis* Z2491), but not in a set of Gram-positive bacteria (*Bacillus subtilis*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium* G37). The input parameters for this session are defined in Table 1. This approach allowed us to select 39 proteins from *E. coli* (see Table 3 for a complete list of selected genes). Logically, the number of gene products that were selected by this approach changes according to the stringency defined by values of the numeric input parameters (selection/exclusion criteria, match number).

Analysis of the results shows that the majority of the 39 proteins are clearly implicated in cell wall processes such as membrane components, including transporter proteins and receptors, enzymes involved in lipid A and lipopolysaccharide biosynthesis, membrane biosynthesis and capsule formation. Only a few proteins with known function do not seem to be related to the cell wall, e.g. Fbp, a fructose bisphosphatase involved in glycolysis, and SpeA, an arginine decarboxylase involved in polyamine synthesis. However, the corresponding biochemical activities are present in the Gram-positive bacterium *B. subtilis*, but the two proteins do not show sequence similarities with their counterparts in *E. coli* (Fujita *et al*., 1998; Sekowska *et al*., 1998). Therefore, FindTarget did not exclude them from the output list. Several proteins of unknown function (YhbX, YhjW, YijP, YjdB) annotated as potential membrane proteins, due to the prediction of transmembrane segments, were also revealed. It will be a challenging task to determine whether these latter proteins are also involved in cell wall processes.

Conclusion

FindTarget is an easy to use and powerful tool for identifying potentially specific genes for one or several species as determined using the similarity criteria selected by the user. The Unix package is available upon request and can be readily installed on cheap Linux personal computers which are now becoming common in molecular biology laboratories.

However, it is important to remember that in some organisms identical biochemical reactions may be catalysed by non-related enzymes. This is non-orthologous gene displacement (Koonin *et al*., 1996) and so the FindTarget user has to be careful with the interpretation of the results as absence of a protein in a genome does not necessarily mean that the corresponding function is missing.

FindTarget includes practical functionalities to analyse the results, such as generation of multiple alignments, reconstruction of phylogenetic trees, similarity searches in local databases and optional links to public databases of annotated genomes. The list of the genes selected during a work session and their coding sequences can be displayed and saved for further analysis. FindTarget quickly produces result outputs. Therefore, it allows successive requests to test several combinations of parameters and to define the most appropriate ones. Finally, results of this *in silico* comparison could be combined with other whole-genome analyses such as transcriptome and two-dimensional gel electrophoresis of proteins. Indeed transcriptome studies often lead to the identification of numerous genes, which cannot be all analysed in depth. A combination of these results with a FindTarget analysis could provide arguments for the selection of genes for further functional analysis. With the growing number of publicly available complete genomes, software tools like FindTarget should provide a rational basis for experimental design in the rapidly expanding field of functional genomics.

ACKNOWLEDGEMENTS

We are grateful to Carmen Buchrieser, Lionel Frangeul for their constant support and helpful discussions. We thank Jerome Gouzy and Daniel Kahn for providing access to the DisplayFam source, and Marc Baudoin and Nicolas Joly for assistance and for valuable suggestions. We acknowledge Roland Brosch, Ivan Moszer, Jean-Marc Reyrat and Timothy Stinear for critical reading of the manuscript.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Brown, N. P., Leroy, C. & Sander, C. (1998). MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics* **14**, 380–381.

Bruccoleri, R. E., Dougherty, T. J. & Davison, D. B. (1998). Concordance analysis of microbial genomes. *Nucleic Acids Res* **26**, 4482–4486.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16**, 10881–10890.

Corpet, F., Gouzy, J. & Kahn, D. (1999). Browsing protein families via the 'Rich Family Description' format. *Bioinformatics* **15**, 1020–1027.

Fujita, Y., Yoshida, K., Miwa, Y., Yanai, N., Nagakawa, E. & Kasahara, Y. (1998). Identification and expression of the *Bacillus subtilis* fructose-1,6-bisphosphatase gene (*fbp*). *J Bacteriol* **180**, 4309–4313.

Huynen, M., Dandekar, T. & Bork, P. (1998). Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett* **426**, 1–5.

Huynen, M. A., Diaz-Lazcoz, Y. & Bork, P. (1997). Differential genome display. *Trends Genet* **13**, 389–390.

Koonin, E. V., Mushegian, A. R. & Bork, P. (1996). Non-orthologous gene displacement. *Trends Genet* **12**, 334–336.

Krause, A., Stoye, J. & Vingron, M. (2000). The SYSTERS protein sequence cluster set. *Nucleic Acids Res* **28**, 270–272.

Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M. & Apweiler, R. (2001). CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res* **29**, 33–36.

Neidhart, F. C., Ingraham, J. L. & Schaechter, M. (1990). *Physiology of the Bacterial Cell*. Sunderland, MA: Sinauer Associates.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**, 2444–2448.

Perriere, G., Duret, L. & Gouy, M. (2000). HOBACGEN: database system for comparative genomics in bacteria. *Genome Res* **10**, 379–385.

Sekowska, A., Bertin, P. & Danchin, A. (1998). Characterization of polyamine synthesis pathway in *Bacillus subtilis* 168. *Mol Microbiol* **29**, 851–858.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V. & 7 other authors (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22–28.

Tatusova, T. A. & Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174**, 247–250.

Wall, L., Christiansen, T. & Schwartz, R. L. (1996). *Programming Perl*, 2nd edition. Sebastopol: O'Reilly & Associates.

Yona, G., Linial, N. & Linial, M. (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res* **28**, 49–55.

Received 12 April 2001; revised 5 July 2001; accepted 13 July 2001.