W470–W474 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 17 May 2008*
doi:10.1093/nar/gkn277

Signature, a web server for taxonomic characterization of sequence samples using signature genes

Bas E. Dutilh*, Ying He, Maarten L. Hekkelman and Martijn A. Huynen

Center for Molecular and Biomolecular Informatics/Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Geert Grooteplein 28, 6525 GA, Nijmegen, The Netherlands

Received January 30, 2008; Revised April 14, 2008; Accepted April 26, 2008

ABSTRACT

Signature genes are genes that are unique to a taxonomic clade and are common within it. They contain a wealth of information about clade-specific processes and hold a strong evolutionary signal that can be used to phylogenetically characterize a set of sequences, such as a metagenomics sample. As signature genes are based on gene content, they provide a means to assess the taxonomic origin of a sequence sample that is complementary to sequence-based analyses. Here, we introduce Signature (http://www.cmbi.ru.nl/signature), a web server that identifies the signature genes in a set of query sequences, and therewith phylogenetically characterizes it. The server produces a list of taxonomic clades that share signature genes with the set of query sequences, along with an insightful image of the tree of life, in which the clades are color coded based on the number of signature genes present. This allows the user to quickly see from which part(s) of the taxonomy the query sequences likely originate.

INTRODUCTION

Signature genes are genes that are unique to a certain taxonomic clade, and are common throughout all its major subclades. The plausible evolutionary explanation for such a phylogenetic distribution is that the gene was invented in the ancestral lineage, and constituted an important innovation that possibly allowed the clade to radiate. The significance of signature genes in biological research is both functional and taxonomic (1). Signature genes contain a wealth of functional signals that are related to clade-specific processes, many of which remain to be described. Furthermore, they hold a strong evolutionary signal that can be used to phylogenetically characterize a set of sequences, such as a metagenomics sample. The recent breakthroughs in nucleotide sequencing (2) disclose a wealth of environmental niches, each harboring an abundance of undiscovered (microbial) life. As the metagenomic sequencing of such niches accelerates (3–5), one of the first challenges lies in the taxonomic characterization of the contributing organisms. Traditionally, sequence analyses are the method of choice in taxonomy, and the first approaches to characterize metagenomic data on the basis of sequence have recently been seeing the light (6–9). An alternative way to approach taxonomic questions is on the basis of signature genes. Signature genes are based on gene content, an intermediate between genotype and phenotype (10,11) that gives a view on phylogeny that is complementary to sequence-based analyses (12). For example, using a prototype of our new signature gene method (13), we showed that the anaerobic ammonium oxidizing bacterium *Kuenenia stuttgartiensis* is closely related to the Chlamydiae, a finding that independently supported the traditional phylogenomic analysis based on a superalignment of 49 proteins.

In the current article, we introduce Signature, an interactive web server that allows a user to identify signature genes in a set of query sequences. The server first assigns each sequence to one of the orthologous groups from the STRING 7.0 database (14) using a sequence similarity search. The phylogenetic distribution of this orthologous group throughout the tree of life is then assessed to determine if the sequence is a signature gene. The server produces a list of taxonomic clades that share signature genes with the set of query sequences, a coverage score for each signature in that clade, and an insightful image of the tree of life, in which the clades are color coded based on the number of signature genes present. This allows the user to quickly see from which part(s) of the taxonomy the query sequences likely originate. Using 1-fold and *k*-fold

*To whom correspondence should be addressed. Tel: +31 024 3619797; Fax: +31 024 3619395; Email: dutilh@cmbi.ru.nl
Present address:
Ying He, VIB Department of Plant Systems Biology, Ghent University, Technologiepark 927, 9052 Gent, Belgium

cross-validation analyses we confirmed that ~93% of all signature genes were re-assigned to their correct clade.

The tool uses MRS (15), a fast data retrieval system that also allows for BLAST searches. Signature currently assigns 2000 unknown protein sequences to orthologous groups in about 10 h, and in 10 s if the sequences occur in the database.

DATA

Signature makes use of two types of information. The first is a protein sequence database containing the proteomes of 373 completely sequenced organisms, in which every protein may or may not be assigned to an orthologous group of proteins (OG). The database presently incorporated in Signature is STRING version 7.0 that contains 43 577 COGs, KOGs and NOGs (14). The second type of information used by Signature is a reference tree of life that defines the evolutionary relationships between the species. It is possible to upload a custom tree of life. As a default option, Signature provides either the tree of life used in STRING 7.0 (consistent with the sequence database), or a recently published phylogenomic tree based on a superalignment of 31 universal protein families (16). If the selected tree of life includes bootstrap values, the user can choose to set a bootstrap threshold, so that only the reliable nodes will be considered and the nodes with lower bootstrap value will be collapsed to form a multifurcating branch.

A TYPICAL SIGNATURE SEARCH

Assigning query sequences to OGs

Typically, a user will have a set of sequences obtained from an unknown organism or environment and wants to know to which taxonomic clade(s) the sample belongs. Alternatively, the user will want to know whether a protein sequence has a signature status in any clade, and may thus represent a process particular for that clade. With either of these questions, the user will go to www.cmbi.ru.nl/signature, enter the amino acid sequences in the input field (in FASTA format) and click 'go'. Signature stores the input sequences and the selected options for the BLAST (17) search and the OG assignment, and then starts by assigning each input sequence to an orthologous group. If, by MD5 checksum and a subsequent sequence identity check, the query sequence is identical to one of the proteins already in the database, it is directly assigned to the corresponding OG(s). Otherwise, the sequence is assigned according to a Cognitor-like rule (18) if the majority of the top N BLAST hits belong to the same OG (this option can be adjusted). Note that a query sequence will be assigned to multiple OGs if they occur as non-overlapping regions (the maximum overlap between two regions can be adjusted, see Figure 1). The BLAST searches can be time-limiting, depending on the number and length of the query sequences. While running, Signature displays an estimate of the duration until the search is finished, based on the running times of searches with a comparable size that were carried out previously. It is always safe to refresh the page for the most up-to-date results. Signature currently identifies the orthologous groups in 2000 unknown protein sequences within 10 h (depending on the length of the sequences), and in 10 s if the sequences occur in the database (by checksum). As the sequence similarity searches are the time-limiting step, we are presently extending the tool with a much faster OG assignment algorithm, making it feasible to analyse metagenomics scale data sets and DNA data sets in six translated frames in a limited time.

Identifying signature OGs in the tree of life

Once the query sequences have been automatically assigned to the appropriate OGs, Signature permits the user to review these assignments and manually change them if desired. It is also possible to include additional OGs for the Signature search in the second step. If the user prefers to rely on a personal OG assignment tool, the first step may be bypassed by leaving the initial input field empty (see previous paragraph), and entering a list of COGs, KOGs or NOGs here. Upon clicking 'go', Signature will assess for each OG whether it is a signature for any clade in the selected tree of life.

A signature OG for a certain clade is defined as an OG that occurs in every daughter lineage of that clade, but nowhere outside it [Figure 2 (1)]. To minimize the number of false negatives, the basic signature definition does not require a signature OG to be present in every species in the clade. We have developed a nested coverage score that indicates the level of species coverage of a signature OG, taking into account potential asymmetrical taxon

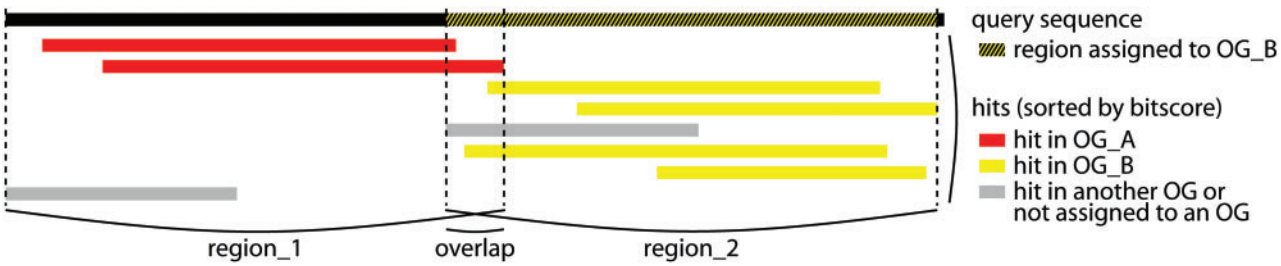

Figure 1. Assignment of a query sequence to OGs using the default OG assignment rule (3 of first 5). Disparate regions are first parsed from all the significant BLAST hits (max 30) if they overlap less than 35 amino acids (all these parameters can be adjusted). Then, each region is assigned to an OG: region_1 is not assigned as only two of the first five BLAST hits belong to the same OG, while region_2 is assigned, as at least three of the first five BLAST hits (in that region) belong to OG_B.

sampling [see the Figure 2 caption for an example (1)]. Using the coverage score cutoff, the user can direct Signature to report only high-scoring signature OGs.

The first result of a Signature search is a list of all the signature OGs contained in the set of query sequences, along with their coverage score. For each taxonomic clade with signature OGs, Signature calculates a log-likelihood significance that is based on the expected number of signature OGs for that clade (see 'Significance score' section below). Furthermore, Signature generates an insightful image of the selected tree of life, where the internal clade branches are color coded based on their number of signature OGs (shades of green), and the species, the terminal leaves in the tree, are color coded based on the cumulative number of signature OGs contained in all the ancestor clades they belong to (shades of red). Thus, the user can easily assess which species are most closely related to the source of the query sequences (Figure 3). It should be noted that Signature can also handle situations where the sequences are derived from several species. Different regions of the tree will then be highlighted.

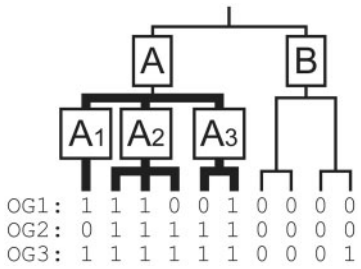

Figure 2. Definition of signature OGs in a (partially unresolved) tree of life. For every species, presence (1) or absence (0) of three OGs is indicated. In this example, only OG1 is a signature for clade A, as it is present in the daughter lineages A1, A2 and A3, but not in clade B. Although OG2 and OG3 are present in more species within clade A, they are not a signature for clade A because OG2 is not present in the daughter clade A1 and OG3 is present outside of clade A. The coverage score of OG1 as a signature for clade A is recursively calculated as the average of the scores in the daughter clades: [(1/1) + (2/3) + (1/2)]/3 = 0.72.

SIGNIFICANCE SCORE

To assess the significance of finding a set of signature genes for a certain clade within a sample, Signature provides a significance score for each clade on the output page. This score is based on the number of expected signature OGs if all OGs were randomly distributed across the genomes. To do this, we composed 1000 randomized sets of genomes, taking care not to place the same OG twice in one genome (1), and calculated the expected number of signatures for each clade. In these randomizations, the number of signature OGs that we find for a clade depends on the number of daughter clades, the genome sizes of the species therein and on the genome sizes of the species outside the clade. To avoid having to compute the expected number of signature OGs by such randomizations at every Signature search, we calculate the expected number of signature genes (e) analytically. We start by calculating the expected the number of OGs present at least once in every daughter lineage, but not in any species outside the clade (Equation 1).

$$e = \prod_{d=1}^{n} \left(\frac{\mathrm{OGs}_d}{\mathrm{OGs}_{\mathrm{total}}} \right) \times \left(1 - \frac{\mathrm{OGs}_{\mathrm{out}}}{\mathrm{OGs}_{\mathrm{total}}} \right) \times \mathrm{OGs}_{\mathrm{total}} \qquad (1)$$

The product runs over all n daughter lineages d of the clade. Note that $n = 2$ in a completely resolved phylogeny, $n \geq 2$ in a tree where nodes with low bootstrap support have been collapsed to a multifurcating branch. OGs_d is the total number of different OGs in all species in daughter d (union), $\mathrm{OGs}_{\mathrm{out}}$ is the number of different OGs outside the clade, and $\mathrm{OGs}_{\mathrm{total}}$ is the number of different OGs in the whole tree.

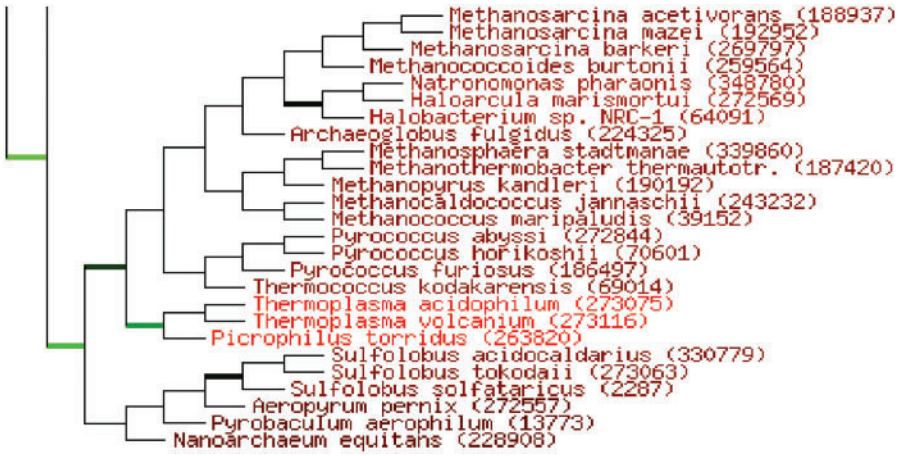

Figure 3. Image of the tree of life that results from a Signature search (cropped to the Archaeal clade). The internal clade branches are color coded based on the number of signature OGs they share with the query (shades of green), the terminal species are color coded based on the sum of the signature OGs of all their parent clades (shades of red). Note that all species in one clade are colored in the same shade of red. This means that an individual species does not necessarily contain all, or even any, of the signature genes assigned to its parental clades, especially if many signatures were found with a low coverage score. Sequences analysed are the *Ferroplasma acidarmanus* type II scaffolds obtained from DNA sequencing of acid mine drainage [see text (4)].

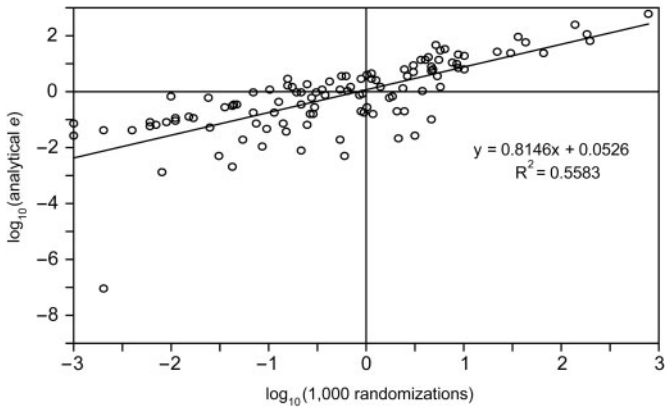

Figure 4. Correlation between analytically calculated e (Equation 1) and the expected number of signature OGs based on 1000 randomized sets of genomes. Equation $y = 0.8146x + 0.0526$ is the formula for the linear regression line plotted in the figure. The results are based on the prokaryotic clades in the Ciccarelli *et al.* (16) tree of life as provided by Signature, where clades with bootstrap values <80% were collapsed (1). The outlier at (−2.70; −7.06) are the Bacteria, a clade with very low bootstrap support for its earliest branches, leading to 13 daughter lineages and an underestimation of e. For the rest of the clades in this plot, we did not find a correlation between the number of daughter lineages and e (data not shown).

Indeed, there is a good correlation between the analytically predicted e and the number of expected signature genes based on randomization (Figure 4). As can be observed in the regression equation, the randomized expected value (x) increases faster than the analytical e (y), due to other factors such as variations in genome sizes. Thus, taking this correction factor of 0.8146 into account, we can calculate the expected number of signatures from the numbers of OGs present in the daughter lineages as $10^{(\log(e)-0.0526)/0.8146}$. This value is based on the total number of different OGs in the whole tree, and can be scaled to the number of query sequences by multiplying it with queries/OGs_{total}. From this expected number of signature OGs, we calculate the observed/expected ratio and present the significance in the Signature search output page as the log odds of drawing the observed number of signature OGs (observed) from the total set of OGs in the tree, given the number of query sequences (queries, Equation 2).

$$\text{significance} = \log_{10}\left(\frac{\text{observed} \times \text{OGs}_{\text{total}}}{\text{queries} \times 10^{(\log_{10}(e)-0.0526)/0.8146}}\right) \quad (2)$$

PERFORMANCE

We assessed the performance of Signature by entering a set of sequences from the first metagenomic sample to be published, obtained from random shotgun sequencing of DNA from a natural acidophilic biofilm (4). From this sample, we used the *Ferroplasma acidarmanus* type II scaffolds. This data set contained 1956 sequences with an average length of 267 amino acids. The sequence-similarity search and OG assignment for these proteins took 9 h and 43 min with default parameters, detecting 974 unique OGs (1531 total) in the data set. Within 30 s, 196 of these 974 OGs were identified to be a signature for a clade in the default tree of life, highlighting Thermoplasmatales as the closest relatives of the species in the metagenomic sample (Figure 3).

DISCUSSION

Signature is a web server that identifies signature OGs within a set of query sequences for any clade in a selected or uploaded tree of life. A signature OG is an orthologous group that is unique for a clade and is retained in all its daughter lineages (Figure 2). Several public databases allow users to download strain-, species- or genus-specific genes, or sets of genes or proteins shared by a chosen set of species (19–21). Signature is unique in that it identifies the signatures in a set of query sequences, for any clade in an adjustable tree of life and places the query in a taxonomic context.

Signature OGs can provide important new functional insights, as they carry out functions that may characterize the clade. They are also interesting from a taxonomic point of view. For example, if the sequences in a sample contain many signature genes for the Thermoplasmatales and its subclades, then the sample is likely taken from a species in that clade. Based on this idea, Signature places the signature OGs it discovers in a taxonomic context and produces an insightful image of the tree of life where the branches and leaves are color coded based on the number of signature genes associated to them (Figure 3). Thus, the user can easily assess from which clade the sample was likely taken.

To test the how well the Signature method assigns species to the correct clade, we did cross-validation analyses (Signature provides the option of leaving species out of the analysis), in which we removed every species from the data set one by one, and identified the signatures among the OGs in the removed genome in a phylogeny that did not contain that genome (1). Each of the removed proteins was either a signature OG for one of the ancestral nodes of the removed species (true positive, *tp*), a signature for another node (false positive, *fp*) or not a signature. In that case, the OG could have been a signature in the situation where no species were excluded (false negative, *fn*) or not (true negative, *tn*). Using these values, we computed sensitivity [$tp/(tp + fn)$; 77.6%], specificity [$tn/(tn + fp)$; 98.8%], precision [$tp/(tp + fp)$; 93.1%] and accuracy (true/all; 95.6%) of the method. Removing 10%, 20% or 30% of the species (k-fold cross-validation) did not change these numbers much.

When interpreting the results of Signature, it is important to realize the issue of horizontal gene transfers [HGTs (22)]. If a signature gene for a certain clade gets transferred to a distantly related organism, it loses its initial signature status, because it will be found outside its taxon in the new situation. An example of such a case might be OG3 in Figure 2, which could have been transferred from a species in clade A to the rightmost species in clade B. In this case, after the HGT, OG3 became a signature for the clade AB, as it is present in both its daughter lineages.

A solution for this is provided by the coverage score threshold. Whereas OG3 would have had a coverage score of $[(1/1) + (3/3) + (2/2)]/3 = 1.00$ in clade A before the HGT event, it now has a coverage score of $\langle\{[(1/1) + (3/3) + (2/2)]/3\} + \{[(0/2) + (1/2)]/2\}\rangle/2 = 0.63$ in clade AB. Thus, by using a (high) coverage score threshold, which is provided as an option in the Signature search input form, the user can filter out signature OGs with a low coverage score that might be the result of HGT. Nevertheless, even without using a coverage score threshold, we find only ~1% false positives and a precision of ~93% in the *k*-fold cross validation analyses (above).

Newly sequenced genomes might of course affect which genes are regarded as HGTs, potentially leading to the loss or change of signatures. If HGT would be widespread, genome sequencing could reach a point where all genes have been transferred outside of the taxon for which they were previously a signature. To assess the severity of this problem, we tested the robustness of the set of signature genes against the addition of new genomes by randomly leaving out 10%, 20% or 30% of the species in the tree, and identifying the overlap between the signature OGs in the original set of species and in the subsampled tree. This mimics the situation where up to 30% of the species are newly added to the current tree. The restricted species sets contained only very few 'new' signature genes: 2.0%, 5.1% and 5.9%, respectively [average of 100 samples (1)]. Based on the results from these cross-validation experiments, we do not expect to lose many of the current signature genes in an extended species set, and the taxonomic characterization of a query will be consistent. We will continue to add new genomes to the database as they become available.

Summarizing, Signature is a user-friendly web server that facilitates the identification of signature OGs within a set of query sequences. Many parameters of the underlying method can be adjusted, providing a comprehensive platform for analyses related to clade-specific genes and processes, both in a functional and in a taxonomic context.

ACKNOWLEDGEMENTS

This work was supported by the European Union's 6th Framework Program, contract number LSHB-CT-2005-019067 (EPISTEM) and contract number LSHG-CT-2003-503265 (BioSapiens), and by the Dutch Science Foundation (NWO) Horizon Project 050-71-058. Funding to pay the Open Access publication charges for this article was provided by the European Union's 6th Framework Program, contract number LSHB-CT-2005-019067 (EPISTEM).

Conflict of interest statement. None declared.

REFERENCES

1. Dutilh,B.E., Snel,B., Ettema,T.J.G. and Huynen,M.A. (in press) Signature genes as a phylogenomic tool. *Mol. Biol. Evol.*
2. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
3. Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J., Detter,J.C. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.
4. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
5. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
6. Abe,T., Sugawara,H., Kinouchi,M., Kanaya,S. and Ikemura,T. (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res.*, **12**, 281–290.
7. Huson,D.H., Auch,A.F., Qi,J. and Schuster,S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.
8. McHardy,A.C., Martin,H.G., Tsirigos,A., Hugenholtz,P. and Rigoutsos,I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72.
9. von Mering,C., Hugenholtz,P., Raes,J., Tringe,S.G., Doerks,T., Jensen,L.J., Ward,N. and Bork,P. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**, 1126–1130.
10. Dutilh,B.E., van Noort,V., van der Heijden,R.T., Boekhout,T., Snel,B. and Huynen,M.A. (2007) Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics*, **23**, 815–824.
11. Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110.
12. Snel,B., Huynen,M.A. and Dutilh,B.E. (2005) Genome trees and the nature of genome evolution. *Annu Rev. Microbiol.*, **59**, 191–209.
13. Strous,M., Pelletier,E., Mangenot,S., Rattei,T., Lehner,A., Taylor,M.W., Horn,M., Daims,H., Bartol-Mavel,D., Wincker,P. *et al.* (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*, **440**, 790–794.
14. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
15. Hekkelman,M.L. and Vriend,G. (2005) MRS: a fast and compact retrieval system for biological data. *Nucleic Acids Res.*, **33**, W766–W769.
16. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
18. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
19. Li,J.B., Zhang,M., Dutcher,S.K. and Stormo,G.D. (2005) Procom: a web-based tool to compare multiple eukaryotic proteomes. *Bioinformatics*, **21**, 1693–1694.
20. Mazumder,R., Natale,D.A., Murthy,S., Thiagarajan,R. and Wu,C.H. (2005) Computational identification of strain-, species- and genus-specific proteins. *BMC Bioinform.*, **6**, 279.
21. Siew,N., Azaria,Y. and Fischer,D. (2004) The ORFanage: an ORFan database. *Nucleic Acids Res.*, **32**, D281–D283.
22. Doolittle,W.F. (1999) Lateral gene transfer, genome surveys, and the phylogeny of Prokaryotes. *Science*, **286**, 1443a.