W518–W523 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 26 May 2011*
doi:10.1093/nar/gkr388

CoMet—a web server for comparative functional profiling of metagenomes

Thomas Lingner[1,*], Kathrin Petra Aßhauer[1], Fabian Schreiber[1,2] and Peter Meinicke[1,*]

[1]Department of Bioinformatics, Institute for Microbiology and Genetics, Georg-August University Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany and [2]Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden

Received February 11, 2011; Revised April 1, 2011; Accepted May 3, 2011

ABSTRACT

Analyzing the functional potential of newly sequenced genomes and metagenomes has become a common task in biomedical and biological research. With the advent of high-throughput sequencing technologies comparative metagenomics opens the way to elucidate the genetically determined similarities and differences of complex microbial communities. We developed the web server 'CoMet' (http://comet.gobics.de), which provides an easy-to-use comparative metagenomics platform that is well-suitable for the analysis of large collections of metagenomic short read data. CoMet combines the ORF finding and subsequent assignment of protein sequences to Pfam domain families with a comparative statistical analysis. Besides comprehensive tabular data files, the CoMet server also provides visually interpretable output in terms of hierarchical clustering and multi-dimensional scaling plots and thus allows a quick overview of a given set of metagenomic samples.

INTRODUCTION

Metagenomics fundamentally changes our view of the microbial world. Instead of isolating a few culturable organisms for a genome-based characterization of single species, the investigation of large amounts of mixed DNA from environmental samples provides a more holistic picture of microbial communities. Furthermore, the development of next-generation sequencing technologies rapidly increases the sequencing coverage and therefore even allows the profiling of highly diverse communities as for instance considered in soil metagenomics (1).

While the taxonomic profiling of metagenomes yields an estimate of the phylogenetic distribution, functional profiling tries to characterize the metabolic potential of a community. In analogy to comparative genomics which aims at the identification of organism specific properties, the comparison of the functional inventory of different metagenomes is crucial for an understanding of community specific properties which are possibly linked with particular environmental factors. Examples are the comparison of communities from different types of environment (2) or from the human gut according to healthy and diseased states (3). In this context the basis for a sequence-based comparison is the analysis of assignments to functional categories. A pipeline for functional profiling of metagenomes typically involves the following steps: first, all open reading frames (ORFs) have to be identified in the DNA sequences. Often ORF finding includes the discrimination of protein coding ORFs from non-coding ones. Subsequently, the ORFs are matched against a comprehensive database of functionally labeled protein sequences or models. In this step a computationally expensive similarity search has to be performed to achieve the final assignments. Common sets of labeled sequences include clusters of orthologous genes [COGs, (4)], the FIGfam protein families (5) and the Pfam domain families (6). Finally, the abundances of functional categories are used for a statistical comparison to distinguish systematic differences from random variation.

Several pipelines have been introduced for functional profiling of metagenomes. Among the web-based platforms, the MG-RAST server (7) provides the most comprehensive analysis of user-supplied data. Under a personal account different sequence data files can be analyzed in terms of BLAST-based assignments to FIGfam (5) protein families. Besides the taxonomic profiles also functional profiling in terms of SEED categories and a KEGG (8) pathway mapping can be used to identify differences among the user samples and to compare the results with pre-computed profiles from public metagenome data. The RAMMCAP web tool (9) which is accessible via the CAMERA portal (10) offers

*To whom correspondence should be addressed. Tel: +49 551 39-13994; Fax: +49 551 39-14929; Email: thomas@gobics.de
Correspondence may also be addressed to Peter Meinicke. Tel: +49 551 39 14925; Fax: +49 551 39 14929; Email: pmeinic@gwdg.de

a wide range of functional profiling methods for analysis of single metagenomic sequence files. However, the comparison of different samples has to be performed offline using for example a local RAMMCAP installation. The focus of WebCARMA (11) is on taxonomic profiling of metagenomes. A functional annotation of the sequences in terms of Pfam and Gene Ontology [GO, (12)] assignments can be obtained to perform an offline comparison of different samples. In addition, several web-based resources for a comparative metagenome analysis exist that do not provide the profiling of large volumes of user-supplied sequence data (13–15). In this case the sequence assignments have to be provided by the user or pre-computed assignments for public metagenomes can be used to perform a comparative analysis.

We here present the CoMet server for fast web-based comparison of metagenomes in terms of their functional profiles. CoMet implements a complete pipeline for comparative functional profiling of multiple sequence data files. The estimation of functional profiles from user-supplied sequence files is based on a computationally efficient assignment of the sequences to Pfam domain families. The resulting domain frequency profiles are then used for a comparative statistical analysis. The CoMet server provides an easy-to-use interface for data submission and interpretable output in terms of figures and downloadable tabular data.

METHODS AND IMPLEMENTATION

Construction of domain frequency profiles

As a first step in the CoMet analysis process, the individual sequence samples are analyzed with respect to significant hits to protein domain families according to version 24.0 of the Pfam database (6). The assignment to Pfam domain families is based on speed-optimized implementations of the Orphelia gene prediction engine (16) and the UFO domain detection approach (17). In the first stage, the Orphelia engine identifies ORFs with a minimum length of 20 amino acids within the sequencing reads [for details see (16)]. In a second stage, the UFO method detects significant Pfam domain hits within these ORFs using an inexact matching of long words [for details see (18)]. The domain detection engine was trained using all unambiguous words extracted from the whole set of domain families contained in the Pfam A 'full' section (17), which in the case of Pfam 24.0 gave rise to $\sim 1.05 \times 10^7$ training sequences. For speed optimization we slightly reduced the UFO word length from 20 to 18 amino acids.

Statistical analysis for comparison of metagenomes

The comparative statistical analysis within CoMet provides a basis for identification of differently abundant Pfam domain families and the associated GO terms in a given set of metagenomic data files. The resulting differences are also used to perform multi-dimensional scaling (MDS) and a hierarchical clustering analysis of the data.

The comparative analysis is based on pairwise tests on the Pfam domain frequency profiles of the metagenomic samples. Here, significantly different families are identified by comparing the domain specific counts in the two samples assuming a binomial distribution model (see Methods in Supplementary Data). Similar tests have been used in (9,19) to identify significantly different domain families and COG clusters in a pair of samples. According to a problem-specific significance level in terms of a P-value threshold (e.g. $P = 0.05$), the tests yield a number of significantly different Pfam domains for each sample pair.

Hierarchical clustering analysis and MDS

The number of significantly different domain families for all sample pairs in a dataset gives rise to a dissimilarity matrix **D**, which can be used for a distance-based analysis. We use **D** to cluster the samples hierarchically according to their ratio of differing to non-differing domains using UPGMA. Further, the distance matrix is also used to perform an MDS (20) analysis to provide an overview of the differences between several metagenomic samples. MDS projects the samples onto a low-dimensional space such that similar samples are close to each other.

WEB SERVER INTERFACE

The CoMet web server (http://comet.gobics.de) provides an easy-to-use interface for upload of sequence data, a result page containing graphical and downloadable output of the various statistical analyses and a detailed help page including example output. In the following, we will describe these elements of the CoMet web server in more detail.

Submission of datasets and jobs

CoMet accepts as input a collection of at most 20 metagenomic sequence files, each of which can contain several million DNA reads of varying length in multi-FASTA format. Because the minimum length of an ORF detected by the Orphelia method is 20 amino acids, the sequence read length should be clearly above 60 bp. Furthermore, the reads should originate from high-quality sequencing methods since frame shifts in the data cannot be detected by Orphelia.

The submission page of the CoMet server allows the upload of metagenomic datasets of up to 500 MB per file, whereby the files maybe compressed in the 'ZIP' format to enable faster upload and the analysis of larger datasets. On the same page the configuration of an analytical task ('job') can be carried out with a few simple steps. More specifically, a P-value for statistical analysis can be selected and an email address for result notification may optionally be specified. For users who want to initially explore the capabilities of the CoMet server, a checkbox allows to use example data for the subsequent analysis. In this case, no dataset has to be uploaded.

Result pages

The output of the CoMet server arises from the statistical analysis associated with a particular job (see 'Methods and Implementation' section). A job-specific CoMet

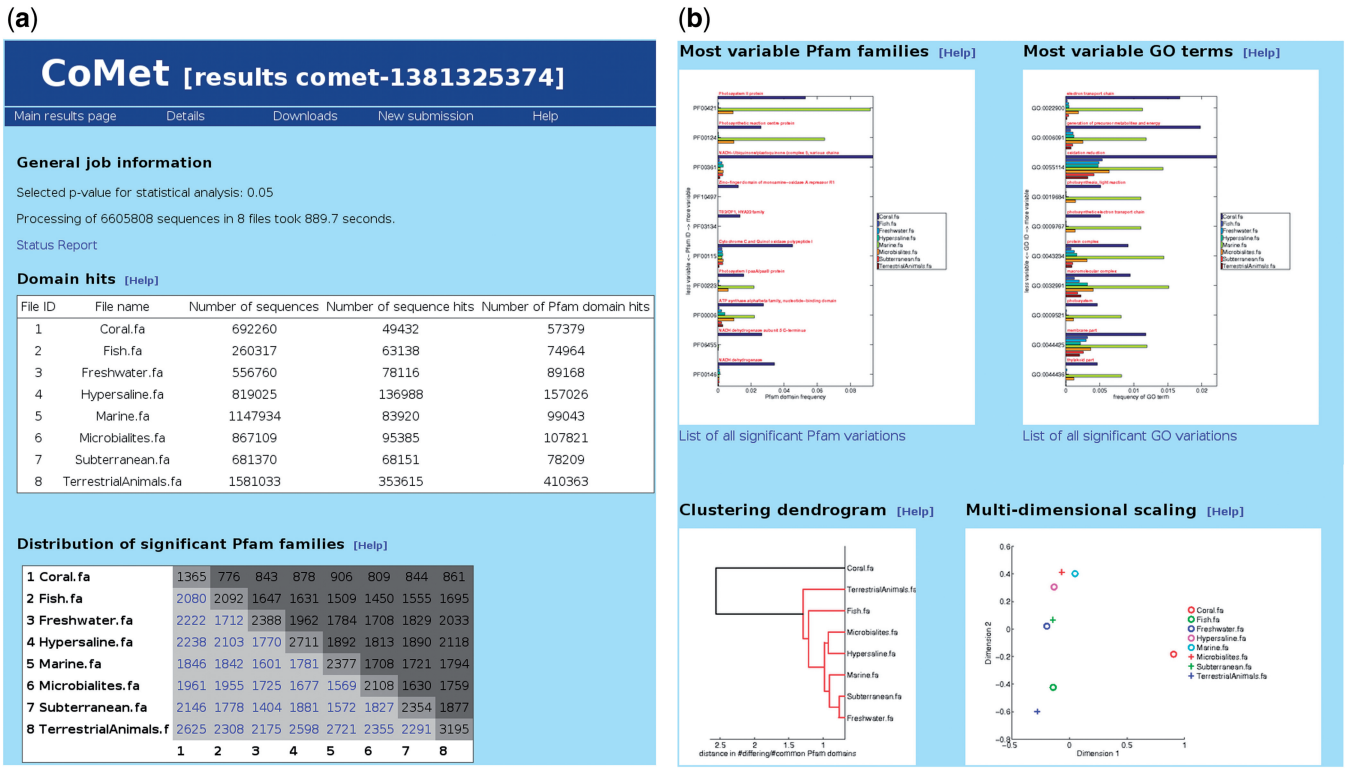

Figure 1. Screenshot of the tabular (**a**) and graphical (**b**) section of the CoMet result page for a metagenomic dataset comprising eight microbial biomes (21).

result page contains different elements, depending on the number of files that have been uploaded for analysis.

Statistics. On top of the main result page some basic information about the job configuration and the Pfam domain detection is displayed (see Figure 1a). A table shows the files that passed the CoMet analysis successfully along with their respective number of sequences, the number of sequences with domain hits and the total number of significant Pfam domains that CoMet detected. In addition, the computation time for the analysis is shown.

If more than one file has been uploaded, the distribution of significant Pfam domain families is displayed in a matrix-like table below the basic summary section. This matrix contains the number of significantly frequent domains for all input files as diagonal elements (see 'Methods and Implementation' section). Furthermore, the entries in the lower triangle of the matrix represent the number of significantly different Pfam domain families for all pairs of input files. The underlying hyperlinks direct to a result page that lists the corresponding domain families along with their Pfam description, their predicted frequency within the samples and the resulting significance *P*-value of the difference. Finally, the upper triangle shows the number of common domain families of two files, i.e. the number of significant families that occur in both samples while not being significantly different in abundance.

Figures and dynamic output

Variation of Pfam and GO terms. If more than one file is uploaded, CoMet determines the variation of each Pfam domain family in terms of its *P*-value distribution across all sample pairs. A bar chart then shows the domain frequencies in all uploaded files for the 10 domain families with the largest variation (see Figure 1b and Supplementary Figure S1). This variation chart allows to identify systematically different functional properties of the samples in terms of their Pfam domain annotation. A complete list of the highly differing Pfam families can be accessed using a hyperlink below the chart. A second bar chart displays the most varying GO categories, whereby the variation is calculated for GO terms that are associated with Pfam domain hits. Note that the hierarchical scheme of GO in general results in more frequent occurrences of top level terms as compared to more specific categories.

Clustering dendrogram. If more than two files are uploaded, CoMet can hierarchically cluster the input data using the UPGMA algorithm. The resulting dendrogram is shown on the CoMet result page (see Figure 1b and Supplementary Figure S2). If a group of nodes is sufficiently unrelated to other groups (group linkage value >70% of maximum linkage value), the groups are displayed using different colors for their associated dendrogram branches.

MDS. For datasets of three or more files CoMet also performs an MDS of the data (see 'Methods and Implementation' section). In the MDS plot, which is displayed next to the clustering dendrogram, the different input files are symbolized using various colors and

symbols according to the legend on the right-hand side of the plot (see Figure 1b and Supplementary Figure S3). The MDS plot allows to identify groups of samples that share functional properties in terms of the domain assignment frequencies.

Details page. CoMet provides detailed result pages for individual samples. On each result page a list of the top 10 Pfam domain families is shown that contains domain families ranked according to their number of hits in the file. Furthermore, a list of at most ten GO terms associated with the most frequent Pfam domains is displayed in descending order. If only one file is uploaded these lists will be displayed on the main result page.

Download of static data. The download section provides plain text files with all sequence-specific Pfam domain assignments and Pfam/GO term frequencies. In principle, the domain assignments can be used as an input to other comparative metagenomics tools that allow the import of pre-calculated assignments to Pfam families. Furthermore, the assignment files allow to estimate the taxonomic composition of a particular metagenome using domain-based profiling as provided by the Treephyler tool (22).

The download page also provides an archive file containing all figures from the result page in EPS format, which allows an easy integration into reports and publications.

Help page

The CoMet server provides a comprehensive help page, which can be accessed at any stage of the CoMet analysis. This page contains a short description of the CoMet server and its usage, an outline of the processing method and a detailed section describing how the results of the CoMet analysis can be accessed and interpreted. Furthermore, a link to example output of the CoMet analysis of three different datasets (2,21,23) can be found on the help page.

EVALUATION AND CASE STUDIES

Prediction performance

The accuracies of the Orphelia ORF finder and the UFO domain detection approach have been evaluated and discussed in (16,17). However, the UFO method within the CoMet server is based on a more recent version of the Pfam database (24.0) as compared to the original method (23.0). Therefore, we measured the prediction accuracy of the updated UFO domain detection in comparison to the widely-used RPS-BLAST (24) and HMMER (http://hmmer.janelia.org/) methods. For this purpose, we evaluated the consensus of domain hits of the three methods on a large metagenomic test dataset (see Methods in Supplementary Data). Here, the consensus sensitivity of the updated UFO method (80.4%) was lower as compared to HMMER (98.1%) and RPS-BLAST (96.6%). However, the consensus specificity of UFO (89.5%) was close to those of HMMER (92.7%) and RPS-BLAST (93.9%).

The statistical analysis within the CoMet server is based on domain frequency profiles rather than single sequence specific domain assignments (see 'Methods and Implementation' section). Therefore, we also measured the similarity of the frequency profiles of the three methods for the same test dataset in terms of Pearson's correlation coefficient ρ. Here, the highest correlation was observed for the two profiles associated with the RPS-BLAST and HMMER method ($\rho = 0.98$). The correlation of the UFO profile with the RPS-BLAST and HMMER profiles ($\rho = 0.93$ for both) was slightly lower.

Case studies using real metagenomic datasets

To validate the CoMet output on a well-studied example dataset, we analyzed metagenomic sequences from different microbial biomes, which have originally been compared in (21). The dataset includes 6.6 million unassembled reads obtained from pyro-sequencing based on the Roche 454 GS20 platform. For the CoMet analysis the data were pooled to compare the eight microbial biome specific profiles.

The resulting CoMet clustering dendrogram as well as the MDS analysis (Figure 1b and Supplementary Figures S2 and S3) both indicate that the functional profile of the coral metagenome is rather different from all other profiles. An exceptional role of the coral metagenome was also found in the original publication, where the authors measured a significantly lower functional diversity of the corresponding SEED profile. In addition, the authors found a salient peak in the relative frequency of respiration related genes for that metagenome. This is also evident from the CoMet output, where protein domains of respiratory chain enzymes (e.g. PF00361, PF00115, PF06455, PF00146) account for the largest differences between profiles (see Supplementary Figure S1). Furthermore, the CoMet overview on the most distinguishing functional categories indicates that photosynthesis related GO terms and protein domains are highly overrepresented in coral and marine metagenomes. A closer look on motility-related genes in the original study showed an overrepresentation of chemotaxis genes for the fish-associated metagenome, which is also highlighted by the significantly differing counts of the methyl-accepting chemotaxis protein (MCP) signaling domain (PF00015) in the pairwise CoMet comparisons. In the pairwise comparisons this metagenome also showed a significant overrepresentation of protein domains associated with transmembrane transport (GO:0055085) and related categories. This finding corresponds well with the original study where the fish-associated metagenome in comparison with the other metagenomes showed the highest fraction of sequences that could be assigned to membrane transport subsystems. An overrepresentation of sulfur metabolism-related genes in that metagenome as reported in the original publication could not be supported by the CoMet analysis. However, an inspection of the pairwise Pfam profile comparisons indicates a slight overrepresentation of taurine catabolism dioxygenase (PF02668) which may possibly reflect an increased utilization of taurine as a sulfur source.

As a second example that has originally been studied in (2), we analyzed a collection of eight environmental shotgun samples [Acid Mine Drainage biofilm, Sargasso sea (3 samples), Whale fall (3 samples), Minnesota farm soil]. The dataset comprises a total number of 1 454 641 sequences. Supplementary Figure S4 shows the clustering dendrogram resulting from a CoMet analysis of the samples using a significance threshold of $P = 0.01$. Analogously to the functional clustering obtained in the original work, a grouping of samples according to their environmental conditions can be observed in this dendrogram as well as in the corresponding MDS plot (Supplementary Figure S5).

Computational efficiency

Our runtime measurements on the 'metaseq' dataset indicate that the domain detection implemented in CoMet is approximately 2000 and 500 times faster than HMMER version 3.0 and RPS-BLAST, respectively. Therefore, the speed-optimized implementations of Orphelia and the UFO domain detection allow the rapid comparative analysis of large metagenomic samples using the CoMet server. For instance, the complete CoMet analysis process for the eight microbiomes dataset from (21) (6 605 808 sequences, 733 MB) took less than 15 min. As a second example, processing of the eight environmental shotgun samples as described in (2) (1 454 641 sequences, 2397 MB) only required about 37 min on the CoMet web server. In order to obtain a quick overview of the data at hand, this speed implies a great advantage over other comparative metagenomics platforms that perform a costly BLAST or HMMER analysis.

CONCLUSION

We presented the CoMet web server, which implements a complete pipeline for functional annotation and comparison of metagenomes. The combination of ORF finding, fast Pfam domain detection and comparative statistical analysis with an easy-to-use web interface allows to obtain a quick overview of putative functional differences in a set of metagenomic samples. In particular, the upcoming analytical challenges in the context of metatranscriptomics will require tools that are designed for large scale comparative analysis.

Although GO categories have only been used in few metagenomic studies so far, our results indicate the potential of GO terms for a comparative metagenome analysis. In addition to the existing GO profiles in CoMet, we are currently working on the integration of a KEGG pathway profiling that builds on the fast detection of Pfam protein domains.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dominic Simm and Rasmus Steinkamp for technical support. We further thank Frank-Oliver Glöckner and Renzo Kottmann for fruitful discussions and two anonymous reviewers for helpful comments.

FUNDING

Deutsche Forschungsgemeinschaft (grant numbers ME3138, LI2050). Funding for open access charge: Deutsche Forschungsgemeinschaft (grant number LI2050).

Conflict of interest statement. None declared.

REFERENCES

1. Daniel,R. (2005) The metagenomics of soil. *Nat. Rev. Microbiol.*, **3**, 470–478.
2. 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.
3. Turnbaugh,P.J., Hamady,M., Yatsunenko,T., Cantarel,B.L., Duncan,A., Ley,R.E., Sogin,M.L., Jones,W.J., Roe,B.A., Affourtit,J.P. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.
4. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
5. Meyer,F., Overbeek,R. and Rodriguez,A. (2009) FIGfams: yet another set of protein families. *Nucleic Acids Res.*, **37**, 6643–6654.
6. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
7. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
8. Aoki-Kinoshita,K.F. and Kanehisa,M. (2007) Gene annotation and pathway mapping in KEGG. *Methods Mol. Biol.*, **396**, 71–91.
9. Li,W. (2009) Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*, **10**, 359.
10. Seshadri,R., Kravitz,S.A., Smarr,L., Gilna,P. and Frazier,M. (2007) CAMERA: a community resource for metagenomics. *PLoS Biol.*, **5**, e75.
11. Gerlach,W., Junemann,S., Tille,F., Goesmann,A. and Stoye,J. (2009) WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, **10**, 430.
12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. and Eppig,J.T. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
13. Markowitz,V.M., Ivanova,N.N., Szeto,E., Palaniappan,K., Chu,K., Dalevi,D., Chen,I.M., Grechkin,Y., Dubchak,I., Anderson,I. *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.*, **36**, D534–D538.
14. Goll,J., Rusch,D.B., Tanenbaum,D.M., Thiagarajan,M., Li,K., Methe,B.A. and Yooseph,S. (2010) METAREP: JCVI metagenomics reports–an open source tool for high-performance comparative metagenomics. *Bioinformatics*, **26**, 2631–2632.

15. Kottmann,R., Kostadinov,I., Duhaime,M.B., Buttigieg,P.L., Yilmaz,P., Hankeln,W., Waldmann,J. and Glöckner,F.O. (2010) Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res.*, **38**, D391–D395.
16. Hoff,K.J., Tech,M., Lingner,T., Daniel,R., Morgenstern,B. and Meinicke,P. (2008) Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics*, **9**, 217.
17. Meinicke,P. (2009) UFO: a web server for ultra-fast functional profiling of whole genome protein sequences. *BMC Genomics*, **10**, 409.
18. Lingner,T. and Meinicke,P. (2008) Word correlation matrices for protein sequence analysis and remote homology detection. *BMC Bioinformatics*, **9**, 259.
19. Rodriguez-Brito,B., Rohwer,F. and Edwards,R.A. (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics*, **7**, 162.
20. Torgerson,W. (1952) Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**, 401–419.
21. Dinsdale,E.A., Edwards,R.A., Hall,D., Angly,F., Breitbart,M., Brulc,J.M., Furlan,M., Desnues,C., Haynes,M. and Li,L. (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.
22. Schreiber,F., Gumrich,P., Daniel,R. and Meinicke,P. (2010) Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, **26**, 960–961.
23. Kunin,V., Raes,J., Harris,J.K., Spear,J.R., Walker,J.J., Ivanova,N., von Mering,C., Bebout,B.M., Pace,N.R., Bork,P. *et al.* (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol. Syst. Biol.*, **4**, 198.
24. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.