Published online 6 June 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W433–W437***
doi:10.1093/nar/gkn284

Comparative Pathway Analyzer—a web server for comparative analysis, clustering and visualization of metabolic networks in multiple organisms

Sebastian Oehm[1,5,*], David Gilbert[2], Andreas Tauch[1,3], Jens Stoye[1,4] and Alexander Goesmann[1,5]

[1]Center for Biotechnology (CeBiTec), Bielefeld University, D-33594 Bielefeld, Germany, [2]Bioinformatics Research Centre, University of Glasgow, Glasgow G12 8QQ, UK, [3]Institut für Genomforschung und Systembiologie, [4]AG Genominformatik, Faculty of Technology and [5]Bioinformatics Resource Facility (BRF), Bielefeld University, D-33594 Bielefeld, Germany

Received February 11, 2008; Revised April 18, 2008; Accepted April 26, 2008

ABSTRACT

In order to understand the phenotype of any living system, it is essential to not only investigate its genes, but also the specific metabolic pathway variant of the organism of interest, ideally in comparison with other organisms. The Comparative Pathway Analyzer, CPA, calculates and displays the differences in metabolic reaction content between two sets of organisms. Because results are highly dependent on the distribution of organisms into these two sets and the appropriate definition of these sets often is not easy, we provide hierarchical clustering methods for the identification of significant groupings. CPA also visualizes the reaction content of several organisms simultaneously allowing easy comparison. Reaction annotation data and maps for visualizing the results are taken from the KEGG database. Additionally, users can upload their own annotation data. This website is free and open to all users and there is no login requirement. It is available at https://www.cebitec.uni-bielefeld.de/groups/brf/software/cpa/index.html.

INTRODUCTION

In molecular biology, the comparative analysis of organisms exhibiting different phenotypes is still mostly performed on the level of genes. However, many genes encode proteins that are involved in complex metabolic pathways, whose functions result from the interplay of all involved enzymes. So to understand the phenotype of any living system, it is essential to not only investigate single genes in isolation, but also the metabolic pathway variant of the particular organisms under study.

Interesting questions to study in this context are 'what are the commonalities and differences within a group of organisms' or more elaborately, if you subdivide the group of organisms of interest into subsets, 'what are the features all organisms of one set share while all members of another set completely lack these functions?' For example, consider analyzing a set of pathogens versus a set of non-pathogens for identifying reactions associated with pathogenicity. Sets of organisms instead of single representatives are used for comparison, because it is not the goal to find special innovations present in one species only, but rather more general differences that could be interpreted as principles of pathogenicity. The set of reactions or respectively their catalyzing enzymes resulting from answering the above questions may serve as candidate set for finding new drug targets. Throughout the rest of this article we will use the term *differential reaction content* to describe the set of reactions that do not occur in all organisms under study.

Forst *et al.* (1) have published an algebraic method for comparing networks that can be used to find the 'metabolic innovations' in a set of organisms as compared to a second set of organisms. They use this method to find those reactions that occur in at least one organism out of a predefined set of organisms and are missing in all organisms of another predefined set. However, it is of interest to also detect all reactions that occur in exactly all organisms of the first set, while missing in precisely all organisms of the second set (which is a subset of the above mentioned 'metabolic innovations' and can also be computed using the algebraic method) and of course vice versa. These reactions will in the following be called *unique reaction content*.

A method for pairwise protein interaction network alignment was published by Kelley *et al.* (2) that aims at

*To whom correspondence should be addressed. Tel: +49 521 106 5166; Fax: +49 521 106 6419; Email: sebastian.oehm@cebitec.uni-bietefeld.de

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

identifying conserved interaction pathways and complexes. Their method combines protein interaction topology and protein sequence similarity in a distance measure. There also exists a web server (3) allowing queries of a short protein interaction path against a target protein–protein interaction network selected from a network database.

The MetaCyc/BioCyc collection of pathway/genome databases (4) relies on a program called 'Pathway Tools' that permits comparative metabolic pathway analysis for two or more organisms. In the web version results are presented as lists, whereas the GUI version not only contains improved comparative methods, but also displays the results visually on pathway maps. However, this version needs to be installed on a local machine along with the organism databases, which requires large amounts of disk space and computational power.

Yet, in order to enable easy access to results, it is essential to visualize the detected reaction content in a graphical way. It should become apparent at one glance whether a reaction occurs in all organisms or only in organisms of one of the sets and in the latter case whether it occurs in all members of this set or only in a subset.

A crucial point in detecting the differential reaction content is the choice of which organisms to put into the two sets to be compared. If not chosen appropriately, reaction content that is unique for one of the sets and thus possibly worth to be further analyzed might not be found (e.g. a reaction shared by all organisms of set 1 and one organism of set 2). In order to identify a good subdivision, one has to find sets of organisms with high similarity in terms of their reaction content within each set and low similarity across the sets.

Furthermore, our analyses showed that the metabolic reaction content of a group of organisms of interest often does not allow for a grouping that makes it possible to detect unique reaction content. When comparing the whole network at once, differences in subnetworks may cancel each other out and thus may lead to groupings with less discriminative power.

In this article, an approach for metabolic pathway analysis and its web implementation is presented. It aims at finding unique metabolic reaction content that is worth to be further analyzed. The website provides a visualization engine for displaying the differential metabolic reaction content resulting from comparing two sets of organisms. Furthermore, the server provides an automated approach based on clustering techniques for finding an appropriate grouping. In order to find sets that are well suited for detecting unique reaction content, we use a subdivision of the overall metabolic network into smaller subnetworks and separately apply the clustering on each of them.

THE CPA WEB SERVER

We have designed and built a web server that supports researchers in comparative metabolic network analysis. It is based on the metabolic reaction annotation provided by the KEGG database (5) and thus currently comprises 155 eukaryotes, 569 bacteria and 49 archaea (30 December 2007). In order to keep the contents up to date, we update our data monthly. Users can additionally upload their own reaction annotation data to include it in their analysis. Currently, two file formats are supported: text files containing one KEGG reaction identifier (e.g. R00001) per line, and files in EMBL format (6) containing EC number annotation.

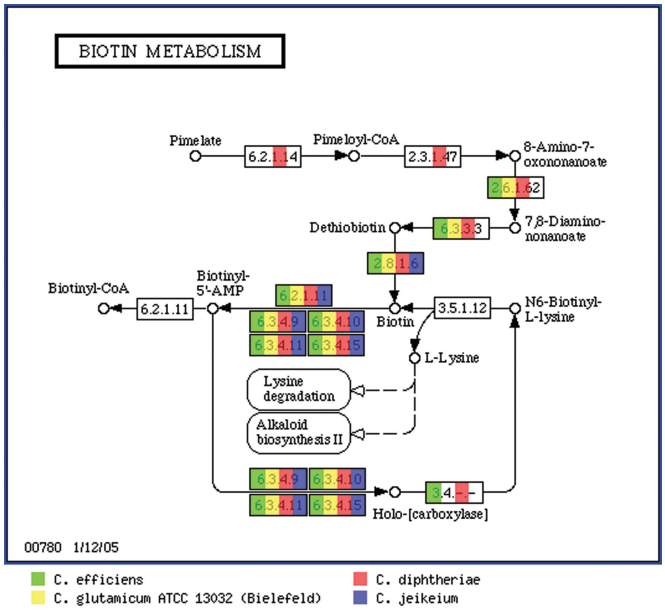

Figure 1. Reaction Content Visualizer showing the reaction content in the biotin metabolism (KEGG) for several Corynebacteria simultaneously.

Displaying reaction content of several organisms simultaneously

One feature of our web server is a visualization engine called 'Reaction Content Visualizer' that—like KEGG—permits the display of the reaction content of organisms on static metabolic pathway maps by coloring the respective enzymes. The pathway maps and reaction data are taken from the KEGG database.

Unlike the KEGG website, our server allows the user to choose several organisms, and displays simultaneously their reaction content using a user-specified color for each organism (Figure 1). Visualizing the reactions within their pathway context and not as plain lists facilitates assessing the functional relevance of missing a certain reaction. Using maps of individual pathways for visualization, instead of one map of the whole metabolic network, allows for easy visual inspection.

Displaying differential reaction content

A second visualization engine called 'Differential Reaction Content Visualizer' allows the user to choose two sets of organisms and a KEGG pathway (Figure 2). Pressing the 'Display pathway' button invokes calculating the differential reaction content, which then will be displayed on the well-known KEGG pathway maps. Here, each reaction is colored according to whether it occurs in both sets of

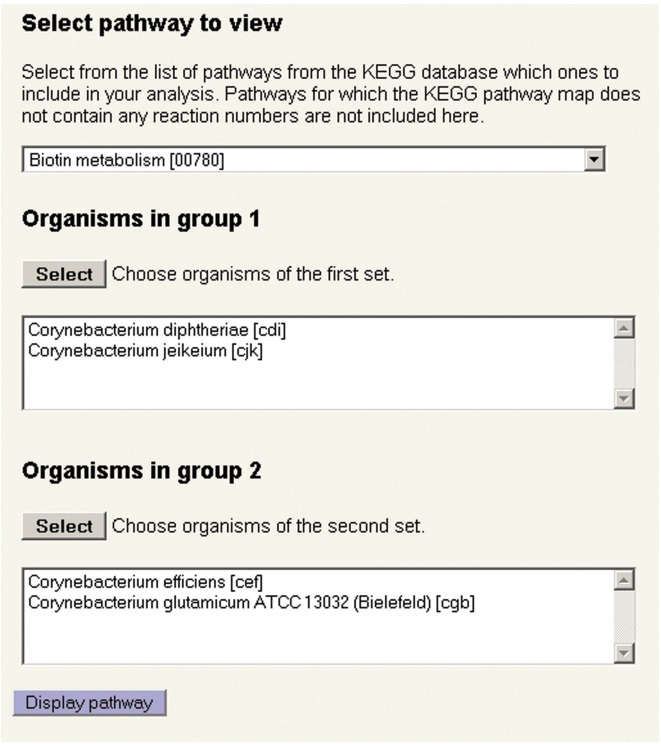

Figure 2. Differential Reaction Content Visualizer: select boxes allow the user to choose pathway, organisms in set one and organisms in set two (from top to buttom). Organisms already selected for the first set are not allowed to be included in the second set and vice versa.

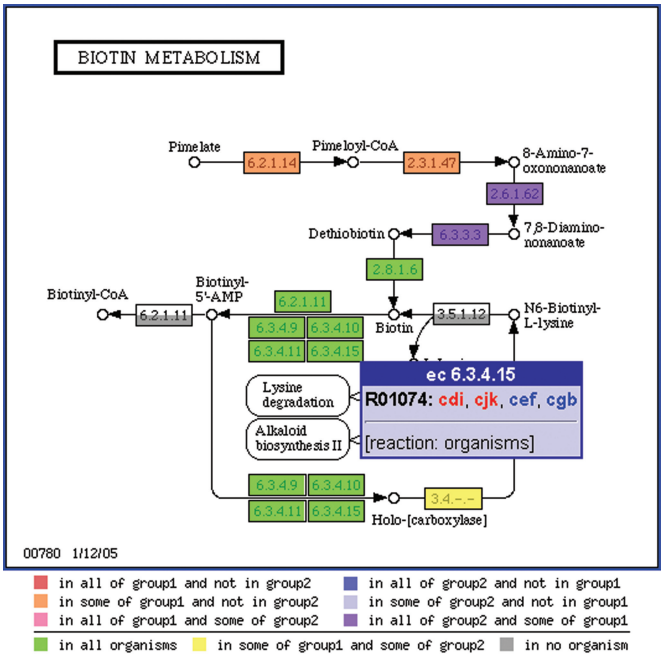

Figure 3. Differential Reaction Content Visualizer applied to the biotin metabolism for the Corynebacteria *C. jeikeium* (cjk) and *C. diphtheriae* (cdi) in the first set compared against *C. glutamicum* (Kyowa Hakko) (cgl), *C. efficiens* (cef) and *C. glutamicum* (Bielefeld) (cgb) in the second set. For each box containing an EC number a tooltip lists all KEGG reactions associated to the respective EC number and all organisms this reaction is annotated for. Red organisms belong to set one and blue organisms to set two.

organisms or only one and for the latter case whether it occurs in all organisms of the respective set or only in a subset (Figure 3).

Subdividing the organisms of interest into two sets

As already indicated, a crucial point for analyzing the differential reaction content is the choice of which organisms to put into the two sets to be compared. To support the user in making this decision, we provide a number of hierarchical clustering methods that automatically group organisms with similar variants of their metabolic network together. Thus, the user can either rely on preliminary knowledge or use the clustering dendrograms to find an adequate grouping of organisms. Unexpected groupings and the corresponding unique reaction content are likely to be detected using the latter.

Another problem arises when organisms are clustered on their overall metabolic reaction network. Our analyses showed that for different parts of the metabolic network in most cases different groupings would be a good choice. In other words, once you fix the grouping, you find some of the discrepancies, but will miss others, which you would only find using another grouping. However, the second grouping will in turn miss the findings enabled by the first grouping. To avoid this, we subdivide the overall metabolic network into smaller networks and apply the clustering analysis separately on each of them. On our web server we use the subdivision provided by the KEGG pathway maps: all reactions on one of these maps constitute one pathway. Additionally, users can define their own pathway by uploading a list of KEGG reaction identifiers. Because in this case the resulting differential reaction content cannot be displayed on KEGG pathway maps, results are provided in tabular form as well as displayed on pathway maps generated by the graph visualization software graphviz (http://www.graphviz.org/).

The start page for the clustering allows the user to choose the organisms of interest and pathways to analyze. The annotation data for the organisms and pathways are taken from the KEGG database. Furthermore, users can upload their own annotation data as well as define their own pathways. Then the user can decide which distance measure to use for calculating the similarity between the pathway variants in two different organisms. Two choices are provided: a reaction edit distance on metabolic network graphs and the normalized network distance defined in Forst *et al.* (1). Pressing the button 'Start calculations' invokes clustering algorithms on our compute cluster.

Once calculations are finished, results are displayed on a new page (Figure 4). On the top of the page the user can select one pathway from the analysis set. After pressing the 'Display results' button, clustering dendrograms of 'average linkage' (7), 'complete linkage' (7) and 'ward' (8) hierarchical clustering are displayed for this pathway. Below each of these images, preselected lists of the automatically detected groupings are presented to the user who can directly invoke the visualization of the

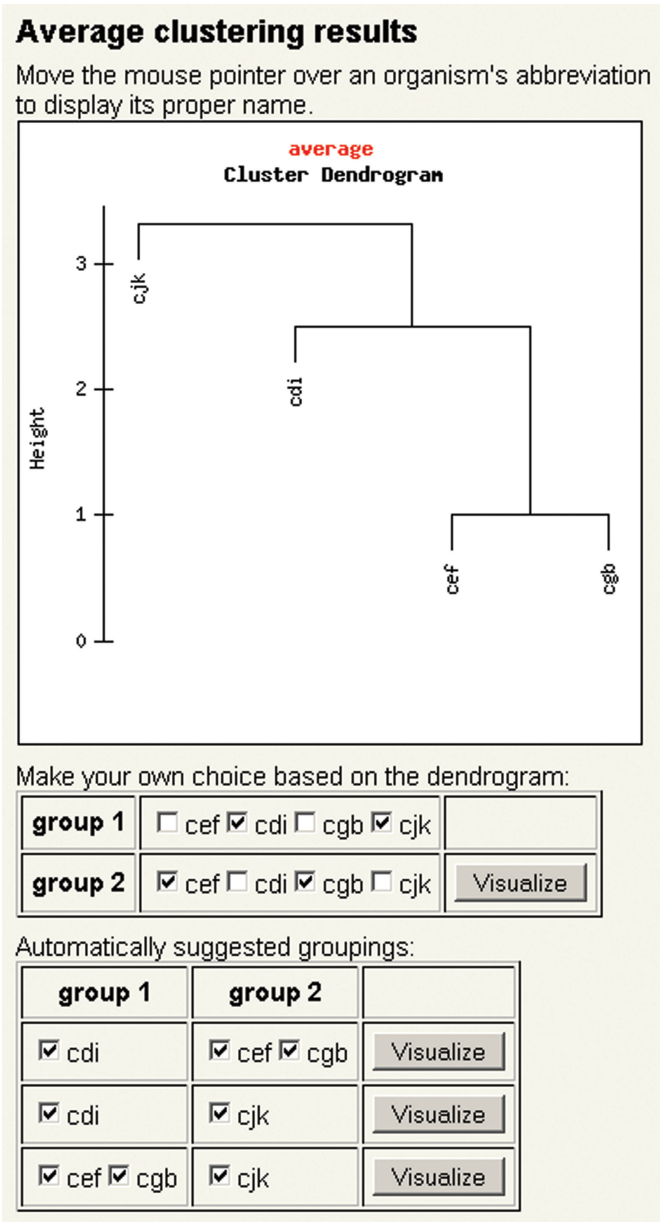

Figure 4. Dendrogram and suggested groupings from applying 'average clustering' on the biotin metabolism for Corynebacteria. cjk: *C. jeikeium*; cdi: *C. diphtheriae*; cgb: *C. glutamicum* (Bielefeld); cef: *C. efficiens*; cgl: *C. glutamicum* (Kyowa Hakko).

differential reaction content based on this grouping. Additionally, the user can make his own selection after visual inspection of the dendrograms. This is useful, since the automatic detection can never be guaranteed to find the optimal solution. Hyperlinks are provided next to each grouping to start the Differential Reaction Content Visualizer using the respective grouping and pathway. On the very top of this page the user can follow a link to an overview page listing all pathways and the respective differential reaction content. The pathways are sorted according to the differential reaction content so that most different pathways appear at the top of the list. This is helpful, because it is often not known which pathway yields interesting results.

We calculate a distance measure to assess how close two organisms are to each other in terms of their metabolic reaction content. We calculate the pairwise distances for each single pathway and then cluster the organisms with average linkage, single linkage and ward hierarchical clustering techniques. We automatically extract groups of organisms from the resulting clustering dendrograms. These groupings are those that maximize the cophenetic correlation coefficient (9), which is a measure for cluster quality.

APPLICATION EXAMPLE

As an application case we analyze biotin metabolism (10) in different Corynebacterium species, with the goal of detecting which species show similar (annotated) reaction content with respect to this pathway. For this task we choose the 'CPA clustering' website, select pathway and organisms, and start the calculations. The resulting clustering dendrograms and suggested groupings are displayed on another web page. In this case, the clustering dendrograms for different clustering techniques are almost identical; they only differ in the height, at which different groups are merged (Figure 4). *Corynebacterium efficiens* (cef) and *C. glutamicum* (Kyowa Hakko) (cgl) have no differences, *C. glutamicum* (Bielefeld) (cgb) is close to them, though not identical, *C. diphtheriae* (cdi) differs and *C. jeikeium* (cgk) is different from both *C. diphtheriae* and the set of the other three organisms. There are three automatically suggested groupings. However, we decide to put *C. jeikeium* and *C. diphtheriae* into a single set and compare this set against the set of the other three. We select the respective checkboxes for 'group 1' and 'group 2' on the website and click 'Visualize'. Now the differential reaction content is calculated and displayed using the Differential Reaction Content Visualizer (Figure 3). One immediately notices that reactions in the upper part of the map are not shared among all organisms. KEGG reaction R03182 (EC 6.3.3.3) and R03231 (EC 2.6.1.62) occur in all organisms of set 1 and *C. diphtheriae*, whereas R03209 (EC 6.2.1.14) and R03210 (EC 2.3.1.47) are only annotated for *C. diphtheriae*. Although this image reflects current knowledge, it has not been proven yet in the wet-lab, whether *C. diphtheriae* actually uses the additional path from pimelate to 8-amino-7-oxononanoate for synthesizing biotin. Note, that the automatic clustering for this pathway correctly suggested not to put the two pathogens among the analyzed organisms, *C. diphtheriae* and *C. jeikeium*, into the same set, as an unexperienced user might have done. Only the green colored reactions are annotated for *C. jeikeium*, which clearly makes it the outsider in this analysis.

Another application case is the comparison of metabolic reaction content of pathogenic bacteria against human. The goal of this pathogen–host comparison can be the identification of reactions that occur only in bacteria but not in human. Bacterial enzymes coding for these reactions may serve as potential drug targets, if they are specific to the bacteria and essential for their survival. Using the Comparative Pathway Analyzer it is easily possible to

screen for such organism specific reactions. As an example, we apply the clustering aproach to *Pseudomonas aeruginosa* PA7, *Pseudomonas aeruginosa* PAO1 and *Homo sapiens*. When screening the results on the overview page, all pathways are of interest for that any reactions are reported to be present in the pathogens only. One example is the KEGG alanine and aspartate metabolism: the detected reactions are the KEGG reactions R00401, R00490 and R00357 (Supplementary Figure 1). R00401, which in *P. aeruginosa* is catalyzed by the alanine racemase, has already been shown by Perumal *et al.* (11) to be a potential drug target in their comprehensive analysis of metabolic enyzmes in *P. aeruginosa*. The conversion between L- and D-alanine is essential for *P. aeruginosa* PAO1 due to the fact that D-alanine is a necessary component of the bacterial cell wall (12). Whether or not the other reactions may serve as potential drug targets remains to be proven.

SUMMARY

We presented CPA, a web server that supports researchers in analyzing the metabolic reaction content of organisms. Analyses are based on the annotation provided by the KEGG database and optionally by additional data sets the user may upload. The server can be used for simultaneously visualizing the reaction content of several organisms. It also permits finding and visualizing the differential reaction content of two sets of organisms. Clustering methods can be invoked for finding appropriate sets of organisms as a basis for detecting the differential reaction content. Visualization and pathway definition are by default based on KEGG pathway maps, but we also offer this functionality for user defined pathways. Note that results of the analyses presented here are not only sensitive to true differences, but are also strongly influenced by the quality of annotation of the organisms under investigation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

S.O. thanks the International Graduate School in Bioinformatics and Genome Research for providing financial support. S.O. was also supported by a student summer scholarship from the Department of Computer Science at the University of Glasgow. This work was funded by Bundesministerium für Bildung und Forschung (0313939A); Deutsche Forschungsgemeinschaft (GRK635). Funding to pay the Open Access publication charges for this article was provided by Department of Genetics, Bielefeld University.

Conflict of interest statement. None declared.

REFERENCES

1. Forst,C.V., Flamm,C., Hofacker,I. and Stadler,P. (2006) Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinform.*, **7**, 67
2. Kelley,B.P., Sharan,R., Karp,R.M., Sittler,T., Root,D/E., Stockwell,B.R. and Ideker,T. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Nat. Acad. Sci.*, **100**, 11394–11399.
3. Kelley,B.P., Yuan,B., Lewitter,F., Sharan,R., Stockwell,B.R. and Ideker,T. (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.*, **32(Suppl 2)**, W83–W88.
4. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **36**, D623–D631.
5. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
6. Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A., Bates,K., Bhattacharyya,S., Bower,L., Browne,P. *et al.* (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.*, **35**, D16–D20.
7. Everitt,B.S., Landau,S. and Leese,M. *Cluster Analysis*. Hodder Arnold, London, 2001.
8. Ward,J.H. Jr. (1963) Hierarchical grouping to optimize an objective function. *J. Amer. Stat. Assn.*, **58**, 236–244.
9. Sokal,R.R. and Rohlf,F.J. (1962) The comparison of dendrograms by objective methods. *Taxon*, **11**, 33–40.
10. Streit,W.R. and Entcheva,P. (2003) Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. *Appl. Microbiol. Biotechnol.*, **61**, 21–31.
11. Perumal,D., Lim,C.S., Sakharkar,K.R. and Sakharkar,M.K. (2007) Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification. *In Silico Biol.*, **7**, 453–465.
12. Strych,U., Huang,H.C., Krause,K.L. and Benedik,M.J. (2000) Characterization of the alanine racemases from *Pseudomonas aeruginosa* PAO1. *Curr. Microbiol.*, **41**, 290–294.