Multi-genome core pathway identification through gene clustering

Dimitrios M. Vitsios[1], Fotis E. Psomopoulos[1,2,*], Pericles A. Mitkas[1],
Christos A. Ouzounis[2]
[1] Dept. of Electrical and Computer Engineering
Aristotle University of Thessaloniki
GR541 24, Thessaloniki, Greece
[2] Institute of Agrobiotechnology
Center for Research and Technology Hellas
GR570 01, Thessaloniki, Greece

[*]corresponding author: fpsom@issel.ee.auth.gr

Abstract. In the wake of gene-oriented data analysis in large-scale bioinformatics studies, focus in research is currently shifting towards the analysis of the functional association of genes, namely the metabolic pathways in which genes participate. The goal of this paper is to attempt to identify the core genes in a specific pathway, based on a user-defined selection of genomes. To this end, a novel methodology has been developed that uses data from the KEGG database, and through the application of the MCL clustering algorithm, identifies clusters that correspond to different "layers" of genes, either on a phylogenetic or a functional level. The algorithm's complexity, evaluated experimentally, is presented and the results on a characteristic case study are discussed.

Keywords: bioinformatics; metabolic pathways, clustering algorithm; phylogenetic analysis

1 Introduction

Metabolomics is the scientific study of chemical processes involving metabolites. Specifically, metabolomics is the "systematic study of the unique chemical fingerprints that are left behind specific cellular processes". The metabolome represents the collection of all reactants (enzymes, proteins or other chemical compounds) in a biological cell, tissue, organ or organism, which are the end products of cellular processes.

In biochemistry, metabolic pathways are series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by a series of chemical reactions. Enzymes catalyze these reactions, and often require dietary minerals, vitamins, and other cofactors in order to function properly. Because of the many chemicals (a.k.a. "metabolites") that may be involved, metabolic pathways can be quite elaborate. In addition, numerous distinct pathways co-exist within a cell. This collection of pathways is called the metabolic network.

For the current study, pathway-related information was retrieved from KEGG Pathway Database (KEGG) [1], a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for various processes and systems including “Metabolism”, “Human Diseases” and “Cellular Processes” among others. The analysis of completely sequenced genomes can yield useful insight into the evolution and multi-level organization of organisms. With the current advances in genomics and proteomics, it has become imperative to explore its impact as reflected in the metabolic signature of each genome [2]. To this end a methodology is presented, which applies a clustering algorithm to genes from different species participating in the same pathway.

2 Problem Description

Considering the fact that the majority of the known genomes from all three domains (Archaea, Bacteria and Eukaryota) use highly similar chemical processes in the form of metabolic pathways, it is a logical step to assume that the dissimilarities in the particular expressions in the metabolic pathways of different genomes can be viewed as the result of the evolution of a reference pathway existent at the common ancestor of those genomes. Specifically, considering that each pathway can be represented by a graph, then, starting from the graph of the reference pathway, the currently formed pathways have emerged essentially after the addition or deletion of some nodes and/or edges from the reference graph, while preserving the core body of the pathway virtually unchanged. In fact, by observing the graph representing a metabolic pathway using several genomes as a reference, several compact sets of edges (i.e. reactions) can be distinguished that seem to be highly conserved in the genomes, whereas at the same time there exist several sub-graphs that have been added or deleted, making evident the differentiation of the pathway expression between the genomes. This observation leads us to the conclusion that the transformations that have been carried out in the gene-level of the genomes through evolution due to mutations of the genome sequences have resulted to changes in the topology of the metabolic pathway while retaining the overall process.

Thus, the objective of the current study is to extract a meaningful clustering of the genes participating in a common metabolic pathway of several genomes by comparing the genes sequences and by evaluating the degree of homology between genes from different genomes. The result of the applied process is the extraction of cohesive gene groups that can give us information about the evolutionary similarity between the genomes examined or the uniqueness of some reactions in particular genomes. This knowledge can then be used in order to correlate some macroscopic differences between the genomes with the differences that arise between them in the pathway level.

2.1 Related work

Genes usually do not act individually but form functional or structure organizations, exemplified by metabolic pathways. As metabolic pathways are essential to the survival of organisms, and their evolution has been under debate for more than half a century [3], a combined phylogenetic and phenetic analysis of pathway topology might expand the understanding of the evolutionary processes molding their form and structure.

Several groups have carried out phylogenetic analyses based on metabolic pathways, deriving phylogenetic trees from the information of individual pathways [4-6], the presence and absence of entire pathways [7], or the reaction content of entire pathways [8]. These studies have provided valuable insight into the evolution of metabolism; however, as phylogenetic trees, they have generally diverged substantially from trees based on 16S rRNA, the most used molecule for phylogeny reconstruction. A common feature of phylogenetic trees based on metabolic information is that, owing to similar evolutionary pressures, organisms in similar habitats tend to be clustered together, and Aguilar et al. [9] therefore regarded such trees as phenetic rather than phylogenetic. Furthermore, one group showed that trees based on different subsets of metabolic networks were different [9], and another result also indicated a similar situation when several different pathways were used to construct trees separately [6].

On the other hand, phylogenetic profiles are commonly used in evolutionary studies, as they are based on sequence similarity. There are several recent approaches that either directly utilize phylogenetic profiles for functional prediction of gene clusters [10] or combine them with other biological data sources for increased sensitivity [11]. A slightly different approach, and an intermediate step towards the work presented in this paper, is to produce a tree-like structure of gene clusters in order to reconstruct the evolutionary relationships between them [12]. However, it is shown that the output of large-scale reconstructions is notably more difficult to interpret biologically. Our approach extends this work by aiming to extract close gene associations from metabolic pathways through unsupervised clustering at a sequence level. This level of association can be enhanced if the phylogenetic relationship of the corresponding genomes is taken under consideration.

3 Algorithm overview

The algorithm accepts as input a KEGG pathway map identifier *mapID*, and a list of ***n*** genome identifiers that comprise the target data set. A specific clustering algorithm (MCL [13] or EM [14]) is also selected for the current run and some parameters can be set i.e. the inflation parameter for MCL and the e-Value threshold for the construction of the homology matrix. We have determined through extensive experimentation the optimal value for the inflation parameter ('12') and the recommended value for the execution of the methodology (Table 1). Specifically, setting the inflation parameter to the optimal value stabilizes the number of clusters produced by the algorithm while at the same time yielding the optimal results (maximum values) with regards to the average intra-cluster similarity and homology.

By retrieving specific data (gene identifiers, FASTA sequences and the corresponding EC identifiers) about the $\boldsymbol{k_i}$ genes ($\boldsymbol{i}$ = 1 … ***n***) from each genome that participate in *mapID*, ***n*** blastable databases are constructed. In the next step, blast searches are performed, leading eventually to a matrix containing the phylogenetic profiles of all genes participating in the study. The homology between genes is determined using the default BLAST values and an e-Value threshold of 10^{-5}. The data in this homology matrix $\boldsymbol{P}$ are then clustered using the MCL or EM algorithm and a custom similarity metric based on the jaccard metric, given from the following equation:

$$\mathrm{sim}_{jac} = \frac{m_{11}+m_{00}}{m_{01}+m_{10}+m_{11}} \qquad (1)$$

where:

$\boldsymbol{m_{11}}$ represents the number of genomes where both genes (the distance of which is calculated) have a homologue,
$\boldsymbol{m_{00}}$ represents the number of genomes where both genes do not have a homologue and
$\boldsymbol{m_{10}}$ ($\boldsymbol{m_{01}}$) represents the number of genomes where the first gene (second gene) has a homologue while the second one (first one) does not.

After thorough experimentation using both the MCL and the EM clustering algorithms (data not shown), we have concluded that the clusters generated with MCL are more robust and more closely correlated to the biological aspect of the problem, while the clusters produced by EM cannot be readily interpreted. Generally, in many cases the EM's results tend to closely resemble with the respective results acquired with MCL. However, although they both generate almost the same number of clusters and have over 70% similarity between the "corresponding" clusters, they do not provide an equivalent level of granularity regarding the resulted clustering. Moreover, in all the experimental setups that were performed, EM cannot attain the distinction in separate groups of the genes belonging to a specific genome from the entire dataset, something that is achieved by MCL. On the other hand, in sharp contrast with EM, the MCL algorithm leads to a degenerate clustering (i.e. singleton cluster) when the organisms selected for a test case have a close phylogenetic relationship. Thus, in case resulting clusters form EM can be sufficiently interpreted from a biological aspect, then we can assume that EM is more advantageous than MCL in cases of limited phylogenetic diversity. For the current study however, the development of the overall methodology is based on the use of MCL as the genes' clustering algorithm.

Following that step, a clustering validation is performed to determine the similarity and consequently the significance of the clusters generated and the FASTA files for each of the clusters are retrieved from KEGG. The homology score for each pair of clusters $\boldsymbol{i}$ and $\boldsymbol{j}$ (at index $\boldsymbol{i,j}$ of the homology matrix) is the total number of homologues found between clusters $\boldsymbol{i}$ and $\boldsymbol{j}$ divided by the product of the total numbers of genes in clusters $\boldsymbol{i}$ and $\boldsymbol{j}$. The homology between each pair of genes is determined by executing blast searches for all pairs of FASTA sequences corresponding to the genomes participating in the current test case and considering that an e-Value score lower than 10^{-5} indicates the existence of homology between the examined genes. The similarity score for each pair of clusters $\boldsymbol{i}$ and $\boldsymbol{j}$ is the sum of the distances (Eq. 1) between all pairs of genes in clusters $\boldsymbol{i}$ and $\boldsymbol{j}$ divided by the product of the total numbers of genes in clusters $\boldsymbol{i}$ and $\boldsymbol{j}$. The validation process is performed by comparing the intra-cluster similarity and homology to the respective inter-cluster values. In all cases (data not shown), it was confirmed that the ratio of intra-cluster to inter-cluster values was in the range of 10, with the notable exception of a few minor clusters were the ratio was closer to 1.

Table 1. Statistical overview of several different runs of the methodology using MCL for various values of the inflation parameter. The reference pathway is the Glycolysis / Gluconeogenesis and the examined organisms are Escherichia Coli K-12 MG1655 (eco), Arabidopsis Thaliana (ath) και Homo Sapiens (hsa).

MCL inflation **Results**	*2*	*6*	*10*	***12***	*16*	*20*	*24*	*30*
Number of clusters	2	3	3	**4**	4	4	2	1
Average inter-cluster similarity	0.29	0.42	0.42	0.48	0.48	0.48	0.29	-
Standard deviation of inter-cluster similarity	0	0.21	0.21	0.28	0.28	0.28	0	-
Average intra-cluster similarity	1.92	1.39	1.39	1.81	1.81	1.81	1.92	0.77
Standard deviation of intra-cluster similarity	1.53	0.41	0.41	0.86	0.86	0.86	1.53	0
Average inter-cluster homology	0.035	0.02	0.02	0.016	0.016	0.016	0.035	-
Standard deviation of inter-cluster homology	0.035	0.017	0.017	0.014	0.014	0.014	0.035	-
Average intra-cluster homology	0.12	0.14	0.14	0.18	0.18	0.18	0.12	0.08
Standard deviation of intra-cluster homology	0.045	0.018	0.018	0.07	0.07	0.07	0.045	0

Finally, a post-processing procedure is applied to the data of the clusters in order to extract additional information regarding the composition of the clusters and the distribution of similar genes across the genomes and the acquired results are used to retrieve the customized colored versions of the examined metabolic pathways by cluster and by genome. The flowchart of the algorithm is shown in Fig. 1.

3.1 Algorithm complexity and efficiency

The maximum data size processed using that algorithm involved 325 genomes and 10,327 genes, due to memory space limits, considering 4 GBs of available memory space. The corresponding execution time of the algorithm reached about 15 hours using a 3.0 GHz Quad-Core Processor. The overall complexity of the implemented methodology is shown in Fig. 2.

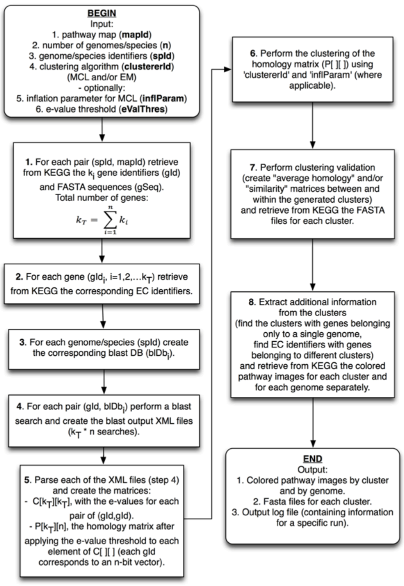
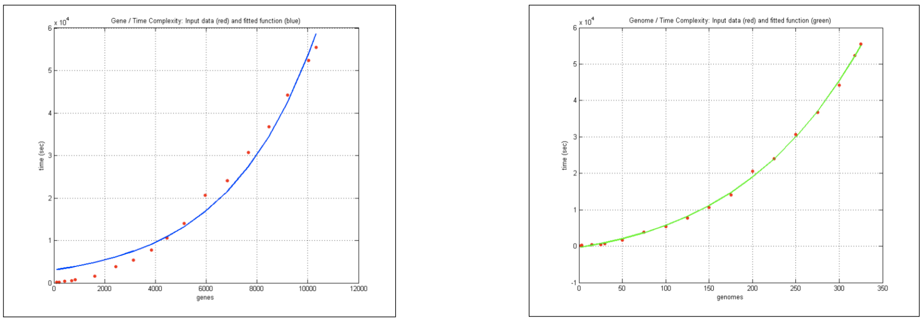

BEGIN Input: 1. pathway map (mapId) 2. number of genomes/species (n) 3. genome/species identifiers (spId) 4. clustering algorithm (clustererId) (MCL and/or EM) - optionally: 5. inflation parameter for MCL (inflParam) 6. e-value threshold (eValThres)
1. For each pair (spId, mapId) retrieve from KEGG the k_i gene identifiers (gId) and FASTA sequences (gSeq). Total number of genes: $k_T = \sum_{i=1}^{n} k_i$
2. For each gene (gId_i, i=1,2,...k_T) retrieve from KEGG the corresponding EC identifiers.
3. For each genome/species (spId) create the corresponding blast DB ($blDb_i$).
4. For each pair (gId, $blDb_i$) perform a blast search and create the blast output XML files (k_T * n searches).
5. Parse each of the XML files (step 4) and create the matrices: - C[k_T][k_T], with the e-values for each pair of (gId,gId). - P[k_T][n], the homology matrix after applying the e-value threshold to each element of C[ ][ ] (each gId corresponds to an n-bit vector).
6. Perform the clustering of the homology matrix (P[ ][ ]) using 'clustererId' and 'inflParam' (where applicable).
7. Perform clustering validation (create "average homology" and/or "similarity" matrices between and within the generated clusters) and retrieve from KEGG the FASTA files for each cluster.
8. Extract additional information from the clusters (find the clusters with genes belonging only to a single genome, find EC identifiers with genes belonging to different clusters) and retrieve from KEGG the colored pathway images for each cluster and for each genome separately.
END Output: 1. Colored pathway images by cluster and by genome. 2. Fasta files for each cluster. 3. Output log file (containing information for a specific run).

Fig. 1. Flowchart demonstrating the steps of the algorithm.



a.

b.

Fig. 2. Execution of the algorithm for different number of genes [a] and genomes [b]. In both cases the exponential character of the algorithm is evident.

3.1 Software implementation

The implementation of the algorithm outlined in the paper incorporates several different software tools and libraries that are combined and interconnected on top of a Java-based application framework. Specifically, data retrieval from KEGG database is achieved through SOAP-based web-services provided by the KEGG API. The construction of the blastable databases and the consequent BLAST searches are performed using the BLAST libraries (version 2.2.25+) provided by NCBI, whereas the output of the BLAST runs is read and organized using a custom XML DOM parser. The gene clustering is performed using the MCL implementation provided by the author of the original paper [15] while the EM's implementation used is the one provided by WEKA Suite [16]. Finally, the application makes use of certain functions derived from the Matlab Bioinformatics toolbox [17] for the calculation of phylogenetic profiles using metabolic data in conjunction with information extracted from the resulted genes clustering. Those functions are accessed via Matlab generated Java components that are executed with MCR runtime engine [18]. The overall application is built on top of a common Java layer which is, besides the combination and coordination of the aforementioned heterogeneous tools and libraries, responsible for several pre/post-processing, evaluation and data mining tasks. Finally, the java tool that has been implemented is publicly available from the following url: http://olympus.ee.auth.gr/~fpsom/alignPaths-pkg_Bundle_v0.9.tar.gz

4 Results overview

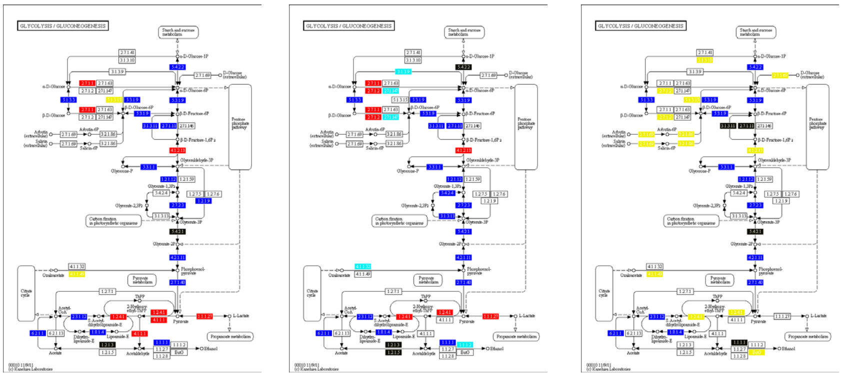
The proposed method was applied on several different pathways and genome sets and validated through both statistical methods and literature reviews. A characteristic test case is presented here, namely the application of the method on the Glycolysis / Gluconeogenesis metabolic pathway (KEGG identifier map00010) which is a well known and extensively documented pathway.

Derived from the Greek stem glyk-, "sweet," and the word lysis, "dissolution", glycolysis is an ancient pathway employed by a host of organisms. It is the sequence of reactions that metabolizes one molecule of glucose to two molecules of pyruvate with the concomitant net production of two molecules of ATP. Glycolysis is an energy-conversion pathway in many organisms, and it is tightly controlled. Gluconeogenesis is the opposite pathway from glycolysis, generating glucose from non-carbohydrate carbon substrates such as lactate, glycerol, and glucogenic amino acids. Gluconeogenesis and glycolysis are reciprocally regulated.

The three genomes participating in the test case are Arabidopsis Thaliana (*ath*), Escherichia Coli K-12 MG1655 (*eco*) and Homo Sapiens (*hsa*), and were selected for sufficient phylogenetic diversity. The total number of genes in this dataset is 209, distributed across the three genomes as follows: ath: 105, eco: 39 and hsa: 65. The selected algorithm was MCL and the results are presented in detail in the following section.

4.1 Experimental results

The results of the test case are presented here along with some interesting observations. Specifically, the pathway images shown below depict the distribution of the genes both by cluster and by genome after the clustering is performed and provide useful information about the internal structure of a single pathway as a result of genomes evolution.

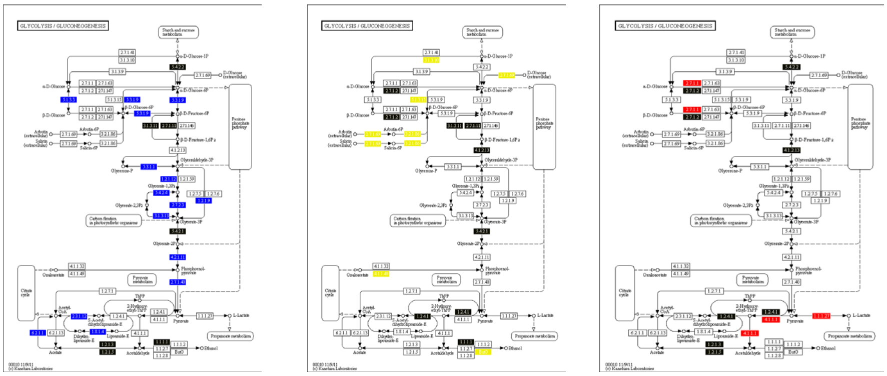


Arabidopsis Thaliana Escherichia Coli K-12 MG1655 Homo Sapiens

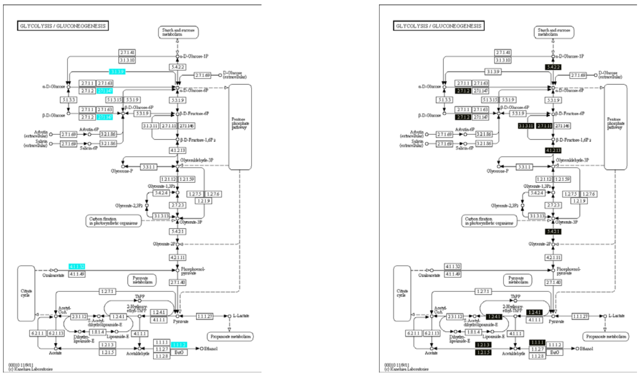
Fig. 3. The Glycolysis/Gluconeogenesis pathway for the three genomes in the case study.

In each pathway image (Fig. 3) the EC identifiers of the corresponding genome are highlighted according to the cluster their genes belong to. A special case is the EC identifiers highlighted in black; they contain genes from more than one cluster. However, it must be noted that each gene is assigned to a single cluster, whereas an EC identifier, corresponding to several genes, may in turn belong to different clusters.

Fig. 4 shows the distribution of the gene clusters across the pathway. It is interesting to note that Cluster 1 contains EC numbers corresponding to genes that constitute the main process of the pathway (core pathway), whereas the EC identifiers of the fourth cluster contain genes only from the human genome. The second cluster contains genes only from Homo Sapiens and Arabidopsis Thaliana, as opposed to Cluster 3 that contains genes only from Escherichia Coli and Arabidopsis Thaliana. Finally, there exist several cases where an EC identifier corresponds to genes that individually belong to different clusters. These cases are shown collectively as a fifth cluster, but are also highlighted in each cluster diagram when the EC identifier contains at list one gene of the specific cluster.



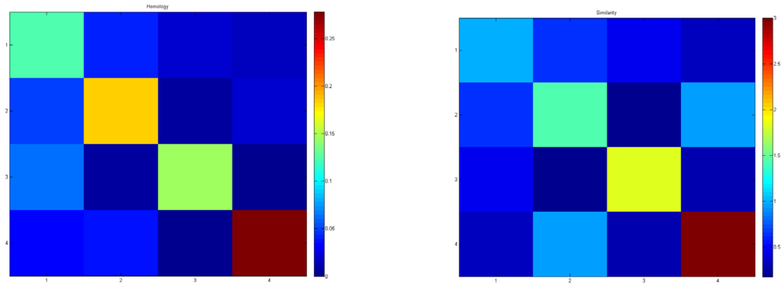
Cluster 1 Cluster 2 Cluster 3



Cluster 4 Multi cluster

Fig. 4. The Glycolysis/Gluconeogenesis pathway for each of the four produced clusters, and the case of EC identifiers with genes from multiple clusters (highlighted in black).

For validation purposes, the produced clusters were evaluated using both the modified jaccard similarity metric (Eq. 1) and the gene homology. The average intra-cluster similarity and homology (≈ 1.813 and 0.182, respectively) was significantly higher than the average inter-cluster similarity and homology (≈ 0.479 and 0.016 respectively). Moreover, every similarity and homology value calculated within a cluster is higher than every corresponding value calculated between the clusters (Fig. 5). The aforementioned results are indicative to the significance and validity of the produced clusters.



a. b.

Fig. 5. Inter-cluster and intra-cluster homology [a] and similarity [b] figure.

5 Discussion

Although these are only preliminary results, some interesting observations can be made. The presented test cases were among several different experimental setups. In all cases however, the first cluster always contained EC identifiers along the main reaction chain of the pathway, leading to the tentative conclusion that it may correspond to the highly conserved genes. Moreover, by superimposing the highlighted pathway diagrams along the implied phylogenetic distance of the genomes, one may infer sub-chains of the pathway that have been transformed or evolved across the species. A thorough investigation of this problem, together with rigorous experimentation on several different sets of pathways / genomes may provide more information in these areas.

6 References

1. M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes”, Nucleic Acids Research, Vol. 28 No. 1, pp. 27-30, 2000
2. Yong Zhang et al, “Phylophenetic properties of metabolic pathway topologies as revealed by global analysis”, BMC Bioinformatics, Vol 7 No 252, pp. 1-13, 2006
3. S. Schmid, S. Sunyaev, P. Bork and T. Dandekar, “Metabolites: a helping hand for pathway evolution?”, Trends in Biochemical Sciences, Vol. 28 No. 6, pp. 336-341, 2003
4. C.V. Forst, K. Schulten, “Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information”, Journal of Computational Biology, Vol. 6 No. 3-4, pp. 343-360, 1999
5. C.V. Forst, K. Schulten, “Phylogenetic analysis of metabolic pathways”, Journal of Molecular Evolution, Vol. 52 No. 6, pp. 471-489, 2001
6. M. Heymans, A.K. Singh, “Deriving phylogenetic trees from the similarity analysis of metabolic pathways”, Bioinformatics, pp. I138-I146, Vol. 19, Suppl 1, 2003
7. L. Liao, S. Kim, J.F. Tomb, “Genome Comparisons Based on Profiles of Metabolic Pathway”, in Sixth International Conference on Knowledge-Based Intelligent Information & Engineering Systems: 16–18 September 2002. Crema, Italy; 2002
8. S.H. Hong, T.Y. Kim, S.Y. Lee, “Phylogenetic analysis based on genome-scale metabolic pathway reaction content”, Applied Microbiology and Biotechnology, Vol. 65 No.2, pp. 203-210, 2004
9. D. Aguila, F.X. Aviles, E. Querol, M.J. Sternberg, “Analysis of phenetic trees based on metabolic capabilites across the three domains of life”, Journal of Molecular Biology, Vol. 340 No.3, pp. 491-512, 2004
10. F. PY Lin, E. Coiera, R. Lan, V. Sintchenko, "In silico prioritisation of candidate genes for prokaryotic gene function discovery: an application of phylogenetic profiles", BMC Bioinformatics 2009, 10:86 doi:10.1186/1471-2105-10-86
11. G. Gupta, A. Liu, J. Ghost, “Automated Hierarchical Density Shaving: A Robust Automated Clustering and Visualization Framework for Large Biological Data Sets”, IEEE Transactions on Computational Biology and Bioinformatics, Vol. 7, Issue 2, pp. 223-237, 2010
12. F. Psomopoulos, P. Mitkas, “Multi Level Clustering of Phylogenetic Profiles”, Proceedings of the IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2010), pp. 308-309, 2010
13. A. J. Enright, S. Van Dongen and C. A. Ouzounis, “An efficient algorithm for large-scale detection of protein families”, Nucleic Acids Research, Vol. 30 No. 7, pp. 1575-1584, 2002
14. A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, Series B (Methodological) Vol. 39 No. 1, pp. 1–38, 1977
15. Stijn van Dongen, *Graph Clustering by Flow Simulation*, PhD thesis, University of Utrecht, May 2000, http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm. http://micans.org/mcl.
16. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1. http://www.cs.waikato.ac.nz/ml/weka
17. Matlab Bioinformatics Toolbox, http://www.mathworks.com/products/bioinfo
18. Matlab MCR Runtime Engine: http://www.mathworks.com/products/compiler