Multi Level Clustering of Phylogenetic Profiles

Article *in* International Journal of Artificial Intelligence Tools · August 2011

DOI: 10.1142/S0218213012400234

2 authors:

Fotis E. Psomopoulos
Aristotle University of Thessaloniki
48 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Pericles A. Mitkas
Aristotle University of Thessaloniki
272 PUBLICATIONS **1,635** CITATIONS
SEE PROFILE

International Journal on Artificial Intelligence Tools
Vol. XX, No. X (2011) 1–16

World Scientific
www.worldscientific.com

MULTI LEVEL CLUSTERING OF PHYLOGENETIC PROFILES

FOTIS E. PSOMOPOULOS

Dept. Electrical and Computer Engineering,
Aristotle University of Thessaloniki,
Thessaloniki GR-54124, Greece
fpsom@issel.ee.auth.gr

PERICLES A. MITKAS

Dept. Electrical and Computer Engineering,
Aristotle University of Thessaloniki,
Thessaloniki GR-54124, Greece
mitkas@eng.auth.gr

Received (30 January 2011)
Revised (06 June 2011)
Accepted (26 August 2011)

The prediction of gene function from genome sequences is one of the main issues in Bioinformatics. Most computational approaches are based on the similarity between sequences to infer gene function. However, the availability of several fully sequenced genomes has enabled alternative approaches, such as phylogenetic profiles. Phylogenetic profiles are vectors which indicate the presence or absence of a gene in other genomes. The main concept of phylogenetic profiles is that proteins participating in a common structural complex or metabolic pathway are likely to evolve in a correlated fashion. In this paper, a multi level clustering algorithm of phylogenetic profiles is presented, which aims to detect inter- and intra-genome gene clusters.

Keywords: Algorithm; Clustering; Phylogenetic profiles.

1. Introduction

Phylogenetic profiles present the expression of genes, or homologs of genes, across a number of fully sequenced species[1]. They have emerged as a representation paradigm for use in defining and discovering functionally linked genes or metabolic pathways[2, 3, 4]. Recent developments in high throughput computational methods have established phylogenetic profiles as an intriguing representation of gene information due to their simple form on one hand, and a greater coverage of detail on the other[5]. A phylogenetic profile of a gene is classically represented as a vector of binary digits. Each value corresponds to the occurrence of orthologs or homologs of a gene in a particular genome; a value of 1 stands for the presence of a gene, whereas 0 shows the absence in a given genome. These values are transformed from BLAST[6] E-values by imposing a threshold for presence.

Phylogenetic profiles have been extensively used for various goals; gene annotation and network inference, gene prediction for orphan metabolic activities, or identification

of functionally linked proteins[7]. Here we propose a method for clustering phylogenetic profiles on multiple levels, in order to detect interesting relationships between genes of either the same or different species.

The availability of an ever increasing number of fully sequenced genomes has enabled the development of high-throughput algorithms and methodologies towards the detection of similarities and abnormalities of genes across species. There exists a plethora of algorithms and methods in data mining literature for gene clustering using their phylogenetic profiles. Some approaches are purely statistical in nature[8, 9], while others are based on pattern recognition techniques[10], kernel trees[11] or combinations of several methods[12, 5, 13]. It must be noted that the aforementioned methods are instances of clustering algorithms specifically tailored to utilize phylogenetic profiles. There exist several protein clustering approaches that use gene data other than phylogenetic profiles[14, 15].

2. Methodology

The problem definition can be stated as follows:

> *Given a set of n profiles p_i, i = 1, ..., n, produce a hierarchical set of clusters C, so that the bottom level clusters contain genes with at least t similarity, and each level subsequently increases this similarity threshold by t' (i.e. every level is a generalization of the previous one).*

The main concept of the algorithm is to produce a tree-like structure of clusters of phylogenetic profiles. Each level of the tree is constructed through application of a similarity measure on the instances of the previous level, continuously relaxing the threshold at the higher levels. Thus, the leaves of the tree (bottom level) consist of singleton clusters whereas the top level of the tree (i.e. the root of the tree) is a single "general" cluster containing all available instances.

Overall, the algorithm is an iterative process with each iteration consisting of five steps:

Step 1: Cluster initialization based on the available data instances, using hamming distance for assigning instances to clusters.
Step 2: Evaluate cluster centroids and merge clusters with 0 distance between the corresponding centroids.
Step 3: Eliminate multiple assignments of the same instance to different clusters, by selecting the most relevant cluster.
Step 4: Re-evaluate cluster centroids and merge clusters with 0 distance between the corresponding centroids.
Step 5: Consolidate final clusters and set the centroids as the data instances for the next iteration.

A visual diagram showing a typical application of the multi level clustering algorithm is presented in Figure 1.

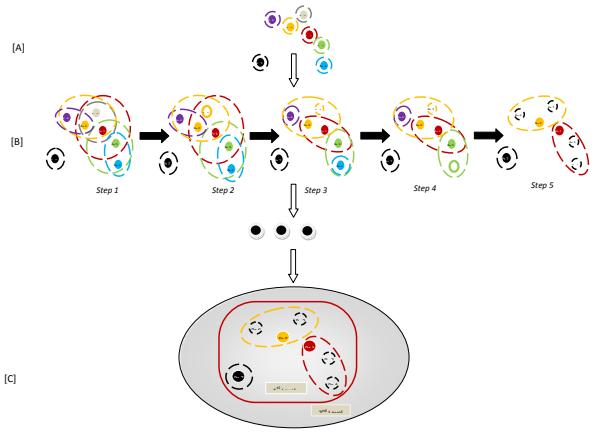

Fig. 1. Example of the Multi Level Clustering algorithm, including the initial singleton clusters (A), the five steps of the first iteration [B], and the final multi level output [C].

In the following subsections, the individual steps of the multi level algorithm will be described in more detail. In order to enhance the readers' understanding, a sample dataset of seven hypothetical phylogenetic profiles (shown in Table 1) produced across 6 hypothetical genomes will be used.

Table 1: Artificial dataset (seven phylogenetic profiles across six genomes)

	Genome #1	Genome #2	Genome #3	Genome #4	Genome #5	Genome #6
Protein 1 (*Pr1*)	1	1	1	0	0	0
Protein 2 (*Pr2*)	1	1	0	0	0	0
Protein 3 (*Pr3*)	1	0	1	1	0	0
Protein 4 (*Pr4*)	1	0	1	1	1	1
Protein 5 (*Pr5*)	0	1	0	0	0	0
Protein 6 (*Pr6*)	0	1	0	0	1	1
Protein 7 (*Pr7*)	0	0	0	1	1	0

2.1. *Step 1: Initial Cluster Construction*

During the first step of the algorithm, a number of clusters are constructed equal to the number of instances available. Obviously, in the first iteration the number of instances equals to number of the phylogenetic profiles. The instances are then assigned to the clusters using the hamming distance measure, which is defined as the number of different values in corresponding positions between two strings of the same length.

The clusters are created using the instances as centroids, and each cluster is labeled according to the label of the instance that created it (Fig. 2). The assignment of the instances in the clusters is achieved by an all-against-all calculation of the hamming distance. In order for an instance to be assigned to a specific cluster, the hamming

distance between the instance and the corresponding centroid should be less or equal to a threshold *t*.

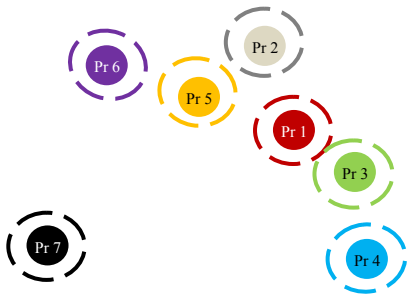

Fig. 2. Cluster initialization. Each cluster is a singleton.

Using the example dataset (Table 1) and a threshold $t = 2$, the hamming distances are the following (Table 2):

Table 2: All-against-all Hamming distances of the artificial dataset. Highlighted are the distances which are less or equal to the threshold $t = 2$.

	Pr1	Pr2	Pr3	Pr4	Pr5	Pr6	Pr7
Pr1	0	1	2	4	2	4	5
Pr2	1	0	3	5	1	3	4
Pr3	2	3	0	2	4	6	3
Pr4	4	5	2	0	6	4	3
Pr5	2	1	4	6	0	2	3
Pr6	4	3	6	4	2	0	3
Pr7	5	4	3	3	3	3	0

Obviously, cases of overlapping clusters are to be expected, a fact evident in the example distances of Table 2. Based on these distances, the initial clusters are created as presented in Figure 3.

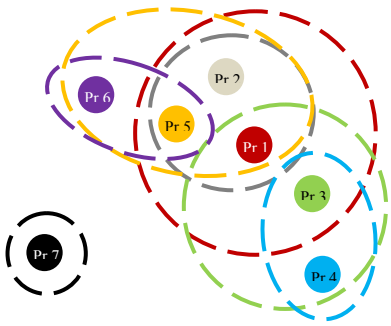

Fig. 3. Initial cluster creation. The distance between the sample instances does not correspond to the actual distances.

2.2. *Step 2: Evaluation of the Cluster Centroids*

The next step in the process is the calculation of the cluster centroids, according to the following equation:

$$C_i = \left[\frac{\sum_{j=1}^{n} p_{i,j}}{n} \right] \quad (1)$$

where i is the cluster ID, n is the number of instances (phylogenetic profiles) in the cluster i. The notation $p_{i,1}$ denotes the first phylogenetic profile that belongs in cluster i. The centroids are subsequently compared using the hamming distance, and the clusters whose corresponding centroids have zero hamming distance are merged. During merging, the largest cluster retains its label, whereas the other clusters are simply absorbed into the larger one.

Using the artificial dataset (Table 1) and the initial clusters constructed at the end of Step 1 (Figure 3), the cluster centroids are the following:

Table 3: Centroids of the 7 initial clusters, produced at the end of Step 1. Each cluster is identified by the label of the instance it was initialized by.

	#1	#2	#3	#4	#5	#6
Pr1	1	1	1	0	0	0
Pr2	1	1	0	0	0	0
Pr3	1	0	1	1	0	0
Pr4	1	0	1	1	1	1
Pr5	1	1	0	0	0	0
Pr6	0	1	0	0	1	1
Pr7	0	0	0	1	1	0

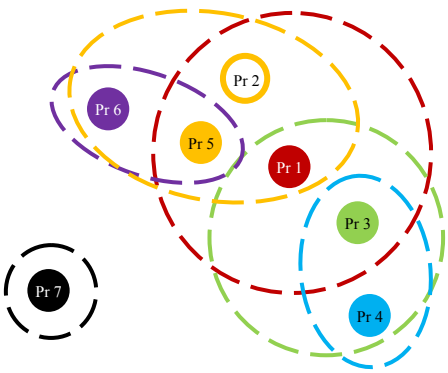

Fig. 4.Clusters at the end of Step 2. Cluster “Pr2” has been merged into cluster “Pr5”. Graphically, the cluster “Pr2” line has been removed, and the corresponding instance has been de-colored.

From Table 3 it is evident that only two clusters have identical centroids, namely clusters "*Pr2*" and "*Pr5*". This means that the two clusters will be merged, and specifically "*Pr2*" will be absorbed into "*Pr5*" which is the largest of the two. The resulting clusters at the end of Step 2 are depicted in Figure 4.

2.3. *Step 3: Removal of multiple instance assignments*

The main goal of this step is to ensure that each instance may be assigned to at most two clusters, and, if that is the case, then one of the two clusters must be the cluster that was initialized by the specific instance. In other words, this step aims to detect the strongest cluster assignment for each instance (except for the cluster it has initialized), and to remove all other assignments.

The cluster assignment strength is evaluated using the unity factor, $U(p_i,k)$, as defined in Equation 2

$$U(p_i,k) = \frac{h(p_i,C_k)}{\tilde{h}(k)} \tag{2}$$

where $h(p_i, C_k)$ is the hamming distance of instance p_i from the centroid C_k of cluster k, and $\tilde{h}(k)$ is the mean of the distances of the other members in cluster k, and is defined as follows:

$$\tilde{h}(k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} h(p_{k,i},p_{k,j})}{n\cdot(n-1)} \tag{3}$$

Using the artificial dataset, the unity factor values are presented in Table 4.

Table 4: The values of the unity factor for the artificial dataset. The highlighted cells correspond to the cluster assignments that are going to be deleted at the end of Step 3.

	$\tilde{h}(k)$	$U(p_1,k)$	$U(p_2,k)$	$U(p_3,k)$	$U(p_4,k)$	$U(p_5,k)$	$U(p_6,k)$
Pr1	2.16	0.00	0.46	0.92	-	0.92	-
Pr3	2.66	0.75	-	0.00	0.75	-	-
Pr4	2.00	-	-	1.00	0.00	-	-
Pr5	2.16	0.46	0.00	-	-	0.00	0.92
Pr6	2.00	-	-	-	-	1.00	0.00

The final clusters at the end of Step 3 using the artificial dataset are presented in Figure 5. Clusters "*Pr5*" and "*Pr7*" have not been changed, whereas cluster "*Pr1*" has lost pr2, cluster "*Pr3*" has lost instance pr1, cluster "*Pr4*" has lost pr3, and cluster "*Pr6*" has lost pr5.

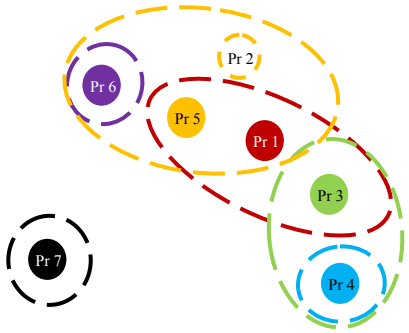

Fig. 5. Clusters at the end of Step 3. All instances now belong to at most two clusters.

2.4. *Step 4: Re-evaluation of the cluster centroids*

The next step is in essence the same as Step 2; the cluster centroids are calculated, and the clusters whose centroids have zero hamming distance are merged. Using the artificial dataset, the new cluster centroids are presented in Table 5.

Table 5: Centroids of the 6 clusters, produced at the end of Step 3. Highlighted are the clusters whose centroids have zero hamming distance.

	#1	#2	#3	#4	#5	#6
Pr1	1	1	1	0	0	0
Pr3	1	0	1	1	1	1
Pr4	1	0	1	1	1	1
Pr5	1	1	0	0	0	0
Pr6	0	1	0	0	1	1
Pr7	0	0	0	1	1	0

Cluster "*Pr4*" is smaller than "*Pr3*", and therefore during merging "*Pr4*" will be absorbed into "*Pr3*". The clusters as produced at the end of Step 4 can be seen in Figure 6.

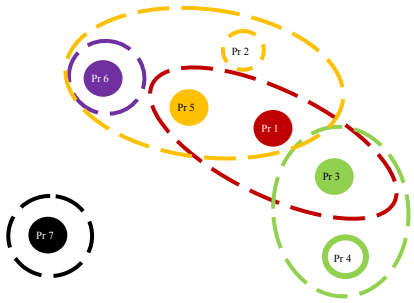

Fig. 6. Clusters at the end of Step 4. All instances now belong to at most two clusters.

2.5. *Step 5: Consolidation of the final clusters*

The final step in the iterative process comprises the detection of any cluster overlapping and the assignment of the instances to the cluster with the strongest correlation. Due to Step 4, in any case of cluster overlapping the instances may belong to at most two clusters, one of which is the cluster initialized by the specific instance. The selection of the final assignment is performed by calculating the hamming distance of the specific instance to the two centroids of the corresponding clusters.

Using the artificial dataset, the final clusters are presented in Figure 7.

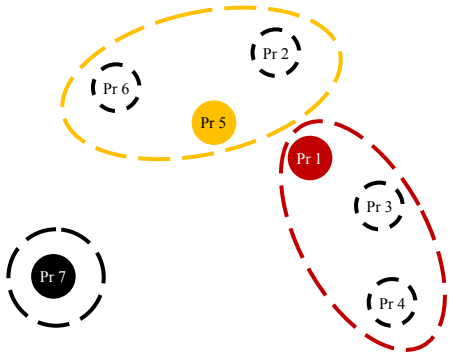

Fig. 7. Final clusters. Each instance belongs now to a single cluster.

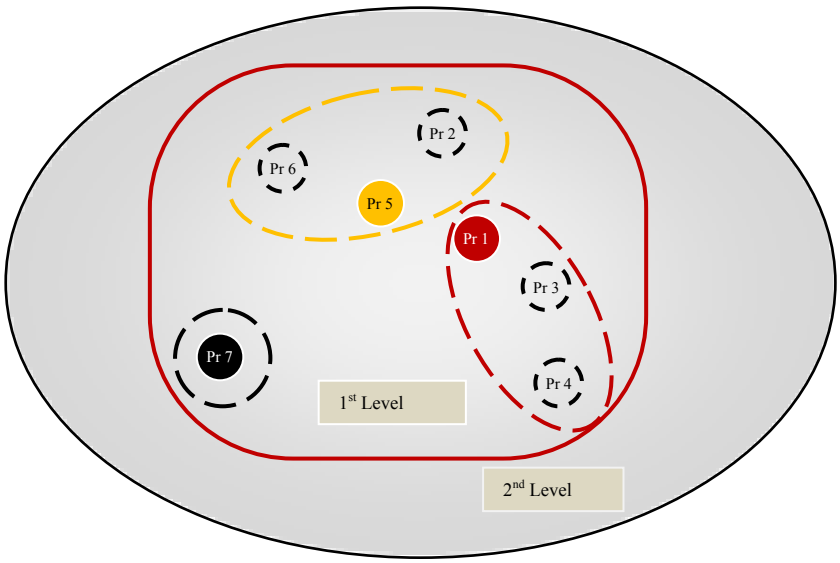

Fig. 8. Final output of the multi level clustering algorithm. For our simplified example, the result contains only two clustering levels.

2.6. *Algorithm Iteration*

The five steps described before are iterated, until all instances are assigned to a single cluster. The cluster centroids are used as input in each iteration, together with a second parameter, *step*, used for the relaxation of the similarity (distance) threshold *t*. The parameter *step* denotes the increase in percentage of the threshold *t* in each iteration. For example, *step* = 100 means that the threshold *t* will be doubled in each iteration of the algorithm.

Using the artificial dataset, and setting parameter *step* = 100, the final output of the multi level algorithm is presented in Figure 8.

3. Experiments and Results

In order to evaluate the algorithm, data from the ProfUse database[16] have been used. The database contains 915,554 phylogenetic profiles of genes from 243 different species. This means that each profile p_i is a binary vector of 243 bits.

For our experiments, a subset of the data has been used, which contains the phylogenetic profiles of 3,896 genes, from 5 different species (Table 6). The species were selected both for their small number of genes and for allowing the best possible representation of the whole database. Specifically, "*Mycoplasma genitalium, G-37*" was selected as the base genome with the smallest number of genes in the database. The next species selected were the smallest possible but in the same family ("*Ureaplasma urealyticum, serovar 3*"), in the same phylum ("*Streptococcus pyogenes M1, SF370*"), in a different phylum ("*Buchnera aphidicola, SG*"), and finally in a different kingdom ("*Nanoarchaeum equitans, Kin4-M*").

Table 6: Input Genomes

Genome ID	Number of Genes	Percentage of Dataset (%)	Relative Position ir profile
MGEN-G37-01	479	12.29	2
UURE-SV3-01	613	15.74	39
SPYO-SF3-01	1696	43.53	50
BAPH-XSG-01	545	13.99	88
NEQU-N4M-01	563	14.45	148

A number of experiments were performed with different parameters. In this paper we show the results obtained with similarity initially set to $t = 5$, and subsequently increased by 80% in each iteration (*step* = 80). Using these parameters, the algorithm produced an 8-level clustering tree (not including the top level), as shown in Figure 9. The horizontal axis corresponds to the 3,896 genes of the dataset, and the vertical axis corresponds to the different levels (iterations) in the clustering process. Each cluster is designated a different

color, and clusters that remain after each iteration (i.e. clusters that have emerged as "parent" clusters) retain their color. Genes are arranged in alphabetical order of the genome names; genes 1 to 545 belong to BAPH-XSG-01, genes from 546 to 1024 belong to MGENG37-01, genes from 1025 to 1587 belong to NEQU-N4M-01, genes from 1588 to 3283 belong to SPYO-SF3-01, and finally genes from 3284 to 3896 belong to UURE-SV3-01.

Applying the Multi-Level Clustering algorithm, the dataset produced five clusters at the 2nd level of the cluster tree. However, two of the five clusters (Fig. 10) contained only 1.46% of the whole dataset, which is peculiar given the highly flexible distance measure at that level (similarity ≥ 31%).

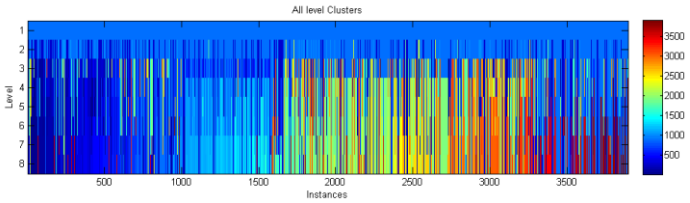

Fig. 9. Clusters of all different levels. Each cluster is designated a distinct color and the genes retain the order of appearance

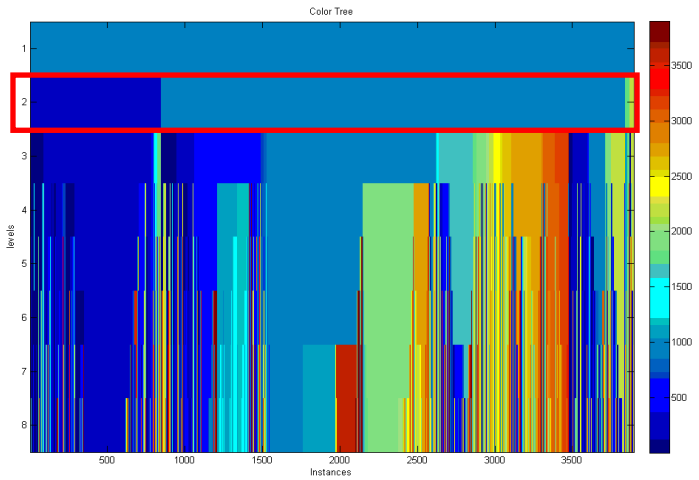

Fig. 10. Clusters of all different levels. Each cluster is assigned a distinct color, but the genes are re-ordered at each level in order to group instances of the same cluster. The five clusters of the second level are marked by a red horizontal bar.

3.1. *Second Level Clusters*

The clusters of the 2nd level present an interesting distribution of instances. At this level, the threshold *t* is equal to 167, which, one would expect, should yield a small number of large clusters. However, as shown in Table 7, out of the five clusters, two contain less than 30 instances each.

Table 7: 2nd Level Clusters' content

Cluster based on protein:	Number of Genes	Percentage of Dataset (%)
BAPH-XSG-01-000264	845	21.7
MGEN-G37-01-000352	2632	67.6
MGEN-G37-01-000444	362	9.29
SPYO-SF3-01-000233	29	0.7
SPYO-SF3-01-000613	28	0.7

Moreover, by comparing the distribution of the instances in each of the five clusters across the five genomes that have been used in the experiment, it is obvious that the two small clusters have concentrated proteins (instances) mainly from a single genome (Table 8).

Table 8: 2nd Level Clusters' distribution across genomes

Cluster based on protein:	BAPH-XSG-01	MGEN-G37-01	NEQU-N4M-01	SPYO-SF3-01	UURE-SV3-01
BAPH-XSG-01-000264	24.73 %	14.56 %	10.77 %	36.33 %	13.61 %
MGEN-G37-01-000352	08.43 %	11.74 %	17.55 %	45.36 %	16.91 %
MGEN-G37-01-000444	30.39 %	12.71 %	00.55 %	42.82 %	13.54 %
SPYO-SF3-01-000233	00.00 %	00.00 %	27.59 %	62.07 %	10.34 %
SPYO-SF3-01-000613	14.29 %	3.57 %	00.00 %	78.57 %	03.57 %

The first three clusters (BAPH-XSG-01-000264, MGEN-G37-01-000352 and MGEN-G37-01-000444) contain the 98.59% of the dataset. However, out of the three clusters, only the first (BAPH-XSG-01-000264) exhibits an almost uniform distribution of the protein-instances across the species, leading to the conclusion that it may contain proteins that perform common functions across all species.

The next two clusters (MGEN-G37-01-000352 and MGEN-G37-01-000444) both show an increased content of genes from the SPYO-SF3-01 genome, but differ at the content of genes from the BAPH-XSG-01 and the NEQU-N4M-01. Given the fact that NEQU-N4M-01 is the genome with the greatest phylogenetic distance from the other 4 species, a fair explanation of these two clusters is that the second (MGEN-G37-01-000444) contains proteins common in the kingdom of the four species (Bacteria), whereas the first contains proteins that are mainly present in the Archaea kingdom and may have homologs in the other kingdoms.

Finally, two small clusters are especially interesting due to the size of the clusters regards to the highly flexible distance measure at the specific level. The small size of the clusters allowed a more thorough investigation of the properties of the proteins involved (a detailed listing of the protein NCBI identifiers is presented in Appendix A). Some preliminary results indicate that both clusters contain highly specialized proteins. The first cluster (SPYO-SF3-01-000233) contains proteins that are ABC-Transporters and are involved in Environmental Information Processing, Membrane Transport and in the Bacterial secretion system. The second cluster (SPYO-SF3-01-000613) contains proteins that are involved in the amino-acid metabolism systems, and specifically in the glycine, serine, threonine, cysteine, and methionine metabolism. However, in any case, both these groups require further study.

3.2. *Cluster Evaluation*

One of the major issues in unsupervised machine learning approaches is the evaluation of the final output. Since the data instances given as input are unlabeled, there is no error or reward signal to evaluate a potential solution. In the case of the multi-level clustering of phylogenetic profiles, there are two approaches towards the evaluation of the produced clusters; the study of the biological relationship among members of the same cluster, which is the most time-consuming but yields the most accurate evaluation, and the use of an arbitrary quantitative metric for the estimation of the intra- and inter-cluster cohesiveness.

The complete implementation of the first approach, i.e. the evaluation of the clusters via the biological information of their members, is beyond the scope of the current paper. However, the two small clusters produced in the second level of the cluster tree can be used as an indicative measure of the biological evaluation. The results of this process are presented in section 3.1.

In order to evaluate the intra- / inter-cluster cohesiveness, the following metric has been utilized:

$$Coh_{I,J} = avg\left(dist(p_i, p_j)\right), p_i \in C_I, p_j \in C_J \tag{4}$$

where the distance measure is jaccard distance.

The metric in Eq. 4 produces an *n* x *n* matrix, where *n* is the number of clusters of the specific level. In each level, the intra-cluster distance is lower than the inter-cluster distance, thus validating the clustering process. The specific results of the cohesiveness metric for the 2nd level clusters are presented in Table 9 and Figure 11.

4. Discussion and Conclusions

The method presented here is experimentally proven to be consistent with the phylogenetic relation and position of the genes involved. By introducing the notion of a hierarchy in clustering phylogenetic profiles, the algorithm succeeds in identifying both

the strong correlation among genes of the same genome, as well as the underlying signal that relates genes from across different species.

Table 9: 2nd Level Clusters' intra- / inter- distances. The minimum values in each row and column are highlighted and correspond to the intra-cluster cohesiveness.

Cluster IDs	BAPH-XSG-01-000264	MGEN-G37-01-000352	MGEN-G37-01-000444	SPYO-SF3-01-000233	SPYO-SF3-01-000613
BAPH-XSG-01-000264	0.20	0.78	0.28	0.63	0.40
MGEN-G37-01-000352	0.78	0.72	0.77	0.84	0.77
MGEN-G37-01-000444	0.28	0.77	0.27	0.63	0.37
SPYO-SF3-01-000233	0.63	0.84	0.63	0.54	0.62
SPYO-SF3-01-000613	0.40	0.77	0.37	0.62	0.31

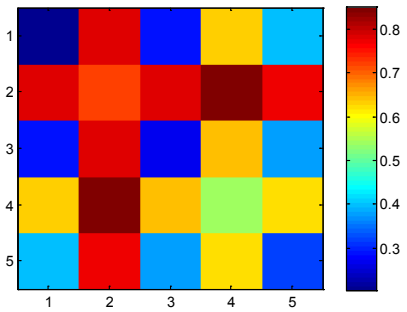

Fig. 11. Intra- / Inter-cluster distances for the 2nd level clusters. The clusters are numbered sequentially in the order of appearance in Table 9.

There exist several approaches towards the clustering of phylogenetic profiles, especially in the computational biology literature. The majority of the cases either address the issue of how to measure the "similarity" between two profiles, in an evolutionary relevant way, in order to develop efficient function prediction methods [17], or focus the research problem into a specific genome (such as E.coli) in order to assess functionally related proteins [18]. Moreover, there exist a plethora of applications that enhance the researchers' ability to quickly visualize and retrieve user-defined subsets of phylogenetic profiles, without any significant processing of the actual data [19] [20]. Finally, there are several approaches that perform the "opposite" process of the multi-level clustering algorithm; using a specific phylogenetic profile as a seed, the algorithm iteratively includes similar profiles and expands the search including the newly added profiles, thus constructing tight clusters of specific gene vectors [21]. However, in most cases the presented approach utilizes additional information, such as metabolic pathway membership, or constrains the process into specific genomes.

The algorithm presented here addresses the problem of gene clustering in a less deterministic way and introduces the notion of levels in gene clustering. This method serves as a general computational tool for the annotation of large numbers of genes by

highlighting evolutionary and functional patterns. The experiments showed that this method spots distinct patterns based on inter and intra-genomic signals. The outcome is a multilevel gene clustering, which attempts to capture at each level the different aspects of the affinity of a protein with another, in the same or in a different species.

Appendix A. Listing of the proteins members in clusters SPYO-SF3-01-000233 and SPYO-SF3-01-000613

In order to provide more insight into the specific details regarding the two small clusters derived at the 2nd level of the experiment, the following tables (Tables 9 and 10) contain the COGENT and NCBI protein identifiers.

Table 9: Members of the SPYO-SF3-01-000233 cluster (29 members)

COGENT Protein Identifier	NCBI Gene Identifier
NEQU-N4M-01-000018	NP_963311
NEQU-N4M-01-000020	NP_963315
NEQU-N4M-01-000022	NP_963317
NEQU-N4M-01-000091	NP_963383
NEQU-N4M-01-000151	NP_963442
NEQU-N4M-01-000351	NP_963630
NEQU-N4M-01-000380	NP_963659
NEQU-N4M-01-000447	NP_963720
SPYO-SF3-01-000233	NP_268657
SPYO-SF3-01-000235	NP_268659
SPYO-SF3-01-000237	NP_268661
SPYO-SF3-01-000426	NP_268850
SPYO-SF3-01-000427	NP_268851
SPYO-SF3-01-000667	NP_269091
SPYO-SF3-01-000812	NP_269236
SPYO-SF3-01-000865	NP_269289
SPYO-SF3-01-001203	NP_269627
SPYO-SF3-01-001204	NP_269628
SPYO-SF3-01-001205	NP_269629
SPYO-SF3-01-001206	NP_269630
SPYO-SF3-01-001209	NP_269633
SPYO-SF3-01-001245	NP_269669
SPYO-SF3-01-001325	NP_269749
SPYO-SF3-01-001335	NP_269759
SPYO-SF3-01-001467	NP_269891
SPYO-SF3-01-001469	NP_269893
UURE-SV3-01-000096	NP_077926
UURE-SV3-01-000099	NP_077929
UURE-SV3-01-000266	NP_078096

Table 10: Members of the SPYO-SF3-01-000613 cluster (28 members)

COGENT Protein Identifier	NCBI Gene Identifier
BAPH-XSG-01-000029	NP_660390
BAPH-XSG-01-000058	NP_660419
BAPH-XSG-01-000297	NP_660658
BAPH-XSG-01-000505	NP_660866
MGEN-G37-01-000117	NP_072777
SPYO-SF3-01-000090	NP_268514
SPYO-SF3-01-000139	NP_268563
SPYO-SF3-01-000142	NP_268566
SPYO-SF3-01-000207	NP_268631
SPYO-SF3-01-000245	NP_268669
SPYO-SF3-01-000261	NP_268685
SPYO-SF3-01-000338	NP_268762
SPYO-SF3-01-000461	NP_268885
SPYO-SF3-01-000462	NP_268886
SPYO-SF3-01-000529	NP_268953
SPYO-SF3-01-000613	NP_269037
SPYO-SF3-01-000670	NP_269094
SPYO-SF3-01-000845	NP_269269
SPYO-SF3-01-000861	NP_269285
SPYO-SF3-01-000950	NP_269374
SPYO-SF3-01-000955	NP_269379
SPYO-SF3-01-001059	NP_269483
SPYO-SF3-01-001242	NP_269666
SPYO-SF3-01-001259	NP_269683
SPYO-SF3-01-001435	NP_269859
SPYO-SF3-01-001443	NP_269867
SPYO-SF3-01-001674	NP_270098
UURE-SV3-01-000083	NP_077913

References

1. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O.Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4285–4288, 1999.
2. L. Chen and D. Vitkup, "Predicting genes for orphan metabolic activities using phylogenetic profiles", *Genome Biology*, vol. 7, no. R17, 2006, doi:10.1186/gb-2006-7-2-r17.
3. S. Cokus, S. Mizutani, and M. Pellegrini, "An improved method for identifying functionally linked proteins using phylogenetic profiles", *BMC Bioinformatics*, vol. 8, no. S7, 2007, doi:10.1186/1471-2105-8-S4-S7.
4. V. Kunin, D. Ahren, L. Goldovsky, P. Janssen, and C. A. Ouzounis, "Measuring genome conservation across taxa: divided strains and united kingdoms", *Nucleic Acids Research*, vol. 33, no. 2, pp. 616–621, 2005, doi:10.1093/nar/gki181.
5. E. S. Snitkin, A. M. Gustafson, J. Mellor, J. Wu, and C. DeLisi, "Comparative assessment of performance and genome dependence among phylogenetic profiling methods", *BMC Bioinformatics*, vol. 7, no. 420, 2006, doi: 10.1186/1471-2105-7-420.
6. S. F. Altschul et al., "Gapped blast and psi-blast: a new generation of protein database search programs", *Nucleic Acids Research*, vol. 25, pp. 3389–3402, 1997.
7. A. Stamatakis, "Parallel and distributed computation of large phylogenetic trees", in *Parallel Computing for Bioinformatics and Computational Biology*, Wiley Series on Parallel and

Distributed Computing, ed. A. Zomaya, (Wiley-Interscience, New Jersey, USA, 2006), pp. 327–346.

8. T. Sato, Y. Yamanishi, M. Kanehisa, and H. Toh, "Prediction of protein-protein interactions based on real-valued phylogenetic profiles using partial correlation coefficient", in *Poster Abstracts of the Genome Information Conference 2004*, 2004, p. 122.
9. J. Wu, Z. Hu, and C. DeLisi, "Gene annotation and network inference by phylogenetic profiling", *BMC Bioinformatics*, vol. 7, no. 80, 2006, doi:10.1186/1471-2105-7-80.
10. A. L. Cuff, I. Sillitoe, T. Lewis, O. C. Redfern, R. Garratt, J. Thornton, and C. A. Orengo, "The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies", *Nucleic Acids Research*, vol. 37, pp. 310–314, 2009.
11. J.-P. Vert, "A tree kernel to analyse phylogenetic profiles", *Bioinformatics*, vol. 18, pp. 276–284, 2002.
12. J. Wu, S. Kasif, and C. DeLisi, "Identification of functional links between genes using phylogenetic profiles", *Bioinformatics*, vol. 19, no. 12, pp. 1524–1530, 2003.
13. R. Jothi, T. M. Przytycka, and L. Aravind, "Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment", *BMC Bioinformatics*, vol. 8, no. 173, 2007.
14. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events", *Nature*, vol. 402, pp. 86–90, 1999.
15. Y. Chen, K. Reilly, A. Sprague, and Z. Guan, "Seqoptics: a protein sequence clustering system", *BMC Bioinformatics*, vol. 7, 2006.
16. L. Goldovsky, P. Janssen, D. Ahr´en, B. Audit, I. Cases, N. Darzentas, A. J. Enright, N. L´opez-Bigas, J. M. Peregrin-Alvarez, M. Smith, S. Tsoka, V. Kunin, and C. A. Ouzounis, "Cogent++: an extensive and extensible data environment for computational genomics", *Bioinformatics*, vol. 21, no. 19, pp. 3806 – 3810, 2005, doi:10.1093/bioinformatics/bti579.
17. Jean-Philippe Vert, "A tree kernel to analyse phylogenetic profiles", *Bioinformatics*, vol. 18, pp. S276 - S284, 2002
18. M. A. Pyatnitskiy, A. V. Lisitsa, A. I. Archakov, "Comparison of Algorithms for Prediction of Related Proteins Using the Method of Phylogenetic Profiles", *Proteomics and Bioinformatics, Biochemistry Supplement Series B: Biomedical Chemistry*, Vol. 4, No. 1, pp. 42 – 48, 2010.
19. Miklos Csuros, "Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood", *Bioinformatics*, Vol. 26, No. 15, pp. 1910 – 1912, 2010.
20. Xuejian Xiong et. al., "PhyloPro: a web-based tool for the generation and visualization of phylogenic profiles across Eukarya", *Bioinformatics*, Vol. 27, No. 6, pp. 877 – 878, 2011.
21. Galina Glazko, Michael Coleman, Arcady Mushegina, "Similarity searches in genome-wide numerical data sets", *Biology Direct*, Vol. 1, No. 13, 2006, doi: 10.1186/1745-6150-1-13