A Hierarchical Clustering Strategy and Its Application to Proteomic Interaction Data

Vicente Arnau[1] and Ignacio Marín[2]

[1] Departamento de Informática, Universidad de Valencia, Campus de Burjassot
Avda. Vicent Andrés Estellés, s/n. 46100 Burjassot, Valencia, Spain
vicente.arnau@uv.es

[2] Departamento de Genética, Universidad de Valencia, Campus de Burjassot
Calle Doctor Moliner, 50. 46100 Burjassot, Valencia, Spain
ignacio.marin@uv.es

Abstract. We describe a novel strategy of hierarchical clustering analysis, particularly useful to analyze proteomic interaction data. The logic behind this method is to use the information for all interactions among the elements of a set to evaluate the strength of the interaction of each pair of elements. Our procedure allows the characterization of protein complexes starting with partial data and the detection of "promiscuous" proteins that bias the results, generating false positive data. We demonstrate the usefulness of our strategy by analyzing a real case that involves 137 *Saccharomyces cerevisiae* proteins. Because most functional studies require the evaluation of similar data sets, our method has a wide range of applications and thus it can be established as a benchmark analysis for proteomic data[1].

1 Introduction

When we can define a distance measure among elements of a set, hierarchical clustering techniques are often very useful to define "natural" groups within that set [4]. However, the ability of such methods to obtain reasonable classifications depend on how are the distances among the elements. For example, when many pairs of elements are at the same distance, it is often impossible to unambiguously define the groups. This problem arises in many cases, as in the characterization of nets of irregular topology, in which distances are generally constrained to values between 1 and 5 [1]. The available data on protein-protein interactions generated in massive proteomic analyses [5-7, 13] can be similarly converted into distances, that measure the degree of metabolic or functional proximity within the cell. Again, those distances are constrained. For both prokaryotic and eukaryotic organisms, it has been found that

[1] This paper is supported by CICYT (Grant No. TIC2000-1151-C07-04), by Fundació *La Caixa* (Grant No. 01/080-00) and Generalitat Valencia (Gran. No. BM-011/2002). It is also part of the *NEUROGENOMICA* project, supported by an MCYT grant (GEN2001-4851-C06-02).

F.J. Perales et al. (Eds.): IbPRIA 2003, LNCS 2652, pp. 62-69, 2003.

distances have very often low values, suggesting that the cellular protein interaction network has "small world" properties, with a high degree of connectivity and closeness among components [8, 11].

It is therefore very interesting to generate methods able to deal with those difficult cases. In this work, we describe a fast, iterative hierarchical clustering algorithm that uses the information provided by the whole database of distances among elements of a set (that we will call from now on as *primary distances, d*) to evaluate the closeness of two particular elements. The algorithm converts the primary distances between two elements into *secondary distances (d')* that reflect the strength of the connection between two elements *relative to all the other elements in the set*. Those secondary distances can then be used again to perform a hierarchical clustering analysis.

In the following section, we will detail the new algorithm and we will show its properties by analyzing a simple case. Then, we will describe the results when the method is applied to a real case (a complex set of 137 interacting proteins of the baker's yeast *Saccharomyces cerevisiae*). The last section contains some concluding remarks about the advantages of this strategy.

2 A New Hierarchical Clustering Strategy

We start by defining the parameters used to perform a typical hierarchical clustering strategy (see also [10]). Let us consider a set of N elements. For each pair of elements, we have determined a distance value, that we will call *primary distance (d)*. Let us now establish in that set a partition P, formed by M clusters (A_1, A_2, …, A_M). Each cluster A_i contains x_i elements (a_1, a_2, …, a_{xi}). We can define then a cluster function for A_i (F[A_i]) as follows:

$$F(A_i) = \sum_{k=1}^{x_i-1} \sum_{j=k+1}^{x_i} d_{a_k a_j} \tag{1}$$

where d_{ij} is the primary distance between element a_i and element a_j. The number of primary distances within this cluster is:

$$n(A_i) = x_i (x_i - 1) / 2 \tag{2}$$

Similarly, we can define a function for the whole partition (F[P]) , that includes the distances among all elements:

$$F(P) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} \tag{3}$$

The number of primary distances for the whole partition is:

$$n(P) = N (N - 1) / 2 \tag{4}$$

We can then define a global function (F[G]) that evaluates, once the clusters have been established, the average of the distances for pairs of elements included in the clusters respect to the average value of distances in the whole partition:

$$F(G) = \left[\frac{\sum_{i=1}^{M} F_{A_i}}{\sum_{i=1}^{M} \frac{x_i(x_i-1)}{2}} \right] / \left[F(P)/n(P) \right] \tag{5}$$

This F(G) value is minimum when the clustering obtained is optimal. Therefore, the problem to solve is to minimize the value of F(G) for a certain set of elements. A typical algorithm of hierarchical clustering is developed in [2]. Starting with N elements, a maximum number of N clusters are established. An F(A_i) value equal to zero is assigned to all single-element clusters (i. e. intraelement distances are zero). Then, the best grouping with N – 1 clusters is determined by examining all possible combinations among the N elements and putting together the two elements that have a minimum distance (equivalent to minimizing F[G] for that particular number of clusters). This procedure can be repeated for N-2, N-3, ..., up to 1 clusters. It is significant that the way that the F(G) values change every time a cluster is eliminated provides a hint of the quality of the clustering. When a large increment is obtained for the F(G) value when we pass from X to X - 1 clusters, we can conclude that the grouping is becoming artificial, i. e. is putting together elements that are too dissimilar for the clustering to be meaningful [3].

Let us consider now the situation when there are many identical primary distances between pairs of elements. This situation causes the additional problem that there are many identically optimal (i. e. with identical F[G] values), but totally unrelated solutions, both when the same or different numbers of clusters are established. A typical example will clearly show how this additional difficulty complicates the clustering procedure. In Table 1, we show a table of distances, generated for illustrative purposes.

In the set shown in Table 1, there are 8 elements, named A to H, and all primary distances have values ranging from 1 to 5. Thus, many of these distances are identical. When we apply the typical clustering strategy described above, we will find that several independent solutions, obtained by connecting elements that are separated by a distance equal to 1, yield identical, optimal F(G) values. Using the data in Table 1, if we make 20 hierarchical clusterings, we obtain four solutions with identical values of F(G) (Table 2, left).

Table 1. Matrix of distances among eight elements (A – H). The distances are constraines to values between 1 and 5

	A	B	C	D	E	F	G	H
A	-	1	1	2	3	4	5	5
B	1	-	1	1	2	3	4	5
C	1	1	-	2	3	4	5	5
D	2	1	2	-	1	2	3	4
E	3	2	3	1	-	1	3	2
F	4	3	4	2	1	-	1	1
G	5	4	5	3	3	1	-	1
H	5	5	5	4	2	1	1	-

Table 2. Four optimal solutions found using Table 1 distances

Optimal clusterings	No. of times found
(A, B, C) (D, E) (F, G, H)	15
(A, C) (B, D) (E, F) (G, H)	2
(A, C) (B, D) (E) (F, G, H)	2
(A, B, C) (D) (E, F) (G, H)	1

Table 3. Secondary distances among the eight elements analyzed

	A	B	C	D	E	F	G	H
A	-	5	1	21	21	21	21	21
B	5	-	5	17	21	21	21	21
C	1	5	-	21	21	21	21	21
D	21	17	21	-	6	21	21	21
E	21	21	21	6	-	18	21	21
F	21	21	21	21	18	-	4	4
G	21	21	21	21	21	4	-	1
H	21	21	21	21	21	4	1	-

The results of the multiple replicates can be used to evaluate the strength of the connection between two elements respect to the connections among all the elements in the partition. For example, if we apply the clustering algorithm 20 times, it is found that the four solutions are generated with different frequencies. One of the solutions is found in 75% of the analyzed cases (Table 2, right). Moreover, connections between particular pairs of elements occur in several final solutions (e. g. elements A and C are together in all 20 solutions shown in Table 2). Thus, the strength of the connection between two elements, respect to the whole set, can be evaluated by considering the number of times those two elements are found together in all alternative solutions and the frequency of each alternative solution. Thus, a new table of *secondary distances* (*d'*) can be generated that contains the number of times that each pair of elements appear together for a large and randomly generated set of alternative optimal solutions. In our example, these secondary distances are shown in Table 3. This secondary distances are simply calculated as the number of times two elements do not appear together plus one. Thus, in our case, all elements that never appear together have a secondary distance of 21 and all those elements that go always together have a secondary distance of 1 (Table 3).

Table 4. Optimal clustering using secondary distances

(A, C) (B) (D) (E) (F) (G) (H)	F(G) = 0.06086957
(A, C) (B) (D) (E) (F) (G, H)	F(G) = 0.06086957
(A, C) (B) (D) (E) (F, G, H)	F(G) = 0.15217391
(A, C, B) (D) (E) (F, G, H)	F(G) = 0.20289855
(A, C, B) (D, E) (F, G, H)	F(G) = 0.22608696
(A, C, B, D, E) (F, G, H)	F(G) = 0.69297659
(A, C, B, D, E, F, G, H)	F(G) = 1

Once these secondary distances are established, we can now use them to make a new cluster analysis. As an example, we show, in Table 4, the groups obtained by taking the secondary distances shown in Table 3 and using the heuristic hierarchical clustering algorithm described above.

In Table 4, the small increments of F(G) up to the establishment to three clusters together with the large jump in the F(G) value, from 0.226 to 0.693, when two clusters are established suggest that three natural clusters are present. In fact, they correspond to those more frequently found in the original analysis using primary distances (Table 2). However, it would be most interesting to be able to *a priori* establish a cutoff value beyond which the clustering results will be considered unreliable. To do so, we have defined an *Affinity Coefficient* (*AC*), as follows:

$$\mathrm{AC} = 100\ \{(1 - \mathrm{F}[\mathrm{G}]) / (1 - \mathrm{F}[\mathrm{G}_{\min}])\} \tag{6}$$

Where $\mathrm{F}(\mathrm{G}_{\min})$ is the minimum value for the F(G) function. We thus will proceed to define a particular value of *AC* and then use it to establish the limit in which the clustering procedure is stopped. If $AC = 100$, then only optimal clustering will be considered. In a case as the one discussed above, that would mean that only elements with distances equal to 1 will be clustered together. However, by relaxing the conditions, that is using $AC < 100$, we will allow some level of non-optimal clustering to occur. As we will show in the next section, relaxing the conditions of clustering may be useful when considering incomplete and/or unreliable data, as those generated in massive proteomic projects.

For a total of R replicates for the hierarchical clustering analysis using primary distances, the clustering strategy may be described as follows:

```
Select AC value
Repeat_from N = 0
     Random ordering of elements;
     Hierarchical Clustering (d, AC);
     Increment d' counters according to the solution found;
     N = N + 1
To N = R
```

3 Application to Real Proteomic Data

Protein-protein interaction data are rapidly accumulating and the analysis of these data may provide very important hints about cellular function. In the yeast *Saccharomyces cerevisiae*, massive interaction data have been obtained using two different strategies, namely massive two-hybrid system analyses [7, 12] and affinity purification of complexes using tagged proteins [5, 6]. However, there are two problems with the information generated using those techniques. On one hand, false positive interactions are generated by proteins that are "promiscuous", that is, able, under the conditions of these experiments, to anomalously bind to multiple partners. The number of false positive interactions may be up to 50% [13]. On the other hand, purification of complexes using tagged proteins is often partial, that is, the complexes obtained do not contain all the proteins that constitute them *in vivo*. This is shown by the fact that different complexes that however share several, often many, subunits are found (data from [5, 6]).

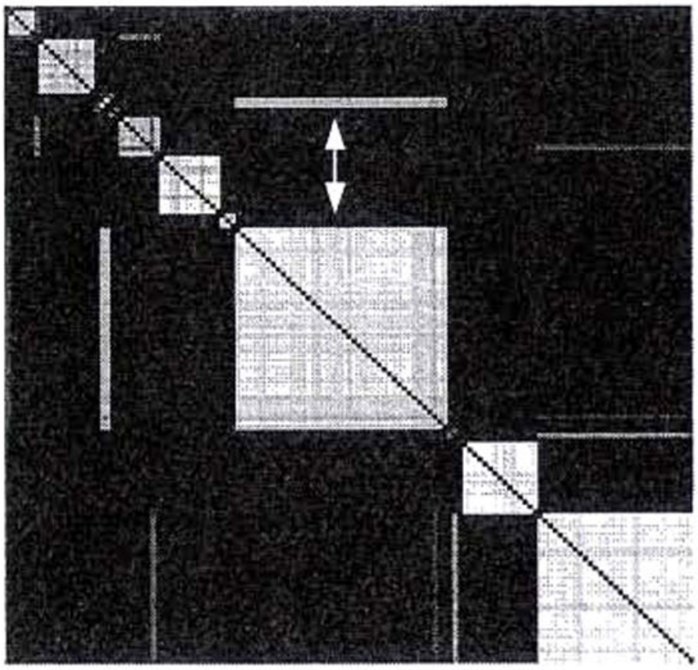

Fig. 1. Summary of results for septin-interacting proteins of *S. cerevisiae*, with $AC = 100\%$. The light gray tones correspond to low secondary distances between proteins. Proteins that are part of a complex are shown consecutively in this figure. Asterisks indicate two examples of putative promiscuous proteins, characterized by having similar secondary distance values with proteins belonging to several different complexes. These results were obtained after 1000 replicates

Our clustering strategy may contribute to the resolution of these problems, especially for the data provided by complex purification experiments. In order to implement this strategy, we started by creating a simple measure of distance among proteins, that ranges from 1 (when direct interactions are known) to 5 (unrelated proteins) (Mars, Arnau and Marín, submitted). Once distances are determined for a set of relevant proteins, the clustering strategy detailed in the previous section allows to determine secondary distances among proteins. When a protein is promiscuous, it has primary distances of 1 with many proteins. This fact determines that the secondary distances of this protein with many others are similar and often much higher than expected for a protein that belongs to a particular complex. On the other hand, when different independent complexes are found that have several proteins in common, and thus most likely correspond to partial purifications of a same, bigger complex, those common proteins obtain values of distance equal to 1. When secondary distances are established, proteins of these complexes with common subunits have values that are much smaller that those found for proteins that belong to independent complexes.

We have used this strategy with the set of proteins that interact with a group of *S. cerevisiae* cytokinesis and cell cycle regulators, the proteins known as septins. Using data obtained by Gavin *et al.* and Ho *et al.* [5, 6], we established that septins interact with a total of 137 proteins that were purified as part of 13 complexes. We then generated a 137 x 137 matrix of distances by compiling all the information available for those proteins, and used our hierarchical clustering strategy to determine secondary distances among proteins under different AC values, ranging from 100% (only distances equal to 1 are used for clustering) to 70% (a much more relaxed condition, when proteins with distances equal to 2 or even 3 were allowed to cluster together). Figure 1 shows our results for *AC* = 100% using gray tones to represent secondary distances.

The first important result is that our analyses allowed the recognition of eleven of the thirteen complexes, demonstrating that the clustering strategy is correctly functioning. Moreover, our results also established the existence of a very strong link between proteins of the remnant two complexes, that suggests these complexes actually may be just partial purifications of a single, larger complex. Examination of the components of those two highly related complexes led us to the finding that they have related functions, and most likely are part of a single complex, which function would be to coordinately generate multiple aminoacyl-tRNAs in order to locally increase protein synthesis. A similar complex had been hitherto characterized in animals U (see [9] and references therein), but never in yeasts as *S. cerevisiae*. In summary, our method has demonstrated its usefulness to deal with real proteomic data, generating significant information to interpret complex interaction results.

4 Conclusions

In this paper, we propose a strategy of hierarchical clustering with two distinctive features: iterative generation of multiple solutions and control of the quality of the clustering, using the *AC* parameter. We also show that it can be used to analyze real proteomic data. It is known that protein complexes are often partially characterized and that a certain amount of false positives are obtained when massive interaction data are generated. Our strategy allows detection of those anomalies.

Our implementation of this method is relatively fast. Data presented above for 137 proteins generated a dataset of 9316 distances. A total of 1000 replicates to obtain reliable secondary distances from that dataset can be obtained in about an hour on an IBM-compatible PC computer running at 1.7 GHz. The examined dataset contains about 2.5 x 10^{-4} of all possible interactions in *S. cerevisiae* (that has about 6000 different protein products) and perhaps about 10^{-6} of all possible interactions in human cells (assuming 100000 different proteins, in part determined by alternative RNA processing). That means that analysis of the whole datasets for eukaryotic species would require parallelizing our algorithms. However, research of most scientists is focused on particular cellular processes that involve limited groups of proteins. Those applications require the analyses of much smaller datasets, as the one showed above, that can be easily performed on a standard personal computer in a short time. Thus, we think our strategy can be of very general use, and its simplicity allows it to potentially become established as a benchmark for proteomic data analysis.

References

[1] V. Arnau, J.M. Orduña, A. Ruiz, and J. Duato. "On the Characterization of Interconnection Networks with Irregular Topology: a New Model of Communication Cost", in Proceedings of the IASTED Internactonal Conference Parallel and Distributed Computing and Systems (PDCS'99) pp. 1-6, Massachusetts, 1999.

[2] R. O. Duda and P. E. Hart. "Pattern Classification and Scene Analysis", John Wiley and Sons, 1973.

[3] B. Everitt, "Cluster Analysis". John Wiley and Sons, New York, 1974.

[4] D. Fasulo. "An Analysis of Recent Works on Clustering Algorithms", Tech. Rep. # 01-03-02. Dpto. of Computer Science & Engineering, University of Washington, 1999.

[5] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch *et al.* (38 authors), "Functional organization of the yeast proteome by systematic analysis of protein complexes", *Nature*, 415, 141-147, 2002.

[6] Y. Ho, A. Gruhler, A. Helibut, G. D. Bader, L. Moore *et al.* (46 authors), "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry", *Nature*, 415, 180-183, 2002.

[7] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", *Proc. Natl. Acad. Sci. USA*, 98, 4569-4574, 2001.

[8] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. "Lethality and centrality in protein networks", *Nature* 411, 41-42, 2001.

[9] L. Nathanson and M. P. Deutscher, "Active aminoacyl-tRNA synthetases are present in nuclei as a high molecular weight multienzyme complex", *J. Biol. Chem.* 41, 31559-31562, 2000.

[10] J.M. Orduña, V. Arnau, and J. Duato. "Characterization of Communications between Processes in Message-Passing Applications", in "IEEE International Conference on Cluster Computing (CLUSTER2000)", pp. 91-98, Chemnitz, Germany, 2000.

[11] J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, et al. (13 authors), "The protein-protein interaction map of *Helicobacter pylori", Nature* 409, 211-215, 2001.

[12] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson et al. (20 authors). "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*", *Nature*, 403, 623-627, 2000.

[13] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions", *Nature*, 417, 399-403, 2002.