Protein Sequence Annotation by means of Community Detection

Giuseppe Profiti[12], Damiano Piovesan[3], Pier Luigi Martelli[23], Piero Fariselli[13], Rita Casadio[23]

[1]Department of Computer Science and Engineering, University of Bologna, via Mura Anteo Zamboni 7, Bologna, Italy
[2]Health Science and Technologies CIRI, University of Bologna, via Tolara di Sopra 50, Ozzano dell'Emilia, Italy
[3]Biocomputing Group, University of Bologna, Via San Giacomo 9/2, Bologna, Italy
giuseppe.profiti2@unibo.it, {piovesan, piero, gigi, casadio}@biocomp.unibo.it

The improvement of sequencing technologies is increasing the volume of biosequences in databases. Experimental validation of genomes and proteomes is however far too slow compared to the pace at which data are being produced and electronic annotation is the current solution to this problem. The annotation of a new sequence is inferred from experimentally validated reference proteins using different algorithms. Recently we developed BAR+ [1], an annotation system that is cluster centric: a protein enters an annotated cluster provided that it shares at least 40% sequence identity over at least 90% of the alignment length with a protein of the cluster. From the cluster the protein inherits all the statistically validated features that characterize the cluster. These can include GO terms, Pfam domains and protein structure. Clusters in BAR+ where generated by splitting the components of graphs where two nodes (two proteins) are linked when they share at least 40% sequence identity over at least 90% of the pairwise sequence alignment [2]. BAR+ clusters are therefore graphs where protein sequences are the nodes and similarity relationships are the edges, with weight equal to the evaluated sequence identity between the pair of nodes. Over 13 million protein sequences have been clustered into 913962 clusters, with cluster size up to 87893 nodes.

Here we enhance the level of detail within BAR+ clusters by applying algorithms used to identify communities in graphs. This is done in order to subcluster sequences that share within the same cluster more specific functional and structural features. A community is defined as a subset of nodes having more edges leading to members of the same community than to other nodes in the graph. The term community comes from the original application of this concept to social networks; however, community detection is now used to assess robustness of network infrastructures and to analyze interaction networks [3], [4]. The definition of community is a bit vague and then a mathematical measure is needed in order to compare different assignment of nodes to communities in a graph. Different approaches to community detection have been developed [5], mostly relying on the maximization of a target function. Other clustering techniques, like spectral methods and k-means, require a-priori knowledge of the number of communities. For our purpose, however, an algorithm capable to automatically detect the communities without the need of setting a parameter is

necessary. We therefore decided to focus on modularity optimization algorithms since they are mostly deterministic and the number of communities is not needed in input.
Given a graph containing nodes belonging to a set of communities, the modularity measure [6] evaluates how well connected the nodes inside a community are in respect of the other nodes. In theory, maximizing the modularity means that the best partitioning of the graph has been found. Maximizing the modularity is not easy from a computational point of view [7] and finding small communities requires additional knowledge [8].
We applied the Louvain method [9] of modularity maximization to all BAR+ clusters. This method was chosen due to its fast execution when compared to other approaches like the Girvan-Newman algorithm [10], that would be unfeasible on graphs with thousands of nodes like the ones of BAR+. Some of the clusters showed a separation of annotations amongst the communities. For the cluster of ABC transporters (87893 nodes) our procedure allows splitting into communities specific for different substrates [11]. In other clusters there is a separation of proteins to be found in the extracellular from those in the intracellular region, or from proteins operating in the nucleus from those located in the cell cytoplasm. These results are statistically validated using a Bonferroni corrected p-value evaluation, the same approach of BAR+ [2]. Our approach apparently is useful to better select within the cluster associated features those that eventually are specific for a given community and therefore make possible an improved electronic annotation of those protein sequences that infer from the cluster their structural and functional characterization.

References

[1] D. Piovesan, P. Luigi Martelli, P. Fariselli, A. Zauli, I. Rossi, and R. Casadio, "BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences," *Nucleic Acids Research*, vol. 39, pp. W197–W202, May 2011.

[2] L. Bartoli, L. Montanucci, R. Fronza, P. L. Martelli, P. Fariselli, L. Carota, G. Donvito, G. P. Maggi, and R. Casadio, "The Bologna Annotation Resource: a Non Hierarchical Method for the Functional and Structural Annotation of Protein Sequences Relying on a Comparative Large-Scale Genome Analysis," *Journal of Proteome Research*, vol. 8, no. 9, pp. 4362–4371, 2009.

[3] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein–protein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006.

[4] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 1128–1133, 2003.

[5] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, Feb. 2010.

[6] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, p. 8577, 2006.

[7] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 2, pp. 172–188, 2008.

[8] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, p. 36, 2007.

[9] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, 2008.
[10] M. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Physical Review E*, vol. 69, Feb. 2004.
[11] D. Piovesan, G. Profiti, P. L. Martelli, P. Fariselli, and R. Casadio, “Extended and robust protein sequence annotation over conservative non hierarchical clusters: the case study of the ABC transporters.,” *Emerging Technologies in Computing Systems, ACM Journal on*, vol. To be published, 2012.