Chinese Whispers for Protein-Protein Interaction Network Analysis to Discover Overlapping Functional Modules

Ying Liu
Department of Computer Science, Mathematics and Science, College of Professional Studies,
St. John's University, Queens, NY 11439
Email: liuy1 [AT] stjohns.edu

Abstract - One of the most pressing problems of the post genomic era is identifying protein functions. Clustering Protein-Protein-Interaction networks is a systems biological approach to this problem. Traditional Graph Clustering Methods are crisp, and allow only membership of each node in at most one cluster. However, most real world networks contain overlapping clusters. Recently the need for scalable, accurate and efficient overlapping graph clustering methods has been recognized and various soft (overlapping) graph clustering methods have been proposed. This paper introduces Chinese Whisper, for protein-protein interaction network analysis to discover overlapping functional modules. The paper illustrated the importance of soft clustering methods in systems biology by giving a few concrete examples of how the biological function of the overlap nodes relates to the functions of the respective clusters.

Keywords: Protein-Protein Interaction networks; Graph Clustering; Chinese Whispers

I. Introduction

Homology based approaches have been the traditional bioinformatics approach to the problem of protein function identification. Variations of tools like BLAST [1] and Clustal [2] and concepts like COGs (Clusters of orthologous Groups) [3] have been applied to infer the function of a protein or the encoding gene from the known a closely related gene or protein in a closely related species. Although very useful, this approach has some serious limitations. For many proteins, no characterized homologs exist. Furthermore, form does not always determine function, and the closest hit returned by heuristic oriented sequence alignment tools is not always the closest relative or the best functional counterpart. Phenomena like Horizontal Gene Transfer complicate matters additionally. Last but not least, most biological Functions are achieved by collaboration of many different proteins and a proteins function is often context sensitive, depending on presence or absence of certain interaction partners.

A Systems Biology Approach to the problem aims at identifying functional modules (groups of closely cooperating and physically interacting cellular components that achieve a common biological function) or protein complexes by identifying network communities (groups of densely connected nodes in PPI networks). This involves clustering of PPI-networks as a main step. Once communities are detected, a hypergeometrical p-value is computed for each cluster and each biological function to evaluate the biological relevance of the clusters. Research on network clustering has focused for the most part on crisp clustering. However, many real world functional modules overlap. The present paper introduces a new simple soft clustering method for which the biological enrichment of the identified clusters seem to have in average somewhat better confidence values than current soft clustering methods.

II. Previous Work

Examples for crisp clustering methods include HCS [4], RNSC [5] and SPC [6]. More recently, soft or overlapping network clustering methods have evolved. The importance of soft clustering methods was first discussed in [7], the same group of authors also developed one of the first soft clustering algorithms for soft clustering, Clique Percolation Method or CPM [8]. An implementation of CPM , called CFinder [9] is available online. The CPM approach is basically based on the "defective cliques" idea and has received some much deserved attention. Another soft clustering tool is Chinese Whisper [10] with origins in Natural Language Processing. According to its author, Chinese Whispers can be seen as a special case of the Random Walks based method Markov-Chain-Clustering (MCL) [11] with an aggressive pruning strategy.

Recently, some authors [12, 13] have proposed and implemented betweenness based [14] Clustering (NG) method, which makes NG's divisive hierarchical approach capable of identifying overlapping clusters. NG's method finds communities by edge removal. The modifications involve node removal or node splitting. The decisions about which edges to remove and which nodes to split, are based on iterated all pair shortest path calculations.

In this paper, we apply Chinese Whispers for protein-protein interaction network analysis to discover overlapping functional modules. Ian the rest of the paper, we first describe Chinese Whispers. The second part of this work aims to illustrate the biological relevance of soft methods by giving several examples of how the biological functions of overlap nodes relate to biological functions of respective clusters.

III. Chinese Whispers

Chinese Whispers [10] is a randomized bottom-up Clustering algorithm with a time complexity of *O(*|E|*)*. In terms of complexity, the algorithm is quasi unbeatable. The Algorithm is outlined as (Figure 1):

```
initialize:
 forall vi in V: class(vi)=i;
while changes:
 forall v in V, randomized order:
 class(v)=highest ranked class
           in neighbourhood of v;
```

Figure 1. Pseudocode of Chinese Whispers

The algorithm is parameter free (there is no need to specify the number of clusters, a threshold, an external stopping condition etc.). There are however several configuration options that can strongly influence its behavior (a node changes its label in an update step differently, depending on chosen options).

The most important one is the choice of how the “highest ranked class” (fifth line in the description of the algorithm, Figure 1) in neighborhood of a vertex is determined.

To explain the difference between the possible choices we use the same example as Chinese Whispers User’s manual [10] and paraphrase it where necessary:

Assume that we want to determine the highest ranked class in neighborhood of node A in Figure 2. Node A is currently labeled (i.e. assigned to community) L1, node B is labeled L4, C and E are assigned to community L3 and D is assigned to community L2. Furthermore link-strengths (weights) and degrees of the nodes are as shown in the figure.

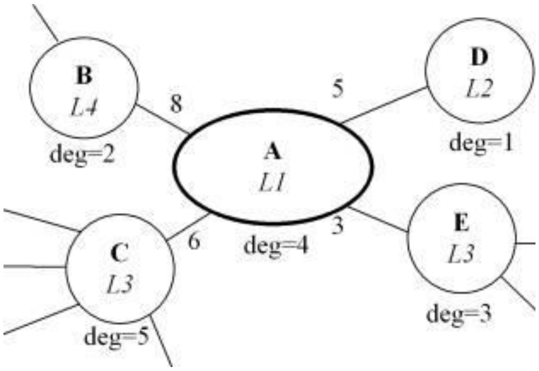

Figure 1: Calculation of Node Labels in Chinese Whisper

The strengths of classes for the situation in figure 1 are, dependent on the algorithm option:

top: strength(**L3**)=9; strength (L4)=8; strength (L2)=5

top sums over the neighbourhood’s classes; there is an edge of weight 6 between A and C, and a Link of weight 3 between A and E. Both C and E have label L3, hence the total strength of L3 at A is 6+3=9. This is larger than the strengths of the two other classes (L2 and L4), so using this option would change A’s Label to L3.

dist nolog: strength (**L2**)=5; strength (L4)=4; strength (L3)=2.2

dist downgrades the influence of a neighbouring node by its degree. For example, the total strength of L3 at A can be computed as: $\frac{6}{5}$ $(for\ node\ C)$+$\frac{3}{3}$ *(for node E)* =2.2. This is smaller than the influence of L2 at A $(\frac{5}{1} = 5)$. Using this option would change A’s label to L2.

dist log: strength (**L4**)=7.28; strength (L3)=5.51; strength (L2)=3.46

The influence of neighbouring nodes is downgraded by their degree, but the penalty is less severe than in the previous case.

vote: strength (**L3**)=0.409; strength (L4)=0.363; strength (L2)=0.227

Setting the algorithm option to ‘vote’ gives essentially the same ranking as top, but expresses the strength of each label at a node as the fraction of their vote in the total vote. In other words, it divides each strength value by the sum of all strength values. Therefore strength of L3 at A, using the vote option is calculated as $\frac{9}{8+5+9}$ =0.409. When using vote as algorithm option, an additional vote threshold must be set. If the vote threshold is set to a value above 0.409, then A keeps its label L1.

Table 1. 38 Clusters with size >=10 identified by Chinese Whispers

Cluster Number	Cluster Size	GO Enriched ?
1	151	Yes
2	59	Yes
3	50	Yes
4	48	No
5	45	Yes
6	33	Yes
7	25	Yes
8	24	No
9	24	Yes
10	21	Yes
11	21	Yes
12	20	Yes
13	20	Yes
14	20	No
15	20	No
16	18	Yes
17	17	No
18	16	Yes
19	16	Yes
20	16	Yes
21	16	No
22	16	No
23	15	Yes
24	15	Yes
25	14	No
26	13	No
27	12	Yes
28	12	Yes
29	12	No
30	12	Yes
31	12	No
32	12	No
33	11	Yes
34	11	Yes
35	11	Yes
36	11	Yes
37	10	Yes
38	10	No

As mentioned before, the algorithm option in ChineseWhispers is the most influential option. But the choices are limited and the algorithm is very fast, so in the worst case, there is the possibility to try out all options and consider only the best results. Furthermore, the decision is by far not as arbitrary as many other parameters that often surface in ML tasks. Using the knowledge from the last chapter, regarding the multi-functionality of highly-connected nodes, we can already speculate that the *dist nolog* Algorithm option will yield better results than the top or the vote option. This idea was confirmed in the analysis of the results on the yeast-PPI-Network.

Other configuration options include a random mutation rate that assigns new classes with a probability decreasing in the number of iterations to avoid premature convergence in small graphs and to further decrease the influence of extraordinary well connected nodes (hubs). Lastly, there is a choice between continuous and stepwise update: in the continuous mode, a nodes label is changed immediately, so that it will participate in any calculation of its neighbors label with its new label. In the stepwise update mode, all class labels are updated at once, after all labels have been computed.

Biemann [10] explains how Chinese Whispers in stepwise mode can be interpreted as a tuned up version of a very popular graph clustering method, namely MCL.

The result of CW is a hard partitioning of the input graph into a number of partitions that emerges in the process – there is no need to specify the number of clusters in advance. The algorithm outputs the two highest ranked classes in the immediate neighborhood of each node. Therefore it is possible to obtain a *soft partitioning* based on the weighted distribution of (hard) classes in the neighborhood of a node in a final step.

IV. Experimental Results and Discussions

There are 38 clusters with more than 10 nodes. We were able to confirm a significant enrichment in Terms of Gene Ontology for 26 of these clusters. Table 1 summarizes the information about size and GO-significance of the clusters.

Table 2. Overlaps between Communities in Clustering Results

Cluster	Cluster	overlaps
1	119	2
1	153	1
19	73	1
19	83	2
19	85	1
19	119	1
19	392	2
22	43	2
22	73	2
22	83	3
22	129	1
22	137	5
22	869	1
43	60	1
43	73	1
43	83	1
43	85	2
43	364	5
60	492	1
65	83	1
65	143	1
65	869	1
73	83	1
73	85	1
73	137	1
73	226	3
83	137	1
83	196	3
83	392	2
83	870	2
85	153	2
85	236	11
119	153	2
119	364	2
129	226	2
137	170	1
143	153	1
143	170	2
150	364	1

In general, our Chinese Whispers cluster sizes are small. Also, the interfaces between clusters – where they exist- tend to be relatively sharp. The 26 clusters of size 10 and larger with significant GO-Enrichment “share” 79 nodes. These are nodes that after the final softening step have one of the clusters as primary and another one of the clusters as secondary class. Talbe 2 summarizes the overlaps between all of those clusters that were deemed biologically significant by GO-Enrichment analysis.

A. Enrichment Analysis

We preformed both GO-Enrichment and MIPS-functional catalogue Enrichment analysis for all clusters of size 10 and larger. Table 3 reviews the results, ordered by p-values. The table lists up to 6 different assignments for each of the clusters. The clusters listed are all clusters with a corr. p-value better than e-12.

B. GO Enrichment Analysis for Overlaps

Interestingly, the two clusters with the highest number of common nodes, namely clusters 85 and 236, have exactly the same GO-ID assigned to them. Here are two more examples of how the enrichment values of overlaps fit into enrichment values of the clusters.

Table 3. Best CW Communities - GO Enrichment

Cluster ID	GO_ID	p_val	cor_pval	hits	Network total	**Description**
distlognm85	377	5.6052E-45	4.8905E-43	31	96	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
distlognm85	398	4.2039E-45	4.8905E-43	31	95	Nuclear
distlognm85	375	7.567E-44	4.2908E-42	31	103	RNA splicing, via transesterification reactions
distlognm83	30163	5.814E-42	1.7498E-39	36	172	Protein
distlognm83	6508	1.4431E-41	2.1717E-39	36	176	Proteolysis
distlognm85	6395	7.5524E-41	3.2098E-39	31	125	RNA splicing
distlognm83	6511	4.8044E-41	3.6152E-39	34	146	ubiquitin-dependent protein catabolic process
distlognm83	19941	4.8044E-41	3.6152E-39	34	146	modification-dependent protein catabolic process
distlognm83	51603	1.0384E-40	6.2512E-39	34	149	proteolysis involved in cellular protein catabolic process
distlognm83	43632	2.8234E-40	1.4164E-38	34	153	modification-dependent macromolecule catabolic process
distlognm85	16071	3.2687E-39	1.1114E-37	34	201	mRNA metabolic process
distlognm85	6397	3.282E-38	9.299E-37	31	149	mRNA processing
distlognm43	42254	2.3371E-31	5.7961E-29	32	327	Ribosome
distlognm364	6365	1.1333E-30	7.5934E-29	20	169	rRNA
distlognm364	16072	2.6754E-30	8.9626E-29	20	176	rRNA metabolic process
distlognm143	7035	3.8047E-28	1.2429E-26	12	24	vacuolar acidification
distlognm143	45851	3.8047E-28	1.2429E-26	12	24	pH
distlognm143	51452	3.8047E-28	1.2429E-26	12	24	cellular pH reduction
distlognm143	51453	7.3133E-28	1.4334E-26	12	25	regulation of cellular pH
distlognm143	30641	7.3133E-28	1.4334E-26	12	25	cellular hydrogen ion homeostasis
distlognm43	22613	1.224E-28	1.5178E-26	32	396	ribonucleoprotein complex biogenesis and assembly
distlognm226	6383	8.4203E-28	2.1051E-26	12	38	Transcription
distlognm143	6885	7.2843E-27	1.1898E-25	12	29	regulation of pH
distlognm119	6402	2.6261E-27	3.6503E-25	14	59	mRNA
distlognm137	6350	1.1452E-26	1.5231E-24	25	546	Transcription
distlognm119	6401	3.0431E-26	2.115E-24	14	69	RNA catabolic process
distlognm137	32774	3.5555E-25	1.4628E-23	24	501	RNA biosynthetic process
distlognm137	6351	2.7814E-25	1.4628E-23	24	496	transcription, DNA-dependent
distlognm137	6366	4.3995E-25	1.4628E-23	22	333	transcription from RNA polymerase II promoter
distlognm364	42254	1.0632E-24	2.3744E-23	20	327	ribosome biogenesis and assembly
distlognm43	42273	3.259E-25	2.6941E-23	18	64	ribosomal large subunit biogenesis and assembly
distlognm60	31123	2.1759E-24	2.263E-22	12	39	RNA
distlognm364	22613	5.3701E-23	8.9949E-22	20	396	ribonucleoprotein complex biogenesis and assembly
distlognm60	31124	2.2489E-23	9.637E-22	11	29	mRNA 3'-end processing
distlognm60	6378	2.7799E-23	9.637E-22	10	18	mRNA polyadenylation
distlognm364	6394	8.0791E-23	1.0826E-21	20	404	RNA processing
distlognm236	398	9.9106E-23	4.4641E-21	14	95	Nuclear
distlognm236	377	1.1595E-22	4.4641E-21	14	96	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
distlognm236	375	3.3191E-22	8.5191E-21	14	103	RNA splicing, via transesterification reactions
distlognm60	43631	1.239E-21	3.2213E-20	10	24	RNA polyadenylation
distlognm236	6395	5.8246E-21	1.1212E-19	14	125	RNA splicing
distlognm236	6397	7.6012E-20	1.1706E-18	14	149	mRNA processing
distlognm153	6810	2.4763E-20	9.5833E-18	72	958	Transport
distlognm119	16071	2.3906E-19	1.1077E-17	14	201	mRNA metabolic process
distlognm153	51234	7.5479E-20	1.4605E-17	72	976	establishment of localization
distlognm60	6379	1.0696E-18	2.2249E-17	9	25	mRNA cleavage
distlognm153	51179	1.7797E-19	2.2958E-17	73	1017	Localization
distlognm137	16070	1.7109E-18	4.5509E-17	24	944	RNA metabolic process
distlognm236	16071	5.736E-18	7.3612E-17	14	201	mRNA metabolic process

distlognm22	6366	7.8764E-19	1.402E-16	21	333	Transcription
distlognm43	16072	3.5612E-18	2.2079E-16	19	176	rRNA metabolic process
distlognm22	6357	4.9211E-18	2.9199E-16	18	215	regulation of transcription from RNA polymerase II promoter
distlognm22	114	3.5293E-18	2.9199E-16	9	14	G1-specific transcription in mitotic cell cycle
distlognm119	43285	1.6413E-17	5.7037E-16	14	270	biopolymer catabolic process
distlognm119	44265	3.3728E-17	9.3763E-16	14	284	cellular macromolecule catabolic process
distlognm22	51318	3.7099E-17	1.3207E-15	10	26	G1 phase
distlognm22	80	3.7099E-17	1.3207E-15	10	26	G1 phase of mitotic cell cycle
distlognm60	6397	8.2237E-17	1.4254E-15	12	149	mRNA processing
distlognm43	6365	4.4492E-17	2.2068E-15	18	169	rRNA processing
distlognm119	9057	1.5382E-16	3.5635E-15	14	316	macromolecule catabolic process
distlognm137	6139	3.7892E-16	8.3994E-15	25	1419	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
distlognm43	42255	3.6745E-16	1.5188E-14	13	64	ribosome assembly
distlognm364	30490	2.1141E-15	1.8998E-14	9	38	maturation of SSU-rRNA
distlognm73	6366	1.4017E-16	2.439E-14	16	333	Transcription
distlognm22	6351	2.9808E-15	8.8432E-14	21	496	transcription, DNA-dependent
distlognm129	6338	8.1485E-15	1.1326E-12	11	149	Chromatin
distlognm226	32774	1.4217E-13	1.1848E-12	12	501	RNA biosynthetic process
distlognm226	6351	1.2588E-13	1.1848E-12	12	496	transcription, DNA-dependent
distlognm129	6323	3.9905E-14	1.3867E-12	12	247	DNA packaging
distlognm129	6366	2.606E-14	1.3867E-12	13	333	transcription from RNA polymerase II promoter
distlognm129	6325	3.9905E-14	1.3867E-12	12	247	establishment and/or maintenance of chromatin architecture
distlognm226	6350	4.0352E-13	2.522E-12	12	546	Transcription
distlognm73	6351	8.0707E-14	5.4854E-12	16	496	transcription, DNA-dependent
distlognm73	32774	9.4575E-14	5.4854E-12	16	501	RNA biosynthetic process
distlognm392	6454	4.8605E-14	5.9784E-12	8	46	Translational
distlognm129	6368	3.2547E-13	9.0481E-12	8	53	RNA elongation from RNA polymerase II promoter

Table 4. Two examples of how the enrichment values of overlaps fit into enrichment values of the clusters

Example 1: 137 and 22 share 5 nodes.

Distlognm137: (25 nodes)

GO-ID	p-value	corr p-value	# selected	# total	Description
6350	1.15E-26	1.52E-24	25	546	transcription
6351	2.78E-25	1.46E-23	24	496	transcription, DNA-dependent
32774	3.56E-25	1.46E-23	24	501	RNA biosynthetic process

Distlognm22:(33 nodes)

GO-ID	p-value	corr p-value	# selected	# total	Description
6366	7.88E-19	1.40E-16	21	333	transcription from RNA polymerase II promoter
114	3.53E-18	2.92E-16	9	14	G1-specific transcription in mitotic cell cycle
6357	4.92E-18	2.92E-16	18	215	regulation of transcription from RNA polymerase II promoter

Overlap of 137 and 22 (5 nodes)

GO-ID	p-value	corr p-value	# selected	# total	Description
6355	1.93E-04	4.40E-03	3	338	regulation of transcription, DNA-dependent
45449	2.41E-04	4.40E-03	3	364	regulation of transcription
122	3.10E-04	4.40E-03	2	60	negative regulation of transcription from RNA polymerase II promoter

Example 2: 43 and 364 share 5 nodes

Distlognm43(45 nodes, hereof 1 un-annotated):

GO-ID	p-value	corr p-value	# selected	# total	Description
42254	2.34E-31	5.80E-29	32	327	ribosome biogenesis and assembly
22613	1.22E-28	1.52E-26	32	396	ribonucleoprotein complex biogenesis and assembly
42273	3.26E-25	2.69E-23	18	64	ribosomal large subunit biogenesis and assembly

Distlognm364(21 nodes):

GO-ID	p-value	corr p-value	# selected	# total	Description
6365	1.13E-30	7.59E-29	20	169	rRNA processing
16072	2.68E-30	8.96E-29	20	176	rRNA metabolic process
42254	1.06E-24	2.37E-23	20	327	ribosome biogenesis and assembly

Overlap of 43 and 364(5 nodes):

GO-ID	p-value	corr p-value	# selected	# total	Description
42254	5.37E-07	1.77E-05	5	327	ribosome biogenesis and assembly
22613	1.41E-06	2.32E-05	5	396	ribonucleoprotein complex biogenesis and assembly
6365	3.32E-06	3.22E-05	4	169	rRNA processing

V. Conclusions

This paper introduced Chinese Whispers [10], a randomized bottom-up Clustering algorithm with a time complexity of *O(*|E|*)*, for protein-protein interaction network analysis to discover overlapping functional modules. In this paper, we first described Chinese Whispers. We further illustrated the biological relevance of soft methods by giving several examples of how the biological functions of overlap nodes relate to biological functions of respective clusters. The paper illustrated the importance of soft clustering methods in systems biology by giving a few concrete examples of how the biological function of the overlap nodes relates to the functions of the respective clusters.

VI. References

[1] Altschul, SF, et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs“. Nucleic acids research 25, no. 17: 3389, 1997.

[2] Thompson, JD, DG Higgins, and TJ Gibson. “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”. Nucleic acids research 22, no. 22: 4673-4680, 1994

[3] Tatusov, R. L., E. V. Koonin, and D. J. Lipman. “A genomic perspective on protein families”. Science 278, no. 5338: 631, 1997.

[4] Hartuv, E., R. Shamir. “A clustering algorithm based on graph connectivity”. Information processing letters 76, no. 4-6: 175-181, 2000.

[5] King, A. D., N. Przulj, and I. Jurisica. “Protein complex prediction via cost-based clustering”. Bioinformatics 20,: 3013-3020, 2004.

[6] Spirin, V., L. A. Mirny. “Protein complexes and functional modules in molecular networks”. Proceedings of the National Academy of Sciences 100, no. 21: 12123-12128, 2003.

[7] Palla, G., I. Derenyi, I. Farkas, and T. Vicsek. “Uncovering the overlapping community structure of complex networks in nature and society”. Nature 435, no. 7043 (Jun 9): 814-818, 2005.

[8] Derenyi, I., et al. “Clique percolation in random networks”. Physical Review Letters 94, no. 16: 160202, 2005.

[9] Adamcsek, B., G. et al. “CFinder: locating cliques and overlapping modules in biological networks”. Bioinformatics 22, no. 8: 1021-1023, 2006.

[10] Biemann, C. “Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems”. In Proceedings of the HLT-NAACL-06 workshop on textgraphs-06, new york, USA, 2006.

[11] Van Dongen, S. “A cluster algorithm for graphs”. Report-Information systems , no. 10: 1-40, 2000.

[12] Pinney, J. W., D. R. Westhead. “Betweenness-based decomposition methods for social and biological networks”. In Interdisciplinary statistics and bioinformatics. Edited by S. Barber, P. D. Baxter, K. V. Mardia and R. E. Walls. Leeds University Press, 2000.

[13] Gregory, S. “An algorithm to find overlapping community structure in networks”. Lecture Notes in Computer Science 4702: 91, 2007.

[14] Girvan, M., M. E. Newman. “Community structure in social and biological networks”. PNAS 99: 7821-7826, 2002.