A Network of Hidden Markov Models and Its Analysis

Liqing Zhang[1]**, Layne T. Watson**[2]**, and Lenwood S. Heath**[1]
[1]Departments of Computer Science, Virginia Tech, Blacksburg, VA, USA
[2]Departments of Computer Science and Mathematics, Virginia Tech, Blacksburg, VA, USA

Abstract— *The Structural Classification of Proteins (SCOP) database uses a large number of hidden Markov models (HMMs) to represent families and superfamilies composed of proteins that presumably share the same evolutionary origin. However, how the HMMs are related to one another has not been examined before. In this work, taking into account the processes used to build the HMMs, we propose a working hypothesis to examine the relationships between HMMs and the families and superfamilies that they represent. Specifically, we perform an all-against-all HMM comparison using the HHsearch program and construct a network where the nodes are HMMs and the edges connect similar HMMs. We hypothesize that the HMMs in a connected component belong to the same family or superfamily more often than expected under a random network connection model. Results show a pattern consistent with this working hypothesis. Moreover, the HMM network possesses features distinctly different from previously documented biological networks, exemplified by the exceptionally high clustering coefficient and the large number of connected components. The current finding may provide guidance in devising computational methods to reduce the degree of overlaps between the HMMs representing the same superfamilies, which may in turn enable more efficient large-scale sequence searches against the database of HMMs.*

Keywords: hidden Markov models, network, centrality, clustering coefficient, tree

1. Introduction

The Structural Classification of Proteins (SCOP) database is a comprehensive protein database that organizes and classifies proteins based on their evolutionary and structural relationships [1], [7], [8]. It is organized into four hierarchical levels: family, superfamily, fold, and classes. At the lowest level (family), individual proteins are clustered into families based on some criteria that may indicate their common evolutionary origin, such as having a pairwise sequence similarity of more than 30% or lower sequence similarity but similar functions and structures. A good example of the latter is seen in globin proteins whose pairwise sequence similarities are much lower than 30% but which have similar protein functions. Next, families are grouped into superfamilies if their structures and/or function features indicate a possible common evolutionary origin. Then superfamilies are clustered into folds if superfamilies share major secondary structures with the same topological arrangements. Finally, different folds are grouped into classes based on their secondary structural compositions. Unlike the other levels, a class might not necessarily imply common evolutionary origins and exists more for convenience than for actual biological implications.

Apart from the hierarchical classification and organization of proteins, the SCOP database employs hidden Markov models (HMMs) to represent superfamilies [4], [5]. The basic procedure of building an HMM for a particular superfamily starts with a seed protein and performs sequence search in a database to obtain other proteins that have sequence similarities above a set threshold. The newly obtained sequences are used to iterate the search for some number of times to obtain additional proteins. Finally, all sequences are aligned and an HMM is constructed for the multiple sequence alignment [4], [5]. It has been shown that different seed proteins might produce HMMs that cover different members of the superfamily [4], [5]. Thus, in order to represent the full set of proteins in a superfamily, multiple HMMs are built for the superfamily using multiple seed proteins. For example, the beta-beta-alpha zinc fingers superfamily has altogether 91 HMMs representing it, and the P-loop containing nucleoside triphosphate hydrolases superfamily has 406 HMMs representing it.

Because each superfamily might be represented by multiple HMMs, there may be a high degree of overlap and redundancy among the models. However, there have not been any studies examining this issue systematically. To understand how the HMMs in the SCOP database are related to one another and the degree of overlap or redundancy among HMMs from either the same or different superfamilies, we perform a detailed analysis of the HMMs in SCOP for their similarity and relationships using a network approach. Specifically, we perform an all-against-all HHsearch for the library of HMMs in the SCOP database. HHsearch is similar to BLAST, except that instead of matching a sequence against a database of sequences, it uses a query HMM or sequence to match against a database of HMMs and identifies the HMMs significantly homologous to the query HMM or sequence [10]. We then construct a network of HMMs, where the link between two HMMs is based on their similarity, and examine some commonly evaluated network properties. We compare the current network with previously documented networks and outline some questions for future research.

2. Methods

The SCOP library of HMMs was downloaded from the SCP website (http://scop.mrc-lmb.cam.ac.uk), where the SCOP version was filtered to 70% maximum pairwise sequence identity. The library contains a total of 13,730 HMMs, from seven classes *a,b,c,d,e,f,g,* where class *a* contains only α (i.e., α helix) proteins, class *b* contains only β (i.e., β sheet) proteins, class *c* contains α and β proteins (mainly parallel β sheets ($beta-alpha-beta$ units)), class *d* contains α and β proteins (mainly antiparallel β sheets, i.e., segregated α and β regions), class *e* contains multi-domain proteins (i.e., folds consisting of two or more domains belonging to different classes), class *f* contains membrane and cell surface proteins, and class *g* contains small proteins. It is useful to mention that the SCOP domain classification ID specifies the entire hierarchy, e.g. c.1.1.1, the first field is for the class *c*, second for the fold, third for the superfamily, and the last for the family.

HHsearch [10] was performed for all-against-all HMMs with the default parameters. HHsearch, similar to BLAST, uses a query that can be either a protein sequence or an HMM to search a database of sequences or HMMs and identify homology between the query and sequences and HMM models in the databases that is above a given threshold. In the current study, the e-value, a measurement of homology similar to BLAST's e-value, was set to 0.001. This e-value cutoff has also been used by Pfam to identify a Pfam clan [2], which is essentially equivalent to the superfamily hierarchy. A total of 13,547 HMMs have matches that met the criterion, with 1,618 having no other matches except themselves. Thus, 11,929 HMMs were used for the subsequent network analysis.

To study the relationship of the HMMs, an undirected network $G = (V, E)$ was constructed, where the vertices V are HMMs, and there is an edge in E between two HMMs if their e-value is below the threshold. General network statistics were computed, and a quadratic function was fitted to the log-log degree distribution. Three common vertex centrality measurements, degree centrality, betweenness centrality, and closeness centrality, were computed to evaluate the importance of vertices in the network. The degree of a vertex a is the number of edges incident on a. Betweenness for a vertex a,

$$b(a) = \sum_{\substack{s,t \in V \\ s \neq a \\ t \neq a}} \frac{\sigma(s, t \mid a)}{\sigma(s, t)}, \tag{1}$$

introduced in Freeman [3], measures roughly the number of shortest paths going through a. $\sigma(s, t)$ is the number of shortest paths between vertices s and t, and $\sigma(s, t \mid a)$ is the number of shortest paths between vertices s and t that go through a. Thus, the higher the betweenness of a vertex, the more central/important the vertex is. In a fully connected network, the betweenness of all vertices is 0.

The closeness centrality measures the number of steps required to access every other vertex from a given vertex, specifically, the closeness of a vertex a, $c(a)$, is computed by

$$c(a) = \frac{|V| - 1}{\sum\limits_{\substack{i \in V \\ i \neq a}} d_{a,i}}, \tag{2}$$

where $d_{a,i}$ is the length of the shortest path between vertex a and vertex i. Closeness ranges from 0 (does not reach 0) to 1; the higher it is for a vertex, the more "central" the vertex is. These centrality measurements have different motivations and show different aspects for the importance of vertices in a network.

The network clustering coefficient, C, also known as transitivity, measured by the ratio between the number of triangles and the number of connected triplets, was computed for the entire network. The number of connected components that are trees, where there are N vertices but only $N - 1$ edges between the vertices, was computed for the entire network as well.

To systematically study the consistency between the e-value cutoffs for the prediction of whether or not HMMs belong to the same hierarchical level and classification of the SCOP database, we examined the Receiver Operating Characteristic (ROC) curves for the prediction of the hierarchical categories of two HMMs provided by different e-value cutoffs. The ROC curve shows how the true positive rate changes with the false positive rate for a classification. Specifically, for example, at the family level, if a sample of two HMMs were classified to the same family by the SCOP database, the prediction based on a specific e-value cutoff is considered to be a false negative (FN) if the e-value similarity of the two HMMs is worse/higher than the e-value cutoff, a true positive (TP) if the e-value is better (i.e., lower) than the cutoff; if the two HMMs were not classified to the same family by the SCOP database, the prediction based on the specific e-value cutoff is considered to be a true negative (TN) if the e-value similarity of the two HMMs is worse/higher than the e-value cutoff, a false positive (FP) if their e-value is better (i.e., lower) than the cutoff. Similar rules were applied to classify each pair of HMMs into the four categories (TP, FP, FN, and TN), for the four hierarchies, class, fold, superfamily, and family. True positive rate (i.e., sensitivity) was calculated as

$$TPR = \frac{TP}{TP + FN}, \tag{3}$$

and false positive rate (i.e., $1-$ specificity) as

$$FPR = \frac{FP}{FP + TN}. \tag{4}$$

An ROC curve was plotted for the four levels (i.e., class, fold, superfamily, and family) with different e-value cutoffs ranging from 10^{-20} to 10^{-3}.

Table 1: The general statistics of the HMMs

Class	Number of HMMs	Number of folds	Number of superfamilies	Number of families
a	1975	157	262	506
b	2590	109	231	485
c	3391	120	194	686
d	2932	223	328	683
e	199	34	34	51
f	145	29	44	50
g	697	49	70	112
All	11929	721	1163	2573

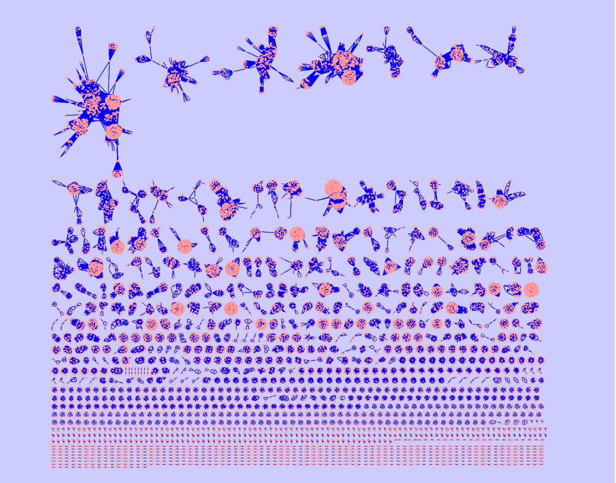

Fig. 1: The HMM network

3. Results

The working hypothesis. Taking into account the processes that built the HMMs and the hierarchical classification of the HMMs in the SCOP database, we hypothesize that the network should reflect this process, i.e., *the HMMs in a connected component belong to the same family or superfamily more often than expected under a random network connection model*.

General statistics of the HMMs and their network. A general description of the HMMs used to construct the network is shown in Table 1. There are seven classes in the collection of HMMs, falling into 721 folds, 1163 superfamilies, and 2573 families. Class *c* has the highest number of HMMs (3391) and class *f* the fewest (145).

The entire HMM network is shown in Figure 1, where the e-value cutoff is 0.001. There are altogether 151,461 edges for the 11,929 vertices. A significant property shown in Figure 1 is that the entire network is highly disconnected, with many much smaller connected components. In fact, there are altogether 1524 connected components (CCs). The smallest CC contains two vertices, the largest 590 vertices, $566/1524 = 37\%$ contain only two vertices and about 73% contain five or fewer vertices. The median CC size is 3 and the mean 7.8. The top 20 largest CCs are listed in Table 2.

Table 2: The 20 largest CCs and their densities

Size rank	Number of vertices	Density
1	590	0.12
2	349	0.21
3	277	0.65
4	155	0.15
5	141	0.38
6	121	0.33
7	120	0.19
8	106	0.72
9	99	0.84
10	90	0.95
11	86	0.99
12	85	0.89
13	81	0.32
14	80	0.83
15	74	0.66
16	73	0.65
17	72	0.16
18	70	1.00
19	69	0.97
20	66	0.40
All	11929	0.002

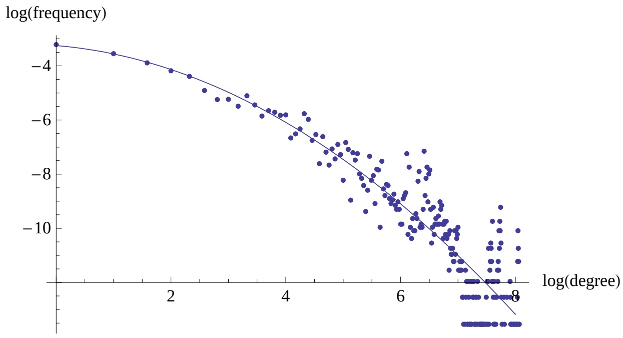

Fig. 2: Log-log degree distribution. The base is 2. The best fitting quadratic curve is $3.2481 - 0.176557x - 0.133088x^2$.

Degree distribution. The degree of the HMM network ranges from 1 to 268, with the average of 26 and median of 10. The log-log degree distribution is shown in Figure 2. It is evident that a power law distribution does not fit the data. The best fitting quadratic curve is also plotted with the data. It provides a relatively good fit for the smaller values of log(degree), and then towards the larger degrees, the fit is not so good.

Network Density. Density, computed as the number of edges over the number of all possible edges (in a fully connected graph), provides some quantitative evaluation on the connectivity of a network. The density of the entire network is low, only $0.002 = 151461/\binom{11929}{2}$. In contrast, individual CCs tend to have high densities, with more than 82.5% of CCs having density greater than 0.95. 1236 CCs are fully connected, i.e., cliques, with the largest clique of size 70.

Thus, individual CCs tend to have very high connectivity, whereas the entire network is not well connected. The

Table 3: The 20 HMMs with highest degree

Rank	HMM ID	SCOP ID	Degree
1	d1n26a1	b.1.1.4	268
2	d1f2qa1	b.1.1.4	265
3	d1qz1a3	b.1.1.4	265
4	d1biha1	b.1.1.4	264
5	d1rhfa1	b.1.1.1	263
6	d1tnna_	b.1.1.4	263
7	d2aw2a1	b.1.1.1	262
8	d1nbqa1	b.1.1.1	262
9	d1x44a1	b.1.1.4	262
10	d1biha3	b.1.1.4	262
11	d1cs6a3	b.1.1.4	262
12	d1f2qa2	b.1.1.4	261
13	d2avga1	b.1.1.4	261
14	d1epfa1	b.1.1.4	261
15	d3b5ha1	b.1.1.4	261
16	d1cs6a2	b.1.1.4	261
17	d1f97a2	b.1.1.4	261
18	d1epfa2	b.1.1.4	260
19	d2dava1	b.1.1.4	260
20	d1f97a1	b.1.1.1	260

Table 4: The 20 HMMs with largest betweenness

Rank	HMM ID	SCOP ID	Betweenness
1	d1bg6a2	c.2.1.6	14915.8
2	d1o8ca2	c.2.1.1	14665.7
3	d1e5qa1	c.2.1.3	14504.0
4	d2bzga1	c.66.1.36	9557.9
5	d3bswa1	b.81.1.8	9168.0
6	d1vj0a2	c.2.1.1	8211.0
7	d1ks9a2	c.2.1.6	7469.9
8	d2bmfa2	c.37.1.14	7439.8
9	d2dt5a2	c.2.1.12	7410.7
10	d1pjca1	c.2.1.4	7325.1
11	d1gtea4	c.4.1.1	7165.3
12	d1gu7a1	b.35.1.2	6768.0
13	d1tt7a1	b.35.1.2	6768.0
14	d2f1ka2	c.2.1.6	5985.2
15	d1ebfa1	c.2.1.3	5959.8
16	d1jqba2	c.2.1.1	5313.1
17	d1gr0a1	c.2.1.3	5220.0
18	d1ye8a1	c.37.1.11	5207.7
19	d1piwa2	c.2.1.1	4556.8
20	d1hdoa_	c.2.1.2	4403.8

density of the 20 largest CCs is shown in Table 2. The largest CC with 590 vertices has the lowest density, and the 18th largest CC with 70 vertices has a density of 1, and is therefore a fully connected component. There is a significant negative correlation between CC size and density (Kendall's rank correlation $\tau = -0.43$, p-value $< 2.2 \cdot 10^{-16}$ for CC size > 2).

Vertex centrality. Two centrality metrics, degree and betweenness, were computed for the vertices in the entire HMM network. Table 3 shows the top 20 HMMs that have the highest degrees. These 20 HMMs all belong to the same superfamily, b.1.1, Immunoglobulin, and also to the third largest CC that has 277 vertices. Thus, these 20 HMMs are connected with almost all other HMMs in the third CC. The HMM d1n26a1 (SCOP ID b.1.1.4, (A:1-93)) has the highest degree, 268, belonging to the Interleukin-6 receptor alpha chain, N-terminal domain (Homo sapiens).

Table 4 shows the top 20 HMMs that have the highest betweenness. Thirteen of the 20 HMMs belong to the superfamily c.2.1 (NAD(P)-binding Rossmann-fold domains), two to the superfamily b35.1.2, and two to the superfamily c.37.1. Eighteen of the 20 HMMs belong to the largest CC and the two remaining (c.37.1.14 and c.37.1.11) to the second largest. The HMM d1bg6a2 (SCOP ID c.2.1.6, (A:4-187)) has the highest betweenness, 14916, belonging to N-(1-D-carboxylethyl)-L-norvaline dehydrogenase (Arthrobacter, strain 1c). Interestingly, there is no overlap of HMMs that have the highest of both degree and betweenness.

Network diameter. The diameter of the largest CC (containing 590 vertices) is 9. The average distance between the vertices is 2.94. We also measured the diameters of all the CCs to see how they change as a function of CC size. Figure 3 shows that larger CCs tend to have larger diameters. However, smaller CCs can have large diameters as well. For example, a CC of size 32 has diameter seven, the same as a CC of size 155; a CC size of 16 has diameter six, the same as a CC of size 121. There are 1236 CCs with diameter 1, corresponding to the number of cliques.

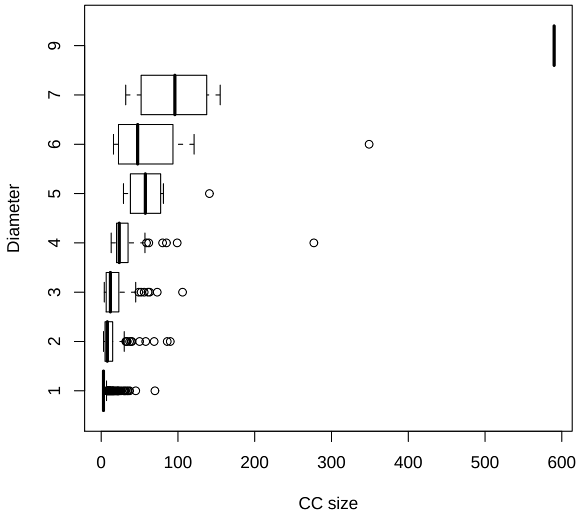

Fig. 3: Boxplot for the diameter of CCs as a function of CC size. The box marks the lower and upper quantile of CC sizes with the same diameter, the dark line marks the median, the whiskers mark the border of lower and upper outliers with the dots outside denoting the outliers.

CCs and hierarchy. Within the CCs, we examined whether the HMM members are from the same family, superfamily, fold, or class. There are altogether 1178 CCs whose members have the same SCOP domain classification

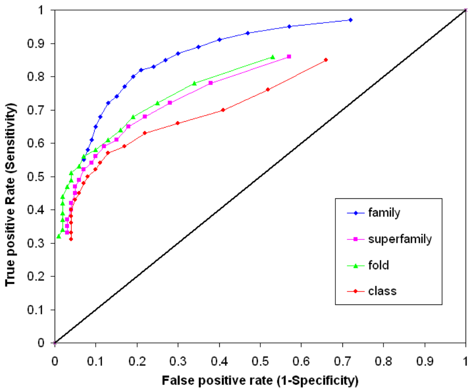

Fig. 4: The ROC curves for family, superfamily, fold, and class with different e-value cutoffs. For each curve, the data points from left to right correspond to the FPR and TPR for the e-value cutoffs from 10^{-20} to 10^{-3}.

(conserved at all hierarchical levels), 271 CCs whose HMMs belong to the same superfamily but to different families, 24 whose members belong to the same fold, but to different superfamilies, 18 whose members belong to the same class but have different folds, and the remaining 33 whose members are from different classes.

The consistency between the prediction of HMM memberships at different hierarchical levels in the SCOP database based on the e-value cutoffs and the classification of the SCOP database was evaluated by ROC curves, shown in Figure 4. We make several observations. First, for all four levels of the hierarchy, the higher the e-value cutoff, the higher the sensitivity (true positive rate), so is the false positive rate, which is expected because higher e-value means a less stringent prediction criterion that in turn leads to a higher number of true positive predictions, and also a higher number of false positive predictions. Meanwhile, the rate of increase in sensitivity outpaces the rate of increase in the false negative rate as the e-value becomes more stringent, suggesting that beyond a certain e-value cutoff, the HMMs belonging to the same hierarchical levels also tend to have high similarity, which make them robust to the e-value cutoff change. Second, the curves for the prediction of fold and superfamily are very similar to each other, indicating that for the same e-value cutoff, the predictions for whether two HMMs belong to the same fold or superfamily are similar. In fact, for the same e-value cutoff, the difference in true positive rate (sensitivity) between the fold and superfamily ROC curves is either 0 or 0.01, and the difference in false positive rate (1-specificity) falls within the narrow range $[0.01 - 0.04]$. Third, the prediction quality is the worst for class as compared to the other three levels, with worst sensitivity and specificity for the same e-value cutoffs. This might not be so surprising as classification at the class level is more for convenience than for biological reasons.

4. Discussion

The important HMMs. In this work, we used three centrality measurements to evaluate the importance of an HMM. The results show that from the entire network, the vertices with the highest degrees do not necessarily have the highest betweenness, and vice versa. Degree measures how many immediate neighbors one HMM has, and therefore, the more it has, the more central it is. The vertices with the 20 largest degrees are all from the third largest CC, and are connected to about 94% of its vertices. The vertices with the 20 largest betweenness values are from either the largest CC or the second largest CC. Since betweenness reflects how essential one vertex is to the connection of any other two vertices in the graph, in the case of HMMs, it may reflect the possibility that one HMM is the *hybrid* of two HMMs, that is, between the two HMMs, there is no significant similarity, but through the one HMM, the HMMs can be linked. Biologically, this idea seems to reflect hybrid or mosaic proteins where one protein contains domains from multiple proteins. To our knowledge, the idea of hybrid HMMs has not been discussed previously and deserves more research attention. Moreover, we hypothesize that the HMMs with high centrality measurements may be better able to pick up the sequences that belong to the superfamily than the more peripheral HMMs. Future studies can be directed to test this hypothesis.

Comparison with other networks. The largest CC (590 vertices) of the current network has a diameter of 9 and the average distance between its vertices is 2.94. This bears some similarity to the protein interaction network [6], whose largest CC (containing 5,128 vertices) also has the same diameter of 9, but a larger average distance of 3.68. Thus, the protein interaction network seems to have more vertices that are a bit more spread out, which contributes to a larger average distance. To this point, it is very interesting that despite the big difference in the sizes of the two CCs of the two networks, the diameters are the same.

It is evident that the HMM network is highly clustered. In fact, its clustering coefficient is 0.85, which, to our knowledge, is the highest among the biological networks that have been studied so far. As shown by Newman [9], the undirected networks that tend to have high clustering coefficients are social networks. For example, the film directors network has a clustering coefficient of 0.20 and coauthorship networks for math, physics, and biology disciplines are 0.15, 0.45, and 0.088, respectively, whereas biological networks such as metabolic network and protein interaction network have only a clustering coefficient of 0.09 and 0.07, respectively. The comparison indicates that the current network has distinct features from the previously characterized real-world networks. Also, consistent with its high clustering coefficient, the network has altogether 585 trees (i.e., the CCs of size n

with $n-1$ edges), most of which (566) are of size 2, 15 of size 3, and four of size 4.

Testing the working hypothesis. The results show strong evidence that HMMs in a connected component tend to represent the same family or superfamily. Among the total 1524 CCs, more than 77% have only members from the same family; more than 95% have only members from the same superfamily. Thus, there is overwhelming evidence supporting our working hypothesis that HMMs belonging to the same family or superfamilies tend to cluster together in the network. However, to formally evaluate this and provide some statistical support, we also simulated 10000 random networks while preserving the degree distribution and the number of connected components. Among the 10000 simulated random networks, the highest proportions of CCs having only members from the same family and superfamily are as low as 0.5% and 0.7%. This shows that in the observed network, the HMMs from the same family or superfamily do have a strong tendency to cluster, agreeing with our working hypothesis.

5. Conclusion

In this paper, we examined the properties of the network constructed for HMM models in the SCOP protein structural classification database. A number of questions remain to be addressed in future research. For example, can we devise a computational method to measure or evaluate the degree of redundancy or overlap between HMM models that are used to represent the same superfamily? This research is meaningful given the ever increasing number of large-scale genomic sequences (thereof more protein sequences). Given that we can measure the redundancy of the HMMs of a superfamily, the logical question becomes, can we computationally reduce the redundancy of the HMM library, e.g., possibly by constructing super-HMMs, each of which represents a collection of redundant HMMs, so that a protein sequence is scanned against a reduced set of HMMs (super-HMMs) rather than the entire set of HMMs that have overlaps and redundancies? Finally, because the HMM network shows distinct properties from many documented networks as discussed above, can we propose a theoretical model to better account for the observations in the current network? Moreover, as our HMM network is also weighted, with edges quantifying the similarity between two HMMs, future proposed models can also consider the incorporation of weighted edges into the network.

Acknowledgment

The authors thank T Murali for discussion. The work was partially supported by NSF Grant IIS-0710945 and AFRL Grant FA8650-09-2-3938 and AFOSR Grant FA9550-09-1-0153.

References

[1] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res*, 36(Database issue):D419–25, 2008.

[2] R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. Pfam: Clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–51, 2006.

[3] L.C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

[4] J. Gough and C. Chothia. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res*, 30(1):268–72, 2002.

[5] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol*, 313(4):903–19, 2001.

[6] E.D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

[7] L. Lo Conte, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2002: Refinements accommodate structural genomics. *Nucleic Acids Res*, 30(1):264–7, 2002.

[8] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, 1995.

[9] M. E. Newman. *Networks: An Introduction*. Oxford University Press. Inc., New York, NY, USA, 2010.

[10] J. Soding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–60, 2005.