In Proceedings of The 2004 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'04), pages 283-289.

New Techniques for Generation and Analysis of Evolutionary Trees

Chang Wang, Stephen D. Scott, Qingping Tao, Dmitri E. Fomenko, and Vadim N. Gladyshev

Abstract

We introduce new distance measures for the construction and analysis of phylogenies, focusing on thioredoxin-fold proteins. Our distance measures for tree construction are based on several criteria, including pairwise alignment of only the thioredoxin fold region of each sequence, Hausdorff distance between sequences represented by sets of real vectors derived from per-residue features of the sequences, and properties of each sequence such as protein function and organism type. We also analyze and compare our trees in several ways. To corroborate the trees, we first compute the distance between the evolutionary trees, and then evaluate the trees based on conditional entropy. We also analyze the trees by finding common subtrees within and between our trees. Finally, biological analysis shows that trees based on our measures yield new information on proteins within the thioredoxin superfamily.

Index Terms

thioredoxin-fold proteins, evolutionary tree, Hausdorff distance, most common sub-tree.

I. Introduction

We propose distance measures for the construction of phylogenies and new techniques for their analysis. Use of these measures in conjunction with existing ones allows us to corroborate phylogenetic trees and build new models of families, which can be used in protein function prediction. We analyze our trees through the use of the conditional entropy (CE) measure and by identifying the most common subtrees within trees and between trees. We also study the trees from a biological perspective. For our experiments, we focus on the thioredoxin-fold (Trx-fold) superfamily.

Oxidation-reduction reactions in cells are catalyzed by various redox proteins, many of which use catalytic cysteine residues. Thiol-dependent redox proteins regulate many basic cellular processes, such as DNA synthesis, apoptosis, signal transduction and transcription [13], [5]. To understand the mechanism of cellular redox regulation, the first step is to characterize the specific functions of these proteins [5], [1]. The thioredoxin superfamily is the major family of thiol-dependent oxidoreductases involved in cellular regulation, and its characterization is important for understanding redox processes. In addition to thioredoxin, it includes protein disulfide isomerases, glutaredoxins, nucleoredoxins, peroxiredoxins, glutathione peroxidases and other redox enzymes. In many proteins of the Trx-fold superfamily, two cysteines separated by two other residues form a redox motif designated the CxxC motif. This motif is conserved in the majority members of the superfamily, including thioredoxins, glutaredoxins, protein disulfide isomerases and other proteins. However, some of the Trx-fold proteins conserve motifs in which only one cysteine is

C. Wang, S. Scott, and Q. Tao {chwang, sscott, qtao}@cse.unl.edu are with the Dept. of Computer Science, University of Nebraska, 115 Ferguson Hall, Lincoln, NE 68588-0115, Phone: 402-472-6994, Fax: 402-472-7767.

D. Fomenko {dfomenko@genomics.unl.edu} and V. Gladyshev {vgladyshev1@unl.edu} are with the Dept. of Biochemistry, University of Nebraska, N151 Beadle Center, 1901 Vine St.,Lincoln, NE 68588-0664, Phone: 402-472-4948.

conserved, and in addition, several proteins are known that lack both cysteines and thus lost the redox function.

Quite a few methods are known to build phylogenies, such as parsimony [9], maximum likelihood method [2] and pairwise distance-based methods [19], which compute pairwise distances between sequences and then build a tree that fits them best. We focus on the pairwise distance-based methods. An established method that we employ for constructing phylogenies is to use hierarchical clustering where inter-sequence distance comes from pairwise alignment scores from ClustalW [20]. The pairwise distance in this tree is based on alignments of the primary structure for the entire sequence. In our new approaches, we base our similarity measures on other criteria: pairwise alignment of only the sequences' thioredoxin fold region, Hausdorff distance between sequences when they are represented by sets of real vectors derived from various per-residue properties of the sequences, and other specific properties of each sequence such as protein and organism type. We build our trees via hierarchical clustering. After building the trees using the new distance measures, we also analyze and compare our trees in several ways. We compute the distance between the evolutionary trees, and then we use a measure called conditional entropy to evaluate how well clusters derived from the trees match various designation of the sequences. To gain further perspective on the sequences, we also analyze the trees by finding common subtrees within and between our trees. These analyses and a subsequently biological analysis show that trees based on our new measures yield new information on the protein families.

The rest of this paper is as follows. In Section II, we describe the distance measures and the algorithms we employ in building our evolutionary trees. In Section III, we discuss on how we analyze the trees by computing distances between trees, computing conditional entropy of clusters derived from our trees, and by finding common subtrees within and between our trees. Then in Section IV we summarize our experimental results. Finally, we conclude in Section V with a discussion of future work.

II. Construction of Evolutionary Trees

We use standard hierarchical clustering to construct our evolutionary trees. The differences between established methods and our approach come from the novel distance measures we employ. An established method that we employ for constructing phylogenies is to use pairwise alignment scores from ClustalW [20]. The pairwise distance in this method is on alignments of the primary structure for the entire sequence. In our new approaches, we base our similarity measures on other criteria: pairwise alignment of only the sequences' thioredoxin fold region, Hausdorff distance between sequences represented by sets of real vectors derived from QFC features of the sequences [16], and other specific properties of each sequence such as protein type and organism type.

A. Distance measures

1) Measure based on primary structure and Trx-fold: We first find the primary sequence motif (e.g. CxxC) in each sequence. Since the Trx-fold of all known thioredoxin-fold proteins extends at most 30 residues upstream and at most 180 residues downstream of the motif [12], we discard subsequences outside that 214-residue window. Then we use dynamic programming [3] to align each pair of sequences using the BLOSUM 62 matrix [11]. The alignment score for each pair of sequences is the distance measure we use in clustering.

2) Measure based on per-residue features and Trx-fold: Our second distance measure comes from the Hausdorff distance [14] on sets of points derived from per-residue properties of each sequence. In the QFC algorithm [16], the physico-chemical properties of the amino acids in the residues are characterized using various indices and standard measurements, such as GES hydropathy index [7], [10], solubility [4], polarity, pI, Kyte-Doolittle index [17], α helix index [6], and molecular weight. A protein sequence is described by a set of variables x_1 through x_n, and for each x_i, there is a value x_{ij} for the ith amino acid index (property) value at the jth position of the sequence. Thus x_{i1} through x_{im} constitutes a profile of the protein in terms of the ith amino-acid property index. For each sequence, we extract the Trx-fold as described above, and then map the residues in each fold to their profiles based on the 7 properties of Kim et al. [16], yielding 7-dimensional data. Since each 7-tuple $x_i = (x_{i1}, \ldots, x_{i7})$ in each profile is tied to a particular residue r_{x_i} in the original sequence, we need to add a coordinate x_{i8} to x_i that corresponds to r_{x_i}'s position in the sequence. We set x_{i8} to the index of r_{x_i} in the fold. The end result is a set of points for each sequence. Since the values of different properties vary significantly, we normalize the properties when computing the Euclidean distance. For example, distance between two amino acids x_i and y_j is defined as $\sqrt{\frac{(x_{i1}-y_{j1})^2}{(Max_{p1}-Min_{p1})^2} + \ldots + \frac{(x_{i8}-y_{j8})^2}{(Max_{p8}-Min_{p8})^2}}$, where Min_{pk} and Max_{pk} are the minimum and maximum values for property k.

The Hausdorff distance (HD) measure is a distance between two sets of points [18]. The classical HD measure between two sets $A = \{a_1, \ldots, a_{Na}\}$ and $B = \{b_1, \ldots, b_{Nb}\}$ of sizes N_A and N_B is defined as $H(A, B) = \max(h(A, B), h(B, A))$, where $h(A, B)$ and $h(B, A)$ represent the directed distances between two sets A and B [14]. For the directed distance from A to B, we use $h_K(A, B) = K^{th}_{a \in A} d_B(a)$, where $d_B(a)$ represents the minimum distance value from point a to the set B, and $K^{th}_{a \in A}$ denotes the K^{th} ranked value of $d_B(a)$. In other words, each point in A finds its nearest neighbor in B and vice-versa. Then these distances are sorted independently for sets A and B, and the Kth largest is chosen from each set (using the Kth largest rather than the largest reduces sensitivity to noise). The largest of these two values is the Hausdorff distance between A and B. This measure needs one parameter, $f \in [0.0, 1.0]$, which we use to calculate $K = fN_A$ for use in $h(A, B)$ (for $h(B, A)$, we use $K = fN_b$). Experimentally, when f is about 0.6, good results are obtained [18].

For each pair of sequences, we extract their folds as described earlier, map them to their sets of points, and compute Hausdorff distance between them, using $f = 0.6$ and normalized Euclidean distance to measure distances between points.

3) Measure based on annotation and fold: We also combine each of the two distance measures described above with sequence annotation information, including protein function assignment and organism type. Here protein function assignment includes Thioredoxin, Glutaredoxin, Protein disulfide isomerase, DsbA, DsbC, DsbD, and DsbG. Organism type includes vertebrates, mammals, invertebrates, plants, fungi, bacteria, archaea and viruses. If two sequences have the same type, we multiply the distance between them by 0.8 (thus if two sequences have the same organism type and protein type, we multiply by 0.64). We combine the sequence type information with both distance measures described above. We also use the above two measures each on its own, yielding four total trees.

B. Evolutionary Tree Construction

The method we use to build trees is very similar to that used by Sokal and Michener [19]. The only difference is that we use the complete link algorithm rather than the single link algorithm they use. We first assign each sequence to its own cluster, and define each of them as a leaf with a height zero. Then we merge the clusters until only one cluster exists. To merge two clusters, we first determine the two clusters C_i and C_j for which the distance D_{ij} between them is minimal. We then merge C_i and C_j into a single new cluster $C_n = C_i \cup C_j$, defining the distance between C_n and any other cluster C_k as $D_{nk} = \max(D_{ik}, D_{jk})$. We then replace C_i and C_j with C_n, whose height is set as $D_{ij}/2$.

III. Evolutionary Tree Analysis

We have 5 evolutionary trees: the tree based on the entire primary structure and ClustalW pairwise alignment score (ClustalW tree)[1]; the tree based on fold, primary structure and dynamic programming (DP tree)[2]; the tree based on fold and QFC features (QFC tree)[3]; the tree based on fold, primary structure, dynamic programming and sequence type; and the tree based on fold, QFC features and the sequence type information. For the rest of this paper, we focus on the first three trees; analysis of the last two trees (using sequence type information) is pending.

We now use several methods to analyze and compare our trees that were built using the distance measures of Section II-A. First we compute a distance between each pair of trees by computing the average distances between their most similar subtrees. Specifically, we do the following to compute the distance between trees T_A and T_B. For every subtree t_a in tree T_A, we find the subtree t_b in tree T_B that is the most similar to t_a based on the tree distance measure of Wang et al. [15] (described below). We then divide the distance between t_a and t_b by the number of nodes in t_a. We then take the average of these distances over all $t_a \in T_A$, and average that with the average of the distances from each $t_b \in T_B$ to its most similar subtree in T_A.

We also analyze our trees by using conditional entropy to assess how well each tree matches various classes of the sequences. Define a sequence's *class* to be an external label assigned to it, e.g. the organism that the sequence came from. Then given m classes and k clusters, for a particular class $i \in [1..m]$ and cluster $j \in [1..k]$, we compute p_{ij}, which is the probability that a member of cluster j belongs to class i ($p_{ij} = n_{ij}/n_j$, where n_j is the number of data points in cluster j and n_{ij} is the number of data points in cluster j belonging to class i). The entropy of the class labels conditioned on a particular cluster j is calculated as $E_j = -\sum_{i=1}^{k} p_{ij} \log(p_{ij})$. The conditional entropy is then defined as: $\sum_{j=1}^{k} \frac{n_j * E_j}{n}$, where n_j is the size of cluster j and n is the total number of instances. The entropy equals zero when each cluster contains instances from the same class [8]. For our class labels, we use protein type, organism type, and the pair (protein type, organism type) taken together.

We also evaluate our trees by studying and comparing their subtrees. For every subtree, we compute the sum of the distance between it and all other subtrees. We define the subtree with the smallest sum of distances as the most common subtree inside a set. Here we use Wang et al.'s method [15] to compute the edit distance between two subtrees. This dynamic programming-based method uses both the tree structure and the label of each node. The edit distance from tree t_1 to tree t_2 is the minimum number of edit operations transforming t_1 to t_2. There are three types of edit operations: relabeling, delete, and insert a node. Here the label is the organism type. Every leaf has a label corresponding to the organism of the sequence on that leaf. Internal nodes have no corresponding sequences, so we use the most common organism in the subtree to label them. For example, in the subtree rooted with node A, there are several types of organisms, but if bacteria is the most common, then we label node A as bacteria.

[1] http://genomics.unl.edu/tree/devel2/

[2] http://genomics.unl.edu/tree/CHANG2/

[3] http://genomics.unl.edu/tree/CHANG1/

TABLE I

Comparison of our trees with conditional entropy. "DP" is for measure based on Primary structure and Trx-fold, "QFC" is for measure based on QFC features and Trx-fold, "ClustalW" is for the measure based on ClustalW pairwise alignment and entire primary sequence, "Protein" means using protein type as class label, "Organism" means using organism type as class label, "Combine" means combination of protein type and organism type as class label.

Conditional	DP	QFC	ClustalW	DP	QFC	ClustalW	DP	QFC	ClustalW
Entropy	Protein Type	Protein Type	Protein Type	Organism	Organism	Organism	Combine	Combine	Combine
K=5	0.541623	0.548408	0.560651	0.636842	0.651389	0.652374	1.10114	1.125980	1.14336
K=10	0.474629	0.431362	0.498964	0.584003	0.594431	0.587185	0.967281	0.945759	1.01590
K=15	0.238583	0.337641	0.49186	0.509938	0.565931	0.568910	0.707798	0.834510	0.987126
K=20	0.200595	0.318217	0.381743	0.498040	0.543341	0.471684	0.661368	0.788463	0.779091
K=25	0.192252	0.29005	0.305821	0.486418	0.503726	0.435822	0.643478	0.718837	0.688479
K=30	0.189096	0.275013	0.204412	0.437526	0.493955	0.407707	0.587621	0.693713	0.56885
K=35	0.172943	0.255603	0.180734	0.420016	0.475313	0.378663	0.561356	0.660215	0.515396
K=40	0.168103	0.234351	0.168921	0.398652	0.447900	0.370802	0.535920	0.609090	0.495375

IV. Experimental Results

We selected 1458 sequences from the Thioredoxin-fold sequences database[4] for our experiments. 98% of these sequences have the CxxC motif, while the others have the CxxS motif or no known motif. For the very few sequences without a known motif, we selected two cysteines that were separated by two residues in the primary structure and assumed that the two cysteines and the two residues between them constitute a motif.

Using the tree comparison method described in Section III, the distance between DP tree and the ClustalW tree is 0.2431. Distance between QFC tree and ClustalW tree is 0.2623. This indicates that the DP tree is more similar to the ClustalW tree than the QFC tree is. This is because both DP and ClustalW are based on similar distance measures (primary structure alignment). Further, both the DP and QFC trees are generally similar to the ClustalW tree in terms of similarities of common subtrees. This supports the idea that each tree (especially DP and ClustalW) generally captures the evolutionary patterns of this superfamily.

We now analyze our trees with conditional entropy. The results are in Table I. In the table, K is the number of clusters that we partitioned each tree into based on the nodes' depths. From Table I we see that in general, the DP tree better matches all three labelings of the sequences than the QFC tree does, regardless of K (as expected, DP and ClustalW have mostly similar values of conditional entropy, especially for large K). In addition, from a biological perspective, the DP tree seems to generally represent the evolutionary and functional classes within the Trx-fold superfamily more accurately. However, we found that the QFC tree is a more convenient tool for visualization of some clusters. It better clustered some subtrees within the family of thioredoxin-fold proteins than the DP tree. For example, viral glutaredoxin-like proteins, which are known to be involved in an oxidative pathway of capsid formation, are co-clustered with bacterial thioloxidases DsbA and several thioredoxin-like proteins of unknown function. Thus, these thioredoxin-like proteins could function as thiol/disulfide oxidoreductases involved in disulfide bond formation. Bacterial disulfide isomerases DsbG are co-clustered with mammalian, plant and invertebrate proteins, suggesting that these proteins might rearrange disulfide bonds.

[4] http://genomics.unl.edu/REDOX/tfpdb.html

Clusters of bacterial oxidative folding proteins are also located closer in the QFC tree than in the DP tree. Further, unlike the DP tree, the QFC tree is enriched in clusters containing representatives from diverse organisms, suggesting that it might be useful for finding functional associations and identification of Trx-fold proteins. Thus the diversity of organisms in the QFC tree's clusters (which caused the larger organism-based conditional entropy numbers in Table I) in fact suggests new functional groups in the Trx-fold superfamiliy.

Finally, we study the most popular subtrees inside each tree and between trees taken pairwise. For each test, we identified the five most popular sub-trees. Two of them are in Figure 1. A biological analysis of these subtrees is going on now.

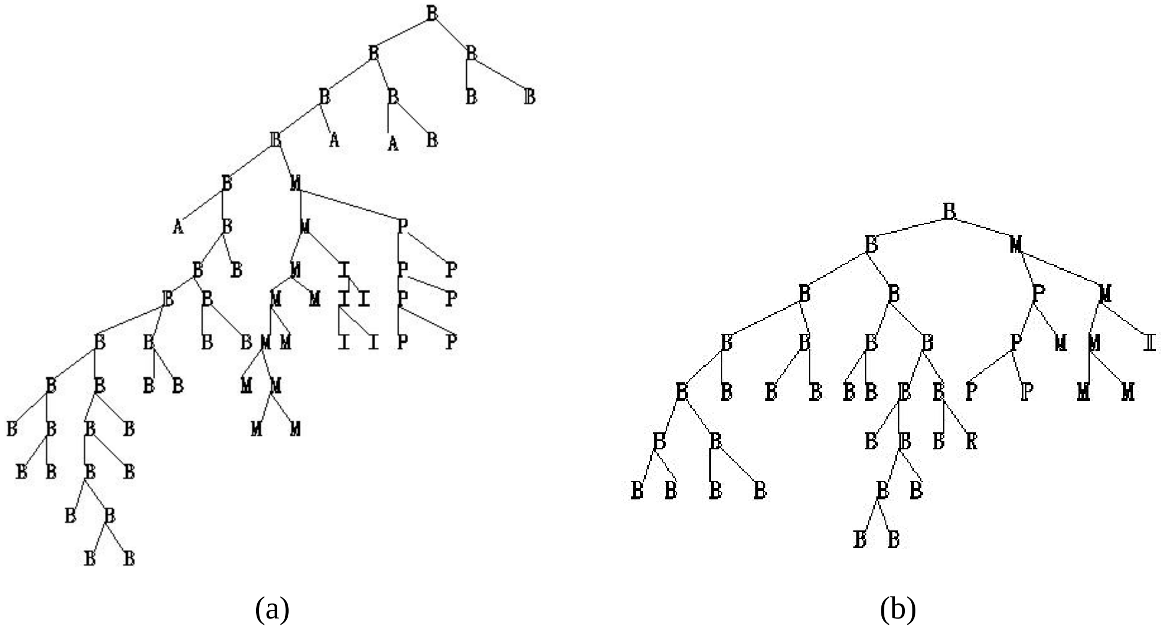

Fig. 1. (a) The most popular subtree in the tree based on ClustalW pairwise alignment. A designates archaea; B bacteria; I invertebrates; P plants; R viruses; M mammals. (b) The most popular subtree in the tree based on QFC features.

V. Conclusion

We proposed methods to construct and analyze phylogenies and applied them to the thioredoxin superfamily. Our methods focus on new distance measures between sequences, computing distances between trees, use of conditional entropy to analyze the clusters implied by trees, and ways to identify common subtrees between trees. Use of these methods in conjunction with existing ones allows us to corroborate existing phylogenetic trees and to build other models of protein families which can lead to prediction of protein function. Future work includes more detailed species-based analysis of subtrees, such as finding subtrees with specific patterns of evolution or representation by specific species or classes of species.

Acknowledgments

The authors thank Gregory Kryukov for his helpful discussions and Etsuko Moriyama for help with QFC. The project described was supported by NIH Grant Number RR-P20 RR17675 from the IDeA program of the National Center for Research Resources. It was also supported in part by NSF grants CCR-0092761 and EPS-0091900.

REFERENCES

[1] F. Aslund and J. Beckwith. The thioredoxin superfamily: redundancy, specificity, and gray-area genomics. *Journal of Bacteriology*, 181:1375–1379, 1999.

[2] D. Barry and JA Hartigan. Statistical analysis of hominoid molecular evolution. *Statistical Science*, (2):191–210, 1987.

[3] R. Bellman. *Dynamic Programming*. Princeton Univversity Press, 1957.

[4] T. Brown. *Molecular Biology Labfax*. Academic Press, second edition, 1998.

[5] L. Debarbieux and J. Beckwith. Electron avenue: pathways of disulfide bond formation and isomerization. *Cell*, 99:117–119, 1999.

[6] G. Deleage and B. Roux. An algorithm for protein secondary structure prediction based on class prediction. *Protein Engineering*, 1:289–294, 1987.

[7] D. M. Engelman, T. A. Steitz, and A. Goldman. Identifying nonpolar traasbilayer helices in amino acid sequences of membrane proteins. *Annual Review of Biophysics and Biophysical Chemistry*, 15:321–353, 1986.

[8] X. Fern and C. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. Wachington DC, 2003. Proceedings of the Twentieth International Conference of Machine Learning (ICML-2003) 186–193.

[9] W.M. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, (20):406–416, 1971.

[10] G. Von Heijne. Membrane protein structure prediction: Hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology*, 225:487–494, 1992.

[11] S. Henikoff and JG. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89:10915–10919, 1992.

[12] J.L. Martin. Thioredoxin–a fold for all reasons. *Structure*, 3(3):245–250, 1995.

[13] A. Holmgren. Thioredoxin and glutaredoxin systems. *Journal of Biological Chemistry*, 264(24):13963–13966, 1989.

[14] D.P. Huttenlocher, G.A. Klanderman, and W.J. Ruchlidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:850–863, 3 1993.

[15] D. Shasha K. Zhang J. T. L. Wang, B. A. Shapiro and K. M. Currey. An algorithm for finding the largest approximately common substructures of two trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):889–895, 1998.

[16] J. Kim, E. N. Moriyama, C. G. Warr, P. J. Clyne, and J. R. Carlson. Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. *Bioinformatics*, 16:767–775, 2000.

[17] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105–132, 1982.

[18] DG Sim, OK Kwon, and RH Park. Object matching algorithms using robust Hausdorff distance measures. *IEEE Transactions on Image Processing*, 8(3):425–429, 3 1999.

[19] R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scienticfic Bulletin*, (28):1409–1438, 1958.

[20] J.D. Thompson, D.G. Higgins, and T.J. Gibson. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, (22):4673–4680, 1994.