BIOMOLECULAR INVARIANTS OF AMINO ACID TREES

By

Debra Knisley

Jeff Knisley

and

Leonard Roberts

IMA Preprint Series # 2190

(February 2008)

INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Biomolecular Invariants of Amino Acid Trees

Debra Knisley[*], Jeff Knisley

Department of Mathematics
East Tennessee State University
Johnson City, TN 37614

Leonard Roberts[†]

Department of Mathematics
Kentucky State University
Frankfort, KY 40601

Abstract

Each of the twenty amino acids are represented by a graph-theoretic tree. Biomolecular descriptors are defined by graphical invariants and are used to determine similarity based on the corresponding Euclidean distances. The results verify that graph-theoretic measures independent of biochemical and biophysical descriptors are valid measures of similarity of amino acids. These results may prove useful for many objectives in proteomics such as protein sequence alignment algorithms and protein folding prediction methods.

1. Introduction

Amino acids have been represented in a number of schematic variations in order to formulate a meaningful method to characterize their physiochemical properties. A large number of these have been compiled and can be accessed in the AAindex database. AAindex is a database of amino acid indices and amino acid mutation matrices. An amino acid index is a numerical value that represents various physicochemical and biochemical properties. For example, alpha-CH chemical shifts [1], the amphiphilicity index [10]and the hydrophobicity indices [2]. In fact, there are 544 indices listed in the AAindex database, all based on some aspect of their biochemical properties. The database may be accessed through the system at GenomeNet [9].

In this work we represent each of the twenty amino acids as a mathematical graph. More specifically, we construct a tree representation for each amino acid. While representing molecules by graphs is certainly not new, as evidenced by the utilization of molecular descriptors or topological indices in the field of computational chemistry, the application of invariants from the mathematical field of graph theory is novel and may prove to be a valuable untapped resource. The tree representations of the amino acids allow us to quantify each amino acid by calculating five graphical invariants. Each of these invariants measure some unique aspect of the graphical structure, both local and global. These invariants are well-known graph theoretic measures that are highly utilized in fields such as computer network design or well studied in graph theory. However, these measures have not previously been utilized as biomolecular descriptors.

We employ what can best be described as a hybrid of techniques from mathematical graph theory and computational chemistry. Since we use graphical invariants as biomolecular descriptors, we refer to the descriptors as biomolecular invariants. This approach proved to be successful in a previous study by the author and others in [7, 8] and is a promising new direction for the modeling and quantification of biomolecules. In this work we show the biomolecular invariant values of the tree representations of the amino acids can correctly classify the amino acids into functional and structural groups that correspond to their known biochemical properties. Since no information regarding their chemical composition, polarity, hydrophocity or other classifiers was utilized, it is apparent that structural properties of the molecules play a dominant role. This is not to say that these biophysical properties are not of utmost importance, but rather that this study may imply that information regarding many biochemical properties may be inherently encoded in and determined by the molecular structure. We show that structure alone, when quantified by graph-theoretic means, is sufficient to provide information about the bioactivity of the amino acids.

In the following section, we describe our modeling method and define the graphical invariants that

[*] Supported in part by the Institute for Mathematics and its Applications with funds provided by the NSF

[†] and by the MAA SUMMA program with funds provided by the NSF, the NSA and the Moody Foundation

we utilize. In section three we analysis the results by defining a graph that represents the amino acid vector space. Each vertex of the graph represents an amino acid and all possible edges belong to the edge set. Each edge is weighted by the Euclidean distance of the invariant vectors. We then analyze the minimum spanning tree of the resulting complete graph and obtain some interesting observations.

2. Biomolecular Invariants of Amino Acid Trees

In this work, we use simple connected graphs that are small ordered trees. Tree representations of biomolecules have been successfully utilize by Schlick et. al [4, 5] and further quantified by Haynes et al, to predict novel RNA structures [7]. Specifically, we use a tree representation of the twenty amino acids. The tree structures can be thought of as a skeletal representation of the structure. Amino acids such as D and E have the same underlying skeleton, whereas T's structure introduces a branch. The structures are further distinguished by weighted vertices. The resulting 20 trees are shown in Figure 1. The dual edge for P is collapsed for the final tree representation. Since all amino acids have the identical backbone, a weight of zero is assigned to the backbone and used as a root of the tree. We assign a weight of two to CH2, and 3 to CH3 while a benzene ring receives a weight of eight and nine if a sulfur molecule is present. We also assign 10 to an -OH group, 11 to a carboxylic group while an amide group is weighed seven. Finally, four, five and six are the weights of long carbon chains containing nitrogen groups.

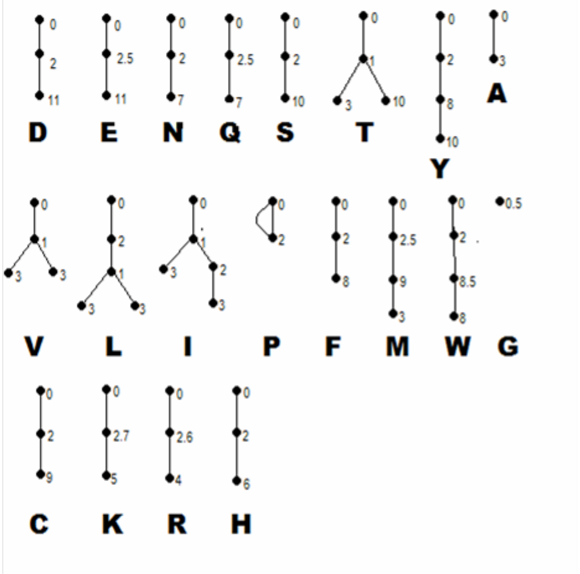

Figure 1: Graphs of Amino Acids

We define five graphical invariants from the weighted trees that are based on the biomolecular structure of each amino acid. The *diameter* of a graph is the length of the longest path in the graph and is labeled I_1. Since the graph is weighted, we sum the vertices along this longest path. The values for the diameter for T, I and D in Figure 2 are 11, 6 and 13 respectively . Hence, by longest path, we mean the path of maximum weight. A vertex of degree one, that is a vertex that is incident to exactly one edge, is called a peripheral vertex. For the second invariant I_2 we calculate the sum of the weights of the peripheral vertices. Hence we obtain I_2 equals 13, 6 and 11 for T, I and D. A *dominating set* of vertices in a graph is a set of vertices with the property that all the remaining vertices in the graph are adjacent to at least one vertex in the dominating set. The *domination number* of a graph is the minimum cardinality among all dominating sets of vertices. The idea of domination is based on sets of vertices that are near (dominate) all the vertices of a graph.

As an example the application of domination in computer network design, suppose each vertex of the graph represents a computer and two computers are adjacent if there is a direct link between them in the network. Some of the computers are designated as file servers to house the programs for the entire network. If the file servers are selected in such a way that every computer is either a file server or has a direct connection to a file server, then the set of file servers is a dominating set. The minimum number of file servers required so that every computer in the network has access to one is the domination number of the associated graph. Since our graph is weighted, the *weighted domination number* is the minimum sum of the vertices in a minimum dominating set and is denoted by I_3. The I_3 values for T, I and D are are 1, 3 and 2. Notice that the tree representation of I has more than one dominating set. However, the sum of 1 and 3 is greater than the dominating set whose sum is that of 1 and 2 and hence we select the minimum weight. The red arrows are pointing to the selected vertices of the minimum weighted dominating set. For more information on the domination number of graphs see[6].

The fourth invariant I_4 is defined as the average of the edge weights where the *weight of an edge* is defined as the sum of the two end vertices of the edge. For example, the average of the edge weights of D is equal to 7.5 since [(0+2) +(2+11)] 2 (or number of edges) = 15/2 =7.5. Finally, we find the minimum weight of all paths of length two in each tree to be I_5

The tree representation of I has four distinct paths of length two with weights of 4,3, 6 and 6. Hence, I_5(I) = 3.

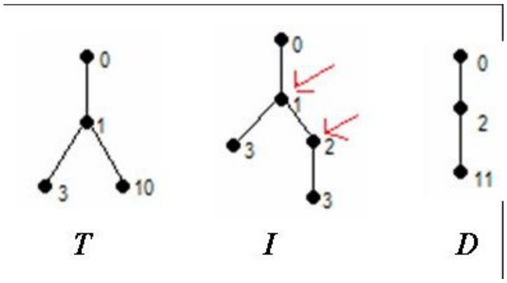

Figure 2: TID

We now us the five graphical invariants to form vectors $\mathbf{v}_a = \langle I_1(a), I_2(a), I_3(a), I_4(a), I_5(a)\rangle$ for each amino acid, a. In order to quantify similarity, we calculate the Euclidean distance between the vectors $\mathbf{v}_a$ and $\mathbf{v}_b$ corresponding to a pair of amino acids, a and b, where the Euclidean distance is given by

$$dist(\mathbf{v}_a, \mathbf{v}_b) = \left(\sum_{j=1}^{5} (I_j(a) - I_j(b))^2\right)^{1/2}$$

We subsequently construct a complete graph with twenty vertices and label each of the edges with the Euclidean distance between the corresponding amino acids. In order to deduce relationships between the amino acids from the distance graph, we implement Kruskal's algorithm to find a minimum spanning tree of the complete graph for the amino acids. We find that an analysis of the minimum spanning trees provides meaningful information about the similarity and differences of the amino acids. Not only are structural similarities revealed, but properties that dictate amino acid substitution preferences are also surpisingly displayed. We discuss this in more detail in the following section.

3. Information revealed by the biomolecular invariants

An analysis of the minimum spanning trees reveals functional as well as structural relationships. The distance measure in Tree 1 utilizes all five of the invariants.

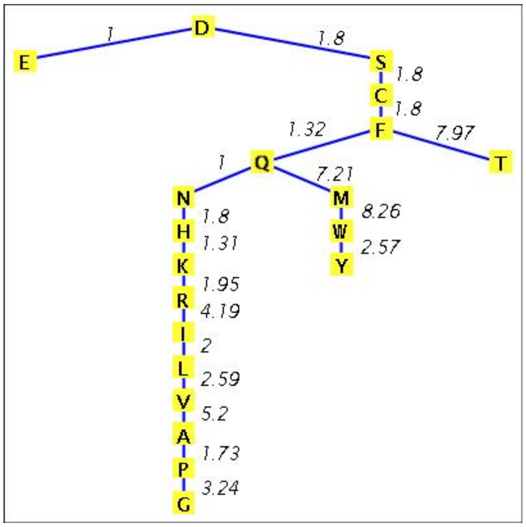

Figure 3: Spanning Tree Incorporating All 5 Invariants.

In contrast, the distance measure in Figure 4 omits the weighted domination,

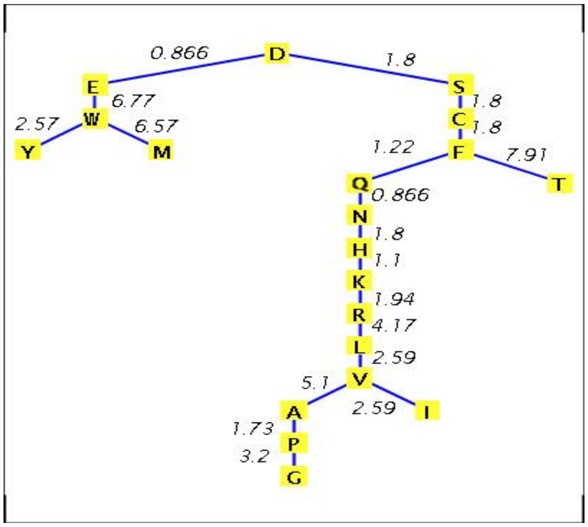

Figure 4: Weighted domination omitted.

the distance measure in Figure 5 omits the diameter

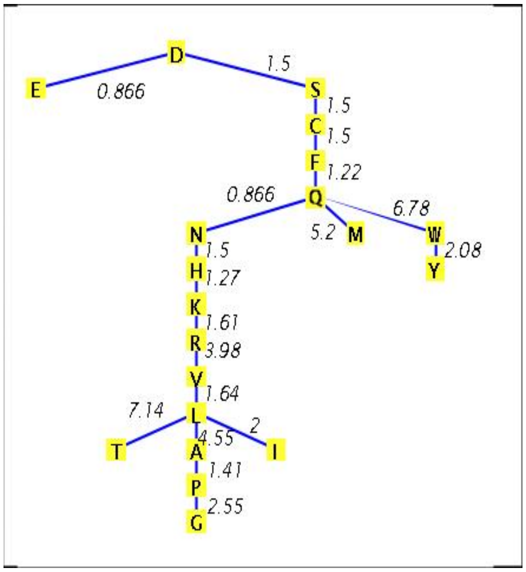

Figure 5: Diameter omitted.

and the distance measure in Figure 6 omits both the diameter and the weighted domination.

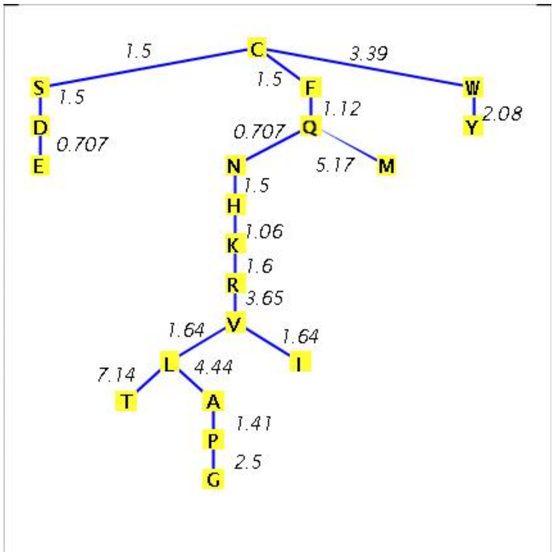

Figure 6: Diameter and Weighted Domination Removed

Figure 3 is the resulting spanning tree when the Euclidean distance is defined in terms of all five invariants. In order to assess the significance of the contributions of each of the invariants, we also calculated the distances by removing each of the invariants, one at a time and recalculating with the remaining four. Figures 4 and 5 are two examples of the distance based on four invariants.

In the following discussion, we primarily analyze the information contained in the spanning tree in figure 6 which is based on the three invariants of average path length, peripheral vertex sum and minimum sum for paths of length two. Figures 3, 4, and 5 are included as illustrations that augment the discussion.

We root the spanning tree with the amino acid Cysteine. We do this because C is unique in that there are two distinct states of C, namely C-SS and C-SH. It is possible for two Cysteines to be bound to each other by disulphide bonds, and hence the C-SS state. These bonds are very unusual in intracellular proteins and thus this state is found primarily in extracellular or membrane proteins. Conversely, we find the free state C-SH primarily in intracellular proteins. In our representation we selected the free state.

As we expected, we see in Figure 6 that W and Y constitute a subtree with the two vertices, labeled W and Y. It is worth noting here that Y and W are the only two non charged polar amino acids. We also find a subtree of S consisting of only D and E and note here that D and E are the only two negatively charged aromatic amino acids. However, no knowledge of polarity was used in defining these invariants. In fact, if we follow the left subtree of C in Figure 6, we note that S is also a very small, polar, non-charged amino acid. S is associated with D in the fact that they are both polar and small and as mentioned earlier, D is associated with E since both D and E are negatively charged. Looking elsewhere in the tree, we observe that F is unique in that it is the only aromatic non polar amino acid. When investigating the properties of Q, we discover that N,H, K and R all are favored subsitutions for Q in membrane proteins, while M is disfavored. What is also noticeable in comparing Figure 3 with Figure 4 is the translocation of the subtree containing M, Y and W. In the tree that omits the domination invariant, these three amino acids are more closely associated with E, whereas in Figure 3 they are more closely associated with Q, F and N. We find that M is a disfavored substitution for E in the membrane and intracellular proteins, but it is considered a neutral substitution for the extracellular proteins. Thus the domination number is indicative of this variation in M with repect to E. Another interesting tree translocation ocurrs when we omit the diameter as in Tree 3. Here we observe that T, which was more closely associated with F is now more closely associated with L, V and I. Since L, V and I are all hydrophobic and F is not, this would indicate that the diameter is an important index to correctly identify characteristics of T. Knowing this, we can then consider why this would be so. A glance at the tree representations reveal that these four all have similar branching and thus we conclude that this was captured by the diameter descriptor. Consequently, the branching is an important factor that we can associate with the hydrophocity of the amino acid. But another perhaps more interesting observation can be made with respect to Figure 6. T has an additional property that if frequently overlooked. Threonine(T) contains two non-hydrogen substituents attached to the C-beta carbon. Interestingly enough, the only other two amino acids with this property are V and I. A suspected consequence of this additional bulkiness near the backbone is the implied restrctions in the conformations the manin chain can adopt. For more information on amino acid properties see [3]

4. Conclusion

In this work we introduced a novel method to quantify amino acids using graph theory to define biomolecular invariants. The invariant values that are calculated for the biomolecules are more than classifiers. Indeed, they reveal potential reasons as to *why* the biomolecules are associated (or not) with each other since the significance of a particular invariant implies the significance of some aspect of the molecular structure. These associations are transitive as well. This highlights a valuable aspect of this line of investigation. Information obtained by the graph-theoretic model may assist in the development of novel algorithms that incorporate structural information in a new way. It has been shown that algorithms based on the minimum free energy assumption alone are not always correct and hence the value of this approach is certainly merited. We hope that this new line of research may provide an entirely different strategy for researchers in proteomics and genomics.

References

[1] Anderson NH, Cao B, Chen C, Peptide/protein structure analysis using the chemical shift index method: upfield alpha-CH va;ies revea; dynamic helices andaL sites, Biochem and Biophys. Res. Comm. **184**1008-1014, 1992.

[2] Argos P, Rao M, Hargrove P, Structural Prediction of Membrane-Bound Proteins, E. Jour. of Biochemistry**128** 565-575, 1982.

[3] Betts M, Russell R, Amino acid properties and consequences of substitutions, IN:Bioinformatics for Geneticists, Barnes and Gray eds, Wiley, 2003.

[4] Gan, Fera, Zorn, Shiffeldrim, Tang, Laserson, Kim and Schlick,RAG: RNA-As-Graphs Database, Bioinformatics **20**, 1285-1291, 2004.

[5] Gan, Pasquali and Schlick, Exploring the repertoire of RNA secondary motifs using graph theory with implications for RNA design, Nucleic Acids Research**31** 2926-2942, 2003.

[6] Haynes T, Hedetneimi S and Slater R, Domination in Graphs, Marcel Dekker, 1998.

[7] Haynes T, Knisley D and Knisley J,Using a Neural Network to Identify Secondary RNA Structures Quantified by Graphical Invariant, MATCH: Communications in Mathematical and Computer Chemistry **60**,2008.

[8] Haynes T, Knisley D, Seier E and Zou Y, A Quantitative Analysis of Secondary RNA Structure using Domination based Parameters on Trees, BMC Bioinformatic**7**,2006.

[9] Kawashima S and Kanehisa M, AAindex Database, Nucleic Acids Research, **28** (http: www.genome.ad.jp/aaindex/)2000.

[10] Mitaku S, Hirokawa T and Tsuji T, Amphiphilicity index of polar amino acids as an aid in the characterization of amino aci preference, Bioinformatics **4** 608-618, 2002.