Adaptively-Branching Fuzzy Greedy K-mean Decision Forest (FGK-DF) Model for Protein Local Tertiary Structure Prediction

Bernard Chen¹*, Cody Hudson¹*, Aaron Crawford¹*, Minwoo Kim¹* and Dongsheng Che²

Received 12 December 2013; Published online 22 March 2014

© The author(s) 2014. Published with open access at www.uscip.us

Abstract

For the past twenty years, protein tertiary structure research has been given much attention, with solutions existing for both wet lab procedures (x-ray crystallography and NMR spectroscopy) and bioinformatics approaches (threading, homology-modeling, and de novo). Unfortunately, each approach has significant shortcomings, such as necessary time, capital, expertise (for wet lab procedures) or restrictions imposed by the method, limiting the resolution or novelty of produced tertiary structures (for bioinformatics approaches). This work propose the Adaptively-Branching Fuzzy Greedy K-means-Decision Forest (FGK-DF) model, which utilizes conserved sequential and structural motifs that transcend protein family boundaries, to predict the local tertiary structure of proteins with unknown structures. In this work, the FGK-DF model is conceptually compared against existing approaches and explicitly compared against the Super Granule Support Vector Machine approach (Super GSVM), with accuracy and coverage results highlighted.

Keywords: FGK Model; Decision Forest; Adaptive Branching; Protein Sequence Motif

1. Introduction

For the last two decades, much attention has been given to the problem surrounding the fine-grained determination of protein function. Maciej et al (2014) mentioned that since the knowledge of the three-dimensional structure of proteins is crucial for understanding many important biological processes. Historically, this process has been accomplished using so-called “wet lab” procedures such as x-ray crystallography by Spec (2004) and nuclear magnetic resonance (NMR) spectroscopy by Sanders et al (1998), which do not determine the function of the proteins directly but rather determine the tertiary structure of proteins. Karp (2009) pointed out that as the function of the protein is almost directly related to its structure, determining the structure of a protein is
closely related to determining its function. Despite this correlation, these historical methods have proven to be slow and expensive to the effect of determining protein function and structures on a large scale.

In response to the difficulty of explicitly determining the tertiary structure of proteins, as mentioned in Nair’s overview (2007), many alternative approaches, which are known as “Protein Tertiary Structure Prediction”, have been proposed in bioinformatics researches; in these approaches, the tertiary structure was not explicitly determined, but approximated using a protein’s primary sequence (i.e. amino acid sequence) and the underlying concept that a protein’s primary sequence almost solely determines is tertiary structure. Development in this field led to the creation of three general families of tertiary structure prediction models, including homology-modeling, threading and de novo, developed by Zhang (2008), Lathrop (1994) and Hardin et al (2004), respectively. Unfortunately, each approach, just as the wet-lab procedures, brought with them a new set of impracticalities, including incomplete modeling, protein size restrictions, inability to generate novel fold predictions, etc.

Thus, given the inherent problems associated with wet-lab procedures and with established bioinformatics models, this paper would propose an alternative solution known as the Fuzzy Greedy K-means Decision Forest model (FGK-DF). In short, the base FGK-DF model is a hybrid template-based approach that uses subtle, conserved primary sequence motifs that transcend protein family boundaries to correlate unknown input proteins to a known tertiary structure. It does this on a local tertiary structure level rather than the conventional global tertiary structure level, which not only affords the algorithm a much higher resolution at the prediction level but also allows the algorithm to provide semi-novel folds, a great improvement over the tremendous weakness of conventional template-based approaches.

The rest of this paper will explore the FGK-DF model in depth, presenting background information concerning protein structure as well as structure elucidation both in terms of historical approaches and through bioinformatics, followed by the specifications of the model’s methodology as well as the experimental setup and results that are used to support the validity and use of this model for the purpose of predicting protein structures. As such, the following sections of the paper will be arranged as follows: background, methods, experimental setup, results, future works, and conclusion.

2. Background

2.1 Protein Structure Taxonomy
Protein structure can be broken down into four primary categories, three of which this research deals with directly: primary sequence, secondary structure, and tertiary structure. The primary sequence of a protein, described as an ordered sequence of chained amino acids, forms the basis for the protein and is a major determinant of the resulting structure of the protein. The arrangement of the constituent amino acids can cause small, highly regular, repeating substructures known as the secondary structure of the protein to develop, often in the form of three overarching motifs: alpha-helix, beta-sheet, and coils. These motifs arise from complex intermolecular reactions between components of the primary structure as they form bonds, creating folds and other such structures that appear as the secondary structure.
Once the secondary structure is stabilized, the tertiary structure, the third category of protein structure, begins to take shape in a process known as folding. A protein folds due to intermolecular forces as well as environmental conditions, such as the presence of water or temperature, and eventually reaches a stable form known as the “native state.” This native state describes the “correct” tertiary structure of the protein. As such, this is the most pivotal category of structure as it is the sole determinant of the function of the protein. It has also been the most elusive, historically, as the next section will explore.

2.2 Wet-Lab Procedures: X-Ray Crystallography and NMR Spectroscopy

X-ray crystallography, the de facto standard approach to protein structure elucidation, relies on firing x-rays through crystallized macromolecules, such as proteins, and recording the resulting diffraction pattern. This pattern can be analyzed mathematically to produce the structure of the macromolecule. This requires that the molecule can be crystallized, which can exclude vast portions of proteins such as those associated with membrane functions. Further requirements include an extensive amount of time, effort, training, and money. Furthermore, Spek (2004) revealed that imperfections in the process can lead to unusable resolutions of the structures, making the already complex process frustratingly impractical in many cases.

Alternatively, NMR spectroscopy utilizes the properties of magnetic fields produced by the spin of charges in atomic structures to produce information concerning their chemical and physical properties. Sanders et al (1998) extended to determine the structure of molecules. Unfortunately, this method can generally only be applied to much smaller molecules and proteins, though it is often times the only method by which one can experimentally determine unstructured proteins. Perhaps more so than x-ray crystallography, this method is extremely expensive, requiring massive machinery and considerable expertise to analyze or even produce the results.

Given the advances of computing power and its utilization in research, X-Ray crystallography and NMR spectroscopy have become unacceptably impractical for the purpose of determining protein structure and function. In response to this impracticality, three core approaches have been developed within bioinformatics that describe the efforts in recent years to develop an automated solution to predict the structure, and thus function, of proteins. It is from these three, described in the following section, which the FGK-DF model extends from.

2.3 Homology-Modeling, Threading, De Novo

According to Nair (2007), there are three distinct approaches to automated protein structure prediction: homology-modeling, threading, and the de novo approach. Each of these approaches, explained in the following paragraphs, have their advantages and, just as with the wet-lab procedures, have their disadvantages.

Zhang (2008) developed a method named Homology-modeling attempts to exploit the mechanism by which protein evolution operates, such that proteins which share an evolutionary ancestor are said to have similar tertiary structures. This is in part due to the fact that there are only three primary ways a protein sequence (and, subsequently, its structure) can change over time, and that is through insertions, deletions, and swaps in the amino acids in its primary sequence. These three types of mutations lead to the various branches within a given protein family, such that all the branches, presumably, lead to a “root” protein (i.e. the shared evolutionary ancestor). Each protein
that is descendant of that “root” protein is just a set number of swaps, insertions, and deletions away from the root. This suggests that to relate one homologue to another one can simply work their way to the root and then back down again to the target homologue through a unique series of mutations.

Given this, homology-modeling attempts to best identify and align those homologues by simulating the mutations through, typically, a variation on string alignment with the primary sequences of target and candidate sequences. This requires a given input protein have an existing homologue to produce a valid prediction. Furthermore, as the structures predicted by homology-modeling are generated from existing structures correlated through protein mutation, no novel structures can be produced from this method.

Threading, conversely, determines a candidate sequence according to similarities in the folding of the tertiary structure between the target and candidate proteins from enormous databases. In more explicit terms, the approach attempts to model the target protein by aligning, or “threading,” an unknown protein’s sequence “to a known structural motif”. This requires one to have a database of “spatial folding templates,” to perform the prediction process. In its purest and most naive form, the unknown protein’s sequence is aligned one amino acid at a time against these templates until a best fit is found. This best fit, being one of the aforementioned templates, has the corresponding “structural motif,” or fold, that is said to be the tertiary structure of the unknown protein. In other words, based on the paper written by Lathrop (1994), the threading approach “recognizes the protein sequences likely to fold into similar structures”. Just as with homology-modeling, no novel structures can be produced through this method.

The other method, de novo modeling developed by Hardin et al (2004), takes a radically different approach to predicting protein structures, in that the mechanisms by which protein folding occurs is simulated rather than the base structures being predicted through structural correlation. The simulation environment can be given an input of only the primary sequence of a protein, from which it simulate the folding of the protein and, thus, its final native form. The simulation environment itself can be generated through sampling a “conformation space,” (i.e. the possible and expected structure of the proteins given constant conditions), from which possible structures are generated, scored, and refined. This, in turn, is supported by a plethora of mathematical models and equations that model physical laws, free energy minimization, water and amino acid hydrophobicity, etc. The sheer complexity and range of approaches contained under the de novo approach prohibits further discussion in this work on the matter, but it should be noted that the same complexity prohibits the de novo approach from being applicable to all but the smallest proteins. Furthermore, template-based modeling, such as the aforementioned threading and homology-modeling, created by Bonneau and Baker (2001) has consistently outperformed de novo approaches in the past.

3. Methodology

In explicit terms, the methodology of the FGK-DF includes the following tasks: the model trains itself on a large, non-homologous training dataset, granulizing and clustering the information in the training set based on the shared presence of sequential motifs. From there, decision trees are trained on each cluster, utilizing both primary and secondary information such that each tree can
discern if a target protein segment contains the sequential motif within its primary sequence. Once the entire forest of decision trees are trained, they can be searched according to primary sequence “distance” to find the best fit, and then using the decision tree to decide if the target protein contains the motif that characterizes the cluster the tree is trained on. If it does, the tertiary structure is predicted to be the average structure of either the cluster or a given branch on the tree, depending on the particular setup of the model. This is repeated for each target protein, such that the end product is a cheap, accurate, and quickly determined tertiary structure for a vast number of protein segments (and thus proteins themselves). This is shown in the following figure:

Thus, given the above, the FGK-DF model operates in five explicit steps: (1) extract the data set, (2) granulate the data set according to primary sequence information from the input set; (3) finely cluster the granules according to primary sequence information such that each cluster is representative of conserved sequential motifs; (4) train decision trees for each cluster that can detect if an input protein segment contains the sequential motif represented by the underlying cluster; and (5), given an input protein segment, search through the forest of decision trees for the tree that is most likely to contain the sequential motif of the input protein segment, and display the predicted tertiary structure based on the representative tertiary structure of the tree. Each of the following sections will explore these five tasks in depth.
3.1 FGK-DF Dataset: PISCES, HSSP, DSSP

At their respective cores, most computational systems that handle protein data owe their information to the existence of the Protein Data Bank (PDB) proposed by Berman (2008). Brought into existence in 1971, the database streamlined the process of structure information exchange by producing a centralized store of standard format, starting an era of structure research based on the free and open exchange of protein information to anyone with an internet connection. Although the PDB started out modestly (in 1976, less than thirty protein structures had been archived in the database), by 2006 the number of structures archived was over 40,000, such that many of the new structures were much more complex and detailed than the older structures.

Like most approaches, the FGK-DF model does not interface with the PDB directly, but makes use of several additional servers and databases to filter, cull, and expand the information provided by the PDB. The services used are based on core data requirements of the FGK-DF approach: primary sequence information of proteins that hail from disparate protein families, that said primary sequence information be in the form of frequency profiles, which is developed by Gribskov et al (1987), and the need for both secondary and tertiary structure information for FGK-DF model training purposes. As such, in order to develop the dataset used for training and validation of the model, the Protein Sequence Culling Server (PISCES) based on Wang and Dunbrack (2003), Homology derived Secondary Structure of Proteins (HSSP) based on Sander and Schneider (1991), and Definition of Secondary Structure of Proteins (DSSP) based on Kabsch and Sander (1983) databases/services are used, respectively, to produce the culled protein segments, frequency profiles, and secondary structures to the develop the necessary dataset for the training and execution of the FGK-DF model. The similar approach is also adopted by Sander et al (2006), and produced promising local protein structure prediction results.

Given the sources for the data used by the FGK-DF model, it is important to understand what each service provides in terms of the actual logic behind the model. One of the core ideas that the FGK-DF model is built upon is that there are sequential motifs that transcend protein family boundaries. As such, the dataset that is being generated must be from proteins that do not share significant homology. The PISCES service can cull proteins based on an upper limit of sequence similarity, or “identity.” In this case, the PISCES service was used to cull 2,710 proteins that share no more than 25% sequence identity. As this is list is culled from the PDB, the data itself contains both primary sequence information and tertiary structure information. In order to transform the primary sequence information into a frequency profile, the HSSP service is necessary.

In a sentence, the HSSP uses multiple sequence alignment to produce the frequency profiles for proteins found in or culled from the PDB. A frequency profile describes the frequency of occurrence of each amino acid at each position in a protein, such that for one position of a protein sequence, there would be twenty (one for each amino acid) values associated with that position describing the relative percentages of occurrence of each amino acid at that particular position. Typically, a frequency profile represents up to w positions, where w is known as the window size of the frequency profile. As the window size is typically less than the size of a protein sequence, a sliding window technique can be utilized, such that i…w positions’ amino acid percentages are recorded in the frequency profile, followed by i+1…w+1 position’s amino acid percentages, and so on until the protein has been completely decomposed into segments. This allows the FGK-DF model to not only directly compare protein segments in terms of segment “distance,” but it also allows the model to
operate on local protein structure rather than global protein structure. In this work, w is set to nine, generating over 560,000 segments from the 2,710 culled proteins from PISCES. The secondary structure, described in terms of helices, sheets, and coils, is generated from the DSSP and concatenated with the frequency profile and tertiary structure information for each segment.

This data process is carried out for both the training and testing set that is generated in this work. While no further processing is performed on the testing data set, the training set is not yet in the necessary organization for utilizing the data. Each of the next three steps describes the necessary processes to generate the underlying model for the FGK-DF, beginning with granulating the data.

3.2 Fuzzy C-Means: Producing Granules
Granule computing, proposed by Lin (2002), is utilized to break a larger set of data into subsets, noted as “information granules,” to allow for parallel execution. This is required in the case of the FGK-DF model due to the size of the dataset (i.e. the 560,000 segments). To generate these information granules, the Fuzzy C-Means algorithm (FCM) created by Bezdek (1981) is used. FCM works much like the popular clustering algorithm K-Means, only that membership to each cluster is determined in a fuzzy manner rather than a static manner. FCM uses two primary equations: an equation (equation 1 below) for determining a degree of belonging to the cluster, and an averaging mechanism (equation 2) for determining the centroid of the cluster. These two equations, respectively, take the following forms:

\[
 u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{1}{\left| x_i - c_k \right|^{2/m}} \right)^{1/m}} 
\]

\[
 C_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} 
\]

**Eq. 1 and 2:** Fuzzy Degree of Belonging and Centroid Calculation

In the above equations, \( u_{ij} \) is the degree of belonging of the data member \( x_i \) in the cluster \( j \). \( C_j \) describes the centroid of the cluster \( j \). \( C \) describes the number of clusters, \( N \) describes the number of data members, and \( m \) is the "fuzzification factor" developed by Dr. Zadeh (1965), which determines the weight of the fuzzy logic as it takes place in the calculations. "\([\ldots]\)" simply describes the distance formula that is used to determine the similarity/dissimilarity of a data member to a given centroid. Just like with K-Means clustering, FCM begins with randomly selected points, determines membership to each of those points, calculates centroids, determines membership, recalculates centroids, and so on until the centroids no longer move or the change falls below a given threshold.

Thus, the FGK-DF uses FCM to produce distinct information granules, setting the fuzzification factor to 1.05 (\( m \) can range from 1.00 to infinity) and the number of information granules to 10, based on the results generated in previous work. The distance formula (equation 3) utilized is based on that used by Baker et al (2001), described as the “city block metric.” This formula is shown below:

\[
 \text{Distance} = \sum_{i=1}^{L} \sum_{j=1}^{N} |F_k(i,j) - F_c(i,j)| 
\]

**Eq. 3:** City Block Metric
This formula, as applied to the extracted frequency profiles, states that distance is equal to the summation of the difference between the frequency value in data member \( F_k \) and \( F_c \) for each of the twenty amino acids (N) in each of the nine positions described by the window size (L). This distance formula is used in the FCM formulas described above, generating the “distance,” or difference that is exhibited between two frequency profiles for two given segments in the training data set. Using this formula, the distance threshold for the FCM (i.e. the allowed difference between two protein sequence segments) is set to 13%, which roughly translates to the omission of 15% of the outlying data that could not be clustered to any of the centroids, resulting in ten trimmed, roughly clustered granules. The next process takes these granules as input and further clusters them, implicitly deriving sequential motifs through the process of Fuzzy Greedy K-Means.

3.3 Fuzzy Greedy K-Means: Producing Clusters

In related work, Zhong et al (2005) proposed an “improved” K-Means algorithm to resolve the initialization problem of traditional K-Means for protein sequence motif. The algorithm had two main steps for generating initial centroids: generate centroids by running traditional K-Means for a fixed number of iterations, then determine if the generated centroids could be added as viable initial centroids based on secondary structural similarity and their distance to other initial centroids. This was run until the number of viable initial centroids was equal to the number of required clusters, \( k \), after which traditional K-Means was run given the selected initial centroids. The distance measure was based on the “city block metric” formula described in the preceding section, while the secondary structure similarity was based on the following equation:

\[
\sum_{i=1}^{ws} \max(p_{iH}, p_{iE}, p_{iC})
\]

**Eq. 4:** Secondary Structure Similarity Measure

In the formula, \( p_{iH} \) describes the frequency of helices in the protein segments in the cluster at position \( i \) for each of the nine positions of the window size \( (ws) \). \( p_{iE} \) and \( p_{iC} \) describe the frequency of sheets and coils, respectively, in the same manner. \( \max(\cdot) \) returns the maximum frequency of the three measures.

Zhong et al.’s Greedy K-Means algorithm served as the basis for the Fuzzy Greedy K-means algorithm developed by Chen et al (2009). More explicitly, the Greedy K-means aspect of the FGK follows the algorithm proposed by Zhong et al., except that a dynamic threshold for required secondary structure similarity is used in place of a static threshold. In addition, the number of iterations is fixed to five runs of traditional K-Means, where each respective run had a secondary structure similarity cutoff of greater than 80%, 75%, 70%, 65%, and finally 60%. Sander and Schneider (1991) claimed that these values were based on two ideas: clusters of protein segments with a secondary structure similarity of greater than 70% can be considered “structurally identical”. In this work, the FGK algorithm is applied to the proteins in each of the generated granules produced by FCM. In order to determine and balance the number of clusters for each granule, an additional equation is needed:

\[
C_k = \frac{n_k}{\sum_{i=1}^{m} n_i} \times p
\]

**Eq. 5:** Granule Size
In this formula, $C_k$ refers to the amount of clusters assigned to a given information granule ‘k,’ where $n_k$ is the number of data members contained within said granule. $M$ refers to the number of granules defined for the FCM run. $P$ refers to the total number of clusters, which is set, in this work, to 799, based on the research and results performed by Zhong et al (2005).

### 3.4 Decision Tree Induction

Given the last step, one could use the clusters generated by the FGK algorithm to generate protein structure predictions, as the clusters themselves denote a shared sequential motif between their member frequency profiles. However, clusters are not practical to perform direct calculations and searches on. As such, the next step in the FGK-DF model is to train a decision tree on each cluster such that the generated tree can decide if an input protein sequence shares the sequential motif denoted by the underlying cluster. To accomplish this, the Itemized Dichotomizer 3 (ID3) decision tree is utilized.

The ID3 algorithm, presented by J.R. Quinlan (1986), builds decision trees based on the minimization of entropy (i.e. the randomness of the dataset) at each branch in the tree. Within the context of the algorithm, there are three main concepts: labels, attributes, and values. The labels, in the case of the FGK-DF model, ultimately state whether or not (“yes” or “no”) a given input sequence belongs to the cluster that the decision tree is trained upon. To decide whether or not a data member in the training set produces a “yes” label or a “no” label, the secondary structure of that data member is compared to the average secondary structure of the cluster to produce an individual secondary structure similarity score (Equation 4) for that data member. If it is greater than a certain threshold, noted as a "label pivot" in this work (set as 70% secondary structural
similarity), it is denoted with a "yes" label. If it is less than the label pivot, it is given a "no" label. The attributes used for the tree are based on the 180 possible positions in the frequency profile, such that the values (0-100) are ranges of the possible values each position in the frequency profile. An example decision tree produced by the FGK-DF model might look like the figure 2.

In figure 2, the value in each circle denotes a position in the frequency profile, the values above each circle denote the attribute range needed to access that leaf in the tree, and the "yes" or "no" in the circles denote the end label for that path. For instance, if an input sequence had a value of 16-29 in position 69 of its frequency profile, and a value of 30-100 in position 78, the end label would be "yes," denoting that the input protein shared a sequential motif with the protein segment cluster represented by the decision tree.

In order to generate the attribute range for each decision tree, a process noted as adaptive branching is utilized, which uses statistical analysis on the frequency profile value range distribution of a given cluster to more accurately represent the underlying data that a decision tree is being trained on. In effect, the algorithm attempts to base data ranges on the percentage of occurrences of values in a particular range relative to the total number of occurrences of all values. That is to say that if frequency profile values 1-5 constitute the first 20% of all data values in a cluster, then the first attribute range would be 1-5. The pseudo code for this is shown in Figure 3 (in this experiment, percentage p is set to 20%).

It should be noted that the first five positions (0-4) of the frequency profile of the entire data set constitute 80% of the data found within the frequency profiles used in this study. As such, the range 0-4 is considered to be its own range and is omitted from the calculations in generating further ranges.

1. INPUT: Percentage p, Cluster frequency profile set F
2. OUTPUT: Attribute range set B
3. T ← createArray() //Array for storing count of occurrences of values in frequency profile (hereafter referred to as the FP)
4. FOR all f in F: //For all frequency profiles in set F
   5. FOR all pos in f: //For all positions in frequency profile fi
   6. T[pos.getValue()]++ //Increment the count corresponding to pos’s value
   7. t ← getTotal(T) //Get the total of all the values in T
   8. a ← 0 //Declare a variable for accumulating values
   9. FOR all count in T: //For all counts of occurrences of frequencies in FP in array T
   10. a += count //Add count to the accumulation variable
   11. IF a/t >= p*B.length(): //If the accumulation value is holding p percentage of the total
   12. B ← i //Add the current position in the array T being examined to the set B
13. return B

Fig 3. Adaptive Branching Pseudo Code
3.5 Tertiary Structure Prediction

The process of performing a prediction first requires an input “unknown” protein segment’s frequency profile. This implies that, in the context of testing the FGK-DF model, this protein’s secondary and tertiary structure is assumed to be unknown until after the prediction is made. Once the unknown/test protein is input, its frequency profile is scanned against the representative frequency profile of each decision tree. The representative frequency profile of each decision tree is determined by averaging the frequency profiles of all of the protein segments that compose the cluster the tree is trained upon. The distance formula used is the city block metric (Equation 3). All 799 decision trees’ representative frequency profiles are compared against the unknown protein segment’s frequency profile, where the “best tree” is determined to be the one with the lowest distance. Once this tree is decided upon, the unknown protein’s frequency profile is run through the decision tree. If the protein segment is found to share the sequence motif represented by the decision tree (i.e. if the decision tree results in a “yes” decision for the unknown protein), then the tertiary structure of that protein segment is said to be equal to the representative tertiary structure of the branch the protein segment followed. If the protein segment is not found to share the sequence motif, the next best tree is found, and so on until either the distance is greater than a given threshold, or a match is made. The accuracy of the prediction can be generated by comparing the predicted tertiary structure, and the “ground truth” tertiary structure extracted from the PDB.

4. Results

To test the effectiveness of the FGK-DF model in predicting the tertiary structure of proteins given only their primary sequence, over 560,000 segments from the 2,710 culled proteins from PISCES (with supplementary data from HSSP and DSSP) were used to generate testing and training sets for the model. The training segments were granulated into 10 granules, from which 799 clusters were generated. For each cluster, a decision tree was trained with the stipulation that those nodes with an entropy of at least 0.75 were considered to be considered classified. Based on the results published by Chen et al (2011), the label pivot for each decision tree was set to 7, and each of the 5 attribute ranges for each tree were determined with the adaptive branching algorithm. Once the model was complete, more than 2400 testing protein primary sequences (none of them were included in the training dataset) were run through the FGK-DF model, producing a predicted tertiary structure given that one of the 799 decision trees’ representative frequency profile’s distance from the input protein’s frequency profile was less than a given distance threshold. This predicted tertiary structure is compared directly with the known tertiary structure of the test protein, giving a root-mean-square deviation (RMSD) measure designed by Maiorov et al (1994) that denotes the quality of the prediction. RMSD effectively finds the weighted macro difference between each element, in this case, in the two tertiary structures. The equation for RMSD is shown below:

\[
\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{1,i} - x_{2,i})^2}
\]

Eq. 6: Root-mean-square Deviation

The tertiary structure of each segment (which, again, corresponds to nine continuous positions within the protein in this experiment) is represented as 36 numbers describing the mutual distance between each component of the protein’s tertiary structure. In other words, because the window size is nine, each of the nine component positions P1...9 has Pi-1 mutual distances relative to the
other eight component positions, such that the first component position is described in terms of eight mutual distances between itself and the other eight component positions of the tertiary structure (the second component would be described in terms of seven distances, and so on). Describing the tertiary structure in terms of mutual distances ignores problems associated with using explicit xyz coordinates, such as inconsistent rotation, mirroring, etc. of the tertiary structures for each protein. As such, $n$ in Equation 6 would equal 36, and $x_1/x_2$ would represent two different tertiary structures (both predicted and ground truth) offset by the current position $i$. The difference between each position's distance is squared, causing more substantial deviations to be more heavily penalized.

In an effort to reduce the complexity and increase the readability of the output of this experiment, three categories of quality have been set forth, given the RMSD values for each tertiary structure prediction (which is, in this case, measured in angstroms, Å). This experiment adopts convention from past research, stating that an RMSD value of at most 1.5 Å is to be considered an acceptable prediction, with good predictions having a maximum of 1.0 Å. This work adds an additional tier, stating that those predictions with at most 0.5 Å RMSD value is to be considered an excellent prediction.

### Table 1 FGK-DF Model Prediction Results using Adaptive Branching

<table>
<thead>
<tr>
<th>Secondary Structure Similarity</th>
<th>&gt;70%</th>
<th>&gt;80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance Coverage 0.5 Å</td>
<td>1.0 Å</td>
<td>1.5 Å</td>
</tr>
<tr>
<td>550</td>
<td>0.16</td>
<td>67.85</td>
</tr>
<tr>
<td>600</td>
<td>0.40</td>
<td>65.05</td>
</tr>
<tr>
<td>650</td>
<td>0.80</td>
<td>61.72</td>
</tr>
<tr>
<td>700</td>
<td>1.39</td>
<td>60.07</td>
</tr>
<tr>
<td>750</td>
<td>2.21</td>
<td>59.08</td>
</tr>
<tr>
<td>800</td>
<td>3.18</td>
<td>58.24</td>
</tr>
<tr>
<td>850</td>
<td>4.26</td>
<td>56.85</td>
</tr>
<tr>
<td>900</td>
<td>5.36</td>
<td>55.54</td>
</tr>
<tr>
<td>950</td>
<td>6.44</td>
<td>54.27</td>
</tr>
<tr>
<td>1000</td>
<td>7.47</td>
<td>52.96</td>
</tr>
<tr>
<td>1050</td>
<td>8.39</td>
<td>51.82</td>
</tr>
<tr>
<td>1100</td>
<td>9.23</td>
<td>50.97</td>
</tr>
<tr>
<td>1150</td>
<td>9.95</td>
<td>50.20</td>
</tr>
<tr>
<td>1200</td>
<td>10.58</td>
<td>49.46</td>
</tr>
<tr>
<td>1250</td>
<td>11.08</td>
<td>48.80</td>
</tr>
<tr>
<td>1300</td>
<td>11.46</td>
<td>48.22</td>
</tr>
</tbody>
</table>

For each prediction run, a distance threshold between the test protein’s frequency profile and the representative frequency profile is specified, such that a given decision tree can only predict the tertiary structure of a test protein if the city block metric distance between the two profiles is less than the given threshold. 16 thresholds are defined, ranging from 550 to 1300 in increments of 50. The output is further partitioned according to the secondary structure similarity of the cluster a given decision tree was generated from, such that all of those predictions generated with trees trained on clusters with a secondary structure similarity of greater than 70% and greater than 80%

are shown. For each distance threshold and secondary structure similarity partition, a percentage of predictions that were acceptable, good, and excellent (1.5 Å, 1.0 Å, and 0.5 Å respectively) was recorded, as well as the percentage of the overall count of test protein segments that were predicted with said secondary structure similarity, distance threshold, and RMSD value. A chart detailing this output is shown in Table 1.

4.1 Analysis of the FGK-DF Model Results
As one can see in Table 1, a combination of a highly restrictive distance threshold (e.g. 550, 600, etc.) coupled with a high secondary structure similarity of the predicting decision tree results in greater percentage of predictions less than 0.5 Å and 1.0 Å. For instance, a distance threshold of 550 and a decision tree with a secondary structure similarity of greater than 80% generates predictions in which 91.873% of the predictions have an RMSD value of less than 1.0 Å. One will note that the coverage (the percentage of proteins predicted with the given constraints) is extremely low at less than 1% of all test proteins. In order to increase coverage (and consequently lower prediction accuracy) one must both accept a higher distance threshold, and a lower tree secondary structure similarity. For instance, the coverage of predictions given a distance threshold of 1300 and a tree secondary structure similarity of greater than 70% is 11.457% of all test segments. The prediction accuracy for RMSD values of 1.0 Å and 1.5 Å remain in acceptable ranges, with percentages of 73.929% and 83.551% respectively.

The FGK-DF model is viable for producing local tertiary structure prediction. Although one cannot easily achieve 100% coverage without sacrificing overall accuracy, the model can accurately produce 10-15% of the tertiary structures for local structures. To extend the FGK-DF model to generate global tertiary structure predictions, protein loop modeling developed by Chung and Subbiah (1996) can be used to generate the unpredicted segment structures much in the same fashion that techniques such as homology-modeling resolve prediction gaps. While this seems that the FGK-DF model is simply providing the same basic solution produced by homology-modeling, the significant strength of the FGK-DF model is that it is not limited to making predictions on only those proteins with existing and identified homologues, on which homology-modeling is based. Furthermore, as the predicted tertiary structure of each segment from the FGK-DF model is the representative (i.e. average) structure of an underlying decision tree, the produced tertiary structure, and thus folds, of the FGK-DF are semi-novel, providing a significant improvement over not only homology-modeling, but other template-based approaches, such as threading.

4.2 Analysis of the Adaptive Branching Methods
The results of Table 1 were generated using the adaptive branching method described in Figure 2. To test the effectiveness of this generalized approach to providing decision tree attribute range sets, the same parametric setup described in section 4.1 was repeated with a static attribute range set (0-4, 5-7, 8-14, 15-29, 30-100) developed by expert opinion on the data set being utilized in this work and applied uniformly to all decision trees produced by the model. The results of this setup are described in Table 2. Table 3 directly compares the static and adaptive branching by examining the accuracy percentages based on fixed coverage (~9% and ~2%).

Table 3 shows the effectiveness of the adaptive branching mechanism in approximating expert opinion on a data set, as the accuracy results are nearly identical to those using static branching. In fact, in some cases, however minute, adaptive branching outperforms the static branching in terms
of accuracy. This suggests that given a vastly different data set (with which the static branches would no longer apply), the adaptive branching method should produce attribute range sets for the underlying decision trees that most accurately reflect and classify the input data. Further testing on the matter, of course, would be needed to verify such.

Table 2 FGK-DF Model Prediction Results using Static Branches

<table>
<thead>
<tr>
<th>Secondary Structure Similarity</th>
<th>Coverage</th>
<th>0.5 Å</th>
<th>1.0 Å</th>
<th>1.5 Å</th>
<th>Coverage</th>
<th>0.5 Å</th>
<th>1.0 Å</th>
<th>1.5 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance 550</td>
<td>0.16</td>
<td>65.93</td>
<td>89.21</td>
<td>94.93</td>
<td>0.06</td>
<td>75.43</td>
<td>90.66</td>
<td>95.85</td>
</tr>
<tr>
<td>600</td>
<td>0.40</td>
<td>64.27</td>
<td>89.43</td>
<td>94.90</td>
<td>0.15</td>
<td>77.47</td>
<td>92.63</td>
<td>96.66</td>
</tr>
<tr>
<td>650</td>
<td>0.80</td>
<td>61.60</td>
<td>87.57</td>
<td>94.45</td>
<td>0.28</td>
<td>74.42</td>
<td>91.74</td>
<td>96.27</td>
</tr>
<tr>
<td>700</td>
<td>1.39</td>
<td>59.91</td>
<td>86.44</td>
<td>94.00</td>
<td>0.47</td>
<td>72.68</td>
<td>91.56</td>
<td>96.28</td>
</tr>
<tr>
<td>750</td>
<td>2.21</td>
<td>58.81</td>
<td>85.78</td>
<td>93.58</td>
<td>0.73</td>
<td>72.39</td>
<td>91.41</td>
<td>96.09</td>
</tr>
<tr>
<td>800</td>
<td>3.18</td>
<td>57.74</td>
<td>84.74</td>
<td>92.76</td>
<td>1.01</td>
<td>72.26</td>
<td>90.87</td>
<td>95.70</td>
</tr>
<tr>
<td>850</td>
<td>4.26</td>
<td>56.32</td>
<td>83.48</td>
<td>91.98</td>
<td>1.31</td>
<td>71.40</td>
<td>90.04</td>
<td>95.16</td>
</tr>
<tr>
<td>900</td>
<td>5.36</td>
<td>55.15</td>
<td>82.46</td>
<td>91.18</td>
<td>1.59</td>
<td>70.19</td>
<td>89.36</td>
<td>94.52</td>
</tr>
<tr>
<td>950</td>
<td>6.44</td>
<td>53.85</td>
<td>81.16</td>
<td>90.17</td>
<td>1.86</td>
<td>68.83</td>
<td>88.28</td>
<td>93.71</td>
</tr>
<tr>
<td>1000</td>
<td>7.47</td>
<td>52.43</td>
<td>79.61</td>
<td>88.99</td>
<td>2.09</td>
<td>67.45</td>
<td>87.10</td>
<td>92.81</td>
</tr>
<tr>
<td>1050</td>
<td>8.39</td>
<td>51.45</td>
<td>78.49</td>
<td>88.05</td>
<td>2.30</td>
<td>66.44</td>
<td>86.25</td>
<td>92.12</td>
</tr>
<tr>
<td>1100</td>
<td>9.23</td>
<td>50.59</td>
<td>77.40</td>
<td>87.08</td>
<td>2.49</td>
<td>65.50</td>
<td>85.27</td>
<td>91.20</td>
</tr>
<tr>
<td>1150</td>
<td>9.95</td>
<td>49.78</td>
<td>76.41</td>
<td>86.08</td>
<td>2.66</td>
<td>64.54</td>
<td>84.25</td>
<td>90.37</td>
</tr>
<tr>
<td>1200</td>
<td>10.58</td>
<td>49.12</td>
<td>75.47</td>
<td>85.13</td>
<td>2.79</td>
<td>63.82</td>
<td>83.48</td>
<td>89.58</td>
</tr>
<tr>
<td>1250</td>
<td>11.08</td>
<td>48.47</td>
<td>74.65</td>
<td>84.31</td>
<td>2.90</td>
<td>63.12</td>
<td>82.78</td>
<td>88.90</td>
</tr>
<tr>
<td>1300</td>
<td>11.46</td>
<td>47.92</td>
<td>74.05</td>
<td>83.70</td>
<td>2.99</td>
<td>62.50</td>
<td>82.18</td>
<td>88.36</td>
</tr>
</tbody>
</table>

Table 3 Comparison of Static and Adaptive Branching Methods

<table>
<thead>
<tr>
<th>Secondary Structure Similarity</th>
<th>Coverage</th>
<th>0.5 Å</th>
<th>1.0 Å</th>
<th>1.5 Å</th>
<th>Coverage</th>
<th>0.5 Å</th>
<th>1.0 Å</th>
<th>1.5 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branching Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive</td>
<td>9.23</td>
<td>50.97</td>
<td>77.31</td>
<td>86.84</td>
<td>2.09</td>
<td>66.58</td>
<td>87.30</td>
<td>93.00</td>
</tr>
<tr>
<td>Static</td>
<td>9.22</td>
<td>50.59</td>
<td>77.40</td>
<td>89.08</td>
<td>2.10</td>
<td>67.45</td>
<td>87.10</td>
<td>92.81</td>
</tr>
</tbody>
</table>

4.3 Super Granule Support Vector Machine Comparison
In order to analyze the effectiveness of the FGK-DF model relative to other predictive models, the results described in section 4.1 are compared against the results of another model known as the Super Granule Support Vector Machine, or Super GSVM, model developed by Chen and Johnson (2009). The two models share the data preprocessing steps described in sections 3.1 to 3.3, diverging only once the 799 clusters described in the prior sections have been generated. Whereas the FGK-DF model trains decision trees on each cluster to generate automated decisions regarding the presence of sequential motifs and, ultimately, the tertiary structure of unknown proteins, the
Super GSVM model uses a Ranking-SVM to improve the secondary structure similarity for each cluster. The improved clusters are then used to predict the tertiary structure in much the same manner as the FGK-DF model's decision trees: if the unknown protein's frequency profile is sufficiently similar to a given cluster's representative frequency profile, then its tertiary structure is said to be similar to the cluster's representative tertiary structure. Chen and Johnson (2009) provide a more comprehensive description of the Super GSVM model.

Table 4 depicts the results of both the Super GSVM and the FGK-DF model (with adaptive branching) in terms of the percentage of tertiary structure predictions that had an RMSD error of less than 1.5 Å and utilized clusters that exhibited greater than 80% secondary structural similarity. As depicted in Tables 1 and 2, the results are partitioned according to the distance measured between the representative frequency profile of a given cluster and the unknown test protein. Other categories of prediction quality results (i.e. categories of predictions of less than 1.0 Å and 0.5 Å RMSD error margins) are not depicted as they were not measured in the study on the Super GSVM model.

**Table 4 Comparison of Protein Tertiary Structure Prediction Percentages of Results Greater than 1.5 Å of Super GSVM and the FGK-DF (Adaptive Branching) Model**

<table>
<thead>
<tr>
<th>Distance</th>
<th>Super GSVM Tertiary Prediction Results</th>
<th>FGK-DF Tertiary Prediction Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>550</td>
<td>71.98</td>
<td>96.47</td>
</tr>
<tr>
<td>600</td>
<td>69.07</td>
<td>96.90</td>
</tr>
<tr>
<td>650</td>
<td>69.08</td>
<td>96.52</td>
</tr>
<tr>
<td>700</td>
<td>69.47</td>
<td>96.46</td>
</tr>
<tr>
<td>750</td>
<td>69.40</td>
<td>96.23</td>
</tr>
<tr>
<td>800</td>
<td>69.85</td>
<td>95.82</td>
</tr>
<tr>
<td>850</td>
<td>70.02</td>
<td>95.31</td>
</tr>
<tr>
<td>900</td>
<td>69.75</td>
<td>94.62</td>
</tr>
<tr>
<td>950</td>
<td>69.53</td>
<td>93.91</td>
</tr>
<tr>
<td>1000</td>
<td>68.88</td>
<td>93.00</td>
</tr>
<tr>
<td>1050</td>
<td>68.26</td>
<td>92.24</td>
</tr>
<tr>
<td>1100</td>
<td>67.63</td>
<td>91.36</td>
</tr>
<tr>
<td>1150</td>
<td>67.09</td>
<td>90.51</td>
</tr>
<tr>
<td>1200</td>
<td>66.54</td>
<td>89.72</td>
</tr>
<tr>
<td>1250</td>
<td>66.11</td>
<td>89.02</td>
</tr>
<tr>
<td>1300</td>
<td>65.73</td>
<td>88.49</td>
</tr>
</tbody>
</table>

Given this, it is clear in Table 4 that the FGK-DF model exhibits consistently higher prediction accuracies than the Super GSVM for the criteria set in the preceding section. This considerable increase in prediction accuracy could be due to a variety of factors, including the greater presence of the decision tree in the FGK-DF’s predictive process relative to the more data pre-processing role of the SVM in the Super GSVM model. The FGK-DF model’s decision tree’s high entropy threshold (set to 0.75) could have also more adequately modeled the underlying patterns within the naturally noisy and sparse protein frequency profile data more so than the mathematically rigid black box
calculations of the SVMs. Regardless, it is clear in terms of RSMD error, that the FGK-DF model performs with a consistently higher prediction accuracy than the Super GSVM model.

Table 5 compares the coverage exhibited by both the Super GSVM model and the FGK-DF model. While the FGK-DF model vastly outperforms the Super GSVM model in terms of accuracy, the Super GSVM model provides a larger coverage. Given the vastly superior accuracy, this reduction in coverage is to be expected and can be easily increased (at the cost of accuracy) by adjusting the various parameters provided by the FGK-DF model.

**Table 5** Comparison of Protein Tertiary Structure Prediction Coverage of Results Greater than 1.5 Å of Super GSVM and the FGK-DF (Adaptive Branching) Model

<table>
<thead>
<tr>
<th>Distance</th>
<th>Super GSVM Coverage Percentage</th>
<th>FGK-DF Coverage Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>550</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>600</td>
<td>0.38</td>
<td>0.15</td>
</tr>
<tr>
<td>650</td>
<td>0.78</td>
<td>0.28</td>
</tr>
<tr>
<td>700</td>
<td>1.37</td>
<td>0.47</td>
</tr>
<tr>
<td>750</td>
<td>2.21</td>
<td>0.73</td>
</tr>
<tr>
<td>800</td>
<td>3.26</td>
<td>1.01</td>
</tr>
<tr>
<td>850</td>
<td>4.45</td>
<td>1.31</td>
</tr>
<tr>
<td>900</td>
<td>5.79</td>
<td>1.60</td>
</tr>
<tr>
<td>950</td>
<td>7.19</td>
<td>1.86</td>
</tr>
<tr>
<td>1000</td>
<td>8.61</td>
<td>2.09</td>
</tr>
<tr>
<td>1050</td>
<td>10.05</td>
<td>2.30</td>
</tr>
<tr>
<td>1100</td>
<td>11.46</td>
<td>2.49</td>
</tr>
<tr>
<td>1150</td>
<td>12.77</td>
<td>2.66</td>
</tr>
<tr>
<td>1200</td>
<td>13.95</td>
<td>2.79</td>
</tr>
<tr>
<td>1250</td>
<td>14.99</td>
<td>2.90</td>
</tr>
<tr>
<td>1300</td>
<td>15.83</td>
<td>2.99</td>
</tr>
</tbody>
</table>

5. Conclusion

This work has explained and examined the Fuzzy Greedy K-means Decision Forest model as an approach to predict tertiary structure of proteins utilizing only their primary sequence. The FGK-DF model makes use of the relations between conserved sequential motifs that transcend protein family boundaries and their associated tertiary and secondary structures within a training set of data. These aspects are used to generate a forest of decision trees that can determine if an input protein contains a given sequential motif. If the decision tree determines the input protein does contain the sequential motif, the tertiary structure of the unknown protein is said to be the same as the representative, or average, tertiary structure of the branch of the decision tree the decision was made with. This work extends the branching mechanism of the previous works on the FGK-DF model to include adaptive branching to determine the attribute ranges of the decision trees.

This work has also shown that the FGK-DF is a viable method for producing local tertiary structure prediction utilizing only an unknown protein’s primary sequence. The model has been tested and verified against a testing set, producing an acceptable accuracy for many of the predictions it can...
generate with confidence (according to the distance threshold). Though the coverage of the FGK-DF model does not encompass a greater portion of the tested proteins, the model could be extended to perform loop modeling. Furthermore, as the tertiary structure predictions produced by the FGK-DF model are semi-novel, they hold a significant advantage over such modeling techniques as de novo modeling and homology modeling, especially in the realm of such fields as personalized medicine. Finally, the model is directly compared against a very similar model known as the Super GSVM, the results of which show that the FGK-DF consistently outperforms the competing model in terms of accuracy.

Granted this, though there is much work to be done to improve the coverage of the FGK-DF model, it can be said that the FGK-DF model is certainly capable of producing highly accurate, local tertiary structure predictions that are not restricted to protein families, to any particular protein size, and that generate novel folds. This research team believes that these merits of the model constitute recognition and further development, such that the shortcomings of other prominent model families can be approached and overcome with the ideas presented in this work.

6. Future Works

One of the more prominent drawbacks of the FGK-DF model is that the sequence profile motif extraction is limited to a static size, the so-called “window size.” This limit enforces the model to assume all motifs that exist within the training and test data are of the window size, which is, in this case, nine. A partial solution to this assumed motif size is the sliding window technique, which generates successive residues for each set of primary sequence data of size ws, where ws is window size. However, while this technique accounts for motifs of size greater than ws, motifs smaller than ws can be overcome by noise (which is only amplified by the sliding window technique). Furthermore, representation of motifs can potentially become distorted given the sliding window can lead to the over-representation of certain motifs while under representing other motifs. The noise and representation distortion can potentially lead to a significantly less effective model.

Given this, the proposed future work of this research is to find an underlying data model that does not enforce a window size, but rather can be clustered and analyzed given any size of protein and/or motif. One prospective data model is the Hidden Markov Model proposed by Baum and Petrie (1966). In the work of Baldi et al (1994), HMMs have been successfully used to model proteins in the past. Effectively, the HMM describes each protein with three primary states: the main state, the delete state, and the insert state, where the first and last emit amino acid sequences (the observable states) and the middle is “mute” (no output). The insert state can be looped indefinitely for multiple insertions. By this, the HMM can abstractly model any protein and, more importantly, be arbitrarily long, such that any motifs extracted from the HMM can likewise be of any length. This arbitrary length eliminates the static window size constraint.

Thus, the next step this research wishes to take is to use the primary and secondary structure information available through the HSSP and DSSP databases to model each protein therein as an HMM. Using hierarchical clustering based on weighted distances between two aligned HMMs, the HMMs generated for each protein can be clustered according to motifs within the primary sequence they model. As the HMMs can be of any size, and the aligned area by which the HMMs are clustered can be of any size, there are no assumptions on the motif length, allowing for direct representation
and detection of any patterns that can exist within the training and testing data. This model can be extended such that the HMMs also model tertiary data, allowing for protein structure prediction via HMM alignment. Given the expected accurate representations of the underlying protein sequence motifs, the predicted tertiary structures using the HMM modeling approach should be more accurate as well.

References

http://dx.doi.org/10.1126/science.1065659

http://dx.doi.org/10.1073/pnas.91.3.1059

http://dx.doi.org/10.1146/annurev.biophys.30.1.173

http://dx.doi.org/10.1214/aoms/1177699147

http://dx.doi.org/10.1107/S0108767307035623

http://dx.doi.org/10.1017/S09692126(96)00119-0


http://dx.doi.org/10.1504/IJCBDD.2009.028822

http://dx.doi.org/10.1016/S0969-2126(96)00119-0


http://dx.doi.org/10.1016/S0959-440X(02)00306-8

http://dx.doi.org/10.1002/bip.360221211


Lathrop, R.H., 1994. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Engineering 7(9), 1059-1068. 
http://dx.doi.org/10.1093/protein/7.9.1059
http://dx.doi.org/10.1023/A:1008384328214

http://dx.doi.org/10.1007/978-3-642-28554-7_2

http://dx.doi.org/10.1006/jmbi.1994.1017


http://dx.doi.org/10.1186/1471-2105-7-14

http://dx.doi.org/10.1007/BF00116251

http://dx.doi.org/10.1002/prot.340090107


http://dx.doi.org/10.1093/bioinformatics/btg224

http://dx.doi.org/10.1016/S0019-9958(65)90241-X

http://dx.doi.org/10.1016/j.sbi.2008.02.004


http://dx.doi.org/10.1109/TNB.2005.853667