TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 08/10 pp635-647
Volume 19, Number 6, December 2014

Hierarchically Clustered HMM for Protein Sequence Motif Extraction with Variable Length

Cody Hudson, Bernard Chen*, and Dongsheng Che

Abstract: Protein sequence motifs extraction is an important field of bioinformatics since its relevance to the structural analysis. Two major problems are related to this field: (1) searching the motifs within the same protein family; and (2) assuming a window size for the motifs search. This work proposes the Hierarchically Clustered Hidden Markov Model (HC-HMM) approach, which represents the behavior and structure of proteins in terms of a Hidden Markov Model chain and hierarchically clusters each chain by minimizing distance between two given chains' structure and behavior. It is well known that HMM can be utilized for clustering, however, methods for clustering on Hidden Markov Models themselves are rarely studied. In this paper, we developed a hierarchical clustering based algorithm for HMMs to discover protein sequence motifs that transcend family boundaries with no assumption on the length of the motif. This paper carefully examines the effectiveness of this approach for motif extraction on 2593 proteins that share no more than 25% sequence identity. Many interesting motifs are generated. Three example motifs generated by the HC-HMM approach are analyzed and visualized with their tertiary structure. We believe the proposed method provides a unique protein sequence motif extraction strategy. The related data mining fields using Hidden Markova Model may also benefit from this clustering on HMM themselves approach.

Key words: Hidden Markov Model; hierarchical clustering; sequential motif; bioinformatics

1 Introduction

Structural genomics is a field of study that strives to derive and analyze the structural characteristics of proteins through means of experimentation and prediction using software and other automatic processes[1]. Alongside implications for more effective drug design[2], the main motivation for structural genomics concerns the elucidation of each protein's function, given that the structure of a protein almost completely governs its function[3]. Currently, structural genomics is supported through a synergetic gambit of processes and applications on both the experimentation and prediction side, including (respectively) "wet lab" procedures such as X-ray crystallography[4] and Nuclear Magnetic Resonance (NMR) spectroscopy[5], and bioinformatics algorithms which include homology-modeling, threading, and de novo modeling[6]. Wet lab procedures drive the process of structural genomics such that "target" proteins are selected and their structures explicitly determined through accurate albeit extremely expensive and time consuming processes. The target proteins are selected in such a manner that allows the predictive algorithms to determine the structure of proteins that are either sequentially or structurally homologous to the target proteins, allowing for accurate structural analysis of most proteins by only explicitly determining the structure of a few.

Granted this, there are significant drawbacks to

• Cody Hudson and Bernard Chen are with Department of Computer Science, University of Central Arkansas, Conway, AR 72034, USA. E-mail: bchen@uca.edu.
• Dongsheng Che is with Department of Computer Science, East Stroudsburg University, East Stroudsburg, PA 18301, USA.
∗ To whom correspondence should be addressed.
Manuscript received: 2014-06-23; accepted: 2014-06-30

this current approach of wet lab driven structural genomics, the most prominent of which being that current predictive algorithms are heavily dependent on the continual explicit determination of protein structures through the resource intensive wet lab procedures. This work would propose and discuss a new predictive algorithm that analyzes protein structure not through strict homologues, but rather seeks to discover sequential patterns, or motifs, that transcend families of homologous proteins. Unknown proteins were analyzed by this approach. This approach allows for the prediction of new protein structures by strictly analyzing the current record of known protein structures for shared motifs that are not aligned alongside protein families, determining the structure generated from each extracted motif, and aligning the motif (and its structure) with the sequence of the new protein.

A large number of services and databases to extract and store motif information have been developed over the years to address the importance of protein sequence motif discovery and analysis. PROSITE[7] provides a service for querying and annotation of conserved regions in sequences using a vast database of "signatures" and annotations describing the functional characteristics of each motif. PRINTS[8] provides a database of "fingerprints", or clusters of conserved motifs, that are extracted from distinct protein structural and functional families. Both of these prominent databases hold serious implications for drawing connections between functional and structural characteristics of motifs, as well as the relationship that exists between those proteins that share "signatures" and "fingerprints". MEME[9] is another popular service for extracting motifs. Other online database/querying services concerning the discovery of motifs include Minimotif Miner[10], which extracts exceptionally short motifs with known functions from protein sequences; Structural Motifs of Superfamilies (SMoS)[11], which provides a database of structural motifs conserved amongst protein domain superfamilies; and Discovery of Linear Motifs (DiLiMot)[12], which discovers and extracts conserved linear motifs[13] from highly irregular sections of protein sequences. However, many of these models also have the further limitation of imposing a motif width limitation and searching the motifs within the same family.

In order to free the constraint of searching a protein family for conserved motifs amongst homologous proteins, Han and Baker's use of the K-means clustering algorithm to detect protein motifs conserved across protein family boundaries[14]. Later additions of granular computing using Fuzzy C-means to reduce data complexity and further refinement of the K-means algorithm to determine initial centroids in a greedy, iterative refinement model resulted in the Fuzzy Greedy K-means (FGK)[15] model for the detection and extraction of protein motifs. However, despite the fact that these models could discover and extract motifs that transcended protein families, each one shared the same limitation as many other motif extraction approaches: an assumed motif size. This can cause motifs that are much larger than the assumed size to be needlessly segmented, and protein motifs that are smaller than the assumed size to be hidden by non-conserved local amino acids.

Granted the drawbacks of popular approaches to motif extraction and the drawbacks of those approaches which attempt to extract motifs that transcend protein family boundaries, this work proposes the Hierarchically Clustered Hidden Markov Model (HC-HMM) approach for discovering and extracting protein motifs. Instead of using HMM to generate clusters, the proposed method clusters models of HMM. Each protein sequence, defined in terms of a frequency profile, is modeled as a Hidden Markov Model and hierarchically clustered according to the minimum distance achievable between given HMMs. Once all HMMs are clustered, those regions with greater than a given threshold of clustered HMM nodes are to be considered protein motifs. No assumption is made on the size of the protein motif, as each sequence is treated as a separate HMM, and the approach can detect protein motifs that transcend protein family boundaries as the model does not rely on protein homologies.

2 Data Representation

The proposed HC-HMM for extracting sequential motifs has four explicit steps: (1) extract and compile the necessary dataset for constructing HMMs using data from the Homology-derived Secondary Structure of Proteins (HSSP)[16] and the Definition of Secondary Structure of Proteins (DSSP)[17], (2) build an unordered set of HMMs using the extracted dataset, (3) hierarchically cluster the models (beginning with the smallest model) using weighted distance calculations against each HMM nodes' attributes, and (4) extract significantly clustered areas as protein sequential

motifs. This section is focused on the first two steps. Step (3) is described in Section 3, and step (4) is included in Section 4.

2.1 Dataset extraction

The primary data source utilized in this work is HSSP[16] database, providing the frequency profile, insertion probability, and deletion probability of the primary sequence of each protein. The HSSP files are derived from the Protein Data Bank (PDB)[18]. Each HSSP file extracted from the database details a single alignment of any number of proteins clustered to a given protein with a known tertiary structure. As such, each HSSP file contains information pertaining to the alignment itself, a set of twenty percentages detailing the occurrence of each of the twenty amino acids that appear in each position in the alignment, the insertion probability at a given position in the alignment, and the deletion probability at a given position in the alignment, as well as other auxiliary information. In this work, only the frequency profile, insertion probability, and deletion probability for each position in the alignment are utilized from the HSSP files to generate the HMM files. For evaluation purposes, the data from the HSSP is merged with data from DSSP[17] which provides secondary structure information for each entry in the PDB (and subsequently the HSSP files) through analysis of said protein's tertiary structure.

In order to satisfy the requirement that the HC-HMM model identifies and extracts motifs that transcend protein family boundaries, the PISCES[19] culling server is utilized to generate a suitable set of HSSP files for processing. PISCES employs PSI-BLAST with position-specific substitution matrices to cull proteins from the PDB (as well as user supplied lists) based on criteria set forth by the user. In this work, the constraints supplied to PICSES ensured that the culled proteins would share no more than 25% sequence identity, resulting in 2593 proteins represented as HSSP files.

2.2 Hidden Markov Model structure and generation

A Markov Model, and by extension a Hidden Markov Model is based on a system of states and probabilities that exist between those states. In Fig. 1, one will note that there is a division between those states that are "observable" ("soggy", "damp", "dryish", and "dry") and the "hidden" states of the model ("sunny", "cloudy", and "rainy"), which are used to build

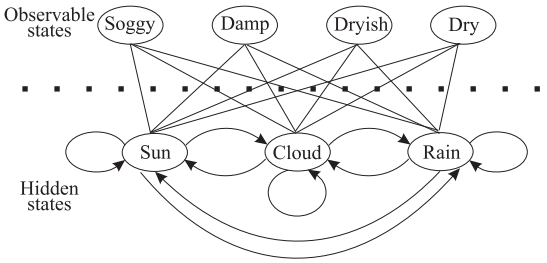


Fig. 1 Hidden Markov Model example.

the model and determine, or output, the observable states. The states "sun", "cloud", and "rain" are "hidden" because the sequence that these states are fired in in order to produce the observable states is unknown; only the output, the "observable" states, can be seen. The HMM can contain multiple hidden levels, where there are probabilities to go from one level to the next, as well as probabilities to output an observable state, making it very flexible and much more representative of how processes in the world actually work[20].

Granted this, the question now becomes how does one use a Hidden Markov Model to more adequately and accurately represent a protein primary sequence? The answer lies in a work by Baldi et al.[21], in which the HMM is structured with five primary states: the start state, the terminal state, the emission state, the insert state, and the delete state, following the evolutionary behavior, such that the traversal each node of the HMM produces an amino acid (or is mute) to build up and represent the overall primary sequence of a protein. A graphical representation of this structure is shown in Fig. 2.

In Fig. 2, state S refers to the aforementioned starting state of the HMM. It produces no output and its transitional probabilities are defined by first node in the protein sequence, where "node" refers to the collection of transitional probabilities $\{p(D_i),\ p(I_i),\ p(E_i)\}$ and states $\{D_i,\ I_i,\ E_i\}$ which describe the behavior and characteristics of the i-th position in a protein sequence. For each node, the state D_i refers to the delete state, which outputs no amino acids. $p(D_i)$ refers to the transitional probability that a given state in node_{i-1} will transition to D_i. State I_i refers to the insertion state and outputs an amino acid based on the frequency profile of node_i, where frequency profile refers to a probability distribution of each possible amino acid appearing at a given position within a given protein sequence. $p(I_i)$ refers to the transitional probability

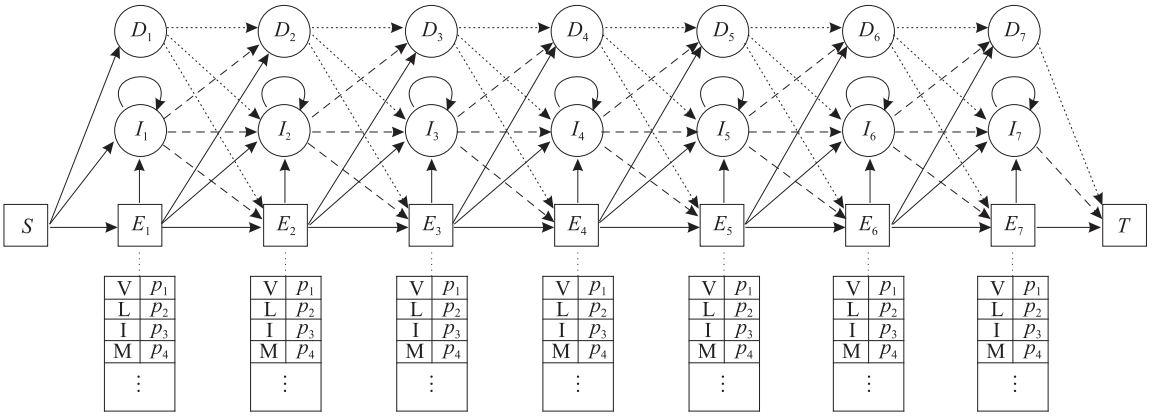


Fig. 2 Example for representation of protein primary sequence with 7 amino acids as a Hidden Markov Model.

that a given state in node$_{i-1}$ will transition to I_i as well as the probability that I_i will transition to itself again (which can be repeated to an arbitrary degree based on said probability). State E_i refers to the emission state, which outputs a single amino acid based on the frequency profile of node$_i$. 20 positions for each state E_i represent 20 different amino acids. p_1-p_{20} refer to the occurring probability for each amino acid that a given state in node will transition to E_i. $p_1+p_2+\cdots+p_{20}$ equals to 1. Finally, state T refers to the terminal state, which marks the end of the Markov chain and produces no output.

Using this structure, any number of protein primary sequences can be easily represented, both structurally and behaviorally, by simply defining the probabilities of each of the three primary states (emission, insertion, and deletion) for each amino acid position in the protein sequence. However, while representing a protein primary sequence using its behaviorally probabilities does more accurately describe the sequential structure and makes no assumptions on motif size (as the protein sequence is in no way segmented), simply representing a protein primary sequence as an HMM does not resolve the problem of being able to *extract* primary sequence motifs without an assumed protein motif size. The solution this work explores to resolve this issue of extracting motifs without an assumed size is to perform hierarchical clustering on the produced HMMs by aligning and clustering two or more HMMs along nodes of highest similarity based on distance calculations and extracting areas with at least m aligned HMMs as sequential motifs. This process is noted as the HC-HMM algorithm, as the next section will explore in greater depth.

3 Hierarchical HMM Model Clustering

As mentioned in the previous section, in order to make no assumptions on the size of the protein sequential motifs that are to be extracted, the HC-HMM uses hierarchical clustering, which builds a *hierarchy* of clusters rather than treating all clusters as distinct, equal entities, such as in K-Means clustering. A simple example of hierarchical clustering is shown in Fig. 3.

The process of hierarchical clustering begins like any other clustering process, with distinct, unclustered data elements. In Fig. 3 above, these data elements constitute a set containing C1, C2, C3, C4, and C5. The clustering process begins in Step 1, such that, using a distance equation or other comparable similarity metric, hierarchical clustering determines the first two data points that are most similar to each other. In the example above, the first two data elements that are most similar to each other are C1 and C2, which are clustered

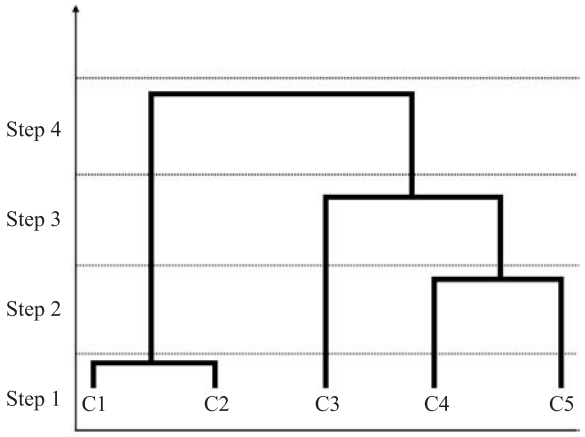


Fig. 3 Hierarchical clustering example.

together as the first level of the cluster hierarchy. The clustering process continues by determining the next two most similar data points, which in this example include C4 and C5. Just as with C1 and C2, these are clustered and added to the hierarchy. This same process is carried out in Step 3, with C3 being determined to be most similar *to the cluster* generated by C4 and C5, creating a new cluster containing a lower level cluster and a data point. This process of determining the similarity between a single data point and cluster can be carried out a great number of ways, one of the more common including averaging all of a cluster's data points into one representative data point and comparing it against the single data point. Finally, the clustering is completed in Step 4 when only one, last cluster is possible to be generated, the one encompassing clusters {C1, C2} and {C3, C4, C5}, which is added at the third and final levels of the hierarchy. The process of hierarchical clustering can be terminated prematurely based on a given threshold or by reaching a certain level in the hierarchy. For instance, the example in Fig. 3 could have been terminated after a certain step (such as Step 3) or once the similarity measures being generated were beyond a given threshold.

Granted the process of hierarchical clustering, HC-HMM attempts to build a hierarchy by comparing each node of an HMM chain against another node in another HMM chain based on weighted distance calculations utilizing each node's emission state, insert state, and delete state probabilities. Those HMM chains containing the nodes that are considered the most similar are clustered as a level in the hierarchy. The clustering process begins with the shortest HMM chain and terminates when all HMM chains have been clustered into one root cluster. The pseudocode for this approach is shown in Fig. 4.

In Fig. 4, α refers to the list of HMM chains generated using the same source of protein primary sequence information described in previous chapters, the HSSP. β refers to the list of processed HMM chains, containing those models that have failed to achieve the minimum distance threshold. Ultimately, all chains will be placed in list β due to the traversal of the chain size hierarchy. The function Find_And_Remove_Shortest_Model() removes the HMM chain with the fewest number of nodes from the list α and stores the removed value in α_i. The local variables minDistance, curDistance, and offset respectively refer to the minimum distance between two HMM chains that has been achieved

```
α = List of generated HMM models
β = List of processed HMM models
WHILE length(α) > 0:
   α_i = Find_And_Remove_Shortest_Model(α)
   minDistance, curDistance, offset = 0
   leastModel = NULL
   FOR each α_j in α:
      FOR each node_k in α_j:
         FOR each node_l in α_i:
            curDistance += Dis(node_{k+l}, node_l)
         curDistance /= length(node_l)
         IF curDistance <= minDistance:
            leastModel = α_j
            minDistance = curDistance
            offset = k
   IF minDistance <= THRESHOLD:
      Add_Model_To_Cluster(leastModel, α_i, offset)
   ELSE:
      β ← α_i
```

Fig. 4 HC-HMM pseudocode.

thus far, the current distance of the current chains being examined, and the number of empty nodes to be inserted at the beginning of chain α_i to achieve the proper clustering with the currently examined chain. The local variable "leastModel" holds a pointer to the HMM chain that currently has the shortest cluster distance with chain α_i. The function Dis(node$_k$, node$_l$) determines the distance between two input nodes using one of the following three equations:

$$\text{Naive}(k,l) = |p(D_k) - p(D_l)| + |p(I_k) - p(I_l)| + \text{FPD}(k,l) \quad (1)$$

$$\text{Mult}(k,l) = (|p(D_k) - p(D_l)|) \times (|p(I_k) - p(I_l)|) \times \text{FPD}(k,l) \quad (2)$$

$$\text{Add}(k,l) = |p(D_k) - p(D_l)| \times \text{FPD}(k,l) + |p(I_k) - p(I_l)| \times \text{FPD}(k,l) + \text{FPD}(k,l) \quad (3)$$

where k and l refer to two nodes from two different HMM chains, $p(D_k)$ refers to the deletion state transitional probability of node k, $p(I_k)$ refers to the insertion state transitional probability of node k, and FPD returns the frequency profile distance between two nodes, defined by the following equations:

$$\text{FPD}(k,l) = \sum_{i=1}^{20} |\text{Freq}_k(i) - \text{Freq}_l(i)| \quad (4)$$

where $\text{Freq}_k(i)$ refers to the probability that amino acid i will be emitted by node k. Equations (1)-(3) are referred to, respectively, as the Naïve, Multiplicative,

and Additive distance equations. The Naïve distance equation lightly penalizes the cluster distance by adding the absolute difference between the insertion and deletion transitional probabilities of node k and node l to the frequency profile difference. The Multiplicative distance equation heavily penalizes the cluster distance by multiplying the absolute difference of each node's deletion and insertion transitional probability plus one (such that if the transitional probabilities are equal, the distance is not penalized at all) with the frequency profile distance. Finally, the Additive distance equation penalizes the cluster distance by separately multiplying insertion transitional probability absolute difference and deletion transitional probability absolute difference with the frequency profile distance.

Once the distance is found for a particular clustering attempt, it is compared against the minDistance. If it is less than the minDistance, the leastModel, minDistance, and offset are all updated appropriately. This is repeated for all possible clusters for a given chain, for all chains. Once the chain with the minimum clustering distance is found, its distance with chain α_i is compared against a set value stored in THRESHOLD. If the distance is less than the threshold, the function Add_Model_To_Cluster() is called, which averages the transitional probabilities and frequency profiles of each node clustered in the chains leastModel and α_i. Each averaged value is weighted by the number of proteins represented by that node, which is extracted from the HSSP data.

This process of removing the smallest HMM chain and attempting to cluster it with the remaining chains is performed until the list α is empty. At this point, the list β will contain all remaining models, including those that have clustered with other chains as well as chains that failed to cluster with any chains. The latter of these are ignored in the final step of the HC-HMM approach, which constitutes the sequential motif extraction.

The final and most pivotal step in the HC-HMM method, motif extraction, is conceptually simple: extract all local sequences with at least m HMM chains clustered at a given position and declare each one to be a sequential motif. This takes advantage of the fact that the HC-HMM compares and clusters HMM chains along their most similar nodes, generating what is effectively an alignment. In a given hierarchy generated by the HC-HMM, there can be a large number of prominent alignments composed of two or more HMM chains overlapping over several nodes. These overlapping alignments composed of at least m HMM chains, again, are to be considered sequential motifs. The process of extracting these motifs can be autonomously performed by iterating over all produce HMM clusters and flagging any contiguous sequences within an HMM cluster that meet the above criteria. To verify that a flagged sequence is a potential motif, visual inspection through an HMM cluster visualizer utility can be performed. An example of the output of the visualizer using a sample HMM cluster is shown in Fig. 5.

In the above output the average frequency profile (Freq Val), the number of clustered HMMs (Count), and the Secondary Structure Similarity (SSS) per node are shown for HMMs that have been successfully

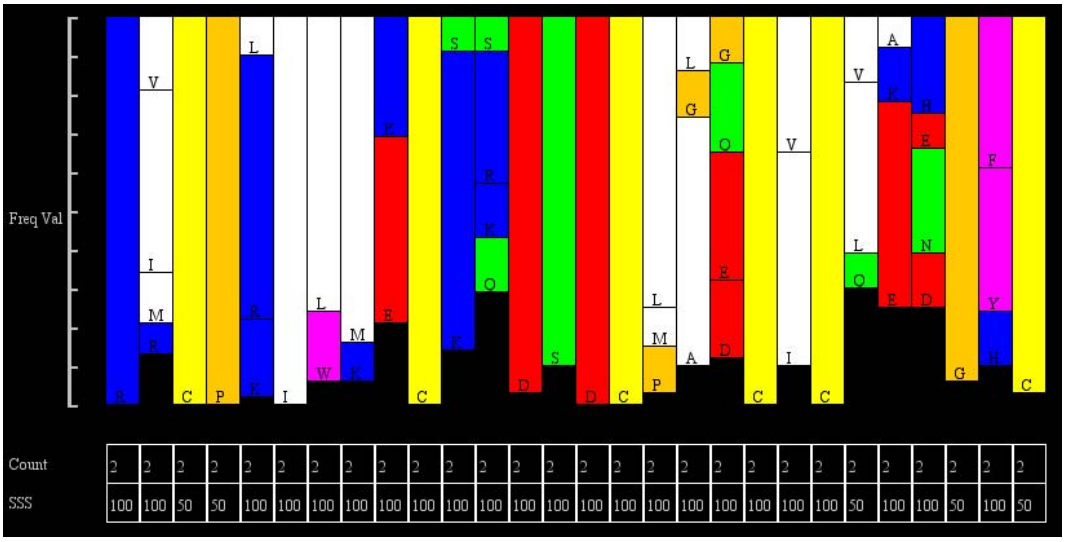


Fig. 5 HC-HMM sequential motif visualizer.

clustered. The average frequency profile per node is shown in terms of single, multi-colored bar denoting values between 0 and 100%. Each color corresponds to a set of amino acids: Amino acids V, L, I, M, and A are white, F, W, and Y are magenta, G and P are orange, S, T, Q, and N are green, C is yellow, H, R, and K are blue, and E and D are red. Note that as certain amino acids share colors in the visualizer, some contiguous blocks of color (such as the R and K or V, I, and M blocks in Fig. 5) are separated by black lines to denote individual amino acid frequencies. For the sake of clarity, amino acids with frequencies of less than 8% are not shown.

In addition to the amino acid frequencies, the count for each node is provided, shown as a number in the first row below the frequency profile data in Fig. 3, which denotes the number of models that were clustered on each node. The secondary structural similarity, shown as a number between 0.0 and 1.0 (with 1.0 denoting complete structural homology) on the bottom row below the frequency profile data in Fig. 2, refers to the overall homology of the secondary structure of each node in a given cluster, computed using the following equation:

$$\frac{\sum_{i=1}^{\text{Count}} \max(p_{i,\text{H}}, p_{i,\text{E}}, p_{i,\text{C}})}{\text{Count}} \tag{5}$$

In the above equation, $p_{i,\text{H}}$ describes the frequency of helices in the protein segments in the cluster at position i for each of the clustered models (of size "count"). $p_{i,\text{E}}$, and $p_{i,\text{C}}$ describe the frequency of sheets and coils, respectively, in the same manner. max() returns the maximum frequency of the three measures.

Thus, all together, the HC-HMM method first takes in protein primary sequence information and generates, for each protein sequence, a Hidden Markov Model. Each of these generated HMM chains are then removed, starting with the smallest chain, and clustered with other HMM chains or HMM chain clusters based on largest nodal similarity utilizing one of the three weighted distance equations listed above. The clustering process terminates once all HMM chains are clustered, at which point sequential motifs can be extracted based on discovering and flagging contiguous sequences of at least m clustered chains. Therefore, given the process involved in the HC-HMM method, the following section will explore the effectiveness of the method in extracting sequential motifs from a set of protein primary sequences, and examine notable motifs extracted by the process.

4 Results

4.1 HC-HMM motif extraction results (data trends)

2593 HSSP files representing proteins exhibiting less than 25% sequence identity were processed by the HC-HMM method utilizing each of the three distance formulas defined in the Methodologies section (Naïve, Multiplicative, and Additive) over a range of distance thresholds normalized between 0 and 1 and a step size of 0.01. Each HSSP file, which contains not only the frequency profile information but also the insertion and deletion probabilities for each amino acid position in the protein primary sequence, was converted into a distinct HMM chain using the structure described in the previous sections. The data was supplemented by the DSSP for secondary structure information strictly for the evaluator purposes as outlined in the previous section. For each produced HMM cluster, motifs were extracted based on a minimum node cluster count, m, such that any contiguous sequence of HMM nodes with at least a node cluster count of m would be considered a motif. The count, average length, and average secondary similarity of the extracted motifs for each application of the HC-HMM method were recorded. This process was executed for values of m ranging between 3 and 5, the results of which are shown in Figs. 6-11.

In Figs. 6-8, the average motif count and secondary structural similarity of each HMM cluster produced by a given threshold (ranging from 0.0 to 0.50, omitting distance thresholds that do not produce HMM clusters) are shown for each of the three distance functions and increasing values of m. Note that in Figs. 6-8, average

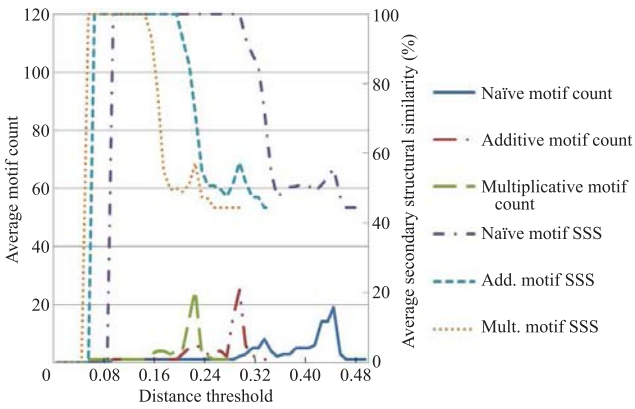


Fig. 6 Motif count and secondary structural similarity when m = 3.

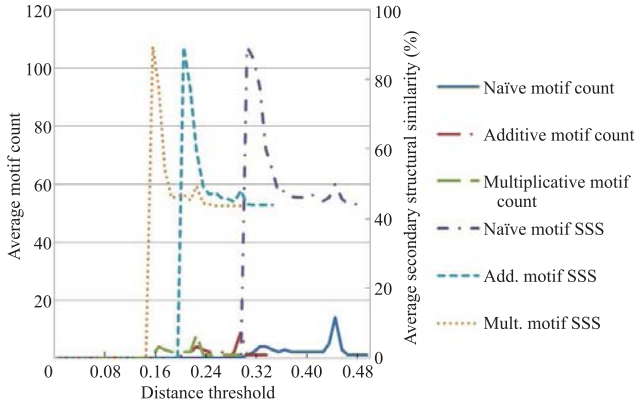


Fig. 7 Motif count and secondary structural similarity when m = 4.

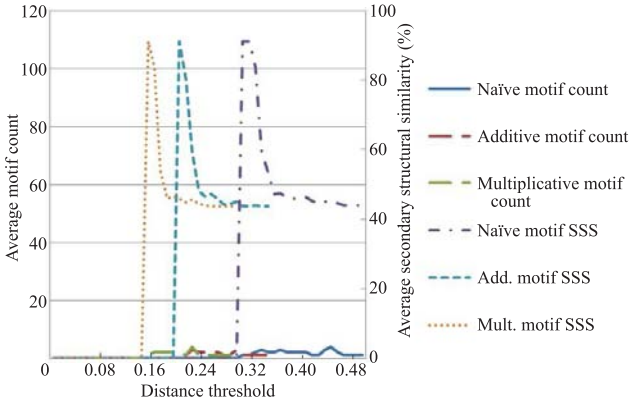


Fig. 8 Motif count and secondary structural similarity when m = 5.

secondary structure similarity is scaled by the right vertical axis while average motif count is scaled by the left vertical axis. A common trend for all values of m shown above is that as distance threshold increases (and thus becomes less restrictive) the motif count, in general, increases as secondary structure similarity decreases. This trend continues until a tipping point in the distance threshold is met, at which all protein data is clustered into one large cluster. At this point, the motif count and secondary structure similarity both spike, producing a significant local maximum for both count and secondary structure similarity. This trend is most apparent when m = 3, growing gradually more subtle as m increases.

A similar trend can be seen in Figs. 9-11, showing the average length of each motif as distance threshold increases for each of the three distance functions. Motif length increases as the distance threshold increases. This is due to the less restrictive distance thresholds, again, causing the HMMs to cluster into one large

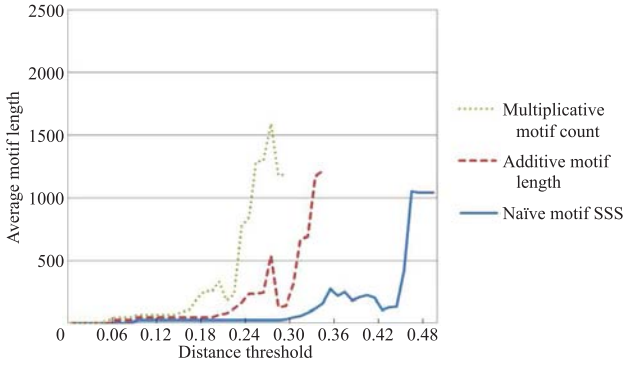


Fig. 9 Motif count length when m = 3.

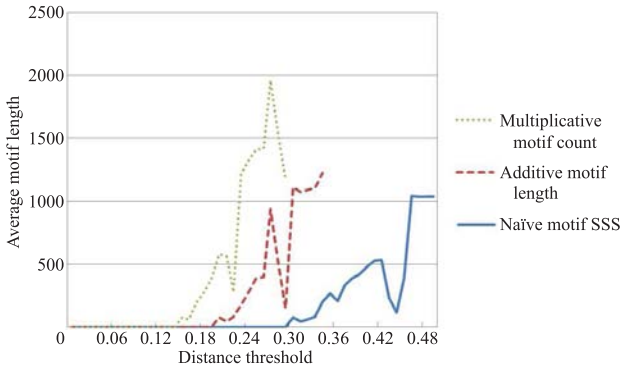


Fig. 10 Motif count length when m = 4.

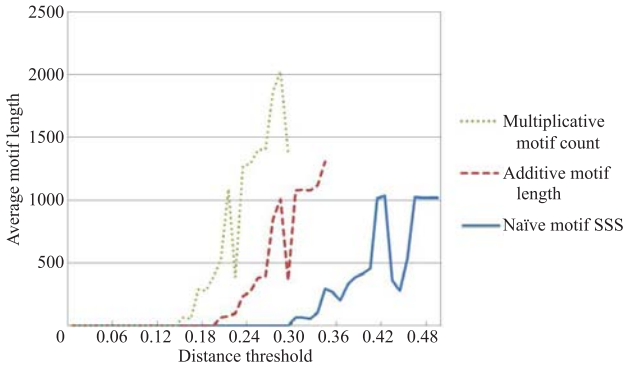


Fig. 11 Motif count length when m = 5.

cluster, increasing the possible length of contiguous sequences. Inverse to what Figs. 6-8 exhibited, the motif length drops to a local minimum as the distance threshold tipping point is reached. It is notable that as m increases, the average length of the motifs also increases. This is most likely due to smaller values of m detecting shorter, sparser motifs, thus lowering the overall average length.

Interestingly enough, all three distance formulas produce roughly identical trends with varying distance threshold scales, suggesting that the primary difference in the three distance formulas is sensitivity, with

Multiplicative being the most sensitive and Naïve being the least sensitive. Given that, the Naïve function will be the only function discussed any further.

4.2 HC-HMM motif extraction results (extracted motifs)

Given the assumption that motifs generated with greater values of m indicate more prominent motifs, and given the assertion that there exists a local maximum for motif secondary structural similarity as distance threshold increases, this work extracts three notable motifs generated with the Naïve distance function with a distance threshold of 0.30 where m =3 and m = 5. These motifs, as depicted by the visualizer, are shown in Figs. 12-14. The proteins used to generate Fig. 12 include 1qsu (chain A), 1q7d (chain A), and 1dzi (chain B). The proteins used to generate Fig. 13 include 1cgd (chain A), 1ei8 (chain A), and 1bkv (chain A). The proteins used to generate Fig. 14 include 1gqe (chain A), 1ic2 (chain A), 1gmj (chain A), 1uuj (chain A), and 1na3 (chain A).

In Fig. 12, there is a clear and relatively consistent three-residue pattern in the extracted motif, with a residue predominately composed of glycine followed by a residue typically composed of proline and ended by a residue composed of alanine, proline, and

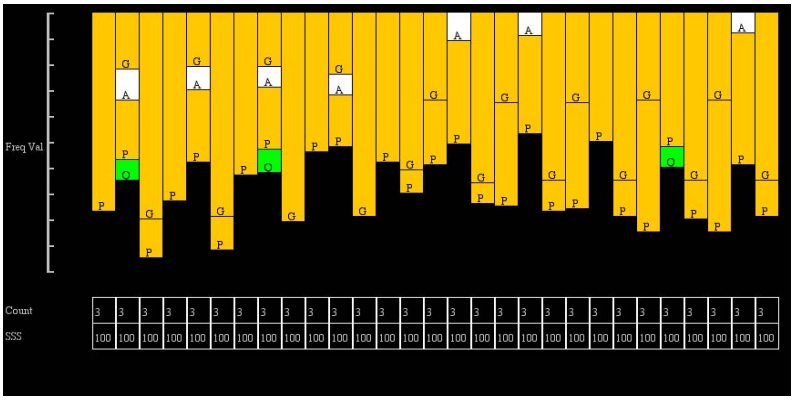


Fig. 12 Motif (29 residues, 100% secondary structural similarity) when m = 3.

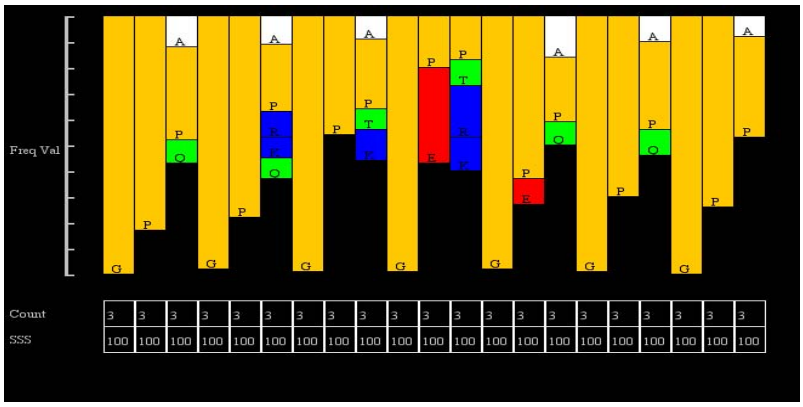


Fig. 13 Motif (21 residues long, 100% secondary structural similarity) when m = 3.

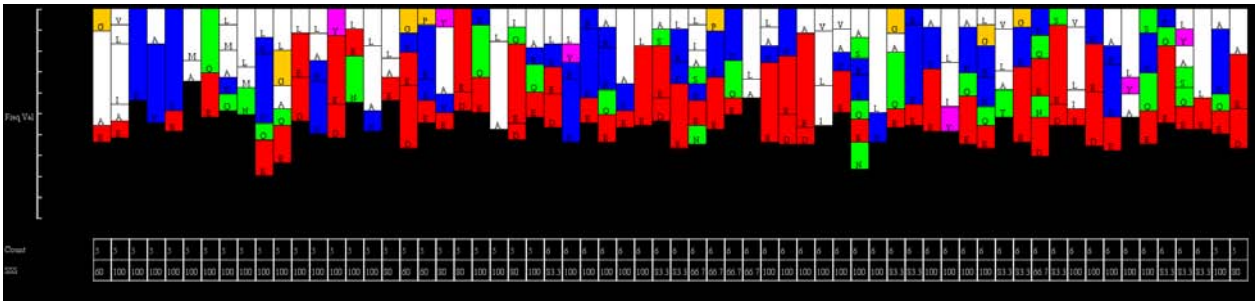


Fig. 14 Motif (64 residues long, 91.09% Sec. Struc. Sim.) when m = 5.

glutamine. This pattern is roughly repeated seven times in the extracted motif. Figure 13 holds a similar repeated pattern structure, with two clear patterns: proline-dominant residue followed by glycine, alanine, and proline-dominant residue followed by a glycine dominant residue, and a pattern defined by proline-dominant residue followed by a glycine-dominant residue followed by roughly equal parts proline and glycine. Each of these two patterns repeat themselves roughly five times within the motif. It is important to note that the average secondary structure similarity of these two motifs is 100%, which suggests that these motifs are significant not only for primary sequence analysis, but structural analysis as well.

Figure 14 denotes a much larger but less regular motif, extracted based on high overall secondary structural similarity (91.09%) as well as its high cluster count, which ranges from 5 to 6. While this motif does not contain any apparent repeating patterns, there are regularities to note. The motif, as a whole, generally exhibits a high frequency of glutamic acid with small but persistent traces of aspartic acid. Though not as consistently present, there is a notable frequency of lysine as well as leucine. Again, given the high cluster count and high secondary structure similarity, this motif has strong implications for both sequential and structural analysis. It is also possible that this motif, given its considerable length, is potentially composed of sub motifs, though further analysis would be required to test this assertion.

To explore the potential of this methodology for structural prediction and analysis, an average tertiary structure for the three motifs shown in Figs. 12-14 was generated and visualized. To generate the tertiary structure information, the base protein models and chain for each extracted motif (described in the prior paragraphs) were used to perform a query on the PDB. The three-dimensional positions for each alpha carbon atom for a given chain of a given protein were recorded, and a mutual distance matrix was calculated between each recorded vertex contained within the generated motif to remove any rotational, mirroring, etc. inconsistencies in the extracted tertiary information. Each mutual distance matrix was then averaged for each protein present in a given motif. The resulting tertiary structures for the motifs denoted by Figs. 12-14 are shown, respectively, by Figs. 15-17.

Thus, taken together, the limitation of the FGK-DF model, as well as many other motif extraction methodologies, was examined, with a focus on an assumed window size. This particular limitation was analyzed and overcome by utilizing the HC-HMM approach by representing protein data as Hidden Markov Models capturing protein behavior and metrics in terms of insertion, deletion, and amino acid

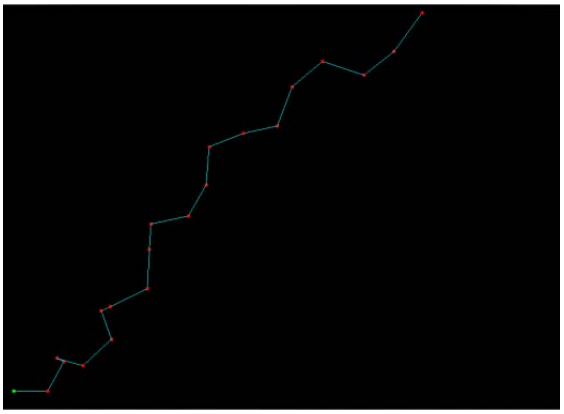


Fig. 15 Visualized tertiary structure of motif containing proteins 1qsu, 1q7d, and 1dzi.

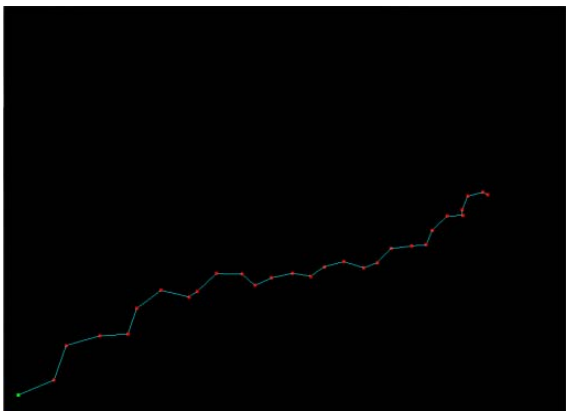


Fig. 16 Visualized tertiary structure of motif containing proteins 1cgd, 1ei8, and 1bkv.

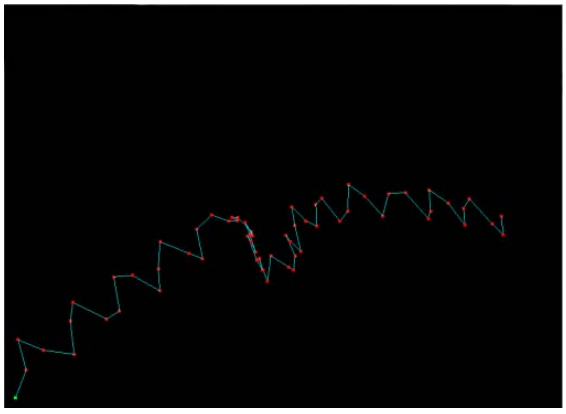


Fig. 17 Visualized tertiary structure of motif containing proteins 1gqe, 1ic2, 1gmj, 1uuj, and 1na3.

probability nodes and hierarchically clustering the resulting HMM chains by minimizing distance between any two given chains. Motifs can then be extracted without any assumption on the length of the motif by analyzing the clusters and extracting contiguous sequences with a given threshold of clustered proteins. The effectiveness of this methodology and various parametric setups were critically examined in terms of the number, quality, and length of the resulting motifs. Furthermore, several example motifs generated by the HC-HMM approach were shown, examined, and visualized in terms of their averaged tertiary structure.

Granted the effectiveness of this approach for eliminating both outlined shortcomings, there is still much that can be improved upon. While the application of the HC-HMM on the outlined data is capable of generating over 100 distinct motifs from the generated clusters, those motifs typically only represent small contiguous segments where $m = 2$. While these motifs still contain valuable information, for the purposes of utilizing the HC-HMM for motif extraction and the FGK-DF for processing said motifs, further improvements are necessary, as the concluding chapter will touch on briefly.

5 Conclusions

Primary sequence motif extraction from protein amino sequences is a field of growing importance in bioinformatics due to its relevance to both sequential and structural analysis. Many approaches for motif extraction include two limitations: a reliance on discovering an existing, known protein homologue to perform motif extraction or structural analysis, and an assumed motif length. In this paper, to tackle both problems, we proposed the HC-HMM approach by representing protein data as Hidden Markov Models capturing protein behavior and metrics in terms of insertion, deletion, and amino acid probability nodes and hierarchically clustering the resulting HMM chains by minimizing distance between any two given chains. Motifs can then be extracted without any assumption on the length of the motif by analyzing the clusters and extracting contiguous sequences with a given threshold of clustered proteins. The input dataset relies on the powers of PISCES to exhibit no more than 25% sequence identity, all motifs that are extracted can be assumed to transcend protein family boundaries. The effectiveness of this methodology and various parametric setups are critically examined in terms of the number, quality, and length of the resulting motifs. Furthermore, three example motifs generated by the HC-HMM approach are shown, examined, and visualized in terms of their averaged tertiary structure.

It is well known that HMM can be utilized for clustering purpose, however, methods for clustering on Hidden Markov Models themselves are rarely studied. We believe that the HC-HMM approach provides a potential solution to various limitations inherent in current motif extraction approaches, and that the techniques employed in this work not only explore novel approaches of analyzing data models through hierarchical clustering, but also hold strong implications for motif extraction processes and protein structural analysis.

6 Discussion and Future Works

As mentioned in the previous sections, one of the most prominent areas requiring improvement is the HC-HMM's ability to extract meaningful motifs both in numerous quantities and higher quality. To reiterate, the HC-HMM generated over 100 distinct motifs from the produced clusters, such that the motifs were usually only short, contiguous segments where $m = 2$. This could be due to a great number of reasons, most prevalent possibly being that a given HMM chain clusters with another HMM chain based on only one node without the possibility of introducing of gaps. This implies that between two HMM chains, there can only exist one motif, which is not a correct assumption given that two or more protein sequences can exhibit more than one motif at a given time.

Therefore, one of the possible improvements to the HC-HMM method is to allow for the introduction of gaps. This can be done in a variety of ways, but the proposed method in this work is to create a mutual distance matrix examining the difference, in terms of the distance equations set forth, of all of the nodes for all of the HMM chains being processed. Those node pairs that exhibit below a given dissimilarity would then be flagged as what would effectively be motif "seeds". These "seeds" could then be grown, from left to right on their respective chains based on a diminishing similarity threshold, such that each subsequently added node onto a given seed would have less stringent similarity requirements.

With this approach, multiple motif "seeds" can

appear in any given HMM chain pair, allowing multiple motifs to be extracted from only one HMM pair (and thus, implicitly, the introduction of gaps). This would allow for more numerous motif extractions and, ideally, the motifs extracted, given the fine grained, node-based similarity measures, would be of much higher quality. Granted such success, this new method could completely replace the FGK portion of the FGK-DF model, in so far as protein sequential motif extraction is concerned. This would allow the newly improved FGK-DF model, trained with extremely accurate and high quality motifs that transcend protein family boundaries, to perform even higher quality tertiary structure predictions with, ideally, higher coverage. With increased coverage, the FGK-DF model could be extended to begin predicting global tertiary structure as well as protein folding (known as quaternary structure). With this in hand, the complete three-dimensional model of the protein can be produced, and thus its function elucidated. This, of course, is too far in the distance to adequately discuss with any true accuracy without first ascertaining the effectiveness of the extended HC-HMM method for protein sequential motif extraction.

Another prominent extension to this methodology is to extend its utility to include tertiary structure prediction. This will involve using the HC-HMM approach to generate a list of known motifs, complimented with tertiary structural data in a manner outlined in the Results section, and declaring the structure of an unknown protein represented as a Hidden Markov Model to be similar to the motif that exhibits the least distance between itself and the unknown protein. This will allow the HC-HMM approach to predict the tertiary structure of an unknown protein using only primary sequence information in terms of sequential motifs that transcend protein family boundaries. Any segments that are not aligned with a known motif can be modeled using a variety of protein fold modeling techniques.

References

[1] J. Chandonia and S. E. Brenner, The impact of structural genomics: Expectations and outcomes, Lawrence Berkeley National Laboratory, California, USA, Dec. 2005.

[2] T. Lengauer and R. Zimmer, Protein structure prediction methods for drug design, *Briefings in Bioinformatics*, vol. 1, no. 3, pp. 275-288, 2000.

[3] G. Karp, *Cell and Molecular Biology: Concepts and Experiments*, 6th ed. New York, USA: John Wiley & Sons Inc, 2009, pp. 52-66.

[4] A. L. Spek, Structure validation in chemical crystallography, *Acta CrystalloGraphica, Section D*, vol. 60, no. 4, pp. 148-155, 2004.

[5] J. K. M. Sanders and B. K. Hunter, *Modern NMR Spectroscopy: A Guide for Chemists*. New York, USA: Oxford University Press, 1998.

[6] A. S. Nair, Computational biology & bioinformatics: A gentle overview, *Communications of the Computer Society of India*, pp. 1-13, 2007.

[7] C. J. A. Sigrist, E. de Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios, New and continuing developments at PROSITE, *Nucleic Acids Research*, vol. 41, pp. 1-4, 2012.

[8] T. K. Attwood, A. Coletta, G. Muirhead, A. Pavlopoulou, P. B. Philippou, I. Popov, C. Romá-Mateo, A. Theodosiou, and A. L. Mitchell, The PRINTS database: A fine-grained protein sequence annotation and analysis resource—its status in 2012, *Journal of Biological Databases and Curation*, vol. 2012, 2012. doi: 10.1093/database/bas019.

[9] T. L. Bailey, MEME SUITE: Tools for motif discovery and searching, *Nucleic Acids Research*, vol. 37, no. 2, pp. 202-208, 2009.

[10] T. Mi, J. C. Merlin, S. Deverasetty, M. R. Gryk, T. J. Bill, A. W. Brooks, L. Y. Lee, V. Rathnayake, C. A. Ross, D. P. Sargeant, C. L. Strong, P. Watts, S. Rajasekaran, and M. R. Schiller, Minimotif Miner 3.0: Database expansion and significantly improved reduction of false-positive predictions from consensus sequences, *Nucleric Acids Research*, vol. 40, pp. 252-260, 2012.

[11] S. Chakrabarti, K. Venkatramanan, and R. Sowdhamini, SMoS: A database of stuctural motifs of protein superfamilies, *Protein Eng*., vol. 16, no. 11, pp. 791-793, 2003.

[12] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell, Systematic discovery of new recognition peptides mediating protein interaction networks, *PLoS Biol*., vol. 3, no. 12, 2005. doi: 10.1371/journal.pbio.0030405.

[13] V. Neduva and R. B. Russell, Linear motifs: Evolutionary interaction switches, *FEBS Lett*., pp. 3342-3345, 2005.

[14] K. F. Han and D. Baker, Recurring local sequence motifs in proteins, *J. Mol. Biol*., vol. 251, pp. 176-187, 1998.

[15] B. Chen, P. C. Tai, R. Harrison, and Y. Pan, FGK model: An efficient granular computing model for protein sequence motifs information discovery, in *Proceedings of IASTED CASB 2006*, Dallas, USA, 2006, pp. 56-61.

[16] C. Sander and R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins Struct. Funct. Genet*., vol. 9, no. 11, pp. 56-68, 1991.

[17] W. Kabsch and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, vol. 22, pp. 2577-2637, 1983.

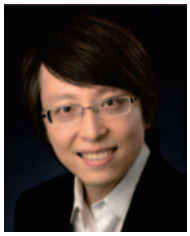
[18] H. M. Berman, The Protein Data Bank: A historical perspective, *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 64, no. 1, pp. 88-95, 2008.

[19] G. Wang and R. Dunbrack Jr., PISCES: A protein sequence culling server, *Bioinformatics*, vol. 19. no. 12, pp. 1589-1591, 2003.

[20] L. R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[21] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure, Hidden Markov models of biological primary sequence information, *Proceedings of Natural Academy of Science, USA*, vol. 91, pp. 1059-1063, 1994.

Cody Hudson graduated as a master student from the Department of Computer Science at University of Central Arkansas. He received his master degree in 2013. His advisor is Dr. Bernard Chen. His research is in bioinformatics. He is currently working for University of Arkansas Medical Science (UAMC) as a research assistant.



Bernard Chen is an associate professor in the Department of Computer Science at University of Central Arkansas. He received his PhD degree in computer science from The Georgia State University in 2008 and has since been on the faculty at UCA. His main research interests are in the broad areas of computational biology/bioinformatics with focus on protein sequence motifs search and protein local tertiary structure prediction. He is interested in data mining, AI, high performance computing, and data science. He has more than 30 publications of his research in bioinformatics journals and conferences including *BMC Bioinformatics*, *Bioinformation*, *BIBE*, and *BIBM*.

Dongsheng Che is an associate professor in the Department of Computer Science at East Stroudsburg University (ESU) of Pennsylvania. He received his PhD degree in computer science from The University of Georgia in 2008 and has since been on the faculty at ESU. His main research interests are in the broad areas of computational biology/bioinformatics with focus on (1) comparative genomics, (2) transcriptional regulatory networks, (3) biological pathways, and (4) structural genomics. He is interested in both development of bioinformatics tools and study of biological problems using *in silico* approaches. He has more than 30 publications of his research in bioinformatics journals and conferences including *Nucleic Acids Research*, *Bioinformatics*, *BMC Genomics*, *Proteins*, and *Pacific Symposium on Biocomputing*.