Identifying Protein Binding Functionality of Protein Family Sequences by Aligned Pattern Clusters

En-Shiun Annie Lee
Systems Design Engineering
University of Waterloo
Waterloo, Canada
annie.lee@uwaterloo.ca

Andrew K. C. Wong
Systems Design Engineering
University of Waterloo
Waterloo, Canada
akcwong@pami.uwaterloo.ca

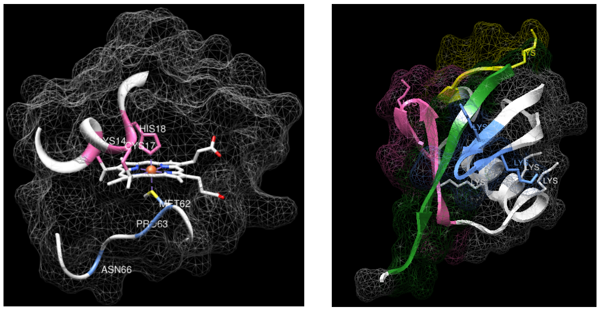
***Abstract*—A basic task in protein analysis is to discover a set of sequence patterns that reflect the function of a protein family. This set of sequence patterns contains non-exact significant residue associations. Currently, the existing combinatorial methods are computationally expensive and probabilistic methods require richer representation of the amino acid associations. To undertake this task, we create a synthesized pattern representation called an Aligned Pattern (AP) Cluster that identifies the residue associations in the binding segment and the site variations in the aligned residues.**

In this paper, our algorithm identifies the binding segments for two protein families: the Cytochrome Complex and the Ubiquitin protein families. For each of the experiments, the AP Clusters obtained correspond to protein binding segments including a few beyond those identified by the other protein databases, PROSITE and pFam. Furthermore, the columns of aligned sites that exist only as a single value in the AP Clusters also corresponds to the binding residues. Additional information retained by the AP Clusters can reveal the amino acid residues of interest, thus averting time-consuming simulations and experimentation.

***Index Terms*—Aligned Pattern Cluster, Protein Function Prediction, Pattern, Hierarchical Clustering**

I. Introduction

Protein functions govern many biological processes of an organism, from enzyme catalysts to ligand binding. These protein functions contribute to the most important human diseases of this decade such as HIV AIDS, cancer, and Alzheimer's. Binding sites are the key functional center of a protein, and therefore, recognizing binding sites can be a key step in protein function analysis. Proteins in the same family perform similar functions, but each protein in the family is encoded by a modified primary sequence. Hence, the significant associations between two consecutive residues within the family of sequences reflect the protein family's functionality. For example, the protein Cytochrome Complex (Cytochrome C.) functions in the electron transport chain of the mitochondria and binds a heme ligand that is involved in electron transfer. In the Cytocrome C. binding site, the protein binds the heme ligand's iron residue from opposite sides by two amino acid residues, called the binding residues. Each of these two binding residues is surrounded by a sequence pattern with variation, called the binding segment. To cite another example, the Ubiquitin protein contains seven lysine binding residues that function by linking individual Ubiquitins. Based on the different types of linkages that are created, the poly-Ubiquitin is recognized by different Ubiquitin binding proteins. In this paper, the Ubiquitin binding residues are the seven lysine residues that link the Ubiquitins, and thus the sequence pattern surrounding that binding residue is the binding segment.



(a) Cytochrome C Binding Residues (b) Ubiquitin Binding Residues

Fig. 1. (a) Cytochrome C. has two binding residues which binds the heme ligand from the proximal and the distal side. (b) Ubiquitin has seven binding residues.

A common approach to identifying the function of a protein family is to use its sequence patterns with variations. Functional patterns are altered through evolutionary mutation; thus, they do not repeat precisely at the exact location of each protein occurrence. Combinatorial and probabilistic methods are the two methods that exist to identify protein function by aligning residues with mutational variations in the family of sequences. The former exhaustively enumerates all possible sequence segments and finds the best consensus alignment. Works reported in [1], [2], [3], [4] create graphs where the vertices are sequence segments and the arcs connect similar sequence segments. The cliques represent the best consensus patterns. These combinatorial methods are computationally expensiveand are NP-Complete [5], [6].

The probabilistic methods calculate the residue distribution at each position to form a consensus sequence. The simplest probabilistic method is the position-specific weighted matrix [7]and another is based on the synthesis of random sequences [8]. Other probabilistic methods make use of the Markov model, such as the pFam database[9]. The probabilistic models

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

compress the data into independent probability distributions and considered the set the number of consecutive positions that are associated. Although each aligned site has its residue distribution, there is no specific way to indicate which residues in the consensus are not statistically or functionally significant. Hence, to efficiently obtain sequence associations of the residues, a knowledge-rich representation of sequence patterns allowing variations is needed.

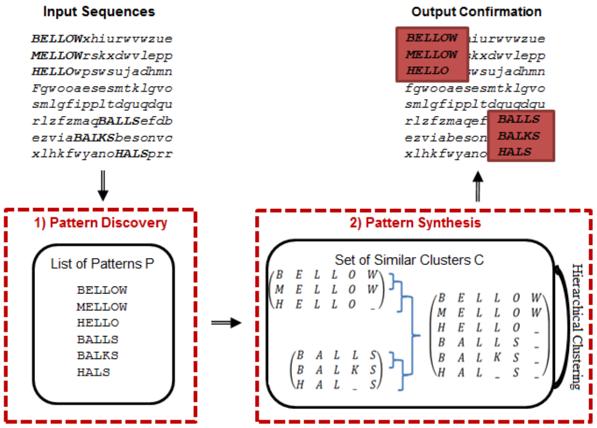


Fig. 2. The overview of the AP Synthesis Process. Our method contains two steps: the Pattern Discovery Step as well as the AP Clustering Step. The results are the AP Cluster.

With this goal in mind, we create what we call Aligned Pattern (AP) Clusters to represent protein functional segments, used for protein binding, by capturing their statistically significant associations of the residues along the sequences as well as the distribution of their occurrence on each aligned site. We align and group similar patterns into AP Clusters and examine whether a cluster of patterns corresponds to the binding segment or to the other protein functional segments. When we applied our AP Synthesis Process to the Cytochrome C. and the Ubiquitin protein family from pFam, we found the AP Clusters for the functional binding segments of the family, and further identified binding residues within the AP Cluster. In our experiment, we intend to find the significant sequence patterns from the processed data of the pFam and discover the functional association patterns in the model. Our AP Synthesis Process discovers the significant sequence pattern while revealing the invariant amino acids which the traditional date mining methods fail to discover. The knowledge-rich representation of AP Clusters can be used to identify binding segments and binding residues.

II. Methodology

Our method (Fig. 2) undergoes two steps: the Pattern Discovery Step and the AP Clustering Step. First, in the Pattern Discovery Step, we discover the most important sequence patterns amongst the family of sequences. These sequence patterns are non-redundant statistically significant patterns. Then, in the AP Clustering Step, we use these discovered patterns as input for aligning and grouping them into an aligned framework. The resulting knowledge-rich representation is called the AP Cluster .

A. Pattern Discovery Pre-Processing

In the Pattern Discovery Step, we use a previously developed pattern discovery and pattern pruning algorithm [10] to obtain a list of significant patterns from a protein family of sequences. Here, we briefly provide the definitions and the background of the algorithm.

Let Σ be an ALPHABET set containing the elements $\{\sigma_1, \sigma_2, \ldots, \sigma_{|\Sigma|-1}, \sigma_{|\Sigma|}\}$. Let $\mathbb{S} = \{S_1, S_2, \ldots, S_{|\mathbb{S}|-1}, S_{|\mathbb{S}|}\}$ be a set of MULTIPLE SEQUENCES. Each sequence is composed of consecutive elements taken from the alphabet Σ as $S_i = s_1^i \ldots s_j^i \ldots s_{|S_i|}^i$, where $s_j^i \in \Sigma$ is the element found in sequence i at position j of that particular sequence.

Definition 1. *An* UNALIGNED PATTERN $\bar{P} = s_1^{\bar{P}} s_2^{\bar{P}} \ldots s_{|\bar{P}|}^{\bar{P}}$ *is a exact substring from* $\mathbb{S}$ *that passes four statistical conditions refined to a score defined in Wong et. al [10].*

B. Pattern Clustering

In the AP Clustering Step, we develop an algorithm which groups a set of similar patterns of different lengths obtained from the Pattern Discovery Step and then align them into a set of APs of the same length by inserting gaps and wildcards. These APs are aligned into a matrix group where corresponding residues amongst the patterns are aligned in the same column, thus implying a common functionality among the APs [11].

An AP Cluster, C, is a group of similar patterns that have been aligned into a set of APs $\mathbb{P} = \{P_1, P_2, ..., P_m\}$ represented by C, which can be expressed as

$$C = \mathrm{LCS}\begin{pmatrix} \bar{P_1} \\ \bar{P_2} \\ \vdots \\ \bar{P_m} \end{pmatrix}. = \begin{pmatrix} s_1^1 & s_2^1 & \cdots & s_n^1 \\ s_1^2 & s_2^2 & \cdots & s_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ s_1^m & s_1^m & \cdots & s_n^m \end{pmatrix}_{n \times m} = \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_m \end{pmatrix}, \quad (1)$$

where $s_j^i \in \Sigma \cup \{\_\}$ is an AP P_i with newly aligned column index j. Each of the m APs in the rows of C is of length n.

Definition 2. *An* ALIGNED PATTERN $P = s_1^P s_2^P \ldots s_{|P|}^P$ *is a subsequence of order-preserving elements maximizing the similarity of the patterns within* $\mathbb{P}$ *with gaps and mismatches so that each* $P \in \mathbb{P}$ *is of length* n.

Definition 3. *An* ALIGNED COLUMN c_j *in* C *represents the* j^{th} *column of characters from the set of APs forming the current AP Cluster. Thus,* $C = \begin{pmatrix} c_1 & c_2 & \ldots & c_n \end{pmatrix}$.

The AP Clustering Step is accomplished by a single-linkage hierarchical clustering algorithm that operates on a list of patterns and synthesized, or more precisely aligns and groups, them into one or more AP Cluster(s). We modified a hierarchical clustering algorithm that synthesized random sequences. The algorithm iteratively MERGEs two AP Clusters in a pairwise-manner based on their SIMILARITY score until the TERMINATION condition is reached. The three key parameters of the hierarchical clustering algorithm are the MERGE

Algorithm 1 The Single-Linkage Hierarchical Clustering Algorithm

```
Require: $\mathbb{P} = \{\bar{P}_1, ..., \bar{P}_n\}$
Ensure: $\mathbb{C} = \{C_1, ..., C_m\}$
  Set all $P_i \in \mathbb{P}$ as $C_i \in \mathbb{C}$
  while (For all pairs of clusters $(C_i, C_j) \in \mathbb{C}$) do
    Calculate SIMILARITY$(C_i, C_j)$
  end while
  while (! TERMINATION) do
    Select max SIMILARITY$(C_{max_i}, C_{max_j})$
    MERGE$(C_{max_i}, C_{max_j}) = C_{new}$
    Update list of clusters $\mathbb{C}$
    while (For all pairs of clusters $(C_{new}, C_i)$) do
      Calculate SIMILARITY $(C_{new}, C_i)$
    end while
  end while
```

algorithm, the SIMILARITY score, and the TERMINATION condition.

a) The MERGE *Algorithm:* The merge algorithm merges two AP Clusters into one iteratively in the hierarchical clustering algorithm. Two possible merge algorithms are considered in this paper: the NeedlemanWunsch global alignment algorithm and the SmithWaterman local alignment algorithm. The MERGE is a dynamic programming algorithm that, first, recursively builds a score table from the optimal sub-scores by forward-scoring and, then, backtracks through the score table from the optimal score to arrive at the final solution. The runtime for computing the score table of two AP Clusters, C and D, in the dynamic programming algorithm is $O(|C||D|)$, where $C = c_1c_2...c_{|C|}$ is an AP Cluster with aligned site c_i, and $D = d_1d_2...d_{|D|}$ is an AP Cluster with aligned site d_j. Note that depending on the type of similarity score used, their runtime is different.

In the resulting dynamic programming table, the total similarity score of a new AP Cluster that is combined from two old AP Clusters, called S_{AP}, is used to select the optimal AP Clusters to MERGE. To cumulate the AP Cluster score, S_{AP}, a similarity score for matching residues in two aligned columns, called S_{col}, is rewarded and a penalty score for gaps is penalized. Note that the symbol ’_’ is used to represent the opening of the pattern by adding an empty null character, which is called the gap. On the other hand, the symbol ’*’ that pads the beginning and end of AP is used to represent any characters, which is called the wild card. The alignment equation below for calculating the S_{AP} corresponds to the global alignment.

$$S_{AP}[i,j] = max \begin{cases} S_{AP}[i-1, j-1] & +S_{col}(c_i, d_j), \\ S_{AP}[i, j-1] & +GapPenalty(\_, d_j), \\ S_{AP}[i-1, j] & +GapPenalty(c_i, \_). \end{cases} \quad (2)$$

b) The SIMILARITY *Score:* Two major categories of SIMILARITY scores, the sum-of-pair scores and the entropy-based scores, are explored for computing the S_{col} of matching the combined aligned columns of two AP Clusters. The sum-of-pair scores has the runtime of $O(m|C|k|D|)$ and the entropy-based scores has the runtime of $O((m+k)|C||D|)$. To give a formal definition for each of the scores, let $c_i = \begin{bmatrix} c_i^1 & c_i^2 & \dots & c_i^m \end{bmatrix}^T$ be an aligned column for C and $d_j = \begin{bmatrix} d_j^1 & d_j^2 & \dots & d_j^m \end{bmatrix}^T$ be an aligned column for D, where each $c_i^k, d_j^l \in \Sigma$.

The sum-of-pair scores compare all pairs of residues from the two AP Clusters' aligned columns and scores them as S_{one}, which is summed to the S_{col}.

$$S_{col}(c_i, d_j) = \sum_{\forall c_i^k \in c_i} \sum_{\forall d_j^l \in d_j} S_{one}(c_i^k, d_j^l). \quad (3)$$

S_{one} is a pre-defined scoring scheme to compare two individual characters. One possible S_{one} is the Hamming distance, which satisfies the metric properties and thus can be summed. The matches are rewarded, and the mismatches and the gaps are penalized. We adapted a variation of the Hamming distance in order to penalize weighted mismatches and weighted gaps. Table I presents the different S_{one}s and their weightings, where w is the weighting placed on the score.

TABLE I
S_{one} FOR SUM-OF-PAIR SCORES

S_{one}	Match	Mismatch	Gap Penalty
Hamming Distance	$+1$	-1	-1
Weighted Gap	$+1$	-1	$-w$
Weighted Mismatch	$+w$	$-w$	-1

The entropy-based scores constitute more variational information than the sum-of-pair scores. Instead, this category of scores uses the probability distribution of the existing character residues occurring at the combined aligned sites. The two different entropy-based scores considered are

- *Information Entropy Score*

$$\begin{aligned} S_{col}(c_i, d_j) &= H(c_i \cup d_j) & (4) \\ &= -\sum_{\sigma \in c_i \cup d_j} Pr(\sigma) \log Pr(\sigma), & (5) \end{aligned}$$

where $Pr(\sigma)$ is the probability distribution of σ from the combined aligned columns, $c_i \cup d_j$.

- *Information Gain Score* [12]

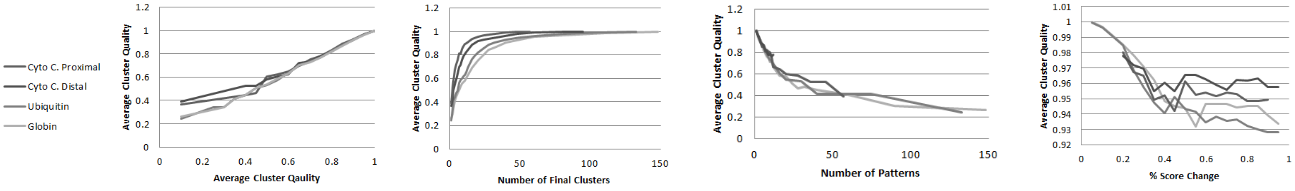
$$S_{col}(c_i, d_j) = w_1 H(c_i) + w_2 H(d_j) - H(c_i \cup d_j), \quad (6)$$

where n is the size of c_i and m is the size of d_j such that $w_1 = \frac{n}{(n+m)}$ and $w_2 = \frac{m}{(n+m)}$.

c) The TERMINATION *Conditions:* The TERMINATION condition of the MERGE algorithms, like the SIMILARITY scores chosen, also determines the quality of the final AP Clusters synthesized. The TERMINATION conditions considered are 1) the threshold on the value of the Average Cluster Entropy, 2) the total number of clusters, 3) the number of patterns in each cluster, and 4) the threshold on the percentage change in the SIMILARITY score.

TABLE II
THE OVERALL AVERAGE CLUSTER QUALITY DEPTH FOR EACH PAIRING OF THE MERGE ALGORITHM COMPARED WITH THE SIMILARITY SCORE

		Sum-Of-Pairs Scores			Entropy Score	
Protein Family	Merge Algorithm	Hamming Distance	Weighted Gap	Weighted Mismatch	Information Entropy Score	Information Gain Score
Proximal Cyto C.	Local	0.0939	0.1267	0.0921	0.0939	0.3502
	Global	0.2420	0.0824	0.6149	0.0838	0.4579
Distal Cyto C.	Local	0.1412	0.1480	0.1412	0.1332	0.3637
	Global	0.3637	0.1332	0.5392	0.1351	0.4728
Ubiquitin	Local	0.1826	0.2223	0.1826	0.1791	0.4605
	Global	0.3938	0.1694	0.5995	0.1359	0.1635



(a) Average Cluster Quality (b) Number of Final Clusters (c) Number of Patterns (d) % Score Change

Fig. 3. Graphs of Termination Conditions compared to Final Cluster Quality, as measured by Average Cluster Quality. Figures (c) to (a) are the Termination Conditions for the Globin Protein, the Ubiquitin Protein, and the Proximal and Distal Binding Segment of the Cytochrome C. Protein.

III. RESULTS AND DISCUSSION

A. Tuning AP Cluster Quality

To determine the parameters that yield the highest quality of AP Clusters, we examined the combinations of the MERGE algorithms, the SIMILARITY scores, and the TERMINATION conditions. We used pFam seed data for this set of tuning experiments: 1) the proximal and 2) distal binding segments of the Cytochrome C., 3) the Ubiuqitin, and 4) the Globin family sequences.

The first set of tuning experiments identifies the optimal combination of the MERGE algorithms with the SIMILARITY scores (Table II). Of the five SIMILARITY scores compared, the sum-of-pairs scores performed better than entropy scores because they use the entire columns of an AP Cluster to account for the scores. Global alignment performs better than local alignment because it aligns the full pattern rather than a sub-sequence of the pattern.

In the second set of tuning experiments, we examined the TERMINATION conditions by fixing the MERGE algorithm to Global Alignment and the SIMILARITY score to Hamming Distance (Fig. 3). We measured the Average Cluster Quality and observed how it varies with the four TERMINATION conditions: 1) the threshold on the value of the Average Cluster Quality, 2) the total number of clusters, 3) the number of patterns in each cluster, and 4) the threshold on the percentage change in the SIMILARITY score. The first TERMINATION condition by Average Cluster Quality (Fig. 3(a)) affects the Average Cluster Quality linearly. The second TERMINATION condition, Number of Clusters (Fig. 3(b)), results in a logarithmic curve, since the threshold point occurs when the quality of the AP Clusters decreases rapidly. In the last two TERMINATION conditions, the Number of Patterns (Fig. 3(c)) and the Percentage Score Change (Fig. 3(d)), loosely fit an inverse function curve, thus demonstrating a threshold point.

B. Biological Experiments

To demonstrate that the binding segments of a protein family can be represented by the AP Clusters, we executed the AP Clustering Step on a list of statistically significant patterns that had resulted from the Pattern Discovery Step. Our goal was to identify protein binding sites within a protein family that correspond to non-exact sequence patterns.

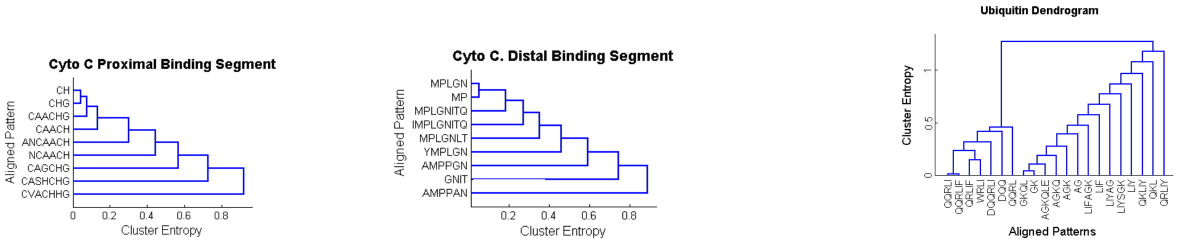
1) Cytochrome C. Results and Discussion: The 238 Cytochrome C. protein pFam seed sequences is uniquely identified by the identification number PF00034. We selectively restricted the two input segments that encloses to the distal and proximal binding residue by length of twenty aligned amino acids each and removed the aligned gaps to create short input segments that surrounded the binding residue.

Table III(a) and Table III(b) show the proximal and distal AP Cluster and their respective dendrogram from the AP Clustering Step. The binding residues that are crucial for the functionality of the Cytochrome C. protein family is in bold.

Our study of the two binding segments of the Cytochrome C. protein family shows that AP Clusters correspond to binding sites. The AP Clusters from these two binding sites are aligned based on their horizontal APs in their rows and are clustered based on the vertical entropy in the aligned columns. First, each AP Cluster contains a set of horizontal APs that are similar to one another. Although these patterns hint at their horizontal significance to the protein, individually, they do not identify the significance of each amino acid residue. Thus, the stability of the aligned columns is important for identifying the binding residue.

Second, the vertical aligned columns anchor their stability and variability in their AP Cluster, which is not possible if each of the horizontal AP is examined individually by itself. Note that the invariant sites of the AP Clusters correspond to binding residues. The three invariant sites in the proximal AP Cluster, His18, Cys17, and Cys14, are essential to the

TABLE III
THE DENDROGRAM AND AP CLUSTERFOR (A) THE PROXIMAL AND (B) THE DISTAL BINDING SEGMENTS OF THE CYTOCHROME C. PROTEIN FAMILY AND (C) THE BINDING SEGMENT OF THE UBIQUITIN PROTEIN FAMILY



$$\begin{pmatrix}
\_ & \_ & \_ & \_ & \_ & C & \mathbf{H} & \_ \\
\_ & \_ & \_ & \_ & \_ & C & \mathbf{H} & G \\
\_ & \_ & C & A & A & C & \mathbf{H} & G \\
\_ & \_ & C & A & A & C & \mathbf{H} & \_ \\
A & N & C & A & A & C & \mathbf{H} & \_ \\
\_ & N & C & A & A & C & \mathbf{H} & \_ \\
\_ & \_ & C & A & G & C & \mathbf{H} & G \\
\_ & \_ & C & A & S & C & \mathbf{H} & G \\
\_ & \_ & C & V & A & C & \mathbf{H} & G
\end{pmatrix}$$

(a) Cyto C. Proximal

$$\begin{pmatrix}
\_ & \mathbf{M} & P & L & G & N & \_ & \_ \\
\_ & \mathbf{M} & P & \_ & \_ & \_ & \_ & \_ \\
\_ & \mathbf{M} & P & L & G & N & I & T \\
I & \mathbf{M} & P & L & G & N & I & T \\
\_ & \mathbf{M} & P & L & G & N & L & T \\
Y & \mathbf{M} & P & L & G & N & \_ & \_ \\
A & \mathbf{M} & P & P & G & N & \_ & \_ \\
\_ & \_ & \_ & \_ & G & N & I & T \\
A & \mathbf{M} & P & P & A & N & \_ & \_
\end{pmatrix}$$

(b) Cyto C. Distal

$$\begin{pmatrix}
\_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & G & \mathbf{K} & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & G & \mathbf{K} & Q & L & \_ & \_ \\
\_ & \_ & \_ & \_ & \_ & \_ & \_ & A & G & \mathbf{K} & Q & L & E & D \\
\_ & \_ & \_ & \_ & \_ & \_ & \_ & A & G & \mathbf{K} & Q & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & \_ & \_ & \_ & A & G & \mathbf{K} & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & \_ & \_ & \_ & A & G & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & L & I & F & A & G & \mathbf{K} & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & L & I & F & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & L & I & Y & A & G & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & L & I & Y & S & G & \mathbf{K} & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & L & I & Y & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & Q & K & L & I & Y & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & Q & K & L & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & Q & R & L & I & Y & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & Q & R & L & I & F & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & Q & Q & R & L & I & F & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & Q & R & L & I & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & Q & Q & R & L & I & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
D & Q & Q & R & L & I & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
D & Q & Q & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ \\
\_ & Q & Q & R & L & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_
\end{pmatrix}$$

(c) Ubiquitin

functionality of the Cytochrome C. protein family for binding the heme ligand (Fig. 1(a)). More precisely, the His18 invariant site acts as the proximal binding residue to the heme iron, and the two Cysteine invariant sites, Cys14 and Cys17, link the two thioether bonds to the two vinyl groups on the heme. Similarly, the Met62 invariant site in the distal AP Cluster acts as the distal binding residue to the heme iron from the opposite distal side of the protein. Our experiments showed that the AP Clusters that are discovered by our AP Clustering Step included individual AP Clusters that contains the protein binding sites. Our proximal AP Cluster for Cytochrome C. is consistent with the PROSITE consensus motif, **`[C]-x(2)-[CH]`**[13] and with the strong pFam emission probability [9]. However, our distal binding AP Cluster for Cytochrome C. was not identified in PROSITE nor pFam, but the invariant site of the AP Cluster is confirmed as the distal binding residue in the three-dimensional structure of the Cytochrome C. protein (Fig. 1(a)). These resulting AP Clusters represent the proximal and distal binding segments, and each AP Cluster contains invariant sites that correspond to the binding residues, which are the main biological function of the Cytochrome C. protein family. The binding residues, represented by invariant sites, are surrounded by horizontal APs that form their corresponding functional binding segments.

2) Ubiquitin Results and Discussion: To establish the iterative steps of the AP Clustering Step as well as to show the resulting AP Clusters, we applied our method to the Ubiquitin protein family. The downloaded input sequences from pFam are from the Ubiquitin protein family, which is uniquely identified in pFam by the family identification PF00240. This family of Ubiquitin sequences contains 78 essential sequences that have a maximum length of 83. We showed the largest resulting AP Cluster, which is a set of patterns that have been grouped and aligned and we displayed its dendrogram to demonstrate the capability of our AP Synthesis Process, where the hierarchical clustering algorithm iteratively merges the AP Clusters to form the final AP Cluster (Table III(c)).

The top eight AP Clusters shown in Fig. 4 correspond to five of the seven binding residues : Lys6, Lys11, Lys33, Lys48, and Lys63 (Fig. 1(b)). The remaining two undiscovered binding residues, Lys27 and Lys29, were discovered in the Pattern Discovery Step with high statistical significance but not aligned and clustered in the AP Clustering Step() due to insufficient variance with these two binding segments. This pattern can be considered an AP Cluster that groups and aligns only one pattern.

For Ubiquitin, our results agreed with the pFam's profile HMM emission probability and not PROSITE's consensus motif, which misses 172 Ubiquitin proteins (Fig. 4). Of the twelve AP Clusters discovered, nine agreed with the pFam HMM emission probability and three did not. Our AP Clusters are short local alignment of patterns, which agree with the pFam global alignments. Further analysis of the sequence position of each occurrence of these three false positive AP Clusters is needed to remove them.

IV. CONCLUSION

In summary, our AP Synthesis Process is able to identify AP Clusters that correspond to protein binding segments for the Cytochrome C. and the Ubiquitin protein families from the pFam model. An AP Cluster represent APs as the rows and the aligned columns as columns, which can be further evaluated for site variation. In fact, for Cytochrome C., the invariant sites in the proximal and distal AP Cluster are the

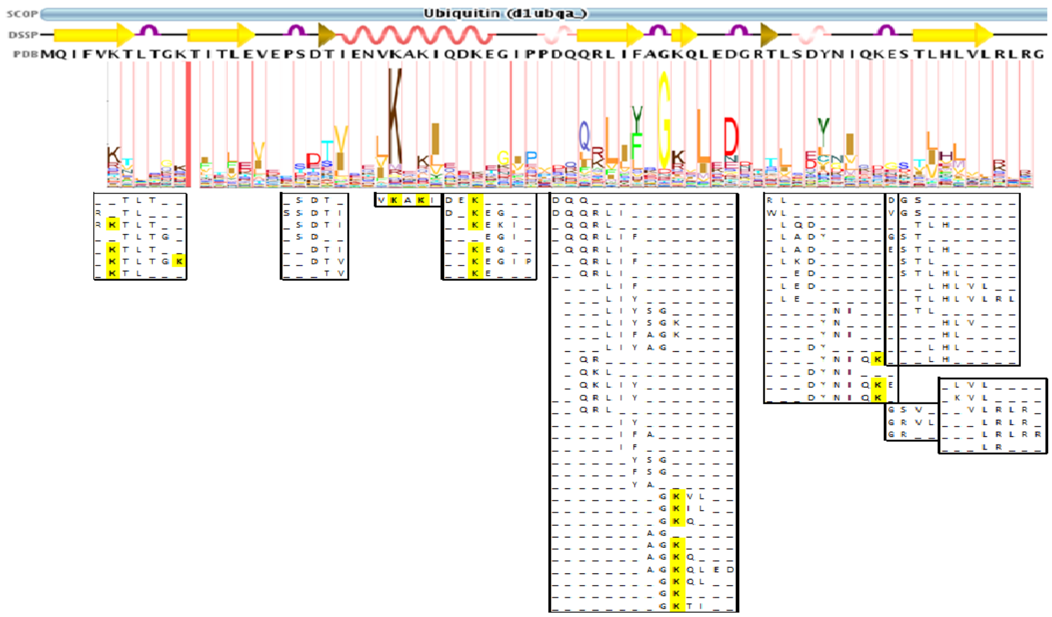

SCOP
Ubiquitin (d1ubqa_)
DSSP
PDB
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRG

Fig. 4. Our resulting twelve AP Clusters are compared to the protein sequence of a single protein sequence (SCOP, DSSP, PDB), 1UBQ from the protein data bank, and the profile Hidden Markov Model logo from pFam. The seven Lys binding residues for the Ubiquitin protein family are highlighted in yellow in the AP Cluster.

binding residues. The AP Clusters in Ubiqiuitin agree with pFam and, furthermore, discover the binding residues from the AP Clusters as invariant sites. The synthetic experimental results demonstrate that the quadratic runtime of our algorithm is faster than the exponential runtime of the combinatorial method, because the search space in the AP Clustering Step starting with input patterns is smaller than the search space for starting with multiple sequences. The AP Cluster quality experiments consider the conditions that are optimal for generating high Average Cluster Entropy. The tuning experiments on Cytochrome C. and Ubiquitin demonstrate that Global Alignment with Hamming Distance yields the optimal Average Cluster Entropy, and that Termination Conditions exhibit logarithmic and linear characteristics.

AP Clusters can be used to search functional domains across different protein families. This knowledge could be essential for understanding the proteins involved in RNA regulation, such as the tryptophan-activated RNA-binding attenuation protein (TRAP), which binds the RNA to regulate gene expression [14].

Acknowledgement

The authors thank the reviewers for their constructive feedback, D. Zhang for efficient implementation of the suffix tree, S. Fung. for running experiments, and D. Leong, C. M. Li, B. Chang, and D. Yuen for helpful feedback on this manuscript. This research is supported by NSERC Post Graduate Scholarship and NSERC Discovery Grant.

References

[1] P. Pevzner and S. Sze, “Combinatorial approaches to finding subtle signals in dna strings,” *In Proc. ISMB*, vol. 2000, pp. 269–278, 2000.

[2] J. Buhler and M. Tompa, “Finding motifs using random projections.” *J Comput Biol*, 2002.

[3] Y. T. Hideya Kawaji and H. Matsuda, “Graph-based clustering for finding distant relationships in a large set of protein sequences,” *Bioinformatics*, vol. 20, pp. 243–252, 2004.

[4] R. Patwardhan, H. Tang, S. Kim, and M. Dalkilic, “An approximate de bruijn graph approach to multiple local alignment and motif discovery in protein sequences,” *Data Mining and Bioinformatics*, vol. 4316, pp. 158–169, 2006.

[5] M. Li, B. Ma, and L. Wang, “Finding similar regions in many strings,” *Journal of Computer and System Sciences*, vol. 65, pp. 73–96, 2002.

[6] P. A. Evans, A. Smith, and H. T. Wareham, “On the complexity of finding common approximate substrings.” *Theoretical Computer Science*, vol. 306(3), pp. 407–430, 2003.

[7] J. C. Jeong, X. Lin, and X.-W. Chen, “On position-specific scoring matrix for protein function prediction,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8(2), pp. 308–315, 2011.

[8] A. K. C. Wong, D. K. Y. Chiu, and S. C. Chan, “Pattern detection in biomolecules using synthesized random sequence,” *Journal of Pattern Recognition*, vol. 29:9, pp. 1581–1586, 1995.

[9] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman, “The pfam protein families database,” *Nucleic acids research*, vol. 38, pp. D211–22, 2010.

[10] A. K. Wong, D. Zhuang, G. C. Li, and E.-S. A. Lee, “Discovery of delta closed patterns and non-induced patterns from sequences,” *IEEE Transactions on Knowledge and Data Engineering Journal*, 2011.

[11] E.-S. A. Lee and A. K. C. Wong, “Synthesizing aligned random pattern digraphs from protein sequence patterns,” *Bioinformatics and Biomedicine Workshops (BIBMW)*, pp. pp. 178 – 185, 2011.

[12] S. C. Chan and A. K. C. Wong, “Synthesis and recognition of sequences,” *IEEE Trans on PAMI*, vol. 13(12), pp. 1245–1255, 1991.

[13] S. CJA, C. L, de Castro E, a. B. V. Langendijk-Genevaux PS, B. A, and H. N., “Prosite, a protein domain database for functional characterization and annotation,” *Nucleic Acids Res*, vol. 38(Database issue), pp. 161–166, 2010.

[14] E. A. Dethoff, J. Chugh, A. M. Mustoe, and H. M. Al-Hashimi, “Functional complexity and regulation through rna dynamics,” *Nature*, vol. 482, pp. 322–330, 2012.