Aligning Discovered Patterns from Protein Family Sequences

En-Shiun Annie Lee, Dennis Zhuang, and Andrew K.C. Wong

University of Waterloo,
Centre of Pattern Analysis and Machine Intelligence,
200 University Avenue West, Waterloo, Ontario N2L 3G1
http://www.pami.uwaterloo.ca

Abstract. A basic task in protein analysis is to discover a set of sequence patterns that characterizes the function of a protein family. To address this task, we introduce a synthesized pattern representation called Aligned Pattern (AP) Cluster to discover potential functional segments in protein sequences. We apply our algorithm to identify and display the binding segments for the Cytochrome C. and Ubiquitin protein families. The resulting AP Clusters correspond to protein binding segments that surround the binding residues. When compared to the results from the protein annotation databases, PROSITE and pFam, ours are more efficient in computation and comprehensive in quality. The significance of the AP Cluster is that it is able to capture subtle variations of the binding segments in protein families. It thus could help to reduce time-consuming simulations and experimentation in the protein analysis.

Keywords: Protein Analysis, Protein Function Identification, Pattern Discovery, Pattern Clustering, Hierarchical Clustering, Motif Finding, Local Alignment, Approximate String Matching.

1 Introduction

Proteins are involved in many biological processes of the organism, from enzyme catalysts to ligand binding. To rapidly and reliably find out from the primary sequence to which known protein family an uncharacterised protein belongs will help to understand its functions and roles in the cellular processes. A protein often assumes a specific function such as binding or enzymatic activity and thus its functionality constrains regions such as binding sites. In addition, domains are less subject to mutations, giving rise to certain discernible conserved segments in the primary sequence. In another word, proteins in the same family can be homologues or distantly related in their primary sequence but they might contain conserved segments (often called motifs, patterns, or fingerprints). It is therefore important to discover such conserved areas that characterize a protein family.

There are different approaches for identifying the similar conserved regions in the protein sequences that characterize a protein family. One approach is multiple sequence alignment, which takes a set of protein sequences aligned by

T. Shibuya et al. (Eds.): PRIB 2012, LNBI 7632, pp. 243–254, 2012.

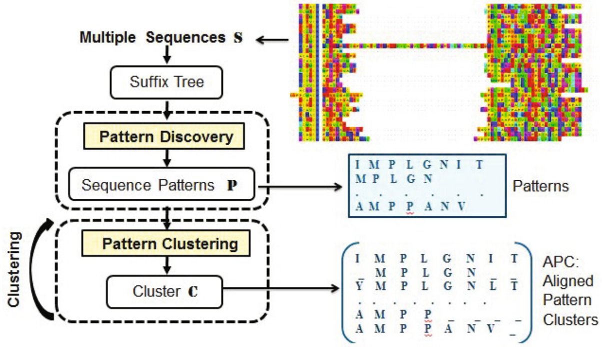


Fig. 1. The overview of the Pattern Alignment (PA) Process. Our method involves two steps: the Pattern Discovery Step and the Pattern Alignment Step. The results are the APC.

dynamic programming to come up with homologous regions, that may be of consequence of functional, structural, or evolutionary relationships. CLUSTAL W [1], T-Coffee [2], DIALIGN [3] and HMMER [4] are the representatives of such approaches. Concerning the computation complexity, it has been shown that finding the global optimal alignment is an NP-complete problem [5]. Even with heuristics, it is still not efficient enough to handle large scale dataset. Moreover, this approach is considered as more suitable for sequences that are globally homologous with have high level of similarity. The result would be unsatisfactory if the sequences are only distantly related or just share local similarities. Another approach is multiple local alignment, unlike the multiple sequence alignment [6] that aligns the whole sequence, attempts to locate and align locally similar subsequences and build up a probabilistic model for describing the conserved regions that represent the motif in the sequences. Hence it is also called the motif finding approach. A commonly used model is position weight matrix (PWM), which assumes independent position in the motif. However such assumption is not realistic in many cases. Furthermore, it is computationally expensive to obtain global optimum. Hence, heuristics such as Expectation Maximization and Gibbs Sampling are used to find the locally optimal model. Two well known methods are MEME [7] and GLAM [6]. This approach often returns one or more highest score solutions. It is likely to miss those motifs that are statistically significant. Furthermore, the reported motifs often have high false positive rate [8]. Another approach is to generate sequence patterns that repeat sufficient times precisely or approximately with variations in the sequences in an exhaustive fashion. YMF and Weeder are two examples for such approach. However, the common problem is that there are usually too many patterns discovered and each pattern often partially characterizes the functional regions in the sequences since even the functional sites may exhibit a certain degree of variability. To overcome this limitation, we present in this paper a new method that groups and aligns the similar patterns discovered by a sequence pattern discovery algorithm into aligned pattern clusters. The aligned pattern cluster is able to align significant patterns while capturing more variability.

Aligned Pattern (AP) Clusters are used to reveal and represent protein functional segments. For eachAPC obtained, We examine whether it corresponds to binding segments or other protein functional segments. When we applied our PA Process to the Cytochrome C. and Ubiquitin protein families, we did find such strong correspondence. Our PA Process is efficient. The results obtained are consistent with the motifs found in the two well known databases: pFam and PROSITE. This shows that the APCs obtained capture the functional regions of a protein family.

2 Methods

Our method (Fig. 1) takes the sequence patterns obtained by a previously developed method as input, and groups and aligns them into aligned pattern clusters. The resulting knowledge-rich representations is abbreviated as APC . We will briefly describe the pattern discovery process, but will focus mainly on the pattern aligning and clustering process.

2.1 Discover Sequence Pattern

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. For protein sequences, the alphabet can be the 20 amino acids. The pattern discovery method in [13] takes the multiple sequences and produces a list of sequence patterns $\mathbb{P} = \{P_1, P_2, \ldots, P_{|\mathbb{P}|-1}, P_{|\mathbb{P}|}\}$. Each pattern P is essentially a substring from the input sequences but passes three conditions. First, it is frequent, that is, it repeats itself sufficiently many times in the input sequences. Second, it is statistically significant, meaning that the pattern is not resulted by the random associations of elements given a random background model. Third, it is not redundant compared against the other patterns in the result set. The information provided by a non-redundant pattern cannot be accounted by other patterns. With these three conditions, a compact yet informative set of patterns are obtained. The running time of the pattern discovery process takes linear time to the input size and thus is efficient. The discovered patterns correspond to potential functional segments in the sequences. We devised a score to rank the patterns according to their interestingness. The score is $s = \frac{q_P}{N} \cdot z_P$, where q_P is the number of sequences where the pattern P appears, N is the number of sequences, and z_P is the statistical significance.

2.2 Aligning Similar Patterns

For the task of pattern alignment, we develop an algorithm which groups a set of similar patterns of different lengths obtained from the pattern discovery process and then align and cluster 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 on the same column, thus implying a common functionality among the APs [9]. An APC, 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

$$C = \texttt{Align}\begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_m \end{pmatrix} = \begin{pmatrix} s_1^1 & s_2^1 & \dots & s_n^1 \\ s_1^2 & s_2^2 & \dots & s_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ s_1^m & s_1^m & \dots & s_n^m \end{pmatrix}_{n \times m} , \quad (1)$$

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.

An ALIGNED PATTERN $P = s_1^P s_2^P ... s_{|P|}^P$ 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. An ALIGNED COLUMN c_j in C represents the j^{th} column of characters from the set of APs forming the current APC. Thus, $C = \begin{pmatrix} c_1 \, c_2 \dots c_n \end{pmatrix}$.

The Alignment Algorithm. The algorithm iteratively ALIGNs two APCs in a pairwise-manner based on their ALIGNMENT score and that they do not lie on the same sequences. The alignment algorithm combines two APC into one iteratively in the hierarchical manner. Two possible alignment algorithms are considered in this paper: the NeedlemanWunsch alignment algorithm, which is global, and the SmithWaterman alignment algorithm, which is local. The ALIGNMENT is essentially 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 APCs, $\mathbb{P}_1$ and $\mathbb{P}_2$, in the dynamic programming algorithm is $O(|\mathbb{P}_1||\mathbb{P}_2|)$. Note that depending on the type of alignment score used, there may be an added linear time of complexity described in the next section.

The Alignment Score. Two major categories of ALIGNMENT scores are explored for computing the score of matching the combined aligned columns of two APCs: the sum-of-pair scores and the entropy-based scores. The sum-of-pair scores has the runtime of $O(m|\mathbb{P}_1|k|\mathbb{P}_2|)$ and the entropy-based scores has the runtime of $O((m+k)|\mathbb{P}_1||\mathbb{P}_2|)$.

Sum-of-Pair Scores. The sum-of-pair scores compare all pairs of residues from the two APCs' aligned columns and scores them using Hamming Distance. In addition to Hamming Distance, we also considered weighting the penalty of the Hamming Distance to prefer gaps or to prefer mismatches.

Entropy-Based Scores. 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 the Information Entropy Score and the Information Gain Score.

The Stopping Conditions. The Stopping condition of the Alignment algorithms, like the Alignment scores chosen, also determines the quality of the final resulting APCs. The Stopping conditions considered are the Number of Patterns per Cluster and the Final Number of Clusters.

3 Synthetic Results and Discussion

For demonstrating the runtime and quality of our method, we created nine sets of synthetic input data containing synthetic patterns of length 10, where each pattern occurs with a frequency of five and pattern has a 10% chance of mutation at a random position from the previous pattern. These nine datasets vary based on the number of synthetic patterns in each set in increments of five.

The Runtime Comparison. To compare the runtime of our PA Process against the combinatorial method, we plotted the experimental runtime of our PA Process. We measured the experimental runtime of our PA Process by counting the number of character comparisons and plot it against the number of synthetic patterns in the dataset. Five Alignment scores are plotted for the global alignment and for the local alignment resulting in ten combinations. As described in the methodology section, the pairwise-sum-of scores performed slower than the entropy scores due to a more complete pairwise comparisons (Fig. 2).

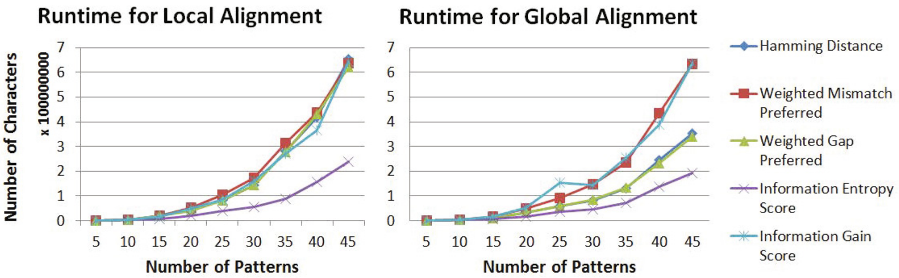


Fig. 2. The five Alignment scores are Hamming Distance, Weighted Mismatch Preferred, Weighted Gap Preferred, Information Entropy Score, and Information Gain Score. The first graph compare the runtime of five Alignment scores while executing local Alignment Algorithm while the second graph executes global Alignment Algorithm.

The Alignment Algorithm and Score. To determine the parameters that yield the highest quality of APCs, we examined the combinations of the Alignment algorithms, and the Alignment scores. We compare the resulting quality of the APC by computing the Average Cluster Entropy of all the normalized entropy of the final clusters and their columns. The first set of tuning experiments identify the optimal combination of the Alignment algorithms with the Alignment scores (Fig. 3). Of the five Alignment scores compared, the sum-of-pairs scores performed better than entropy scores because they exhausively compare all pairs of amino acids from both aligned columns and take longer to execute. These observations indicate that sum-of-pair scores tend to perform better than entropy-based scores because these scores use the full residue and take longer to run. Global alignment performs better than local alignment because it aligns the full pattern rather than a sub-sequence of the pattern.

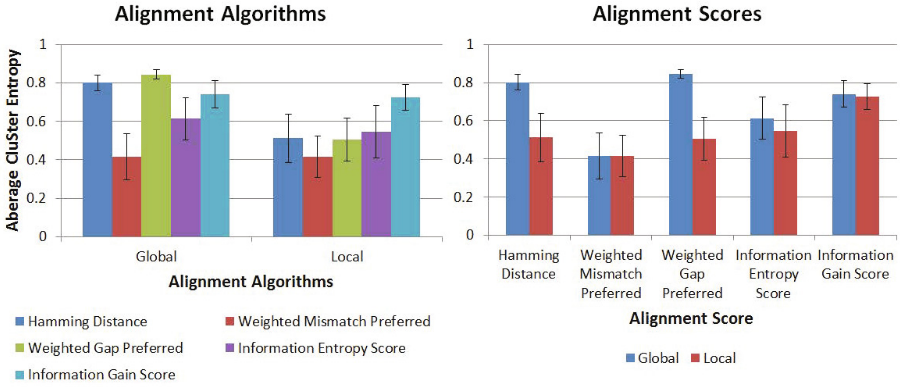


Fig. 3. The first graph divides the ten Average Pattern Quality into the two Alignment Algorithms. For global alignment, the Hamming Distance is the best Alignment Score; for local alignment, the Information Gain Score is the best. The second graph divides the ten Average Pattern Quality into the five Alignment Scores. Of the two Alignment algorithms compared, the global alignment results in a better APC than the local alignment.

The Stopping Conditions. To examine the properties of the Stopping conditions, we fixed the Alignment algorithm to Global Alignment and the Alignment score to Hamming Distance. We measured the Average Cluster Quality and observed how it varies with the two Stopping conditions (Fig. 4): 1) the Number of Patterns per Cluster, and 2) the Final Number of Clusters. The threshold is adjusted for each set of synthetic patterns. The first Stopping condition, the Final Number of Clusters, results in an inverse exponential curve, since the threshold point occurs when the quality of the APCs decreases rapidly.

There is an ideal threshold point where the quality of the APC is close to the optimal value of one and increases slowly. The Second Stopping condition by the number of clusters fits a logarithmic curve, because decreasing the number of clusters also increases the number of patterns which in turn increases the cluster entropy.

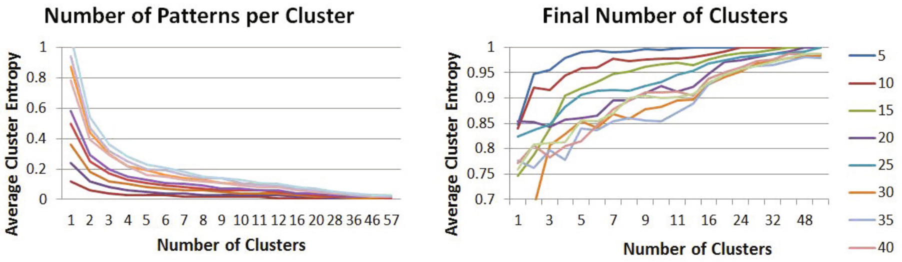


Fig. 4. Finally the two conditions considered for stopping the alignment are: 1) the Number of Patterns per Cluster, which fits a inverse exponential curve, and 2) the Final Number of Clusters, which fits a logarithmic curve

4 Biological Results and Discussion

4.1 Cytochrome C. Results and Discussion

To demonstrate that the binding segments of a protein family can be represented by APCs, we executed the pattern alignment method on a list of patterns that had resulted from the pattern discovery process. The downloaded input sequences from pFam are from the protein family Cytochrome C., which is uniquely identified by the pFam family identification number PF00034. The pFam seed sequences of the Cytochrome C. contains 238 essential sequences. The two binding residues in the Cytochrome C. protein that binds the heme ligand are (1) the proximal binding segment that binds the heme ligand from the proximal side of the protein and (2) the distal binding segment that binds the opposite side of the heme ligand from the distal side of the protein.

Table 1 shows the top ranked patterns. Most of them correspond to the proximal and distal binding segments for the Cytochrome C. protein family. Nineteen of these top twenty patterns contain the binding residues that is crucial for the binding functionality of the Cytochrome C. family protein. However, each pattern, on its own, has a small fraction of supporting sequences and hence a single pattern alone cannot represent variety of the functional binding segments in the protein sequences. However,the APC, containing a set of similar patterns with variations, provides a much more detailed description of the binding segments and are able to capture their variability.

Table 1. Top 20 Patterns in the Full Sequences of the Cytochrome C. Family

Rank	Pattern	Frequency	Score	Binding Residues
1	ADRGEKLYQKVGCV	8	1179941.62	
2	CSMC**H**AREPVW	6	55750.35	H18
3	GRCSMC**H**AREP	6	23786.79	H18
4	RCSMC**H**AREP	8	12410.76	H18
5	I**M**PLGNITQMT	5	11628.94	M62
6	CSMC**H**AREP	11	5021.18	H18
7	**M**PLGNITQMT	6	3763.97	M62
8	GRCSMC**H**A	11	928.88	H18
9	RCSMC**H**A	16	576.93	H18
10	MC**H**AREP	13	250.46	H18
11	**M**PLGNITQ	7	202.92	M62
12	CSMC**H**A	19	174.14	H18
13	SHA**M**PPAN	6	117.56	M62
14	GVSHA**M**PP	6	117.14	M62
15	HA**M**PPANV	5	79.82	M62
16	**M**PLGNIT	8	57.37	M62
17	HA**M**PPAN	8	47.54	M62
18	MC**H**AAEP	6	33.14	H18
19	S**H**AMPP	12	32.01	M62
20	CAAC**H**	22	27.97	H18

Table 2. Top 20 Patterns in the Full Sequences of the Ubiquitin Protein Family

Rank	Pattern	Frequency	Score	Binding Residues
1	TLHLVLRL	5	161.28	
2	DYNIQ**K**E	5	104.63	Lys63
3	DYNIQ**K**	7	55.28	Lys63
4	AG**K**QLED	5	53.62	Lys48
5	QQRLIF	7	39.87	
6	LIFAG**K**	7	39.25	Lys48
7	YNIQ**K**	9	23	Lys63
8	DQQRLI	6	19.96	
9	LIYSG**K**	5	17.47	Lys48
10	QQRLI	11	16.88	
11	IFAGK	8	16.81	
12	QRLIF	9	16.61	
13	**K**EGIP	9	15.66	Lys33
14	**K**TLTG**K**	6	13.49	Lys6, Lys11
15	V**K**A**K**IQ	5	13.26	Lys27,Lys29
16	LHLVL	10	11.85	
17	QRLIY	7	10.95	
18	LIYSG	7	10.36	
19	LIYAG	6	9.51	
20	ESTLH	6	7.1	

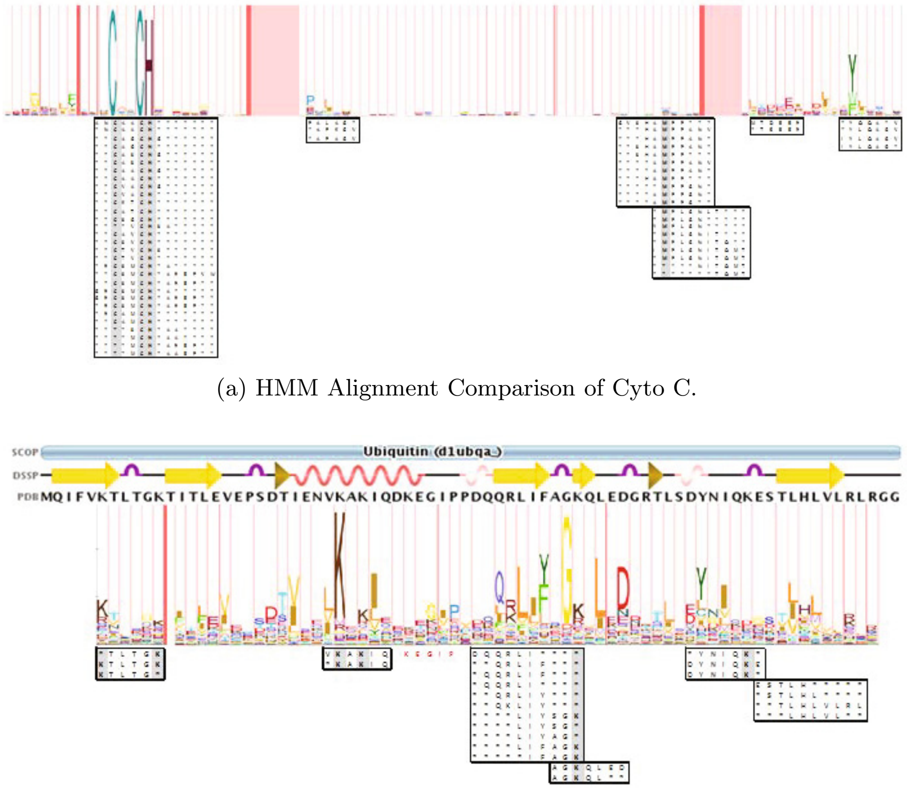
In our first biological study, we showed that protein functional segments can be represented by a set of patterns called an APC , built using our PA Process. The set of discovered APCs are displayed with pFam's alignment represented by HMM Logo (Fig. 5a). The APCs contained invariant sites in their columns and APs in its rows. For the rightmost proximal APC, the three top invariant sites, His18, Cys17, and Cys14 in their proper location, are essential to the functionality of the Cytochrome C. protein family for binding the heme ligand. More precisely, the His18 invariant site acts as the proximal binding residue to the heme iron, and the two Cysteines 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 APC acts as the distal binding residue to the heme iron from the opposite distal side of the protein. These resulting APCs contain invariant sites corresponding to the binding residues, which are the main biological function of Cytochrome C. protein family. Also, the binding residues, represented by invariant sites, are surrounded by APs that form the functional binding segment.

Our discovered proximal APC for Cytochrome C. is consistent with the proximal binding motif [C]-x(2)-[CH] from PROSITE [10,11]and the strong emission probability from pFam [12,13]. Moreover, our method identified the distal binding APC , whereas PROSITE does not annotate this APC as a binding motif and pFam only identifies it as a weak emission probability.

4.2 Ubiquitin Results and Discussion

We applied our method to the Ubiquitin protein family. The input sequences from pFam are from the Ubiquitin protein family, which is uniquely identified in pFam by the family identification PF00240 and contains 78 essential sequences that have a maximum length of 83. Table 2 shows that many top ranked patterns correspond to the seven binding residues of the Ubiquitin protein. Other patterns correspond to the conserved elements around the binding residues. Though the discovered patterns do indicate some important functional signals in this family of Ubiquitin proteins, each pattern on its own has only a small fraction of supporting sequences and thus achieve a low sensitivity in representing the binding segments of this protein family. Proteins often exhibit great variability and thus APC would represent its functional sites more effectively and explicitly.

In our Ubiquitin experiment, we executed our PA Process on the multiple unaligned sequences of the Ubiquitin protein family. The Ubiquitin contains seven lysine residues, Lys6, Lys11, Lys27, Lys29, Lys33, Lys48, and Lys63 that can be linked to another Ubiquitin to form a poly-Ubiquitin chain [14–18]. The six APCs contain five out of the seven binding residues, however two remaining binding residues, Lys27 and Lys29, was not sufficient variants to be aligned and grouped into APCs in the pattern alignment process (Fig. 5b). For Ubiquitin, our results did not agree with the PROSITE consensus motif for the Ubiquitin domain signature, K-x(2)-[LIVM]-x-[DESAK]-x(3)-[LIVM]-[PAQ]-x(3)-Q-x-[LIVM]-[LIVMC]-[LIVMFY]-x-G-x(4)-[DE], which misses 172 Ubiquitin proteins. However, our results did agree with the profile HMM's emission probability in pFam.



(a) HMM Alignment Comparison of Cyto C.

(b) HMM Alignment Comparison of Ubiquitin

Fig. 5. In figure (a) two of the largest resulting APCs represent the proximal and distal binding segments of the Cytochrome C. are compared to the HMM logo from pFam. In the largest APC Cys14, Cys17, and His18 are identified as the invariant sites. In the second and third largest APCs that overlap, Met62 is the invariant site of the distal binding segment where Met62 binds the heme iron. In Figure (b) the four resulting binding segments for the Ubiquitin protein family are compared to the HMM logo from pFam. The six discovered APCs contain five out of the seven binding residues: Lys6, Lys11, Lys33, Lys48, and Lys63.

5 Conclusion

In summary, our PA Process is able to identify APCs that correspond to protein binding segments for the Cytochrome C. and the Ubiquitin protein family. The APCs shows APs as its rows and residue variations in its aligned columns, which captures binding segment variations. In fact, for Cytochrome C., the invariant sites in the proximal APC are the binding residues as identified in PROSITE and pFam. However, the distal APC identifies an invariant site as the binding residue which is not identified in PROSITE. Hence, APCs can render much more effective protein analysis by automatically finding and grouping similar patterns from the sequences and narrowing down the important segments to be examined.

Acknowledgment. The authors thank the reviewers for their constructive feedback, Dr. K. Durston for insightful analysis of molecular biophysics, and D. Leong, C. M. Li, B. Chang, and D. Yuen for reading this manuscript. This research is supported by an NSERC Post Graduate Scholarship, Ontario Graduate Scholarship and an NSERC Discovery Grant.

References

1. Thompson, J.D., Higgins, D.G., Gibson, T.J.: Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22(22), 4673–4680 (1994)
2. Notredame, C., Higgins, D.G., Heringa, J.: T-coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302(1), 205–217 (2000)
3. Subramanian, A.R., Kaufmann, A.M., Morgenstern, B.: Dialign-tx: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol. Biol. 3, 6 (2008)
4. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press (1998)
5. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. Journal of Computational Biology 1(4), 337–348 (1994)
6. Frith, M.C., Hansen, U., Spouge, J.L., Weng, Z.: Finding functional sequence elements by multiple local alignment. Nucleic Acids Res. 32(1), 189–200 (2004)
7. Bailey, T.L., Elkan, C.: Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning 21(1/2), 51–80 (1995)
8. Pisanti, N., Crochemore, M., Grossi, R., Sagot, M.F.: Bases of motifs for generating repeated patterns with wild cards. IEEE/ACM Transactions on Computational BIology and Bioinformatics 2(1), 40–50 (2005)
9. Lee, E.-S.A., Wong, A.K.C.: Synthesizing aligned random pattern digraphs from protein sequence patterns. In: Bioinformatics and Biomedicine Workshops (BIBMW), pp. 178–185 (2011)
10. Bairoch, A.: Prosite: a dictionary of sites and patterns in proteins. Nucleic Acids Research 19, 2241–2245 (1991)
11. Sigrist, C.J.A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., Hulo, N.: Prosite, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38(Database issue), 161–166 (2010)

12. Sonnhammer, E.L., Eddy, S.R., Durbin, R.: Pfam: A comprehensive database of protein domain families based on seed alignments. PROTEINS: Structure, Function, and Genetics 28, 405–420 (1997)
13. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A.: The pfam protein families database. Nucleic Acids Research 211, D211–D222 (2010)
14. Peng, J., Schwartz, Elias, Thoreen, Cheng, Marsischky, Roelofs, et al.: A proteomics approach to understanding protein ubiquitination. Nature Biotechnology 21(8), 921–926 (2003)
15. Xu, P.P.: Characterization of polyubiquitin chain structure by middle-down mass spectrometry. Analytical Chemistry 80(9), 3438–3444 (2008)
16. Kirisako, T., Kamei, K., Kato, M., Fukumoto, Kanie, Sano, Tokunaga: A ubiquitin ligase complex assembles linear polyubiquitin chains. The EMBO Journal 25(20), 4877–4887 (2006)
17. Kim, H., Kim, Lledias, Kisselev, S., Skowyra, Gygi, Goldberg: Goldberg: Certain pairs of ubiquitin-conjugating enzymes (e2s) and ubiquitin-protein ligases (e3s) synthesize condegradable forked ubiquitin chains containing all possible isopeptide linkages. The Journal of Biological Chemistry 282(24), 17375–17386 (2007)
18. Ikeda, F.: Dikic: Atypical ubiquitin chains: new molecular signals. 'protein modifications: Beyond the usual suspects' review series. EMBO Reports 9 (6), 536–542 (2008)