SMGroup

SM Journal of Bioinformatics and Proteomics

Research Article

Unsupervised Pattern Discovery in Biosequences Using Aligned Pattern Clustering

En-Shiun Annie Lee*, Antonio Sze-To, Andrew KC Wong and Daniel Stashuk

Department of Systems Design Engineering, Centre for Pattern Analysis and Machine Intelligence, University of Waterloo, Canada

Article Information

Received date: Jul 21, 2016
Accepted date: Aug 10, 2016
Published date: Aug 12, 2016

***Corresponding author**

En-Shiun Annie Lee, Department of Systems Design Engineering, Centre for Pattern Analysis and Machine Intelligence, University of Waterloo, Canada, Email: annie.lee@uwaterloo.ca

Abstract

Bioseqeunces such as protein, RNA and DNA, are made up of sequences of amino acids/nucleotides. The binding of biosequences among themselves is important for governing many biological processes of a living organism. The bindings are maintained by short segments of these biosequences, known as functional elements. Due to the importance of these functional elements, their presence is well conserved throughout evolution, allowing them to be discovered as patterns. As sequencing technologies continue to improve, the amount of biosequences is available in abundance. It is thus convenient and cost-effective if functional elements can be discovered from biosequences data computationally in an unsupervised manner without the need of prior knowledge or costly pre-preprocessing. In this paper, we aim to give a brief review of an unsupervised pattern discovery tool known as Aligned Pattern Clustering (or its software WeMine™). It is developed to facilitate the discovery and analysis of patterns in biosequences, and has been applied in1) unsupervised identification of protein binding sites; 2) revealing functioning subgroup characteristics; and 3) identification of intra-protein, inter-protein and protein- DNA binding sites. In the era of ever-expanding biosequence data, we believe that this unsupervised pattern discovery approach would render a reliable, robust, and scalable method for scientific discovery and applications through leveraging the ever expanding volume of biosequences.

Introduction

Protein, RNA and DNA are made up of sequences of amino acids/nucleotides, which interact among themselves via binding. For example, (1) protein-DNA binding regulates gene transcription [1]; and (2) Protein-protein binding plays important roles in cell cycle control and signal transduction [2].The binding is maintained by either the direct participation or assistance of conserved short segments of biosequences called functional elements. Because of their importance in preserving function, they are well conserved throughout evolution. Their recognition is therefore essential for an in-depth understanding of the biological mechanisms [3] such as inhibitor design [4]. Although these functional elements could be discovered from the three-dimensional structural forms of the biosequences, the applicability is limited due to the high experimental cost. With the advent of new sequencing technologies [5], it is preferable to discover, directly from the abundant biosequence data, functional elements where many of them are short with variable length, like Short Linear Motifs (SLiMs [6]) which play important roles in protein-protein interaction but are only 3 to 15 amino acids in length. Such short elements could not be captured well by the popular position weight matrices [7]. In this paper, we aim to briefly review an unsupervised pattern discovery tool known as Aligned Pattern Clustering (or its software WeMine™) [8-11] which is developed to facilitate the discovery and analysis of patterns in biosequences. Its applications include 1) identifying functional elements in protein sequences [8-11,2] revealing functioning subgroup characteristics of functional elements [12-14,3] identifying co-occurring intra-protein [15,16], inter-protein [17] and protein-DNA functional elements [18,19].

Methodology

Our Aligned Pattern Clustering algorithm [8-11] is packaged as a software tool named WeMine™. We first discuss its rationale, and then briefly introduce its methodology followed by illustrative applications.

Rationale

Figure 1 gives an overview of how a set of biosequence data could be turned into useful knowledge. In the knowledge discovery sense, both Pattern Discovery and AP Clustering find the "what" and "where" of biologically conserved functional units/regions from purely sequence data without relying on prior clue or knowledge. We refer the "what" as the pattern space which reveals statistically significant residue/nucleotide associations and the "where" as the data space to demarcate the location of patterns or APCs in a set of biosequences. Table 1 depicts the role of our pattern-data

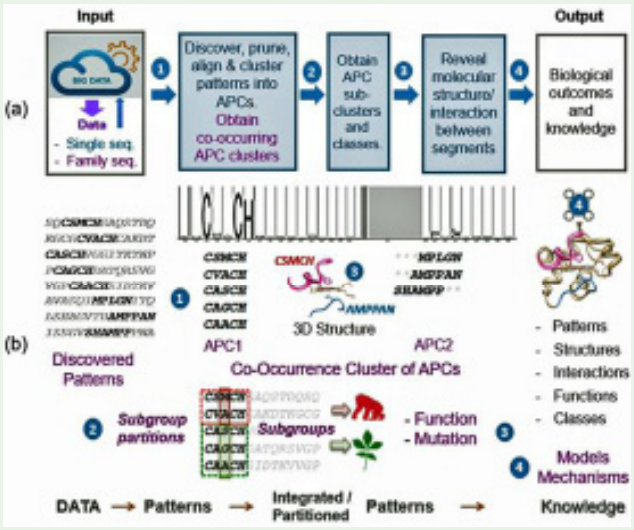

Figure 1: Overview of WeMine: a software for discovering, pruning, aligning and clustering sequence patterns to reveal elements, conserved domains and their characteristics with local and co-occurring biological functionality. **(a)** Describes the major steps of WeMine. It begins with a data set of a sequence family. If only a single sequence is available, it could draw data close to or within its family from the cloud using PSI-BLAST [26]. From the data, in step 1, it uses our PD algorithm to discover statistically significant sequence patterns and prune the redundant ones; it then aligns and clusters homologous patterns into Aligned Pattern Clusters (APCs) and, if desired, find clusters of co-occurring APCs (cAPCs) to reveal co-binding relation and subgroup characteristics (elaborated in Section 2.1,2.2 and Figs 2, 3 and 4). In step 2, for an APC, it uses information measures to reveal sub-cluster/ class characteristics (Section 2.2 and Figure 3). In steps 3 and 4 (Section 2.3, Figures: 4,5 and 6) from the explicit findings, it helps domain experts to relate/conjecture models/mechanisms for explanation, validation and action from established knowledge or via wet-lab experiments. **(b)** Provides simple examples of the steps to show how data can be turned into patterns and then knowledge (models/mechanisms).

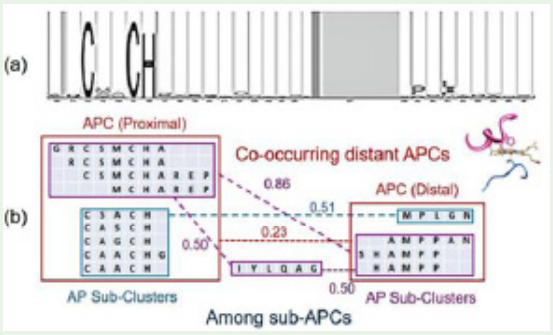

Figure 2: Identification of binding sites through clustering Co-occurrence Aligned Pattern Clusters (cAPCs): Since the number of significant patterns discovered could be large, both APCs and cAPCs help to bring similar patterns and co-occurrence patterns into smaller groups of homologous and co-occurrence patterns. The co-occurrence (the degree of co-occurring patterns) is measured by juccard coefficient defined as the ratio of the number of patterns co-occurring (or co-supportive) on the same sequence over the total number of sequences induced by both APCs. Clusters of cAPCs suggest joint functionality of the domains spanned by the APCs. Here **(a)** shows the HMM of a cytochrome c segment; and **(b)** shows that two APCs (proximal and distal) are sharing 23% of co-occurring patterns. It also shows that the existence of two different co-occurring clusters of sub-APCs in the proximal and distal domains of cytochrome c indicate that a certain degree of difference of interacting functionality may govern different sub-classes of proteins.

dual space (what and where) approach and its applications. From the "what" and "where" discovered, we explore, analyze, synthesize, and organize the APCs to uncover complex relations, which we consider as complex knowledge. Knowing the "what" and "where" will help exploring the "how" and the "why", assisted by the ever expanding public data banks and knowledge base in the intranets and the cloud.

The Aligned Pattern Clustering Algorithm

Aligned Pattern Clustering [8-11] is a novel computationally efficient method for discovering, pruning, representing and ranking homologous sequence patterns with variations in the form of APC [8-11]. Figure 1 gives an overview. Its input is a set of biosequence data. Then, based on linear time and space suffix tree and suffix links, a Pattern Discovery (PD) algorithm [9] is adopted to discover, prune (removing redundancy) and locate sequence patterns from biosequence data. Next, an effective algorithm AP Clustering is used to align similar patterns and cluster them into an Aligned Pattern Cluster (APC) [8-11] as output to represent a group of homologous sequence patterns with variable length and pattern variation. APC has three unique properties. First, it adopts alphabet representation based on strong statistically significant sequence association which naturally preserves column-wise associations. Second, it allows

Table 1: The significance of Pattern and Data Space (the "what" and "where" of the essentials).

Role	Pattern Space	Data Space
What	Patterns (statistically significant sequential base/residue association) with statistical significance residuals and ranking. Aligned Pattern Clusters (APCs) and Clusters of APCs (cAPCs) revealing conserved local and interacting functional domains respectively.	Data sequence segments or arrays of segments spanned by all patterns discovered; instances of all data within a sequence data block spanned by each of the aligned patterns in APCs and/or cAPCs.
Where	Patterns/APCs with statistical ranking in a functional conserved domains (local/distant) which could be located in a single and/or multiple interacting biosequences and/or sequence domains/regions.	Data covering all the patterns with sequence ID and location within and between sequences in the sequence data of all functionally related and/or interacting patterns and/or regions.
How	Interpreting and assessing association patterns and sites. Looking for biologically relevant functioning/interacting macromolecules and sites and obtaining additional supporting and explanatory evidence.	Revealing and interpreting functional characteristics of conserved regions (local/distant) for useful actions for different classes, groups, samples or individual biosequences. .
Why	Seeking explanation/confirmation via Patterns /APCs/cAPCs relating to homologous functionality from known counterparts in established knowledge bases (in the cloud). Based on collected and integrated evidences, conjecture functional models/ mechanisms for further exploratory validation or helping in design of web-lab experiments for the final verification.	Data of discovered patterns/APCs/cAPCs within and between sequences provides a statistical and functional base to validate the underlying models/mechanisms. Locating functional sequences and regions could narrow down the data and scope for further search/validation as well as design of specific experiments for wet-lab verification.

Table 2: Qualitative and Quantitative Experimental Comparison.

Description	Qualitative Comparison	Quantitative Comparison
Pattern Discovery [9]		
In real world, a large set of patterns could be discovered yet many of them are redundant, thus degrading the output quality. We improve the output quality by removing two types of redundant patterns: Delta-closed and Statistically Non-induced Patterns.	Comparing to existing algorithms, the benefits of our algorithm are: (1) The use of a generalized suffix tree to discover and locate patterns in linear time. (2) Both redundant delta closed item-sets and statistically induced patterns are pruned to render a smaller set of quality patterns.	- Faster run-time (up to 7X) comparing to CISP mining, Gap BIDE, and DDCP - An average percentage (70%) of reduction in terms of the number of homologous pattern
Aligned Pattern Clustering (APC) [8-11]		
An APC represents a set of sequence patterns that have been grouped due to their aligned similarities. Aligned patterns plus flexible mutations, capture the vertical similarity of amino acids between the patterns. The effectiveness and efficiency of the algorithm rely upon pattern-based clustering.	Comparing with its counterparts, our algorithms are: (1) faster since patterns are aligned and clustered, not driven by site similarity and alignment;(2) flexible in pattern length, mutation and data coverage; (3) more compact as they group multiple patterns into one group (4) showing statistical significance, ranking and AA distributions in APCs; (5) finding motifs missed by others.	- Faster run-time (up to 616x) comparing with MEME in identifying protein binding site. - More precise (up to 50%) comparing with MEME in protein site identification. - More compact homologous pattern reduction (upto82.1 %) compared to rigid pattern discovery
Protein-DNA Co-Occurring APC Discovery [18,19]		
Protein-DNA Co-Occurring APC Cores allowing minor mutations was developed to represent TF-TFBS binding cores shown in the format such as: TF: {FCNRRQK,FQNRRMK,FQNRRAK} with{TTATTTG, TTAATTG}as TFBS.	(1) Protein-DNA Co-Occurring APC allows mutations on both TF and TFBS binding segments while traditional methods do not.(2) Since APCs are obtained in pattern space rather than data space, the runtime required to obtain cAPCs is much faster than traditional methods requiring exhaustive search for potential DNA and proteins binding pairs.	- Our results have higher consistency (~20%) to those obtained 3D structures by comparing t - the latest published binding core discovery algorithm - Our approach has a speed-up of over 1600X comparing with the latest published binding discovery algorithm
Predicting Protein-protein interaction (PPI) [17]		
Predicting Protein-protein interaction (PPI) is important for discovering molecular interaction mechanisms. WeMine-P2P is developed to predict PPI based on only sequences leveraging APCs to construct feature vectors to represent interaction of protein sequence.	Unlike SVM-based black-box methods, WeMine-P2P renders interpretable biological features from which more discriminative co-occurring sequence patterns can be observed from the compositional bias regions.	WeMine-P2P (1) outperformsPIPE2 [23,24] which also uses co-occurring AA sequence segments but does not allow variation of pattern content / length;(2) achieves PPI prediction comparable to the SVM-based methods with a potential 1280x reduction of feature dimension.

patterns in a single representation with variable lengths. Third, it tracks sequence locations of all patterns in the APC. Their roles in pattern and data spaces, as well as qualitative and quantitative importance are exemplified in Tables1 and 2 respectively.

Application 1: Identifying Functional Elements in Protein Sequences

Aligned Pattern Clustering helps to identify and locate compact functioning elements in biosequence domains and capture the amino acid associations, conservations and variations there in. It analyzes, synthesizes, and reveals functional information of protein families. To further allow more coverage of sequences, we developed an algorithm [17] to extend the APCs containing only highly statistically significant patterns to their variants with minor mutations. Figure 2 shows how APCs and cAPCs from a cytochrome C protein family are discovered and brought into cAPCs based on juccard index [15-16]. It shows two APCs (proximal and distal) sharing 23% of co-occurring patterns. Furthermore, it shows two cAPC subgroups, indicating that protein sub-class might interact differently. Its capability in uncovering the essential binding sites in protein families of cytochrome c, ubiquitin, and trios phosphate isomerase is demonstrated in [8-11].

Application 2: Revealing Functioning Subgroup Characteristics

The biological function of protein families and their class characteristics can be discovered in APCs as demonstrated by a protein known as class-A scavenger receptor (Figure 3).Once the highly correlated information is brought into an APC, two types of information measures (Figure 4) can be used to reveal the group/subgroup characteristics: **(i)** data measures computed from input sequences; and **(ii)** class measures computed using a priori class groupings to reveal class (subgroup) functional characteristics. Using known and putative sequences of two proteins belonging to a relatively uncharacterized protein family, we can group evolutionarily related sequences and identify conserved regions within individual proteins via their family data. An initial synthetic demonstration with in silico a result [12-14] reveals that **(i)** the data measures are unbiased; and **(ii)** our class measures can be used to accurately rank the quality of the evolutionarily relevant groupings [12-14]. Furthermore, combining these measures allows us to interpret the results by inferring regions of biological importance within the binding domains of these proteins. Compared to popular supervised methods, ours has a superior

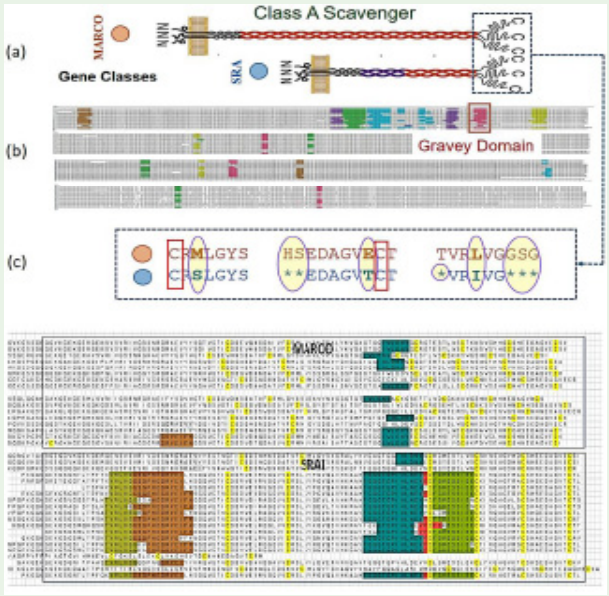

Figure 3: Biological Interpretation of APCs and cAPCs in Class A Scavenger Proteins. **(a)** The MARCO and SRA of Class A Scavenger proteins varying in length (500+) and the number of domains were collected. **(b)** From our unsupervised methods, cAPCs were discovered in certain regions of the protein as shown. One could see that islands of significant functional domains are revealed in the form of cAPCs in the data space. On the top row of cAPCs (Figure 3(b)), we observe that the highest ranking APCs correspond to the Gravey Domain. **(c)** Three top APCs discovered in the SRCR binding domains in the right flanks (dotted box) show the existence of disulfide bonds C (the two shown here are enclosed in red rectangle). We also observe that correlated patterns of amino acids from specific sites reveal distinct class separation and group characteristics as shown by their high SR2 values (not given in the diagram) where the correlated sites depicting gene class separation (not taxonomic classification). (d) APCs in domain 4 of MARCO and Domain 5 specific to SRAI were clustered together in a cAPC. It embodies the GRAVEY domain (right) as well as the "remaining 2 cysteines".

	424	425	426	427	428	429	430	431	432	433	434	435	436		MARCO	SRA	H
1	*	*	*	*	*	*	C	R	M	L	G	Y	S		17	0	0
2	*	*	*	*	*	*	C	R	S	L	G	*	*		0	15	0
3	*	*	*	*	V	F	C	R	M	L	G	Y			18	0	0
4	S	D	A	T	V	F	C	R	M	L	G	Y	S		10	0	0
5	*	D	A	T	V	F	C	R	M	L	G	Y	S		15	0	0
														TOTAL	19	15	
AA	*	*	*	*	*	*	C	R	M	L	G	Y	S				
IG	0.95	0.74	0.74	0.74	0.34	0.34	0.99	0.99	0	0.99	0.99	1	0				
AA	S	D	A	T	V	F			S			*	*				
IG	0	0	0	0	0	0			0			0	0.52				
SR2	0.22	0.23	0.27	0.26	0	0.36	0	0	0.36	0	0	0.05	0.41				
IG	0.32	0.58	0.58	0.58	0.84	0.84	0	0	1	0	0	0.05	0.74				

Figure 4: Subgroup/class characteristics and patterns revealed by class and data information measures. In this figure, an APC is obtained from Class A Scavenger Protein. The APC is represented in table form. Each numbered row associates with a statistically significant pattern and each numbered column is an aligned column of amino acids, which are either conserved or mutated. The instance counts of the patterns for MARCO and SRAI are listed on the right hand columns labeled with MARCO in violet and SRAI in green. The Class Entropy of each pattern and the APC is displayed in the last column with the heading H and each amino acid is displayed below the patterns; in addition, the SR2 and IG is summarized at the bottom. In this case, since each pattern distinctly pertains to only one class, the pattern Class Entropy is 0 in view of total certainty. Tabulated at the bottom rows are the Class Information Gain (IG) of each individual amino acid on a column (site) and also the IG of each column is tabulated in the last row labeled IG. More specifically, the aligned column 432 [MS] is a strong class partition due to its perfect IG of 1 where M corresponds to MARCO and S to SRAI. The sum of normalized mutual information (SR2) which measures for each column the interdependence of the column site with all the other sites is also given at the bottom of the table. Since SR2 does not dependent on class labels, it is unbiased by the labels but reveals the intrinsic subgroups correlation characteristics. In majority cases, it is an excellent classifier for column information in APCs since in APC the functional correlated information is brought in through its inherent association patterns. Next, we evaluate the top SR2 class measure for each APC to help interpret the data partition. The aligned column 432 with high SR2 is attributed by a [S] amino acid mutation.

runtime (16X faster than SVM and 37X faster than HMM [14]) with comparable accuracy (a higher minimum of 10% better than those of SVM and HMM [14]) while not relying upon or biased by inadequate ground truths --- a challenge when the data gets big and diverse.

Application 3: Identifying co-occurring intra-protein, protein-DNA and protein-protein functional elements

Co-occurring intra-protein functional elements: Patterns from two conserved regions with patterns co-occurring frequently on the same sequences suggest joint functionality. Using WeMine™ [8-11], we discovered cAPCs in protein families associating to biological significant regions validated by three-dimensional contact closeness and biological functionality found from established work [15-16]. Figure 3 shows how APCs and cAPCs discovered can reveal the local and distant functional and interactive relations in Class A Scavenger proteins and in their gene class separation. Our methods of discovering significant regions between proteins, antibodies and ncDNA will play an important role in drug and treatment discovery for preclinical studies and personalized treatments.

Co-occurring protein-dna functional elements: The regions between a protein and a DNA in close contact (<3.5A° [20]) are referred to as binding cores [21]. Understanding binding cores is important for deciphering Protein-DNA (TF-TFBS) binding and gene regulation. In Figure 5b, our algorithm takes protein and DNA sequences from TRANSFAC (a Protein-DNA Binding Database documenting which protein (~500 residues) sequence interacts with which DNA sequences (~25 bp) but provides no information on what and where of the binding cores) [22] as input; discovers from both sequence sets conserved regions in the form of APCs; associates them as Protein-DNA cAPCs; ranks them according to their co-occurrence, and among the top ones, finds 3-dimensional structures to support each binding core candidate. If successful, candidates are verified as binding cores. Otherwise, homology modeling is applied to their close matches in Protein Data Bank (PDB) [23] to attain new chemically feasible binding cores. Our algorithm obtains binding cores with higher precision in much faster runtime (>=1600x) than those of its contemporaries [18-19].

Co-occurring protein-protein functional elements: Predicting Protein-Protein Interaction (PPI) is important for making new discoveries in molecular mechanisms. WeMine-P2P (Figure 6) is applied to predicting PPI based only on sequence data via the co-occurrence of APCs [17]. Through 40 independent experiments, we showed that **(1)** WeMine-P2P outperforms the well-known algorithm, PIPE2 [24-25], which also utilizes co-occurring amino acid sequence segments but does not allow variation of patterns and lengths; **(2)** it achieves satisfactory PPI prediction performance,

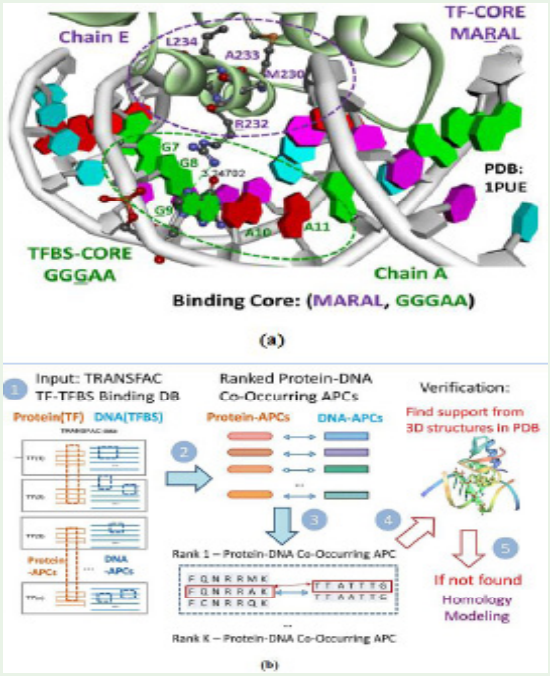

Figure 5: The Overview of the Protein-DNA Binding Core and its Discovery Process [18,19]. **(a)** A protein-DNA (TF-TFBS) binding core with TF-Core: MARAL and TFBS-Core: GGGAA. They both contain less than 10 residues and nucleotides respectively. To look for binding cores in TRANSFAC **[22]** containing proteins with 500 residues (on average) and DNA with 25 bp (on average) is indeed challenging. **(b)** A Protein-DNA Binding Core Discovery algorithm [18,19] is developed to overcome this hurdle. It is a process with 5 major steps as exemplified by the following five items as depicted in circled indexed steps in the figure 1) The input is TRANSFAC [22], a database of Protein-DNA (TF-TFBS) binding sequences; 2) An Aligned Pattern Clustering algorithm [18] is applied to discover Protein-DNA cAPCs and rank them according to their co-occurrence. 3) For the top-ranking Protein-DNA cAPCs, binding core candidates are enumerated. 4) Each candidate is then checked if support can be found in PDB. If found, the candidate is ascertained as a binding core. 5) If not found, homology modeling is conducted to an existing 3D structure closely matching to the candidate to check if the binding mechanism is chemically feasible.

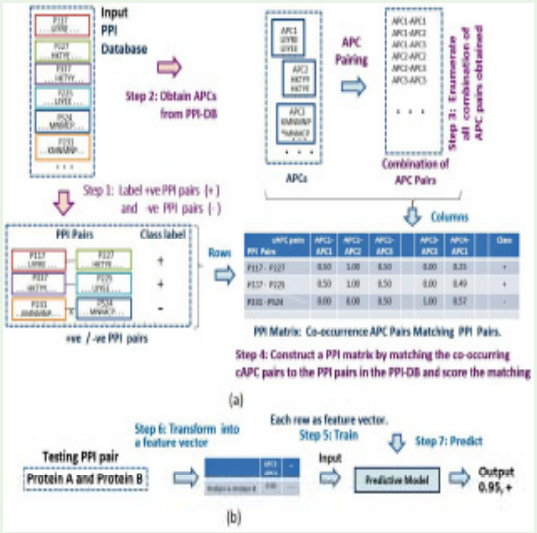

Figure 6: WeMine-P2P: a PPI Predictor **[17].** The input dataset, denoted as PPI Database (PPI-DB), consists of a set of protein sequences, as well as positive (binding) and negative (non-binding) PPI pairs. Each protein sequence has a unique ID, e.g. P117, P227...etc. For illustration, only some segments on a protein sequence are shown. To train a predictive model, positive and negative PPI pairs are labeled by "+" and "-" labels respectively (Step 1). For extracting features, APCs are obtained from PPI-DB using WeMine Aligned Pattern Clustering algorithm (Step 2). All possible pair wise combination of APCs is then enumerated as cAPC pairs (Step 3). To construct a PPI matrix, cAPC pairs are then matched to the PPI pairs taken from the PPI-DB and the matchings are scored by the APC-PPI Score (Step 4). A predictive model is trained on the PPI matrix, where each of its rows is a feature vector with a class label (+) or (-) as its last element (Step 5). Any protein pair can be turned into a feature vector by computing and concatenating the APC-PPI Score of all cAPC pairs to it. To train the predictor (Step 5), the feature vectors from the PPI Matrix with APC-PPI Score are used. To predict whether a protein pair will interact (Step 6), we input it into the predictor after converting it to a feature vector to obtain the PPI classification results.

comparable to the SVM-based methods particularly among unseen protein sequences with a potential reduction of feature dimension of 1280x; **(3)** unlike SVM-based methods, it renders interpretable biological features revealing co-occurring sequence patterns from the compositional bias regions are more discriminative.

Conclusion

Throughout our research on discovering complex knowledge, we have demonstrated that our sequence-based methods on discovering and locating functional elements are preferred over structure-based methods and superior to its counterparts. We have shown that our methods can discover and locate functional elements in protein and DNA sequences and reveal biological joined functionality through cAPCs. Hence, it can be used for identifying co-occurring intra-protein, inter-protein and protein-DNA functional elements. With the rapid advent of sequencing technologies, our sequence-based methods are surely important in the era of big data.

References

1. YH Cai and H Huang. Advances in the study of protein-DNA interaction. Amino Acids. 2012; 43: 1141-1146.
2. EM Phizicky and S Fields. Protein-protein interactions: methods for detection and analysis. Microbiol Rev. 1995; 59: 94-123.
3. M Baker. Proteomics: The interaction map. Nature. 2012; 484: 271-275.
4. P Chène. Inhibition of the p53-MDM2 interaction: targeting a protein-protein interface. Mol Cancer Res. 2004; 2: 20-28.
5. C Winter, A Henschel, A Tuukkanen and M Schroeder. Protein interactions in 3D: From interface evolution to drug discovery. Journal of Structural Biology. 2012; 179: 347-358.
6. RJ Edwards and N Palopoli. Computational prediction of short linear motifs from protein sequences. Computational Peptidology. 2015; 89-141.
7. X Xia. Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction. Scientifica (Cairo). 2012; 2012.
8. ESA Lee and AKC Wong. Identifying protein binding functionality of protein family sequences by Aligned Pattern clusters. Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference. 2012; 1-6.
9. AKC Wong, D Zhuang, GCL Li and ESA Lee. Discovery of delta closed patterns and non induced patterns from sequences. IEEE Trans Knowl Data Eng. 2012; 24: 1408-1421.
10. ES Lee and AK Wong. Ranking and compacting binding segments of protein families using aligned pattern clusters. Proteome Sci. 2013; 11: S8.
11. AKC Wong and ESA Lee. Aligning and clustering patterns to reveal the protein functionality of sequences. IEEE/ACM Trans Comput Biol Bioinforma. 2014; 11: 548-560.
12. EA Lee, DME Bowdish, FJ Whelan, AKC Wong, DME Bowdish and AKC Wong. Characterizing Amino Acid Variations of Scavenger Receptors by Class Information Gain. Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. 2013; 818.
13. EA Lee and AKC Wong. Classifying proteins by amino acid variations of sequential patterns. Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. 2013; 276.
14. ESA Lee, FJ Whelan, DME Bowdish and AKC Wong. Partitioning and correlating subgroup characteristics from Aligned Pattern Clusters. Bioinformatics. 2016; 211.
15. EA Lee, S Fung, HY Sze-To and AKC Wong. Confirming biological significance of co-occurrence clusters of aligned pattern clusters. Bioinformatics and Biomedicine (BIBM) 2013 IEEE International Conference. 2013; 422-427.
16. ESA Lee, S Fung, HY Sze-To and AKC Wong. Discovering co-occurring patterns and their biological significance in protein families. BMC Bioinformatics. 2014; 15: S2.
17. A Sze-To, S Fung, ESA Lee, AKC Wong, HY Sze-To, S Fung, et al. Predicting Protein-Protein Interaction Using Aligned Pattern Clusters. IEEE Bioinforma Biomed. 2015; 55-60.
18. ESA Lee, KS Leung, HY Sze-To, TCK Lau, MH Wong and AKC Wong. Discovering protein-dna binding cores by aligned pattern clustering. IEEE/ACM Trans Comput Biol Bioinforma. 2015.
19. EA Lee, HY Sze-To, MHH Wong, KSS Leung, TC Lau, AKC Wong, et al. Discovering Protein-DNA Binding Cores by Aligned Pattern Clustering. IEEE Bioinforma Biomed. 2014; 125-130.
20. S Ahmad, MM Gromiha and A Sarai. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics. 2004; 20: 477-486.
21. KS Leung, KC Wong, TM Chan, MH Wong, KH Lee, CK Lau, et al. Discovering protein-DNA binding sequence patterns using association rule mining. Nucleic Acids Res. 2010; 38: 6324-6337.
22. V Matys, E Fricke, R Geffers, E Gößling, M Haubrock, R Hehl, et al. TRANSFAC®: Transcriptional regulation, from patterns to profiles. Nucleic Acids Research. 2003; 31: 374-378.
23. HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28: 235-242.
24. S Pitre, C North, M Alamgir, Jessulat M, Chan A, Luo X, et al. Global investigation of protein–protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. Nucleic acids. 2008.
25. S Pitre, M Hooshyar, A Schoenrock, B Samanfar, M Jessulat, JR Green, et al. Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps. Sci Rep. 2012; 2: 239.
26. SF Altschul, TL Madden, AA Schäffer, J Zhang, Z Zhang, W Miller, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. 1997; 25: 3389-3402.