8

WeMine Aligned Pattern Clustering System for Biosequence Pattern Analysis

En-Shiun Annie Lee[1] • Peiyuan Zhou[2] • Andrew K. C. Wong[2]

[1]Department of Computer Science, University of Toronto, Toronto, ON, Canada;
[2]Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada

Author for correspondence: En-Shiun Annie Lee, Department of Computer Science, University of Toronto, Toronto, ON, Canada. Email: annie.lee@cs.toronto.edu

Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch8

Abstract: A major challenge in bioinformatics is discovering functional regions in biosequences. These regions may correspond to folded structures, physicochemical functionality, or mutation hotspots. The identification of functional regions in biosequences is essential to better understand biological mechanisms, design new drugs, and uncover novel knowledge concerning sporadic and genetic diseases. Pattern analysis and WeMine aligned pattern clustering (APC) systems enable the discovery of conserved regions with adaptive width and mutations, including frameshift, without relying on prior knowledge or exhaustive search. They align and rank patterns in local and distant correlated regions with statistical support within, and between, sequences. This chapter provides an overview of the WeMine APC and its utility in identifying functional regions such as protein binding sites, predicting pairwise interactions between protein-DNA and protein-protein network, and finding correlations among patterns and residues with class labels. Pattern analysis and WeMine APC could play an important role in personalized medicine, gene therapy, biomarker identification and drug discovery.

Keywords: aligned pattern clustering; biosequence pattern analysis; class association; protein-protein interaction; WeMine system

In: *Bioinformatics*. Nakaya HI (Editor). Exon Publications, Brisbane, Australia.
ISBN: 978-0-6450017-1-6; Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021

INTRODUCTION

One of the major challenges in proteomics is to understand the functional regions in protein sequences. Such knowledge, if translated into explicit interpretable and succinct forms, could reveal the biological functionality crucial for a wide variety of applications. It not only will enable the understanding of the biological mechanisms, but also will support drug and vaccine discovery for curing genetic and epidemic diseases. While an explosive amount of proteomic data is streaming in ubiquitously from research laboratories, the challenge remains to create effective scalable data-driven algorithmic methods for deciphering the data.

Protein families may consist of many members, and the similarity and dissimilarity of the functional regions between their sequences becomes less clear with greater evolutionary distance. Evolutionarily conserved amino acids within proteins characterize functional or structural regions. Conversely, less conserved amino acids within these regions are generally areas of evolutionary divergence. Existing unsupervised sequence analysis methods such as multiple expectation maximizations for motif elicitation (MEME) (1) and gapped local alignment of motifs (GLAM2) (2) can neither handle frameshift mutations effectively, nor find rare mutations occurring in a single or a few sequences. These short comings are: (i) when mutations such as substitution, insertion, deletion, and frameshifts occur in these functional regions, they are difficult to be handled with fixed motifs without proper alignment; (ii) though patterns are conserved within a family of sequences, their locations vary, and homologous sites need to be aligned; (iii) certain new/rare mutations are often detected only in a few sequences in an available ensemble, hence, they are not easily seen/noticed by MEME; and (iv) if the sample sequences pertain to a close family, their similarity is high with only a few break points due to strong homology over a large domain, hence, it is difficult to segment them to reveal the local conserved functional domains.

Due to the reason (iv) stated above, we need to identify spots to delimit local regions to highlight their local conserved functional homology. Hence, we need to get additional sequences from a broader family (via blasting) with variations of less dominating functionality or different related sequence families with common significant/dominating functionality. These are hard to find from BLAST (3, 4) or other existing sequence analysis methods. Hence, it is a surmountable challenge to discover local conserved regions effectively from sequence data alone.

Due to these and other shortcomings, we developed the WeMine System for pattern analysis based on intricate pattern-data duality via a novel pattern-data representation called aligned pattern cluster (APC), which utilizes the correspondence of data and patterns to drive the algorithm using effective data structures and statistical measures. The definition and concept of pattern-data space duality and the WeMine System are explained in greater details in the next section. The WeMine System produces a pattern representation that is more accurate due to precision obtained from the patterns with variable length, allowing gaps between sub-patterns and more diverse types of local mutations. It is faster due to compression of pattern supported by statistical data measures, and interpretable due to disentanglement minimizing bias and subtle entangled factors. The WeMine System has been applied to discover functional regions in

protein families (cytochrome *c* and class A scavenger receptors) as well as transcription factor binding sites (TFBS). Furthermore, the algorithm of discovering co-occurring patterns has been applied to discover binding cores between transcription factor proteins and their complementary DNA sequences, intra-protein interacting sites of various protein families, and protein-protein interaction (PPI) prediction.

The pattern analysis method demonstrates faster runtime than its contemporaries, such as CISP (Contiguous Item Sequential Pattern Mining) (5), Gap BIDE (Gap BI-Directionall-Extension –based frequent closed sequence mining) (6, 7), MEME (665x) (1) and GLAM2 (10x) (2) with an average reduction (70%) in number of homologous patterns. It is more accurate in identifying protein binding sites, up to 50% when compared with MEME and GLAM2 (2). In terms of biological application in co-occurrence and classification, our method displays the following features: (i) discovers binding cores (protein-DNA) at a higher consistency (~20%) and a faster computation time speed-up (1600X); (ii) outperforms PPI prediction compared to PIPE2 (Protein Interaction Prediction Engine 2) (8, 9) and achieves 1280x feature dimension reduction compared to Support Vector Machine (SVM) method; and (iii) ranks residue mutations correlated with class in unsupervised manner faster than Hidden Markov Model (HMM) (100x) and SVM (14x). Due to the accessibility of protein sequences on the internet, to achieve the significant task of identifying functional regions on proteins for proteomic research and drug discovery, it is more economical and effective to first look for conserved segments from the data of a set of functionally similar protein sequences than to perform laborious and time-consuming experiments and computationally intensive modeling. Thus, our WeMine system has great implications in drug discovery and protein analysis.

THE WEMINE SYSTEM

At the core of the sequence analysis system, we introduce the concept of pattern-data space duality (10). To put simply, the pattern space consists of patterns that are statistically significant, highlighted from the sea of data, and the data space is the instantiations (occurrences) of these patterns in the data. They are used to compute statistical scores and measures. After pattern discovery, all patterns in the family, from low to higher order, are discovered with their addresses (sequence ID and location) registered in an Address Table (AT). We then acquire, align and grow an APC by extending each pattern therein if found in the AT. If not found, a mutation may occur at the end (referred to as a pattern breakpoint) of the pattern. From its pattern location in the AT, we look for possible mutation (substitution, insertion or deletion) at the breakpoint in the data space. If found, we jump over the breakpoint to see if it is a pattern in the AT. From its frequency of occurrences obtained and the statistical testing, we will include the mutated pattern in the growing APC. If not, we will place it to a rare mutant pool. The functional hypothesis is that each summarized APC has its corresponding data occurrences for instantiating patterns in the functional region, thus allowing verification of the biological function at pattern, data and knowledge level. Each step in the WeMine System makes use of the pattern-data space duality in a certain manner.

It organizes the discovered patterns in the pattern space and makes computations in the data space to reveal the hidden patterns and their statistics within the data. The pattern-data space duality is further elaborated in Table 1.

Figure 1 gives an overview of the WeMine System. It consists of three major modules: pattern discovery, pattern summary (called Aligned Pattern Clustering shortened as APC), and pattern refinement (representation of APCs and co-occurrence APCs). Figure 1A gives a schematic description. It takes a family of biosequences, discovers the statistically significant sequence patterns, aligns and clusters them into APCs, and refines their patterns. Figure 1B gives more details. The fundamental data structure of sequence pattern discovery is the suffix tree with suffix links representing both the patterns (paths) and the data space (the leaves in the suffix tree data structure stores the pattern IDs and addresses). APC aligns and clusters patterns into APCs and refine the aligned patterns including gaps, mutations and their addresses in the data. The extended pattern and data spaces allow in-depth analysis of subgroup characteristics in the functional domains. Figure 1C presents the biological applications of the WeMine System as three types of outcomes: (i) patterns co-occurrence through co-occurring APCs, obtained within proteins (APCs sharing co-occurring patterns on same sequences to reveal intra-protein three-dimensional interaction proximity); (ii) between biosequences (discovering protein-DNA binding cores): and (iii) protein-protein interaction prediction. Table 1 further elaborates each step of the WeMine System as well as the biological applications with two examples of co-occurrence APC and one example of class partitioning APC. WeMine System takes a set of multiple sequences of a protein family as input and discovers patterns, aggregates similar patterns, aligns and refines patterns, and uncovers associations.

Pattern Discovery

The pattern discovery takes advantage of a fast-linear run time and space-efficient algorithm we developed (12). It ingeniously uses a generalized suffix tree to efficiently identify the proper superpatterns and subpatterns and obtains a compressed or pruned statically ranked list of patterns with corresponding data space that are statistically significant and nonredundant. The suffix tree with suffix links (Figure 1B) stores patterns with statistical weights and addresses on the leaves of the suffix tree while removing pattern redundancy by avoiding multiple listing of patterns in the sub-patterns. Its early success was demonstrated in finding TFBS using DNA sequences from the promoter regions of the yeast, *Saccharomyces cerevisiae*. It generated a relatively small set of binding sites and achieved best overall results when compared with other motif discovery methods (12). It can retain patterns associated with conserved functional units in the promoter regions and drastically reduce the pattern set. This biological problem of transcription factor binding is revisited later for finding co-occurring APC pairs between protein-DNA pairs, specifically for finding complementing binding regions between the transcription factor protein and its companion TFBS DNA.

Pattern summarization

Since similar patterns reflect homologous functionality, a pattern summarization process is used to align and cluster homologous patterns in an array referred to as APC to represent a local functional region. The pattern summarization algorithm

TABLE 1 **Terminologies used in this chapter and their definitions**

Pattern Space	Data Space	Problem Definition (Input & Output)	Biological Application	Algorithm Process	Findings & Significance
WeMine System of Pattern Analysis (discovering, summarizing, and refining patterns representations that are compressed, flexible, and robust)					
Pattern Discovery					
Patterns that are statistically significant and sequence-based (residue association)	Address location (sequence ID, position number) of each instance or occurrence of the pattern (called induced data space of the sequence pattern)	Given a set of DNA sequences, find consecutive high-order sequence patterns	Discover sequence patterns as Transcription Factor Binding Sites (TFBS) using only DNA sequences	Build suffix tree, compute statistical measure on the nodes and leaves.	– Faster run-time (up to 7X) vs CISP mining, Gap BIDE, and DDCP. – An average reduction (70%) in number of homologous patterns
Pattern Summary					
Aligned Pattern Clusters (APCs) for conserved local functional/ structural regions.	Address locations spanned by all the patterns summarized within the APC (called induced data space of all the patterns summarized by the APC).	Given a list of patterns, find groups or clusters of patterns that are aligned based on sequence similarity.	Reveal binding sites and binding domain of the Cytochrome *c* protein family.	Dynamic programming with single linkage hierarchical clustering.	– Faster run-time (up to 616x) vs MEME in identifying binding sites. – More precise (up to 50%) vs MEME in identifying protein site. – Pattern reduction (up to 82.1 %) vs rigid pattern discovery.

Table continued on following page

TABLE 1 Terminologies used in this chapter and their definitions (Continued)

Pattern Space	Data Space	Problem Definition (Input & Output)	Biological Application	Algorithm Process	Findings & Significance
Pattern Refinement					
Seed pattern for initialization and extending seed pattern with gaps using breakpoints mutations.	Induce the data space of the extended patterns to find rare mutations that matches the main extended pattern.	Given a list of seed patterns, find the refined aligned patterns with gaps and a pool of rare mutations	Cytochrome C protein family and Synthetic datasets	Pattern extension of gaps and data space with frameshift and rare mutation	– Faster run-time vs MEME (665x) and GLAM2 (10x). – Higher recall and F-measure than MEME; higher F-measure and precision and F-measure than GLAM2
Biological Applications of Pattern Associations					
Pattern Pair					
Co-occurring APC (cAPC) pair with many-to-many associations	Used to compute the co-occurrence between the two APCs to reveal the protein-DNA binding core	Given database with protein-DNA binding pairs, find pattern-pair association (called binding cores)	Finding binding cores from Protein-DNA pairs from the TFBS TRANSFAC database (11)	Compute co-occurrence between protein pattern segments binding to DNA pattern segments	– Higher consistency (~20%) – Results supported by 3D structures – Computation time speed-up (1600X) vs the latest published binding site discovery algorithm

Table continued on following page

TABLE 1 **Terminologies used in this chapter and their definitions (Continued)**

Pattern Space	Data Space	Problem Definition (Input & Output)	Biological Application	Algorithm Process	Findings & Significance
Intra-protein Interaction					
Graph network of co-occurring APC (cAPC) pairs	Used to compute the co-occurrence measure to reveal the interacting functional domains within the protein molecule	Given a database with protein-DNA binding pairs, find a pattern-pair association (called binding cores)	cAPC residues discovered from the sequences of the cytochrome *c* protein family are closer in 3D space and form chemical bond	Cluster cAPCs using spectral clustering with thresholds on the co-occurrence measure to form interaction graph network	– Unsupervised method using sequence data only to identify biologically intriguing pattern associations
Protein Interaction Prediction					
Co-occurring patterns	Used to compute co-occurrence measure for predicting protein-protein interaction	Given a protein-protein interaction database, build a predictive model	Predict whether two protein sequences will interact from the PPI database	Label PPI pairs and use cAPCs to construct interaction matrix to train a predictive model	– Outperforms PIPE2 (8, 9) – Achieves PPI prediction with 1280x feature dimension reduction vs SVM method

Table continued on following page

TABLE 1 **Terminologies used in this chapter and their definitions (Continued)**

Pattern Space	Data Space	Problem Definition (Input & Output)	Biological Application	Algorithm Process	Findings & Significance
Pattern-Class Association					
Association of residue mutation of aligned sites with class labels	Used to compute information measures to reveal residue variance of an aligned site and residue correlation with other sites	Given a set of input sequences with class labels, partition the data space according to pattern-class correlation	Residue mutation within the APCs for protein families (i.e. cytochrome c and scavenger receptor) reveal corresponding to class (i.e. taxonomic groups or protein family sub-grouping) partitions	Simultaneously create APCs with sites mutations and correlated those mutations to class labels using unbiased information measures	– Unsupervised method for ranking residue mutation correlated with class. – Faster runtime vs HMM (100x) and SVM (14x)
Other Associations					
Association of residue mutation of aligned sites within APCs	Used to compute statistical correlation of residues with other sits and/or with the class label	Given input sequences with class labels, obtain APCs with residue mutation associated with class partition	Cytochrome C protein family sequences residue associate patterns/ pattern clusters with taxonomic groups/ classes and physico-chemical characteristics	Build frequency matrix between residue occurrences to compute statistical significance, then use principal component decomposition and vector reprojection. to cluster residue association groups	– Unsupervised discovery and disentanglement of attributes-values/ residue to associate with class partition or biological characteristics

APC, aligned pattern cluster; BIDE, bi-directional extension; cAPC, co-occurring APC; CISP, contiguous item sequential pattern; DDCP, discovery of delta closed patterns; HMM, hidden Markov model; PIPE2, protein information and property explorer 2; PPI, protein to protein interaction; SVM, support vector machine; TFBS, transcription factor binding sites.

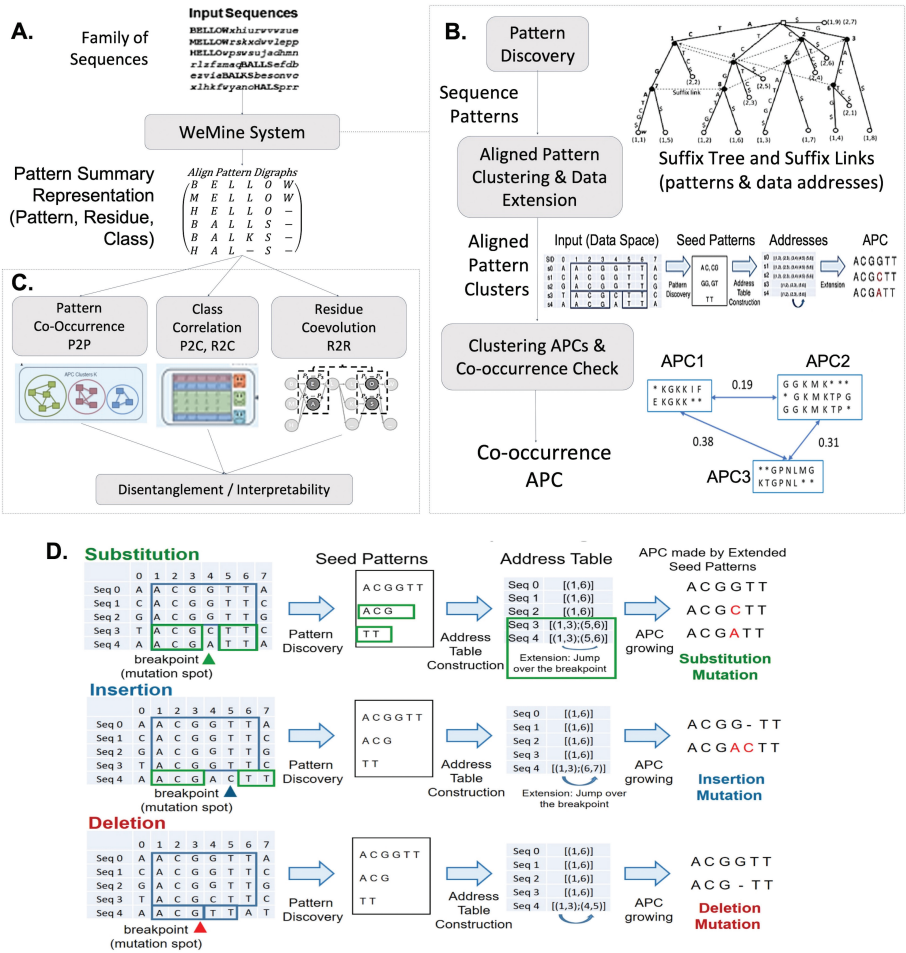

Figure 1. An Overview of the WeMine System. A. Schematic view with input, throughput and application outcomes. **B.** a brief description of pattern discovery, pattern summarization and pattern refinement. **C.** Biological applications of the WeMine system as three types of outcomes. **D.** The growing of APC via finding seed patterns, extending aligned patterns and identify and locate the breakpoints, the types of mutation. APC, aligned pattern cluster; P2C, protein to class; P2P, protein-to-protein; R2C, residue to class; R2R, residue to residue.

aligns and clusters the discovered patterns into APCs (Figure 1B and Figure 2 A-E). It groups the patterns of different lengths obtained into arrays of homologous aligned patterns to maintain the same length by inserting gaps and mutations/wildcards. The amino acids within the patterns are aligned at the same location, reflecting its regional functionality of the pattern and its mutation within the sequence. The key rationale behind summarizing the patterns (by aligning and clustering them) is that once an APC with its relative position is obtained, it will reflect the statistically significant residue association with each other in the

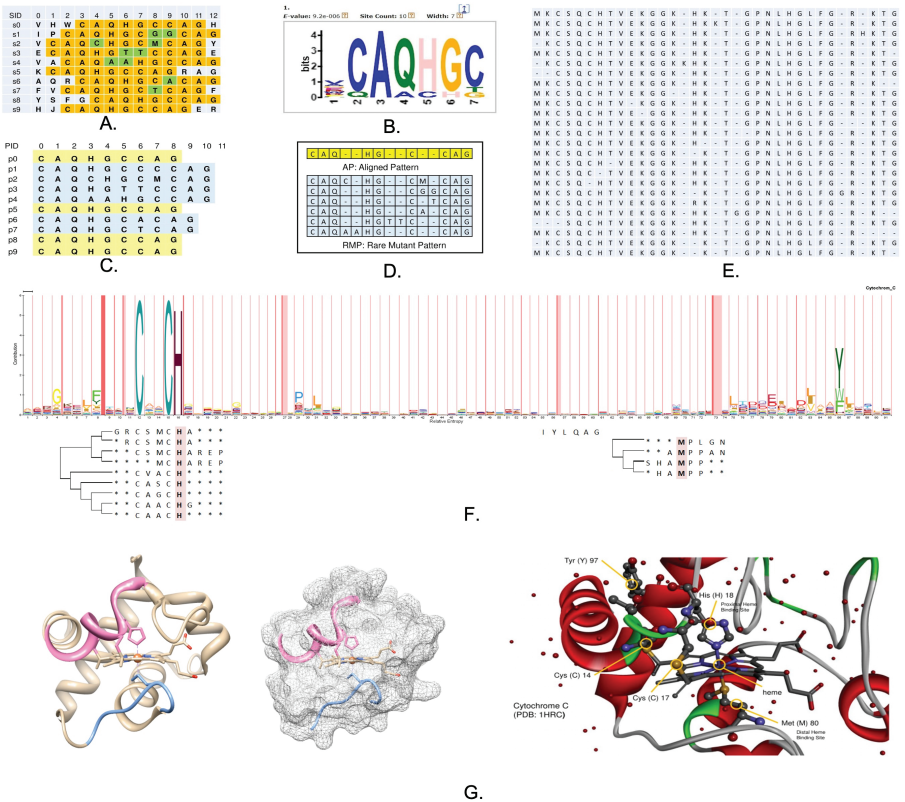

Figure 2. Pattern Summarization and Refinement. Aligned pattern clustering, motif discovery, aligned pattern clusters and the 3D structures they represent. **A.** Patterns with mutation embedded in the sequences of the cytochrome c family. **B.** Probabilistic Weighted Matrix PWM and its Logo. **C.** APC obtained from WeMine. **D.** Pattern space of aligned patterns and rare mutant patterns. **E.** Data space of the pattern directed aligned pattern clustering result. **F.** An aligned pattern cluster obtained from the sequences of a cytochrome c family. **G.** 3D structure and its representation of a binding site discovered by WeMine APCn, showing the binding of the bio- molecule to the complex of the heme and its iron atom at the center.

patterns (with variations) and also the amino acid distribution of each of its aligned sites (columns) to reveal the functionality of the protein family within the regions spanned by the patterns in the APCs with statistical ranking and support. APCs represent functionally homologous regions of protein patterns, specifically binding segments, wherever they are in the input sequences of the protein family. Algorithmically, this step takes a reduced list of patterns (obtained from step 1 of pattern discovery) as input and clusters/groups them using a single-linkage hierarchical clustering algorithm that iteratively aligns them, using dynamic programming, into one or more APCs. The algorithm iteratively merges two APCs in a pairwise-manner based on their similarity scores until one of the termination conditions is reached. The three key parameters of the algorithm are the Merge

Algorithm, the Similarity Score, and the Termination Condition. APCs are ranked according to their statistical significance computed from their corresponding data space.

Proteomic application

We applied WeMine to the cytochrome *c* protein family and obtained the APCs that correspond to the functional binding segments and its binding residues. The cytochrome *c* protein covalently binds the heme (13) attached to two cysteine residues. The heme's iron ion is chemically bonded to two binding residues from the opposite sides of the protein, each of them is surrounded by a sequence pattern with variations, within the discovered APC, referred to as the binding segment. The APCs discovered in each cover significant binding sites. WeMine APC runs faster than other motif finding algorithms while not restricted by parameters such as motif width and number of variations. Figure 2A shows part of the protein sequences containing the pattern "CAQHGCCA" and their mutations; 2b is the visual display of Position Weighted Matrix (PWM) obtained by MEME showing the amino acid probability distribution-independent sites; and 2c shows our Phase 1 Pattern Directed APC (PD-APC) results. Figure 2D gives the full result; 2E displays the APC data space, with an adapted width of 35 amino acids, consisting of 25 aligned segments where the 7 discovered patterns listed in 2c are embedded. Figure 2F gives the Pfam representation of a long cytochrome *c* segment, and the hierarchical structures of two APCs discovered by APC corresponding to the proximal and distal binding segments of cytochrome *c* to the heme. The larger APC contains C17 and H18 that binds, respectively, to the heme (a molecular complex) and the iron atom at the center of the heme (Figure 1G). M62 in the distal APC binds to the iron molecule in the heme. The results of APC on ubiquitin have been described previously (14, 15). Figure 2G displays their 3D binding configurations.

Pattern refinement for revealing pattern gaps and mutations

To overcome certain problems encountered in WeMine, we developed a third algorithm for pattern refinement called pattern directed align pattern clustering (PD-APC) (Figures 1B and 1D, and Figure 2A, C-E). It uses seed patterns discovered, extends gaps by pattern breaking points, and uncovers rare mutation patterns. It contains two steps: (i) the use of suffix trees to discover seed patterns (Figure 1B); and (ii) the growing of patterns (Figures 1D and Figures 2 C-E). Herein we provide a brief description of the algorithm. Given a set of sequences, PD-APC step 1 discovers small seed patterns leveraging the pattern discovery algorithm (12). Using the suffix tree, seed patterns are discovered and located which are then extended to super-patterns by "jumping over" the breakpoint mutations (substitution, insertion, and deletions). From the top ranked seed patterns, it extends them using step 2 (14, 15), and grows these extended patterns from the induced data space to the APC and rare mutants to a separate pool. This step uses the seed patterns (with high frequency of occurrences) discovered (Figure 1D) to initiate an APC. To grow the APC, it extends each pattern therein from its location in the data space using information directly from the

Address Table (AT). If the extended pattern is not found, consider the gap as a breakpoint. It then jumps over the gap to find a pattern. If found, it then determines the type of mutation (Figure 1D) taken place and identifies the mutated pattern candidate. If the candidate passes the pattern hypothesis testing, it will be included in the APC. Otherwise, place it to the rare mutant pool. This process is iterated until termination when no more extended pattern could be found.

This algorithm is based on two important concepts. The first is the use of the breakpoint since some mutated patterns, when fragmented, could not be discovered by the pattern discovery due to the low frequency of occurrences of the entire mutational pattern. Hence, if we have the address location of the low frequency sub-patterns, we consider the mutation spot between them as a breakpoint. By jumping over it, the mutated variants and the rare mutant patterns can be discovered from the data space. The second concept is the seed pattern extension introduced to increase the coverage of the growing APC. We observed that the width of seed patterns is inherent in data, unaffected by the algorithmic process and/or the width parameters. We apply the same procedure of "jumping over" a breakpoint to obtain full coverage. When the seed width is changed, the same full coverage remains, indicating pattern width adaptation without exhaustive search.

Application

The pattern extension method uses a systematic process to determine the representation model width adaptively from data without exhaustive search, and discovers rare mutational patterns with substitution, frameshift, insertion and deletion. We evaluated our method against MEME (1) and GLAM2 (2) via three synthetic datasets, where the conserved region positions are a priori known and considered as the ground-truth. The discovered conserved regions output could then be compared with the ground-truth quantitatively. Dataset 1 is composed of 500 protein sequences containing a mutated protein segment with 30 amino acids but no noise. A larger noise-free Dataset 2 consists of 1000 protein sequences with a mutated segment with 30 amino acids. Dataset 3 consists of 2000 protein sequences where 1000 sequences contain a mutated segment with 30 amino acids where the remaining were noise sequences. The results showed that our method is faster than MEME (665x) and GLAM2 (10x) and has a higher F-measure than MEME and GLAM2 (Table 2).

APPLICATIONS OF PATTERNS WITH VARYING ASSOCIATION RELATIONSHIPS

This section explores the biological application of patterns to various associated relationships, such as binding, three-dimensional closeness, interaction, and class partitioning. The folding of protein sequences renders tertiary structures and physicochemical conditions for site (amino acid residues represented as aligned column) and segment (domain represented as patterns), interaction within (Figure 2F) and between proteins, or between proteins and other biomolecules,

TABLE 2 **A comparison of WeMine with MEME and GLAM2. The WeMine system is faster than MEME (665x) and GLAM2 (10x) and has higher F-measure than MEME and GLAM2**

Performance evaluation of PD-APC on Dataset 1 (500 sequences)			
	Precision	**Recall**	**F-measure**
GLAM2 (nMotifs=1) (2)	0.37840	**1.00000**	0.54904
GLAM2 (nMotifs=2) (2)	0.34745	**1.00000**	0.51572
GLAM2 (nMotifs=3) (2)	0.33325	**1.00000**	0.49991
MEME (nMotifs=1) (1)	0.99839	0.49630	0.66301
MEME (nMotifs=2) (1)	0.99261	0.77936	0.87315
MEME (nMotifs=3) (1)	**0.99269**	0.78816	0.87868
PD-APC (ω_{seed} = 3, gap_{break} = 2)	0.96348	0.89905	0.93015
PD-APC (ω_{seed} = 3, gap_{break} = 3)	0.96335	0.91655	**0.93942**
Performance evaluation of PD-APC on Dataset 2 (1000 sequences)			
	Precision	**Recall**	**F-measure**
GLAM2 (nMotifs=1) (2)	0.46781	**1.00000**	0.63742
GLAM2 (nMotifs=2) (2)	0.41305	**1.00000**	0.58462
GLAM2 (nMotifs=3) (2)	0.35262	**1.00000**	0.52139
MEME (nMotifs=1) (1)	**0.97967**	0.39232	0.56028
MEME (nMotifs=2) (1)	0.97922	0.84919	0.90958
MEME (nMotifs=3) (1)	0.97930	0.85249	0.91151
PD-APC (ω_{seed} = 3, gap_{break} = 2)	0.96541	0.89065	0.92092
PD-APC (ω_{seed} = 3, gap_{break} = 3)	0.96462	0.91266	**0.93792**
Performance evaluation of PD-APC on Dataset 3 (2000 sequences)			
	Precision	**Recall**	**F-measure**
GLAM2 (nMotifs=1) (2)	0.61117	**1.00000**	0.75867
GLAM2 (nMotifs=2) (2)	0.59827	**1.00000**	0.74865
GLAM2 (nMotifs=3) (2)	0.54501	**1.00000**	0.7055
MEME (nMotifs=1) (1)	0.99898	0.48957	0.65711
MEME (nMotifs=2) (1)	0.99261	0.77936	0.87315
MEME (nMotifs=3) (1)	**0.93682**	0.83278	0.88426
PD-APC (ω_{seed} = 3, gap_{break} = 2)	0.92997	0.89605	0.91269
PD-APC (ω_{seed} = 3, gap_{break} = 3)	0.93039	0.91266	**0.92149**
Runtime Comparison of PD-APC on Dataset 1, 2 and 3			
GLAM2 (nMotifs=3) (2)	202.074s	334.273s	228.779s
MEME (nMotifs=1) (1)	368.401s	2315.512s	15721.029s

Table continued on following page

TABLE 2 **A comparison of WeMine with MEME and GLAM2. The WeMine system is faster than MEME (665x) and GLAM2 (10x) and has higher F-measure than MEME and GLAM2 (Continued)**

MEME (nMotifs=2) (1)	471.633s	2749.722s	17437.620s
MEME (nMotifs=3) (1)	570.683s	3155.81s	18786.427s
PD-APC (ω_{seed} = 3, gap_{break} = 2)	**4.759s**	**12.531s**	**28.104s**
PD-APC (ω_{seed} = 4, gap_{break} = 2)	5.143s	13.466s	30.309s
PD-APC (ω_{seed} = 5, gap_{break} = 2)	5.213s	13.997s	33.232s
PD-APC (ω_{seed} = 3, gap_{break} = 3)	4.843s	12.999s	28.232s
PD-APC (ω_{seed} = 4, gap_{break} = 3)	5.193s	13.653s	30.454s
PD-APC (ω_{seed} = 5, gap_{break} = 3)	5.726s	14.070s	33.696s

such as DNA and RNA (Figure 3 Part I). Those interacting segments/sites correspond to sequence patterns co-occurring in their primary structures or within their interacting environment. As APCs reflect functional regions, distant APCs with patterns co-occurring on the same sequence or within an interacting environment between biosequences, can be identified through the co-occurring patterns discovered among distant APCs. We refer to such configurations as co-occurrence APCs (cAPCs). These enable to explore interacting sites within or between interacting biosequences.

Co-occurring pattern association (protein-DNA interaction) for TFBS protein-DNA interaction

For this application, we return to the biological problem of TFBS first introduced in the pattern discovery step, but instead of only studying the TFBS in DNA sequences we consider the protein-DNA sequences for the binding interaction. For protein-DNA binding, the regions between a transcription factor (TF) and a TFBS in close contact, <3.5Å (16, 17), are referred to as protein-DNA binding cores (18, 19) (Figure 3 Part I). Sequence- specific binding is the ability of a TF protein to distinguish different DNA sequences where the TF protein's binding domain can recognize a collection of similar TFBS DNAs.

Figure 3 shows how APC (19) is able to obtain cAPCs related to binding/interaction sites/regions within, and between bio-sequences. Part I of Figure 3 depicts a 3D configuration of a protein-DNA binding core between a TF and a TFBS DNA. It describes the algorithmic process we developed to identify and locate the binding core (19). In step 1, we obtained a public file, Transfac, listing TF proteins that bind to strands of DNA but not the binding sites which are difficult to find since an ordinary TF protein may consist of over 150 residues and the DNA TFBS may have 5 to 12 bases. The APCs in TF proteins and the DNA strands were discovered and ranked, generating a set of cAPCs consisting of matching pairs of protein-APC

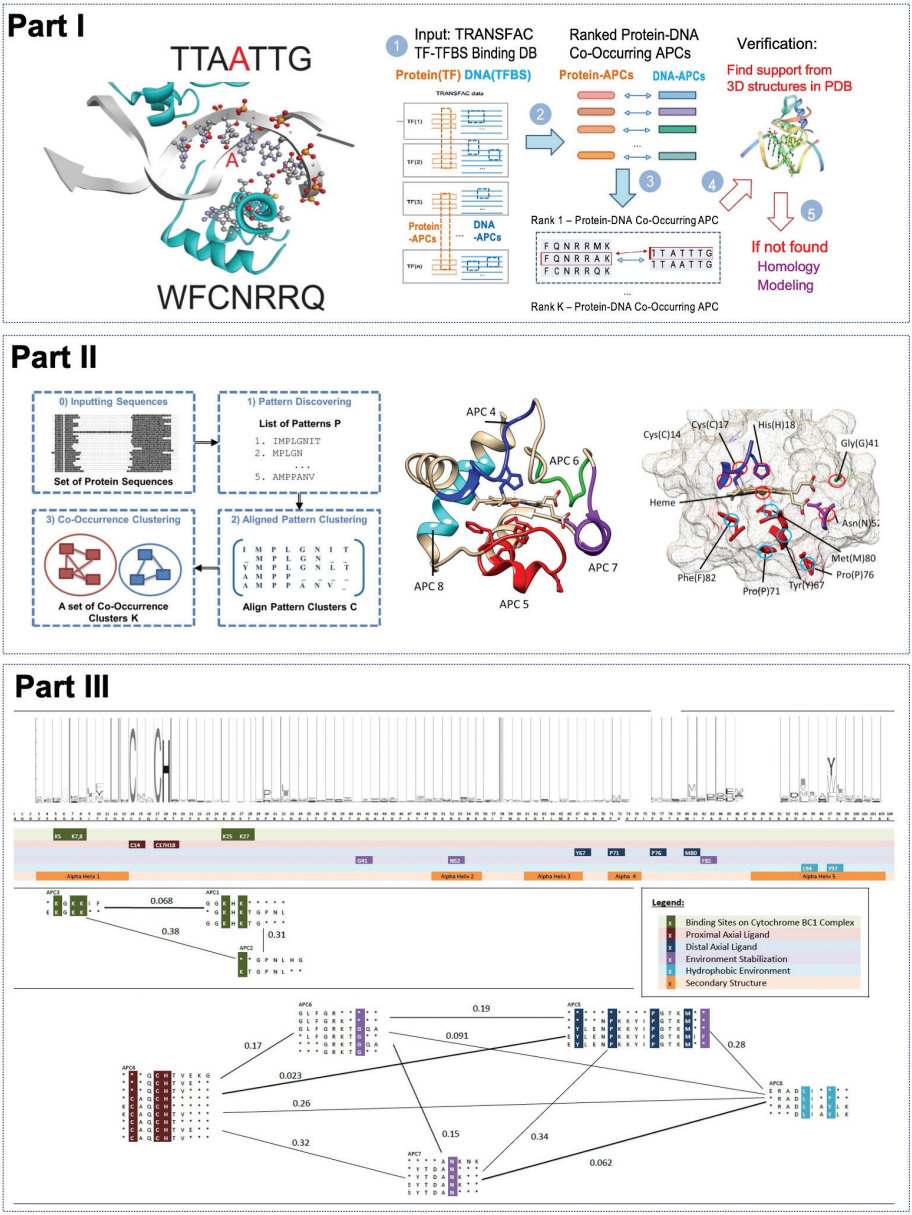

Figure 3. Discovering protein-DNA binding cores and protein interaction regions via cAPCs. **Part I.** A 3D configuration of a protein-DNA binding core between a TF and a TFBS DNA. **Part II.** The three-step process of how cAPC network is constructed from a set of protein family sequences and the protein's 3D configuration with the resulting cAPC highlighted in color. **Part III.** The Pfam representation of a cytochrome c segments with two cAPC. TF, transcription factors; TFBS, Transcription factors binding sites.

and DNA-APC. For each cAPC we find a set of patterns in the TF proteins that co-occur with patterns in the DNA in the Transfac file. Of these co-occurring APCs, some might have already been reported in Protein Data Bank (PDB). If there is, the binding core will be confirmed. If they are not found, we could use a technique known as homology modelling to transform the closet known pairs into the candidate pairs. If the physicochemical homologous transformation succeeded, we would include that pair as a new binding core.

Part II of Figure 3 shows how a cAPC network is constructed from a set of protein family sequences and their corresponding 3D configuration. The cAPC co-occurrence scores are used as similarity measure to obtain the graphical theoretic clusters of APCs, the model (network) that makes up the cAPC. Part III of Figure 3 shows the Pfam representation of a cytochrome *c* segments with two cAPC, one consisting of three APCs and the other made up of five APCs. The edge weights of these graphical representations are the co-occurrence scores between a pair of APCs

Predicting the likelihood of protein-protein interaction

Protein-protein interaction prediction refers to predicting if one protein will interact with another. It enhances our understanding of the molecular mechanisms inside the cell (20) and is particularly useful for discovering unknown functions of a protein (21), particularly for prediction based only on sequence data. Our protein-protein interaction prediction method, the WeMine-Protein2Protein (P2P), is based on a biologically interpretable features in conserved functional regions, and a biologically realistic algorithm in finding binding segments with variable width and mutations. WeMine-P2P is not only able to yield superior or comparable predictive results but can also discover knowledge for PPIs through analyzing the interpretable discriminative features with significant feature dimension reduction. The knowledge discovered in the interpretable feature space is useful for building better predictive models. Through the results of 40 independent experiments, it has been shown that: (i) WeMine-P2P outperforms the well-known algorithm, PIPE2, which also utilizes co-occurring amino acid sequence segments but does not allow variable lengths and pattern variations; (ii) WeMine-P2P achieves satisfactory PPI prediction performance, comparable to the SVM-based methods particularly among unseen protein sequences with a potential reduction of feature dimension of 1280x; and (iii) unlike SVM-based methods, WeMine-P2P renders interpretable biological features from which co-occurring sequence patterns from the compositional bias regions are more discriminative. Since no prior information on PPI is incorporated, WeMine-P2P is extendable to other biosequence applications in the future.

Class association

To explore how the summarized coded patterns, reflect biological functionality, we used unsupervised algorithm to partition the sequence patterns and their residues in APCs and cAPCs without prior knowledge to observe what the partitions reveal. From a set of input sequences, class labels were removed, and unsupervised algorithm was used to examine sub-sequence segments with strong

statistical association with distinct subgroups. The quality of these discovered patterns was ranked and grouped based on their inherent characteristics using our information measures which account for amino acid conservation of a site and their correlation with those on other sites and/or class within the APCs (22). We tried to find out whether our algorithm could identify the discovered patterns and the partitioned mutated subgroups that match the pre-existing groups as reflected by the class labels. Experiments were conducted on known and putative sequences of two proteins belonging to a relatively uncharacterized protein family. We could group taxonomy-related sequences and identify conserved regions with strong homologous association patterns within individual proteins and across the members of these families. Our results revealed that the data information measures are unbiased, and our class information measures can accurately rank the quality of the taxonomic relevant groupings. Furthermore, by combining our data and class measures, we were able to interpret the results by inferring regions of biological importance within the binding domain of these proteins. Compared to popular supervised methods, our algorithm has a superior runtime and comparable accuracy (22).

Figure 4 shows APCs obtained by our unsupervised methods using data and class information measures. In Figure 4, Part I shows the steps that take the protein family sequences with embedded class patterns, mammal (red), plant (green) or fungi (blue) and output APCs (step 2) which produce clusters in the APC dataspace in an unsupervised manner. It validates the class partition results after the putting back the class labels to the samples data in the APC dataspace. Step 3a shows how the sites with conserved or mutated residues can be identified using a data information measure R1 that reveals the degree of residue conservation of the site. For the first, fourth and fifth columns of 3a, R1=1 indicates invariant sites. For columns 2 and 3 with residue variation, R1 between 1 (invariant) and 0 (uniform distribution). R1 measures the conservation of a site. We also use another data measure known as sum of mutual information (SR2) to account for the strength of interdependence of a site with all other sites in the APC, indicating the functional significance of the sites. In step 3b, we use a class information measure to account for the correlation of a site with the class labels, like E in column 3 are associated with plants and Q with mammals.

Figure 4 Part II shows that the top-ranking APC contains cysteine from the scavenger receptor cysteine rich (SRCR) binding domain. It also indicates that macrophage receptor with collagenous domain (MARCO), and scavenger receptor class A I (SRAI) have unique expressed functions sharing highly conserved domains with variations. Top APCs discovered correspond to two binding domains of SRCR having relatively high SR2 and Immunoglobulin (IG) (yellow ellipses) which separated the APCs into the MARCO and SRAI classes. The sites with C-C disulfide bonds are highlighted in the yellow columns (Figure 4 Part II). Note that the second APC SDATVFCR[MS]LGYS consists of numbered rows associated with a statistically significant pattern and columns representing either conserved or mutated sites. The instance counts of the patterns for MARCO and SRAI are listed on the right-hand columns labelled with MARCO (in pink) and SRAI (in green). The class entropy of each pattern is displayed in the last column with the heading H and each amino acid is displayed below the patterns, where the SR2 and IG are summarized at the bottom.

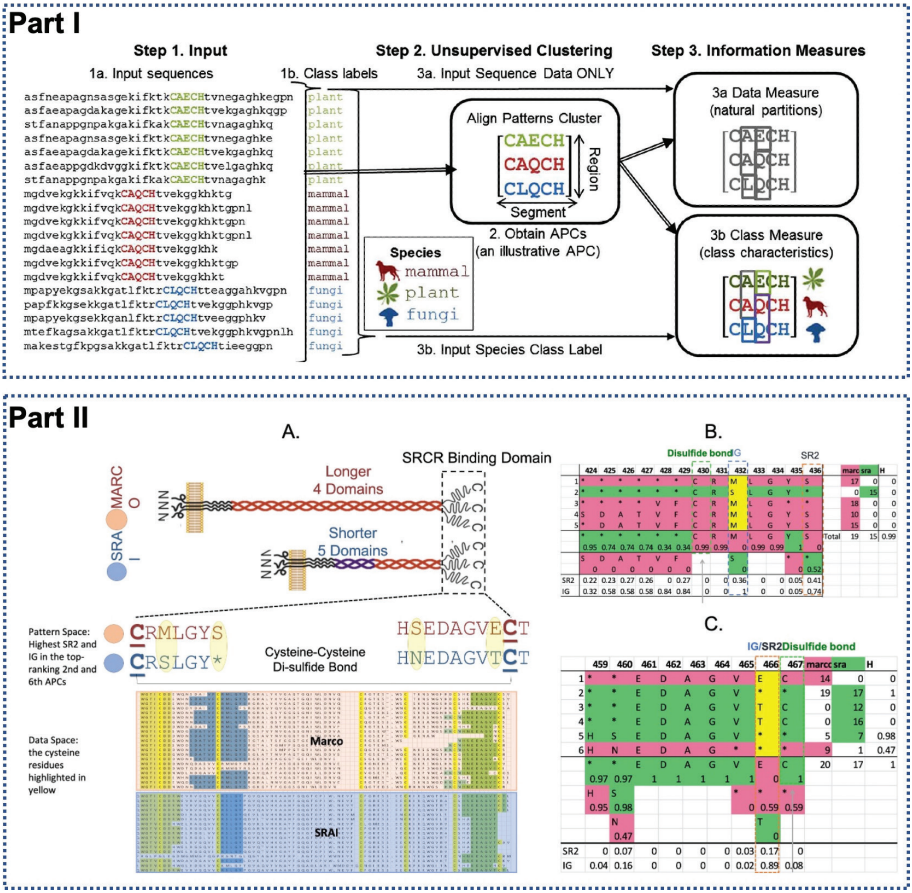

Figure 4. Partitioning and Analyzing APCs obtained by our unsupervised methods using data and class information measures. Part I. The steps in the APC process that takes as input the protein family sequences with embedded class patterns and outputs the final APCs with their associated classes. **Part II**. Examples of the top-ranking APCs contain cysteine from the SRCR binding domain. **A.** The top 2nd and 6th APCs in MARCO and SRAI. **B.** Top 2nd APC. **C.** Top 6th APC. SRCR, scavenger receptor cysteine-rich.

CONCLUSION

We describe our novel sequence pattern analysis system called WeMine System for discovering, summarizing, and refining patterns representations. Our method can reveal biological function of DNA sequences, protein domains in protein families (cytochrome *c* and class A scavenger receptor), protein-DNA binding cores, and protein-protein interactions. By utilizing insights from the pattern-data space duality, our results are rendering a more precise prediction due to flexible representations, a faster runtime due to compressed statistical results, and unbiased

disentangled interpretations of the results due to robust associations. We thus believe that PD-APC would be important for the discovery of new functional regions from protein family sequences, as well as rare mutants that will be significant to drug discovery and personalized medicine in the future.

Acknowledgment: Publication costs were funded by NSERC Discovery Grant (xxxxx 50503-10275 500).

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s).

REFERENCES

1. Bailey , Johnson J, Grant CE, Noble WS. The MEME suite. Nucl Acids Res. 2015;43(W1): W39–W49. https://doi.org/10.1093/nar/gkv416
2. Frith MC, Saunders NF, Kobe B, Bailey TL. Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol. 2008;4(5): e1000071. https://doi.org/10.1371/journal.pcbi.1000071
3. Altschul SF, Madden TL, Schaffer, AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res. 1997;25(17): 3389–3402. https://doi.org/10.1093/nar/25.17.3389
4. Basic Local Alignment Search Tool. [Online]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome.
5. Chen J. Contiguous item sequential pattern mining using UpDown Tree. Intell Data Anal 2008;12(1):25–49. https://doi.org/10.3233/IDA-2008-12103
6. Wang J, Han J. BIDE: Efficient mining of frequent closed sequences. In Proceedings. 20th international conference on data engineering; 2004. p.79–90. https://doi.org/10.1109/ICDE.2004.1319986
7. Li C, Wang J. Efficiently mining closed subsequences with gap constraints. In proceedings of the 2008 SIAM International Conference on Data Mining; 2008. p.313–322. https://doi.org/10.1137/1.9781611972788.28
8. Pitre S, Hooshyar M, Schoenrock A, Samanfar B, Jessulat M, Green J, et al. Short co-occurring polypeptide regions can predict global protein interaction maps. Sci Rep 2012;2:239. https://doi.org/10.1038/srep00239
9. Pitre S, North C, Alamgir M, Jessulat M, Chan A, Luo X, et al. Global investigation of protein--protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. Nucl Acids Res. 2008; 36(13):4286–4294. https://doi.org/10.1093/nar/gkn390
10. Wong AKC, Li GCL. Simultaneous pattern and data clustering for pattern cluster analysis. IEEE Trans Knowl Data Eng. 2008;20(7):911–923. https://doi.org/10.1109/TKDE.2008.38
11. Matys V, et al. TRANSFAC (R) and its module TRANSCompel (R): 767 transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34:D108–768. https://doi.org/10.1093/nar/gkj143
12. Wong AKC, Zhuang D, Li GCL, Lee ESA. Discovery of delta closed patterns and noninduced patterns from sequences. IEEE Trans Knowl Data Eng. 2011;24(8):1408–1241. https://doi.org/10.1109/TKDE.2011.100
13. Colon W, Wakem LP, Sherman F, Roder H. Identification of the predominant non-native histidine ligand in unfolded cytochrome c. Biochemistry. 1997;36(41):12535–12541. https://doi.org/10.1021/bi971697c

14. Lee ESA, Wong AK. Ranking and compacting binding segments of protein families using aligned pattern clusters. Proteome Sci. 2013;11(S1):S8. https://doi.org/10.1186/1477-5956-11-S1-S8
15. Wong AK, Lee ESA. Aligning and clustering patterns to reveal the protein functionality of sequences. IEEE/ACM Trans Comput Biol Bioinform. 2014;11(3):548–560. https://doi.org/10.1109/TCBB.2014.2306840
16. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics. 2004;20(4):477–486. https://doi.org/10.1093/bioinformatics/btg432
17. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. Bioinformatics. 2007;23(13):i347–i353. https://doi.org/10.1093/bioinformatics/btm174
18. Chan TM, Li G, Leung KS, Lee KH. Discovering multiple realistic TFBS motifs based on a generalized model. BMC bioinformatics. 2009;10(1): 321. https://doi.org/10.1186/1471-2105-10-321
19. Lee ESA, Sze-To HYA, Wong MH, Leung KS, Lau TCK, Wong AK. Discovering protein-dna binding cores by aligned pattern clustering. IEEE/ACM Trans Comput Biol Bioinform. 2015;14(2):254–263. https://doi.org/10.1109/TCBB.2015.2474376
20. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002;415(6868):141–147. https://doi.org/10.1038/415141a
21. Hu L, Chan KCC. Discovering variable-length patterns in protein sequences for protein-protein interaction prediction. IEEE Trans Nanobioscience. 2015;14(4): 409-416. https://doi.org/10.1109/TNB.2015.2429672
22. Lee ESA, Whelan FJ, Bowdish DM, Wong AKC. Partitioning and correlating subgroup characteristics from Aligned Pattern Clusters. Bioinformatics. 2016; 32(16): 2427–2434. https://doi.org/10.1093/bioinformatics/btw211