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 these biosequences, known as 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, and has been applied in 1) 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.

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

<table>
<thead>
<tr>
<th>Role</th>
<th>Pattern Space</th>
<th>Data Space</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What</strong></td>
<td>Patterns (statistically significant sequential base/residue association) with statistical significance residuals and ranking.</td>
<td>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.</td>
</tr>
<tr>
<td><strong>Where</strong></td>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td><strong>How</strong></td>
<td>Interpreting and assessing association patterns and sites. Looking for biologically relevant functions/interacting macromolecules and sites and obtaining additional supporting and explanatory evidence.</td>
<td>Revealing and interpreting functional characteristics of conserved regions (local/distant) for useful actions for different classes, groups, samples or individual biosequences.</td>
</tr>
<tr>
<td><strong>Why</strong></td>
<td>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.</td>
<td>Data of discovered patterns/APCs/cAPCs within and between sequences provides a statistical and functional base to validate the underlying models/mechanisms.</td>
</tr>
</tbody>
</table>

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...
and trios phosphate isomerase is demonstrated in [8-11].

essential binding sites in protein families of cytochrome c, ubiquitin, sub-class might interact differently. Its capability in uncovering the Furthermore, it shows two cAPC subgroups, indicating that protein APCs (proximal and distal) sharing 23% of co-occurring patterns. and brought into cAPCs based on juccard index [15-16]. It shows two patterns to their variants with minor mutations. Figure 2 shows how APCs to construct feature vectors to represent interaction of protein sequence.

Alignment 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 jaccard 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 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 jaccard 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 results [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

Table 2: Qualitative and Quantitative Experimental Comparison.

<table>
<thead>
<tr>
<th>Description</th>
<th>Qualitative Comparison</th>
<th>Quantitative Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern Discovery [9]</td>
<td>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.</td>
<td>- 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</td>
</tr>
<tr>
<td>Aligned Pattern Clustering (APC) [8-11]</td>
<td>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.</td>
<td>- 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</td>
</tr>
<tr>
<td>Protein-DNA Co-Occurring APC Discovery [18,19]</td>
<td>(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.</td>
<td>- Our results have higher consistency (~20%) to those obtained 3D structures by comparing to the latest published binding core discovery algorithm - Our approach has a speed-up of over 1600X comparing with the latest published binding discovery algorithm</td>
</tr>
<tr>
<td>Predicting Protein-protein interaction (PPI) [17]</td>
<td>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.</td>
<td>WeMine-P2P (1) outperforms PIPE2 [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.</td>
</tr>
<tr>
<td>Predicting Protein-DNA Co-Occurring APC Cores allowing minor mutations was developed to represent TF-TFBS binding cores shown in the format such as: TF: [FCNRRLK,FQNRMK,FQNRRAK] with TTATTTG, TTAATTG as TFBS.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
Co-occurring intra-protein functional elements: Patterns from two conserved regions with patterns co-occurring frequently on the same sequences suggest joint functionality. We discovered 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”.

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 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