Using Fixed Point Theorems to Model the Binding in Protein–Protein Interactions

Jinyan Li and Haiquan Li

***Abstract*— The binding in protein–protein interactions exhibits a kind of biochemical stability in cells. The mathematical notion of fixed points also describes stability. A point is a fixed point if it keeps unchanged after a transformation by a function. Many points may not be a fixed point, but they may approach to a stable status after multiple steps of transformation. In this paper, we define a point as a protein motif pair consisting of two traditional protein motifs. We propose a function and propose a method to discover stable motif pairs of this function from a large protein interaction sequence dataset. There are many interesting properties for this function (for example the convergence). Some of them are useful for gaining much efficiency in the discovery of those stable motif pairs; some are useful for explaining reasons why our proposed fixed point theorems are a good way to model the binding of protein interactions. Our results are also compared to biological results to elaborate the effectiveness of our method.**

***Index Terms*— Bioinformatics (genome or protein) database, mining methods and algorithms, generating functions, stability and instability, biology and genetics.**

I. Introduction

LET f be a function and x be a point in its domain, if $f(x) = x$, then x is called a *fixed point* for f. A famous fixed point theorem in modern mathematics, proposed by L. Brouwer in 1911, says that any continuous function $f : B \to B$, where B is a closed ball in R^n, has at least one fixed point [1]. An easy example of fixed points is $x = 1$ for $f(x) = 2x - 1$. Hence, the idea of fixed points is to find conditions under which a function possesses a point that maps into itself. An interesting instantiation of this mathematical notion is in life science: The DNA of a cell can be split into two parts, then they grow, in two separate cells, to become the same DNA as the original one after self-replicating. In this example, the x is the DNA, and the $f(x)$ is the laws of physics and chemistry applied to the DNA.

Recently, we made an important discovery for fixed points at protein type level [2]. The study is on genomic sequences of a gene family. This family of genes is called C2H2 Zinc-Finger genes, consisting of 226 members. A characteristic of this gene family is the frequent presence of tandem repeats. An interesting problem about these genes is whether they can be translated into the same type of protein before and after a *frameshift*. We found 12 of them that can be each translated into the same type of protein after frameshifts. Again, this is a fixed point phenomenon. The x is the protein type, the function $f(x)$ is the frameshift.

In this paper, we apply fixed point theorems to model the binding in protein–protein interactions, where we define a point as a protein motif pair [22], [23] consisting of two traditional protein motifs. To transform starting motif pairs to become stable motif pairs, we propose a function $f_{\mathcal{D}}$ where $\mathcal{D}$ is a protein interaction sequence dataset. Next, we explain why we choose a motif pair instead of a traditional single motif as a point, and why this *in-silico* study is important.

A protein is a complex, high molecular weight organic compound that consists of linear *amino acids* joined by peptide bonds. Proteins are essential to the structures and functions of all living cells and viruses. Many proteins are enzymes or subunits of enzymes. Other proteins play structural or mechanical roles. Since a protein is a chain of amino acids, it can be mathematically represented by a *string* of the abbreviations[1] of the 20 standard amino acids, allowing repetitions. Life of cells depends on the interactions of proteins [3]. The interactions are through the so-called *binding motifs* [4], each a region on a protein, to connect pairs of proteins.

In the biology field, it is a challenging problem to identify binding motifs. A commonly-used way is to examine the 3-D structure of the so-called *protein complex* data [5] generated by X-ray crystallography [6], [7] or by multidimensional nuclear magnetic resonance (NMR) [8], [9]. But, these methods are time-consuming and expensive. However, it is relatively easy and economical to get the amino acid sequence data (strings of amino acid letters) for a pair of interacting proteins, and these interaction sequence data have been shown to be useful for discovering single binding motifs. (See Brazma et al. [4], [10] for a good survey about the algorithms to discover binding motifs.)

In this paper, we are more interested in binding motif pairs consisting of two traditional protein motifs, and try to discover them using fixed point theorems from large amount of protein interaction sequence data. A recent study reported that protein interactions could be determined by correlated mutations during evolution [11]. For example, the co-evolution of interacting protein pairs has long been observed in such well-known interacting protein pairs as dockerins and cohesins [12], as well as insulin and its receptors [13]. These mutations are thought to be interactively happening between the binding sites of a pair of interacting proteins: if a residue [2] change incurred in one protein disrupts its interaction with its

Manuscript received 19 July 2004; revised 30 October 2004.

Jinyan Li and Haiquan Li are with Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613. Email:{jinyan,haiquan}@i2r.a-star.edu.sg

Haiquan Li is also with School of Computing, National University of Singapore, Singapore,119260.

[1] These abbreviations are a, c, d, e, f, g, h, i, k, l, m, n, p, q, r, s, t, v, w, and y.

[2] An equivalent name to an amino acid.

partner, some compensatory residue changes must also occur in its interacting partner in order to sustain the interaction, otherwise, they will be selected against and be eliminated. Therefore, a more proper way to study the binding of protein interactions is to focus on binding motif pairs instead of only those individual binding motifs.

The correlated mutations in the evolution imply a chain of binding motif pairs. We can assume that the recently survived binding motif pairs should occur more frequently than those ancient binding motif pairs, and should be more frequent than those non-binding motif pairs. Also, the recent survived binding motif pairs should be more stable than others. Otherwise, they would be mutated further. Based on these ideas and assumptions, we emulate the transformation in fixed point theorems to model the evolution of binding sites, and use fixed points to model the survived binding sites. As will be seen in Section VII, such discovered stable motif pairs are biologically interesting.

The remaining of the paper is organized as follows: In Section II, we define basic notations. In Section III, we give a formal description of the problem. In Section IV, we introduce a function $f_{\mathcal{D}}$ that is closely related to a sequence dataset $\mathcal{D}$ of protein interactions. The function will be used to transform protein motif pairs such that they can become stable ones. In Section V, we prove and discuss the properties of $f_{\mathcal{D}}(X)$, including the convergence property and the forest-like decomposition of its domain. In Section VI and Section VII, we introduce a method to select good starting point X, and apply our ideas to a massive real-life protein interaction sequence data to find meaningful fixed points. We also give full details of some fixed points and explain their biological meanings to show the significance of our model. We conclude this paper in Section VIII.

II. Basic Notations

We use Σ to denote the alphabet set of the 20 standard amino acids. All the amino acids are denoted by lower-case letters; but proteins and amino acid patterns are denoted by capital letters. A protein P is defined as *a sequence* (a string) of amino acids. For example, P can be $a_1a_2\cdots a_v$, where $a_i \in \Sigma$ for $i = 1, \cdots, v$. This P is also called a v-length protein. A *segment* of a protein P is a substring of P where amino acids are connected continuously.

An *amino acid pattern*, or called a protein *motif*, is defined as a sequence (a string) of subsets of Σ. Hence, a motif M can be written in the form $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$, where $\mathcal{A}_i \subseteq \Sigma$ for $i = 1, \cdots, k$.

The following is an example of protein motifs that was found to be biologically important in signal transduction [14], [15]. This protein motif is $\{p\}\Sigma\{l\}\{p\}\Sigma\{kr\}$ that binds to the SH3 domain of the protein *CrkA*. The length of this motif is 6; the second position of this motif is the whole alphabet set, meaning "don't care what is matched". It can also be written as $\{p\} * \{l\}\{p\} * \{kr\}$ in a traditional way by replacing Σ with the sign "*".

Definition 1: Let a motif M be $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$, where at least one $\mathcal{A}$ is not $\emptyset$. M is defined to be *contained* in a protein $P = a_1a_2\cdots a_v$ if there exists a k-length segment of P, denoted $a_{i+1}a_{i+2}\cdots a_{i+k}$ for some i, such that $a_{i+j} \in \mathcal{A}_j$ for all $\mathcal{A}_j$, $1 \leq j \leq k$, that are not $\emptyset$. If a motif is a sequence of only empty sets, we define that there is no protein containing such a motif.

A motif M contained in a protein P is denoted by $M \subseteq P$, and the segment $a_{i+1}a_{i+2}\cdots a_{i+k}$ is said to *match* the motif M.

Next, we give definitions related to interactions. A pair of interacting proteins P_1 and P_2 is called a *protein pair PPr*. This pair is denoted by the set of the two proteins, that is, $PPr = \{P_1, P_2\}$. A *motif pair*, denoted *MPr*, is a set of two motifs. One of the most important definitions used in this paper is about the inclusion relationship between a motif pair and a protein pair.

Definition 2: Let $MPr = \{M_1, M_2\}$ be a motif pair and $PPr = \{P_1, P_2\}$ be a protein pair. *MPr* is *contained* in *PPr*, denoted $MPr \subseteq PPr$, if (1) $M_1 \subseteq P_1$ and $M_2 \subseteq P_2$, or (2) $M_1 \subseteq P_2$ and $M_2 \subseteq P_1$.

Let two proteins: $P_l = eanftw$, $P_r = wefc$, and three motifs: $M_1 = \{ard\}\{nc\}$, $M_2 = \{e\}\{f\}$, and $M_3 = \{ard\}\emptyset\{nc\}$. Then the protein P_l contains the motif M_1, i.e. $M_1 \subseteq P_l$. This is because there exists a 2-length segment *an* in P_l such that $a \in \{ard\}$ and $n \in \{nc\}$. Similarly, $M_2 \subseteq P_r$. Hence, the motif pair $\{M_1, M_2\}$ is contained in the protein pair $\{P_l, P_r\}$.

However, the motif $M_3 = \{ard\}\emptyset\{nc\}$ is not contained in any of the two proteins because there does not exist any 3-length segment in P_l or P_r that can match M_3. Therefore, motif pairs $\{M_1, M_3\}$ or $\{M_2, M_3\}$ cannot be contained in the protein pair $\{P_l, P_r\}$. But, if M_3 is changed to $M_3' = \{erd\}\emptyset\{nc\}$, then both P_l and P_r contain M_3'. Note that the empty set $\emptyset$ in M_3 or M_3' has the same semantic meaning as that of Σ in this case (See Definition 1).

We denote a *sequence dataset* $\mathcal{D}$ of n protein pairs by $\{PPr^i = \{P_1^i, P_2^i\}, i = 1, \ldots, n\}$, where P_1^i and P_2^i have interactions.

Definition 3: The support of a motif pair $MPr = \{M_1, M_2\}$ in a protein sequence dataset $\mathcal{D}$ is defined as the number of protein pairs in $\mathcal{D}$ that contain *MPr*, denoted by $|\{PPr^i |\ PPr^i \in \mathcal{D}\ , MPr \subseteq PPr^i\}|$.

III. Problem Statement

Let $\mathcal{D}$ be a sequence dataset of interacting protein pairs, the problem studied in this paper is to design a function $f_{\mathcal{D}}$ that is closely related to $\mathcal{D}$, and then to discover stable motif pairs that are fixed points with regard to $f_{\mathcal{D}}$.

The domain of the function $f_{\mathcal{D}}$ is the set of all possible motif pairs. Let us first discuss the possibilities of single motifs. Recall that a motif is a sequence of subsets of Σ, denoted by $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$, where $\mathcal{A}_i \subseteq \Sigma$ for $i = 1, \cdots, k$. Hence, if $k = 1$, then the set of all possible motifs is the power set of Σ, denoted $\mathcal{P}(\Sigma)$. Then, possibilities of k-length motifs $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$ can be represented by the following set union:

$$\bigcup\{\mathcal{A}_1\cdots\mathcal{A}_k \mid \mathcal{A}_i \in \mathcal{P}(\Sigma) \text{ for } i = 1, \cdots, k\}$$

Since motif pairs are pairs of motifs, the set of all possible motif pairs has a much larger size than the domain of single motifs. We use $\mathcal{M}$ to denote all possibilities of motif pairs.

Therefore, in a formal way, the problem can be described as follows. Let $\mathcal{D}$ be a sequence dataset of protein pairs, our objective is to design a function

$$f_{\mathcal{D}} : \mathcal{M} \to \mathcal{M},$$

and to find those stable motif pairs X such that

$$f_{\mathcal{D}}(X) = X$$

by using an efficient algorithm.

IV. Our Proposed Function $f_{\mathcal{D}}$

As discussed, the function $f_{\mathcal{D}}$ is to transform a motif pair MPr through an interaction sequence dataset $\mathcal{D}$, and to make it become a different motif pair MPr' at most cases. Ideally, for any motif pair X, the following motif pairs, $f(X)$, $f(f(X))$, $\cdots$, $f(\cdots f(X))$, should converge to a stable motif pair. We will show our proposed $f_{\mathcal{D}}$ satisfies these conditions.

Given a motif pair $MPr = \{M_1, M_2\}$, our proposed $f_{\mathcal{D}}$ involves three steps to transform MPr. In the first step, it discovers a *subset* of $\mathcal{D}$ such that for every protein pair PPr in this subset, PPr contains the given motif pair MPr. We denote this subset by

$$s_{\mathcal{D}}^{MPr} = \{PPr \mid PPr \in \mathcal{D}, MPr \subseteq PPr\}. \qquad (1)$$

In the second step, $f_{\mathcal{D}}$ moves to extract a *segment pair* from every protein pair in $s_{\mathcal{D}}^{MPr}$. Let $Y = \{P_l, P_r\} \in s_{\mathcal{D}}^{MPr}$, then $MPr \subseteq Y$. Therefore, there must exist: (1) a segment in P_l that matches M_1 and a segment in P_r that matches M_2, or (2) a segment in P_r that matches M_1 and a segment in P_l that matches M_2. If the both cases are true, we choose either of them. In any case, we denote the segment that matches M_1 by $segment_1$, and the segment that matches M_2 by $segment_2$. Observe that M_1 and $segment_1$ have the same length, and so for M_2 and $segment_2$. Suppose there are u protein pairs in $s_{\mathcal{D}}^{MPr}$, then we can get u number of $segment_1$ and u number of $segment_2$. Let the length of $segment_1$ be w. Then, the u $segment_1$ can be represented as the following matrix $[a_{ij}]$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1w} \\ a_{21} & a_{22} & \cdots & a_{2w} \\ & & \cdots & \\ a_{u1} & a_{u2} & \cdots & a_{uw} \end{bmatrix}$$

This matrix is denoted by $aln_s^{M_1}$. It is called the *alignment* of M_1 with regard to $s_{\mathcal{D}}^{MPr}$ in the bioinformatics literature. Similarly, we can represent those u $segment_2$ as another matrix, denoted by $aln_s^{M_2}$.

In the third step, our $f_{\mathcal{D}}$ moves to find a *consensus pattern* from the matrix $aln_s^{M_1}$ and a consensus pattern from the matrix $aln_s^{M_2}$. In the matrix $aln_s^{M_1}$, for every column j, denoted by $[a_{ij}], i = 1, \cdots, u$, we choose those a_{ij}, whose population in this column is larger than a *threshold*, to form a set denoted by $\mathcal{A}_j$. If none of these a_{ij} satisfies the condition, we set this position as $\emptyset$. Then the sequence $\mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_w$, a motif, is called the consensus pattern of M_1. This consensus pattern is denoted by M_1'. Similarly, we can find the consensus pattern M_2' for M_2. Then $\{M_1', M_2'\}$ is a transformed motif pair for $MPr = \{M_1, M_2\}$ by $f_{\mathcal{D}}$. Therefore, we can write $f_{\mathcal{D}}(\{M_1, M_2\}) = \{M_1', M_2'\}$.

The threshold for the amino acids' population in a column is important for the consensus pattern discovery. In this paper, we use 20%, a percentage value, as the threshold. That is, if the occurrence rate of an amino acid at a column is less than 20%, then we drop it, not allowing it to get into the consensus pattern. Absolute support numbers are also possible for the threshold, but we explain later why percentage thresholds are better than absolute ones.

The discussion above assumes that $s_{\mathcal{D}}^{MPr}$ is non-empty. To let $f_{\mathcal{D}}$ be well-defined, we define the following extreme case for $f_{\mathcal{D}}$: Given a motif pair $X = \{M_1, M_2\}$, if $s_{\mathcal{D}}^{X} = \emptyset$, we define $f_{\mathcal{D}}(X) = \{\emptyset \cdots \emptyset, \emptyset \cdots \emptyset\}$, where the number of empty sets in the first sequence is the length of M_1, and the number of empty sets in the second sequence is the length of M_2. Note that if a motif pair $X = \{\emptyset \cdots \emptyset, \emptyset \cdots \emptyset\}$, then $f_{\mathcal{D}}(X) = X$. Such a motif pair is a trivial fixed point for $f_{\mathcal{D}}$.

Next, we use an example to show how $f_{\mathcal{D}}$ proceeds. Let a motif pair X be $\{M_1, M_2\}$, where $M_1 = \{a\}\{g\}\{g\}\{g\}\{iy\}$ and $M_2 = \{fv\}\{g\}\{ek\}\{ae\}\{ens\}\{il\}\{a\}$. Let $\mathcal{D}$ be a sequence dataset of interacting protein pairs. Suppose $s_{\mathcal{D}}^{X}$ contains the following 7 protein pairs

$$\begin{array}{ll} \{qqq\mathbf{agggi}yy, & eei\mathsf{fgkasia}ss\} \\ \{aa\mathsf{fgkasia}yy, & sss\mathbf{agggy}qy\} \\ \{yy\mathbf{agggi}qqq, & vx\mathsf{fgkasia}kk\} \\ \{kks\mathbf{agggy}ssa, & ggq\mathsf{vgeaeia}ii\} \\ \{vv\mathbf{agggi}yy, & iii\mathsf{vgeaeia}sss\} \\ \{qqq\mathsf{vgeaeia}kk, & yyy\mathbf{agggi}qqq\} \\ \{qqq\mathbf{agggy}qqq, & qqq\mathsf{vgeenla}yy\}. \end{array}$$

Then $aln_s^{M_1}$—the segments from the 7 protein pairs that match M_1—is the following matrix:

$$\left[\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ \hline a & g & g & g & i \\ a & g & g & g & y \\ a & g & g & g & i \\ a & g & g & g & y \\ a & g & g & g & i \\ a & g & g & g & i \\ a & g & g & g & y \end{array}\right]$$

The consensus pattern M_1' for this matrix is

$$\{a\}\{g\}\{g\}\{g\}\{iy\}.$$

Observe that M_1' is equal to M_1. This is because that at the fifth column of this matrix, both i and y occur more than 20%. Hence, they are kept in the consensus pattern.

Similarly, $aln_s^{M_2}$—the segments that match M_2—is the following matrix:

$$\left[\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline f & g & k & a & s & i & a \\ f & g & k & a & s & i & a \\ f & g & k & a & s & i & a \\ v & g & e & a & e & i & a \\ v & g & e & a & e & i & a \\ v & g & e & a & e & i & a \\ v & g & e & \mathbf{e} & \mathbf{n} & \mathbf{l} & a \end{array}\right]$$

The consensus pattern M_2' for this matrix is

$$\{fv\}\{g\}\{ke\}\{a\}\{se\}\{i\}\{a\}.$$

Note that M_2' is not equal to M_2. Also observe that the amino acids e, n, l at columns 4, 5, and 6 (in bold font) respectively are dropped. Therefore, they do not appear in the fourth, fifth, and sixth set of M_2'.

Since $f_{\mathcal{D}}(\{M_1, M_2\}) = \{M_1, M_2'\}$, $X = \{M_1, M_2\}$ is not a fixed point of $f_{\mathcal{D}}$.

This example has illustrated that $f_{\mathcal{D}}$ uses three steps—discovery of a subset of $\mathcal{D}$, extraction of segments from this subset, and discovery of consensus patterns—to transform a given motif pair.

V. Properties of $f_{\mathcal{D}}$

This section presents some important properties of $f_{\mathcal{D}}$. At first part, we prove the convergence property of $f_{\mathcal{D}}$ for any starting motif pair, and also discuss the forest structure of the domain of $f_{\mathcal{D}}$. At the second part, we discuss some specific properties of $f_{\mathcal{D}}$ when the consensus pattern threshold is set as percentage values or set as absolute numbers. At the third part, we explain why using percentage thresholds is a better choice than using absolute numbers for our fixed point theorems to model the binding in protein–protein interactions.

A. Convergence properties

Proposition 1: Given a motif pair Y and a sequence dataset $\mathcal{D}$ of interacting protein pairs, let $X = f_{\mathcal{D}}(Y)$ and $X' = f_{\mathcal{D}}(X)$, then $s_{\mathcal{D}}^{X'} \subseteq s_{\mathcal{D}}^{X}$.

Proof: If $s_{\mathcal{D}}^{X'} = \emptyset$, of course, $s_{\mathcal{D}}^{X'} \subseteq s_{\mathcal{D}}^{X}$. Next we prove this proposition for $s_{\mathcal{D}}^{X'} \neq \emptyset$. Denote $X = \{M_1, M_2\}$, $M_1 = \mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_v$, $M_2 = \mathcal{B}_1\mathcal{B}_2 \cdots \mathcal{B}_w$; $X' = \{M_1', M_2'\}$, $M_1' = \mathcal{A}_1'\mathcal{A}_2' \cdots \mathcal{A}_v'$, $M_2' = \mathcal{B}_1'\mathcal{B}_2' \cdots \mathcal{B}_w'$. Because X is a motif pair resulting from Y after a transformation by $f_{\mathcal{D}}$, then $\mathcal{A}_i' \neq \emptyset$ and also $\mathcal{A}_i' \subseteq \mathcal{A}_i$ for those i satisfying $\mathcal{A}_i \neq \emptyset$. Similarly, $\mathcal{B}_i' \neq \emptyset$ and also $\mathcal{B}_i' \subseteq \mathcal{B}_i$ for those i satisfying $\mathcal{B}_i \neq \emptyset$. That is, if $\mathcal{A}_i \neq \emptyset$ (respectively $\mathcal{B}_i \neq \emptyset$), $\mathcal{A}_i'$ (respectively $\mathcal{B}_i'$) would never become an empty set under the percentage thresholds such as 20% used in this paper. (Note that this is not true when X is an arbitrary motif pair. That is why we need to set $X = f_{\mathcal{D}}(Y)$ for any Y.)

Let $PPr \in s_{\mathcal{D}}^{X'}$, we prove $PPr \notin \mathcal{D} - s_{\mathcal{D}}^{X}$. Assume $PPr \in \mathcal{D} - s_{\mathcal{D}}^{X}$, then $PPr \not\supseteq X$. Therefore, for any two segments from PPr, they cannot match M_1 and M_2 at the same time. Therefore, they cannot furthermore match M_1' and M_2' at the same time. This is because $\mathcal{A}_i' \subseteq \mathcal{A}_i$ for those i satisfying $\mathcal{A}_i \neq \emptyset$, and $\mathcal{B}_i' \subseteq \mathcal{B}_i$ for those i satisfying $\mathcal{B}_i \neq \emptyset$. Here is a contradiction. Thus our assumption, that $PPr \in \mathcal{D} - s_{\mathcal{D}}^{X}$, must be false. Therefore, we can conclude that $PPr \in s_{\mathcal{D}}^{X}$. ■

This proposition is useful for efficiently computing $s_{\mathcal{D}}^{X'}$. By definition, $s_{\mathcal{D}}^{X'}$ is a subset of $\mathcal{D}$ in which every protein pair contains the motif pair X'. Therefore, a naive way to compute $s_{\mathcal{D}}^{X'}$ is to check whether every protein pair in $\mathcal{D}$ contains X'. Having the proposition, this naive method becomes unnecessary because the check within $s_{\mathcal{D}}^{X}$ is sufficient. Since $s_{\mathcal{D}}^{X}$ is much smaller than $\mathcal{D}$, we can gain much efficiency.

Theorem 1: Let $\mathcal{D}$ be a sequence dataset of interacting protein pairs. Then for any starting motif pair X, $f_{\mathcal{D}}(X)$ converges to a fixed point X_F. That is, there exists an integer $t_0(\geq 1)$ such that $f_{\mathcal{D}}^{(t_0)}(X) = X_F$, and $f_{\mathcal{D}}(X_F) = X_F$, where $f_{\mathcal{D}}^{(1)}(X)$ represents $f_{\mathcal{D}}(X)$, $f_{\mathcal{D}}^{(2)}(X)$ represents $f_{\mathcal{D}}(f_{\mathcal{D}}(X))$, and $f_{\mathcal{D}}^{(t+1)}(X)$ represents $f_{\mathcal{D}}(f_{\mathcal{D}}^{(t)}(X))$.

Proof: Denote $X^{(0)} = X$, $X^{(1)} = f_{\mathcal{D}}^{(1)}(X)$, $\cdots$, $X^{(t)} = f_{\mathcal{D}}^{(t)}(X)$.

By Proposition 1, we know that $s_{\mathcal{D}}^{X^{(t+1)}} \subseteq s_{\mathcal{D}}^{X^{(t)}}$ for any $t \geq 1$. Since $s_{\mathcal{D}}^{X^{(1)}}$ is a limited set, there must exist a $t \geq 1$ such that $s_{\mathcal{D}}^{X^{(t)}} = s_{\mathcal{D}}^{X^{(t+1)}}$. Therefore, the consensus pattern from $s_{\mathcal{D}}^{X^{(t)}}$ is equal to the consensus pattern from $s_{\mathcal{D}}^{X^{(t+1)}}$. Because the consensus pattern from $s_{\mathcal{D}}^{X^{(t)}}$ is represented as $X^{(t+1)}$, and the consensus pattern from $s_{\mathcal{D}}^{X^{(t+1)}}$ is represented as $X^{(t+2)}$, we have $X^{(t+1)} = X^{(t+2)}$. That is, $f_{\mathcal{D}}(X_F) = X_F$, where $X_F = X^{(t+1)}$, as desired. ■

From this theorem, we can understand: (1) that any starting motif pair will converge to a fixed point (likely an empty pattern) and (2) that different starting motif pairs may converge to the same fixed point. Therefore, the domain of $f_{\mathcal{D}}$ can be partitioned into non-overlapping clusters with each cluster corresponding to one fixed point. More specifically, each cluster is a tree, as proved by the following proposition. Which trees are interesting and biologically meaningful? In the next section, we provide a heuristics.

Proposition 2: The domain (search space) of $f_{\mathcal{D}}$ is a forest, with each root node as a fixed point (a stable motif pair).

Proof: We denote a motif pair X as a node. If an edge is set from all possible X to $f_{\mathcal{D}}(X)$, the search space can be viewed as a graph. Since $f_{\mathcal{D}}(X)$ is an unique motif pair, the out-degree of each node should be no more than one. Meanwhile, it is impossible to have a circle in the graph. Assume $X_0, X_1 \ldots X_k, X_0$ is a circle. According to Proposition 1, $s_{\mathcal{D}}^{X_0} \supseteq s_{\mathcal{D}}^{X_1} \ldots \supseteq s_{\mathcal{D}}^{X_t} \supseteq s_{\mathcal{D}}^{X_0}$. Then $s_{\mathcal{D}}^{X_0} = s_{\mathcal{D}}^{X_1} \ldots = s_{\mathcal{D}}^{X_t} = s_{\mathcal{D}}^{X_0}$. Therefore, $X_0 = X_1 = \ldots = X_t$. Hence, X_0 is a fixed point. Thus it is impossible to have an out edge to X_1. Also, by Theorem 1, any motif pair can lead to a fixed point, with the out degree as zero, which is the corresponding root of that tree. ■

B. Specific properties

Recall that the definition of $f_{\mathcal{D}}$ involves a step for consensus pattern discovery. To find consensus patterns, we need a threshold to filter out those minor amino acids from the alignments. As mentioned, we have two options to select the threshold: one is to use percentage values as the threshold; the other is to use absolute numbers. We denote the former approach as $f_{(\%,\mathcal{D})}$, and the latter as $f_{(\pi,\mathcal{D})}$.

The following proposition shows that the stability of a fixed point of $f_{(\pi,\mathcal{D})}$ can be transferred to its sub-motifs. Here, a motif M' is a sub-motif of motif M if M' is a segment of M.

Proposition 3: Let a motif pair $X = \{M_1, M_2\}$ be a fixed point of $f_{(\pi,\mathcal{D})}$, then any of its sub-motif pairs $X' = \{M_1', M_2'\}$ is a fixed point of $f_{(\pi,\mathcal{D})}$ as well, where M_1' is a sub-motif of M_1, and M_2' is a sub-motif of M_2.

Proof: Because X' is a sub-motif pair of X, for $\forall \mathit{PPr} \in s_{\mathcal{D}}^{X}$, we have $\mathit{PPr} \in s_{\mathcal{D}}^{X'}$, i.e. $s_{\mathcal{D}}^{X} \subseteq s_{\mathcal{D}}^{X'}$. Since X is a fixed point of $f_{(\pi,\mathcal{D})}$, $\forall a_{ij} \in \mathcal{A}_i$ either from M_1 or from M_2, its population in $s_{\mathcal{D}}^{X}$ must be above the threshold. Since any occurrence of a_{ij} in $s_{\mathcal{D}}^{X}$ is also an occurrence of a_{ij} in $s_{\mathcal{D}}^{X'}$, the occurrence of $\forall a_{ij}$ in X' is also above the threshold. Therefore, X' is also a fixed point of $f_{(\pi,\mathcal{D})}$. ■

Proposition 3 says that the fixed points of $f_{(\pi,\mathcal{D})}$ satisfies the famous Apriori-property [16] known in data mining field. That is, if a sub-motif pair of a motif pair is not a fixed point, the motif pair is impossible to be a fixed point. Therefore, the mining of fixed points of $f_{(\pi,\mathcal{D})}$ should be similar to those algorithms for mining frequent itemsets.

Note that Proposition 3 does not hold if we replace $f_{(\pi,\mathcal{D})}$ with $f_{(\%,\mathcal{D})}$.

Proposition 4: Let X and Y be two equal-length stable motif pairs of $f_{(\pi,\mathcal{D})}$, where $X = \{M_{X1}, M_{X2}\}$, $Y = \{M_{Y1}, M_{Y2}\}$, $|M_{X1}| = |M_{Y1}|$ and $|M_{X2}| = |M_{Y2}|$. Then the union motif pair $X + Y = \{M_{X1} + M_{Y1}, M_{X2} + M_{Y2}\}$ is also a fixed point of $f_{(\pi,\mathcal{D})}$. The union operation $'+'$ of two motifs is defined as follows: suppose $M = \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$, and $M' = \mathcal{A}_1'\mathcal{A}_2'\cdots\mathcal{A}_k'$, then $M + M' = \mathcal{A}_1''\mathcal{A}_2''\ldots\mathcal{A}_k''$, where $\mathcal{A}_i'' = \mathcal{A}_i \cup \mathcal{A}_i'$, $1 \leq i \leq k$.

Proof: Observe that $\forall \mathit{PPr} \in s_{\mathcal{D}}^{X}$, then $\mathit{PPr} \in s_{\mathcal{D}}^{X+Y}$. Hence, we have $s_{\mathcal{D}}^{X} \subseteq s_{\mathcal{D}}^{X+Y}$. Similarly, we can get $s_{\mathcal{D}}^{Y} \subseteq s_{\mathcal{D}}^{X+Y}$. Since X and Y are fixed points of $f_{(\pi,\mathcal{D})}$, for $\forall a_{ij} \in \mathcal{A}_i$ either from M_{X1} or from M_{X2}, its support in $s_{\mathcal{D}}^{X}$ is above the threshold. Since any occurrence of a_{ij} in $s_{\mathcal{D}}^{X}$ is also an occurrence of a_{ij} in $s_{\mathcal{D}}^{X+Y}$, the occurrence of $\forall a_{ij}$ in $X + Y$ is also above the support threshold. Therefore, $X + Y$ is also a fixed point. ■

Note that this proposition may not hold if replacing $f_{(\pi,\mathcal{D})}$ with $f_{(\%,\mathcal{D})}$. This is because the occurrence of the union motif pairs not only covers the occurrences of the two original fixed points, but also covers some occurrences from new combinations. Therefore, it is difficult to determine whether the occurrence rate is still above the percentage threshold. Another interesting thing is if X is not a fixed point, $X + Y$ is not impossible to be a fix point of $f_{(\pi,\mathcal{D})}$.

Proposition 5: Let $f_{(\%,\mathcal{D})}$ be the $f_{\mathcal{D}}$ under the percentage threshold in the consensus pattern discovery. Let a motif pair $X = \{M_1, M_2\}$, where $M_1 = \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_v$, $\mathcal{A}_i \subseteq \Sigma$, for $i = 1, \cdots, v$; $M_2 = \mathcal{B}_1\mathcal{B}_2\cdots\mathcal{B}_w$, $\mathcal{B}_j \subseteq \Sigma$, for $j = 1, \cdots, w$. If all $\mathcal{A}_i$ and $\mathcal{B}_j$ are singleton sets, and $s_{\mathcal{D}}^{X} \neq \emptyset$, then X is a fixed point of $f_{(\%,\mathcal{D})}$.

Proof: Denote $\mathcal{A}_i = \{a_i\}$ for $i = 1, \cdots, v$, and $\mathcal{B}_j = \{b_j\}$ for $j = 1, \cdots, w$. Suppose $s_{\mathcal{D}}^{X}$ contains m protein pairs $\mathit{PPr}^i, i = 1, \cdots, m$. Then the segment from the protein pair PPr^i for every i that matches M_1 must be $a_1a_2\cdots a_v$; Similarly, the segment from the protein pair PPr^i for every i that matches M_2 must be $b_1b_2\cdots b_w$. Therefore, the two alignments $aln_s^{M_1}$ and $aln_s^{M_2}$ are the following two special matrixes:

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_v \\ a_1 & a_2 & \cdots & a_v \\ & & \cdots & \\ a_1 & a_2 & \cdots & a_v \end{bmatrix}$$

and

$$\begin{bmatrix} b_1 & b_2 & \cdots & b_w \\ b_1 & b_2 & \cdots & b_w \\ & & \cdots & \\ b_1 & b_2 & \cdots & b_w \end{bmatrix}$$

Then, the consensus pattern for $aln_s^{M_1}$ and $aln_s^{M_2}$ are $\{a_1\}\{a_2\}\cdots\{a_v\}$ and $\{b_1\}\{b_2\}\cdots\{b_w\}$ respectively, under percentage threshold, as the occurrence rate is 100% in this case. Hence, we can see that X is a fixed point of $f_{(\%,\mathcal{D})}$. ■

C. The function $f_{(\%,\mathcal{D})}$ better than $f_{(\pi,\mathcal{D})}$

In this subsection, we give a comparison between $f_{(\%,\mathcal{D})}$ and $f_{(\pi,\mathcal{D})}$, and explain the reasons for that $f_{(\%,\mathcal{D})}$ is better than $f_{(\pi,\mathcal{D})}$ to model the binding in protein–protein interactions.

First, let us examine the most likely lengths of fixed points derived by $f_{(\%,\mathcal{D})}$ and $f_{(\pi,\mathcal{D})}$. According to Proposition 3, for a *long* stable motif pair X of $f_{(\pi,\mathcal{D})}$, all sub-motif pairs of X are also fixed points of $f_{(\pi,\mathcal{D})}$. In extreme cases, those many 1-1 pairs are stable motif pairs. In biology, they are called residue–reside interaction pairs [17]. Though they may be fundamental components of some binding sites, they may have very high false positive rate. One way to solve this problem is to discover only those *maximal fixed points* of $f_{(\pi,\mathcal{D})}$ which are similar to a well studied data mining concept called maximal frequent patterns [18], [19]. On the other hand, both very short and very long motif pairs are unlikely to be fixed points of $f_{(\%,\mathcal{D})}$ due to the equal possibility for short motif pairs and rare possibility for long motif pairs. This property of $f_{(\%,\mathcal{D})}$ is very consistent with the observations in biology [20] that most binding sites generally include more than 10 but less than 20 residues. In fact, the lengths of our discovered stable motif pairs of $f_{(\%,\mathcal{D})}$ match very well with those of real motif pairs.

Secondly, let us discuss the union $('+')$ operation for $f_{(\%,\mathcal{D})}$ and $f_{(\pi,\mathcal{D})}$. According to Proposition 4, the union of *any* two equal-length fixed points of $f_{(\pi,\mathcal{D})}$ is also a fixed point of $f_{(\pi,\mathcal{D})}$, but this flexibility does not hold for fixed points of $f_{(\%,\mathcal{D})}$. In the real biology circumstances, this union property does not usually hold for binding sites either. For example, a study on active sites [21] shows that only specially selected amino acids (not arbitrarily united) are possible to compose a binding site or an active site. The union property of fixed points of $f_{(\pi,\mathcal{D})}$ also leads to another bad consequence: the motif pairs with large set in all positions are more likely to be fixed points. In the extreme case, the motif pairs which contain only full alphabet sets in each position are most likely to be fixed points. It is obviously meaningless from biology perspective. However, $f_{(\%,\mathcal{D})}$ does not produce such fixed points.

Hereby, $f_{(\%,\mathcal{D})}$ is better than $f_{(\pi,\mathcal{D})}$ for modeling the binding in protein–protein interactions, as it reflects more properties of the real binding sites. However, $f_{(\%,\mathcal{D})}$ has the singleton problem as discussed in Proposition 5. By this proposition, every segment pair from any protein pair of $\mathcal{D}$ is a fixed point of $f_{(\%,\mathcal{D})}$. Hence, it seems that there are many easy fixed points for $f_{(\%,\mathcal{D})}$. Therefore, we need other statistical measurements to remedy this, for example, using the support level or P-score of these fixed points in $\mathcal{D}$, or biological evidence as discussed in our another paper [22] to filter out some easy ones. In the remaining of the paper, any $f_{\mathcal{D}}$ refers back to $f_{(\%,\mathcal{D})}$.

VI. Selection of Starting Points for $f_{\mathcal{D}}$

Starting from any motif pair, we have already known (by Theorem 1) that this motif pair will become a fixed point after a number t_0 times of transformation by $f_{\mathcal{D}}$. Since the domain of the function $f_{\mathcal{D}}$ is huge, in this section we discuss a method to select good candidates for starting motif pairs, so that the resulting fixed points can have good biological significance.

As discussed in the introduction of this paper, protein interaction data are categorized into two types: protein interaction sequence data and protein complex data. Existing biotechnologies can generate high-throughput protein interaction sequence data efficiently. But, it is expensive and time-consuming to generate protein complex data. However, only protein complex data contains clear 3-D structure information for interacting proteins. From a protein complex, the exact locations of binding sites of the interacting proteins can be determined by calculating the distances between amino acids in a pair of proteins in this complex.

Hereby, in this paper, we use protein complex data as our platform because these data can provide important clues to guide the selection of meaningful starting motif pairs. We first discover binding sites from this kind of biologically reliable data. Then, we generalize these binding sites, and then transform those generalized patterns by our $f_{\mathcal{D}}$ to get stable motif pairs.

In one of our previous studies [23], we proposed a method to discover binding sites from protein complex data. These binding sites are called *maximal contact segment pairs* [23]. Two segments from two proteins are a *contact segment pair* if every residue in one segment can find at least one contact residue in the opposite segment, where the contact of two residues means that at least one of their atom pairs has an Euclidean distance less than a threshold. A contact segment pair is maximal if no any other contact segment pair in the same protein pair contains both segments of this contact segment pair, capturing contact segment pairs as lengthy as possible. The maximal contact segment pairs are then generalized into starting motif pairs. The formal definitions and explanations about maximal contact segment pairs and the search algorithms can be found in our previous work [23].

VII. Some Real-Life Examples

In this section, we report some fixed points of $f_{\mathcal{D}}$ discovered from a real-life sequence dataset $\mathcal{D}$ of interacting protein pairs. This sequence dataset is constructed by von Mering [24]. It consists of 78390 non-redundant interactions, containing almost all the latest interacting protein pairs in yeast genome produced by various experimental and high-confident computational methods. The lengths of these proteins are typically from hundreds to thousands. The data is also available at our website (`http://sdmc.i2r.a-star.edu.sg/BindingMotifPairs`).

Our starting motif pairs are also discovered from a real-life protein complex dataset. This protein complex dataset is derived from PDB (`http://www.rcsb.org/pdb/`). It consists of 1533 entries that have at least two chains, by using online search tools in PDB-REPRDB (`http://mbs.cbrc.jp/pdbreprdb-cgi//reprdb_query.pl`). In this complex dataset, the maximum pairwise sequence identity between any two complexes is 30% and each complex has a structure of resolution 2.0 or higher.

From this protein complex data, we identified 1222 starting motif pairs. After transformation by $f_{\mathcal{D}}$, 913 of them become fixed points that are not empty patterns. (That is, 309 of the 1222 starting motif pairs become the empty pattern: $\{\emptyset \cdots \emptyset, \emptyset \cdots \emptyset\}$.) Most of the 913 stable motif pairs have a length between 10 and 20. About 30% of these stable motif pairs have a support of at least 10 in $\mathcal{D}$.

Table I gives an example showing the transformation from a starting motif pair to a fix point, where three rounds of transformations by $f_{\mathcal{D}}$ are reported.

Next, we give full details for one of the 913 stable motif pairs to see how it is discovered, where its origin is, and what its biological significance is. This stable motif pair is

$$\{\{g\}\{ly\}\{d\}\{iy\}\{iv\}, \{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}\},$$

denoted by $MPr_{example} = \{M_1, M_2\}$, where $M_1 = \{g\}\{ly\}\{d\}\{iy\}\{iv\}$ and $M_2 = \{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}$.

Its origin is located at the so-called pdb1ors protein complex [25]. Specifically, the motif $M_1 = \{g\}\{ly\}\{d\}\{iy\}\{iv\}$ is evolved from the segment $gydyf$ at the chain B of the pdb1ors complex. These five amino acids are indexed from 99 to 103 residues in the chain B. See Figure 1. To combine these amino acids and their positions together, this segment is sometimes written as $[g99, y100, d101, y102, f103]$.

The motif $M_2 = \{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}$ is rooted at the segment $aglglfrl$ at the chain C of the pdb1ors complex. These eight amino acids are indexed from 111 to 118 residues in the chain C. This segment is sometimes written as $[a111, g112, l113, g114, l115, f116, r117, l118]$ to combine the amino acids and their positions together.

The segment pair, $[g99, y100, d101, y102, f103]$ and $[a111, g112, l113, g114, l115, f116, r117, l118]$, is a maximal contact segment pair. We use Figure 2, abstracted from Figure 1, to demonstrate it.

Using our method proposed in [23], this maximal segment pair $\{gydyf, aglglfrl\}$ is generalized to the following starting motif pair X,

$$X = \{\{g\}\{ly\}\{d\}\{fiy\}\{fiv\}, \{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{dr\}\{fi\}\}$$

for the function $f_{\mathcal{D}}$.

TABLE I

A STARTING MOTIF PAIR $X = \{\{ek\}\{g\}\{l\}\{l\}, \{k\}\{ek\}\{ek\}\Sigma\{g\}\{iv\}\}$ BECOMES A FIXED POINT OF $f_{\mathcal{D}}$ AFTER THREE ROUNDS OF TRANSFORMATION BY THIS FUNCTION.

convergence	motif pairs X											$\lvert s_{\mathcal{D}}^{X}\rvert$
starting	{ek}	{g}	{l}	{l}	,	{k}	{ek}	{ek}	Σ	{g}	{iv}	31
$X^{(1)}$	{ek}	{g}	{l}	{l}	,	{k}	{ek}	{ek}	{a}	{g}	{iv}	11
$X^{(2)}$	{ek}	{g}	{l}	{l}	,	{k}	{e }	{ k}	{a}	{g}	{ v}	10
M_{fixed}	{ek}	{g}	{l}	{l}	,	{k}	{e }	{ k}	{a}	{g}	{ v}	10

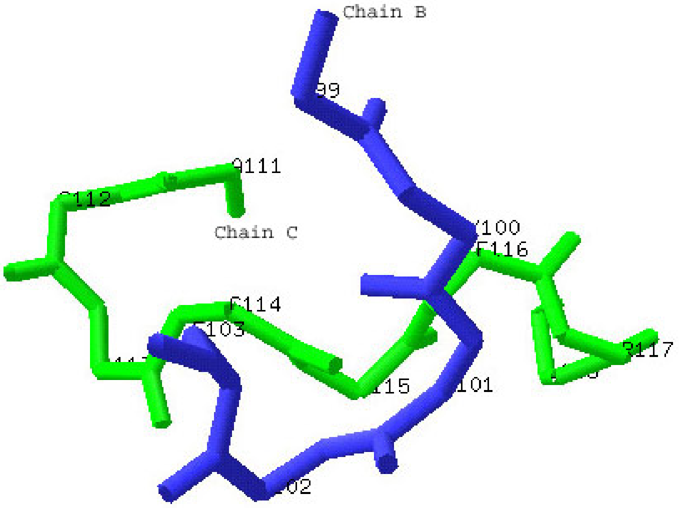

Fig. 1. 3-D structure of a binding site in the pdb1ors protein complex, a complex between the kvap potassium channel voltage sensor and an fab in species mouse and E. Coli., where Chain B is in blue color, and Chain C is in green color.

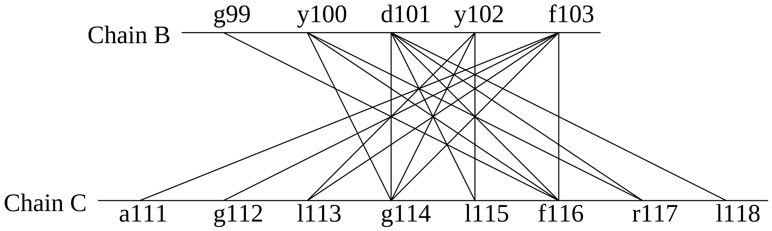

Fig. 2. A maximal contact segment pair discovered from the pdb1ors complex. A line between Chain B and Chain C represents that the two corresponding amino acids are close in distance.

After one step of transformation by $f_{\mathcal{D}}$, this starting motif pair X becomes the fixed point $MPr_{example}$, i.e. $f_{\mathcal{D}}(X) = MPr_{example}$.

We also found that this stable motif pair $MPr_{example}$ is statistically significant after examining its support level against random motif pairs. The support of motif $\{g\}\{ly\}\{d\}\{iy\}\{iv\}$ is 15 in yeast protein set (not the protein interaction sequence dataset $\mathcal{D}$), and the support of motif $\{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}$ is 2 with respect to the same protein set. The support of $MPr_{example}$ as a pair is 6 in the protein interaction sequence dataset $\mathcal{D}$. Then, we generated 1000 random motif pairs according to $MPr_{example}$, where each random motif pair is generated by substituting every residue in $MPr_{example}$ with a random residue. Therefore, the random motif pairs have the same length as $MPr_{example}$. The distribution of the randomly generated residues follows the same distribution of all the residues in the whole yeast genome. For these 1000 random motif pairs, the average support of the random motifs corresponding to $\{g\}\{ly\}\{d\}\{iy\}\{iv\}$ is 11.14, the support of every random motif corresponding to $\{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}$ is 0. Consequently, the support for any of those 1000 motif pairs is also 0 in the protein interaction sequence dataset $\mathcal{D}$. From these statistical numbers of $MPr_{example}$ and its equal-length 1000 random motif pairs, we can see that $MPr_{example}$ has occurrence much more than its random expectation in single motifs or in pairs. Therefore, the stable motif pair $MPr_{example}$ is not a random result indeed.

We also found some biological significance of the motif pair $MPr_{example}$. In biology, Pellicena and Miller [26] studied a protein motif $M_{PM} = \{y\}\{d\}\{y\}\{v\}$ within the protein p130Cas of v-Src transformed cells. This motif was biologically confirmed to bind to the Src homology 2 (SH2) domain that is a protein domain with about 100 amino-acid residues in many intracellular signal-transducing proteins [27]. We had the following observations after comparing these biological literature results with our computational results:

- $M_{PM} = \{y\}\{d\}\{y\}\{v\}$ is similar to the left motif $\{g\}\{ly\}\{d\}\{iy\}\{iv\}$ of our motif pair $MPr_{example}$.
- The segment $lvrf$ in the SH2 domain partially matches to our right motif $\{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}$ of $MPr_{example}$. The precise location of the segment $lvrf$ is from positions 118 to 121 at the SH2 domain of the protein $SH2A\_HUMAN$, and from positions 139 to 142 at the SH2 domain of the protein $SH2A\_MOUSE$. At the left side of the matched segments in the SH2 domain, there is a segment $qgcy$ from 114 to 117 in $SH2A\_HUMAN$. The residue q at position 114 of this segment is a structure interchangeable residue of r [28]; the residue g at position 115 exactly matches with the second residue in our motif; at position 116, both residue c and l are hydrophobic residues that imply some structure similarity; at position 117, both residue y and residue g are surface residues (charged/polar residues). Similarly, we find a segment gcy from 136 to 138 in $SH2A\_MOUSE$. Hereby, the right motif of $MPr_{example}$ has five positions which are exact matches and two positions which are compatible with the biological protein sequences (from a domain of 92 residues).
- There are total 295 proteins containing SH2 domains, where the segment lvr occurs in 139 of them. (This can be seen from the prosite: `http://tw.expasy.org/prosite/`.) Moreover, the segment lvr locates near the most conserved region in the domain, where the most conserved region is just between g—the second residue and r—the last second residue. (See `http://tw.expasy.org/cgi-bin/aligner?psa=PS50001&color=1&maxinsert=10&linelen=0`). This implies that the motif pair we discovered is likely to be the most critical factor for the binding between the $\{y\}\{d\}\{y\}\{v\}$ motif in p130Cas and SH2 domain.

Finally in this section, we describe two more examples to explain the biological significance of our discovered fixed points. Vancompernolle [29] reported a result that protein actobindin contains an actin-binding motif $\{v\}\{th\}\{v\}\{k\}\{k\}\{v\}$. From our discovered 913 stable motif pairs, we observed that there are three motif pairs containing motifs that are similar to the actin-binding motif $\{v\}\{th\}\{v\}\{k\}\{k\}\{v\}$. The left side and right side of the three motif pairs are listed in the second and third column of Table II respectively. A more interesting observation is that the three right-side motifs are all contained in the sequence of the protein actin or its associated proteins.

Kay et al [15] had a study on the interaction of proline-rich motifs in signaling proteins with their cognate domains. Four binding motifs (called binding consensus sequences in [15]) are listed in the first column of Table III. From our discovered binding motif pairs, we observed that there are 4 motif pairs containing a motif that is similar to one of the 4 binding motifs. The 4 motif pairs are listed in the second and third columns of Table III. Another observation is that our right-side motifs are all contained in the proteins in the last column of Table III which are reported to bind to the corresponding consensus sequences in the first column [15]. (Note that similar results have been obtained by using emergence significance measurement in our previous work [23].)

These observations indicate that the stable motif pairs discovered by our fixed-point based method would possess strong biological meaning. An important implication of this is that our discovered binding motif pairs are likely to be real biological binding sites. Therefore, this computational method would have a potential guidance role to play for the identification of real biological binding sites.

VIII. Conclusion

In this paper, we have proposed a fixed point theorem to model the binding in protein–protein interactions where a point is defined as a protein motif pair consisting of two traditional protein motifs. The transformation by a function emulates the evolution of binding sites, while the fixed points of the function models the binding sites. To discover stable motif pairs from the sequence data of interacting protein pairs, we proposed a mathematical function $f_{\mathcal{D}}$. The transformation of a motif pair by $f_{\mathcal{D}}$ involves three steps: the discovery of a subset of $\mathcal{D}$, the extraction of alignments from this subset, and the discovery of two consensus patterns. We have proved that $f_{\mathcal{D}}$ is a convergent function for any starting motif pairs. In this paper, we have also discussed that $f_{(\%,\mathcal{D})}$ is better than $f_{(\pi,\mathcal{D})}$ for modeling the binding in protein–protein interactions, as it reflects more properties of the real binding sites. We applied our method to a huge real-life dataset and found many biologically interesting motif pairs. As future work, we will collaborate with biologists to confirm our results using wet experiments. Meanwhile, we are also working on different functions $f_{\mathcal{D}}$ to see whether it can be optimized.

References

[1] A. K. Mohamed and A. K. William, *An introduction to metric spaces and fixed point theory*. John Wiley & Sons, 2001.

[2] S. W. Meng, Z. Zhang, and J. Li, "Twelve c2h2 zinc finger genes on human chromosome 19 can be each translated into the same type of protein after frameshifts," *Bioinformatics*, vol. 20, pp. 1–4, 2004.

[3] A. Bruce, B. Dennis, L. Julian, R. Martin, R. Keith, and W. James, D., *Molecular Biology of the Cell*, 4th ed. New York: Garland Science, 2002.

[4] P. Bork and E. Koonin, "Protein sequence motifs," *Curr. Opin. Struct. Biol.*, vol. 6, no. 3, pp. 366–76, 1996.

[5] A. Dziembowski and B. Seraphin, "Recent developments in the analysis of protein complexes," *FEBS. Lett.*, vol. 556, no. 1-3, pp. 1–6, Jan 2004.

[6] J. D. Bernal, I. Fankuchen, and M. F. Perutz, "An x-ray study of chymotrypsin and haemoglobin," *Nature*, vol. 141, pp. 523–524, 1938.

[7] J. Drenth, *Principles of Protein X-ray Crystallography*. Springer Verlag, 1994.

[8] K. Wuthrich, *NMR of Proteins and Nucleic Acids*. New York: John Wiley and Sons, 1986.

[9] A. Wand and S. Englander, "Protein complexes studied by nmr spectroscopy," *Curr. Opin. Biotechnol.*, vol. 7, no. 4, pp. 403–8, 1996.

TABLE II

THE COINCIDENCE BETWEEN OUR MOTIF PAIRS AND ACTIN-MOTIF BINDING PAIRS.

Actobindin Motif	Our Left Motif	Our Confirmed Right Motif in Actin
{v}{th}{v}{k}{k}{v}	{iv}{t}{iv}{k}	{a}{ek}{iv}{fl}{g}{kr}
{v}{th}{v}{k}{k}{v}	{iv}{ek}{k}{flv}{de}	{ek}{il}{l}{p}
{v}{th}{v}{k}{k}{v}	{ek}{iv}∅{ilv}{ek}	{g}{k}{k}{il}{v}{s}

TABLE III

THE COINCIDENCE BETWEEN OUR MOTIF PAIRS AND PEPTIDE-PROTEIN BINDING PAIRS.

Consensus Sequence	Left motif	Right Motif	Binding Protein
{p}∅{l}{p}∅{kr}	{p}{ek}∅{p}	{g}{v}{fi}{s}	CRK A
{rkh}{p}{p}{ailvp}{p}{ailvp}{k}{p}	{p}{iv}{ep}{iv}{a}	{a}{a}{s}{fi}	Cortactin
{r}{l}{p}∅{l}{p}	{p}{ek}∅{p}	{g}{v}{fi}{s}	Synaptojanin I
{rkh}{p}{p}{ailvp}{p}{ailvp}{k}{p}	{p}{iv}{dp}{p}{fv}	{p}{iv}{dp}{p}{fv}	Shank

[10] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert, “Approaches to the automatic discovery of patterns in biosequences,” *J. Comput. Biol.*, vol. 5, no. 2, pp. 279–305, 1998.

[11] F. Pazos and A. Valencia, “Similarity of phylogenetic trees as indicator of protein-protein interaction,” *Protein Engineering*, vol. 14, pp. 609–614, 2001.

[12] S. Pages, A. Belaich, J. Belaich, E. Morag, R. Lamed, Y. Shoham, and E. Bayer, “Species-specificity of the cohesin-dockerin interaction between clostridium thermocellum and clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain,” *Proteins*, vol. 29, no. 4, pp. 517–27, Dec. 1997.

[13] K. Fryxell, “The coevolution of gene family trees,” *Trends Genet.*, vol. 12, no. 9, pp. 364–9, Sep. 1996.

[14] A. B. Sparks, J. E. Rider, and et al., “Distinct ligand preferences of src homology 3 domains from src, yes, abl, cortactin, p53bp2, plcgamma, crk, and grb2,” *Proc. Natl. Acad. Sci.*, no. 4, pp. 1540–4, 1996.

[15] B. K. Kay, M. P. Williamson, and M. Sudol, “The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains,” *FASEB J.*, vol. 14, no. 2, pp. 231–41, 2000.

[16] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in *Proc. of the 20th Int’l Conference on Very Large Databases*, Santiago, Chile, Sep. 1994, pp. 487–499.

[17] F. Glaser, D. Steinberg, I. Vakser, and N. Ben-Tal, “Residue frequencies and pairing preferences at protein-protein interfaces,” *Proteins*, vol. 43, no. 2, pp. 89–102, 2001.

[18] D. Burdick, M. Calimlim, and J. Gehrke, “Mafia: A maximal frequent itemset algorithm for transactional databases,” in *ICDE*, Heidelberg, Germany, 2001, pp. 443–452.

[19] G. Grahne and J. Zhu, “Efficiently using prefix-trees in mining frequent itemsets,” in *(FIMI’03) Workshop on Frequent Itemset Mining Implementations*, Melbourne, FL, Nov 2003.

[20] S. Sheu, D. J. Lancia, K. Clodfelter, M. Landon, and S. Vajda, “Precise: a database of predicted and consensus interaction sites in enzymes,” *Nucleic Acids Res*, vol. 33, no. Database Issue, pp. D206–11, 2005.

[21] B. Doray and S. Kornfeld, “Gamma subunit of the ap-1 adaptor complex binds clathrin: implications for cooperative binding in coated vesicle assembly,” *Mol. Biol. Cell.*, vol. 12, no. 7, pp. 1925–35, 2001.

[22] H. Li and J. Li, “Discovery of stable and significant binding motif pairs from pdb complexes and protein interaction datasets,” *Bioinformatics*, vol. 21, no. 3, pp. 314–324, 2005.

[23] H. Li, J. Li, S. H. Tan, and S. K. Ng, “Discovery of binding motif pairs from protein complex structural data and protein interaction sequence data,” in *Proceeding of the Ninth Pacific Symposium on Biocomputing (PSB)*, USA, 2004, pp. 312–323.

[24] C. von Mering, R. Krause, and et al., “Comparative assessment of large-scale data sets of protein-protein interactions,” *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[25] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B. T. Chait, and R. Mackinnon, “X-ray structure of a voltage-dependent k+ channel,” *Nature*, vol. 423, no. 6935, pp. 33–41, 2003.

[26] P. Pellicena and W. Miller, “Processive phosphorylation of p130cas by src depends on sh3-polyproline interactions,” *J. Biol. Chem.*, vol. 276, no. 30, pp. 28 190–28 196, 2001.

[27] R. Russell, J. Breed, and G. Barton, “Conservation analysis and structure prediction of the sh2 family of phosphotyrosine binding domains,” *FEBS. Lett.*, vol. 304, no. 1, pp. 15–20, 1992.

[28] E. Azarya-Sprinzak, D. Naor, H. J. Wolfson, and R. Nussinov, “Interchanges of spatially neighbouring residues in structurally conserved environments,” *Protein Eng.*, vol. 10, no. 10, pp. 1109–22, 1997.

[29] K. Vancompernolle, J. Vandekerckhove, M. R. Bubb, and E. D. Korn, “The interfaces of actin and acanthamoeba actobindin. identification of a new actin-binding motif,” *J. Biol. Chem.*, vol. 266, no. 23, pp. 15 427–31, 1991.

Jinyan Li is Lead Scientist at the Institute for Infocomm Research, Singapore. He obtained the Ph.D. degree from the Department of Computer Science and Software Engineering, the University of Melbourne in 2001. His research interests include data mining, machine learning, and bioinformatics. He is currently working on fundamentals of emerging patterns, committees of decision trees, and biomarkers’ discovery for *in-silico* cancer diagnosis.

Haiquan Li is currently a PhD student in School of Computing, National University of Singapore, attached in Institute for Infocomm Research. He received his B.Eng. and M.Eng. degrees both in computer science from Huazhong University of Science and Technology, China, in 1996 and 1999. His research interests include bioinformatics and data mining. His current work is focused on the discovery of binding motif pairs, odds ratio patterns, generators and closed patterns.