

# Discovering Motif Pairs at Interaction Sites from Protein Sequences on a Proteome-Wide Scale

**Limsoon Wong**

**(joint work with Haiquan Li & Jinyan Li)**



# Plan

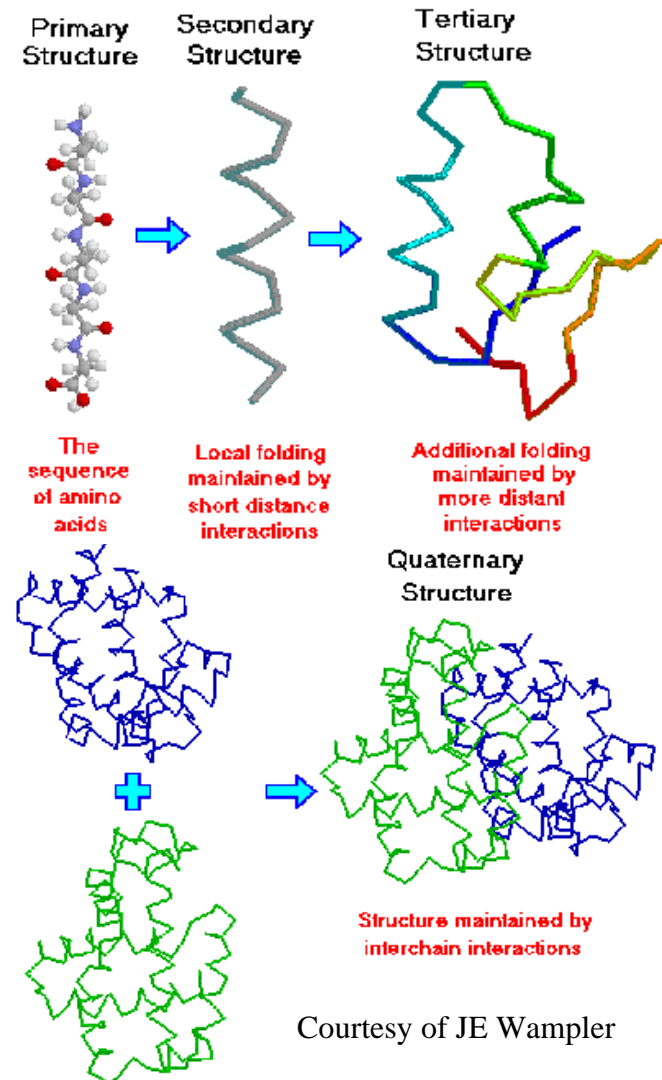
- **Problem statement:**
  - Discover binding motif pairs
- **Transform to a graph problem:**
  - Enumerate max complete bipartite subgraphs
- **Transform to a data mining problem:**
  - Mine closed patterns
- **Generate motifs from blocks**
- **Verify using known data**

# Problem Statement: Discover Binding Motif Pairs



# Proteins & Their Interactions

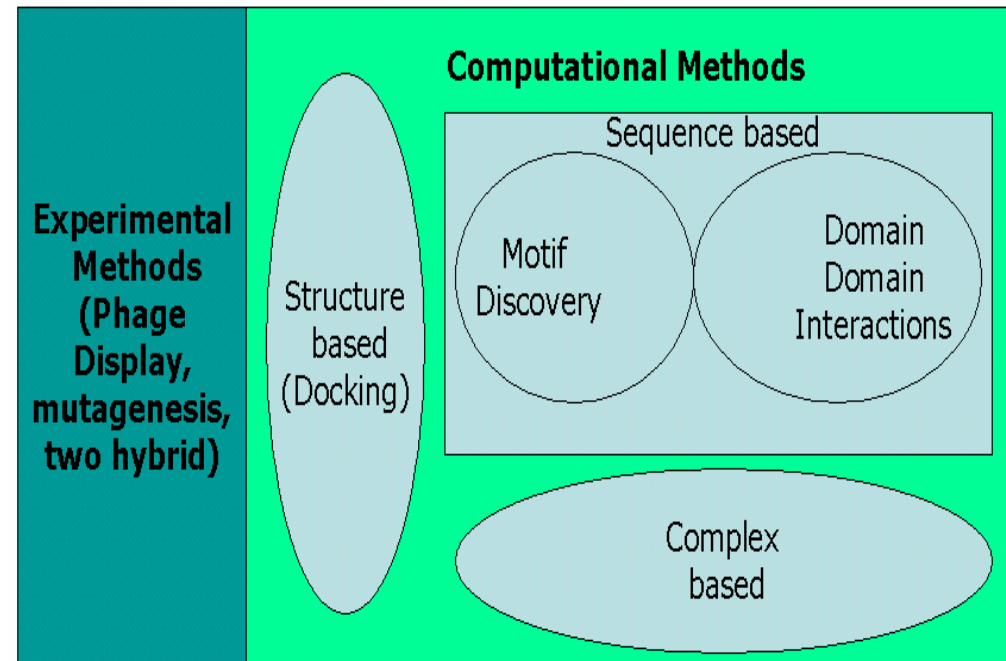
- 4 types of reps for proteins: primary, secondary, tertiary, & quaternary
- Protein interactions play impt role in inter cellular communication, in signal transduction, & in the regulation of gene expression



Courtesy of JE Wampler

# Binding Sites

- **Discovery of binding sites is a key part of understanding mechanisms of protein interactions**
- **Structure-based approaches**
  - E.g., docking
  - Relatively accurate
  - Struct must be known



⇒ **Sequence-based approaches**

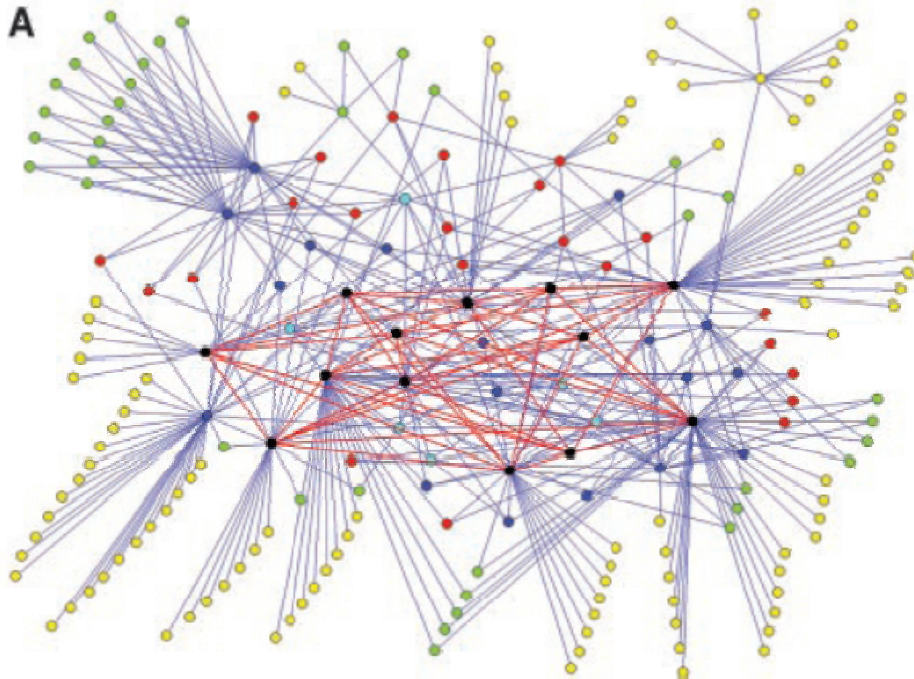
# Typical Sequence-Based Approach

- **Typical seq-based approaches have two steps:**
  - Use pattern discovery algorithms to discover domains and/or motifs of a group of proteins
  - Use domain-domain interaction discovery methods (e.g., domain fusion) to discovery interacting domains
- **Shortcomings:**
  - Protein interaction information is not used by motif discovery algorithms
  - Exact positions of binding sites often not recognized

## How about ...

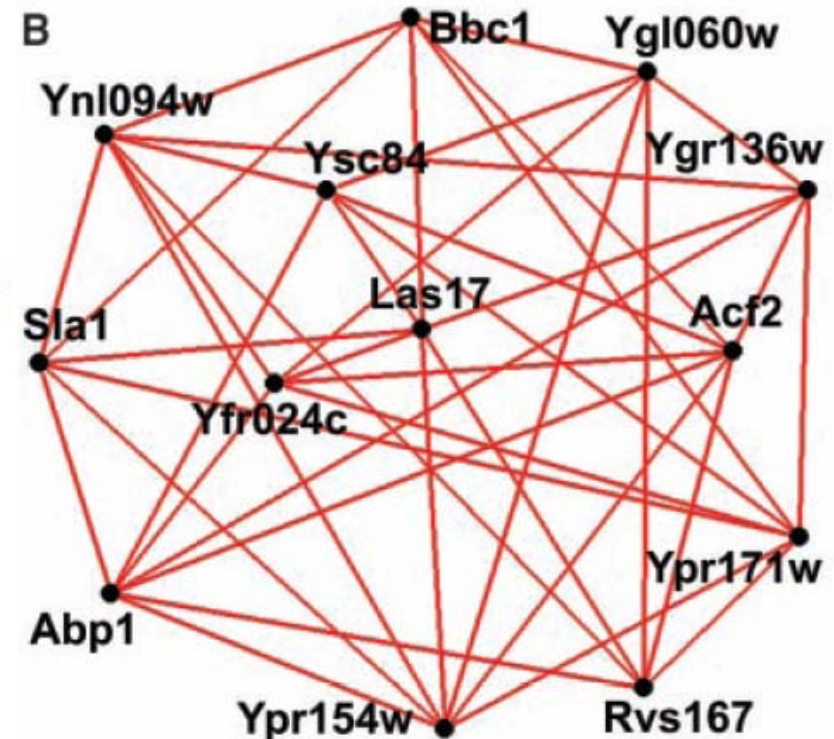
- **How about making use of known protein-protein bindings to guide the discovery of binding motifs?**

# Protein Interaction Graphs



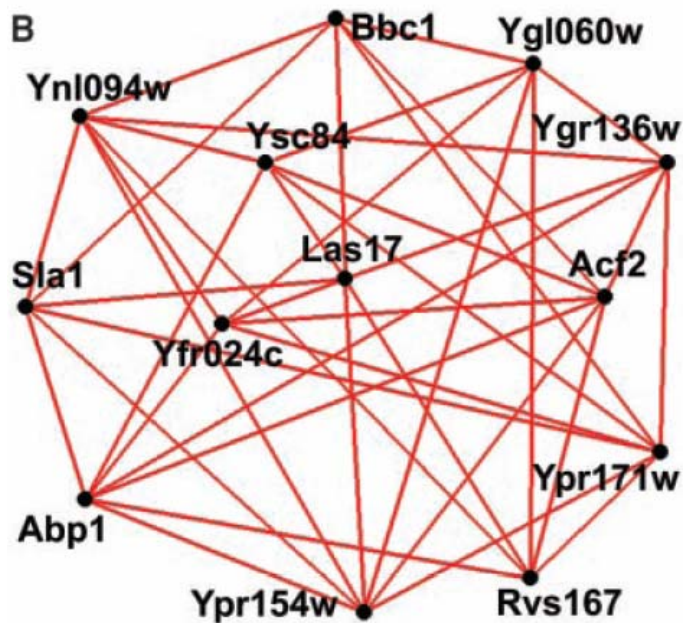
Yeast SH3 domain-domain  
Interaction network:  
394 edges, 206 nodes

Tong et al. *Science*, v295. 2002



8 proteins containing SH3  
5 binding at least 6 of them

# Bipartite Subgraphs



SH3 Proteins

Yfr024c ●

Yvs167 ●

Ysc84 ●

Ygr136w ●

Ypr154w ●

Bdc1 ●

SH3-Binding  
Proteins

● Las17

● Ypr171w

● Acf2

● Ynl094w

The larger this group,  
the more likely their  
active sites will show  
up clearly in a multiple  
alignment?

# Problem Statement

**Given a PPI expt E, the problem is**

**(1) To find all pairs X, Y of interacting protein groups,  
so that**

**(1.1) X and Y have full mutual interactions**

**(1.2) X and Y are as large as possible**

**&**

**(2) To identify “good” binding motif pairs from  
these pairs of interacting protein groups**

# Transform to a Graph Problem: Enumerate Max Complete Bipartite Subgraphs



## PPI Expt As a Graph

- PPI expt **E** as undirected graph  $G^E = \langle V^E, D^E \rangle$ ,
  - where  $V^E$  are the proteins and  $D^E$  the edges,
  - so that two proteins are connected in  $G^E$  iff there is a binding betw them in PPI expt **E**
- Let  $\beta^E(p)$  denote **neighborhood** of protein  $p$  in  $G^E$
- Let  $\beta^E(P) = \bigcap_{p \in P} \beta^E(p)$  denote the **common neighborhood** of all proteins in  $P$  in  $G^E$

# Maximality

- **Proposition 2.1**

Let  $E$  be a PPI expt.

Let  $X, Y$  be a pair of protein groups so that

$$X = \beta^E(Y) \text{ and } Y = \beta^E(X).$$

Let  $X', Y'$  be another pair of protein groups so that

$$X' = \beta^E(Y'), Y' = \beta^E(X'), X' \subseteq X, \text{ \& } Y' \subseteq Y.$$

Then  $X = X'$  and  $Y = Y'$ .

⇒ In other words, if  $X = \beta^E(Y)$  and  $Y = \beta^E(X)$ , then  $X, Y$  is a maximal pair of protein groups that have full interactions

## Problem Statement

Given a PPI expt  $E$ , the problem is

(1) To find all pairs  $X, Y$  of interacting protein groups, so that

(1.1)  $X$  and  $Y$  have full mutual interactions

(1.2)  $X$  and  $Y$  are as large as possible

&

(2) To identify “good” binding motif pairs from these pairs of interacting protein groups

## Maximality

### Proposition 2.1

1.1

Let  $E$  be a PPI expt.

Let  $X, Y$  be a pair of protein groups so that

$X = \beta^E(Y)$  and  $Y = \beta^E(X)$ .

1.2

Let  $X', Y'$  be another pair of protein groups so that

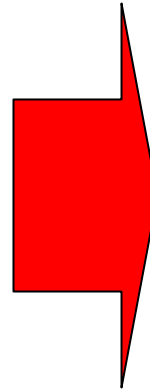
$X' = \beta^E(Y')$ ,  $Y' = \beta^E(X')$ ,  $X' \subseteq X$ , &  $Y' \subseteq Y$ .

Then  $X = X'$  and  $Y = Y'$ .

$\Rightarrow$  In other words, if  $X = \beta^E(Y)$  and  $Y = \beta^E(X)$ , then  $X, Y$  is a maximal pair of protein groups that have full interactions

# Recasting to Graph Theory

- $X, Y$  is a pair of interacting protein groups in PPI expt  $E$  iff  $X = \beta^E(Y)$  and  $Y = \beta^E(X)$



# Max Complete Bipartite Subgraph



- A graph  $H = \langle V_1 \cup V_2, D^H \rangle$  is a maximal complete bipartite subgraph of  $G$  iff
  - $H$  is a subgraph of  $G$ ,
  - $V_1 \times V_2 = D^H$ ,
  - $V_1 \cap V_2 = \{\}$ , &
  - There is no  $H' = \langle V'_1 \cup V'_2, D^{H'} \rangle$  with  $V_1 \subset V'_1$  &  $V_2 \subset V'_2$  that has the same properties above

## Max Complete Bipartite Subgraph

- A graph  $H = \langle V_1 \cup V_2, D^H \rangle$  is a **maximal complete bipartite subgraph** of  $G$  iff
  - $H$  is a subgraph of  $G$ ,
  - $V_1 \times V_2 = D^H$ ,
  - $V_1 \cap V_2 = \{\}$ , &
  - There is no  $H' = \langle V'_1 \cup V'_2, D^{H'} \rangle$  with  $V_1 \subset V'_1$  &  $V_2 \subset V'_2$  that has the same properties above

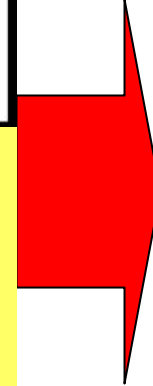
Copyright © 2005 by Limsoon Wong.



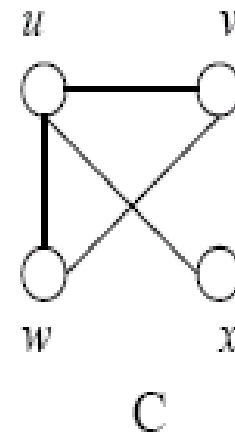
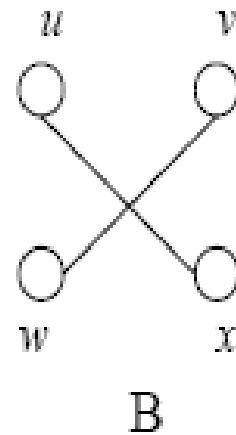
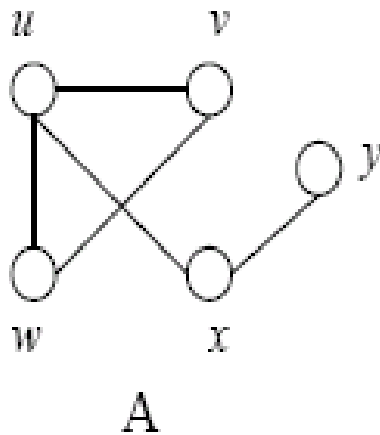
## Connection to Graph Theory

- $X, Y$  is a pair of interacting protein groups in PPI expt  $E$  iff  $H = \langle X \cup Y, X \times Y \rangle$  is max complete bipartite subgraph of  $G^E$

- Let  $H = \langle X \cup Y, D^E|_{X \cup Y} \rangle$  be subgraph of  $G^E$  with  $X, Y$  a pair of interacting protein groups
  - $\Rightarrow X = L^E(Y)$  and  $Y = L^E(X)$
  - $\Rightarrow$  Full interactions betw  $X$  &  $Y$
  - $\Rightarrow X \times Y = D^E|_{X \cup Y}$
  - By excluding self-binding, we have  $X \cap Y = \{\}$
  - By Prop 2.1, we have  $H$  is max



We are talking about subgraphs, not vertex-induced subgraph .....



- B is a subgraph of A, but it is not a vertex-induced subgraph
- C is a subgraph of A, and it is a vertex-induced subgraph

## Therefore ... But ...

- Therefore, to find pairs of interacting protein groups, we can use algorithms from graph theory for enumerating maximal complete bipartite subgraphs
- According to Eppstein 1994, this has complexity  $O(a^3 2^{2a} n)$ , where “a” is the average degree of the graph and “n” the number of vertices
- This is inefficient because “a” is often around 10-20 in practice

# Transform to a Data Mining Problem: Mine Closed Patterns



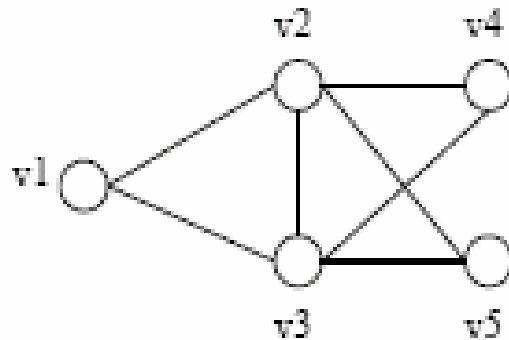
## From PPI Expts To Transactions

- In PPI expt  $E$ , we obtain for each protein  $p$ , a list  $\beta^E(p)$  of proteins that bind  $p$ 
  - assume  $p \notin \beta^E(p)$ , as such expts are not intended to detect self-binding
  - assume  $q \in \beta^E(p)$  implies  $p \in \beta^E(q)$ , as binding is symmetric
- $\beta^E(p)$  can be thought of as a transaction &  $p$  as the “id” of this transaction

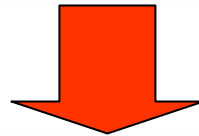
$\Rightarrow$   $E$  can be thought of as generating a db of transactions  $D^E = \{\beta^E(p_1), \dots, \beta^E(p_k)\}$ , where  $p_1, \dots, p_k$  are all the proteins involved in  $E$

$\Rightarrow$  a set of proteins  $X$  can be thought of as a pattern in  $D^E$  if there is  $p \in D^E$  st  $X \subseteq \beta^E(p)$

# Example



A graph  $G$



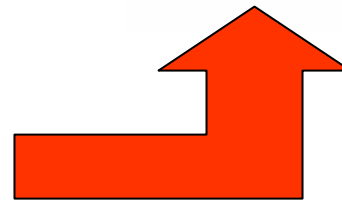
	$v_1$	$v_2$	$v_3$	$v_4$	$v_5$
$v_1$	0	1	1	0	0
$v_2$	1	0	1	1	1
$v_3$	1	1	0	1	1
$v_4$	0	1	1	0	0
$v_5$	0	1	1	0	0

its adjacency matrix

We use the protein  $v$  to be  $id(\beta(v))$

$id(T)$	$T$	items				
		$v_1$	$v_2$	$v_3$	$v_4$	$v_5$
$v_1$	$\beta(v_1)$	0	1	1	0	0
$v_2$	$\beta(v_2)$	1	0	1	1	1
$v_3$	$\beta(v_3)$	1	1	0	1	1
$v_4$	$\beta(v_4)$	0	1	1	0	0
$v_5$	$\beta(v_5)$	0	1	1	0	0

transformation to  $DB_G$



# Occurrence Set

- The **occurrence set** of a pattern  $P$  in database  $D$  is defined as

$$\begin{aligned} \text{occ}^D(P) &= \{ \text{id}(T) \mid T \in D, P \subseteq T \} \\ &= \{ \text{id}(T) \mid T \in f^D(P) \} \end{aligned}$$

- **Proposition**

$$\text{occ}^{D^E}(P) = \beta^E(P)$$

# Closed Patterns

- **Let**
    - $I$  be a set of items and  $D$  a transaction db on  $I$
    - $f^D(P) = \{T \in D \mid P \subseteq T\}$ ,
    - $g(D') = \bigcap_{T \in D'} T = \bigcap D'$ , for  $D' \subseteq D$
- Then  $CL^D(P) = g(f^D(P))$  is the closure of  $P$ ,  
 and  $P$  is called a **closed pattern** iff  $P = CL^D(P)$

- **Proposition**  
 **$CL^D$  is a closure operation. That is,  $CL^D$  is monotonic, idempotent, and inflationary**

- **Proposition**  

$$CL^{D^E}(P) = \beta^E(\beta^E(P))$$

# Proof of $CL(P) = \beta(\beta(P))$

- $\beta(\beta(P))$ 
  - $= \beta(\mathbf{occ}(P))$ , since  $\beta(P) = \mathbf{occ}(P)$
  - $= \bigcap_{\mathbf{id}(T) \in \mathbf{occ}(P)} \beta(\mathbf{id}(T))$ , defn of  $\beta(\cdot)$
  - $= \bigcap_{\mathbf{id}(T) \in \mathbf{occ}(P)} \mathbf{T}$ , defn of  $\mathbf{id}(\cdot)$  and  $\beta(\cdot)$
  - $= \bigcap_{T \in f(P)} \mathbf{T}$ , since  $\mathbf{occ}(P) = \{\mathbf{id}(T) \mid T \in f^D(P)\}$
  - $= \mathbf{g}(f(P))$ , defn of  $\mathbf{g}(\cdot)$
  - $= \mathbf{CL}(P)$ , defn of  $\mathbf{CL}(\cdot)$

- **Proposition**

Let  $C_1, C_2$  be closed patterns in  $D^E$

Then  $C_1 = C_2$  iff  $\text{occ}(C_1) = \text{occ}(C_2)$

- **Proposition**

If  $C$  is closed pattern in  $D^E$ , then  $C \cap \text{occ}(C) = \{ \}$

- **Proposition**

Let  $C$  be a closed pattern in  $D^E$ .

Then  $\text{occ}(C)$  is a closed pattern in  $D^E$

- **Corollary**

The number of closed patterns in  $D^E$  is even

# Proofs

**Proposition 7** *Let  $G$  be a graph. Let  $C_1$  and  $C_2$  be two closed patterns of  $DB_G$ . Then  $C_1 = C_2$  iff  $occ^{DB_G}(C_1) = occ^{DB_G}(C_2)$ .*

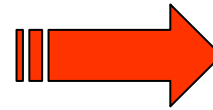
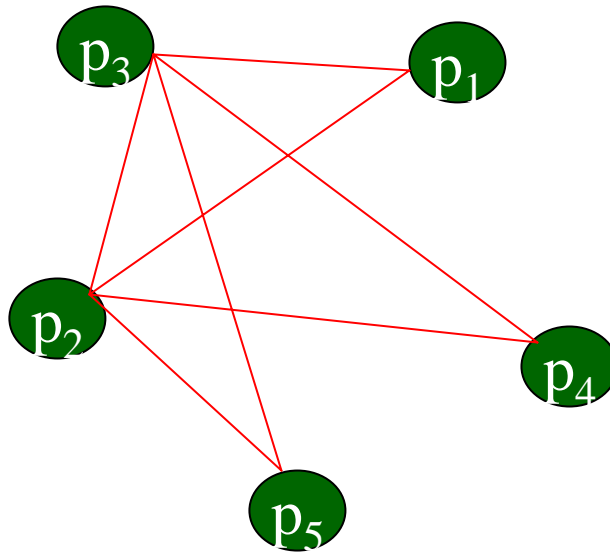
**Proof:** *The left-to-right direction is trivial. To prove the right-to-left direction, let us suppose that  $occ(C_1) = occ(C_2)$ . It is straightforward to see that  $id(T) \in occ(P)$  iff  $T \in f(P)$ . Then we get  $f(C_1) = f(C_2)$  from  $occ(C_1) = occ(C_2)$ . Since  $C_1$  and  $C_2$  are closed patterns of  $DB_G$ , it follows that  $C_1 = g(f(C_1)) = g(f(C_2)) = C_2$ , and finishes the proof.  $\square$*

**Lemma 1** *Let  $G$  be a graph. Let  $C$  be a closed pattern of  $DB_G$ . Then  $f^{DB_G}(occ^{DB_G}(C)) = \{\beta^G(c) \mid c \in C\}$ .*

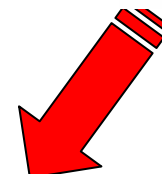
**Proposition 9** *Let  $G$  be a graph and  $C$  a closed pattern of  $DB_G$ . Then  $occ^{DB_G}(C)$  is also a closed pattern of  $DB_G$ .*

**Proof:** *By Lemma 1,  $f(occ(C)) = \{\beta(c) \mid c \in C\}$ . So  $CL(occ(C)) = g(f(occ(C))) = \bigcap f(occ(C)) = \bigcap_{c \in C} \beta(c) = \beta(C)$ . By Proposition 5,  $\beta(C) = occ(C)$ . Thus  $occ(C)$  is a closed pattern.  $\square$*

# Example



	$v_1$	$v_2$	$v_3$	$v_4$	$v_5$
$\beta(v_1)$	0	1	1	0	0
$\beta(v_2)$	1	0	1	1	1
$\beta(v_3)$	1	1	0	1	1
$\beta(v_4)$	0	1	1	0	0
$\beta(v_5)$	0	1	1	0	0



$sup(X)$	<i>closed pattern X</i>	$Y = occ(X)$	$sup(Y)$
3	$\{v_2, v_3\}$	$\{v_1, v_4, v_5\}$	2
4	$\{v_2\}$	$\{v_1, v_3, v_4, v_5\}$	1
4	$\{v_3\}$	$\{v_1, v_2, v_4, v_5\}$	1

# Isomorphism

- **Theorem**

Let  $G^E$  be PPI graph, and  $C$  closed pattern of  $D^E$ . Then  $H = \langle C \cup \text{occ}(C), C \times \text{occ}(C) \rangle$  is a max complete bipartite subgraph of  $G^E$

- **Theorem**

Let  $H = \langle V_1 \cup V_2, E' \rangle$  be max complete bipartite subgraph of  $G^E$ . Then  $V_1, V_2$  are closed pattern of  $D^E$ ,  $\text{occ}(V_1) = V_2$ , and  $\text{occ}(V_2) = V_1$

⇒ **An isomorphism exists betw max complete bipartite subgraphs of  $G^E$  & closed patterns of  $D^E$**

# Proofs



**Theorem 1** *Let  $G$  be an undirected graph without self-loop. Let  $C$  be a closed pattern of  $DB_G$ . Then the graph*

$$H = \langle C \cup \text{occ}^{DB_G}(C), C \times \text{occ}^{DB_G}(C) \rangle$$

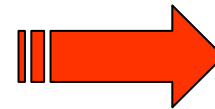
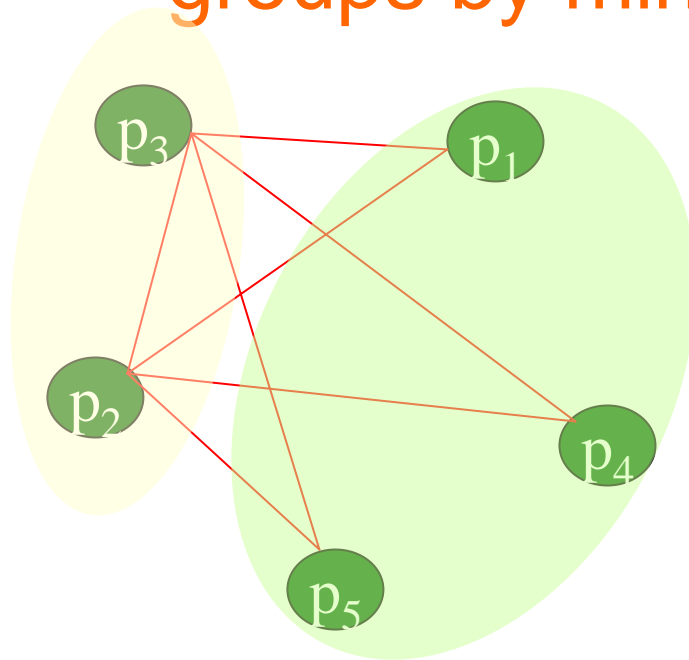
*is a maximal complete bipartite subgraph of  $G$ .*

**Proof:** *By assumption,  $C$  is non-empty and  $C$  has a non-zero support in  $DB_G$ . Therefore,  $\text{occ}(C)$  is non-empty. By Proposition 8,  $C \cap \text{occ}^{DB_G}(C) = \{\}$ . Furthermore, for every  $v \in \text{occ}(C)$ ,  $v$  is adjacent in  $G$  to every vertex of  $C$ . So,  $C \times \text{occ}(C) \subseteq E^G$ , and every edge of  $H$  connects a vertex of  $C$  and a vertex of  $\text{occ}(C)$ . Thus,  $H$  is a complete bipartite subgraph of  $G$ . By Proposition 5, we have  $\text{occ}^{DB_G}(C) = \beta^G(C)$ . By Proposition 6,  $C = \beta^G(\beta^G(C))$ . By Proposition 5, we derive  $C = \beta^G(\text{occ}^{DB_G}(C))$ . So  $H$  is maximal. This finishes the proof.  $\square$*

**Theorem 2** *Let  $G$  be an undirected graph without self-loop. Let graph  $H = \langle V_1 \cup V_2, E \rangle$  be a maximal complete bipartite subgraph of  $G$ . Then,  $V_1$  and  $V_2$  are both a closed pattern of  $DB_G$ ,  $\text{occ}^{DB_G}(V_1) = V_2$  and  $\text{occ}^{DB_G}(V_2) = V_1$ .*

**Proof:** *Since  $H$  is a maximal complete bipartite subgraph of  $G$ , then  $\beta(V_1) = V_2$  and  $\beta(V_2) = V_1$ . By Proposition 6,  $CL(V_1) = \beta(\beta(V_1)) = \beta(V_2) = V_1$ . So,  $V_1$  is a closed pattern. Similarly, we can get  $V_2$  is a closed pattern. By Proposition 5,  $\text{occ}(V_1) = \beta(V_1) = V_2$  and  $\text{occ}(V_2) = \beta(V_2) = V_1$ , as required.  $\square$*

Thus, can mine protein interaction groups by mining close patterns



	$v_1$	$v_2$	$v_3$	$v_4$	$v_5$
$\beta(v_1)$	0	1	1	0	0
$\beta(v_2)$	1	0	1	1	1
$\beta(v_3)$	1	1	0	1	1
$\beta(v_4)$	0	1	1	0	0
$\beta(v_5)$	0	1	1	0	0



$sup(X)$	closed pattern $X$	$Y = occ(X)$	$sup(Y)$
3	$\{v_2, v_3\}$	$\{v_1, v_4, v_5\}$	2
4	$\{v_2\}$	$\{v_1, v_3, v_4, v_5\}$	1
4	$\{v_3\}$	$\{v_1, v_2, v_4, v_5\}$	1

## An Extension

- **Not all interacting protein groups  $X, Y$  are equally interesting**
    - $X$  and  $Y$  are both singleton, vs
    - $X$  is a large group,  $Y$  is small group, vs
    - $X$  is a large group,  $Y$  is a large group
- ⇒ **Set “interestingness” threshold on  $X, Y$  st a pair of interacting protein groups  $X, Y$  is interesting only if  $|X| \geq p$  and  $|Y| \geq q$**

## An Optimization

- A max complete bipartite subgraph  $H = \langle V_1 \cup V_2, E' \rangle$  is **(p,q)-large** if  $|V_1|$  or  $|V_2|$  is at least p, and the other is at least q

- **Theorem**

Let  $G^E$  be PPI graph, and C closed pattern of  $D^E$ .

Then  $H = \langle C \cup \text{occ}(C), C \times \text{occ}(C) \rangle$  is (p,q)-large iff C occurs at least p times in  $D^E$  and  $|C| \geq q$

- $\Rightarrow$  To mine interesting pairs X, Y of interacting protein group in expt E st  $|X| \geq p$  and  $|Y| \geq q$ , it suffices to mine closed patterns X that appears  $\geq q$  times in  $D^E$  and  $|X| \geq p$

# Proofs

**Corollary 3** *Let  $G$  be a graph. Then  $H = \langle C \cup \text{occ}^{DB_G}(C), C \times \text{occ}^{DB_G}(C) \rangle$  is a  $(p, q)$ -large maximal complete bipartite subgraph of  $G$  iff  $C$  is a closed pattern such that  $C$  occurs at least  $p$  times in  $DB_G$  and  $\text{occ}^{DB_G}(C)$  occur at least  $q$  times in  $DB_G$ .*

**Theorem 5** *Let  $G$  be a graph. Then  $H = \langle C \cup \text{occ}^{DB_G}(C), C \times \text{occ}^{DB_G}(C) \rangle$  is a  $(p, q)$ -large maximal complete bipartite subgraph of  $G$  iff  $C$  is a closed pattern such that  $C$  occurs at least  $p$  times in  $DB_G$  and  $|C| \geq q$ .*

**Proof:** *Suppose  $H = \langle C \cup \text{occ}^{DB_G}(C), C \times \text{occ}^{DB_G}(C) \rangle$  is a  $(p, q)$ -large maximal complete bipartite subgraph of  $G$ . By Theorem 2,  $C = \text{occ}(\text{occ}(C))$ . By definition of  $\text{occ}(\cdot)$ ,  $\text{sup}(\text{occ}(C)) = |\text{occ}(\text{occ}(C))| = |C|$ . Substitute this into Corollary 3, we get  $H$  is a  $(p, q)$ -large maximal complete bipartite subgraph of  $G$  iff  $C$  is a closed pattern such that  $C$  occurs at least  $p$  times in  $DB_G$  and  $|C| \geq q$  as desired.  $\square$*

# Closed Patterns

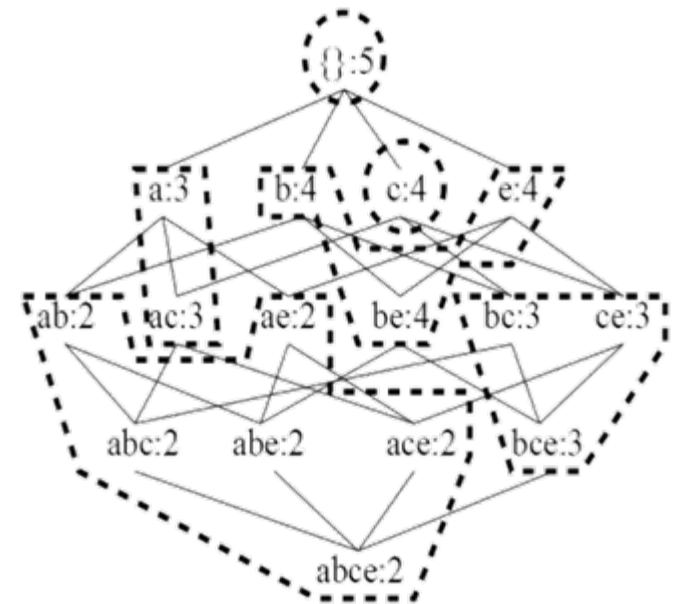
- Let  $[X]^D = \{Y \mid f^D(Y) = f^D(X)\}$  denote the equivalence class of the pattern  $X$  in  $D$
- Then  $\max [X]^D = \{CL^D(X)\}$

⇒ A closed pattern is the most specific pattern in its equivalence class

⇒ To mine patterns, it is sufficient & more efficient to mine just the closed patterns

Table 1: A transaction database  $T$

Transaction-id	Items
$T_1$	$a, c, d$
$T_2$	$b, c, e$
$T_3$	$a, b, c, e, f$
$T_4$	$b, e$
$T_5$	$a, b, c, e$



Support threshold = 2

# Closed Pattern Mining Algorithms

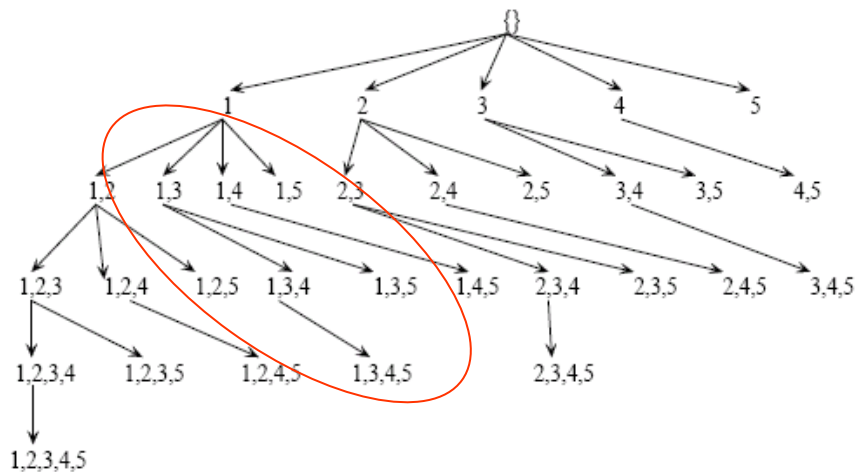
- **CLOSET, Pei et al. 2000**
- **CARPENTER, Pan et al. 2003**
- **FPclose\*, Grahne & Zhu 2003**
- **LCM, Uno et al., 2004**
- **GC-growth, Li et al. 2005**
- ...

⇒ **We have efficient algo for mining closed patterns**

- **But these algo have size constraint only on one side---occ(C), and do not pair up closed patterns**

# Pruning Small Max Complete Bipartite Subgraphs

- Search space of typical closed pattern mining algo



- An order is assumed on items
- Only items after last item in  $X$  can appear in sub search space of  $X$
- E.g.,
  - 4 is in  $\text{tail}\{1, 3\}$
  - 2 is not in  $\text{tail}\{1,3\}$

- To find closed pattern  $Y$  st  $|Y| \geq q$  &  $\text{sup}(Y) \geq p$

- Itemset  $Y$  in sub search space of  $X$  is subset of  $X \cup \text{tail}(X)$

$\Rightarrow$  Skip if  $|X \cup \text{tail}(X)| < q$

- Itemset  $Y$  in sub search space of  $X$  st  $|Y| \geq q$  &  $\text{sup}(Y) \geq p$  is subset of  $X \cup \{x \in \text{tail}(X) \mid \text{sup}(X \cup \{x\}) \geq p\}$

$\Rightarrow$  Skip if there is less than  $p - |X|$  items  $x \in \text{tail}(X)$  st  $\text{sup}(X \cup \{x\}) \geq p$

# Pruning Duplicate Max Complete Bipartite Subgraphs

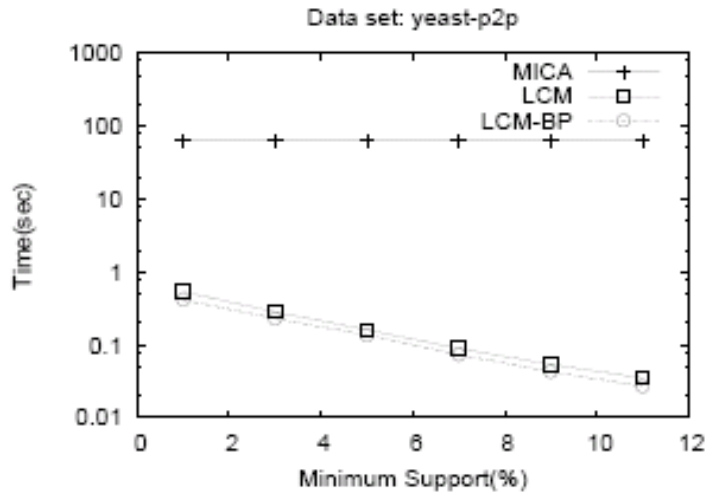
- Max complete bipartite subgraphs are generated twice if we output all closed patterns
- Set sup threshold to  $\max(p, q)$ 
  - ⇒ **Maximize pruning power**
- Do not extend if size of closed pattern exceeds its support
  - ⇒ **Max complete bipartite subgraphs w/ vertex sets of diff sizes enumerated only once**
- Do not output closed pattern if smaller than its occurrence set
  - ⇒ **Max complete bipartite subgraphs w/ vertex sets of same size enumerated only once**

MICA: a previous consensus-based max complete bipartite mining algo

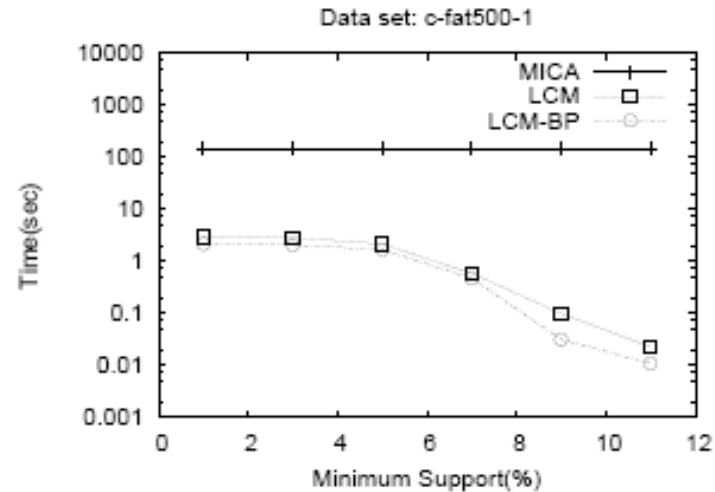
LCM: state-of-art closed pattern mining algo

LCM-BP: our modified LCM

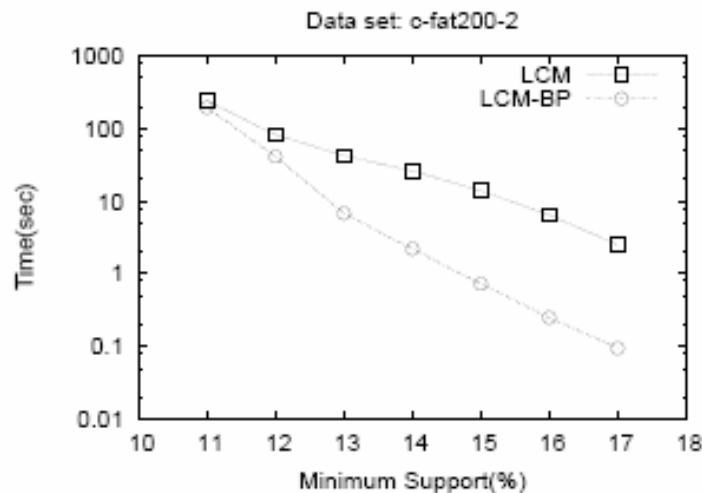
# Effectiveness



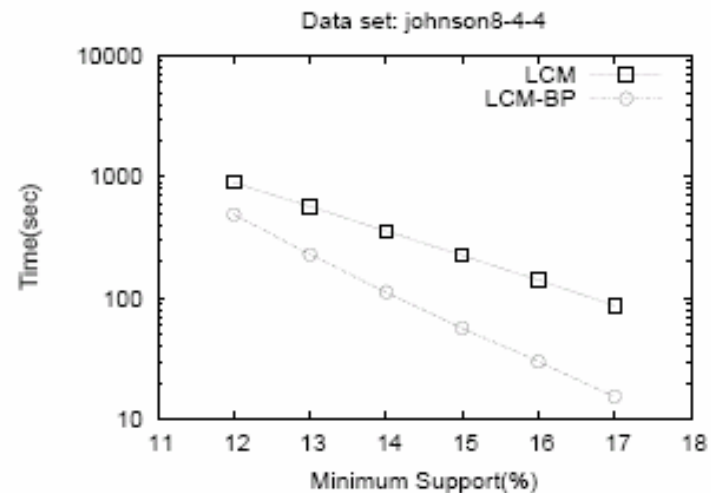
(a) yeast-p2p



(b) c-fat500-1



(e) c-fat200-2



(f) johnson8-4-4

# Experiment

- So let's use LCM-BP to mine interesting protein interaction groups ...
- Consider the yeast PPI graph from Breitkreutz et al, *Genome Biology*, 4, R23, 2003
  - 4904 vertices
  - 17440 edges (after removing 185 self-loops, 1413 redundant edges)
  - Ave number of interactions per protein = 3.56

# Resulting Max Complete Bipartite Subgraphs

support threshold ( <i>ms</i> )	# of frequent closed patterns	# of qualified closed patterns	time in sec.
1	121314	121314	0.59
2	117895	114554	0.50
3	105854	95920	0.44
4	94781	80306	0.40
5	81708	60038	0.36
6	66429	36478	0.32
7	50506	15800	0.25
8	36223	3716	0.21
9	25147	406	0.11
10	17426	34	0.07
11	12402	2	0.06
12	9138	0	0.05

# Generate Motifs From Blocks , Verify Binding Motif Pairs



# Many Motif Discovery Methods

- **MEME**, Bailey & Elkan 1995
  - **CONSENSUS**, Hertz & Stormo 1995
  - **PROTOMAT**, Henikoff & Henikoff 1991
  - **CLUSTAL**, Higgins & Sharp 1988
  - ...
- 
- **For illustration, we use PROTOMAT here**

# PROTOMAT

- **Core of Block Maker, a WWW server that return blocks (ungapped multiple alignments) for any submitted set of protein sequences**
- **Comprises 2 steps:**
  - MOTIF, Smith et al. 1990
    - **Look for spaced triplets in given set of proteins**
  - MOTOMAT, Henikoff & Henikoff 1991
    - **Merge overlapping blocks produced by MOTIF**
    - **Extend blocks in both directions until similarity falls**
    - **Determine best set of blocks that are in the same order and do not overlap**

we treat every block, instead of whole set of blocks generated by PROTOMAT, as a binding motif

## Example, Breitkreutz et al, Genome Biology, 4, R23, 2003

- Comprises 17440 genetic and physical interactions in yeast among 4904 proteins
- Look for interesting pairs with  $p = q = 5$
- $<1s$  to generate 60k closed patterns
- ⇒ Too many for PROTOMAT. So consider only maximal closed patterns, giving 7847 pairs
- PROTOMAT produces 17256 left blocks and 19350 right blocks after 6 hours
- Most groups yield 1 to 3 blocks
- Ave length of blocks = 11.696, std dev = 5.45

## Databases Used for Validation

- **BLOCKS**, Pietrokovski et al. 1996
- **PRINTS**, Attwood & Beck 1994
- **Pfam**, Sonnhammer et al. 1997
- **InterDom**, Ng et al. 2003

	BLOCKS	PRINTS	Pfam	InterDom
Version	14.0	37.0	16.0	1.1
Num. of groups / families	4944	1850	7677	3535
Num. of entries	24294	11170	7677	30037

# Validation for Single Motifs

- **Compare all single motifs in our discovered motif pairs with all domains of specific domain db**
  - LAMA, Pietrokovski 1996
  - transform blocks into position-specific scoring matrices (PSSM)
  - run Smith-Waterman to align pairs of PSSM using Pearson correlation coefficient to measure similarity betw 2 columns
  - a block is mapped to another block if 95% of positions in a block occurring in the optimal alignment is common to another block and Z-score is  $> 5.6$ , where Z-score is the number std dev away from the mean generated by millions of shuffles of the BLOCKS database
- **Determine number of motifs that can be mapped to these domains and the overall correlation in the portions that are mapped**

## Results for Single Motifs

	Mapped / total num. in BLOCKS	Mapped / total num. in PRINTS	Mapped / total num. in BOTH
Unique blocks	8401 / 24294	2872/ 11170	11273/ 35464
Unique groups	3568 / 4944	1325/ 1850	4893 / 6794

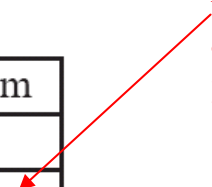
- **Our blocks map to 32% of blocks in BLOCKS and PRINTS, yet motifs from our blocks cover 72% of domains in BLOCKS and PRINTS**
- ⇒ **Maybe most domains in BLOCKS and PRINTS have less than half a block as binding motifs, or may not be related to binding behaviour**

## Validation for Motif Pairs

- Map our motif pairs into domain-domain interacting pairs
- Determine the number of overlaps between our motif pairs and those in the domain-domain interaction database
- Use InterDom as the domain-domain interaction database

	BLOCKS	PRINTS	Pfam	InterDom
Version	14.0	37.0	16.0	1.1
Num. of groups / families	4944	1850	7677	3535
Num. of entries	24294	11170	7677	30037

30037  
interactions  
among  
3535 domains



# Linking Our Motif Pairs to InterDom

- **InterDom represents domains by Pfam entries**
- ⇒ **To x-link, we have to**
  - Map our motifs to blocks in BLOCKS and PRINTS
  - Link from BLOCKS and PRINTS to InterPro
  - Link from InterPro to Pfam
  - Match Pfam to InterDom

# Results for Motif Pairs

Domain-domain interactions  
 inferred from protein complexes  
 or from interactions between  
 single domain proteins

	BLOCKS overlaps	PRINTS overlaps	Combined overlaps	Confident overlaps	Complex confirmed
Domain pairs	862	26	1163	396	241

Both sides  
 mapped to BLOCKS

Both sides  
 mapped to PRINTS

One side mapped to PRINTS,  
 one side mapped to BLOCKS

## Example Confirmed Binding Motif

- 1 of the 241 binding motifs we found that can be confirmed using protein complexes is #1781...

ID none; BLOCK

AC 1781xxxxxx; distance from previous block=(26,378)

DE none BL GNL motif=[5,0,17] motomat=[1,80,-10] width=14 seqs=6

YBL026W (27) GTLQSV DQF LNLKL  
 YCR077C (379) GNSS QDNKQ ANTVL  
 YER112W (27) GI LTNVDNWMNLT L  
 YER146W (32) GTLVGFDDF VNVIL  
 YNL147W (42) GVLKGYDQL MNLVL  
 YOL149W (129) GKTL SGKDI YNYGL

gdb1mgq\_A (38) GVLKSF D1 h MNLVL

ID none; BLOCK

AC 1781xrihg; distance from previous block=(2,316)

DE none BL LDN motif=[4,0,17] motomat=[1,80,-10] width=9 seqs=4

YDR378C (75) LESIDGFMN  
 YGL173C (317) LLHTDGYI N  
 YJL124C (68) LRTFDQYA N  
 YJR022W (46) LNGFDKNT N

pdb1mgq\_B (40) Lk SFD1 hMN

As shown in the next slide, this pair corresponds to interaction sites between LSM domains. E.g., all 7 pairs of adjacent LSM domains of pdb1mgq exhibits it.

# Example: LSM Domains of pdb1mgq

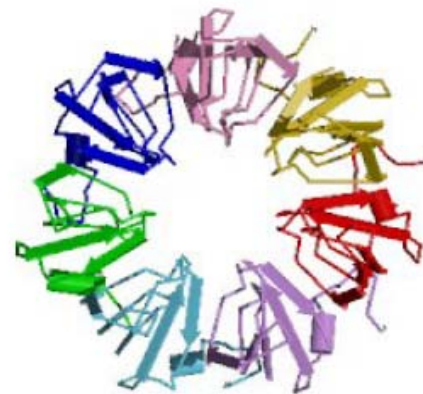


Fig. 2. The structure of the complex pdb1mgq consisting of 7 LSM domains corresponding to chain A, Chain B, ..., to chain G.

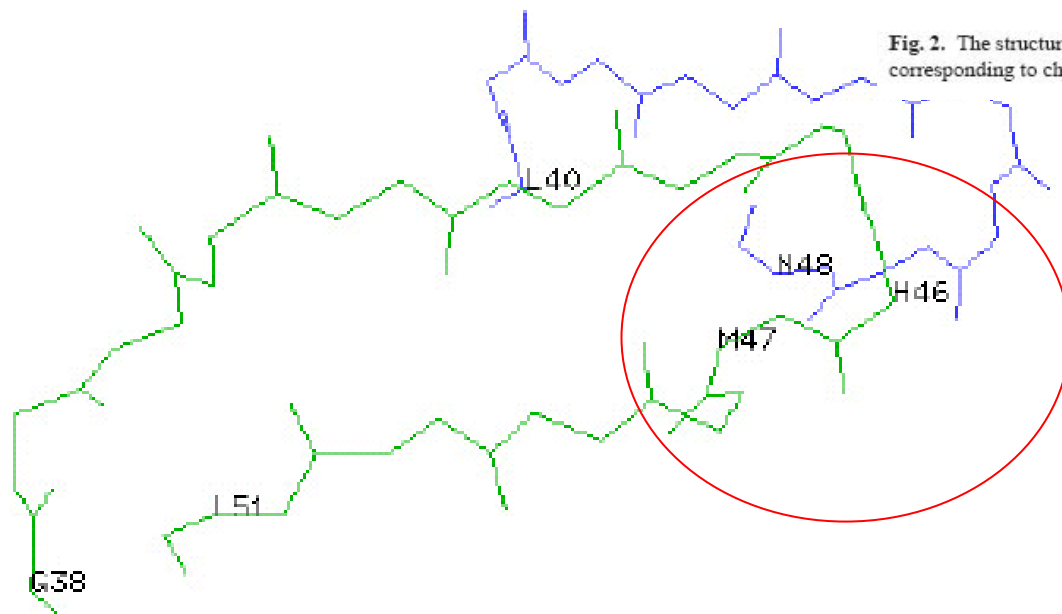
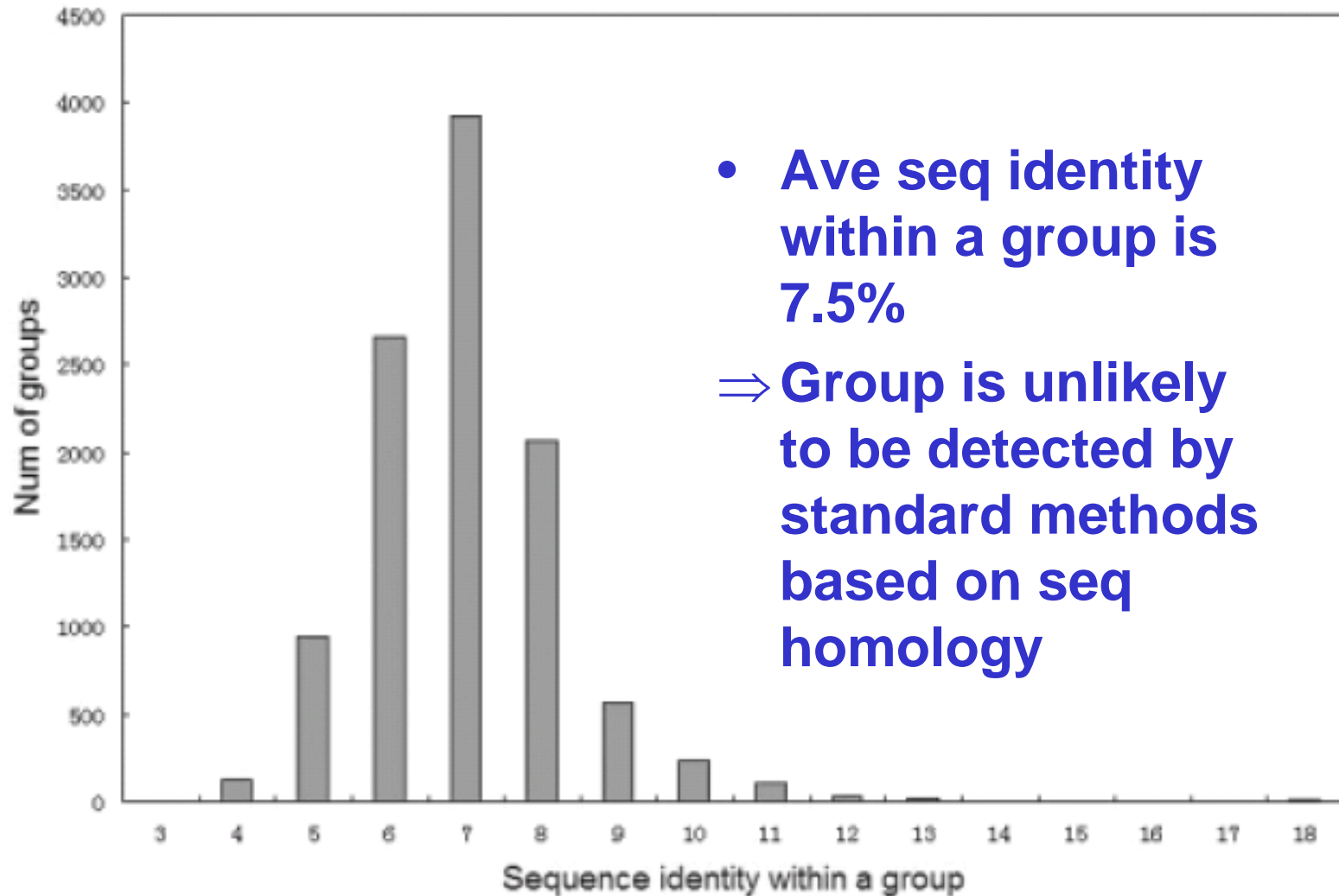


Fig. 3. Interactions between segment [38G, 51L] of LSM A and segment [40L,48N] of LSM B in the complex pdb1mgq (showing only the backbone).

# Sequence Identity Within a Group



- Ave seq identity within a group is 7.5%
- ⇒ Group is unlikely to be detected by standard methods based on seq homology

## Conclusions

- **Connection between maximal complete bipartite subgraphs and closed patterns**
  - ⇒ **Closed pattern mining algorithms can be used to enumerate maximal complete bipartite subgraphs efficiently**
- **Connection between pairs of interacting protein groups and closed patterns**
  - ⇒ **Discovery of binding motifs is accelerated because we need not execute expensive motif discovery algorithms on insignificant groups**

# References

- Haiquan Li, Jinyan Li, Limsoon Wong. Discovering Motif Pairs at Interaction Sites from Protein Sequences on a Proteom-Wide Scale. *Bioinformatics*, 22(8):989--996, 2006
- Jinyan Li, Haiquan Li, Donny Soh, Limsoon Wong. A Correspondence Between Maximal Complete Bipartite Subgraphs and Closed Patterns. *Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 146--156, Porto, Portugal, October 2005
- Smith et al., Finding Sequence Motifs in Groups of Functionally Related Proteins. *PNAS*, 87:826-830, 1990
- Henikoff & Henikoff. Automated assembly of protein blocks for database searching. *NAR*, 19:6565-6572, 1991