Computational Methods for the Prediction of Protein-Protein Interactions

Concettina Guerra

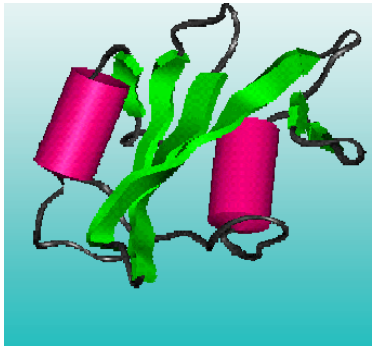
University of Padova, Italy

and Georgia Tech, USA

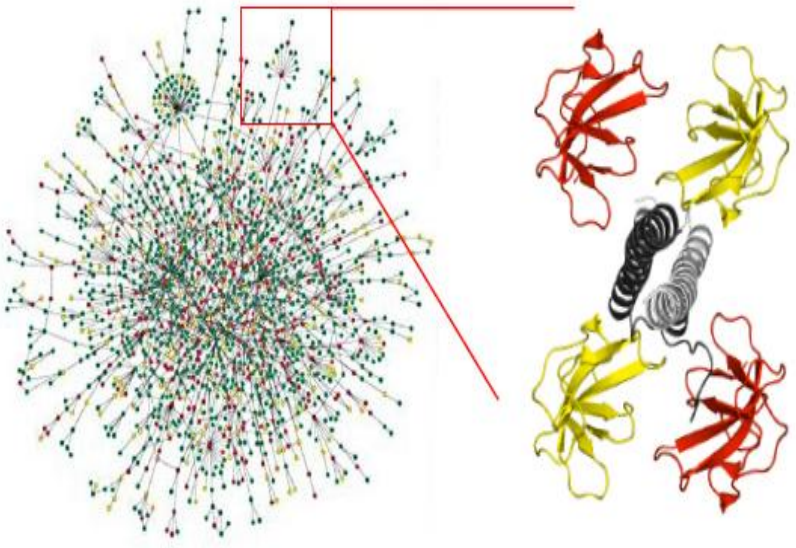
Different sources of information

- Sequences

- Folds (3D structures)



- PPI networks

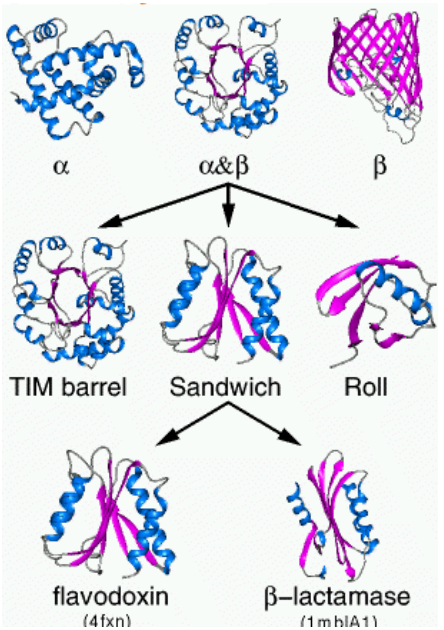


Inferring Protein Function - 1

Recognize a **sequence or fold similarity** with a protein whose function is known.

However:

- similar fold does not necessarily imply a similar function.
- proteins with different folds can share the same function.

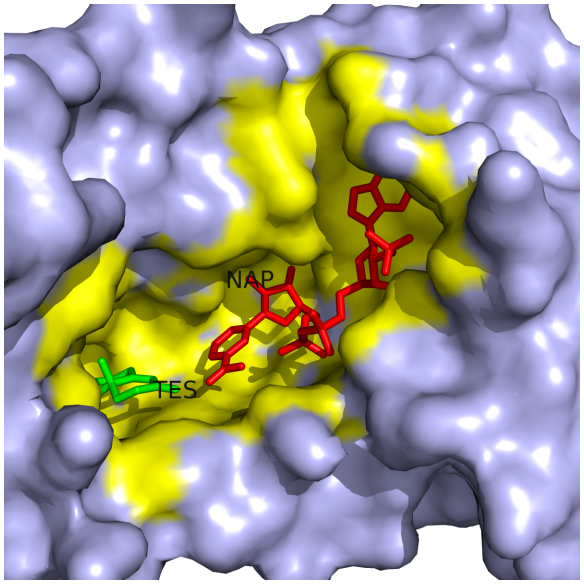


M. Comin, C.Guerra, G. Zanotti (2004). *J. of Computational Biology*.

Inferring Protein Function - 2

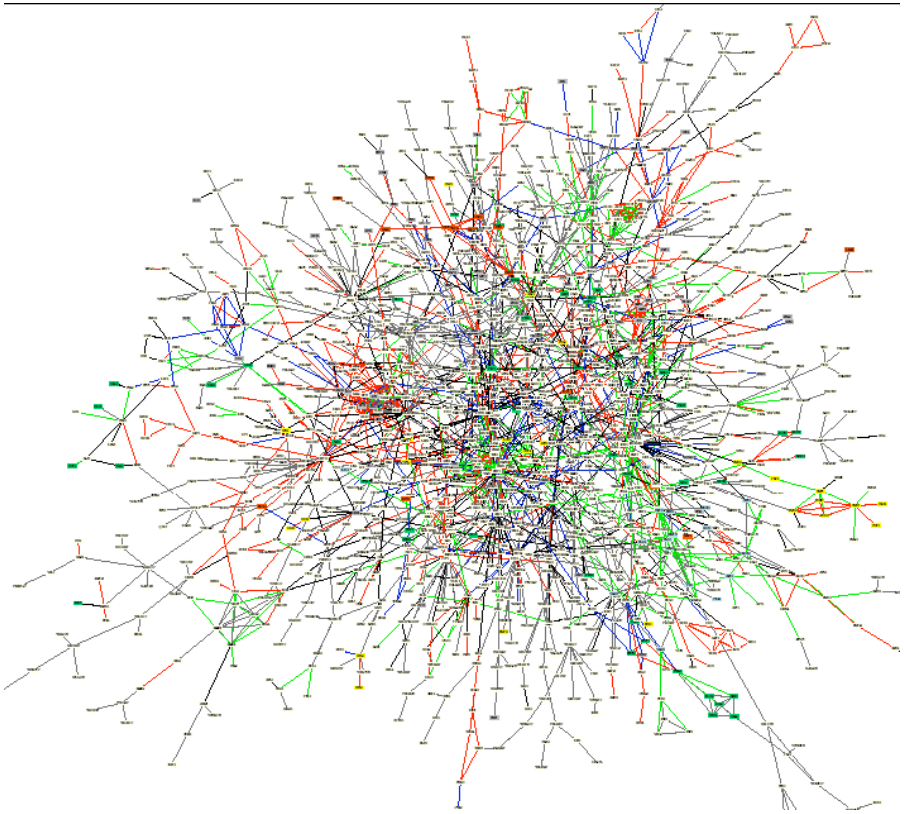
Prediction of binding sites

Proteins are assumed to perform similar functions if they **share similar binding patterns**



Inferring Protein Function - 3

From protein-protein interaction networks

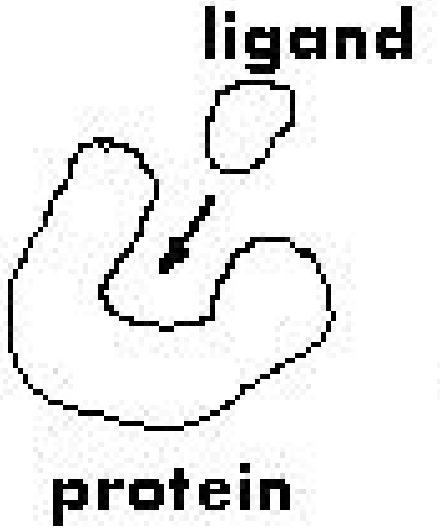


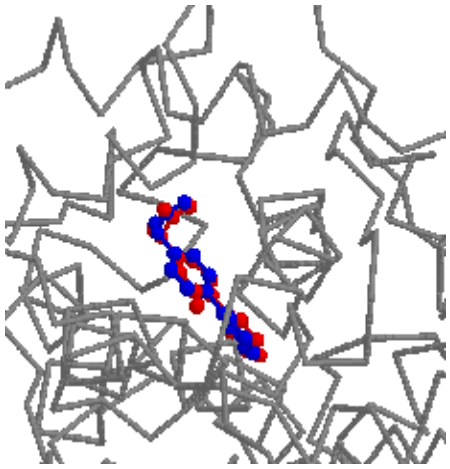
Overview

- Protein-protein interaction basics
- Protein surface comparison/docking
- Geometric shape descriptors
- Shape matching algorithms
- Network-based function prediction
- Network alignement

Protein-ligand docking

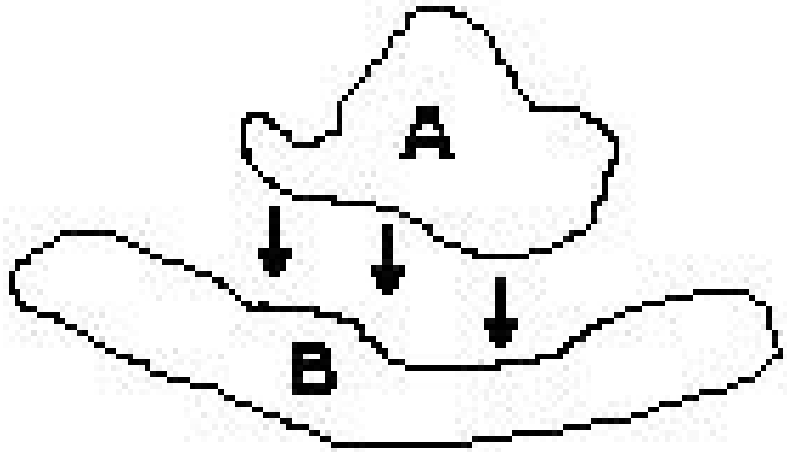
- A large molecule (receptor) and a small molecule (ligand) docking in a cavity.
- Key in Lock

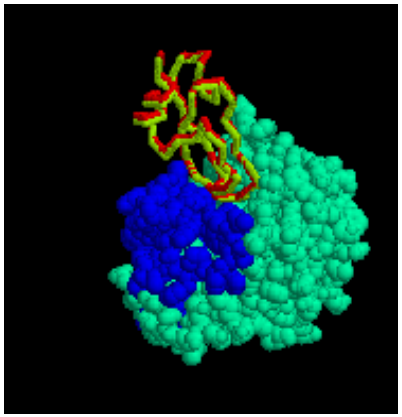


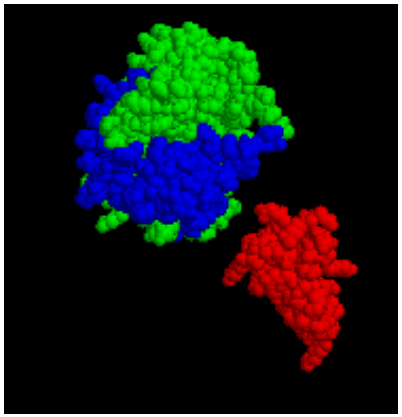


Protein-Protein Docking

- Two proteins approx the same size
- Tipically the docking site is a planar surface rather than a cavity.







Interface Characterization

Jones S., Thornton J.M., 97 -2000
Lo Conte L. , Chothia C. Janin J. 1999

- Interaction surfaces have few differential characteristics that can be captured by statistical methods
- No single parameter absolutely differentiate the interfaces from all other surface patches

Protein surface comparison

Three instances of the comparison problem:

(i) comparison of two binding sites

(ii) searching the surface of a protein (or one of its cavities) for a given binding site

(iii) given two complete protein surfaces find similar patches on the two surfaces

Geometry

Align two surface patches by finding the rigid transformation that best superimposes their atoms/residues

Surface representation

based on shape descriptors such as:

- Spin images
- Shape Contexts
- Spherical Harmonics
- …

Physico-chemical properties

Atoms are labeled for instance as

- hydrogen-bond donor
- hydrogen-bond acceptor
- mixed donor/acceptor
- hydrophobic aliphatic and aromatic(pi) contacts

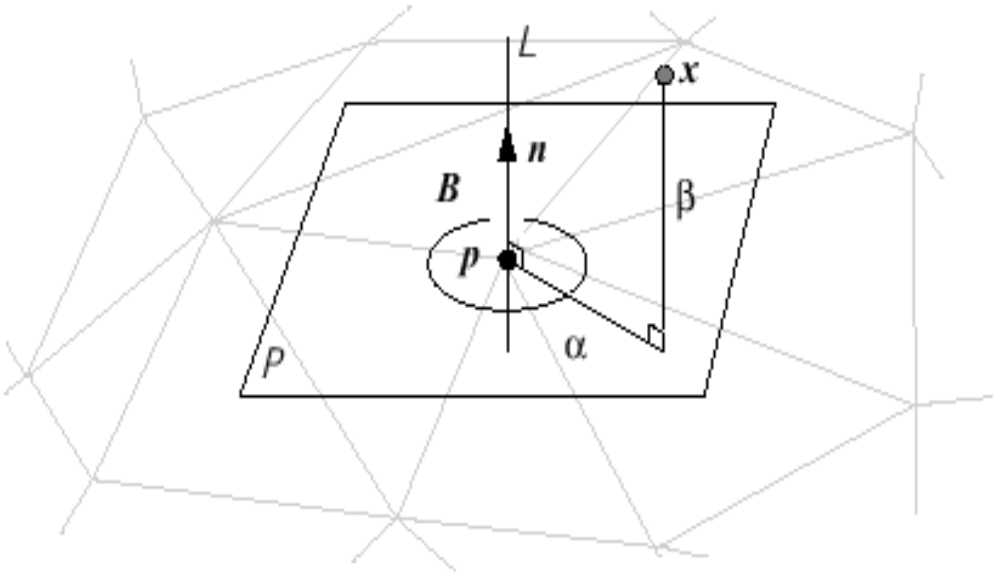
Schmitt et al., JMB 2002

Protein surface comparison using Spin Images

- A surface representation that uses 2D images to describe 3-D oriented points (Jonhson, Hebert, 1997)
- It allows to apply powerful techniques from 2-D template matching and pattern classification to the problem of 3-D surface recognition.

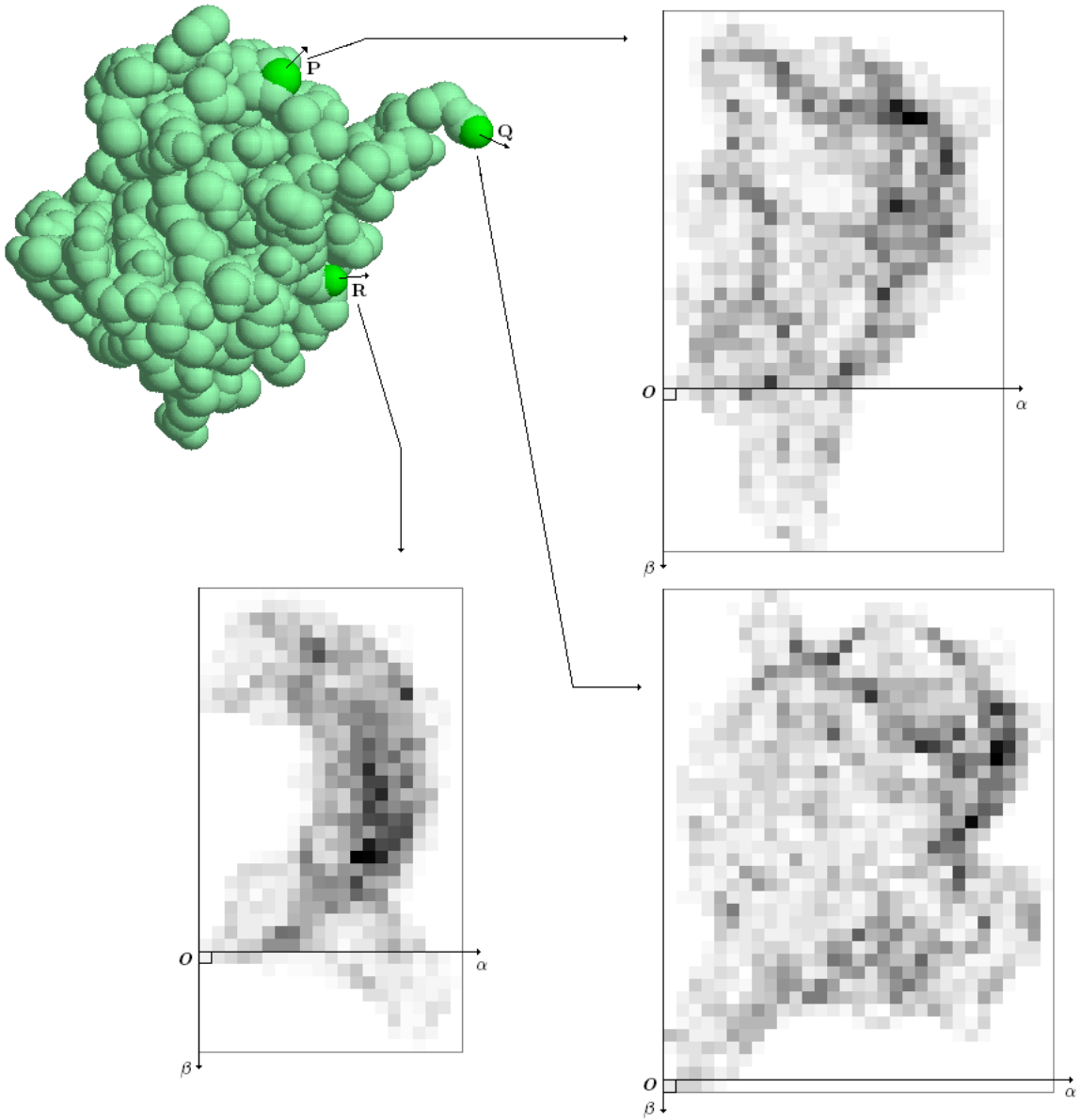
M. E. Bock, C. Garutti, C. Guerra, Discovery of similar regions on protein surfaces. J. of Computational Biology, 2007.

An oriented point basis



$$S_O : R^3 \to R^2$$

$$S_O(x) \to (\alpha, \beta) = \left(\sqrt{\|x-p\|^2 - (n\cdot(x-p))^2},\ n\cdot(x-p)\right)$$


P
Q
R
O
α
β
O
α
β
O
α
β

Correlation of spin images

- Given two spin-images *P* and *Q* with *N* bins each, the linear correlation coefficient *R(P,Q)* is:

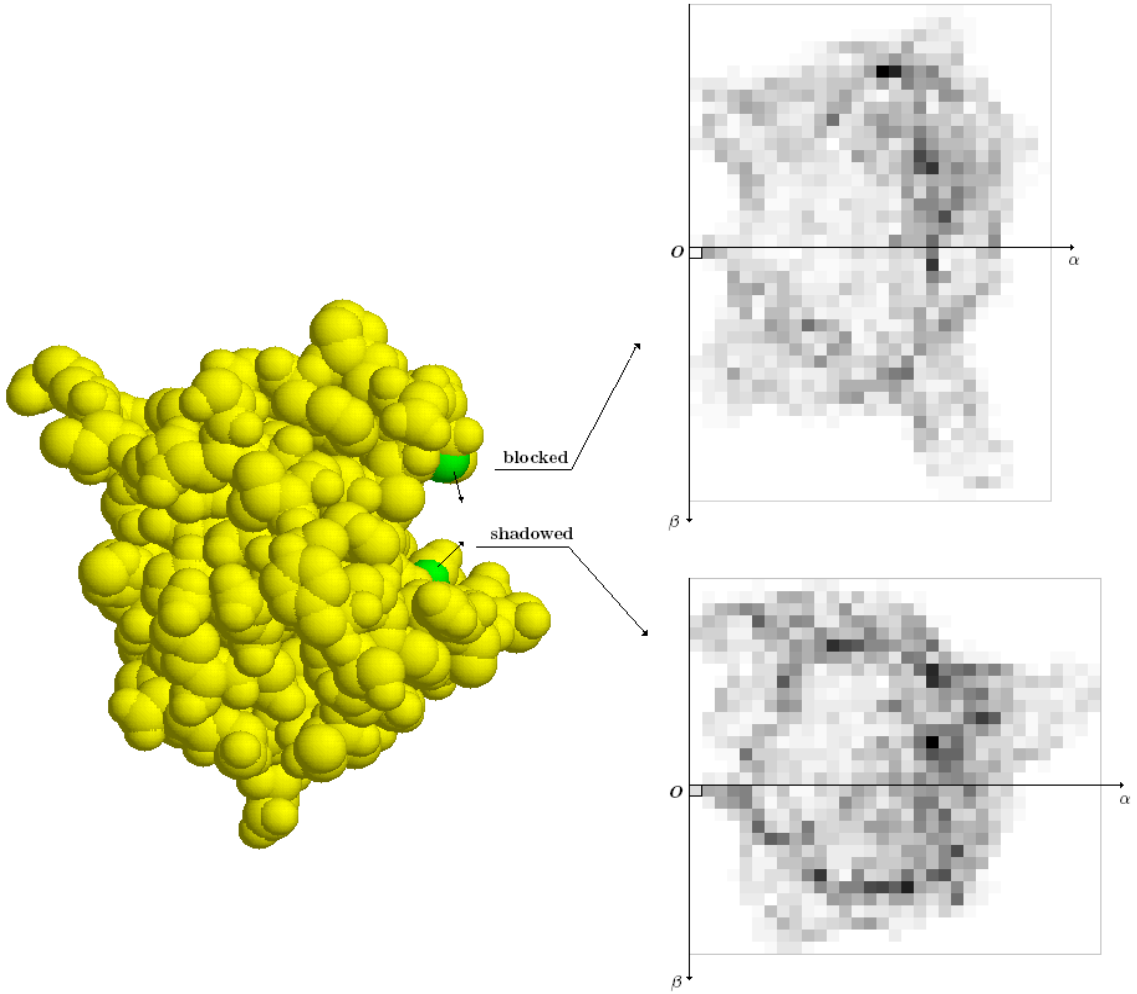
$$R(P,Q) = \frac{N\sum p_i q_i - \sum p_i \sum q_i}{\sqrt{(N\sum p_i^2 - (\sum p_i)^2)(N\sum q_i^2 - (\sum q_i)^2)}}$$

Labeling surface points

Protein surface points are lebaled as *blocked, shadowed,* or *clear.*

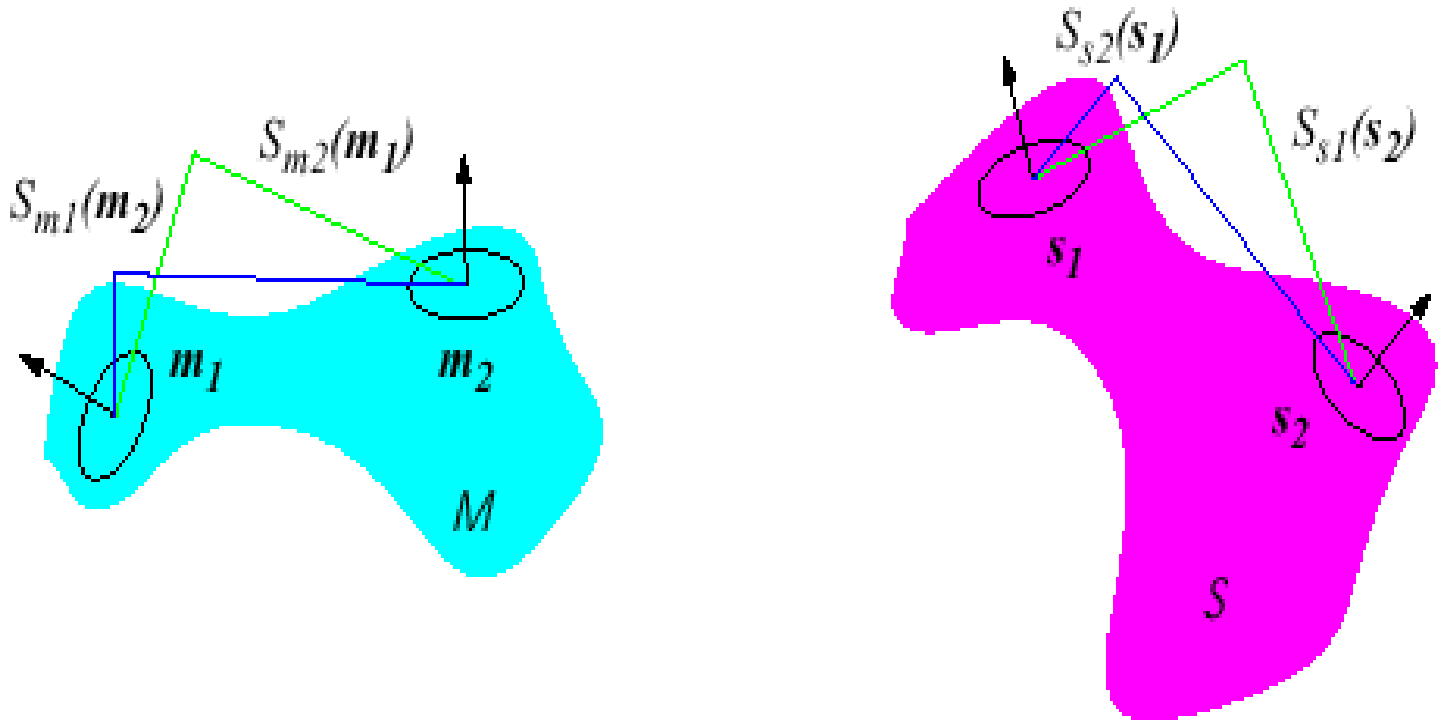
- A surface point *P* with normal *n* is *blocked* if *n* intersects the surface at some other point lying above the tangent plane at P perpendicular to *n*.
- The unblocked points that belong to the convex hull of the protein surface are labeled *clear* points all others are shadowed.

Examples of blocked and shadowed points



Grouping Point Correspondences for surface matching

The grouping criterion is the Geometric Consistency of distances and angles of corresponding points



Geometric Matching

A three-step procedure:

1. Establish individual point correspondences based on the correlation of the spin images and of spin image profiles.

   Correspondences are restricted to points with the same label

2. Group point correspondences using a geometric consistency criterion

   Use a grid-based search algorithm that grows regions around selected point correspondences

3. Score each group by the number of pairs of corresponding points.

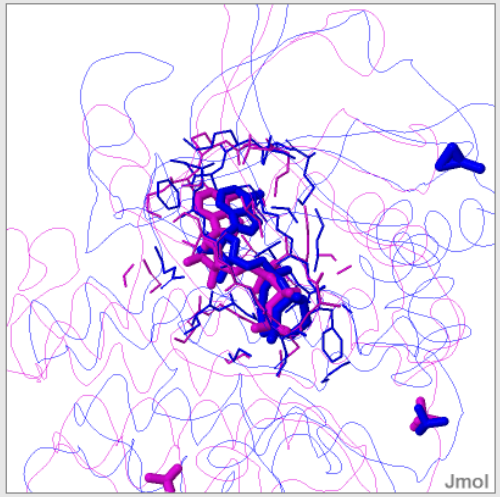
MolLoc:

a web server for local alignment of molecular surfaces

Statistics	
1st structure	*1atp,E*
2nd rototranslated structure	*1csn,A*
Number of selected atoms in the 1st structure	159
Number of selected atoms in the 2nd structure	131
Superimposed surface in the 1st structure	868/901 AA (96%)
Superimposed surface in the 2nd structure	795/820 AA (96%)
Number of atom correspondeces	41
RMSD of atoms correspondences	0.11
Matrix of rototranslation in DaliLite format	*Download*
PyMol script	*Download*
Table of correspondeces	*Download*

You can later view this page for about 24 hours through *this link*

1atp,E			1csn,A			Check to show atoms in the picture	
RES.	*RES.NUM.*	*ATOM*	*RES.*	*RES.NUM.*	*ATOM*	*TYPE*	☐
VAL	57	CG1	ILE	26	CG2	AL	☐
VAL	57	CB	ILE	26	CB	AL	☐
VAL	123	CG1	LEU	88	CD2	AL	☐
VAL	123	CG2	LEU	88	CD1	AL	☐
VAL	57	CG2	ILE	26	CG1	AL	☐
TYR	122	O	LEU	87	O	AC	☐
THR	51	O	GLU	20	O	AC	☐
THR	51	C	GLU	20	C	PI	☐
THR	51	N	GLU	20	N	DO	☐
SER	53	O	SER	22	O	AC	☐
SER	53	OG	SER	22	OG	DA	☐
SER	53	N	SER	22	N	DO	☐
PHE	54	N	PHE	23	N	DO	☐
PHE	54	O	PHE	23	O	AC	☐
PHE	54	C	PHE	23	C	PI	☐
MET	120	SD	ILE	85	CD1	AL	☐



S. Angaran, M.E. Bock, C. Garutti and C. Guerra (2009). *Nucleic Acids Research*.

Experimental Results

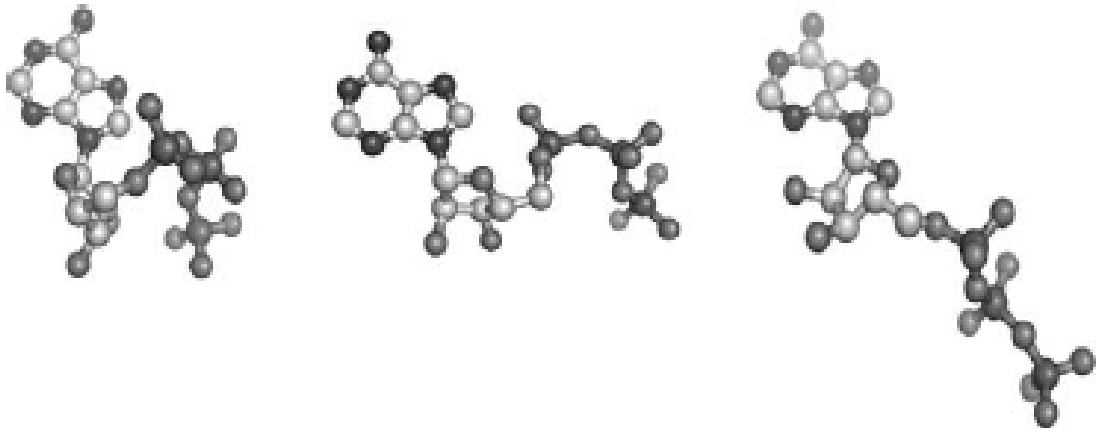
Data set of protein complexes

(Wolfson et al, 2005)

Protein family	PDB id
Adenine-binding	1ads 1byq 1bv4 1bx4 1byq 1kpf 1mmg 2src 1zin 9ldt
ATP binding proteins	1a82 1atp 1csn 1e2q 1f9a 1hck 1j7k 1jjv 1mjh 1nhk 1nsf 1phk
Serine proteases	1abi 4sgb 4tgl
Fatty acid binding proteins	1b56 1kqw 1lib 2cbr
Estradiol	1a27 1e6w 1fds 1luh 1qkt 3ert
Anhydrase	1jd0
Retinoic acid-binding	1gx9
Antibiotics	1alq 1bt5 1dcs
HIV-1	1mu2
Viral proteinase	1cqq 1mbm 1q2w
Chorismate mutase	1fnj

Different conformations of ATP

compact **intermediate** **extended**



Conformational Diversity of Ligands Bound to Proteins
Stockwell, Thornton *J. Mol. Biol. (2006)*

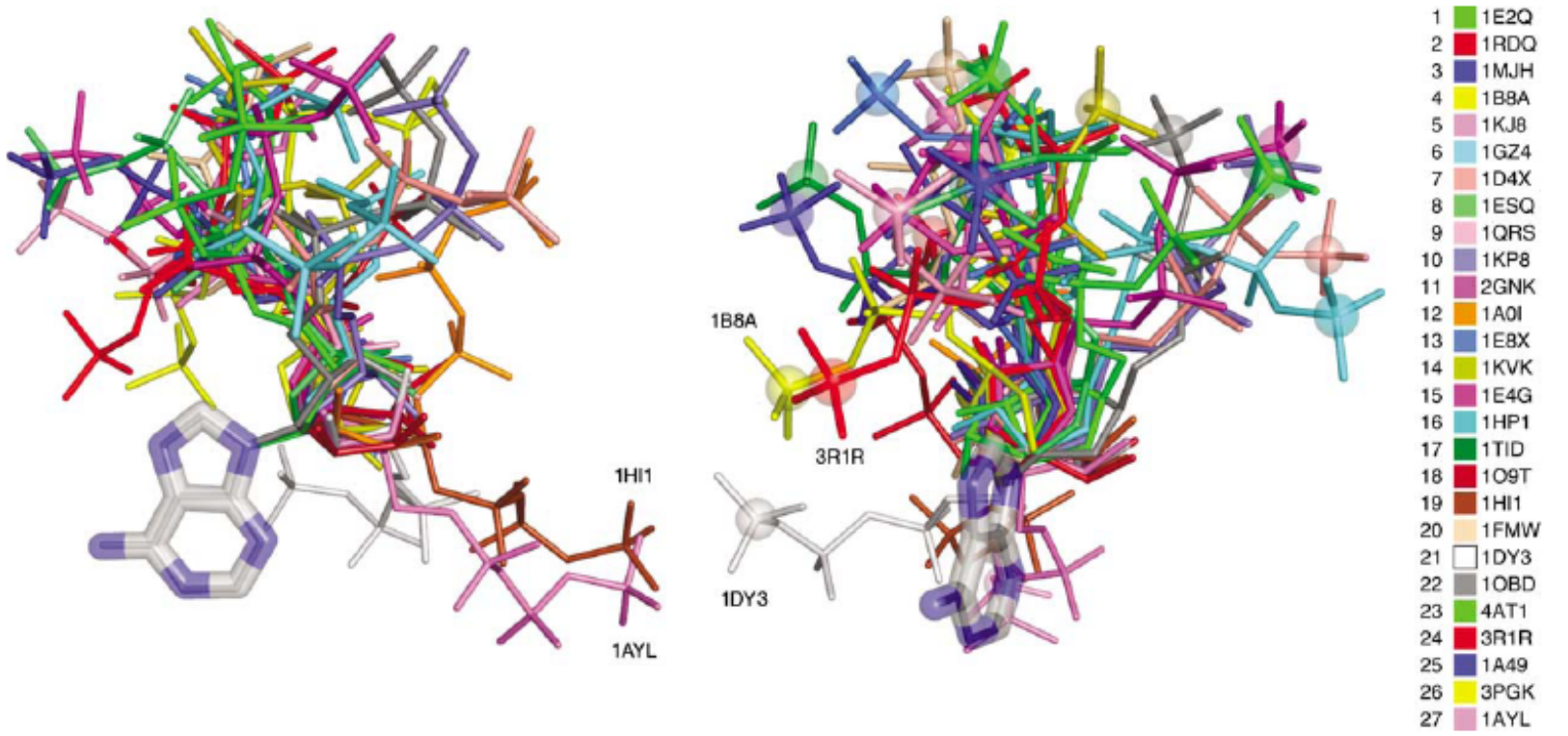


Figure 1. Superposition of the 27 ATP cluster representatives on their adenine rings (highlighted). In the second image, the gamma phosphate atoms are shown with translucent spheres, to highlight the broad range of conformations adopted by the triphosphate tail. The key shows from which PDB entry each molecule was taken. Several particularly unusual conformations are indicated with labels on the plots themselves.

Results
Method Based on Spin-Images (SIM)

Rank	PDB:chain	Protein	Fold	# Corr.	Ligand	Rmsd
1	1phk	g-Subunit of glycogen phosphorylase kinase	Protein-kinase	190	ATP	1.1
2	1csn	Casein kinase-1, CK1	Protein-kinase	92	ATP	1.9
3	1mjh:B	"Hypothetical" protein MJ0577	Adenine nucleotide a hydrolase-like	56	ATP	0.7
4	1g5y:B	Retinoid-X receptor alpha	Nuclear receptor ligand-binding domain	55	REA	1.0
5	1bx4:A	Human Adenosine Kinase	Ribokinase-like	46	ADN	1.8
6	1b4v:A	Cholesterol Oxidase	FAD/NAD(P)-binding domain	46	FAD	1.8
7	2src	Tyrosine-protein Kinase SRC	Protein kinase-like (PK-like)	44	ANP	1.3
8	1hck	Cyclin-dependent PK	Protein-kinase	43	ATP	2.6
9	1nsf	Hexamerization domain of N-ethylmalemide-sensitive fusion protein	P-loop containing nucleoside triphosphate hydrolases	43	ATP	1.4
10	1f9a:A	"Hypothetical" Protein MJ0541	Adenine nucleotide alpha hydrolase-like	43	ATP	0.9

Table 2: High scoring pair-wise comparisons with 1atp:E.

SiteEngine
(Wolfson et al, 2004)

Table 3. Recognition of ATP-binding sites by searching the database of active sites

Rank	PDB	Protein	Fold	Sequence similarity (%)	Match score	Ligand	Run time (seconds)
1	1mjh	Hypothetical protein MJ0577	Adenine nucleotide alpha hydrolase-like	100	100	ATP	4
2	9ldt	Lactate dehydrogenase	NAD(P)-binding Rossman-fold domain	6	36	NAD	7.8
3	1atp	cAMP-dependent PK, catalytic subunit	Protein kinase-like (PK-like)	8	35	ATP	6.4
4	1b4v	Cholesterol oxidase of GMC family	FAD/NAD(P)-binding domain	11	34	FAD	6.8
5	1a27	Human estrogenic 17beta-hydroxysteroid dehydrogenase	NAD(P)-binding Rossman-fold domain	12	34	FAD	9.6
6	1nsf	Hexamerization domain of *N*-ethylmalemide-sensitive fusion (NSF) protein	P-loop containing nucleotide triphosphate hydrolases	10	34	ATP	5.8
7	1a82	Dethiobiotin synthetase	P-loop containing nucleotide triphosphate hydrolases	5	34	ATP	6.3
8	1hsh	HIV-1 protease	Acid proteases	6	33	MK1	8.3
9	1e8x	Phoshoinositide 3-kinase (P13K) helical domain	Alpha–alpha superhelix	6	33	ATP	7
10	1a49	Pyruvate kinase	PIK beta-barrel domain-like	10	32	ATP	6.4
11	2src	c-src Tyrosine kinase	Protein kinase-like	10	32	ATP	7.5
12	1csn	Casein kinase-1, CK1	Protein kinase-like	14	32	ATP	6
13	1hck	Cyclin-dependent PK	Protein kinase-like	10	31	ATP	6.1
14	1zin	Adenylate kinase	P-loop containing nucleotide triphosphate hydrolases	6	31	ATP	6.8
15	1bx4	Adenosine kinase	Ribokinase-like	5	31	ATP	5.6

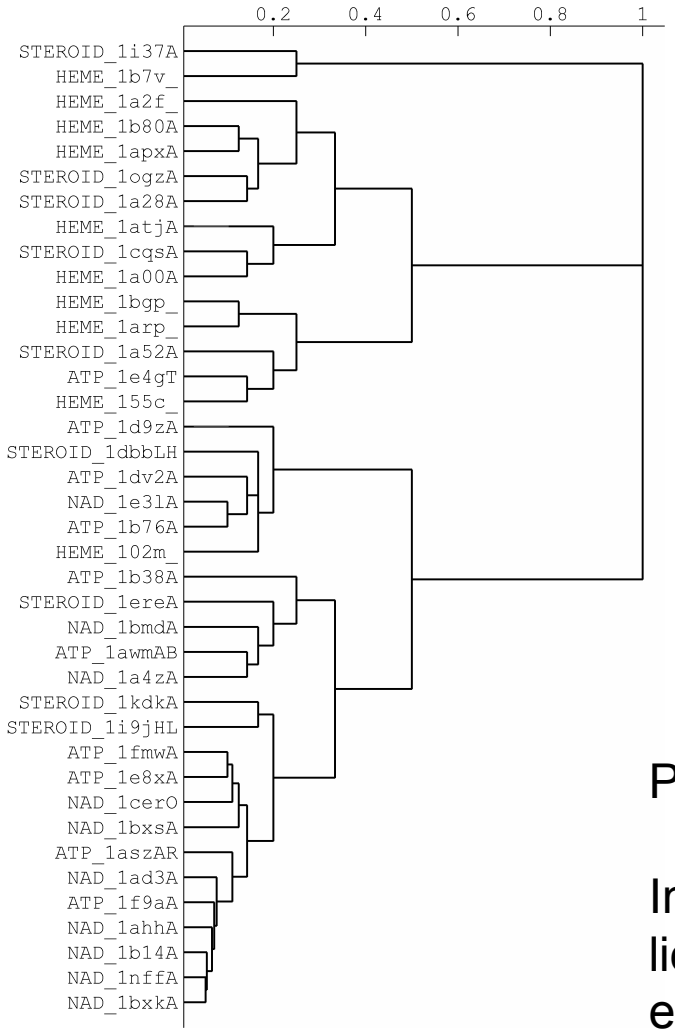
Results

Pdb ID	# residues in binding site	Coverage (Bock 2007)	Coverage Cavity comparison	Accuracy Cavity comparison
1atp	23	78%	91%	80%
1phk	26	69%	90%	76%

Pdb ID	# residues in binding site	Coverage (Bock 2007)	Coverage Cavity comparison	Accuracy Cavity comparison
1atp	23	43%	60%	93%
1nsf	23	35%	43%	76%

All-To-All Pairwise Cavity Comparison

Clustering of solutions



Protein 1jtv binding a steroid is misclassified

In this protein, the binding site of the steroid lies into a big cavity, which hosts also a more extended NAD(P) binding site

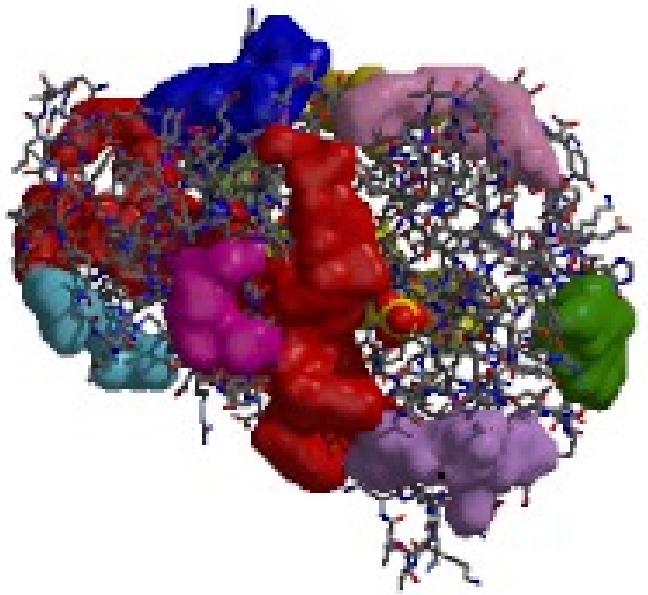
Finding surface cavities and binding pockets

For protein/drug interaction

- **SPHGEN**, **Surfnet**
determine sphere clusters

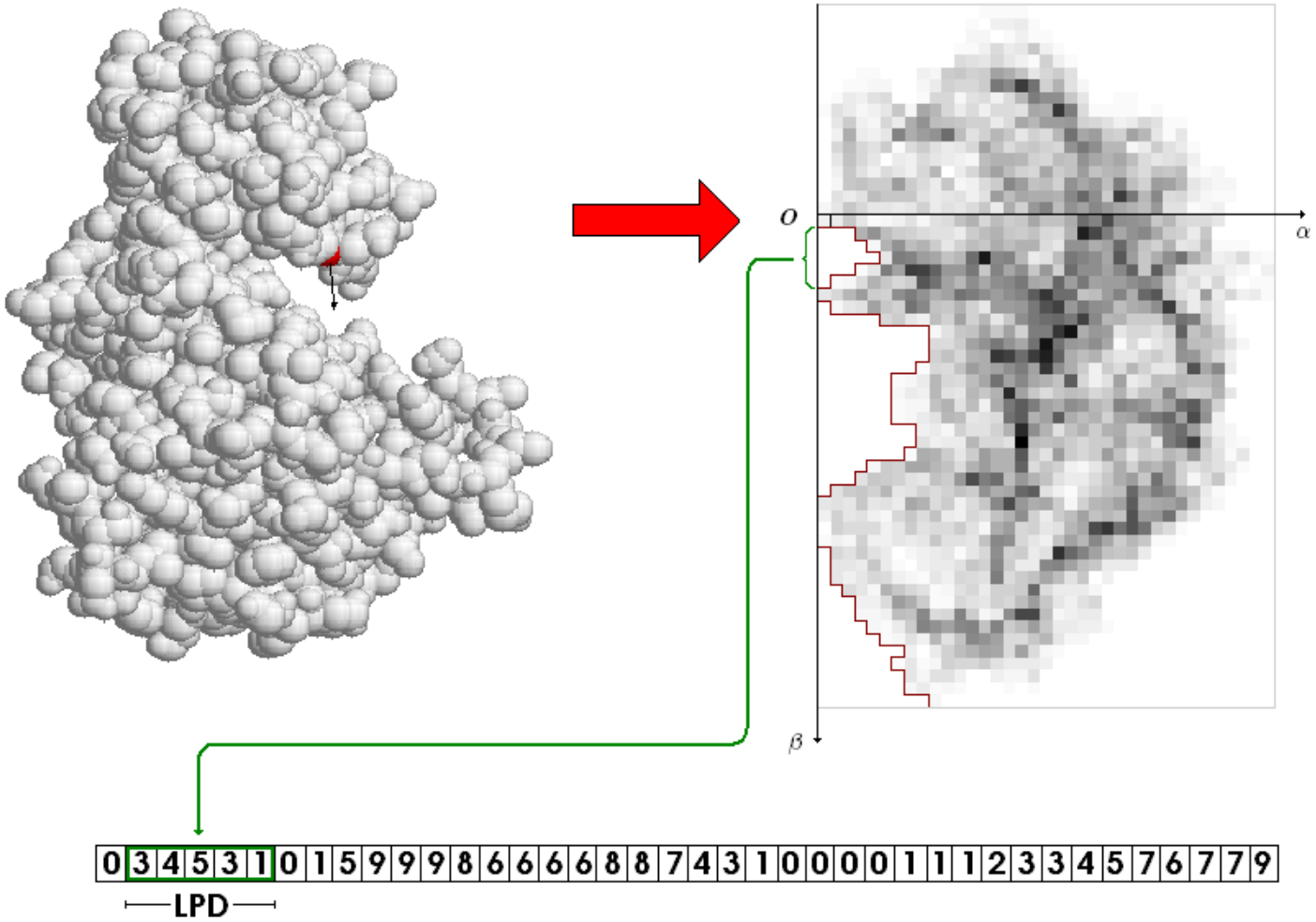
- **CastP**
Alpha Shapes

- **SpinImages**



Cavity detection using spin image profiles

Find the largest sphere that can fit into the empty space



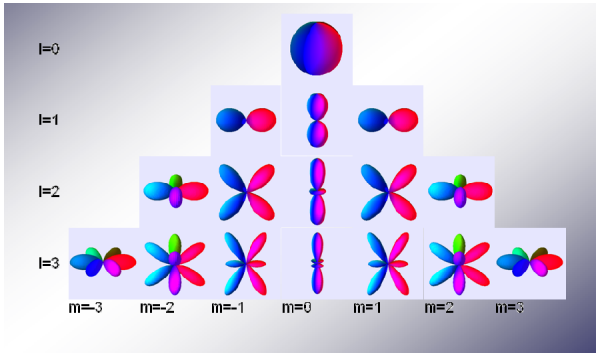
M.E. Bock, C. Garutti, C. Guerra, Proc. CSB, San Diego, USA 2007

Binding Site Recognition using Spherical harmonics

Quickly identify promising binding sites, either in a protein cavity or on an entire protein surface

no explicit alignment

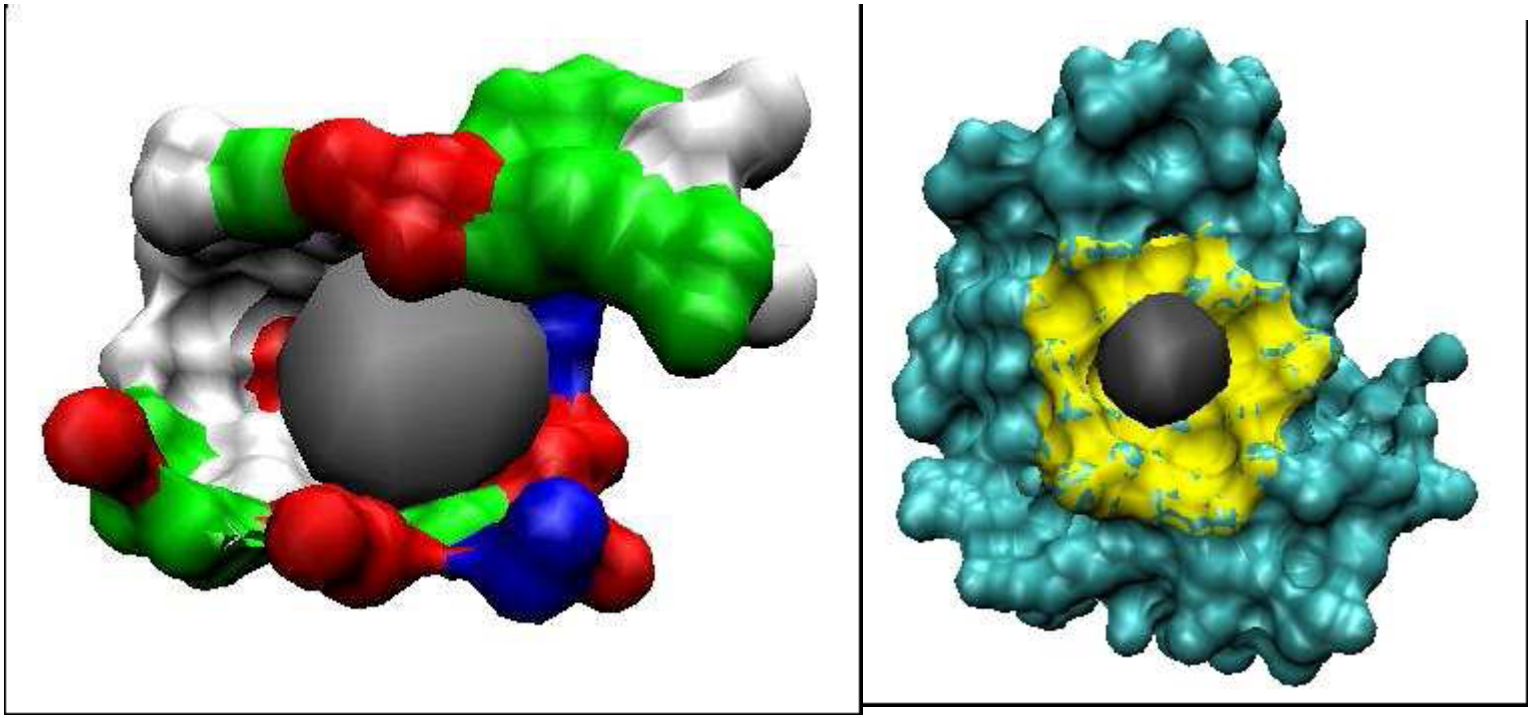
this method can save up to 40% in time compared with traditional approaches.



M. Comin, F. Dellaert, C. Guerra. Binding Balls: Fast detection of Binding Sites using a property of Spherical FourierTransform. J. of Computational Biology, 2009.

Binding Balls

Fast detection of Binding Sites using a property of Spherical Fourier Transform.



Global Optimization by controlled-random search

Determine the best rotation that superimposes two surface patches

Similar to Iterative Closest Point ICP method used in computer vision.

ICP however converges to a local minimum

P. Bertolazzi, C. Guerra, G. Liuzzi (2009), Proc. CWS, Washington, USA .

A new dissimilarity measure

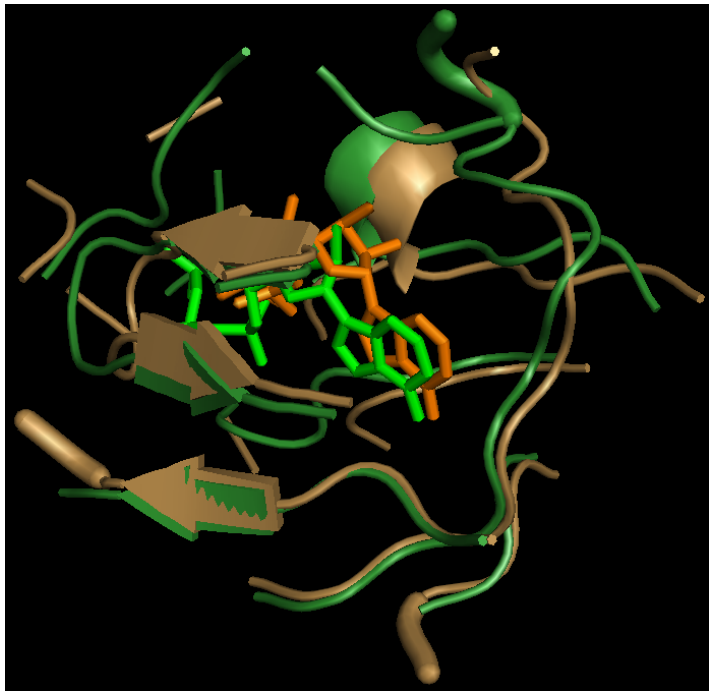
based on the solution of an

Asymmetric Assignment Problem

on a bipartite graph associated to the matching problem.

The matching takes into account physico-chemical constraints

Example of superposition



Good performance in classifying proteins based on functional classes

Assessment of existing methods

At date, no systematic and comprehensive evaluation exists of methods for binding site recognition

(Unlike methods for protein structure alignment see M. Levitt et al, 2005)

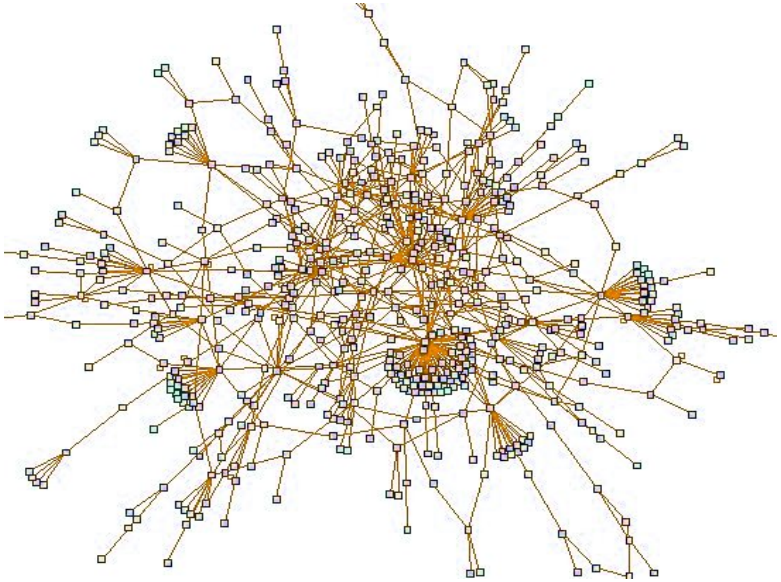
Difficulty arise because of:

- **different instances of comparison problems**

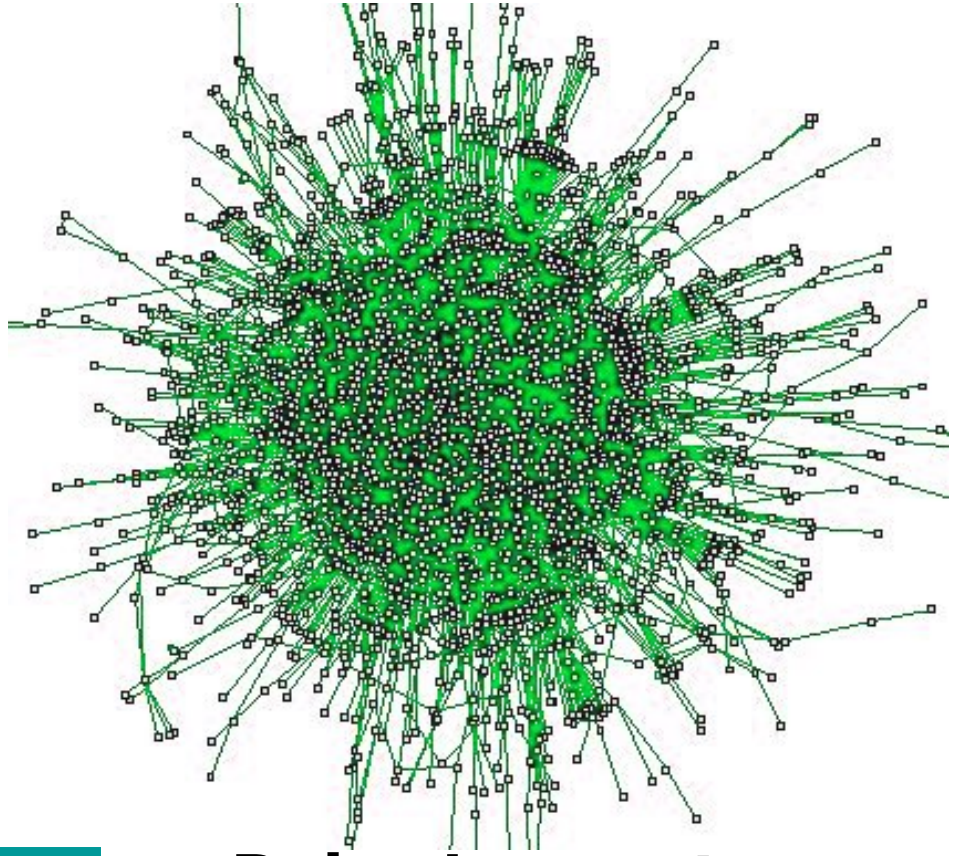
and because of the use of:

- **different surface representations**
- **different native score**

PPI networks



Bacterial pathogen (*Helicobacter pylori*)



Baker's yeast (*Saccharomyes cerevisiae*)

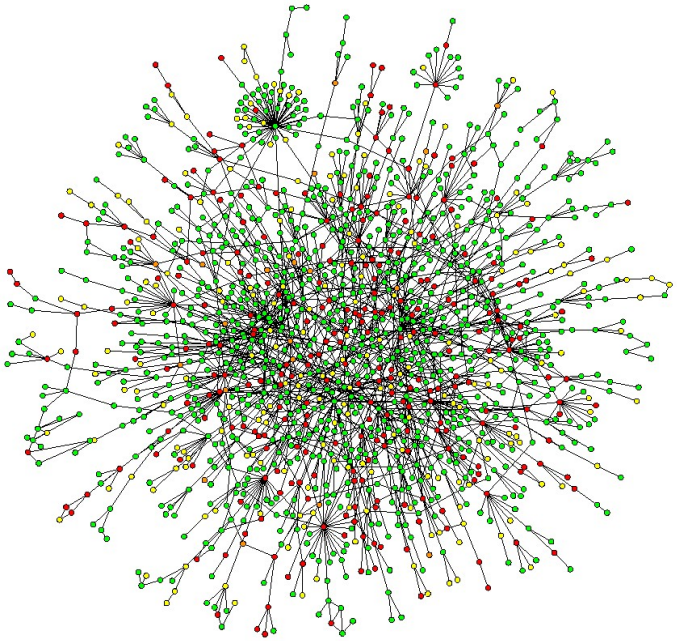
Interaction Data
is noisy (~50% false positives) and incomplete

Models and Properties

- Global Properties
  - Degree Distribution
    - Scale-Free
    - Random
    - ..
- Local Properties
  - Motifs
  - Motif Profiles
  - ..

Network-based prediction of protein function

From a single species



Methods

- **Neighbor-based**
- **Module-assisted**

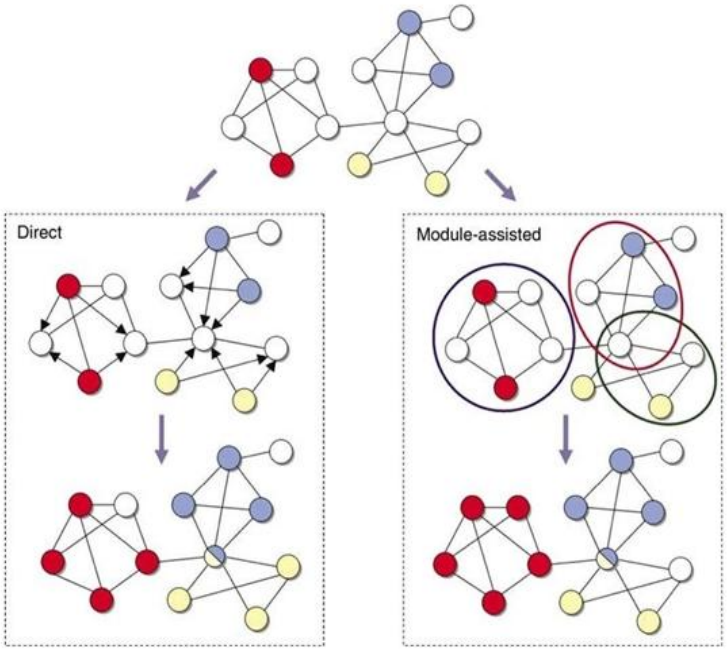


Figure from R.Sharan-Shamir, MSB, 2007

Functional information from multiple networks

Identify evolutionary conserved modules via the integration of networks from multiple organisms

Paradigm: Evolutionary conservation implies functional significance

Network Comparison or Alignment

Conservation: similarity in sequence and interaction topology.

Problem Formulation

Input: two PPI networks *G*1 and *G*2.

- Each edge *e* may have a weight *w*(*e*)
- Other measures of similarity between the nodes may be available
  - BLAST similarity scores
  - co-expression
  - ….

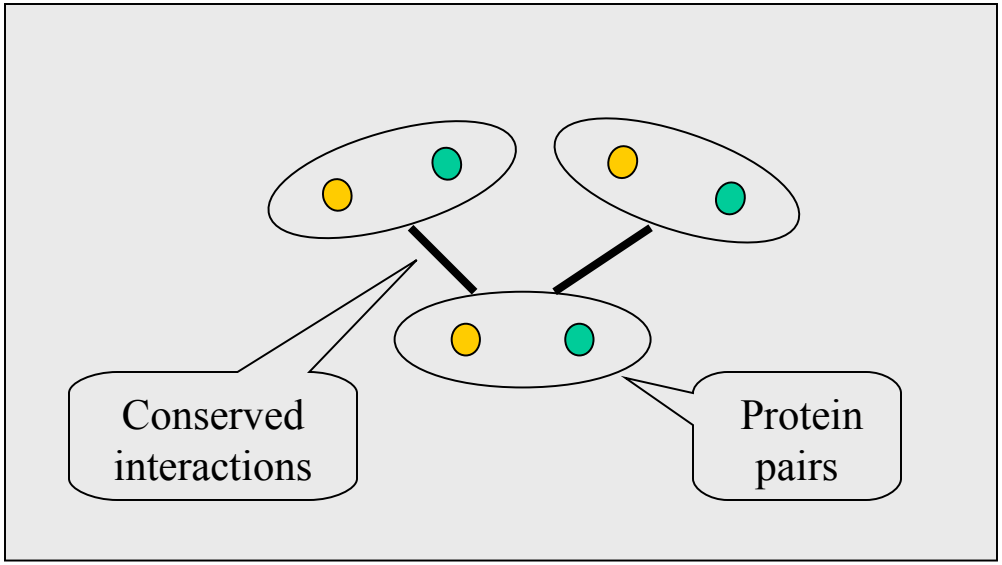
Output:

the maximum common subgraph (MCS) between *G*1 and *G*2

NP-complete problem

Note: By incorporating sequence data, the global alignment problem is no longer a pure MCS problem.

Alignment graph:



Nodes: pairs of orthologous proteins, one per species.
Edges: conserved interactions.

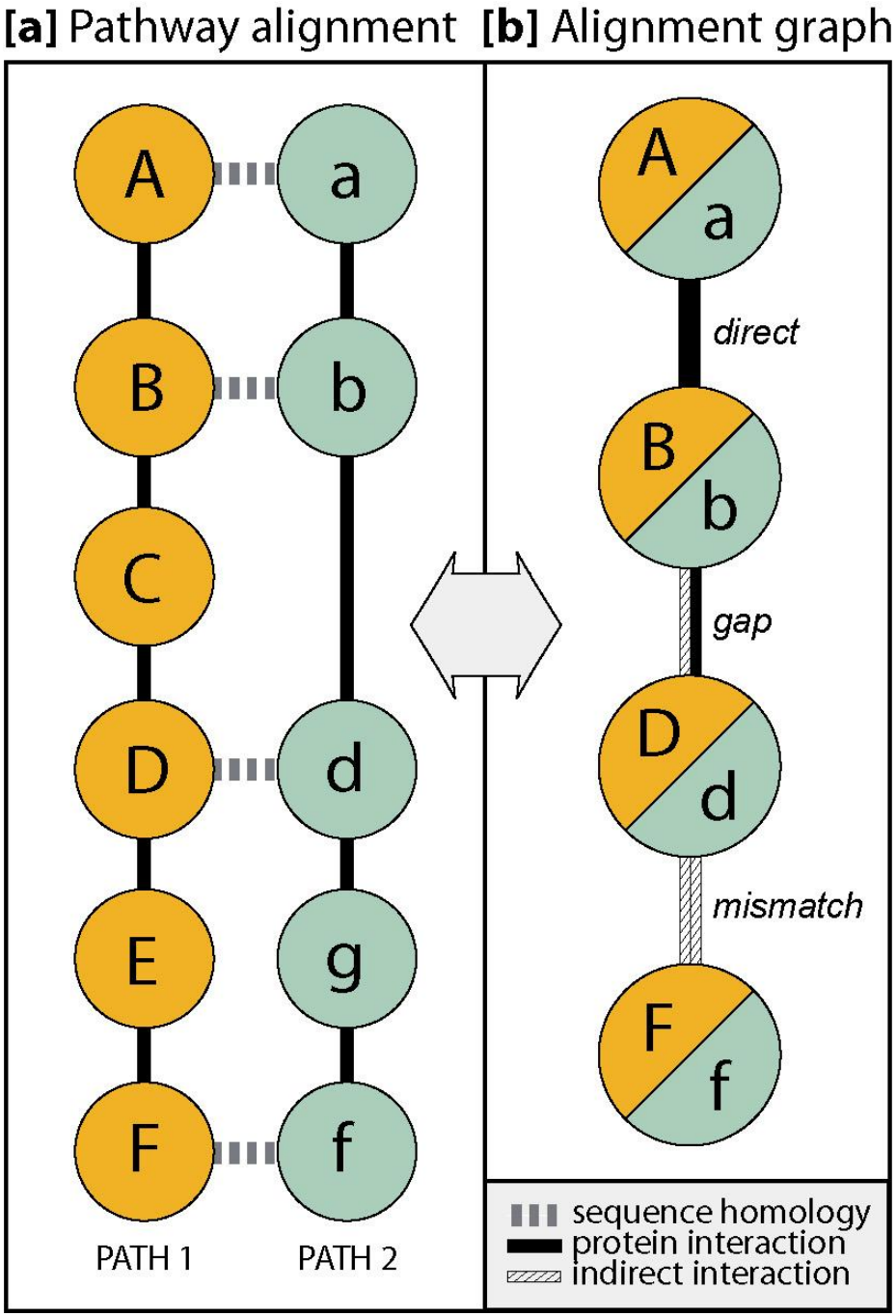
PathBLAST

Reduction to finding **paths** in an "alignment" graph.

- Repetitions are possible.

NetBLAST

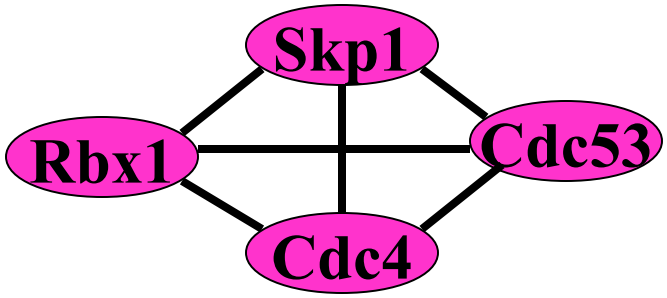
Generates arbitrary sub-graphs

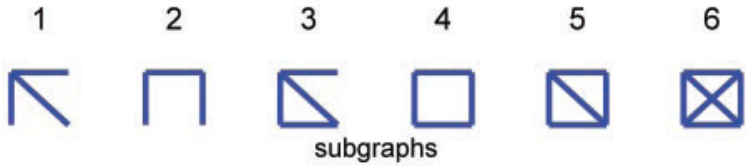


Kelley et al., PNAS'03

Identifying conserved subgraphs
Our approach

Identify high scoring graphlets in each PPI





How to devise a scoring scheme and search for high-scoring conserved subnetworks?

G. Ciriello, C. Guerra, P. Guzzi (2010) in preparation

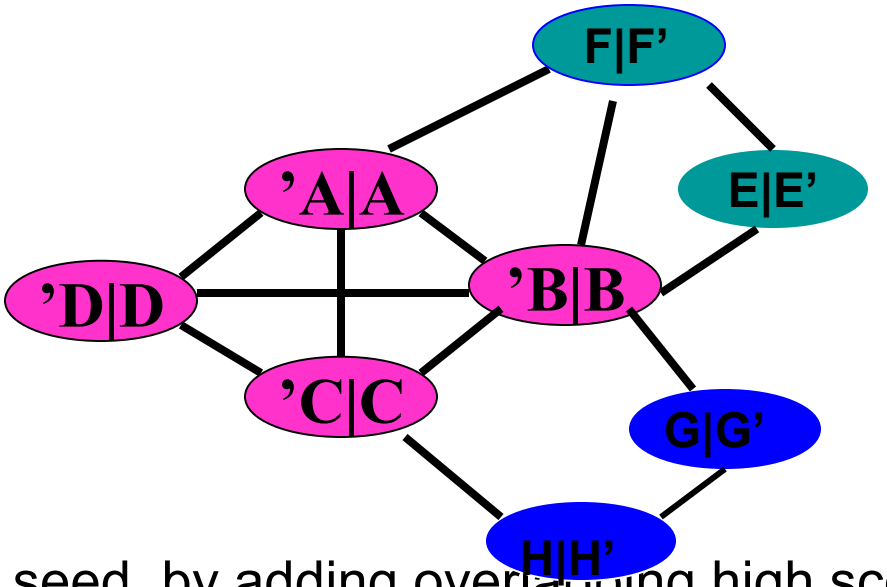
Preprocessing

- Interactions in a graphlet are scored based on:
  - Type of experimental observation
    (Y2H, CoIP, small scale, ect)
  - Number of experimenal observations
  - Number of references in scientific literature

- Orthologous proteins in the two species are established based on sequence similarity (Inparanoid)

Greedy Search

- Construct a high-scoring seed formed by pairs of graphlets containing orthologous proteins



- Expand seed by adding overlapping high score graphlets containing orthologous proteins

Preliminary Results

Generally the common subgraphs are

larger and more dense than any common subgraph identified using

- Netblast (Sharan et al, 2006)
- Isorank (Berger et al, 2009)
- Mawish (Koyouturk et al, 2005)

Summary and Conclusions

Functional prediction from:

- sequences
- structures
- PPI networks

Most importantly by the integration of the above and other data sources such as:

co-expression data

evolutionary profiles

multi-level information

Collaborators

- Univ. Padova
  - Matteo Comin
  - Giovanni Ciriello
  - Claudio Garutti
  - Giuseppe Zanotti

- Purdue Univ.
  - Mary Ellen Bock
- IASI – CNR
  - Paola Bertolazzi
  - Giampoalo Liuzzi

Univ. Magna Grecia
Pietro Guzzi

- Georgia Tech
  Frank Dellaert