Methods for Comparing Protein Surfaces and their Application to Binding Site Recognition

Concettina Guerra

University of Padova, Italy

and Georgia Tech, USA

Overview

Protein-protein and Protein-Ligand interactions

Protein surface comparison

– Geometric shape descriptors

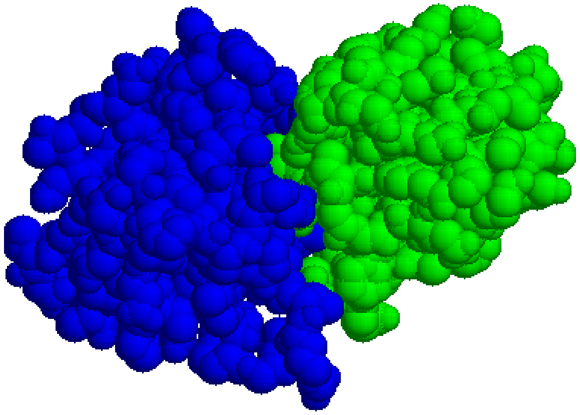
– Shape matching algorithms

Protein-RNA interactions

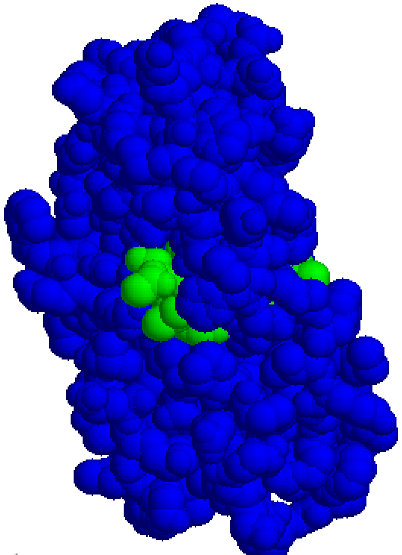
Protein-protein interactions

Given a pair of molecules represented by their 3D structures.

- **Decide whether the molecules will interact/bind**
- **Predict the 3D structure of the complex.**
- **Derive function.**



a) protein-protein

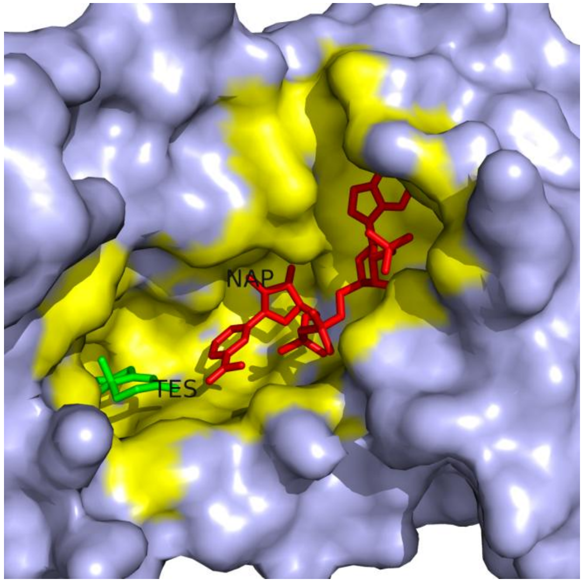


b) receptor-ligand

Prediction of binding sites of proteins

To infer protein function

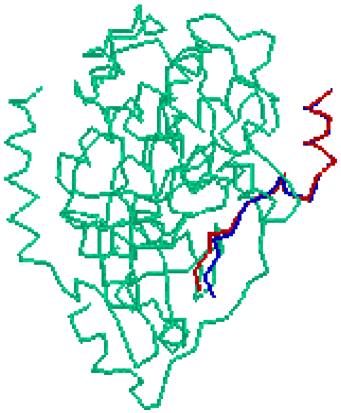
Proteins are assumed to perform similar functions if they **share similar binding patterns**

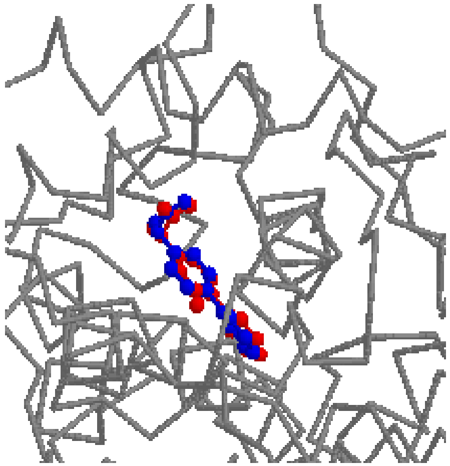


Protein-ligand docking

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

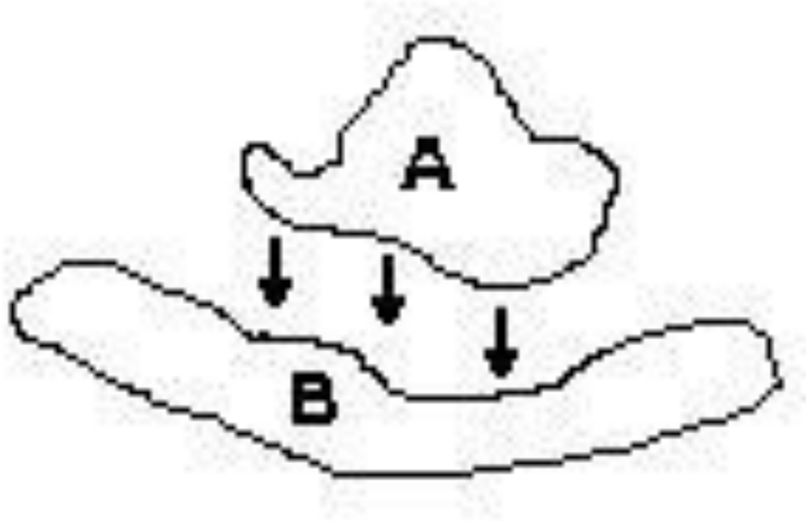
ligand
protein

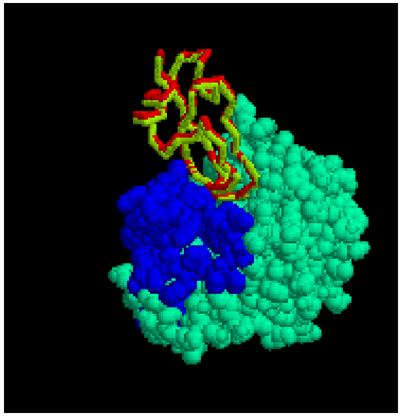


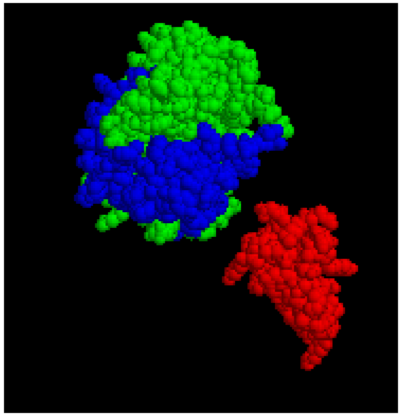


Protein-Protein Docking

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







Interface Characterization

- 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

Jones S., Thornton J.M., (2000)

Lo Conte L. , Chothia C. Janin J. (1999)

Surface patches

Surface residue – relative accessible surface area (ASA) > 5%

Patch – central surface accessible residue and *n* nearest surface accessible neighbors, where *n* – number of residues in the observed interface

Interface patch – those residues with side-chains possessing an ASA that decreased by > 1Å^2

Properties

- Residue interface propensity
- Hydrophobicity
- Planarity
- Protrusion
- Accessible surface area
- ….

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
- Pseudo-centers
- Spherical Harmonics

Physico-chemical properties

Atoms are labeled as

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

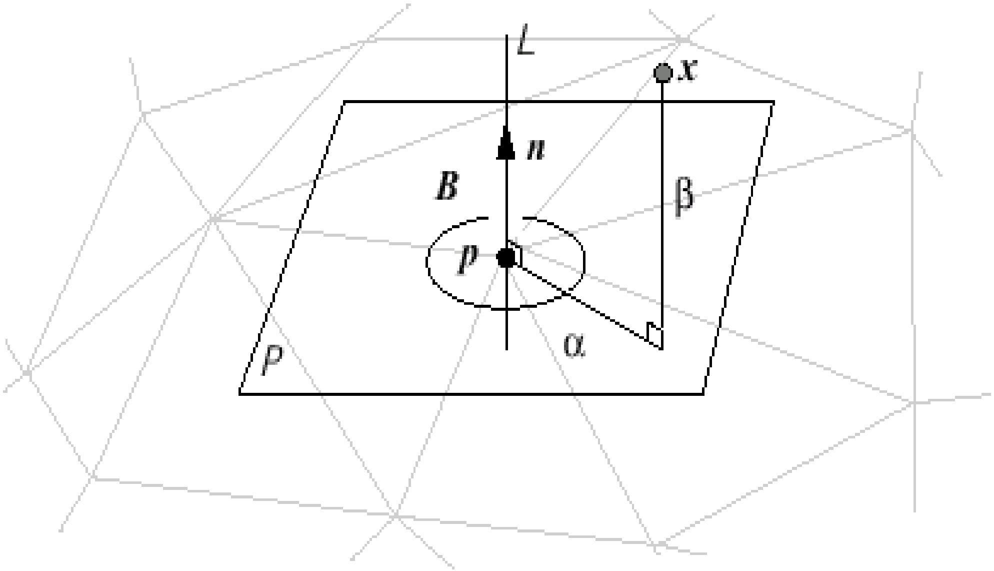
Schmitt et al., (2002) JMB

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.

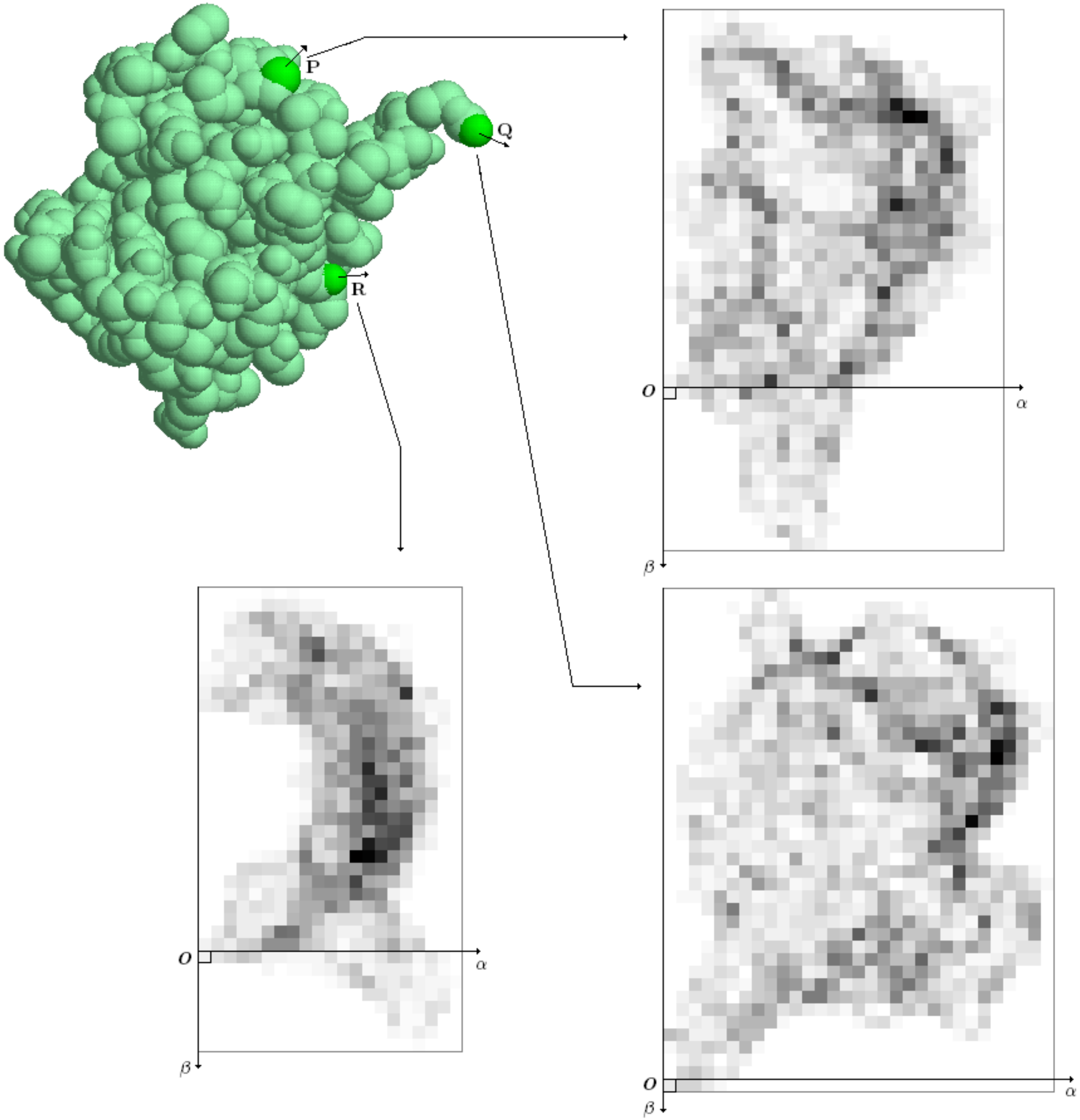
M. E. Bock, C. Garutti, C. Guerra, J. of Computational Biology, 2007.

An oriented point basis



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

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


P
Q
R
O
α
β
O
α
β
O
α
β

Comparing spin images

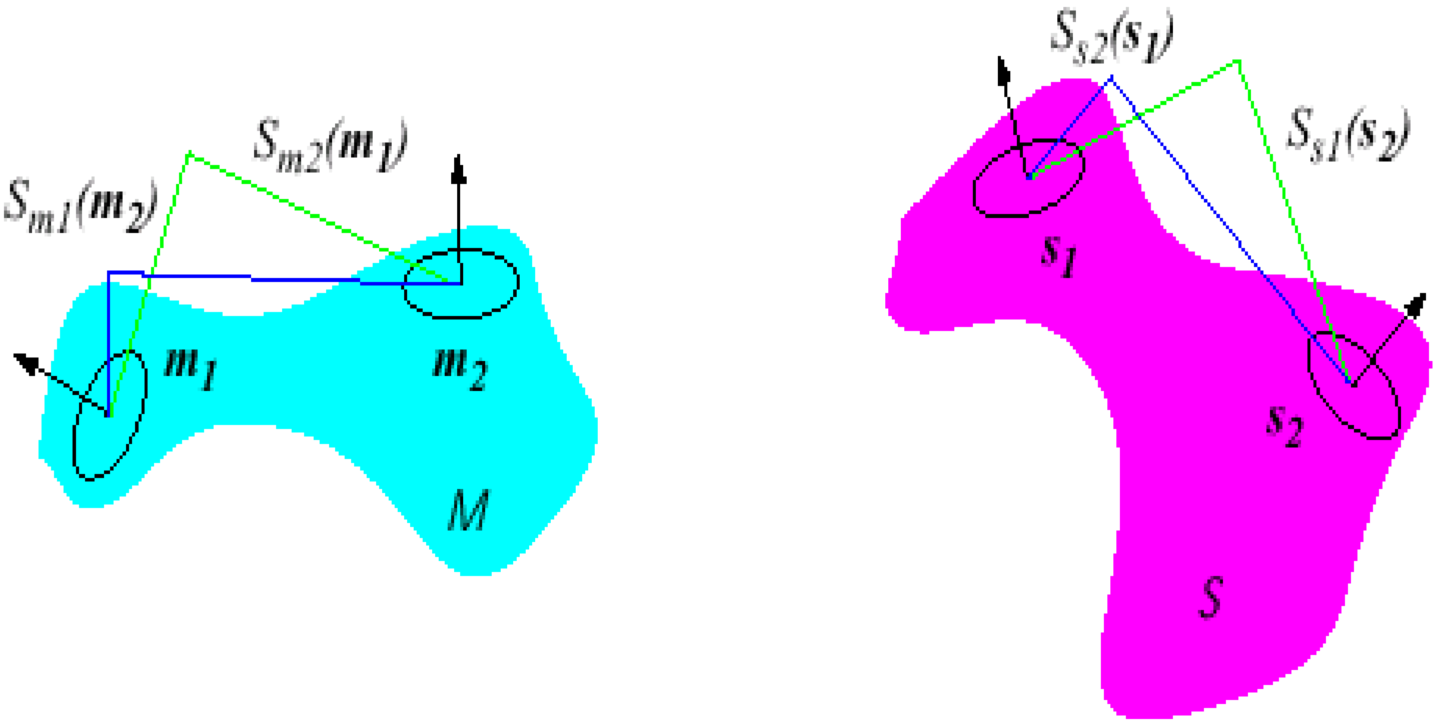
Surfaces with similar shape tend to have similar spin images

Given two spin-images *P* and *Q* with *N* bins each, compare them using

- correlation coefficient
- *Euclidean distance*

Grouping Point Correspondences for surface matching

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



Geometric Matching 1

A three-step procedure:

1. Establish individual pointcorrespondences based on the correlation of the spin images
2. Group point correspondences using a geometric consistency criterion

   Use a greedy algorithm that grows regions around selected point correspondences

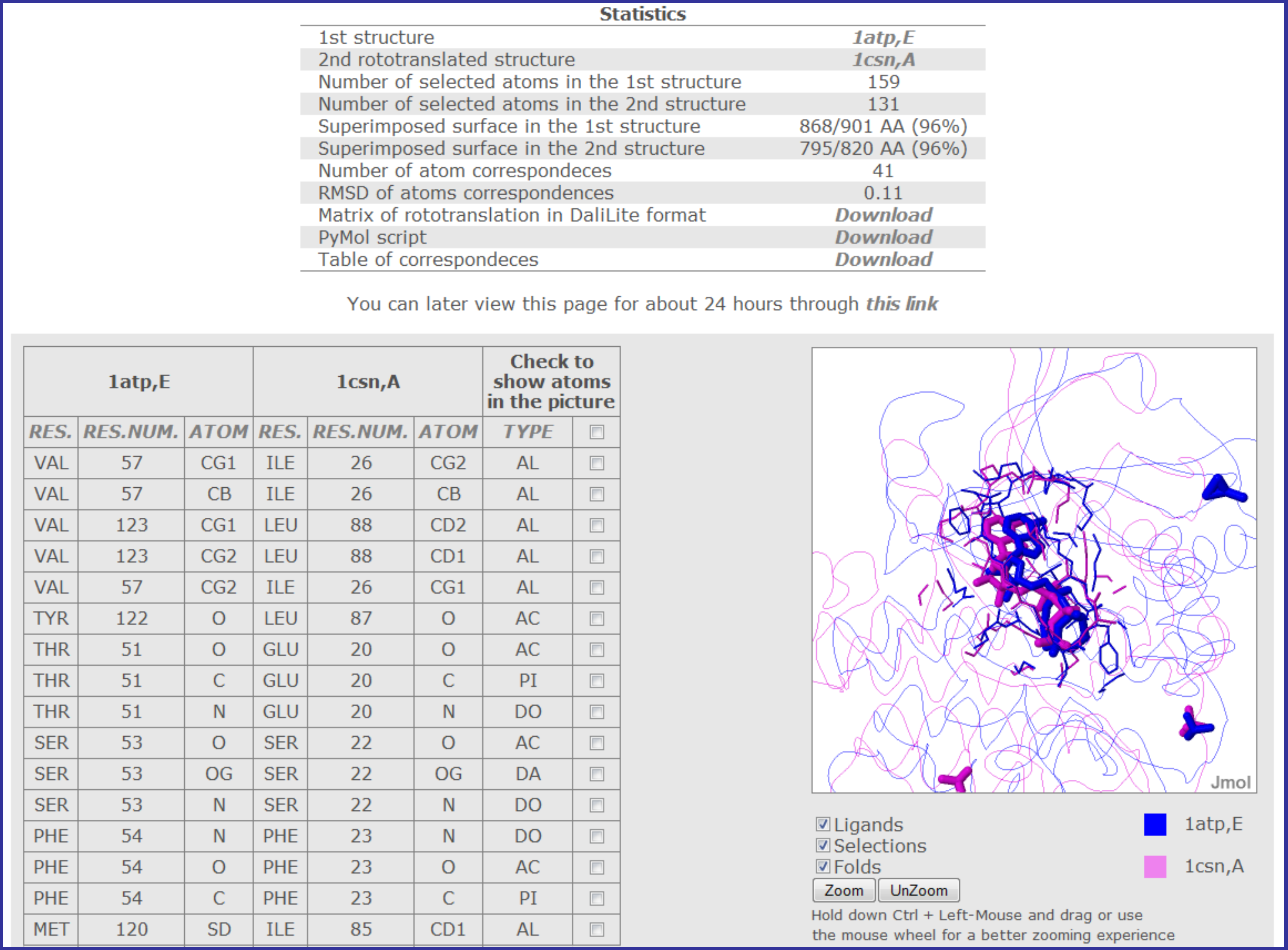
3. Score each group by the number of pairs of corresponding points.

Geometric Matching 2

As above, but borrespondences are restricted to points with the same physico-chemical properties

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	☐

☑ Ligands
☑ Selections
☑ Folds

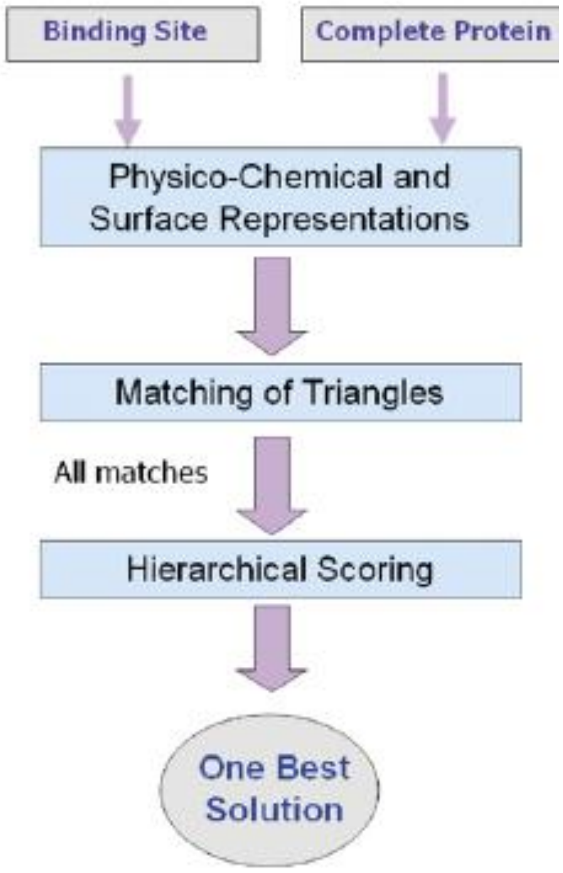
🟦 1atp,E
🟪 1csn,A

Hold down Ctrl + Left-Mouse and drag or use the mouse wheel for a better zooming experience

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

SiteEngine: Functional Site Recognition

- based on hashing of triangles of centers of physico-chemical properties.



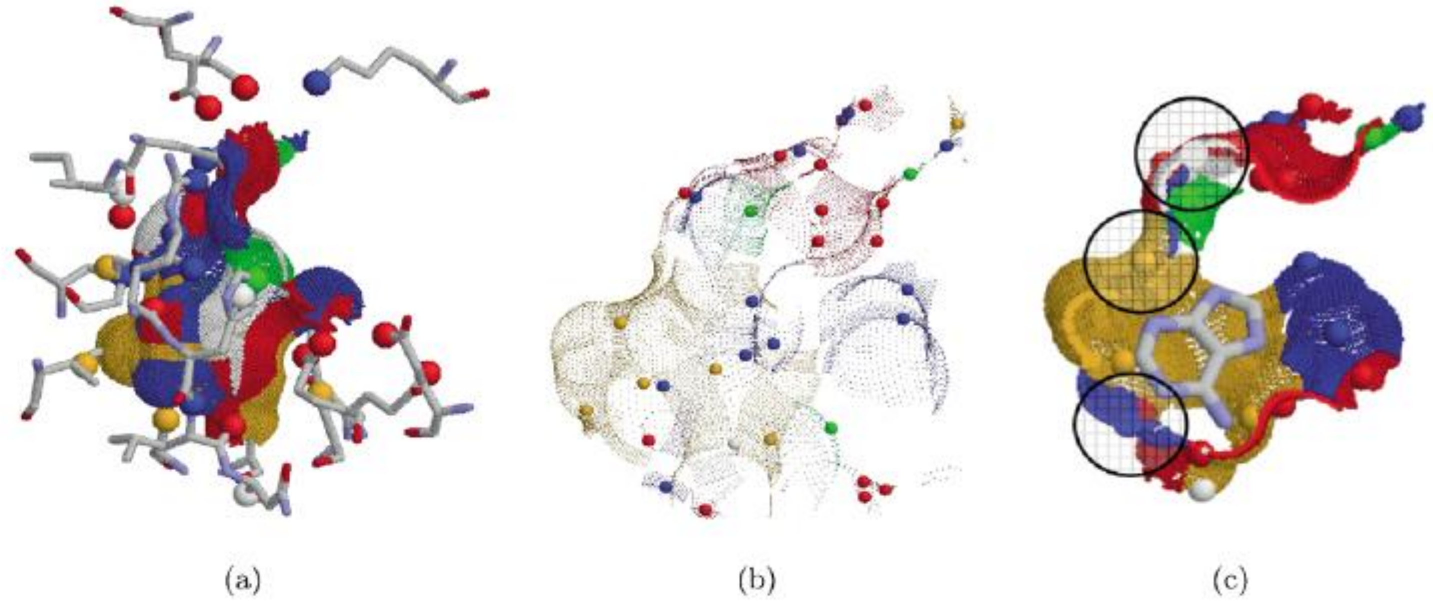
A Shulman, R Nussinov H. Wolfson, JMB, 2004

Pseudo-centers

3D points of residues representing one of the properties:

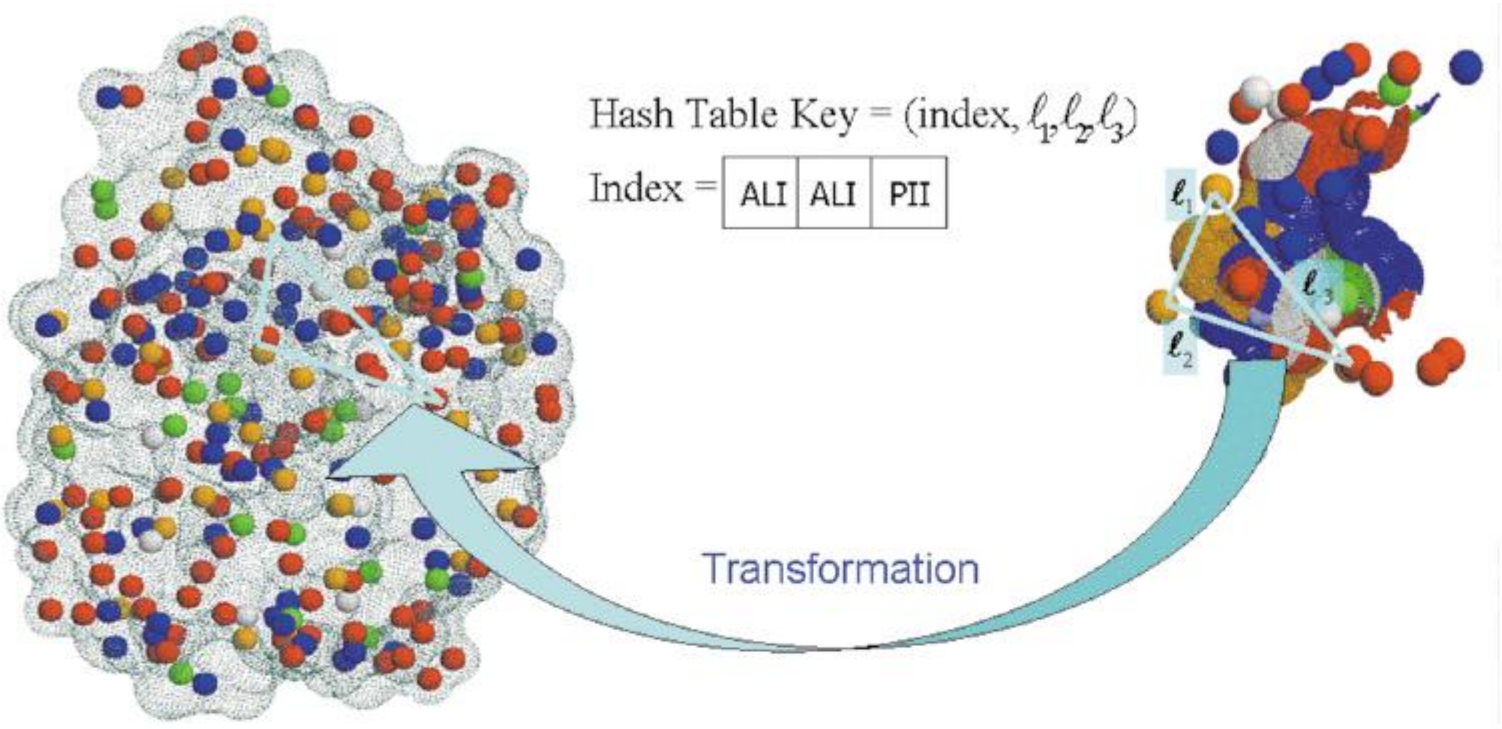
- hydrogen-bond donor
- hydrogen-bond acceptor
- mixed donor/acceptor
- hydrophobic aliphatic and aromatic(pi) contacts.

Physico-chemical representation by pseudocenters



Hydrogen-bond donors
acceptors I
donors/acceptors
hydrophobic aliphatic in orange and aromatic

Geometric Hashing of Triangles



Hash Table Key = (index, ℓ_1, ℓ_2, ℓ_3)

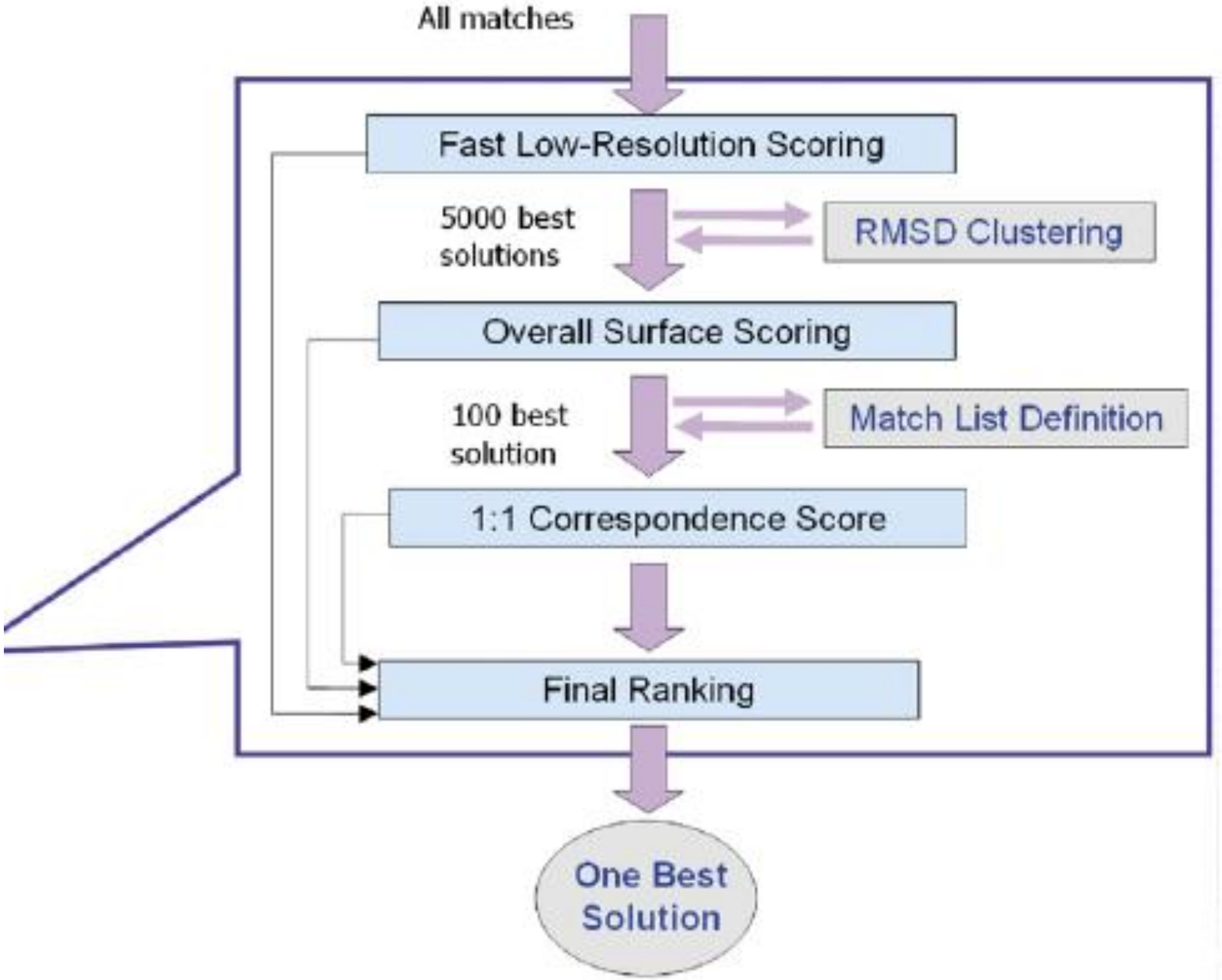
Index = | ALI | ALI | PII |

Hashing

Consider triplets of non-ordered non-colinear pseudocenters

- Triplets that form triangles with side lengths within a predefined range are stored in a hash table.
- A key to the hash table consists of the three parameters of side lengths of a triangle and of an additional physico-chemical index

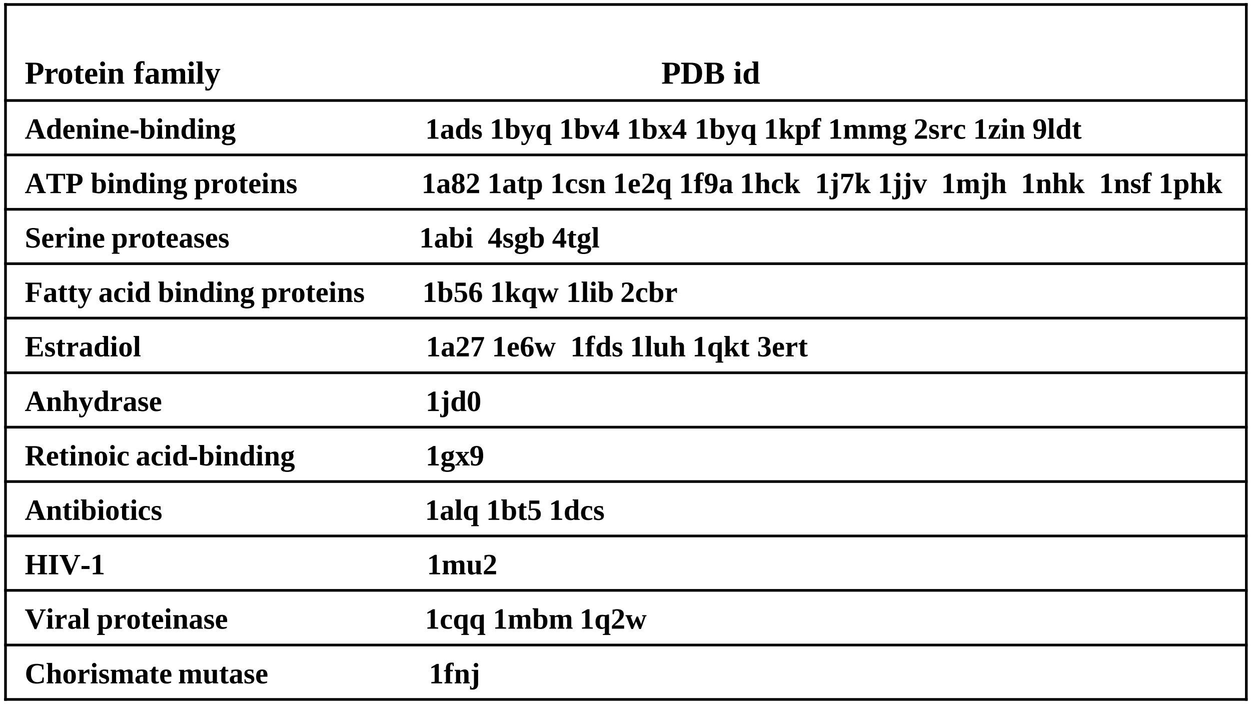
Hierarchical Scoring for local & global similarity



Experimental Results

Data set of protein complexes

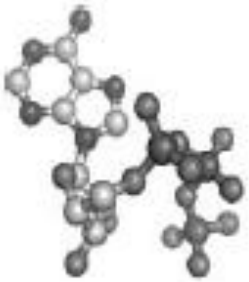
(Wolfson et al, 2005)



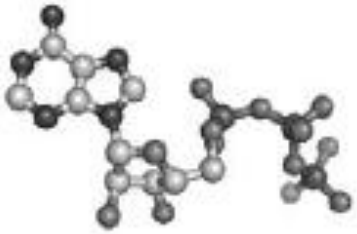
Protein family	PDB id
Adenine-binding	1ads 1byq 1bv4 1bx4 1byq 1kpf 1mmg 2src 1zin 9ldt
ATP binding proteins	1a82 1atp 1csn 1e2q 1f9a 1hck 1j7k 1jjv 1mjh 1nhk 1nsf 1phk
Serine proteases	1abi 4sgb 4tgl
Fatty acid binding proteins	1b56 1kqw 1lib 2cbr
Estradiol	1a27 1e6w 1fds 1luh 1qkt 3ert
Anhydrase	1jd0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

Different conformations of ATP

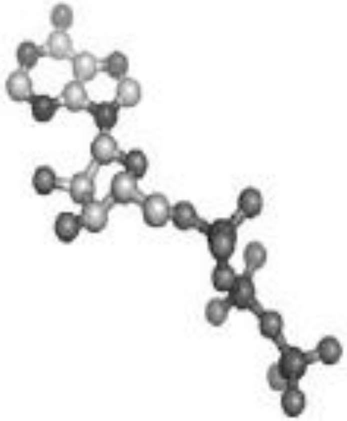
compact



intermediate



extended



Conformational Diversity of Ligands Bound to Proteins

Stockwell, Thornton *J. Mol. Biol. (2006)*

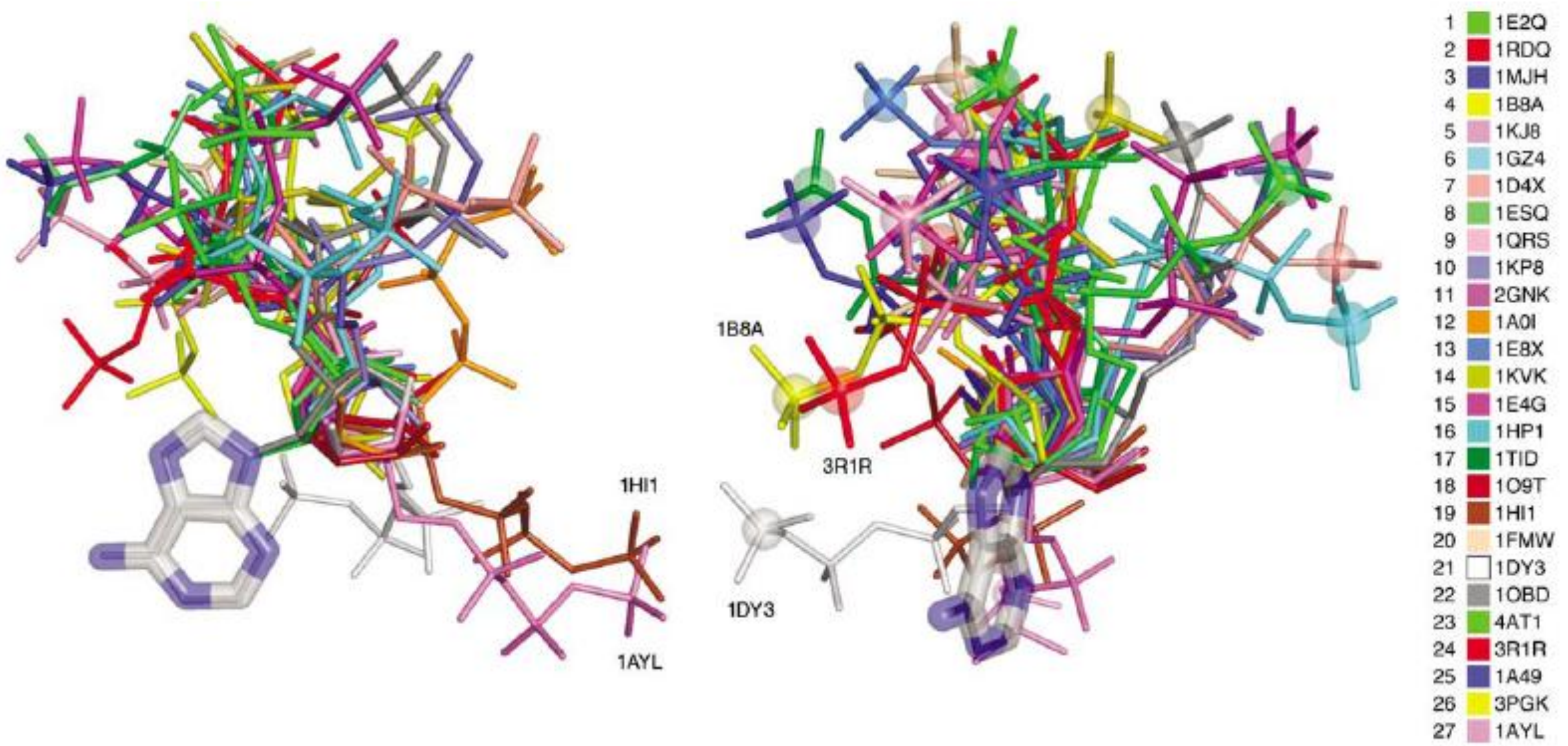
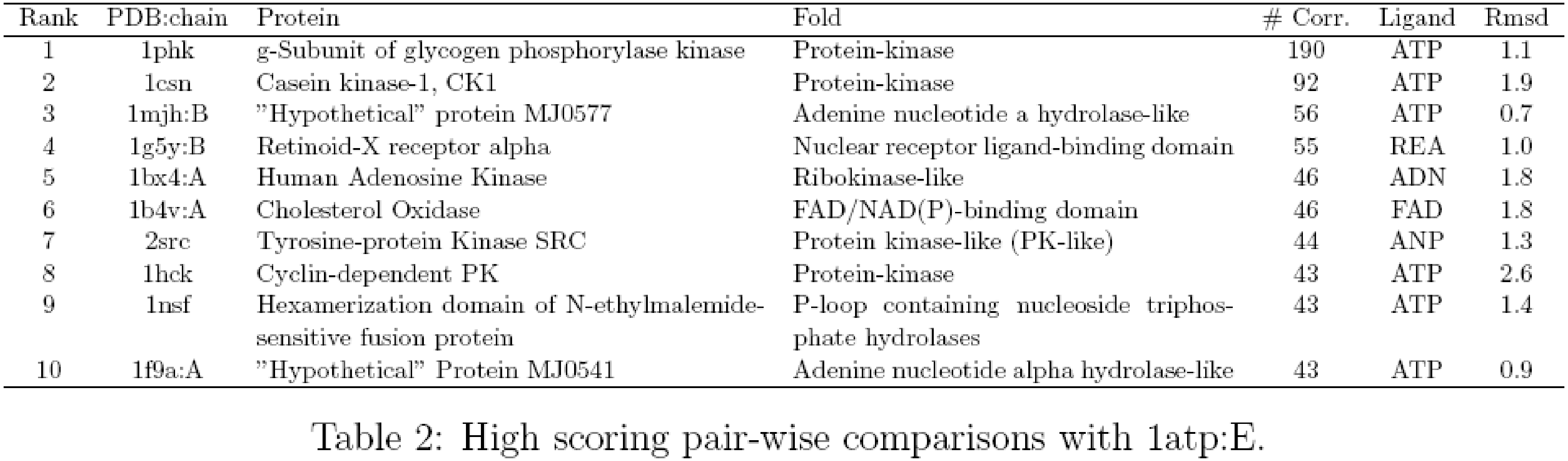


Figure 1. Superposition of the 27 ATP cluster representatives on their adenine rings (highlighted). In the second image, the gamma phosphate atoms are shown with translucent spheres, to highlight the broad range of conformations adopted by the triphosphate tail. The key shows from which PDB entry each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

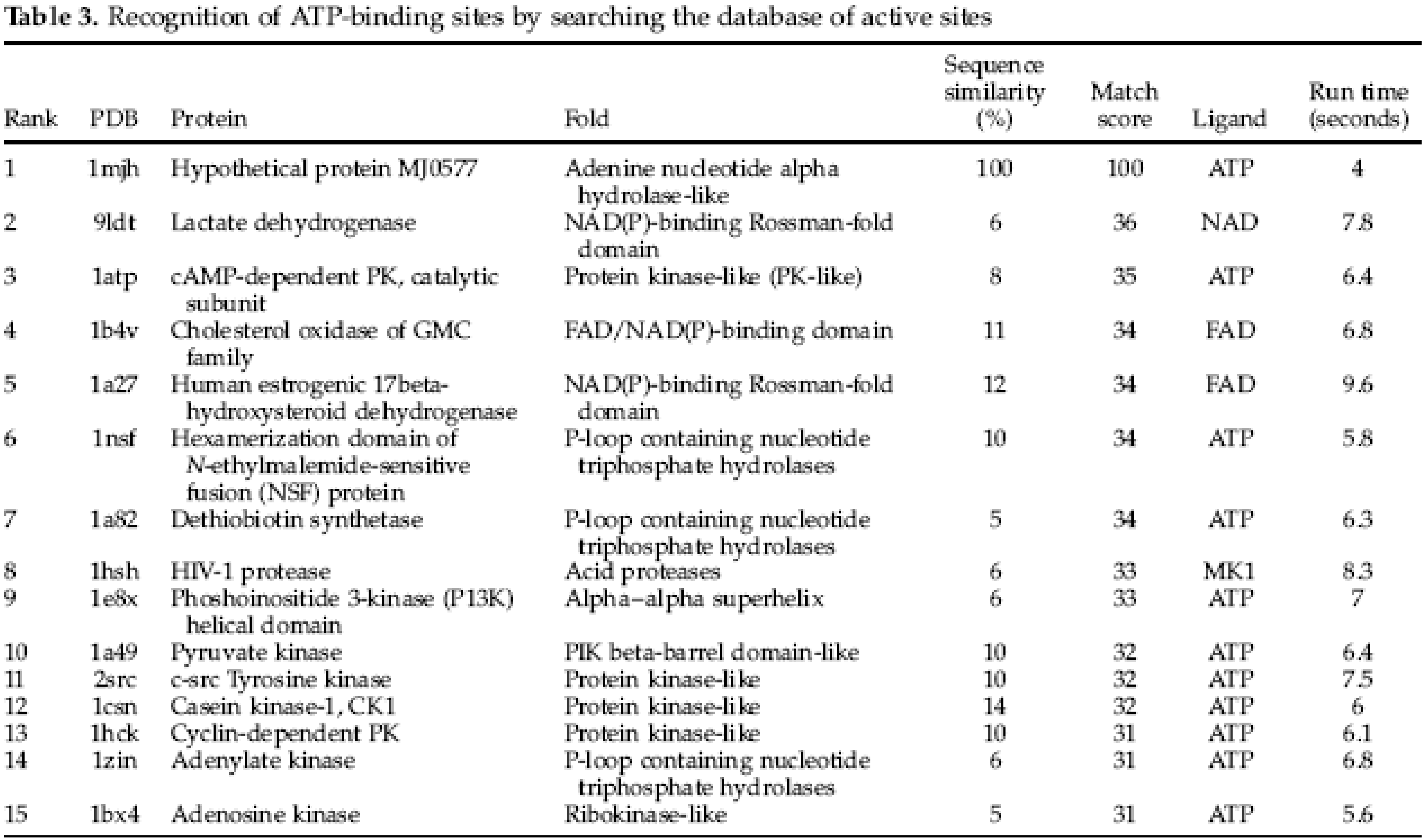


Table 3. Recognition of ATP-binding sites by searching the database of active sites

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

How to evaluate the results of a classifier?

- Accuracy/ coverage
- ROC curves
- Distance matrices

Accuracy vs coverage

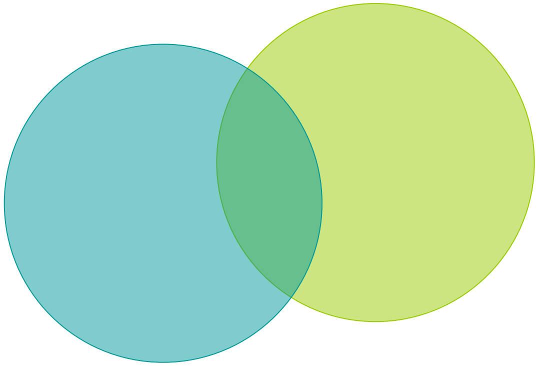
- Accuracy: how many of the solutions found were correct?

  $A= (F \cap T) /F$

- Coverage: How many of the correct solutions were found?

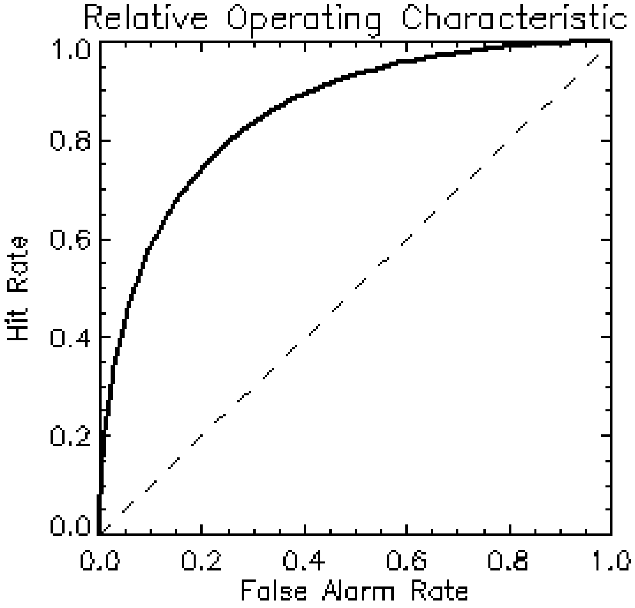
  $C= (F \cap T) /T$

T: correct sol. F:solutions found

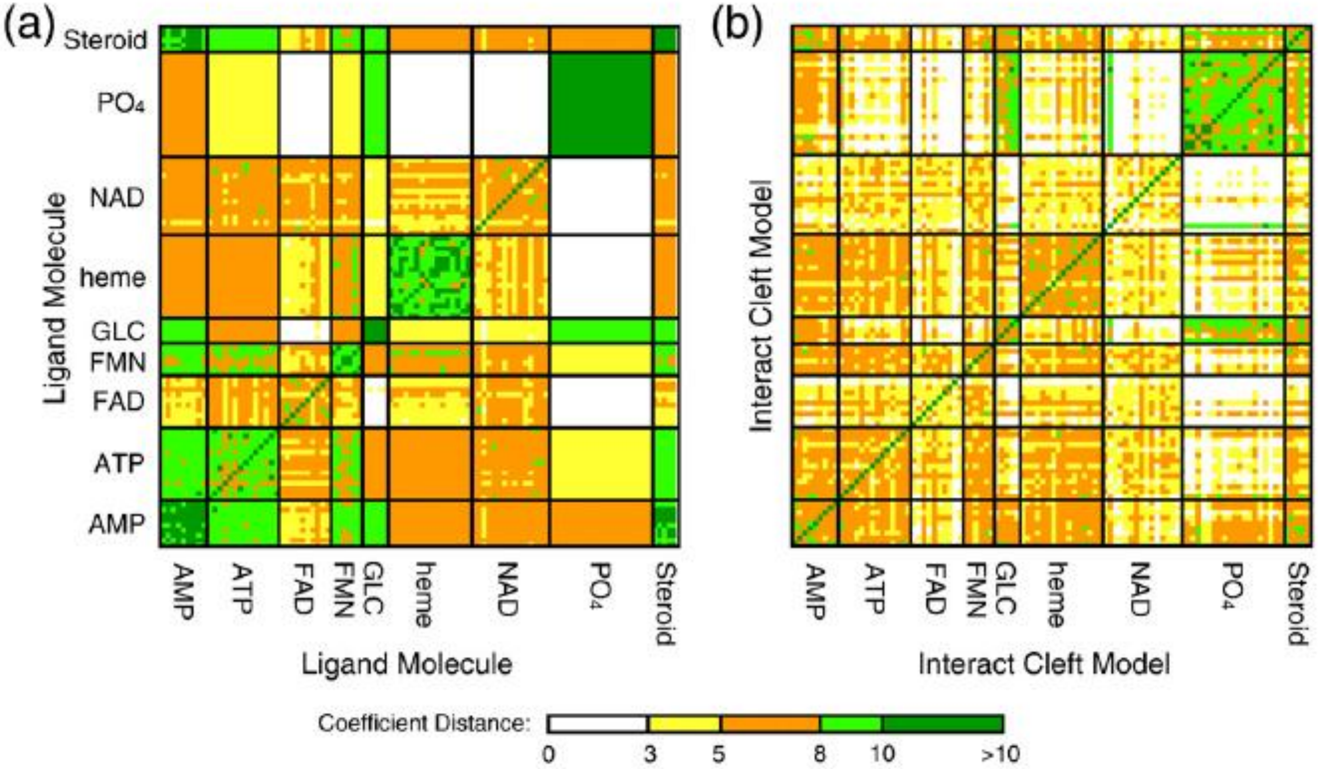


Receiver Operating Characteristic (ROC) curves

The ROC curves display the fraction of true positives or correct answers versus the fraction of false positives for all positions of the ranked solutions

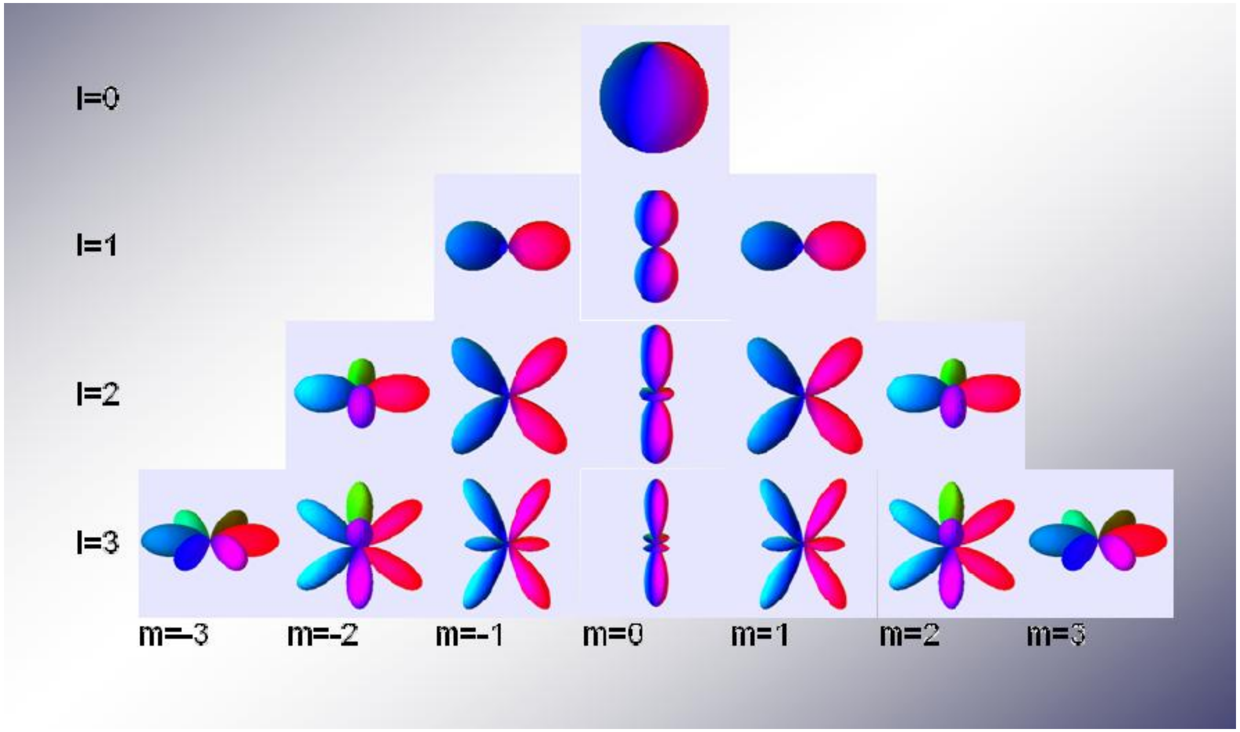


Distance matrices
All-against-all

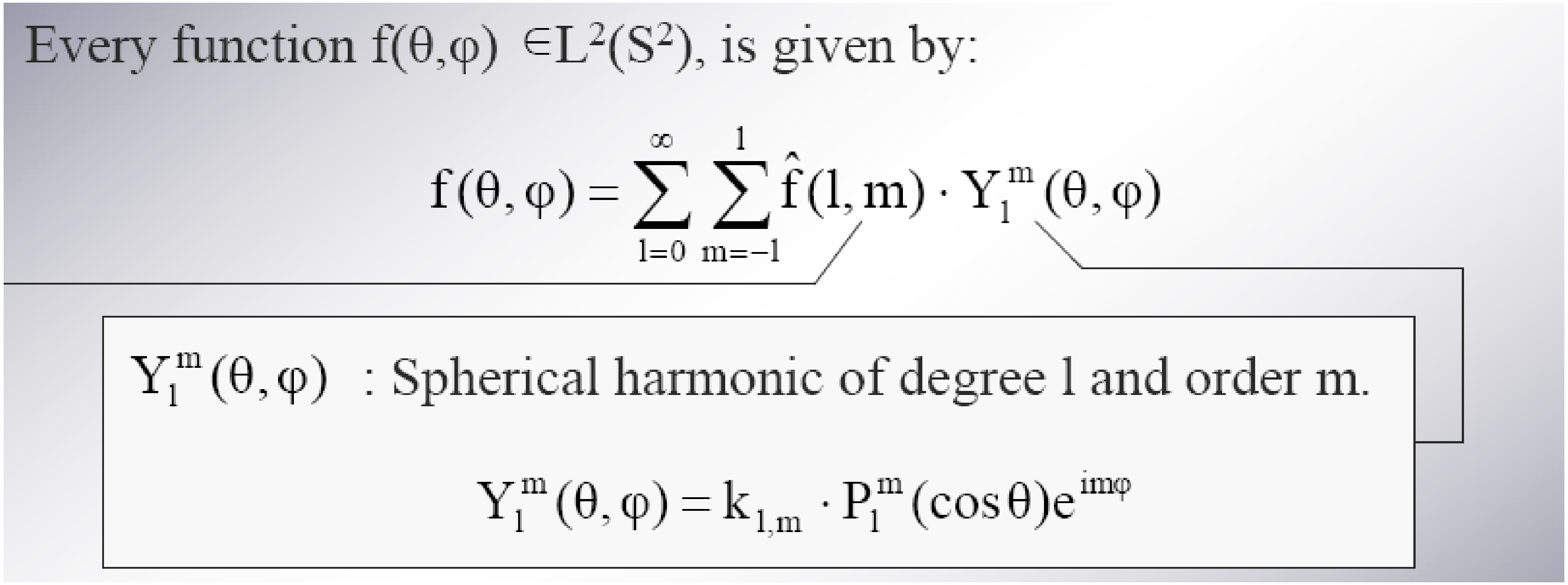


Binding Site Comparison by Spherical harmonics

Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons,



Spherical harmonics



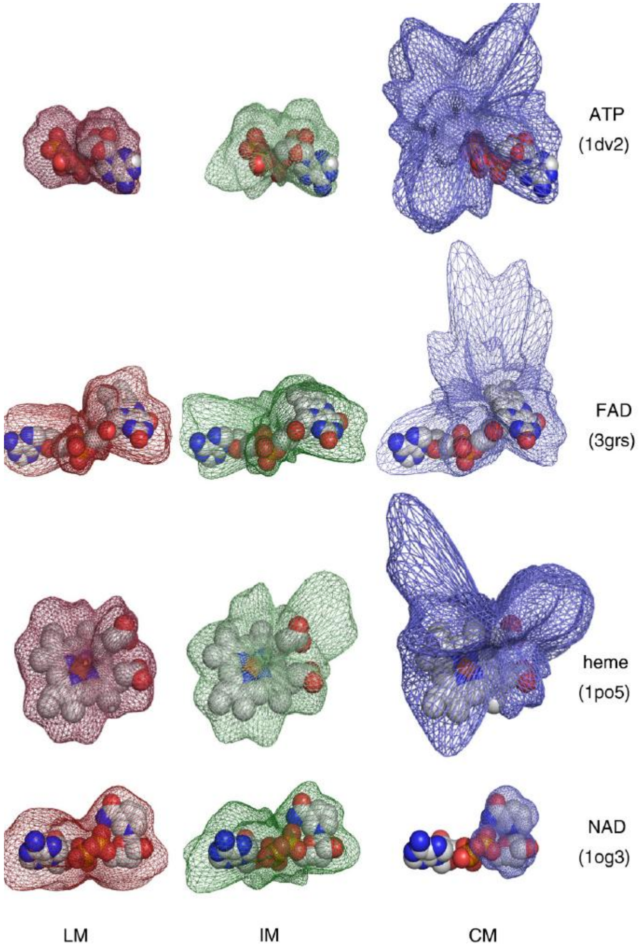
Every function $f(\theta,\varphi) \in L^2(S^2)$, is given by:

$$f(\theta,\varphi) = \sum_{l=0}^{\infty} \sum_{m=-1}^{1} \hat{f}(l,m) \cdot Y_1^m(\theta,\varphi)$$

$Y_1^m(\theta,\varphi)$: Spherical harmonic of degree l and order m.

$$Y_1^m(\theta,\varphi) = k_{1,m} \cdot P_1^m(\cos\theta)e^{im\varphi}$$

Results with l=14



R.J. Morris, R.J. Najmanovich, A. Kahraman and J.Thornton (2005), Bioinformatics

Clustering Proteins based on Shperical Harmonics

The expansion coefficients can be used as a feature vector or shape signature.

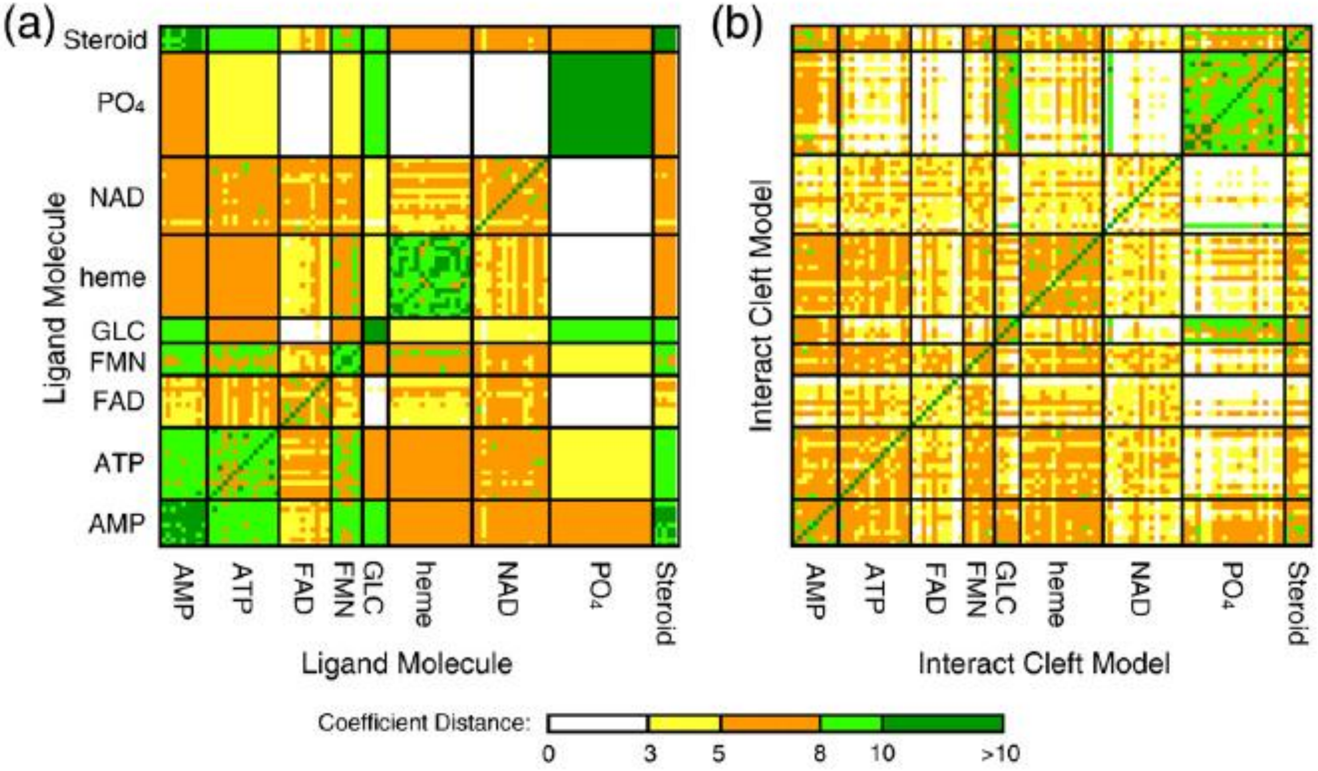
Protein shapes are classified based on the L2 distance in coefficient space

A registration phase is used to align two binding sites prior to comparing them

Cai, W., Shao, X., and Maigret, B. (2002). J. Mol. Graphics Modelling.

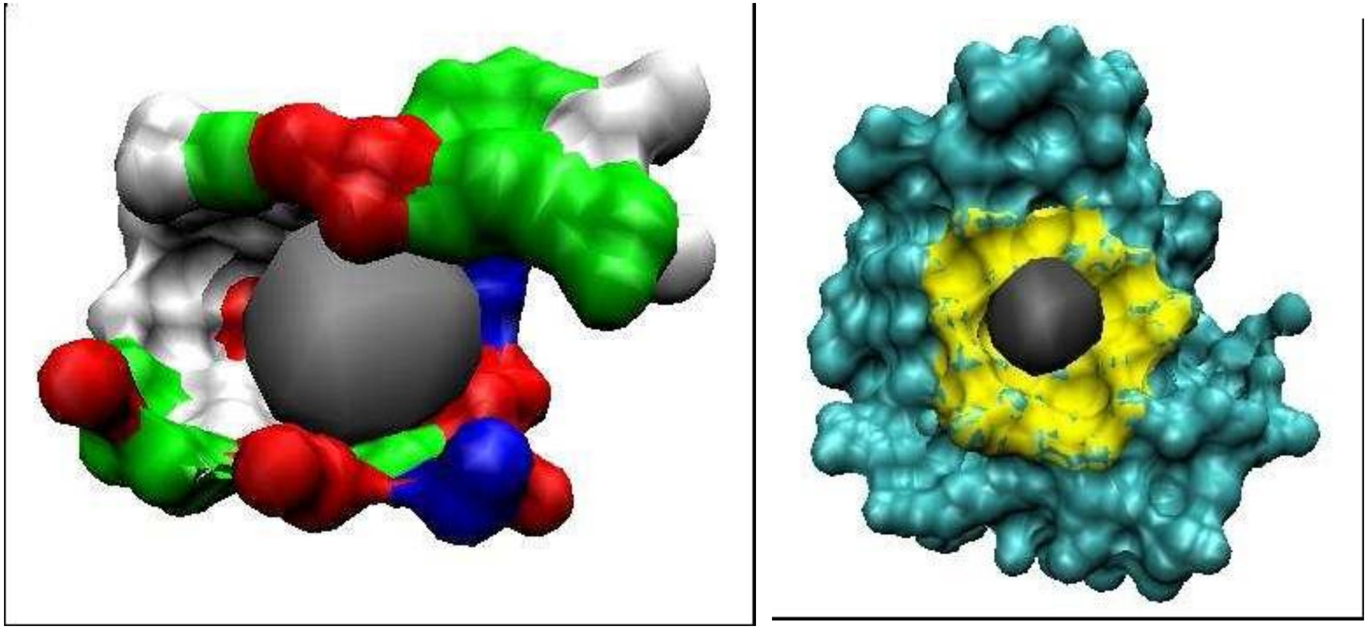
Leicester, S., Finney, J., and Bywater, R. (1994b).J. of Math. Chemistry

All-against-all comparison of binding sites



Binding Balls

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



M. Comin, F. Dellaert, C. Guerra. J. of Computational Biology, 2009.

Binding Site Recognition using Spherical harmonics and Binding Balls

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.

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 (2010), BMC Bioinformatics (to appear)

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

Geometric Matching 1

A two-step procedure:

- an initial population of points (defining roto-translations in three-dimensional space) is generated by randomly sampling a sufficiently large set of points
- At every iteration, a new point is generated and the population is updated if this new point improves on the worst point of the population.

More details
Search Phase

- N + 1 points are randomly chosen in the set S. Then,
- (a) the weighted centroid a_c of the N + 1 points is computed;
- (b) the new trial point a* is computed by a weighted reflection of the centroid onto the worst point among the selected N + 1 points.

Finding surface cavities and binding pockets

For protein/drug interaction

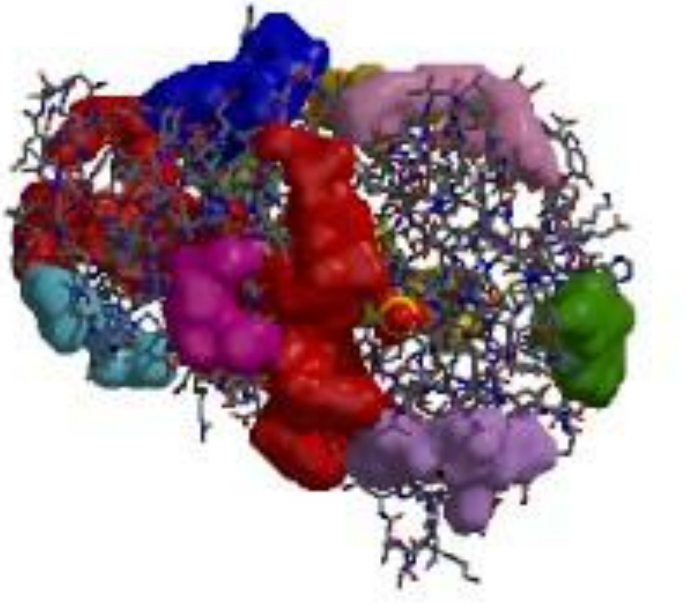
- **SPHGEN**, **Surfnet**

determine sphere clusters

- **CastP**

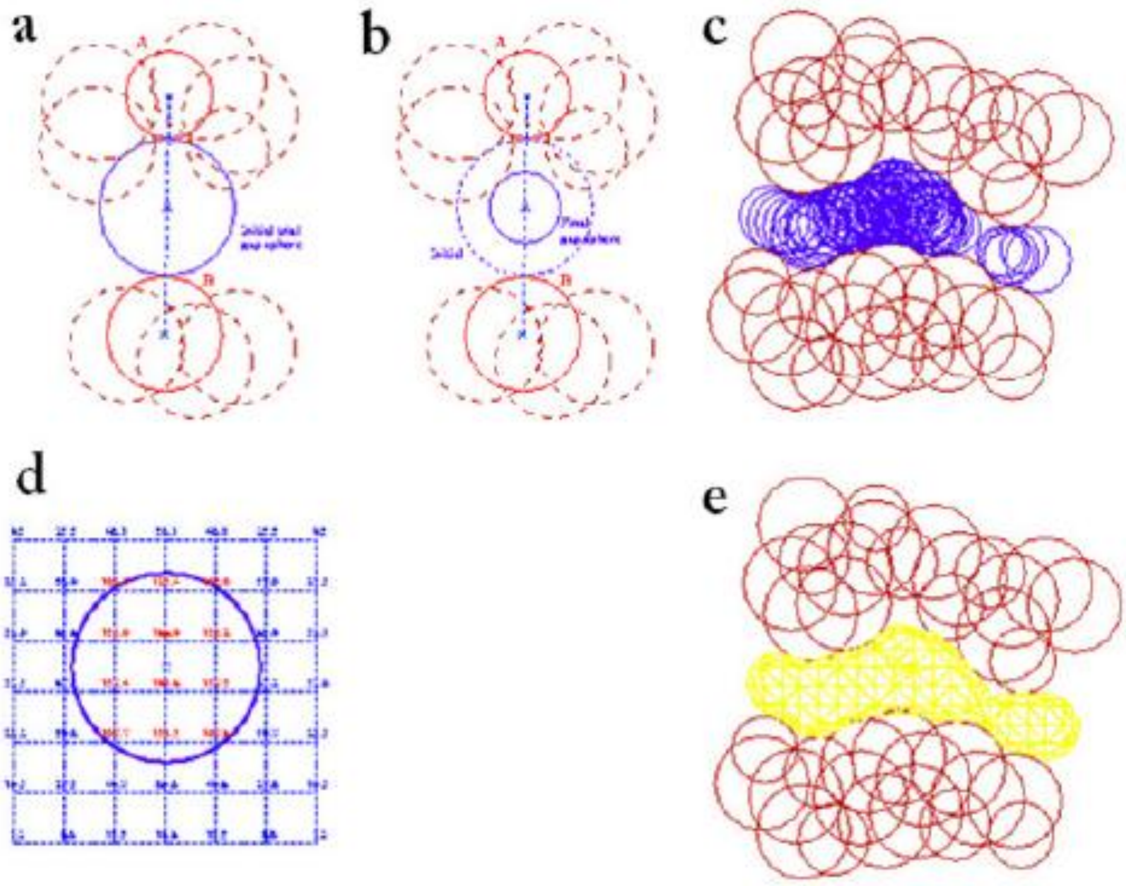
Alpha Shapes

- **SpinImages**



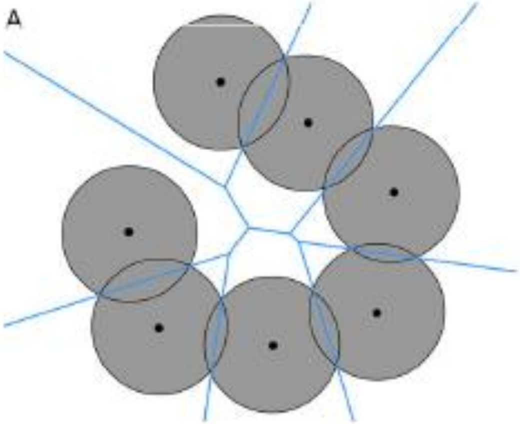
Surfnet

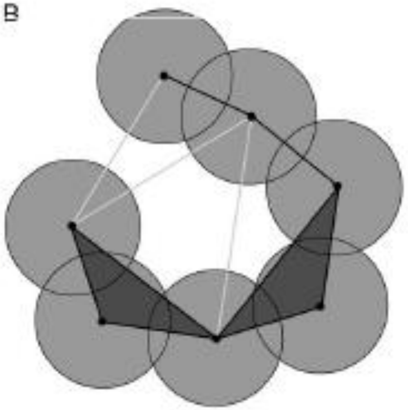
(Laskowski et al 2005)

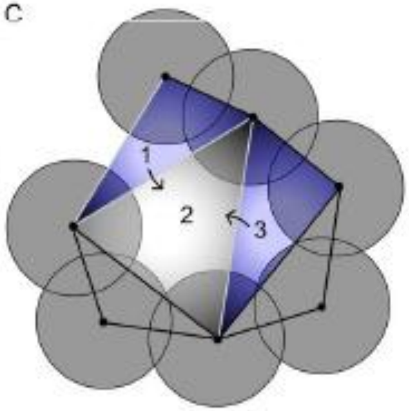


CastP
(Binkowski et al 2003)

Based on alpha-shapes



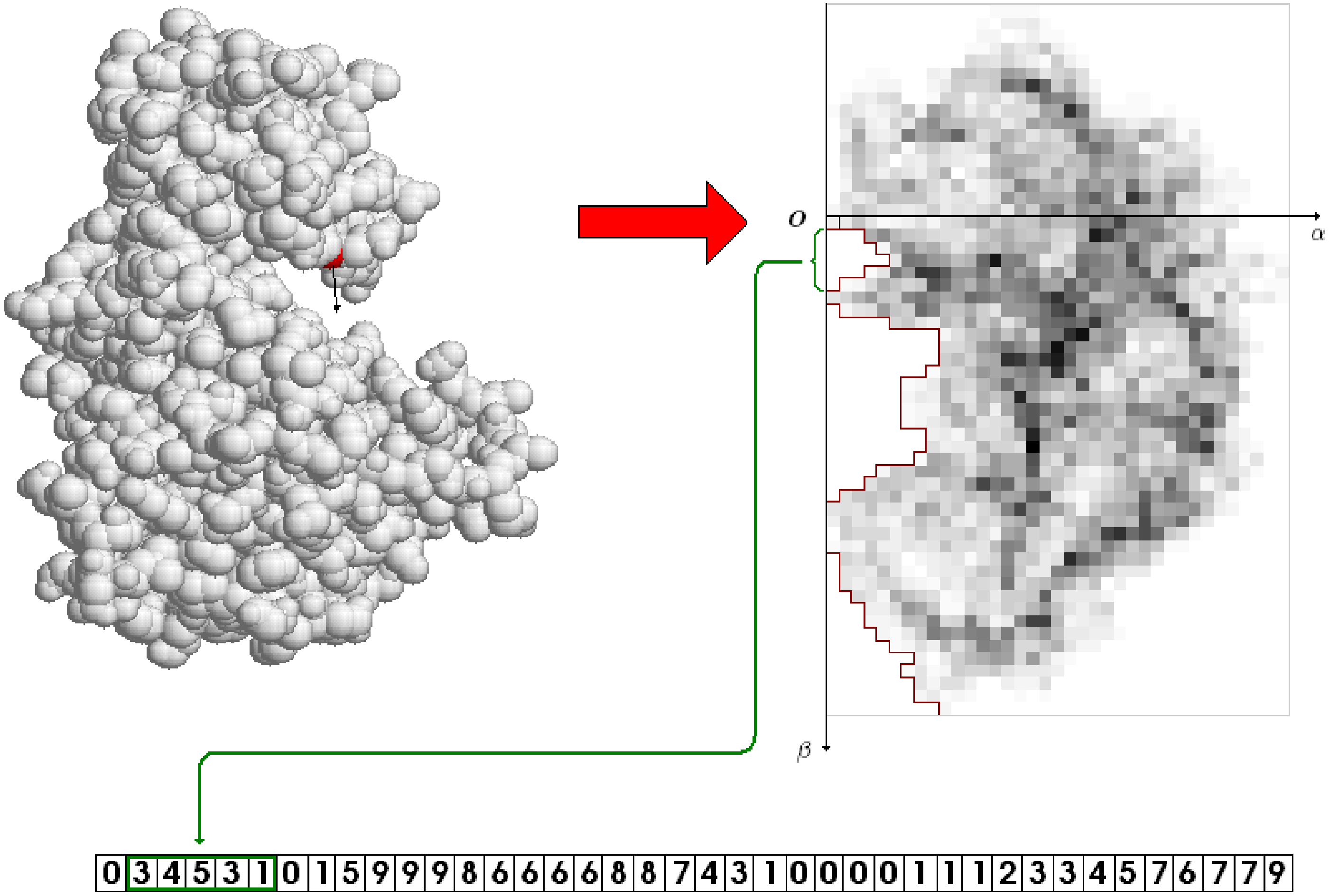




Cavity detection using spin image profiles

(Bock et al 2007)

Find the largest sphere that can fit into the empty space



0	3	4	5	3	1	0	1	5	9	9	9	8	6	6	6	6	8	8	7	4	3	1	0	0	0	0	1	1	1	2	3	3	4	5	7	6	7	7	9

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**

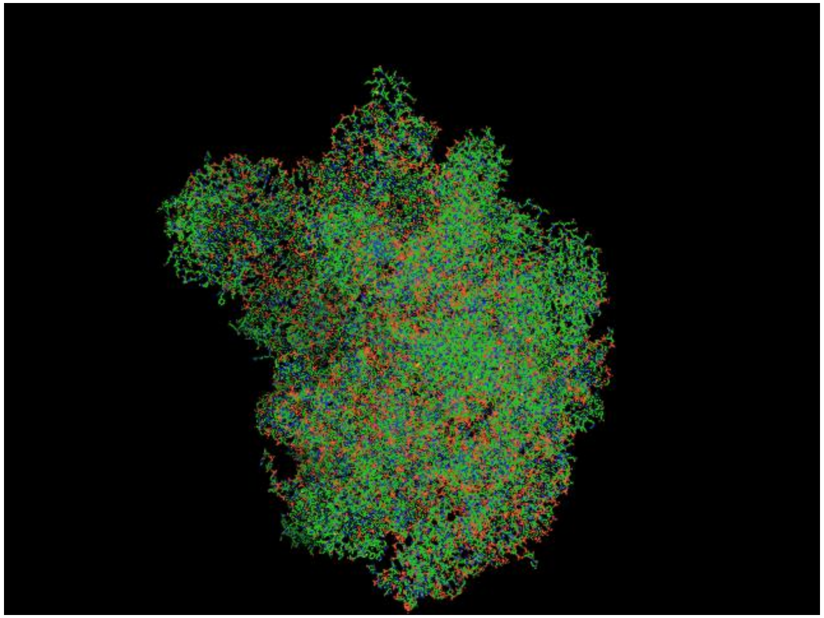
Protein-ligand Interactions Conclusion

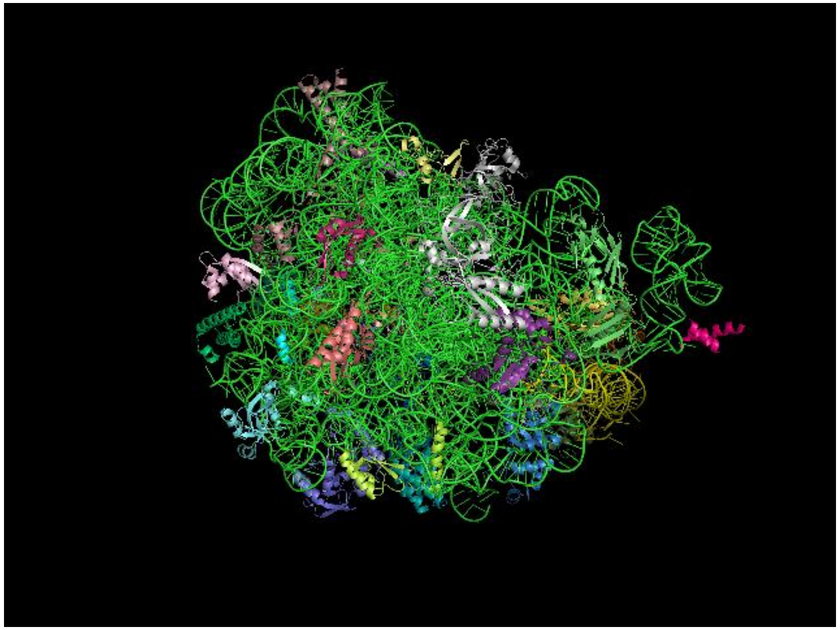
- Variety of shape descriptors and shape matching methods developed in computer vision
- Adaptation to protein analysis far from trivial
- Results on protein surface comparison based on geometry only comparable to those based on a combination of geometry and physico-chemical properties.

Interactions of ribosomal RNA with proteins

Ribosomal RNA of Haloarcula Marismortui

- Two subunits: 23S e 5S
- 28 Ribosomal proteins (r-proteins)





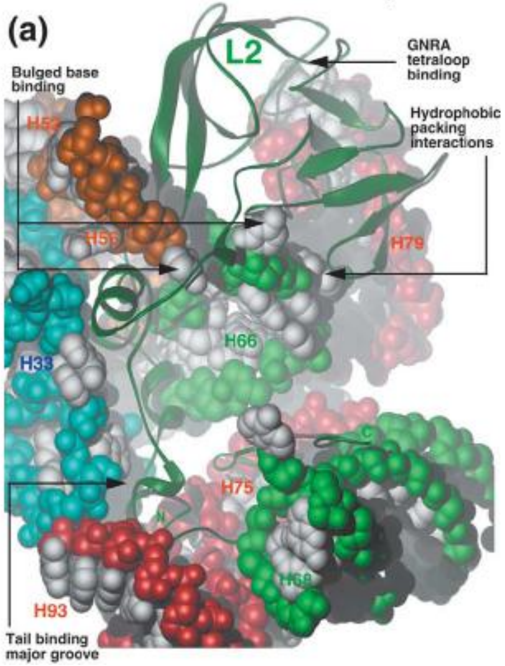
The 28 protein have different colours

The structure of ribosomal proteins

The protein structures fall into six groups based on their topology.

The single most striking feature of the r-proteins in the large subunit are the many long extensions

- they represent only 18% of the proteins
- but are responsible for 44% of the RNA buried surface area



D. J. Klein, P. B. Moore, T. A. Steitz (2004), JMB.

The function of ribosomal proteins

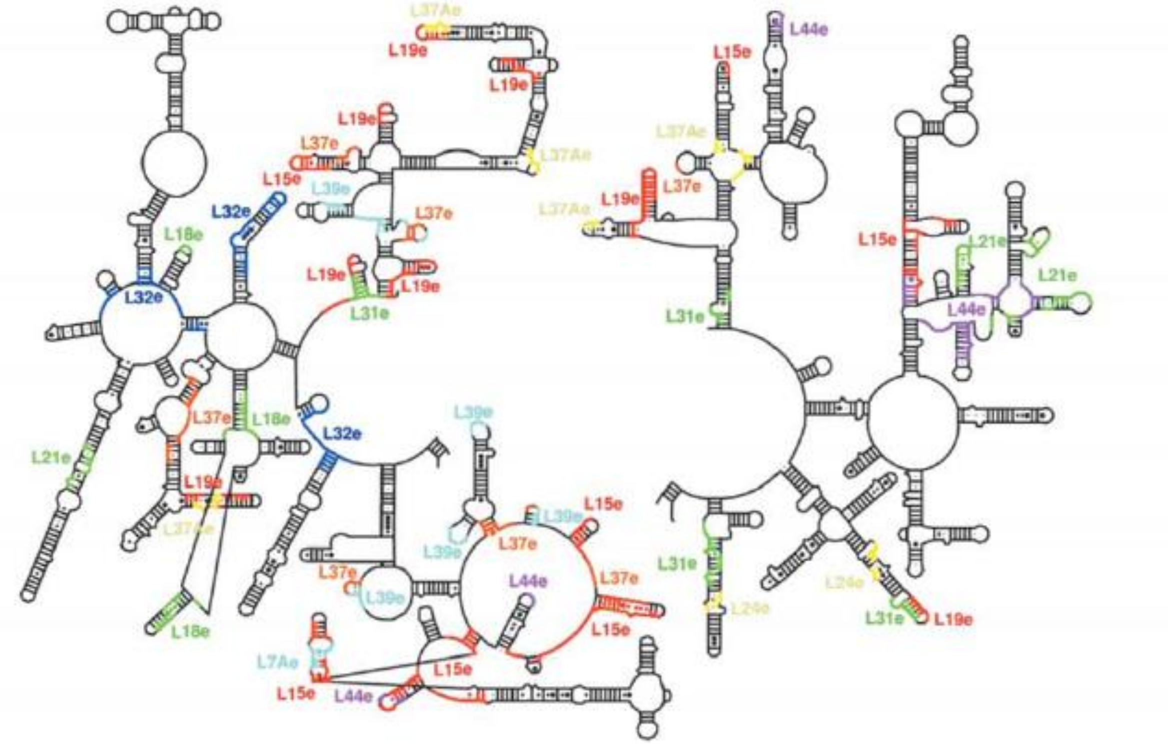
The 50 S subunit proteins function primarily

to stabilize inter-domain interactions that are necessary to maintain the subunit’s structural integrity.

- Understand the assembly process

- Provide insight into ribosome evolution

Proteins typically contact sites in several domains



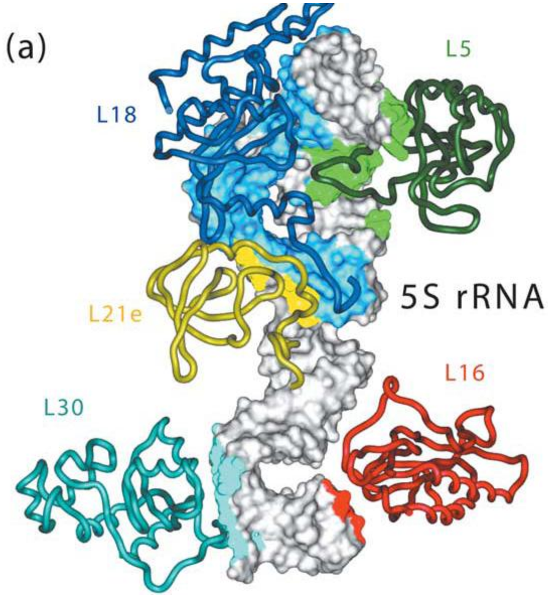
Sites are distinguished by color

Figure from: D. J. Klein, P. B. Moore, T. A. Steitz (2004), The Roles of Ribosomal Proteins in the Structure Assembly, and Evolution of the Large Ribosomal Subunit, JMB.

RNA-binding sites of r-proteins

High variety of protein–RNA interactions is observed in Haloarcula Marismortui

The size of the buried surface area varies greatly among the r-proteins



Role of RNA structural motifs in the interaction of proteins

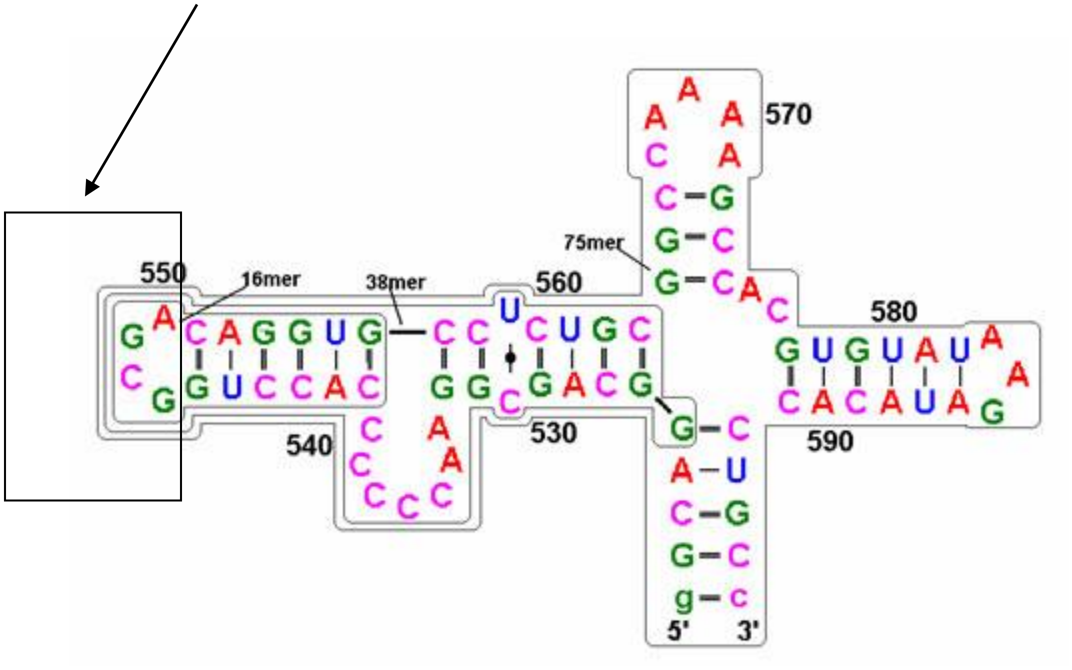
Consider the interface regions involving motifs such as **tetraloops, kink turns and single extruded nucleotides** and analyze their

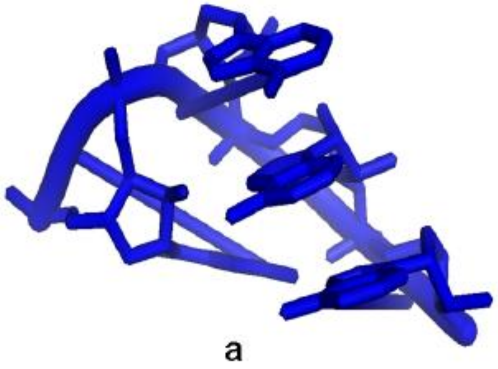
- composition
- local geometries
- 3D conformation

Ciriello,G., C. Gallina, C, Guerra,C., (2010),, BMC Bioinformatics

RNA motifs - Tetraloops

A tetraloop is a contiguous fragment of 4 nt of non-helical RNA which terminates a single helix.

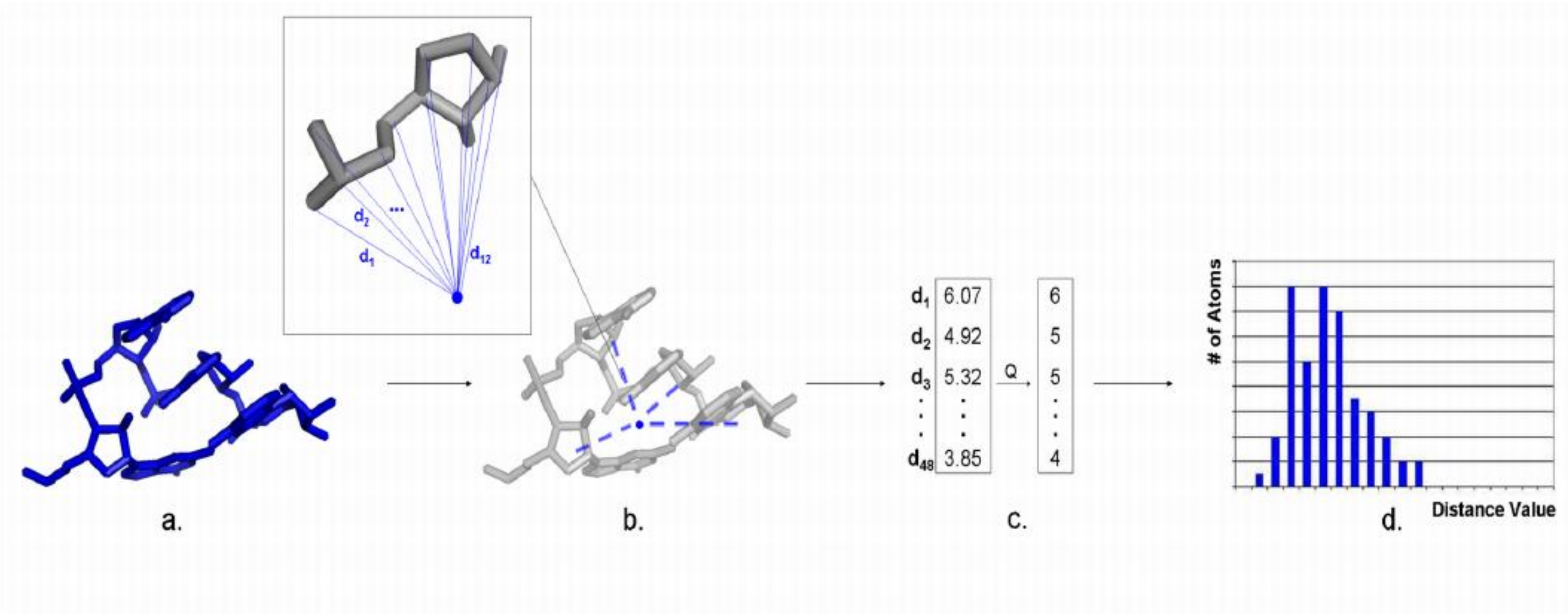




Secondary structure representation

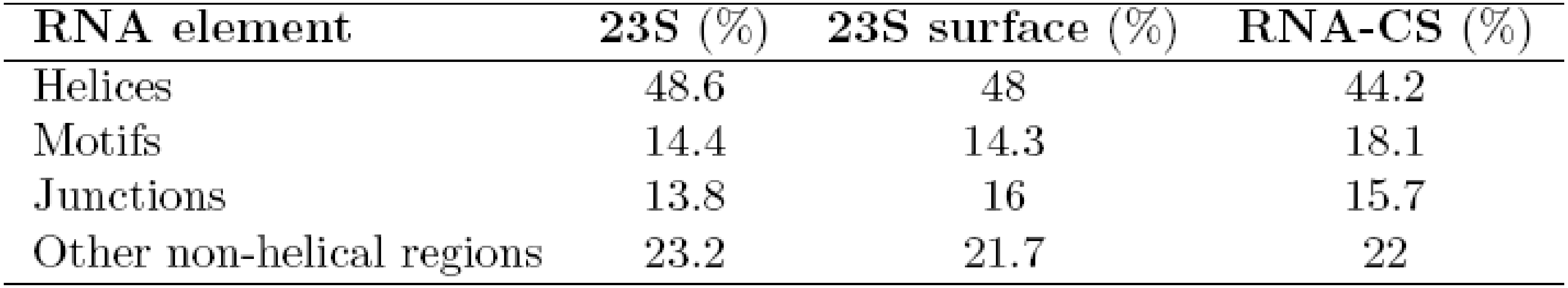
3D representation

Finding 3D motifs in ribosomal RNA structures



Apostolico,A., Ciriello,G., Guerra,C., Heitsch,C.E., Hsiao,C. and Williams, L.D. (2009), NAR.

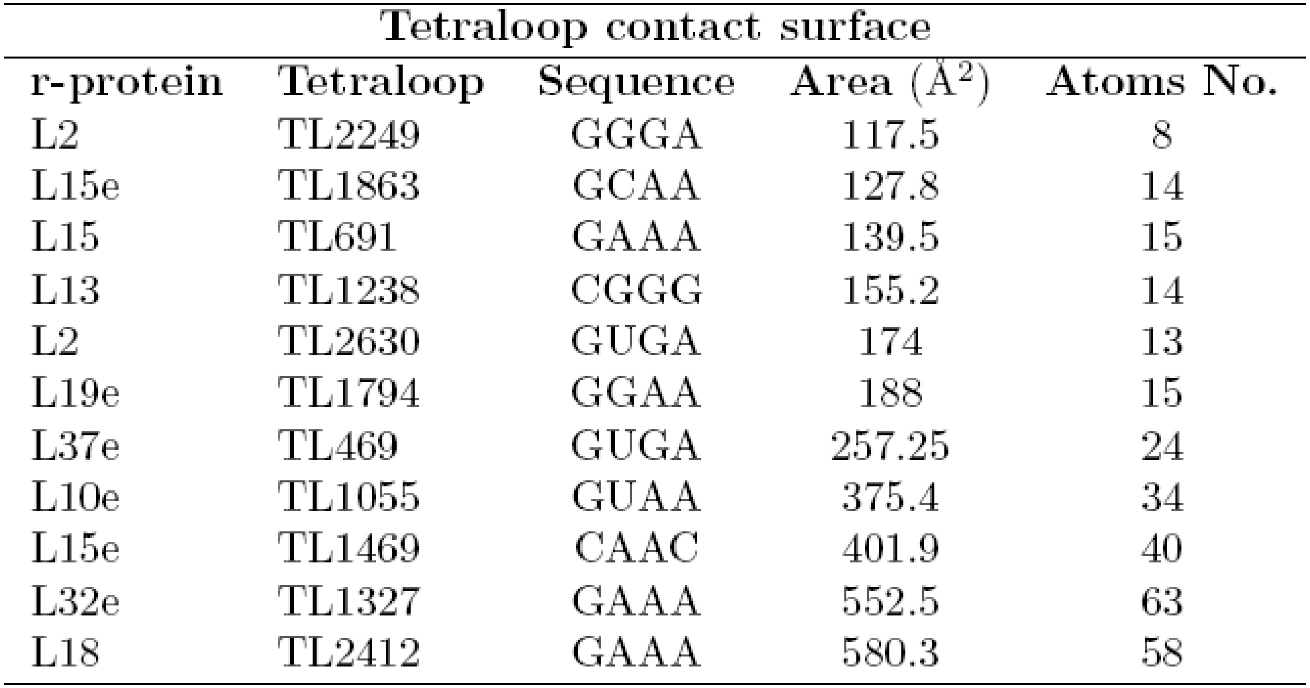
Frequency of Structural Motifs at interfaces



RNA element	23S (%)	23S surface (%)	RNA-CS (%)
Helices	48.6	48	44.2
Motifs	14.4	14.3	18.1
Junctions	13.8	16	15.7
Other non-helical regions	23.2	21.7	22

RNA-CS = RNA contact surfaces

Geometry of Interfaces with Tetraloops



Tetraloop contact surface

r-protein	Tetraloop	Sequence	Area (Å^2)	Atoms No.
L2	TL2249	GGGA	117.5	8
L15e	TL1863	GCAA	127.8	14
L15	TL691	GAAA	139.5	15
L13	TL1238	CGGG	155.2	14
L2	TL2630	GUGA	174	13
L19e	TL1794	GGAA	188	15
L37e	TL469	GUGA	257.25	24
L10e	TL1055	GUAA	375.4	34
L15e	TL1469	CAAC	401.9	40
L32e	TL1327	GAAA	552.5	63
L18	TL2412	GAAA	580.3	58

Chemical Composition

RNA side- distribution of phosphate-ribose-base atoms

- 80% of interacting atoms are backbone atoms, i.e. P and R
- 73% of interacting atoms in regions consisting of structural motifs are backbone atoms

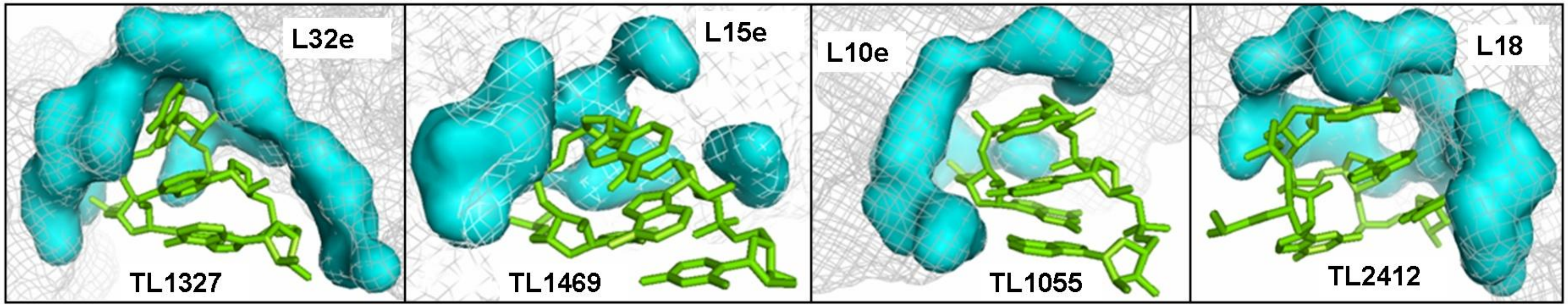
Protein side - amino acid composition

Interfaces with tetraloops:

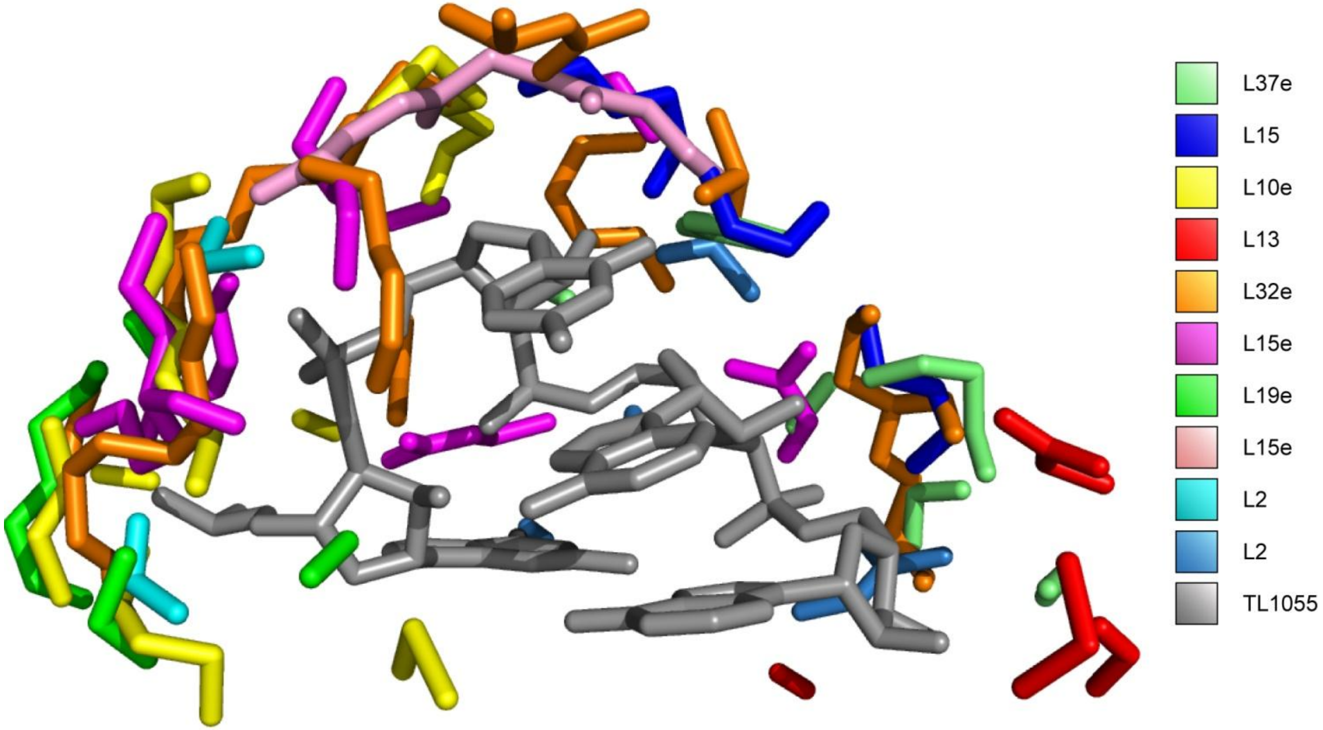
A signicant preference for Arg, 31% (20.6% on the entire contact area)

A decrease in Lys with 4.45% (13.3% on the entire contact area)

Tetraloop contact surfaces

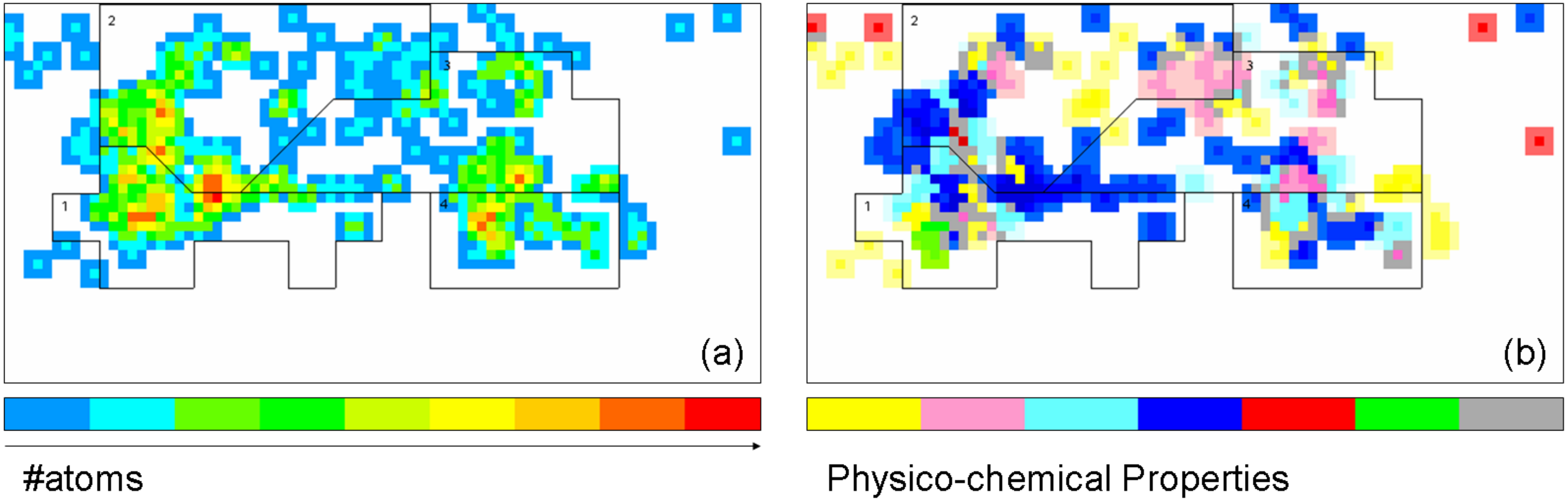


3D conformation of interfaces



Interaction Maps

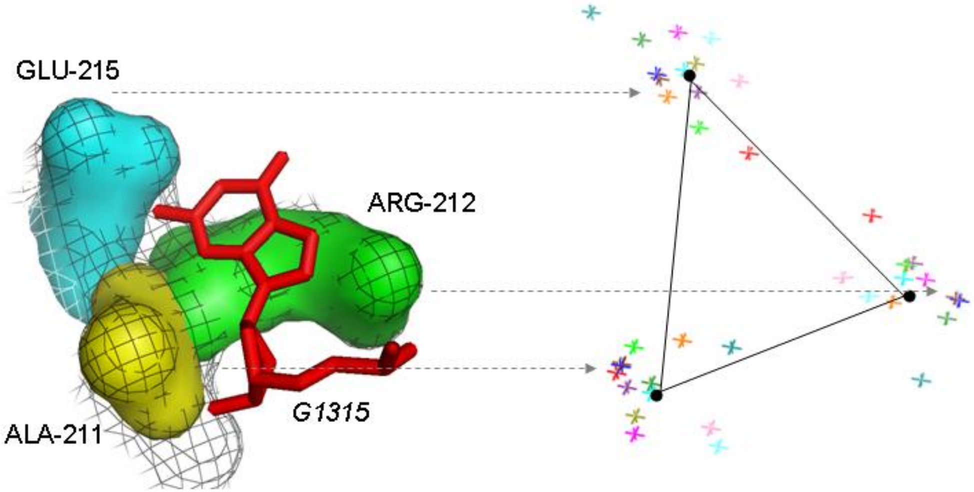
Graphical representation of superimposed tetraloop interfaces in polar coordinates



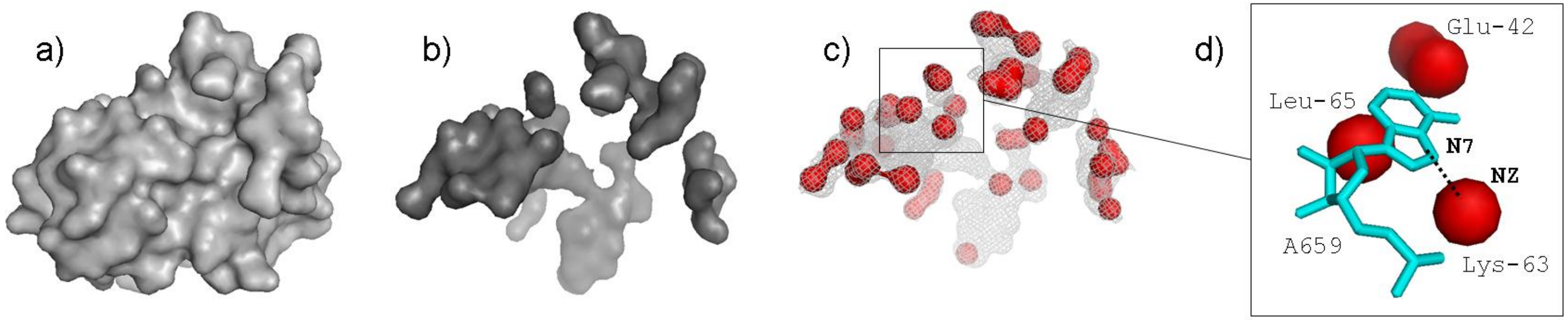
The density of the atoms is higher between the first and the second nucleotide and between the third and the fourth

A characteristic shape: the tripod

An extruded nucleotide mainly interacts with three aminoacids of a protein



Tripods: fingerprint and search



Triplets found at step c) are filtered based on their P-B-R composition

13 instances of the tripod were identified on the ribosome

Protein-RNA interactions
Final considerations

Difficulty of the analysis

- Limited amount of 3D data
- Great conformational variability of interfaces

Main Challenge

Prediction of protein-RNA interfaces

Important fact: The existence of non-homologous proteins that bind the same sites in both archaeal and eubacterial large subunits