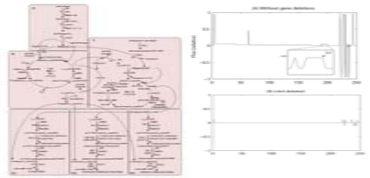
Genome scale Comparison of ligand binding sites in protein structures: Algorithms and applications in drug discovery

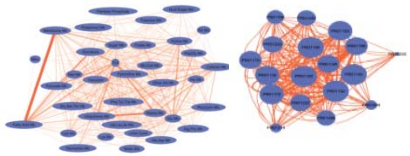
Nagasuma Chandra
Indian Institute of Science
Bangalore, India

Nagasuma Chandra, Research Overview

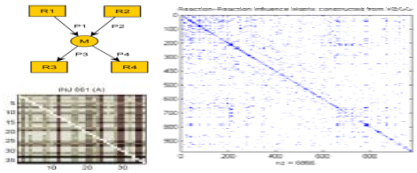
Pathway Modelling



Reactome Modelling



Protein-Protein influences



Interactome Modelling

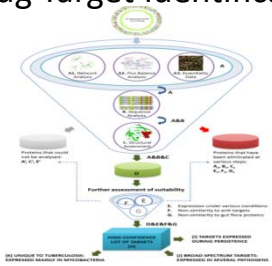
SYSTEMS BIOLOGY

Algorithms for Structure Analysis

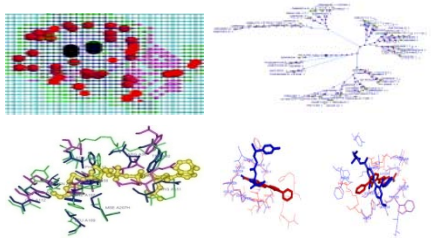
STRUCTURAL BIOINFORMATICS

Drug Discovery

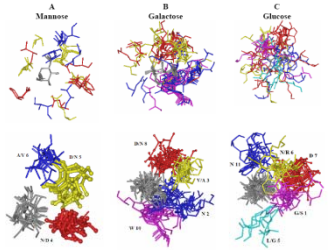
Drug Target Identification



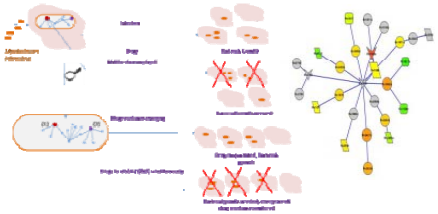
Druggability Assessment



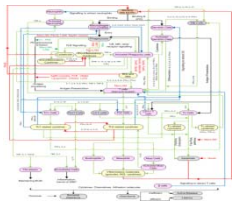
Lead Identification



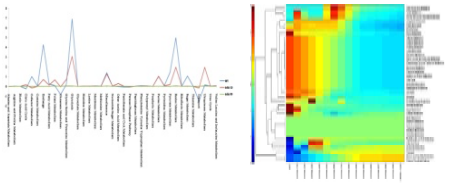
Drug Resistance



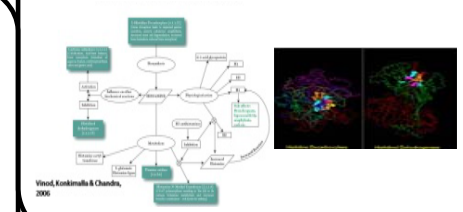
Host-Pathogen Modelling



Modelling drug effects

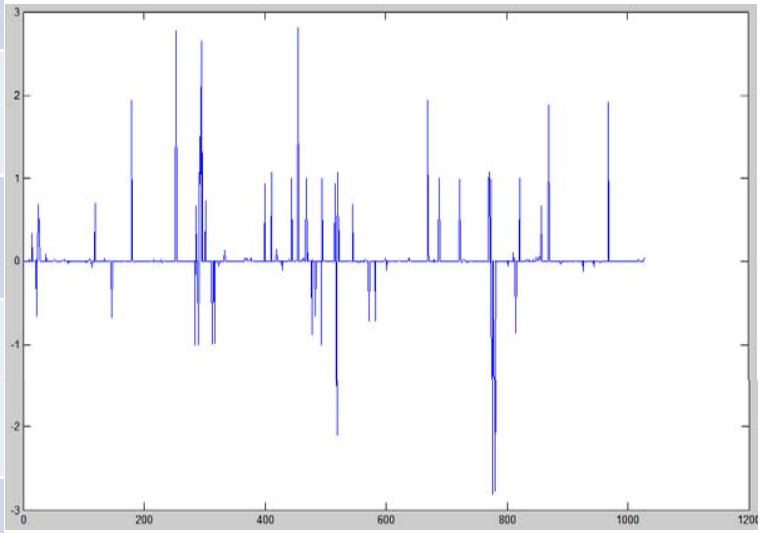


Drug failure Analysis



Modeling Metabolism in *M. tuberculosis*

Genes	661
Proteins	543
Reactions (Intra Systems)	939
Reactions (Exchange)	88
Gene Association Reaction	77%
Metabolites	828
Average Confidence Level	2.31



Mtb Reactome Flux profile

Growing bacteria *In silico* under different media

In silico gene deletions

Essential nutrients required for growth

Response to other Perturbations- Nutrient Uptake

Insights obtained

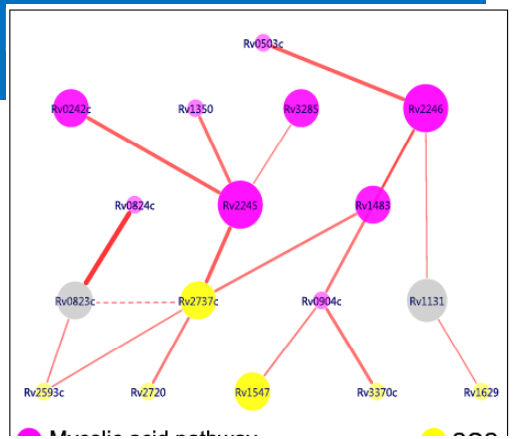
- Gene Essentiality: 220 essential genes
- Nutrient essentiality:
- Hard coupled reaction sets: groups of reactions that are forced to operate in unison due to mass conservation and connectivity constraints)
- Fatty Acid Metabolism & Lysine Metabolism

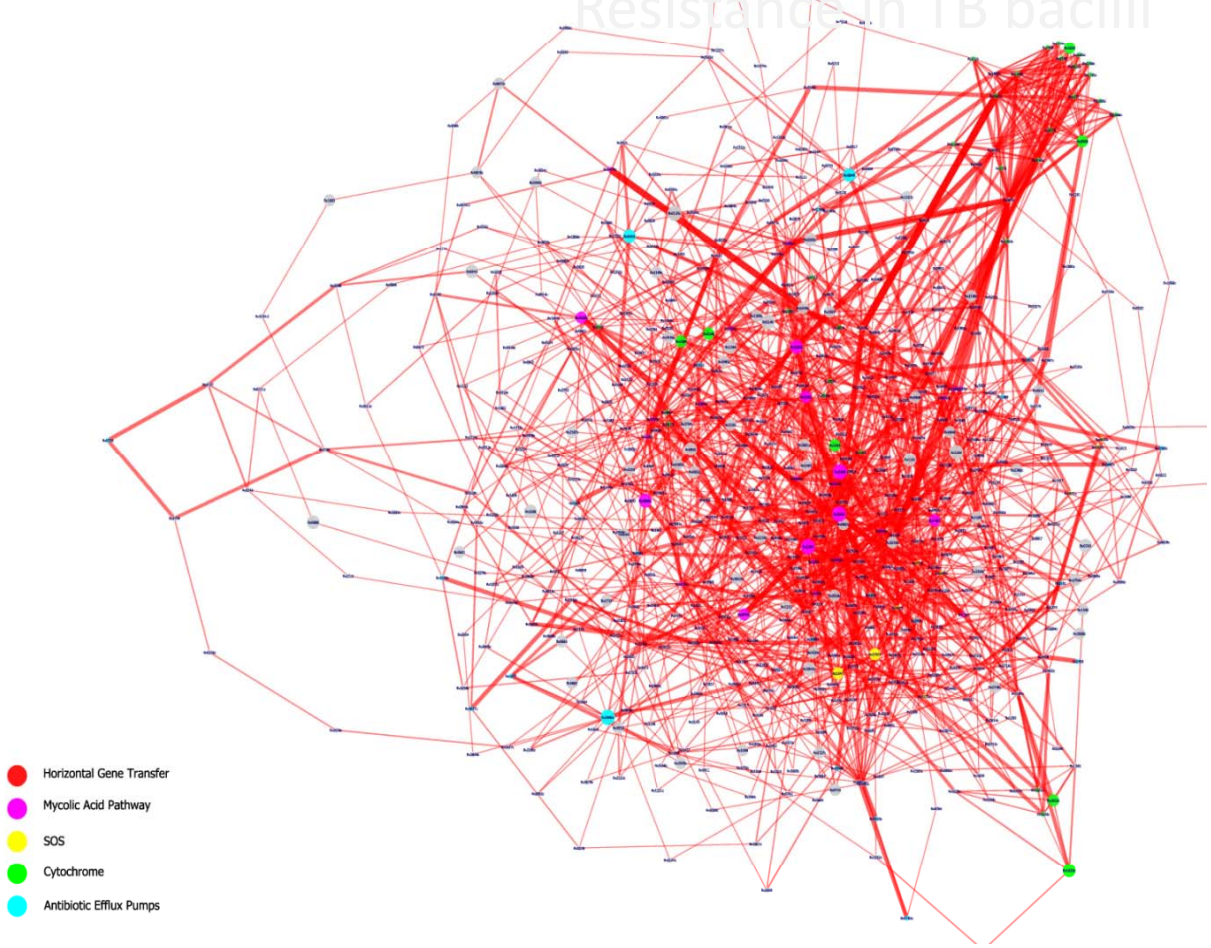
Raman *et al.*, 2008

Drug Resistance Pathways

Abstraction of the flow of information that

Resistance in TB bacilli

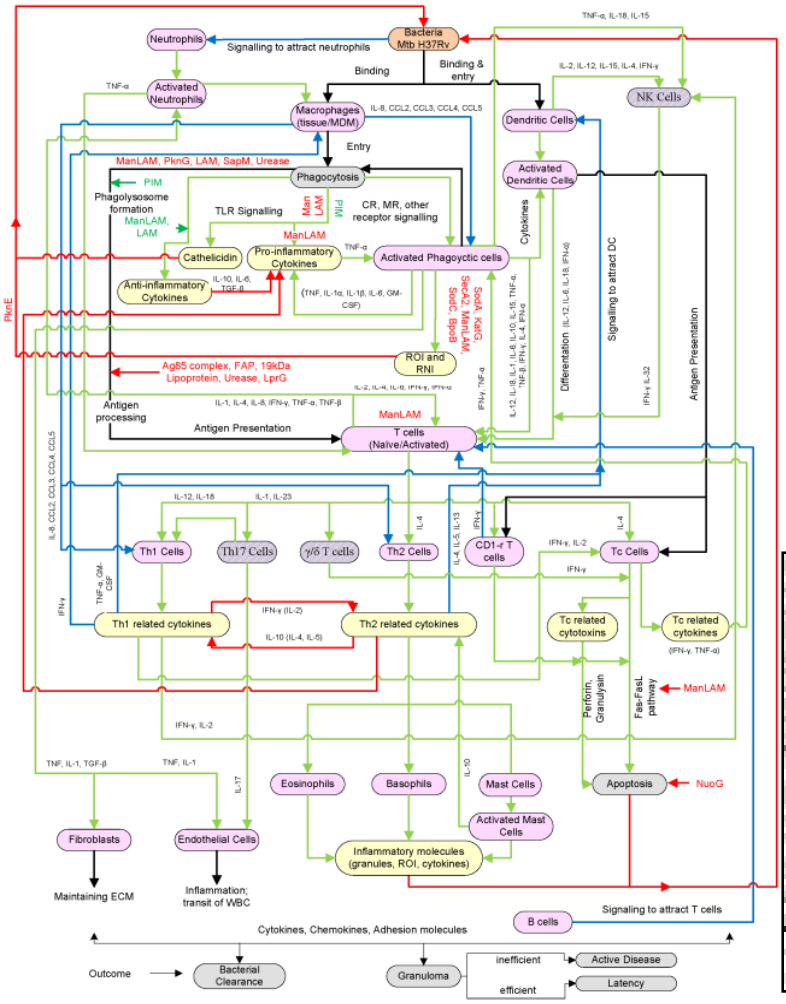


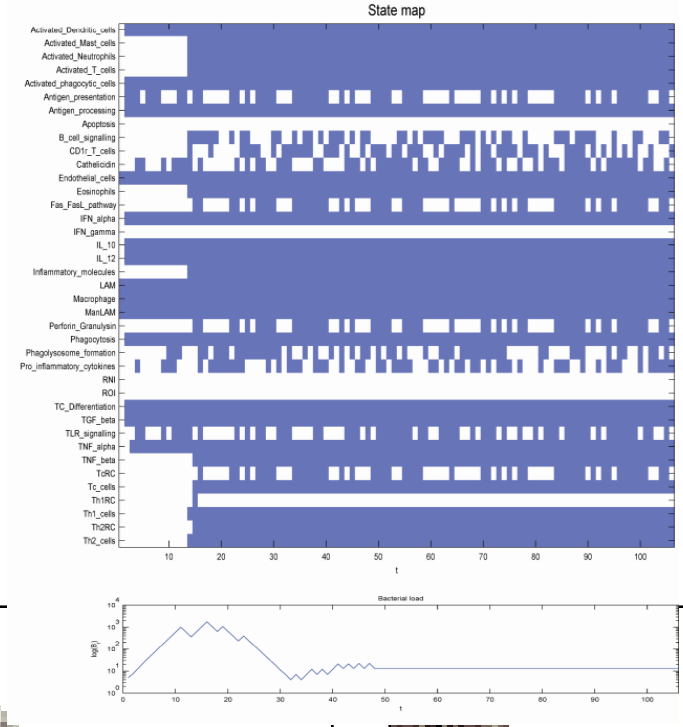


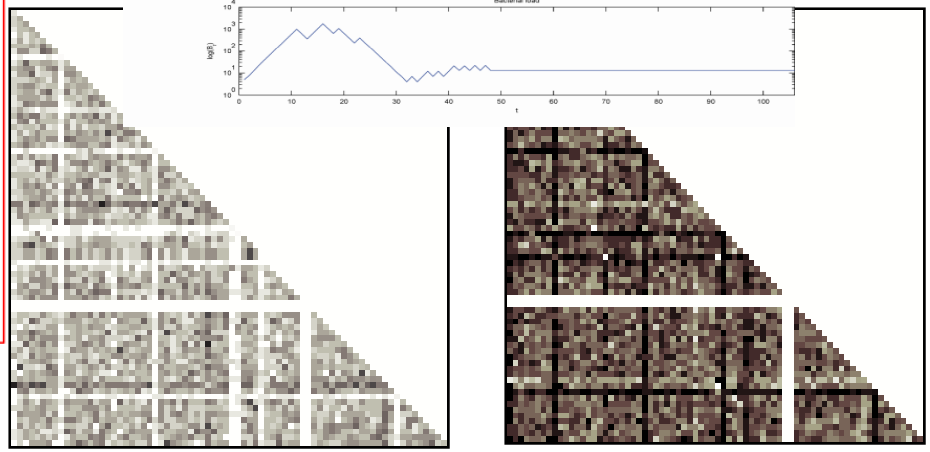
- Network of shortest paths from MAP to Resistance Genes
- 616 nodes and 1,683 edges
- Paths scored based on edge frequency, up-regulation of source and target nodes

Raman and Chandra, 2008, BMC Microbiology

Host-Pathogen Interaction Network







A Boolean model of HPIs developed, Simulations to capture a variety of scenarios

Raman, Bhat & Chandra, Mol. Biosyst, 2010

Modelling iron homeostatis in M.tuberculosis

Soma Ghosh, KVS Prasad, Sarswathi Vishveshwara, Nagasuma Chandra

Figure 4:

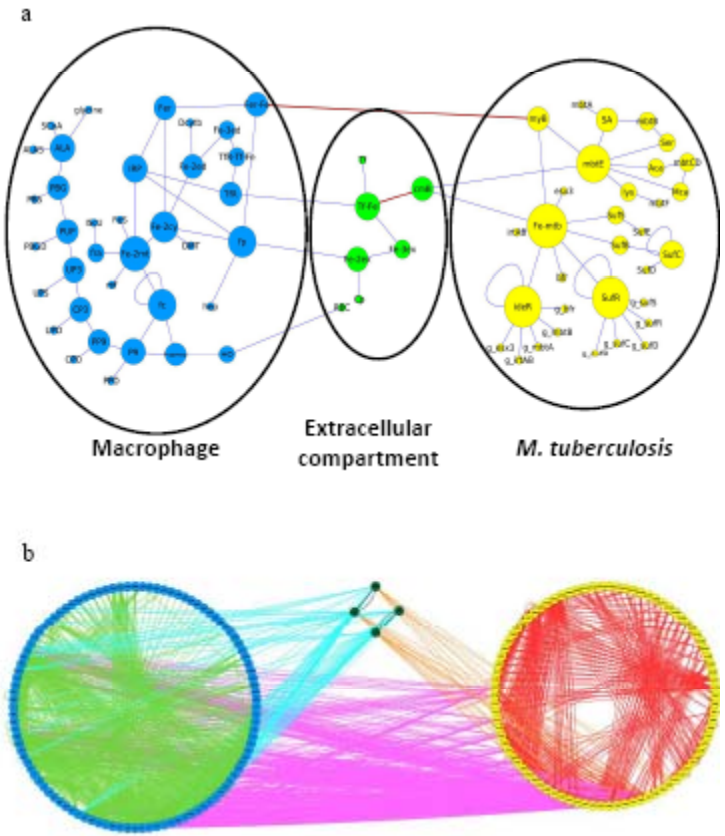
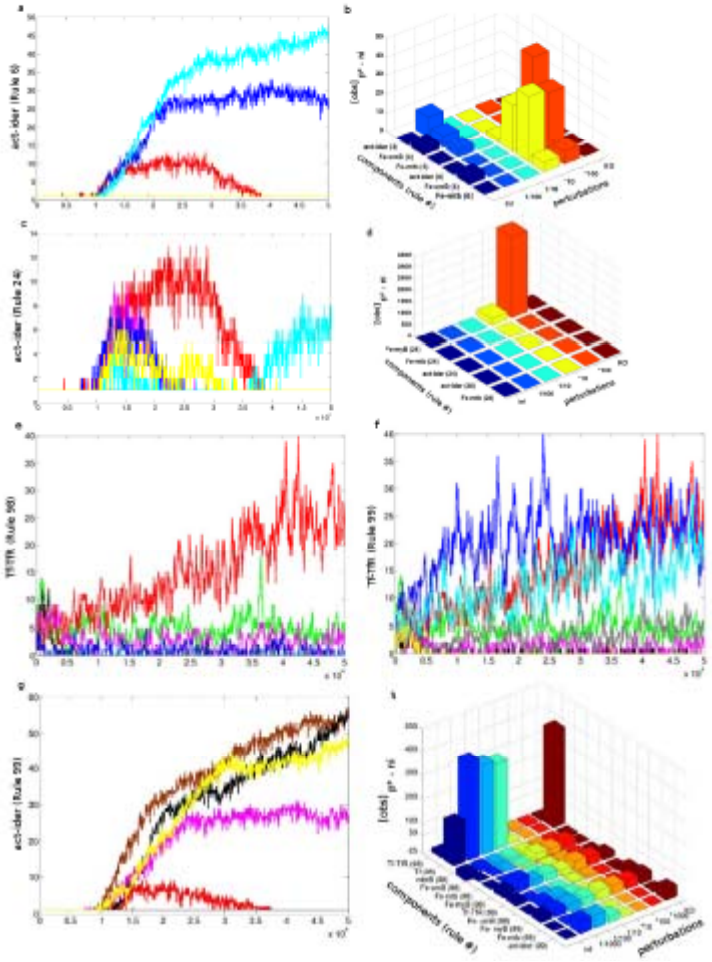


Figure 7:



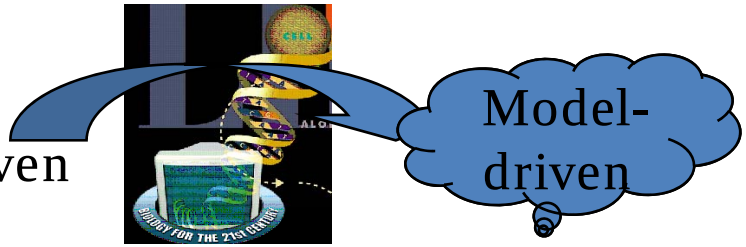
Ghosh et al., Mol Biosyst, 2011, In Press

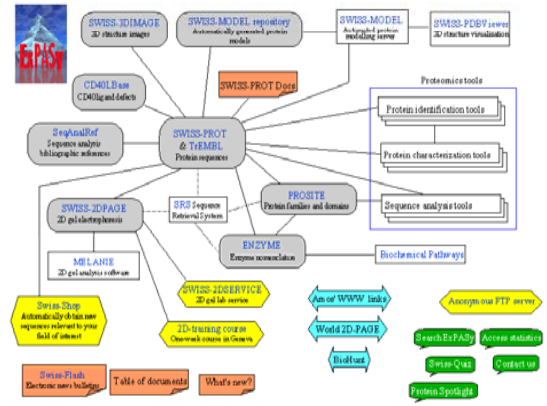
Integration of Systems Perspective with Structural level detail

Raman, Kalidas, Chandra, 2008
Model Driven Drug Discovery: Principles and Practices In: Biological Database Modeling, Artech House ISBN 1596932589

Comparison of protein molecules

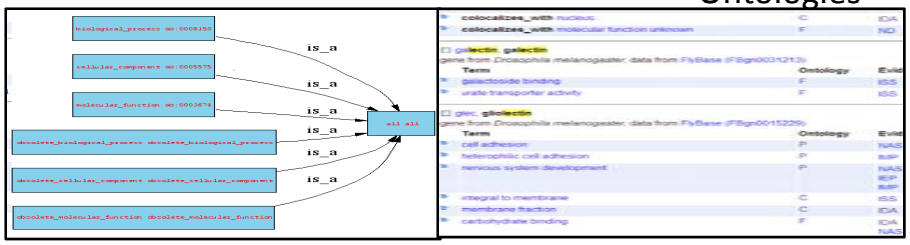
Data-driven





Databases

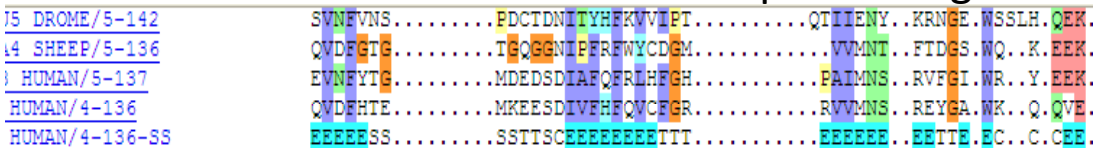
Gene Ontologies



How to compare?

- Annotation- keyword
- Function identification
  - Molecular level
  - Cellular level
- Sequence - Alignment
- Structure – Fold comparison

Sequence Alignment



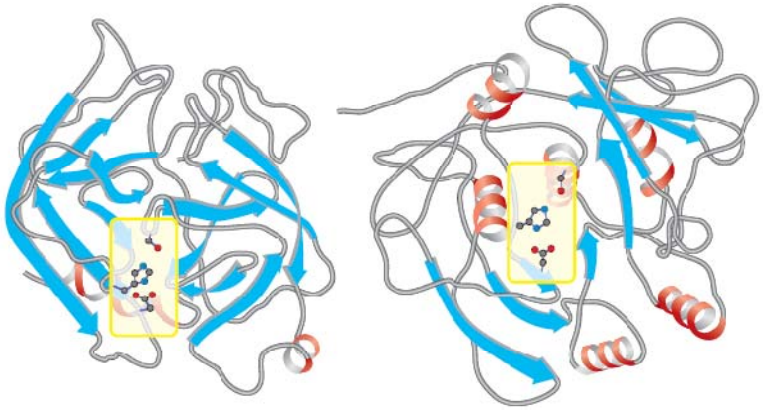
Structure Alignment

Problems with these approaches

- Dissimilar Sequences- Similar Function → benzoylformate decarboxylase (BFD) and pyruvate decarboxylase (PDC)
- Similar sequences- Different Function → Steroid delta isomerase; nuclear transporter 2; scytalone dehydratase

Similar Structures- Different Function → triose phosphate isomerase and FMN-linked oxidoreductases

Dissimilar Structures – Similar Function → ATP binding proteins from different SCOp families; C-type lectin and bulb lectin



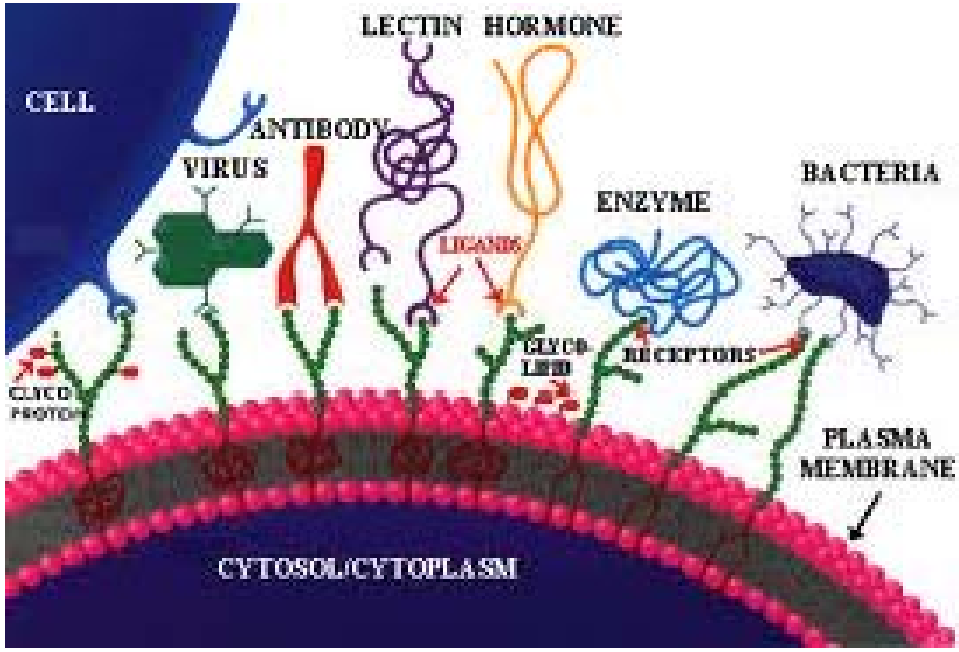
Chymotrypsin & Subtilisin

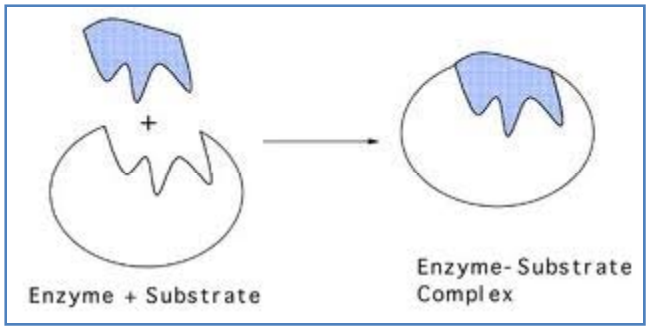
What really matters for a protein molecule is its function and not what means it uses to achieve it !

It’s the **meaning** that counts….

Whether two proteins can recognize the same molecules

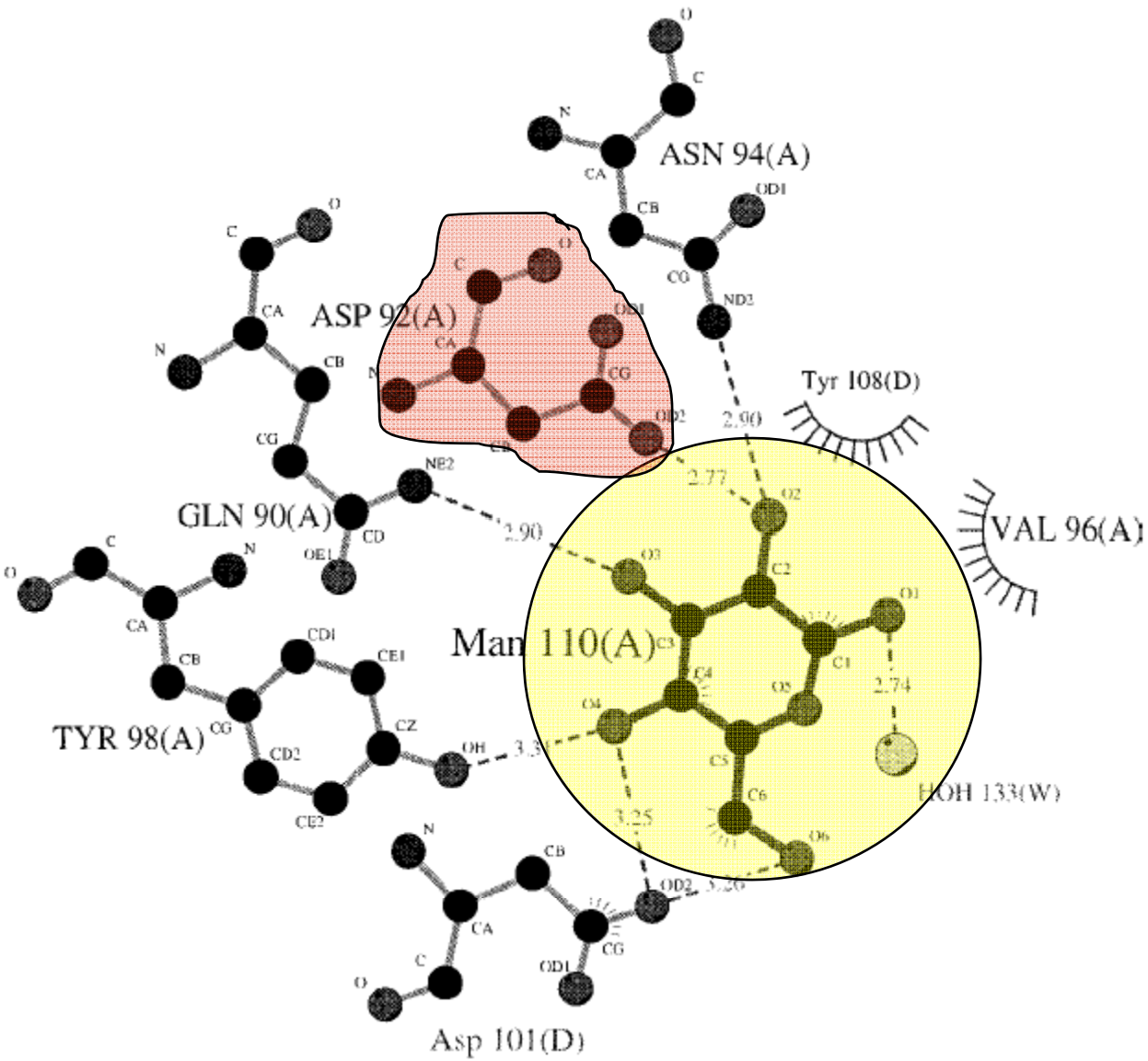
Molecular recognition



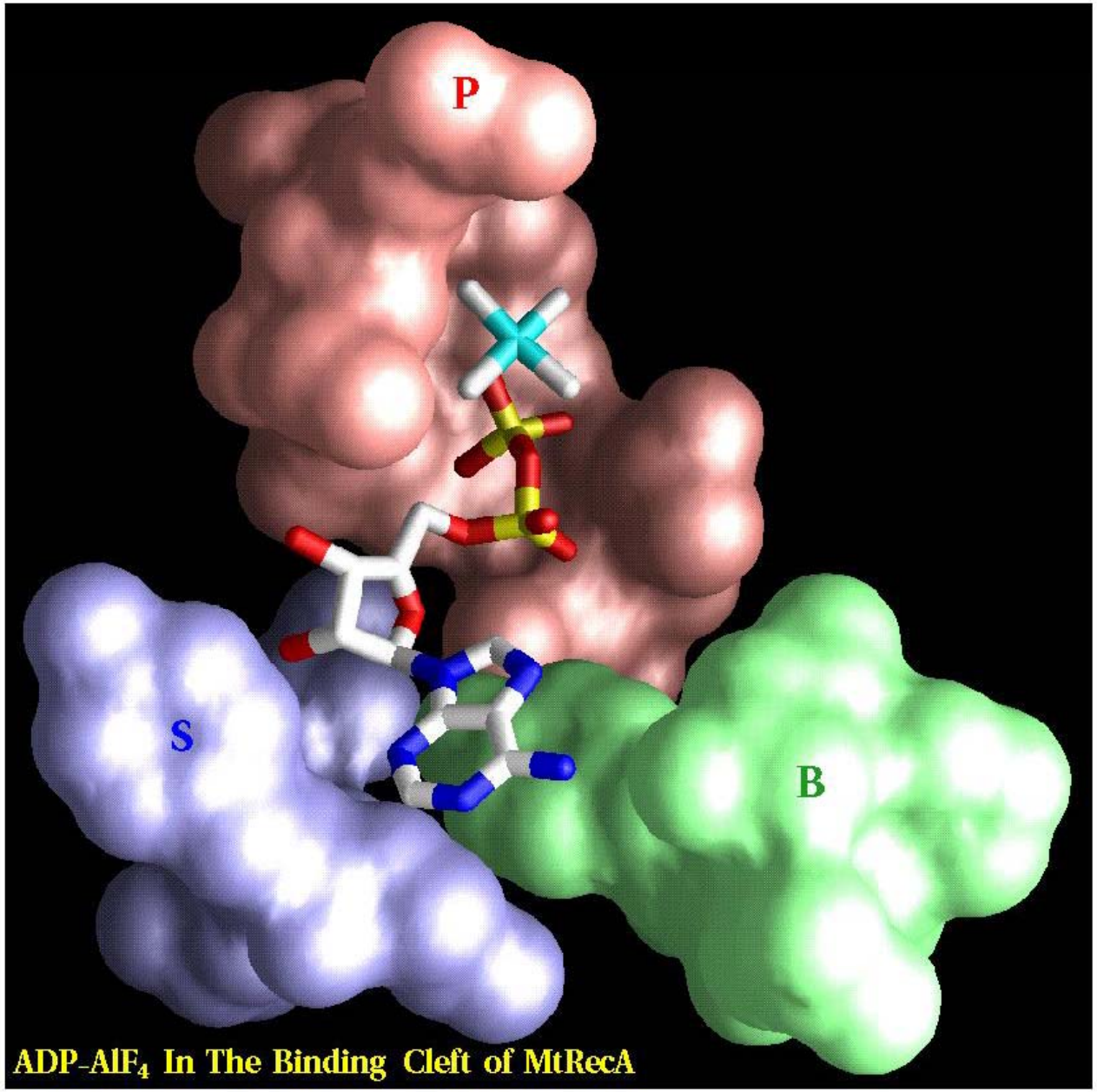


```
GARL  1    39  -YAGQSLDVEPYHLIMQEDCNLVLYDHS-TAVWTSNTDIPG---
GARL  2    71  -SNTDIPGKKGCKAVLQSDGNFVVYDAEGRSLWASHSVRG----
GARL  3   103  ASHSVRGN-GNYVLVLQEDGNVVIYG---SDIWSTNTYKGSAR-
SNOW  1    34  -STGEFLNYGSFVFIMQEDCNLVLYDVD-KPIWATNTGGLS---
SNOW  2    65  ATNTGGLS-RSCFLSMQTDGNLVVYNPSNKPIWASNTGGQ----
SNOW  3    97  ASNTGGQN-GNYVCILQKDRNVVIYG---TDRWATGTHTGLVG-
HIPP  1     1  ----------------MQQDCNLVLYDVD-KP------------
HIPP  2    18  ASNTGGLA-RGCHLSMQSDGNLVVYSPSNSPIWASNTQGE----
HIPP  3    50  ASNTQGEN-GNYVCIVQKDRNVVIYG---TARWATGTYTGTVG-
NARC  1    35  -STGQFLSYGSYVFIMQEDCNLVLYDVD-KPIWATNTGGLS---
NARC  2    66  ATNTGGLS-SDCHLSMQTDGNLVVYSPQNKAIWASNTDGE----
NARC  3    98  ASNTDGEN-GHFVCILQKDRNVVIYG---TDRWATGTYTGAVG-
ONIO  1    26  -YAGQSLVVEQYTFIMQDDCNLVLYEYS-TPIWASNTGVTG---
ONIO  2    58  -SNTGVTGKNGCRAVMQADGNFVVYDVKGRAVWASNSRRG----
ONIO  3    90  ASNSRRGN-GNYILVLQKDRNVVIYG---SDIWSTGTYRKKVG-
RAMS  1    71  -SNTGVSGRNGCRAVMQADGNFVVYDSNSRAVWASQSRRG----
RAMS  2   103  ASQSRRGN-GNYILALQEDRNVVIYG---TDIWSTGTYRRGVG-
URSI  1    38  -YAGQSLEEGPYKLIMQEDCNLVLYEHS-TPVWATNTGVTG---
URSI  2    70  -TNTGVTGRNGCKAVMQADGNFVVYDSNGRAVWASNSIKG----
URSI  3   102  ASNSIKGN-GNYILVLQNDRNVVIYG---SDIWSTGTYIRGVG-
LEEK  1    41  -YAGQSLDVEQYKFIMQDDCNLVLYEYS-TPIWASNTGVTG---
LEEK  2    73  -SNTGVTGKNGCRAVMQKDGNFVVYDVNGRPVWATNSVRG----
LEEK  3   105  ATNSVRGN-GNYILVLQQDRNVVIYG---SDIWSTGTYRRSAG-
SHAL  1    37  -YAGQSLVEEQYTFIMQDDCNLVLYEYS-TPIWASNTGITG---
SHAL  2    69  -SNTGITGKNGCRAVMQPDGNFVVYNVKGRAVWASNSRRG----
SHAL  3   101  ASNSRRGN-GNYILVLQKDRNVVIYG---SDIWSTGTYRKKVG-
TULI  1    54  -YGGQSLTWESYTFIMQTDCNLVLYEGN-GPIWASGSNDLG---
TULI  2    85  ASGSNDLG-SGCYVTMQKDGNLVIYSKSGNSVWASQTHQA----
TULI  3   117  ASQTHQAE-GNYVLVLQKDRNVVIYG---PSLWATNTDQFSLT-
CLIV  1    38  -SPGESLSHGRYVFIMQEDCNLVLYDVD-RPIWATNTGGIS---
CLIV  2    69  ATNTGGIS-HGCHLSMQADGNLVVYSQRNNPIWASNTGGE----
CLIV  3   102  -SNTGGENDANYVLILQKDRNVVIYG---PARWATGTYTGIVG-
EPIP  1    39  -GTGGSLALGGYIFKIQADCNLVLYDNN-RAVWASGTNGRA---
EPIP  2    70  ASGTNGRA-TDCILSMQRDGNLVIYS-GSRVIWASNTNRQS---
EPIP  3   101  ASNTNRQS-GNYYLILQRDRNVVIYDNSNNAIWATGTNVG----
LIST  1    38  -NTGQSLTDGGNAFIMQEDCNLVLYESS-RPTWASGTYHRG---
LIST  2    69  ASGTYHRG-SGCYLAMQNDGNLVVYDNRNRAIWASQTDRQ----
LIST  3   102  -SQTDRQNVGTVILILQKDHNVVLYT---NPIWATGTNRFGSP-
CYMB  1    40  -NPGQSLTSGNVDLAMQYDCNLVLYDNG-KSIWSSGTYGSG---
CYMB  2    71  SSGTYGSG-SGCYVTLQTDGNLVIYDNTNNPLWASNTSGE----
CYMB  3   103  ASNTSGEN-GNYILILQKDRNLVIYS---HPIWATGTNHAGSV-
CUCK  1    37  -NTDGRLTNGELTLIMQVDCNLVLYHG----GWQSNTANNGRDC
CUCK  2   157  -YGDGTLSARNHKLVMQGDCNMVLYGGK-Y-GWQSNTHGNGK--
POLY  1    40  -PSGHSLNTGSYRLIMQADCNLVVYDSG-KPVWATNTGGLA---
POLY  2   103  QTNTNEKE-DHYVLVLQQDRNVVIYG---PAVWATGSGPAVGL-
ALOE_1    11  -HENQYISYGPYEFIMQHDCNLVLYESG-NPTWASNTGGLA---
```

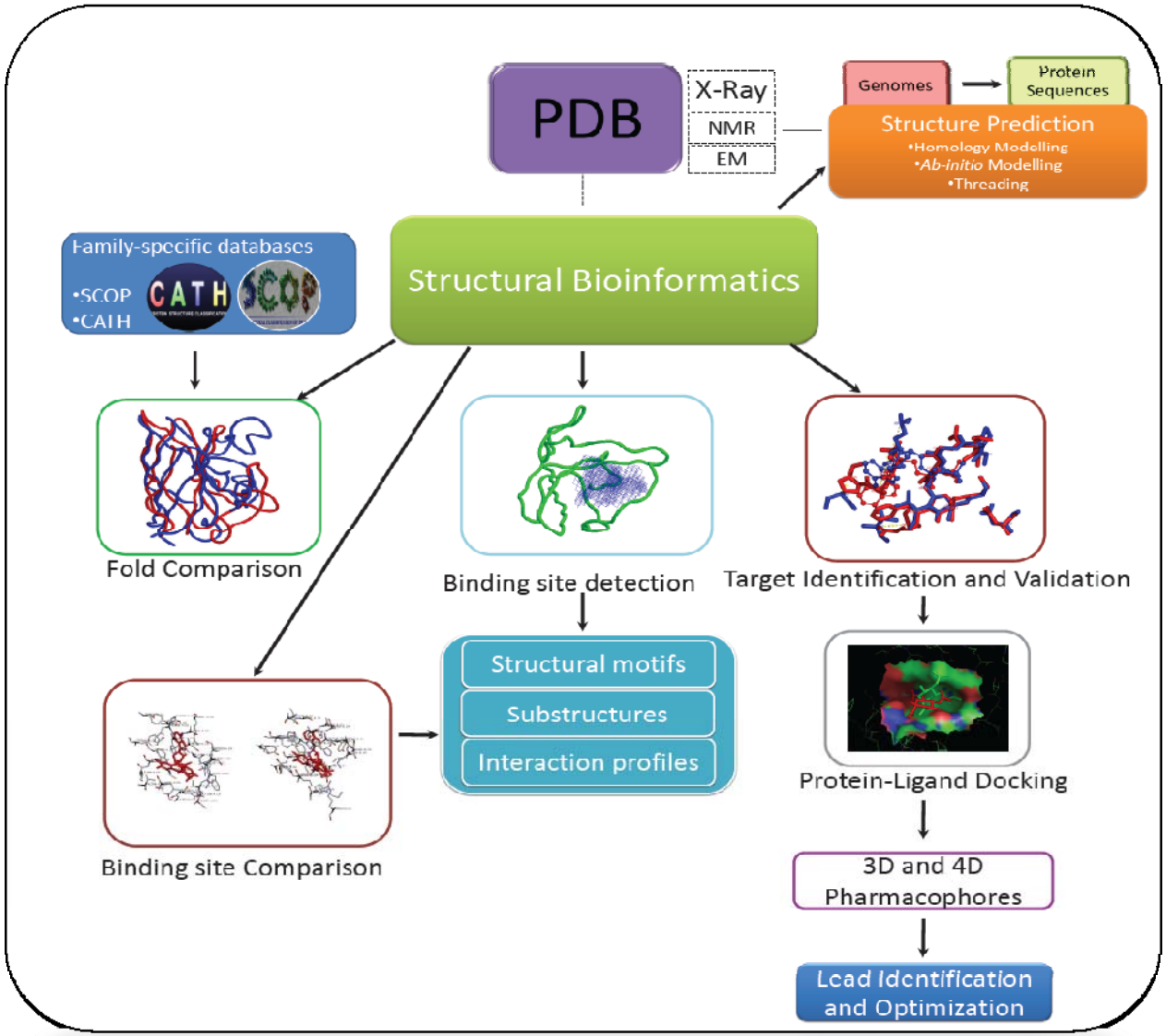
Multiple alignment of bulb lectins indicating the mannose binding sequence motif



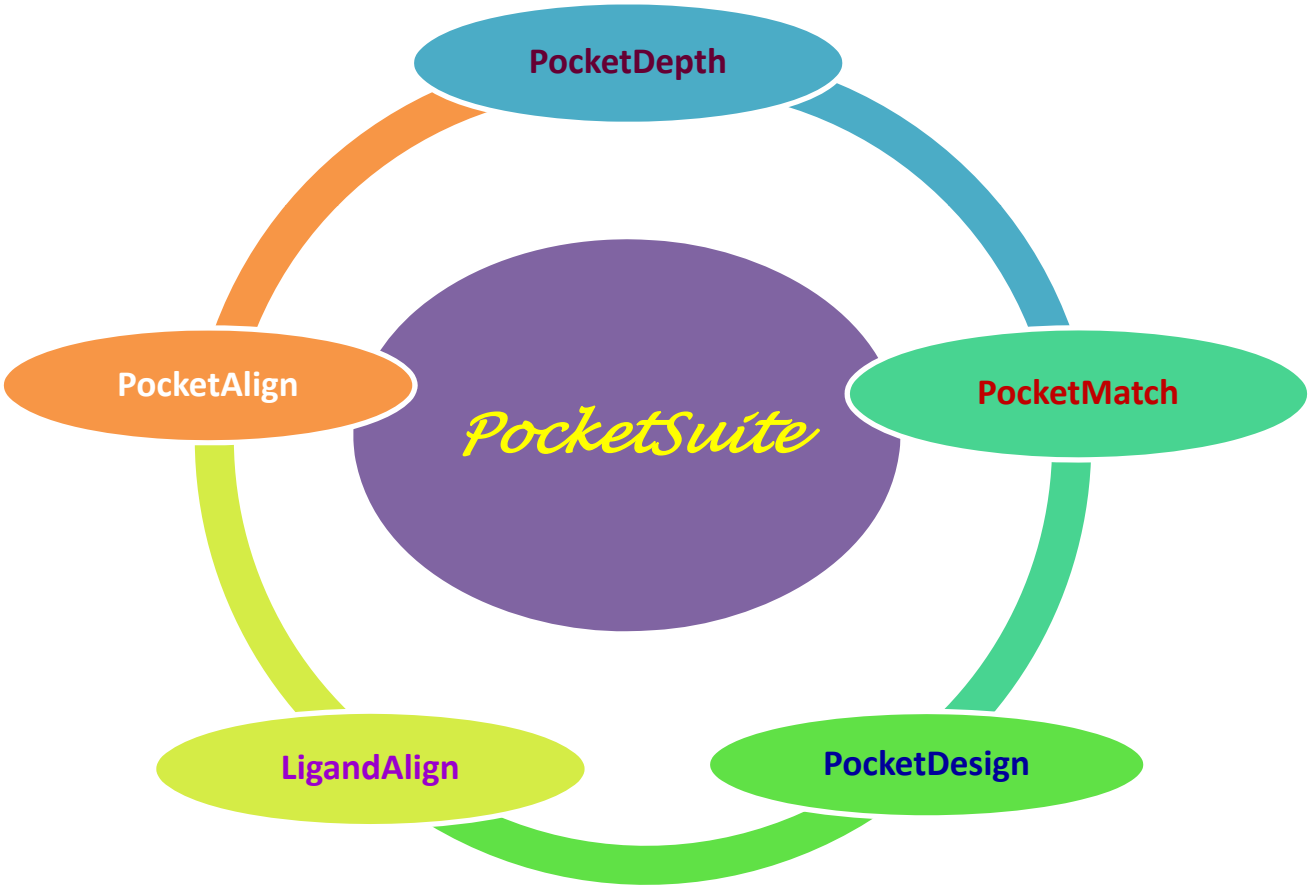
A structural motif specific for mannose recognition


P
S
B
ADP-AlF$_4$ In The Binding Cleft of MtRecA

Structural Bioinformatics



Structural Bioinformatics: *Development of 5 novel algorithms integrated into PocketSuite*



Yeturu and Chandra; 2008, 2009, 2010, 2011

Binding site Prediction Methods:

Homology-based methods

- Alignment with known sites
- Conservation

Sequence-based methods

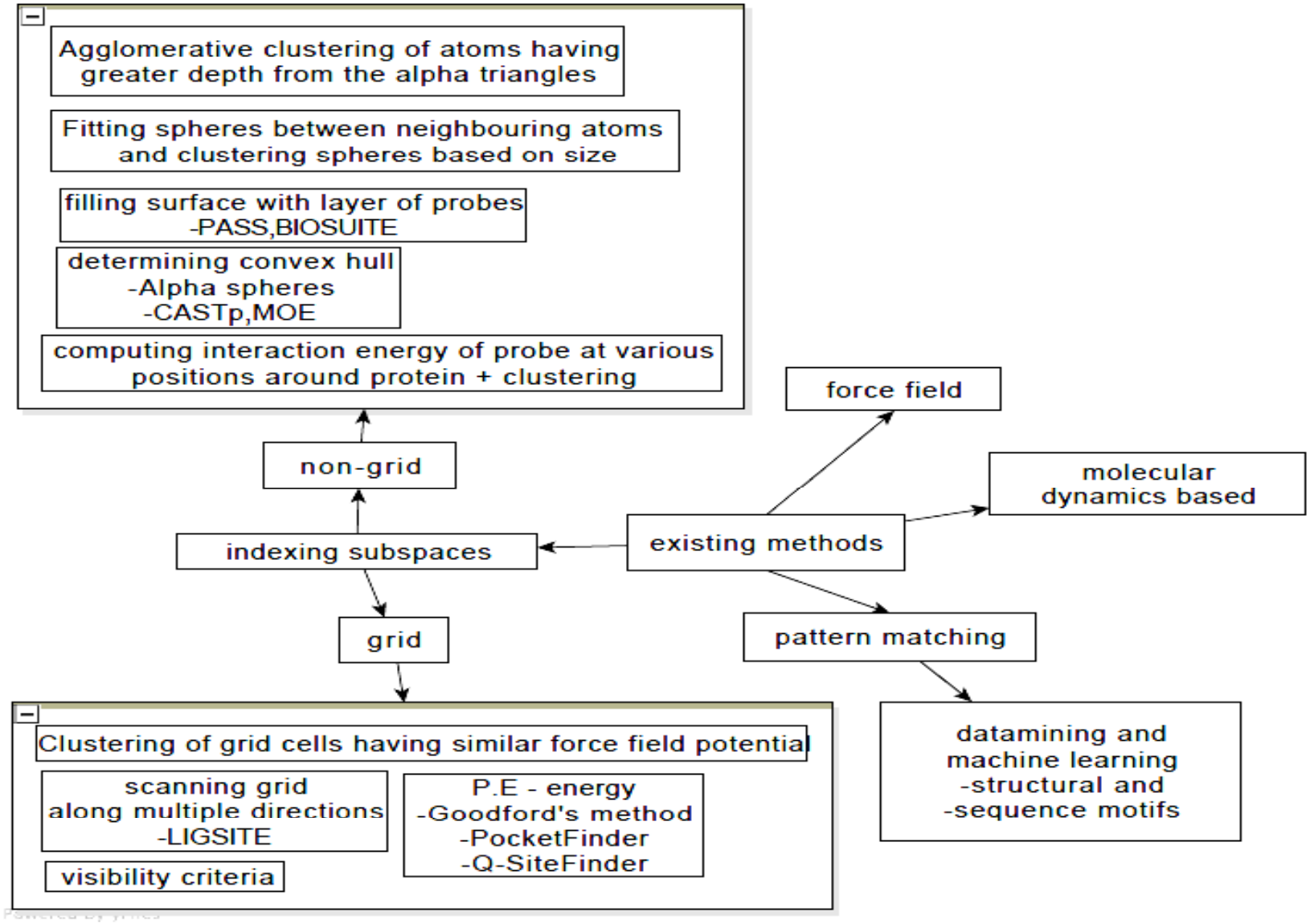
- Motifs

Structure-based methods

- Geometric Chemical
- Hybrid methods

Machine learning

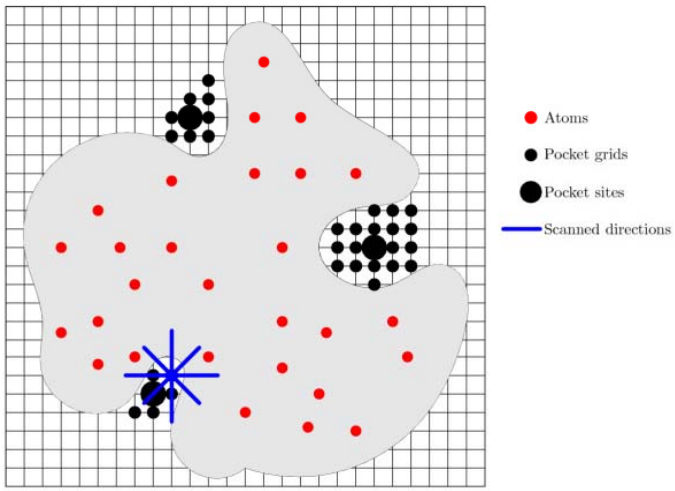
Binding site prediction Methods



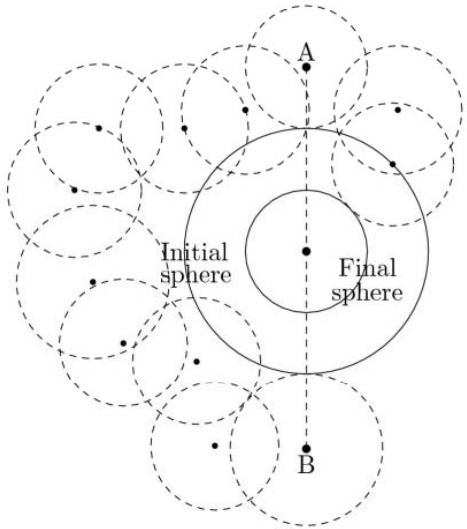
Method	Algorithm	Resource
POCKET	Scan the grid along 3 dimensions	(Levitt and Banaszak, 1992)
Ligsite	Scanning grid along 3 axes and 4 diagonals	(Hendlich *et al.*, 1997)
LigSite ^{CSC}	Similar to Ligsite but with residue conservation information for each set of residues to occur in site	(Huang and Schroeder, 2006)
LigandFit	Eraser to swipe the grid cells to demarcate cells belonging to a grove	(Venkatachalam *et al.*, 2003)
PASS	Filling up surface of protein by multiple layers of probes and retaining probes with high burial count	(Brady and Stouten, 2000)
CASTp	Fill the interatomic regions by spheres and cluster moderately sized spheres	(Liang *et al.*, 1998)
VOIDOO	Similar to VOIDOO	(Kleywegt and Jones, 1994)
SURFNET	Determine depressions on the surface of protein by placing spheres between pairs of atoms.	(Glaser *et al.*, 2006)
APROPOS	Find clusters of atoms with depth from surface of protein	(Peters *et al.*, 1996)
Goodford's method Q-SiteFinder PocketFinder	Clustering of grid cells with higher energy values	(Goodford, 1985) (Laurie and Jackson, 2005) (Jianghong *et al.*, 2005)

Geometry based

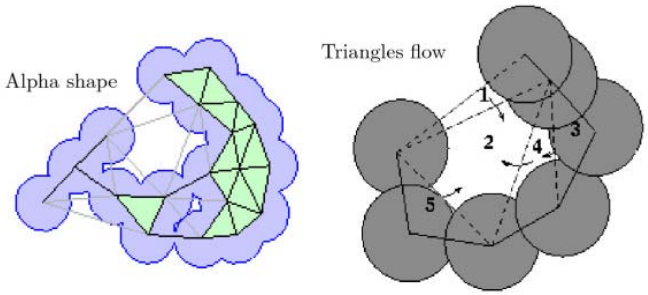
a. POCKET, LIGSITE, LIGSITE^{csc}



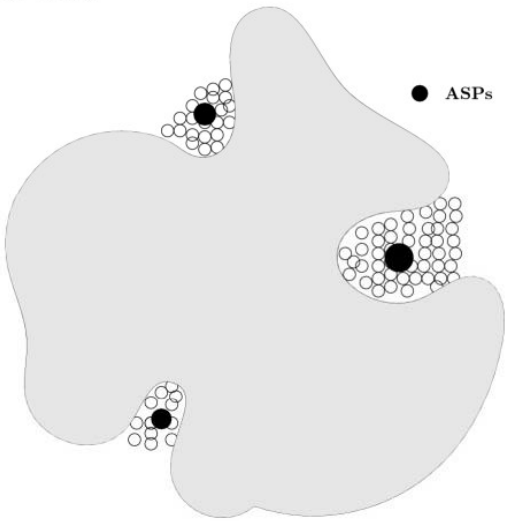
b. SURFNET



c. CAST

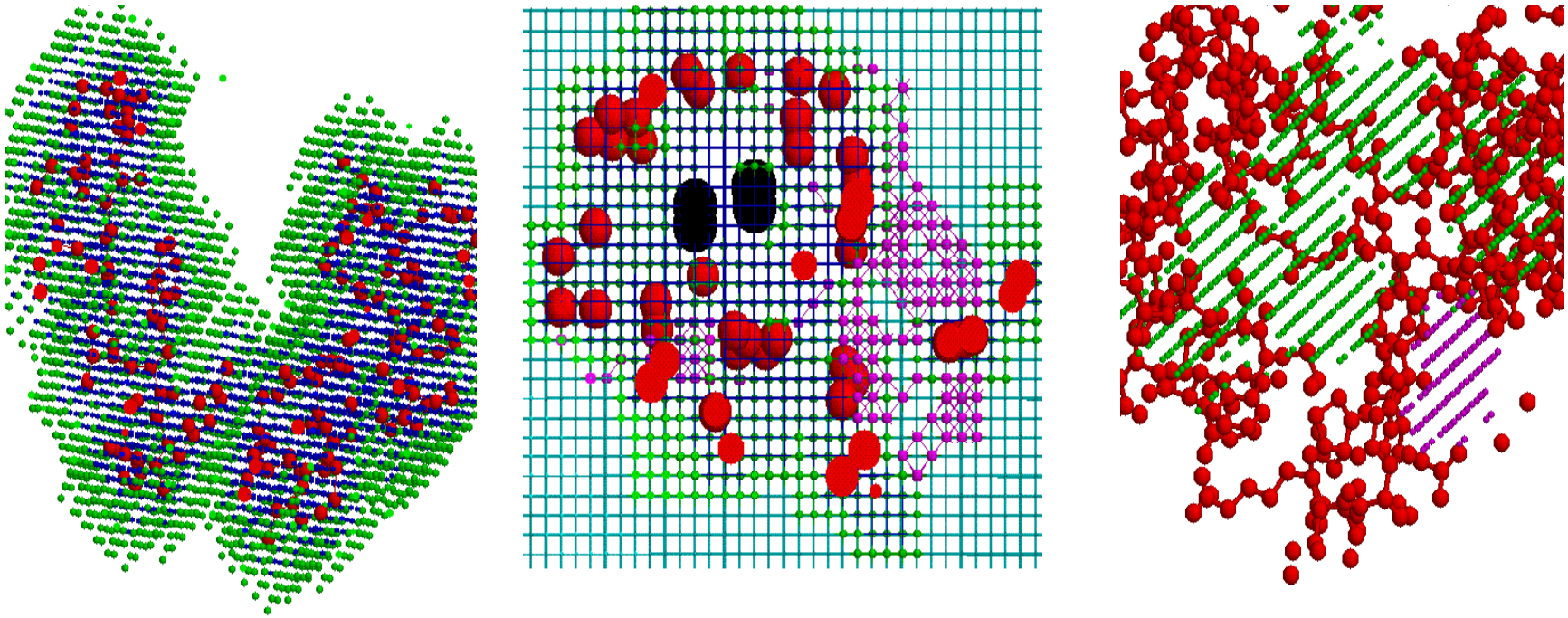


d. PASS



PocketDepth
Grid based binding site prediction method

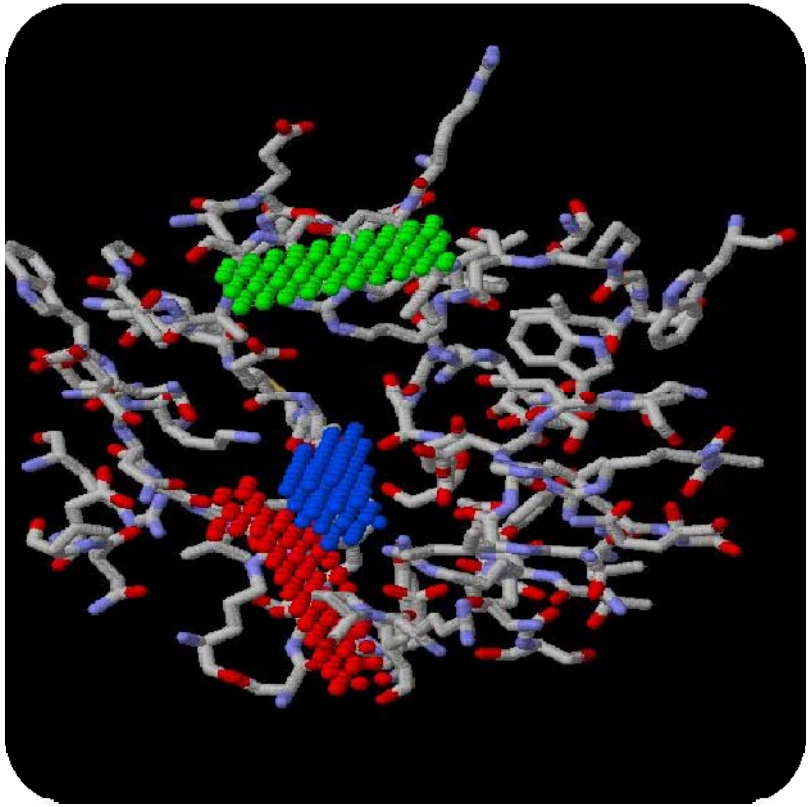
Figure 1

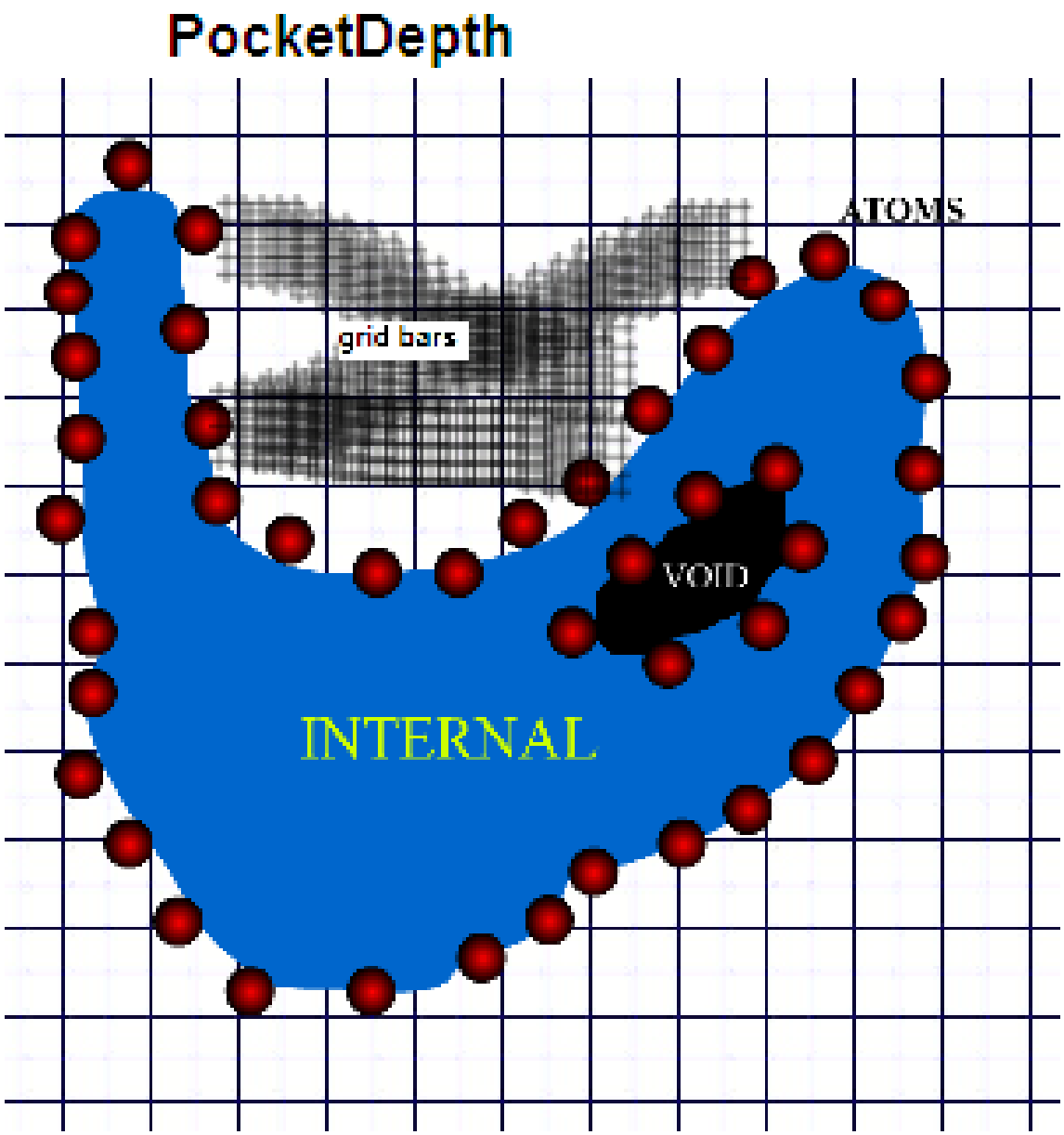


(a) (b) (c)

Grid Bar Generation

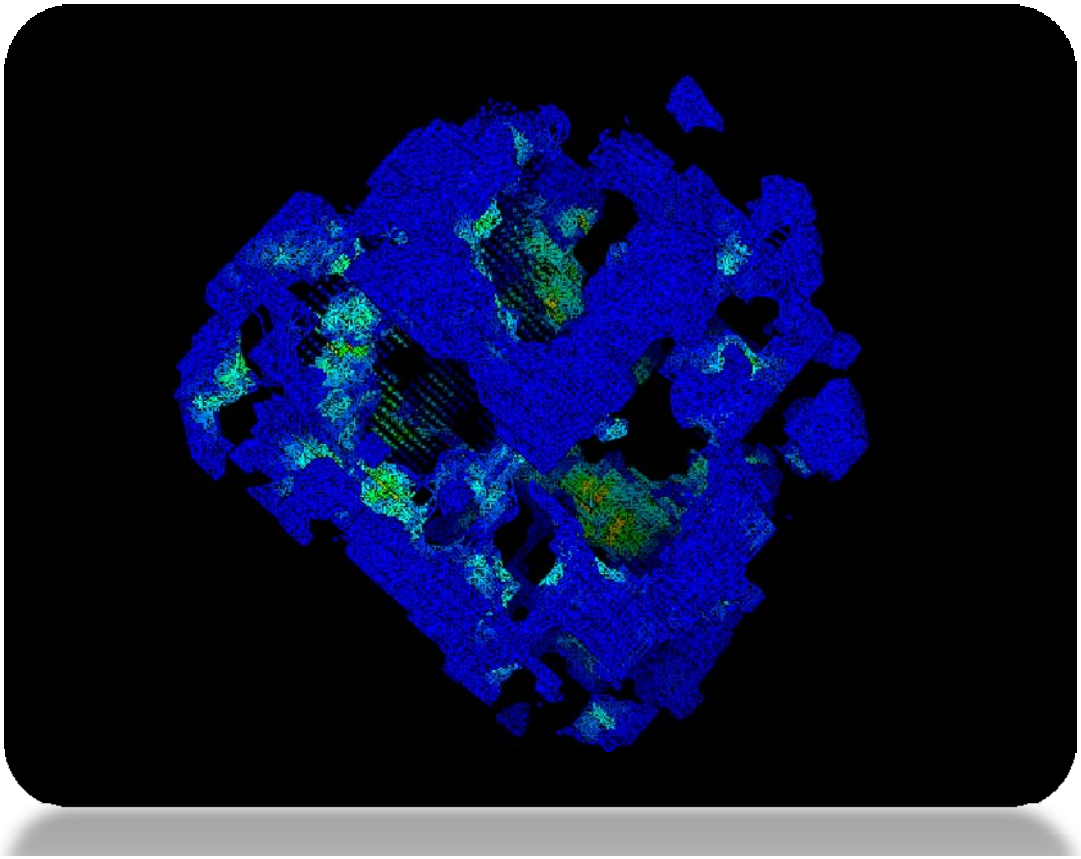
Fill inter-atomic (surface atoms) regions with grid bars.
A grid bar $\{x, y\} \in A : GB(x, y) \subset G$ between pair of atoms x, y
A grid bar is valid only if does not intersect an atom
Obtain set of all valid grid bars
$\{(\forall a, b \in S)(\nexists c \in (A - S) : cell(c) \in GB(a, b))\}$



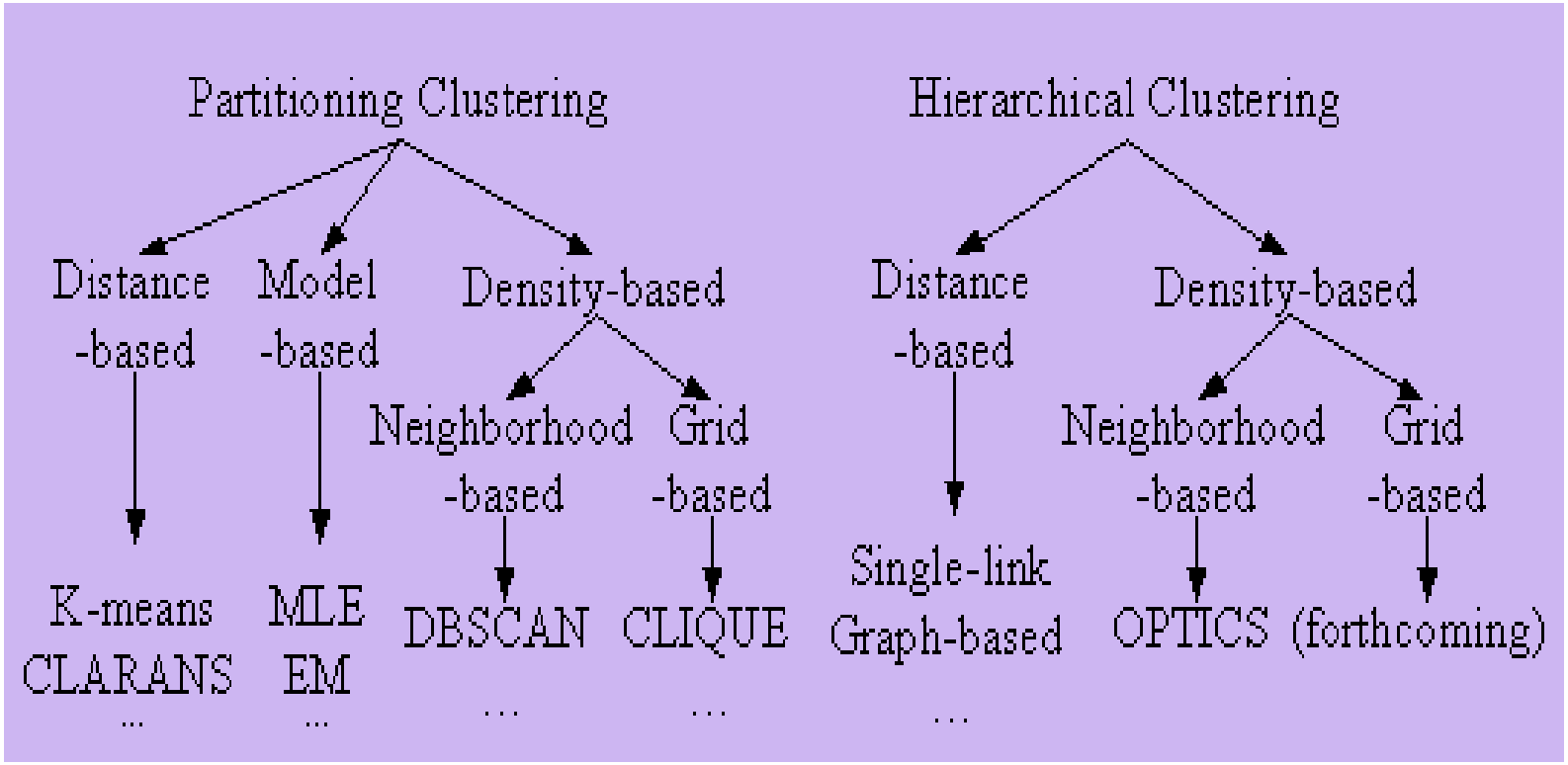

PocketDepth
ATOMS
grid bars
VOID
INTERNAL

Rendering of DepthFactor as temperature

Update traversal counter, called Depth Factor, of each grid cell in a valid $GB(a, b)$
$(\forall c \in GB(a, b)) c.depth \leftarrow c.depth + 1$
Cluster grid cells based on Depth Factor and spatial proximity (DBSCAN)
Partition the whole of the set of grid cells G into non-overlapping clusters
$S^C = C_1, ...C_n : C_i \cap C_j = \emptyset$ where S^C denotes a set of clusters $1...n$



CLUSTERING METHODS



DBSCAN clustering

FUNCTION DBSCAN(point p, c, N)

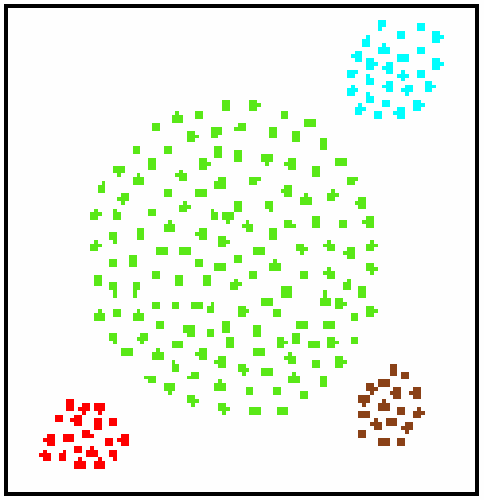
p is a point and c is cluster number

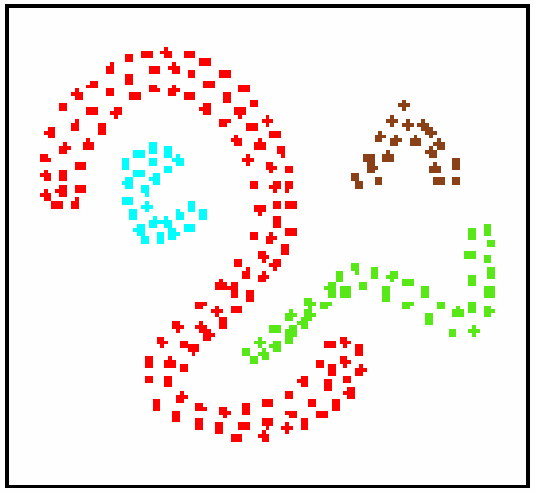
if $|S = q : d(q,p) \leq d_{threshold} \wedge q_c = \epsilon| \geq N$ **then**

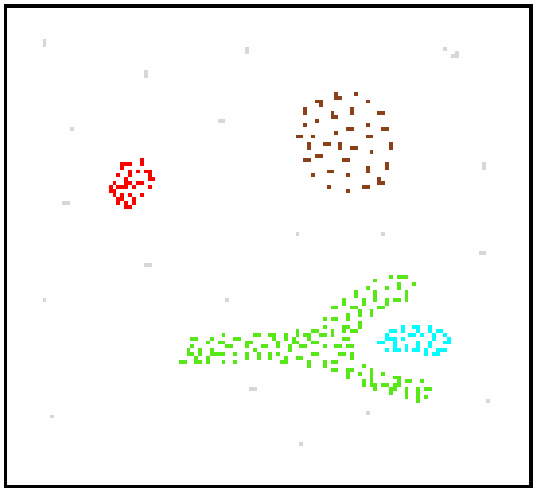
$p_c \leftarrow c$

$call\ DBSCAN(q) : (q \in S)$

end if







Clustering based on Depth Factor

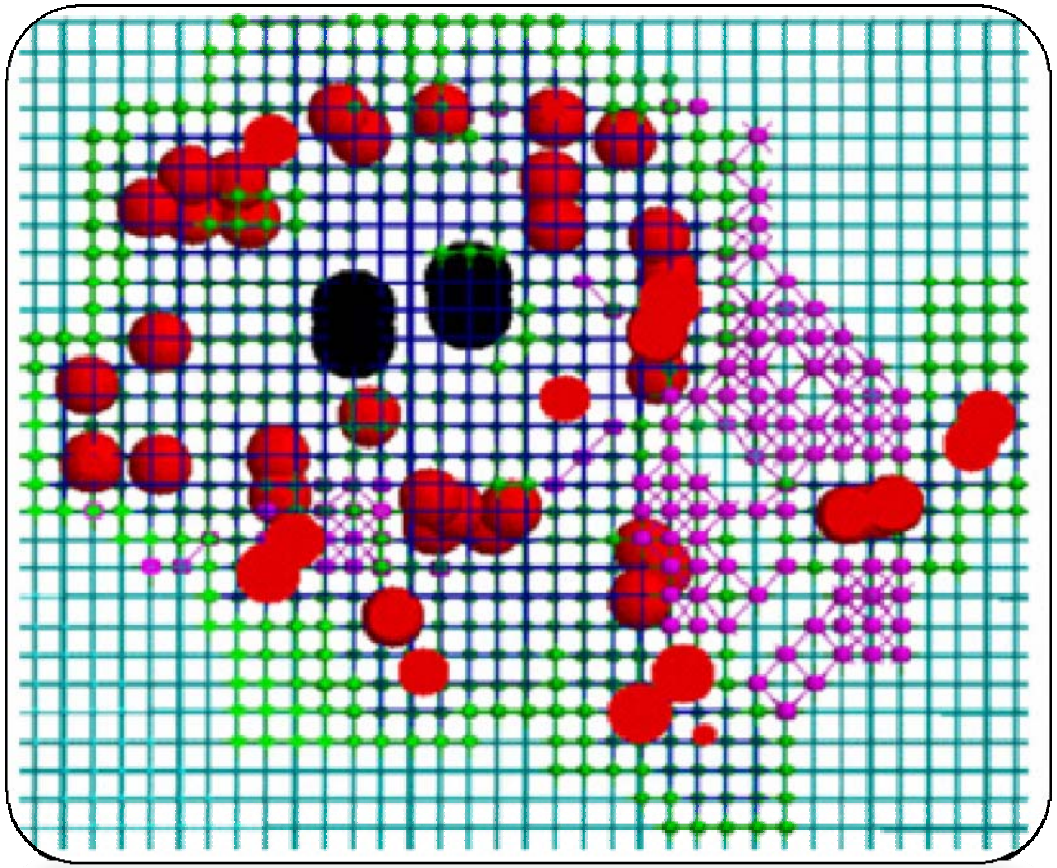
Each grid cell in a cluster $C_i(\forall i \in [1...n])$ satisfies the depth and density requirements

$(\forall(c \in C_i) : |\{(\forall c' \in C_i) distance(c, c') \leq \rho\}| \geq N \wedge DF(c) \geq \Delta$

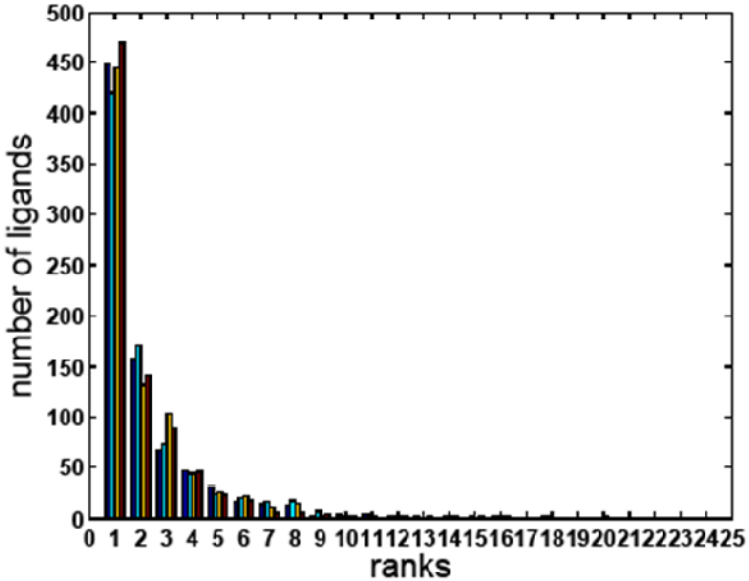
where $\wedge$ denotes logical AND

where ρ, N are radius and number of points within radius (DBSCAN parameters);

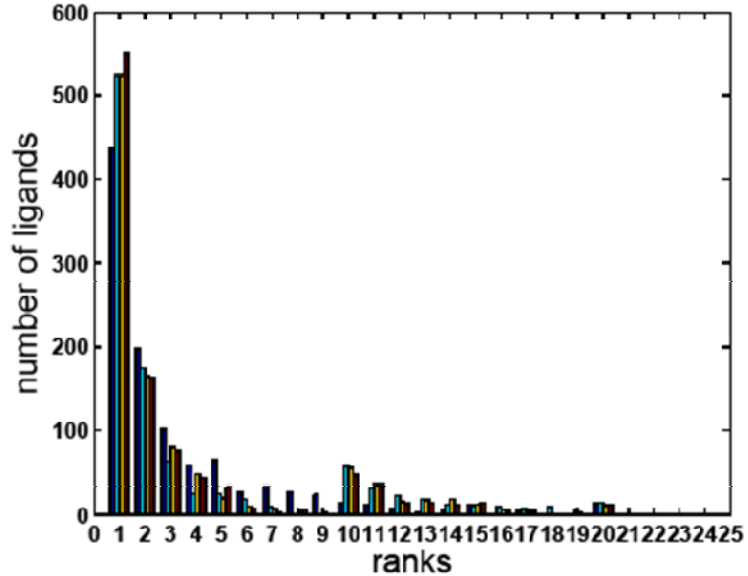
and $DF(c)$ is the Depth Factor and Δ is the imposed threshold



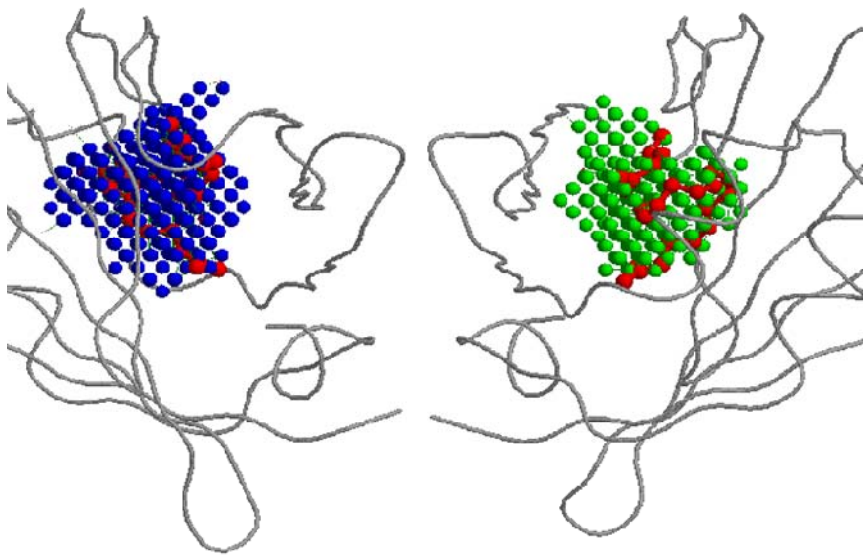
In 82% and 94% of proteins (*PDBBind* 1091) top 5 and 10 ranked clusters overlapped with crystal ligand



(4A1)



(4A2)

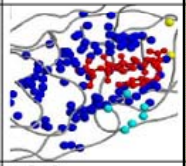
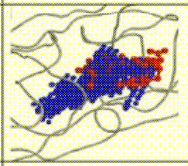
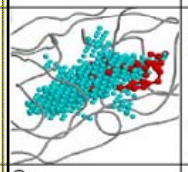
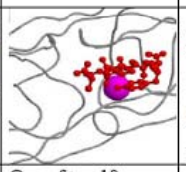
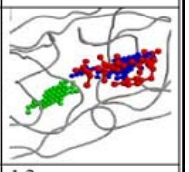
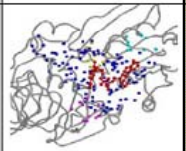
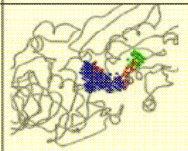
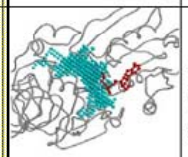
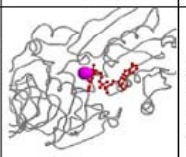
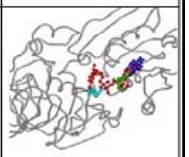
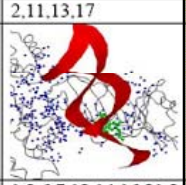
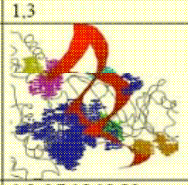
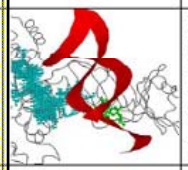
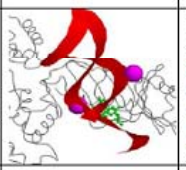
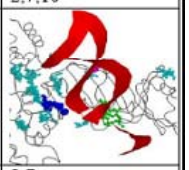
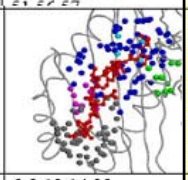
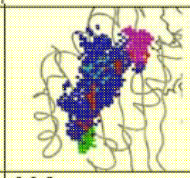
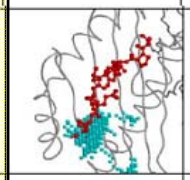
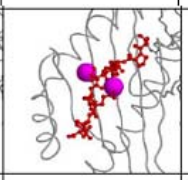
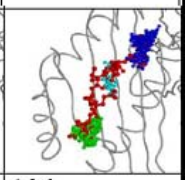
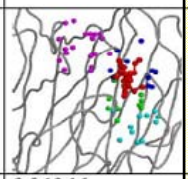
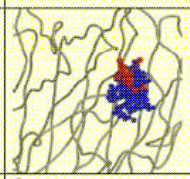
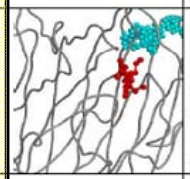
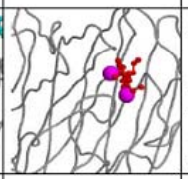
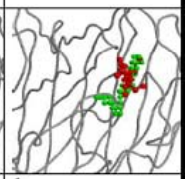
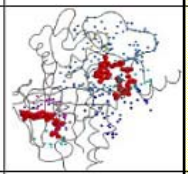
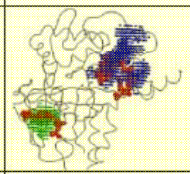
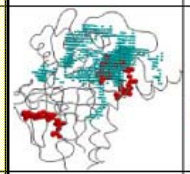
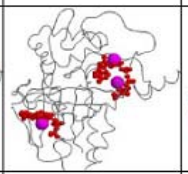
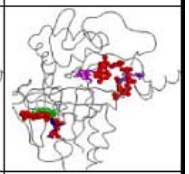


In dimers ranks 1 and 2 corresponded to ligand locations

(4B)

Kalidas & Chandra (2007) JSB (in press) PocketDepth..

Binding Site Prediction Algorithms: PocketDepth Performance

PDB	CASTp	PocketDepth	LigandFit	LigSite^{CSC}	QSiteFinder
1sp5					
ranks	1,5.12	1	One	One of top 10	1,2
1a72					
ranks	2,11,13,17	1,3			2,7,10
1ais					
ranks	1,2,6,7,12,14,16,21,25,28,34,39,41,	1,3,6,7,15,19,20			2,7
1a1m					
ranks	2,3,13,14,28	2,3,5			1,2,6
2pel					
ranks	2,5,13,16	1			1
2g88					
ranks	1,2,3,4,7,30,48,50	1,2,5			2,5,7

Kalidas & Chandra, 2008
JSB

Binding Site Comparison

Need..

Binding site comparison can

- Predict Important residues in a protein binding site
- Predict the function of a hypothetical protein.
- Predict the similarity between proteins.

Context of Binding Comparison...

- Sequence or structural similarity && Not same molecular function
- Same function && No fold similarity

  Nicola D.G & Richard M.J, (2006) J. Mol. Biol.
- Necessity of binding site comparison methods
  - Understanding protein function
  - Understanding side effects of drugs

Challenges in site comparison

- **Point set superposition**
- Binding site → Set of points (atoms/residues)
- Determining **point-point correspondences**
- **Topology Undefined; Size of 'match' small** & unknown ('Indels' possible)
- Involves costly least squares evaluation of **rotation & translation** matrices
- **Many** possible correspondences

Geometric Hashing; Maximal Common Sub-graph Search; Depth First Traversal (incrementally determine correspondences)

Comparison of a pair of binding sites involves three aspects:

(a) representation of each site as sorted lists of distances between chosen points,

(b) alignment of two sets of distance lists and

(c) choosing a scoring scheme for arriving at a final score

Description of the site

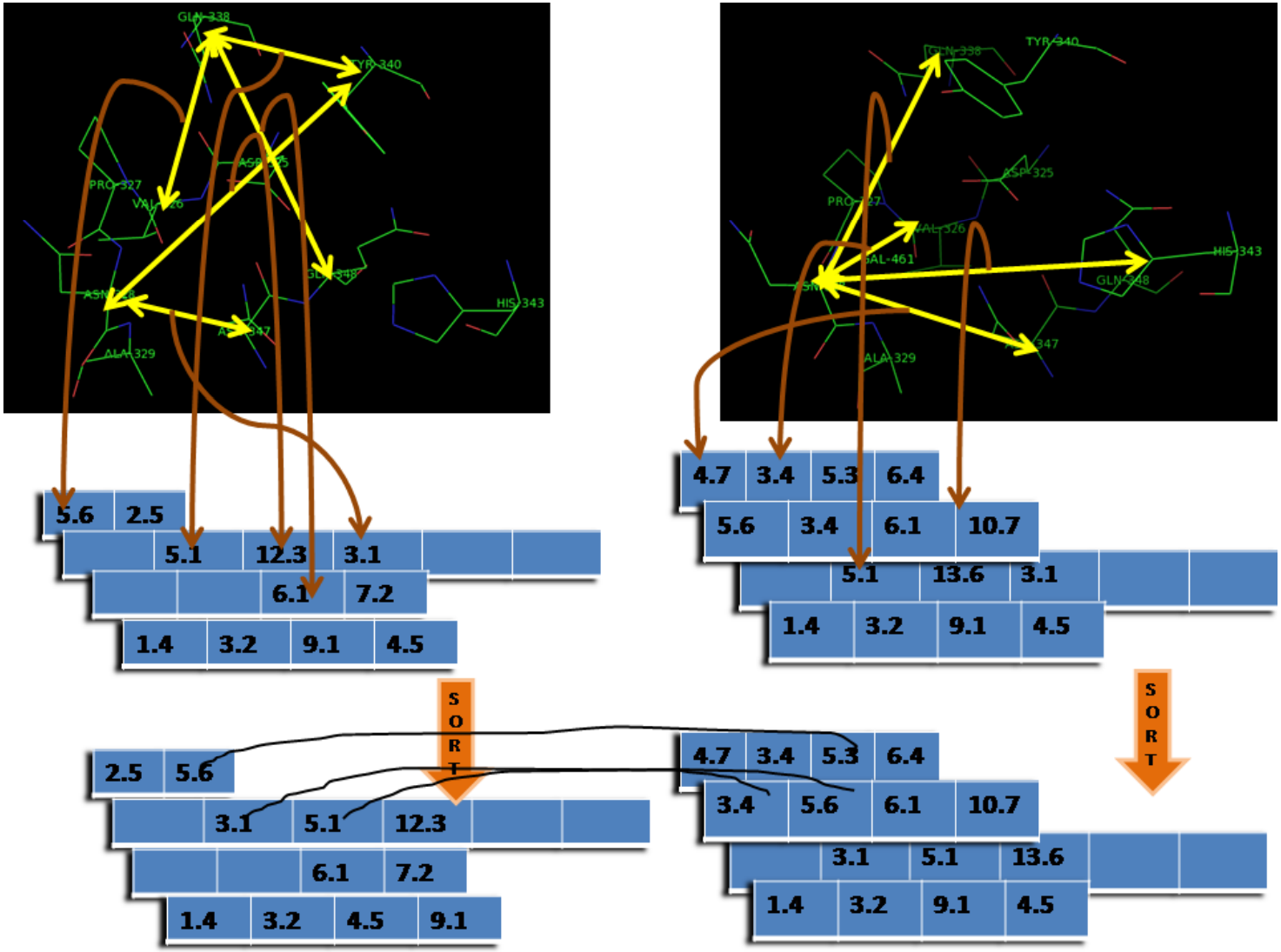
- Global features
  - Volume
  - Surface area
  - Number of polar/non-polar atoms/residues
- Shape Descriptors; Frame-invariant representations
  - Image moments
  - Spherical harmonics
  - All pair sorted distance sequences (*PocketMatch*)

(Morris et al., 2005, Bioinformatics ; Gold and Jackson, 2006, NAR; CavBase - Kuhn et al., 2007, CHEMMEDCHEM; PINTS – Stark et al., 2003, NAR; SPASM & RIGOR – Gerard et al., 1999, JMB; Binkowski et al., 2003, JMB; Morris et al., 2005, Bioinformatics; Nagano et al., 2002, JMB; Kunin et al., 2001, JMB; Campbell et al., 2003, An et al., 2005)

Tools for Binding site comparison :

- **PocketMatch-** A new algorithm to compare binding sites in protein structures
- **CavBase**
- **SitesBase-** a database for structure-based protein–ligand binding site comparisons
- **CPASS** - Comparison of Protein Active-Site Structures
- **PINTS**- Patterns in Non-homologous Tertiary Structures
- **Spasm/RIGOR**
- **SMAP-WS Pairwise Comparison /SMAP-WS Database Search**
- **SiteSorter™ -**N-by-N Binding Site Similarity Assessment
- **SLiC -**Site-Ligand Contact Analysis and Binding Mode Similarity Assessment
- **MAPPIS(M**ultiple **A**lignment of **P**rotein-**P**rotein **I**nterface**S (PPIs))-**
  Recognizes spatially conserved chemical interactions shared by a set of PPIs
- **MULTIBIND(Multi**ple Alignment of Protein **Bind**ing Sites)-
  Recognizes Spatial Chemical Binding Patterns Common to a Set of Protein Structure

PocketMatch Algorithm

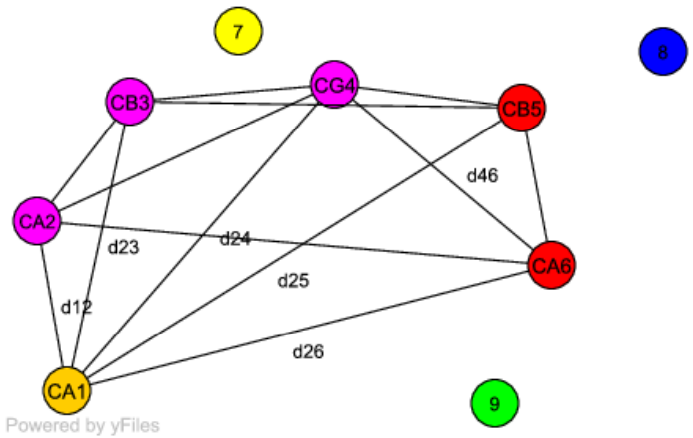


Number of matching distance elements

$$PMScore = \frac{\Sigma_{i=1}^{90} Count_i}{maximum(|S_1|,|S_2|)}$$

PocketMatch Algorithm

http://proline.physics.iisc.ernet.in/pocketmatch/



Representation of the binding site :

$$NGP$$
$$NTP$$

$$\left\{ \begin{array}{lllll} ND_1 & & & & \\ d_1, & d_2, & \cdots & d_j, & \cdots \end{array} \right\}$$

$$\left\{ \begin{array}{lllll} ND_2 & & & & \\ d_1, & d_2, & \cdots & d_j & \cdots \end{array} \right\}$$

$$\vdots$$

$$\left\{ \begin{array}{lllll} ND_{90} & & & & \\ d_1, & d_2, & \cdots & d_j & \cdots \end{array} \right\}$$

Where, NGP : Number of pairs of group-types, NTP : Number of pairs to point-types, ND_i : Number of distances in the ith bin, d_j : distance between j^{th} pair of points.

Sub-routine 1 Alignment of a pair of sorted distance sequences

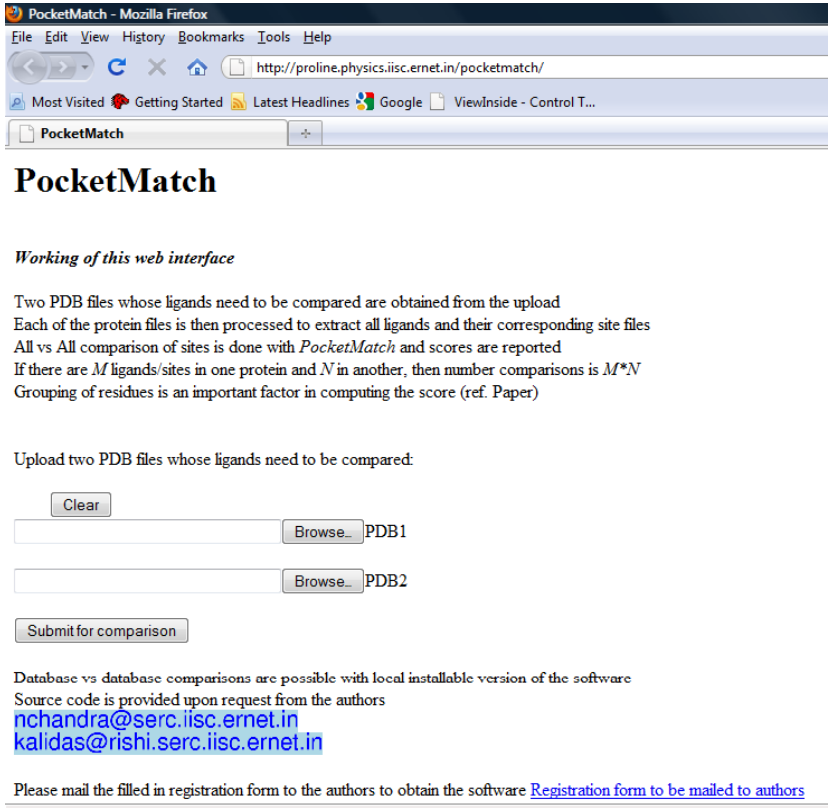
```
i=0; j=0; counter=0;
while (i ≤ m) ∧ (j ≤ n) do
  if |S1[i] − S2[j]| ≤ τ then
    i ← i + 1; j ← j + 1
    counter ← counter + 1
  else
    if S1[i] < S2[j] then
      i ← i + 1;
    else
      j ← j + 1;
    end if
  end if
end while
```

3 types of points (CA,CB,CNTR)
5 types of residue groups (AVILGP; KRH; DE; YFW; CSTQN)
(3*(3-1)/2+3)*(5*(5-1)/2+5) → 90 lists
120 lists are possible and yielded similar results.

PocketMatch: A new algorithm to compare binding sites in protein structures
Kalidas Yeturu and Nagasuma Chandra, 2008, BMC Bioinformatics

PocketMatch implementation

- Sites extracted → around (4Å) ligand/predicted pocket
- Complete residues → representative points → Sorted Distance lists
- MPI version (C language)
- Run on **IBM Bluegene** utilizing 1024 processors



Perturbation studies

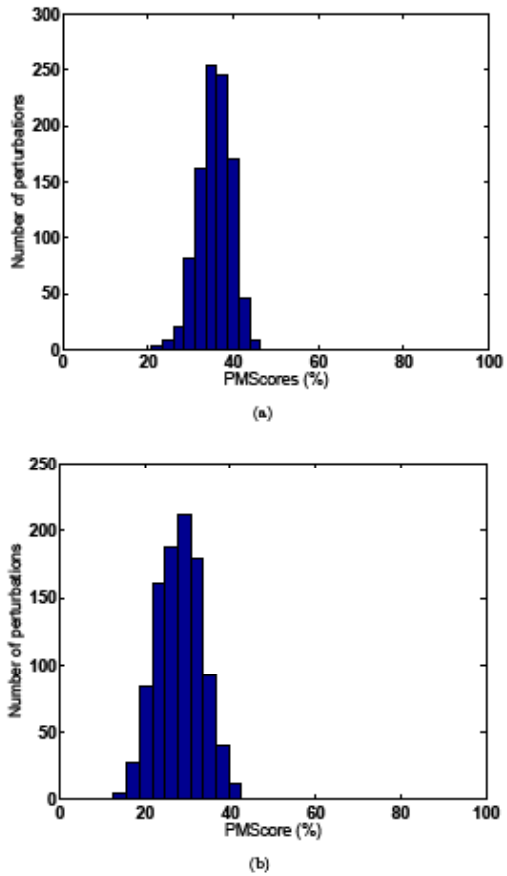


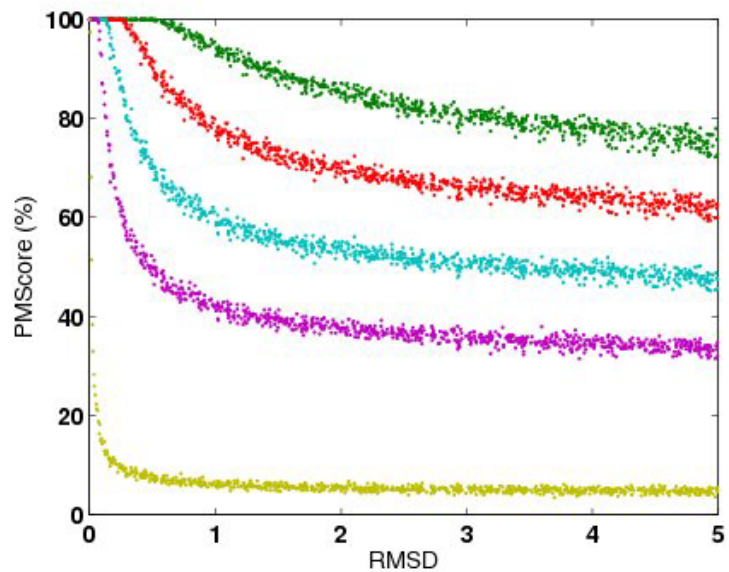
Figure 6:

Type perturbation

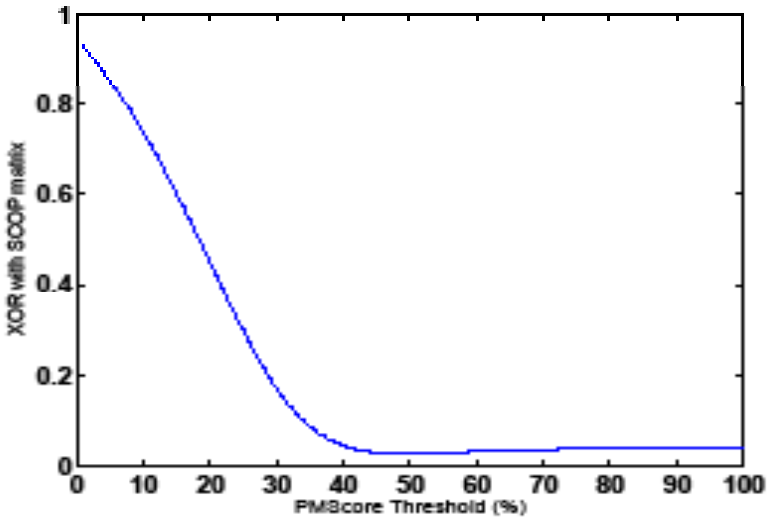
Superposition of sites with (a) High PMScores (80.9% for 1H8H-ATP and 1W0K-ADP) and (b) low \ PMScores (25.8% for 1H8H-ATP and 1H8H-ADP)

Validation with respect to random perturbation of positions of site-points

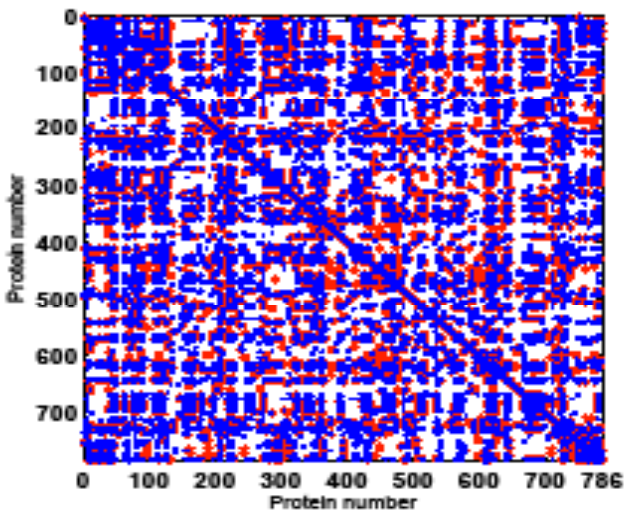
Random perturbations of site points for (a) ligand(PP8) with 54 PMScores for perturbed sites with respect to its original site for different extents of perturbations(RMSD) are shown at different values of (1.0-green,0.5-red,0.25-cyan,0.125-blue,0.01-yellow)



Validation

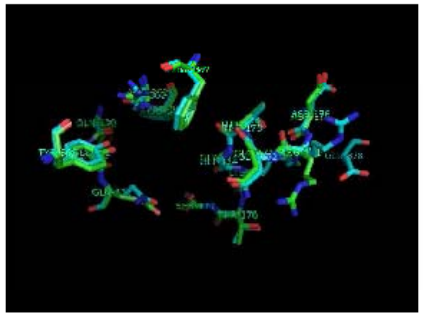


(a)

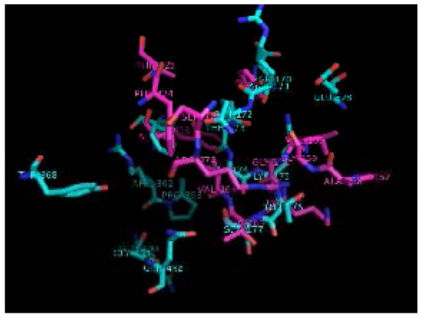


(b)

SCOP VS PM

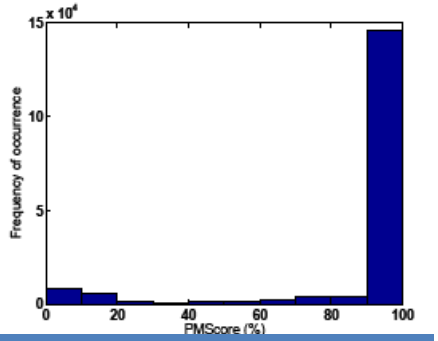


(a)



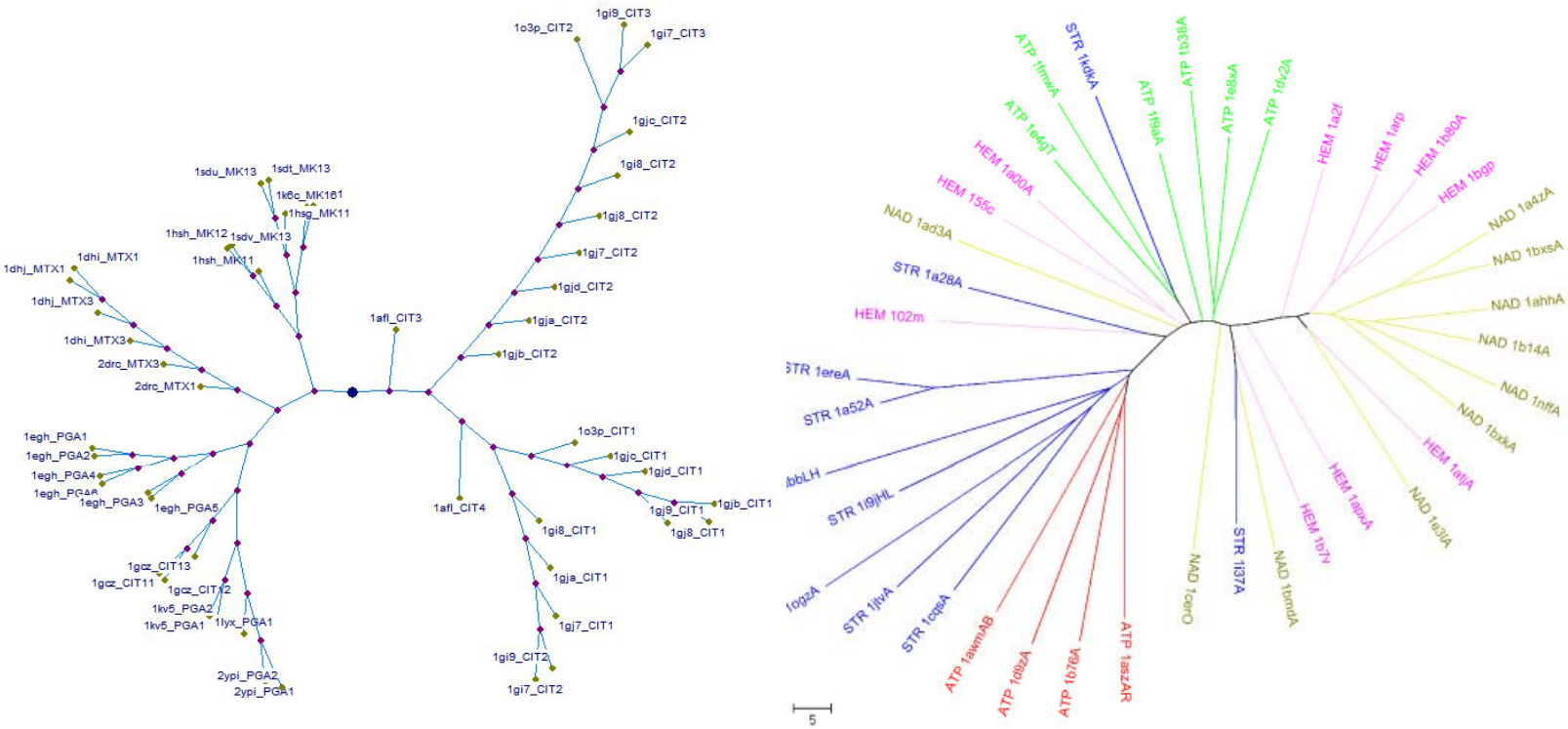
(b)

ATP-ADP Similar and dissimilar sites



Known similarities in tetramers

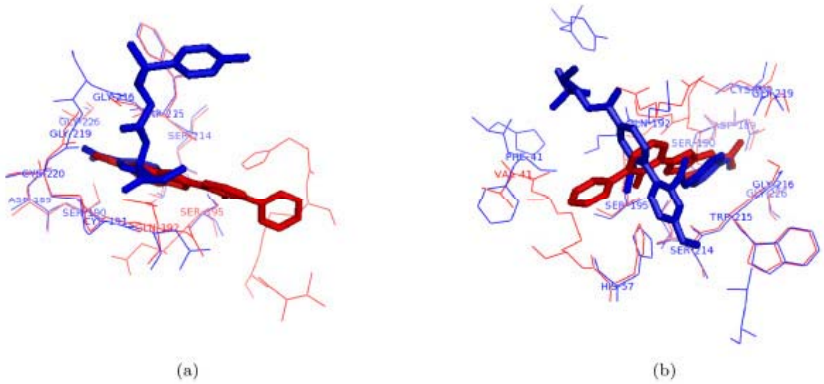
Cladogram based validation



Cladogram for pairs of sites

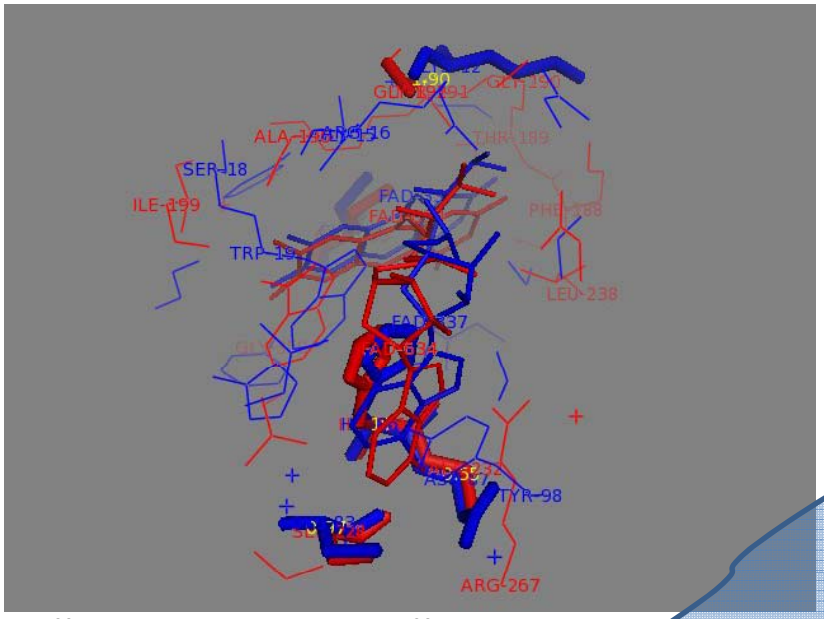
Detection of part-similarities by PocketMatch.

Examples illustrating binding of different ligands in essentially the same binding pocket, but with different orientations. The part-similarities in these were identified correctly by PocketMatch. Binding of different trypsin inhibitors (stick models) complexed to trypsin variants (wire) as in PDB entries (a) 1GJC and 1V2Q and (b) 1GJC and 2AYW.

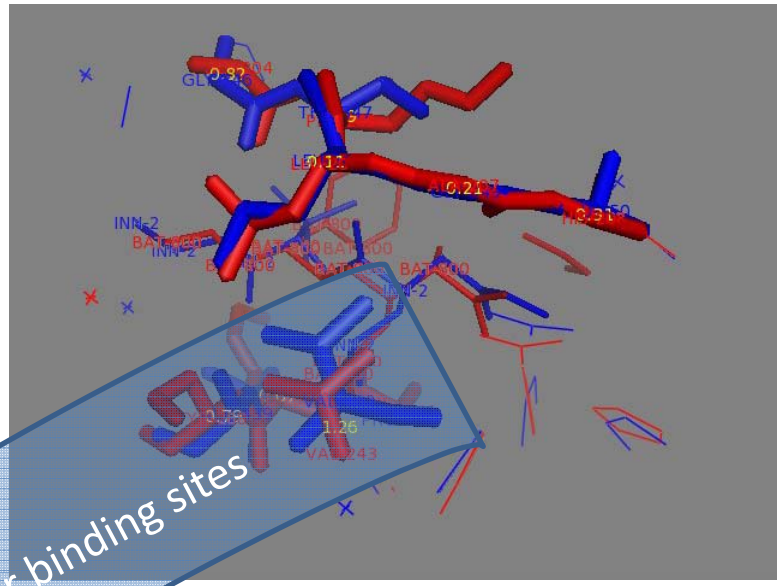


Need for site alignment

- Different folds exhibit similar binding sites
  - Ex – cofactor binding sites HEM, NAD, FAD
- Difficult to detect local similarities by human – error prone
- Structural motifs – determining function

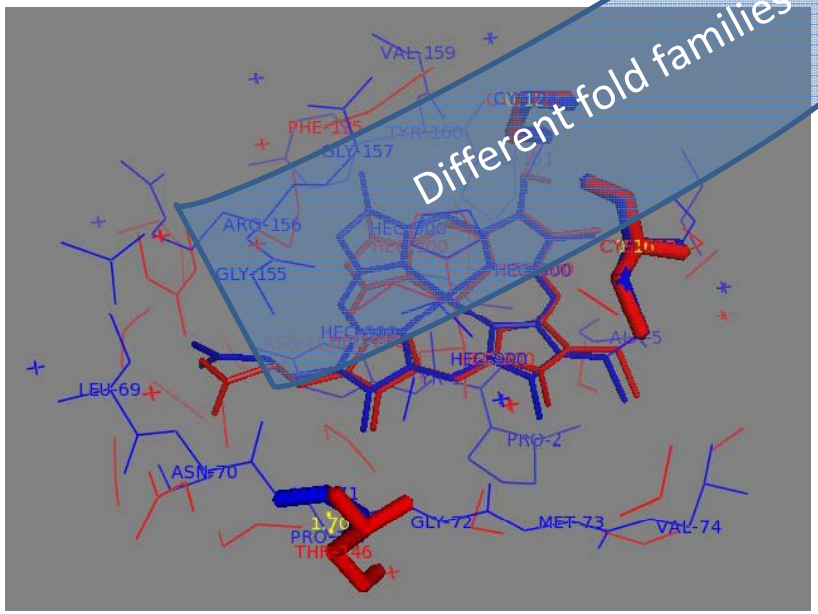


pdb1rp4.ent_FAD-pdb1jra.ent_FAD-pa

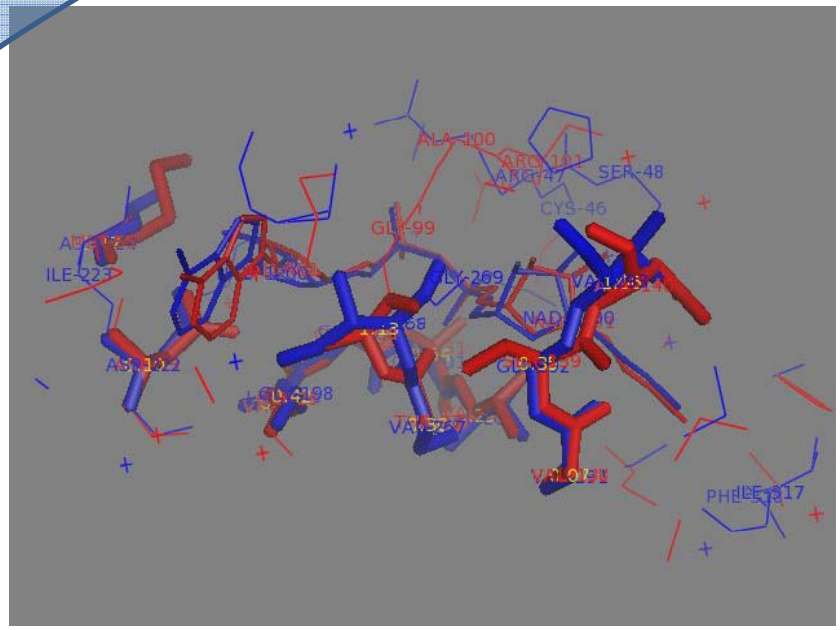


pdb1rm8.ent_BAT-pdb1bkc.ent_INN-pa

Different fold families – similar binding sites



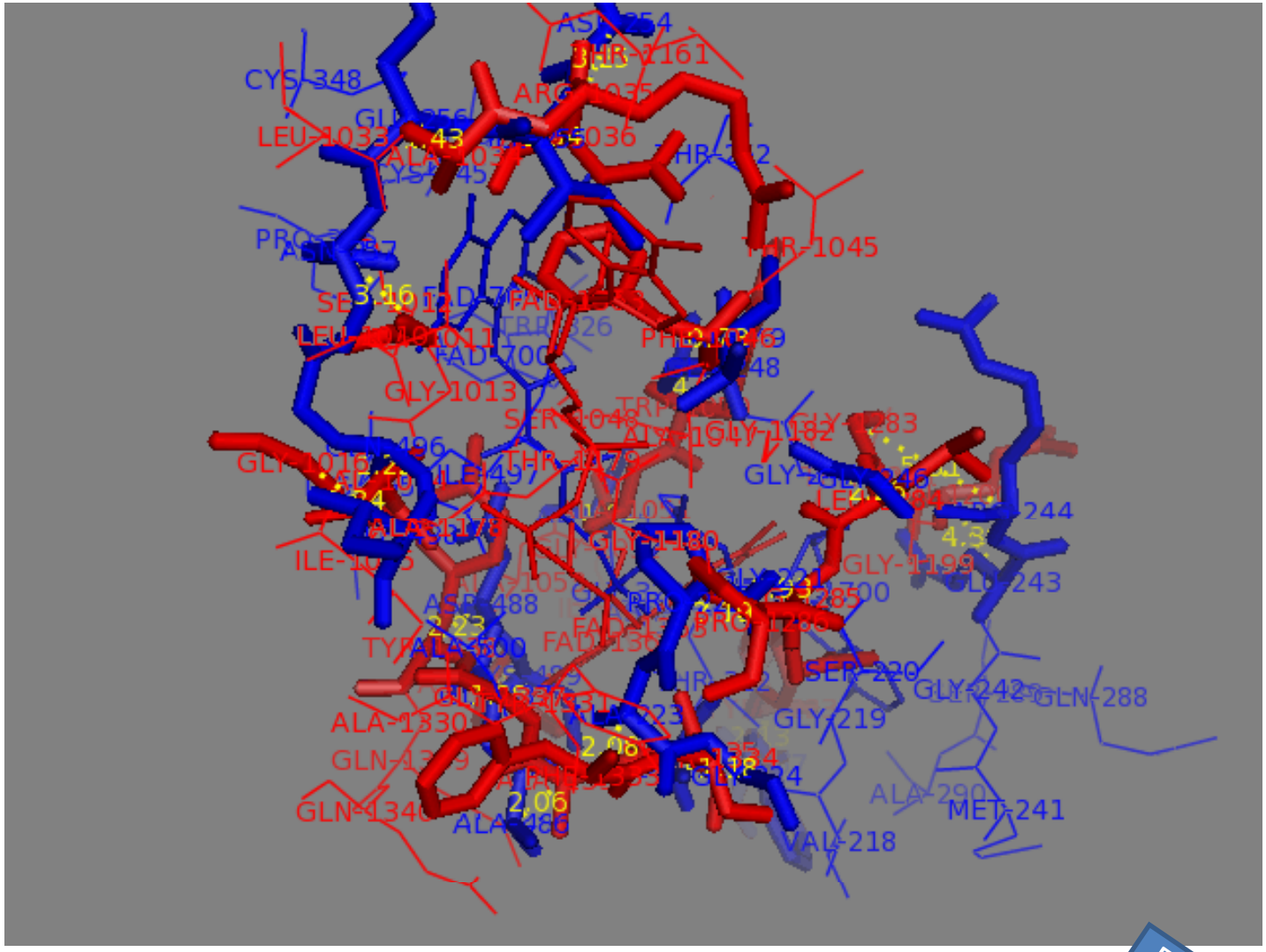
pdb1m6z.ent_HEC-pdb1e2z.ent_HEC-pa



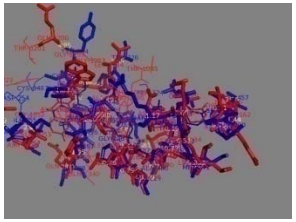
pdb9ldt.ent_NAD-pdb1ee2.ent_NAD-pa

Challenges in alignment

- Many possible local similarities exist
  - Exhaustive enumeration is impractical
- Finding out the best is tough
  - Quantify when does an expert call superposition 'good'
- What level to consider structural match
  - Atomic, Residue, C-alpha



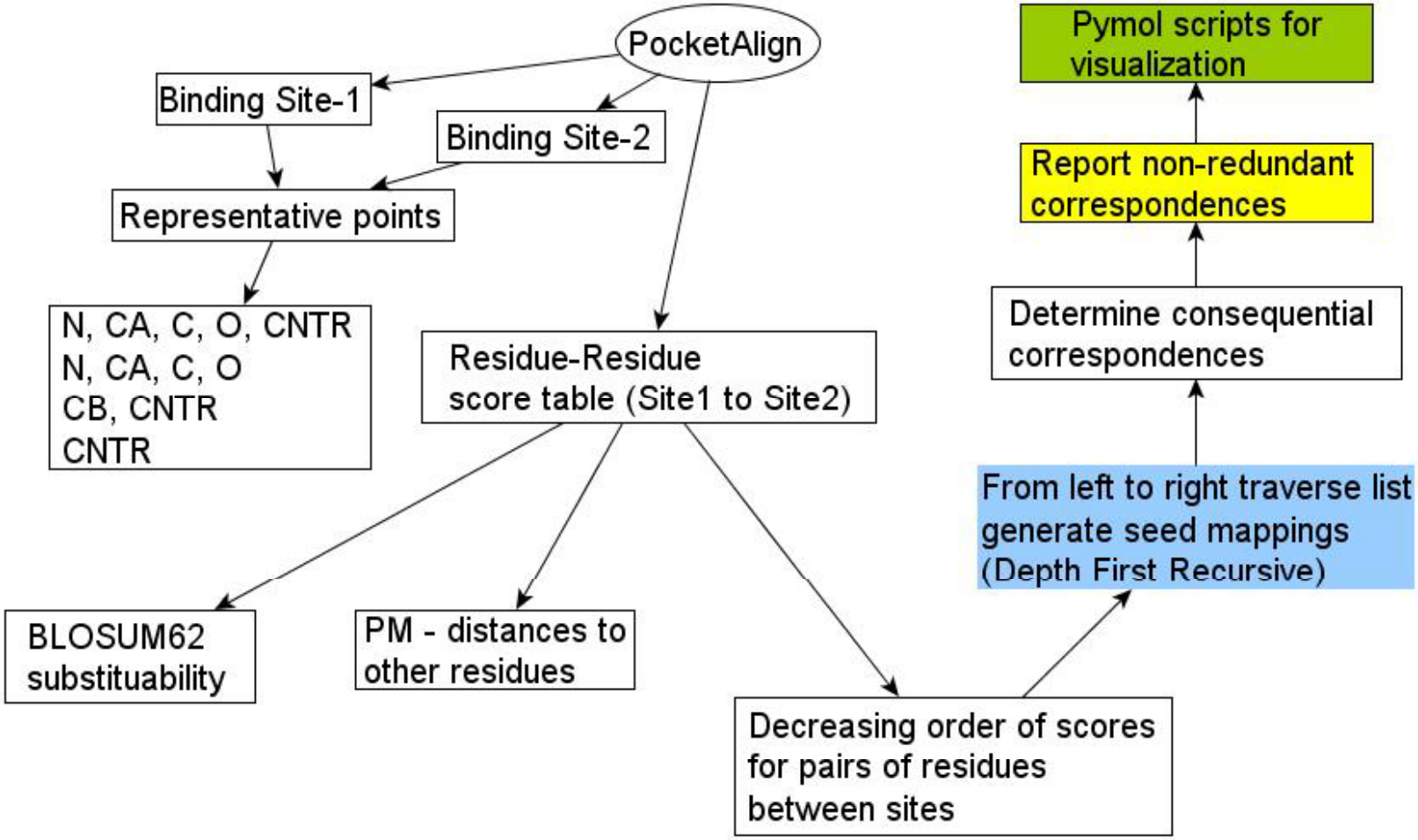
Extract



Superpositions of FAD binding sites
1C0P and 1HYU

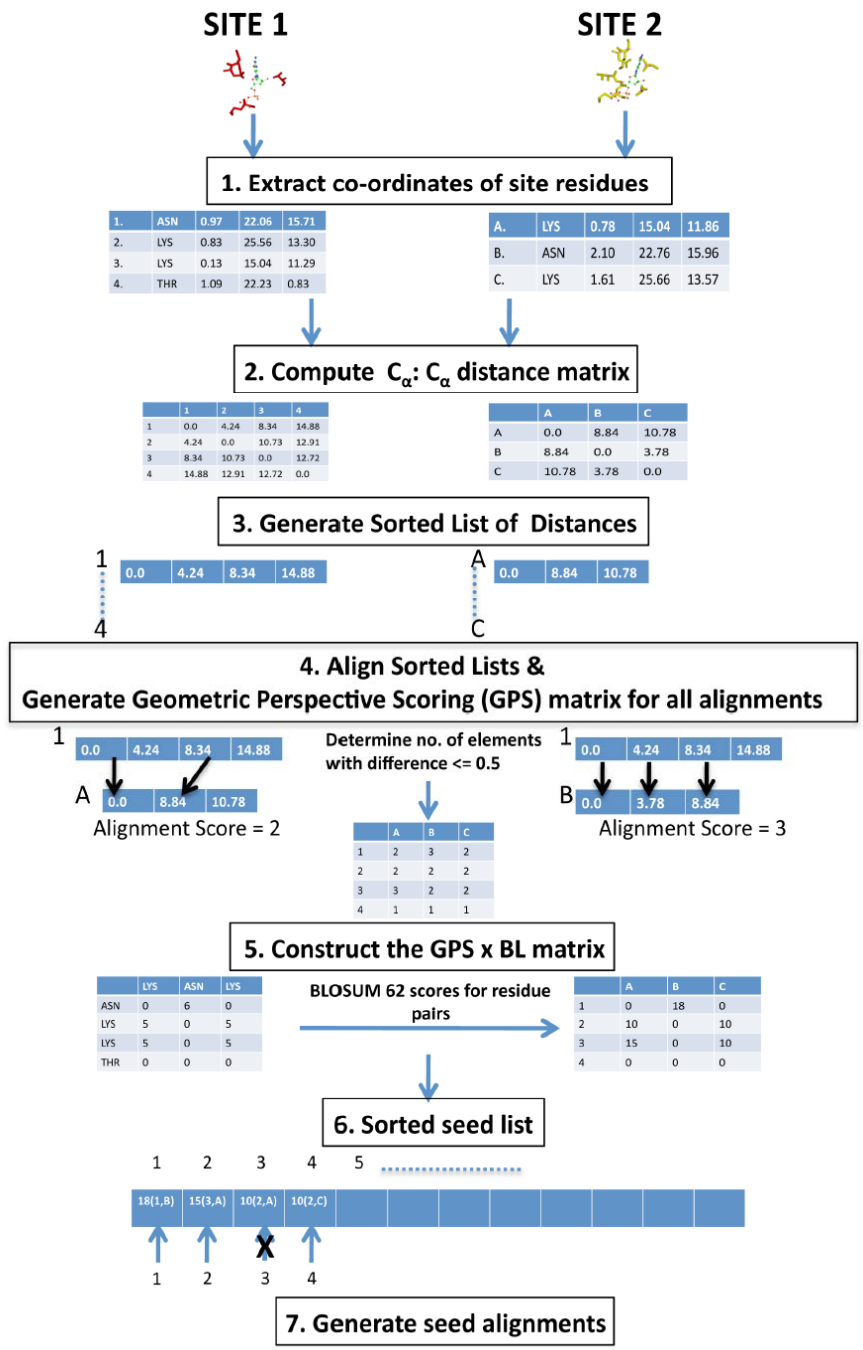
POCKETALIGN

PocketAlign
Alignment of binding sites

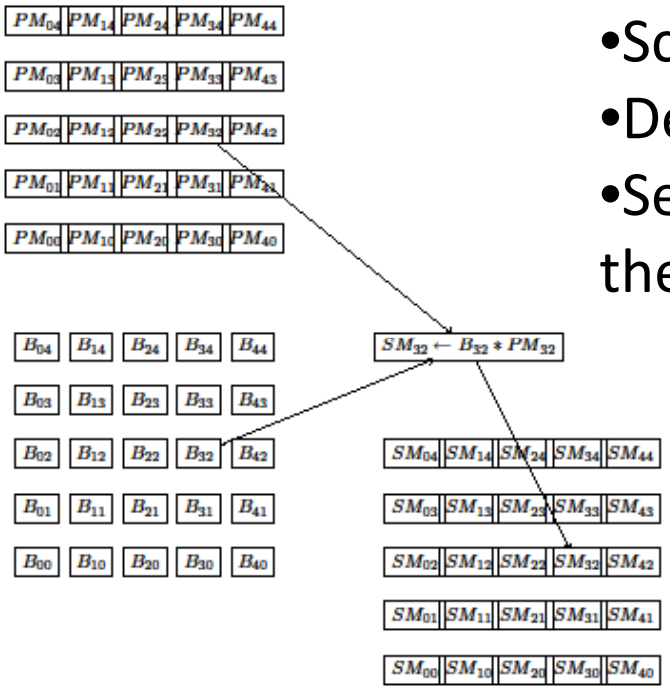


Yeturu and Chandra, JCIM (ACS), 2011, In press

Schematic



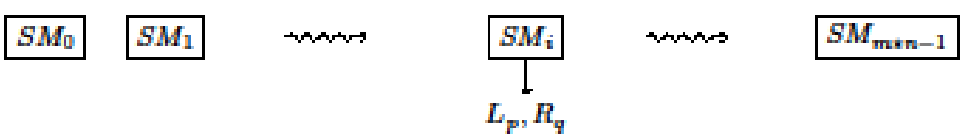
Entries 1, 2 and 4 yield a seed alignment.



- Scores between residue pairs
- Descending order sorted pair-scores
- Selection of top pair from left moving righ on the string

Algorithm 1 Generation of seed alignments

```
for i = 1 to m * n do
  for j = i to m * n do
    p ← SMM[j].residue[0]
    q ← SMM[j].residue[1]
    if p and q are not already mapped then
      if RMSD criteria is met for map ∪ ⟨p, q⟩ then
        {Above check considers current alignment type}
        map ← map ∪ ⟨p, q⟩
      end if
    end if
  end for{#j}
  Update database with map
end for{#i}
```

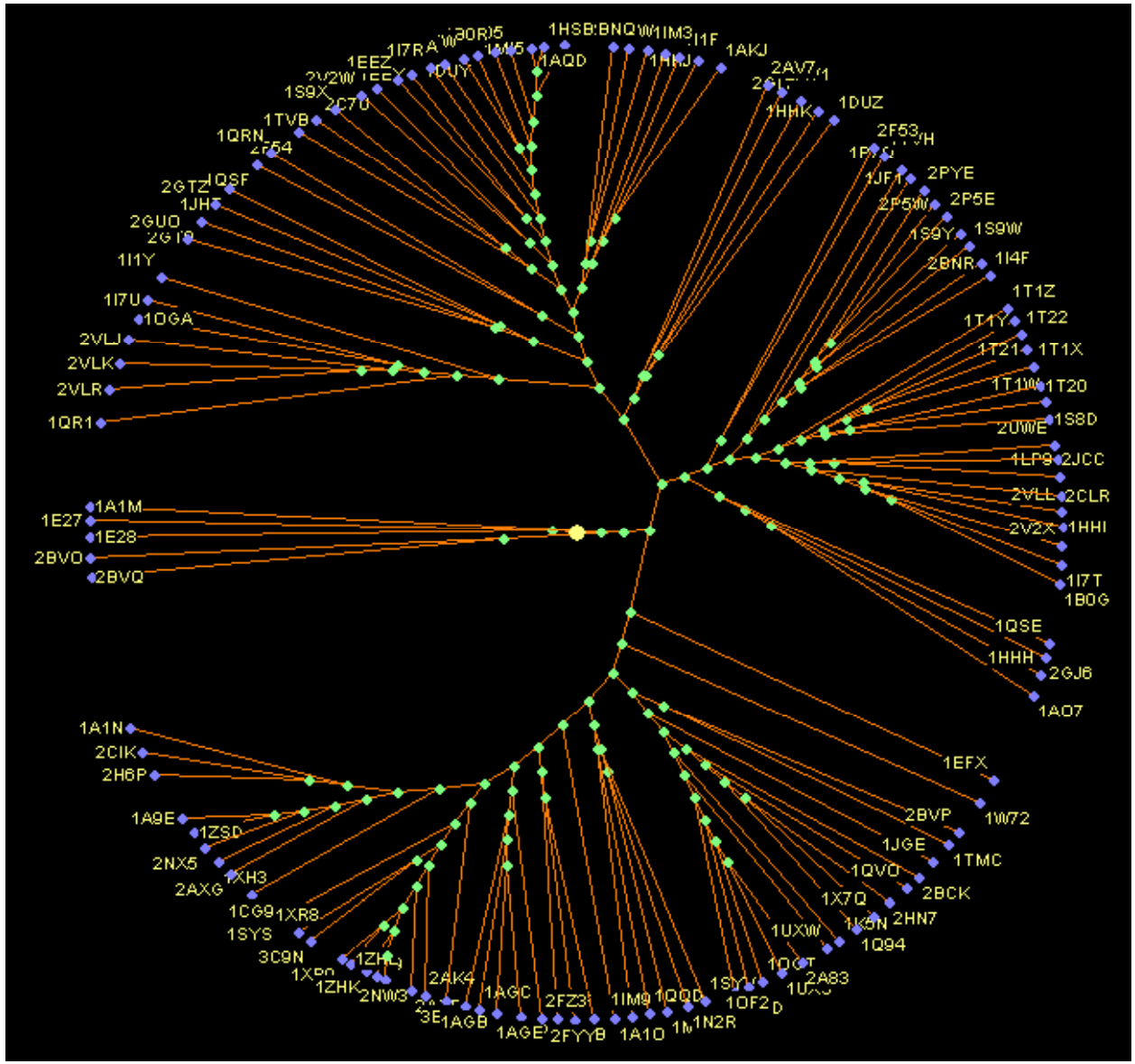


Two binding sites are represented as sets of residues
$S = \{R_1 ... R_m\}$ where R_i is i^{th} residue of first site
Each residue defines a partitioning of the set of atoms, A
$R = \{a \in A\} \subset A$
$R_i \cap R_j = \emptyset (\forall i \neq j \in |S|)$
Where $|S|$ denotes cardinality of the set, S
Similarly second site is represented by $S' = \{R'_1 ... R'_n\}$ on set of atoms, A'
Chemical similarities are denoted by a function $BL : S \times S' \rightarrow N$
Geometric similarities (GPS) are denoted by $GPS : S \times S' \rightarrow N$
A combination scoring scheme is defined $GPS \times BL_{ij} \rightarrow GPS_{ij} * BL_{ij}$
A linearization of $GPS \times BL$ is performed
A one-to-one function is defined $L : [1...m] \times [1...n] \rightarrow [1...m * n]$
SeedList is created by obtaining values from $GPS \times BL$
$SeedList^V_{L(i,j)} \leftarrow GPS \times BL_{ij}$ for storing the values
$SeedList^P_{L(i,j)} \leftarrow (i, j)$ for storing the residue pairs
SeedList is sorted such that $(\forall p \leq q) SeedList^V_p \geq SeedList^V_q$
A mapping is defined as residuewise correspondences between the two sites
A one-to-one function, for a mapping $M : [1...m] \rightarrow [1...n]$
Seed mapping or alignment B is derived by traversal of SeedList
$B \leftarrow \{(p, q)\} \subset SeedList^P$

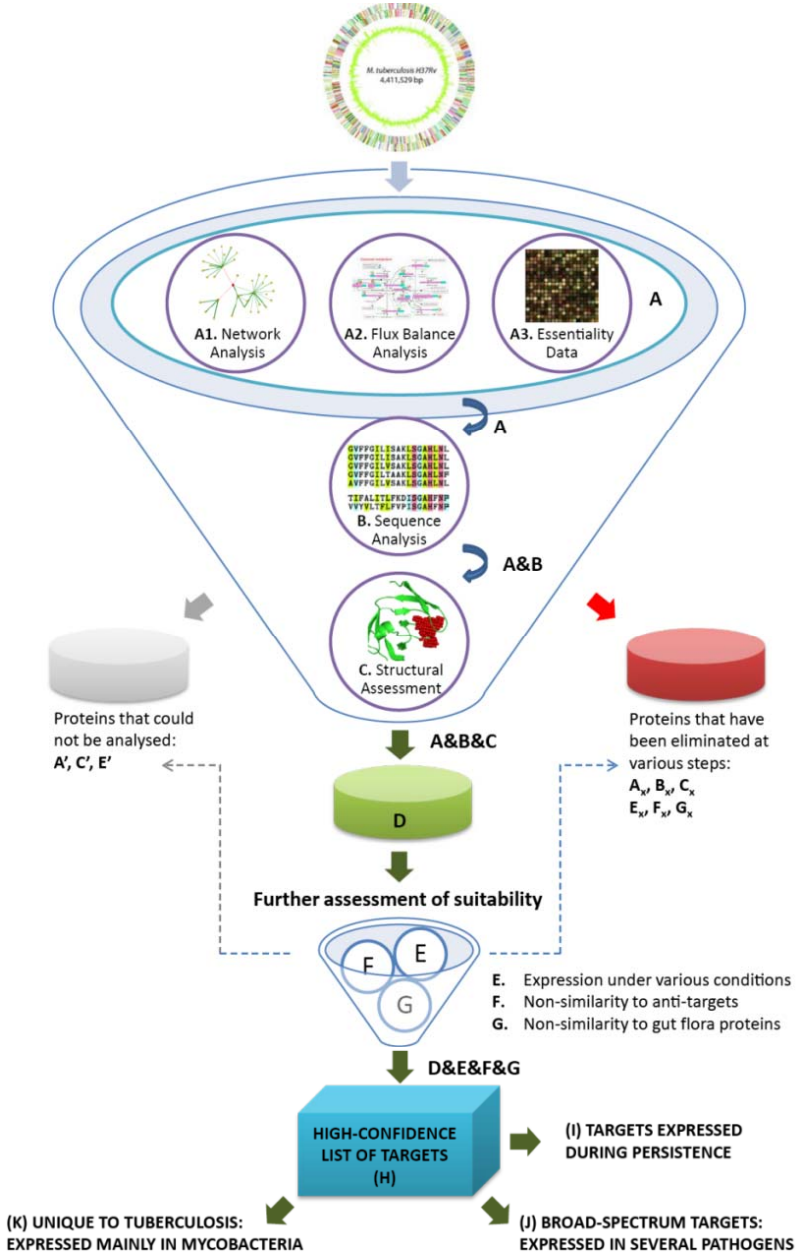
PocketAlign (Validation & New Results)

- Ran for a set of 34 pairs of sites known to be similar
- Encouraging results obtained from a set of 143 pairs of histamines, 29 pairs of lectins, 209 pairs of sites of carbohydrate (GAL, GLC and MAN) sites and ATP binding sites

Cladogram of MHC class I binding sites of 120 molecules from various species



targetTB – Target Identification Pipeline



(A&B) Systems and Sequence Level Filters

- A1 Node deletions on STRING + Metabolic Influences network
- A2 Essential genes from *Mtb iNJ*661, GSMN-TB
- A3 High-throughput Transposon Site Hybridisation (TraSH) Mutagenesis study
- B Eliminated proteins with close homologues in human proteome

(C) Structural Assessment of Targetability

- Binding site prediction and comparison – *Mtb* vs. *Hsa*
- Structural models obtained from ModBase
- Binding sites identified using PocketDepth and compared using PocketMatch (cut-off: 0.80) (A&B&C ⇒ D)

Other Filters (applied to (D))

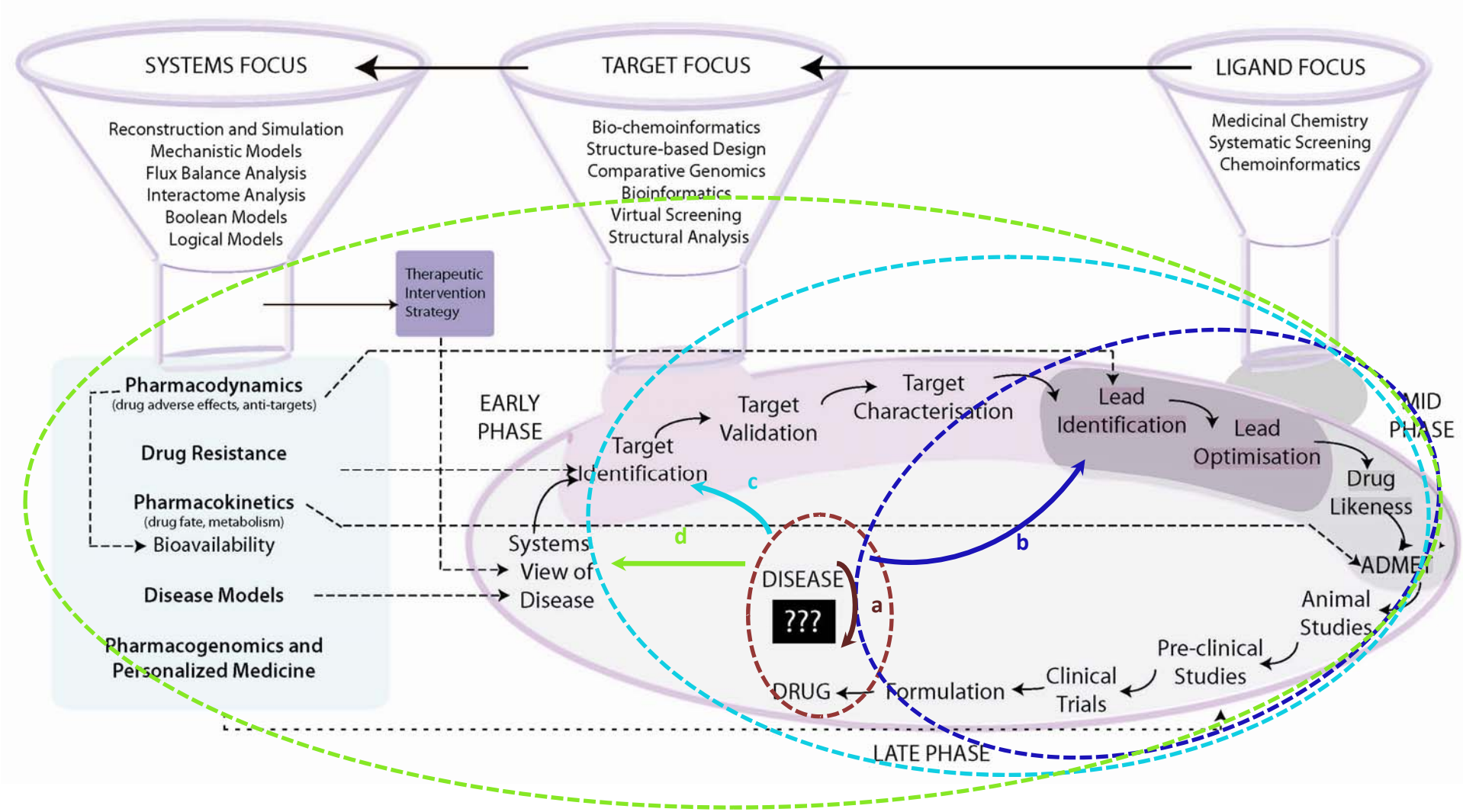
- E Expression of target (Microarray data)
- F Non-similarity to human 'anti-targets'
- G Non-similarity to gut flora proteins
- Paths to resistance mechanisms

Multiple Lists of Targets

- H Passing filters A–G
- I H-List targets upregulated in persistence
- J H-List targets that can serve as broad-spectrum targets
- K H-List targets, unique to *Mtb*.

Raman et al ;. 2008, BMC Syst. Biol

THE NEW DRUG DISCOVERY PIPELINE



Chandra, Expert Opin Drug Disc. 2009


SUMA LAB

Thank You