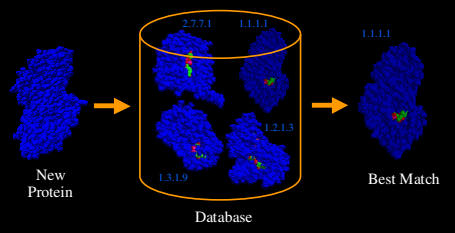
Representing and Matching Binding Sites with Grids

Thomas Funkhouser
Princeton University
CS597A, Fall 2005

Goal

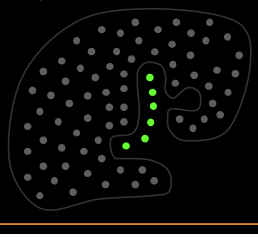
Match binding site to others in a database of proteins with known biochemical function



Binding Site Representations

Possible binding site descriptions

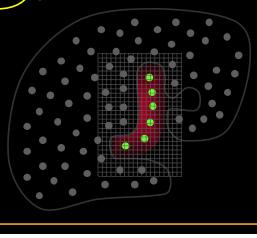
- Points (atoms, residues, pseudo-centers, critical points, …)
- Surfaces (spheres/tori, meshes, radial extent, …)
- Volumes (alpha shapes, grids, …)



Binding Site Representations

Possible binding site descriptions

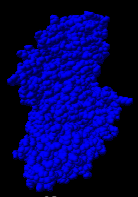
- Points (atoms, residues, pseudo-centers, critical points, …)
- Surfaces (spheres/tori, meshes, radial extent, …)
- Volumes (alpha shapes, grids, …)



Grid Representations

Binding site …

- Detection
- Modeling
- Alignment
- Matching
- Classification
- Analysis

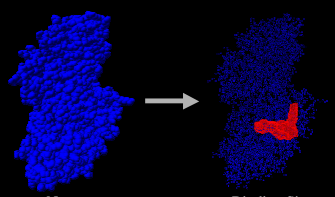


Grid Representations

Binding site …

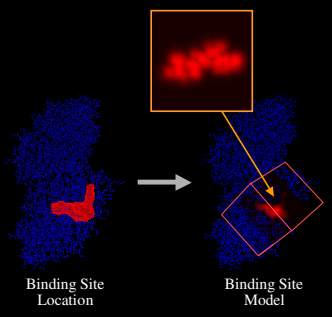
Ø Detection

- Modeling
- Alignment
- Matching
- Classification
- Analysis



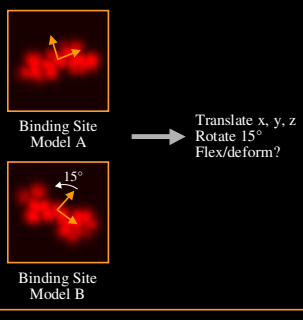
Grid Representations

Binding site …
- Detection
- Ø Modeling
- Alignment
- Matching
- Classification
- Analysis



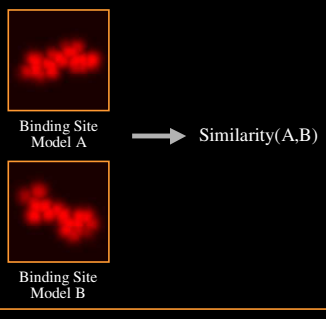
Grid Representations

Binding site …
- Detection
- Modeling
- Ø Alignment
- Matching
- Classification
- Analysis



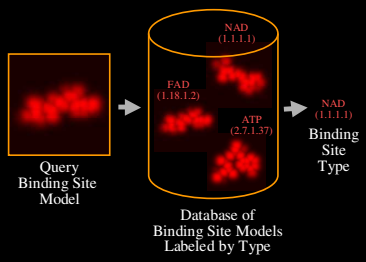
Grid Representations

Binding site …
- Detection
- Modeling
- Alignment
- Ø Matching
- Classification
- Analysis



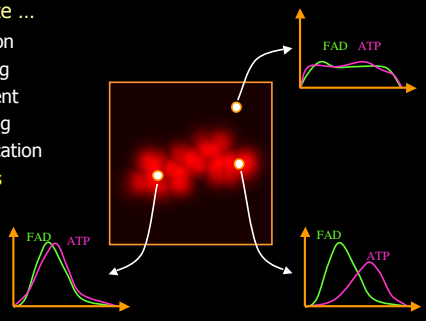
Grid Representations

Binding site …
- Detection
- Modeling
- Alignment
- Matching
- Ø Classification
- Analysis



Grid Representations

Binding site …
- Detection
- Modeling
- Alignment
- Matching
- Classification
- Ø Analysis



Outline

Introduction

Binding site modeling with grids
- Simulation-based
- Knowledge-based

Binding site matching with grid correlation
- Fast rotational matching

Searching a database with grid signatures
- Power spectrum signature

Results

Discussion

Outline

Introduction

Binding site modeling with grids

Ø Simulation-based ←

- Knowledge-based

Binding site matching with grid correlation

- Fast rotational matching

Searching a database with grid signatures

- Power spectrum signature

Results

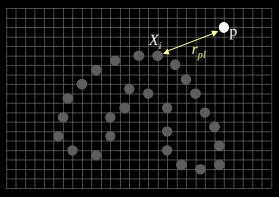
Discussion

Simulation-Based Modeling

Compute propensity for different "probes" to appear at grid points

$$P_p^0 = \sum_{i=1}^{N}\left(\frac{A_{X_i C}}{r_{pl}^{12}} - \frac{B_{X_i C}}{r_{pl}^{6}}\right)$$

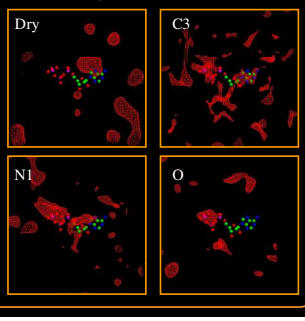
Lenert-Jones Potential



[At05]

Simulation-Based Modeling

Compute propensity for different "probes" to appear at grid points



Predicted Binding Site Model for 1kp8-1-H-ATP-1-_ using GRID

[Goodford85]

Simulation-Based Modeling

Grids of this type are sometimes used to accelerate computation of scoring functions in docking programs

$$\begin{aligned}\Delta G_{bind} = {} & C_{lipo-lipo}\sum f(r_{lr}) + \\ & C_{hbond-neut-neut}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{hbond-neut-charged}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{hbond-charged-charged}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{max-metal-ion}\sum f(r_{lm})h(\Delta\alpha) + \\ & C_{rotb}H_{rotb} + \\ & C_{polar-phob}V_{polar-phob} \\ & C_{coul}E_{coul} + \\ & C_{vdW}E_{vdW} + \\ & solvationterms\end{aligned}$$

[Friesner04]

Outline

Introduction

Binding site modeling with grids

- Simulation-based

Ø Knowledge-based ←

Binding site matching with grid correlation

- Fast rotational matching

Searching a database with grid signatures

- Power spectrum signature

Results

Discussion

Knowledge-Based Modeling

Simulation from physical principles is difficult

$$\begin{aligned}\Delta G_{bind} = {} & C_{lipo-lipo}\sum f(r_{lr}) + \\ & C_{hbond-neut-neut}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{hbond-neut-charged}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{hbond-charged-charged}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{max-metal-ion}\sum f(r_{lm})h(\Delta\alpha) + \\ & C_{rotb}H_{rotb} + \\ & C_{polar-phob}V_{polar-phob} \\ & C_{coul}E_{coul} + \\ & C_{vdW}E_{vdW} + \\ & solvationterms\end{aligned}$$

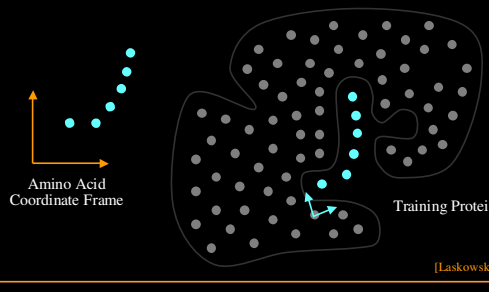
Knowledge-Based Modeling

Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites

[Laskowski96]

Knowledge-Based Modeling

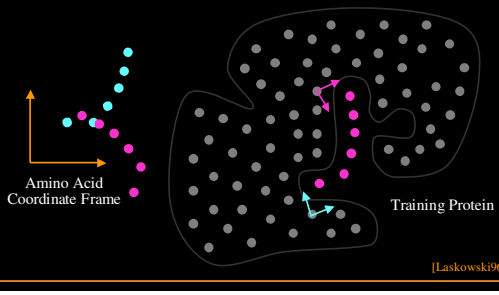
Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



[Laskowski96]

Knowledge-Based Modeling

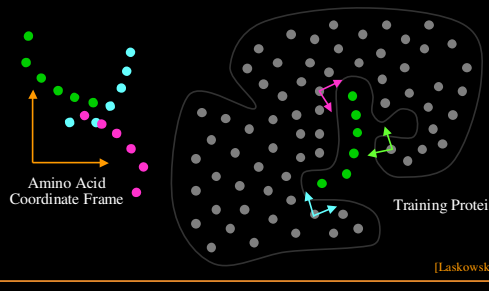
Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



[Laskowski96]

Knowledge-Based Modeling

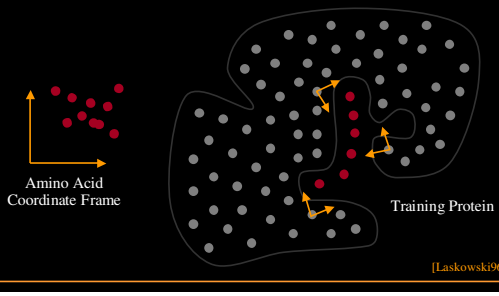
Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



[Laskowski96]

Knowledge-Based Modeling

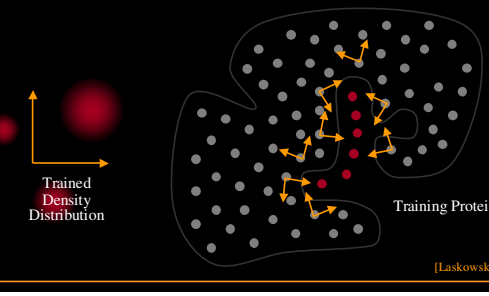
Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



[Laskowski96]

Knowledge-Based Modeling

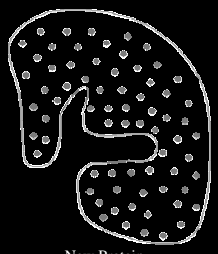
Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



[Laskowski96]

Knowledge-Based Modeling

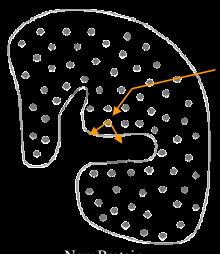
Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



New Protein [Laskowski96]

Knowledge-Based Modeling

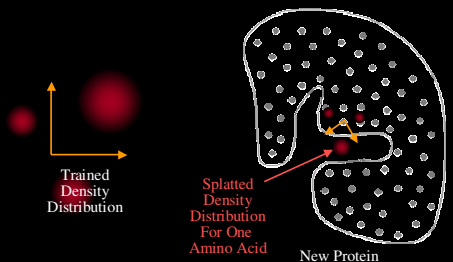
Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



New Protein [Laskowski96]

Knowledge-Based Modeling

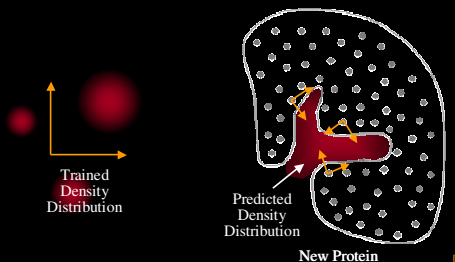
Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



New Protein [Laskowski96]

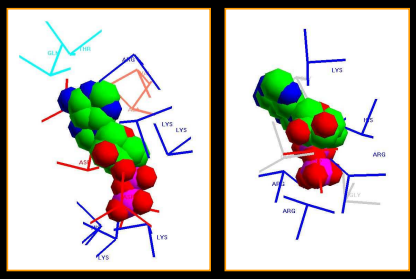
Knowledge-Based Modeling

Train on distributions of ligand atoms in bound proteins to develop predictive model for new binding sites



New Protein [Laskowski96]

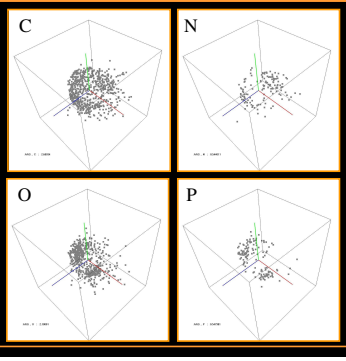
Knowledge-Based Modeling



1mxb-1-A-ADP-385-_ 4pfk-1-A-ADP-326-_

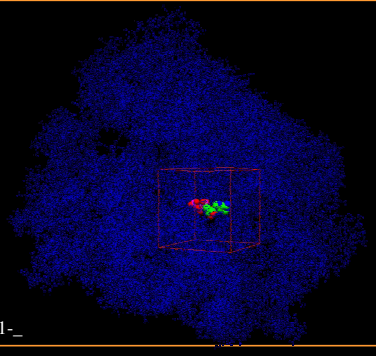
Residue Coordinate Frames

Knowledge-Based Modeling



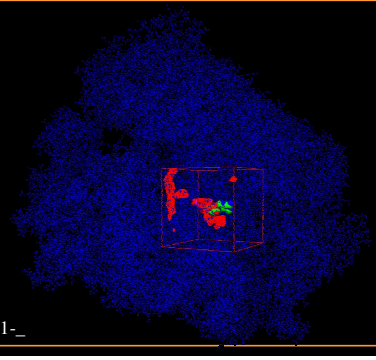
Trained
Density Distributions
for Argenine

Knowledge-Based Modeling



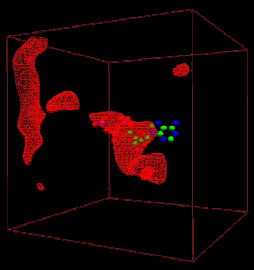
Predicted Binding Site Model for 1kp8-1-H-ATP-1-_

Knowledge-Based Modeling



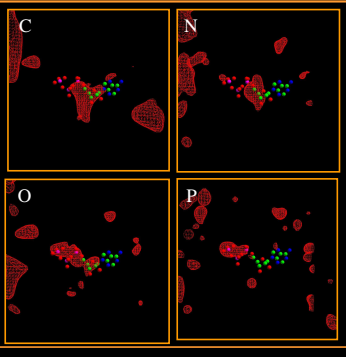
Predicted Binding Site Model for 1kp8-1-H-ATP-1-_

Knowledge-Based Modeling



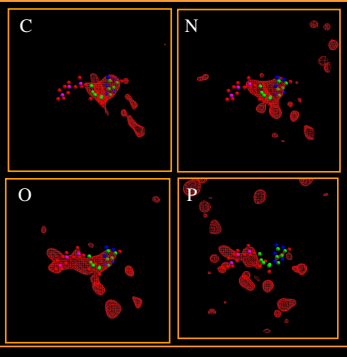
Predicted Binding Site Model for 1kp8-1-H-ATP-1-_

Knowledge-Based Modeling



Predicted Binding Site Model for 1kp8-1-H-ATP-1-_

Knowledge-Based Modeling



Predicted Binding Site Model for 1jlr-1-_-GTP-300-_

Knowledge-Based Modeling

Class	H	C	N	O	P	S	Avg
HEM	-0.745799	1.97266	-0.0882554	0.905717	0	0	**1.61848**
ATP	0	1.85844	-0.0841064	1.12217	-0.517955	0	**1.0064**
ADP	0	1.85449	-0.109433	1.07771	-0.497058	0	**1.02891**
AMP	0	1.90591	-0.0765689	0.888299	-0.600485	0	**1.07227**
MES	0	1.80551	-0.179099	0.908781	0	-0.888176	**1.11674**
EPE	0	1.81995	-0.267681	0.888044	0	-0.865757	**1.11404**
TRS	0	1.7987	-0.228735	0.887196	0	0	**1.20346**
MPD	0	1.91391	0	0.673981	0	0	**1.60392**
Avg	**-0.745799**	**1.8662**	**-0.147697**	**0.918988**	**-0.535804**	**-0.876966**	1.22053

Average #stddevs above/below mean of predicted binding site at positions of ligand atoms for each ligand and element type

Outline

Introduction

Binding site modeling with grids

- Simulation-based
- Knowledge-based

Binding site matching with grid correlation ←

- Fast rotational matching

Searching a database with grid signatures

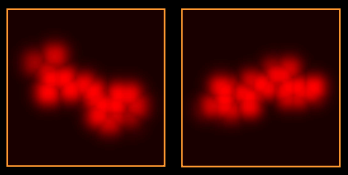
- Power spectrum signature

Results

Discussion

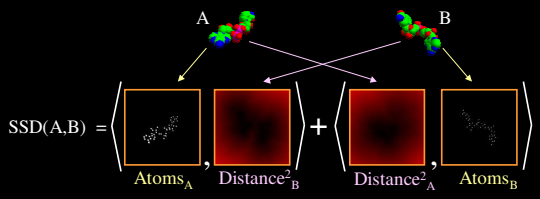
Matching with Grid Correlation

The correlation (dot product) between two grids provides a good measure of similarity



Matching with Grid Correlation

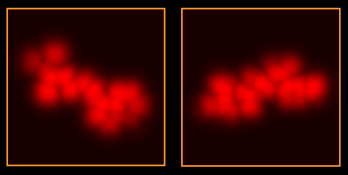
The correlation (dot product) between two grids provides a good measure of similarity



$$SSD(A,B) = \sum_{A_i \in A} \|A_i - B\|^2 + \sum_{B_i \in B} \|A_i - B\|^2$$

Fast Rotational Matching

Goal: given two sets of grids (representing molecules), find the maximal correlation over all rotations



[Kovacs et al., 2002]

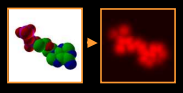
Fast Rotational Matching (2D)

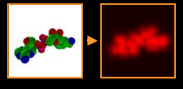
Given two molecules …

Molecule

Fast Rotational Matching (2D)

1. Compute grids representing molecules

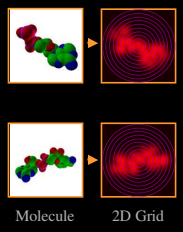




Molecule 2D Grid

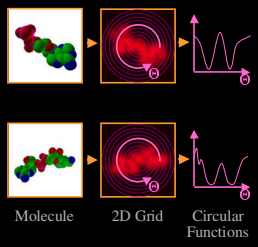
Fast Rotational Matching (2D)

2. Decompose grids into concentric circles



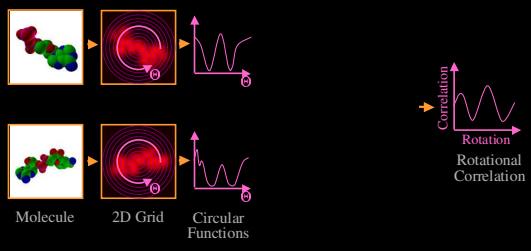
Fast Rotational Matching (2D)

3. Represent grids as 1D functions of angle for all circles



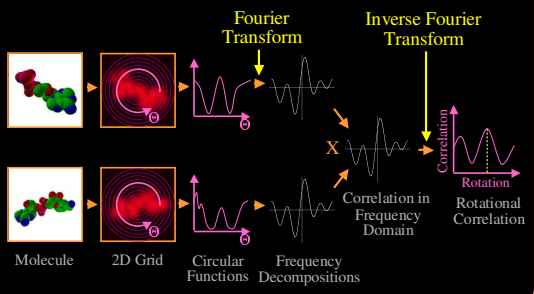
Fast Rotational Matching (2D)

4. Convolve to compute correlation for all rotations



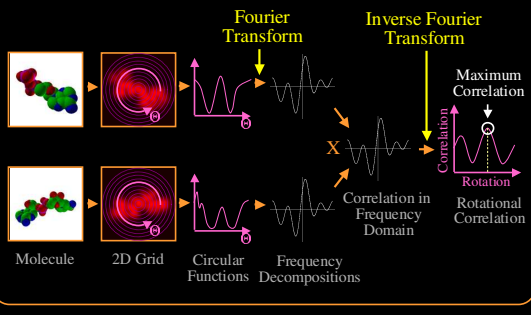
Fast Rotational Matching (2D)

4. Convolve to compute correlation for all rotations



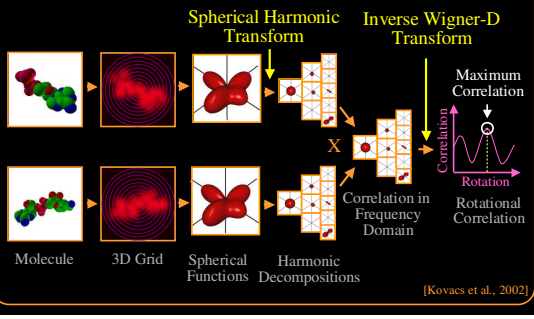
Fast Rotational Matching (2D)

5. Check correlation at every rotation to find maximum



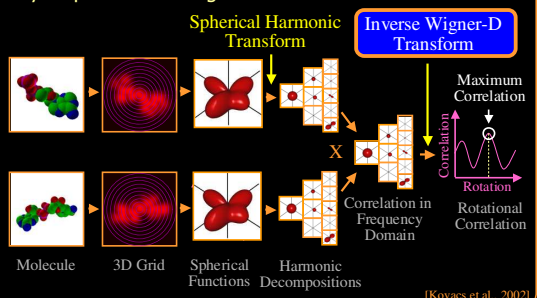
Fast Rotational Matching (3D)

Similar to 2D, but spheres instead of circles



Fast Rotational Matching (3D)

Key step is inverse Wigner-D transform



[Kovacs et al., 2002]

Fast Rotational Matching (3D)

Theoretical complexity:

- Complexity is $O(N^4)$ for NxNxN grid, rather than $O(N^6)$
- Complexity of Wigner-D^{-1} independent of #fields/molecule
- Complexity independent of #atoms

Practical complexity (times in seconds):

Grid Resolution (voxels)	Max Error (degrees)	Per Field			Per Pair		
		Spherical Grid	Spherical Harmonics	Per Field Total	Cross-Multiply	Wigner-D^{-1} Transform	Per Pair Total
32x32x32	5.5	0.02	0.01	0.03	0.01	0.02	0.03
64x64x64	2.8	0.18	0.10	0.28	0.19	0.28	0.47
128x128x128	1.4	2.47	1.17	3.64	5.55	4.45	10.00

Fast Rotational Matching (3D)

Theoretical complexity:

- Complexity is $O(N^4)$ for NxNxN grid, rather than $O(N^6)$
- Complexity of Wigner-D^{-1} independent of #fields/molecule
- Complexity independent of #atoms

Practical complexity (times in seconds):

Grid Resolution (voxels)	Max Error (degrees)	Per Field			Per Pair		
		Spherical Grid	Spherical Harmonics	Per Field Total	Cross-Multiply	Wigner-D^{-1} Transform	Per Pair Total
32x32x32	5.5	0.02	0.01	0.03	0.01	0.02	0.03
64x64x64	2.8	0.18	0.10	0.28	0.19	0.28	0.47
128x128x128	1.4	2.47	1.17	3.64	5.55	4.45	10.00

Fast Rotational Matching (3D)

Theoretical complexity:

- Complexity is $O(N^4)$ for NxNxN grid, rather than $O(N^6)$
- Complexity of Wigner-D^{-1} independent of #fields/molecule
- Complexity independent of #atoms

Practical complexity (times in seconds):

Grid Resolution (voxels)	Max Error (degrees)	Per Field			Per Pair		
		Spherical Grid	Spherical Harmonics	Per Field Total	Cross-Multiply	Wigner-D^{-1} Transform	Per Pair Total
32x32x32	5.5	0.02	0.01	0.03	0.01	0.02	0.03
64x64x64	2.8	0.18	0.10	0.28	0.19	0.28	0.47
128x128x128	1.4	2.47	1.17	3.64	5.55	4.45	10.00

Outline

Introduction

Binding site modeling with grids

- Simulation-based
- Knowledge-based

Binding site matching with grid correlation

- Fast rotational matching

Searching a database with grid signatures ←

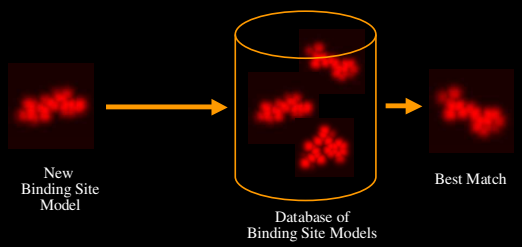
- Power spectrum signature

Results

Discussion

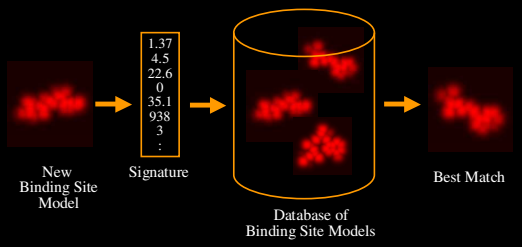
Searching a Database

One-by-one matching is slow for large database



Searching a Database

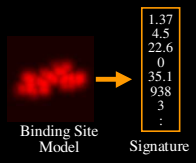
Compute signature for binding site and search for matches to signatures in database



Searching a Database

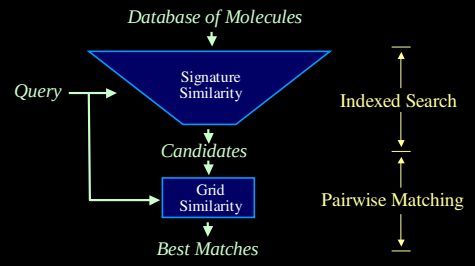
Desirable properties for molecular signatures:

- Concise
- Quick to compute
- Indexible
- Insensitive to noise
- Invariant to transformations
- Discriminating
- Conservative



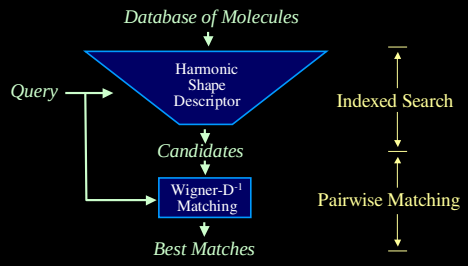
Searching a Database

Use quick, conservative cull to find possible matches and perform detailed match only for the best candidates



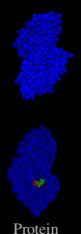
Searching a Database

Use quick, conservative cull to find possible matches and perform detailed match only for the best candidates



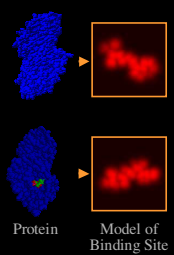
Power Spectrum Signature

Given two proteins …



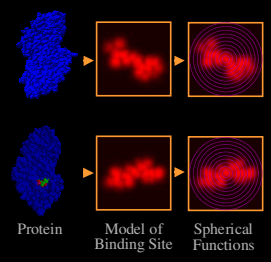
Power Spectrum Signature

1. Build models of binding sites



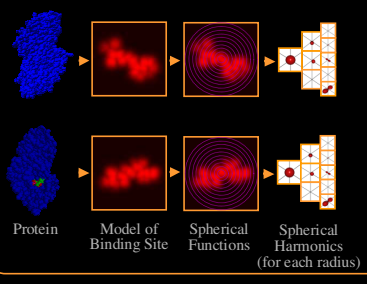
Power Spectrum Signature

2. Build spherical functions for concentric shells at different radii



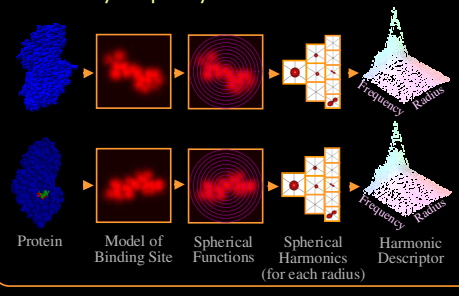
Power Spectrum Signature

3. Decompose the spherical function at each radius into spherical harmonics



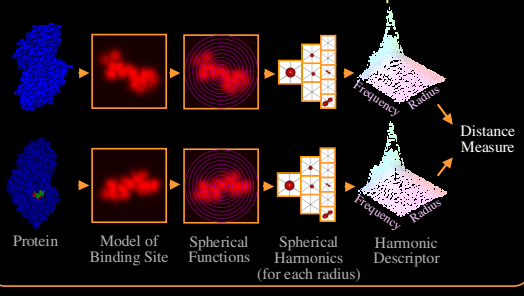
Power Spectrum Signature

4. Store amplitude of spherical harmonic coefficients for every frequency and radius in harmonic descriptor



Power Spectrum Signature

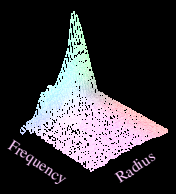
5. Define distance (dissimilarity) between binding sites as L^2 distance between harmonic descriptors



Power Spectrum Signature

Main properties:

- Fast, concise, robust, indexible, etc.
- Invariant to grid orientations
- Works for multiple grids per binding site
- Conservative approximation



Outline

Introduction

Binding site modeling with grids

- Simulation-based
- Knowledge-based

Binding site matching with grid correlation

- Fast rotational matching

Searching a database with grid signatures

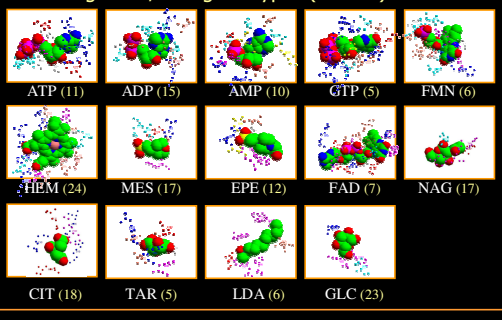
- Power spectrum signature

Results ←

Discussion

Test Data Set

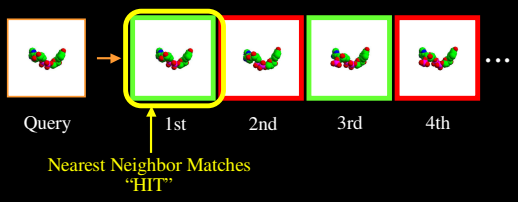
176 binding sites / 14 ligand types (classes)



Classification Experiment

"Leave-one-out" classification experiment

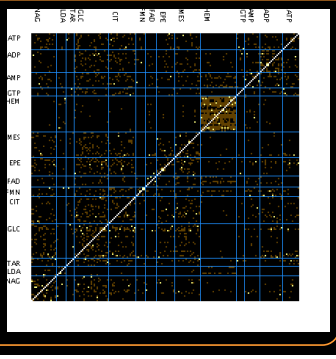
- Match every ligand against all the others in data set
- Log a "hit" when best match performs same reaction
- Report percentage of hits (correctly classified ligands)



Classification Results

Classification rate:

SiteAtoms	= 35%
SCOP	= 17%
CATH	= 13%
CE	= 11%
FASTA	= 10%
Random	= 5%

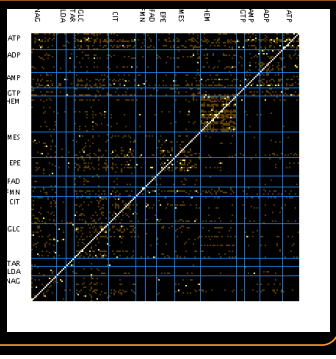


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Classification Results

Classification rate:

SiteAtoms	= 35%
SiteGrid	= 32%
SCOP	= 17%
CATH	= 13%
CE	= 11%
FASTA	= 10%
Random	= 5%

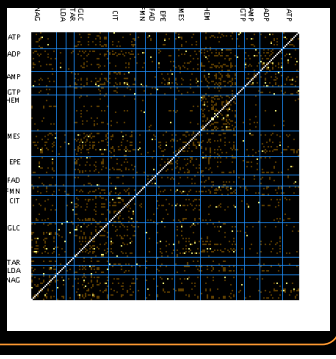


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Classification Results

Classification rate:

SiteGrid	= 32%
SiteSig	= 20%

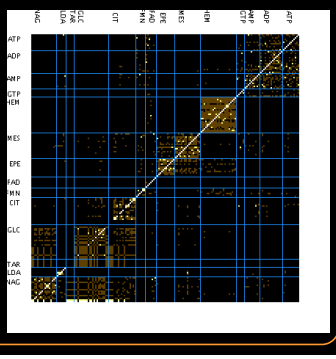


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Classification Results

Classification rate:

ClippedGrid	= 65%
ClippedSig	= 71%
SiteGrid	= 32%
SiteSig	= 20%

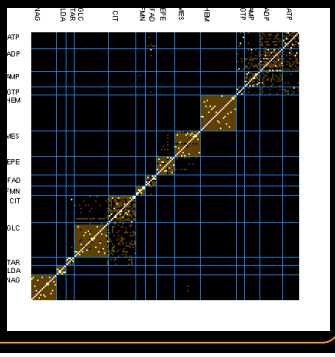


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Classification Results

Classification rate:

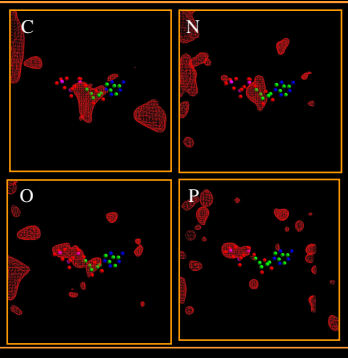
LigandGrid	= 94%
LigandSig	= 89%
ClippedGrid	= 65%
ClippedSig	= 71%
SiteGrid	= 32%
SiteSig	= 20%



Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Classification Results

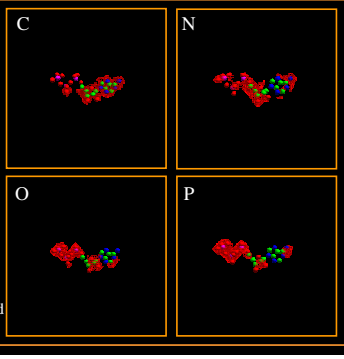
SiteGrid



Raw Knowledge-Based Distributions Predicted for Box around 1kp8-1-H-ATP-1-_

Classification Results

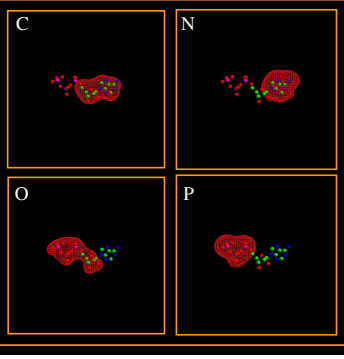
ClippedGrid



Normalized Knowledge-Based Distributions Predicted for Clipped Volume of Ligand 1kp8-1-H-ATP-1-_

Classification Results

LigandGrid



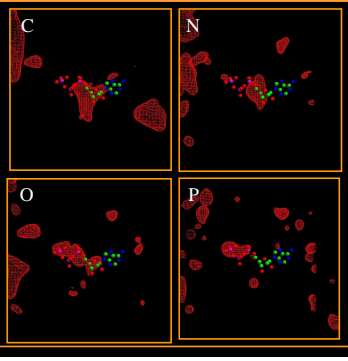
Raw Distributions Observed for Ligand 1kp8-1-H-ATP-1-_

Conclusion

Grid-based representations of binding sites is interesting, but needs work …

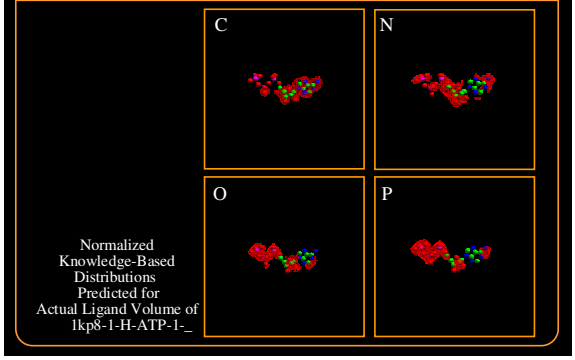
- Grid matching algorithms work pretty well when given a good model and the right center
- Bottlenecks right now are mainly segmentation and modeling

Conclusion



Raw Knowledge-Based Distributions Predicted for Box around 1kp8-1-H-ATP-1-_

Conclusion



Discussion

?