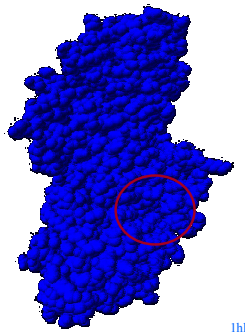
# Detection of Protein Binding Sites II

Thomas Funkhouser
Princeton University
CS597A, Fall 2007

## Introduction

Goal:
- Given a protein structure, predict where a ligand might bind



1hld

## Outline

Introduction

Binding site detection methods
- Geometric, chemical, evolutionary properties, etc.
- Machine learning
- Optimization
- Docking

Evaluation methods

Discussion

## Outline

Introduction

Binding site detection methods
- ➢Geometric, chemical, evolutionary properties, etc.
- Machine learning
- Optimization
- Docking

Previous Lectures

Evaluation methods

Discussion

## Site Properties

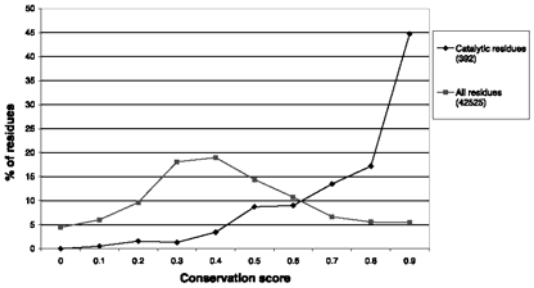
Learned distributions of properties:

| Surface cavity property | Category | Drug-binding cavities | Non drug-binding cavities |
|---|---|---|---|
| Cavity rank[a] | Size | $1.89 \pm 2.07$ | $8.88 \pm 5.4$ |
| Number of residues[b] | Size | $22.8 \pm 14.3$ | $7.31 \pm 5.4$ |
| Number of atoms[c] | Size | $85.0 \pm 62.4$ | $18.7 \pm 21.2$ |
| Smallest moment of inertia[d] | Size/shape | $1.7 \times 10^3 \pm 2.5 \times 10^4$ | $1.2 \times 10^3 \pm 8.3 \times 10^3$ |
| Depth standard deviation[e] | Size/shape | $2.3 \pm 1.1$ (Å) | $0.75 \pm 0.45$ |
| Maximum depth[f] | Size/shape | $10.5 \pm 4.0$ (Å) | $4.75 \pm 1.67$ |
| Average depth[g] | Size/shape | $5.3 \pm 1.9$ (Å) | $3.2 \pm 0.7$ |
| Normalized smallest moment of inertia[h] | Shape | $17.0 \pm 11.7$ | $3.9 \pm 5.3$ |
| Proportion of cavity at depth between [6.5, 6.75][i] | Shape | $0.02 \pm 0.013$ | $0.003 \pm 0.001$ |
| Largest moment of inertia[j] | Size/shape | $1.6 \times 10^5 \pm 8.4 \times 10^4$ | $2.8 \times 10^3 \pm 1.6 \times 10^4$ |
| Average side-chains residual entropy[k] | Rigidity | $-0.41 \pm 0.18$ (kcal) | $-0.55 \pm 0.25$ |
| Average curvature[l] | Shape | $-49.0 \pm 8.3$ | $-57.0 \pm 13.1$ |
| Maximum curvedness[m] | Shape | $6.4 \pm 2.9$ | $4.9 \pm 4.9$ |
| Maximum mean curvature[n] | Shape | $5.3 \pm 2.6$ | $3.5 \pm 4.2$ |
| Curvedness < 0.5[o] | Shape | $0.35 \pm 0.04$ | $0.29 \pm 0.08$ |
| Proportion of proline[p] | Amino acid composition | $0.019 \pm 0.028$ | $0.04 \pm 0.09$ |
| Proportion of cavity with logP between [−1, 0][q] | Hydrophobicity | $0.09 \pm 0.07$ | $0.15 \pm 0.16$ |
| Side-chain residual entropy standard deviation[r] | Rigidity | $0.43 \pm 0.18$ (kcal) | $0.55 \pm 0.17$ |

[Nayal06]

## Site Properties

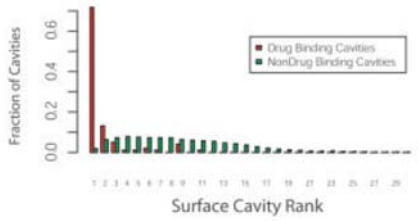
Example: conservation




[Bartlett02]

## Site Properties

Example: cavity rank




[Nayal06]

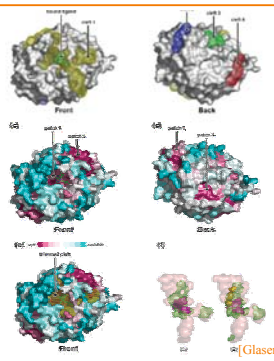
## Combining Multiple Properties

Example:

- Conservation + cavity size

Method:

- Remove portions of cavity (predicted by Surfnet) more than X angstroms from the closest residue with conservation above Y

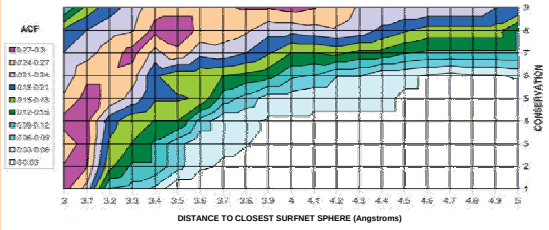


[Glaser06]

## Combining Multiple Properties

Example:

- Conservation + cavity size

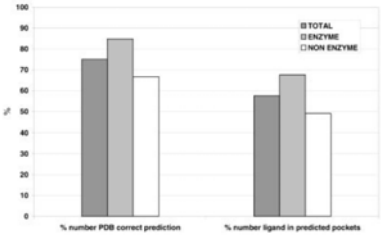



[Glaser06]

## Combining Multiple Properties

Example:

- Conservation + cavity size



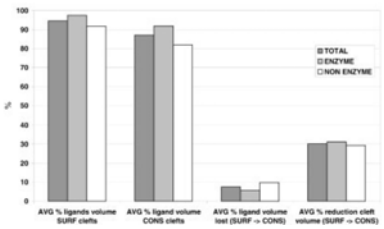

[Glaser06]

## Combining Multiple Properties

Example:

- Conservation + cavity size

"AVG % ligands volume SURF clefts" is the average percentage of ligands volume included in the four biggest clefts produced by the SURFNET program. "AVG % ligand volume CONS clefts" is the average percentage of ligands volume included in the four biggest trimmed clefts. "AVG % ligand volume lost (SURF 3 CONS)" is the average ligand volume lost during the trimming procedure. "AVG % reduction cleft volume (SURF 3 CONS)" is the average volume cleft reduction of clefts including a ligand during the trimming procedure.




[Glaser06]

## Outline

Introduction

Binding site detection methods

- Geometric, chemical, evolutionary properties, etc.
  - ➢Machine learning
- Optimization
- DockingEvaluation methods

Discussion

## Machine Learning

Build classifier to recognize functional residues/sites from multiple properties:

- Depth
- Solvent accessibility
- Propensity
- Conservation
- Hydrophobicity
- Secondary structure type
- Pocket size
- Amino acid
- etc.

## Machine Learning

Example classifiers:

- Naïve Bayes
- Decision trees
- Neural nets
- Support vector machines
- etc.

## Machine Learning [Gutteridge03]

Data type

- Protein residues

Properties

- Many (next slide)

Training set

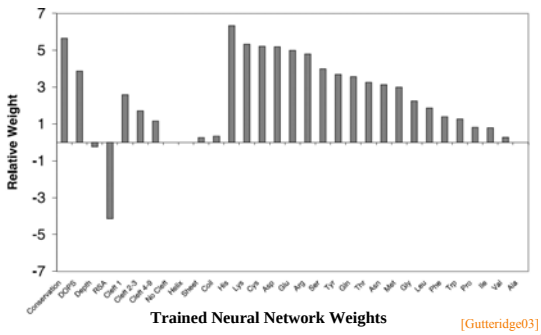
- 159 crystallized proteins
- 55,000 non-catalytic residues, 550 catalytic residues

Classification methods

- Neural network
- Spatial clustering

[Gutteridge03]

## Machine Learning [Gutteridge03]

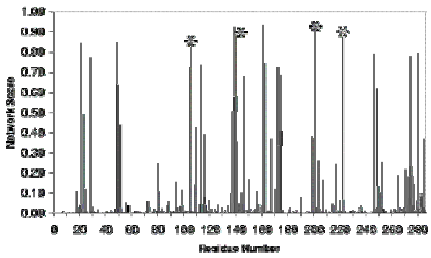



Trained Neural Network Weights

[Gutteridge03]

## Machine Learning [Gutteridge03]

Neural network classifier

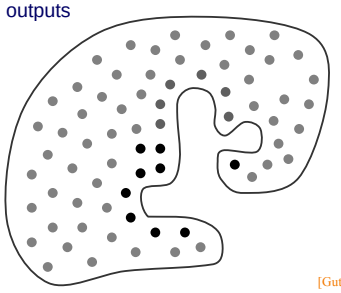



[Gutteridge03]

## Machine Learning [Gutteridge03]

Spatial clustering

- Compute spheres around clusters of nearby residues with high NN outputs

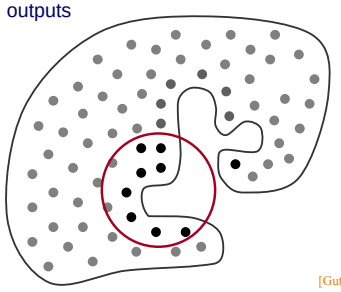


[Gutteridge03]

## Machine Learning [Gutteridge03]

Spatial clustering

- Computer spheres around clusters of nearby residues with high NN outputs

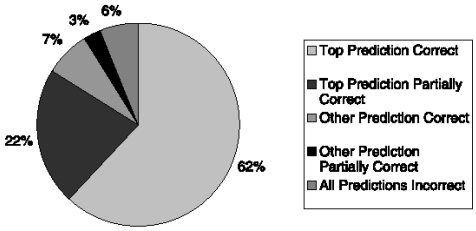


[Gutteridge03]

## Machine Learning [Gutteridge03]

Evaluation

- Sphere overlaps known actives site by > 50%




[Gutteridge03]

## Machine Learning (FEATURE)

Data type

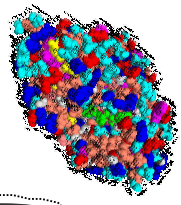
- Points in/around protein

Properties

- Many properties logged in histograms for shells around point (on right)

Classification methods

- Naïve Bayes



[Bagley95]

## Machine Learning (FEATURE)

Data type

- Points in/around protein

Properties

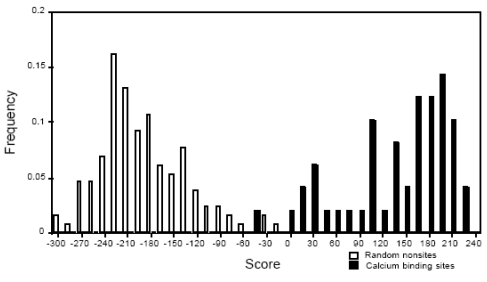
- Many properties logged in histograms for shells around point (on right)

Classification methods

- Naïve Bayes

[Bagley95]

## Machine Learning (FEATURE)




[Bagley95,Wei98]

## Machine Learning [Nayal06]

Data type

- Cavity surfaces

Properties

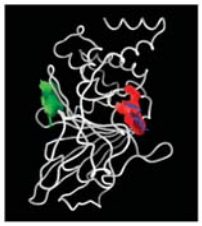
- Many (next slide)

Training set

- 1347 cavities
- 99 non-redundant proteins

Classification methods

- Random forests



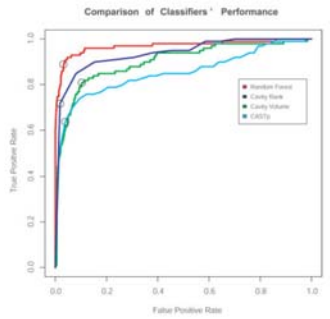
[Nayal06]

## Machine Learning [Nayal06]

| Surface cavity property | Category | Drug-binding cavities | Non drug-binding cavities |
|---|---|---|---|
| Cavity rank[a] | Size | 1.89 ± 2.07 | 6.88 ± 5.4 |
| Number of residues[b] | Size | 22.8 ± 14.3 | 7.31 ± 5.4 |
| Number of atoms[c] | Size | 85.0 ± 62.4 | 18.7 ± 21.2 |
| Smallest moment of inertia[d] | Size/shape | $1.7 \times 10^4 \pm 2.5 \times 10^4$ | $1.2 \times 10^3 \pm 8.3 \times 10^3$ |
| Depth standard deviation[e] | Size/shape | 2.3 ± 1.1 (Å) | 0.75 ± 0.45 |
| Maximum depth[f] | Size/shape | 10.5 ± 4.0 (Å) | 4.75 ± 1.67 |
| Average depth[g] | Size/shape | 5.3 ± 1.9 (Å) | 3.2 ± 0.7 |
| Normalized smallest moment of inertia[h] | Shape | 17.0 ± 11.7 | 3.9 ± 5.3 |
| Proportion of cavity at depth between [6.5, 6.75][i] | Shape | 0.02 ± 0.043 | 0.003 ± 0.001 |
| Largest moment of inertia[j] | Size/shape | $1.6 \times 10^4 \pm 8.4 \times 10^4$ | $2.8 \times 10^3 \pm 1.6 \times 10^4$ |
| Average side-chains residual entropy[k] | Rigidity | −0.41 ± 0.18 (kcal) | −0.55 ± 0.25 |
| Average curvature[l] | Shape | −49.0 ± 8.3 | −57.0 ± 13.1 |
| Maximum curvedness[m] | Shape | 6.4 ± 2.9 | 4.0 ± 4.9 |
| Maximum mean curvature[n] | Shape | 5.3 ± 2.6 | 3.5 ± 4.2 |
| Curvedness < 0.5[o] | Shape | 0.35 ± 0.04 | 0.29 ± 0.08 |
| Proportion of proline[p] | Amino acid composition | 0.019 ± 0.028 | 0.04 ± 0.09 |
| Proportion of cavity with logP between [−1, 0][q] | Hydrophobicity | 0.09 ± 0.07 | 0.15 ± 0.16 |
| Side-chain residual entropy standard deviation[r] | Rigidity | 0.43 ± 0.18 (kcal) | 0.55 ± 0.17 |

[Nayal06]

## Machine Learning [Nayal06]




[Nayal06]

## Outline

Introduction

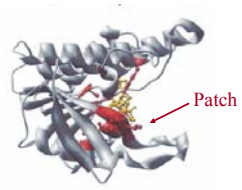
Binding site detection methods
- Geometric, chemical, evolutionary properties, etc.
- Machine learning
- ➢Optimization
- Docking

Evaluation methods

Discussion

## Optimization

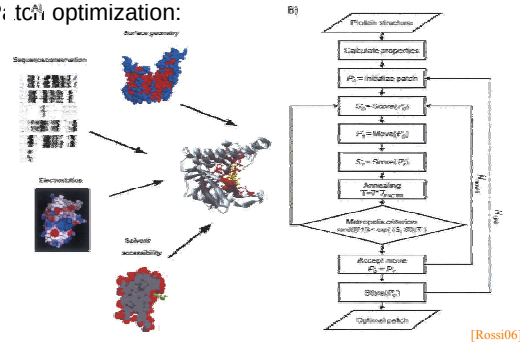
Patch optimization:
- Define patch as set of contiguous residues
- Compute patch properties
- Compute patch score
- If not optimal, grow/shrink patch and iterate




[Rossi06]

## Optimization

Patch optimization:




[Rossi06]

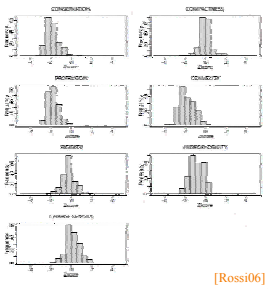
## Site Optimization

Patch properties:
- Z-score in context of training set

$z_p = (value_p - mean_p / \sigma_p$

Patch score:

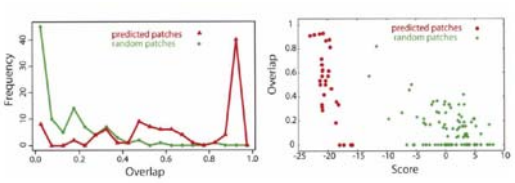
$score = \sum_p w_p z_p$



[Rossi06]

## Optimization

Results:

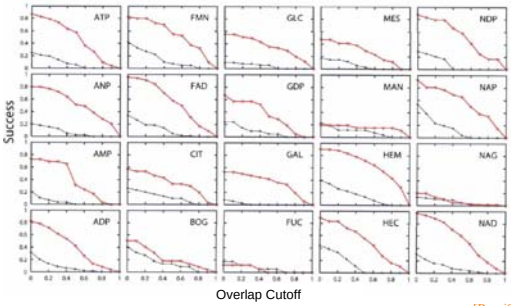



$$\text{Overlap} = \frac{N_1 \cap N_2}{\min(N_1, N_2)}$$

[Rossi06]

## Optimization

Results:




[Rossi06]

## Outline

Introduction

Binding site detection methods

- Geometric, chemical, evolutionary properties, etc.
- Machine learning
- Optimization
- ➢Docking

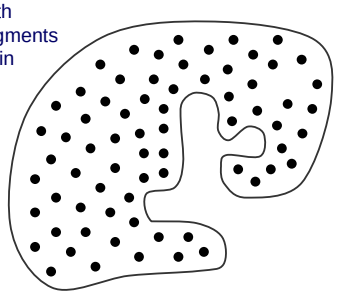
Evaluation methods

Discussion

## Docking for Site Prediction

General idea:

- Compute map with distribution of fragments docked into protein

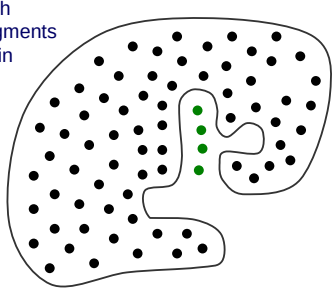


[Miranker91], [Mattos96]

## Docking for Site Prediction

General idea:

- Compute map with distribution of fragments docked into protein

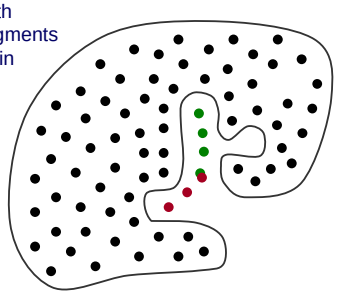


[Miranker91], [Mattos96]

## Docking for Site Prediction

General idea:

- Compute map with distribution of fragments docked into protein

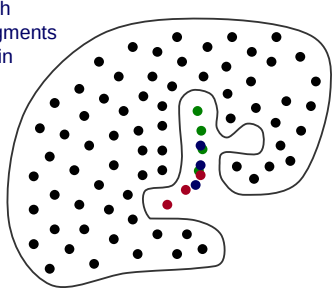


[Miranker91], [Mattos96]

## Docking for Site Prediction

General idea:

- Compute map with distribution of fragments docked into protein

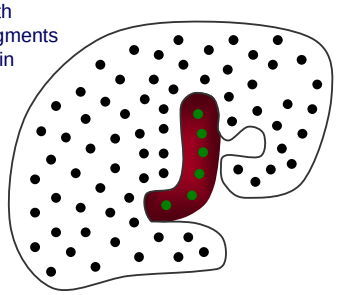


[Miranker91], [Mattos96]

## Docking for Site Prediction

General idea:

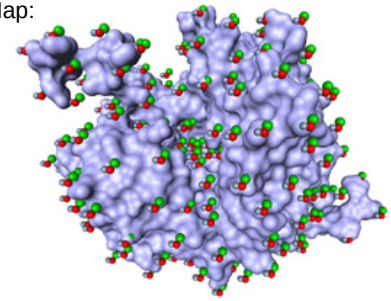
- Compute map with distribution of fragments docked into protein



[Miranker91], [Mattos96]

## Docking for Site Prediction

CS-Map:

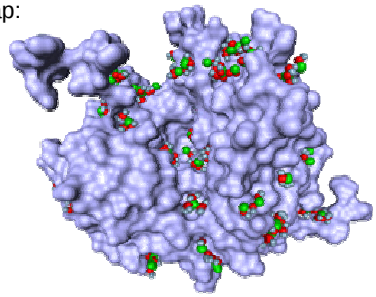


Step 1: Place probes

[Vajda; Silberstein; Landon et al.]

## Docking for Site Prediction

CS-Map:

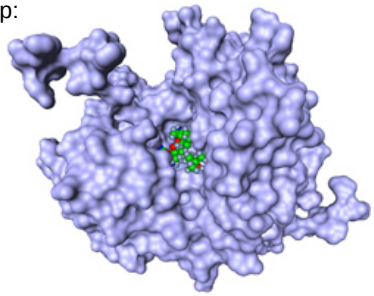


Step 2: Move the probes around to find binding positions

[Vajda; Silberstein; Landon et al.]

## Docking for Site Prediction

CS-Map:

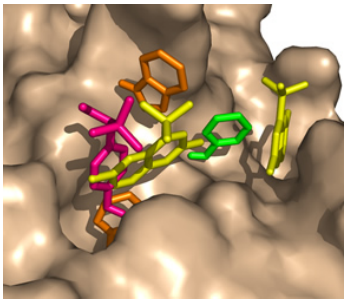


Step 3: Remove high energy clusters of the ligand

[Vajda; Silberstein; Landon et al.]

## Docking for Site Prediction

CS-Map:

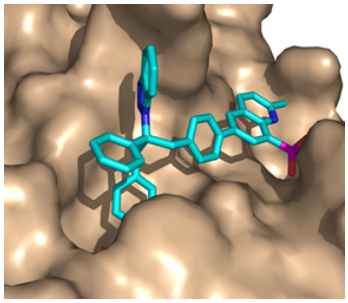


Step 4: Repeat mapping with a number of fragments

[Vajda; Silberstein; Landon et al.]

## Docking for Site Prediction

CS-Map:

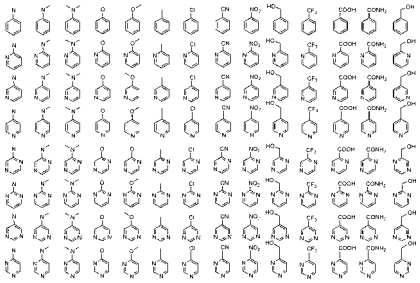


Step 5: Combine fragment into potential ligand molecules

[Vajda; Silberstein; Landon et al.]

## Docking for Site Prediction

CS-Map:

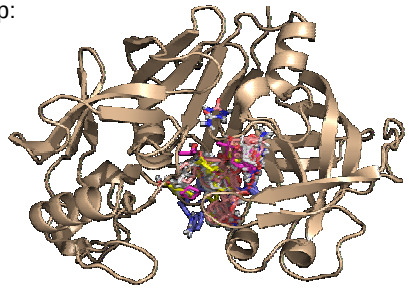


Prototype fragment library

[Vajda; Silberstein; Landon et al.]

## Docking for Site Prediction

CS-Map:

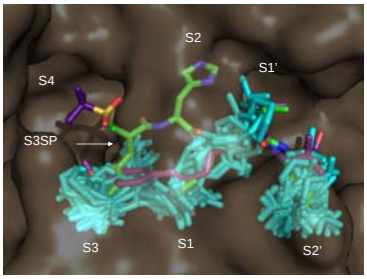



Top two consensus sites for each structure are in binding pocket

[Vajda; Silberstein; Landon et al.]

## Docking for Site Prediction

CS-Map:




Identification of Preferred Binding Modality of Aliskiren using CSMap

[Vajda; Silberstein; Landon et al.]

## Outline

Introduction

Binding site detection methods
- Geometric, chemical, evolutionary properties, etc.
- Machine learning
- Optimization
- Docking

Evaluation methods ←

Discussion

## General Evaluation Method

Gather a set of PDB files
- Both bound and unbound (with homologues)

Predict binding sites (clefts, pockets)
- Output is usually grid, polyhedron, set of spheres

Report results
- Measure properties of predicted binding sites
- Test how well predictions match bound ligands

## PocketFinder

Liganded-pocket data set

- Consider all protein-ligand complexes from PDB
- Eliminate frequent co-factors (HEM, etc.)
- Eliminate ligands far from protein (>3.5Å)
- Eliminate ligands in seams between assymmetric units
- Eliminate “duplicates” (?)
- 50 < protein residues < 2000
- 6 < ligand atoms
- 2.5Å < resolution

5,616 bound binding sites

[An04]

## PocketFinder

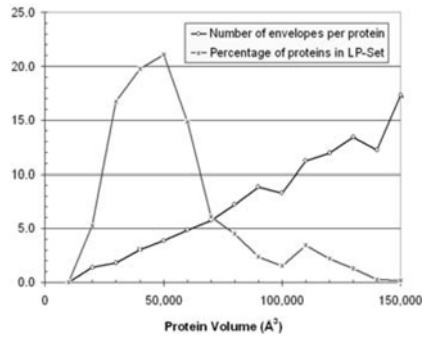
Unliganded-pocket data set

- Align unliganded PDB files with liganded ones
- Single chain proteins
- 95% < sequence identity
- No mutations on surface within 8Å of ligand
- No other ligands within 8Å of ligand
- 2.5Å < resolution

11,510 unbound binding sites

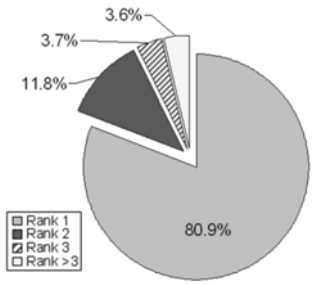
[An04]

## PocketFinder




[An04]

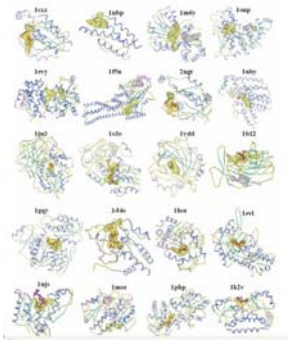
## PocketFinder




Rank of the real binding sites in the predicted putative binding site lists.

[An04]

## PocketFinder



Two largest predicted envelopes (1st:yellow, 2nd:gray)

[An04]

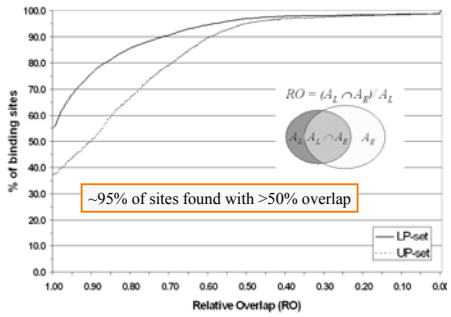
## PocketFinder

Accuracy measured by overlap of protein atoms in contact with ligand and protein atoms in contact with predicted envelope.

$$RO = (A_L \cap A_E) / A_L$$

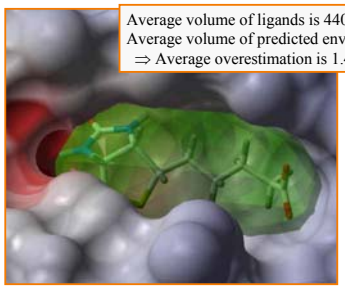
$A_L$ = solvent accessible surface area of protein atoms within 3.5Å of bound ligand
$A_E$ = solvent accessible surface area of protein atoms within 3.5Å of predicted envelope

## PocketFinder




[An04]

## PocketFinder

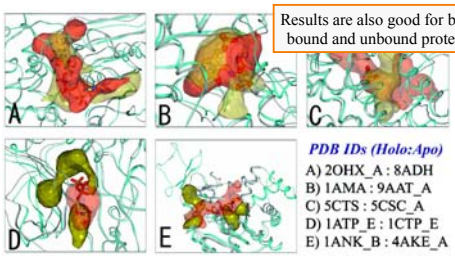


Average volume of ligands is 440Å
Average volume of predicted envelopes is 611Å
⇒ Average overestimation is 1.4x (?)

Biotin-streptavidin binding site predicted with PocketFinder

[An04]

## PocketFinder




Results are also good for both bound and unbound proteins

*PDB IDs (Holo:Apo)*
A) 2OHX_A : 8ADH
B) 1AMA : 9AAT_A
C) 5CTS : 5CSC_A
D) 1ATP_E : 1CTP_E
E) 1ANK_B : 4AKE_A

Effect of conformational changes on predictions for bound (holo) and unbound (apo) proteins
*(enzymes: gray-holo, green-apo)*
*(envelopes: red-holo, yellow-apo)*

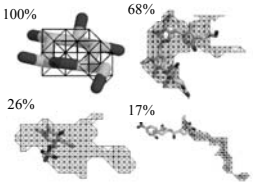
[An04]

## Q-SiteFinder

Test set:
- 134 bound proteins (GOLD test set)
- 35 unbound proteins (homologues to bound proteins)

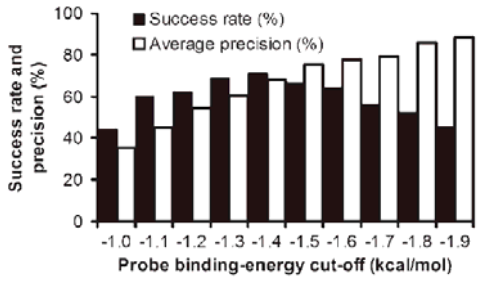
Metric:
- Precision = % predicted site within 1.6Å of ligand

Success:
- Precision >25%



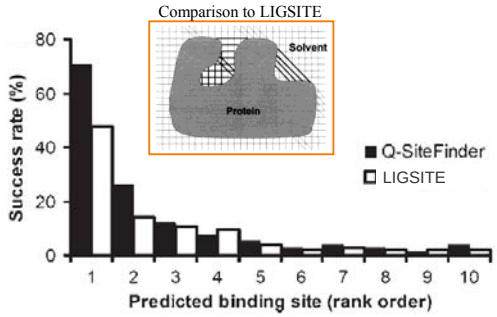

[Laurie05]

## Q-SiteFinder




[Laurie05]

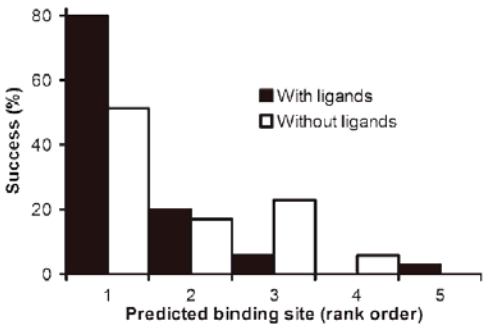
## Q-SiteFinder




Q-SiteFinder cutoff = -1.4 kcal/mol, LIGSITE threshold = 5

[Laurie05]

## Q-SiteFinder




[Laurie05]

## References


[An04] J. An, M. Totrov, R. Abagyan, "Comprehensive Identification of ``Druggable" Protein Ligand Binding Sites," Genome Informatics, 15, 2, 2004, pp. 31-41.

[Bartlett02] G.J. Bartlett, C.T. Porter, N.Borkakoti, J.M. Thornton, "Analysis of catalytic residues in enzyme active sites," J. Mol. Biol, 324, 1, 2002, pp. 105-121.

[Bate04] P. Bate, J. Warwicker, "Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods," J Mol Biol, 340, 2, 2004, pp. 263-276.

[Campbell03] S.J. Campbell, N.D. Gold, R.M. Jackson, D.R. Westhead, "Ligand binding functional site location, similarity and docking," Curr Opin Struct Biol, 13, 2003, pp. 389-395.

[Elcock01] A.H. Elcock, "Prediction of functionally important residues based solely on the computed energetics of protein structure," J. Mol. Biol., 312, 4, 2001, pp. 885-896.

[Gutteridge03] A. Gutteridge, G.J. Bartlett, J.M. Thornton, "Using a neural network and spatial clustering to predict the location of active sites in enzymes," J Mol Biol, 330, 2003, pp. 719-734.

[Laurie05] A.T.R. Laurie, R.M. Jackson, "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites," Bioinformatics, 2005.

[Nimrod05] G. Nimrod, F. Glaser, D. Steinberg, N. Ben-Tal, T. Pupko, "In silico identification of functional regions in proteins," Bioinformatics, 21 Suppl., 2005, pp. i328-i337.

[Silberstein03] Michael Silberstein, Sheldon Dennis, Lawrence Brown III, Tamas Kortvelyesi, Karl Clodfelter, Sandor Vajda, "Identification of Substrate Binding Sites in Enzymes by Computational Solvent Mapping," J. Mol. Biol., 332, 2003, pp. 1095-1113.

[Young94] L. Young, R.L. Jernigan, D.G. Covell, "A role for surface hydrophobicity in protein-protein recognition," Protein Sci, 3, 5, 1994, pp. 717-29.