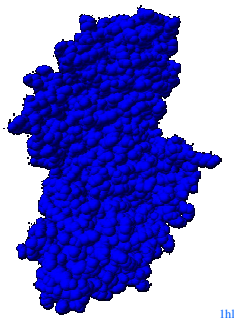
Representation and Matching of Ligand Binding Sites I

Thomas Funkhouser
Princeton University
CS597A, Fall 2007

Introduction

Goal:
- Given a protein structure, predict its ligand bindings

Applications:
- Function prediction
- Drug discovery
- etc.

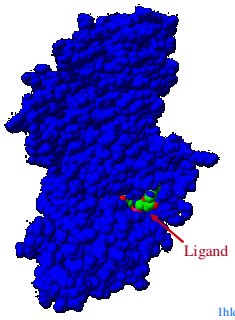


1hld

Introduction

Questions:
- Where will the ligand bind?

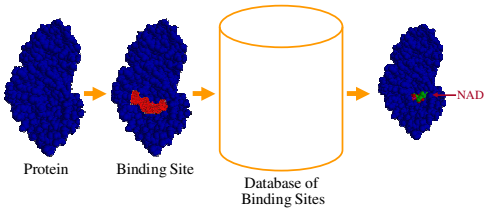
ØWhich ligand will bind?
- How will the ligand bind?
- When?
- Why?
- etc.



1hld

Which Ligand Will Bind?

Possible matching strategies

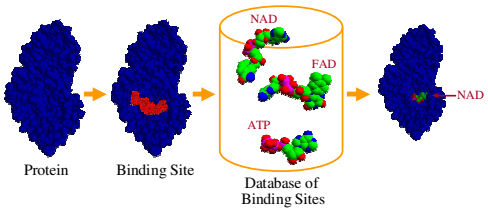


Which Ligand Will Bind?

Possible matching strategies

ØBinding site → Ligands

Protein-Ligand Docking
(after fall break)

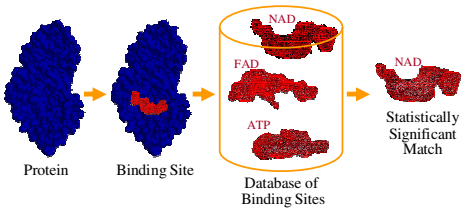


Which Ligand Will Bind?

Possible matching strategies
- Binding site → Ligands

ØBinding site → Binding sites

Binding Site Matching
(next few lectures)

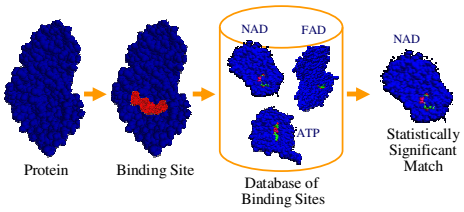


Which Ligand Will Bind?

Possible matching strategies

- Binding site → Ligands
- Binding site → Binding sites
- ØBinding site → Proteins

Binding Site Search
(in two lectures)



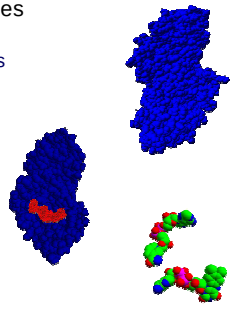
Which Ligand Will Bind?

Possible matching strategies

- Binding site → Ligands
- Binding site → Binding sites
- Binding site → Proteins

- Protein → Ligands
- Protein → Binding sites
- Protein → Proteins

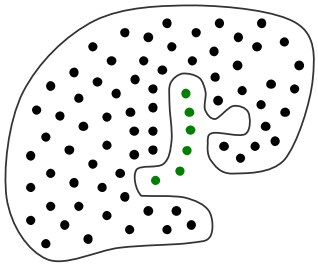
- Ligand → Ligands
- Ligand → Binding sites
- Ligand → Proteins



Binding Site Representation

Possible descriptions:

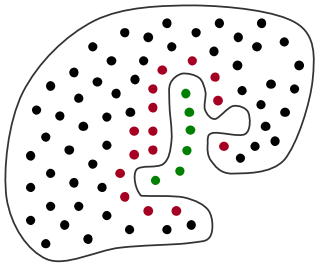
- Point set
- Surface
- Volume



Binding Site Representation

Possible descriptions:

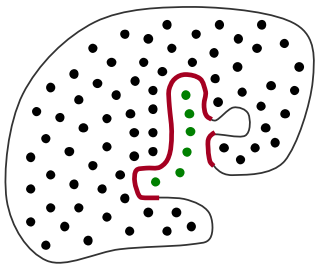
- ØPoint set
- Surface
- Volume



Binding Site Representation

Possible descriptions:

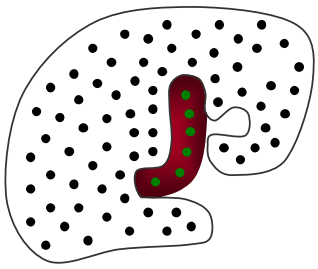
- Point set
- ØSurface
- Volume



Binding Site Representation

Possible descriptions:

- Point set
- Surface
- ØVolume



Outline

Introduction

Point set representations

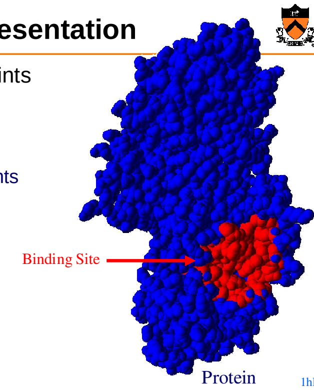
Point set matching
- Brute force search
- RANSAC
- Geometric hashing
- Association graphs
- Iterative closest point

Evaluation

Discussion

Point Set Representation

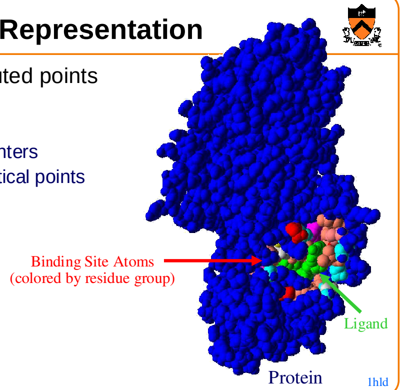
Set of attributed points
- Atoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.



Point Set Representation

Set of attributed points

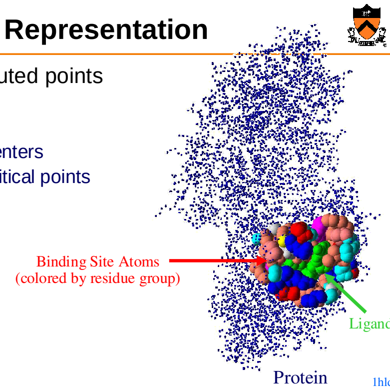
ØAtoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.



Point Set Representation

Set of attributed points

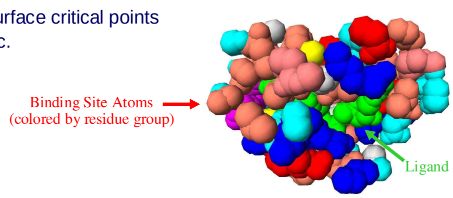
ØAtoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.



Point Set Representation

Set of attributed points

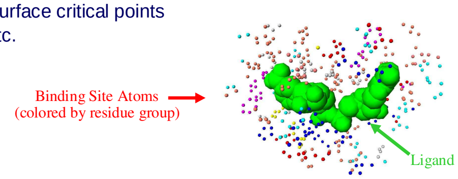
ØAtoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.



Point Set Representation

Set of attributed points

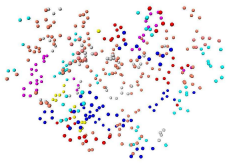
ØAtoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.



Point Set Representation

Set of attributed points
- Atoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.

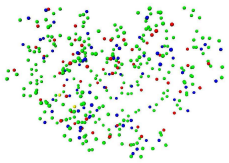
Binding Site Atoms (colored by residue group)



Point Set Representation

Set of attributed points
- Atoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.

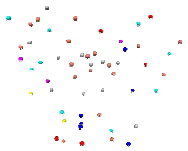
Binding Site Atoms (colored by element type)



Point Set Representation

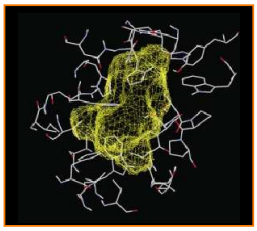
Set of attributed points
- Atoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.

Binding Site Residues (colored by residue group)



Point Set Representation

Set of attributed points
- Atoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.

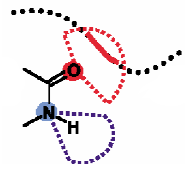


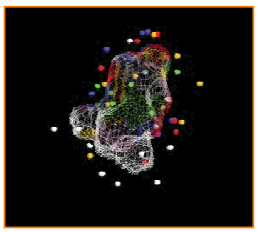
Residues Surrounding Binding Site

[Schmitt02]

Point Set Representation

Set of attributed points
- Atoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.



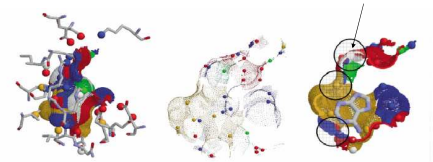


Residues Surrounding Binding Site

[Schmitt02]

Point Set Representation

Set of attributed points
- Atoms
- Residues
- Pseudo-centers
- Surface critical points
- etc.



Represent Chemical and Geometric Properties of Surface

[Shulman-Peleg04]

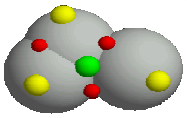
Point Set Representation

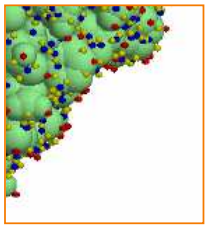
Set of attributed points

- Atoms
- Residues
- Pseudo-centers

ØSurface critical points

- etc.





Critical Points on Surface of Binding Site

[Lin94] [Wolfson]

Outline

Introduction

Point set representations

Point set matching ←

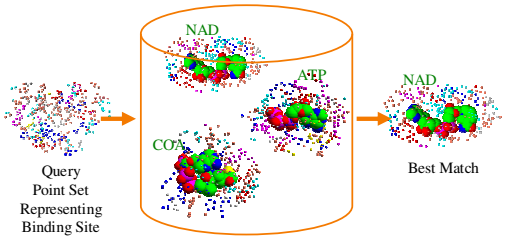
- Brute force search
- RANSAC
- Geometric hashing
- Association graphs
- Iterative closest point

Evaluation

Discussion

Point Set Matching

Detecting similarities in point sets may reveal functional similarities



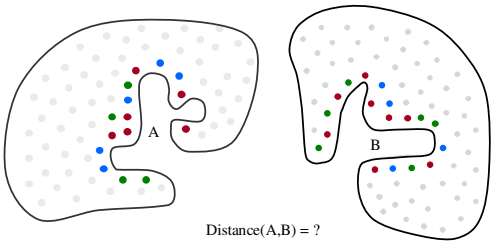
Query Point Set Representing Binding Site

Best Match

Database of Point Sets Representing Binding Sites

Point Set Matching

Goal is to compute a similarity measure for a pair of attributed point sets

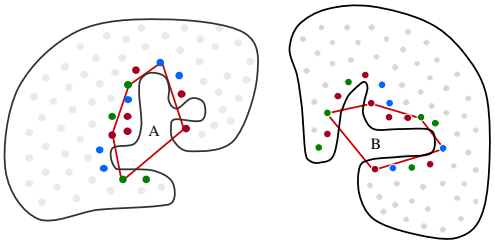


Distance(A,B) = ?

Point Set Matching

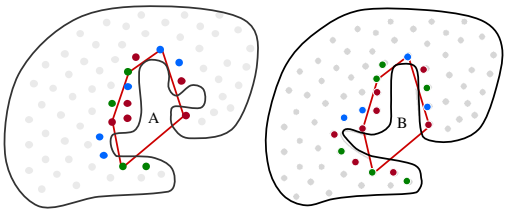
Challenge is to find corresponding points

- "Subset of points in A" may match "subset of points in B"



Point Set Matching

Calculating a superposition and distance measure is easy if correspondences are known (proposed)

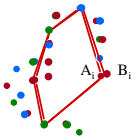


Point Set Matching

Calculating a superposition and distance measure is easy if correspondences are known (proposed)

Point Set Matching

Calculating a superposition and distance measure is easy if correspondences are known (proposed)

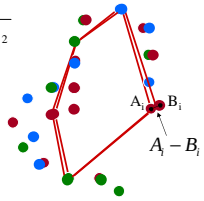


Least-squares optimal superposition of corresponding points

Point Set Matching

Calculating a superposition and distance measure is easy if correspondences are known (proposed)

$$RMSD(A,B) = \sqrt{\sum_{i=1}^{N} (A_i - B_i)^2}$$



Distance(A,B) = RMSD(A,B) + OtherTerms …

Outline

Introduction

Point set representations

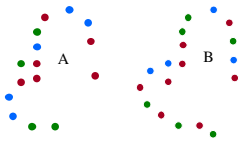
Point set matching
- ØBrute force search
- RANSAC
- Geometric hashing
- Association graphs
- Iterative closest point

Evaluation

Discussion

Brute Force Search

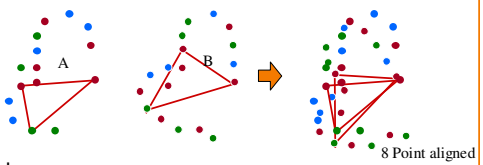
Simple method:
- Try all possible sets of point correspondences
- Score the alignment for each one



Problem:
- $O(n^m)$ possible sets of m correspondences among n points

Brute Force Search

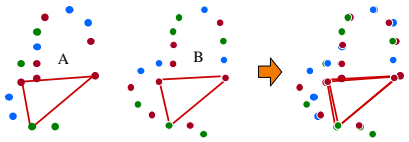
Simple method:
- Try all possible sets of point correspondences
- Score the alignment for each one



Problem:
- $O(n^m)$ possible sets of m correspondences among n points

Brute Force Search

Simple method:
- Try all possible sets of point correspondences
- Score the alignment for each one (e.g., RMSD)



All points aligned
RMSD = 0.2

Problem:
- $O(n^m)$ possible sets of m correspondences among n points

Outline

Introduction

Point set representations

Point set matching
- Brute force search

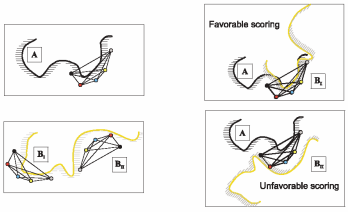
ØRANSAC
- Geometric hashing
- Assocation graphs
- Iterative closest point

Evaluation

Discussion

RANSAC

Randomly sample set of possible correspondences
- Randomly generate a small set of point correspondences
- Compute the aligning transformation for correspondences
  Score how well other points align after that transformation



[Schmitt02]

Outline

Introduction

Point set representations

Point set matching
- Brute force search
- RANSAC

ØGeometric hashing
- Assocation graphs
- Iterative closest point

Evaluation

Discussion

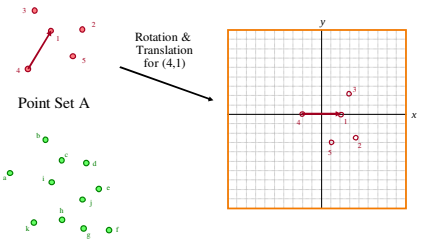
Geometric Hashing

Discretize transformations and scoring

[Wolfson97]

Geometric Hashing

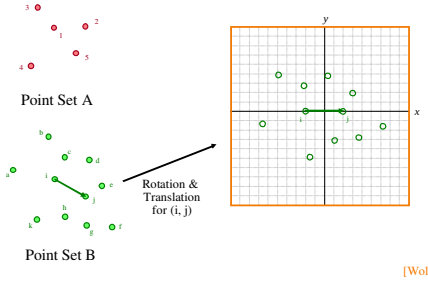
Discretize transformations and scoring



[Wolfson97]

Geometric Hashing

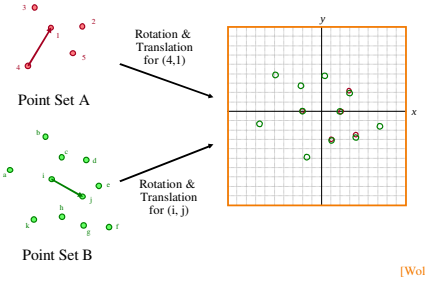
Discretize transformations and scoring



[Wolfson97]

Geometric Hashing

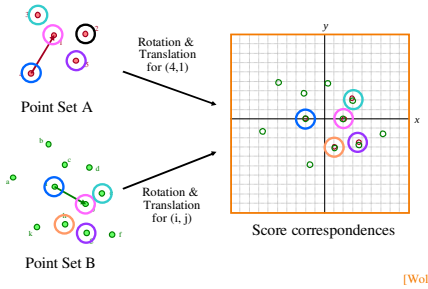
Discretize transformations and scoring



[Wolfson97]

Geometric Hashing

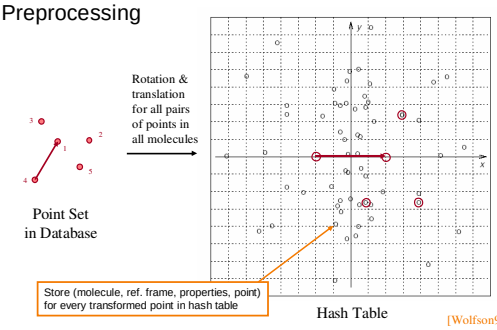
Discretize transformations and scoring



[Wolfson97]

Geometric Hashing

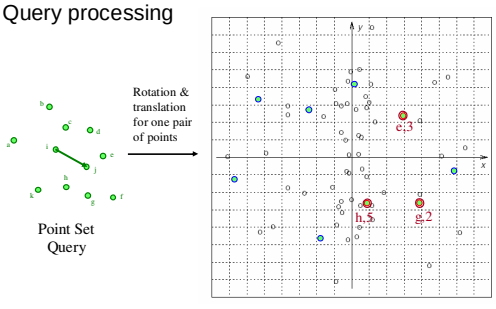
Preprocessing



[Wolfson97]

Geometric Hashing

Query processing

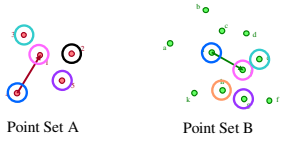


[Wolfson97]

Geometric Hashing

Further processing

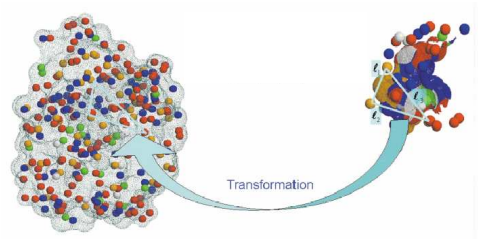
- Refine alignment based on computed correspondences



[Wolfson97]

Geometric Hashing

Create transformations for triples of points in 3D



[Shulman-Peleg04]

Geometric Hashing

Preprocessing
- For each triple of points
  - Compute reference frame
  - For each point
    - Transform point into reference frame
    - Hash (molecule, ref. frame, properties, point)

Query processing
- Choose any triple of points
- Compute reference frame
- For each point
  - Transform point into reference frame
  - For each entry in hash bin for transformed point
    - Check point properties
    - Vote for (molecule, ref. frame)

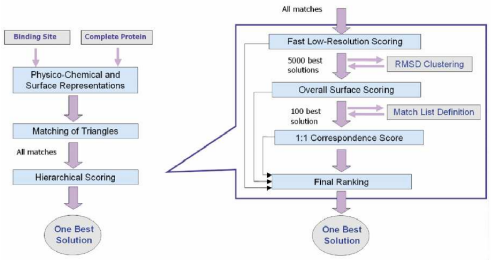
Geometric Hashing

Preprocessing complexity
- $O(n^4)$ for n points per binding site
  - § $O(n^3)$ possible triples * O(n) transformations per triple

Query complexity
- O(m) * binsize for m points in query binding site
  - § 1 triple * O(m) transformations per triple * binsize hash processing per transformation

[Wolfson97]

Shulman-Peleg et al. 2004



[Shulman-Peleg04]

Outline

Introduction

Point set representations

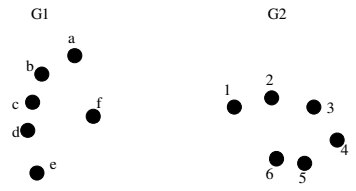
Point set matching
- Brute force search
- RANSAC
- Geometric hashing

ØAssociation graphs
- Iterative closest point

Evaluation

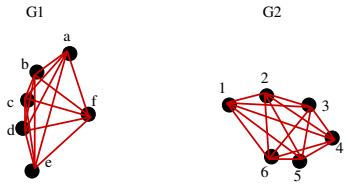
Discussion

Association Graphs



[Schmitt02, Brown82]

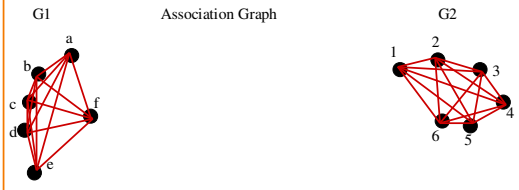
Association Graphs



Represent both points sets as complete graphs (G1 and G2).
(edges connect all pairs of vertices within each point set)

[Schmitt02, Brown82]

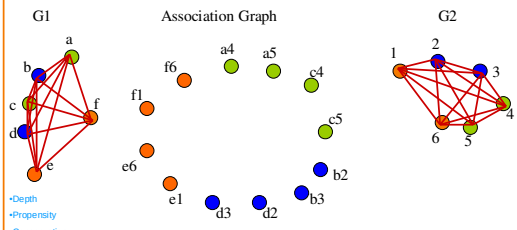
Association Graphs



Create vertices in the association graph for all compatible pairs of vertices in the original graphs.
This can lead to a large number of vertices.

[Schmitt02, Brown82]

Association Graphs

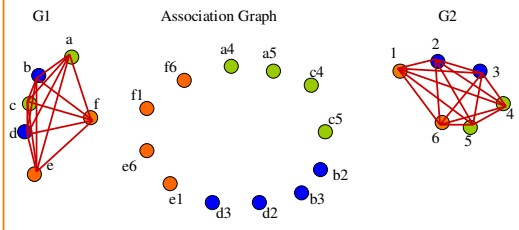


- Depth
- Propensity
- Conservation
- Charge
- Hydrophobicity
- Secondary structure type
- Destabilization

Create vertices in the association graph for all compatible pairs of vertices in the original graphs.
Compatibility could refer to chemical properties.

[Schmitt02, Brown82]

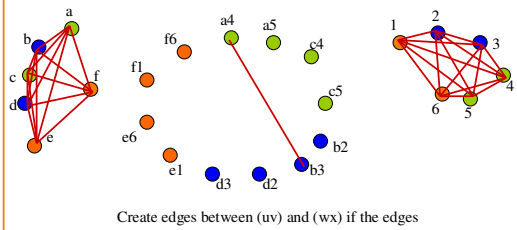
Association Graphs



Create edges between (uv) and (wx) if the edges between (u) and (w) as well as between (v) and (x) match.

[Schmitt02, Brown82]

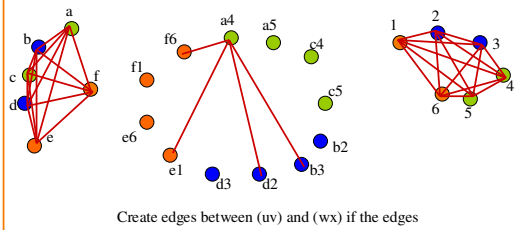
Association Graphs



Create edges between (uv) and (wx) if the edges between (u) and (w) as well as between (v) and (x) match.
For this example, edge length is the only consideration

[Schmitt02, Brown82]

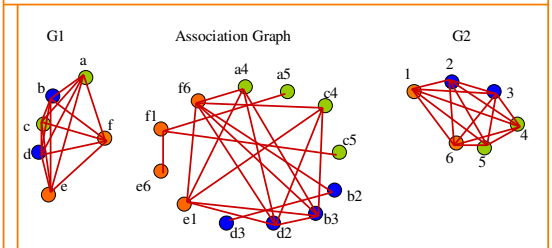
Association Graphs



Create edges between (uv) and (wx) if the edges between (u) and (w) as well as between (v) and (x) match.
For this example, edge length is the only consideration

[Schmitt02, Brown82]

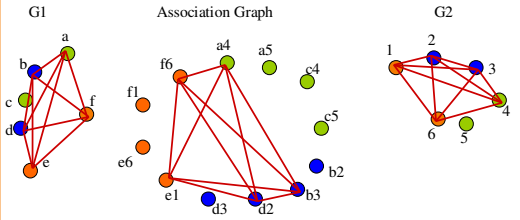
Association Graphs



Create edges between (uv) and (wx) if the edges between (u) and (w) as well as between (v) and (x) match.
For this example, edge length is the only consideration

[Schmitt02, Brown82]

Association Graphs



Finding correspondences: The largest set of corresponding nodes in the same configuration is the maximal clique in the association graph

[Schmitt02, Brown82]

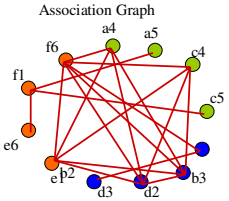
Association Graphs

Computational complexity:

- $O(2^n)$ for n points
- NP-complete
- Branch and bound algorithms

```
Find the Maximal Clique{
    return Cliques(empty, all nodes)
}

Cliques(X, Y){
    if (no node in Y-X is connected to all of X){
        return X;
    }else{
        y = node in Y connected to all of X;
        return Largest(Cliques(X union y, Y},
                       Cliques{X, Y-y});
    }
}
```



[Schmitt02, Brown82]

Outline

Introduction

Point set representations

Point set matching

- Brute force
- RANSAC
- Geometric hashing
- Assocation graphs

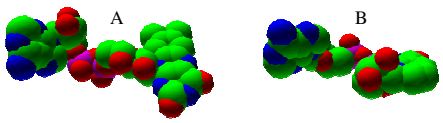
ØIterative closest points

Evaluation

Discussion

Iterative Closest Points

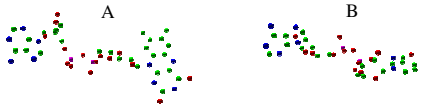
Given two molecules



[Besl92]

Iterative Closest Points

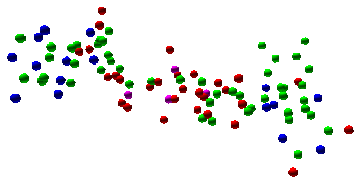
Given two molecules



[Besl92]

Iterative Closest Points

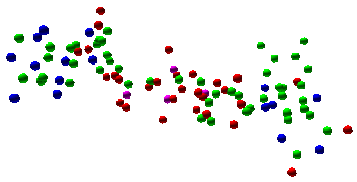
Given two molecules and an initial guess for the transformation that aligns them



[Besl92]

Iterative Closest Points

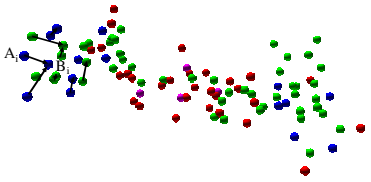
Assume closest points correspond



[Besl92]

Iterative Closest Points

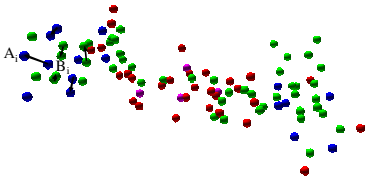
Assume closest points correspond: A→B



[Besl92]

Iterative Closest Points

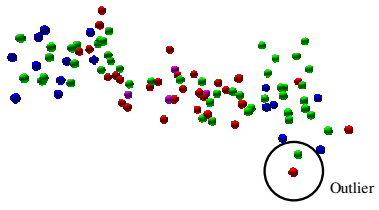
Assume closest points correspond: A→B and B→A



[Besl92]

Iterative Closest Points

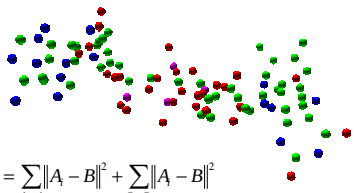
Rejecting outliers



[Besl92]

Iterative Closest Points

Find the transformation that optimally aligns proposed correspondences (superposition)



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

[Besl92]

Iterative Closest Points

Iterate until convergence

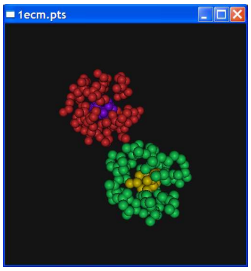
1. Select source points (from one or both molecules)
2. Match to points in the other molecule
3. Weight the correspondences
4. Reject outlier point pairs
5. Compute an error metric for the current transform
6. Minimize the error metric w.r.t. transformation

Computational complexity

- O(k * nlogn) for n points per binding site and k iterations
  - § k iterations * O(n) points * O(logn) to find closest point

Slide courtesy of Szymon Rusinkiewicz

Iterative Closest Points



Demo courtesy of Szymon Rusinkiewicz

Summary

Brute force
- Accurate, slow

RANSAC
- Approximate

Geometric hashing
- Fast query, after slow preprocessing
- Distance threshold implicit in hash bucket sizes

Association graphs
- Expensive for large point sets
- Distance threshold for "associations"

Iterative closest points
- Fast, in practice
- Requires good initial guess

Outline

Introduction

Point set representations

Point set matching
- Association graphs
- Geometric hashing
- Iterative closest point

Evaluation ←

Discussion

Evaluation

Questions:
- How well can the types of bound ligands be predicted from the positions of protein atoms near its binding site using standard point matching algorithms?
- What types of information (element type? residue type?) must be included with the atom positions in order to get good classification performance?

Binding Site Test Set

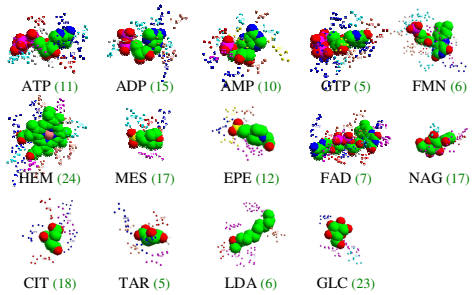
Protein-ligand complexes from PDB
- Crystallization resolution ≤ 3Å
- Ligands ≥10 HETATOMS

Remove homologous protein domains
- No two ligands contact same CATH superfamily
- No two ligands contact same SCOP family
- No two ligands from same PDB file

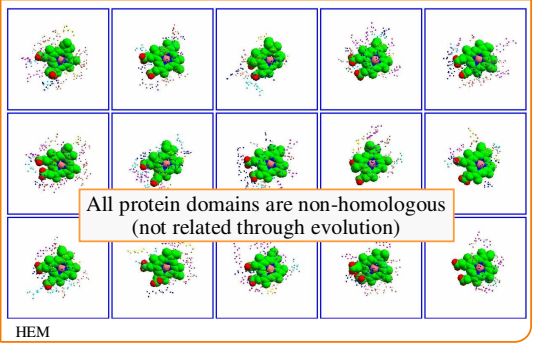
Select groups for classification experiment
- Classified by bound ligand type (e.g., ATP, NAD, etc.)
- Keep all classes with at least four members

Binding Site Test Set

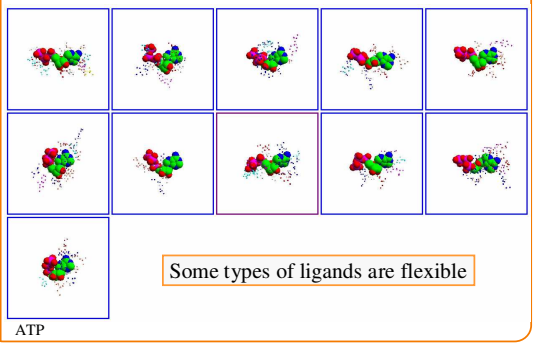
176 binding sites / 14 ligand types (classes)



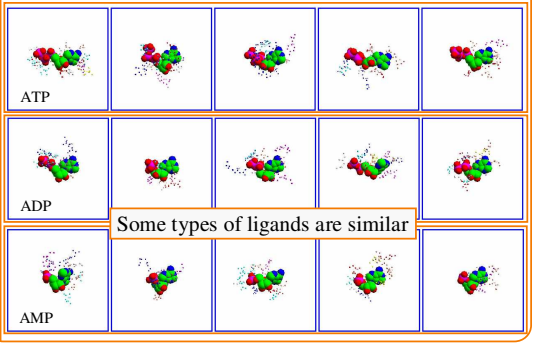
Binding Site Test Set



Binding Site Test Set

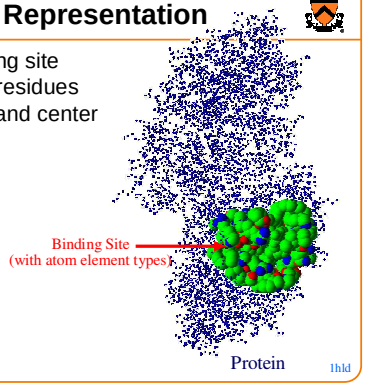


Binding Site Test Set



Binding Site Representation

Represent binding site
by set of atoms/residues
within 10Å of ligand center



Point Set Matching Method

Two steps:

1. Initial alignment computed with fast rotational matching
   (future lecture)
2. Final alignment computed with iterative closest points
   (last lecture)

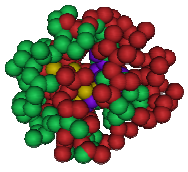
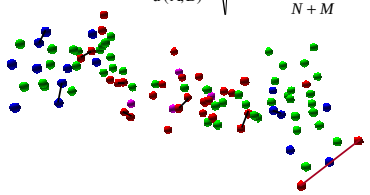


Image courtesy of Szymon Rusinkiewicz

Point Set Matching Method

Score is RMSD between corresponding points of the same type within 3Å of each other

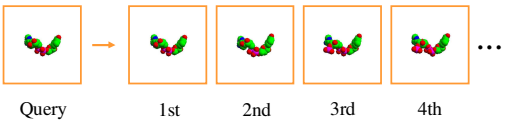
$$d(A,B) = \sqrt{\frac{\sum_{i=1}^{N}\|A_i - B\|^2 + \sum_{i=1}^{M}\|A - B_i\|^2}{N+M}}$$



Evaluation Method

"Leave-one-out" classification experiment

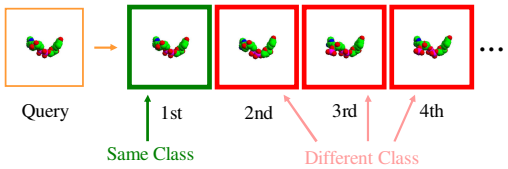
- ØMatch every ligand against all the others in data set
- Log a "hit" when best match performs same reaction
- Report percentage of hits (correctly classified ligands)



Evaluation Method

"Leave-one-out" classification experiment

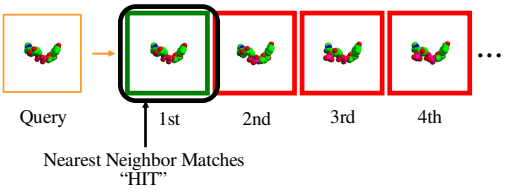
- ØMatch every ligand against all the others in data set
- Log a "hit" when best match performs same reaction
- Report percentage of hits (correctly classified ligands)



Evaluation Method

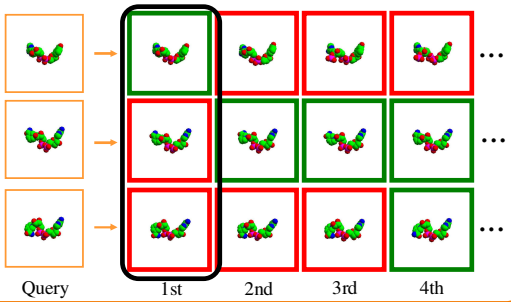
"Leave-one-out" classification experiment

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



Evaluation Method

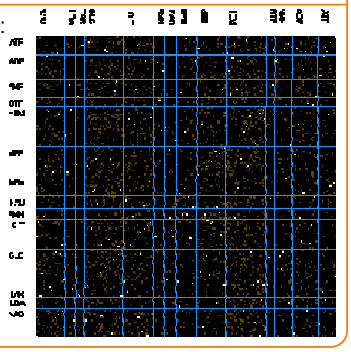
Classification rate is 33% is this example



Random

Classification rate:

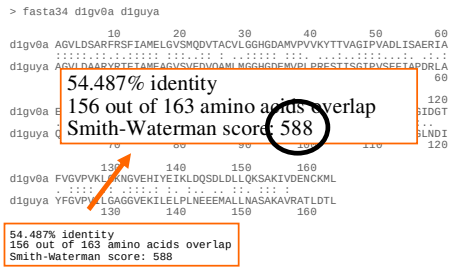
Random = 4.5%



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

FASTA

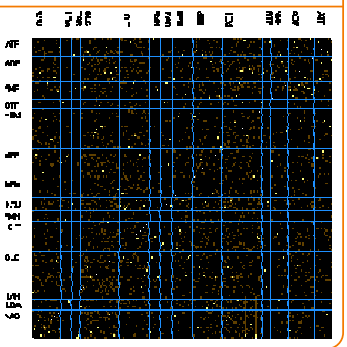
Distance between proteins is least Smith-Waterman score for pairs of SCOP domains contacting ligands



FASTA

Classification rate:

- FASTA = 9.7%
- Random = 4.5%



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

CE

Similarity is highest Z-score computed with CE for pairs of SCOP domains contacting ligands

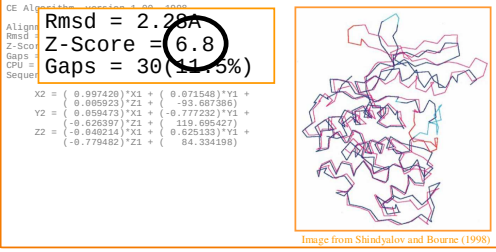
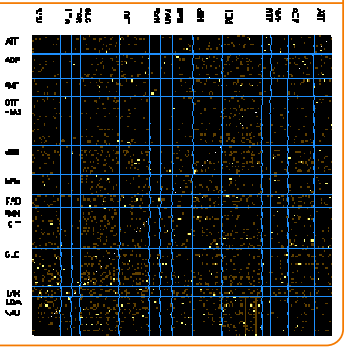


Image from Shindyalov and Bourne (1998)

CE

Classification rate:

- CE = 10.8%
- FASTA = 9.7%
- Random = 4.5%



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

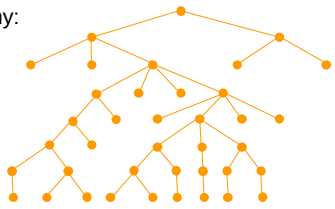
CATH

Distance measure is proximity in CATH hierarchy

- D(A,B) = least #levels to common ancestor in hierarchy for any pair of contacting chains

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S95
- S100



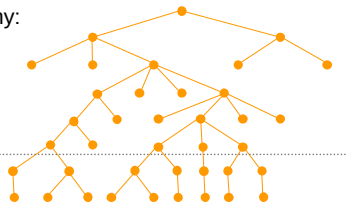
CATH

Distance measure is proximity in CATH hierarchy

- D(A,B) = least #levels to common ancestor in hierarchy for any pair of contacting chains

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S95
- S100



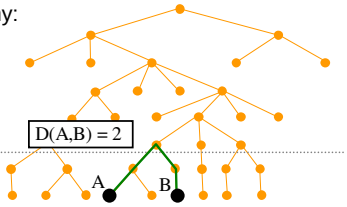
CATH

Distance measure is proximity in CATH hierarchy

- D(A,B) = least #levels to common ancestor in hierarchy for any pair of contacting chains

CATH hierarchy:

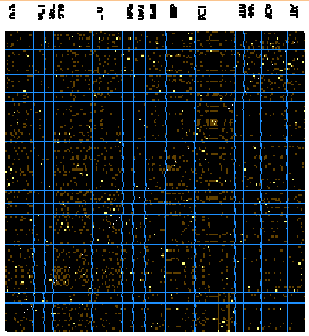
- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S95
- S100



CATH

Classification rate:

- CATH = 12.5%
- CE = 10.8%
- FASTA = 9.7%
- Random = 4.5%

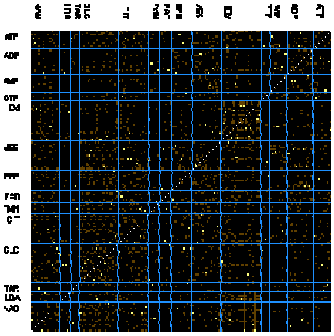


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

SCOP

Classification rate:

- SCOP = 17.0%
- CATH = 12.5%
- CE = 10.8%
- FASTA = 9.7%
- Random = 4.5%

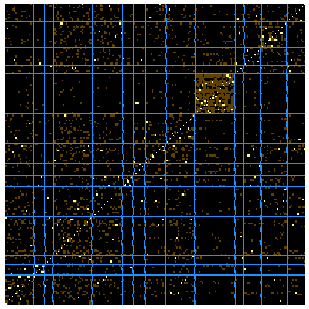


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Binding Site Atoms (Position only)

Classification rate:

- SiteAtoms = 34.7%
- SCOP = 17.0%
- CATH = 12.5%
- CE = 10.8%
- FASTA = 9.7%
- Random = 4.5%

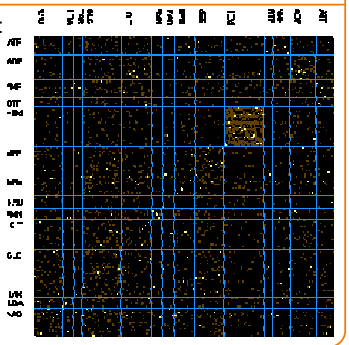


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Binding Site Atoms (w/ Elements)

Classification rate:

- SiteAtoms = 34.7%
- SiteAtomsE = 26.7%
- SCOP = 17.0%
- CATH = 12.5%
- CE = 10.8%
- FASTA = 9.7%
- Random = 4.5%

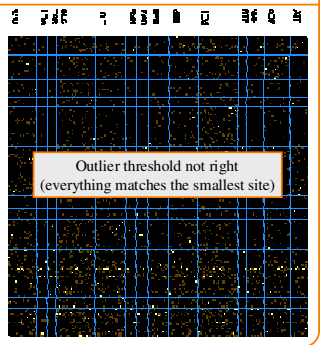


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Binding Site Atoms (w/ Elements)

Classification rate:

- SiteAtoms = 34.7%
- SiteAtomsE = 26.7%
- SCOP = 17.0%
- CATH = 12.5%
- CE = 10.8%
- FASTA = 9.7%
- Residues = 5.7%
- Random = 4.5%

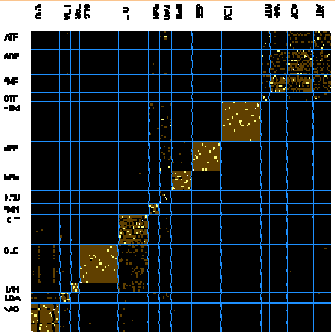


Best Matches:
(White = Self match)
(Yellow = Nearest Neighbor)
(Orange = 1st tier match)

Binding Site Atoms (w/ Elements)

Classification rate:

- LigAtoms = 88.1%
- LigAtomsE = 88.1%
- SiteAtoms = 34.7%
- SiteAtomsE= 26.7%
- SCOP = 17.0%
- CATH = 12.5%
- CE = 10.8%
- FASTA = 9.7%
- Residues = 5.7%
- Random = 4.5%



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

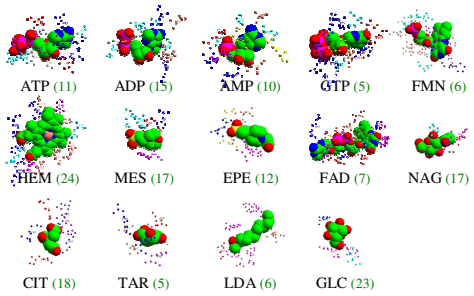
Conclusions (1 of 4)

Current point matching methods may be useful for classifying binding sites by ligand type when given the correct location for the center of the ligand

- 34.7% of ligand types classified correctly, as compared to 9.7%-17.1% with other methods

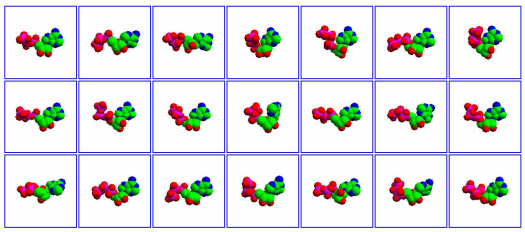
Conclusions (2 of 4)

Shape is sufficient to classify most ligands (88%)



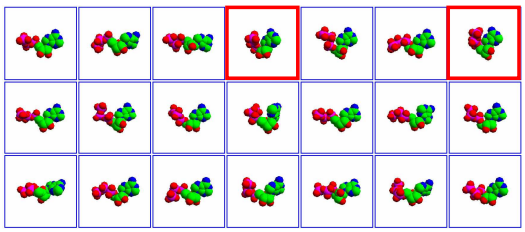
Conclusions (3 of 4)

The conformational variation of ligands bound to proteins in the PDB usually is not so great that it thwarts a rigid shape matching algorithm



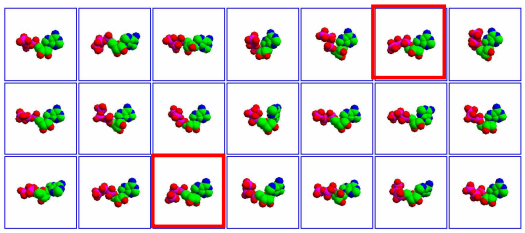
Conclusions (3 of 4)

The conformational variation of ligands bound to proteins in the PDB usually is not so great that it thwarts a rigid shape matching algorithm



Conclusions (3 of 4)

The conformational variation of ligands bound to proteins in the PDB usually is not so great that it thwarts a rigid shape matching algorithm



Conclusions (4 of 4)

Considering chemical properties (element type, residue type) did not help in this experiment

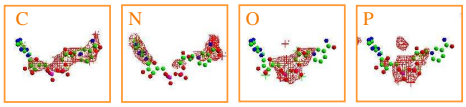
LigAtoms	= 88.1%
LigAtomsE	= 88.1%
SiteAtoms	= 34.7%
SiteAtomsE	= 26.7%
SCOP	= 17.0%
CATH	= 12.5%
CE	= 10.8%
FASTA	= 9.7%
Residues	= 5.7%
Random	= 4.5%

Still to Do …

Investigate impact of parameters
- Site representation, outlier rejection

Investigate other point properties
- Conservation, charge, etc.

Investigate other binding site representations
- Templates, surfaces, grids, etc.



Grid-based model of binding site (with atom types) predicted by XSITE

Discussion

?

References

[Besl92] P.J. Besl and N.D. McKay, "A method for registration of 3d shapes", IEEE Transactions on PAMI, 14, 1992, pp. 239-256.

[Brakoulias04] A. Brakoulias,, R.M. Jackson, "Towards a structural classification of phosphate binding sites in proteinnucleotide complexes: an automated all-against-all structural comparison using geometric matching," Proteins-Structure Function and Genetics, 56, 2004, pp. 250-260.

[Lin94] S.L. Lin, R. Nussinov, D. Fischer, H.J. Wolfson, "Molecular-Surface Representations By Sparse Critical-Points," Proteins-Structure Function and Genetics, 18, 1994, pp. 94-101.

[Pennec98] X. Pennec,, N. Ayache, "A geometric algorithm to find small but highly similar 3D substructures in proteins," Bioinformatics, 14, 1998, pp. 516-522.
[Schmitt02] S. Schmitt, D. Kuhn, G. Klebe, "A new method to detect related function among proteins independent of sequence and fold homology," J Mol Biol, 323, 2002, pp. 387-406.

[Shulman-Peleg04] A. Shulman-Peleg, R. Nussinov, H.J. Wolfson, "Recognition of functional sites in protein structures," J Mol Biol, 339, 2004, pp. 607-633.

[Wolfson97] H.J. Wolfson and I. Rigoutsos, "Geometric hashing: an overview," IEEE Computational Science & Engineering, 4(4), 1997, pp. 10-21

[Weskamp04] N. Weskamp, D. Kuhn, E. Hullermeier, G. Klebe, "Efficient similarity search in protein structure databases by k-clique hashing," Bioinformatics, 20, 2004, pp. 1522-1526.