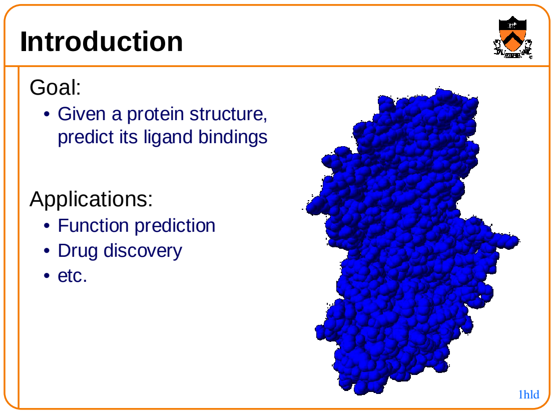
Representation and Matching of Ligand Binding Sites I

Thomas Funkhouser

Princeton University

CS597A, Fall 2005



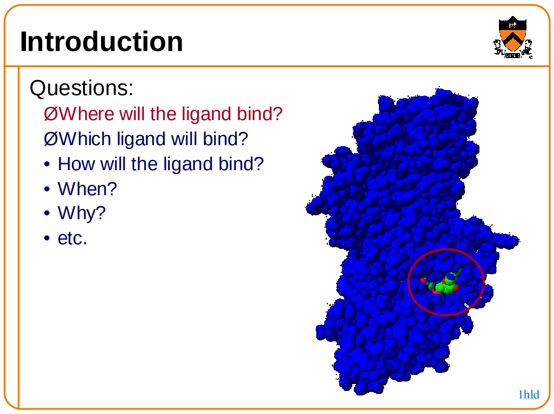
Introduction

Goal:

- Given a protein structure, predict its ligand bindings

Applications:

- Function prediction
- Drug discovery
- etc.



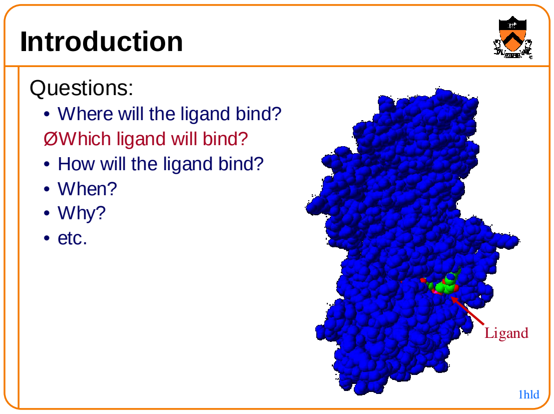
Introduction

Questions:

ØWhere will the ligand bind?

ØWhich ligand will bind?

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



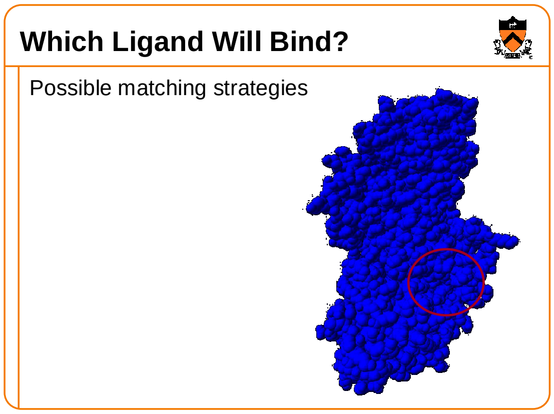
Introduction

Questions:

- Where will the ligand bind?

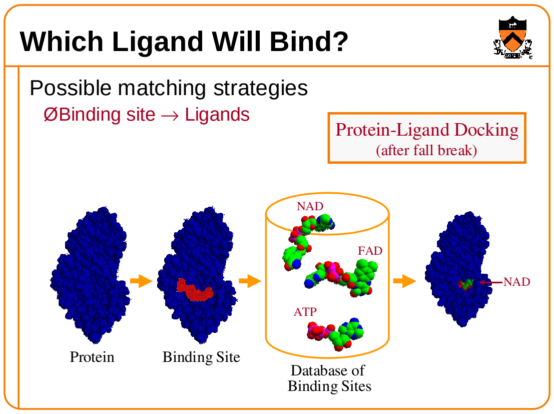
ØWhich ligand will bind?

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



Which Ligand Will Bind?

Possible matching strategies

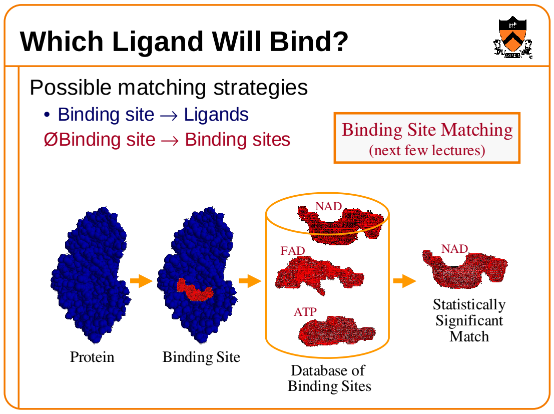


Which Ligand Will Bind?

Possible matching strategies

ØBinding site → Ligands

Protein-Ligand Docking
(after fall break)



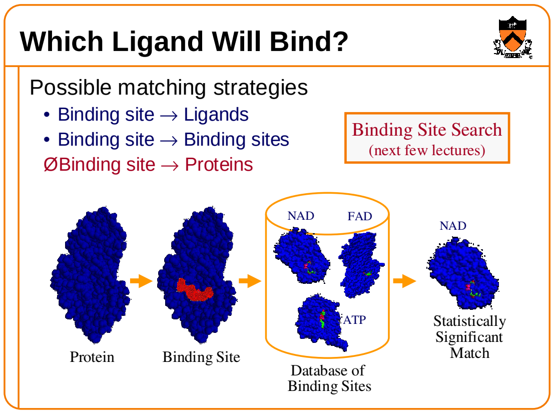
Which Ligand Will Bind?

Possible matching strategies

- Binding site → Ligands

ØBinding site → Binding sites

Binding Site Matching
(next few lectures)



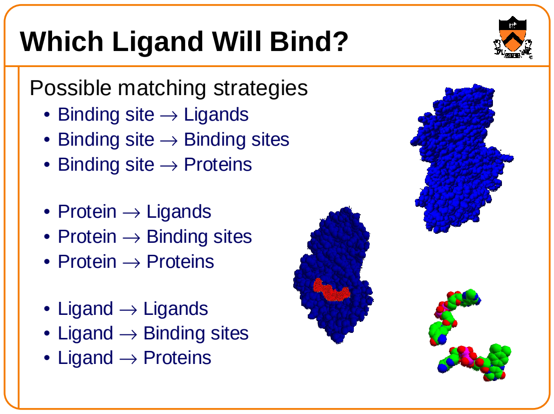
Which Ligand Will Bind?

Possible matching strategies

- Binding site → Ligands
- Binding site → Binding sites

ØBinding site → Proteins

Binding Site Search
(next few lectures)



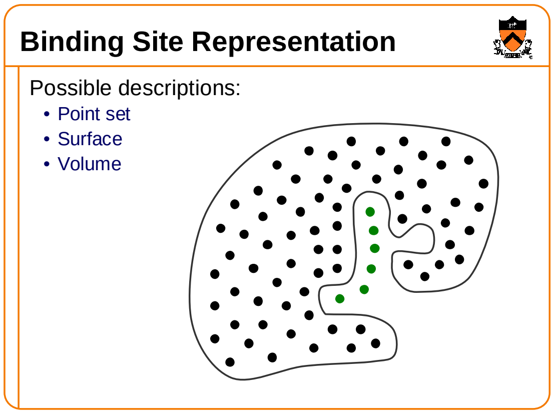
Which Ligand Will Bind?

Possible matching strategies

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

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

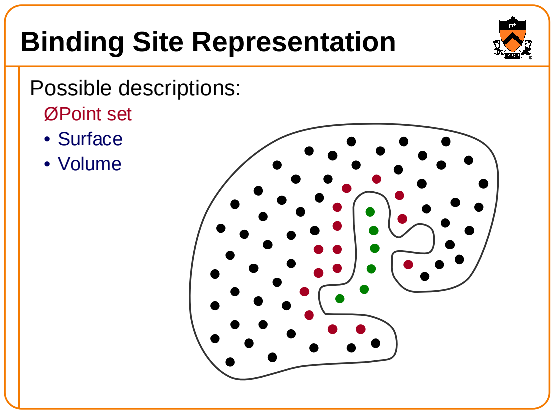
- Ligand → Ligands
- Ligand → Binding sites
- Ligand → Proteins



Binding Site Representation

Possible descriptions:

- Point set
- Surface
- Volume

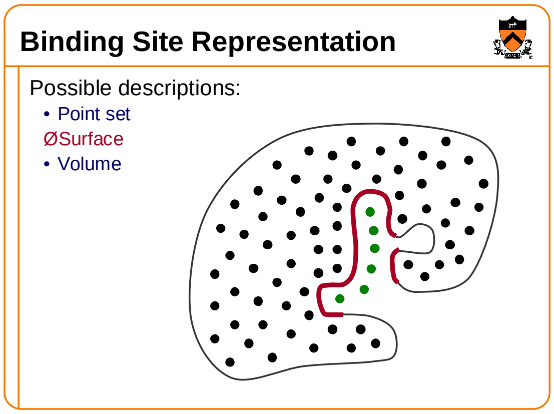


Binding Site Representation

Possible descriptions:

ØPoint set

- Surface
- Volume



Binding Site Representation

Possible descriptions:

- Point set

ØSurface

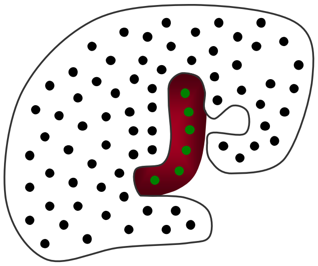
- Volume

Binding Site Representation

Possible descriptions:

- Point set
- Surface

ØVolume



Course Schedule

Next few lectures

- 10/10: Binding site point sets
- 10/12: Discussion with Helen Berman
- 10/17: Binding site templates
- 10/19: Project proposals
- 10/24: Binding site surfaces
- 10/26: Binding site volumes

After fall break

- Protein-ligand docking
- Protein-protein docking
- Drug screening and design
- Structure determination

Course Schedule

Next few lectures

- 10/10: Binding site point sets

Ø10/12: Discussion with Helen Berman

- 10/17: Binding site templates
- 10/19: Project proposals
- 10/24: Binding site surfaces
- 10/26: Binding site volumes

After fall break

- Protein-ligand docking
- Protein-protein docking
- Drug screening and design
- Structure determination

Course Schedule

Next few lectures

- 10/10: Binding site point sets
- 10/12: Discussion with Helen Berman
- 10/17: Binding site templates

Ø10/19: Project proposals

- 10/24: Binding site surfaces
- 10/26: Binding site volumes

After fall break

- Protein-ligand docking
- Protein-protein docking
- Drug screening and design
- Structure determination

Course Schedule

Next few lectures

Ø10/10: Binding site point sets

- 10/12: Discussion with Helen Berman
- 10/17: Binding site templates
- 10/19: Project proposals
- 10/24: Binding site surfaces
- 10/26: Binding site volumes

After fall break

- Protein-ligand docking
- Protein-protein docking
- Drug screening and design
- Structure determination

Outline

Introduction

Point set representations ←

Point set matching

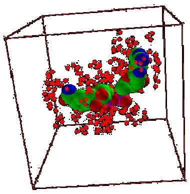
- Association graphs
- Geometric hashing
- Iterative closest point

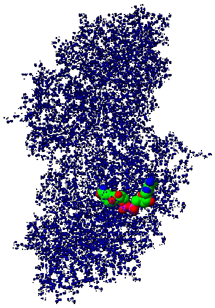
Evaluation

Discussion

Point Set Representation

Set of attributed points

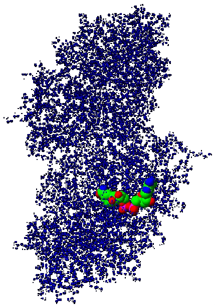




1hld

Point Set Representation

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

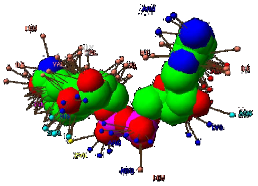


1hld

Point Set Representation

Set of attributed points

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



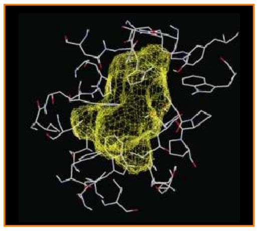
Key Atoms Surrounding Binding Site

1hld

Point Set Representation

Set of attributed points
- Atoms

ØResidues
- Pseudo-centers
- Surface critical points
- etc.



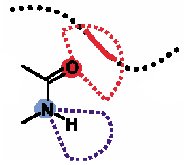
Residues Surrounding Binding Site

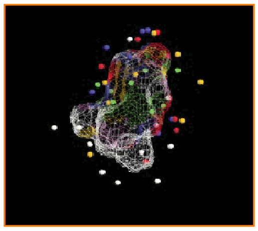
[Schmitt02]

Point Set Representation

Set of attributed points
- Atoms
- Residues

ØPseudo-centers
- Surface critical points
- etc.





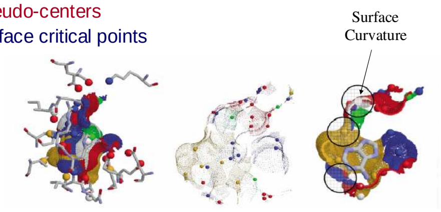
Residues Surrounding Binding Site

[Schmitt02]

Point Set Representation

Set of attributed points
- Atoms
- Residues

ØPseudo-centers
- Surface critical points
- etc.



Represent Chemical and Geometric Properties of Surface

[Shulman-Peleg04]

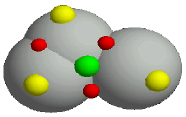
Point Set Representation

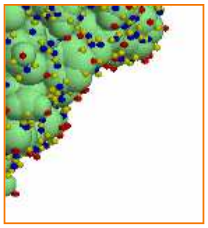
Set of attributed points

- 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 ←

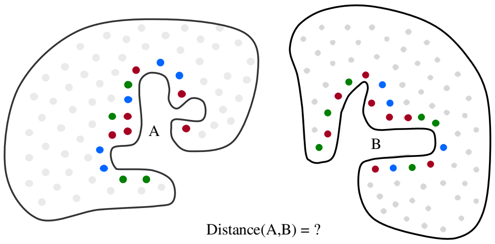
- Association graphs
- Geometric hashing
- Iterative closest point

Evaluation

Discussion

Point Set Matching

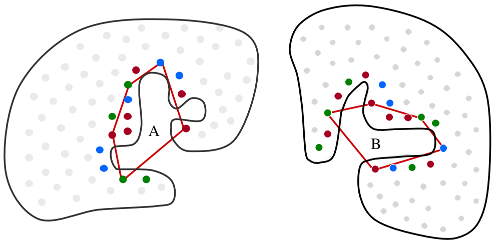
Goal is to compute a distance measure for a pair of attributed point sets



Point Set Matching

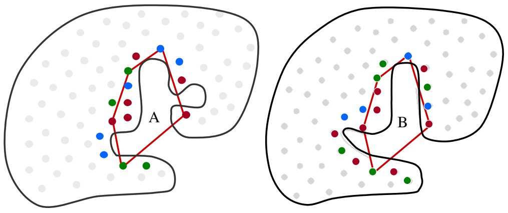
Challenge is to find corresponding points

- “Subset of points in A” may match “subset of points in B”



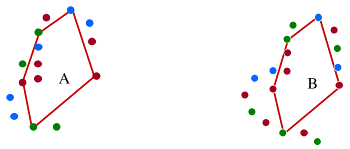
Point Set Matching

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



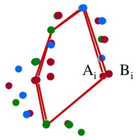
Point Set Matching

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)

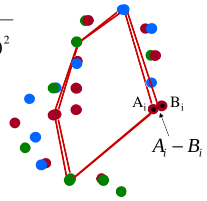


Least-squares optimal superposition of corresponding points

Point Set Matching

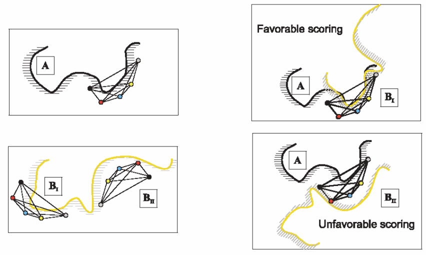
Calculating a superposition and distance measure is easy if correspondences are known (proposed)

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



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

Point Set Match Scoring



Aligned points may be examined further to compute scoring function (e.g., check if surfaces align)

[Schmitt02]

Outline

Introduction

Point set representations

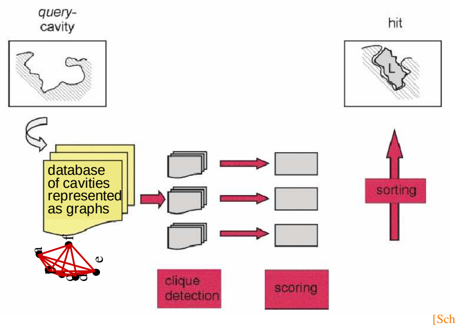
Point set matching

ØAssociation graphs
- Geometric hashing
- Iterative closest point

Evaluation

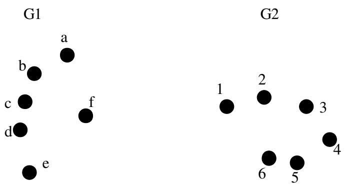
Discussion

Matching with Association Graphs

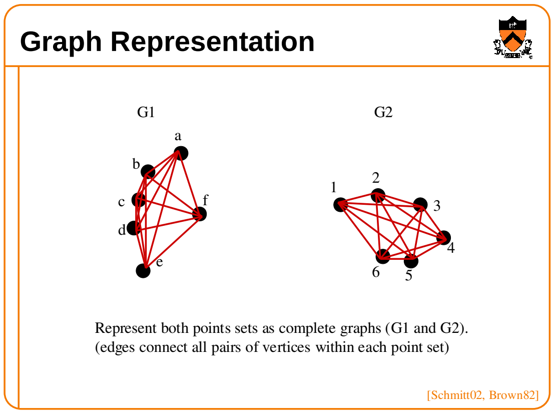


[Schmitt02]

Graph Representation



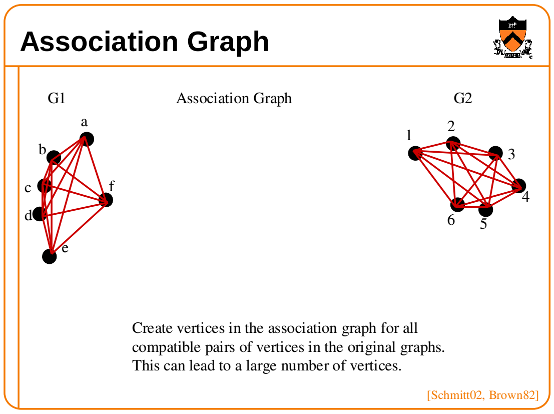
[Schmitt02, Brown82]



Graph Representation

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

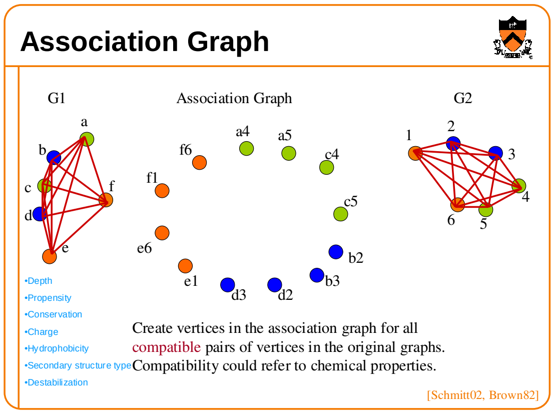
[Schmitt02, Brown82]



Association Graph

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]

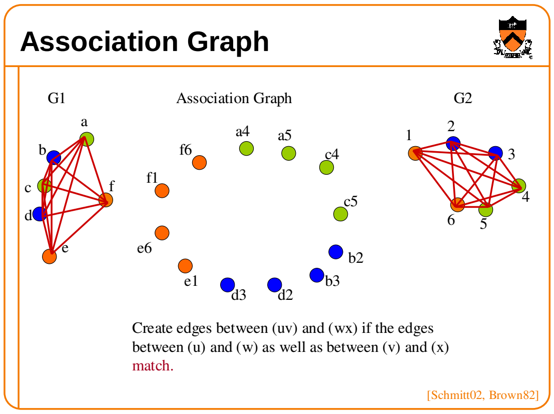


Association Graph

- •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.

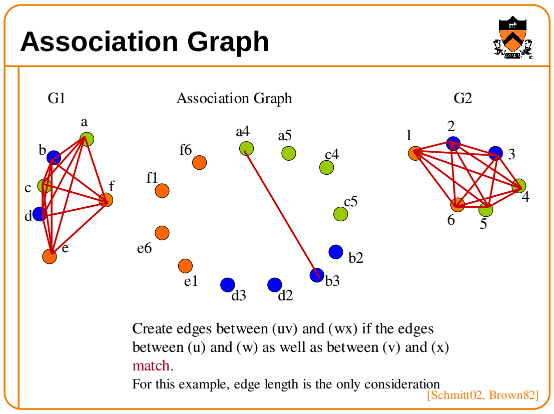
[Schmitt02, Brown82]



Association Graph

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

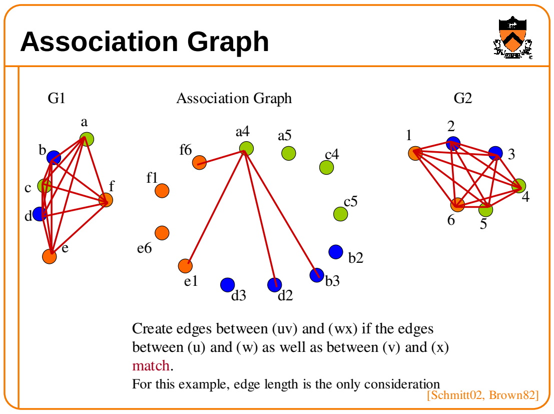
[Schmitt02, Brown82]



Association Graph

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

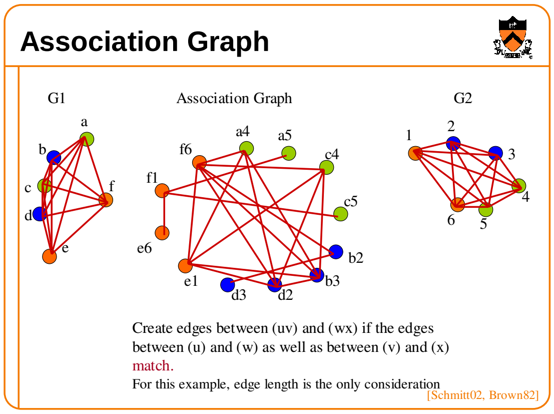
[Schmitt02, Brown82]



Association Graph

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]

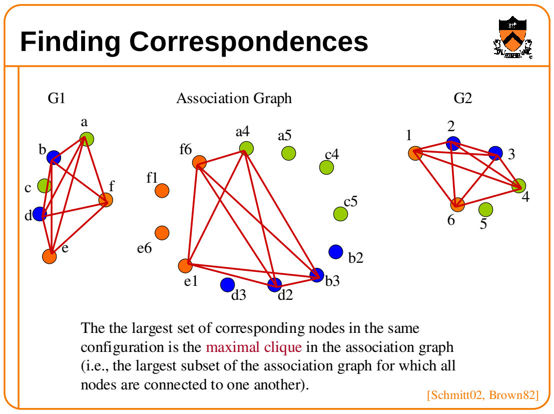


Association Graph

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

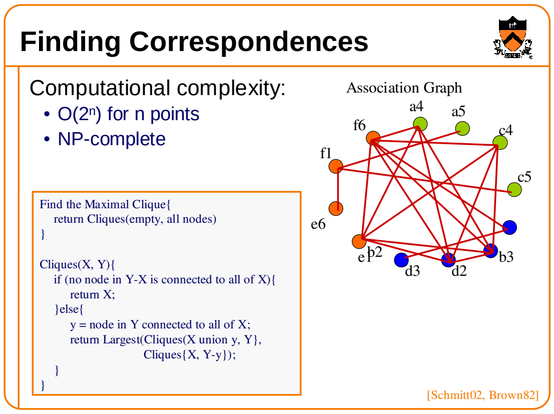
[Schmitt02, Brown82]



Finding Correspondences

The the largest set of corresponding nodes in the same configuration is the maximal clique in the association graph (i.e., the largest subset of the association graph for which all nodes are connected to one another).

[Schmitt02, Brown82]



Finding Correspondences

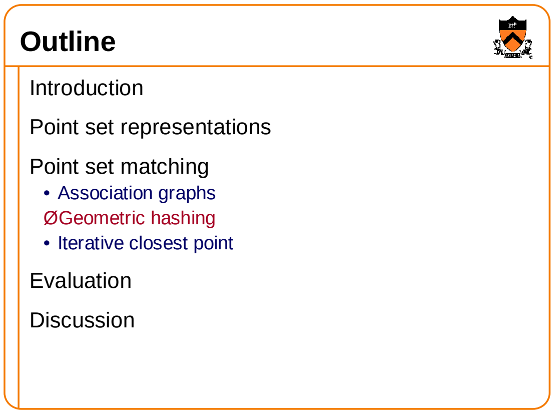
Computational complexity:

- $O(2^n)$ for n points
- NP-complete

```
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

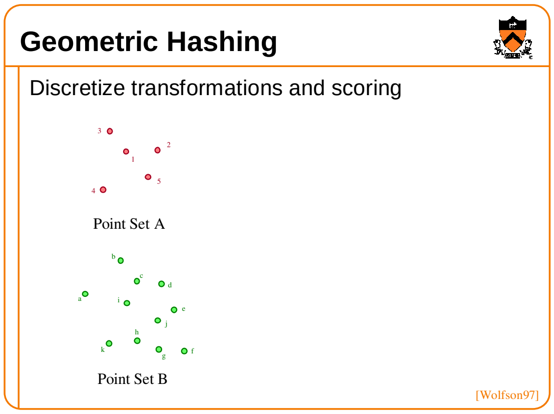
Point set representations

Point set matching

- Association graphs
ØGeometric hashing
- Iterative closest point

Evaluation

Discussion



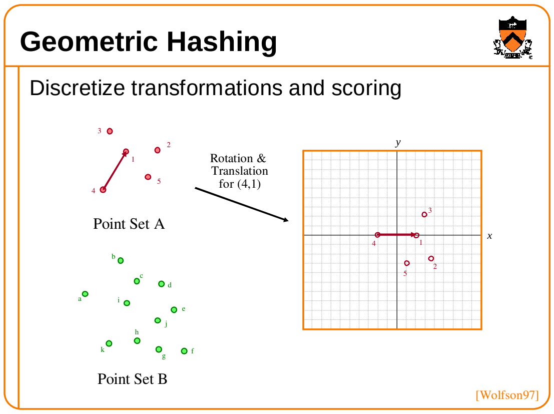
Geometric Hashing

Discretize transformations and scoring

Point Set A

Point Set B

[Wolfson97]



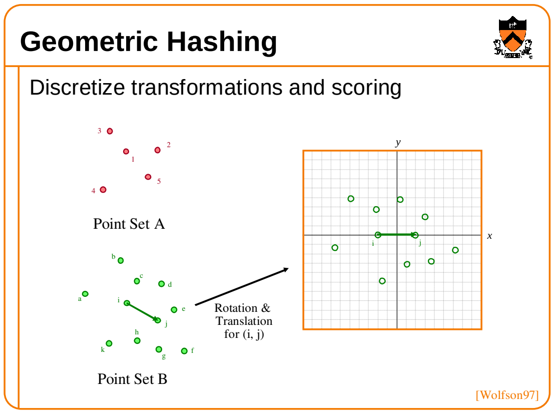
Geometric Hashing

Discretize transformations and scoring

Point Set A

Point Set B

[Wolfson97]



Geometric Hashing

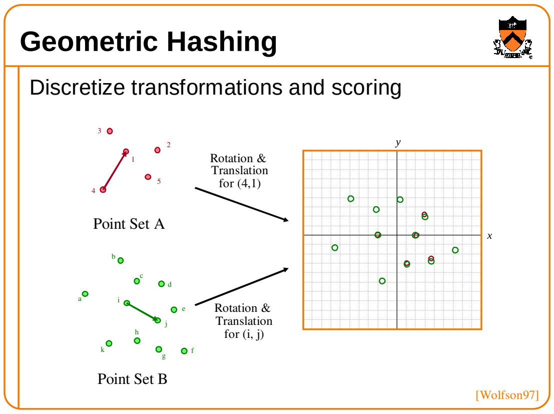
Discretize transformations and scoring

Point Set A

Point Set B

Rotation & Translation for (i, j)

[Wolfson97]



Geometric Hashing

Discretize transformations and scoring

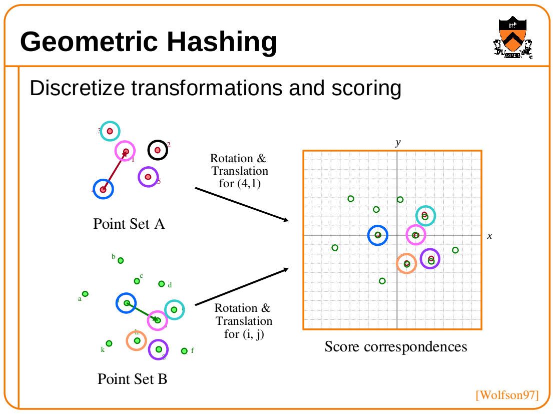
Rotation & Translation for (4,1)

Point Set A

Point Set B

Rotation & Translation for (i, j)

[Wolfson97]



Geometric Hashing

Discretize transformations and scoring

Rotation & Translation for (4,1)

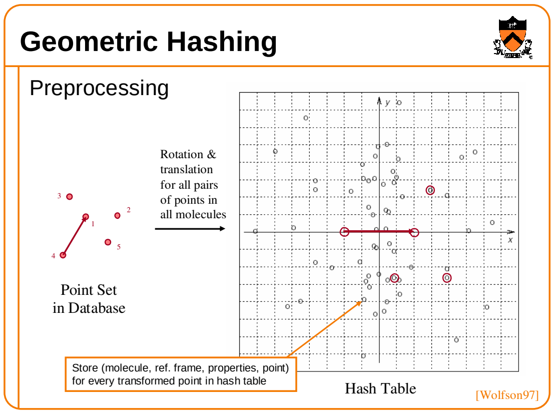
Point Set A

Point Set B

Rotation & Translation for (i, j)

Score correspondences

[Wolfson97]



Geometric Hashing

Preprocessing

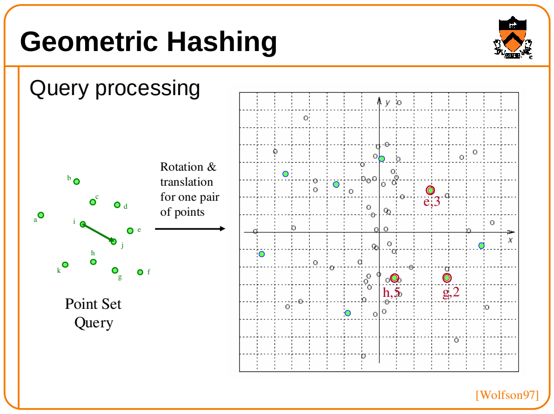
Rotation & translation for all pairs of points in all molecules

Point Set in Database

Store (molecule, ref. frame, properties, point) for every transformed point in hash table

Hash Table

[Wolfson97]



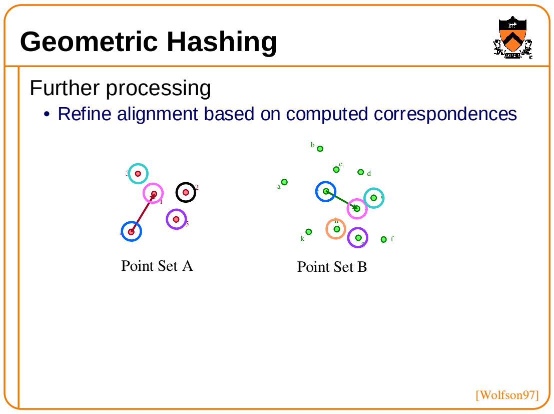
Geometric Hashing

Query processing

Rotation & translation for one pair of points

Point Set Query

[Wolfson97]



Geometric Hashing

Further processing

- Refine alignment based on computed correspondences

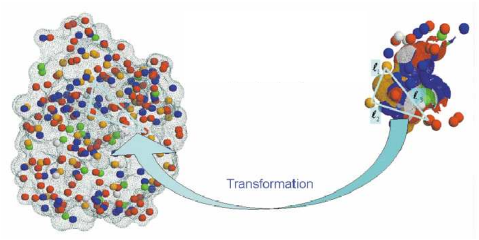
Point Set A

Point Set B

[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

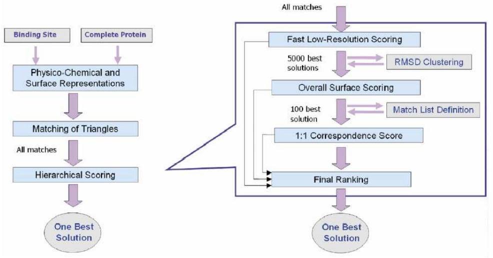
- O(n^4) for n points per binding site
  - O(n^3) possible triples * O(n) transformations per triple

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

- Association graphs
- Geometric hashing

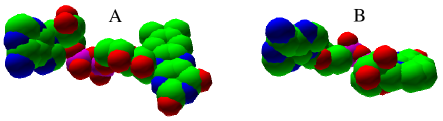
ØIterative closest point

Evaluation

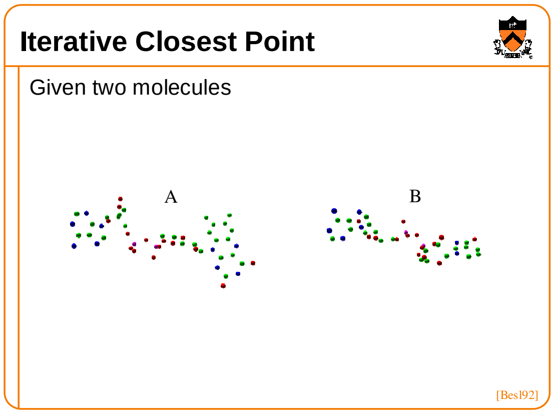
Discussion

Iterative Closest Point

Given two molecules



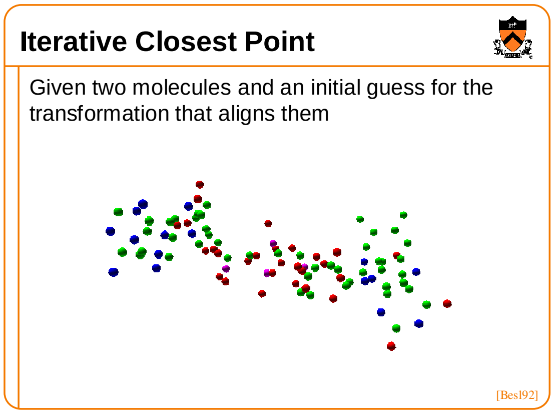
[Besl92]



Iterative Closest Point

Given two molecules

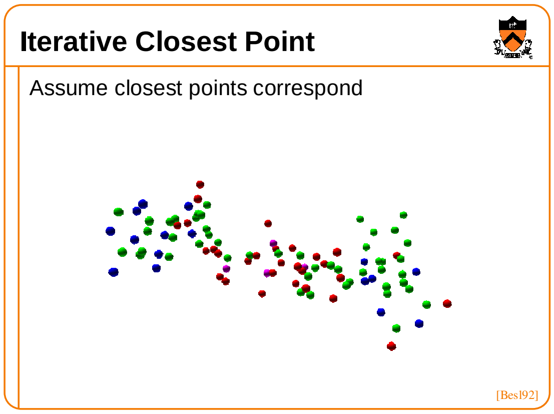
[Besl92]



Iterative Closest Point

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

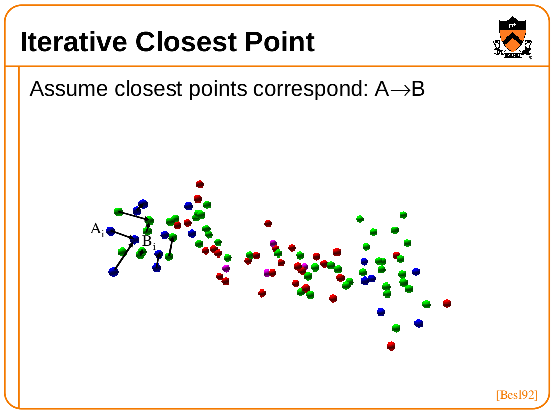
[Besl92]



Iterative Closest Point

Assume closest points correspond

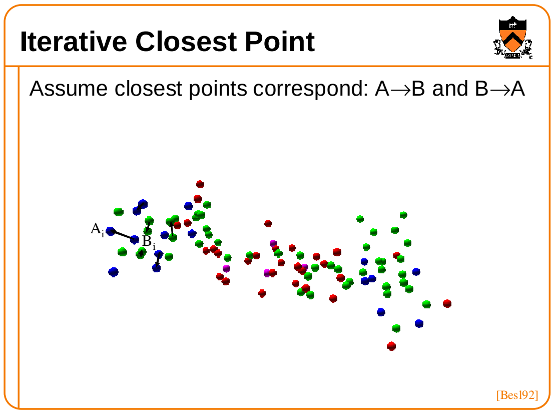
[Besl92]



Iterative Closest Point

Assume closest points correspond: A→B

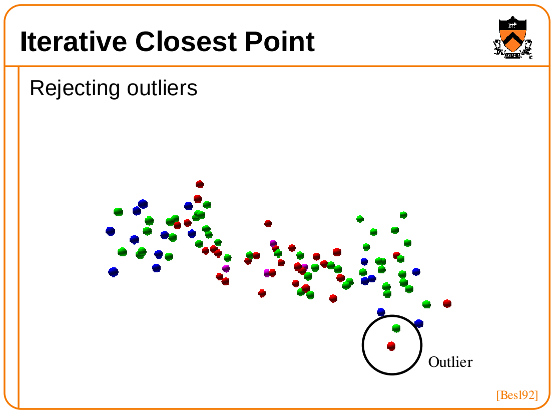
[Besl92]



Iterative Closest Point

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

[Besl92]



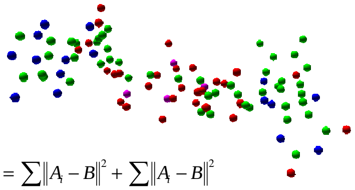
Iterative Closest Point

Rejecting outliers

[Besl92]

Iterative Closest Point

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 Point

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 Point

Demo

Demo courtesy of Szymon Rusinkiewicz

Summary

Association graphs

- Expensive for large point sets
- Distance threshold for “associations”

Geometric hashing

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

Iterative closest points

- Fast, in practice
- Allows soft scoring functions
- Requires good initial guess

Outline

Introduction

Point set representations

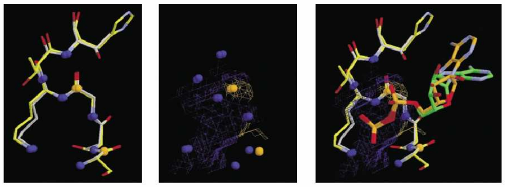
Point set matching

- Association graphs
- Geometric hashing
- Iterative closest point

Evaluation ←

Discussion

Aligned Points



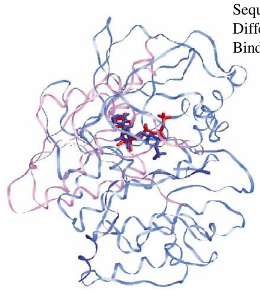
Equivalent phosphate binding areas in the binding pockets of uridylate kinase (1ukz) and the structure of a kinesine-type domain (3kar).

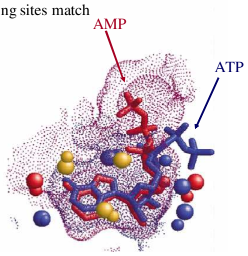
Sequence identity is very low (SW=41)
Different functions
Binding sites match

[Schmitt02]

Aligned Points

Sequence identity is very low (SW=41)
Different functions
Binding sites match



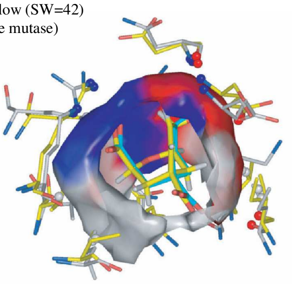


Statistical significance?

[Shulman-Peleg04]

Aligned Points

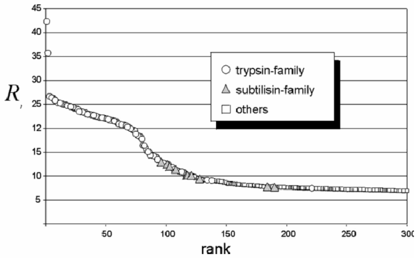
Sequence identity is very low (SW=42)
Same function (chorismate mutase)
Binding sites match



Superposition of the binding pockets from the chorismate mutases 1ecm and 4csm

[Weskamp04]

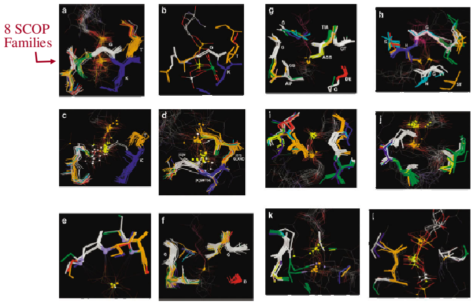
Ranked Matches



Results of query with binding site of trypsin structure (1tpo)

[Schmitt02]

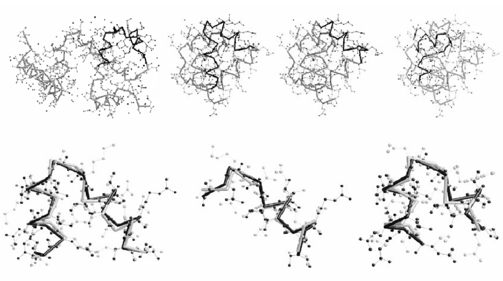
Clusters of Point Sets



Clusters of 476 phosphate binding sites

[Brakoulias04]

Detected Structural Motifs



[Pennec98]

Generally Speaking …

Small sets of proteins
- Serine proteases (catalytic triad)
- Adenine-binding proteins (largest source of data)

Focus on true positives
- False positives, false negatives?
- Aggregate statistics for large set of queries?
- Statistical significance? [Stark03]

Rarely provide comparison to related approaches
- Comparison to sequence-based matching methods?
- Comparison to other local structure matching methods?

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.