

Structure Alignment, Structure Prediction & Protein Folding

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (1)

Structure Alignment Objectives

- Detect evolutionary relationships
- Find possible active sites
- Locate most stable parts of structure
- Assemble templates for structure prediction
- Increase understanding of protein architecture

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (2)

What's the Problem?

- Find rotation matrix R and translation vector T for which:

$$Y = R \cdot X + T$$

- NP hard!
- No known deterministic algorithm

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (3)

Algorithm Terminology

- NP := non-deterministic polynomial time
- “guesses” can be checked in polynomial time
- NP-hard := NP problem at least as hard or harder as all other NP problems
- “order” of algorithms = max. time needed:
 - e.g. $O(N)$, $O(N^2)$, $O(\log(N))$, $O(e^N)$...
 - Want $O(N)$ not $O(N^2)$ or even $O(N^N)$!!!
 - Polynomial time (P): $aN + bN^2 + cN^3 + \dots$

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (4)

Two Main Issues

- 1) Measure used to quantify difference, i.e. a similarity score
- 2) Combination of non-locality of scoring function and existence of gaps and insertions

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (5)

Similarity Measures

Root Mean Square Deviation (RMSD):

$$R_{ms} = \sqrt{\sum_{i=1}^n \frac{(X_{Ai} - X_{Bi})^2}{n}}$$

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (6)

RMSD

- Penalizes worst fitting atoms
- Contributions of individual atoms not discernable
- Similarity Scores:

$$S = \sum_{i,j} S(i,j) - nG$$

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (7)

Other measures

- Differences of Distance maps
 - DALI (distance matrix alignment program)
- Contact Map overlay
- Secondary structure element (SSE) representations
 - VAST
 - CATH

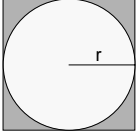
Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (8)

Optimization Algorithms Used

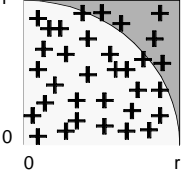
- Dynamic programming [CATH]
- Monte Carlo [DALI]
- 3D clustering
- Graph theory [VAST]
- Combinatorial Extension [CE]
- Combinations

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (9)

Monte Carlo



$A_S = 4r^2$
 $A_C = \pi r^2$
 $\pi = 4A_C/A_S$



$A_S = A_S + 1;$
 if (hit in circle) { $A_C = A_C + 1;$ }
 $\pi = 4A_C/A_S$

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (10)

Monte Carlo [cont.]

$\pi = 3.141592654$

1000	3.26	3.204	3.176	3.144
10000	3.1672	3.1428	3.1488	3.1448
100000	3.15764	3.14788	3.14732	3.14028
1000000	3.142848	3.142864	3.141488	3.141152
10000000	3.141352	3.142390	3.141398	3.141341
20000000	3.141586	3.142083	3.141652	3.141601
30000000	3.141718	3.141860	3.141456	3.141669
40000000	3.141715	3.141879	3.141336	3.141444
50000000	3.141698	3.141900	3.141479	3.141451
60000000	3.141806	3.141898	3.141597	3.141458
70000000	3.141974	3.141647	3.141531	3.141373
80000000	3.141938	3.141636	3.141504	3.141393
90000000	3.141868	3.141696	3.141478	3.141412
100000000	3.141822	3.141734	3.141466	3.141453

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (11)

Structure Prediction

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (12)

Structure Prediction

Sequence ➡ Structure ➡ Function

- ❑ Many times more sequences than structures
- ❑ Structure most conserved during evolution
- ❑ Sequence alignment methods inadequate at low identity levels
- ❑ Structure prediction "holy grail" of structure biology community

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (13)

Structure Prediction [2]

- ❑ 3 main areas:
 - ❑ Homology modeling
 - ❑ Fold recognition
 - ❑ *Ab initio* prediction
 - ❑ Automated servers
- ❑ Bi-Annual meeting: Critical Assessment of Structure prediction (CASP), Assilomar, CA

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (14)

Homology Modeling

- ❑ 6 Steps:
 - ❑ Align target sequence on the backbone of parent structure
 - ❑ Choice of core structure
 - ❑ Construction of core side chains
 - ❑ Building the loops
 - ❑ Refinement of the model
 - ❑ Estimation of reliability

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (15)

Alignment

- Map each target residue onto residue closest in space in each parent
- Usually done with sequence alignment methods
- Very susceptible to % identity
- Produces more or less correct results

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (16)

Choice of Core Structure

- Which part of which parent structure to use
 - Which parent?
 - Where does the similarity end?
- Often less overall homology better choice
- Local seq. similarity often misleading

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (17)

Core Side Chains

- Number of algorithms build back side chains with high accuracy
- Not translated into prediction arena
- 50% of χ angles > 30° in error
- Accuracy of side-chain builder deteriorates rapidly with increasing main-chain errors

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (18)

Building the loops

- ❑ After core completion, short regions of chain remain to be built
- ❑ Resembles *ab initio* problem
- ❑ Biggest error source with predictions
 - ❑ Erroneous placement of core residues
 - ❑ Intramolecular interactions of loops

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (19)

Refinement of the model

- ❑ Can the initial model be improved?
- ❑ Energy minimization & molecular dynamics
- ❑ Typically models do not improve but deteriorate!!
- ❑ Large structural rearrangements required?
- ❑ Abandoned for most part at CASP3

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (20)

Reliability Estimation

- ❑ Normally not possible
- ❑ Statistics show correlation of model's quality with homology between target and parents
- ❑ No model better than parents!!!!

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (21)

Why bother then?

- ❑ Some results approach 2Å RMSD or better in small regions
 - ❑ Identification of potential active site residues
 - ❑ Fold determination
 - ❑ Watch out for residues in loops!!
- ❑ At higher RMSD:
 - ❑ Gross topological location of residues
 - ❑ General chain path

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (22)

Fold Recognition

- ❑ Best method: Alexey Murzin's brain
- ❑ Comparison of target sequence with library of known folds
- ❑ At CASP3: 13 out of 17 correct by at least one group
- ❑ Human expertise clearly important
- ❑ Identified parents used with limited success for comparative modeling

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (23)

Ab Initio Prediction

- ❑ Build a model without specific template structures
- ❑ Use of secondary structure prediction methods!
- ❑ Knowledge based energy functions (HMM or neural nets)
- ❑ RMSD's usually around 6Å

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (24)

Protein Folding

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (25)

Protein Folding

- Levinthal Paradox:
 - $t(\text{folding}) \propto x^n$
 - x =: degrees of freedom
 - n =: number of amino acids
- There is insufficient time for a protein to explore the entire conformational space via random search!
- Proteins must fold through some directed process!

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (26)

Protein Folding [2]

- Current view: statistical ensembles of states described by complex multidimensional potential energy functions
- “folding funnel”:

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (27)

Protein Folding [3]

- ❑ Decrease in Energy to compensate for decrease of entropy
- ❑ Statistical mechanical treatments describe funnel shape and flatness by few critical parameters
- ❑ Converted to computer simulations using simple lattice models

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (28)

Protein Folding [4]

- ❑ Experimental advances:
 - ❑ Support for secondary structure formation as early events
 - ❑ Alternative: hydrophobic collapse and formation of SSE's from there
 - ❑ Isolated helices make use of nucleation propagation mechanisms
- ❑ Implications of strong SSE formation propensities for SS prediction!

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (29)

Protein Folding [5]

- ❑ Use of Molecular Mechanics calculations to understand stability
- ❑ Also in combination with Molecular Dynamics
- ❑ Importance of solvation effect! (Poisson-Boltzmann)

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (30)

Protein Folding [6]

- ❑ Folding Pathways involve transitional SSE's
- ❑ Small differences in free Energy between different SSE states
- ❑ SS propensities provide small set of "correct" SSE's for given sequence
- ❑ Tertiary interactions lead to final selection of actual native topology

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (31)

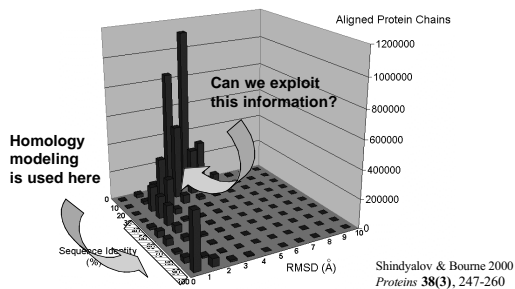
All x All Structure Comparison



- ❑ Combinatorial Extension (CE)
- ❑ 11,748 chain in the PDB (1/98)
- ❑ 1868 representatives
- ❑ 24,000 Cray T3E CPU hours
- ❑ <http://cl.sdsc.edu/ce/>
- ❑ Shindyalov & Bourne 1998 *Protein Engineering* 11(9) 739

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (32)

Structure Alignments using CE




Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (33)

CKAAPs

- Conserved Key Amino Acid Positions
- Sequence based analysis:
 - CE alignments compared with non-redundant protein sequence database
 - Subsequences with 4-90% sequence identity are grouped and the raw amino acid count at each position calculated
- Identified AAs important for folding and structure stability

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (34)

CKAAPs In Comparison



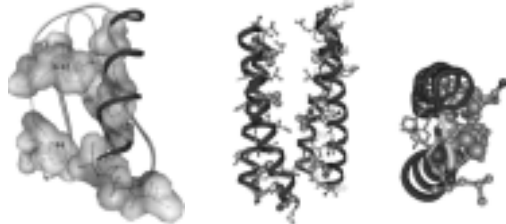
- Red
Clarke et al, 1999 (9 aa)
- Green
Mirny et al, 1998, 1999 (8 aa)
- Blue
CKAAPs (18 aa)
- Magenta: Red + Blue
- Yellow: Red + Green

CKAAPs Includes structural core residues (folding nucleus) and positions at turns.

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (35)

Paracelsus Challenge

- 28 mutations to the B1 domain convert it to a ROP like domain (Dalel et al.)
- includes all 12 CKAAPs
- CKAAPs important to the helix-helix interface in ROP not found in B1



Protein G B1 domain HTH protein (ROP) Top view (monomer)

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (36)
