

Secondary Structure Calculation & Structure Prediction/Protein Folding

Introduction to Structural Bioinformatics

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (1)

Overview

- o Secondary Structure Calculation
 - o Kabsch & Sander
- o Protein Folding
- o Structure Prediction
 - o Homology Modeling
 - o Fold recognition
 - o Ab initio methods

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (2)

Secondary Structure Calc.

- o 4 major methods:
 - o P-Curve [Sklenar et al. (1989), Proteins 6: 46-60]
 - o DEFINE [Richards & Kundrot (1988), Proteins 3: 71-84]
 - o STRIDE [Frishman & Argos (1995), Proteins 23: 566-579]
 - o DSSP [Kabsch & Sander (1983), Biopolymers 22: 2577-2637]
- o Most widely used: DSSP!

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (3)

DSSP Algorithm

- o "Define Secondary Structure of Proteins"
- o Automated processing of PDB files and archived at ftp://ftp.embl-heidelberg.de/pub/databases/protein_extra/dssp
- o Defines elementary hydrogen-binding patterns "turn" and "bridge"
- o Defines cooperative 2°structure as repeats of elementary units:
 - o Multiple Turns := "helices"
 - o Multiple Bridges := "ladders"
 - o Interconnected Ladders := "sheets"
- o Also defines geometric structure (torsion and curvature) as well as solvent exposure

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (4)

DSSP [2]

- o Hierarchical definition of patterns:
 - o H-bonds
 - o Turns & bridges
 - o α-helices, β-ladders, kinks & bulges
- o Structural features are defined independently
- o Single state assigned to each residue after resolving overlaps

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (5)

H-bonds

- o Described by electrostatic model only:

$$E = q_1 q_2 \left(\frac{1}{r(ON)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)} \right) \cdot f \quad [kcal/mol]$$

$r(AB)$ = distance between A and B

$q_1(C, O) = 0.42e$; $q_2(N, H) = 0.20e$; $f = 332$

- o Good bond energy: -3 kcal/mol
- o Threshold: $E < -0.5$ kcal/mol

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (6)

H-bonds [2]

- o Problems:
 - o Where are H placed??
 - o What about quantum-mechanics??
- o Ergo: no perfect answer, only good approximation!
- o Need also allow for coordinate errors!

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (7)

Turns

- o Basic pattern: single H-bond between residues i and $i+n$ " $(i, i+n)$ "
- o Assigned for H-bonds between CO(i) and CN($i+n$)
- o $n=3,4$ or 5
- o Residues between and including H-bonding ones are denoted as "T" unless part of helix

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (8)

Bridges

- o Two non-overlapping triplet patterns of H-bond types $(i-1,i,i+1)(j-1,j,j+1)$
- o Parallel bridge(i,j) =: $[(i-1,j) \& (j, i+1)]$ or $[(j-1,i) \& (i, j+1)]$ -> lower case
- o Antiparallel bridge (i,j) =: $[(i,j) \& (j,i)]$ or $[(i-1,j+1) \& (j-1,i+1)]$ -> upper case

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (9)

Helices

- o Minimal helix is defined as two n -turns:
 n -helix($i, i+2$) =: [n -turn($i-1$) & n -turn(i)]
- o Longer helices:
 - o 3_{10} -helix =: long 3-helix -> "G"
 - o α -helix =: long 4-helix -> "H"
 - o π -helix =: long 5-helix -> "I"

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (10)

Ladders & Sheets

- o Ladder =: one or more consecutive bridges of the same type
- o Sheet =: set of one or more ladders connected by shared residues
- o Single bridges (ladders of length 1) -> "B"
- o Extended ladders -> "E" (β -strands)

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (11)

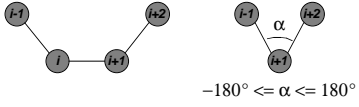
Irregularities

- o Implicit for helices:
 - o Overlapping helices
 - o missing H-bonds (kinks due to Pro ?)
- o Explicit for bridges:
 - o Bulge-linked ladder: 2 (perfect) ladders or bridges of the same type connected by one single residue on one strand and less than 5 residues on other
- o Both are marked as consecutive in one line summary!

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (12)

Geometric Structure

- o Bends ("S"):

$$\text{Bend}(i) =: \left[\Phi \left\{ \begin{matrix} x \\ y \\ z \end{matrix} \right\}_{C_{\alpha(i)}} - \begin{matrix} x \\ y \\ z \end{matrix} \right\}_{C_{\alpha(i-2)}} \left\{ \begin{matrix} x \\ y \\ z \end{matrix} \right\}_{C_{\alpha(i+2)}} - \begin{matrix} x \\ y \\ z \end{matrix} \right\}_{C_{\alpha(i)}} \right] > 70^\circ]$$
- o Chirality: sign of dihedral angle α


$-180^\circ \leq \alpha \leq 180^\circ$

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (13)

Geometric Structure [2]

- o Most helices are "right handed", i.e. of positive chirality
- o Most sheets are "left handed", i.e. of negative chirality
- o Exceptions: thermolysin (only "-" helices)

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (14)

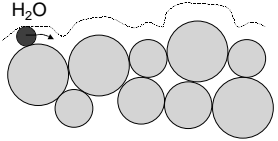
Structure Summary

- o Given as single line aligned with sequence
- o Priorities for overlaps:
H > B > E > G > I > T > S
- o Blanks: no H-bonds, low curvature
- o Most people forget to look at detailed description!!

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (15)

Solvent Exposure

- o "Rolling sphere" surface:



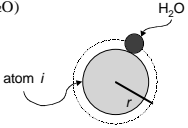
Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (16)

Solvent Exposure [2]

- o Mathematically:

$$\int f(x) d(x) \text{ with } f(x) = \begin{cases} 1 & \text{if water sphere at } x \text{ does not intersect other atoms} \\ 0 & \text{if it does} \end{cases}$$

$x =$ all points on sphere with radius r around atom i
 $r = r(\text{atom}) + r(\text{H}_2\text{O})$




Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (17)

Solvent Exposure [3]

- o Algorithm:
 - o Sum over 20, 80 or 320 approximately equal triangles
 - o use triangle centers as points x and area of triangle as weight
 - o Generate polyhedron iteratively:
 - o Start with icosahedron
 - o Dived each triangle into 4 by connecting midpoints and project the 3 new vertices onto sphere

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (18)

Solvent Exposure [4]



- o First iteration gives 20 points, next 80, next 320
- o Accuracy: within 4\AA^2 for 80, within 1\AA^2 for 320 points

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (19)

Solvent Exposure [5]

- o Average number of water:

$$W = \frac{\text{Surface Area}}{V(\text{H}_2\text{O})^{2/3}} \approx \frac{\text{Surface Area}}{10}$$
- o Since Area \cong Volume \cong ave. # of H_2O and $V(\text{H}_2\text{O}) = 30\text{\AA}^3$ ($30^{2/3} \approx 10$)
- o NOTE: solvent area differs between monomers and dimer!

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (20)

Overview

- o Secondary Structure Calculation ✓
 - o Kabsch & Sander
- o Protein Folding
- o Structure Prediction
 - o Homology Modeling
 - o Fold recognition
 - o Ab initio methods

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (21)

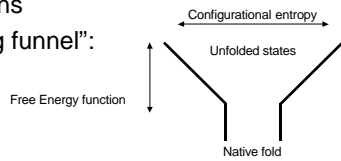
Protein Folding

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (22)

Protein Folding [2]

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



Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (23)

Protein Folding [3]

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (24)

Protein Folding [4]

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (25)

Protein Folding [5]

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (26)

Protein Folding [6]

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (27)

Overview

- o Secondary Structure Calculation ✓
 - o Kabsch & Sander
- o Protein Folding ✓
- o Structure Prediction
 - o Homology Modeling
 - o Fold recognition
 - o Ab initio methods

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (28)

Structure Prediction

Sequence ➡ Structure ➡ Function

- o 400 times more sequences than structures
- o Structure most conserved during evolution
- o Sequence alignment methods inadequate at low identity levels
- o Structure prediction “holy grail” of structure biology community

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (29)

Structure Prediction [2]

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (30)

Homology Modeling

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (31)

Alignment

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (32)

Choice of Core Structure

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (33)

Core Side Chains

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (34)

Building the loops

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (35)

Refinement of the model

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (36)

Reliability Estimation

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (37)

Why bother then?

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (38)

Fold Recognition

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

Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2001 (39)

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, 2001 (40)
