

Secondary Structure Calculation & Structure Classification

Introduction to Structural Bioinformatics

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

Overview

- Secondary Structure Calculation
 - Kabsch & Sander
- Structure Classification
 - SCOP
 - CATH

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

Secondary Structure Calc.

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

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

DSSP Algorithm

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

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

DSSP [2]

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

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

H-bonds

- 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
 $q_1(C,O) = 0.42e$; $q_2(N,H) = 0.20e$; $f = 332$
- Good bond energy: -3 kcal/mol
- Threshold: $E < -0.5$ kcal/mol

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

H-bonds [2]

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

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

Turns

- Basic pattern: single H-bond between residues i and $i+n$ (" $i, i+n$ ")
- Assigned for H-bonds between CO(i) and CN($i+n$)
- $n=3,4$ or 5
- 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, 2002 (8)

Bridges

- Two non-overlapping triplet patterns of H-bond types $(i-1,i,i+1)(j-1,j,j+1)$
- Parallel bridge(i,j) =: [($i-1,j$) & ($j, i+1$)] or [($j-1,i$) & ($i, j+1$)] -> lower case
- 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, 2002 (9)

Helices

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

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

Ladders & Sheets

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

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

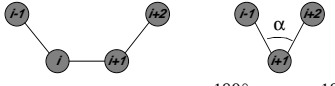
Irregularities

- Implicit for helices:
 - Overlapping helices
 - missing H-bonds (kinks due to Pro ?)
- Explicit for bridges:
 - 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
- Both are marked as consecutive in one line summary!

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

Geometric Structure

- Bends ("S"):

$$\text{Bend}(i) =: \left[\Phi \left\{ \begin{matrix} x \\ y \\ z \end{matrix} \right\}_{C\alpha(i)} - \begin{matrix} x \\ y \\ z \end{matrix} \right\}_{C\alpha(i-2)} \left\{ \begin{matrix} x \\ y \\ z \end{matrix} \right\}_{C\alpha(i+2)} - \begin{matrix} x \\ y \\ z \end{matrix} \right\}_{C\alpha(i)} \right] > 70^\circ]$$
- 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, 2002 (13)

Geometric Structure [2]

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

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

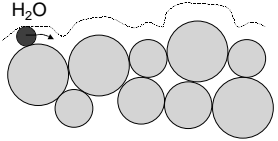
Structure Summary

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

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

Solvent Exposure

□ “Rolling sphere” surface:



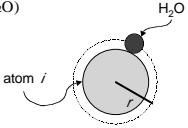
Introduction to Structural Bioinformatics • <http://www.sdsc.edu/~helgew/ISB/> • © Helge Weissig, 2002 (16)

Solvent Exposure [2]

□ 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, 2002 (17)

Solvent Exposure [3]

□ Algorithm:

- Sum over 20, 80 or 320 approximately equal triangles
- use triangle centers as points x and area of triangle as weight
- Generate polyhedron iteratively:
 - ◊ Start with icosahedron
 - ◊ 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, 2002 (18)

Solvent Exposure [4]

- ❑ First iteration gives 20 points, next 80, next 320
- ❑ 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, 2002 (19)

Solvent Exposure [5]

- ❑ Average number of water:

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

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

Evolution of Structure

- ❑ Evolutionarily related proteins retain memory of relationship through sequence, structure and function
- ❑ String sequence similarity considered sufficient evidence for common ancestry
- ❑ Close structural and functional similarity **together** also evidence common ancestry
- ❑ Structure is more conserved than sequence in distantly related proteins

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

Protein Folds

- ❑ Similarities in secondary structure element assembly
- ❑ Topological units of polypeptide chains
- ❑ Regularities arise from intrinsic physical and chemical properties
- ❑ Folds are the units of protein function, structure, and evolution
- ❑ Examples: propeller, horseshoe, TIM barrel...

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

SCOP

- ❑ Structure Classification of Proteins
- ❑ Based on evolutionary relationships
- ❑ Generated through visual comparison and inspection of automated structure alignments
- ❑ Provides links to coordinates of domains, images and sequence data
- ❑ <http://scop.mrc-lmb.cam.ac.uk/scop/>

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

SCOP Hierarchy

- ❑ Class
 - ❑ secondary element composition
 - ❑ All α , all β , α/β , $\alpha+\beta$, some others
- ❑ Folds
 - ❑ Common core structures
 - ❑ 138, 93, 97 & 184 respectively for each class
- ❑ Superfamily
 - ❑ Share common structure and function
- ❑ Family
 - ❑ Share clear common evolutionary origin

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

SCOP Hierarchy [2]

- Fold classification most difficult step
- Differences between mixed classes:
 - α/β
 - ◊ Principally single β -sheet with α -helices joining the individual strands
 - ◊ Two subclasses: β -sheet barrel surrounded by α -helices and planar β -sheet flanked on either side by α -helices
 - $\alpha+\beta$
 - ◊ α and β units largely separated
 - ◊ Antiparallel strands usually joined by hairpins
 - ◊ Small clusters of helices tightly packed against sheet

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

CATH

- Class, Architecture, Topology and Homologous Superfamily
- Also based on evolutionary relationships
- Automated generation with validation of ambiguities in assignments
- http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html

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

CATH Hierarchy

- Class
 - Determined by secondary structure composition and packing
 - mainly- α , mainly- β and α - β .
- Architecture
 - Description of orientation of secondary structures regardless of connectivity
 - Assigned manually

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

CATH Hierarchy [2]

- Topology
 - Regards both secondary structure orientation and connectivity
- Homologous Superfamily
 - Evolutionary grouping based on structure, sequence and/or functional similarity
 - Proteins are clustered into sequence families at different levels of sequence identity (35%, 60%, 95%, 100%)

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

CATH Update Strategy

1. Identify close relatives using pairwise sequence alignments
2. Detect distant relatives using sequence profiles and structure comparisons
3. Examine unclassified structures using both automatic and manual procedures to determine domain boundaries
4. Reiterate over steps 2 and 3
5. Manual assignment to existing or new architectures

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

Homework

- Pet protein project:
 - Add general information
 - Add Rasmol generated images
 - Add a section on SCOP and CATH classifications including images of the domain(s) in your structure
- Review lectures 1 and 2, specifically SCOP and CATH

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