Content

• Definition of the problem
• Possible approaches
• DSSP / PSI-BLAST
• Generalization
• Results
Definition of the problem

• Massive amounts of data from DNA sequencing available
• Protein sequences can be derived
• Knowing the tertiary structure is useful in drug design etc.
Primary structure

Amino Acids

Primary protein structure is sequence of a chain of amino acids.
Secondary structure
Tertiary structure
Quaternary structure

- Two and more protein molecules assembled
- Called protein subunits
Zur Anzeige wird der QuickTime™ Dekompressor „YUV420 codec“ benötigt.

What we are looking for...
Possible Approaches

- 2D prediction
- Ab initio structure prediction
- Homology based modeling
2D prediction

• Classical approach
• X-ray crystallography
• NMR techniques
Zur Anzeige wird der QuickTime™ Dekompressor „GIF“ benötigt.

Myoglobin
2D prediction limitations

- High experimental costs
- Large computing power required
- Depending on parameters: time, chemical environment, location in the metabolical pathway
Ab initio structure prediction

- Predicts 3D structure from 1D sequence data
- Ignores additional informations
- Target function (Free energy minimalization, molecular dynamics)
Levinthal paradox

• The number of possible protein sequences and its folding possibilities are astronomical

• Even real proteins cannot try all possible confirmations during folding

• Paradox: In nature, the folding is very quick and reliable
Further problems

• The physics behind structural stability not completely understood

• Primary sequence may not completely define tertiary structure
Our goal: to understand protein folding, protein aggregation, and related diseases

What are proteins and why do they "fold"? Proteins are biology's workhorses - its "nanomachines." Before proteins can carry out their biochemical function, they remarkably assemble themselves, or "fold." The process of protein folding, while critical and fundamental to virtually all of biology, remains a mystery. Moreover, perhaps not surprisingly, when proteins do not fold correctly (i.e. "misfold"), there can be serious effects, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's, and Parkinson's disease.

What does Folding@Home do? Folding@Home is a distributed computing project which studies protein simulations of villin.
Homology based modeling

• Reduce to a known problem

• Similar amino sequence yields similar protein structure
Templates

- Known structure of another protein
- Sequence alignment called threading
Threading

- Extension of sequence analysis
- Predicts the course of the protein backbone
- Aligns a sequence to the templates structure
- Scoring function
Folding

• 3D alignment of secondary structures

• In nature: folds more important than exact protein sequence
DSSP

• Application in Bioinformatics

• “Define Secondary Structure Of Proteins”

• Objective classification

• No prediction
Protein Data Bank
Application

• Input: 3D coordinates of atoms

• Output: secondary structure

• Usage: dssp [-na] [-v] pdb_file [dssp_file]
Codes

- **H**: alpha-helix
- **B**: beta-bridge
- **E**: beta-sheet, extended strand
- **G**: 3-turn-helix
- **I**: 5-turn-helix
- **S**: bend
•...;...1....;...2....;...3....;...4....;...5....;...6....;...7...
  .-- sequential resnumber, including chain breaks as extra residues
  |-- original PDB resname, not nec. sequential, may contain letters
  |  |-- amino acid sequence in one letter code
  |  |  |-- secondary structure summary based on columns 19-38
  |  |  |  | xxxxxxxxxxxxxxxxxxxxxxx recommend columns for secstruc details
  |  |  |  | .-- 3-turns/helix
  |  |  |  | .-- 4-turns/helix
  |  |  |  | .-- 5-turns/helix
  |  |  |  | .-- geometrical bend
  |  |  |  | .-- chirality
  |  |  |  | .-- beta bridge label
  |  |  |  | .-- beta bridge label
  |  |  |  | .-- beta bridge partner resnum
  |  |  |  | .-- beta bridge partner resnum
  |  |  |  | .-- beta sheet label
  |  |  |  | .-- solvent accessibility

# RESIDUE AA STRUCTURE BP1 BP2 ACC
| 35   47   I  E     +     0   0    236   48   R  E >  S- K   0  39C  97
| 36   48   R  E >  S- K   0  39C   97 (example from 1EST)
| 37   49   Q  T 3  S+     0   0   34
| 38   50   N  T 3  S+     0   0   34
| 39   51   W  E <  -KL  36  98C  6
Stereoscopic rendering
PSI-BLAST

- Application in Bioinformatics
- “Position-Specific Iterative BLAST”
- Iterative use of BLAST on a specific database
Application

• Input: Protein sequence
• Output: Protein family / sequence profile
• Usage: Web based
Protein family

• PSI-BLAST helps to find homologies
• Finds distant relatives
• Generates multiple alignments aka. sequence profile
Generalization

• Input: Target sequence
• Output: Template structure
Measurement

• Two values to be optimized:
  • Minimum $rms$ between matched points
  • Maximum number of matched residues
Scoring

- Determine the best alignment
- Minimize free energy
- Pair-interaction and contact capacity are important
Scoring function calibration

• Different factors weighted

• I.E. sequence, secondary structure, contact capacities...

• Substitution matrices, linear programming, supervised learning, support vector machines... help to improve the scoring function
Secondary structure prediction

• The better the secondary structure prediction, the better the tertiary structure prediction

• In special cases knowing secondary structures don’t help

• Requires more computing power
Assessment

• Simple calculation:

\[
\frac{\text{Number of matching residues}}{\text{Total number of residues}}
\]
Tertiary structure prediction

• Brute force
• Sequence methods
• Sequence and secondary structure
• Tertiary structure based methods
Brute force

\[ f(n, n) \sim \frac{(1 + \sqrt{2})^{2n+1}}{\sqrt{n}} \]
Sequence methods

• PSI-BLAST and fold libraries allow for efficient searching

• Other methods: FASTA, HMM...

• Include secondary structures for further information
Tertiary structure based methods

• Two main types:
  Profile method
  Threading approaches
Profile threading methods

• Many variations, based on different input data

• Example: 3D-PSSM
Other methods

• Optimal threading via exhaustive enumeration

• Branch-and-bound algorithms for fragment threading

• Heuristic search procedures based on simulations

• Dynamic programming / Double dynamic programming

• Recursive dynamic programming
Algorithms and support

• Major databases: ~40
• Tools: ~50
Evaluation

• Every two years: CASP

• Algorithms are tested against resolved, but unpublished proteins
Results

• CASP history teaches: Progress is being made, but humans can still outperform computers

• Not everything can be solved
CASP 3

• 43 targets

• 15 easy ones solved

• More than half of the 21 difficult ones solved by at least one method (completely automatic)