Protein Structure Prediction

BMI/CS 776

www.biostat.wisc.edu/~craven/776.html

Mark Craven

craven@biostat.wisc.edu

April 2002

The Protein Folding Problem

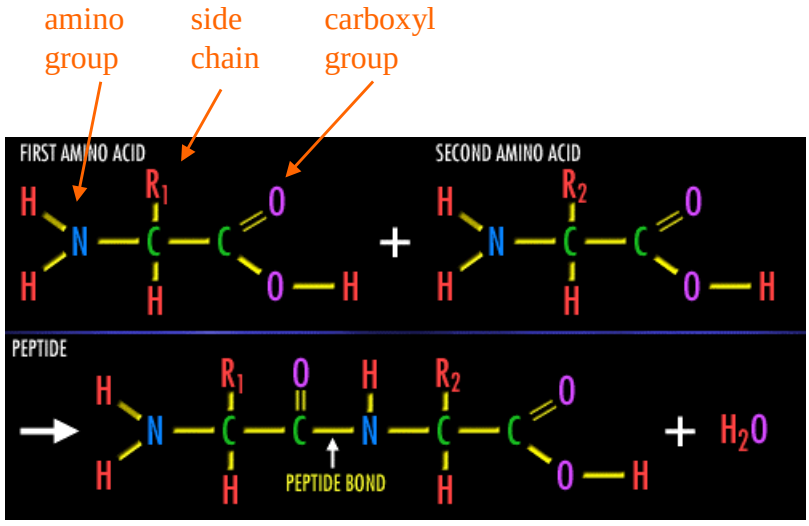
- we know that the function of a protein is determined by its 3D shape (*fold*, *conformation*)
- can we predict the 3D shape of a protein given only its amino-acid sequence?

- in general, NO!
- but methods that give us a *partial* description of the 3D structure are still helpful

Protein Architecture

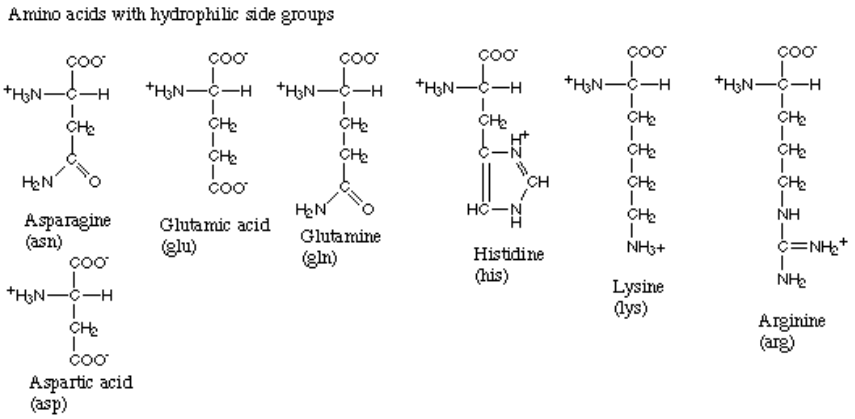
- proteins are polymers consisting of amino acids linked by *peptide* bonds
- each amino acid consists of
  - a central carbon atom
  - an amino group NH_2
  - a carboxyl group $COOH$
  - a side chain
- differences in side chains distinguish different amino acids

Peptide Bonds



Amino Acid Side Chains

- side chains vary in: shape, size, polarity, charge



What Determines Fold?

- in general, the amino-acid sequence of a protein determines the 3D shape of a protein [Anfinsen et al., 1950s]
- but some exceptions
  - all proteins can be denatured
  - some molecules have multiple conformations
  - some proteins get folding help from *chaperones*
  - *prions* can change the conformation of other proteins

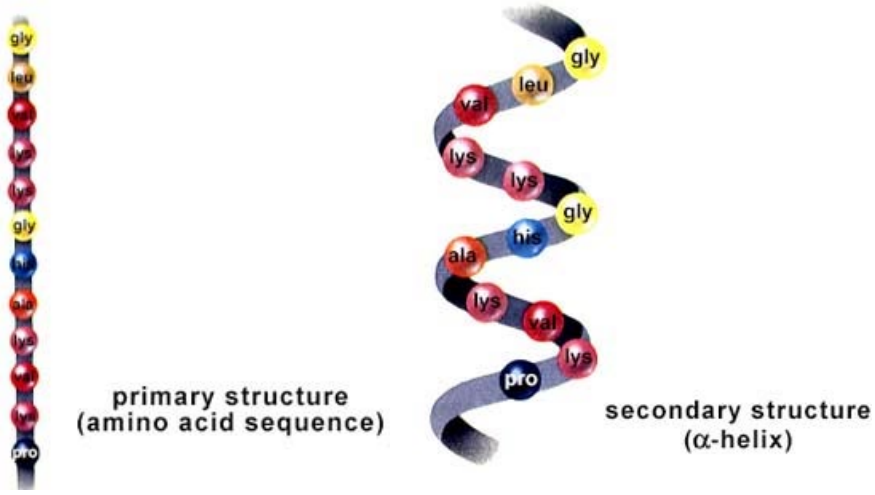
What Determines Fold?

- what physical properties of the protein determine its fold?
  - rigidity of backbone
  - interactions among amino acids, including
    - electrostatic interactions
    - van der Waals forces
    - volume constraints
    - hydrogen, disulfide bonds
  - interactions of amino acids with water

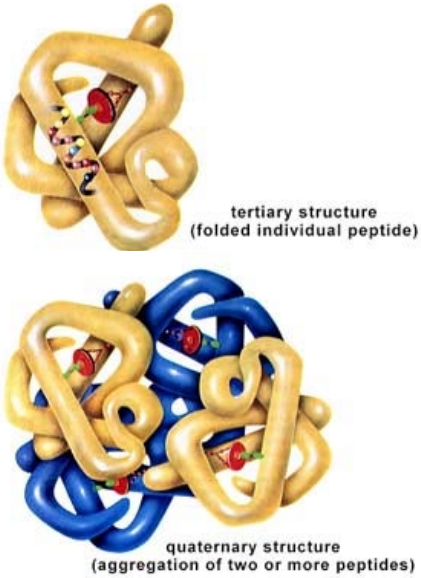
Levels of Description

- protein structure is often described at four different scales
  - primary structure
  - secondary structure
  - tertiary structure
  - quaternary structure
- don’t confuse these with Rost’s references to structure prediction in “1D”, “2D”, and “3D”

Levels of Description



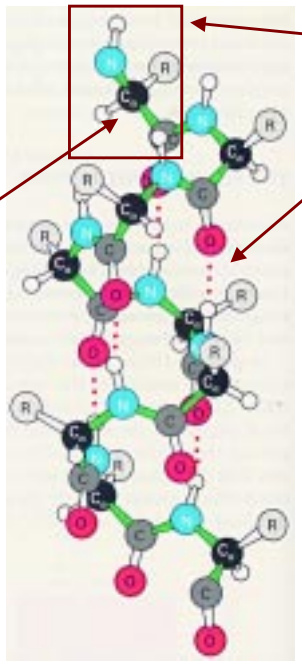
Levels of Description



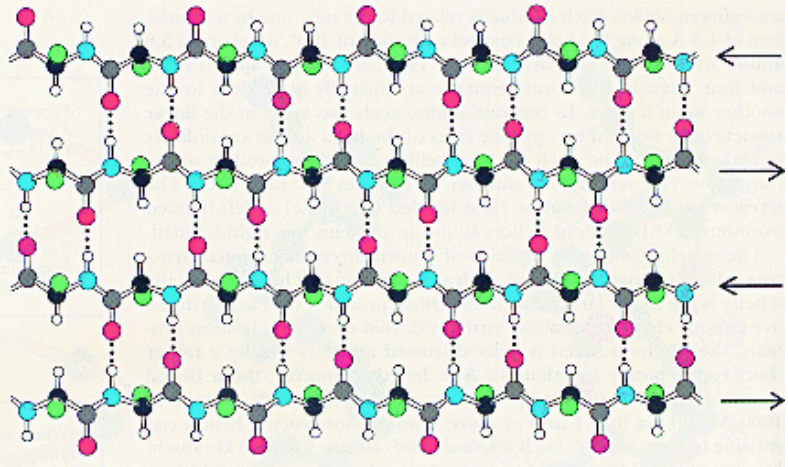
Secondary Structure

- secondary structure refers to certain common repeating structures
- it is a “local” description of structure
- 2 common secondary structures
  - α helices
  - β strands
- a 3rd category, called *coil* or *loop*, refers to everything else

α Helices



β Strands

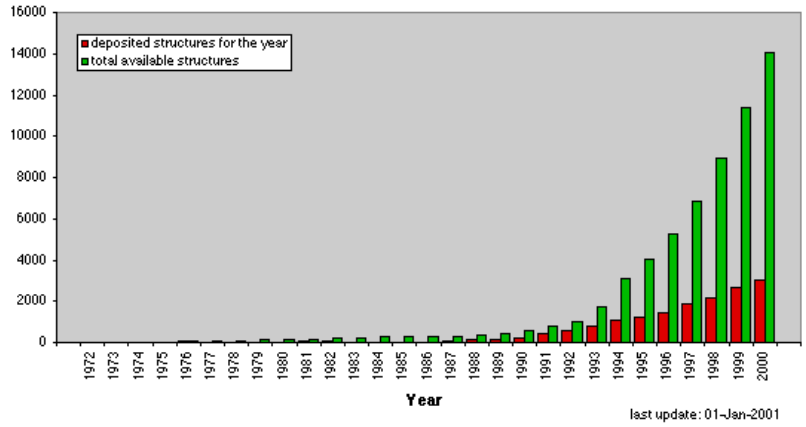


Ribbon Diagram Showing Secondary Structures

Determining Protein Structures

- protein structures can be determined experimentally (in most cases) by
  - x-ray crystallography
  - nuclear magnetic resonance (NMR)
- but this is very expensive and time-consuming
- can we predict structures by computational means instead?

PDB Content Growth



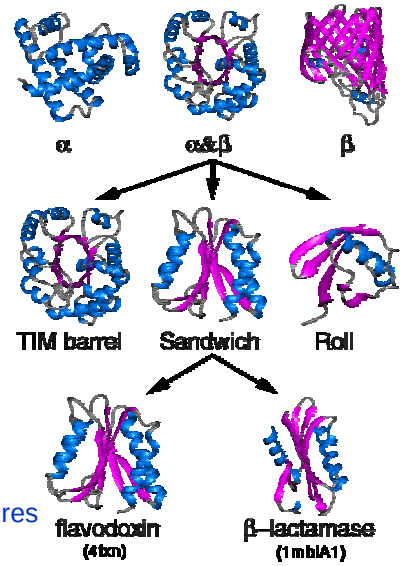
- the 4/12/01 release of SWISS-PROT, in contrast, has entries for 94,743 protein sequences

Top Levels of CATH Taxonomy

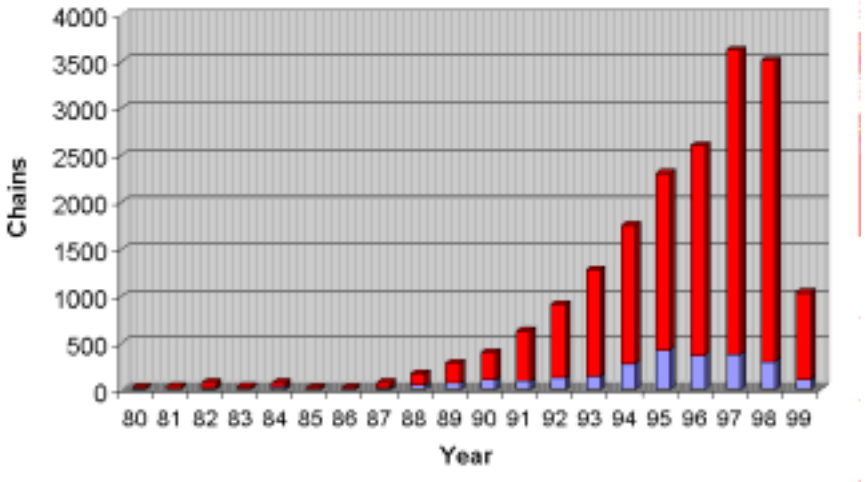
class:
defined by secondary structure composition

architecture:
defined by overall shape of domain structure

topology (fold):
defined by overall shape and connectivity of domain structures



PDB Growth in New Folds



- old folds are shown in red, new folds in blue

Approaches to Protein Structure Prediction

- prediction in 1D
  - secondary structure
  - solvent accessibility
  - transmembrane helices
- prediction in 2D
  - inter-residue/strand contacts
- prediction in 3D
  - homology modeling
  - fold recognition (e.g. via threading)
  - *ab initio* prediction (e.g. via molecular dynamics)

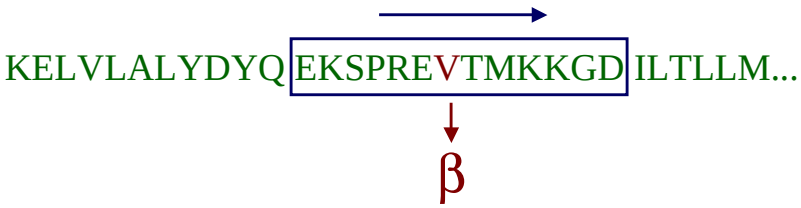
Secondary Structure Prediction

- given: an amino-acid sequence
- do:predict a secondary-structure state (α, β, coil) for each residue in the sequence

KELVLALYDYQEKSPREVTMKKGDILTLLM...
cccβββββccccccccccccβββββccccccβββββββ...

Secondary Structure Prediction

- one common approach:
  - make prediction for a given residue by considering a window of n (typically 13-21) neighboring residues
  - learn model that performs mapping from window of residues to secondary structure state

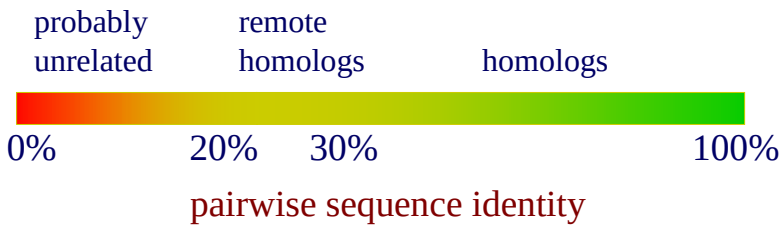


Homology Modeling

- observation: proteins with similar sequences tend to fold into similar structures
- given: a query sequence Q, database of protein structures
- do:
  - find protein P such that
    - structure of P is known
    - P has high sequence similarity to Q
  - return P's structure as an approximation to Q's structure

Homology Modeling

- most pairs of proteins with similar structure are remote homologs (< 25% sequence similarity)
- homology modeling usually doesn't work for remote homologs ; most pairs of proteins with < 25% sequence identity are unrelated



Protein Threading

- generalization of homology modeling
  - homology modeling: align sequence to sequence
  - threading: align sequence to *structure*
- key ideas
  - limited number of basic folds found in nature
  - amino acid preferences for different structural environments provides sufficient information to choose among folds

Components of a Threading Approach

- library of core fold templates
- objective function to evaluate any particular placement of a sequence in a core template
- method for searching over space of alignments between sequence and each core template
- method for choosing the best template given alignments

A Core Template

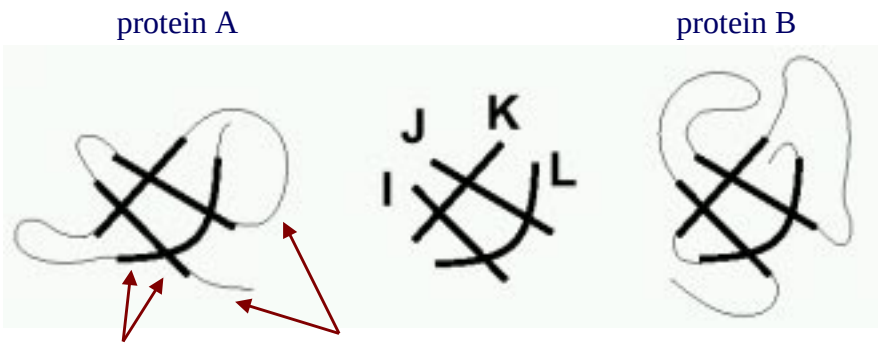


Figure from R. Lathrop et al, "Analysis and Algorithms for Protein Sequence-Structure Alignment" in Computational Methods in Molecular Biology, Salzberg et al. editors, 1998.

Objective Functions

- the objective function scores the sequence/structure compatibility between
  - sequence amino acids
  - their corresponding positions in the core template
- it takes into account factors such as
  - a.a. preferences for solvent accessibility
  - a.a. preferences for particular secondary structures
  - interactions among spatially neighboring a.a.'s

Core Template with Interactions

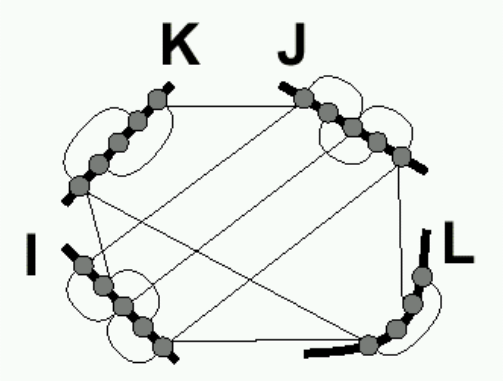


Figure from R. Lathrop et al, "Analysis and Algorithms for Protein Sequence-Structure Alignment"

- small circles represent amino acid positions
- thin lines indicate interactions represented in model

One Threading

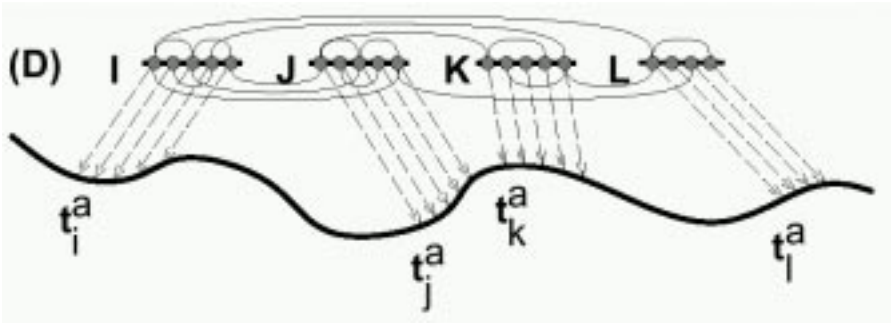


Figure from R. Lathrop et al, "Analysis and Algorithms for Protein Sequence-Structure Alignment"

- a threading can be represented as a vector $\vec{t}$, where each element indicates the index of the amino acid placed in the first position of each core segment

Possible Threadings

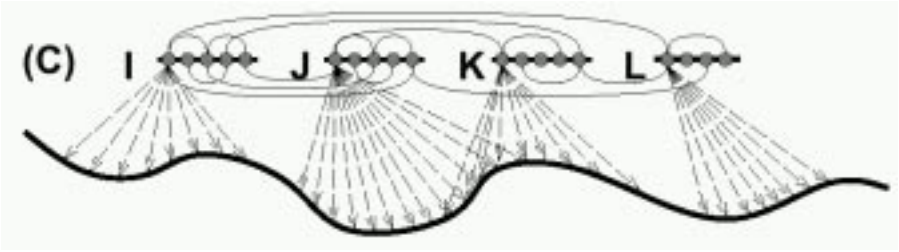


Figure from R. Lathrop et al, "Analysis and Algorithms for Protein Sequence-Structure Alignment"

- finding the optimal alignment is NP-hard in the general case where
  - there are variable length gaps between the core segments
  - the objective function includes interactions between neighboring amino acids

A Typical Pairwise Objective Function

$$f(\vec{t}) = \sum_{v \in V} f_{\text{vertex}}(v, \vec{t}) + \sum_{\{u,v\} \in E} f_{\text{edge}}(\{u,v\}, \vec{t}) + \sum_{\lambda \in \lambda_i} f_{loop}(\lambda_i, \vec{t})$$

$\vec{t}$ a vector characterizing a threading (each element indicates sequence position that starts each segment)

u, v amino acid positions in the core template

Searching the Space of Alignments

- higher-order interactions not allowed
  - dynamic programming
- higher-order interactions allowed
  - heuristic methods
    - fast
    - might not find the optimal alignment
  - exact methods (e.g. branch & bound)
    - will find the optimal alignment
    - might take exponential time

Branch and Bound Search

initialize Q with one entry representing the set of all threadings

repeat

$l \leftarrow$ set in Q with lowest lower bound

if l contains only 1 threading

return l

else

split l into smaller subsets

compute lower bound for each subset

put subsets in Q sorted by lower bound

Branch and Bound

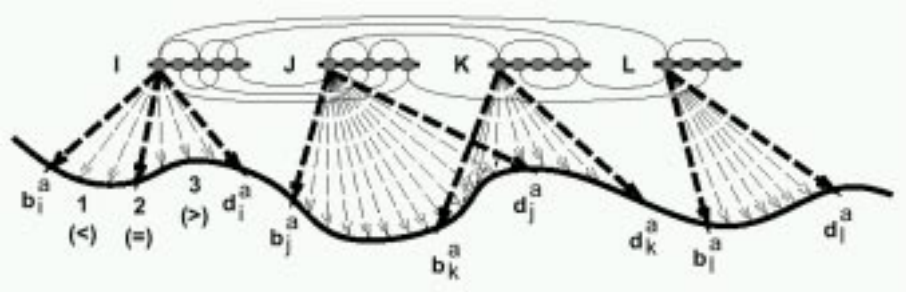


Figure from R. Lathrop et al, "Analysis and Algorithms for Protein Sequence-Structure Alignment"

A Lower Bound

- the general objective function with pairwise interactions is:

$$f(\vec{t}) = \underbrace{\sum_i g_1(i, t_i)}_{\text{scores for individual segments}} + \underbrace{\sum_i \sum_{j>i} g_2(i, j, t_i, t_j)}_{\text{scores for segment interactions}}$$

- the lower bound used by Lathrop et al. is: $\min_{\vec{t} \in T} f(\vec{t}) \geq$

$$\min_{\vec{t} \in T} \sum_i g_1(i, t_i) + \underbrace{g_2(i-1, i, t_{i-1}, t_i)}_{\text{interaction with preceding segment}} + \underbrace{\min_{\vec{u} \in T} \sum_{|j-i|>1} \frac{1}{2} g_2(i, j, t_i, u_j)}_{\text{best case interaction with other segments}} |$$