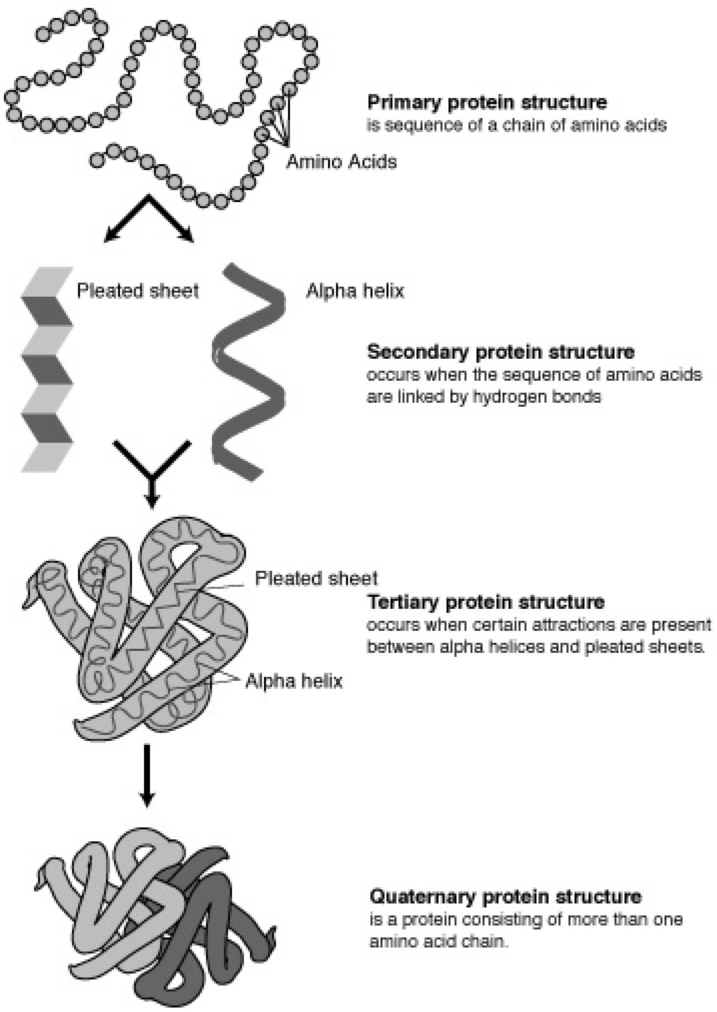
CSE 397-497:
Computational Issues in Molecular Biology

Lecture 19

Spring 2004

Protein structure


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

http://crystal.uah.edu/~carter/protein/protein.htm

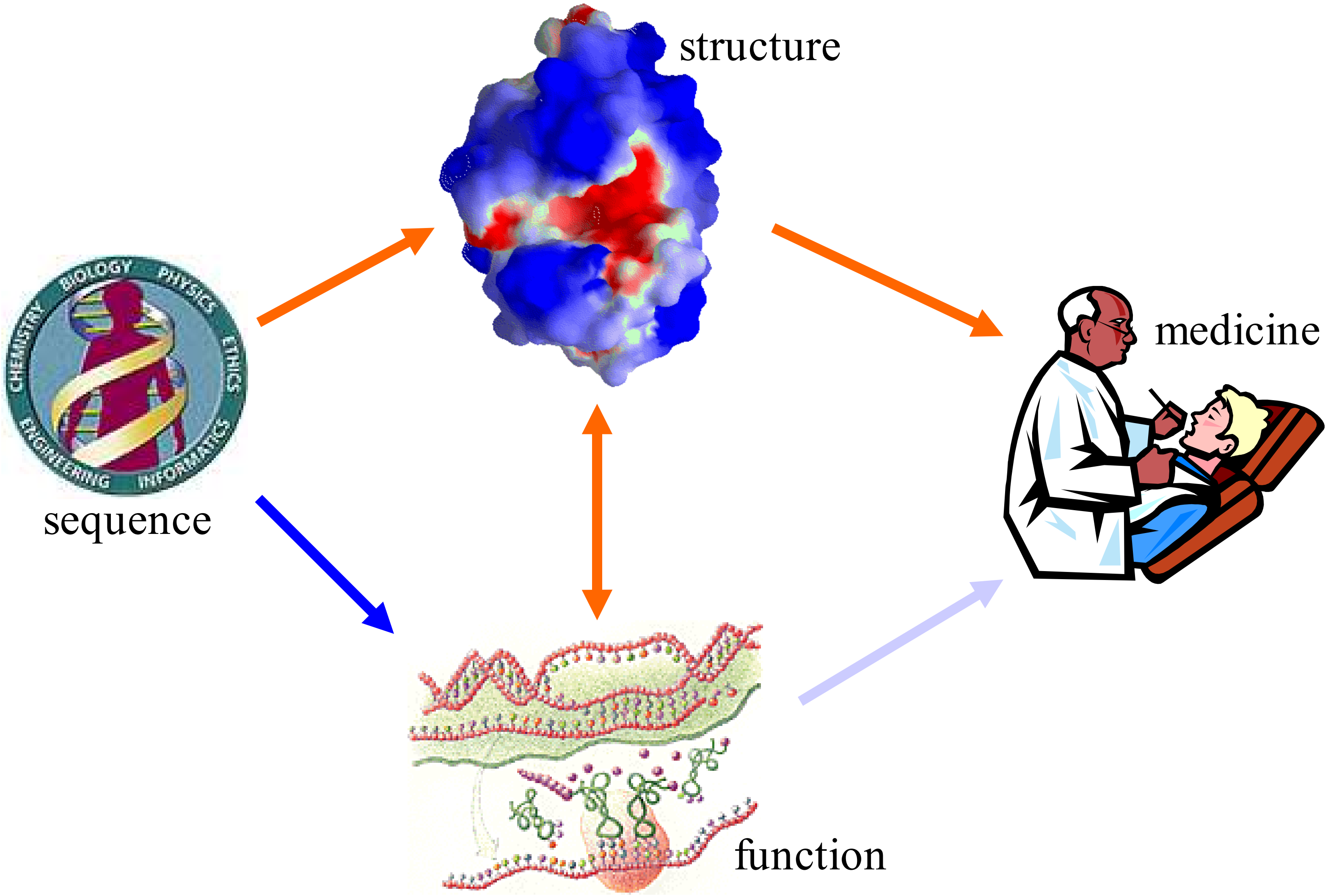
Primary structure of protein is determined by number and order of amino acids within polypeptide chain.

Protein's *secondary structure* is defined as local conformation of its backbone, which consists of molecules that make up an amino acid's frame excluding side chains. Two common motifs include beta-pleated sheets and alpha helices.

Tertiary structure is formed when attractions of side chains and secondary structure combine to form distinct 3-dimensional structure. This gives protein its specific function.

Sometimes distinct proteins must combine to form correct 3-dimensional structure for a particular protein to function properly. E.g., hemoglobin is made of four similar proteins that combine to form its *quaternary structure*.

Sequence → structure → function



structure

sequence

medicine

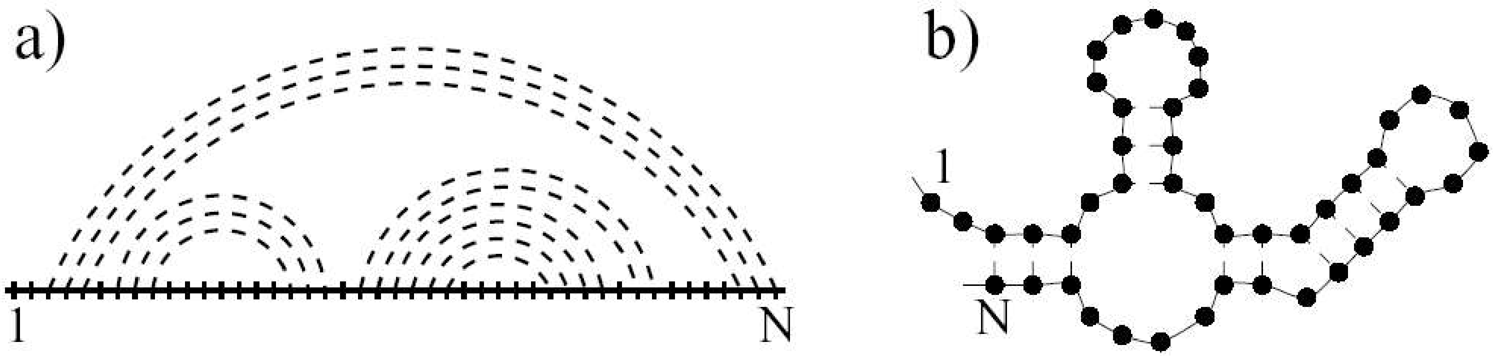
function

http://www.bioinformatics.uwaterloo.ca/~j3xu/CS882/CS882-ProteinStructurePrediction.ppt

Recall from last time ...

Situation for RNA structure (somewhat similar):

- Tertiary structure is difficult to model and compute.
- Determining secondary structure is more amendable to a solution. While only an approximation, it gives good hints.
- No knots ⇒ planar graph.

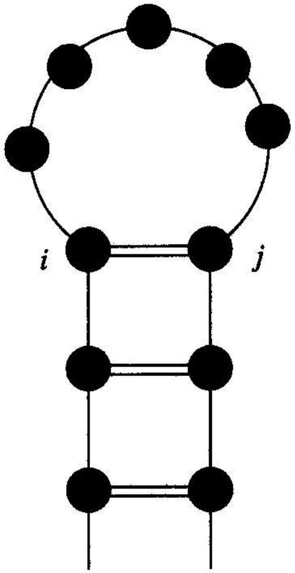


Solved using dynamic programming in $O(n^3)$ time.

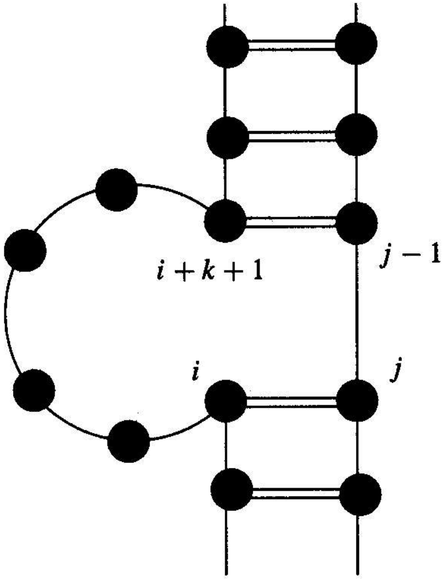
http://matisse.ucsd.edu/~hwa/pub/goletter.pdf

Recall from last time ...

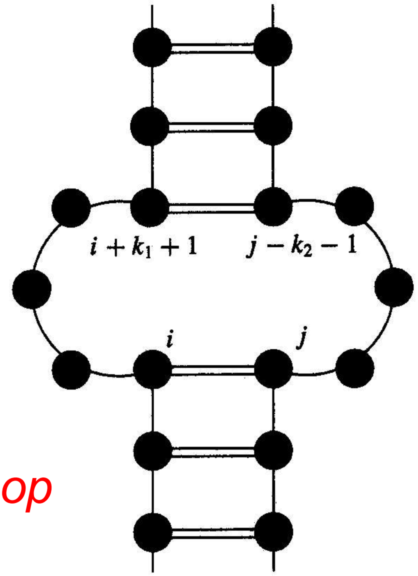
Better results can be obtained by modeling loops. This problem is also solvable in $O(n^3)$ time using some tricks.



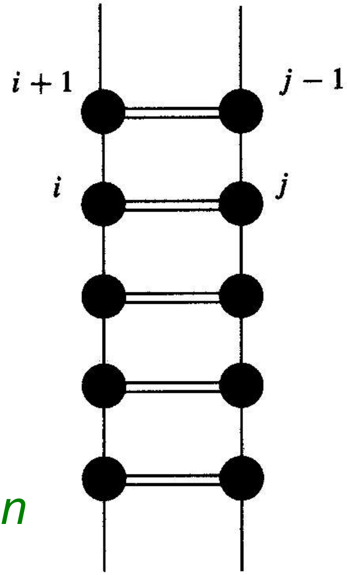
Hairpin loop



Bulge



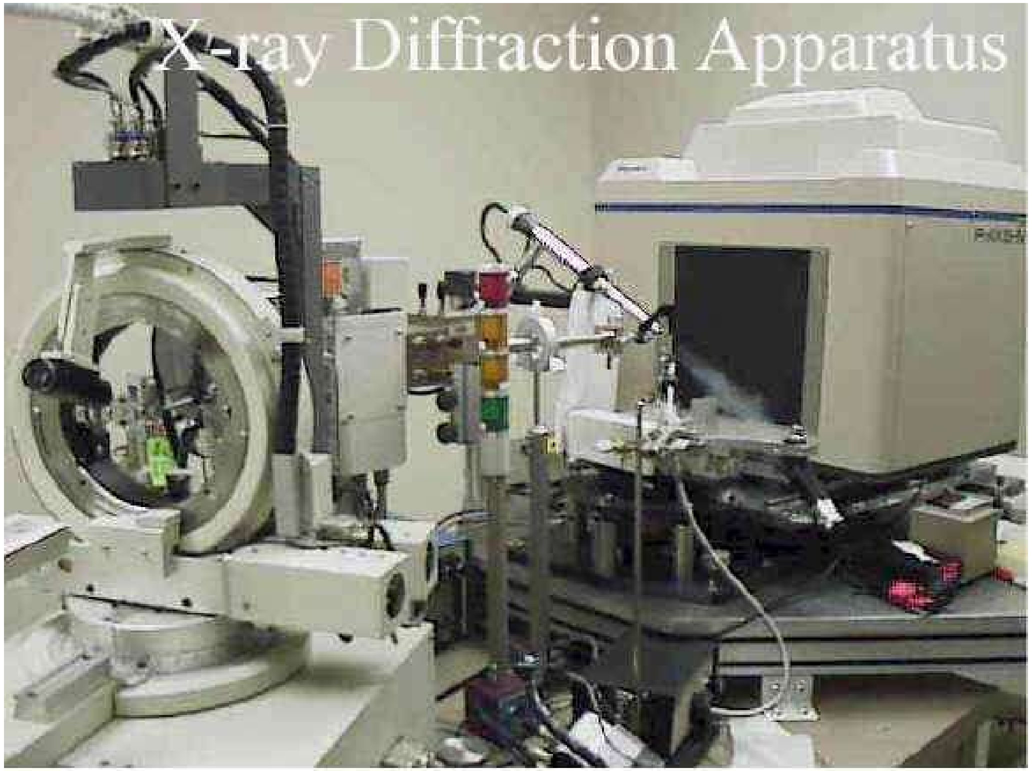
Interior loop

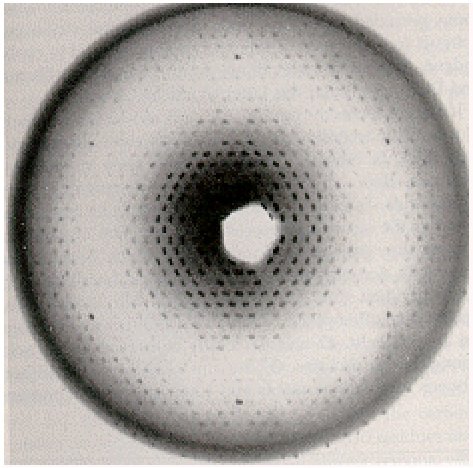


Helical region

How is true 3-D structure determined?

- As of today, must be determined experimentally.
- Techniques include x-ray crystallography and NMR.

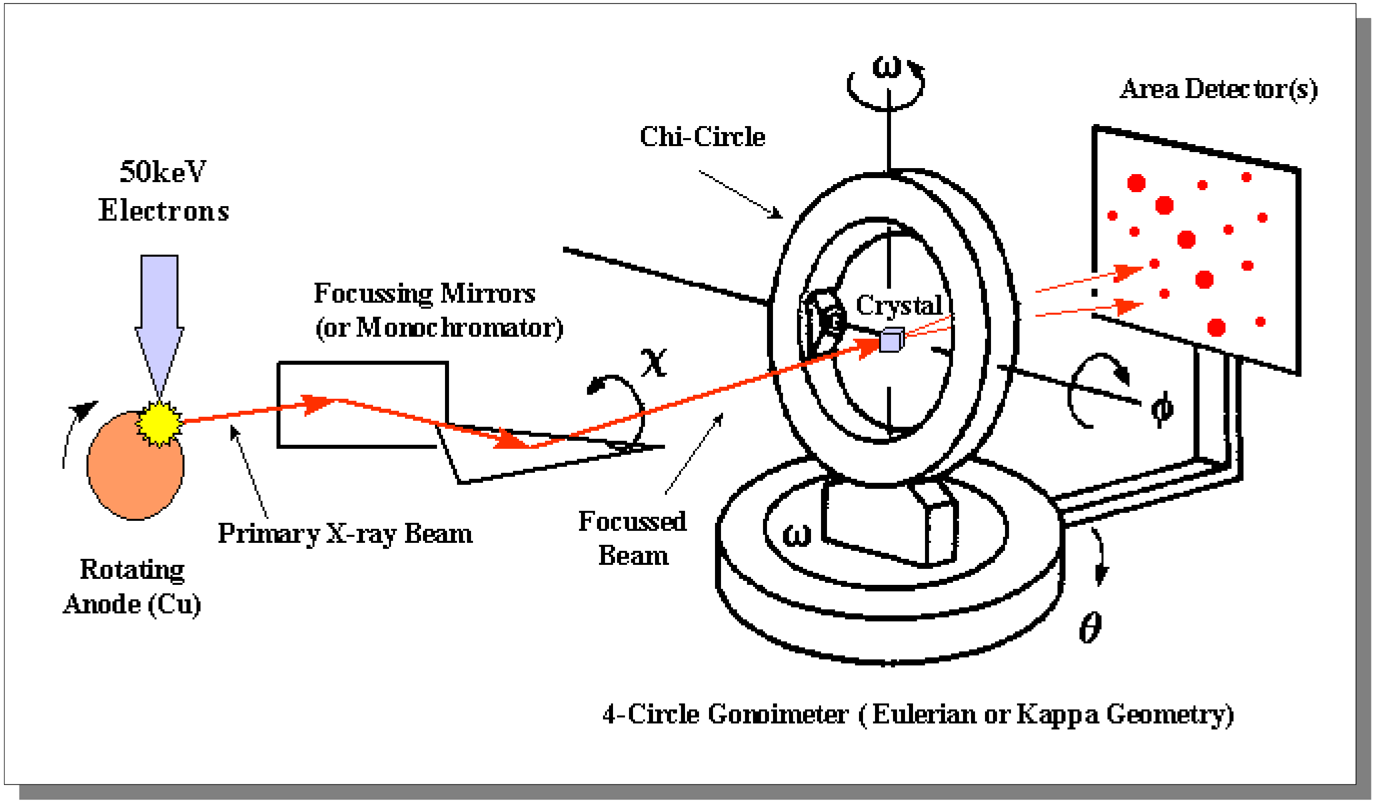


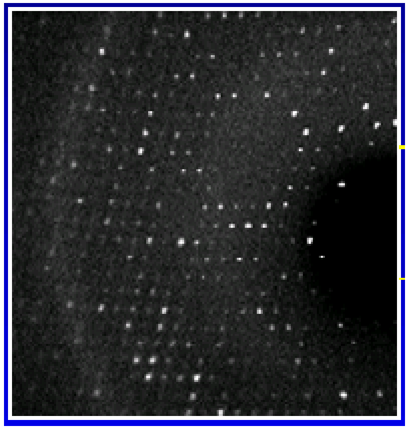


diffraction pattern

http://crystal.uah.edu/~carter/protein/xray.htm

X-ray crystallography

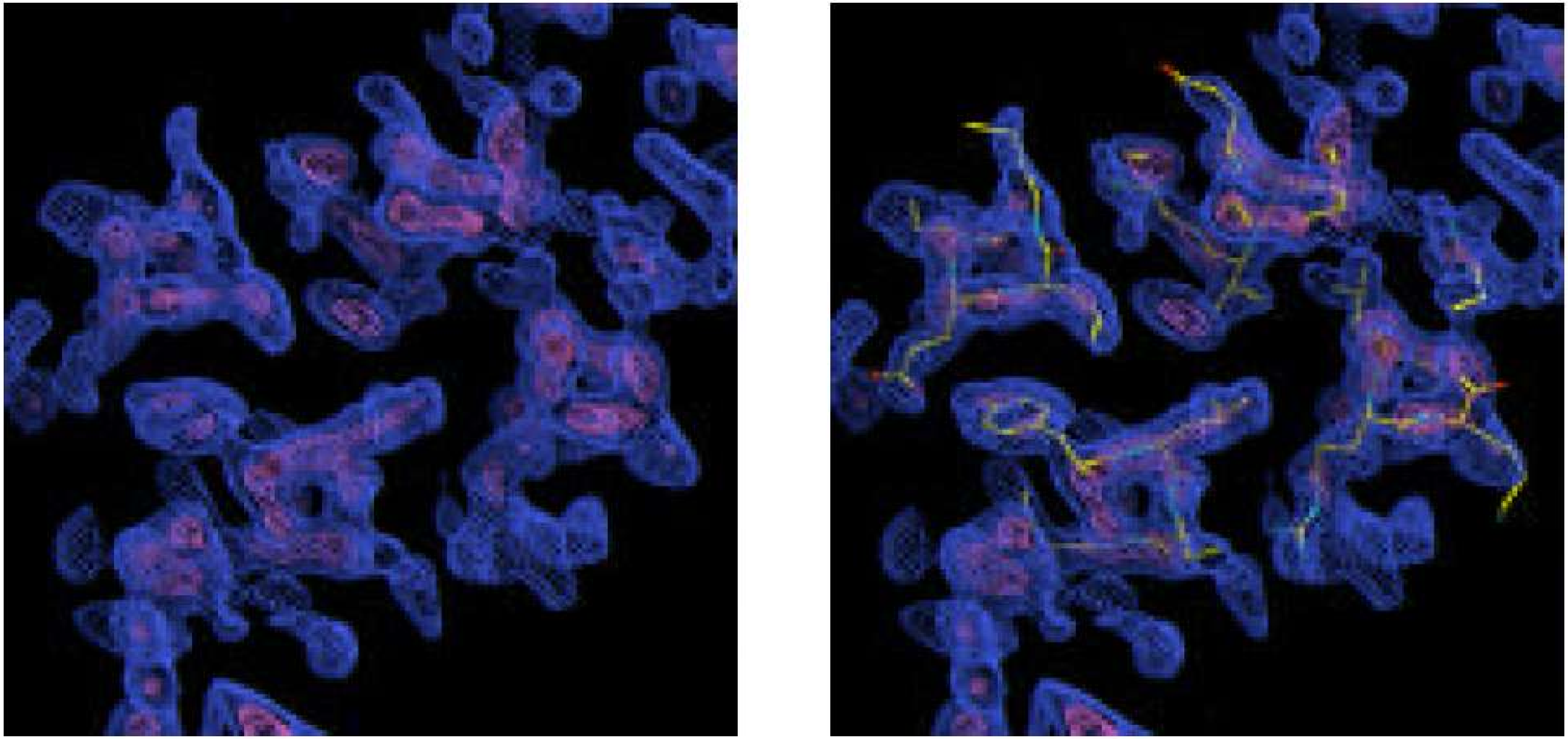




A diffraction pattern: the white spots are the reflections.

http://www-structure.llnl.gov/Xray/xrayequipment.htm

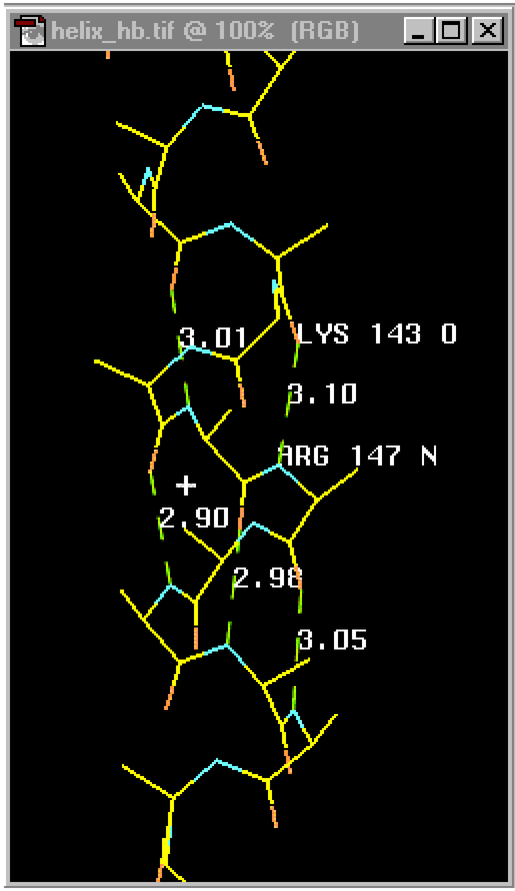
From electron density map to structure



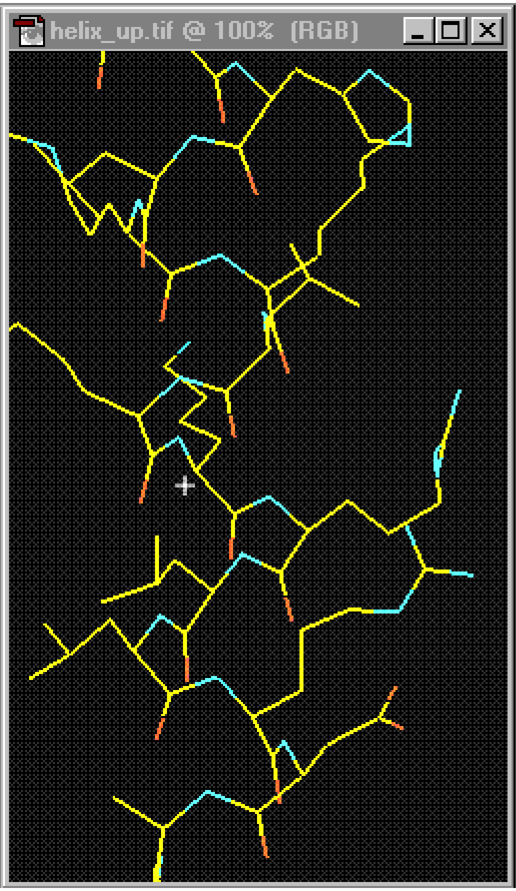
Experimental electron density map and model fitting

http://www.ib3.gmu.edu/vaisman/csi731/lec04f02.pdf

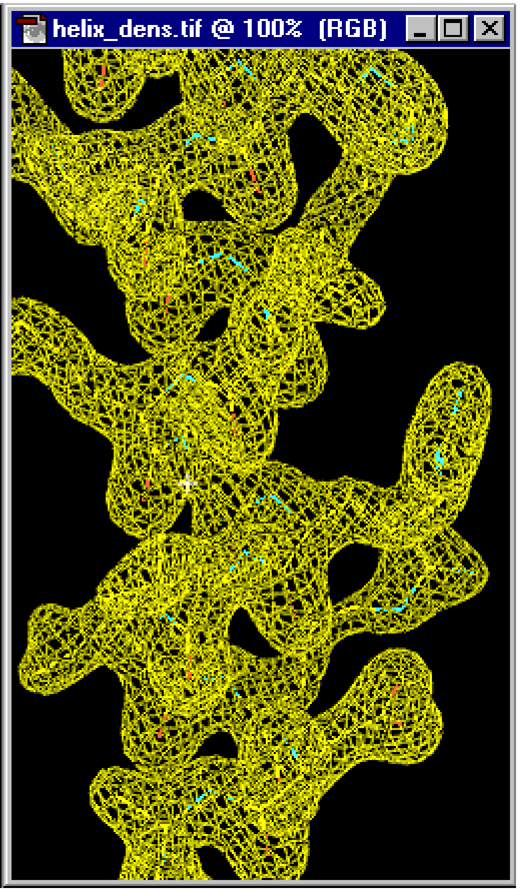
Protein structure determination



backbone



… w/ side chains



electron density map

http://www-structure.llnl.gov/Xray/tutorial/protein_structure.htm

Two common protein secondary structures

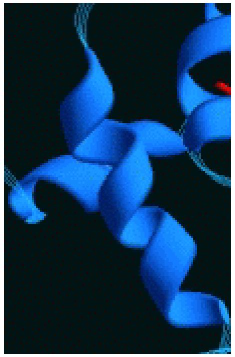
Alpha Helix

R groups of amino acids all extend to outside.

Helix makes a complete turn every 3.6 amino acids.

Helix is right-handed; it twists in clockwise direction.

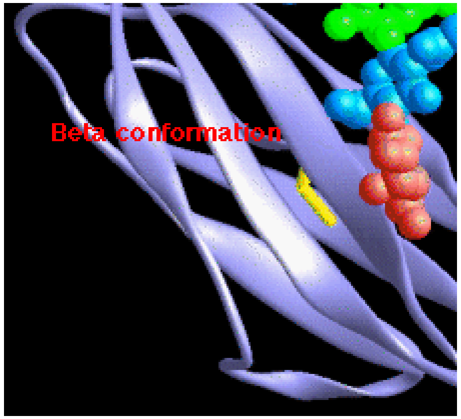
Carbonyl group (-C=O) of each peptide bond extends parallel to axis of helix and points directly at -N-H group of peptide bond 4 amino acids below it in helix. A hydrogen bond forms between them [-N-H·····O=C-].



Beta Conformation

Consists of pairs of chains lying side-by-side and stabilized by hydrogen bonds between carbonyl oxygen atom on one chain and -NH group on adjacent chain.

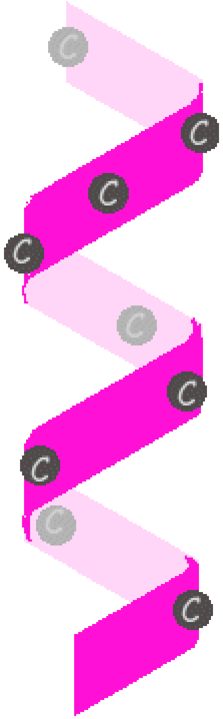
Chains are often "anti-parallel"; N-terminal to C-terminal direction of one being reverse of other.



http://www.rothamsted.bbsrc.ac.uk/notebook/courses/guide/protalpha.htm

Alpha helix

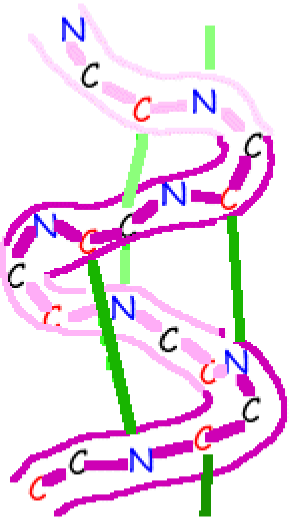
Alpha-helix (also written α-helix) is rod-like structure stabilized by hydrogen bonds between CO and NH groups of main chain.



1: Helical Ribbon

Ribbon representation of right-handed alpha-helix with only the alpha carbons represented.

Examining backbone structure, note that alpha carbons spaced three and four in linear sequence are actually quite close together in helix structure. The hydrogen bonds are shown in green; all main chain CO and NH groups are hydrogen bonded. This structure is quite sturdy.



2: Backbone:

N Nitrogen

C Alpha Carbon

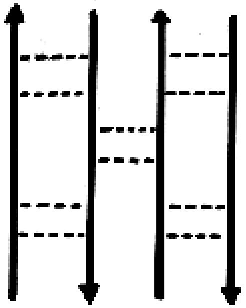
C Carboxyl Carbon

🟩 Hydrogen bond

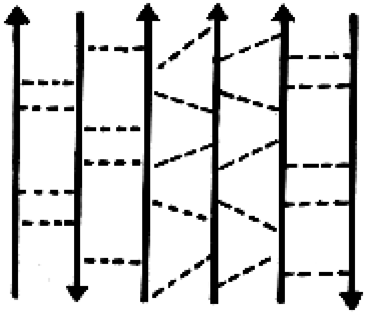
http://www.rothamsted.bbsrc.ac.uk/notebook/courses/guide/protalpha.htm

Beta sheet

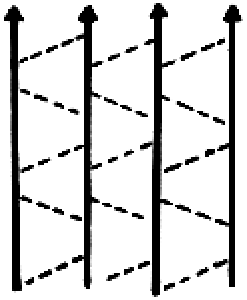
Antiparallel beta-sheet



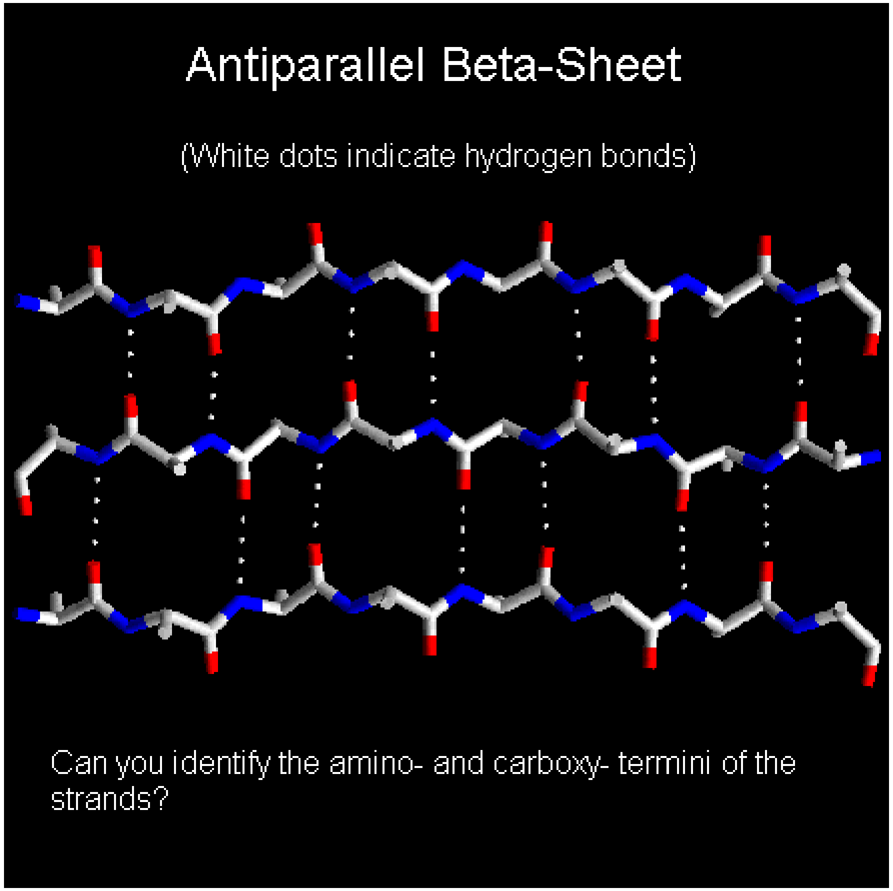
The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.



Mixed beta-sheet



Parallel beta-sheet



Antiparallel Beta-Sheet

(White dots indicate hydrogen bonds)

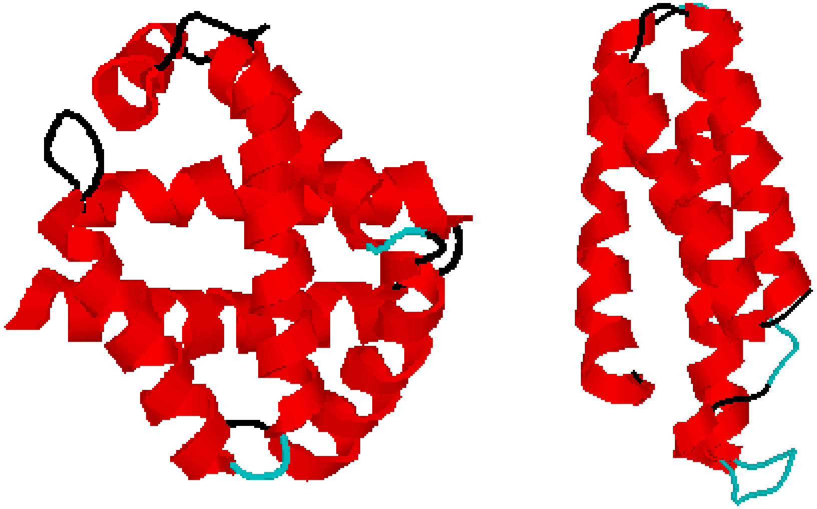
Can you identify the amino- and carboxy- termini of the strands?

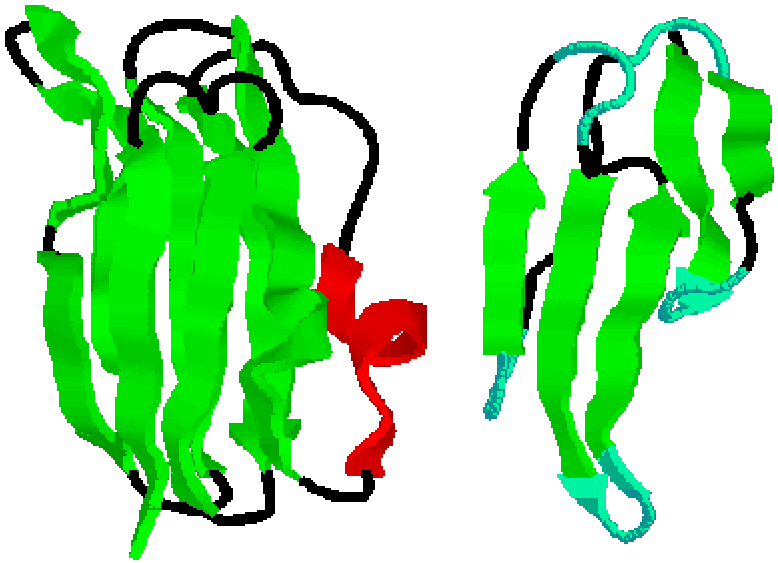
http://www.cryst.bbk.ac.uk/PPS2/course/section3/sheet.html

Protein structure

Some proteins are made up of mostly alpha helicies.

Both marine bloodworm hemoglobin (left) and E. coli cytochrome B562 (right) are composed of mostly alpha helicies. The 4 helix bundle of the cytochrome is a common motif.





Some are mostly beta sheet.

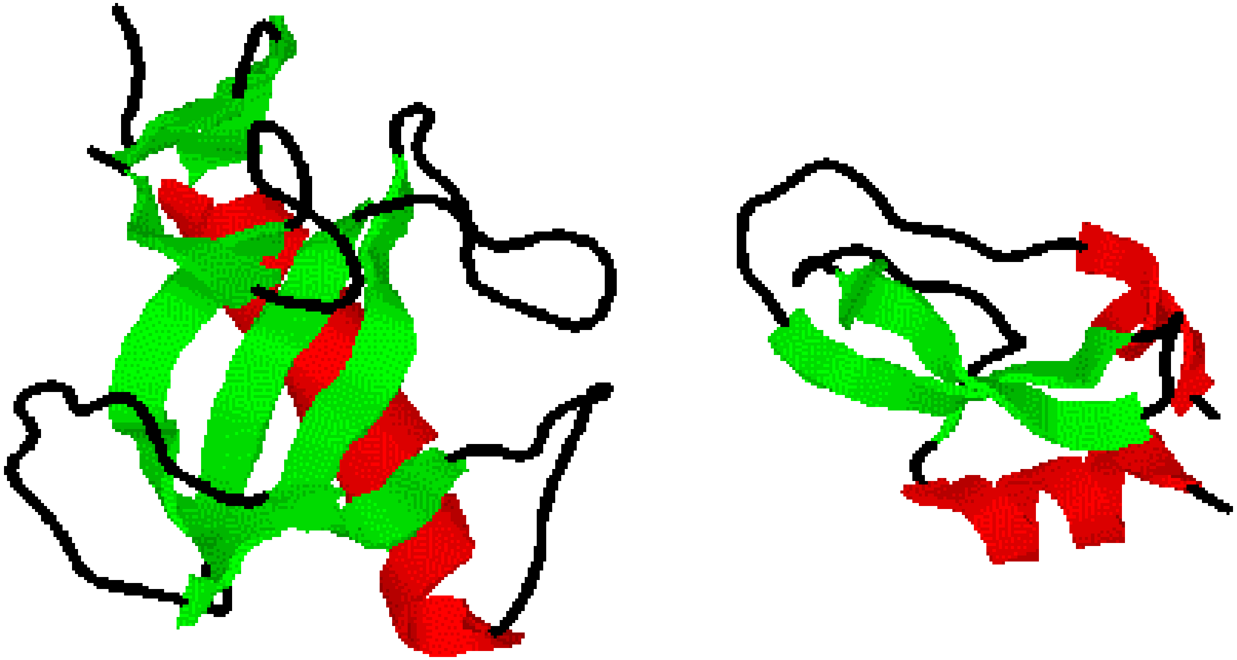
The green alga plastocyanin (left) and sea snake neurotoxin (right) are mostly beta sheets.

Red = alpha helix
Green = beta sheet
Black = misc. loops

http://bmbiris.bmb.uga.edu/wampler/tutorial/prot3.html

Protein structure

Many proteins are a mix of alpha helicies and beta sheets.



Two simple proteins with a mix of 2^o components:
ribonuclease T_1 (left) and pancreatic trypsin inhibitor (right).

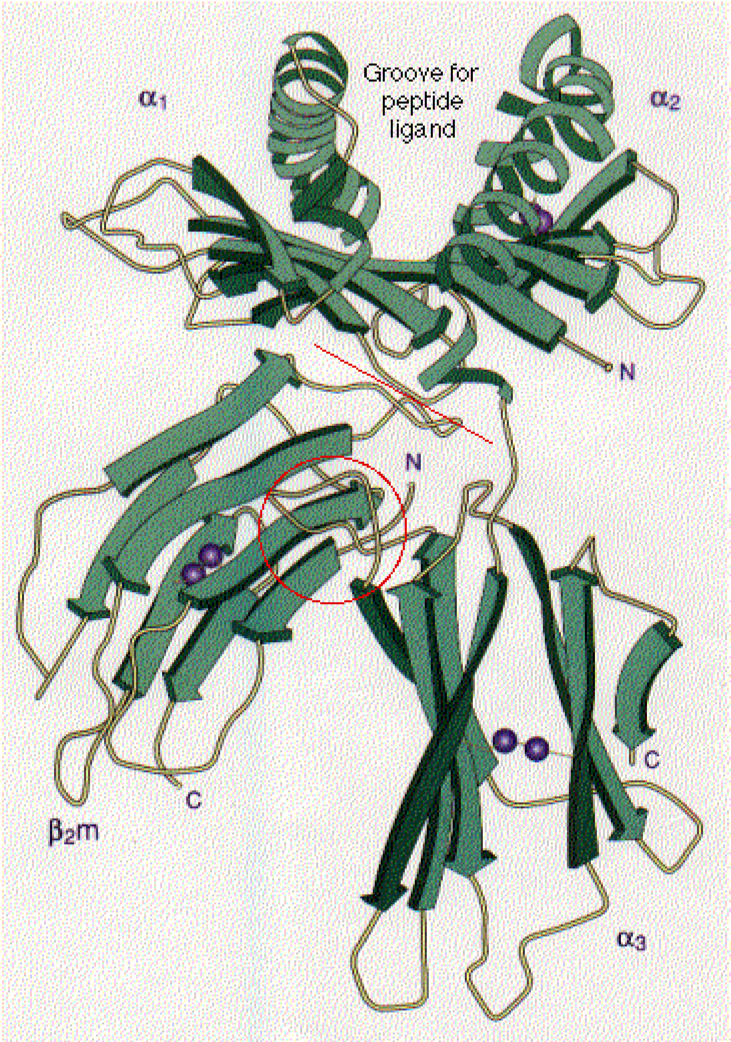
http://bmbiris.bmb.uga.edu/wampler/tutorial/prot3.html

Protein Domains

Tertiary structure of many proteins is built from several domains. Often each has a separate function to perform, such as:

- binding a small ligand (e.g., a peptide in the molecule shown here)
- spanning the plasma membrane (transmembrane proteins)
- containing the catalytic site (enzymes)
- DNA-binding (in transcription factors)
- providing a surface to bind specifically to another protein.

In some cases, each domain is encoded by a separate exon in the gene. The histocompatibility molecule shown here has three domains: α1, α2, and α3 are each encoded by its own exon.



http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/T/TertiaryStructure.html

Moving towards protein structure prediction ...

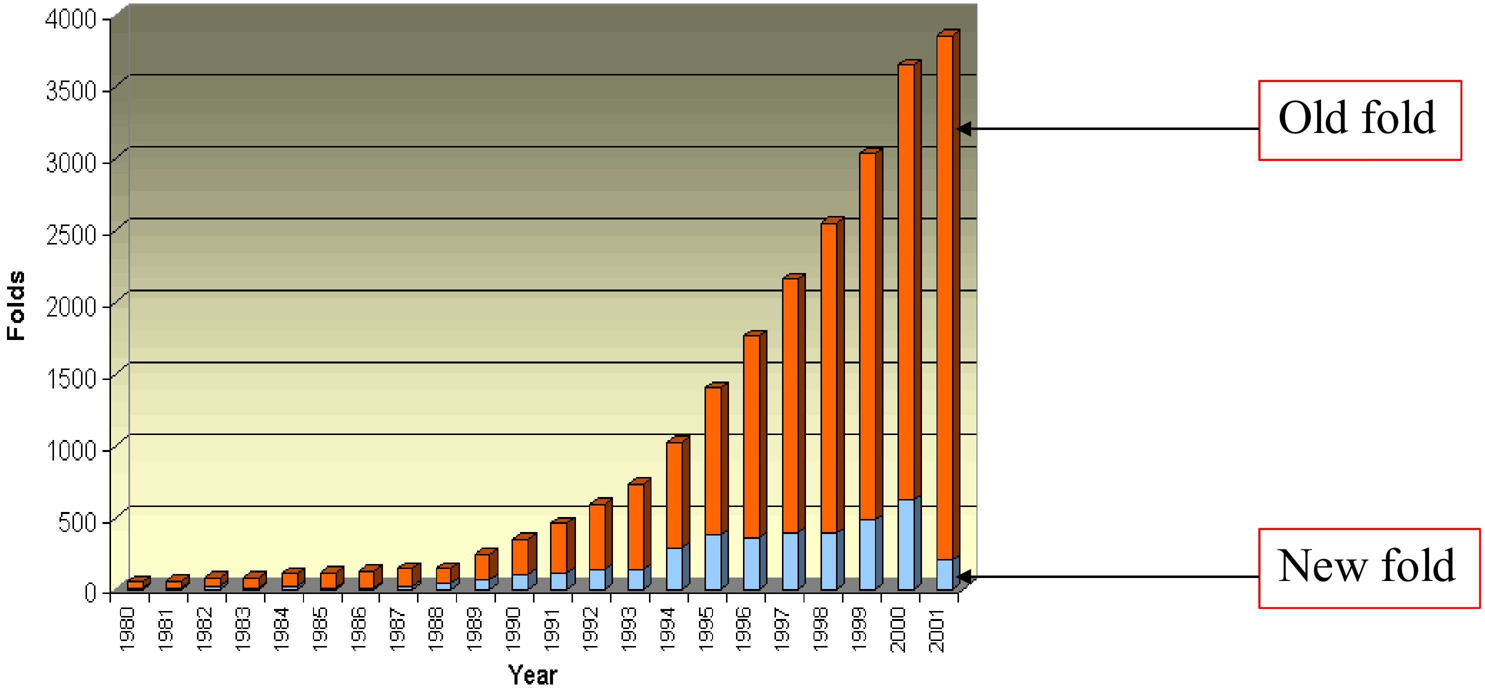
- Most important information seems to be contained in alpha helices and beta sheets (which form *core*), not in loops.
- Given amino acid sequence, we want to determine locations of helices, sheets, and loops, and their arrangements.

How to do this?

- Experimental techniques are expensive and time-consuming.
- Exhaustive enumeration at molecular level (taking structure with smallest free energy)? Nah ...
- As of 2002, the NIH protein structure database contained approximately 15,000 entries. Hmm ...
- Idea: given sequence, see if it could fit a known structure.

This is known as *protein threading*.

PDB new vs. old folding growth



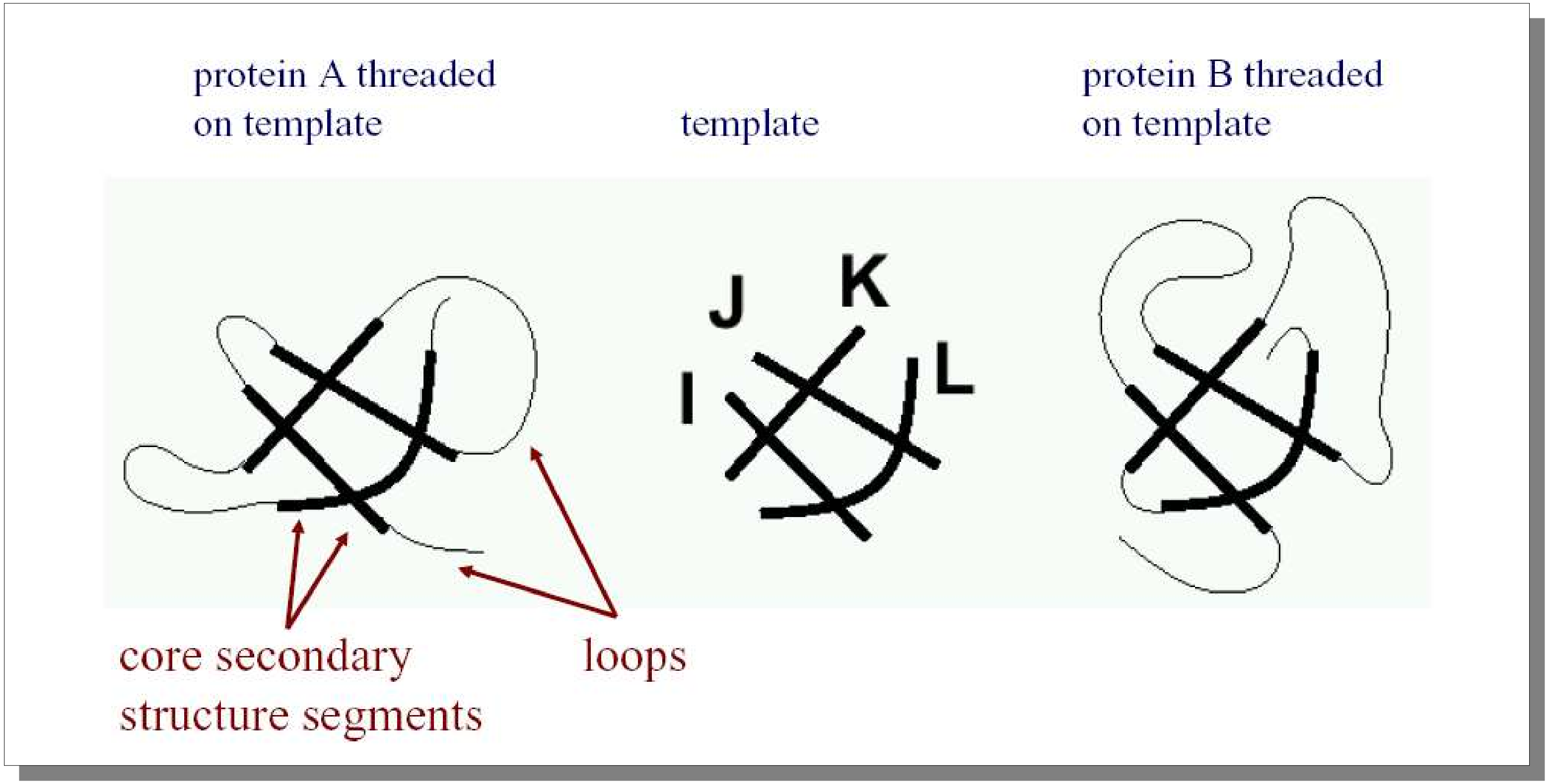
- Number of unique folds in nature is fairly small (possibly a few thousands).
- 90% of new structures submitted to PDB in the past three years have similar structural folds in PDB.

http://www.bioinformatics.uwaterloo.ca/~j3xu/CS882/CS882-ProteinStructurePrediction.ppt

Protein threading

Somewhat similar to sequence alignment we studied earlier:

- homology modeling: align sequence to sequence,
- threading: align sequence to structure (templates).

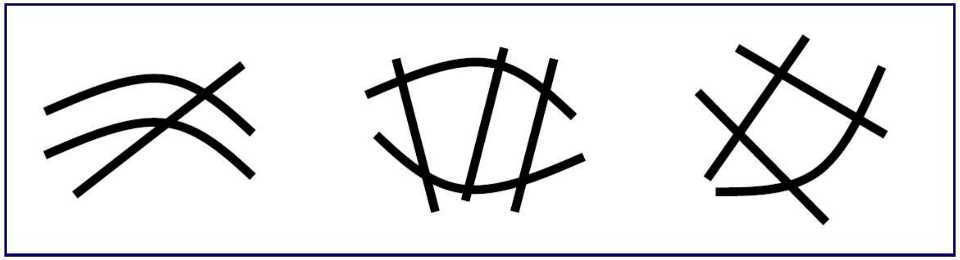

protein A threaded on template
template
protein B threaded on template
J
K
I
L
core secondary structure segments
loops

http://www.biostat.wisc.edu/bmi576/lecture16.pdf   http://www.ics.uci.edu/~rickl/publications/1998-salzberg-chapter.pdf

The protein threading problem

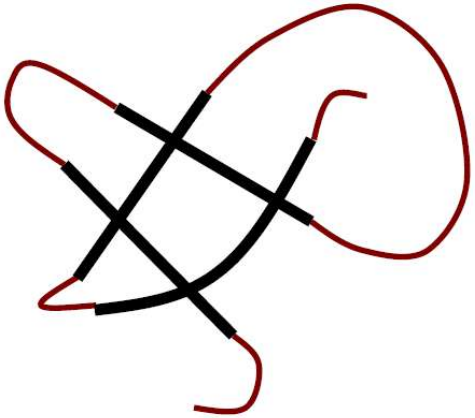
Given:

- new protein sequence, and
- library of templates:



Find:

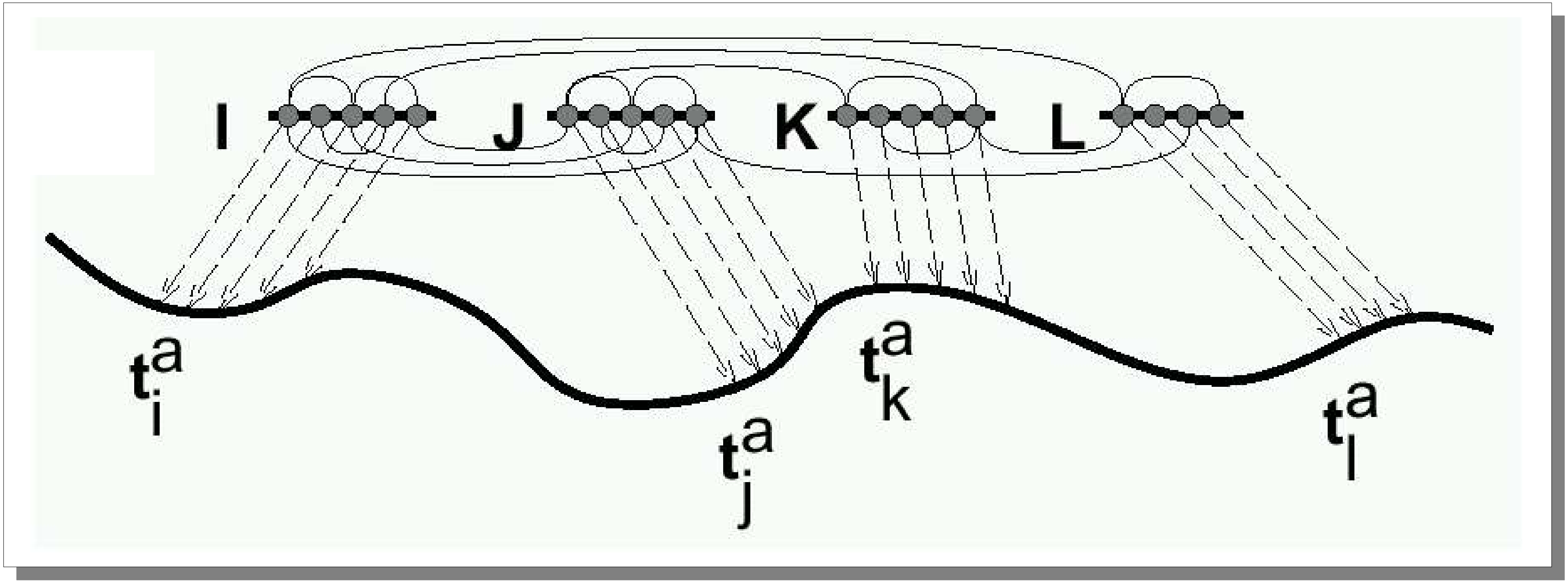
- best alignment of sequence to some template.



http://www.biostat.wisc.edu/bmi576/lecture16.pdf

The protein threading problem

One possible threading (note non-local interactions):



http://www.dcs.kcl.ac.uk/teaching/units/csmacmb/DOC/lecture17b.pdf
http://www.ics.uci.edu/~rickl/publications/1998-salzberg-chapter.pdf

The protein threading problem

Input:

1. Protein sequence A with n amino acids a_i.
2. Core structural model C, with m core segments C_i. Also:

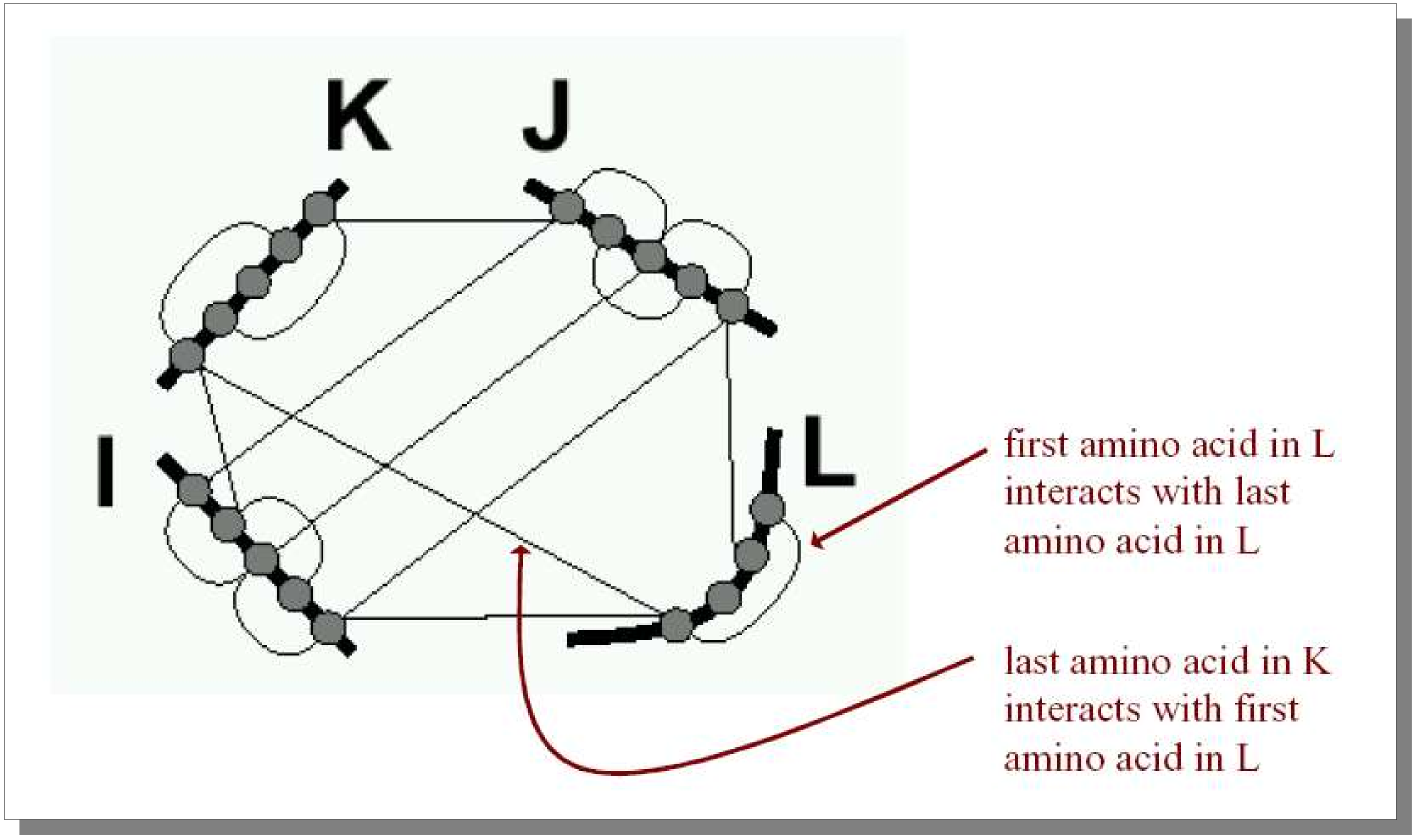
   (a) Length c_i of each core segment.

   (b) Core segments C_i and C_{i+1} are connected by loop λ_i for which we know max ($lmax_i$) and min ($lmin_i$) lengths.

   (c) Structural environment for each amino acid position.
3. Scoring function $f(T)$ to evaluate each threading T.

Output: set of integers $T = \{t_1, t_2, ..., t_m\}$ such that value of t_i indicates which amino acid from A occupies first position in core segment i.

Core templates with interactions



first amino acid in L interacts with last amino acid in L

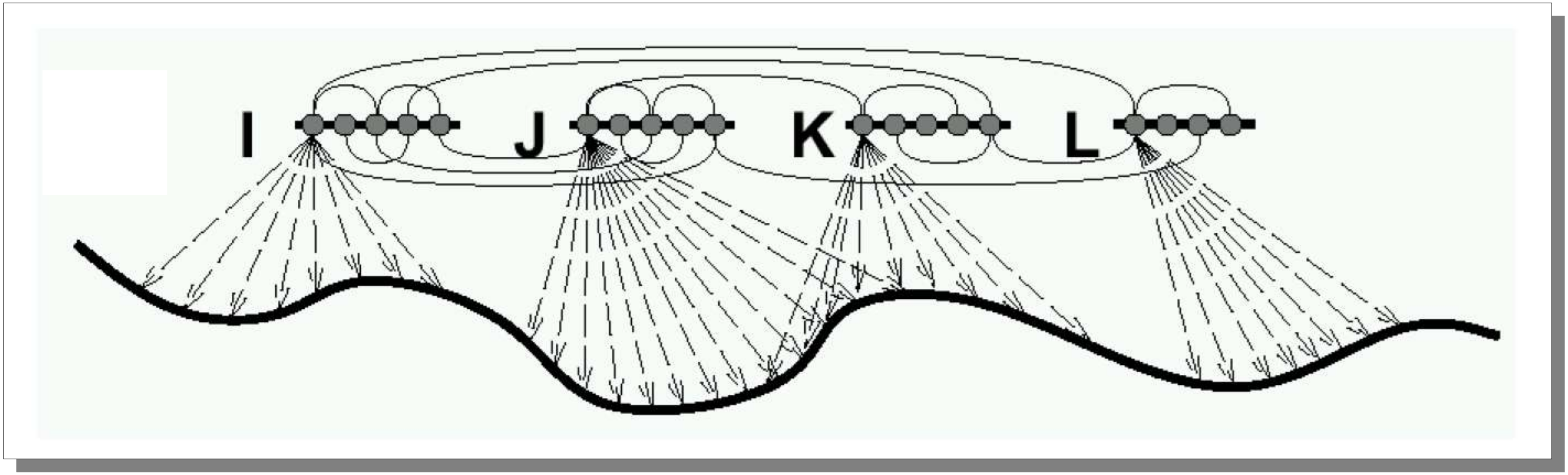
last amino acid in K interacts with first amino acid in L

- Small circles represent amino acid positions.
- Thin lines indicate interactions represented in model.

http://www.biostat.wisc.edu/bmi576/lecture16.pdf http://www.ics.uci.edu/~rickl/publications/1998-salzberg-chapter.pdf

The protein threading problem

Possible threadings:



Unfortunately, due to variable-length gaps between core segments and non-local interactions, this problem is NP-hard.

Fortunately, it is amenable to solution by a general-purpose optimization strategy known as *branch-and-bound*.

http://www.biostat.wisc.edu/bmi576/lecture16.pdf http://www.ics.uci.edu/~rickl/publications/1998-salzberg-chapter.pdf

Digression: branch-and-bound

Basic idea:

- Partion solution space into distinct sets.
- Compute lower bound that applies to all solutions in given set.
- If we can find a solution that is better than this lower bound, we don't need to explore any solution in that set.

Important note:

branch and bound will find optimal solution (it's not heuristic)

... but ...

it might take exponential time to do it.

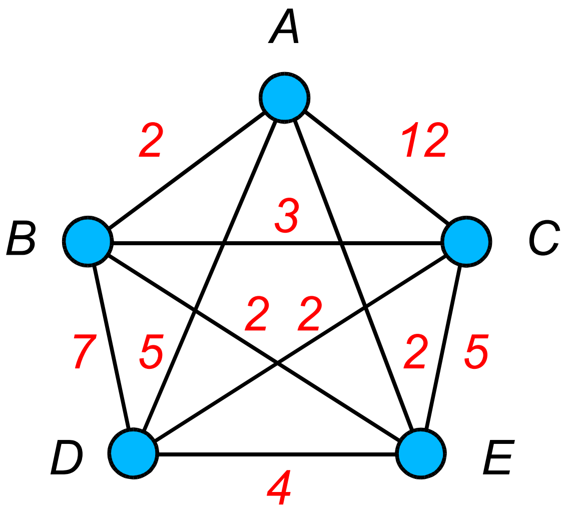
Still, it is often much faster than naive exhaustive search.

Digression: branch-and-bound

Let's see how this works for the traveling salesman problem, which we know is also NP-complete (protein threading is a bit too complicated for now).

Given: a set of cities and costs to travel between them.

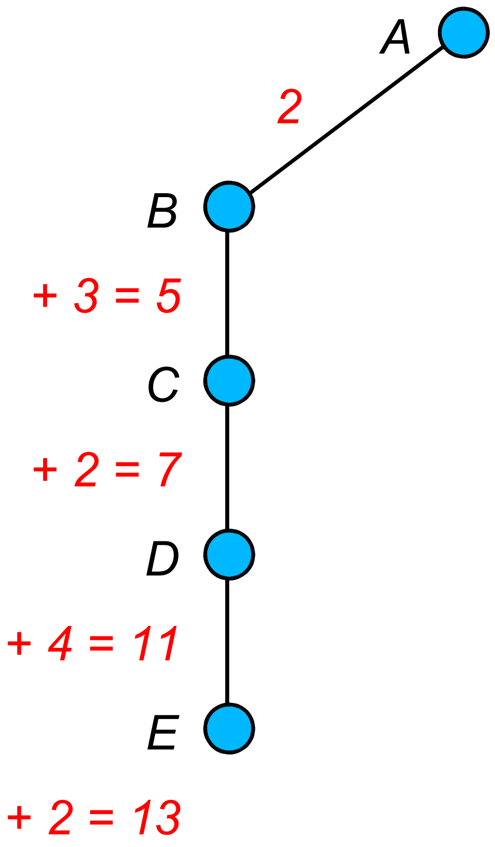
Find: a minimum cost tour that visits each city once.



	A	B	C	D	E
A	-	2	12	5	2
B	2	-	3	7	2
C	12	3	-	2	5
D	5	7	2	-	4
E	2	2	5	4	-

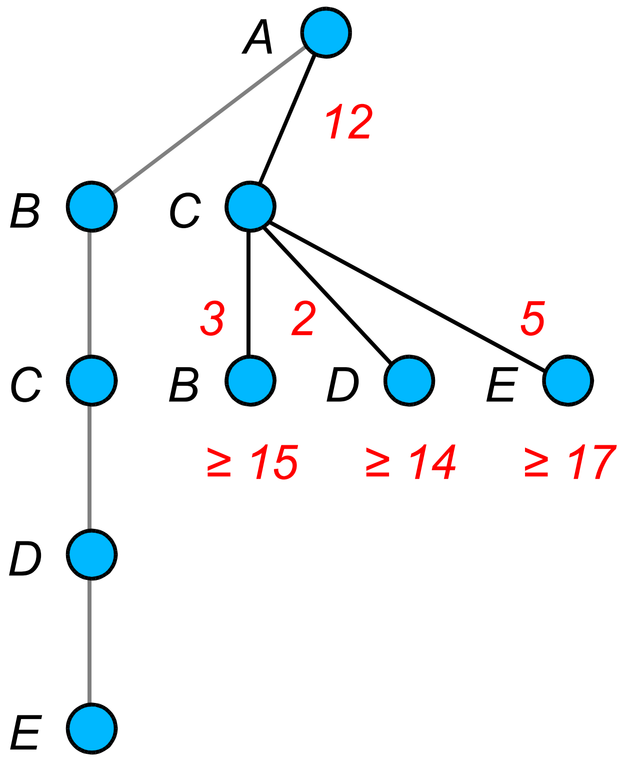
Branch-and-bound

Start search at *A*:



Total cost for this tour is 13.

Now let's try going from *A* to *C*:



No point in exploring any of these subtrees any further!

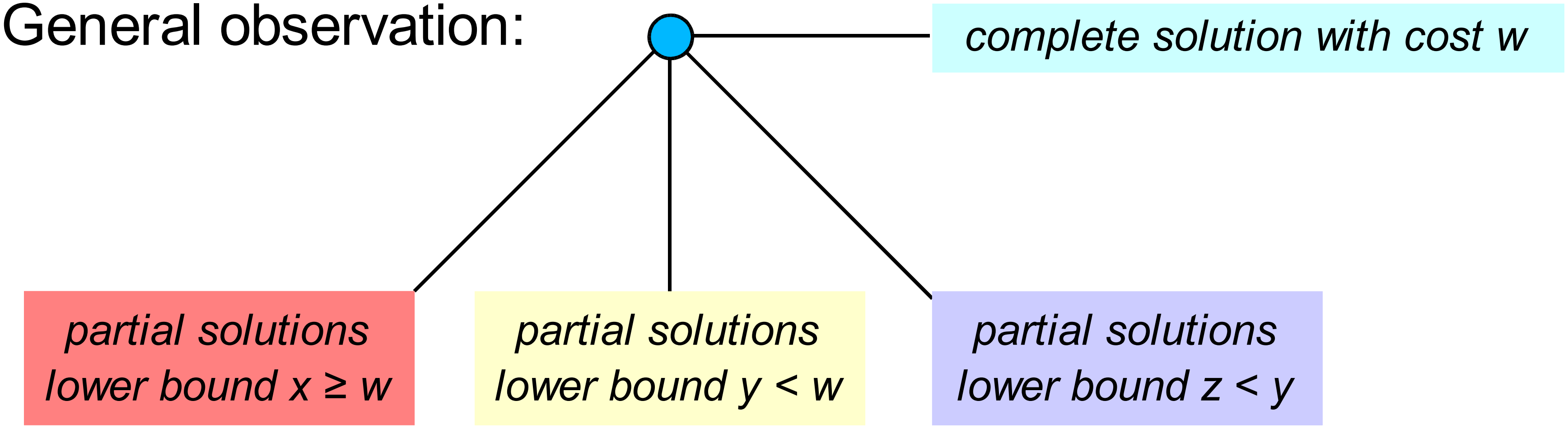
Branch-and-bound

In traveling salesman, we were able to eliminate from consideration all tours starting with:

A-C-B-... and *A-C-D-...* and *A-C-E-...*

because we knew they could never be optimal; we already had a tour with total cost less than their partial costs.

General observation:



complete solution with cost w

partial solutions lower bound x ≥ w

partial solutions lower bound y < w

partial solutions lower bound z < y

don't bother exploring this

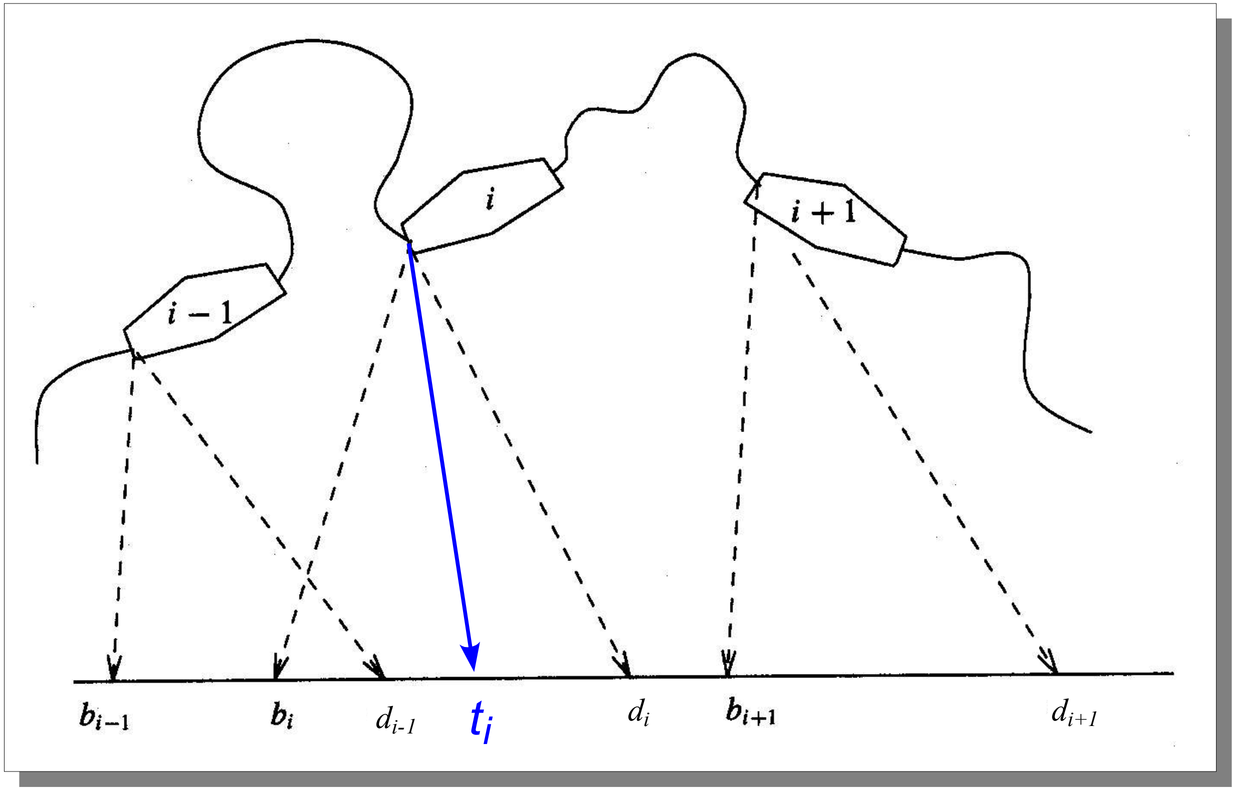
explore this if/when bound becomes best

explore this first

Back to protein threading ...

Recall that t_i indicates which amino acid occupies the first position in core segment *i*. Our scoring function is:

$$f(T) = \sum_{i} g_1(i, t_i) + \sum_{i} \sum_{j>i} g_2(i, j, t_i, t_j)$$



As we know sizes of core segments and min and max lengths for loops, we can determine ranges for t_i's.

This will form basis for branch-and-bound.

Branch-and-bound for protein threading

Given a set of threadings T^*, the optimization problem is:

$$\min_{T \in T^*} f(T) = \min_{T \in T^*} \sum_i g_1(i, t_i) + \sum_i \sum_{j>i} g_2(i, j, t_i, t_j)$$

$$= \min_{T \in T^*} \sum_i \left[g_1(i, t_i) + \sum_{j>i} g_2(i, j, t_i, t_j) \right]$$

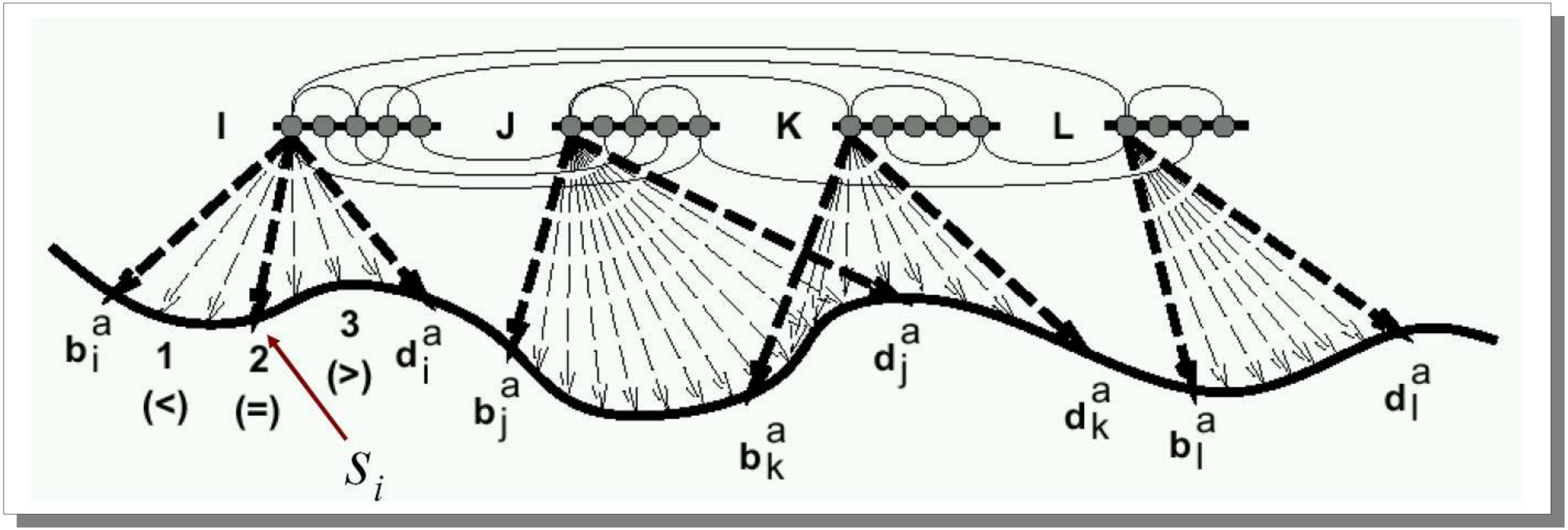
What's a lower bound we can use?

$$\geqslant \sum_i \left[\min_{b_i \leqslant x \leqslant d_i} g_1(i, x) + \sum_{j>i} \min_{\substack{b_i \leqslant y \leqslant d_i \\ b_j \leqslant z \leqslant d_j}} g_2(i, j, y, z) \right]$$

Note this is determined by the interval $[b_i, d_i]$ that t_i may fall in.

Branch-and-bound for protein threading

Now we must split solution space into disjoint sets. Do this by selecting largest current interval for a t_i and cutting it in half.



$$T = \{\, \vec{t} \mid b_i \le t_i \le d_i,\ b_j \le t_j \le d_j,\ b_k \le t_k \le d_k,\ b_l \le t_l \le d_l \,\}$$

splitting into three sets

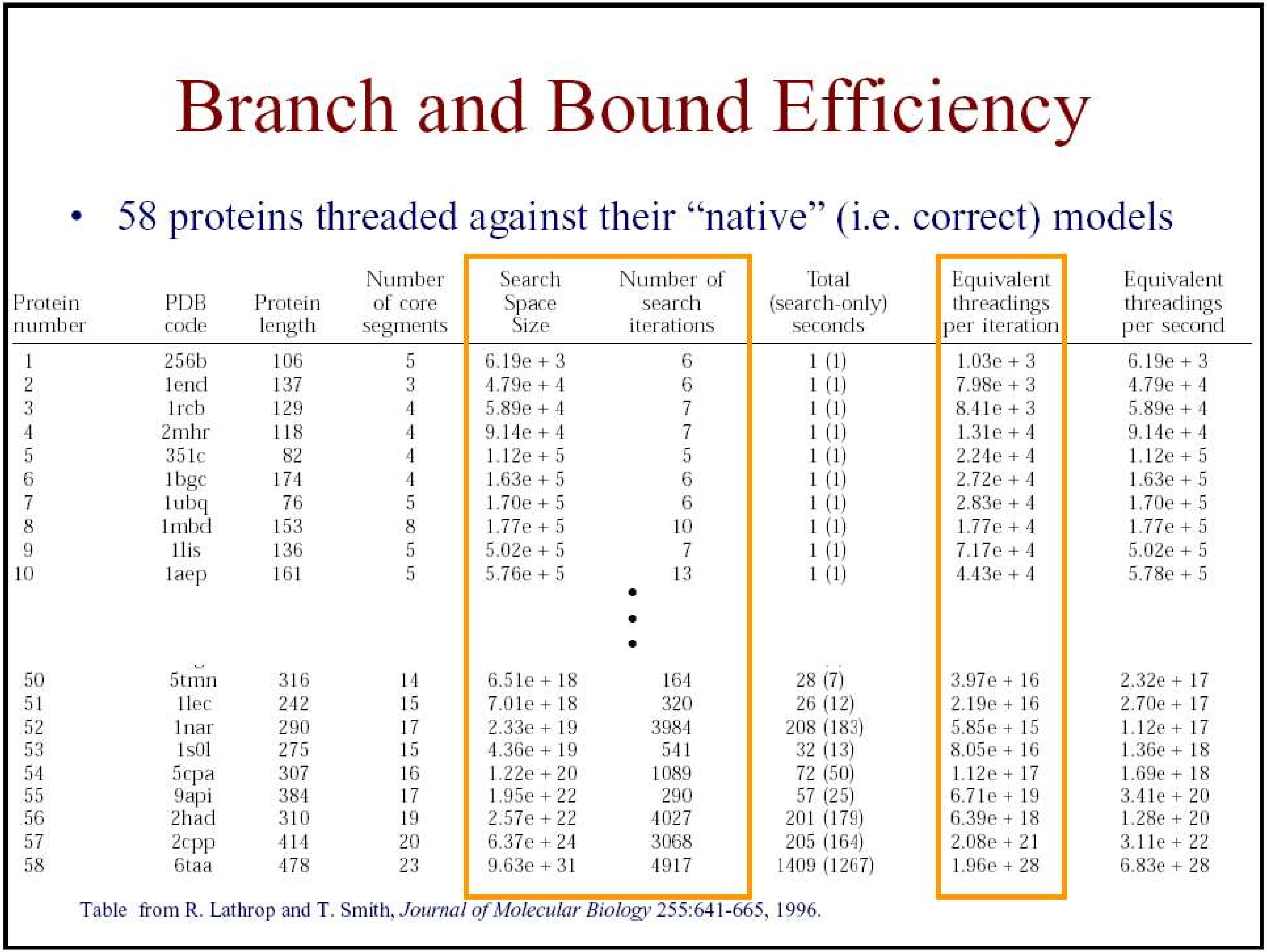
1 $T = \{\, \vec{t} \mid b_i \le t_i < s_i,\ \cdots \}$

2 $T = \{\, \vec{t} \mid t_i = s_i,\ \cdots \}$

3 $T = \{\, \vec{t} \mid s_i < t_i \le d_i,\ \cdots \}$

http://www.biostat.wisc.edu/bmi576/lecture16.pdf

Branch-and-bound for protein threading



Branch and Bound Efficiency

- 58 proteins threaded against their "native" (i.e. correct) models

Protein number	PDB code	Protein length	Number of core segments	Search Space Size	Number of search iterations	Total (search-only) seconds	Equivalent threadings per iteration	Equivalent threadings per second
1	256b	106	5	6.19e + 3	6	1 (1)	1.03e + 3	6.19e + 3
2	1end	137	3	4.79e + 4	6	1 (1)	7.98e + 3	4.79e + 4
3	1rcb	129	4	5.89e + 4	7	1 (1)	8.41e + 3	5.89e + 4
4	2mhr	118	4	9.14e + 4	7	1 (1)	1.31e + 4	9.14e + 4
5	351c	82	4	1.12e + 5	5	1 (1)	2.24e + 4	1.12e + 5
6	1bgc	174	4	1.63e + 5	6	1 (1)	2.72e + 4	1.63e + 5
7	1ubq	76	5	1.70e + 5	6	1 (1)	2.83e + 4	1.70e + 5
8	1mbd	153	8	1.77e + 5	10	1 (1)	1.77e + 4	1.77e + 5
9	1lis	136	5	5.02e + 5	7	1 (1)	7.17e + 4	5.02e + 5
10	1aep	161	5	5.76e + 5	13	1 (1)	4.43e + 4	5.78e + 5
					⋮			
50	5tmn	316	14	6.51e + 18	164	28 (7)	3.97e + 16	2.32e + 17
51	1lec	242	15	7.01e + 18	320	26 (12)	2.19e + 16	2.70e + 17
52	1nar	290	17	2.33e + 19	3984	208 (183)	5.85e + 15	1.12e + 17
53	1s01	275	15	4.36e + 19	541	32 (13)	8.05e + 16	1.36e + 18
54	5cpa	307	16	1.22e + 20	1089	72 (50)	1.12e + 17	1.69e + 18
55	9api	384	17	1.95e + 22	290	57 (25)	6.71e + 19	3.41e + 20
56	2had	310	19	2.57e + 22	4027	201 (179)	6.39e + 18	1.28e + 20
57	2cpp	414	20	6.37e + 24	3068	205 (164)	2.08e + 21	3.11e + 22
58	6taa	478	23	9.63e + 31	4917	1409 (1267)	1.96e + 28	6.83e + 28

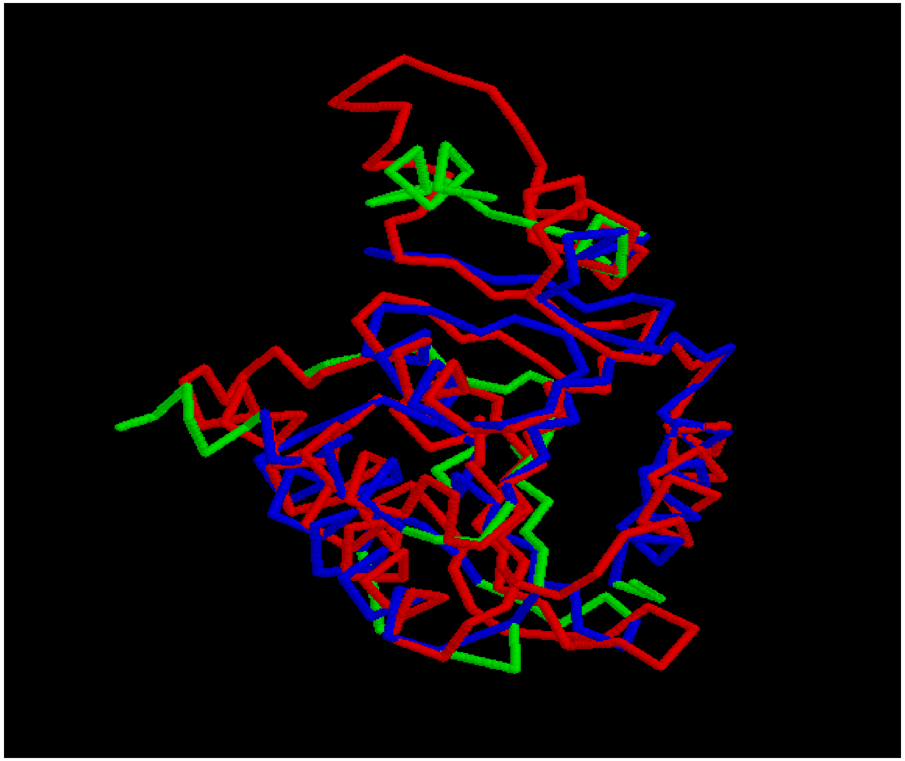
Table from R. Lathrop and T. Smith, *Journal of Molecular Biology* 255:641-665, 1996.

http://www.biostat.wisc.edu/bmi576/lecture16.pdf

CAFASP3 Example

CAFASP: Critical Assessment of Fully Automated Structure Prediction

CAFASP3 evaluated by MaxSub, a computer program. Predicted structures are superimposed to the experimental structures to see how long is superimposable.



Red: Experimental Structure Blue: Correct Prediction Green: Incorrect Prediction

http://www.bioinformatics.uwaterloo.ca/~j3xu/CS882/CS882-ProteinStructurePrediction.ppt