

CMSC 858P

Lecture 25: RNA folding cont'd; Protein folding

5/1/2008

# RNA Folding – covariance models

- Based on stochastic context free grammars (SCFGs)

$W \rightarrow (P|L|R|B|S|E)$

$P \rightarrow aWb$       *pair (a is paired with b)*

$L \rightarrow aL$       *left (a unpaired on the left)*

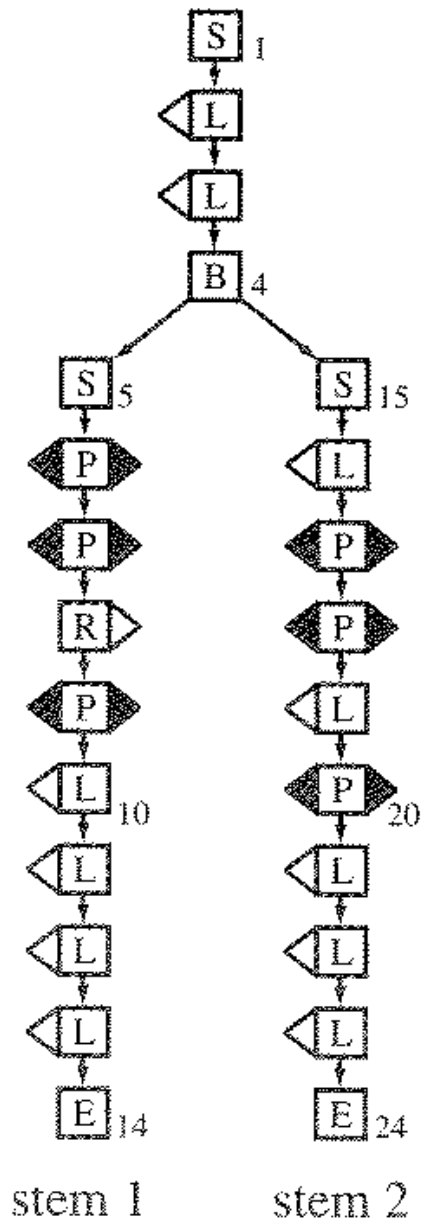
$R \rightarrow Lb$       *right (b unpaired on the right)*

$B \rightarrow SS$       *bifurcation*

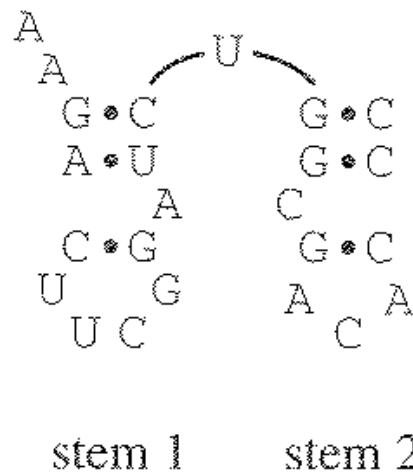
$S \rightarrow W$       *start*

$E \rightarrow \varepsilon$       *end*

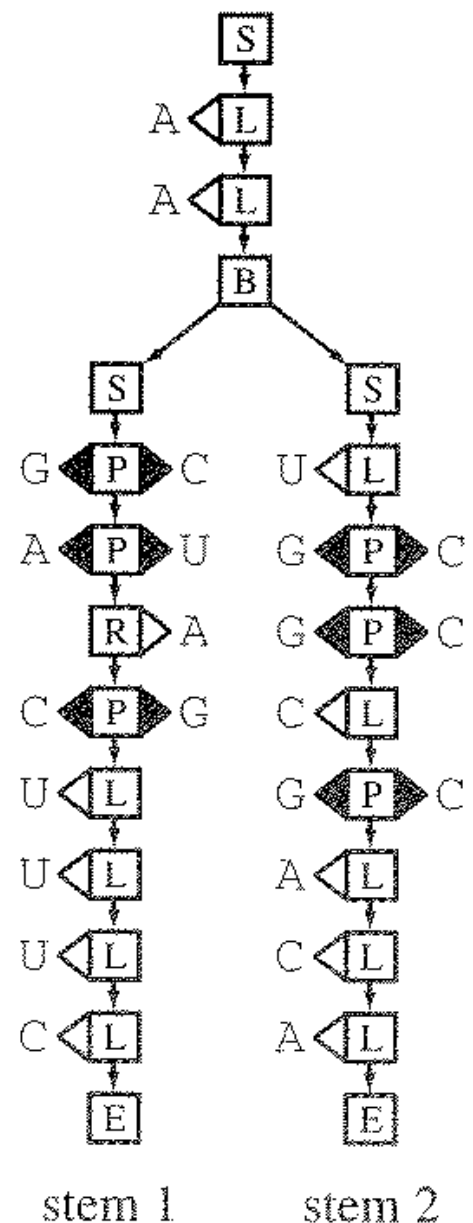
- State transitions associated with transition probabilities



SCFG



RNA structure



parse tree

# Parsing problem

- How likely is it that the RNA sequence observed was generated by the covariance model (CM)?
- Scoring (calculating this probability) can be done with dynamic programming (inside/outside/CYK/forward/backward, etc.)
- High-scoring regions of the genome are likely to be RNAs with the structure encoded in the CM.
- tRNAscan-SE – finds transfer RNAs
- More on machine learning techniques in CMSC 828N Spring 2007

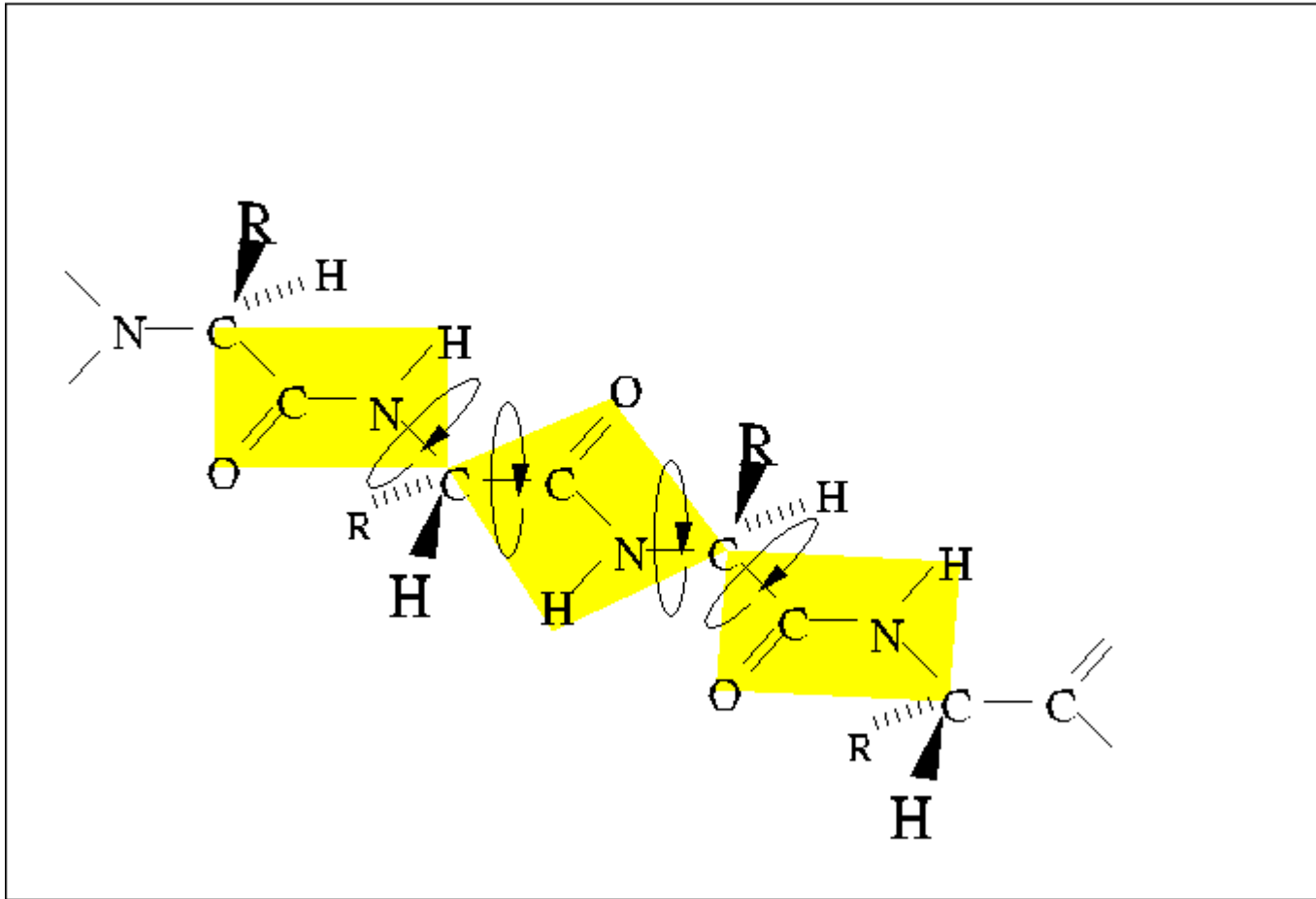
# Protein folding

- Protein shape determines protein function
- Protein sequence determines protein shape (Anfinsen's experiment)
- Levinthal's paradox – space of possible protein conformations is exponentially large, yet proteins fold fast (usec – minutes).
- Corollary: proteins must “know” how to fold (i.e. they don't search the entire space of conformations)

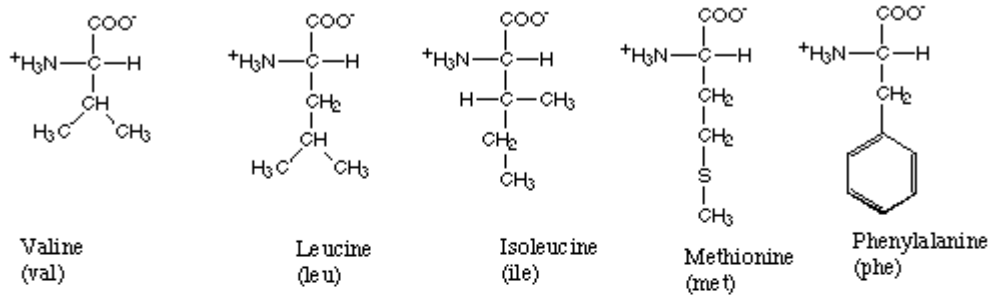
# Protein folding

- Note: mis-folded proteins may cause disease (e.g. Creutzfeldt-Jakob a.k.a. mad cow)
- Drugs (e.g. antibiotics) often inhibit protein function – knowing structure can help design drugs
- Folding@home – lend your computer's unused cycles to help fold proteins (like SETI@home) (do you believe in evolution or aliens ?)

# Protein structure (primary structure = sequence)

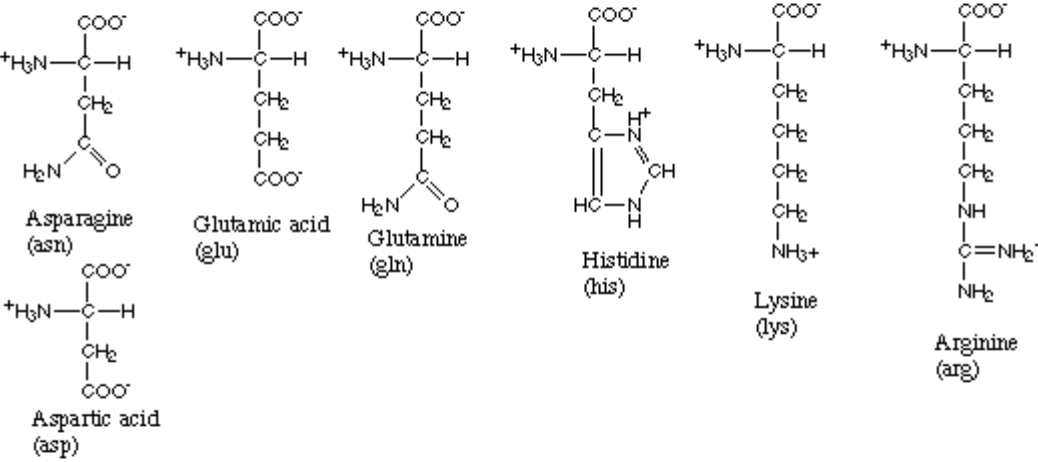


Amino acids with hydrophobic side groups



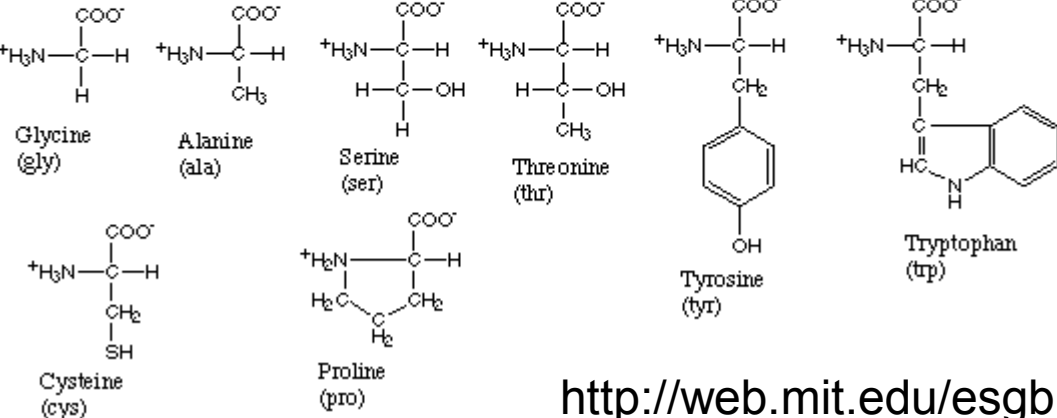
hate water

Amino acids with hydrophilic side groups



like water

Amino acids that are in between

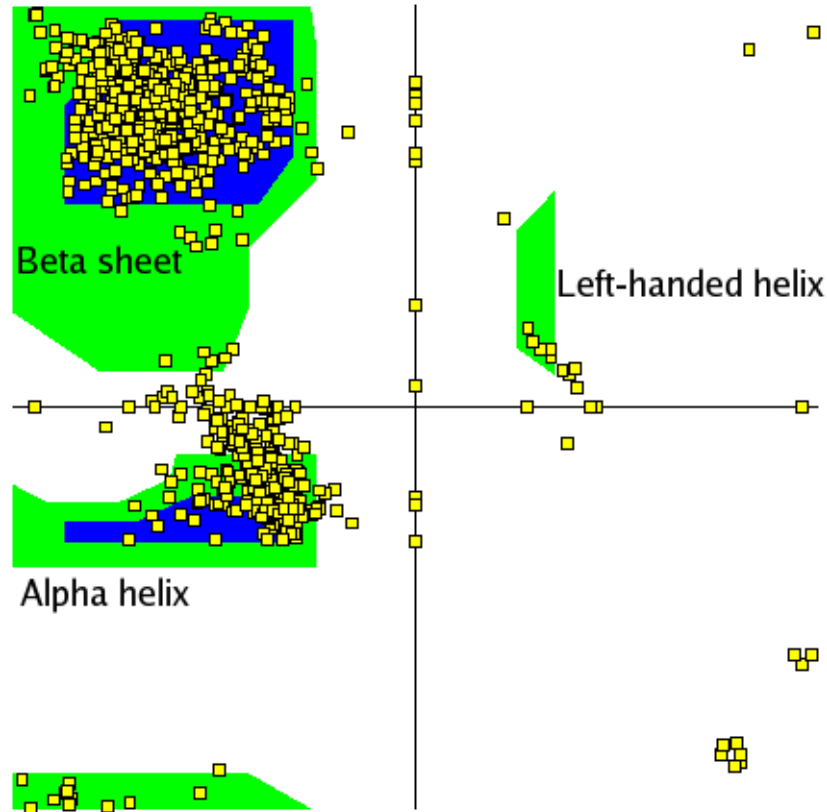


can't decide



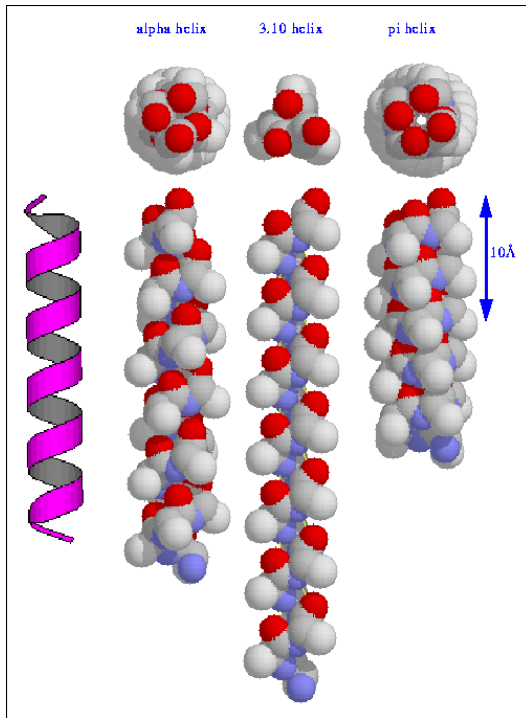
# Not all bends equally likely

## Ramachandran plot

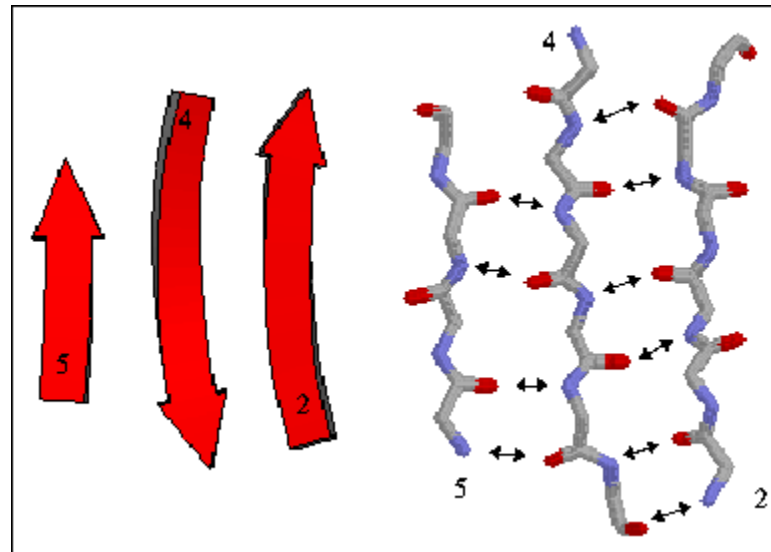


# Secondary structure (motifs)

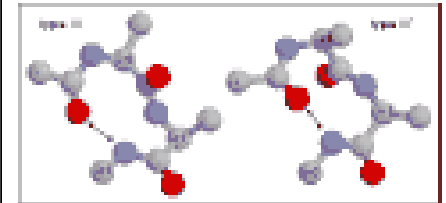
helix



sheet

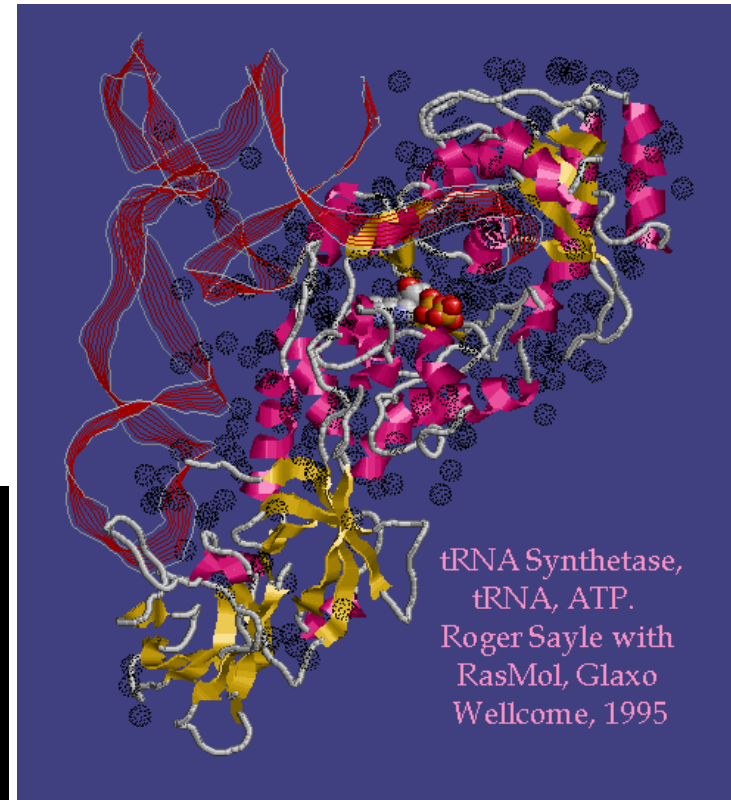
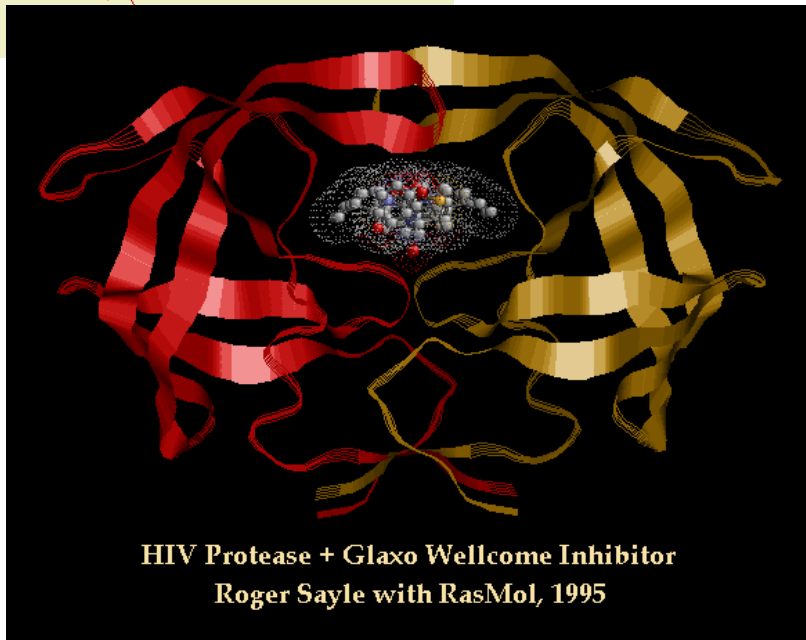
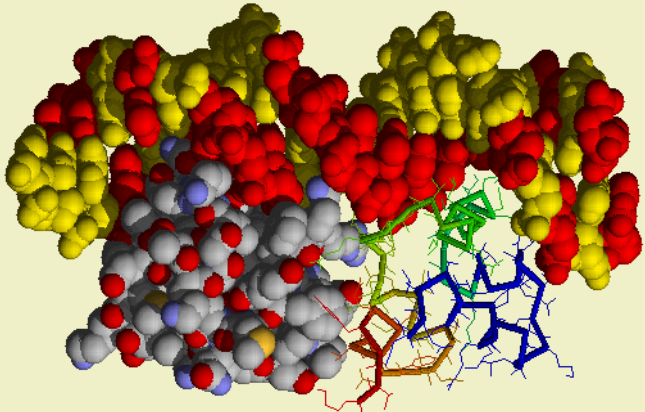


turn



# Tertiary structure (3D shape)

Phage CRO Repressor on DNA. Andrew Coulson & Roger Sayle with RasMol, University of Edinburgh, 1993

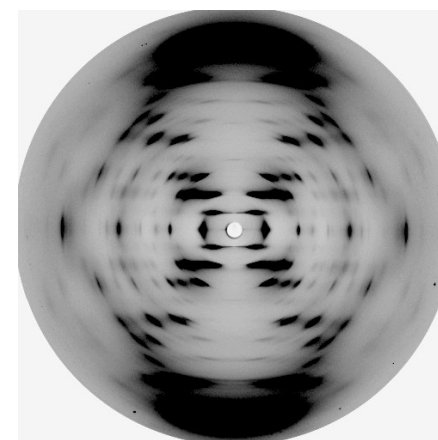


# Folded shape: lowest free energy

- Energy components
  - electrostatic ( $\sim 1/D^2$ ) ( $n^2$  terms)
  - van der Waals ( $n^2$  terms)
  - hydrogen bonding ( $n$  terms)
  - “bending” ( $n$  terms)
  - solvent (water/salt) (?? terms)
  - exclusion principle (no two atoms share same volume)
- Energy minimization
  - small perturbations & computation: hill climbing, simulated annealing, etc.
- Molecular dynamics

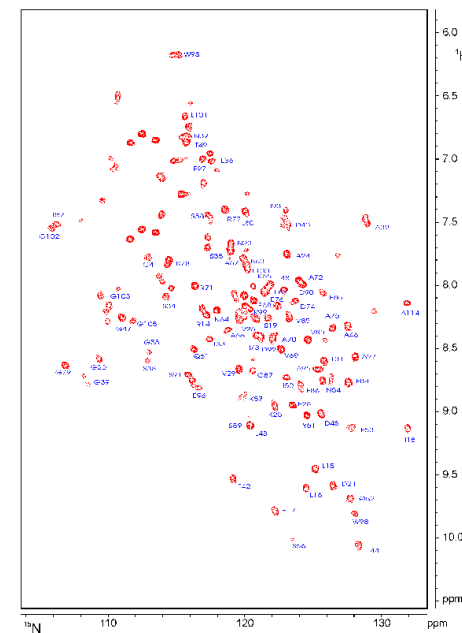
# How do we know the truth?

- X-ray crystallography
  - crystallize protein
  - shine X-rays
  - examine diffraction patterns



[http://www.cryst.bbk.ac.uk/BBS/whatis/cryst\\_an.html](http://www.cryst.bbk.ac.uk/BBS/whatis/cryst_an.html)

- Nuclear Magnetic Resonance (NMR)
  - no crystallization necessary
  - magnetic field “vibrates” hydrogen atoms
  - Nobel prize: Kurt Wuethrich



<http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/2dnmr.htm>

# Simpler problems

- Secondary structure prediction
- Side-chain conformation (assuming fixed backbone)
- Protein docking (how do proteins interact)
- Database searches (protein threading)
  
- Simpler energy functions
- Folding on a lattice (theoretical approximation)
  
- Critical Assessment of Fully Automated Structure Prediction – competition on proteins with unpublished 3D structure

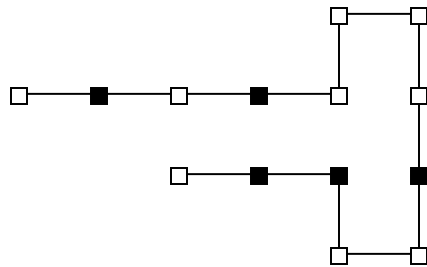
# Secondary structure prediction

## Chou-Fasman algorithm

- Estimate amino-acid propensities for helix/sheet structures (from known structures)
  - mostly found in helix/often found in helix
  - mostly found in sheet/often found in sheet
  - ambiguous
- Find helix/sheet “seeds” - regions with many “mostly” AAs
- Extend seeds while overall propensity/likelihood of structure is good
- Clean up prediction (e.g. overlapping modules)

# Folding on a lattice

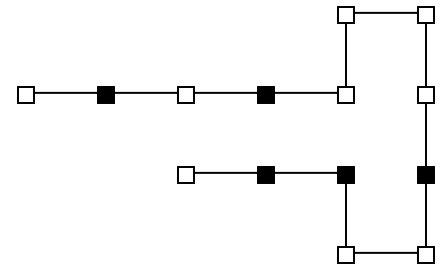
- Protein – colored beads on a string
- Lattice – beads can occupy nodes in a 2D/3D lattice (not necessarily square lattice)
- Hydrophobic (black or 1) / hydrophilic (white or 0) model
- Objective: maximize # of contacts between hydrophobic beads
- NP-hard, constant approximation computable in linear time





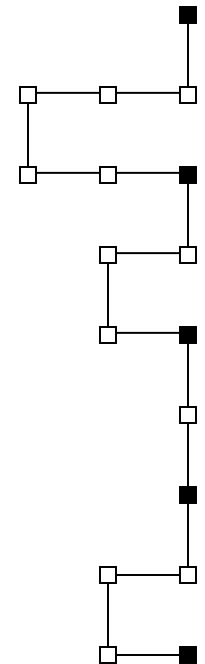
# Folding on a lattice

- Note: sheets are reason why RNA folding dynamic programming algorithm doesn't work (lots of pseudo-knots in proteins)
- Residues  $i$  and  $j$  are adjacent iff  $|j - i|$  is odd
- Block decomposition:
  - $b = 1$  or  $1Z_11Z_21Z_3\dots1Z_k1$   $Z_i$  – odd # of 0s
  - blocks separated by even # of 0s
- Properties:
  - 1s from a same block cannot be paired
  - 1s from even blocks can only be paired with 1s from odd blocks



# Block decomposition

- Odd blocks – X blocks, even blocks – Y blocks (1s from X blocks can only line up to 1s from Y blocks)
- Normal form – 1s in a line separated by single 0s. Block separators fall to the side (the “face”)
- X – super-block structure – treat Y blocks like 0s
- Y – super-block structure – treat X blocks like 0s



# Approximation algorithm

- Decompose protein into blocks
- Find the optimal “folding” place
- Build “super-block” structure for each half (one half as an X super-block, the other as a Y super-block)
- Fold the halves onto each other
  
- Claim:  $1/4$  approximation in 2D,  $1/8$  in 3D. iterative algorithm leads to  $3/8$  approximation. All algorithms run in linear time!!
- Proof: read the paper (careful counting of contacts)