CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 23

Protein folding

Internship opportunity

Genetic Data Curation Intern

The Consortium for the Barcode of Life (CBOL) housed at the Smithsonian Institution's National Museum of Natural History is seeking a summer intern (16 weeks) to work with data curation and quality assurance of large databases of gene sequence data. The intern will also be responsible for the compilation and confirmation of summary data on museum repositories around the world.

The intern should be an advanced undergraduate or recent graduate with a biology background, to include course work in genetics and some experience with bioinformatics. The intern must have strong computer skills and ideally some knowledge of GenBank.

Housing is not provided. Stipend is $400/week.

To apply, please email a cover letter, CV, and a list of relevant coursework with grades earned to Katie Ferrell, project manager, at ferrellk@si.edu.

For more information on CBOL please visit our website at http://barcoding.si.edu/

Internship opportunity

Intern Project Description, Consortium for the Barcode of Life

The Consortium for the Barcode of Life (CBOL) is catalyzing the creation of a global database of database records that link together gene sequences in GenBank, species names in several other global databases, and specimen records stored in museum databases. CBOL has created a set of data standards that ensure the quality and consistency of these data records. For example, all "official" barcode records must have:

(1) a structured link for a voucher specimen;

(2) a link to a species name that can be retrieved from a published source;

(3) links to trace files in the NCBI Trace Archive;

(4) primer sequences; and

(5) a minimum length of 500 base-pairs with less than 1% ambiguous sites.

These barcode records are being assembled and submitted by research labs all around the world. Many of these labs are using the Barcode of Life Data Systems at the University of Guelph in Ontario as a workbench for assembling, testing, correcting, and finally uploading the records to GenBank. Other labs are using local data management tools to prepare their data for submission to GenBank.

In this project, the intern will analyze the most common errors in the BOLD and GenBank data and in the data being assembled by some of the major barcoding labs. The intern will explore the sources of these common errors and will suggest informatics solutions for them.

Folded shape: lowest free energy

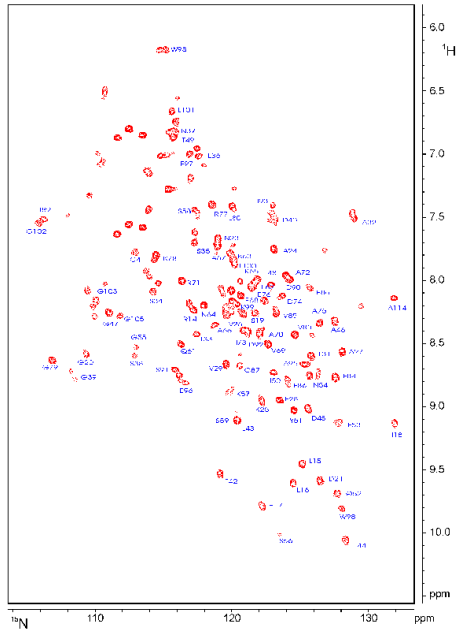
- Energy components
  - electrostatic (~$1/D^2$) (n^2 terms)
  - van der Waals (n^2 terms)
  - hydrogen bonding (n terms)
  - “bending” (n terms)
  - solvent (water/salt) (?? terms)
  - exclusion principle (no two atoms share same volume)
- Energy minimzation
  - small perturbations & computation: hill 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

- 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

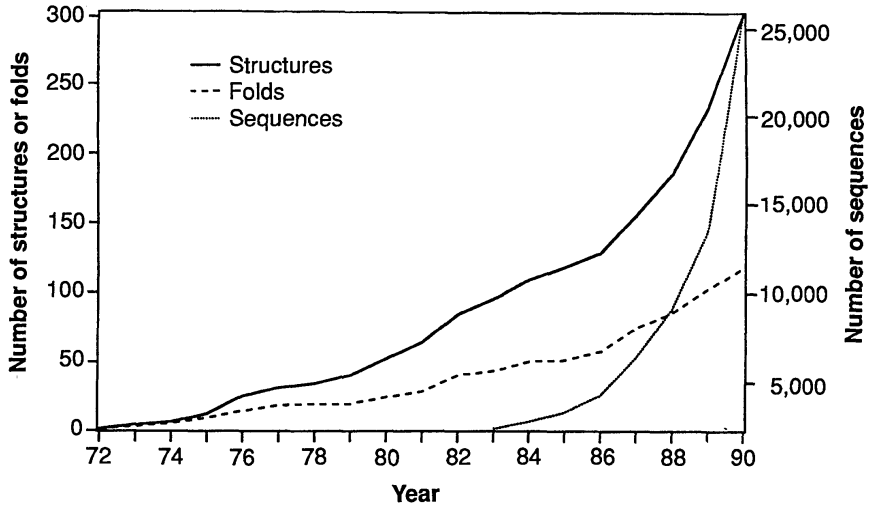
Secondary structure prediction

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)

Threading: reverse structure prediction

- Main hypothesis: while there are many protein sequences, there are much fewer folds. I.e. nature keeps reinventing useful structures



- Given a database of structures and a query string, find which structure “fits” the string best

Initial idea: 3D-1D scores

- From a 3D structure, determine “environment” for every amino-acid
  - buried (inside the protein)
  - outside
  - inner side of helix
  - outer side of helix
  - etc...
- Annotate each position in protein with the environment information
  ACKCAHGT -> $E_1E_2E_1E_3E_4E_2E_3E_1E_4$
- Why this is reasonable? Amino-acids have “preference” for specific environments

Alignment to an environment string

- Idea: use gapped alignment algorithm to estimate how likely it is for a sequence to conform to a structure (represented as an environment string)

- $E_1E_2-\ E_1E_3-\ -\ E_4E_2-\ E_3E_1E_4$
  A G H - K T G A L K M N G

- Question: what is the score of aligning an amino-acid to an environment?

Answer: use statistics

- For each environment – calculate likelihood (observed frequency) of all amino-acids based on known structures
- For each environment – empirical estimation of gap opening/extension penalties
- Alignment algorithm – use Gribskov's profile method: replace each environment character with the amino-acid frequency table for that environment

E_1

A 0.22

K 0.15

W 0.08

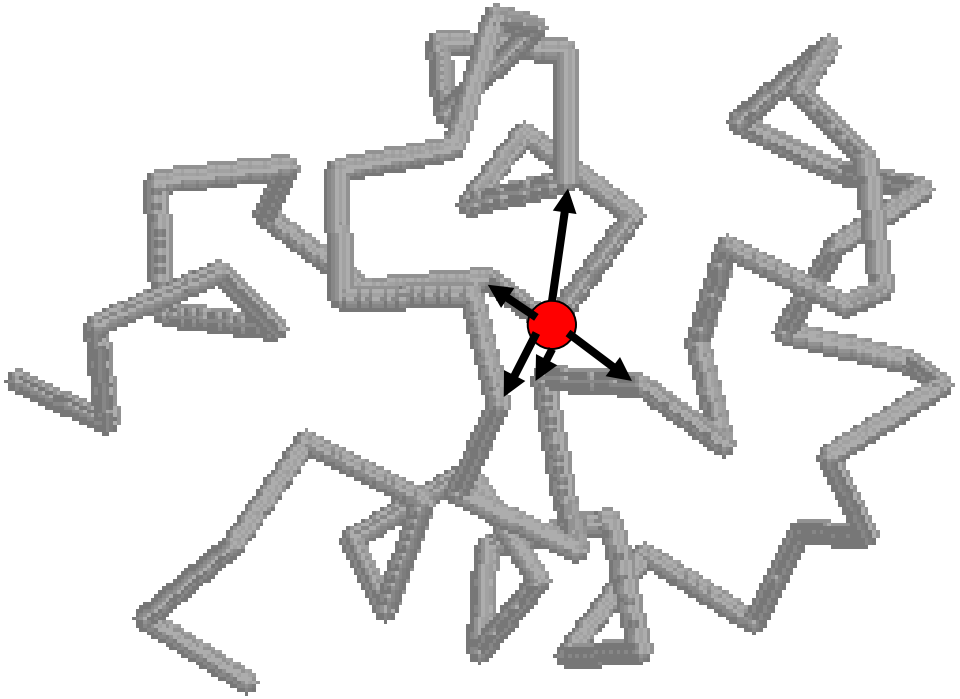
...

$S(E_1, G) = \sum_{AA} S(AA, G) * freq_{E1}(AA)$

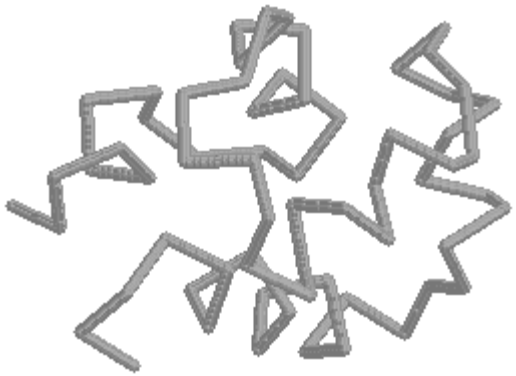
S(AA, G) – e.g. from BLOSUM matrix

Environments – not good enough

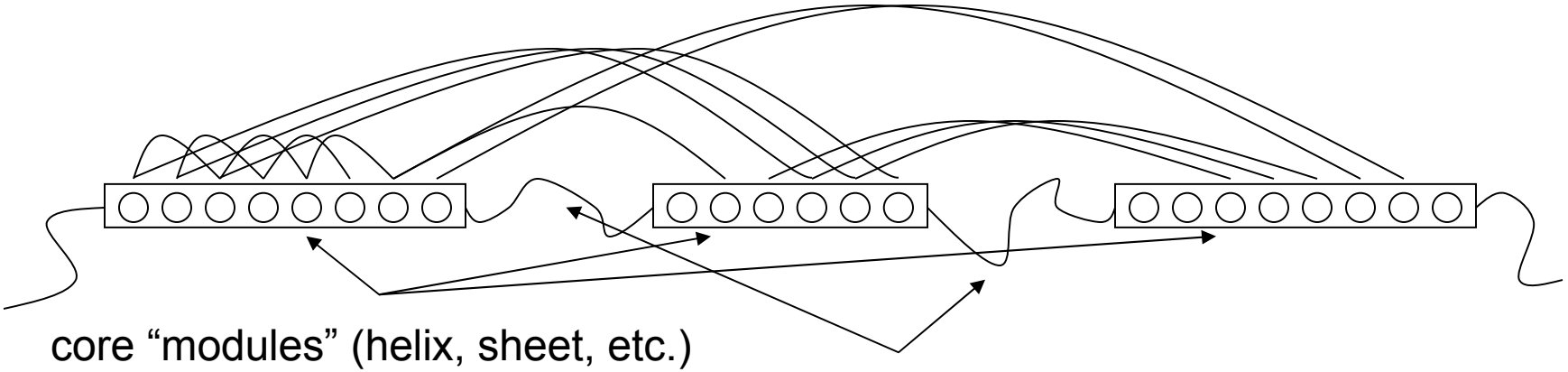
- Each amino-acid may have multiple contacts



A better model



residue interactions (and associated energy parameters)



core “modules” (helix, sheet, etc.)

variable length connections (gaps)

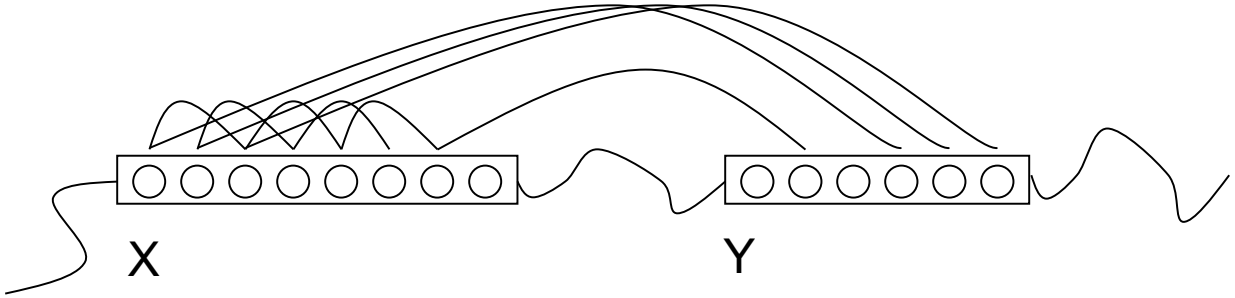
The threading problem

- Model assumptions:
  - loop AA composition and length contributes to energy score (note: can also place restrictions on minimum/maximum size in gaps)
  - interactions are pair-wise: interaction energy depends on at most two AAs
  - individual AAs in core modules also contribute to energy due to local environment
- Thread a protein sequence through a structure model s.t.
  - the place-holders are filled with amino-acids
  - a variable number of amino-acids fall in the gaps
  - overall energy is minimized
- Easy to say, hard to do: Thus defined (variable length gaps AND pair-wise interactions) - NP-hard!

NP-hard => heuristics

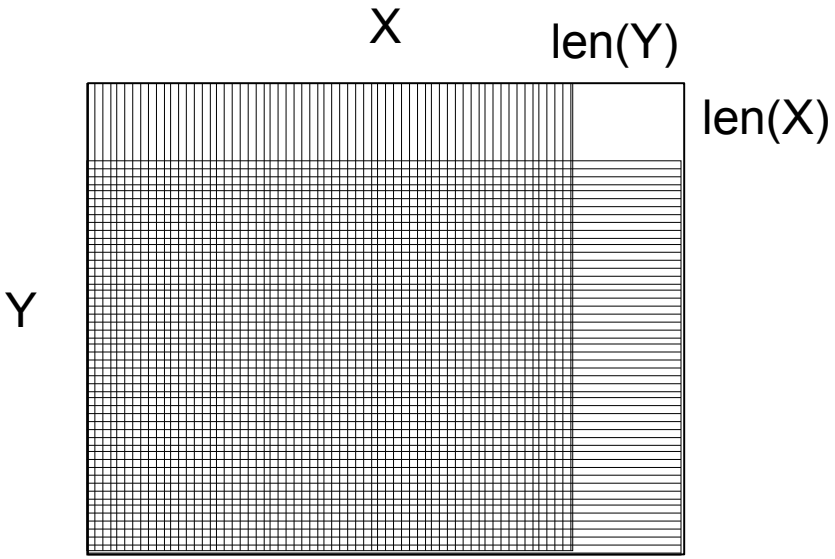
- Branch and bound (Lathrop, Smith)
  - Represent all possible folds (search space) s.t. it is easy to compute a lower bound on the score
  - Note: a threading is uniquely defined by the coordinates of the core elements – a set of threadings is a hyper-rectangle in a C-dimensional space where C is the # of core elements
  - Divide search space and compute energy lower-bounds on each sub-division (choose a dimension (core) and a coordinate and split hyper-rectangle at that location)
  - Recurse on sub-division with lowest lower-bound

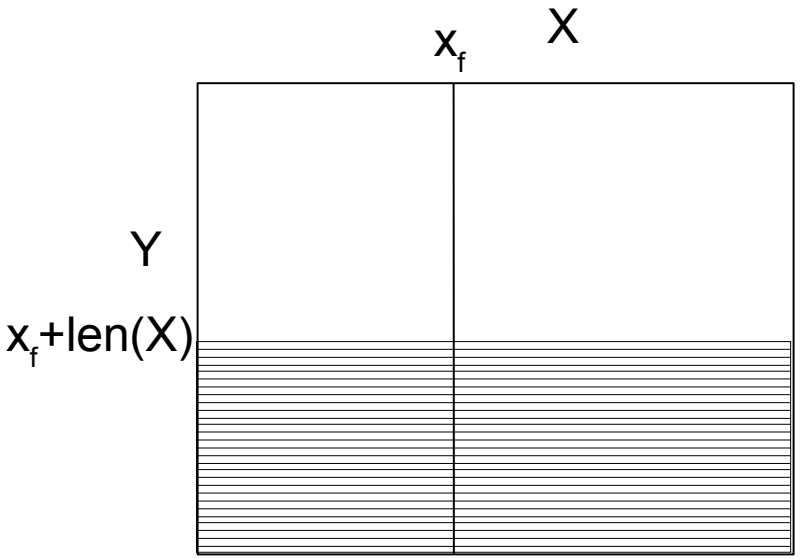
Hyper-rectangle heuristic



Each “module” corresponds to a dimension - offset of module in the protein

Fixing one module restricts the flexibility in assigning the remaining modules (imagine beads on a string)





Proteomics

- Large-scale analysis of proteins
  - protein-protein interactions (e.g. yeast 2-hybrid)
  - 2D gels (mass vs. isoelectric point)
  - Mass-spectrometry
  - Protein microarrays
  - etc.