From: Steve Lewis
Subject: CSE 527 Notes, Lecture 3 Wed Oct 4

Protein Data Bank - large collection of 3-d protein structures:
http://www.rcsb.org

JMol Viewer
TopoIsomerase is a protein which wraps around the DNA helix
Viewer supports 3D view with rotation

http://www.rcsb.org/pdb/explore.do?structureId=1YUA

With TopoIsomerase (or any protein ) there are active regions and less critical regions
Question - can we determine critical regions by looking at variation among species for conserved regions

Mapping amino acid sequence to protein structure
3D structure is largely determined by X-Ray crystallography.

Review dynamic programming model

Question - Needleman Wunsch algorithm - widely called dynamic programming algorithm
traceback follows multiple alignments

Graph representation of all paths may be shown in quadratic time -
My Question - is it possible that less symmetric scoring might reduce tying paths.

Homework? find the sequence best path.

Question - can an approximation of the best path be made in subquadratic time?
Answer - not sure.

Needleman Wunsch algorithm - assumes independence of score of previous data
Possible solution - change the scoring

Global alignment might miss if there are large insertions.

Local alignment - find substrings of S and T with maximal global alignment

V(i,j) max value of opt global alignment of suffix of S[1],S[2]...S[i] with T[1],T[2]...T[j]
assume aligning with a blank is negative
V(i,0) - empty suffix costs 0
V(0,j) = 0;

optimal suffix has similar 3 choices plus empty suffix
4 possibilities align values , align
in the matrix score can never fall below 0

goal is maximum entry with any cell in the table

Traceback all paths to a zero.

Question - difference between empty set and a space - optimal local alignment will not start with a space in either String.

Local alignment Smith-Waterman
Global alignment Needleman-Wunsch

Alignment with gap penalties - might want to penalize multiple gaps -
biologically there is a cost for a gap somewhat independent of length

score for gap function of gap length - function increasing (convex)
affine beginning penalty then lower cost for extension -
cost g to start a gap s cost to continue a gap
cost[0] = x, cost[i > 0] A * x 0 < a < 1
V[i,j] optimal alignment
G[i,j] last pair matches
E[], F[], gap at end of one string or other

3 cases to track G,F,E case with a gap in S,T, Neither also V is a max value

Problem with bookkeeping is that V(i,j) is not unique sequence

cost of affine only requires knowing whether gap is being started
or continued not tracking the precise gap length.

cost of general - non increasing
convex can use binary search

Nobel prize in medicine went to work on RNA interference

Chemistry - detail of RNA transcription

DNA replication-
in soup: nucleotide triphosphates
DNA polymerase grabs a fitting nucleotide and ratchets forward
cleaving the phosphate provides the energy
ATP GTP ... provide energy (interesting since there are cells energy)
construction 5' to 3'

Starting process is hard - requires special catalyst - primase -
primers are first few nucleotides

problem - still accuracy
primase makes an RNA strand

Problem - unwinding the helix: helicase unwinds the helix
on leading strand synthesis follows helicase

on the other strand replication has to run away from unwinding double
helix (lagging strand). DNA polymerase runs backwards so many
segments (Okazaki fragments) about 200 pairs
Then gaps fused by ligase

problem - primer produces RNA -
nuclease chews up RNA - then polymerase fills in gap
because RNA fragment is thrown away primase does not need to make a high fidelity copy

Question - leading strand start with a primer which must be patched -
There are multiple sites of replication in a DNA strand

Timing of triggering of multiple sites of replication is choreographed

helicase, priming and leading strand replicase are one complex

problem - helix needs to spin to unwind - TopoIsomerase I nicks DNA to make a swivel
TopoIsomerase II allows two strands to pass through each other

proofreading
total error rate 10^-9
polymerase error 10^-4
other enzymes look for bulges
somehow they find which is original
bacteria methylate 'A' nucleotides - allowing strand recognition

Eucaryotes - new strand has nicks - not old strand

Time scale – replicate 500 base pairs per sec procaryotes
Eucaryotes 50 base pairs per sec