CSE 527 - Lecture 10, Monday October 30, 2006
Class notes, Hanna Filipsson

Gibbs sampler:

EM and Gibbs are similar methods, biggest difference is that EM uses hill climbing to find max, and Gibbs samples from the different instances

Extensions
basic approach is simple but doesn't always work
tricks to help:

- Phase shift
  - once you've found an alignment, look around to see if shifting over gives better results
  - similar --> broaden search
  - not similar --> narrow search
- Algorithmic adjustment of pattern width
  - add/remove flanking positions (to maximize average relative entropy positions)

Example (of algorithmic adjustment of pattern width):

AATAAA
AATAAG
AATAAG

The algorithm might settle on the last 5. Theory says that if you run the algorithm long enough, it will find the better alignment of the first 5. But it requires a series of unlikely choices. We dont want to wait that long. Instead, try move instance right and left.

Methodology

There are several tools to use (to find transcription factor binding sites).
The paper mentioned on the slides (Assessing computational tools for the discovery of transcription factor binding sites) presents a research done to see if any tools (algorithms) are better than others:

A group of experts were asked to find transcription factor binding sites.

species used: human, mouse, fly, yeast

datasets:
'real' - collect real data
'generic' - upstream regions, random

'markov' - built third order markov of upstream regions
they created datatypes of all 3 kinds for all species

Unfortunately nature isn't as simple as the sequences we model with this, and we might bias some algorithms.
Using the real data might result in that an algorithm is penalized for making false predictions that
are actually good (because we didnt know about it when making our model).
The generic isn't perfect either since we don't know the exact stochastic process.
There are arguments against all 3 methods, therefore they used all 3 to increase their odds of doing good.

They let the experts of each algorithm perform the algorithm and took only the top prediction.

(Table of how well the different methods did)
The different methods did quite similarly good.
Weeder did suprisingly good, but it is conservative (got few wrongs predictions but
also few right)
They all have their pros and cons.
One might think a fancier algorithm is better, but it doesn't look like that in this table.

Consensus - greedy algorithm, doing okay here
(Gibbs was originally for protein alignment, but here adjusted for DNA alignment)
many use a variant of Gibbs
many use a variant of EM
MEME, MEME3, variants of the same thing
run by different groups, got similar results

(Table where the results are broken into datatypes, i.e. mouse, human, yeast, fly)
Yeast got high score
we know a lot about yeast
probably overfit the model to yeast (yeast is what we created the models from)

remember that models are missing things
for ex. a position can affect a place 100 000 positions away, this is not in any model

----

Comparative genomics:

cross-species comparison

Phylogenies (aka Evolutionary trees):

Complex question: given data (sequence, anatomy), find the phylogeny
talked about for a long time
can make big mistakes (ex. looks like a fish but isn't one)

Simpler question: Given data and phylogeny, how well does the tree fit the data
(this will be our focus here)

Parsimony:

The general idea is like Occam's Razor: "All things being equal, the simplest solution tends to be the best one."
Given data where change is rare, prefer an explanation that requires few events (mutations).

We compare sequences from different species and look at the places they differ. If the species are related,
there has been a mutation somewhere. We want to construct a tree from this, to see how closely different
species are. Construction the tree is however difficult (complex question above), so instead we
look at how good an existing tree is (simpler question above).

The tree with the smallest number of required mutations is probably the right one.
(It is "more likely that animals with an eyeball came from the same heritage, than several inventing
the eyeball seperately".)

Given a tree, we count the number of required mutations. For example:

(1)	(2)
A	T
A	T
A	G
C	G
C	T

(1) and (2) might look equally good, but with the given tree (1) requires 1 change
and (2) requires 2 changes (see slides for the tree).

This method works when change is rare.

Counting events parsimoniously:

use Sankoff & Rousseau algorithm
it is dynamic programming, but instead of using a matrix (like last homework), we use a tree

Maximum likelihood is usually considered to be a better way to evaluate a phylogeny, but
parsimony is a natural approach and fast.

Phylogenic footprinting:

goal: identify regulating elements

functional sequence evolve slower than non-functional ones
non-functional has no selection effect (and may spread fast)
if functional, then usually this is already the best and usually another is worse and will not spread

consider a set of orthologues (evolved from common ancestor) sequences from different species
identify unusually well conserved regions (hint that it is functional)