Charles Stewart
CSE 527
Lecture 11 notes

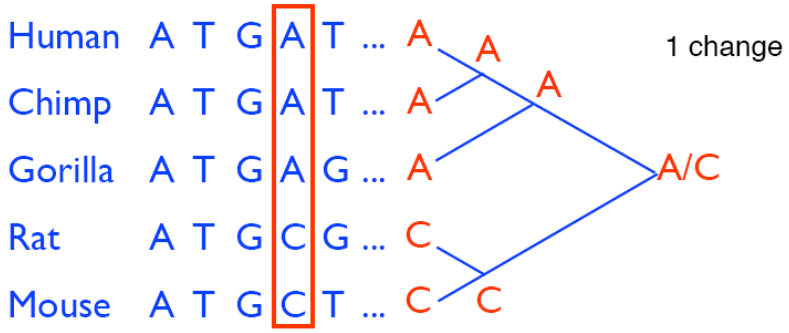
Definitions
Phylogenetics: The study of evolution relatedness among organisms. We take a group of species, look backwards for attributes that are similar enough to be related, but distinct enough to find mutations.
Homologues: sequences that are similar by descent from an ancestor
Paralogues: Homologues in the same species
Orthologues: Genes of different species thought to have evolved from a common ancestor. These genes may or may not be responsible for a similar function

A Complex Question:
Given data (sequences, anatomy, ...) infer the Phylogeny

A Simpler Question:
Given data *and a phylogeny*, evaluate "how much change" is needed to fit data to tree

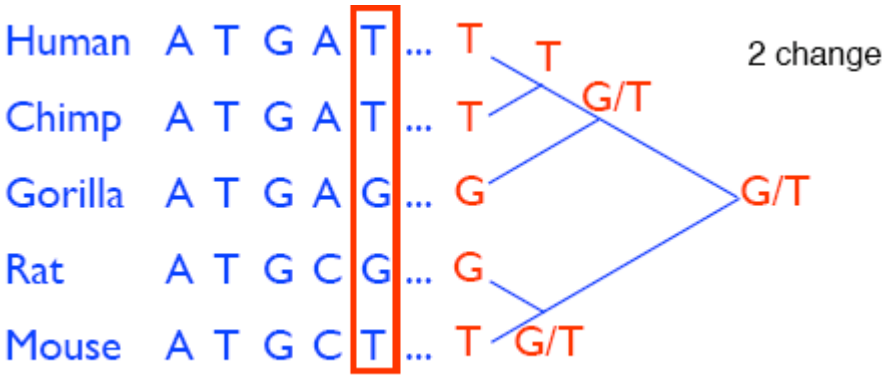
What we would like to do do:
We would like to create an ancestral evolutionary tree by finding the mutations between known organisms. To do this, we create a set of possible trees, and go with the tree that is accurate, but not complex.

Parsimony
Parsimony is similar to Occam's razor, the simplest explanation is most likely the correct one.



When given this set of 5 organisms, we analyze their sequences of ACTGs in hopes of finding where they differ. If they have a common ancestor, there was a mutation at some point. When created a tree to tie together these five organisms, the most accurate tree is probably the one with little change. In this case, when crossing the forth column, we see that the minimum number of changes to explain the tree is one. One change is necessary to differ these organisms.

Charles Stewart
CSE 527
Lecture 11 notes



With the next column, two changes (/) are required.

With this approach, the tree with the best minimum number of changes is most likely to be an accurate tree of the given set's evolution. While finding the best tree is a difficult (easier with Humans / Chimps than Humans / Elephants) , finding the minimum number for a given tree is easy to produce using Sankoff & Rousseau algorithm. This algorithm is well documented in the class slides and previous notes.

It should be emphasized that likelihood based methods are widely accepted as being more accurate than parsimony, but the later is a simple approach that still has applications, including the following.

Mathieu Blanchette & Martin Tompa: Phylogenetic Footprinting Algorithm
"Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting". Genome Research 2002.

Our goal is to identify regulatory elements in DNA sequences. These are binding sites for proteins, located up to 1000 nucleotides from gene. We notice that functional sequences in the DNA evolve much slower than non-functional sequences due to selective pressure. This difference in mutation rates is what allows Phylogenetic footprinting to be successful, while MEME and Gibbs sampler neglects this advantage.

By looking at orthologues between related species, we are identifying unusually well preserved regions. With this we can assume that there are fewer mutations, and a lower minimum number of changes to the ancestral tree. This is helpful when comparing a broad range of organisms, as orthologues of closely related organisms may not have any changes at all due to the danger of mutating functionally relevant sequences.

Substring Parsimony problem
Given:

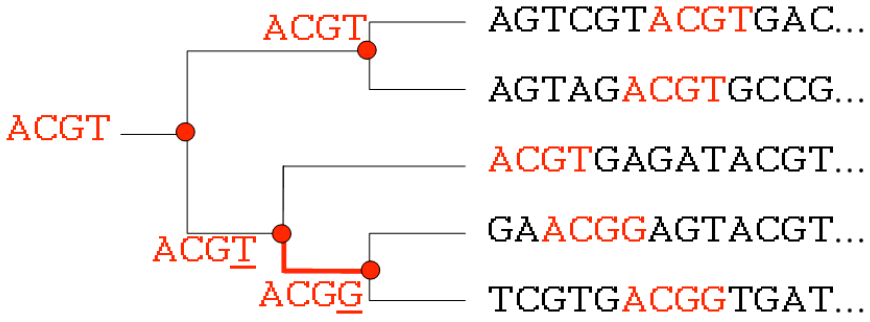
- A phylogenic tree *T*, a set of orthologous sequences at leaves of *T*,
- A motif of length *k*,
- a threshold d

Charles Stewart
CSE 527
Lecture 11 notes

Find each set S of k-mers, one k-mer from each leaf, such that the "parsimony" score of S in T is at most d.

Another example of parsimony, looking for a motif of size 4

Guessing common ancestors from like length 4 motifs



The parsimony score from a 4-long motif is one mutation.

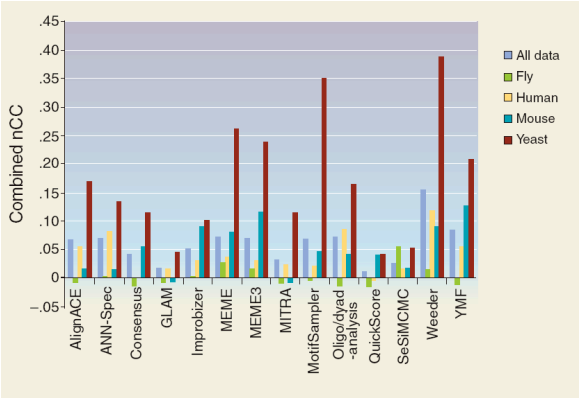
Using a variation of the Sankoff and Rousseau algorithm we find the total time is:
$O(n\ k\ (4^{2k} + l\ )$

- n = number of species in the tree
- k = motif length
- l = average sequence length

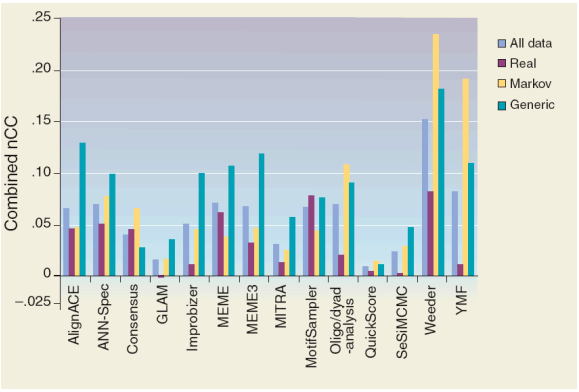
Notes on Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites. *Nature Biotechnology*, vol. 23, no. 1, January 2005, 137 - 144

In this paper, a group of experts on 13 varying computer programs were challenged to find transcription factor binding sites when given a variety of data sets. Three data sets

were used on the premise that all three are somewhat flawed. For example, a synthesized data set was created by stochastic methods, but the correct stochastic process is not known. Also, a data set collected from real data is flawed because a program could accurately locate a previously unknown binding site and be penalized. The remaining data is taken from known transcription factors, and all data is separated into Fly, Human, Mouse and Yeast sequences. With this basis and the following diagrams, we contemplate the possible strengths and weaknesses of the tested programs.



In this diagram, we compare the programs effectiveness on different organisms. We see that in nearly every program, the programs are able to most accurately find the binding sites in yeast. This could be for a few different reasons: Perhaps systems are simpler on Yeast. Perhaps Yeast has more data. Or perhaps the programmers of these 13 algorithms used yeast to help train their programs.



In this diagram, we compare the programs effectiveness based on the different data sets provided. In general, these programs are not very effective when dealing with the Real data set. Perhaps this is because the programs are actually finding binding sites that are not yet known, or perhaps they simply perform better on data that is synthetically produced with strict rules.

Also noted in the paper is the dramatic improvement of the MEME (and possibly other) algorithms when allowed to fine tune the parameters of the algorithm. It was also mentioned that there may be some improvement by allowing a given tool to predict two motifs rather than just one. Simply put, there is much room for improvement, and it is an exciting time to work in this field.