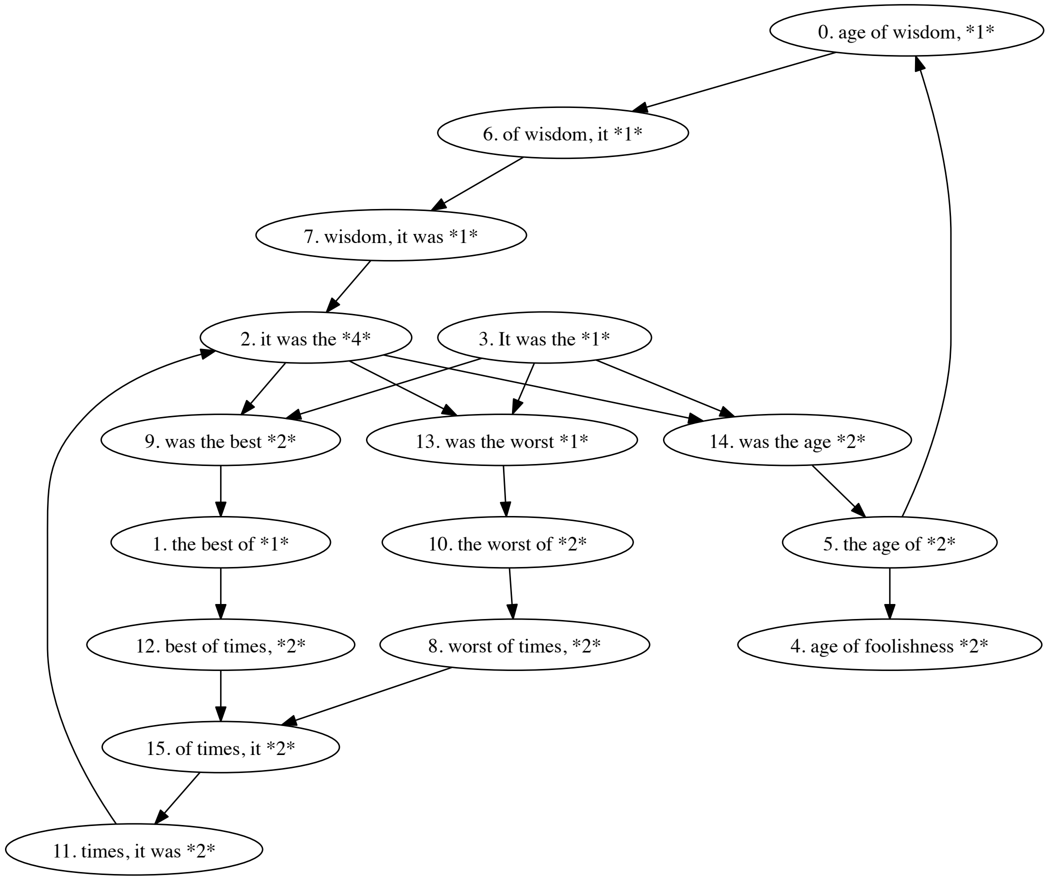
Ethan Sherbondy
Problem Set 2

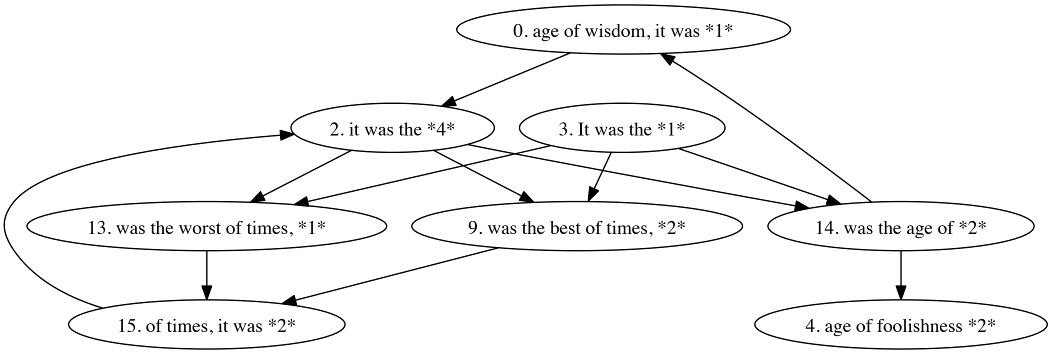
6.047 Fall 2012
10 October 2012

1. String Graphs and Manuscript Assembly



(a)

Above is the string graph pre-chain-collapse with transitive overlaps removed. Note that phrase frequencies are *starred*.


0. age of wisdom, it was *1*
2. it was the *4*
3. It was the *1*
13. was the worst of times, *1*
9. was the best of times, *2*
14. was the age of *2*
15. of times, it was *2*
4. age of foolishness *2*

(b)

And this is the string graph post-chain-collapse. The effective vertex count has been reduced from 16 to 8.

(c) The following **9** sequence reconstructions are all viable:

i. "It was the best of times, it was the best of times, it was the age of wisdom, it was the age of foolishness."

ii. "It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness."

iii. "It was the best of times, it was the age of wisdom, it was the best of times, it was the age of foolishness."

iv. "It was the best of times, it was the age of wisdom, it was the worst of times, it was the age of foolishness."

v. "It was the worst of times, it was the best of times, it was the age of wisdom, it was the age of foolishness."

vi. "It was the worst of times, it was the age of wisdom, it was the best of times, it was the age of foolishness."

vii. "It was the age of wisdom, it was the best of times, it was the best of times, it was the age of foolishness."

viii. "It was the age of wisdom, it was the best of times, it was the worst of times, it was the age of foolishness."

ix. "It was the age of wisdom, it was the worst of times, it was the best of times, it was the age of foolishness."

The algorithm is incapable of discerning which of these distinct resolutions is the true sequence due to repeats in the original dataset. We know with certainty that the text starts with "*It was the...*" and ends with "*age of foolishness*" because of in-degree/out-degree.

2. **Markov chains for finding conserved regions**

(a) It's much more likely that Alignment 1 came from N (1000x) and that Alignment 2 came from C (70x):

	P(N I ...)	P(C I ...)
P(... I a1)	**1.31E-07**	9.49E-11
P(... I a2)	2.50E-08	**1.75E-06**

(b) The probability of a false positive p(Clsequence) > p(Nlsequence) is surprisingly high (> 10%):
Relative frequences ~ **N = 8745 : C = 1255**

(c) Likewise: N = 1416, C = 8584

(d) From b and c we can conclude that there's a **substantial** (~ 10%) chance for error when classifying based on a simple likelihood estimate.

(e) A prior on how often a sequence was part of a conserved region would be extremely useful. This would allow us to make better approximations using Bayes' rule:

The prior described = P(C). P(N) = 1 - P(C), and from here, since we know P(seql C) and P(seq|N), we could determine the conditional probabilities P(Clseq) and P(Nlseq):

$$P(N|seq) = \frac{P(seq|N)P(N)}{P(seq)}$$

3. **HMMs for GC-rich regions: State durations and limitations**

(a) Expected state duration =

$$\frac{1}{(1 - a_{kk})}$$

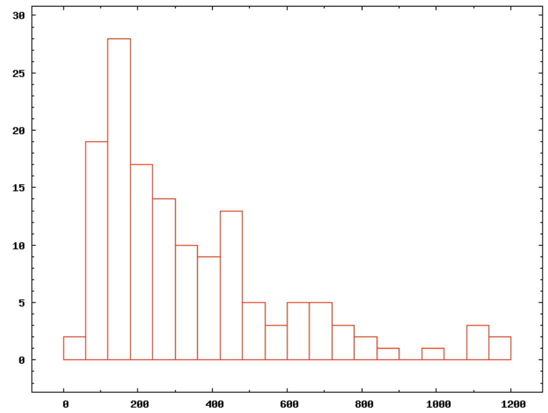
The distribution of state durations is represented by:

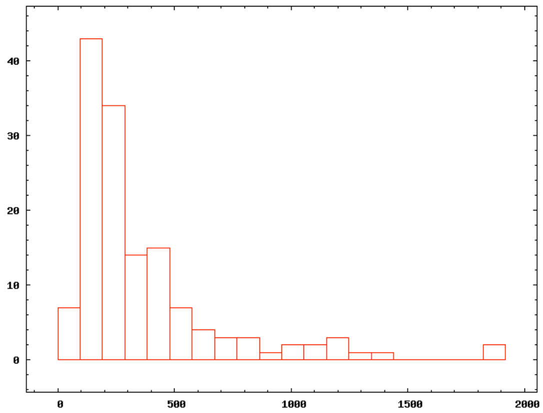
$$P(D_k = d) = (a_{kk})^d (a_{kl})^{n-1}$$

That is to say, in our simple 2-state system: the probability that after n transitions the state duration will be exactly d is equal to multiplying a_kk d times by the product of transitioning away from k multiplied n-1 times. Naturally, you can determine the distribution of probabilities for d <= some value by summing over [0...d].

(b) The expected state durations for high GC and low GC regions are identical, since tr[0][0] = tr[1][1]. The value = 1/(1-0.99) = **100**.

My implementation of the Viterbi algorithm yields **83.36%** accuracy for the hmmgen dataset. Here are my graphs for the *high_gc* data and *low_gc* data (left and right, respectively):





(c) The table below summarizes the differences between the three mystery datasets:

	mystery1	mystery2	mystery3
High-GC mean length	100	100	100

High-GC base composition	A=20.21%
	G=29.45%
	C=29.60%
	T=20.74%
		A=19.85%
		G=29.78%
		C=30.07%
		T=20.30%
			A=19.81%
			G=29.71%
			C=30.56%
			T=19.91%

Low-GC mean length	101	99	100

Low-GC base composition	A=29.87%
	G=20.27%
	C=19.73%
	T=30.13%
		A=29.84%
		G=19.86%
		C=19.99%
		T=30.31%
			A=29.56%
			G=20.09%
			C=20.11%
			T=30.24%

Viterbi Accuracy	71.89%	68.66%	67.85%

As you can see, the base composition and mean lengths for both types of regions were nearly identical across the three mystery samples. **But** in examining the charts, I see that the *distribution* of composition varies immensely between the three samples. In **Mystery1**, both + and - lengths are distributed squarely between 60 and 140, whereas the *majority* of the + and - regions in **Mystery2** are at 100, with a steep symmetric bell-curve level off. The regions in **Mystery3**, on the other hand, are all of identical length, with *no* regions of size < 100. **Just looking at the means is deceptive**! Adding a standard deviation statistic would help to elucidate the data at a glance.

The Viterbi prediction consistently yielded a mean-region length around 200: double the true mean length for the training data.

(d) Doing supervised learning on the HMM parameters using the mystery sequences as training data would likely result in overfitting and could actually further skew the data, since the three mystery sequences have such distinct distribution shapes. The base emission probabilities already perfectly match the training data, and the transition probabilities already correspond precisely to the average mean-lengths. I think the best bet for improving the HMM is by *changing* the model to increase the state-space. Instead of having just two states +, and -, we could, as discussed in lecture, retrain the data on a state-space of + and - conditioned on the previous base being A, T, C, or G, increasing the size of Q from 4 (2*2) to 16 (4 * 4).

Extra Credit: Just for fun, I experimented with adjusting (decreasing) the self-transition probabilities on the intuition that the initial hard-coded probabilities were yielding an average mean region length of double the authoritative value. I found $a_{++} = a_{--}$ = **0.9** to be idea, yielding an average accuracy around **84%** for both the mystery data and the hmmgen training set. This is a substantial boost in accuracy given that I adjusted a single parameter. Somewhat unintuitive-ly, this corresponds to an expected mean region length of 10, but consistently yielded an actual length close to 100. This indicates to me that the current model consisting of two states with fixed transition probabilities does *not* accurately portray the true transitions.

(e) Relevant quotes from the Burge paper:
"Similarly, we use a general three-periodic (inhomogeneous) fifth-order Markov model of coding regions rather than using specialized models of particular protein motifs or data base homology information."

Essentially, their state space is substantially larger and attempts to map states to the actual underlying biology, with different C + G % values corresponding to four distinct **Isochores**. For each group, a unique *q* parameter was estimated.

4. **Final Project Preparation**

(a) Done.

(b) Very interested in *Study of Local Variation of Mutation Rates and its Mechanism:* would be fun to expand this premise beyond DNA even, to examine RNA virus mutation rates and differences. Examining the mutation of cancer cells would also be revealing. Also could see expanding this to investigate which regions of plant genomes mutate most readily under artificial evolutionary pressure via seed irradiation.

Intrigued by the general idea presented in *Reconstructing ecological environments from microbial genomes* of using biological data to predict *abiotic* factors. Could see approaching the reverse problem: using environmental data to predict variations in microbe genomes across the planet (e.g. variations in cyanobacteria in parts of the ocean with varying light/temperature/CO_2 concentrations).

(c) Very interested in improving the tooling for visualizing genomic data. A paper presented in BMC Bioinformatics, *A web-based multi-genome synteny viewer for customized data*, presents a new approach for interactively examining genes within a chromosome. It expands the notion of synteny to allow for comparison between chromosomes of multiple genomes simultaneously. This is a nice first step, and I'd like to see these tools become increasingly available, both for use by the research community and as learning tools for newcomers. I'm surprised by the general lack of self-descriptive visual tools for qualitatively examining genomic data. I'd like to expand this to allow for visual-representation of HMMs, phylogenetic-tree parameters, and the likes.

The paper, *piRNA-mediated transgenerational inheritance of an acquired trait*, in the latest edition of *Genome Research* piqued my interest after hearing about piRNA in class. Exciting to see ongoing new discovery of such a fundamental and still poorly understood component of genetic regulation. I'd like to take what we know and start building models of how piRNA behaves from a systems perspective, but I'm guessing that experimental data may still be lacking here.

(d) Ongoing...

(e) I'll probably approach: Sarah Brockmueller, Stephen Serene, and *Lydia Krasilnikova*. Would be happy to explore the strange world of sex determinism in other organisms, e.g. honeybees.