[redacted]

CS121 / Chem 160A Quiz

Release v10

Christopher Lee

November 27, 2017

Contents

Instructions

- write your name and university ID number at the top right of each page.
- **IMPORTANT**: if you are an undergraduate, your exam should be titled "CS 121 / Chem 160A Quiz". If you are a graduate student, your exam should be titled "CS 221 / Bioinfo 260A Quiz".
- For full credit, show your work and briefly explain your reasoning where appropriate.
- note that there is a glossary and equation sheet included at the end of the quiz.

[redacted]

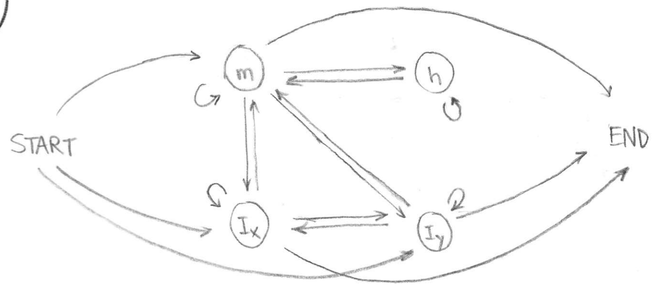
1. **Detecting hyperconserved regions in a pair of genomes**

Genome alignments reveal so-called *hyperconserved regions*, segments with a much higher level of nucleotide identity than found on average in the rest of the genome alignment. We can seek to identify such hyperconserved regions during alignment of a pair of genome sequences $\vec{X}, \vec{Y}$, as follows.

- in addition to the standard pairwise alignment states m (normal match state that emits aligned letter pair X_t, Y_u), I_x (emits one letter X_t and no letter of Y), I_y (emits one letter Y_u and no letter of X), we define a hyperconserved match state h, which acts like m except that it allows a much lower probability of $X_t \neq Y_u$ than m does).
- as usual, we allow all possible transitions between m, I_x, I_y, but only allow h to transition to/from itself and m.
- as usual, we are given the emission probabilities, transition probabilities and priors for all four states.
- we consider all possible alignments of sequences $\vec{X}, \vec{Y}$ using this four state HMM, and interpret letters emitted by the h state as hyperconserved.

(a) Draw the state graph for this HMM.

5



[redacted]

(b) Say we are given two sequences $\vec{X}, \vec{Y}$ and wish to find the optimal (maximum probability) alignment path through our HMM for these sequences. Write the Viterbi optimization rule for the *m* state to emit letter pair X_t, Y_u, explicitly showing what set of incoming Viterbi values it must consider.

3.5

Call the emitted letter pair O_t. Then the Viterbi optimization rule is

$X \quad Y \quad t$

$$V(m @ O_t) = \max_{\pi_{m@O_t}} P(O_t | m)\, P(m | m @ O_t)\, V(\pi_{m@O_t}).$$

-1.5

where the $V(\pi_{m@O_t})$ term takes on incoming Viterbi values. Namely:

$\{ V(m @ O_{t-1}), V(h @ O_{t-1}), V(I_x @ O_{t-1}), V(I_y @ O_{t-1}), V(\text{START} @ O_{t-1}) \}$.

↳ empty case

where $O_{t-1} = (X_{t-1}, Y_{u-1})$.

-1

$\vec{X}, \vec{Y}$ allows t and $u-1$ @ I_x

$t-1$ and u @ I_y

[redacted]

(c) Say you want to calculate the probability that letter pair X_t, Y_u are aligned as part of a *hyperconserved region*. Define what specific conditional probability you would calculate, and briefly explain your reasoning.

~~2~~

1.5

Call the emitted letter sequence $\vec{o}$. I would find the posterior probability $P(\theta_t = h | \vec{o})$ because the question asks for the hidden state probability given the observations.

-0.5

not necessarily $= t$ in X_t

(d) Briefly explain an efficient method for calculating this probability, indicating what specifically makes it computationally efficient, and its computational complexity in big-O notation, in terms of the length *L* of the sequences X,Y, and the number of states *S* in the HMM. You do *not* need to derive equations for the algorithm.

5

$$P(\theta_t = h | \vec{o}) = \frac{P(\theta_t = h, \vec{o})}{P(\vec{o})}$$

To compute this, we need the forward and backward probabilities at each time. Each of those 2L probabilities can be computed in S^2 time, so the complexity is $O(LS^2)$. This is efficient because the computation can use prior results (dynamic programming).

1 Basic Genetics

DNA A DNA sequence is a string consisting of a four letter alphabet (A, C, G, T). The four "letters" are called nucleotides or bases.

gene a specific substring of the DNA that encodes a specific functional unit called a protein.

genome The total DNA sequence of an organism. The genome of a small bacterium is about 4 million letters long and contains about 4000 genes. The human genome is about 3 billion letters long, and contains about 25,000 genes. The length of a DNA string is given in "base-pairs" (bp; just means a single letter); kilobases (Kb, 1000 letters) or megabases (Mb, a million letters).

chromosome A large genome usually consists of several separate pieces (each a single DNA chain) called chromosomes. e.g. the biggest human chromosome (chromosome 1) is 247 Mb long, and humans have 23 chromosomes.

genome copy number Some organisms keep a single copy of the genome in each individual, e.g. bacteria. Animals and plants generally have two distinct copies of the genome in each individual. These two copies can have (somewhat) different sequences, see *polymorphism*.

mutation an alteration of the DNA sequence that changes one or more letters in the DNA string. Mutation occurs at a low rate over time (roughly 10^{-8} letters change per generation). Some mutations change only a single letter, while other mutations might insert or delete a substring of the DNA.

sequence specifies function The specific DNA sequence of a gene is essential for its normal biological function; if that sequence is altered by a mutation it may not be able to perform that function normally or at all.

phenotype A specific, observable change in function due to a change in the DNA sequence.

fixation Over time, a mutation can either be permanently lost from the population (due to the individual(s) with that mutation for some reason not passing it on to the next generation), or it may become the majority of the population or even found in *all* individuals in the population. In the latter case it is said to be "fixed" in the population.

polymorphism Mutations in a population that have not yet been lost or "fixed" are called "polymorphisms", which simply means that some individuals have the mutation while others do not.

allele frequency the probability of finding a specific polymorphism on any given copy of the genome in a population.

SNP A single nucleotide polymorphism (SNP) is a single letter in the DNA sequence that differs between individuals.

polymorphism copy number For species that have more than one copy of the genome in each individual, the number of copies in an individual that have a specific polymorphism is called its copy number. In people the possible values are 0, 1, or 2.

recessive mutation Because of this, most mutations that damage a gene function are "recessive", which means that *both* copies of the gene must have a damaging mutation, in order to produce the mutation's phenotype (i.e. disease symptoms). An individual with one mutated copy plus one normal copy would simply have a normal phenotype (i.e. no disease symptoms).

dominant mutation A mutation that can cause its phenotype even if it is present on only one of the two copies is said to be dominant.

Mendelian inheritance the standard pattern of gene transmission in plants and animals, in which each individual has two copies of a gene (one copy from each of its two parents), and passes on one of its copies (chosen at random) to each of its children.

clone If two individuals are genetically identical, we say they are *clones* (of each other). Asexual organisms (such as bacteria) typically have only one copy of the genome, and reproduce clonally, i.e. the "daughter" cell is an exact genetic copy of the "mother" cell. Biologists often separate a complex population of cells by "plating" them at low density (say 1 cell / cm^2) and letting them grow; each cell will make a colony; each colony is clonal. Note that human clones (i.e. exact twins) are not very rare.

2 Basic Molecular Biology

5' end, 3' end A DNA chain has inherent asymmetry, and its two different ends are referred to as 5' (pronounced "five-prime") vs. 3'. By convention, DNA sequences are always written from 5' to 3'.

base pairing the physical attraction and binding of two DNA chain segments that are reverse-complements of each other. This can occur between two separate DNA molecules, or (due to the physical flexibility of the DNA chain) between two separate segments of a single DNA molecule.

reverse complement the specific DNA string that will naturally base-pair with a given DNA sequence: it is formed by the following rule: A pairs with T; G pairs with C; and the paired DNA strings run in opposite directions, i.e. the reverse complement of ATTGC is GCAAT, with the A in the first string base-paired with the T in the second string, and the C in the first string base-paired with the G in the second string.

double-stranded DNA that is base-paired with its reverse complement is said to be "double-stranded"; DNA in a cell is usually found in this form. This is the famous "double helix".

hybridization The process of single-stranded nucleotide sequence encountering a reverse-complement nucleotide sequence, and binding to it via base pairing (resulting in a double-stranded complex). Note that even in a mixture of a huge number of different sequences, those that are reverse-complementary will quickly find and bind to each other (i.e. "hybridize").

DNA microarray To measure the individual amounts of many specific sequences in a sample, a microscopic dot of a specified reverse-complement ("probe") sequence is attached to the surface of a glass slide; millions of such dots are arrayed microscopically. A fluorescently labeled DNA sample is then allowed to hybridize to the array; each dot will only bind its specific target sequence. A laser scans the array to measure the fluorescence bound to each dot.

RNA a variant of DNA whose chemical structure is just slightly different. Whereas DNA is the "permanent storage" of the genetic code, RNA usually operates as an "active form" of the genetic code e.g. for producing proteins coded by the gene sequence.

transcription the process of copying the DNA sequence representing a gene, to an RNA sequence, often referred to as the "messenger RNA" or "mRNA" for that gene.

gene expression regulation of the activity (or "expression") of a gene by increasing or reducing the amount of mRNA for that gene.

protein a string composed of a 20 letter alphabet (the amino acids) encoded by translating the DNA sequence of a gene. Proteins perform most of the biochemical activities in a cell, and a specific protein sequence carries out a specific activity, thanks to its unique physical structure determined by its sequence.

codon a group of three consecutive nucleotides that encode a single amino acid according to the universal genetic code.

translation the process of making a protein, by making the string of amino acids specified by a string of codons in a gene.

short read sequencing a recent technology for rapid sequencing of entire genomes. Current systems can read up to 500 million short sequence fragments each approximately 150 nucleotides in length, in one machine run. Due to the intrinsic error rates in the sequencing process, many fragment reads must be compared to arrive at a reliable "base call" at each letter position in the genome. Experimentally, DNA is randomly fragmented, and the resulting "short read" sequence strings must be aligned to each other (computationally, based on matching strings from overlapping reads) as the starting point for all analysis.

sequencing coverage the average number of independent reads covering any position in the genome (or region) being sequenced. Typically, sequencing projects aim for coverage of 70x or higher.

3 Basic Probability

$$p(a) = \frac{|a \cap S|}{|S|}$$

$$p(a|b) = \frac{|a \cap b \cap S|}{|b \cap S|}$$

$$\sum_b p(a \cap b) = p(a)$$

$$p(X_1, X_2, X_3, \cdots, X_n) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2) \cdots p(X_n|X_1, X_2, \cdots, X_{n-1})$$

$$p(H|O) = \frac{p(O|H)p(H)}{\sum_H p(O|H)p(H)}$$

$$\log(p_1 + p_2) = \log p_1 + \log(1 + \exp(\log p_2 - \log p_1))$$

binomial distribution

$$p(m|\theta, n) = \binom{n}{m} \theta^m (1-\theta)^{n-m}$$

where

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

poisson distribution

$$p(k|\lambda) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Hypergeometric Distribution

$$p(m|N, M, n) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

p-value (for observed value k given hypothesis h_0)

$$p_> = p(K \geq k | h_0)$$

Bonferroni correction for N p-value tests

$$\alpha = \frac{\beta}{N}$$

4 HMM Equations

discrete Markov chain transition matrix

$$\vec{p}_{u+t} = \vec{p}_u T^t$$

balance equation

$$\vec{\pi} = \vec{\pi} T$$

detailed balance

$$\pi_i \tau_{ij} = \pi_j \tau_{ji}$$

Viterbi optimality

$$p^*(\vec{X}^t, \vec{\theta}^t) = \max_{\theta_{t-1}} p^*(\vec{X}^{t-1}, \vec{\theta}^{t-1}) p(\theta_t|\theta_{t-1}) p(X_t|\theta_t)$$

This can also be written

$$V(s_i@X_t) = \max_{\pi_{s_i@X_t}} V(\pi_{s_i@X_t}) p(s_i|\pi_{s_i@X_t}) p(X_t|s_i)$$

Forward-Backward Posteriors

$$p(\theta_t = s_i|\vec{X}^n) = \frac{p(\vec{X}^t, \theta_t) p(\vec{X}^n_{t+1}|\theta_t)}{p(\vec{X}^n)}$$

This can also be written

$$p(s_i@X_t|\vec{X}_{[1,n]}) = \frac{p(s_i@X_t, \vec{X}_{[1,t]}) p(\vec{X}_{(t,n]}|s_i@X_t)}{p(\vec{X}_{[1,n]})}$$

continuous-time Markov chain transition matrix

$$\vec{p}_{u+t} = \vec{p}_u e^{\Lambda t}$$

$$e^{\Lambda t} = I + t\Lambda + \frac{t^2\Lambda^2}{2!} + \frac{t^3\Lambda^3}{3!} + ...$$