Ben Siranosian
April 4, 2014

CSCI1820 MIDTERM

Problem 0

International Human Genome Sequencing Consortium Assembly

The International Human Genome Sequencing Consortium (IHGSC) produced a draft sequence of the human genome in 2001 using a bacterial artificial chromosome (BAC) based method. Four main steps were required to produce the draft sequence: generating BAC clones, selecting the clones to be sequenced, sequencing the selected clones, and assembling the draft genome.

BAC and P1-derived artificial chromosome libraries from several sources were used in the project, constructed through digestion of human DNA with restriction enzymes and incorporation into vectors. The libraries represented 65-fold coverage of the human genome overall.

A subset of clones was then selected to be sequenced. Digestion of a clone with the restriction enzyme HindIII produced an agarose gel unique to the clone and data that could be used to assemble clones into "fingerprint clone contigs." Fingerprint clone contigs were then ordered along the genome using previously generated maps and sequence tagged sites (STS, short regions easily PCR-amplified and used for mapping). A subset of clones was selected to be a minimally overlapping set of the genome where possible.

Shotgun sequencing was used to sequence the individual clones. The sequencing effort was spread out over many labs around the globe, each with a different sequencing protocol. However, it was agreed that the collection of draft clones sequenced by each lab should have an average coverage above fourfold, with no clone below threefold. Other labs used whole-genome shotgun sequencing to generate additional data that could be used to identify SNPs and provide linking information for assembly.

Assembling sequences from many different labs into a draft genome was no small task. First, the data was filtered contamination from other organisms and different clones. Next, sequenced clones were mapped along the length of the genome. In-silico digestion with restriction enzymes and comparison with the fingerprint clone contigs along with end sequences from the BACs was used to create an ordering along the genome. Fingerprint clone contigs were assigned specific chromosomes and positions through the use of STSs and preciously established maps. GigAssembler was used to assemble the segments into a draft genome. The program first assembles merged sequence contigs by considering sequences that overlap at one end or both ends (when one sequence is contained entirely in another). It then selects a path through the sequence contigs and orders them through use of additional data, such as paired end reads and ESTs. These are linked together to form sequence-contig scaffolds, which can be ordered within fingerprint clone contigs generated earlier.

Craig Ventner and Celera Genomics Assembly

Craig Ventner and Celera Genomics used a very different whole-genome shotgun sequencing approach to generate their draft sequence of the human genome. First, plasmid libraries were generated from the genome of each human donor. Libraries were generated at sizes of 2kb, 10kb and 50kb. Three qualities were representative of a high quality library: equal representation of all parts of the genome, few clones

without inserts, and minimal contamination from outside sequences. Next, each end of the plasmid inserts were sequenced with the didexoy method, yielding over 27 million reads with an average length of 543bp.

The brunt of the Celera project was assembling the results from shotgun sequencing. Both the Celera shotgun library described above and data from the BAC clone sequencing from the Human Genome Project were combined in two different assembly methods. In the first, HGP data was processed into a synthetic shotgun dataset and combined with Celera data to form an 8x covering of the genome. In the second, Celera and HGP data were combined into chromosomal segments that were independently assembled.

The actual Celera assembler algorithm has five main stages.

- **Screener:** Filters the data for repetitive sequences. Microsatellite repeats less than 6pb are marked but still used in the next steps. Interspersed repeats are screened out of the next steps but can be part of an overlap between unfiltered matching sequences.
- **Overlapper:** Compares every read against every other read in the dataset. Read ends that overlap by at least 40bp and fewer than 6% mismatches are considered overlaps. The overlapper will produce "repeat-induced overlaps" from sequences that are repeated a few times in the genome and must be filtered in the next step.
- **Unitigger:** First identifies contigs that are uncontested by other reads in the dataset, calls them unitigs. Then compares the coverage of unitigs to the average sequencing coverage. If coverage is too high, the contig must have come from a repeat and is filtered out. Contigs with high but not certain probability of being correct are kept if they consistently scaffold.
- **Scaffolder:** Uses mate-pair information to assemble subcontigs into scaffolds. Subcontigs with 2 or more matches from the 2kb or 10kb mate pairs are assumed to be correctly spaced and ordered. Larger scaffolds are formed with information from the 50kb mate pairs and BAC sequences.
- **Repeat Resolver:** The majority of gaps between scaffolds correspond to repeats. First, contigs with a high but not certain probability of being correct are placed in gaps where at least mate pairs indicate their position. Next, when a minority of reads placed in a given gap don't agree with the consensus they are removed. Finally, remaining gaps are filled with BAC data from HGP.

Comparison of the Two Assemblies

The IHGSC and Celera used vastly different methods to arrive at a draft sequence of the human genome. Both groups used three categories of methods: clone generation and selection, sequencing and assembly. In a broad perspective, the IHGSC focused most of their effort in the steps before sequencing while Celera developed a complicated assembly algorithm to do most of the work. Differences in the methods produced draft human genome sequences with different characteristics. For example, the IHGSC method was better at sequencing exact repeats and produced a higher quality assembly for the X and Y chromosomes. Overall, though, the Celera assembly performed better than the IHGSC effort. When compared to an updated NCBI-34 sequence by Istrail et. al., the Celera draft placed at least 79% of the sequence in the correct order and orientation, compared to 74% for the IHGSC.

Ben Siranosian
April 4, 2014

Repeats and Genes

The IHGSC estimated at least 50% of the genome was repetitive sequences, composed of transposable elements, processed pseudogenes, simple repeats, segmental duplications and tandem repeats at certain chromosomal locations. The Initial Integrated Gene Index was estimated to contain about 24,500 true genes.

Celera found about 35% of the genome to be repetitive, a much lower estimate than the IHGSC. They noted that some repetitive sequences might be underrepresented because of the repeat resolution algorithm, however, and suggested that much of the gaps in the assembly would be repetitive sequences. Celera reported 26,383 high confidence genes, but noted that the number could have an upper limit around 40,000.

The number of human genes is surprisingly low. Scientists originally thought organisms of increasing complexity would have a proportional increase in the number of genes. However, humans only have about twice the number of genes as the worm or fly, as noted in the IHGSC paper. Additionally, scientists thought the large size of the human genome would correspond with a proportionally large number of genes, but this was not the case. Repetitive sequences can account for much of the increase in length. A high number of splice variants and greater degree of gene regulation can be used to explain the complexity of humans without an increasingly large number of genes.

Additional Questions

Large regions of the genome remained unassembled in the draft presented by both groups. Theoretically, these regions could have been sequenced, especially with the whole genome shotgun method (there are probably difficulties incorporating repetitive, centromeric and telomeric sequences into vectors, though).

Difficulties arose in both the wet lab and computational steps. The IHGSC had to coordinate the sequencing effort of several labs simultaneously, each using their own sequencing protocol. On the assembly side, repetitive sequences posed the largest challenges.

The Celera assembly depended on the accuracy of the algorithm and statistical models behind it. Nobelist Hamilton Smith was worried about UV light introducing any un-accounted for errors into the sequencing process. Errors not modeled in the algorithm could reduce the quality of the assembly.

Problem 1

I assembled the following super-contigs from the provided data. Contigs are enumerated by their zero-indexed position in the original file.

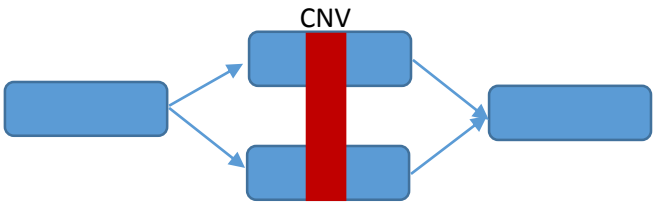
0 -> 7 -> 5 -> 6 -> -4 -> 2 -> 10

1 -> 9 -> 3 -> -8

1) I was able to assemble all contigs into two super-contigs.
2) I am confident in the ordering and orientation of contigs within my super-contigs. However, it would be good to find a link between the two. To do this we would have to sequence a mate pair that maps to a location in both super-contigs. Increasing sequencing coverage would be the best way to do this – by increasing coverage, you increase the probability that a read will span the gap between the two super-contigs (provided they are close enough to be captured by a

mate pair, increasing coverage won't help if there's a centromere between the two or something).

3) If mate pairs were generated from a diploid genome it adds an extra layer of complexity. Instead of being from either the forward or reversed strand, the read could come from either strand on either chromosome copy. It's fair to assume most of the genome is homozygous, so this wouldn't be a problem in most cases. Heterozygosities like SNPs or copy number variations would complicate things, though. A scaffold could have 'bubbles' similar to contig assembly (see picture below).



4) I used blastn against the nr database to look for similarity. These contigs come from ebolavirus, one of the most deadly viruses in the world. This project interestingly correlates with the outbreak happening in Africa right now…

Problem 2

1) a) The evolutionary distance between the two species may not accurately reflect the divergence time of the amino acid sequence. This would be more common with horizontal gene transfer between bacteria. Incorrectly estimating evolutionary distance would lead to choosing the wrong substitution matrix and being biased towards a result with more accepted substitutions in the case of a later horizontal gene transfer. Additionally, the organisms being compared may be biased for certain amino acids or substitution types.

   b) One could calculate a substitution matrix specifically for the two organisms being compared, much like Dayhoff did with the original PAM matrix, and use this to align the amino acid sequences.

2) a) Are we trying to figure out if certain sequences are selected for or against in secondary structures with the first claim? If so, bias could come from proteins selected for a particular function. The second claim could be biased by a number of biological factors, such as availability of amino acids in different organisms or different nutritional settings or tRNA bias in an organism. Some amino acids are coded for by several codons, while others like tryptophan are represented by only a single codon.

   b) Bias in the first claim could be eliminated by looking at proteins that are not selected for a particular function. If all amino acids form secondary structures, selection on function is driving the high frequency. If only the high frequency amino acids form secondary structures, sterics are probably hindering the less frequent amino acids from forming them. Bias in the second claim could be eliminated by taking extra parameters, like the number of codons coding for an amino acid, into account. It might also help to build a model for each different organism being compared.

3) a) The exon shuffling hypothesis could be biased because "of course most proteins came from common ancestors, because all organisms came from common ancestors!" You'd be biased towards predicting this result. The real question would be to ask if more proteins have been created from evolution from archetypal proteins than would be expected, given the evolutionary conservation between organisms.

   b) Correcting for bias in the exon shuffling hypothesis would be difficult. Between two organisms, a certain part of the genome is expected to be explained by evolutionary conservation and the other part by random events. The claim could be verified by comparing the conservation in new protein sequences against the conservation in sequences genome wide.
4) a) Percent identity can depend on the substitution matrix used. A certain percent identity with a stringent substitution matrix on a short sequence may be more significant than a alignment with a less stringent matrix but longer sequence.

   b) When talking about percent identity it is also necessary to mention the substitution matrix that is used, which should eliminate the bias described above.
5) a) The protein data bank stores structures of sequences previously researched by scientists, meaning they were found "interesting" enough to compute the crystal structure. Comparing a random protein against this database could be biased for structures that are similar in some way or come from the same few (model) organisms.

   b) To verify bias in the Protein Data Bank, one could compare a database of proteins from randomly selected genomic sequences. If proteins from certain organisms or with certain features show up more frequently than expected under random chance, the bias would be verified. To correct for this bias

Problem 5

a) $I(p_{xi})$ takes on a minimum value of 0 when $p_{xi} = 0.25.$ 0.25 is the minimum value p_{xi} can have because it is the proportion of sequences containing the consensus – if it were reduced, another base would rise in frequency and become the consensus. When $p_{xi} = 0.25,$ zero bits of information are recorded because the consensus sequence is no better than random. Every base has the same frequency in this case.

b) $I(p_{xi})$ takes on a maximum value of 2 when $p_{xi} = 1.0$. In this situation the consensus agrees completely and a maximum amount of information is recorded. Every base other than the consensus has zero frequency in this case.

c) $I(p_{xi})$ takes on a value of 1 when $p_{xi} \approx 0.68342$. Half as much information is recorded in this case compared to a completely agreeing consensus. Unlike the other two cases, the frequencies of the other bases are free to vary within a total of 1- $p_{xi.}$ Encoding one bit of information isn't a special case like the last two.

Problem 6

First, I identified three discrete syllables that make up the Oriole birdsongs - high, medium and low (*h,m,l* from now on). The syllables are easy to differentiate (the alert song begins with *hhhmmmll*, for example) and represent a good choice for emissions from the HMM.

Next, I had to identify the state space. Two structures of multiple syllables were observed in each song type. In the normal song: *mlmhl* and *lml*. In the alert song: *hhhmmm* and *llmmmmhhhmmmm.* These four structures represent the state space, called *n1, n2, a1, a2* from now on. Thus, we have a four state HMM with states *n1* and *n2* corresponding to the normal call and states *a1* and *a2* corresponding to the alert call.

Transition probabilities: Ideally, each state should persist for the number of syllables in the structure it was identified from. The second most probable state transition should be to the other state from the same song. We don't have any training data about transitioning from one song to another, but I assume this will happen with a small but nonzero probability. The following is an estimated transition matrix. Diagonals are (number of syllables in the structure normally emitted by the state) / (number of syllables in the structure normally emitted by the state +1)

	n1	**n2**	**a1**	**a2**
n1	0.833	0.20	0.025	0.025
n2	0.117	0.75	0.025	0.025
a1	0.025	0.025	0.857	0.031
a2	0.025	0.025	0.093	0.929

Emission probabilities: Each state has emission probabilities calculated by the frequency of the syllable in that state.

		States			
		n1	**n2**	**a1**	**a2**
Syllables	**h**	0.2	≈ 0	0.5	0.23
	m	0.4	0.33	0.5	0.62
	l	0.4	0.67	≈ 0	0.15

To find the most probable sequence of hidden states given a sonogram and my HMM parameters, I would use the Viterbi algorithm.

Problem 7

To determine what inspired Coco to sing, we're going to need high quality samples from the sources we think inspired him. So, cockatiel songs, people with Albuquerque accents, Beatles records, samples from TV, other birdsongs, and samples of Sorin's beautiful voice. We're also going to need audio from Coco singing. These samples should be digitized and discretized in some way as to pick out the individual symbols. Looking at the sonograms like problem 6 would be a good place to start.

Next, I would build a local alignment algorithm that takes sonogram data and compares the syllables. It might be useful to train this algorithm on language data from other sources, such as matching people with the same accent. It's important that this is a local alignment algorithm because different parts of Coco's

song may match different sources with gaps and mismatches. With some smart coding this algorithm could guess what inspired the bird to sing!