6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

Lecture 15 - Comparative Genomics I: Genome annotation

11/23/08

1 Introduction

This lecture and the next will discuss the recent and current research in comparative genomics being performed in Professor Kellis' lab. Comparative genomics allows one to infer understanding of genomes from the study of the evolution of closely related species, and vice-versa. This lecture will discuss the use of evolution to understand genomes, and lecture 16 will deal with using genomes to better understand evolution. By understanding genomes, we mean primarily to annotate the various parts: protein coding regions, regulatory motifs, etc. We'll see later that comparative genomics also allows us to uncover completely new ways various elements are processed that we would not recognize using other methods.

In Dr. Kellis' lab, mammals, flies and fungi are studied. Slide 6 shows the many species that are part of the data sets that are analyzed. We want to study a wide variety of organisms to observe elements that are at different distances from humans. This allows the study of processes at different ranges of evolution (different snapshots in time based on divergence point). There are several reasons why it is important to have closely related species as well as more distantly related species. More closely related species should have very similar functional elements and randomness in the non-functional elements. This is because selection weeds out disrupting mutations in functional regions, and mutations accumulate in the non-functional regions. More distantly related species will likely have significant differences in both their functional and non-functional elements. Phylogeny allows observation of individual events that may be difficult to resolve in species that are more separated. However, our signal relies on the ability to identify/observe an evolutionary event, thus if we look only at species that are close, there won't be enough changes to discriminate between functional and non-functional regions. More distantly related species allow us to better identify neutral substitutions.

2 Preliminary steps in comparative genomics

Once we have our sequence data (or if we have a new sequence that we wish to annotate), we begin with multiple alignments of the sequences. We BLAST regions of the genome against other genomes, and then apply sequence alignment techniques to align individual regions to a reference genome for a pair of species. We then perform a series of pairwise alignments walking up the phylogeny until we have an alignment for all sequences. Because we can align every gene and every intergenic region, we don't just have to rely on conserved regions, we can align every single region regardless of whether the conservation in that region is sufficient to allow genome wide placement across the species. This is because we have 'anchors' spanning entire regions, and we can thus infer that the entire region in conserved as a block and then apply global alignment to the block.

3 Evolutionary signatures

Slide 10 shows results for nucleotide conservation in the DBH gene across several species (note that other species that are not show on the slide were also used to calculate conservation). To calculate the degree of conservation a hidden Markov model (HMM) was used with two states: high conservation and low conservation. The Y-axis shows the score calculated using posterior decoding with this model. There are several interesting features we can observe from this data. We see that there are blocks of conservation separated by regions that are not conserved. The 12 exons are mostly conserved across the species, but certain exons are missing (e.g. zebrafish is missing exon 9). Certain intronic areas have stretches of high conservation as well. We also note the existence of lineage-specific conserved elements. If there's a region that's thought to be intronic but still appears to be highly conserved, then this is evidence for that region being functional.

We want to develop evolutionary signals for each of the functional types in the genome. The specific function of a region results in selective pressures which give it a characteristic signature of insertions/deletions/mutations. Protein-coding genes exhibit particular frequencies of codon substitution as well as reading frame conservation. RNA structures have compensatory changes to maintain their secondary structure. μRNAs look different from RNA genes, here paired regions are not undergoing the compensatory changes that occurred above, they are very highly conserved. Intermediate regions are able to diverge. Regulatory motifs are not conserved at the exact same position, they can move around since they only need to recruit a factor in a particular region. They show an increased conservation phylogenetically across the tree, while showing small changes that preserve the consensus of the motif, while the primary sequence can still change. This lecture will discuss further how to determine protein-coding signatures.

4 Protein-coding signatures

Slide 12 shows a region of a gene near a splice site. The same level of conservation exists on both sides of the splice site, but we notice significant differences between the sequences to the left and to the right of the splice site. Recognizing these differences allows us to construct our signature. To the right of the splice site, gaps occur in multiples of three (thus conserving the frame), whereas to the left of the splice site frame-shifting occurs. There is also a distinct pattern to the mutations on the right side, as the mutations are largely 3-periodic and certain triplets are more frequently exchanged. As a bonus, by being able to recognize the change in regions, our splice site becomes immediately obvious as well. By testing for (i) reading-frame conservation and (ii) codon-substitution patterns, we can identify protein coding regions very accurately.

4.1 Reading-Frame Conservation

By scoring the pressure to stay in the same reading frame (i.e. no gaps that are not multiples of three), we can easily quantify how likely a region is protein-coding or not. Staying in the same frame is obviously important in a protein coding region, as a frame shift would completely alter the amino-acid structure of the protein. There are two methods that can be used to do this: a cutoff method or a scoring method. In the cutoff method we penalize gaps if they are not in multiples of three. We can do this by selecting a segment of the gene to see how many gaps satisfy this condition. We need to perform this three times over all possible offsets and then choose the best alignment. When penalizing the gaps, it is important to skip gaps that are not multiples of three if they have compensating gaps in their pre-specified neighborhood such that the sum of the gaps is still a multiple of three. In the scoring method we count the number of nucleotides that are out of frame and normalize by the length of the gene. The percentage of nucleotides out of frame is very high in the non-coding regions, while the number is very low in the coding regions. These methods do not work as well if there are sequencing errors. We can compensate for these errors by using a smaller scanning window and observing local reading frame conservation.

This method has been applied to the yeast genome, with very good results. The method was shown to have 99.9% specificity and 99% sensitivity, and when applied to 2000 hypothetical in yeast ORFs (open reading frames)[1], it rejected 500 of them as false positives. When applied to human genome, 4000 genes were rejected[2]. Both finding have been validated experimentally.

4.2 Codon-Substitution Patterns

The second signature is somewhat more elaborate, and measures the exchange rate of different codons. A 64 × 64 codon substitution matrix (CSM) is created to measure how often a specific triplet is exchanged in species 1 for another triplet in species 2. Slide 15 shows the CSM for genes and for intergenetic regions. A number of salient features present themselves in the gene CSM. Note that the main diagonal element has been removed, because the frequency of a triplet being exchanged for itself will obviously be much higher than any other exchange. We still see a strong diagonal element in the protein coding regions. We also note certain high-scoring off diagonal elements, these are substitutions that are close in functional spaces rather than in sequence space; either 6-fold degenerate codons, or very similar elements. We also note stripes of low values. These correspond to stop codons, so substitutions to this triplet would significantly alter protein function and thus are strongly selected against. One thing to note regarding this image: There is a CpG dinucleotide mutational bias in the human genome (due to methylation sites). The authors needed to guess the correct CpG frequency rate and subtract that from the image. There could be some remnants of this correction in the image. In intergenic regions the exchange rates are more uniform. In these regions, what matters is the mutational pattern, i.e. whether a change is one or more mutations away. Therefore, intergenic regions are dictated by mutational proximity whereas genetic regions are dictated by selective proximity. We can use these two matrices to create a likelihood ratio matrix. Slide 17 applies this test to aligned sequences, and this makes the protein coding regions immediately obvious.

An interesting feature of this method is that it automatically infers the genetic code from the pattern of substitutions that occur, simply by looking at the high scoring substitutions. In species with a different genetic code, for example in Candida albumin (for which CTG codes for serine (polar) rather than leucine (hydrophobic)), the patterns of codon exchange will be different.

[1] Kellis M, Patterson N, Endrizzi M, Birren B, Lander E. S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Science. 423: 241-254.

[2] Clamp M et al. 2007. Distinguishing protein-coding and noncoding genes in the human genome. PNAS. 104: 19428-19433.

However no knowledge if this is required by the method. Instead, we can deduce this *a posteriori* from the CSM.

Regarding implementation: These methods can be implemented using a HMM or conditional random field (CRF). CRF allows the integration of diverse features that do not necessarily have a probabilistic nature, whereas HMMs require us to model everything as transition and emission probabilities. CRF will be discussed in an upcoming lecture. One might wonder why these more complex methods need to be implemented, when the simpler method of checking for conservation of the reading frame worked well. The reason is that in very short regions, insertions and deletion will be very infrequent, thus there won't be enough signal to make the distinction between protein-coding and non-protein-coding.

5 Protein-coding evolution and nucleotide conservation

Finally, on slide 18 we look at simultaneous observation of protein-coding signatures and conservation. We see regions that have low conservation, but a large protein-coding signature, as well as the reverse. Identification of regions tagged as being genes but that did not have high protein-coding signatures helped to strongly reject 414 genes in the fly genome previously classified as CGid-only genes which led FlyBase curators to delete 222 of them and flag another 73. In some cases there were definite false negatives, as functional evidence existed for the genes under examination. In the data, we also see regions with both conversation as well as a large protein-coding signature, but that are not marked as being parts of genes. Some of these have been experimentally tested and have been show to be parts of new genes or extensions of existing genes.

Finally, comparative genomics allows the identification of new mechanisms of regulation. 150 genes in the fly and 5 in the human possess regions where a stop codon exists inside a region with a large protein-coding signature, and a second stop codon follows shortly after it. These are in brain proteins and ion channels. The reason for this is still unclear. There may be an increased conservation of secondary structure that favors translational read-through, or A-to-I editing of the stop codon.