

Eric discussed two topics on the broad theme of "signal to noise extraction":

- 1) Comparative genome analysis
- 2) Gene functionation

Both of these projects have grown out of projects that are on-going in the laboratory.

1. COMPARATIVE GENOME ANALYSIS

How much of the human genome matters? We don't know, but we could do systematic mutagenesis to discover what is important.

However, the natural mutation rate is 2×10^{-8} per base - about 10^2 mutations in human across generations. In 100 million years, you have 10 million generations, so about 20% of the bases change in a single lineage, or about 50% in an individual that is a product of two lineages. Thus, if we look at what has been conserved across time, we can hypothesize it is important.

Another method is to align genomes from mouse and human into synteny blocks where genes are in the same order (say 10-20 MBases). We can then find exons (150 bp) say eight, that make up a typical gene. Splice signals are weak, and thus it is difficult to recover genes using HMMs. However, if we consider mouse and human together, we can use a pair HMM that will take consistent information into account, and find the genes with higher specificity and sensitivity to 96% fidelity.

For example, about 170 features are conserved between the human and mouse HOX locus, one every 5KB, is conserved, from 100 - 1KB. These are not exons. In the synteny map between human and mouse there are about 250K conserved exons, and 250K regulatory features. We know what only a handful do. How do we discover function? The rat genome will be available by Christmas.

They have sequenced four Yeast species, with excellent synteny. The genes can be aligned quite well, and this alignment used for analysis. For example, there are about 400 fewer genes that have been suspected because they are not supported by other organisms. They are short, and only code for a few hundred amino acids.

The cross-species Yeast alignment provides sequence islands that are conserved and match up with known regulatory elements. For example, Gal4 motifs are four times more likely to be conserved in intergenic regions as compared with coding regions. If we create synthetic motifs (all triplet dimers with fixed gap) most are more conserved in coding regions, but certain of them are more conserved in intergenic regions. What do they do? We can see what genes they relate to, detect commonality of expression patterns from mRNA arrays. For example, we could test if all of the genes that are purported to be influenced by a given motif have significant variation as a group.

2. GENE FUNCTIONATION

How can we learn what the genes we discover do?

Looking at Cytochrome c Oxidase (COX) deficiency in Canada - children die by age 12 when they are recessive for the mutant allele. The product of this gene is used in the mitochondria, and is important for energy production. The actual allele is an assembly factor for COX. Mitochondria was originally an endosymbiote, and has been kicking genes out to the host genome. With about 12

generations, we would expect about 8% of the genome is retained around each marker, and thus we only need a few hundred markers to scan the genomes of our population for markers that correlate with the disease. We can look for ORFs in the regions that correlate with the disease. They identified all of the genes in the affected region, and detected which of these genes clustered with mitochondrial genes. They wound up purifying mitochondrial complexes, digesting them, and sequencing them on mass spec. He used the ~1000 conserved islands between mouse and human across the conserved locus for the DB for the mass spec digest analysis, and he found exactly one gene that co-expressed with known mitochondrial genes. Once he was done to the gene, he sequenced the gene and found the mutation.

In Yeast, they have been able to discover a motif that is upstream of mitochondrial genes. In human they have not been able to do this yet.

Their basic idea is that evolution is their laboratory, and they need to decide what species to sequence. They have chosen species at different distances, but they do not know what the optimum is. In the species tree, how can you describe what kinds of signals can be detected? What are the best tools? How can we attach meaning to these features. This are is an interesting area for computational research.

What we need to do is collect data sets under a wide variety of conditions, cluster genes, and attach meanings to those clusters. This would be an essential resource for future research.

Define a cluster C as a meta-gene. Might we derive meta-genes from quite a bit of data, and then explain the results of other studies in terms of these meta-genes? Such meta-genes might be far less noisy than an individual gene, and might permit more accurate explanations. Might these meta-genes directly represent biological pathways?

QUESTIONS

Why are these conserved non-exons? answer: enhancers, other regulatory regions, insulators, RNA genes, RNAi, small nuclear RNAs. "other stuff". Can we cluster these elements to find out what is in common among different genes?

How can we use intersecting lists? answer: the mouse/human alignment is just six weeks old, and only 24K genes have been placed on it. There is quite a bit of low hanging fruit, and is anybody interested in working on it?