Searching a genome

By Kin Chan

Identifying functional elements in a genome is important to biology. By tradition, this is done experimentally. In practice, however, experimentally determining gene function for an entire genome becomes impractical due to the shear number of genes for each species. This is further complicated by many regulatory elements per gene. Identifying all functional elements experimentally is a major undertaking for just a single species and likely to be cost prohibitive.

High quality sequences for entire genomes are now available. Being able to locate the functional elements by solely examining the genomic DNA sequence with computerized algorithms is a valuable technique.

Genomic elements required to sustain life is likely to be conserved across related species. Most random mutations tend to be lethal. Thus, evolutionary conserved sequences are likely to be functional elements. Furthermore, the greater the number of species sharing the same sequence, the more likely the sequence will be a functional element.

The researchers in this paper (Kellis, 2003, P. 241) used comparative genomic analysis to look for genes and other genomic elements by comparing the entire genome of several related yeast species. *Saccharomyces cerevisiae* was chosen because it is the

best known Eukaryote. Three related species were used for comparison, *S. paradoxus, S. mikatae* and *S. bayanus*.

The researchers first generated the genomic sequences that were not already available using a whole genome shotgun plasmid approach. The Arachne computer program was used to assemble a draft sequence.

The *S. cerevisiae* genome was aligned with the other three species of yeast. Attempts were made to make global alignments of the known genes of *S. cerevisiae* to the other three genomes to create landmarks between species. Of the 6,235 ORFs most have one to one correspondence to the other genomes and only 211 have ambiguous correspondences and they are mostly clustered in the telomeric regions of the 16 chromosomes.

With landmarks established, genes are found by observing conserved ORFs in analogous regions between the previously suspected genes in *S. cerevisiae* and the other three yeast species. For each gene, each of the three other species votes as to the validity of the gene. The other species votes yes, abstain or no based on level of conservation. The results indicate that the vast majority of the previously suspected genes are indeed genes with a unanimous vote.

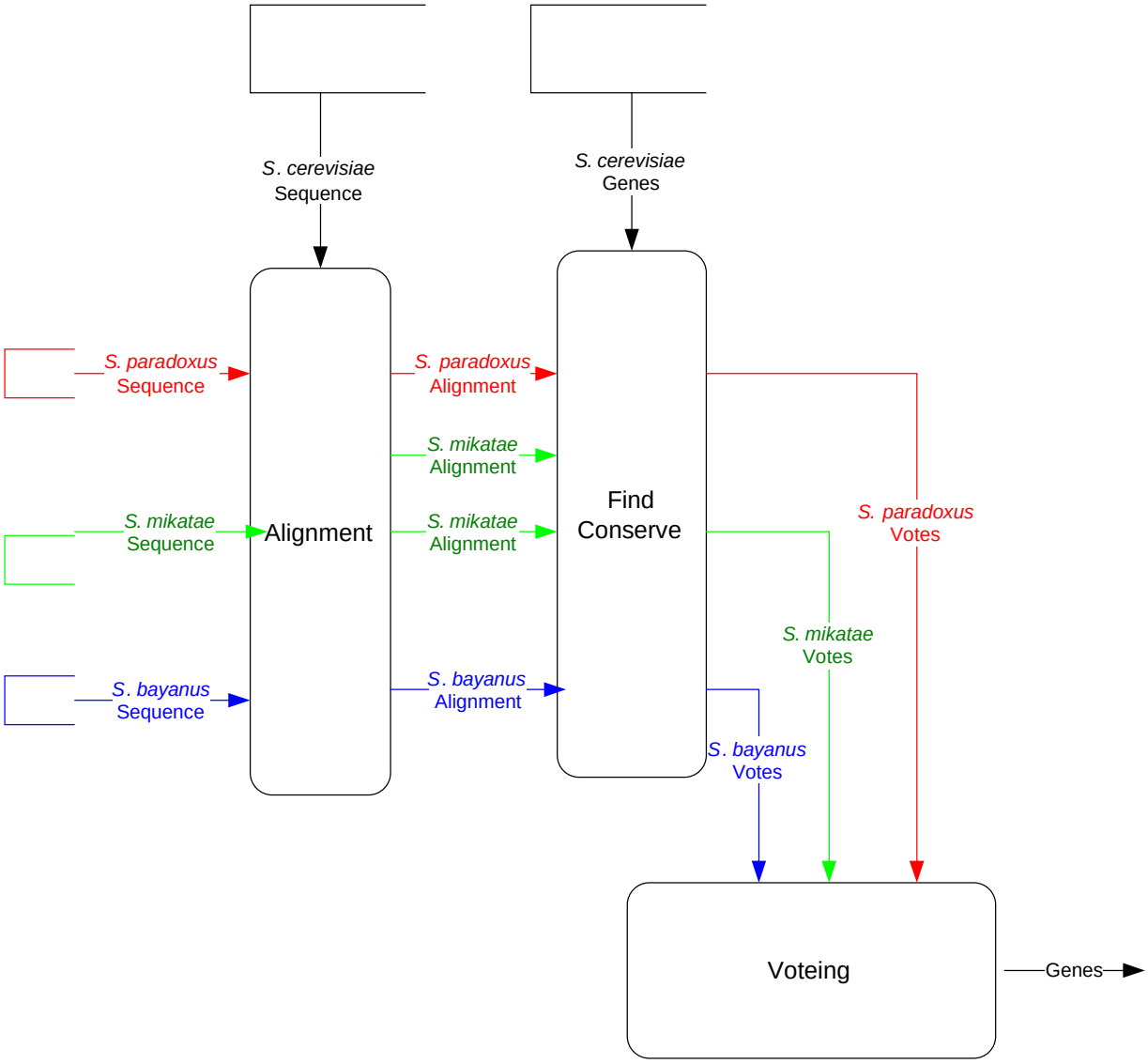
Comparative genomic analysis also finds new introns as well as better locating the start and stop codon. The donor, branch point and acceptor sites were all found to be highly

conserved for introns. Another result is that the location of the stop codon is more variable than the start codon. Comparative genomic analysis also showed that protein coding sequences evolve at a slower rate than non-coding regions.

Regulatory elements are binding sites for transcription factors and are harder to find. They are short and follow no known rules. The conserved non-coding regions can be compared to binding sites of known transcription factors. Another method is to look for conserved and similar short sequences near the start of genes that have common function. But this requires prior knowledge of gene function.

The overall C+G content in *S. cerevisiae* is only 38%(Normore, 1976). It is unclear what the CG content in a yeast random sequence would look like, but it is probably closer to 38% rather than 50%. Thus the estimate for the number of random stop codons would not be one in twenty as stated in the paper. Also the other probabilities will also changed base on the stop codon frequency. The clarity and completeness of the paper will be improved if the method of derivation of the probabilities were included. This is not a big deal since it is not the major point of the paper.

Overall, comparative genomic analysis is a very useful method to determine the functional elements of entire genomes without much prior knowledge of the genome.



This is a data flow of the voting algorithm. The data flow in colors, red, green and blue are independent from each other until it reaches the voting process.

Further Investagations

Here I describe a project in three phases to build on the work from the paper. I give the rationale and approach for the work.

1 What does changing the number of comparison species do to functional element finding? Perform the same analysis as described in the paper, except with just one of the species. Do this three times, an analysis for each species of yeast. Then three more times with the combinations of tow of the three yeast. Finally perform the analysis with the sequenced genome most related to *S. cerevisiae* along with the other three yeast so that there will be four separate related species for comparison. The reason for doing this is to gage the number of species required to get reasonable results. If similar results are achieved with just the evolutionally most distant species, we will have confidence in the technique when comparing say human and mouse. On the other hand, if we get tangibly better results when comparing with four species, we will be confident only when more mammals have been sequenced.

2 Increase sensitivity and selectivity by computing a score from the conservation step rather than a vote. The paper points out that some genes had conflicting votes when some of the species diverged and another didn't. The likely hood is that this is a real gene especially when the conserved version has a high conservation value. I am

thinking that an optimal scoring scheme can be derived from a statistical analysis of the genomes between the two species involved.

There are several types of mutations, insertions, deletions, inversions and single nucleotide mutations. The total number of each type of mutations from the base genome to the comparison genome can be computed given the sequence and alignment of the two genomes. This over the entire genome can be considered the average for the entire genome for that specific type of mutation. The idea is that functional sequences will have a below average number of mutations where non-functional sequences will have an above average number of mutations. Also, the mutation rate of any nucleotide to any other nucleotide can be computed from the base genome to the compared genome.

We can compute the mean and standard deviation for each type of mutation between the two genomes. For example, if there are five insertions in a million base pairs, the mean probability for insertions is 5/1000000. The standard deviation is computed by determining the variance in the distribution of the number of insertions in a sequence of the length of the gene. The mean and variance may be weighted by the length of the insertion. This is done for deletions, single nucleotide mutations and other type of mutations. There will be a weight for the different mutations because some types of mutations will be more common. Also some parts of the chromosome mutates at higher rates. The mutation probabilities should be specific for chromosome region.

Given a sequence in each of the species under comparison, compute the likelihood that one sequence would have mutated to the other based on genome wide statistics. This is the score.

Part of the future work is to clarify and refine the algorithm. The goal is to compute a p-value that describes the likelihood of conservation between regions in the two species. A high likelihood p-value in the most related species should be sufficient to declare a found gene regardless of the score from the other species.

The donor, branch point and acceptor sites for introns are conserved. We can use the same probability scheme to look for these elements. The same holds true for transcription factor binding sites.

3 Of course the paper discusses applying the method to the human genome. With the chimpanzee genome now sequenced (Mikkelsen, 2005, p. 69) along with mouse (Botcherby, 2002, p. 226); we can apply the analysis described in the paper on these genomes to discover the quality of the human annotation. The scoring method outlined in item 2 can be used instead of the voting scheme.

The technology to find binding sites have improved since 2003 (Fogel, 2005, p. 137). We now have about 800 known transcription factor binding sites for vertebrates. Each of these binding sites has a probability matrix describing the likelihood of each

nucleotide in each position. These algorithms can be combined with the genome wide search to find transcription factor binding sites. In addition, with so many binding sites already known, conserved regions in non-protein coding sequences near the start of genes that do not conform to known transcription factor binding sites are potential binding sites to undiscovered transcription factors.

References

Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren & Eric S. Lander (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements, NATURE, VOL 423, 15 MAY 2003 pp 241 – 251

Normore, W. M., Shapiro, H. S., and Setlow, P., (1976) CRC Handbook of Biochemistry and Molecular Biology. (Ed. G.D. Fastman), CRC Press.

Tarjei S. Mikkelsen et.al., Initial sequence of the chimpanzee genome and comparison with the human, genome, Nature, Vol 437, 1 September 2005 pp 69 – 87

Marc Botcherby, Just click on the mouse!, Briefings in Functional Genomics and Proteomics, Volume 1, Number 3, 1 October 2002, pp. 226-229(4)

Gary B. Fogel, Dana G. Weekes, Gabor Varga, Ernst R. Dowb, Andrew M. Craven, Harry B. Harlow, Eric W. Su, Jude E. Onyi, Chen Su, A statistical analysis of the TRANSFAC database, BioSystems 81 (2005) 137–154