ThaiBinh Luong
CBB 752
12/10/05

Sequencing and comparison of yeast species to identify genes and regulatory elements

Although genomes have become fairly easy to sequence, the ability to directly interpret genomes is still undeveloped. For example, we have known the complete genome sequence of Saccharomyces cerevisiae for nearly ten years. However, the number of biologically significant open reading frames is still uncertain. The identification of regulatory elements is even more of a challenge. Currently, the best approach is to cluster genes into functionally related subsets, then search for common sequence motifs in the general vicinity of the genes. Unfortunately, this requires extensive prior knowledge about the gene function.

Comparative genome analysis of related species provides a powerful and general approach for identifying functional elements without much prior knowledge of function. Past genomic comparisons have been used to identify supposed genes or regulatory elements in small genomic regions. The authors of this paper do a whole-genome comparative analysis of Saccharomyces cerevisiae and three of its related species: S. paradoxus, S. mikatae, and S. bayanus. These specific species were chosen because their evolutionary distances are close enough to S. cerevisiae to have sufficient sequence similarity for orthologous regions to be aligned, while at the same time, there is enough divergence to recognize functional elements. Each ORF in S. cerevisiae was compared to the genome of the other three species to determine the presence of an ortholog. In the cases in which the ORFs did not have a one-to-one match, most these "ambiguous" cases occurred in the telomeric regions. Overall, the four genomes showed a high level of conservation and using the one-to-one ORF matches as orthologous landmarks, they can be aligned at the nucleotide level.

In the identification of genes, the authors created a reading frame conservation (RFC) test to classify whether or not an ORF in S. cerevisiae is biologically meaningful. If an ORF is

conserved in a species, that species casts a "vote" towards the significance of the ORF. Most of the ORFs tested from the Saccharomyces Genome Database were strongly validated by this test. The test also evaluated ORFs encoding less than 100 amino acids. The results of the RFC test suggested the elimination of many ORFs in the existing yeast gene catalog, but also presented some new genes. In addition to the presence of ORFs, comparative genome analysis found cases in which the translation start in S. cerevisiae did not correspond to the first in-frame start codon of the other species. Similarly, there were cases in which the supposed translational stop codon did not correspond to the other species.

Potential problem areas for comparative analysis involve gene identification in cases of rapid evolution such as: the acquisition of entire genes, the loss of entire genes, the rapid divergence of nucleotide sequence, or the presence of large insertions. The RFC test incorrectly rejected meaningful ORF because of the numerous insertions and deletions that happen due to rapid evolvement.

To identify regulatory elements, the authors of the paper looked at the binding site of the Gal4 transcription factor. After examining the frequency and conservation of Gal4-binding sites across the aligned genomes, they noticed certain properties. These properties would serve as the basis for the conservation criteria for mini-motifs:

1) The mini-motif shows a significantly high conservation rate in intergenic regions

2) The mini-motif shows significantly higher conservation in intergenic regions than in genic regions

3) The mini-motif shows significantly different conservation rates when it occurs upstream compared with downstream of a gene

The mini-motifs that pass these criteria are extended with additional conserved bases and are merged to form a full motif.

To assign functions to the newly discovered motifs, the Gal4 motif was once again used as a model. It was determined that the function of the Gal4 motif could be inferred from the function of the genes adjacent to its conserved occurrences. A collection of yeast gene categories was compiled based on functional and experimental data. Most of the discovered motifs that corresponded to known motifs showed strong category correlation.

Following this, the authors explored whether additional motifs could be found by searching specifically for conservation within individual gene categories rather than overall conservation across the genome. Although several new motifs were found, this method contributed very little to the study in comparison to the genome-wide analysis.

With the aim of constructing co-occurrence networks and possible biological pathways, the authors searched for motifs that occurred together in the same intergenic regions more frequently than would be expected by random chance. They were able to find several examples of significant correlations.

Theoretically, this comparison analysis can be applied to any organism by finding an appropriate set of species. This set is dependent on the branch length between species, the total branch length, and the amount of noise in the genome.

In my opinion, this method isn't as generalized and applicable as the authors would like to think. It seemed as though after classifying ORFs they would cite rules to justify an accepted or rejected ORF, as with the rapidly evolving YBR184W ORF. A significant number of ORFs were unclassifiable based on their RFC test, but they examined each case individually and made

their own decisions. Although they judged based on the conservation of amino acids, start and stop codons, and the presence of indels, it seemed that they were not hesitant to reject ORFs.

The authors used observations of Gal4 to create a set of rules for motif discovery. Although Gal4 is one of the best-documented transcription factors, it is possible that not all motifs follow the same rules as Gal4. I think they should have used an additional regulatory element to ensure that are being more inclusive of motifs.

In the section concerning the rapidly evolving genes, the authors noted that multiple alignment of all four species simultaneously improves the alignment sufficiently to allow the YBR184W gene to pass the RFC test. I think this begs the question: why didn't they do a multiple alignment in the first place? Will a multiple alignment help classify other ORFs as well, or only rapidly evolving genes? They never address this issue throughout the rest of the paper. Perhaps the multiple alignment after the paper was written, and they added it at the last minute? It would have been nice to know why they did not pursue this further.

This evaluation of ORFs resulted in a 15% change in the yeast genome catalog, most of which were removals of spurious ORFs. I would have liked to see the authors' predictions for the future of the yeast genome catalog. From the point when the yeast genome sequence was completed to the point after this paper, the database was continually shrinking due to reanalysis. At what point would we be closest to the actual number of genes. Currently, the number of ORFs (verified, uncharacterized, and dubious) in the S. cerevisiae database is higher than what the authors presented.

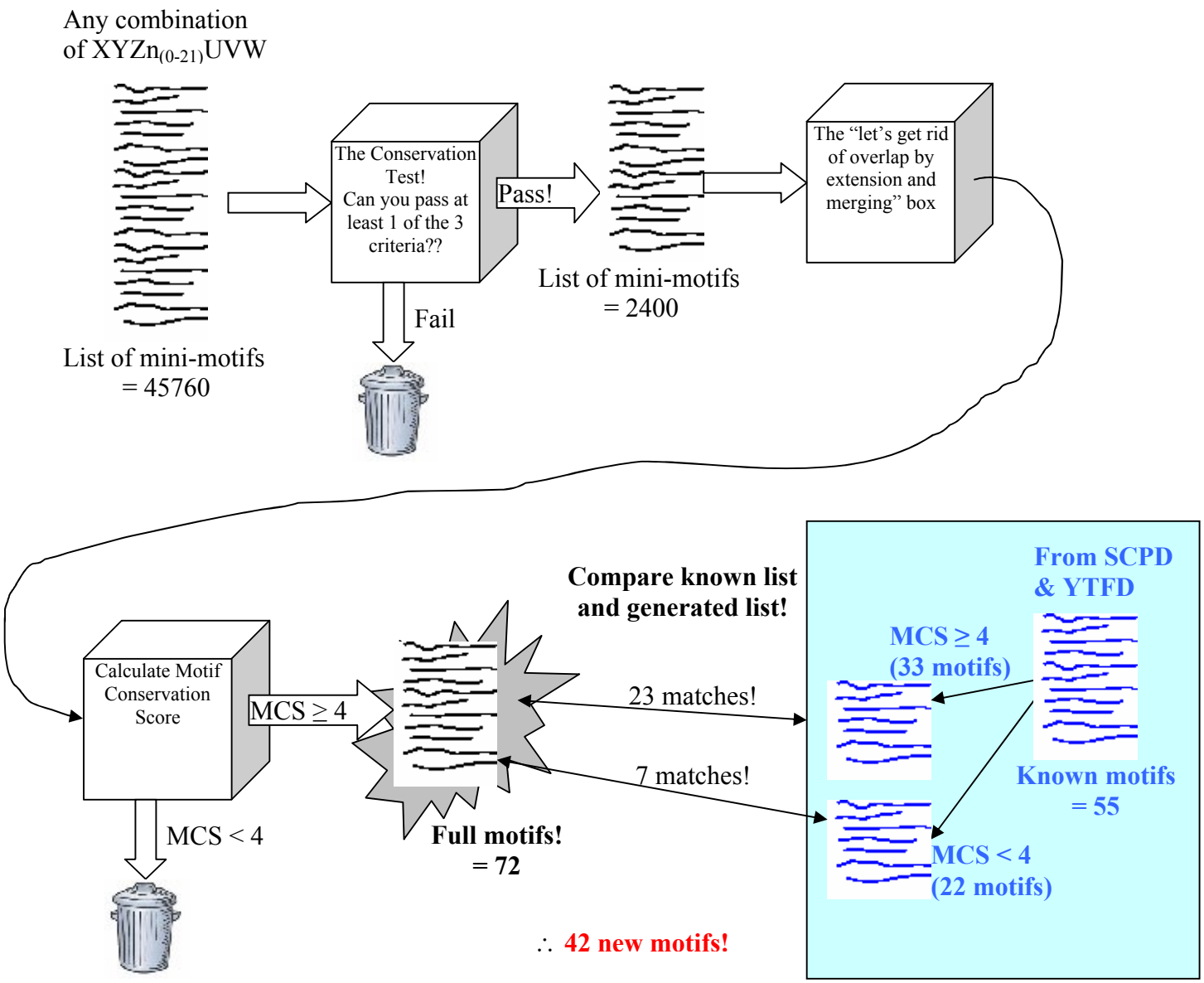

Figure 1 How the mini-motifs are generated, tested, and narrowed down to the resulting 42 new motifs.

The sequence divergence between S. cerevisiae and S. bayanus is similar to that between human and mouse[1]. A natural extension would be to do a comparative analysis between human and mouse. Unfortunately, a whole-genome multiple alignment of the rat, mouse, and human has already been done[2].

A recently published paper presented a draft genome of the chimpanzee, and compared that genome with the human genome[3]. Because the chimpanzee is very evolutionarily close to the human, nearly all the bases will be identical and sequences can be readily aligned, so a comparative analysis will not produce very useful results in terms of gene conservation. An interesting point brought up in this chimpanzee paper was that although the sequence difference between human and chimpanzee is similar to that of the mouse species Mus musculus and Mus spretus, the phenotypic variation is much greater between human and chimpanzee. An additional point of interest is that dogs shows considerable phenotypic variation despite having very little overall sequence variation (0.15%). No other species displays the large range of diversity that is seen in purebred dogs[4].

This seems to imply that there is not a linear correlation between percentage of sequence difference and amount of phenotypic variance. As a project, I would like to find out whether there is a relationship between percentage of genotypic and phenotypic differences, and what would cause those differences to vary between species.

A possible explanation for a great variance in phenotype between close evolutionary relatives could be that there were a small number of genotypic changes, but these changes were drastic in gene expression. Another reasoning could be that a specific region is very prone to changes, so there may have been a large number of changes, but they occur in the same overlapping regions.

The length of the branch of a species in the evolutionary branch may be an indicator as to which option is more likely. If a branch is longer, it has probably gone through more mutations, but these changes have small, trivial effects on gene expression. On the other hand, if there is great variance in phenotype and the branch is short, then the species have probably gone through fewer but more phenotypically significant mutations.

By doing a comparative analysis among a number of closely related species, we might be able to find out which genes are most likely to evolve rapidly and where they are located. As with any comparative analysis, we would want a handful of strategically chosen species to compare. The human and chimpanzee are very close together, so finding the differences in sequence are more useful and meaningful than the similarities. If a number of species also close to the human and chimpanzee (ex. orangutan, gorilla) can be sequenced and compared, a large number of the differences may occur in specific regions. This would point to a region of high mutation, and can be further examined for its function.

After comparing genomic regions for handful of human-like species, it might be interesting to build a profile from these similar species. The same can be done with the mouse species and the dog species. There has been a study using comparative analysis for human, mouse, rat, and dog genomes to find regulatory motifs[5]. It is possible that using a more generalized form of the species could yield different motifs.

A common concern would be whether all these comparisons are computationally doable. The whole-genome multiple alignment of the rat, mouse, and human was done in less than 1 day on a 24-node computer[2]. However, this experiment would require first creating the sequences of the similar species. Currently, only major species (particularly the model species) have been sequenced. Sequencing a draft of another species could easily take more than a year. In the

sequencing and comparison of yeast species paper[1], the yeast genomes were small enough such that they could feasibly be sequenced for the purpose of the paper, and the high signal-to-noise ratio allowed a draft to be used. Sequencing multiple species at the level of human complexity, especially a draft, solely for the purpose of comparison is not practical. If one day more sequences are available for complex species such as humans, it would make sense to use established data to do this experiment. Additionally, I don't personally own a 24-node computer.

References

1. Kellis, Manolis et al. “Sequencing and comparison of yeast species to identify genes and regulatory elements.” Nature 423 (15 May 2003): 241-54.

2. Brudno, Michael et al. “Automated Whole-Genome Multiple Alignment of Rat, Mouse, and Human.” Genome Research 14 (01 Apr 2004): 685-92.

3. The Chimpanzee Sequencing and Analysis Consortium. “Initial sequence of the chimpanzee genome and comparison with the human genome.” Nature 437 (01 Sep 2005): 69-87.

4. Irion, D N et al. “Analysis of genetic variation in 28 dog breed populations with 100 microsatellite markers.” Journal of Heredity 94 (2003): 81-87.

5. Xie, Xiaohui et al. “Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals.” Nature 434 (17 Mar 2005): 338-45