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

Pavithra Shivakumar

12/10/2005

Identification of genes based on information in the DNA sequence has been an area of active research. This paper presents a comparative genomics approach to this problem by sequencing and then comparing three other *Saccaromyces* species (S. *paradoxus*, S. *mikatae*, S. *bayanus*) to S. *cerevisiae*. They do this under the assumption that genes or regulatory elements that are required for the function of these species would be under a selective pressure to resist mutations when compared to DNA with no apparent function. Thus, comparing the sequences of these four yeast genomes should give insights into which genes and intergenic elements were conserved, and whether they have a biological role.

The four yeast genomes were sequencing and were aligned both on a large scale (entire genomic regions) and on a local nucleotide level (around each orthologous ORF). The alignments were done pair-wise between S. *cerevisiae* and the three other genomes. Although most of the ORFs aligned had a clear one-to-one match in each of the species, there were a small number where the correspondence was ambiguous, and they markedly clustered in telomeric regions. Sequence alignments at the nucleotide level showed tremendous conservation of synteny and thus help in the identification of conserved elements.

The *Saccaromyces* Gene Database (SGD) was used to identify 6062 ORFs encoding ≥ 100 amino acids (of which 2096 are uncharacterized). A Reading Frame Conservation (RFC) test was developed to classify each ORF as either biologically meaningful or meaningless on the basis of the proportion of the ORF over which the reading frame is conserved in each of the other three species. Of the 5945 ORFs that were tested, the RFC analysis strongly validated 5550.

Most of the rejected genes were justified as having a lack of experimental evidence on then. The one exception is the YBR184W gene, which was assumed to be rejected because it is evolving rapidly. On the basis of this analysis, a revised yeast gene catalogue of 5538 ORFs (≥100 amino acids) was proposed. This catalogue also contains188 short genes (<100 amino acids), of which 43 are new. The comparative genome analysis also helped identify 210 cases in which the presumed translation start in S. *cerevisiae* did not correspond to the first in-frame start codon in at least two of the three other species, and 330 cases in which the presumed stop codon did not correspond to the first in-frame stop codon in two of the other species. It was realized that in 92% of the cases, sequences at the donor, branch point, and acceptor sites at the exon-intron junction were conserved. They were also able to predict 58 new introns by searching the genome for conserved splicing signals.

Genes that were identified in the other three species but not in S. *cerevisiae* can be thought of as species-specific. This included five unique to S. *paradoxus*, eight unique to S. *mikatae*, and 19 unique to S. *balanus*. Rapidly evolving genes like YBR184W are rejected by the RFC analysis, but multiple alignment allows the gene to pass. On the other end of the spectrum, the MATa2 gene seems to be perfectly conserved in all the species, although its function is yet to be known.

Similar to identifying genes by comparing genomes, analyzing the genome for conserved intergenic regions might give insight into regulatory elements that control the transcription of genes. The Gal4-binding site was used as an example to construct a catalogue of 55 known regulatory sequence motifs from public databases. These were collected based on observations about the conservation properties of Gal4, namely high intergenic conservation rate (when compared to random motifs), high intergenic occurrences (when compared to genic occurrences), and high conservation rate in divergent as compared to convergent intergenic regions. A Motif Conservation Score (MCS) was developed to indicate the conservation rate of the motif in intergenic regions.

Two types of methodologies were used to identify motifs – genome-wide motif discovery and category-based identification of motifs. In genome-wide motif discovery, the genome was first searched for conserved mini-motifs, and these were then construct full motifs (by extension). Each full motif was assessed for conservation using MCS, and 72 full motifs with MCS ≥ 4 were retained. Further analysis of Gal4-binding motif suggested that the function of the motif could be inferred from the function of genes adjacent to its conserved occurrences. This idea was extended to find functions associated with the discovered motifs, by trying to find genes adjacent to their conserved occurrences with known gene categories. Most of the discovered motifs strongly corresponded to known motifs, and several new ones were described. The second method, category-based identification of regulatory elements was used to explore whether additional motifs could be found by searching specifically for conservation within individual gene categories. In all, 46 out of the 55 known motifs were found by either genome-wide or category-based analysis.

Several regulatory elements control the transcription of more than one gene, and many genes are controlled by more than one regulatory element. Cross-species conservation decreases random noise in a genome and reveals more biologically meaningful correlations. Several pairs of motifs were found to co-occur across genomes (Ste12 and Tec1, Leu3 and Gcn4, Met31 and Cbf1, etc.).

The paper aims to extract useful information from the comparison of closely related genomes, that would help in identification of both functional and regulatory elements. The species are far enough that there is a difference between the number of intergenic mutations and the number of mutations inside genes, but close enough that most of the functional ORFs are conserved to a certain extent. The ideas presented could be further extended to the comparison of other genomes including human and mouse.

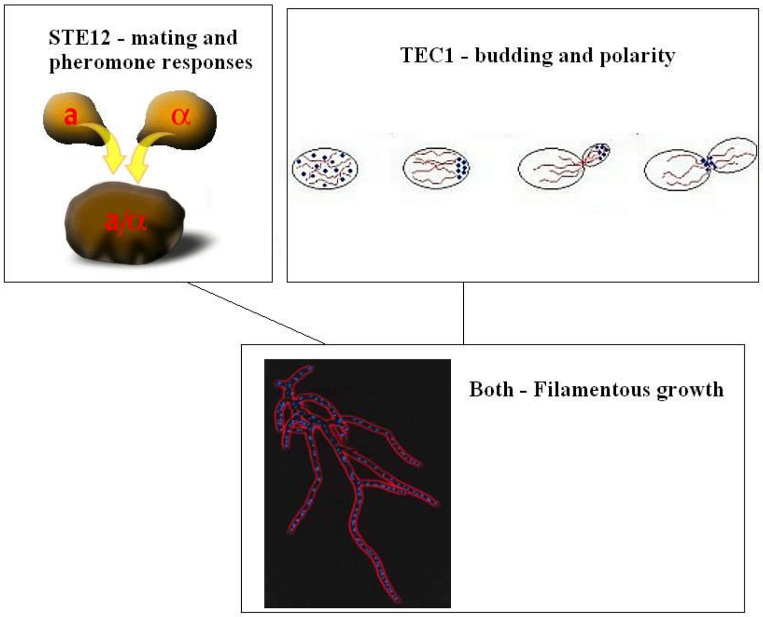
STE12 - mating and pheromone responses
a
α
a/α
TEC1 - budding and polarity
Both - Filamentous growth

Figure 1 – Combinatorial control. The transcriptional activator motif Ste12 is involved in mating and pheromone responses, while the motif Tec1 is involved in yeast budding and polarity. Comparing the yeast genomes showed that genes that have conserved occurrences of both motifs were also involved in filamentous growth

The comparative genomics idea that this paper presents doesn't seem novel. Genes from different species have always been compared to see which residues are most conserved. Comparing whole genomes just seems like a minor extension. When the paper mentions that many of the ambiguous alignments occur in the telomeric regions, the reason why this is so is not mentioned. The paper also fails to mention why only three other yeast species are chosen for the genomic comparisons. It is not clear whether using more species would be better, or worse since the conserved regions would be more obvious, but on the other hand there might not be enough ORFs shared between S. *cerevisiae* and the all the other genomes. Most of the rejections from the RFC analysis were justified as being non-functional because they have no experimental evidence. But having no data does not validate their prediction, since not knowing that ORF might be a reason that there is no data on it. The paper does not present a validation of their results. They do not propose any experiments or tests for their newly discovered motifs.

Comparing human and mouse genomes to identify and characterize regulatory elements

Regulatory elements (promoters, enhancers, silencers) in the human genome are found in the non-coding parts of the DNA (namely, intergenic regions and sometimes introns). Although, a lot of effort has been put into identifying genes by comparing human and mouse genomes, insight into the location and function of regulatory elements is lagging behind. (Guigo et al., 2003) Since any region that serves a functional purpose in the genome (both genes and regulatory elements) is under selective pressure to resist null mutations, the same principles applied to identifying genes using comparative genomics can be used to identify conserved motifs between the human and mouse genome. Earlier studies have confirmed that the mouse and human genomes share enough synteny that comparing their intergenic regions would provide a wealth of information about conserved regulatory elements. (Hardison et al., 1997) Another study done by Wasserman et al. in 2000 found muscle-specific transcription factors by the comparative analysis of the human and mouse genomes. These principles can therefore be applied to find any type of regulatory element in the entire human genome.

The main goal of the project is to not only identify conserved regulatory motifs in the human genome by comparing it to the mouse genome, but also to find ways to characterize the function of those motifs. The first step would be to download the complete sequences of both mouse and human from the University of California (Santa Cruz) genome browser (http://genome.ucsc.edu). Alignment of the intergenic regions of the two genomes to identify segments of conservation would be starting point for building a database of motifs. For each conserved segment of residues, the three conservation rules mentioned in Kellis et al, 2003 will be applied (1. Conservation rate of the motif in intergenic regions when compared to equivalent random motifs; 2. Conservation rate in intergenic occurrences when compared to genic occurrences; 3. Conservation rate in divergent when compared to convergent intergenic regions). A motif conservation score (MCS) based on the conservation rate of that motif will also be calculated, and motifs with MCS ≥ 4 will be retained in the database.

The core of the project would involve trying to characterize the function of the identified motifs using various methods. In Kellis et al., 2003, it was mentioned that regulatory elements in yeast are usually within a few 100 Kbs from the gene they control. But in mammalian genomes, this is not the case. Promoters, that help in transcription initiation, are usually adjacent to the gene, while enhancers or silencers are several 1000 Kbs away from their respective genes, or sometimes in introns. Most of the elements control the function of more than one gene, and there are many genes that are controlled by more than one regulatory element. (Pennacchio and Rubin, 2001). Therefore, multiple methods will be used to identify candidate functions for the newly discovered motifs.

Comparing the list of motifs to already known transcriptional factor binding motif databases can help to identify their functions. Namely, the databases for regulatory motifs (TRANSFAC, TRRD, and COMPEL) will be used to try and identify previously defined transcriptional factor binding motifs from the list of conserved elements. This will help in narrowing down the list to those motifs that still need to be defined using other techniques. Some motifs might be known to bind specific DNA-binding proteins. This property can also be used to infer the function of the motif from the function of its binding protein. (Wingender et al., 1997)

Since most genes are controlled by more than one regulatory element, finding two or more segments of DNA that share a common conservation pattern might be useful. If the function of one of those conserved elements is known, the function of the other can be assumed to be similar. A similar approach can be used to analyze genes that share a conservation pattern with a certain regulatory element. Since some regulatory elements control more than one gene, it can be expected that those genes would coevolve with the regulatory motif.

Another computational approach to deciphering the function of these elements is searching the literature for previously identified regulatory elements that control a certain locus. In this method, a certain locus (for example, vertebrate SCL locus as used by Gottgens et al., 2000) is used to search the literature for earlier studies that have identified regulatory elements affected the transcription of that gene. This information can then be incorporated into the database.

Comparison of the human genome with species other than mouse can give further insights into which motifs are conserved and their possible function. Comparison to species closer than mouse (like primates) can give information about motifs that have evolved later, and comparison to species that are more distant (like fish) helps to see which motifs have been conserved the most and thus are more important. If the function of the motif is known in any of the other genomes, it's function in humans can be inferred. Multiple aligning the human genome with more than two species can also help to identify conserved regions that are sometimes not detected in pair-wise comparisons. (Thomas et al., 2003) Multi-LAGAN, a tool for efficiently aligning multiple genomes will be used. (Brudno et al., 2003)

Some of the conserved motifs might not necessarily represent regulatory elements. For example, parts of the genome involved in chromosomal assembly are highly conserved. Similarly, intergenic sequences involved in the replication of DNA are also conserved. Motifs that belong to these classes need to be excluded from the list of regulatory motifs. (Pennacchio et al., 2001)

One way of validating the function of some motifs is finding their orthologues in yeast. Since regulatory motifs would affect the amount of transcript, changes in its sequence should result in a change in the amount of mRNA of the gene that it controls. This can be done by mutating the sequence of the motif in yeast and measuring the amount of mRNA using microarray chips. If the amount of mRNA increased due to the mutation, the regulatory element could be characterized as a silencer. If the amount of mRNA decreases as a result of the mutation, the motif could be an enhancer. The motif could also be a promoter if no transcript is produced as a result of the mutation.

A combination of several of the above mentioned techniques should give a comprehensive understanding of the functionality of the discovered motifs. Knowing the function of these motifs will help in further elucidating their roles in diseases. They would also help in identifying the proteins that bind to these motifs, and therefore help in designing drugs that target these proteins, and can thus affect transcription of a particular gene.

References

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., Batzoglou, S., and NISC Comparative Sequencing Program. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13:** 721-731

Gottgens, B., Barton, L.M., Gilbert, J.G., Bench, A.J., Sanchez, M.J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M. 2000. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.* **18:** 181-186

Guigo, R., Dermitzakis, E.T., Agarwal, P., Ponting, C.P., Parra, G., Reymond, A., Abril, J.F., Keibler, E., Lyle, R., Ucla, C., et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci.* **100**: 1140–1145

Hardison, R., J. Oeltjen, and W. Miller. 1997a. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7**: 959-966

Pennacchio, L. A. & Rubin, E. M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100-109

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analysis of multi-species sequences from targeted genomic regions. *Nature* **424:** 788-793

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet*. **26**: 225-228

Wingender, E., Kel, A.E., Kel, O.V., Karas, H., Heinemeyer, T. et al. 1997. TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. *Nucleic Acids Res.* **25**: 265−268