Christina Agapakis
Genomics and Bioinformatics
December 10, 2005

Comparative Genome Analysis and p53: From Lab Bench to Computer and Back Again

Manolis Kellis and his colleagues in "Sequencing and comparison of yeast species to identify genes and regulatory elements" compare the genomes of four closely related yeast species—*Saccharomyces paradoxus*, *S. mikatae*, *S. bayanus*, and *S. cerevisiae*—in order to identify new genes and define their structure, find areas of fast or slow evolutionary change in the genome, find gene regulatory elements and understand how these elements work in a combinatorial fashion to control gene expression[1]. Despite having a complete genome sequence of *S. cerevisiae* since 1996, the number of open reading frames (ORFs) present in the yeast genome is still debated[2]. Comparison of conserved sequences across species in this paper indicates which ORFs are likely to be actual genes and which are spurious. This method is a powerful one, and has comparable or even greater power to identify genes and regulatory elements than high-throughput experimental methods. Does this signify the end of experimentation in genomics and biology in general? Perhaps, but more likely experimental biologists will take this information and continue to work on experiments that will improve our understanding of biology.

The basic method throughout the paper applied to identifying all parts of the gene has to do with comparing sequences aligned across the four yeast genomes. The species are separated by 5-20 million years in evolution and share 62% sequence identity. There is a huge amount of synteny across the species, with only 211 of the 6,234 ORFs in the current *Saccharomyces* Genome Database (SGD) not having a one-to-one orthologous match across

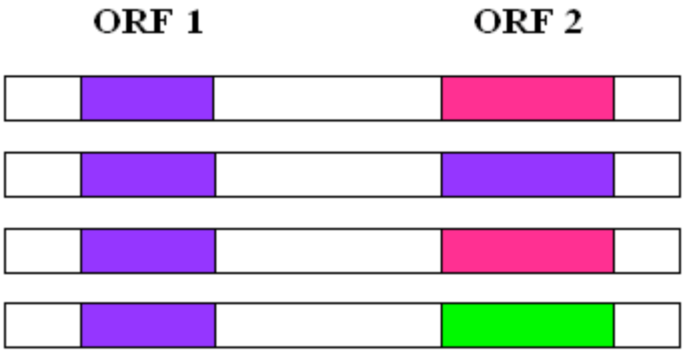

Figure 1 Schematic of RFC in Kellis et. al. ORF 1 is highly conserved across all four yeast species, while ORF 2 is not. By this test it is likely that ORF 1 is a true gene while ORF 2 represents a spurious reading frame in *S. cerevisiae* caused either by random chance or mistakes in genome sequencing.

all of the species. Of those 211, 80% were in telomeric regions, known to be areas of rapid change. For gene identification the ORFs themselves were compared across the species with a method they called the "reading frame conservation" (RFC) test. ORFs that have highly conserved sequences (like ORF 1 in figure 1) are likely to be true genes because evolution would want to maintain the sequence intact over time.

Comparative genome analysis can also test small ORFs that do not make it into the SGD. Forty three new genes expressing proteins between 50 and 99 amino acids were identified by the authors with this new method. More interestingly the structure of the genes in terms of introns and exons could by identified by comparing the donor, branchpoint, and acceptor sites for splicing. New introns can be found by searching for such sites that are conserved across species.

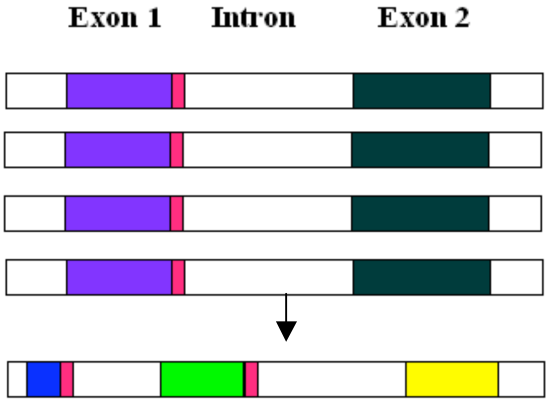

Figure 2 Gene structure identification by genome comparison. The introns splice sequences (pink rectangles) are conserved across all four species being tested. New introns can be found by searching for such conserved splicing sequences throughout the genome.

Comparing the aligned sequences across species can not only tell if an ORF represents a true gene and how that gene is organized, but can also indicate the level of evolutionary change in the gene over time. There are genes that change rapidly, as defined by the ratio of amino acid altering substitutions to silent substitutions. The proteins that were found to be most rapidly changing are likely to be involved in sporulation, consistent with previous biological investigation into positive selection. The slowest evolving gene found in this study was *MATa2*, which amazingly had 100% sequence conservation in all four species. This fact suggests useful information for understanding the function of the gene, which was unknown at the time that this study was published.

Regulatory sequences are much harder to identify than genes, but comparative sequence analysis is able to not only identify real regulatory elements (rather than random a appearance of a motif) that bind known transcription factors, but was actually able to discover new regulatory motifs. These regulatory elements are much more likely to be in intergenic than genic regions and are under selective pressure to resist mutations the same way that genes are. The " motif conservation score" (MCS) system is based on observations from the Gal4 regulatory element, a very well characterized motif in the yeast genome controlling expression of genes involved in galactose metabolism. Using this scoring method the researchers searched through the 45,760 possible motifs with the structure $XYZn_{(0-21)}UVW$ and identified 72 which had a high enough MCS score to be considered actual regulatory elements. Of these 42 had never been found before, an amazing feat since before this paper the only way to identify regulatory elements was to cluster genes with similar function and look for conserved sequences in the vicinity of the genes. Without knowing anything about the function of the genes Kellis et. al. found many new regulatory sequences and was able to propose a regulatory function for them. They could also use comparative genome analysis to determine whether two or more regulatory elements could work together to regulate gene expression.

This paper is amazing in its scope and implications for biology research. The authors cite several experimental biology papers which have tried to do what they were able to do with a relatively simple bioinformatics comparison of yeast sequences. In particular, there are a myriad of papers simply trying to decide how many ORFs there are in the yeast genome such as Kowalczuk et. al.,[3] with more than 15% of all previously catalogued yeast ORFs were cast into suspicion by the comparative sequencing. Also, there are many papers that took the assumed ORFs and used various techniques in an attempt understand splicing[4] or regulatory networks[5,6] which Kellis et. al. could render almost useless all at once. Not to mention that the alignment of the four genomes in itself represents a tremendous feat in terms of sequencing and alignment technology.

Is this paper a bad sign for experimental biology fans out there? In my opinion, no. While this paper does counter the work of many labs, it also depends on the work of many more experimental biologists, and creates a huge amount of data that can be used for future lab work and it does this all in a very clear, well-written way that you do not have to be a yeast geneticist in order to understand. Kellis et. al. do a much better job at identifying which ORFs encode genes than studies that systematically knock out ORFs looking for a phenotype and can even determine the potential function of the gene product, but need the old experiments in order to find homologs and to really understand the biology of the gene products in any meaningful way. Likewise, the sequence comparison method can find new regulatory motifs, but these are meaningless without biochemical verification. The combination of the bioinformatics and the biochemistry will hopefully one day lead to a full understanding of gene regulation.

Like all yeast experiments, one of the most important parts is how the method can be applied to understanding human biology. Comparative genome analysis can be used to understand the human genome; the human and mouse genomes share 66% sequence identity, close enough to be able to align orthologous sequences but still far enough to be able to make reasonable claims about positive selection for sequence conservation. Several groups are trying to align human and mouse sequences for these types of bioinformatics experiments. This project is still in early stages given the enormity of the human genome and a multitude of other problems. A brand new paper in *Bioinformatics* suggests methods for aligning orthologous sequences by anchoring sequences with the WU-BLAST algorithm[7].

One of the biggest problems in trying to use this method in humans is the very small signal-to-noise ratio. While nearly 70% of the yeast genome is protein coding, only about 2% of the human genome is expressed. The rest is a mess of repeated sequences frequently referred to as "junk DNA." However, as Tom Fagan put it, "A tour of almost any neighborhood on trash day reveals how loosely people define the word junk…in the genome, as in the trash, what seems of little use to some often turns out to be surprisingly valuable.[8]" In a relatively recent paper by Cawly et. al. hundreds of new transcription factor binding domains were found in the "junk" regions using tiled microarray technology in conjunction with chromatin immunoprecipitation (ChIP)[9]. Comparative genome analysis could be a great way to verify that these are indeed conserved regulatory elements and improve our understanding of transcriptional control in humans.

There are of course many other problems with human comparative genome analysis. Junk DNA may be full of potential treasure in the form of regulatory sequences, but these regulatory sequences themselves are full of junk that can make pure sequence based analysis impossible. Most human transcription factor binding sequences contain built in degeneracy, which is hypothesized to affect expression by allowing for different degrees of transcription factor binding[10]. As a result, algorithms searching for conservation have to consider the pattern of the regulatory site rather than the exact sequence of bases.

An approach much like that used in the Kellis paper may be useful here. Kellis et. al. use the Gal4 motif as their model for finding all regulatory motifs in yeast. This motif and its control of sugar metabolism are well known, and has become a useful tool in molecular biology experiments. In the comparative analysis, the Gal4 sequence was found to be perfectly conserved across all four species in its spot between the *GAL10* and *GAL1* genes that it controls. From this they extrapolated that it is likely that if a motif is in fact regulating gene expression it will be conserved over time. They then found all the conserved Gal4 sequences that appear throughout the yeast genome and were able to make generalizations about the location and structure of such genes that they could apply to their MCS test about finding new regulatory motifs.

For something as complicated as the human genome, any generalizations we can make based on known regulatory motifs would greatly increase the power of a comparative genome analysis. One transcription factor that can be used for making these kinds of generalizations is p53, which regulates expression of proteins important for the G_2 cell cycle checkpoint. Its role in DNA repair makes p53 is a tumor suppressor gene, and it is perhaps the most commonly mutated protein in human cancers. As a result, p53 has been extensively studied in protein biochemistry, molecular biology, and bioinformatics experiments. All of the information generated from these experiments can be used in order to build a model for comparative sequence analysis which can in turn lead to more information about p53 and gene regulation that can be used in cancer research.

The DNA sequence that p53 is thought to bind to is actually a pair of degenerate ten base pair palindromic sequences. It can be schematized as “→← … →←”, where “→” represents the sequence PuPuPuC(A/T), “←” represents the palindromic sequence (T/A)GPyPyPy, and “…” represents a spacer region that ranges from zero to fourteen base pairs in length[10]. Other putative binding sequences have been identified relating to DNA damage[11]. Different p53 binding sequences have been found and verified using standard biochemistry experiments as well as ChIP technology[9] or computer algorithms searching through the genome[10]. Of the thousands of possible sequences with the →← … →← structure there are only a handful that appear to be realistic p53 binders from these experiments.

Comparison of these binding sequences with the aligned mouse genome can indicate whether these motifs are likely to be true regulatory sequences and define conservation rates that will be useful for further analysis of the genome. For example, the rate of conservation at intergenic as opposed to genic regions, the rate at conservation in upstream versus downstream regions, and more also just the conserved sequence itself can be easily measured by this type of aligned analysis. These generalizations can be then applied to searching for other p53 binding sequences in the human genome.

A list of putative binding sequences will provide insight into how p53 binds to DNA and controls genes and how these genes can play a role in cancer progression. More generally, a method that can identify transcription factor binding sequences through a comparative genomics analysis will be very useful in determining how human gene regulation occurs on a much wider scale. This type of experiment will always have to follow the chain of lab bench to computer and back to the lab bench again. The data inputted into computer programs for identifying genes, regulatory elements, or whatever other part of the genome has to first come from experiments with real cells and real DNA. Once this information is tabulated and arranged the bioinformatics stage can begin. Out from this comes more information which is only as useful as the implications it has on future experiments.

Kellis et. al. have written a remarkable paper. Not only have they almost definitively found the number of ORFs in the yeast genome and made progress into defining regulatory sequences for transcriptional control of these genes, but they have created a model for genome interpretation that can be applied to all the sequenced genomes, including the human. By starting to study regulatory sequences such as those binding the p53 transcription factor in comparative studies between the human and mouse genomes we will soon be able to understand the specifics of p53 control as well as human gene expression across the genome.

References

1. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature.* **423**, 241-254 (2003).
2. Harrison, P. M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30**, 1083-1090 (2002).
3. Kowalczuk, M., Mackiewicz, P., Gierlik, A., Dudek, M. R. & Cebrat, S. Total number of coding open reading frames in the yeast genome. *Yeast* **15**, 1031-1034 (1999).
4. Clark, T. A., Sugnet, C. W. & Ares, M. Jr Genome-wide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296**, 907-910 (2002).
5. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281-285 (1999).
6. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799-804 (2002).
7. Sauer, T., Shelest, E., & Wingender, E. Evaluating Phylogenetic Footprinting for human-rodent comparisons. *Bioinformatics.* E-publication. (2005).
8. Fagan, Tom. "Study Finds Regulated Transcription of Novel RNAs: Noncoding Transcripts May Be Widespread in Genome." News from Harvard Medical, Dental, and Public Health Schools. March 5, 2004. <http://focus.hms.harvard.edu/2004/March5_2004/biological_chemistry.html>
9. Cawley, S., Bekiranov, S., Ng, H. N., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K. & Gingeras, T. R. Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs. *Cell.* **116**, 499-509 (2004).
10. Hoh, J., Jin, S., Parrado, T., Edington, J., Levine, A.J. and Ott, J. The p53MH algorithm and its application in detecting p53-responsive genes. *Proc. Natl. Acad. Sci. USA.* **99**, 8467–8472 (2002).
11. Walter K, Warnecke G, Bowater R, Deppert W, Kim EL. Tumor suppressor p53 binds with high affinity to CTG-CAG trinucleotide repeats and induces topological alterations in mismatched duplexes. *J Biol Chem.* E-publication. (2005).