John Mission

December 10, 2005

MB&B 452

Professor Gerstein

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

In their review paper, Kellis et al. attempt to characterize and analyze the genome of the frequently used model organism, S. cerevisiae by using newly developed methods of comparative genomics. Comparing the genome to that of three evolutionary next-door neighbors, S. paradoxus, S. mikatae, and S. bayanus, they performed the following tasks: an alignment of their genomes; a characterization of their evolution that explores the specific mechanisms of this evolution; the development and utilization of new methods to recognize genes and regulatory motifs, which is especially problematic and difficult (Mathé, 4103); and to analyze the genes and motifs found. In summary, their results suggested a major revision of the existing annotations and functions for genes listed in the currently existing databases.

Using the shotgun approach, they assembled draft sequences for each of the genomes that were previously listed in the introduction section; this data is currently available on the web at http://genome-www.stanford.edu/Saccharomyces/. After sequencing, they aligned the sequences, using one-to-one matches of open reading frames (ORFs) as landmarks to help define the large-scale alignment, then proceeding to generate local alignments around each orthologous ORF. Across all four species, they found a conservation of synteny, serving as a good measure that these species were closely related enough to justify an analysis.

Specifically, Kellis et al. noted that a specific gene family whose ORFs showed significantly more expansion or contraction in cross-species comparisons clustered in the regions surrounding the telomeres. They posit that the rapid change in these regions is supposed to create phenotypic diversity over evolutionary time and suggest a well-developed mechanism for this change, citing that the 20 inversions were flanked by tRNA genes in opposite translational orientation, with the tRNA genes usually expressing the same isoacceptor type. At the nucleotide level, they see that the probability for finding different nucleotides in nongenic regions is twice that of finding different nucleotides in gene-coding regions. Along with the strength of alignment and synteny in these four species, they cite this observation as a justification for using the species which they had chosen.

In this section, they demonstrate a clear knowledge of the biology which they are discussing. From this point of view, one might guess that their backgrounds are more grounded in biology than they are in statistics or computational analysis. In this sense, I think they gain more credibility, seeing as we have seen repeatedly in the papers that we have read this semester that often computational analysis without biological backing might be problematic (Lakshminarayan, et al.). One criticism is that it might be useful

The next task attempted was a *de novo* identification of the protein-coding regions of genes, which is simultaneously an extremely important yet daunting task in the current field of genomics. They describe the current rationale for ORF notation, which suggests that some ORFs are too long to have occurred by chance. Here, I take issue with the mode of probabilistic reasoning that underlies this approach. Some biological knowledge (knowledge of start and stop codons) guides this method of annotation, but I still think that this reductive approach leaves too much room for error. Anecdotally, the existence of life on earth itself is very “improbable.”

Several studies examined in this year (i.e. Hutchinson, et al.) exemplify the weakness seen in such an approach.

However, the approach that Kellis, et al. use buttresses the pre-existing mode of reasoning for gene annotation by making cross-species analyses to test if “ORFs” seen in one species have orthologs in the other species. To do this, they develop a reading frame conservation test (RFC) to classify each ORF as meaningful or meaningless. Although they don’t address the issues of false positives in this section, I am still not at all surprised that their annotation method finds fewer ORFs than those using the previous methodology. Such results that are well-guided by biology are more likely to be reliable. Also, they do make an attempt at addressing another assumption in the preceding method with which I also take issue: the lower limit of 100 amino acids for protein size. They find 43 new ORFS with high RFC scores whose lengths fall between 50 and 99 amino acids; shorter ones could not be measured because of the lack of statistical strength of their methods.

In the section where they evaluate their annotated ORFs, 117 could not be analyzed because of either a lack of overlap among species or because of an excessive overlap between species. The former case might result from rapid evolution while the evolutionary proximity that suggests the latter might produce such a corresponding sequence by chance. A possible limitation of this method is that it might not work in characterizing species that are extreme evolutionary outliers, meaning that they are exist in an environment that fostered the rapid development of one species alone or that encouraged dramatic speciation to fit into niches. Such sequences with very specific function like those seen in extreme thermophiles (Vieille, et al.), might contain too many singletons to be analyzed in such a manner. As if they read my mind, however, Kellis et al. do bring up this point as a weakness, which further lends support to their analytical capacities.

The next section attempts a *de novo* discovery on a genome-wide level of the regulatory elements that exist within S. cerevisiae. They claim to have discovered 54 new introns in the yeast genome and cite how their findings corroborate with the findings of their collaborators working with microarrays. Such an interdisciplinary and collaborative approach deserves praise. In discovering other elements, they start with a well-studied example, the Gal4-binding site, and cite three observations that lend to the strength of the applicability of their approach to Gal4: the higher rate of conservation of this motif in comparison to equivalent random motifs, the higher occurrence in intergenic regions as opposed to intergenic regions, and the higher rate of conservation in divergent compared to convergent intergenic regions. Turning to the identification of other motifs looking at these same characteristics, constructing a motif conservation score (MCS) that also looks for three similar characteristics: higher levels of intergenic conservation, higher levels of intergenic vs. genic conservation, and higher upstream conservation than downstream conservation.

Here, I think the paper begins to falter in strength. Despite the fact that the authors chose to search for the 55 motifs from public databases that have the most support on data from public databases, I still feel uneasy about the fact that the results of these experiments rely on previously annotated genes; this study has already suggested the modification of current databases, and I would prefer to see the experimental results that support these data before I believe their findings. Perhaps a citation of the motifs chosen would be sufficient. Despite this, their results do strengthen some of the results of previously cited genes, and their discovery of new possible motifs does warrant some experimental study.

Overall, the execution of the paper itself is excellent. I considered it an easy enough read for even novice students interested in comparative genomics. Whether the background of the reader lies in molecular biology, evolutionary biology, or statistics and probability, this paper keeps its explanations very concise and coherent while providing the interdisciplinary biological background information underlying their analytical thought that remains necessary for the elementary reader to make sense of their work. However, the extent of their descriptions might leave the computational biologist itching for the detailed specifics of the analytical methods they used in their work. Despite this criticism, the techniques and methodologies presented and executed in the paper are powerful and useful. After reading this article, I am eager for the chance to make an attempt at implementing their work in designing a project that will extend their work.

Figure:

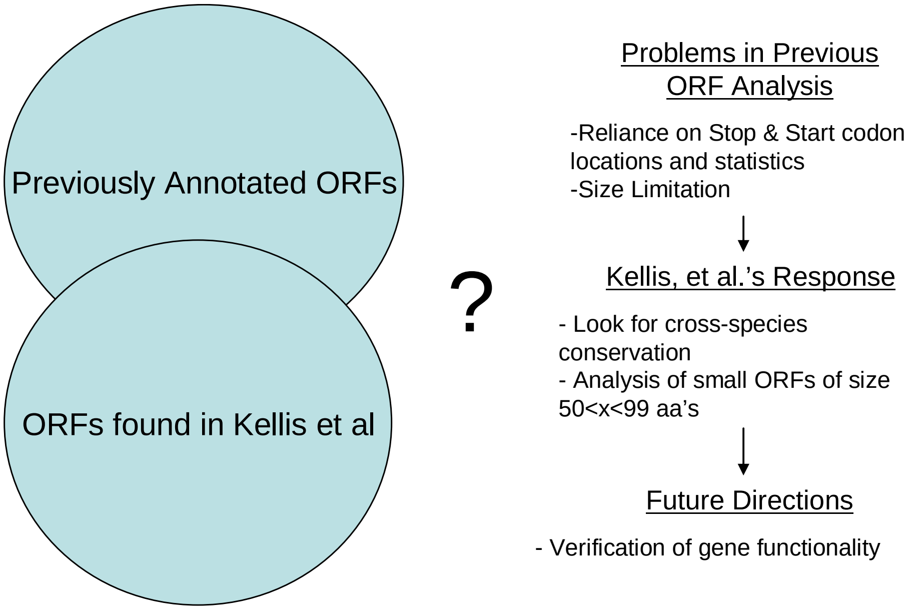
Previously Annotated ORFs
ORFs found in Kellis et al
?
Problems in Previous ORF Analysis
-Reliance on Stop & Start codon locations and statistics
-Size Limitation
Kellis, et al.'s Response
- Look for cross-species conservation
- Analysis of small ORFs of size 50<x<99 aa's
Future Directions
- Verification of gene functionality

Figure 1. Schematic of previous ORF annotation, Kellis, et al's ORF annotation methods, and future directions in this area.

An indendent project proposal: The Transposon Mutagenesis of Yeast and Microarray Technology: A Logical Extension of Kellis, et al.

Question:

What are the functions of the new genes and regulatory sites in S. cerevisiae that are suggested by “Sequencing and comparison of yeast species to identify genes and regulatory elements” by Kellis, et al.?

Background Information

The most pressing question that came to mind after reading Kellis, et al.’s paper was whether or not the new method of ORF annotation utilized by Kellis, et al. actually yielded biologically significant results. The Saccharomyces Genome Database (SGD) collects DNA and protein information from sources like GenBank, EMBL, DDBJ, SwissProt, and PIR and presents the results in datasets on their webpage (Cherry, et al.). The source of their complete genomic sequence is that of strain S288C and resulted from a collaboration from around the world that was completed in April 1996, with several revisions being made as more information became available. In terms of the annotated ORFs included in the SGD, the creators of the SGD recognize that many of the sequences annotated in their database have tenuous possibilities, it is their policy to keep these ORFs in the SGD until these ORFs are otherwise refuted by experimental evidence.

Kellis, et al.’s method of ORF annotation produced a smaller number of ORFs, providing 5550 ORFs, some of which were previously not annotated in the SGD while others were already present in the database. But are such differences strain-specific? Is it sufficient to assume that the genes corroborated by Kellis’ study are not simply conserved by chance because of the evolutionary proximity of the species measured? This project will use various techniques of genomics to attempt to approach this question.

Goals

This project specifically will look at three different sets of ORFs: previously annotated ORFs in the SGD that were corroborated by the Kellis, et al study (a further search for false positives), of which there are 5550; previously annotated ORFs from the SGD that were rejected, deadlocked, or unclear (a search for false negatives) of which there were 512, and an analysis of the 43 new genes identified in Kellis, et al. that were approximately between 50 and 99 amino acids in side.

The approach involves building a compendium of profiles similar to that used by Hughes, et al. The fact that a compendium of profiles for this species has already been produce provides support for executing a similar study to test for ORF functionality and biological relevance. A compendium of profiles will be constructed in S. cerevisiae by taking as many deletion mutations of yeast with experimentally established functions in the SGD and using a two-color cDNA microarray hybridization assay in order to characterize the expression levels of mRNA within each of these specific strains of yeast. For the groups of ORFs in question in this project, deletion mutation strains will be constructed by the systematic knockout method (it is assumed that this project has unlimited resources for funding). If a change in the mRNA expression profile is seen in the knockout strain compared to the wild-type strain, then the results of experiment provide elementary evidence that a gene function exists for the ORF that has been systematically knocked out.

Some weaknesses exist within this approach. For example, the characterization of a gene’s function might not be seen with clarity if the unknown ORF encodes a type of protein of still uncharacterized function. For such unique proteins, a different method of analysis would have to be implemented. Another limitation of this work is that deletions within the ORF in question might involve genes with redundant functions; in other words, other genes might encode proteins with nearly identical functions to those of the experimental mutant. Such proteins might be able to compensate for the missing hypothetical gene product. Furthermore, the potential for false positives also exists within this method of functional notation.

Despite this, the method can still serve to be of particular use as corroborating evidence for ORFs in each of the three categories under investigation. In the category of ORFs for which both the latest revision of the SGD and Kellis et al.'s study are in agreement, the use of such expression profile-matching experiments might be able to assign new functions to previously annotated genes. Just because an ORF has been previously annotated and one of its functions has been evaluated and established experimentally does not have to rule out the possibility that other functions for the gene product of this ORF might exist. For the ORFs which were annotated in the SGD but which have no previous assigned function, a compendium study might provide the proper divining rod to guide the direction in which gene function research should proceed. It should be noted that this experiment is not intended to serve as the ultimate authority in determining gene function. However, performing such studies would provide corroborating evidence for guiding future work just as was the case for the Kellis, et al. study.

The execution of the experiment described above would ultimately help to establish whether or not the ORF annotation method developed by Kellis, et al. yields more specific and meaningful results. If this study fails to find an exceedingly high number of false positives, then this would suggest the potential application of this comparative genomics approach of ORF analysis to other organisms.

References

Cherry, J. Michael et al., “SGD: Saccharomyces Genome Database.” Nucleic Acids Research 26.1 (1998): 73-79.

Hughes, TR et al. “Functional Discovery via a Compendium of Expression Profiles.” Cell 102 (2000): 109-126.

Hutchinson, Clyde et al. “Global Transposon Mutagenesis and a Minimal Mycoplasma Genome.” Science 286 (1999): 2165-2165.

Kellis, Manolis et al. “Sequencing and comparison of yeast species to identify genes and regulatory elements.” Nature 423 (2003): 241-254.

Lakshminarayan, M. Iyer et al. “Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences.” Biology 2.12 (2001): http://genomebiology.com/2001/2/12/research/005.1 - 0051.11

Levitt, Michael and Mark Gerstein. “A unified statistical framework for sequence comparison and structure comparison.” Proc. Natl. Acad. Sci 95 (1998) 5913-5920.

Mathé, Catherine et al. “Current methods of gene prediction, their strengths and weaknesses.” Nucleic Acids Research 30.19 (2003): 4103-4117

Vielle, Claire and J. Gregory Zeikus. “Thermozymes: identifying molecular determinants of protein structural and functional stability.” Tibtech 14 (1996): 183-190.