Meisha Bynoe
MB&B 752a Final Paper
December 10, 2005

Summary: Kellis, M.* et al. Nature** *423,**241-253.

Background

In 1996, the genome of *Saccharomyces cerevisiae* was completed due to the work of approximately 600 scientists world-wide. This group of researchers set the number of ORFs of the single-celled eukaryote at 5885. They initially came up with 6275 ORFs putatively encoding proteins of ≥100 residues, but reasoned that 390 of these ORFs were unlikely to give rise to a protein product[1]. The publishing of the yeast genome revolutionized work with this eukaryote as it allowed members of the scientific community to systematically knock-out its genes in an effort to discover their functions, as well as in other studies. Since the publishing of the genome, however, studies have arisen redefining the size of the genome of *S. cerevisiae*[2, 3, 4].

As time passes, the number of genomes sequenced increases exponentially as it becomes easier and cheaper to sequence a genome. This capability has led researchers to begin comparing the genome sequences of related organisms. At least one comparison of the genomes of different strains of *Helicobacter pylori* has been published[5]. Sequence comparisons of the genomes of the human malarial parasite *Plasmodium falciparum* and the rodent version of the pathogen *Plasmodium yoelii yoelii* have allowed researchers to identify orthologs between the two species, and thus, facilitated the exploration of vaccine targets against malaria in a rodent model[6].

The discrepancies between the proposed numbers for the functional ORFs in *S. cerevisiae* and the apparent success of comparative genomics led the Lander group to explore genomic comparison of *S. cerevisiae* with three of its relatives, *S. paradoxus*, *S. mikatae* and *S. bayanus* as a possible method for determining functional ORFs and regulatory elements in *S. cerevisiae*[7]. The results obtained from the study demonstrated that such a use for the data being generated with the sequencing and characterization of genomes is feasible.

Summary of Methods and Results

The group opted to compare the genome of *S. cerevisiae* with *S. paradoxus*, *S. mikatae* and *S. bayanus* based on their phylogeny. The three species used in the comparison were closely related enough to *S. cerevisiae* to allow for meaningful comparison, but their genomes were sufficiently diverged to avoid meaningless sequence identity/similarity. The sequences used in the study gave full coverage of *S. cerevisiae*, and over 90% coverage of the other three species. The researchers aligned the genetic material, and deduced that, for the most part, there were one-to-one orthologous matches among the four genomes. There were also ambiguities, mainly due to gene family expansion or contraction. The most variable regions of the genome were the telomeric regions, as has been observed in other organisms. At the nucleotide level, the variability of the nucleotides in the genic regions was 30%, showing high conservation of the nucleotide sequence. In addition, the occurrence of insertions or deletions (indels) was only 1.3% in the genic regions (figure 1).

The authors then proceeded to identify genes *de novo*. They developed a reading frame conservation (RFC) test to determine if a given ORF was biologically meaningful or not. Complications that were dealt with were (i) the overlapping nature of 948 ORFs and (ii) the presence of exons and introns in the genomic sequences. The authors opted to use ORFs with a certain proportion of unique sequence, and the largest exon of coding sequences that included introns. They tested the algorithm on 340 intergenic sequences, 96% of which were rejected, demonstrating the sensitivity of the test. All 6062 ORFs of *S. cerevisiae* as denoted in the *Saccharomyces* Genome Database (SGD) were tested, with 5,458 ORFs being unanimously validated. The authors determined 15 rejected ORFs to be spurious. Further analysis led them to

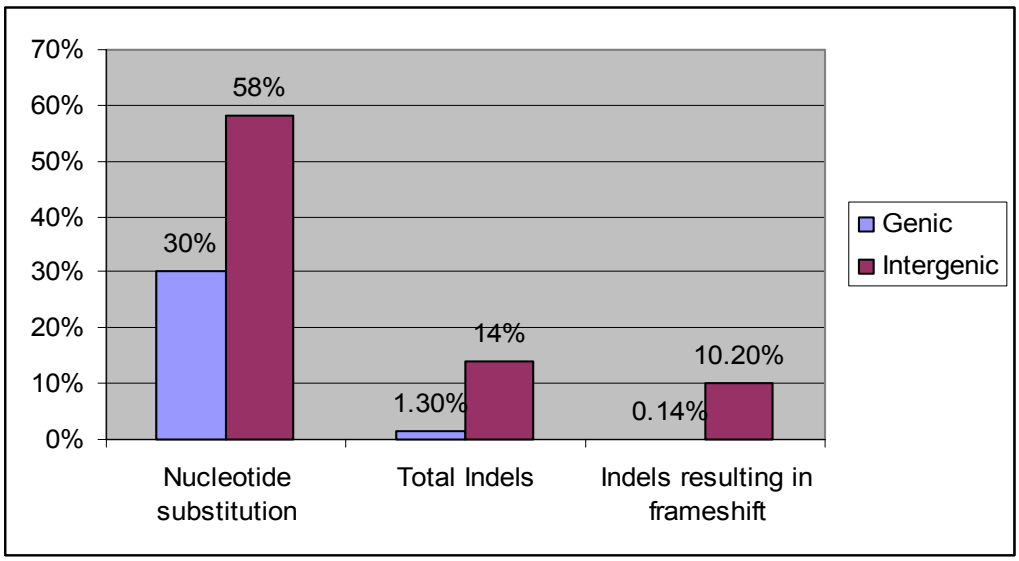

Figure 1: Differences observed in the genomes of the four species of Saccharomyces being compared.

propose the number of ORFs in *S. cerevisiae* to be 5,538. The algorithm, however, needs to be improved to confidently identify small ORFs encoding fewer than 100 residues.

Gene structures were also examined. For the most part, the findings of this study coincided with previous data. However, the authors discovered errors in the published *S. cerevisiae* genome regarding stop and start codons, as well as splicing sites.

By comparing the genomic sequences, the authors were able to find genes that were subject to rapid and, in other cases, slow, evolution. A total of 34 genes encoding ≥200 amino acids were found to not be present in all four genomes.

The identification of regulatory elements *de novo* was done in this study as well. The parameters for the task were based on characteristics of the Gal4 motif, to which the well-studied Gal4 transcription factor binds. The criteria were (i) high conservation rate in intergenic regions, (ii) higher conservation rates in intergenic regions compared with genic regions, and (iii) higher conservation rates when the motif appeared upstream, as opposed to downstream, of a gene.

45,760 mini-motifs were first identified. Scoring of these motifs and combining overlapping and very similar/identical mini-motifs gave rise to the finding of 72 motifs, of which 42 were new. The comparison allowed for 25 to be assigned a biological role. The authors also checked for motifs that worked in combination with others.

Review

The paper itself was well structured. There was, however, a simple mathematical error in calculating the number of genes not found in all four species of *Saccharomyces* – the sum of the numbers given was 34, not 35 as stated. In addition, the authors mentioned a few times data from other studies, but did not give any references, which would have been useful for the reader to have.

The authors proposed that, given the success of the comparative genomics on the four *Saccharomyces* strains, the technique could be extended to humans. They suggested that, given the similarity in the divergence between *S. cerevisiae* and *S. bayanus*, and the mouse and human, the latter pair could be compared. However, the success of this study was based on the use of *S. bayanus* alongside two other species more closely related to *S. cerevisiae*. Other organisms of similar or less divergence relative to humans might be needed to increase the likelihood of deriving informative data from such a study involving the human genome.

As noted by the authors, the human genome would produce a low signal-to-noise ratio due to the small proportion of coding sequences in the genome. Thus, it is presumptuous to suggest the study be applied to humans right away. A group of genomes of intermediate complexity should be analyzed first. The *Saccharomyces* genomes used contain very few introns, and sequences that do not encode proteins. The human genome lies on the other end of

the spectrum. Should the approach be successful with genomes of intermediate complexity, it would be more likely to work on human genome comparisons.

The idea of using comparative genomics as done in this study can prove useful in the daunting task of annotating genomes. However, there needs to be some fine-tuning of the process, and further testing to ensure that the algorithms, parameters and such will give rise highly informative data.

Proposal: Further study into the uncharacterized genes of Saccharomyces cerevisiae

Introduction

In May 2002, the genome of *Saccharomyces cerevisiae* was said to contain 6062 ORFs, 3966 of which had been assigned a function. Thus, 2096 of the ORFs remained uncharacterized[7]. Other estimations of the size haves also been made: Kellis *et al* and Wood *et al* have proposed that the actual size of the *S. cerevisiae* genome is in the range of 5538-5570 ORFs. These numbers were derived from *de novo* gene prediction[7] and re-annotation of the genome[3]. Comparison of the genes identified in these two studies would give intriguing results, and could provide valuable information regarding the characterization of the genome of *S. cerevisiae*.

In addition to *de novo* ORF identification, Kellis *et al* were able to pinpoint, *de novo*, regulatory elements throughout the genome of *S. cerevisiae*. The group used an algorithm to identify mini-motifs, which they scored and combined accordingly to derive 72 motifs, 30 of which corresponded to known, characterized regulatory elements. Of the remaining 42, the authors were able to putatively assign 25 of the motifs as being bound by certain proteins, or functioning in certain pathways. They were unable to categorize 17. In a subsequent study,

Harbison *et al* attempted to identify the binding sites, and possible combinatorial effects, of 203 transcriptional regulators of *S. cerevisiae*[8]. They do not outrightly note the connection of the sequences identified in their study with those motifs found in the material presented by Kellis *et al*. Such correlation of the studies would be useful, to show if any of the categorized motifs without specific assignments in the Kellis *et al* paper were identified as binding sites for the regulators analyzed by Harbison *et al*.

In addition to combining the studies as proposed, it is suggested that the assignment of putative structures to uncharacterized proteins be performed, as some of these proteins may bind to the uncharacterized motifs found by Kellis *et al*, which might not have been assigned in the Harbison *et al* study. These proteins may have functions in transcriptional and other regulatory mechanisms in *S. cerevisiae*. Such a study would prove useful as, in the annotation of genomes, the major method of determining the functionality of an ORF is to check for sequence homology with sequences known to encode proteins. Structural homology is not typically checked, but may provide invaluable information regarding the possible biological functions of putative proteins.

Proposed Experimentation

There are various methods of assigning structure to protein sequences. One of the methods that can be applied in this study would be testing the uncharacterized sequences on known folds. Known DNA-binding proteins can be analyzed and trends in their folding patterns determined. These fold patterns can then be assigned to the proteins hypothesized to arise from the uncharacterized ORFs of *S. cerevisiae*.

It is possible that a variety of folds may be found upon analysis of the known DNA-binding proteins. However, the chance of finding great variability in the structures of such proteins is very low, due to the structure of DNA and that fact that, typically, DNA binding proteins interact with the major groove of duplex DNA using helical domains[9]. Given such and other information, it should be very possible to score the putative proteins with the known folds, and thus deduce if they bind DNA.

To determine the scoring of the putative proteins with the DNA-binding folds, threading of the amino acid sequences through the folds would be the best way to begin. This method involves the passing of the amino acid sequences encoded by the uncharacterized ORFs through the known folds, and evaluating the feasibility of the sequence giving rise to such a structure. The scoring is based on interactions of different residues, their chemical properties and their relative positions in the protein structure on a whole (e.g. buried versus exposed). For this method, a scoring scheme would be developed based on the fold being analyzed and the sequences that conform to this fold. The advantage of this method is that it was engineered for assigning structure to proteins that have very little (<30%) sequence identity with proteins that are known to conform to the structure being fitted to the uncharacterized protein[10].

Other methods, as detailed and evaluated by Ginalski *et al*, can be used. None of the methods based on sequence similarity can be used in this study, however, as the uncharacterized ORFs were already scanned against databases, with no significant hits based on their sequences. *Ab initio* methods are employable. However, use of these methods would likely make the study more time-consuming. Such methods involve the simulation of protein folding, to give rise to a number of energetically favorable conformations. As a large number of DNA binding proteins are already known, it is simpler to apply the threading method in lieu of *ab initio* methods.

In evaluating the folds associated with known DNA-binding proteins, attention would be paid to the sized of the proteins giving rise to the folds to be used. By taking into account the size range of these DNA-binding proteins, it might be possible to reduce the large pool of uncharacterized proteins. For instance, if the vast majority of proteins known to bind DNA are large, it would highly unlikely that a very small *S. cerevisiae* ORF would encode a protein that has this capability.

Extension of Study

Upon finding proteins encoded by the uncharacterized ORFs of *S. cerevisiae* that fit the folds defined by known DNA-binding proteins, analyses can be done to assess whether these proteins do in fact bind any of the uncharacterized, or known, motifs. Based on the results of such studies, further analysis could allow for the assignment of these motifs and proteins to cellular pathways and processes. This information would be useful in future genomic comparisons with newly sequenced genomes, and in, thus, the annotation of such.

References

1. Goffeau, A. *et al*. "Life with 6000 Genes." *Science* **274**, 546-567 (1996).
2. Kowalczuk, M. *et al*. "Total Number of Coding Open Reading Frames in the Yeast Genome." *Yeast* **15**, 1031–1034 (1999).
3. Wood, V. *et al*. "A Re-annotation of the *Saccharomyces cerevisiae* Genome." *Comp. Funct. Genomics* **2**, 143-154 (2001).
4. Velculescu, V.E. *et al*. "Characterization of the yeast transcriptome." *Cell* **88**, 243-251 (1997).

5. Alm, R.A. *et al*. "Genomic- sequence comparison of two unrelated isolates of the human gastric pathogens *Helicobacter pylori*." *Nature* **397**, 176-180 (1999).

6. Carlton, J.M. *et al*. "Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*." *Nature* **419**, 512-519 (2002).

7. Kellis, M. *et al*. "Sequencing and comparison of yeast species to identify genes and regulatory elements." *Nature* **423**, 241-254 (2003).

8. Harbison, C.T. *et al*. "Transcriptional regulatory code of a eukaryotic genome." *Nature* **431**, 99-104 (2004).

9. Martin, A.C.R., *et al*. "Protein folds and functions." *Structure* **6**, 875-884 (1998).

10. Ginalski, K. *et al*. "Practical lessons from protein structure prediction." *Nuc. Acids Res.* **33**, 1874-1891 (2005).