Sebastian Szpakowski
12/10/2005

1. Review

Kellis et al. (2003) present a compelling case for the use of comparative genomics to aid sequence analysis and annotation efforts. The comparison of four closely related species of yeast on genomic level allowed the authors to gain insight about the gene localization and their composition. Additionally, a similar approach was used in intergenic regions, where a new method was capable of distinguishing several new, genome-wide, well conserved motifs. The reported outcome of the analysis suggests that the current annotation of yeast genome could be enhanced and cleansed of existing errors using comparative genomics approach.

The experiment is preceded by whole-genome sequencing of *S. paradoxus (Sp)*, *S. mikatae (Sm)* and *S. bayanus (Sb)*, which belong to the *Sacharomyces sensu stricto* family. The method described in the article required that the analyzed genomes were as evolutionary proximal as possible to represent genomic syntheny required to build cross-species orthologue landmark maps. At the same time, however, the sequences should demonstrate enough evolutionary divergence in intergenic regions to be able to distinguish between noise and signal or evolution and conservation, respectively. Another requirement imposed by the authors requires the compared species to come from a narrow taxon. This can be helpful in assigning a function to unclassified genes and functional elements, based on an assumption that related organisms share much of the biology of their physiological processes. Knowledge of a function of a gene in one species can then be hypothetically extended to an orthologous region in a related organism.

The three species sequenced were chosen based on these criteria with respect to *Sacharomyces cerevisiae (Sc).* The authors suggest that a evolutionary distance between human and mouse genomes is similar to that of *Sc* and *Sb*, implying that a similar comparative genomics study could be extended to investigate more complex organisms. The required criteria are not too stringent. However, human and mouse are not as close to each other in taxonomy as the two species of yeast, so perhaps some parameters, and statistics of the models used in the article would have to be adjusted to fit the latter case.

Not much is said about the specifics of genomes of the three chosen species of yeast. For example, table one of the article reports that a sevenfold redundant coverage of the genome of *Sm* for example, reconstructed only 93% of the sequence with many gaps left which, presumably can skew the orthologue matching across the species. My Figure 1 shows which regions in each of the three species were not sequenced (white areas). Data to generate the plot was acquired from the supplemental material page given in the article. It would be beneficial to the study to analyze if the gaps among contigs occur by chance, or perhaps certain specific regions are more difficult to sequence. Individual alignments with completely sequenced and assumed to be correct genome of *Sc* could shed some light as to which parts are missing in the other 3 species. The supplemental area does not contain chromosomal positions of the aligned elements. Therefore the use of BLAST would be necessary to reconstruct the missing information.

The analysis method used in the study involved first aligning *Sc* to each of the other 3 genomes individually. To accomplish that, the sequences of known ORFs of *Sc* were searched against *Sm*, *Sb* and *Sp* genomes using BLAST. It is noted that the method will not return meaningful results for ORFs shorted that 50 nucleotides, so that initial list of ORFs has been filtered accordingly. The resulting high density landmark maps were used with CLUSTAL W program to generate multiple alignments of all 4 sequences for the entirety of available sequence. Consequently, each aligned ORF among 4 species was evaluated for conservation and consistency among the 4 species. Well conserved regions allowed examining of the validity of existing ORF boundary predictions. If an ORF showed a significant number of conserved

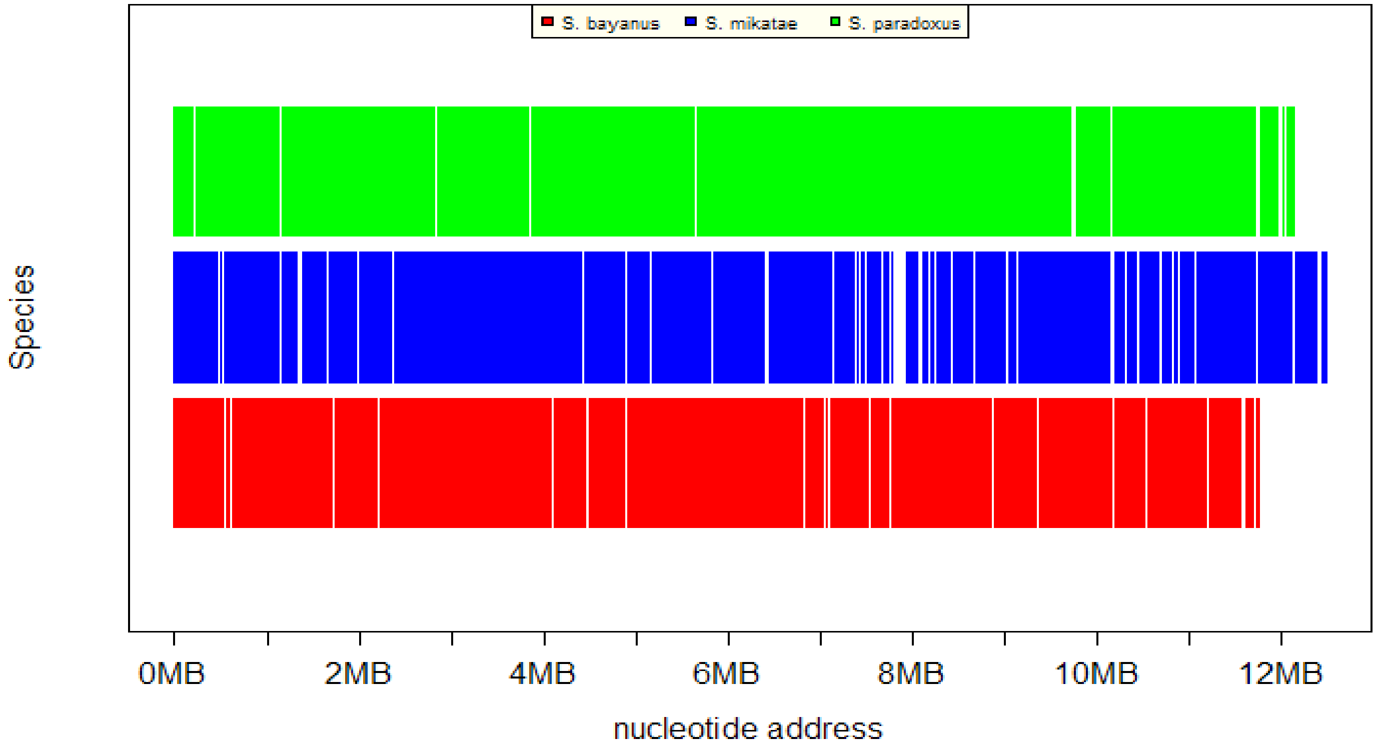


Figure 1. Comparison of sequencing effort among the three species of yeast used in the article. The white spaces indicate regions that were not sequenced. Whereas *S. Bayanus* and *S. Paradoxus* (green and red bars respectively) show at this resolution a random distribution of gaps across their genomes *S. mikatae* (blue bar) shows at least 3 regions with systematic sequencing problems: at 1.2MB, at 8MB and at 10.4MB.

nucleotides across its length among the 4 species it was considered valid. Otherwise the method indicated such an ORF as spurious. In general the method showed a weakness for fast evolving genes (telomeric regions and YBR184W involved in gamete production) indicating them as spurious based on their lack of conservation among 4 species of yeast. Overall the authors claim the false positive rate to be about 1%. It would be helpful to investigate if a larger genome with more ORFs would help lower the number.

The method cannot evaluate the ORF boundaries for a few species-specific genes since they would not have orthologues, however knowing the functions of proximal conserved genes allowed presuming that the unknown gene could have a similar function as the cluster of genes to which it belongs.

One of the most remarkable findings was that the most species-unique changes were observed in the telomeric regions of chromosomes. These regions contain non-unique gene sequences which appear to have evolved rapidly. In addition, new sequences, some protein coding, and numerous translocations were observed as well. Even the transposable Ty elements did not introduce as much variation in the remaining parts of the genome as was noted to exist at the telomeric regions. Another finding mentions a species-specific gene for all four species of yeast inserted among two genes that appear to have been conserved.

The second part of the article discusses a method that allows investigating the existence and localization of novel regulatory elements in the intergenic regions of genomes. This is a more challenging task than that described in the first part of the article, as regulatory elements, i.e. binding sites for transcription factors, are relatively short motifs. Additionally they have not been studied as extensively as genes, leaving a potential for many unknown rules that govern these elements. Comparative genomics, again, can help. This time, by evaluating if a conserved region that is not an ORF could contain regulatory elements. This method leaves a lot of speculations regarding the functions and validity of predictions. More experiments would have to follow to confirm any of the findings.

Because the knowledge of regulatory elements in comparative genomics is limited, the authors used a Gal4-binding motif to build a set of rules that could then be applied in search of other (similarly acting) elements. By comparison of conservation rates of a random motif to that corresponding to Gal4 binding site in coding and non-coding regions, surprisingly non-coding regions showed a much higher conservation than coding regions containing the motif. This is opposite to a random motif.

The method searches for mini-motifs of a certain form which are then iteratively extended until the significance drops below a certain level based on a degeneracy matrix which includes observed random mutations in the genome among species. Besides defining four conservation criteria (based on previously observed characteristics of Gal4 binding domain), the article does not extensively discuss the creation of data structures and statistical model used for this analysis, leaving the reader with general concepts. The outcome motifs are then ranked based on the motif conservation score as a function of conservation criteria. The rank cutoff was tweaked based on the rediscovery of known motifs in the genome. The use of cross species comparison allowed to distinguish between random noise, and conserved biological correlation in establishing predictions of potential regulatory motifs.

Overall the article seems to present a well organized set of methods and results. The definitions and concepts are accessible to most readers with some background in computational analysis of sequences. Beyond conceptual explanations, however, not many details are revealed about the methods neither in the article nor in the supplemental section online, though the latter seem to contain a little more information. Anyone interested in comparative genomics should find this article interesting. The audience does not have to exclusively include enthusiasts of yeast research. The article mentions several avenues in which similar techniques can be applied to other organisms. Several noted observations involving evolutionary aspects of DNA conservation and speciation provide yet another reason not to overlook this paper.

2. New Ideas

Why is the DNA sequence at the telomeric ends of chromosomes so variable when compared across related species of yeast, while the middle of chromosomes remains largely intact? How come the Ty elements with a tremendous potential of genome rearrangement do not wreak as much havoc in the genome? What is the probability that a conserved region among four different genomes will be split by a gene unique to each species? How accurate and useful is the prediction of mutation rates of intra- and intergenic regions given the knowledge of annotated functionality of certain parts of the sequence? Can such predictions be used to ascertain what will happen to a genomic sequence in a few (hundred) generations? Is there a limit to how far an evolution can go? When does speciation occur? Perhaps some of the above or similar questions could be answered using an agent based simulation system? Following there is a draft of such a system.

A growing and extensively developed for the number of years Recursive Porous Agent Simulation Toolkit (REPAST) could help conceptualize and understand the way genome changes and adapts providing more insight into comparative genomics, and similar attempts to “reverse engineer” the evolution and find patterns.

Summarizing very quickly, repast is a robust and extensible set of java programming interfaces that provides framework for various agent based simulators designed initially for social networks. Repast is not the only available platform for agent based simulation. In fact it is based on another toolkit: SWARM. (Other options could include NetLOGO.) The choice of repast as a platform to use, at least initially, is based most of all on its maturity but also its extensibility and portability provided by underlying java technology. Potential parallelization or distribution of java objects onto nodes of a cluster (would there be such a demand based on the complexity of the problem) should not be too difficult to implement with repast.

A simulation environment would define “nucleotide” as an agent in the simulation. Each agent can store the nucleotide letter it represents, and perhaps some functional association such as a gene id for tracking of the changes. Depending on the available computational resources certain simplifications can be made [1] in the implementation of the system. An agent could represent several nucleotides, decreasing genetic resolution and precision. Consequently, using a machinery of network display provided by an appropriate repast class, a networked chain of agents can be created to represent a “genome”. Since repast uses a concept of ticks of time to propel the simulation forward, at every tick of time, each one of the agents will have a certain probability of participating in a genomic event: of change (mutation), relocation (transposon activity/crossing over), disappearance (deletion), emergence (insertion) etc. A certain number of ticks could approximate organism’s development, after which an “offspring” DNA is created. At this point there are two individual DNA sequences each one undergoing its own metamorphosis, and (if implemented) with certain probability interact with other genomes to create more variation, so that the pool of available genomes increases.

The genomic events do not have to occur on a single agent level. In fact we can use our current knowledge of locations of functional and non-functional elements within a genome of an organism such as yeast to group our agents into virtual functional units. Certain units can be LTR parts of Ty elements, genes, regulatory motifs etc. A random genome generator can create an initial sequence of a given length and randomly assign all known yeast ORFs to agents, a) based on real coordinates, b) clustering the ORFs according to functional similarity or c) scattering them completely at random, in all cases making sure that non-unique genes have several copies.

[1] Although by 2011, the projected graduation date of the author when he could even start considering becoming a postdoc, the available desktop computers will be incomparably more powerful than these available today in 2005.

The sequence would then be filled up with transposons, transcription factors' binding domains, and so on. During the simulation, an activated Ty element for example, would drag along with itself a chain of agents of sequence to insert and assimilate somewhere else in the genome. Appendix A contains some concept art of the design and basic functionality. Snapshots of all or a subset of existing genomes at a given time can be aligned using multiple alignment tools, aided by using the high density orthologue landmark maps described in the article (Kellis, 2003). The correlation of the history of recorded repast genomic events and the clustered alignments perhaps can help in identifying the fate of short ORFs or ambiguously randomized telomeres.

The process of establishing the rules that would make the simulator realistic is probably the most challenging part of the design and will be a significant milestone of this research project before any meaningful predictions could be produced. Among the simplest rules for example there would be the random mutation rate at which a nucleotide can flip into another letter. These can be established from numerous observations included in the nucleotide substitution matrices BLOSUM or PAM, or from comparative genomics studies such as the one discussed earlier (Kellis, 2003). Gross Chromosomal Rearrangements, however, and spontaneous mutations do not occur too often in yeast (Liti et al, 2005) which partially explains the stability of middle parts of chromosomes.

The existence of genome functional annotation databases and a wealth of research available on yeast organisms (SGD, YMMR) could help establish certain rules, about which mutations in a genome are lethal. If upon creation of an "offspring" genome repast would determine that such an offspring would not have been able to survive based on lack or defects of certain genes, the "offspring" would not have been included in the pool of genomes available for future perturbation. In addition, the parent who also contains the lethal mutation would be declared "extinct". Potential other information taken from the annotation databases could include how rapid is evolution of a given gene, which could then modify the basic mutation rate parameter in the simulator.

Having established the basic rules for the majority of the sequence, certain rules have to be developed for agents located towards the "virtual telomeres" that could accommodate for the reported disarray of yeast genes in telomeres. Telomeres are highly repetitive sequences. During cell replication the repetitive DNA sequences in there undergo rapid elongation prone to errors. The yeast mutants unable to elongate their telomeres replicate a few times until a cell cycle arrest (Teixeria, 2005). Several genes that mediate the extension of telomeres will be included in the list of essential genes. Several models for potential implementation of telomere functionality are discussed by Fajkus et al (2005).

A recent study by Patel et al (2005) indicates a potential rule of creating an "offspring" DNA sequence using random origins for replications, which could potentially introduce more variability and transcriptional mistakes to account for. Implementing the random firing using an agent based simulation platform would not pose a problem, but perhaps could help with better understanding of variations in certain parts of genome.

An additional component of the simulator would be implementing of a functionality allowing for the interaction of two genomes. So far the design discusses a genome copying itself with minor modifications due to transcription errors. Allowing certain genomes to exchange information this way could approximate the formation of ascopores. As reported by Liti et al (2005) there is an increased Ty activity with this kind of a genomic event. Further complexity to the simulation could be obtained by coiling the virtual DNA strand onto histones and thus blocking the strand's accessibility at certain times.

It is not the purpose of this simulator to create a collection of genomes best suited for a particular environment. Rather, the perturbations of genetic fragments should create a range of genomes each producing a viable phenotype capable of surviving in an environment, based on the rules that have been observed experimentally. A clever literature search along with data mining of the known information resources can help establish certain rules that could be implemented in an agent based system. The bioinformatic tools could then help studying the arrangements of genomic elements.

There are several potential uses of such a dataset. Whereas the proposed simulation system deals primarily with information for yeast as a model organism, genome sequences of other organisms or perhaps tissue samples could be used as well. Understanding of the propagation of genetic changes to predict the future generations' genomes could be useful in simulating of tumorigenesis for example. Overall the described system can help in our understanding to what degree is the emergence of conserved patterns stochastic. A dataset of several "related" and complete (i.e. without gaps) sequences generated using this simulator can be analyzed using the methods described by Kellis et al (2003) to check whether the predictions would hold with fabricated data as well. Perhaps having a little more control over the intermediate evolutionary states of the sequences can help to get rid of the ambiguities irresolvable otherwise.

Acknowledgments:
I'm thankful to Dr. Paul Lizardi for the inspiration and ideas regarding the concept of simulator codenamed *chromosome shuffle*.

References:

Kellis M., Patterson N., Endrizzi M., Birren B., Lander E. S. (2003), *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature. **423**(6937) pp. 241-54.

Supplemental information at: http://www.nature.com/nature/journal/v423/n6937/suppinfo/nature01644.html

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990), *Basic local alignment search tool*. J Mol Biol. 1990 Oct 5;**215**(3):403-10.

Thompson JD, Higgins DG, Gibson TJ. *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res. 1994 Nov 11;**22**(22):4673-80.

Patel PK, Arcangioli B, Baker SP, Bensimon A, Rhind N, (2005), *DNA Replication Origins Fire Stochastically in Fission Yeast*. Mol Biol Cell. 2005 Oct 26; [Epub ahead of print] (http://www.molbiolcell.org/cgi/reprint/E05-07-0657v1)

Fajkus J, Sykorova E, Leitch AR. (2005), *Telomeres in evolution and evolution of telomeres.* Chromosome Res. 2005;**13**(5):469-79.

Teixeira MT, Gilson E. (2005), *Telomere maintenance, function and evolution: the yeast paradigm*. Chromosome Res. 2005;**13**(5):535-48.

Liti G, Peruffo A, James SA, Roberts IN, Louis EJ., (2005), *Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the Saccharomyces sensu stricto complex*. Yeast. 2005 Feb;**22**(3):177-92.

Repast Organization for Architecture and Development (2003-2005) at http://repast.sourceforge.net

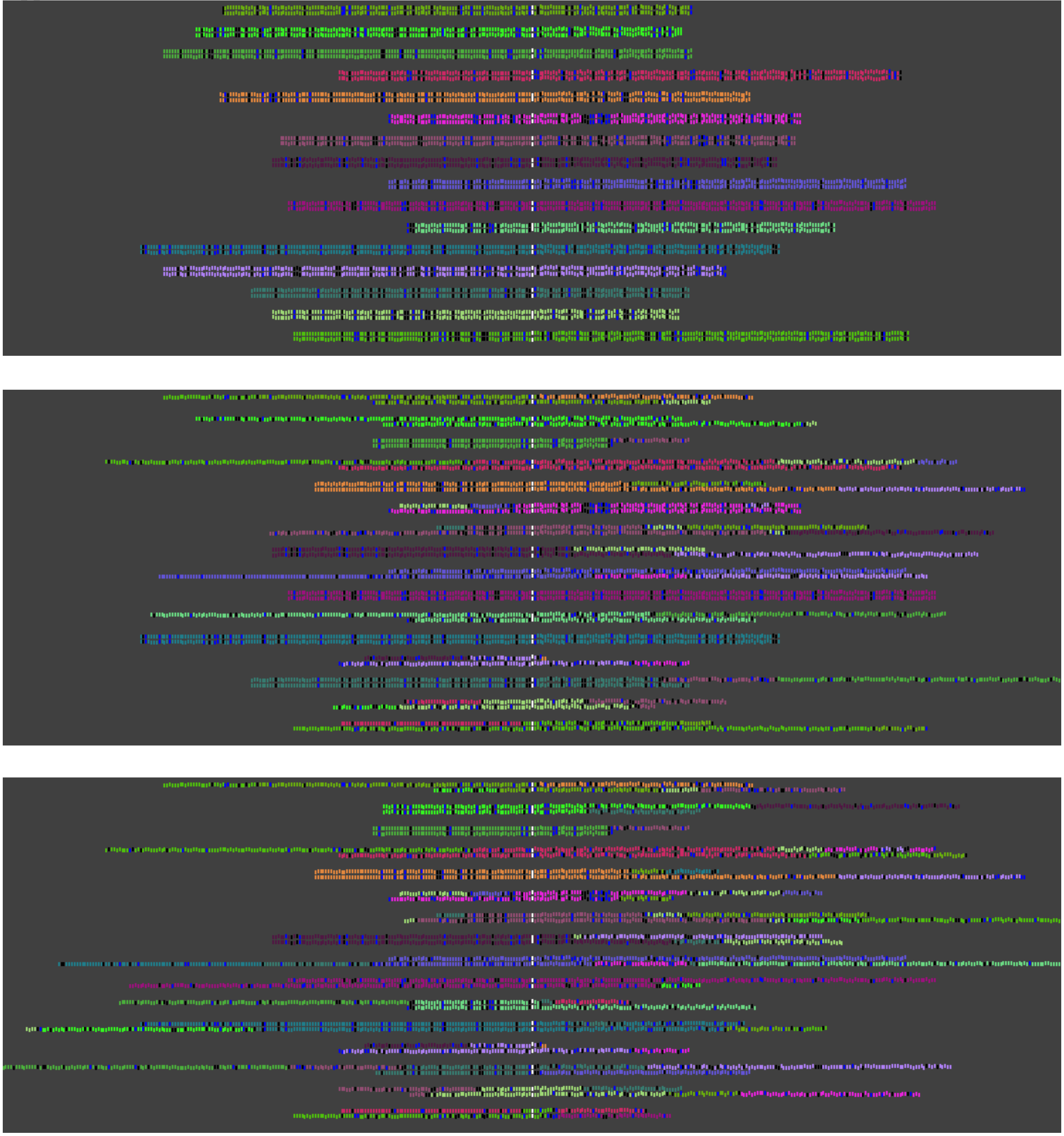
Swarm Development Group (2004) Swarm 2.2, Available at http://wiki.swarm.org (Aug. 2004)

Yeast resources:

Saccharomyces Genome Database (SGD) Available at http://www.yeastgenome.org/

Yeast Metabolism Model Repository (YMMR), Available at http://jjj.biochem.sun.ac.za/database/index.html

Appendix A



[Concept art] Time progression of one genome in the simulation. Top figure represents the very beginning with 16 diploid yeast chromosomes, each chromosome shown in different color, additionally dark blue and black rectangles represent Ty and regulatory elements respectively. White rectangle indicates a centromere. Middle and bottom figures represent snapshots of nonconsecutive iterations conceptualizing an exaggerated retrotransposon mediated rearrangement of chromosomes. Initially uniformly colored chromosomes become multicolored to represent the peculiar shuffle /translocations of portions of chromosomes reported by Kellis et al (2003) in figure 2.