Modelling The Fitness Landscapes of a SCRaMbLEd Yeast Genome

Bill Yang
ICOS
School of Computing
Newcastle University
1, Urban Sciences Building
Science Square, Newcastle upon Tyne, UK

Goksel Misirli
School of Computing and Mathematics
Keele University, UK
g.misirli@keele.ac.uk

Anil Wipat
ICOS
School of Computing
Newcastle University
1, Urban Sciences Building
Science Square, Newcastle upon Tyne, UK
orcid.org/0000-0001-7310-4191

Jennifer Hallinan
BioThink
Brisbane, Australia
orcid.org/0000-0002-2860-1022

***Abstract*—The use of microorganisms for the production of industrially important compounds and enzymes is becoming increasingly important. Eukaryotes have been less widely used than prokaryotes in biotechnology, because of the complexity of their genomic structure and biology. The Yeast2.0 project is an international effort to engineer the yeast *Saccharomyces cerevisiae* to make it easy to manipulate, and to generate random variants using a system called SCRaMbLE. SCRaMbLE relies on artificial evolution *in vitro* to identify useful variants, an approach which is time consuming and expensive. We developed an *in silico* simulator for the SCRaMbLE system, using an evolutionary computing approach, which can be used to investigate and optimize the fitness landscape of the system. We applied the system to the investigation of the fitness landscape of one of the *S. cerevisiae* chromosomes, and found that our results fitted well with those previously published. Our simulator can be applied to the analysis of the fitness landscapes of any organism for which SCRaMbLE has been implemented.**

Keywords*— SCRaMbLE, yeast, simulation, fitness landscape

I. Introduction

The Synthetic Yeast 2.0 (Sc2.0) project is an international effort aimed at engineering a eukaryotic genome, that of the Baker's yeast *Saccharomyces cerevisiae*. The project involves eleven institutions from five countries. *S. cerevisiae* is widely recognized as a model organism, is generally regarded as safe, and hence has been studied in considerable detail, and is extensively used in industry [1].It is therefore an ideal organism for genome-scale engineering of a eukaryote [2]. The ultimate aim of Sc2.0 is to reconfigure the yeast genome in such a way that it is easier to understand and manipulate, using procedures including the deletion of all known genome destabilizing elements (transposons and sub-telomeric repeat regions); the insertion of symmetrical *loxP* (loxPsym) recombination sites immediately downstream of all non-essential genes; conversion of rarely used stop codons, such as TAG, to the major stop codon TAA, to free up a codon; the watermarking of all protein coding sequences by synonymous base changes, so that they can be identified as synthetic genes by PCR amplification; the removal of all tRNA genes; and the removal of the majority of the 250 introns.

The insertion of the loxPsym sites is of particular importance, since these sites become the locations of genome reshuffling with the addition of Cre recombinase [3]. This system is known as SCRaMbLE: Synthetic Chromosome Rearrangement and Modification by loxPsym-mediated Evolution. The loxPsym sites themselves are too short, at only 34 bp, to participate in homologous recombination, so the SCRaMbLE system is only induced by the addition of Cre recombinase [4]. When the SCRaMbLE system is induced, not all loxPsym sites will be activated. The stretch of DNA between two active loxPsym sites is referred to as a segment, and may include several ORFs.

The ability to generate multiple variations from a wild-type chromosome, via insertion, deletion, translocation, or inversion of existing genes, means that it is possible to produce thousands or millions of novel genomes. Most of these genomes will, of course, be non-functional, and the Sc2.0 project aims to use directed evolution to select colonies with desirable characteristics. On solid medium, a primary metric for fitness *in vitro* is colony size. Growth in liquid media can also be measured. Fitness *in vitro* is often measured as the ability to produce a substance at enhanced levels. Of these metrics, only the last is amenable to evaluation using computational simulation.

Directed evolution can be an efficient approach to identifying desired variants of a wild-type organism. However, by applying only directed evolution, many interesting and potentially useful genotypes will be missed. Further, directed evolution is a time-consuming and wasteful process, which cannot fully explore the genomic richness generated by the SCRaMbLE system.

There is a large body of work into the interaction between evolutionary processes and the fitness landscape generated by individuals in a population [5-9]. Evolution has been shown to occur more efficiently—that is, more of the possible phenotypes are explored in a shorter time—upon a relatively smooth fitness landscape than on a jagged surface, in which the fitness of one individual is largely unrelated to that of an individual close in genotype [10]. At present, this more theoretical view of the potential of the SCRaMbLE system is largely ignored, the assumption being that if enough recombinant chromosomes are generated, individuals with desired phenotypes can be identified via screening.

Computational modeling and analysis of the fitness landscape generated by the SCRaMbLE system offers the prospect of identifying system parameters which can produce a smooth fitness landscape, hence improving the efficiency of the directed evolution process, and of improving our understanding of the biology of an important

978-1-6654-0112-8/21/$31.00 ©2021 IEEE

model eukaryote. The stochastic nature of an evolutionary algorithm, combined with genome-wide data on mutations, means that multiple runs can explore different areas of the evolutionary landscape.

In this paper we report the development of a computational model of the SCRaMbLE system, and the fitness landscape generated thereby. The model was parameterized on both genome-scale experimental mutant data and a computational yeast metabolic model. Each set of recombined chromosomes generated from a chromosome pre-processed by the SCRaMbLE system is considered as comprising a genetic landscape. Each newly generated chromosome has a genetic distance, *D*, from its parent and all other chromosomes in the population, and a fitness, *f*. By combining these two metrics, a fitness landscape can be constructed and explored. This model allows different configurations of a SCRaMbLEd chromosome to be explored.

II. Methods

A. Algorithm

Because of the extensive processing and modularization of yeast chromosomes, as described briefly above, it is reasonable to consider each chromosome as a linear vector of genes. We chose to simulate the synthesized right arm of Chromosome IX (synIXR), because it is relatively short [3], allowing detailed manual checking of the results, and because *in vitro* data from SCRaMbLE experiments is available for this chromosome, permitting comparison of simulation results with laboratory data.

The simulated chromosome was initialized as a list of strings of 43 segments, reflecting the relative position of each segment in the targeted chromosome. Segments were separated by the loxPsym site immediately after every non-essential ORF. The probablity of a Cre recombinase binding to a breakpoint is *scrProb*. A pseudo-random number generator was used to determine whether a Cre recombinase binds to a breakpoint.

B. Distance metric

The genetic distance between each pair of chromosomes in the population was calculated using the Levenshtein distance [11]. This distance metric measures the number of edits required to convert one string into another, using insertion, deletion, or substitution. The genetic distance can be considered to be a measure of the similarity between evolved chromosomes.

In order to determine which of the mutation operations are optimal at any point in the chromosome, we need a cost value for each modification. If we were working with DNA sequence information, the use of a substitution matrix, such as PAM [12] or BLOSUM [13], might be appropriate. However, in this simulation, we handled segments as indivisible units, as dictated by the SCRaMbLE system, so all operations were assigned an equal weight of 1.0. This weighting assumes that all operations are equally likely, an assertion which could be modified in the light of experimental data.

Fitness is an abstract concept and is very difficult to apply in practice. The fitness function used here was based on two types of data: Single gene deletion/overexpression fitness data (experimental fitness) and flux balance analysis results (FBA fitness).

An *in vitro* project described previously used colony size as a measure of the fitness of gene deletion and gene overexpression strains [14]. In a similar fashion, SCRaMbLEd chromosomes which produce colonies at least equal in area to those of the wild-type can be considered to be fit. However, this approach is clearly infeasible for a simulated system. Another important concept related to fitness is gene essentiality. Of the entire genetic complement of an organism, only some genes are essential for life [2]. However, this concept is also fraught with difficulty, particularly for unicellular organisms. Which genes are essential depends largely upon the environment, and an organism grown in rich media is likely to require fewer genes for survival than one grown in minimal media.

Further, genes do not act in isolation; a gene may be essential only in the absence of one or more other genes [15]. Synthetic lethality occurs when either of two genes is sufficient for viability alone, but the organism becomes inviable when both genes are knocked out [16]. Although most of the research into synthetic lethality has been performed in the context of two-gene interactions, most genes and their products interact with multiple other genes and gene products, in a plethora of ways [17]. There is a very large body of research into complex genetic networks and their robustness or otherwise in the face of internal and environmental challenges, based largely upon the work of Paul Erdös in the 1950s [18], but blossoming in a genomic context in the early 2000s [19-21], and to which we have contributed [22-24]. However, because of the nebulous nature of gene essentiality, and the preliminary nature of this work, we chose to apply a naïve definition of essentiality. For our chromosome, genes were deemed essential if they were identified as such in any description in the *Saccharomyces* Genome Database [25]. Of the 43 segments on synIXR, 7 ORFs (YBL112C, YIR006C, YIR008C, YIR010W, YIR016W, and YIR023W, located on segments 2, 7, 9, 10, 12, and 20 respectively) were essential, and the segments carrying them were identified as essential, using this criterion. Any SCRaMbLEd chromosome not carrying all seven essential genes was deemed to be non-viable.

Whilst the absence of essential genes results in the failure of the cell to grow under certain conditions, some genes, especially those encoding enzymes which carry out key processes in metabolic networks, only result in a reduced growth rate when absent. The contribution of these enzymes, and therefore their genes, can be modelled using genome scale metabolic modelling. Flux balance analysis (FBA) is a common approach for simulating metabolic networks [26]. FBA determines the flow of metabolites through a given metabolic network, and can be used to predict the growth rate of an organism under a given growth regime. In this work we reconstructed the yeast metabolic network for each of the SCRaMbLEd chromosome variants, taking into account deleted and duplicated enzyme-encoding genes.

The FBA-related fitness was based on the latest consensus yeast metabolic model [27], while the constraint file was created from simulated SCRaMbLEd results in which ORFs on deleted segments are set to 0. The fitness of the SCRaMbLEd genomes, F_s, was then calculated by running a flux balance analysis, using FlexFlux software

[27, 28]. The results were normalized by considering the wild-type fitness, *Fw*, which was calculated as 88120.37 mmol/gDW/h by running the no-constraint FBA of the original yeast metabolic model (Eq. 1).

$$Normalized\ FBA\ Fitness = \frac{Fs - Fw}{Fw} \quad (1)$$

Both the single gene deletion fitness data and the single gene over-expression fitness data were obtained from published data [14]. We analyzed the distribution of growth rates in the set of mutants from the experimental data, including the deletion and duplication data, to determine how these data could be used to parameterize the fitness function. The variability between the deletion and duplication datasets was found to be high, while the variability between those two data groups was relatively low, with both datasets showing a high proportion of mutants, peaking at a fitness of 3-3.5 as measured by the growth rate (Figure 1). The culture media used for both groups were similar, but differed slightly due to the strategies used for the selection of mutated strains. Using the hypothesis that the fitness distributions of deletion and duplication mutations are similar, we applied quantile normalization to the fitness score of the deleted and duplicated ORFs in the two datasets (Figure 1).

Using this approach, given a SCRaMbLEd genome, an experimental fitness score (EFS) could be calculated by averaging the normalized fitness scores of the deleted or duplicated ORFs.

$$EFS = \bar{X} , \quad (2)$$

where *x* is the normalized fitness of a mutated ORF

A comprehensive fitness value was then calculated by multiplying the FBA fitness and the EFS. ORFs not included in the metabolic model or wet-lab data were considered to not affect the fitness.

$$Fitness = EFS * normalized\ FBA\ Fitness \quad (3)$$

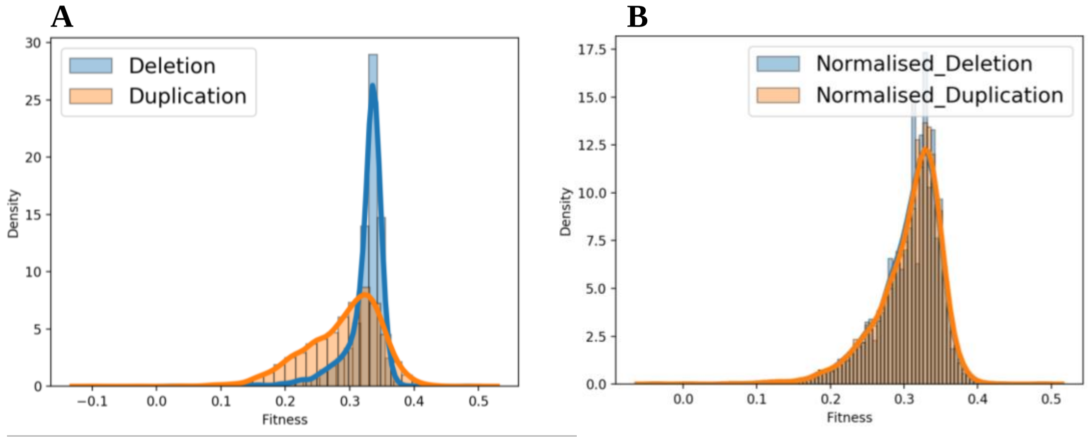

Fig. 1. A) Growth rates of a systematically mutated set of yeast strains obtained experimentally by Yoshikawa et al. There is high variability within the deletion and duplication datasets, but relatively low variability between datasets, allowing the application of quantile normalisation to both datasets. B) Quantile normalised distribution of single ORF deletion and duplication data.

C. Flux balance analysis with FlexFlux

We developed an algorithm to incorporate flux balance analysis when determining fitness values. The algorithm was implemented in Java, and relies on FlexFlux, a steady-state based metabolic network research tool for flux balance analysis [28]. FBA models were represented using the Systems Biology Markup Language (SBML) [29]. The implementation takes a list of deleted genes from the Chromosome class and runs FlexFlux to simulate gene knockout on the latest yeast consensus SBML genome-scale model, yeast_8.3.5 [27]. First, it creates a text-based constraint file which contains an objective function for maximizing the biomass. Next, the unique ID of every deleted ORF is obtained from the SBML model file. These IDs are written into the constraint file, and their status is set as “0” to represent deletion. Finally, FlexFlux is called with the constraint file and generates a result document. Some of the deleted ORFs might not be included in the SBML model, indicating that such ORFs are not involved in the well-understood metabolic network. In these cases, these ORFs are not written into the constraint file. If none of the deleted ORFs are included in the SBML model, the method returns wild-type fitness.

D. Parameterisation

A recombinase protein, Cre, randomly binds to a loxPsym site and initiates SCRaMbLing. We simulated this process using a parameter, *scrProb*, the probability of a Cre protein binding to a loxPsym site and triggering deletion or duplication. *scrProb* was estimated based on experimental data. On average, for chromosome synIXR, around six SCRaMbLE events occurred following four hours of induction with 1 μM estradiol [3]. Using this information, we estimated the probability of an event, using a simple simulation to investigate the correlation between *scrProb* and the average number of SCRaMbLEevents.

The simulation results indicated that when *scrProb* was between 0.2 and 0.4, the number of SCRaMbLE events was about six (Figures 2 and 3), which is the average number of SCRaMbLE events of surviving strains identified from the experimental data. *scrProb* could be further refined using additional experimental results, which are not currently available. With different *scrProb*, the survival rates of SCRaMbLE strains were different. If *scrProb* = 0.2, the survival rate was 147/1000; while with *scrProb* = 0.4, the survival rate was 207/1000. Hence, given further data about the survival rate, which could be obtained by running a simple wet-lab experiment comparing colony numbers between a SCRaMbLEd culture and a negative control, we could produce a more accurate estimate of the possibility of a loxPsym site being involved in a SCRaMbLE event. In this work, we set the *scrProb* to 0.2.

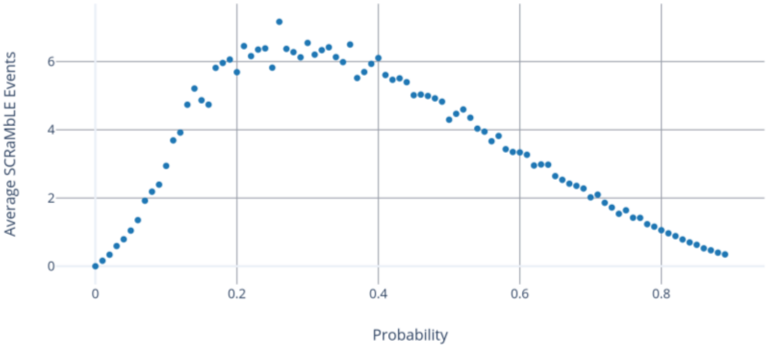

Fig. 2. Single point breaking probability versus the average number of SCRaMbLE events. When the possibility of a Cre binding to a loxPsym was around 0.2 to 0.4, the average number of SCRaMbLE events in chromosome synIXR was about six per surviving strain, which is the average number of SCRaMbLE events determined *in vivo*.

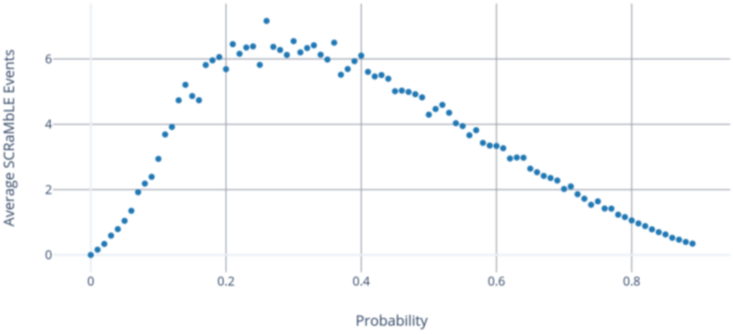

Fig. 3. Single point breaking probability versus the number of survivors. When the probability of a loxPsym binfing a Cre was around 0.2 to 0.4, the survival rate was between 14.7% and 20.7%.

E. *Fitness landscape analysis*

We simulated 4,280 strains, using a pseudo random number generator. The resulting dataset was used for fitness landscape analysis and further investigation.

Chromosome synIXR has loxP flanked 43 segments, making it difficult to visualize.We therefore used a dimension reduction algorithm named t-Distributed Stochastic Neighbor Embedding (t-SNE) to convert the 43-dimensional input into a two-dimensional array representing the genotype of a genome for every genome in the simulation [30]. t-SNE is a non-linear dimension reduction algorithm and is implemented by minimizing the Kullback-Leibler divergence between two similarity distributions: the pairwise similarities of high dimensional data points, and the corresponding low-dimensional embedded output points [31].

The two-dimensional array produced by t-SNE was used as the *x* and *y* axes of the fitness landscape, with the fitness score of each strain as the *z* axis. A tunable parameter of t-SNE, *perplexity*, balances the attention between local and global data by estimating the number of close neighbors of each point in the landscape. t-SNE was optimized by comparing the results produced by our simulator, resulting in a *perplexity* value of 40.

For a larger chromosome or a genome with much higher dimensions, due to pairwise similarities, computation is expensive, and t-SNE is inefficient. A linear dimension reduction algorithm such as Principal Components Analysis could be applied to reduce the dimensionality to under 50, followed by the use of t-SNE [32]. A potential approach would be to use LargeVis [33] instead of t-SNE for dimension reduction. LargeVis constructs *K*-nearest neighbor graphs more efficiently, and uses a principled probabilistic model for graph visualization. Another dimension reduction solution is to use UMAP, which is believed to preserve a better global structure than t-SNE [34].

III. Results

In this work we used an *in silico* evolutionary approach to develop a model describing the evolution of a population of yeast mutants whose genomes were perturbed using the SCRaMbLE system. We validated the model of deletions by comparing *in silico* and *in vivo* ORF deletions in mutant populations, validated the fitness function by reference to experimental data, and finally analysed the fitness landscape of the populations generated *in silico*, using the model.

A. *Deletion patterns in SCRaMbLEd genomes in silico and in vivo.*

To evaluate whether the results of the SCRaMbLE simulation were comparable with the wet lab data, we ran a simulation investigating deletion patterns on the circular chromosome synIXR. A random simulation dataset with 80 surviving strains was generated, and compared with a wet-lab experimental dataset with 64 surviving strains [3](Figure 4).

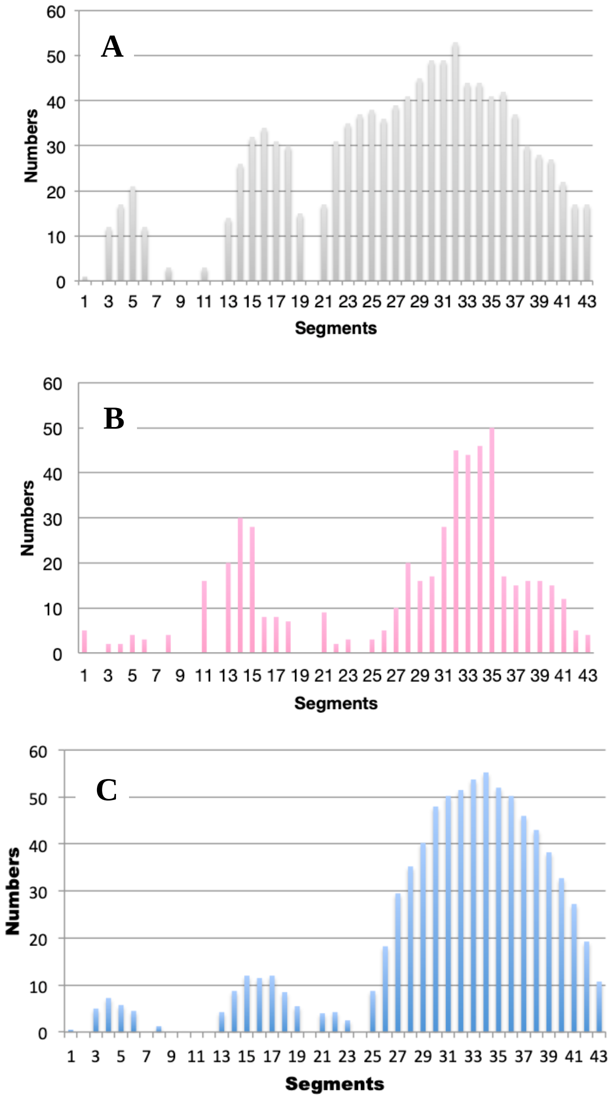

Fig. 4. A) Simulated deletion patterns using the modelling system, with deletion probability = 0.2. B) Deletion patterns obtained *in vivo* by Shen and co-workers [3]. C) Simulated deletion patterns with Segment 24 as an essential segment.

Comparing the above two figures (Figure 4A and 4B), we observed similar deletion patterns. Both experimental and simulation data have two subsequent deletion patterns ranging from Segment 13 to Segment 18, and Segment 20 to Segment 43 respectively. The peaks of these deletion patterns are similar, with around 30 deletions for pattern Segment 13-18 and around 50 deletions for pattern Segment 20-43. However, for the simulation results, the second pattern, between Segments 20 and 43, was much smoother than its counterpart. If we combined the fitness score of the simulations by isolating strains with a relatively high fitness score, we might produce a different perspective, and the results may be even more similar to those of the real data. However, since the relevant fitness data was not published with the experimental deletion patterns for synIXR [3], we could not make this comparison. The number of deletions of ORFs between Segments 21 and 31 was much higher than in the wet lab results. Further data about gene functions, from the *Saccharomyces* Genome Database (SGD)[35], suggested that null mutants of Segment 24 (carrying ORF YIR026C) decreased the competitive fitness of growth rate, which may explain the difference described above (Figure 4).

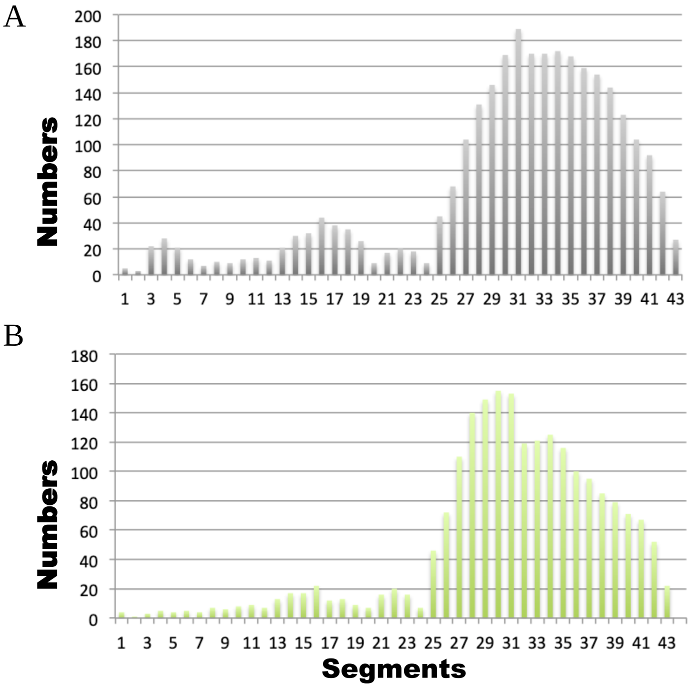

Fig. 5. *In silico* deletion patterns of all genomes (Figure 5A) and high fitness genomes (Figure 5B). The Y axis indicates the number of genomes with a related deleted segment in a dataset generated by SCRaMbLE simualation . The most significant difference between all strains and high fitness strains was observed in the deletion patterns between Segments 3 and 6, which were absent in the high fitness results, suggesting that the ORFs on these segments are related to high fitness.

TABLE I. IN SILICO FITNESS VERSUS QUALITATIVE EXPERIMENTAL FITNESS FOR THREE DELETION MUTANTS AND THE WILD-TYPE STRAIN. THE FITNESS FUNCTION DESCRIBED IN EQ3 WAS USED TO CALCULATE THE FITNESS SCORES OF SIMULATIONS. THE SIMULATION RESULTS WERE CONSISTENT WITH EXPERIMENTAL RESULTS.

Strain	Simulation Fitness	Experimental Fitness
Wild-type	1	Wild-type
ΔYIR004W	0.886	Slow-growth
ΔYIR005W	0.797	Slow-growth
ΔYIR020C	0.977	Slow-growth

According to the findings of Deutschbauer and co-workers, the ORF YIR026C is vital for strains competing with each other, due to the decreased competitive fitness of the null mutant [34]. Hence, strains SCRaMbLEd and sequenced from colonies in a wet lab are very likely to contain this ORF. This observation indicates that Segment 24 is non-essential when there are no other strains competing against it. However, due to its reduced growth rate, the YIR026C null mutant could not survive in competition with other scrambled strains on a plate, resulting in the deletion patterns being shifted from Segment 34 to Segment 30. Thus, YIR026C on Segment 24 is an essential ORF in a multicellularconsortium. Adding Segment 24 as an essential unit in the simulation produced results similar to the wet-lab results (Figure 4C).

Genomes with a higher fitness score than the wild-type genome were filtered for further analysis. The most significant differences between all genomes and high fitness genomes was observed in the deletion patterns between Segments 3 and 6, which were absent in the high fitness results, suggesting that the ORFs on these segments are related to high fitness (Figure 5).

Together, these results suggest that the SCRaMbLE simulator models deletion events with reasonable accuracy. Since the simulation is based on random numbers, those simulation results provide further evidence that the SCRaMbLE deletion process is largely random, but is constrained by its metabolic and phenotypic effects on the resulting mutant strains.

B. *Fitness function validation*

To validate the final fitness function (Eq. 3) used for the evolutionary process *in silico*, we calculated the fitness of the genomes of three reported slow-growing single deletion mutants: YIR004W on Segment 5, YIR005W on Segment 6, and YIR020C on Segment 18 (Table 1). These ORFs, which are supported by experimental evidence ($p < 0.05$), are all on the SynIXR chromosome. [3] These results (Table 1) were consistent with experimental results.

C. *Fitness distance correlation*

Fitness distance correlation (FDC) is usually used for optimizing genetic algorithms and analyzing the ruggedness of fitness landscapes. We applied it to the analysis and comparison of the fitness landscapes generated by the simulated SCRaMbLEd yeast mutant populations. FDC samples data points on the fitness landscape, and calculates the correlation between the measured fitness and the distance to the global optimal fitness. We used it to investigate the topology of the *in silico* landscape, and for studying the shape and size of the evolutionary search space.

The dataset, Fitness Dataset 1, with 4280 strains generated for fitness landscape analysis was also used here (Methods section F). We also constructed a smaller dataset, Fitness Dataset 2, derived from Fitness Dataset 1 by removing strains with mutated genes whose products were not modelled in the metabolic network. Usually, when the FDC value is between -0.15 and 0.15, optimization is difficult, because the fitness landscape is very rough. The FDC coefficient of 4,280 simulated scrambled genomes in Fitness Dataset 1 was 0.07. However, the FDC rose slightly to 0.09 for FDC Dataset 2. In this case, only strains with a fitness with a contribution from FBA contributed to the landscape. Scatter plots (Figure 6 and Figure 7) show the structure of FDC Dataset 1 and FDC Dataset 2. For FDC Dataset 2, a weak tendency could be observed (Figure 6) for fitness to increase with distance from the wild type. There was no significant structure in the correlation of fitness and distance for FDC Dataset 2, in which all mutants were retained (Figure 7). The results shown in the scatter plots are consistent with the FDC values. The slight difference between Fitness Dataset 1 and the Fitness Dataset 2 is probably because, while the whole fitness landscape is rugged, some patterns still exist in the metabolic-genes-mutated subset, since only three ORFs from synIXR are involved in the metabolic network. The generated fitness landscape has high ruggedness, based on FDC analysis.

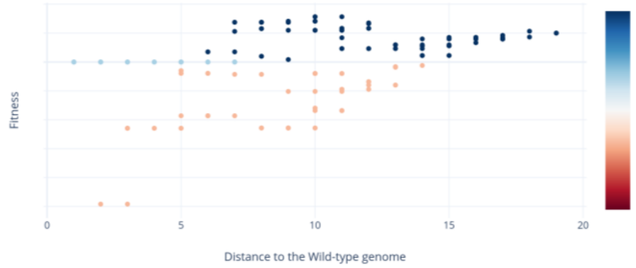

Fig. 6. Scatter plot of fitness-distance correlation of Fitness Dataset 2, with 237 genomes whose mutant gene products all featured in the metabolic network used to calculate genome fitness.

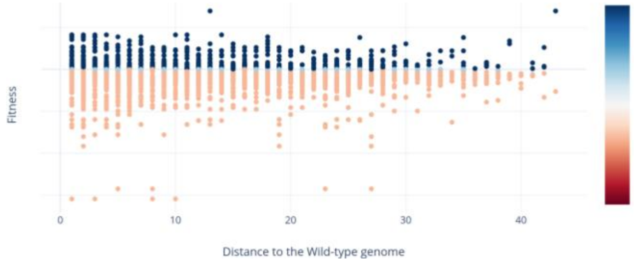

Fig. 7. Scatter plot of fitness-distance correlation of 4,280 simulated scrambled genomes. Mutants were included even if the mutant gene products did not feature in the metabolic model used to calculate the genome fitness, using FBA.

Fitness Dataset 1 was used for visualizing the fitness landscape (Figure 8A). Due to the lack of data points of mutated genomes with ORFs encoding enzymes, the SCRaMbLE simulator was also used to generate 401 SCRaMbLEd genomes with mutated genes whose products contributed to the metabolic network used to evaluate fitness (Figure 8B). All genomes were divided into four groups based on fitness: High, Wild-type, Low, and Dead (not shown). The dimension reduction algorithm t-SNE was used to convert high dimensional data to two dimensions. The fitness of most of the computationally SCRaMbLEd genomes was lower than that of the wild-type genome. A large number of genomes were inviable, due to deletion of essential genes. For Figure 8B, clear boundaries could be observed between each group of genomes, indicating that there are obvious patterns of genomes with mutations in genes contributing to metabolism. Mutations in these ORFs redirect the flux in the FBA model of the metabolic networks, and thus lead to changes in the fitness score. Some ORFs play a key role in the metabolic networks. By altering these key ORFs, the flux of the FBA model changed significantly. Although single gene deletion and duplication experimental data were integrated into the fitness function (Eq. 3), the fitness scores of genomes with and only with mutated ORFs encoding metabolic enzymes were fully dependent on the FBA results of the metabolic model. Since only three ORFs from synIXR were involved in the metabolic model, t-SNE easily captured the key features necessary to distinguish groups with different fitness.

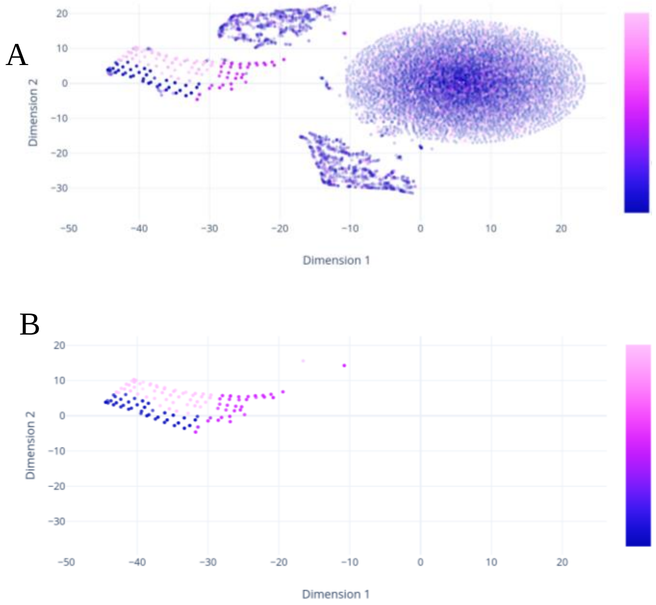

Fig. 8. Scatter plot of fitness distance correlation of 4,280 simulated genomes. Mutants were included even if the nutant gene products did not feature in the metabolic model used to calculate the genome fitness using FBA fitness landscape visualization.

For the other genomes generated by the simulator, including those with non-metabolic-related mutants, no clear patterns regarding fitness were observed (Figure 8A). These results, together with the results from the FDC analysis, indicated that the fitness landscape of SCRaMbLEd yeast had a high degree of ruggedness, although the metabolic enzyme-encoded mutation fitness subset might be smoother.

IV. Discussion

The aim of the Yeast2.0 project is to produce eukaryotic chromosomes which are easily manipulable, and which can produce millions of variants on the original, naturally evolved, genomes, which can then be searched for genomes which are viable, relatively easily cultivated, and have biological characteristics which are desirable for use in application areas such as the production of drug precursors, biofuels, and industrial enzymes. The choice of *S. cerevisiae* as the first target was due to its known safety, easy culture conditions, and well-studied biology, but this approach could be applied, in principle, to any other eukaryote.

The identification of valuable variants, in the original project, was reliant upon evolution *in vitro*. This approach, while demonstrably valuable, has several drawbacks. The most obvious issue is that evolution *in vitro* requires time, expertise, and technology, and is costly. More fundamentally, however, this use of evolution takes no account of the fitness landscape of the system; it essentially considers each variant as an individual entity, without considering the relationships between variants, their mutations, and their places in the fitness landscape.

The concept of a fitness landscape was first suggested by Sewall Wright in 1932 as a unifying concept in evolutionary theory [37]. The concept is based upon the observation that individuals with similar genomes tend to have similar phenotypes, and therefore similar fitnesses, although it is acknowledged that small changes in a genome can lead to large changes in the phenotype, and *vice versa*. A fitness landscape can therefore be characterized in term of its ruggedness, with smooth landscapes facilitating the evolutionary process, and making the prediction of fitness relatively easy, whereas rugged landscapes hamper evolution, and are hard to predict [9].

An understanding of the characteristics of the fitness landscape generated by the SCRaMbLE system will be of interest from a purely theoretical perspective because, in conjunction with the data produced in the biology laboratories of the consortium members, it provides an unparalleled opportunity to explore a real, extensive fitness landscape, and assess our understanding of this process by developing and evaluating simulation approaches. This project also has more directly practical applications. The ability to simulate the fitness landscape generated by the SCRaMbLE system may allow us to investigate the parameters of the system, and identify combinations of parameters which could lead to the generation of smooth fitness landscapes *in vivo*, thereby facilitating the process of artificial evolution *in vitro*, and saving time and money when identifying valuable variants. In the future, it may even be possible to develop genetic circuits to modulate the *in vivo* SCRaMbLE system to bias the evolutionary landscape in an optimal direction through the repression or enhancement of the recombination of particular loxP segments.

In this study, we developed a system for the simulation of SCRaMbLE *in silico*, including metrics for the distance between chromosomes, and for the fitness of the variants. These two metrics allow us to generate a fitness landscape for SCRaMbLE run with a specific set of parameters. We applied our simulator to a single landscape, identified clear clusters of variants, and evaluated the ruggedness of the landscape of the chromosome we used.

We found that the simulation results of the deletion patterns of synIXR we obtained were consistent with real-world data, a finding which confirms that the SCRaMbLE process tends to be random. By testing various values of the breakpoint possibility of the simulator, we inferred that the real-world possibility of a loxPsym site being involved in SCRaMbLE ranges from 0.2 to 0.4. This value could be narrowed down to a more precise number by running simple wet-lab experiments. We also found that the fitness landscape tends to be rugged, a finding which indicates that we may be able to improve the efficiency of the artificial evolution process by identifying changes which can be made to the system.

This work will form the basis for an extended study of simulation of the SCRaMbLE system. In future work, we will apply the system both to other chromosomes and to individual chromosomes repeatedly, to investigate whether these preliminary results apply to other chromosomes, and to evaluate the extent of variability between the landscapes that can be generated from a single chromosome. As discussed above, one of the most important aspects of the research will be the evaluation of the effects of modification of the parameters on the ruggedness of the landscape. It is highly likely that these parameters do not interact in a linear fashion, making it unlikely that optimal parameter settings can be achieved by chance in the laboratory. We also envisage being able to improve the parameterization for our model as new data emerges from future wet-lab SCRaMbLE studies.

In summary, we developed a simulator for the SCRaMbLE system, which has the potential to provide both theoretical and practical insights into this exciting new approach to bioengineering.

Acknowledgments

B.Y. acknowledges funding from the China Scholarship Council (CSC).

References

[1] J. N. Strathern, E. W. Jones, and J. R. Broach. Molecular biology of the yeast *Saccharomyces*. Cold Spring Harbor Laboratory, 1982.

[2] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. V´eronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre,et al., "Functional profiling of the *Saccharomyces cerevisiae* genome," Nature, vol. 418, no. 6896, pp. 387–391,2002.

[3] Y. Shen, G. Stracquadanio, Y. Wang, K. Yang, L. A. Mitchell, Y. Xue,Y. Cai, T. Chen, J. S. Dymond, K. Kang,et al., "Scramble generatesdesigned combinatorial stochastic diversity in synthetic chromosomes, "Genome research, vol. 26, no. 1, pp. 36–49, 2016.10

[4] J. Dymond and J. Boeke, "The *Saccharomyces cerevisiae* scramble systemand genome minimization," Bioengineered, vol. 3, no. 3, pp. 170–173, 2012.

[5] E. Pitzer and M. Affenzeller, "A comprehensive survey on fitness landscape analysis," Recent Advances in Intelligent Engineering Systems, pp. 161–191,2012.

[6] D. J. Earl and M. W. Deem, "Evolvability is a selectable trait," Proceedings of the National Academy of Sciences, vol. 101, no. 32, pp. 11531–11536,2004.

[7] J. A. G. De Visser and J. Krug, "Empirical fitness landscapes and the predictability of evolution," Nature Reviews Genetics, vol. 15, no. 7, pp. 480–490, 2014.

[8] M. Ueda, N. Takeuchi, and K. Kaneko, “Stronger selection can slow downevolution driven by recombination on a smooth fitness landscape,” PloS One, vol. 12, no. 8, p. e0183120, 2017.

[9] S. Kauffman and S. Levin, “Towards a general theory of adaptive walks onrugged landscapes,” Journal of Theoretical Biology, vol. 128, no. 1, pp. 11–45, 1987.

[10] D. J. Kvitek and G. Sherlock, “Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape,” PLoS Genet, vol. 7, no. 4, p. e1002056, 2011.

[11] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in Soviet Physics Doklady, vol. 10, pp. 707–710, Soviet Union,1966.

[12] R. Schwartz and M. Dayhoff, “Detection of distant relationships based on point mutation data,” Evolution of protein molecules (eds. H. Matsubaraand T. Yamanaka), pp. 1–16, 1978.

[13] S. Henikoff and J. G. Henikoff, “Amino acid substitution matrices fromprotein blocks,”Proceedings of the National Academy of Sciences, vol. 89, no. 22, pp. 10915–10919, 1992.

[14] K. Yoshikawa, T. Tanaka, Y. Ida, C. Furusawa, T. Hirasawa, andH. Shimizu, “Comprehensive phenotypic analysis of single-gene deletion and overexpression strains of *Saccharomyces cerevisiae*,” Yeast, vol. 28,no. 5, pp. 349–361, 2011.

[15] I. Ulitsky and R. Shamir, “Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction networks,” Molecular Systems Biology, vol. 3, no. 1, p. 104, 2007.

[16] S. L. Ooi, X. Pan, B. D. Peyser, P. Ye, P. B. Meluh, D. S. Yuan, R. A.Irizarry, J. S. Bader, F. A. Spencer, and J. D. Boeke, “Global synthetic-lethality analysis and yeast functional profiling,” Trends in Genetics,vol. 22, no. 1, pp. 56–63, 2006.11

[17] J. Weile, K. James, J. Hallinan, S. J. Cockell, P. Lord, A. Wipat, and D. J. Wilkinson, “Bayesian integration of networks without gold standards,” Bioinformatics, vol. 28, no. 11, pp. 1495–1500, 2012.

[18] P. Erdos and A. Renyi, “On random graphs i,”Publ. Math. Debrecen, vol. 6, pp. 290–297, 1959.

[19] S. H. Strogatz, “Exploring complex networks,” Nature, vol. 410, no. 6825,pp. 268–276, 2001.

[20] I. J. Farkas, I. Der´enyi, A.-L. Barab´asi, and T. Vicsek, “Spectra of” real-world” graphs: Beyond the semicircle law,” in The Structure and Dynamicso of Networks, pp. 372–383, Princeton University Press, 2011.

[21] A.-L. Barabasi and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization,” Nature Reviews Genetics, vol. 5, no. 2, pp. 101–113,2004.

[22] J. S. Hallinan, G. Misirli, and A. Wipat, “Evolutionary computation for thedesign of a stochastic switch for synthetic genetic circuits,” in 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 768–774, IEEE, 2010.

[23] J. Hallinan, K. James, and A. Wipat, “Network approaches to the functional analysis of microbial proteins,” Advances in Microbial Physiology,vol. 59, pp. 101–133, 2011.

[24] K. James, J. R. Tarn, S. Al-Ali, J. Hallinan, D. A. Young, and W. F.Ng, “Integration of gene expression data with interaction and annotationdata reveals patterns of connection between primary sjogren’s syndromeassociated genes and immune processes,” Rheumatology, vol. 53, pp. i136–i136, 2014.

[25] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley,E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel,et al., “*Saccharomyces* genome database: the genomics resource of budding yeast,” Nucleic Acids Research, vol. 40, no. D1, pp. D700–D705, 2012.

[26] J. D. Orth, I. Thiele, and B. Ø. Palsson, “What is flux balance analysis?,” Nature Biotechnology, vol. 28, no. 3, pp. 245–248, 2010.

[27] H. Lu, F. Li, B. J. S´anchez, Z. Zhu, G. Li, I. Domenzain, S. Marcisauskas,P. M. Anton, D. Lappa, C. Lieven, et al., “A consensus *S. cerevisiae* metabolic model yeast and its ecosystem for comprehensively probing cellular metabolism,” Nature Communications, vol. 10, no. 1, pp. 1–13, 2019.

[28] L. Marmiesse, R. Peyraud, and L. Cottret, “Flexflux: combining metabolicflux and regulatory network analyses,” BMC Systems Biology, vol. 9, no. 1,p. 93, 2015.12

[29] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, et al., “The systems biology markup language (SBML): a medium for representation andexchange of biochemical network models,” Bioinformatics, vol. 19, no. 4,pp. 524–531, 2003.

[30] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” Journal of Machine Learning Research, vol. 9, no. Nov, pp. 2579–2605, 2008.

[31] S. Kullback and R. A. Leibler, “On information and sufficiency,” Annals Of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, 1951.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel,M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., “Scikit-learn: Machine learning in python,” Journal of Machine Learning Research, vol. 12,no. Oct, pp. 2825–2830, 2011.

[33] J. Tang, J. Liu, M. Zhang, and Q. Mei, “Visualizing large-scale and high-dimensional data,” in Proceedings of the 25th International Conference on the World Wide Web, pp. 287–297, 2016.

[34] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng,F. Ginhoux, and E. W. Newell, “Dimensionality reduction for visualizing single-cell data using UMAP,” Nature Biotechnology, vol. 37, no. 1, p. 38,2019.

[35] S. A. Chervitz, E. T. Hester, C. A. Ball, K. Dolinski, S. S. Dwight, M. A.Harris, G. Juvik, A. Malekian, S. Roberts, T. Roe,et al., “Using the *Saccharomyces* genome database (SGD) for analysis of protein similarities and structure,” Nucleic Acids Research, vol. 27, no. 1, pp. 74–78, 1999.

[36] A. M. Deutschbauer, D. F. Jaramillo, M. Proctor, J. Kumm, M. E. Hillen-meyer, R. W. Davis, C. Nislow, and G. Giaever, “Mechanisms of haploin-sufficiency revealed by genome-wide profiling in yeast,” Genetics, vol. 169, no. 4, pp. 1915–1925, 2005

[37] S. Wright, “The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution,” 1932.