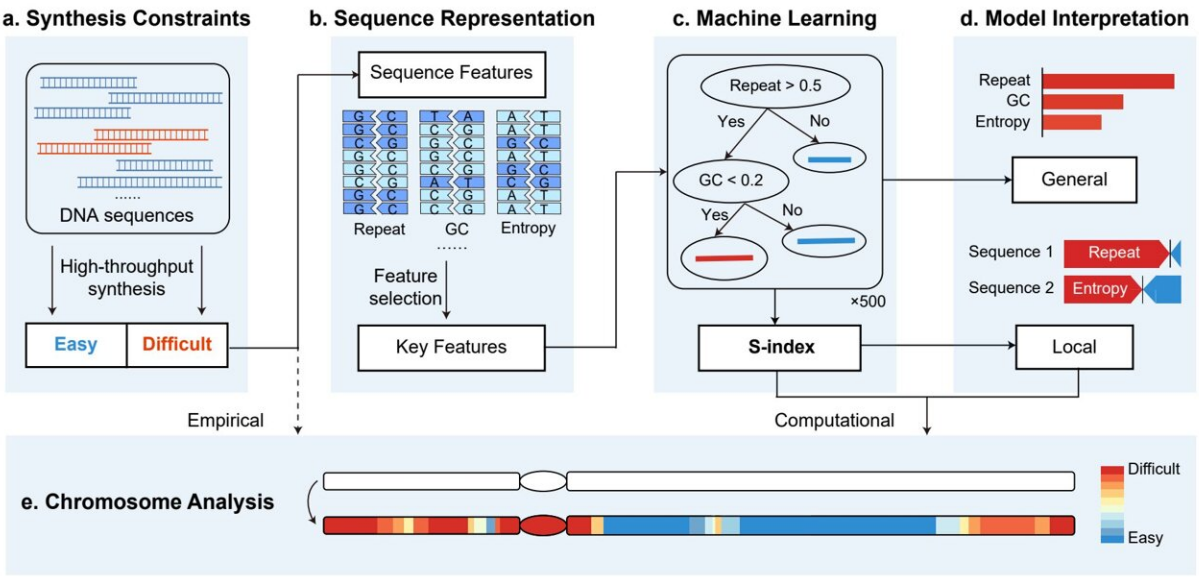
A machine learning framework to predict and quantify synthesis difficulties for designer chromosomes

March 27 2023



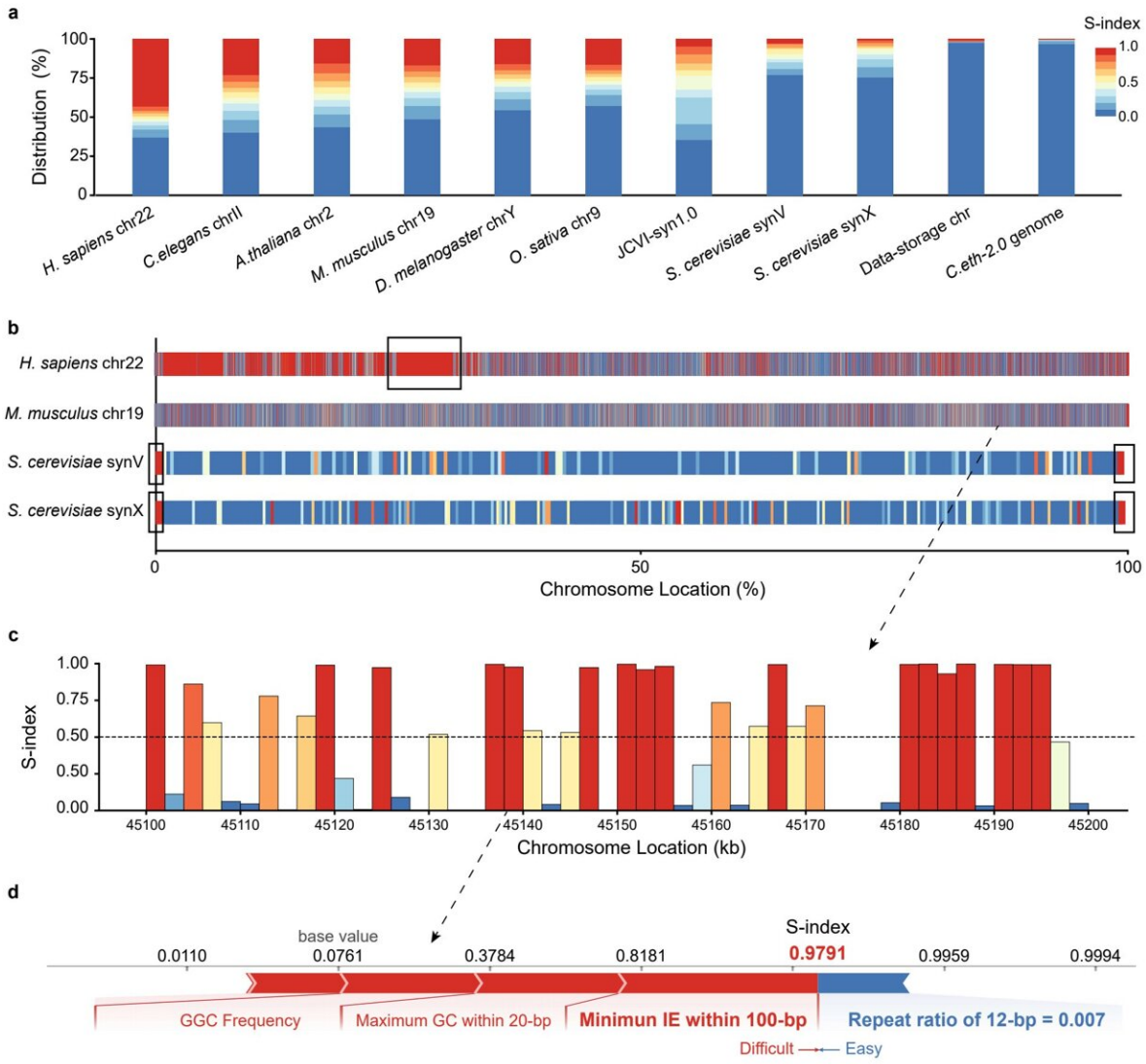
A, Collection of the DNA sequences obtained from high-throughput synthesis. The sequences were classified into easy-to-synthesize (blue) or difficult-to-synthesize (red). B, Graphical representations of DNA sequences: repeat, GC content, information entropy and other types of features. Key features were identified from these sequence features by machine learning methods. C, The XGBoost algorithm utilized to build the classification model and calculate the S-index. D, Methods used to interpret the model. The feature contributions were quantified according to the global importance scores and local SHAP explanations. e, Application of the S-index on a specific chromosome. The heatmap indicates the synthesis difficulties for the different fragments, which range from difficult (red) to easy (blue). The white sequences indicate the

unanalyzed chromosome sequence. Credit: Science China Press

Artificially synthesizing genomes has broad prospects in fields such as medical research and developing industrial strains. From the synthesis of the artificial life JCVI-syn1.0 by Craig Venter's team in 2010, to the rewriting and synthesis of the prokaryotic E. coli genome, and to the Sc2.0 project's artificial synthesis of the yeast genome, researchers are constantly advancing in the depth and breadth of genome design and synthesis.

However, there are still difficulties in synthesizing certain gene segments , ultimately leading to the inability to complete artificial chromosomes, which limits the application and promotion of artificial genome synthesis technology. To address this issue, the team of Professor Yingjin Yuan from Tianjin University has developed an interpretable machine learning framework that can predict and quantify the difficulty of chromosome synthesis, providing guidance for optimizing chromosome design and synthesis processes.

The research team designed an efficient feature selection method by analyzing data of a large number of known chromosome fragments, and identified six key sequence features that cover energy and structural information during DNA chemical synthesis and assembly. Based on these results, the team developed an eXtreme Gradient Boosting (XGBoost) model that can effectively predict the synthesis difficulties of chromosome fragments.



A, The distribution of DNA sequences with different S-index for the natural and synthetic chromosomes and genomes. The heatmap shows the S-index for the different sequences and the color has the same meaning in B and C. B, The difficulties of synthesizing DNA sequences for the different locations within the chromosomes. The black boxes mark the centromeric satellite of Homo sapiens chromosome 22 and telomeres of synV and synX. c, The S-index for the 45,100-45,200-kb region of M. musculus chr19. D, Force plot for 45,138-45,140 kb sequence of M. musculus chr19. The feature with a positive effect value is highlighted in red, and the feature with a negative effect value is highlighted in blue. Photo credit: Yan Zheng. Credit: Yan Zheng

The model achieved an AUC (area under the receiver operating characteristic curves) of 0.895 in cross-validation and an AUC of 0.885 on an independent test set in collaboration with a DNA synthesis company, demonstrating a high accuracy and predictive ability.

The research team proposed a Synthesis difficulty Index (S-index) based on the SHAP algorithm to evaluate and interpret the synthesis difficulties of chromosomes. The study found that there were significant differences in the synthesis difficulties of different chromosomes, and the S-index could quantitatively explain the causes of synthesis difficulties for some gene fragments, providing a basis for chromosome sequence design and synthesis and improving the efficiency and success rate of designer chromosome synthesis.

This achievement provides a practical tool for researchers in chromosome engineering and genome rewriting, and is expected to provide more comprehensive guidance and support for chromosome design and synthesis.

The paper is published in the journal *Science China Life Sciences*.

More information: Yan Zheng et al, Machine learning-aided scoring of synthesis difficulties for designer chromosomes, *Science China Life Sciences* (2023). DOI: 10.1007/s11427-023-2306-x

Provided by Science China Press