A Human-Computation Platform for Multi-Scale Genome Analysis

Akash Singh*, Chris Drogaris*, Elena Nazarova, Mathieu Blanchette and **Jérôme Waldispühl**
School of Computer Science, McGill University
jeromew@cs.mcgill.ca

Anjum Ibna Matin*, Mardel Maduro and **Olivier Tremblay-Savard**
Department of Computer Science, University of Manitoba
tremblao@cs.umanitoba.ca
(*equal contribution)

Abstract

We introduce a citizen science framework for a collective curation genomic annotation at multiple levels of the genome organisation. Currently our system aims to integrate a fully new version of Phylo solving the Multiple Sequence Alignment (MSA) problem, with a new game aiming to understand the evolution of large scale genomic regions.

Introduction

Human computation tasks stream appear in many scientific problems, such as in astronomy (Skibba et al. 2012), molecular biology (Cooper et al. 2010; Kawrykow et al. 2012), neuroscience (Kim et al. 2014), and even quantum physics (Lieberoth et al. 2015). Phylo (`http://phylo.cs.mcgill.ca`), is a citizen science game which aims to help us improve the accuracy of the comparison of DNA data (Kawrykow et al. 2012; Kwak et al. 2013). Phylo annotations aim to help geneticists in various tasks such as studying evolution and understanding mutations that cause genetic disorders. First, we present a new design of Phylo which features novel functionalities such as transcription factors annotations, better feedback to users, and enhancements in game metrics based on the observation made on the data collected by Phylo since 2010. Next, we introduce a new game addressing a different problem in comparative genomics occuring at a higher level of the genomic organization: the genome sorting problem. In the latter, the user aims to find the minimum number of evolutionary events (e.g. duplications, deletions, inversions) needed to transform one genome into the other.

Phylo

Proposed Methods

Phylo will still primarily aim to solve the multiple sequence alignment problem, used to reveal conserved DNA sequence across species. In addition, our novel implementation will feature improved design, enhanced portability, and new functionalities.

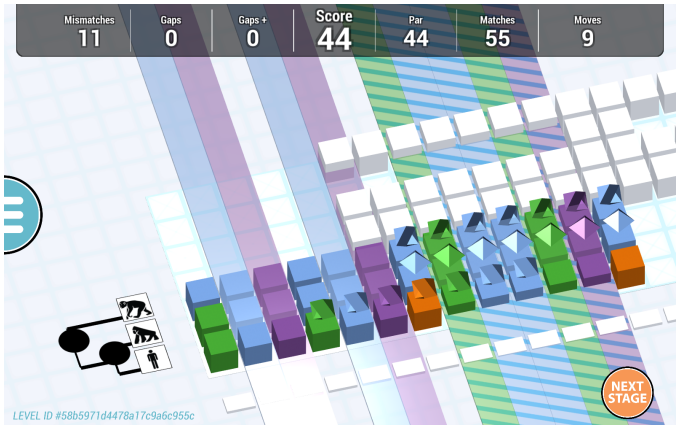

Figure 1: Phylo novel interface. Each color of cube represents a nucleotide in DNA sequence. Users can move the bricks horizontally, and try to maximize similarity within a column. The diamonds on top of the cubes indicate the putative presence of a functional motifs.

1. *Motif based alignments*: The DNA binding specificity of transcription factors (TF) are generally highly conserved among related organisms. Database of known TF motifs is now maintained in Phylo (Heinz et al. 2010). But this information is not highly specific. Users are shown with a roster of putative TF motifs present in the puzzle. The alignment of these motifs will allow the users to collect rewards and help us to false positive annotations.
2. *Enhanced feedback to players:* We now made the game more informative so as to educate the players simultaneously by including the exact disease causing mutation, motif sequences and its annotation in the human gene sequence which will be shown to them once they complete the puzzle.
3. *Evaluation of users based on their click-stream data*: Our systems will include mechanisms to analyze user tactics through an analysis of click-stream data. This version of Phylo will feature an inbuilt storyline, helping us to introduce advanced concepts to the users.
4. *Transition to RNA:* This version of Phylo will also

take a transition for solving MSA of RNA sequences. The massive player count of Phylo will definitely help in enhancing RNA alignments.

Enhancements in Metrics

1. *Puzzle Extraction:* The puzzle database is obtained from Ensembl genome browser(Aken et al. 2016). Human genes associated to diseases and their mutations were noted from HGMD(Lualdi et al. 2017). Puzzles are extracted using EnsEMBL Compara. Puzzle extraction criteria have also been updated by additional constraints to increase the accuracy of Phylo.
2. *Puzzle Difficulty:* Difficulty level of puzzles is important in order to route it to its correct difficulty in the game and also routing it based on the players skill level. We now propose to use the neural network regression model using a vector of 11 features: (1-4) proportions of A, C, G, and T in S, (5) proportion of gaps, (6) mean GC content (this relates to the structural properties of DNA), (7) mean entropy of alignment columns, (8) average length of sequences, (9) number of sequences, (10) tree-entropy, and (11) score of machine-computed alignment.

Evolution Game

Genome Sorting Problem

Similarly to the MSA problem, the genome sorting problem has received a lot of attention from the comparative genomics community in the last decades. Given two genomes (represented by gene orders), the goal is to find the shortest sequence of biological events to transform one genome into the other. When both genomes have exactly one copy of each gene, the problem is simpler. For example, in this context (one copy of each gene), sorting genomes by reversals (inversions of segments of genes) can be solved in polynomial time (Hannenhalli and Pevzner 1999; Tannier, Bergeron, and Sagot 2007). However, it is often the case that genomes have multiple copies of certain genes. In this case, the genome sorting problem becomes NP-hard in the case of sorting by reversals (Christie and Irving 2001), or sorting by reversals and duplications (Chen et al. 2005) for example.

Objectives

As a first step towards the development of new algorithmic methods to study genome evolution in highly divergent genomes with duplicate genes, our goal is to develop a new crowdsourcing and human computing game that will ask players to find optimal evolutionary scenarios transforming one genome into another. Our game takes the form of a puzzle game and will be targeting both casual players and biology students, just like Phylo. In addition to obtaining optimal evolutionary scenarios from the players, we will record every move that the players make. This will allow us to analyze the strategies employed by the players to solve those problems, and this information will then be used as inspiration to develop new algorithmic methods.

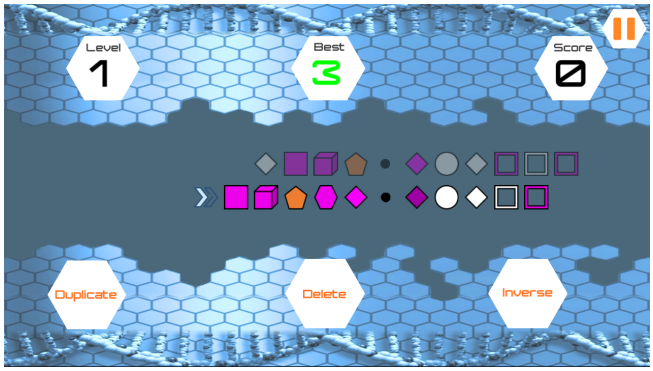

Figure 2: Evolution Game interface. Each colored shape represents one gene and the black dot in the middle represents the terminus of replication. The top genome is the target: players have to apply evolutionary events to the bottom sequence to transform it into the target.

Input Data and Possible Events

Our game will first focus on sorting bacterial genomes, which are simpler because they have only one chromosome. The list of possible evolutionary events (i.e. available player moves) will be the three types of events that are mostly observed in bacterial genomes: duplications of genes, deletions of genes, and inversions of segments of genes. Bacterial genomes also have an interesting characteristic: inversion events occur mostly around the terminus of replication, which is located approximately in the middle of the genome. Consequently, only inversion events of segments of genes that are overlapping the terminus will be considered as valid moves.

Interface

The interface of the Evolution Game is shown in Figure 2. At the very top of the screen, information on the level, best score (out of all the other players for this level) and the current score of the player is shown. The player's score on a level is simply the total number of evolutionary events that was used up to this point. The middle part of the screen shows the two genomes, represented as sequences of colored shapes (i.e. genes). The top genome is the target, whereas the bottom one is the mutable genome. Finally, the bottom panel has three buttons, which correspond to the three possible evolutionary events. In order to apply one of these events, the player has to select the genes of the mutable sequence that will be modified, and then click on the event type. In the case of a duplication, there is one last step after clicking the duplicate button: select where the genes will be duplicated in the genome. Once both genomes are identical, the player will be prompted to continue to the next level, or to restart the level if the player wants to improve his/her score.

References

Aken, B. L.; Achuthan, P.; Akanni, W.; Amode, M. R.; Bernsdorff, F.; Bhai, J.; Billis, K.; Carvalho-Silva, D.; Cummins, C.; Clapham, P.; et al. 2016. Ensembl 2017. *Nucleic acids research* gkw1104.

Chen, X.; Zheng, J.; Fu, Z.; Nan, P.; Zhong, Y.; Lonardi, S.; and Jiang, T. 2005. Assignment of orthologous genes via genome rearrangement. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2(4):302–315.

Christie, D. A., and Irving, R. W. 2001. Sorting strings by reversals and by transpositions. *SIAM Journal on Discrete Mathematics* 14(2):193–206.

Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popović, Z.; et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466(7307):756–760.

Hannenhalli, S., and Pevzner, P. A. 1999. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM (JACM)* 46(1):1–27.

Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y. C.; Laslo, P.; Cheng, J. X.; Murre, C.; Singh, H.; and Glass, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell* 38(4):576–589.

Kawrykow, A.; Roumanis, G.; Kam, A.; Kwak, D.; Leung, C.; Wu, C.; and Zarour, E. 2012. Phylo players, luis sarmenta, mathieu blanchette, and jérôme waldispühl. phylo: a citizen science approach for improving multiple sequence alignment. *PloS one* 7(3):e31362.

Kim, J. S.; Greene, M. J.; Zlateski, A.; Lee, K.; Richardson, M.; Turaga, S. C.; Purcaro, M.; Balkam, M.; Robinson, A.; Behabadi, B. F.; et al. 2014. Space-time wiring specificity supports direction selectivity in the retina. *Nature* 509(7500):331–336.

Kwak, D.; Kam, A.; Becerra, D.; Zhou, Q.; Hops, A.; Zarour, E.; Kam, A.; Sarmenta, L.; Blanchette, M.; and Waldispühl, J. 2013. Open-phylo: a customizable crowd-computing platform for multiple sequence alignment. *Genome biology* 14(10):R116.

Lieberoth, A.; Pedersen, M. K.; Marin, A. C.; Planke, T.; and Sherson, J. F. 2015. Getting humans to do quantum optimization-user acquisition, engagement and early results from the citizen cyberscience game quantum moves. *arXiv preprint arXiv:1506.08761*.

Lualdi, S.; Zotto, G. D.; Zegarra-Moran, O.; Pedemonte, N.; Corsolini, F.; Bruschi, M.; Tomati, V.; Amico, G.; Candiano, G.; Dardis, A.; et al. 2017. In vitro recapitulation of the site-specific editing (to wild-type) of mutant ids mrna transcripts, and characterization of ids protein translated from the edited mrnas. *Human Mutation*.

Skibba, R. A.; Masters, K. L.; Nichol, R. C.; Zehavi, I.; Hoyle, B.; Edmondson, E. M.; Bamford, S. P.; Cardamone, C. N.; Keel, W. C.; Lintott, C.; et al. 2012. Galaxy zoo: the environmental dependence of bars and bulges in disc galaxies. *Monthly Notices of the Royal Astronomical Society* 423(2):1485–1502.

Tannier, E.; Bergeron, A.; and Sagot, M.-F. 2007. Advances on sorting by reversals. *Discrete Applied Mathematics* 155(6):881–888.