Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
An Adjunct to the Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing

Phylo and Open-Phylo: A Human-Computing Platform for Comparative Genomics

Jérôme Waldispühl, Mathieu Blanchette
School of Computer Science, McGill University
3480 University Street
Montreal, QC H3A 0E9

Abstract

Comparative genomics is a field of research that aims to provide us accurate mappings between the genetic material of multiple species. These techniques are useful for biomedical research and evolutionary studies. We present Phylo and Open-Phylo, an open citizen science platform and human computing-game for comparative genomic studies that reached more than 300,000 users.

Introduction

In molecular biology, comparative genomics is a field of research that aims to provide us accurate mappings between the genetic material of multiple species. In practice, genomes are compared and mapped by solving the multiple sequence alignment (MSA) problem. It consists in aligning multiple DNA sequences in order to reveal similar patterns conserved into the genome of multiple organisms. With applications ranging from evolutionary studies to the analysis of genetic diseases, it is one of the oldest and still one of the most powerful technology used to analyze genomic data.

Since the early 70's, the MSA problem has been the object of many studies. Although it has been demonstrated to be $\mathcal{NP}$-hard (Wang and Jiang 1994), we can today rely on efficient heuristics to compute "good" alignments of several dozens of complete genome sequences (Blanchette 2007). Nonetheless, the optimality of these alignments is not guaranteed. Worse, there is no clear consensus on the objective function to use for measuring the quality of an alignment.

Crowdsourcing biological problems in scientific games is becoming increasingly popular. Scientific games like Foldit (Cooper et al. 2010) and Phylo (Kawrykow et al. 2012) illustrate the potential of crowd-sourcing and human-computing techniques in solving hard problem arising in Biology.

Our group was the first to bring citizen science to the field of genomics when, in 2010, we released Phylo (`http://phylo.cs.mcgill.ca`) – a human-computing framework to solve the MSA problem. Phylo is a game-with-a-purpose (von Ahn and Dabbish 2008), which converts the MSA problem into a casual puzzle game that can be played by ordinary web users without prior knowledge of the biological context. In our original study (Kawrykow et al. 2012), the puzzles were extracted from a 44-species MSA stored at the UCSC genome browser (Kent et al. 2002), and the best solutions have been re-inserted at their original locations to produce a higher quality MSA. Compared to its predecessors, one of the main innovations of Phylo was to push the gamification aspect to its limits and propose a casual game that does not require any scientific background.

In 2013, we expanded the concepts of citizen science open our system to the whole scientific community with our new citizen science platform Open-Phylo (Kwak et al. 2013). Unlike classical crowdsourcing projects and scientific games that are all built to the benefit of the researchers who designed the system, i.e. only they can select the problems that are submitted to the community, and access and analyze the solutions submitted by the players, Open-Phylo breaks this paradigm and opens human-computing resources to the whole scientific community. The unique feature of Open-Phylo is that it is not only computing *by the people*, but also *for the people*.

In this paper, we describe Phylo and its latest development Open-Phylo. We provide recent usage statistics, and illustrate the scientific impact of this technology on genomic research.

Methods

Phylo

Phylo is a casual web game that looks like a classical puzzle game such as Tetris (See Fig. 1). Genetic sequences are represented by chains of bricks of 4 colours representing the 4 bases of the genetic code: A, C, G and T. Ecah row store the genetic sequence of a different species. By moving the bricks horizontally, the players try to create column with identical colours while avoiding as much as possible colour mismatches and gaps.

Phylo aims to improve alignments that have been precomputed by state-of-the-art algorithms. Indeed, genomes are huge (up to billions of bases). We cannot ask single users to align them from scratch. Instead, we designed an hybrid system taking advantage of the specificities of human and computer capacities. First, computers calculate an initial alignment. Next, we screen these alignments and extract smaller regions where the program appears to have dif-

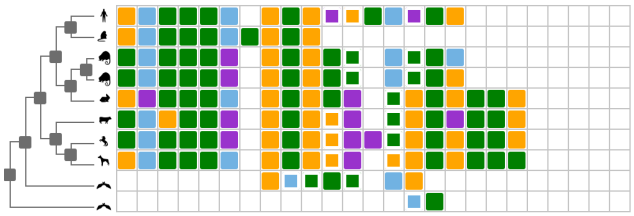

Figure 1: Screenshot of Phylo.

ficulties to find a good alignment. Finally, we ask web user to solve these puzzles on our website and re-insert the best solutions in the original genome alignment in order to build a better alignment.

The original puzzles are played on a relatively small grid (10 rows and 25 columns). Eventually, the smaller puzzles could be solved with exhaustive searches (Knudsen 2003). Nonetheless, this approach would require important computing resources and is not scalable (Altschul 1989). To demonstrate the efficiency of our techniques and make a better use of the most experienced players, we developed an expert version of the game which proposes to any (trained) user who completed at least 20 classical puzzles to play larger puzzle (currently on grids with up to 350 columns). In addition, there, several gamers can collaborate to solving the puzzles.

Open-Phylo

Human-computing games such as Phylo and Foldit use crowds to help solving well-defined scientific problems. The success of these games can be, at least partially, attributed to the broad scientific interest of the problems they are trying to solve. However, despite the potential usefulness of these techniques for the whole community, the original design of these citizen science systems is inherently asymmetric: Only the owners of the games can choose and control the instances to be solved.

We addressed this issue when in October 2013, we released Open-Phylo (Kwak et al. 2013), a submission interface (`http://phylo.cs.mcgill.ca/submit`) that enables geneticists to upload their own sequence data, select the portions they wish the player community to focus on, and manage the effort of the crowd of gamers playing their alignment. In other words, Open-Phylo is an open access, human-powered crowd-computer accessible to the whole community through a web interface.

Results

Our first experiments aimed to analyze promoter regions (i.e. sequences controlling the expression of the gene) of 739 disease-related genes from the UCSC Genome Browser (Kent et al. 2002). Our results showed that the solutions submitted by players contributed to improving the accuracy of up to 70% of these alignments (Kawrykow et al. 2012). More recently, we applied our techniques to study alignments of promoter regions of three major cancer-related genes (Kwak et al. 2013). In contrast with our first study, we showed that Phylo improves the solution calculated by any of the most recent alignment algorithms (MULTIZ, PRANK, MUSCLE and T-Coffee) that are all using distinct objective functions. Moreover, we also showed that expert players can significantly improve the performance of the system when casual gamers fail to improve an alignment.

Usage statistics

Between November 29^{th} 2010 and July 1^{st} 2014, Phylo collected 808225 classic puzzle solutions. To this, we must add 4651 large puzzles solved by expert players. 31784 users registered during this period. However, Phylo also allows users to play anonymously. Our statistics suggest a ratio of 1 registration for 10 to 15 visitors. We estimate that between 300000 and 500000 web users from more than 180 different countries visited our website. In average, guests play 1 or 2 games, but register users play up to 40 puzzles.

Open-Phylo is currently tested by researchers from the North Carolina museum of natural science to align genomes of lanternflies. These data aim to provide us the material needed to have a build better phylogenies of these insects.

References

Altschul, S. F. 1989. Gap costs for multiple sequence alignment. *J Theor Biol* 138(3):297–309.

Blanchette, M. 2007. Computation and analysis of genomic multi-sequence alignments. *Annu Rev Genomics Hum Genet* 8:193–213.

Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popović, Z.; and Players, F. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466(7307):756–60.

Kawrykow, A.; Roumanis, G.; Kam, A.; Kwak, D.; Leung, C.; Wu, C.; Zarour, E.; Phylo players; Sarmenta, L.; Blanchette, M.; and Waldispühl, J. 2012. Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS One* 7(3):e31362.

Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; and Haussler, D. 2002. The human genome browser at ucsc. *Genome Res* 12(6):996–1006.

Knudsen, B. 2003. Optimal multiple parsimony alignment with affine gap cost using a phylogenetic tree. In *WABI*, 433–446.

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 Biol* 14(10):R116.

von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Commun. ACM* 51(8):58–67.

Wang, L., and Jiang, T. 1994. On the complexity of multiple sequence alignment. *J Comput Biol* 1(4):337–48.