Lessons from an online massive genomics computer game

Akash Singh, Faizy Ahsan, Mathieu Blanchette and **Jérôme Waldispühl**
School of Computer Science
McGill University
jeromew@cs.mcgill.ca

Abstract

Crowdsourcing through human-computing games is an increasingly popular practice for classifying and analyzing scientific data. Early contributions such as Phylo have now been running for several years. The analysis of the performance of these systems enables us to identify patterns that contributed to their successes, but also possible pitfalls. In this paper, we review the results and user statistics collected since 2010 by our platform Phylo, which aims to engage citizens in comparative genome analysis through a casual tile matching computer game. We also identify features that allow predicting a task difficulty, which is essential for channeling them to human players with the appropriate skill level. Finally, we show how our platform has been used to quickly improve a reference alignment of Ebola virus sequences.

1 Introduction

Human computation has emerged as a popular approach to solve large-scale scientific problems 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). With more applications of this technology underway, it is important to identify the factors that contributed to the successes of this approach, and improve the aspects that did not work as well as expected. To this end, the analysis of the data collected by the earliest systems may reveal important patterns that could benefit to the next generation of scientific games.

In 2010, we released a game-with-a-purpose named Phylo (`http://phylo.cs.mcgill.ca`), which aims to help us improving the accuracy of the comparison of DNA data (Kawrykow et al. 2012; Kwak et al. 2013). This problem, known as the multiple sequence alignment (MSA), is an essential piece of a vast body of biological studies (Edgar and Batzoglou 2006). It aims to help revealing conserved patterns that may have a functional role. Phylo's tasks are presented as in a casual tile matching game, where alignment problems are embedded in a puzzle accessible to any player, including those without any prior training in biology or computer science. Most importantly, the game is intuitive and allows the players to play Phylo without understanding the underlying biology, and without completing any tutorial. This broadens the spectrum of participants and taps into the computing power generated by regular, non-scientist human computers. Special care is taken to present the puzzles in a fun, exciting, and accessible manner while retaining the scientific interpretability of the data.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Formally, a multiple sequence alignment is represented as a matrix with a motive to place in the same column characters that are homologous (i.e. derived from a same common ancestor), possibly inserting gap characters to account for the presence of insertions and deletions. Although the problem of pairwise sequence alignment can be solved optimally in quadratic time (Feng and Doolittle 1987), calculating an optimal MSA is an $\mathcal{NP}$-hard problem (Akutsu, Arimura, and Shimozono 2000). A large number of fast and efficient heuristics have been developed to align genomic DNA sequences (Blanchette 2007), but the solutions returned by these algorithms are potentially suboptimal. It is thus common for biologists to manually tweak alignments produced by computer programs in order to increase their accuracy (Carrillo and Lipman 1988).

With the rapid expansion of genome sequencing technologies, the shear quantity of DNA sequences to be aligned (potentially hundreds of sequences of several billion characters each) makes the task of producing and maintaining highly accurate MSA intractable for small groups of experts. Because manual curation is a necessary step to guarantee the quality of biological sequence alignments, a crowdsourcing solution appears to be a perfect strategy to address this bottleneck.

Phylo aims to improve MSA solutions already precalculated by state-of-the-art algorithms. Eventually, in the case a MSA cannot be improved, it can also serve as a certificate to validate the input data. Phylo starts from a computationally calculated MSA of multiple vertebrate genomes (Rhead et al. 2009), identities portions of the alignment that are potentially sub-optimal, and transforms them into small puzzles that are dispatched to players (See Figure 1). Once a player has completed a puzzle, the solution found is returned to our server for evaluation. If the alignment found by the player is deemed superior to the original computer-produced alignment (based on a parsimony-based scoring scheme (Wang and Jiang 1994)), it replaces it in the global alignment. Phylo thus contributes to improving a resource (multiple genome alignment) that is used every day by re-

searchers in biology and genetics (e.g. detection of important DNA motifs associated with a biomolecular function, inference of ancestral genomes), while being a fun and educational game for non-experts.

In this paper, we analyze the data (i.e improvements of MSA generated by gamers, user statistics) collected through Phylo since November 2010. It constitutes a unique corpus of 7 years of data, which involved more than 400,000 unique users and generated more than 1 million of puzzle (MSA) solutions. Importantly, since 2014 Phylo also features an educational portal that enables us to compare the performance of casual users vs students who played Phylo as part of their curriculum.

Our study uses the data collected through Phylo to investigate multiple key aspects of a human-computing pipeline:

- (Task difficulty) We identify which features are most useful to predict if a computer-generated puzzle (i.e. alignment) can be improved (usefulness), and further how many solutions need to be collected to improve a MSA (difficulty).
- (Task Aggregation) We describe our pipeline filtering and assembling the solutions from users, and quantify the magnitude of the improvement resulting from this approach. We also estimate the practical efficiency of our system and show that significant improvements can be obtained by looking at the top 30% of the solutions collected within the first 100 days after the release of a puzzle.
- (User profiling) We analyze the individual performance of each player, and characterize the performance of casual vs assiduous players. In particular, we show that recurrent users are performing much better on difficult puzzles.
- (Prior knowledge) We compare the accuracy of alignments of casual gamers (playing the game without prior introduction to the biological motivation of the puzzle) vs students who played Phylo in an educational setting. Interestingly, our results suggests that both populations are performing equally well.

In addition to this study, we also use our system to improve the reference alignment of Ebola virus sequences available on the UCSC Genome browser (Kent et al. 2002). We show that in a relatively short period of time our platform has been able to mobilize citizens for curating scientific data. Our results are publicly available at `http://csb.cs.mcgill.ca/phylo`.

In summary, this paper addresses several key questions such as, deciding the accuracy with which we can decide the difficulty of puzzles and most importantly, it defines the average time, average number of expected attempts, and accuracy with which we can achieve a good alignment using human computers. In addition, we also check whether there is an impact of players playing more number of times in their aligned result, also if players have prior knowledge of biology in context with the game, will they perform better? It also evaluates the aggregated result obtained by placing the human aligned puzzle back into the broader MSA they were extracted from. All these scenarios have been addressed in this paper using analytics performed on the data generated by human computers. The rest of the paper engages in resolving such challenges. Section 2 describes the methods adopted for puzzle extraction, routing, alignment, aggregation, and user analysis. Then, we present our results in Section 3, and discussed them in Section 4.

2 Methods

2.1 Data sets

We analyzed data collected over 5 years by Phylo. The complete data set consists of 1907 alignments puzzles extracted from 575 genomic regions (alignment blocks). Each puzzle consists of a set of 2 to 8 DNA sequences from vertebrate species, including human, of length 10 to 21. The puzzles were played on by different players, for a total of 465027 puzzle solutions, i.e. 243 solutions per puzzle on average.

2.2 Scoring alignments

Each puzzle and alignment block was aligned computationally using several tools (Multiz (Blanchette et al. 2004), T-coffee (Notredame, Higgins, and Heringa 2000), MAFFT (Katoh et al. 2002)). The highest scoring of the machine-computed alignments is called the machine-computed alignment. This score is used as a "par" score that players are challenged to beat. For each puzzle, each solution submitted by a player is evaluated and the highest scoring solution is retained. Its score is called the human-computed alignment score.

Difference between an alignments' best score and its original score is referred as the scope of improvement. We use this difference to obtain the normalized scores for each puzzle or alignment. Puzzles with the least possibility of improvement were considered as difficult puzzles, so if the average normalized scores of puzzles were less than or equal to 0.33 and greater than 0, then they were considered as difficult to align. Similarly, a normalized score value between 0.33 and 0.66 was considered as medium difficulty, whereas, a value greater than 0.66 represents easy puzzles.

Robinson-Foulds distance: To further compare the accuracy of machine and human computed alignments, we evaluated the extent to which phylogenetic trees inferred from these alignments are consistent with the known phylogeny of the species involved. For each puzzle, the neighbor-joining algorithm (Saitou and Nei 1987) was used to produce a tree, and this tree was compared to the correct phylogenetic tree available through the UCSC Genome Browser (Karolchik et al. 2003). The tree comparison was carried out using the Robinson-Foulds (RF) tree-to-tree distance (Robinson and Foulds 1981; Waterman and Smith 1978) (implemented in (Huerta-Cepas, Serra, and Bork 2016)), which is defined as the number of subtree prune-and-regraft operations needed to transform one tree into the other.

Entropy: The entropy of an alignment column is a classical measure of conservation (Lin 1991; Valdar 2002a; Henikoff and Henikoff 1994), with perfectly conserved columns having an entropy of zero whereas columns where many different characters are represented in roughly equal proportions have high entropy. If we define P_a as the proportion of character $a \in \{A, C, G, T, -\}$ in a given align-

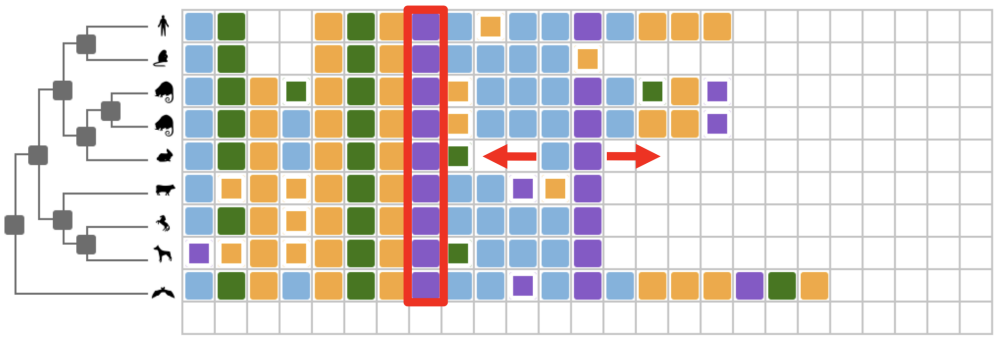

Figure 1: Interface of Phylo (2013 release). Each row is composed of a sequence of bricks of 4 different colors representing the 4 nucleotides A, C, G, and T. These sequences have been extracted from the DNA of different species represented with a icon on the left of the grid. Participants move the tiles left or right in order to maximize the number of color matches in each column. Although, the order of the bricks cannot be changed. Color mismatches and gaps are thus unavoidable and bring penalties. The phylogenetic tree on the left indicates the priority in which the rows should be aligned.

ment column, then the entropy of that column is defined as $\Sigma_{a\in\{A,C,G,T,-\}} P_a \log(P_a)$. This operation is performed for individual columns and later an average of all gives the comprehensive entropy value of the puzzle.

2.3 Prediction of puzzle difficulty levels

In order to train a machine learning predictor to recognize difficult puzzles, these puzzles need to be represented using a vector of features. We extracted and evaluated the following 11 features calculated from the machine-computed MSA: (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, i.e. entropy of the frequency distribution of the four nucleotides and gaps, (8) average length of sequences, (9) number of sequences, (10) tree-entropy based on depth of the leaves of phylogenetic tree of S, calculated by creating a vector of depth of all the leaf nodes in a tree and then calculating its entropy; we used tree-entropy as feature to account for phylogenetic tree information, (11) score of machine-computed alignment.

For each puzzle, we also compared the score of the best human-computed alignment to the score of the machine-computed alignment, and defined the score gain as the difference between the two. Positive score gains correspond to alignments that were better aligned by (some) humans than by algorithms. We further subdivided puzzles based on the value of score gains to obtain two equal size classes. The puzzles with a score gain greater or equal to 17 were assigned to the positive class (i.e. the class of puzzles where humans have produced significantly improved alignments). The rest were assigned to the negative class (little or no improvement).

We trained and benchmarked multiple binary classification algorithms: logistic regression, neural network, extra tree classifier, random forest classifier, Ada boost classifier, gradient boost classifier and decision tree classifier (all implemented in scikit-learn). We partitioned the dataset into 60% training set to learn model parameters and 40% testing set to evaluate the learned models. The hyper parameters were selected using 10-fold cross validation on the training data. Area Under the Curve of the Receiver Operating Curve (AUC ROC) on testing set was used to measure models accuracy.

In many cases, it can be useful to go beyond a binary classification problem and instead predict the expected value of the alignment score gain that can be expected for a given alignment. This can for example be useful to properly assign puzzles to users with the right level. We experimented with different machine learning regression models to attempt to predict the score gain from the set of 11 features, and eventually selected neural networks because of their superior performance. We use 10-fold cross-validation on the training data for hyper parameter search and R^2 measure on the testing set for accuracy measurement.

2.4 Aggregation

When multiple puzzles are extracted from the same alignment region, their solutions need to be reinserted in the alignment in order to properly evaluated the improvement obtained. This is a process we call aggregation. We selected alignment solutions from puzzles with the best score and least entropy. Then the selected alignments were reinserted and compared with the entropy of revised alignment block to the entropy of the original alignment block.

2.5 User efficiency

The efficiency of human-computing systems depends of the expertise of participants. Characterizing the precision of answers from their level of expertise and background knowledge is thus essential to understand the capacity and behavior of the system. In fine, this knowledge can also help us to improve the effectiveness through better routing of the tasks, but also to enhance the satisfaction of the users, hence sustainability of the system.

In this paper, we analyze taxonomies that can distinguish educational players versus game players, as well as rookies

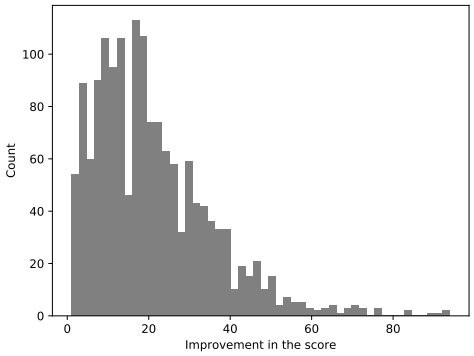

Figure 2: Histogram plot of the improvement in the score for 1556 puzzles, where improvement is measured as the difference between the best score produced by a human player and the initial computer-based alignment.

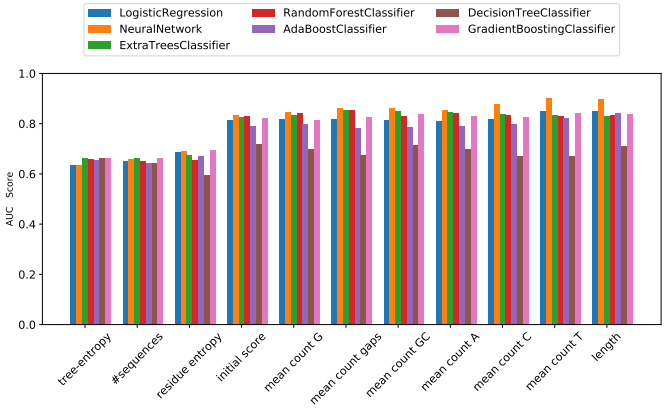

Figure 3: Measuring effects of feature combination on models performances. The features from the ordered list (based on individual feature importance in decreasing order) are added incrementally to train models.

versus expert players. For each solution, we quantify the improvement as the difference between the score obtained and the initial score of the computer alignment. Then, we define the efficiency of a player as the average of the normalized alignment scores of the solutions returned. The results are distributed in 3 categories representing the difficulty of the puzzles (i.e. easy, medium, or hard).

3 Results

We analyzed the data collected by Phylo over the last 5 years in order to deduce lessons that can help us to design more efficient human computing systems in the future. This data set consists of more than 465,000 solutions for nearly 2000 puzzles collected from November 2012 to December 2016.

3.1 Prediction of puzzle difficulty

A key step in Phylo is to assess the difficulty of a puzzle in order to route it to players with the appropriate skill level. Ideally, the difficulty should be estimated automatically, before any player has tried the puzzle. Here, we study how a difficulty level (assessed retrospectively based on the ability of players to improve the alignment score) can be predicted by machine learning algorithms.

Figure 2 shows the distribution of the improvement of alignment scores (as calculated in the game) produced by human players over the machine-computed alignments. Although in many cases improvements are modest, the tail of this distribution highlights a significant number of puzzles with very large improvements. It shows that puzzles are of unequal difficulty – a phenomenon we study further in this section.

We labeled the puzzles as easy or hard based on the player's success rate (See Section 2.3). Then, we trained different types of machine learning classifiers to predict a puzzle's label based on a variety of features (see Methods). Figure 4 shows the classification accuracy. For each predictor, we calculated the Area Under the ROC Curve (AUC), based on 10-fold cross-validation, using a single feature at

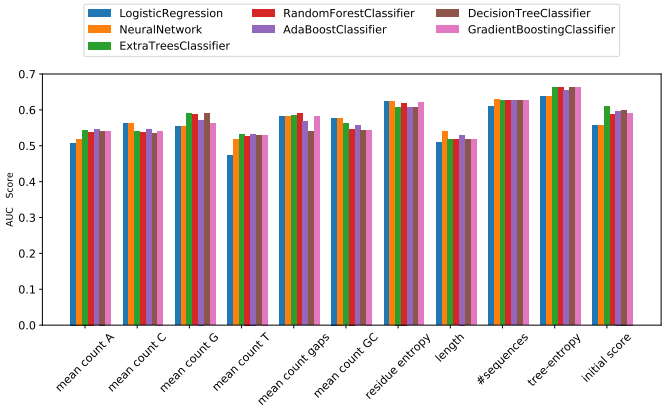

Figure 4: Individual feature performance is measured through AUC values on testing set. Each model is trained using one feature at a time.

a time. As anticipated, we observe that the most informative features are those capturing the number of sequences to be aligned, but also their dissimilarity (tree-entropy and mean entropy of column-wise residues). Then, we trained and evaluated multi-feature models by incrementally adding features in decreasing order of their value (Figure 3). Remarkably, it is the inclusion of the feature corresponding to the score of the initial computer-calculated alignment that provides the largest improvement in prediction accuracy. This is not unexpected, as alignments whose initial scores are already high are difficult to improve.

We then considered the regression version of the problem, where the goal is to predict the quantitative improvement in score one can expect for a given puzzle. The best predictions, with R^2 value of 72% were obtained using a neural network regression model whose hyper parameter values with L2-norm regularization regularization (weight of 10^{-3}) and 300 nodes in each of the two hidden layers. This signifi-

Difficulty (Number of Puzzles)	N_u	N_w	N_{opt}	Gain of Efficiency
Easy (14)	4123	2100	1752	115.4%
Medium (80)	19653	13600	12501	48.4%
Difficult (466)	121531	89472	63205	50.7%

Table 1: Estimate of the total number of solutions that needs to be collected to obtain the highest score in the uniform (N_u), weighted (N_w) and optimal (N_{opt}) routing schemes. Puzzles are sorted in 3 categories (easy, medium, and difficult) representing the observed difficulty for players to achieve the highest score.

cantly outperformed other predictors such as a ridge regression model with L2-norm regularization, which obtained an R^2 value of only 56%. This is because none of the predictor variables used in the regression model are highly correlated.

The last and the most important step is to analyze the efficiency of the proposed model with respect to the current scheme that uniformly distributes the puzzles based on the number of solutions collected. The preliminary step for this comparison is to obtain the number of times a puzzle in each category should be sent to get good alignments. To obtain an accurate estimate of this, we identified the top 25% of the puzzles in each category which took maximum attempts to reach to its highest score and calculated their average. Top 25% were taken to ensure significant number of spare attempts for achieving good alignments. We obtained these values (n_0) as 150, 170 and 192 for easy, medium and difficult puzzles respectively.

Next, we calculate the efficiency as $\frac{(N_u - N_w)}{N_{opt}}$, where N_u is the total number of puzzles sent in the uniform model (all puzzles have the same weight), N_w is the total number of puzzles sent using labels produced by our classifier, and N_{opt} is the minimum number of puzzles that has to be sent to obtain all high scores. N_w is calculated as the product of the number of puzzles in a particular category, and the expected number of solutions we need to collect to obtain the highest score (n_0). Table 1 shows the gain in efficiency obtained for each category of puzzles (i.e. easy, medium, or hard). The data sets are described in Section 2.3.

Lesson 1: Our results suggests that the difficulty of puzzles can be relatively well predicted. These predictions can be used to significantly improve the routing of puzzles (i.e. tasks), and thus the number of task to complete to obtain the best answer. The gain in efficiency is more pronounced on easy puzzles, which suggests that this results is primarily obtained by avoiding unnecessary calculations on easiest tasks.

3.2 Evaluation of the accuracy of solutions

In Phylo, the puzzles played by the participants are extracted from a larger genome alignment. Before addressing the analysis of the performance of the aggregation of all solutions, we want to assess the quality of the improvements obtained by humans regardless of the origin of the alignments.

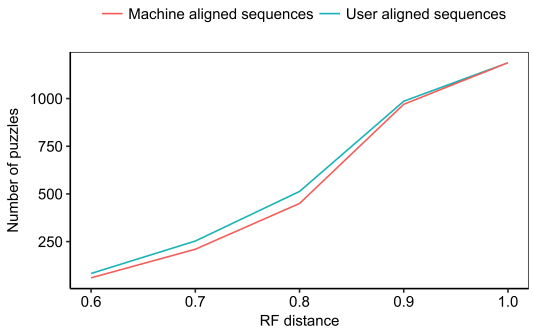

Figure 5: Cumulative distributions of the Robinson-Foulds distances between the true phylogenetic tree and trees inferred from either the machine-computed or human-computed alignments.

Alignments with known phylogenetic tree Although different alignments can be compared on the basis of various alignment scoring schemes, their true accuracy cannot be directly assessed because the biologically correct alignments are not known. Here, we use an indirect, biologically motivated approach in which phylogenetic trees are inferred for each candidate alignment using neighbour-joining algorithm and compared to the correct phylogenetic tree (Kent et al. 2002) for the set of species considered. This method enable us to identify alignments improving their fitness to the only reliable data we have about them.

Figure 5 shows the cumulative distribution of the Robinson-Foulds distance between the inferred and true trees. We observe a slightly larger proportion of small RF distances among trees derived from human-computed alignment, suggesting that those alignments may be more accurate the machine-computed alignments. However, this difference is not statistically significant (based on a Kolmogorov-Smirnov test), in part because several puzzles had machine-computed and human-computed solutions that were identical. 6.3% of the alignments provide phylogenetic tree closer to the known phylogenetic than those aligned via machine. This number could seem relatively low, but determining a phylogeny is a difficult problem and our results show that significant improvements can be obtained using human alignments.

Alignments with unknown phylogenetic tree When the true phylogeny is unknown, the entropy emerges as a reasonable criterion for estimating the information content at each position (i.e. column) of the alignment. First, we compared the sum of column entropy and gaps (Valdar 2002b) of the alignments produced by human players to those calculated by computers. Remarkably, 57% of the user aligned puzzles were either better or at par with the machine aligned sequences, and in many cases by a large margin. This suggests again that although users are not always able to improve alignments, they do so in a non-negligible fraction of the puzzles.

Lesson 2: There is no gold standard score to estimate the quality of an alignment. Phylo uses a simplified consensus scheme to guide the participants. Humans appear to generate improvements that increase the likelihood of their alignments based on multiple external (biologically motivated) criteria that were not presented to them. This finding suggests that we can capture the intuition that humans developed to identify good alignments.

3.3 Aggregated alignment score improvement

Puzzles are small alignments that have been extracted from a (much) larger alignment we are trying to improve. Therefore, many puzzles originate from the same genomic region (alignment block). The aggregation mechanism aims to find the best collection of non-overlapping puzzle solutions that can be re-inserted to improve the original alignment block. Once again, since there is no gold standard metric to compute the quality of an alignment, we estimate the performance of the aggregation process using indirect measures.

Aggregation with known phylogenetic tree The observed alignments with smaller RF distances among trees derived from human-computed alignment were inserted back into the original blocks. Then an analysis of the change in RF distance which was calculated as the difference between the RF distance of machine aligned blocks and aggregated user aligned blocks. Using this metric, we achieved an enhancement of 36.7% of these blocks. The remaining blocks could not be improved by the players, either because the original alignment was already optimal, or because the enhancement was not good enough to show a correct prediction of phylogenesis.

Aggregation with unknown phylogenetic tree In this case, the best scoring alignments calculated by humans with least entropy were inserted back into the original blocks. Then, we performed an analysis of the change in entropy value, which was calculated as the difference between the entropy of machine aligned blocks and aggregated user aligned blocks. We achieved an enhancement of 66.3% of these blocks, with eventually some dramatic improvements (Figure 6). The remaining blocks could not be improved by the players, either because the original alignment was already optimal, or because the problem was too hard for players to obtain a good alignment.

Lesson 3: The aggregation of solutions selected through a basic (simplified) scoring scheme available to users, is sufficient to yield significant improvements of a global problem (i.e. alignment).

3.4 User Analysis

We analyze the performance of the various types of users. Since its launch in 2010, Phylo has now 35,913 registered users (Note: participants can also play anonymously without registration), for which we recorded the number and difficulty of puzzles played. Moreover, since 2014 Phylo also features an educational interface (`http://phylo.cs.mcgill.ca/submit/`) that enables instructors to register students and track their performances. This information allows us to study the impact of a prior knowledge about the multiple sequence alignment problem on the quality of solutions submitted.

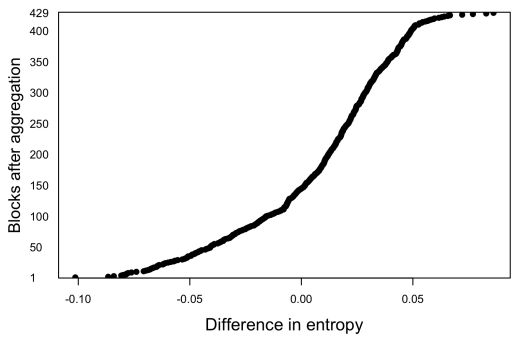

Figure 6: Change of entropy after aggregation of user aligned puzzles to the initial alignment. The x-axis represents the difference between the entropy of the machine aligned sequences and user aligned sequences. Positive values indicate an improvement of the alignment.

Rookies vs experts First, we aim to determine if recurring users gain experience and generate better alignments than casual gamers. Figure 7 compare the performance of rookies ($\leq$ 20 puzzles played) and experts ($\geq$ 40 puzzles completed) on puzzles with various difficulties (easy, medium, and hard). Interestingly, we note that rookie players perform similarly to experts on easy puzzles. However, when the difficulty of the puzzles raises to medium or hard, expert players clearly outperforms their less experienced counterparts. A F-test suggests a P-value of 0.483, 0.94 and 0.018 for easy, medium and hard puzzles respectively between expert and rookie users. It confirms that entry levels are sufficiently accessible to allow participants to learn the rules, and that our human-computing system can benefit of recurring users.

Background knowledge Two types of players have been using Phylo in the past few years: (i) casual gamers, who play just for fun, and (ii) educational players, who play in the context of a high-school or university biology/bioinformatics course through our educational interface. A total of 170 puzzles were played by both types of users, which we used to assess the performance of each group.

We observe a P-value of 0.47 while performing T-test on the dataset of educational players and casual players. It demonstrates that the benefit of a background knowledge toward the production of better alignment is difficult to assess. In figure 8, we show the performanc achieved by educational players playing for education against casual on puzzles with various difficulties. Although educational users seem to initially benefit of their background knowledge, the difference disappears when the difficulty increases.

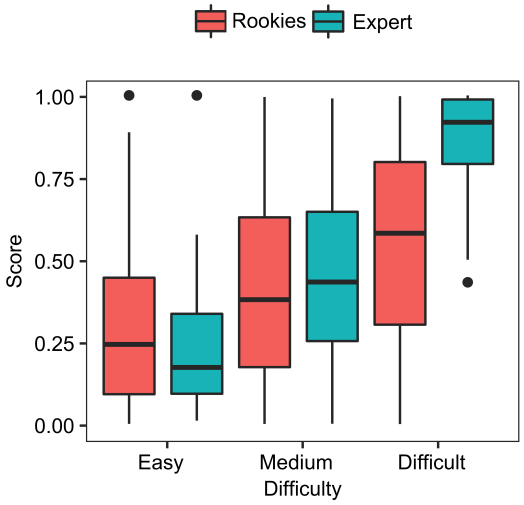

Figure 7: Average performance of rookies ($\leq$ 20 puzzles) and expert players ($\geq$ 40 puzzles) on easy, medium, and hard puzzles. The black dots above the box represent outliers with more than 3/2 times of the upper quartile. The black dots below represent outliers with less than 3/2 times of the lower quartile.

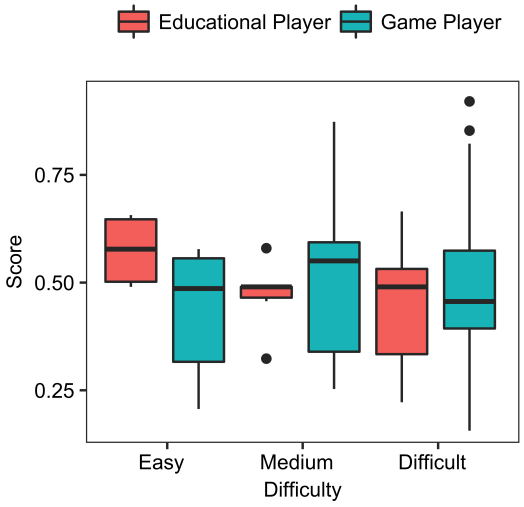

Figure 8: Average performance of casual (no prior training) and educational users (benefiting of background knowledge) on easy, medium, and hard puzzles. The black dots above the box represent outliers more than 3/2 times of the upper quartile while the ones below represent outliers less than 3/2 times of the lower quartile.

Difficulty	Number of Puzzles	Number of Solutions	Rank of highscore μ	Rank of highscore Q_3
Easy	25	2075	83	150
Medium	255	21678	85	170
Difficult	1117	103635	93	192

Table 2: Number of puzzles collected and rank of highest scoring alignments.

Lesson 4: The level of expertise accumulated by the users through their long-term engagement results in enhanced performance in the most difficulty tasks (i.e. puzzles). It follows that recurring users are a precious resource for GWAPs. By contrast, a prior introduction to the principles of the game (i.e. here a multiple sequence alignment) does not seem to provide a long-term advantage. This confirms the importance of design techniques in GWAPs.

3.5 Latency and robustness of solutions

One important aspect to guarantee the efficiency of human-computing systems at solving an optimization problem, is to make the best possible estimate of the number of solutions that needs to be collected to offer some confidence in the quality of the solution returned by the system. In particular, we need to identify the parameters that influence these estimates.

Figure 9 shows that with an average of 320 puzzles solved per day (average over the last 6 years) most of the improvements in alignment scores are obtained relatively quickly, within the first $\approx$ 100 days. Past this age, the accumulation of solutions appears to be redundant and the puzzle can be retired.

Then we investigate the impact of the difficulty of puzzles on these statistics. Table 2 shows the number of attempts taken to get to the highest score. In other words, the number of solutions we need to collect to get one that is not further improved by players. On an average, the ranks of the highest scoring alignment are comparable, albeit a slight (expected) increase of the rank is observed on the most difficult puzzles. The length of the period during which we collected solutions enables us to offer some guarantees on the robustness of these estimates. In average, the highest scores got reached after collecting only $\approx$ 30% of the total number of solutions.

These statistics allow us to estimate the volume of data that can be treated by our system per. With an average of 320 solutions collected per day, we can currently envision to solve 2-4 puzzles per day depending of their complexity and the target level of confidence in the results. Thus, a typical genomic bloc within a week. Of course, regular or occasional growth of the traffic will increase these numbers (Sauermann and Franzoni 2015).

Lesson 5: The efficiency of a human-computing system relies on its capacity to quickly obtain a robust solution. This data enables us to estimate the practical efficiency of the system.

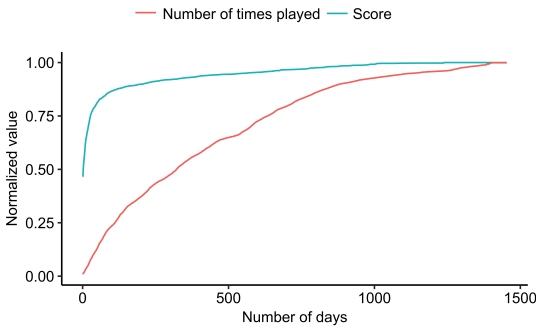

Figure 9: Progression of high scores. The orange line shows the normalized high scores since the release of the puzzles. The blue curve plots the number of solutions collected.

3.6 Case Study: Ebola virus

In 2015, the Ebola epidemics hit West Africa. The genomes of several strains of the virus were sequenced and aligned (WHO Ebola Response Team 2014) in order to study the virus' function and evolution. At that time, the Phylo player community was enlisted to help improve the accuracy of these alignments. Improved alignment scores (measured based on decrease in entropy of the revised alignments) were obtained for 144 of the 411 alignment blocks considered. The best scored alignments were selected and replaced in the original block to obtain the enhanced version of DNA alignments for Ebola virus. Consistent with observations made on larger sets of puzzles, optimal solutions were generally found relatively quickly (within 95 days) and longer periods of puzzle availability did not result in significant alignment score gains. Improved alignments are publicly available at `http://csb.cs.mcgill.ca/phylo/`.

Lesson 6: Human-computing systems can be used to quickly mobilize citizen scientists and produce valuable scientific data in case of emergency crisis.

4 Discussion

In this paper, we analyzed the data collected during the last 6 years (2010-2016) through Phylo, a GWAP and human-computing system designed for solving a fundamental problem in bioinformatics: The multiple sequence alignment problem (Blanchette 2007; Edgar and Batzoglou 2006).

In practice, multiple sequence alignments calculated by state-of-the-art alignment programs are often manually curated. The increasing amount of data generated by novel DNA sequencing technologies reinforces the relevance of a human-computing approach, and suggests the broad impact that a sustainable human-computing systems may have in biology (Good and Su 2013).

An asset of this study resides in the volume and broad spectrum of available data. Although early reports demonstrated the capacity of our system to improve alignments (Kwak et al. 2013; Kawrykow et al. 2012), until now the nature and magnitude of the factors influencing the performance of our system were not characterized.

This study focuses on Phylo, but many of the observations and conclusions we made here can be transposed to other GWAPs and scientific games. They will also help us as to design a novel generation of crowd-computing systems for genomics. The principal findings of this work have been summarized in 6 lessons.

First, an accurate estimate of the complexity of a task can lead to better routing strategies and thus more performant human-computing systems (Yang et al. 2016). Lesson 1 revealed that the difficulty of our puzzles can be accurately predicted and used to significantly reduce the amount of work needed to complete our tasks. Symmetrically, at the other end of the pipeline, the aggregation stage can benefit of a precise estimate of the confidence in the solutions returned by the users (Cheng, Teevan, and Bernstein 2015). Lesson 5 enabled us to validate this concept, and use it to increase the computing capacity of our system. To some extent, our data can be compared to user statistics collected with another similar scientific game (Rallapalli et al. 2015).

Another intriguing aspect of our framework resides in the impact of the decomposition of the initial problem into simple tasks on the performance of the reconstruction (Law and Ahn 2011). Lessons 2 and 3 enabled us to validate that a simplified presentation of the alignment problem is able to capture the wisdom of the crowd, and global intuition.

Characterizing the performance of participants based on their experience and background is also key to improve the efficiency and sustainability of the system (Law and Ahn 2011). In addition to enhance the scientific productivity, since 2013 Phylo also contributes to educate the public about genomics through its educational interface. The analysis of these statistics enables us to investigate the benefits of prior knowledge of the tasks. Overall, our observations (Lesson 4) follow previous conclusions made on other platforms suggesting that (i) recurrent users outperforms rookies on complex tasks only (Papoutsaki et al. 2015), and (ii) a prior training does not provide an significant long-term advantage (Andersen et al. 2012; Horowitz et al. 2016). However, recent studies suggest that customized training mechanisms could take advantage of this situation (Lee et al. 2016).

Finally, we confirmed that human computers can actually enhance the alignment accuracy of both vertebrate and Ebola virus genomes. Our case study on Ebola virus is also a proof of concept that citizens can be quickly mobilized to conduct scientific data curation in case emergency crisis (Lesson 6).

Eventually, the data collected through Phylo could be used to help us designing novel alignment heuristics. Such approach has already been successfully applied with Foldit (Khatib et al. 2011). However, this enterprise would require to collect more data on the progression of players while solving a puzzle.

References

Akutsu, T.; Arimura, H.; and Shimozono, S. 2000. On approximation algorithms for local multiple alignment. In *Proceedings of the fourth annual international conference on Computational molecular biology*, 1–7. ACM.

Andersen, E.; O'Rourke, E.; Liu, Y.-E.; Snider, R.; Lowdermilk, J.; Truong, D.; Cooper, S.; and Popovic, Z. 2012. The impact of tutorials on games of varying complexity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 59–68. ACM.

Blanchette, M.; Kent, W. J.; Riemer, C.; Elnitski, L.; Smit, A. F.; Roskin, K. M.; Baertsch, R.; Rosenbloom, K.; Clawson, H.; Green, E. D.; et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research* 14(4):708–715.

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

Carrillo, H., and Lipman, D. 1988. The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics* 48(5):1073–1082.

Cheng, J.; Teevan, J.; and Bernstein, M. S. 2015. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1365–1374. ACM.

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.

Edgar, R. C., and Batzoglou, S. 2006. Multiple sequence alignment. *Current opinion in structural biology* 16(3):368–373.

Feng, D.-F., and Doolittle, R. F. 1987. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution* 25(4):351–360.

Good, B. M., and Su, A. I. 2013. Crowdsourcing for bioinformatics. *Bioinformatics* 29(16):1925–33.

Henikoff, S., and Henikoff, J. G. 1994. Position-based sequence weights. *Journal of molecular biology* 243(4):574–578.

Horowitz, S.; Koepnick, B.; Martin, R.; Tymieniecki, A.; Winburn, A. A.; Cooper, S.; Flatten, J.; Rogawski, D. S.; Koropatkin, N. M.; Hailu, T. T.; Jain, N.; Koldewey, P.; Ahlstrom, L. S.; Chapman, M. R.; Sikkema, A. P.; Skiba, M. A.; Maloney, F. P.; Beinlich, F. R. M.; Foldit Players; University of Michigan students; Popović, Z.; Baker, D.; Khatib, F.; and Bardwell, J. C. A. 2016. Determining crystal structures through crowdsourcing and coursework. *Nat Commun* 7:12549.

Huerta-Cepas, J.; Serra, F.; and Bork, P. 2016. Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution* 33(6):1635–1638.

Karolchik, D.; Baertsch, R.; Diekhans, M.; Furey, T. S.; Hinrichs, A.; Lu, Y.; Roskin, K. M.; Schwartz, M.; Sugnet, C. W.; Thomas, D. J.; et al. 2003. The ucsc genome browser database. *Nucleic acids research* 31(1):51–54.

Katoh, K.; Misawa, K.; Kuma, K.-i.; and Miyata, T. 2002. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research* 30(14):3059–3066.

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 research* 12(6):996–1006.

Khatib, F.; Cooper, S.; Tyka, M. D.; Xu, K.; Makedon, I.; Popovic, Z.; Baker, D.; and Players, F. 2011. Algorithm discovery by protein folding game players. *Proc Natl Acad Sci U S A* 108(47):18949–53.

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.

Law, E., and Ahn, L. v. 2011. Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5(3):1–121.

Lee, D. J.-L.; Lo, J.; Kim, M.; and Paulos, E. 2016. Crowdclass: Designing classification-based citizen science learning modules. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

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*.

Lin, J. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory* 37(1):145–151.

Notredame, C.; Higgins, D. G.; and Heringa, J. 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* 302(1):205–217.

Papoutsaki, A.; Guo, H.; Metaxa-Kakavouli, D.; Gramazio, C.; Rasley, J.; Xie, W.; Wang, G.; and Huang, J. 2015. Crowdsourcing from scratch: A pragmatic experiment in data collection by novice requesters. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Rallapalli, G.; Fraxinus Players; Saunders, D. G.; Yoshida, K.; Edwards, A.; Lugo, C. A.; Collin, S.; Clavijo, B.; Corpas, M.; Swarbreck, D.; Clark, M.; Downie, J. A.; Kamoun, S.; Team Cooper; and MacLean, D. 2015. Lessons from fraxinus, a crowd-sourced citizen science game in genomics. *Elife* 4:e07460.

Rhead, B.; Karolchik, D.; Kuhn, R. M.; Hinrichs, A. S.; Zweig, A. S.; Fujita, P. A.; Diekhans, M.; Smith, K. E.; Rosenbloom, K. R.; Raney, B. J.; et al. 2009. The ucsc

genome browser database: update 2010. *Nucleic acids research* gkp939.

Robinson, D. F., and Foulds, L. R. 1981. Comparison of phylogenetic trees. *Mathematical biosciences* 53(1-2):131–147.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 4(4):406–425.

Sauermann, H., and Franzoni, C. 2015. Crowd science user contribution patterns and their implications. *Proc Natl Acad Sci U S A* 112(3):679–84.

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.

Valdar, W. S. 2002a. Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics* 48(2):227–241.

Valdar, W. S. 2002b. Scoring residue conservation. *Proteins: Structure, Function, and Bioinformatics* 48(2):227–241.

Wang, L., and Jiang, T. 1994. On the complexity of multiple sequence alignment. *Journal of computational biology* 1(4):337–348.

Waterman, M. S., and Smith, T. F. 1978. On the similarity of dendrograms. *Journal of Theoretical Biology* 73(4):789–800.

WHO Ebola Response Team. 2014. Ebola virus disease in west africa – the first 9 months of the epidemic and forward projections. *N Engl J Med* 2014(371):1481–1495.

Yang, J.; Redi, J.; Demartini, G.; and Bozzon, A. 2016. Modeling task complexity in crowdsourcing. In *Proceedings of The Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016)*, 249–258. AAAI.