

Serious Games for Solving Protein Sequence Alignments - Combining Citizen Science and Gaming

Martin Hess¹, Josef Wiemeyer³, Kay Hamacher², and Michael Goesele¹

¹ Department of Computer Science - Technische Universität Darmstadt

² Department of Biology - Technische Universität Darmstadt

³ Department of Human Science - Technische Universität Darmstadt

Springer LNCS 8395. This is the author version of the publication. The final publication is available at <http://link.springer.com/book/10.1007%2F978-3-319-05972-3>

Abstract. A fundamental task in computational biology is the identification of similarities between multiple protein sequences, to get insight into their functional, structural and evolutionary relationships. These similarities can be revealed by aligning the sequences. This alignment is a NP-hard problem.

In this paper, we present an improved game approach to solving large protein sequence alignments with the concept of citizen science. We abstract the alignment problem into a kind of puzzle, that can be solved even by non-expert players. Due to the early stage of development our approach is primarily focused on usability and game experience. We evaluated our game prototype by a user study with a convenience sample of 20 persons. This study showed, that our prototype is already capable to elicit fun and to deliver a true game experience despite its early stage. Some issues have been revealed requiring further improvement.

Keywords: Serious Games, Bioinformatics, Multiple Sequence Alignments, Protein Sequences, Citizen Science, Game Experience, Game Evaluation

1 Motivation

One of the most fundamental tasks in computational biology is the identification of similarities between a set of DNA, RNA or protein sequences. These similarities provide insight into functional, structural or evolutionary relationships between the sequences, which forms the basis for many important biological applications like evolutionary heritage [1].

The similarities are usually revealed by aligning identical or similar parts of the sequences and thereby adding gaps into the sequences. The quality of such an alignment is determined by an evolutionary model, which „rewards“ matching regions in the sequences and penalizes mismatching areas and gaps.

To get a reliable similarity analysis of the sequences for a given evolutionary model, it is necessary to find a good alignment. However the alignment of multiple sequences (MSA) is computationally expensive and an NP-hard problem [2]. Therefore many MSA algorithms like ClustalW [3], MUSCLE [4], Probcons [5] and MAFFT [6] rely on heuristics to find a good alignment [7].

Due to this probabilistic approach, the quality of the resulting alignments is questionable. One approach to improve the quality is to refine the alignments by hand using the human pattern recognition capability. Humans can easily detect visual patterns, e.g. by applying Gestalt principles or other organization principles of visual perception [8].

One concept to use human capabilities in solving scientific problems or improving their results is called *citizen science* [9]. In this concept, research tasks are outsourced to a volunteer crowd of amateur or non-professional „scientists“. Due to the large number of potential volunteers, this concept of doing research can be very useful for computationally expensive problems. Examples for successful citizen science approaches are Galaxy Zoo [10], Fold it [11] and Phylo [12].

In this paper, we present an improved game approach to solving large protein sequence alignments (currently up to 50 sequences of variable length) by using the concept of citizen science. Inspired by Phylo [12], a computer game for solving DNA sequence alignments, we implemented a computer game prototype in order to enhance game experience by improved game mechanics. In our game prototype, the players have to solve puzzles of colored spheres, which are related to protein sequences alignments. The usability and the game experience of our prototype game was evaluated by a user study with a sample of 20 persons.

The main contributions and application benefits of the paper are:

1. Our approach is capable of dealing with a large number of protein sequences (i.e. up to 50) with variable length. The playing field is not restricted in its dimensions, which enables players to handle the sequences with a high degree of freedom. According to Self-Determination Theory [13], increased options for self-regulation enhance intrinsic motivation, well-being, and consequently game experience.
2. Our computer game prototype is designed to be visually appealing and to deliver a true game experience to the player. In addition, our game runs natively on many platforms like Windows, Mac, Linux and Android devices and can be played using mouse and keyboard as well as touch interfaces. These aspects allow to involve a huge amount of possible players, which ensures a high „human computing power“.
3. Inspired by Phylo, we abstract sequence alignments to puzzles of colored spheres, which enables non-experts to solve alignments without further information or experience.
4. Analogous to Phylo, we use a high score system to ensure that the game is challenging and is capable of refining alignment solutions to get the best results.

2 Related Work

Our work aims at combining the principles of citizen science and computer games. The main idea is to bundle the strengths of both approaches, i.e., exploiting the capabilities of many people and the fun of gaming.

A well-known example of a successful citizen science approach is *Galaxy Zoo* [10]. *Galaxy Zoo* aims at identifying and classifying galaxies in millions of pictures by using the human visual pattern recognition. The users determine the shapes of galaxies presented to them and classify certain features.

Fold it is another citizen science approach, that deals with protein folding problems [11]. Unlike *Galaxy Zoo*, *Fold it* utilizes multiplayer game concepts to attract more volunteer „scientists“. In *Fold it*, the players see the structure of a protein in a 3D rendered context. The players can manipulate the 3D structure to optimize the computed energy using direct manipulation tools and a collection of algorithms. In addition, *Fold it* features the creation of interaction macros [14]. These so called recipes are collected in a cookbook and are available to all *Fold it* players. Recent results show, that some solutions found by *Fold it* players using these concepts outperformed their computer-based counterparts [15,16].

Phylo is a citizen science approach for solving DNA multiple sequence alignments [12]. Analogous to *Fold it*, *Phylo* features a multiplayer game approach. In *Phylo* the players solve puzzles of colored squares. Each square corresponds to a nucleotide inside a DNA sequence, while the color represents its type. The players improve the alignments by dragging squares inside their row. The puzzle states are scored by an evolutionary model, that penalizes gaps and misaligned squares, whereas matching regions are rewarded. A puzzle is solved, if the player beats a given minimum score. Unlike *Fold it*, *Phylo* hides more of the scientific context from the player allowing even non-expert users to produce good results. This promising approach also has some limitations. *Phylo* is only capable to deal with DNA or RNA sequences and supports only 10 sequences at once. The size of the playing field is fixed, which constrains the players in their movements. They also have few help mechanics supporting them during the game such as a consensus view of the columns.

3 Approach

Our prototype facilitates the alignment of large sets of protein sequences. A single amino acid is represented by a colored sphere inside the puzzle (see Fig. 1). Each amino acid has its own unique color. As a first approach to distinguish the different amino acids, we selected 20 different colors from the whole color spectrum related to the Windows Phone 8 accent color palette (see Fig. 2). We note, that the selection of the colors is arbitrary and may change in the future for a better discrimination.

On the top left side of the screen, the player can see all relevant information about the current progress. The four icons on the left show the score rewards and penalties for the current level (see [1] in Fig. 1). Matching spheres in each column are rewarded, mismatching spheres and gaps are penalized. In the example level in Fig. 1, each matching sphere is rewarded by one point, mismatching spheres are penalized by one point, and gaps are penalized with 5 points and one additional point for each length unit but the first. This gives the player direct information about the current scoring scheme. We note, that for reasons of simplification, our prototype does not take into account any mutation probabilities of the amino acids, as they are, e.g, represented in the BLOSUM matrices [17]. Two spheres are therefore only considered as a match, if they have exactly the same color. We will change this scoring scheme in the future to represent evolutionary models used in real applications.

Next to the penalties, the high score and the player's current score are shown (see [2] in Fig. 1). A puzzle is solved, if the player has beaten the high score at least once by improving the alignment.

To improve the alignment, the player has to align identically colored spheres in each column. On the other side, the player should minimize the amount of differently colored spheres in the columns and the number of empty cells.

By clicking and holding the left mouse button, the player can drag a sphere within its row (see Fig. 3). On touch devices, the player can drag a sphere with the finger. The currently selected sphere is highlighted with a white ring for an easy identification (see [3] in Fig. 1). If the sphere collides with other spheres, these will be moved too. It is not possible to switch spheres within a row due to the constraints of sequence alignments. Each moving step enforces the recalculation of the current score. The player can use this feature to determine if the movement results in a score improvement or not.

To grant the players full control over their movements, we implemented an undo/redo functionality. By clicking the arrow buttons on the top right of the screen (see [4] in Fig. 1), the player can revert each movement step done so far, as well as redo a reverted movement. In addition the player can save and load different puzzle states via the game menu, which can be accessed by clicking the monitor button on the upper right of the screen (see [4] in Fig. 1).

As mentioned above, the size of the alignment can be very large. To support the player in handling large alignments, the player can zoom the alignment by rotating the mouse wheel as well as scroll the current view by clicking and holding the right mouse button. On touch devices, the zoom feature is bound to a two finger multi touch gesture. If the distance between the two touches is reduced, the view will be zoomed out, otherwise the view will be zoomed in. For scrolling the view, the player can make a wipe gesture on the screen edges.

To support easy learning of the game mechanics and controls, we implemented a short stepwise tutorial. In each step, the player learns a new aspect of the game mechanics and can directly try out the learned content inside the tutorial level.

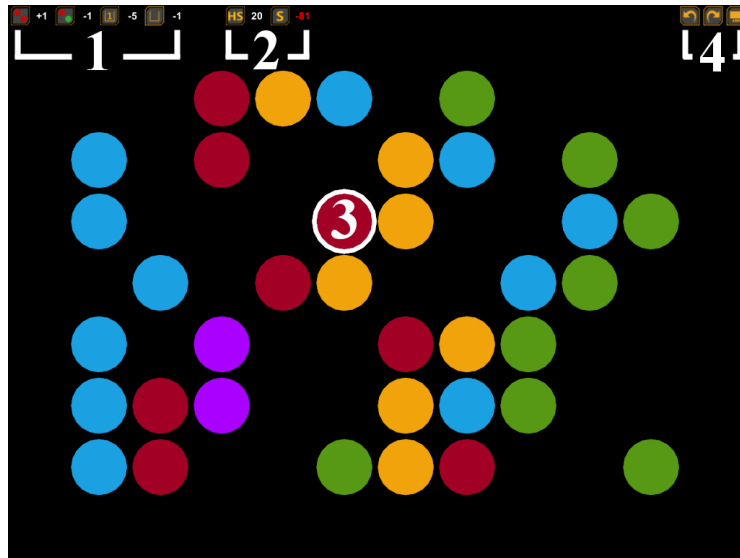


Fig. 1. Game prototype: Ingame view of the tutorial level. Each sphere corresponds to a single amino acid. The color indicates the acid type. [1] shows the scoring scheme, [2] the high score and current score, [3] a selected sphere, and [4] the undo/redo and menu buttons

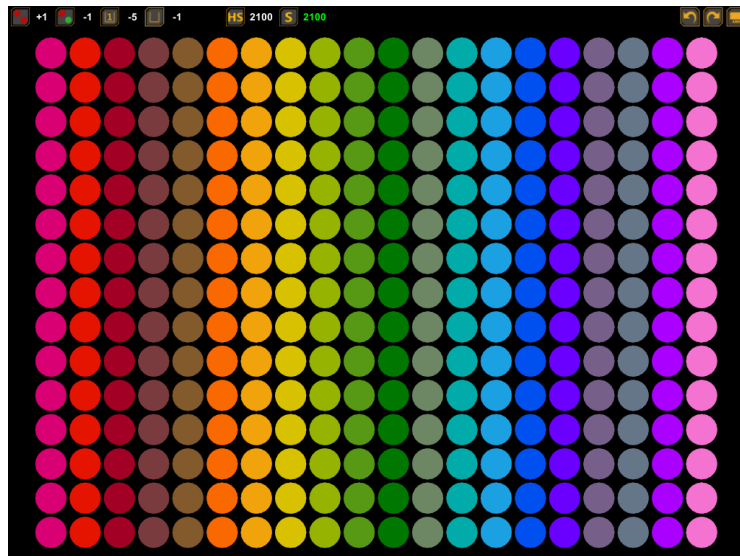


Fig. 2. Color scheme for the 20 different amino acids represented as a game level.

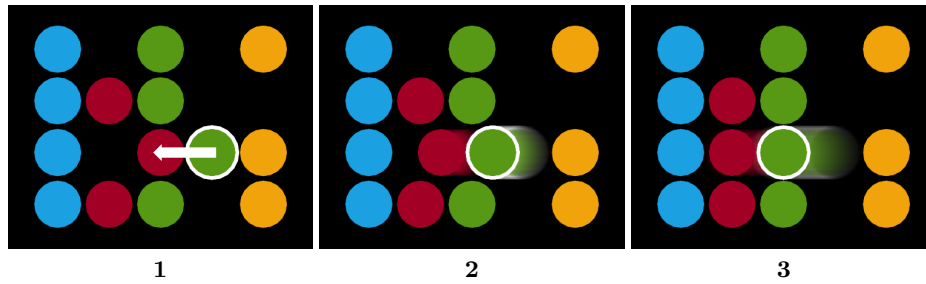


Fig. 3. Aligning by drag and drop. The green sphere with the white outline is dragged one position to the left. It collides with the neighbored red sphere which is also moved one position left.

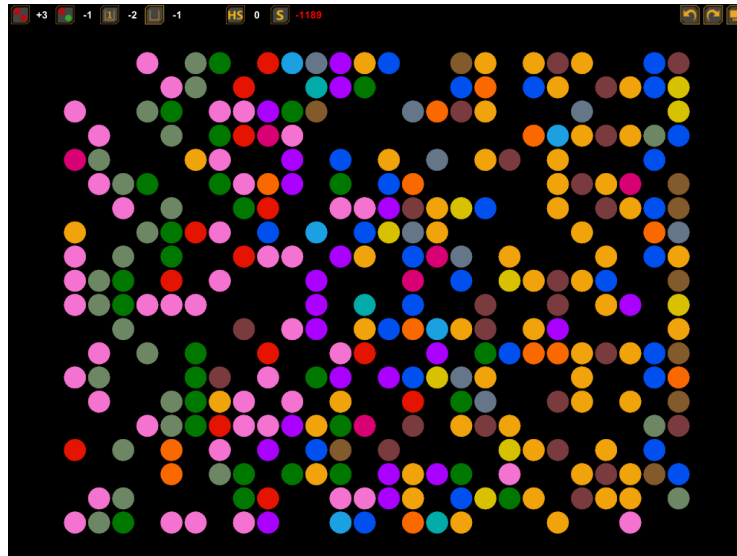


Fig. 4. One of the larger levels played in the user study.

4 Evaluation

4.1 Methods

We evaluated our game prototype in a user study with a convenience sample of 20 persons (6 women and 14 men; age: 22 to 34 yrs). The participants were colleagues and students from the biology domain ($n=11$), from the computer science domain ($n=8$), and from the educational science domain ($n=1$). Five participants had no prior experience in computer games.

The purpose of the evaluation was to test the prototype concerning the quality of the tutorial, the game mechanics, and the game experience.

We implemented 10 different puzzles with increasing difficulty on basis of artificial sequence alignments. With each difficulty step, the number of sequences and their length were increased. On the other hand, the similarity of the sequences was decreased and the scoring scheme changed with each level. For practical reasons, the biggest levels had 20 sequences with a length of 16 amino acids (see Fig. 4).

Each test person had a maximum of one hour to play. There were no further restrictions, therefore playing the tutorial or specific levels were completely optional as well as the minimal playing time. Each participant was encouraged to report her or his immediate game impressions and to identify issues while playing the game. After playing, the participants had to answer a questionnaire consisting of 38 items based on a Likert scale. Each item had 5 possible answers, ranging from „fully disagree“ to „fully agree“.

The first 14 items represented a short version of the Game Experience Questionnaire (GEQ)¹ [18]. The GEQ measures game experience on seven categories: Immersion, Flow, Competence, Tension, Challenge, Positive and Negative Affects. In the short version (ingame GEQ; iGEQ), each category is measured by two items.

The items 15 to 29 were related to usability aspects regarding the design of the user interface and the controls according to the ISO 9241/(1)10 norm. These items were designed in correspondence to the Isometrics Questionnaire [20]. The last 9 items addressed the usability and helpfulness of the tutorial.

We analyzed the different items by coding their correspondent answers with a integer value between 0 (fully disagree) and 4 (fully agree). The results shown in the following sections correspond to the mean value of the coded answers.

4.2 Results

Game Experience The results for the GEQ categories are represented in Fig. 5. Our prototype performed very well in the category of positive affect with a mean of 3.03. The players had fun playing the game and felt pleased. They also had a substantial feeling of competence (2.6) and challenge (2.65) in the game. The majority of the participants reported a loss of their sense of time during the play, which relates to a considerable flow (2.55).

In the immersion category, our prototype performed moderately with a mean value of 2.22. While the majority of the participants were aesthetically pleased by the prototype (2.7), they were only moderately impressed (1.74).

Relating to negative impressions, our prototype showed also very good results. The majority of the players were neither bored nor found the game tiresome, which results in a mean of 0.57 in the category of negative affect. They experienced only a very low level of tension (0.61).

¹ Unfortunately, the original GEQ document [18] is no longer available. The GEQ can be found in [19].

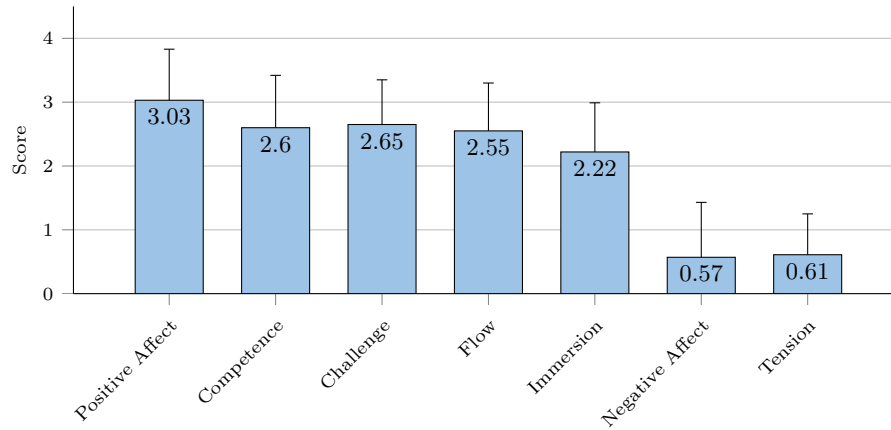


Fig. 5. Results for the GEQ items. The bars indicate the mean values for the single categories. The error bars represents the corresponding standard deviations.

Tutorial The evaluation of the tutorial showed also reasonable results (see Fig. 6). All but one participant used the tutorial. The majority found the tutorial comprehensive (3.3), well structured (3.06) and helpful (2.79). They had the feeling, that they learned all relevant parts of the game (3.06). The possibility to instantly practice the learned content inside the tutorial level was also considered very positive (3.5). Some participants rated the tutorial length as too long (0.97) and would appreciate a shorter version with fewer text components.

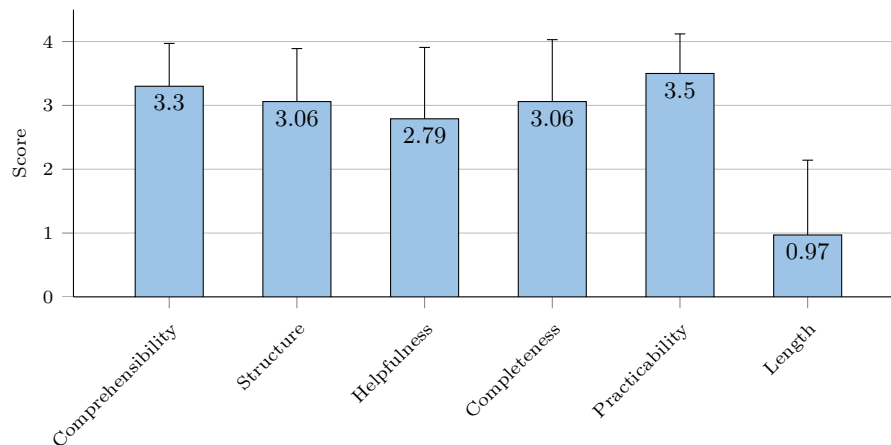


Fig. 6. Results for the tutorial items. The bars indicate the mean values for the single categories. The error bars represents the corresponding standard deviations.

Usability and game mechanics The results of the usability study and the game mechanics are represented in Fig. 7. The design of the graphical user interface (GUI) was rated very well with a mean of 2.95. The participants were very pleased with the screen layout and they estimated the graphical user interface very comprehensive.

The controls showed even better results with a mean of 3.37. The participants stated, that the controls were very intuitive, easy to learn, good to memorize, and precise.

Contrary to the tutorial, the comprehensiveness of the scoring system received only moderate ratings (1.95). Some of the participants had problems identifying the reasons for an increase or decrease of the score, especially with gaps involved. This was primarily due to a lack of transparency of the interdependency of matches, mismatches and gaps and the different scoring schemes per level. Another problem reported by some test users was the discrimination of some sphere colors like different types of green. Due to this issue, the rating for the color discrimination was only moderate with a mean of 2.2.

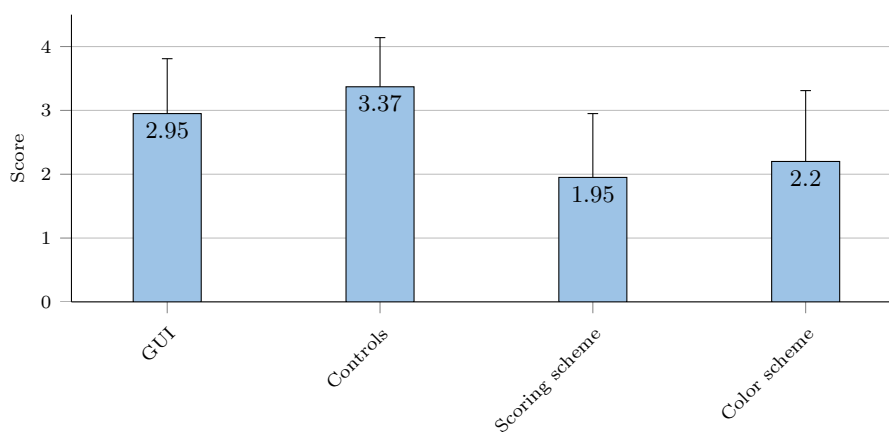


Fig. 7. Results for the usability and game mechanic items. The bars indicates the mean values for the single categories. The error bars represents the corresponding standard deviations.

5 Discussion and Future Work

As mentioned above, our prototype already performed reasonably well regarding usability and game experience. However, there is no 'ground truth' available because game experience was not measured directly in previous studies. Furthermore, comparing our game to a purely entertaining game or sequence alignment applications would be equally unfair. However, the results of this first study can

serve as a reference for follow-up studies. Some major issues are the comprehensiveness of the scoring scheme and the discrimination of the sphere colors, especially in alignments with twenty different amino acids. According to flow theory [21], a negative effect of these issues on game experience would have been reflected in low flow and high tension scores. As illustrated in Fig. 5, this is obviously not the case.

In addition some participants felt a little constrained in their game control due to the lack of a multi-selection feature. The players' immersion should be also improved in the future.

To address these issues, we are planning to implement several new features. We will change the scoring scheme to take into account biologically derived mutation probabilities (e.g. the BLOSUM matrices [17]). Further, we will implement highlighting techniques to improve the understanding of the new scoring schemes. The sphere colors will also be rearranged to improve their discrimination. Another intermediate task, is the development of improved interaction techniques. On the long run, we are planning to improve the prototype by adding sound, a more game like visual design and multiplayer features. We are also planning to integrate the game into a story context to improve the players' immersion.

6 Conclusion

We presented an improved game approach to solving protein sequence alignments using an integration of the concepts of citizen science and gaming. Our approach is capable of dealing with a large number of sequences with variable length without any restrictions on the playing field dimensions. We implemented a game prototype with a major focus on usability, visually appealing appearance and fun of gaming. The evaluation of our prototype regarding these aspects showed promising results. In spite of the early developmental state of our game, the players already had much fun in playing it, especially in large levels. They rated the game prototype as visually appealing, controllable, and challenging. In the future, we will improve our game concept and will evaluate, under which conditions the players are able to solve real alignments.

7 Acknowledgment

The authors are grateful for a FiF (Center of Interdisciplinary Research) grant by Technische Universität Darmstadt supporting this study.

References

1. Mount, D.W.: *Bioinformatics: Sequence and genome analysis*. 2 edn. Cold Spring Harbor Laboratory Press (2004)
2. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. *Journal of computational biology* **1**(4) (1994) 337–348
3. Higgins, D.G., Sharp, P.M.: Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**(1) (1988) 237–244
4. Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5) (2004) 1792–1797
5. Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglu, S.: Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome research* **15** (2005) 330–340
6. Katoh, K., Misawa, K., Kuma, K.i., Miyata, T.: Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research* **30**(14) (2002) 3059–3066
7. Notredame, C.: Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol* **3**(8) (2007) e123
8. Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M.A., Singh, M., von der Heydt, R.: A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological Bulletin* **138**(6) (2012) 1172–1217
9. Hand, E.: People power. *Nature* **466**(7307) (2010) 685–687
10. Land, K., Slosar, A., Lintott, C., Andreescu, D., Bamford, S., Murray, P., Nichol, R., Raddick, M.J., Schawinski, K., Szalay, A., Thomas, D., Vandenberg, J.: Galaxy zoo: the large-scale spin statistics of spiral galaxies in the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society* **388**(4) (2008) 1686–1692
11. Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., et al.: Predicting protein structures with a multiplayer online game. *Nature* **466**(7307) (2010) 756–760
12. Kawrykow, A., Roumanis, G., Kam, A., Kwak, D., Leung, C., Wu, C., Zarour, E., Sarmenta, L., Blanchette, M., Waldispühl, J.: Phylo: A citizen science approach for improving multiple sequence alignment. *PLoS ONE* **7**(3) (03 2012) e31362
13. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* **55**(1) (2000) 68–78
14. Cooper, S., Khatib, F., Makedon, I., Lu, H., Barbero, J., Baker, D., Fogarty, J., Popović, Z., players, F.: Analysis of social gameplay macros in the foldit cookbook. In: *Proceedings of the 6th International Conference on Foundations of Digital Games*. FDG '11, New York, NY, USA, ACM (2011) 9–14
15. Khatib, F., Cooper, S., Tyka, M.D., Xu, K., Makedon, I., Popovi, Z., Baker, D., Players, F.: Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* **108**(47) (2011) 18949–18953
16. Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., Jaskolski, M., Baker, D.: Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology* **18**(10) (2011) 1175–1177
17. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**(22) (1992) 10915–10919

18. IJsselsteijn, W., Poels, K., de Kort, Y.: The game experience questionnaire: Development of a self-report measure to assess player experiences of digital games. TU Eindhoven, Eindhoven, The Netherlands (2008)
19. Nacke, L.: Affective ludology: Scientific measurement of user experience in interactive entertainment. (2009)
20. Gediga, G., Hamborg, K.C., Düntsch, I.: The isometrics usability inventory: an operationalization of iso 9241-10 supporting summative and formative evaluation of software systems. *Behaviour & Information Technology* **18**(3) (1999) 151–164
21. Csikszentmihalyi, M.: *Beyond boredom and anxiety*. Jossey-Bass (2000)