Evaluating the Effectiveness of Color to Convey Alignment Quality in Macromolecular Structures

Julian Heinrich[†], Sandeep Kaur[*] and Seán I. O'Donoghue[†‡]
[†] CSIRO, Australia
[*] UNSW, Australia
[‡] Garvan Institute of Medical Research, Australia

***Abstract*—We investigated the effectiveness of color mapping as a means to convey uncertainty in sequence-to-structure alignments for proteins. To this end, we evaluated a color modulation scheme to encode residue conservation by collecting human preference data for a set of structures and alignment qualities. To this end, we conducted a user study on Amazon Mechanical Turk with more than 40 participants expressing a preference in the choice of two images encoding different alignment qualities of the same structure. The results of this study suggest that there is a strong correlation between human preference of an image of a structure and the alignment quality encoded in that image.**

I. Introduction

The Protein Data Bank (PDB) [1] currently comprises the most extensive protein structure resource listing more than 100,000 structures as of June 8, 2015. In contrast, the number of known protein sequences is substantially larger (the UniProt database [2] currently lists more than 50 million entries), as sequencing technology has advanced at a much faster rate than the technology required for the experimental determination of 3D structures. In fact, the number of known protein sequences is much larger, as individual entries in UniProt may represent many variants of the same sequence, for example as a result of splicing or mutations.

For sequences whose structure is unknown, numerous protein structure prediction methods have been developed [3]. Aquaria [4], developed by our group, is one such tool that does template-based homology modelling and enables visualisation of the resulting 3D structure models. In homology modelling, the protein structure of an amino acid sequence of unknown structure is inferred from the similarity to a sequence of amino acids with known structure. The key step in homology modelling is an *alignment* of the sequences with known and the unknown protein structure - the optimal alignment is found via a complex process that includes gathering all known related sequences and building a profile, thereby including evolutionary information[1]. The quality of such a sequence-to-structure alignment is usually indicated by calculating the number of identical amino-acid matches in the alignment, normalised by the length of the matching residues. The resulting identity score is what we will refer to as *alignment quality*, which gives an indication of the model *uncertainty*. The database underlying Aquaria reports alignment identity scores as a percentages, which can take values between 0% and 100%.

In Aquaria, users search for protein sequences either by name or identifier. Once a protein sequence is found, the system the system then automatically retrieves for all structures in the database with reliable matches to the query sequence; these are presented to the user in a concise visual overview of all matching sequence-to-structure alignments. This is shown in the 'Matching Structures' part of Figure 1, where each lane represents a *cluster* of structures that match the specified sequence at the indicated region. By default, the structure with the largest number of identical residues when aligned onto the specified sequence is first shown in an interactive 3D view. The user can then choose to display the best matching structure within a cluster by selecting that cluster; alternatively, by clicking on the number at the end of each cluster, the user can select and visualise any of the matching structures.

In addition to the overall quality of an alignment, Aquaria further allows the user to visually inspect the *location* of alignment matches or mismatches. A mismatch of residues may fall into one of three categories:

1) *Conserved substitution*: the residue in the structure has similar chemical properties to the one in the sequence.
2) *Not conserved substitution*: the residue in the structure has very different is chemical properties to the one in the sequence.
3) *Insertion*: a region of one or more residues in one sequence that are not aligned to the other sequence.

In order to convey this information to the user, Aquaria employs saturation and brightness modulation for conserved substitutions (i.e. reduces saturation and brightness of the default color for that residue) and shows non-conserved substitutions in dark grey and insertions in light grey [5]. This coloring scheme results in an increasing number of unsaturated and grey colors in the 3D visualisation of a structure with decreasing quality of the respective alignment (see Figure 2) and thus provides a means for conveying the uncertainty associated with the structure matching the query sequence to the user. This is a critical piece of information, as it enables biologists using Aquaria to get a sense for the extent to which the structure they are looking at represents the protein they were looking for.

The main contribution of this work is an investigation of the effectiveness of this color mapping in conveying alignment quality to the user. More precisely, we are interested in

[1]Note that sequence alignments are non-trivial and can be computed in various ways. As it is out of the scope of this article to discuss the optimal choice of alignment algorithm for template-based homology modelling, we refer the reader to O'Donoghue et al. [4] for a description of the algorithms used in Aquaria, on which this work is based.

Figure 1: Screenshot of the Aquaria page after querying for 'BLK' in the search box on the top left. Matching structures are listed according to their alignment position with respect to the query sequence (the sequence of 'BLK', as shown below the 3D view). The coloring is depicted in the legend to the right of the structure. Alignments are sorted by alignment quality, which is also given as a numeric value to the left of an alignment as a percentage of the match.

investigating the correlation between the *alignment quality* of a sequence-to-structure alignment and the *perceived quality* of an image of the respective protein structure. By applying this color map, we hypothesise that high-quality alignments will produce images that are more likely to be perceived as aesthetically pleasing to humans than images of low-quality alignments.

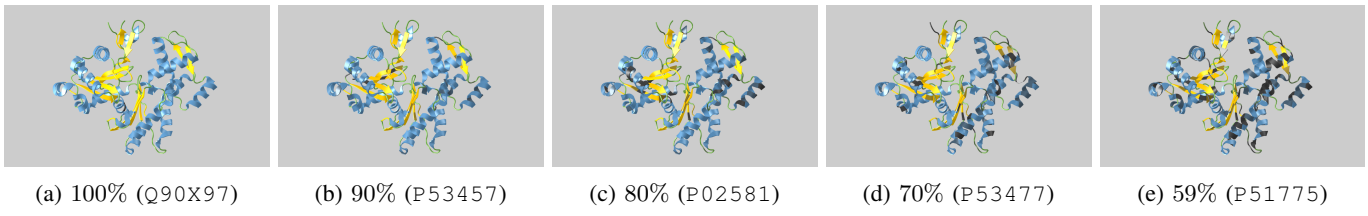
II. Related Work

The visualization of macromolecular structures is an active area of research in visualization [6] with many applications in molecular biology [7]. While there is a large body of knowledge about rendering techniques for biomolecules [6], we are not aware of any work that has been done in evaluating how rendered, static images of such molecules are perceived by the user.

This work is based on Aquaria [4], a resource that focuses on making structures accessible to a wide audience by incorporating visualization design principles [5] and user-interface design into a website. While the coloring of structures by sequence features has been investigated before [8], coloring by alignment quality was introduced by Aquaria.

Crowdsourcing using Amazon's Mechanical Turk service has become a common method to conduct cost-effective user-studies, with examples in information visualization [9] or computer graphics [10]. The results suggest that MTurk can be effective at obtaining good results [9] if some basic pitfalls such as gaming the system can be prevented [11].

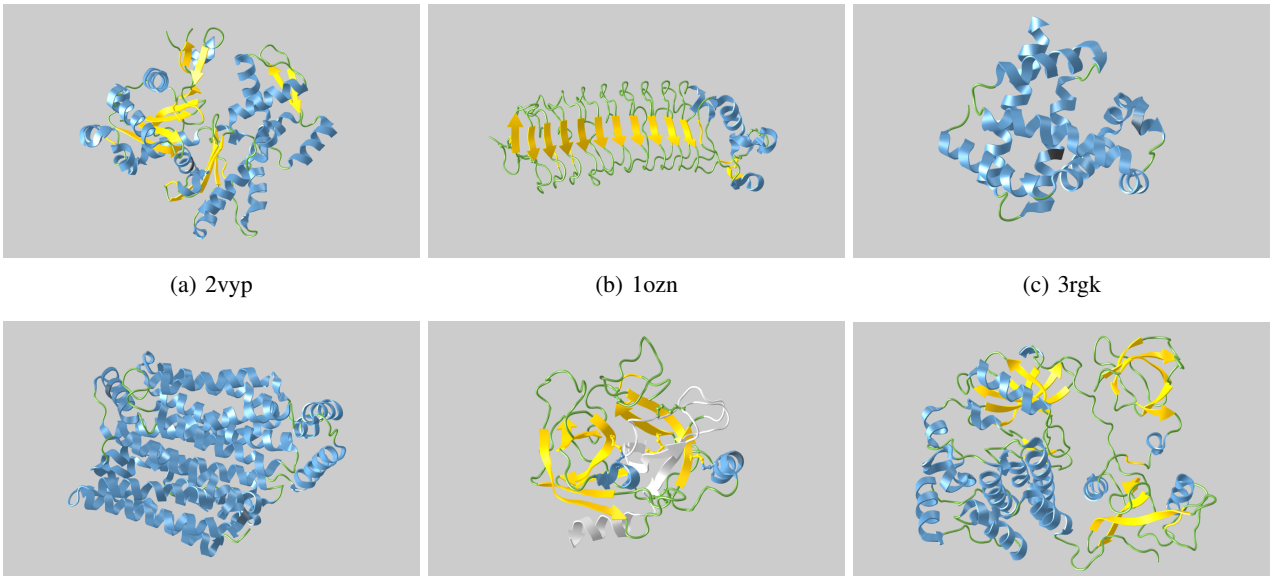
There are different approaches to obtain a ranking of stimuli from a user study [12], [13]. Here, we decided to employ a two-alternative forced choice methodology, as it was found to be very effective in terms of the effects that can be measured given the little amount of work required by participants [12]. To further reduce the workload per participant, we use a sorting algorithm to establish a ranking from a series of images without the need to compare every image against every other [14].



(a) 100% (Q90X97) (b) 90% (P53457) (c) 80% (P02581) (d) 70% (P53477) (e) 59% (P51775)

(f) 49% (P53489) (g) 40% (Q8TC94) (h) 30% (Q9H9F9) (i) 19% (Q9US07) (j) 12% (Q0VA61)

Figure 2: Alignment qualities for a single structure (PDB id: 2vyp), aligned to 10 different protein sequences. The aligned sequence's UniProt identifier and alignment identity percentage are shown with the image.



(a) 2vyp (b) 1ozn (c) 3rgk

(d) 4pyp (e) 2tbs (f) 1qcf

Figure 3: Protein structures used in the study. These images show the highest quality alignment for each of the structures. Note that the structure 1qcf (f) was only used in the pilot.

III. Study Design

The aim of our study was to test if the color map described in the previous section for coloring uncertain regions of a protein structure displayed can be used to effectively convey alignment quality. Our hypothesis is that images of high-quality alignments are more likely to be perceived as high-quality images and as such are preferred by humans over images of structures with poor or low-quality alignments. Our overall strategy to test this hypothesis is to establish a ranking of the perceived quality of images with varying alignment quality by collecting human preference data and to investigate how well this data matches the order of alignment qualities.

While there exist different approaches to establishing such a ranking, Mantiuk *et. al.* [12] found that two-alternative forced-choice (2AFC) experiments offer a good tradeoff between statistical power and time investment required by the experimentor and the participants. In the forced-choice design, participants are shown two images side by side (as illustrated in Figure 4), and are asked to choose one of them. By comparing every image to every other in the set of available images, one can then infer a ranking of the full set from the total number of "votes" per image.

If the number of stimuli n is large, however, the "full experiment" comparing every image to every other can become unfeasible for a within-subjects design, as it requires $n(n-1)$ comparisons per subject. Noting that the number of comparisons should be higher for pairs of stimuli that are similar to each other, Silverstein and Farell [14] proposed a

Figure 4: Screenshot of a single task in the two-alternative forced choice study design.

method to reduce the number of comparisons to $n \times log_2(n)$ per participant. Their approach works by assuming transitivity in the ranking of stimuli, i.e. if image A is better than image B and B is better than C, then A is better than C. Then, a sorting algorithm can be used to choose a pair of images to compare in an online fashion, based on the previous choice. While this technique tends to concentrate comparisons around very similar images (which are most sensitive to subjective variations), and thus may slightly reduce precision, the overall accuracy was found to be very high [12] when compared to alternative methods and taking into account the investment of time for both analysts and participants. In our study, we implemented this approach using a binary search tree as described in detail by Silverstein and Farrel [14].

A. *Choice of Stimuli*

In order to obtain a representative set of stimuli for our study, we selected six protein structures based on three requirements:

1) structures should have an approximate uniform distribution of alignment qualities (from 100% to approximately 10%).
2) all structures should represent a diverse set of sizes, shapes, and secondary structure compositions (such as primarily alpha sheets, beta sheets or mixed), as well as
3) structural classifications (such as globular or membrane proteins).

Based on these requirements, one of the authors who is an expert in protein structures (SIOD) manually chose molecules from the Aquaria database with the following ids: `2vyp`, `1ozn`, `4rgk`, `4pyp`, `2tbs` and `1qcf`. These structures are shown in Figure 3.

Images of these structures with different coloring according to their alignment qualities were created using the Aquaria database, by searching sequences that matched the sequence of the respective PDB entry with a minimum match length of 70%. For every structure, we obtained a range of sequences with alignment qualities from 100% to approximately 10%. From these sequences, we manually picked nine with an approximate equi-distant alignment quality. The aligned sequence identifiers were appended to Aquaria's URL, as "http://www.aquaria.ws/P46896/4pyp/", where for example `P46896` is the sequence that aligned to `4pyp` with 88% identity. On typing in the URL, Aquaria returns an image based on the original `4pyp` PDB file (an example display by Aquaria is shown in Figure 1). The returned image shows full colour as per the Aquaria colour scheme where the sequence aligns, and where the image does not align, the hue and saturation are reduced (light grey or dark gray as discussed in section I). As an example, all images with varying alignment qualities for the protein with PDB id `2vyp` are shown in Figure 2.

The images returned by Aquaria were further processed. We hid the displayed macro-molecular bound ligands. The structure was then zoomed into. The browser was then made full-screen and a screen shot was taken. To ensure that all images, across all structures, had the same dimensions (and minimal padding), we set the screen-shot dimensions to initial values of ($x = 769$, $y = 923$), and ending values of ($x = 1062$, $y = 647$). Note that images for the `1qcf` structure, used in the pilot were saved directly from Aquaria, screen-shots were not taken.

B. Experimental Setup

We utilised Amazon's Mechanical Turk to run our study. Requesters post so called "HITS" (Human Intelligence Tasks), which workers can complete and get paid for. To host the study, we developed our own system (using PHP and MySQL, and hosted on our server) which linked externally from Mechanical Turk. Linking our website via Mechanical Turk ensured that we were able to easily recruit (anonymous) participants to sign up for our study, and hosting the study on our own servers gave us more control on the display and nature of the information collected.

To allow users to perform our tasks, we set a "qualification task", where users had to agree to a Participant Informed Statement and Consent (PISC). A qualification task is Mechanical Turk terminology for a pre-task that users must complete successfully to qualify for the actual study. Once the users successfully accepted the PISC, the Mechanical Turk workers could participate in our study.

We ran an initial pilot, using the structural alignments of `1qcf`, for 12 participants to test our setup. We then created five different tasks on Mechanical Turk, one for each of the structures `2vyp`, `1ozn`, `4rgk`, `4pyp`, and `2tbs`. We requested fifty participants for each of the tasks and offered a reward of 0.68 cents for the completion of each task. This amount was computed from the US minimum wage of 8 USD/hour, assuming that a single choice can be expressed in 10 seconds. During the study, as discussed previously, we used the forced-choice method, to display images and used a binary tree representation to save the responses. Interleaved in this process, at regular intervals (determined by a binomial distribution), two images, which had previously been shown were shown again, and the response was recorded. This was done to test the attentiveness of participants during the study and to filter for participants that appeared to choose images at random.

As we are interested in conveying alignment quality with the first (intuitive) impression users get when they see a structure in Aquaria, we instructed our participants to express a choice as quickly as possible. In order to be able to relate their preference of an image to alignment quality, we further instructed them to chose an image based solely on their personal subjective judgement of the *aesthetic* quality of an image. More precisely, we used the following wording which was inspired by the study conducted by Secord *et. al.* [10]:

> Suppose that you had to choose one of the images to appear in a magazine or product advertisement. Do not worry if neither of them is ideal. Just choose the one that you think is best.

After a worker agreed to the PISC and read the instructions, we started the experiment and recorded responses. A screenshot of a question is shown in Figure 4.

All Mechanical Turk operations (such as downloading results and requesting HITs) were performed using the AWS Mechanical Turk command line reference tools and the R interface.

IV. Results

In this section, we present the data obtained from the pilot and the main study and analyse the relation of perceived quality to alignment quality using logistic regression and analysis of ranks.

A. Pilot Study

In order to test the experimental setup, we first ran a pilot study on Mechanical Turk on one structure (`1qcf`, see Figure 3f) with 9 levels of alignment quality. We assigned 5 participants to the task as described in the previous section and obtained 131 choices for pairs of images. For every image, we computed the probability of success, i.e. the probability an image would be preferred over *any* other image, as the total number of votes divided by the total number of views. We then ranked all images according to that probability such that the image on rank 1 is the most likely to be preferred over any other image and the image on rank 10 is the least likely to be preferred over any other image. Figure 5 shows the resulting ranking in relation to the alignment quality. From this plot, a clear linear relationship between both axes emerges. Except for two images, the resulting ranking matches the order of alignment qualities *exactly*.

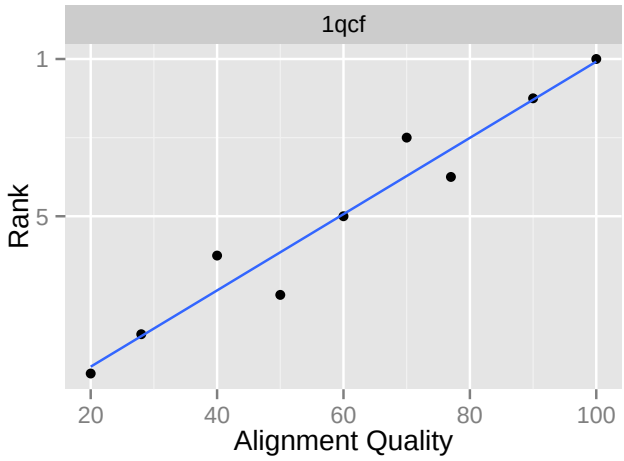


Figure 5: Alignment quality and rank for the pilot study with 9 images of the structure shown in Figure 3f. Note that with exception of two images (on ranks 3 and 6), all ranks match the order of alignment qualities.

From the pilot study we can therefore infer that alignment quality is a good predictor for the ranking of images according to the human preference of these images.

B. Main Study

In order to extend the results we obtained from the pilot study to other structures, we ran another experiment on Mechanical Turk with the 5 structures and 10 levels of alignment qualities as described in the previous section. This time, we obtained a total of 5,478 judgements from 59 participants overall. Figure 6 shows the results of the main study after ranking images as described before for the pilot. Again, a clear

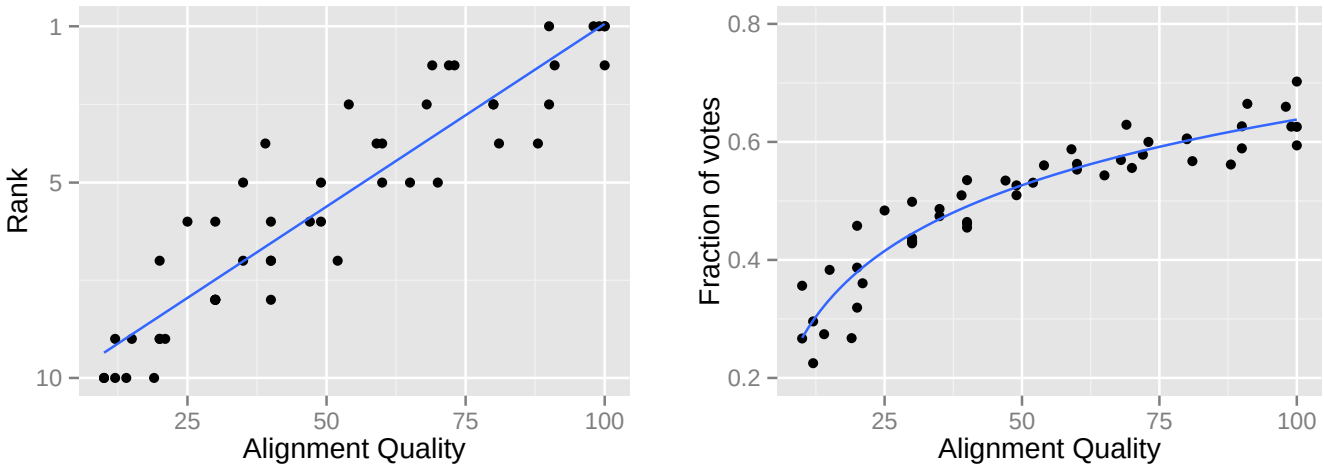


Figure 6: Alignment quality versus rank (left) and fraction of votes (right) per image, for all images in the main study. The best-fit regression curves are shown in blue.

Coefficient	Estimate	S.E.	t-value
Intercept	-0.1	0.031	-3.254
log(quality)	0.16	0.008	19.488

Table I: Coefficients, standard errors, and t-values of the linear regression on the logarithmic transform of alignment quality. Both the intercept and the coefficient for the log-transform are significantly different from zero ($p < 0.01$). Overall, the model fits the data very well ($R^2 = 0.88$).

trend emerges that suggests that alignment quality performs well at predicting the ranking of images as judged by humans based on the perceived quality.

In addition to our investigation of ranks, we further analysed how well alignment quality predicts the probability of an image being preferred over any other image. Figure 6 shows alignment quality and the probability as the fraction of votes obtained for every image and all structures. In contrast to the ranks, the overall shape of this correlation is logarithmic. Differences in low-quality alignments (up to approximately 25%) seem to result in larger differences in the respective probability of getting picked. We therefore conducted a linear regression on the logarithmic transform of alignment quality of the form:

$$P(q) = a \log(q) + b, \tag{1}$$

where P is the probability that a human prefers an image of a structure with alignment quality q, while a and b are the unknown regression coefficients. The best-fitting curve is shown on the plot (Figure 6), and Table I lists the results of the regression. Both the intercept (b) and the slope (a) are significantly different from zero ($p < 0.01$). Overall, the model fits the data very well ($R^2 = 0.88$).

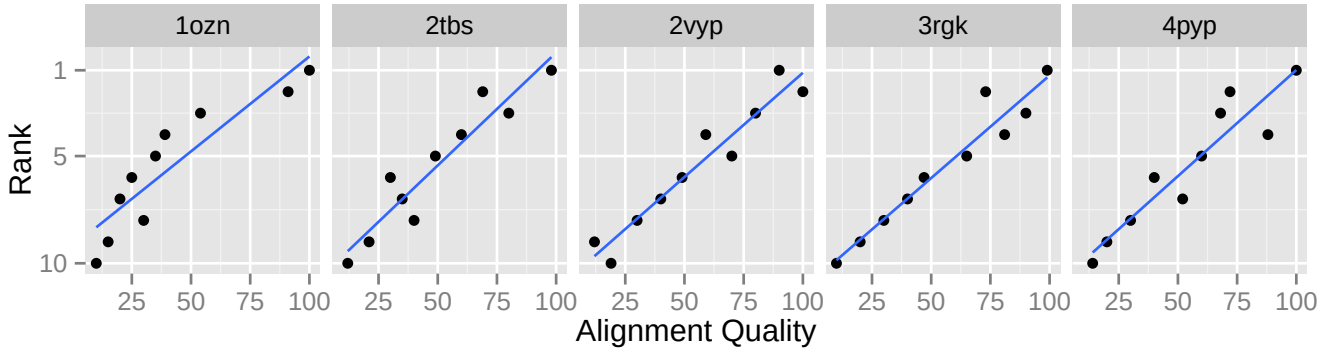
To further analyse the data, we repeated the same analyses for every structure individually. Figure 7 shows the alignment quality, ranks, and probabilities for each structure of the main study. From these plots, the overall trend as seen in Figure 6 can be confirmed for all tested structures as well. For some structures, however, the color mapping seems to work better than for others: while the ranks for structures with ids `3rgk` and `1ozn` match the order of alignment qualities almost perfectly, there are up to three images ranked differently for the structure with id `2vyp`.

Figure 7b also shows the fraction of votes in favor of each image versus the respective alignment quality. All plots show the same logarithmic correlation between both variables. The most frequently chosen image correlates with the highest quality alignment in all but 1 cases (`2vyp`). Similarly, the image participants liked the least agrees with the lowest alignment quality for all but one (`2vyp`) structure. While there is some slight variation in between these two extremes, a clear trend emerges from the plots suggesting that there is a strong correlation between alignment quality and human preference of the corresponding image.

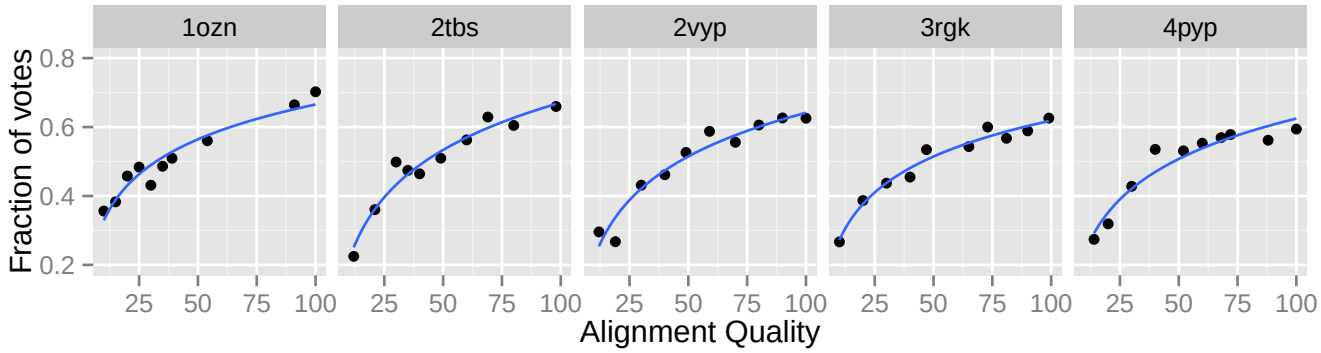
V. Discussion and Future Work

Our results indicate a strong correlation between the quality of sequence-to-structure alignments and the perceived quality of images of molecular structures created with our colormapping. By applying this coloring scheme, high-quality alignments are on average perceived as aesthetically more pleasing than low-quality alignments, thus confirming our initial hypothesis. Furthermore, we found that, overall, participants ranked images according to the respective alignment qualities. This result is important for practical applications, as it confirms that our approach can be used *effectively* to convey alignment quality via aesthetic properties of an image created using our colormap.

The coloring scheme applied in Aquaria to convey alignment quality systematically reduces saturation and brightness with an increasing number of residue substitutions and insertions. Thus, with increasing alignment quality, the overall image becomes "brighter" and more saturated. Given that the intensity of the physical sensation of brightness follows a power law with a similar shape as the curves shown in



(a) Alignment quality versus rank for each structure.



(b) Alignment quality versus fraction of votes for each structure.

Figure 7: Ranks (a) and fraction of votes (b) for each structure in the main study. While there seems to be good agreement of alignment quality with perceived quality overall, there is some variation between structures. For the structures with ids `3rgk` and `1ozn`, only one image does not match the order of alignment qualities, while for `2vyp` and `4pyp`, up to three images have been ranked differently than the alignment quality suggests.

Figures 6 and 7, our results might indicate a possible relation between the overall brightness of the image and its aesthetic quality.

We were aiming at instructing participants to express a choice based on the aesthetic preference for an image. However, we did not explicitly use the word "aesthetic" in the instructions, as each individual's interpretation of this term can be different. Instead, we used the phrasing as suggested by Secord et al. [10] to ensure that participants use a common baseline for their judgements. However, it is important to be aware of the fact that the exact phrasing of instructions can change the outcome of such a study.

Finally, while our results suggest that the coloring scheme used in Aquaria is effective in communicating alignment quality, we can only claim so for the set of structures investigated in this study. However, these results appear to be very promising and represent a starting point to investigate human preference of molecule images further. Of particular interest is the difference in the perception of these images between non-experts and experts in molecular structures. We are planing to investigate this by inviting experts to participate in the study and extending our protocol with pre-tests of participants' knowledge.

VI. Conclusion

We presented an evaluation of the colormapping used in Aquaria to visually convey the quality of protein-to-structure sequence alignments. Our results are based on human preference data for five structures with 10 alignments each and suggest a significant correlation between the alignment quality and the perceived quality of images created from the respective alignment. While this relationship is non-linear, overall the ranking of images as determined by participants matches the order of alignment qualities. We can therefore conclude that images of high-quality alignments in Aquaria are more likely to be perceived as of high quality by a user of the resource.

Acknowledgment

This work was supported by CSIRO's OCE Science Leader program and its Computational and Simulation Sciences platform. The authors would like to thank Tim Peters for his valuable feedback on the statistical analysis of the results of this study.

REFERENCES

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[2] The UniProt Consortium, "Activities at the universal protein resource (UniProt)," *Nucleic Acids Research*, vol. 42, no. D1, pp. D191–D198, 2014.

[3] A. Pavlopoulou and I. Michalopoulos, "State-of-the-art bioinformatics protein structure prediction tools (review)," *International Journal of Molecular Medicine*, vol. 28, no. 3, pp. 295–310, 2011.

[4] S. I. O'Donoghue, K. S. Sabir, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, M. Roos, N. Perdigao, F. A. Buske, J. Heinrich, B. Rost, and A. Schafferhans, "Aquaria: simplifying discovery and insight from protein structures," *Nature Methods*, pp. 98–99, 2015.

[5] C. Stolte, K. S. Sabir, J. Heinrich, C. J. Hammang, A. Schafferhans, and S. I. O'Donoghue, "Integrated visual analysis of protein structures, sequences, and feature data," *BMC Bioinformatics*, vol. 16, no. Suppl 11, p. S7, 2015.

[6] B. Kozlikova, M. Krone, N. Lindow, M. Falk, M. Baaden, D. Baum, I. Viola, J. Parulek, and H.-C. Hege, "Visualization of biomolecular structures: state of the art," in *Eurographics Conference on Visualization (EuroVis) - STARs*, R. Borgo, F. Ganovelli, and I. Viola, Eds. The Eurographics Association, 2015.

[7] S. I. O'Donoghue, D. S. Goodsell, A. S. Frangakis, F. Jossinet, R. A. Laskowski, M. Nilges, H. R. Saibil, A. Schafferhans, R. C. Wade, E. Westhof, and A. J. Olson, "Visualization of macromolecular structures," *Nature Methods*, vol. 7, pp. S42–S55, 2010.

[8] E. C. Meng, E. F. Pettersen, G. S. Couch, C. C. Huang, and T. E. Ferrin, "Tools for integrated sequence-structure analysis with UCSF Chimera," *BMC Bioinformatics*, vol. 7, p. 339, 2006.

[9] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using Mechanical Turk to assess visualization design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 203–212.

[10] A. Secord, J. Lu, A. Finkelstein, M. Singh, and A. Nealen, "Perceptual models of viewpoint preference," *ACM Transactions on Graphics*, vol. 30, no. 5, pp. 109:1–109:12, 2011.

[11] R. Kosara and C. Ziemkiewicz, "Do Mechanical Turks dream of square pie charts?" in *Proceedings of the 3rd BELIV'10 Workshop: BEyond Time and Errors: Novel evaLuation Methods for Information Visualization*, ser. BELIV '10. New York, NY, USA: ACM, 2010, pp. 63–70.

[12] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[13] C. Demiralp, M. Bernstein, and J. Heer, "Learning perceptual kernels for visualization design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1933–1942, 2014.

[14] D. A. Silverstein and J. E. Farrell, "Efficient method for paired comparison," *Journal of Electronic Imaging*, vol. 10, no. 2, pp. 394–398, 2001.