Structural Biology: Assignment 3

University of Cambridge

Henrik Åhl

June 15, 2017

Preface

This is an assignment report in connection to the *Structural Biology* module in the Computational Biology course at the University of Cambridge, Lent term 2017. All related code is as of June 15, 2017 available through a Github repository by contacting hpa22@cam.ac.uk.

Exercises

1 Figure 2a shows the frequency distribution of the weighted pairwise matches between sequences, given that their similarity fraction is $> 70\%$. For every sequence, the number of *other* sequences with a > 70 % similarity to this adds an increment of one to the parameter d_i in the equation

$$W_i = \frac{1}{1 + d_i}.$$

That is, we downscale for weights for the sequences that are high in abundance (given our similarity threshold), since they add little additional information to what is already known. This means that the histogram we see in fig. 2a is given in fractions of $1, 1/2, 1/3, \ldots$, as this shows the frequency distribution of these weights.

Figure 2b shows the predicted protein contact map between the residues in each sequence, i.e. how correlated certain positions are with each other based on the evolutionarily inferred contact (EIC) scores. Figure 2c in turn depicts the minimal distance between the amino acids in the crystal as a function of the corresponding DI rank. Lastly, fig. 2d shows the overlap between the experimentally observed structure and the predicted tertiary interactions.

2 The accuracy of the method is directly dependent on the quality of the sequence alignment. As these often utilize multiple heuristic simplification in order to improve speed, longer sequences are bound to align improperly. While functionally important segments should be better conserved and align better, there is always the risk of error in this, which makes the method as a whole reliant both on the alignment method used, and the numbers of sequences aligned. Using more elaborate approaches such as direct dynamic programming is instead feasible when analysing smaller and/or fewer proteins. While cumbersome and time-consuming, one can also, by using prior information of how the sequences in question differ, tailor the alignment according to data at hand in order to improve upon this. Another obvious problem is the quality of the experimental correlations, which depends inherently on the sample size and resolution of the protein in question. This might obfuscate lowly correlated but functionally important residues, which affect the overall structure and stability by long-range interactions or simply aid the protein in folding into the native state. Lastly, sequence alignment data does not take into account environmental factors, such as how certain parts of the protein might interact with other molecules in a way such that protein functionality invisible in the primary sequence is conserved or maintained. If this is the case, and this functionality is not con-

served spatially in relation to other residues, this might as well be obfuscated. In other words, as a high score indicates contact, a low score must not necessarily mean no contact between residues.

3

4 The parameter θ sets, as previously hinted, the similarity percentage cutoff between sequences. In our case, as it is set to $\theta = 0.3$, only sequences with more than $1.0 - 0.3 = 0.7 = 70$ % similarity will be processed when calculating the weights. This makes it so that, according to the aforementioned equation, sequences which are very similar will add less to the corresponding weights, as the information is duplicated in several parts of the input data, which here are the homologues. By doing so, we smoothen our distribution of weights, and thus makes it more variable due to smaller variation.

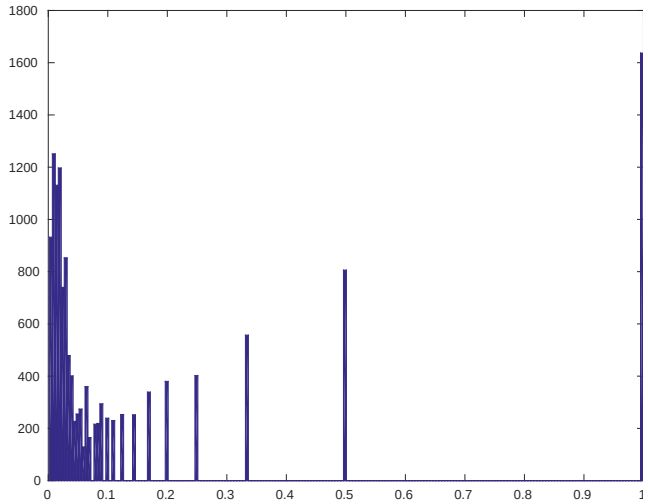
5 Pseudocounts are used for smoothing distributions in particular when some specific probabilistic outcomes are unlikely to occur in relation to the size of the data set. In the case of amino acids, this could be to add and increment to the number of times a residue has been observed. For each amino acid, a different increment would be attributed, based on a prior estimates of how likely they are to appear. This approach is done in order to avoid sharp peaks in the distribution of values in for example a position weight matrix, where small changes in the number of observations would have large effects the resulting value, in the case where pseudocounts are not used. In practice, pseudocounts can be set to calculate the posterior estimator of the mean as

$$\theta_i^{PME} = \frac{n_i + \alpha_i}{|\mathbf{n}| + |\alpha|}$$

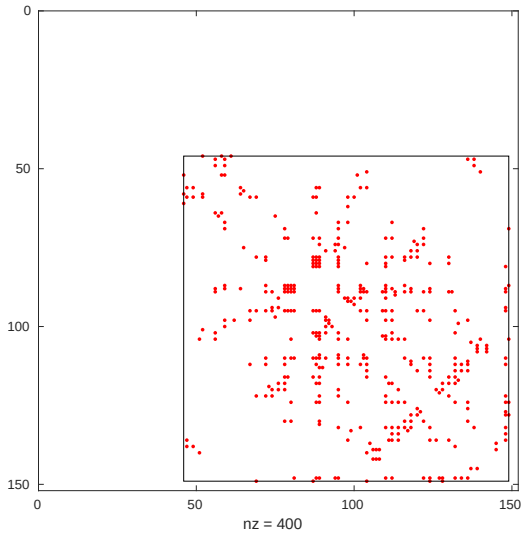
where n_i is the number of observations of component i, and α is the pseudocounts for that component [1].

References

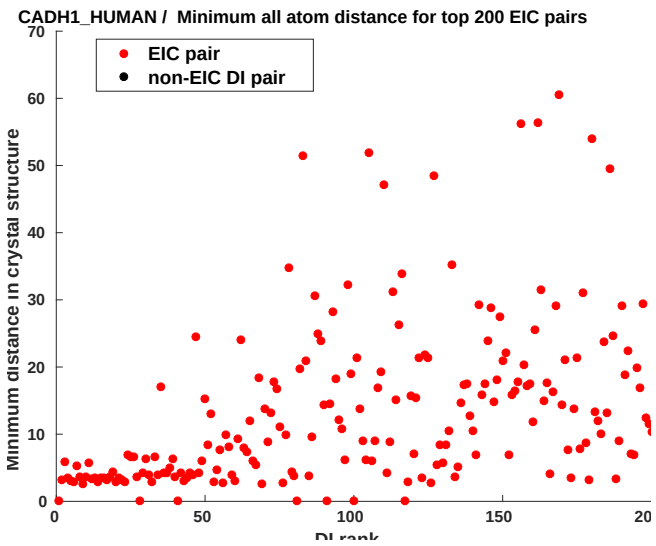
[1] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.



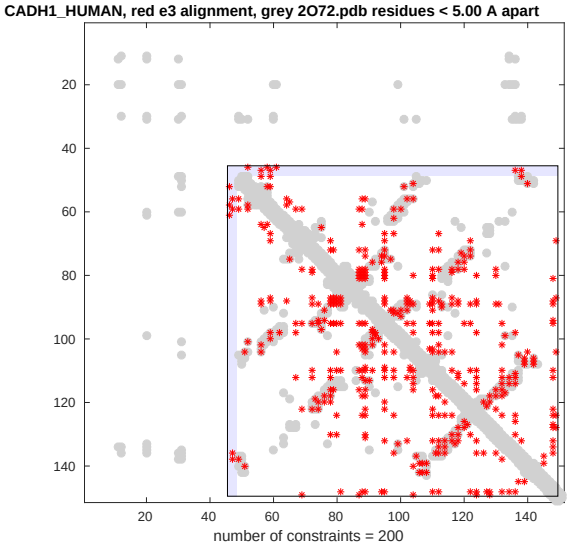
(a)



(b)

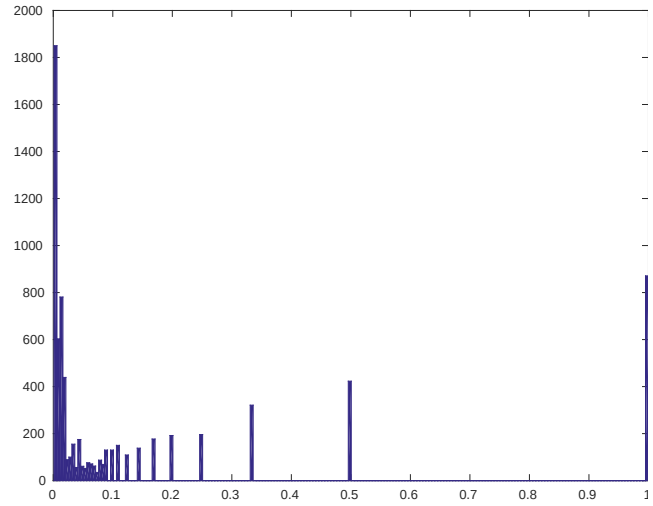


(c)

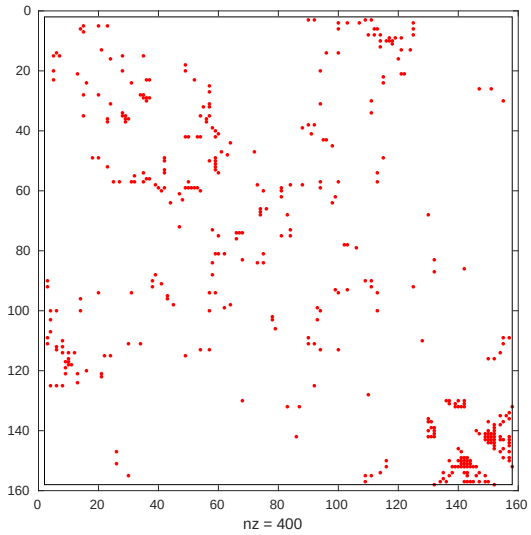


(d)

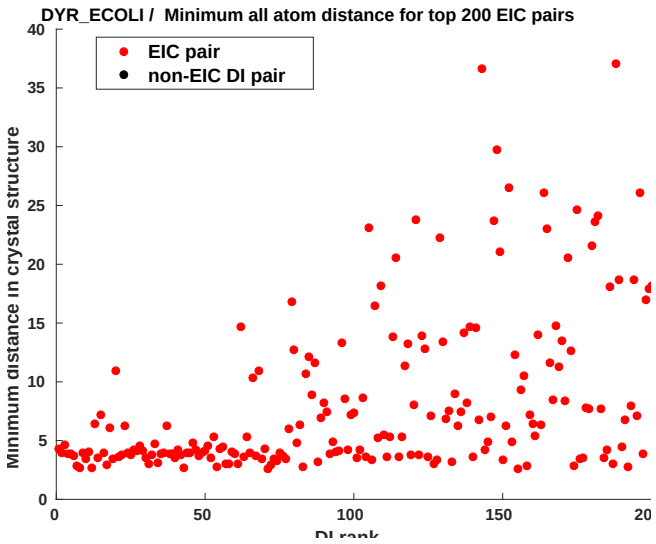
Figure 1: Figure of figures



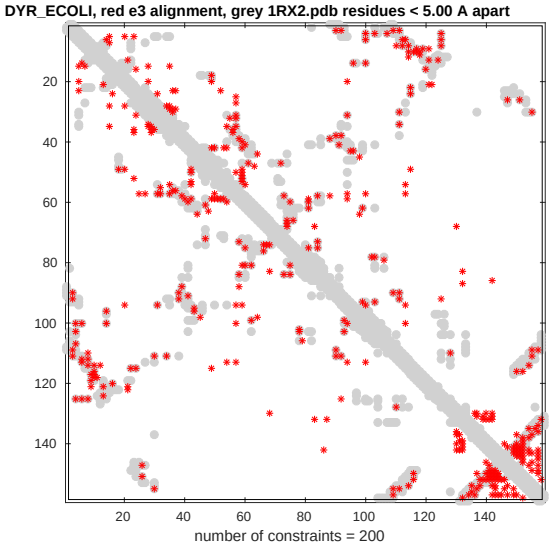
(a)



(b)



(c)



(d)

Figure 2: Figure of figures