Using Structural Protein Features for Long- and Medium-Range Contact Prediction

Heye Vöcking 376154

Abstract—Protein contacts have proven to be a very valuable tool helping to solve the problem of *ab initio* protein structure prediction. Based on the idea that contact-maps of different proteins may in part be similar, we are evaluating the MaDCaT library as a search tool for improving predicted contact-maps. Furthermore, we evaluate the robustness of the method regarding noisy native contact-maps.

Index Terms—protein structure prediction, contact prediction, contact-maps, MaDCaT, PDB

1 Introduction

The well-established method of template-based modeling using fragments from the Protein Data Bank (PDB) [1] has shown that it holds valuable information that can be used for protein structure prediction. In this work, we want to make use of this information from another perspective, by mining the PDB for residue-residue contacts that may help improve predicted contact-maps in the future. Because this search is not based on the sequence of the protein it may be able to find similarly structured proteins that have a completely different sequence and therefore would have not been considered by a sequence-based method.

March 18, 2019

1.1 Related Work

The idea of using contacts for predicting three-dimensional (3-D) models of proteins has been around since more than two decades ago [2], [3] and has gained more and more traction in the recent years [4], [5]. Several methods for protein structure prediction exist, in particular, we want to mention PSICOV [6] here, which uses sparse inverse covariance estimation, as we will be using it in combination with MaDCaT [7] which uses distance-maps to search for similar protein backbone arrangements.

1.2 Motivation

We are using predicted PSICOV contact-maps as the query for MaDCaT to search for similar contacts in a databank of native contact-maps. We then replace the contacts of the prediction with those found by MaDCaT in hope of improving the accuracy and thus improving the quality of the contact-map, to better aide protein structure prediction.

1.3 Definition of Contacts

When folded, proteins form a 3-D structure, which can be expressed in x, y, and z coordinates of the amino acid residues in the form of a PDB-file. Two residues are said to be *in contact* if the Euclidean distance of their specific atoms in 3-D space in the folded protein structure is less than a distance threshold, the threshold used in this work is 8 Ångstrom. See figure 7. When we refer to a *contact* in this work we are speaking of a pair of residues fulfilling the aforementioned criterion. Therefore, based on the positional distance of the amino acids on the backbone, we differentiate between short-range (<12 residues in between), medium-range (12 to 23 residues in between), and long-range (>24 residues in between) contacts. Short-range contacts are not very interesting in this scenario since they are in contact automatically (depending on the secondary structure) due to their placement on the backbone. However, medium- and long-range contacts give very valuable information in regard to the overall structure of the protein.

A contact-map is a sparse matrix represented by a list of residue pairs (a and b) that encode a contact from residue a to b. Since this matrix is mirrored on the diagonal a contact from a to b implies a contact from b to a, which therefore must not explicitly be expressed, that means the contact-map only defines the upper part of the matrix, so for each entry $a > b$ holds true.

MaDCaT uses distance-maps, which are basically contact-maps with an additional value expressing the inverse distance d^{-1}. Where d is the Euclidean distance measured in Ångstrom. To make a contact-map compatible with a distance-map we add a value for d^{-1} of 1.0.

1.4 Contact Accuracy

An important tool for comparing the quality of contact-maps is the measure of contact accuracy. That is the number of contacts that are correct versus the total number of contacts.

$$accuracy = \frac{correct\ contacts}{total\ number\ of\ contacts}$$

Note, whenever we talk of accuracy in this work, we exclude short-range contacts and contacts with coil regions.

2 Methods

At the heart, our approach consists of slicing up the contact-map of the *target* (the protein for which we want to find contacts) to build several queries. We then use the MaDCaT library to search the database of contact-maps with each of the generated queries. Finally, we reassemble the returned matches and calculate their accuracy.

2.1 Query

To build a query, the input contact-map is getting broken up into ranges, defined by secondary structures, as shown in figure 1. We are using the protein 1bkr as an example since it has a reasonable length and the best accuracy compared to its random baseline result.

To simplify the search, coils are excluded, so only alpha helices and beta sheets are considered (figure 6 and 7).

Searching over the whole protein takes quite a long time, therefore we build triplets of these structures by combining the contact-maps of the broken up structures with each other in all possible combinations. An example of one of those triplets extracted from the contact-map is shown in figure 8 and one of those parts colored in a 3-D structure is shown in figure 9. Each combination is then passed to MaDCaT as input query to search for the match with the best score in the database. MaDCaT returns a list of matches, each with a score indicating how well it matches the query. We always take the match with the best score and discard all others. We then combine the results from all queries back together. The retrieved and merged contact-map is shown in figure 2. We then calculate the accuracy of this combined contact-map against the native contact-map.

2.2 Dataset

We use native structures provided by the PDB and precompute contact-maps of these. Our dataset is based on the "all.list" provided on the website of MaDCaT [8]. We clean it by filtering out undesired entries, using the following criteria: We only use A-chains of the proteins and exclude those entries without A-chains completely. Furthermore, we exclude all proteins with less than 80 or more than 200 residues. We also exclude all structures not analyzed using X-Ray crystallography or with a resolution lower than 2Å. All PDB-files downloaded are fixed and checked for errors. We exclude those we cannot process properly. The resulting dataset has 15,723 contact-maps.

We differentiate between *database* and *dataset*. *Dataset* refers to all of the aforementioned 15,723 contact-maps and *database* refers to a subset of contact-maps of the *dataset*. For the experiments, we created 3 different kinds of databases: one containing only the native contact-map of the target which is used as a query, which we call DB_1. Another database named DB_2, which contains the whole dataset, including the native contact-map of the current target. And a third database called DB_3, containing all contact-maps from the dataset with the exclusion of the contact-map for the current target. See table 1.

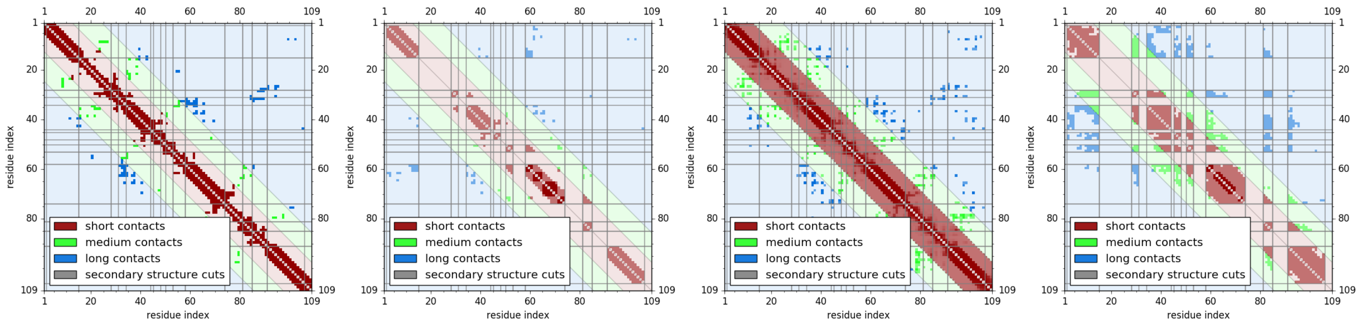

Fig. 1: Native contact-map

Fig. 2: All retrieved contact-maps merged together of exp3.

Accuracy: 0.574

Fig. 3: PSICOV contact-map, with all short-, $1L$ medium- and $1L$ long-range contact confidences.

Accuracy: 0.257

Fig. 4: All retrieved contact-maps merged together of exp6.

Accuracy: 0.012

Fig. 5: Different contact-maps of 1bkr. The gray lines indicate borders of secondary structures. The darker the color of the square in figure 2 and figure 4, the more matches indicate a contact at this position. The darker the color of the square in figure 3, the higher the confidence for it being a real contact.

Name	Description
DB_1:	Only native of the target
DB_2:	Wohle dataset
DB_3:	Whole dataset without native

TABLE 1: Types of databases used

Name	Query	DB type
exp1	Native	DB_1
exp2	Native	DB_2
exp3	Native	DB_3
exp4	PSICOV	DB_1
exp5	PSICOV	DB_2
exp6	PSICOV	DB_3

TABLE 2: Configurations of experiments conducted

Target	exp1	exp2	exp3	exp4	exp5	exp6	PSICOV
1bkr	0.574	0.574	0.574	0.092	0.092	0.012	0.257
1e6k	0.356	0.281	0.279	0.068	0.071	0.071	0.504
1f21	0.607	0.617	0.609	0.131	0.119	0.119	0.594
1h0p	0.333	0.253	0.160	0.058	0.058	0.068	0.528
1hzx	0.750	0.750	0.094	0.004	0.013	0.013	0.046
1odd	0.538	0.538	0.357	0.093	0.033	0.033	0.313
1r9h	0.798	0.288	0.297	0.141	0.158	0.158	0.713
1rqm	0.485	0.266	0.175	0.112	0.084	0.084	0.473
1wvn	1.000	0.518	0.509	0.251	0.035	0.035	0.482
2hda	0.506	0.366	0.366	0.293	0.302	0.302	0.849
2it6	0.391	0.222	0.222	0.106	0.077	0.077	0.500
2o72	0.285	0.123	0.119	0.089	0.076	0.076	0.683
5p21	0.439	0.319	0.394	0.085	0.072	0.072	0.516
5pti	1.000	1.000	0.964	0.750	0.122	0.122	0.686

TABLE 3: Accuracy for all experiments with sampling value of 1.0 and their respective PSICOV contact-map, sorted by target name

3 Results

The results look promising when querying with native contact-maps but did not yield improvement for predicted contact-maps. Therefore, we analyzed what may be the cause of the lack of improvement.

3.1 Experiments

We conducted 6 experiments each using a different combination of query and database ordered from the simplest case (exp1) to the most difficult (exp6), see table 2 for an overview.

The simplest experiment *exp1* is using a contact-map generated from the native protein structure for building the queries and a database consisting of only that same exact contact-map. In the results of this experiment, we can see that the algorithm finds matches with varying levels of accuracy.

Though one could argue that the algorithm didn't really have to ”search” the database as there was only one (and perfect) entry. Therefore we performed exp2, which used the whole dataset, also known as DB_2, as the database. The results decreased substantially in some cases, which leads to the assumption that the algorithm is deceived into preferring non-optimal results.

Finally, we conducted exp3, removing the native from the database. Interestingly the performance is still quite similar compared to exp2 in most cases, even though the native contacts were not present in the database. Therefore we can conclude that the algorithm is able to find contacts similar to contacts in other proteins.

Results shown in figure 14, 15 and table 3.

3.2 Further Investigation

In order to determine how easy it is for the algorithm to find correct contacts by chance, we used randomly generated contact-maps to get a baseline for each target.

To test robustness of the algorithm contacts are removed (subsampling), shown in figure 11

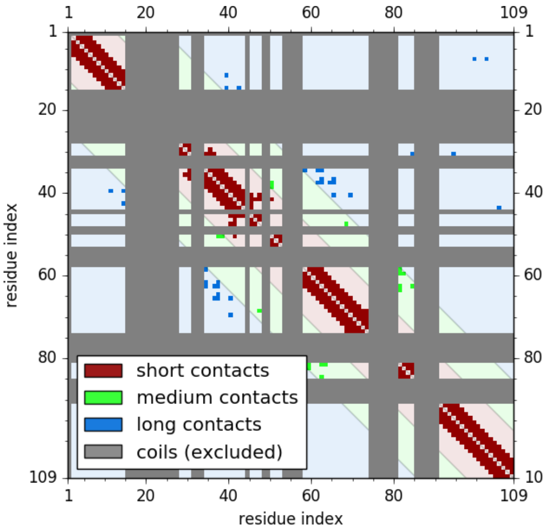

Fig. 6: Contact-map of 1bkr without coils

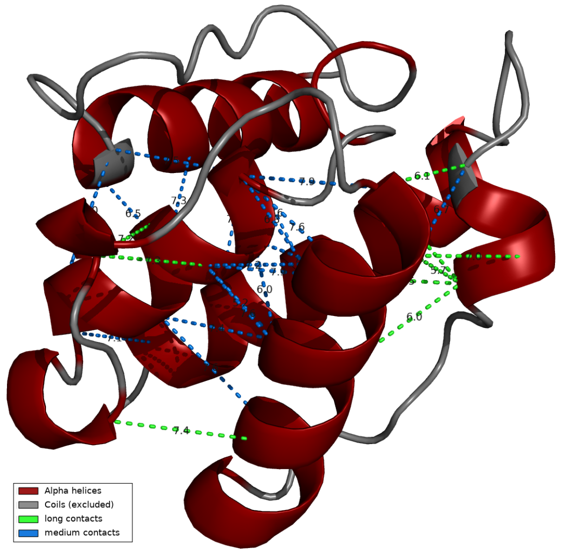

Fig. 7: Contacts of 1bkr in 3-D structure

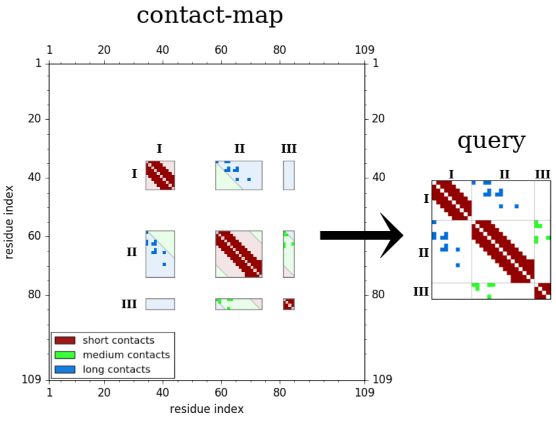

Fig. 8: Query built from selected ranges in contact-map

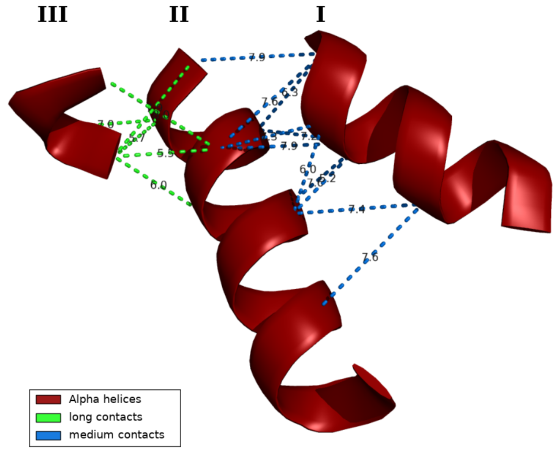

Fig. 9: 3-D representation of query with contacts

Fig. 10: A query built from the contact-map of 1bkr made up of a secondary structure triplet from ranges I: 34-43, II: 58-73, and III: 81-84

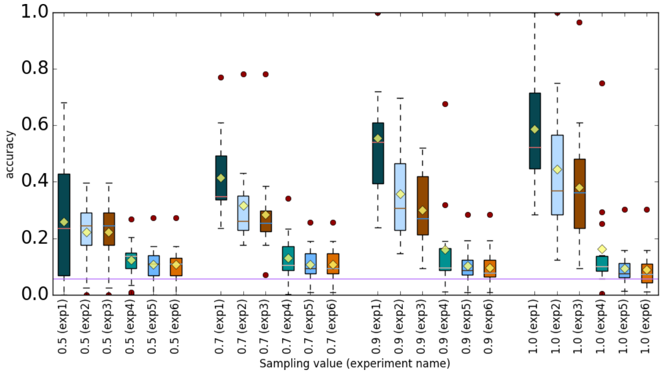

Fig. 11: Subsampling

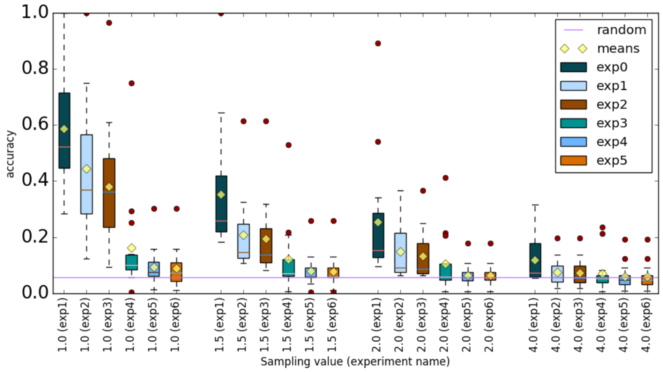

Fig. 12: Added noise

Fig. 13: Comparison of accuracy over all experiments. The purple line indicates the mean accuracy achieved with random contact-maps over all targets. A sampling value X of less than 1 indicates removal of contacts (only a ratio of X contacts of the original set are still present). A value of more than 1 indicates randomly added contacts (the total number of contacts is X times the original value). A sampling value of $X = 1.0$ indicates that the original contact-map was used without added or removed contacts.

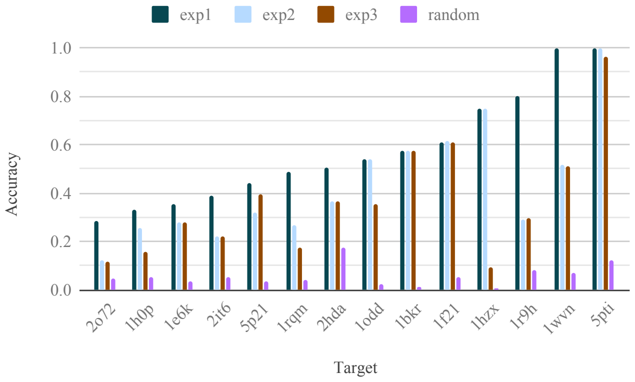

Fig. 14: Native contact-map results, sorted by accuracy of exp1

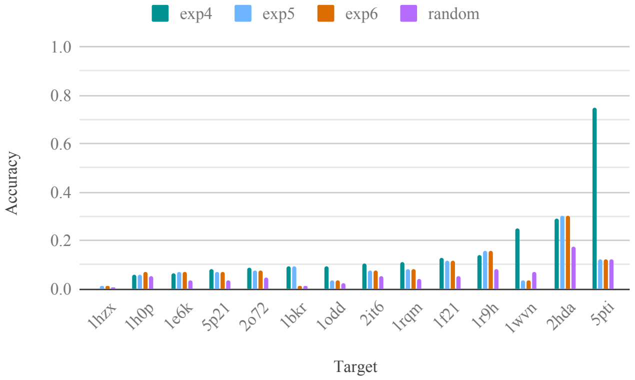

Fig. 15: PSICOV contact-map results, sorted by accuracy of exp4

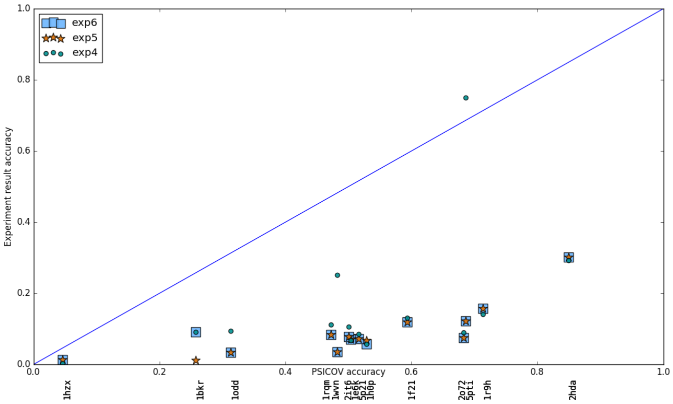

Fig. 16: Achieved accuracies in exp4, exp5, and exp6 compared to accuracies of the PSICOV predictions used for querying

and noise is added (figure 12). Even though the accuracy decreases the more we deviate from the native, it is still well above the baseline.

3.3 Predicted Contact-Maps

In exp4, exp5, and exp6 we used predicted instead of native contact-maps. The results are shown in figure 15 and 16, and in table 3. It is apparent that the accuracy decreased to very low numbers ranging from 0.004 (1hzx) to 0.750 (5pti). 1hzx is by far the longest protein with 340 residues and 5pti is the shortest with 58. The next highest accuracies are achieved by the 2nd shortest (2hda, 59 residues) and 3rd shortest (1wvn, 74 residues) targets. 1bkr achieved the best result in comparison to its random baseline in all PSICOV experiments.

For predicted contact-maps we also conducted experiments with noise and subsampling (results also in figure 11 and 12).

This shows that the of the performance algorithm does not substantially decrease depending on the grade of deviation from the predicted contact-map. We can conclude that the algorithm is not suited to improve predicted contact-maps.

4 Future Work

Despite not being able to achieve a real improvement for state of the art contact predictions, we still think there is some potential in this method. PSICOV predicted contact-maps, can basically be seen as a kind of noisy native contact-map. Since we have shown that noisy native contact-maps work compared to predicted contact-maps, further investigation eg. by combining native and predicted contact-maps, may reveal the reason for the lack of improvement.

References

[1] H. M. Berman, “The Protein Data Bank,” vol. 28, no. 1, pp. 235–242.

[2] L. Mirny and E. Domany, “Protein fold recognition and dynamics in the space of contact maps,” vol. 26, no. 4, pp. 391–410.

[3] M. Vendruscolo and E. Domany, “Protein folding using contact maps,” in *Vitamins & Hormones*. Elsevier, vol. 58, pp. 171–212.

[4] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, “Protein Structure Prediction Using Rosetta,” in *Methods in Enzymology*. Elsevier, vol. 383, pp. 66–93.

[5] D. T. Jones, “Predicting novel protein folds by using FRAGFOLD,” vol. 45, no. S5, pp. 127–132.

[6] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, “PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments,” vol. 28, no. 2, pp. 184–190.

[7] J. Zhang and G. Grigoryan, “Mining Tertiary Structural Motifs for Assessment of Designability,” in *Methods in Enzymology*. Elsevier, vol. 523, pp. 21–40.

[8] ——. MaDCaT - protein structure search tool. [Online]. Available: https://grigoryanlab.org/madcat/