Decomposing structural response due to sequence changes in protein domains with machine learning

Patrick Bryant[a] and Arne Elofsson[a]

a) Dep of Biochemistry and Science for Life Laboratory, Solna, Sweden
Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden
e-mail: patrick.bryant@scilifelab.se

Corresponding Author: Arne Elofsson, e-mail: arne@bioinfo.se, tel: +468-16 10 19, Postal address: Institutionen för biokemi och biofysik 106 91 Stockholm, Sweden

Abstract

How protein domain structure changes in response to mutations is not well understood. Some mutations change the structure drastically, while most only result in small changes. To gain an understanding of this, we decompose the relationship between changes in domain sequence and structure using machine learning. We select pairs of evolutionarily related domains with a broad range of evolutionary distances. In contrast to earlier studies, we do not find a strictly linear relationship between sequence and structural changes. We train a random forest regressor that predicts the structural similarity between pairs with an average accuracy of 0.029 lDDT score, and a correlation coefficient of 0.92. Decomposing the feature importance shows that the domain length, or analogously, size is the most important feature. Our model enables assessing deviations in relative structural response, and thus prediction of evolutionary trajectories, in protein domains across evolution.

Introduction

The structure of a protein is encoded by its amino acid sequence [1]. Amino acid sequences from proteins that share a common ancestor are homologous, and homologous proteins often have similar structures [2]. However, similar structures can be those that reach sufficient similarity in terms of a structural measurement, which is quite an elusive concept. A frequently-used definition of structural similarity is having two protein structures that share at least 70 % secondary structure identity and at most 2.5 Å root mean square deviation between alpha carbons in tertiary structure[3], although no consensus exists in this matter.

Nature appears to be reusing certain self-sustaining segments of proteins deemed protein domains [4]. Furthermore, protein evolution seems to be proceeding more by creating new combinations of these domains and less by creating actual new domains [5]. The frequency of the creation (and mechanisms) of new domains is debated. However, what is clear is that a large fraction of all proteomes is made up of a quite limited set of domains. This is less clear for sequences and proteins consisting of multiple domains, mainly present in Eukaryotes and higher organisms, which appear more evolutionarily active [6]. Domain structures considered homologous are grouped together in different databases such as CATH [7], SCOP [8] and ECOD [9], in order to utilize sequence-structure homology relationships. Today, the structural coverages of the best-studied organisms is rather complete and the regions that can not be modelled, often referred to as the dark proteome (PMID:26578815), are enriched in non-structured and low-complexity regions.

The connection between sequence and structure can be used to infer changes in structural similarity from changes in sequence similarity. This relationship has been extensively studied [8, 9, 10], but is still far from being elucidated. There is currently no method that accurately predicts how sequence changes will affect the structure in evolution. Understanding the relationship between protein sequence and structure in evolution is essential to understand proteins and their most probable evolutionary changes and directions.

There are indications of the difficulty for already stable folds to further evolve into new ones. A model of the connectivity between sequences in nature and the variation of folds has shown that groups of highly connected sequences have less structural variability. This has led to the suggestion that there exists a sequence barrier that makes it difficult for some sequences to evolve into new folds, making the overall sequence space disconnected [11]. This sequence barrier may be created by fitness barriers [12], meaning folds with high fitness have a lower tendency to be subject to structural changes.

Studies of the structural protein universe have instead observed overlaps in fold space, suggesting it is possible to go from one fold to another by the addition of structural motifs [13]. Such additions are seldom tolerated in structural cores and are instead more often added to domain surfaces. In fact, a comparison of structural overlaps has found 32 % of nonredundant structures within CATH superfamilies to overlap with different folds [14]. This is apparent when assessing global domain structure, but not when only structural cores are considered [14]. Due to the possible transitions, it may be better to represent the structural universe as a continuum [15, 16].

An early study reported a nonlinear relationship between protein sequence identity and protein structure RMSD [2]. Later, protein cores were studied by measuring sequence similarity as the number of mutations (evolutionary distance) and not sequence identity [17]. Changes in protein structure were then found to have a linear relationship with changes in the protein sequence. Analyzing the slope of the linear relationship resulted in a suggestion of protein structure to be 3-10 times more conserved than the sequence in protein cores [17], depending on what structural similarity measure was used.

Explanations of why the evolutionary rates of proteins in both sequence and structure vary are lacking. Previously, evolutionarily conserved proteins, where conservation was defined as having a BLASTP match of Evalue < 10^{-6}, have been reported to be longer on average than poorly conserved ones [18], indicating that there is a relationship between evolutionary conservation and sequence length. Domain length has also been found to be correlated with evolutionary rate measured as the number of nonsynonymous substitutions per site, dN, in yeast [19].

Contact density appears to influence the evolutionary rates of domains. Even within the same protein, domains with lower contact density have been found to evolve slower than those with higher [20]. Contact density has also been correlated with length [18], with longer sequences having higher contact densities. This is contradictory since longer sequences were found to be more conserved in proteins, but higher contact density related to higher evolutionary rates in domains. The reasoning for the positive correlation between contact density and evolutionary rate is that proteins with more buried residues can be encoded by more sequences, enabling them to evolve faster. This is only true when measuring evolutionary rates in terms of sequence changes alone (dN). It may be that the structures encoded by the rapidly evolving sequences change very slowly, why the relationship between sequence and structure evolution has to be assessed simultaneously.

Here we analyze the full structures, not only cores, to explain the complete sequence-structure relationship across different kinds of evolutionarily related domains. We study domains of different evolutionary distances, thus enabling bridging of the gap between structural classification and the assessment of the transition of folds. The correlation between relative conservation in terms of deviation from the average sequence-structure relationship is used to enable a more thorough evaluation in terms of both sequence and structural conservation. We decompose the observed variations with machine learning to test the possibility of different combinations of features explaining observed variance. This yields insight into why a similar amount of mutations results in different structural responses across different domains. The trained model can be used to predict and infer both evolutionary sequence- and structural changes of domains.

Results

Comparing sequence and structural workflows

We analyzed sequence-structure relationships using both initial sequence- and structural alignments, here called the sequence- and structural workflow respectively. Figure 1 shows all pairs between AA20 ED 0-6 for both workflows. The Pearson correlation coefficients between the sequence and structural workflows are 0.88 for AA20 ED and 0.89 for lDDT score. This makes the two workflows interchangeable to a large extent. The workflows only deviate at higher EDs, where the point spread and uncertainty in both AA20 ED and lDDT score become greater.

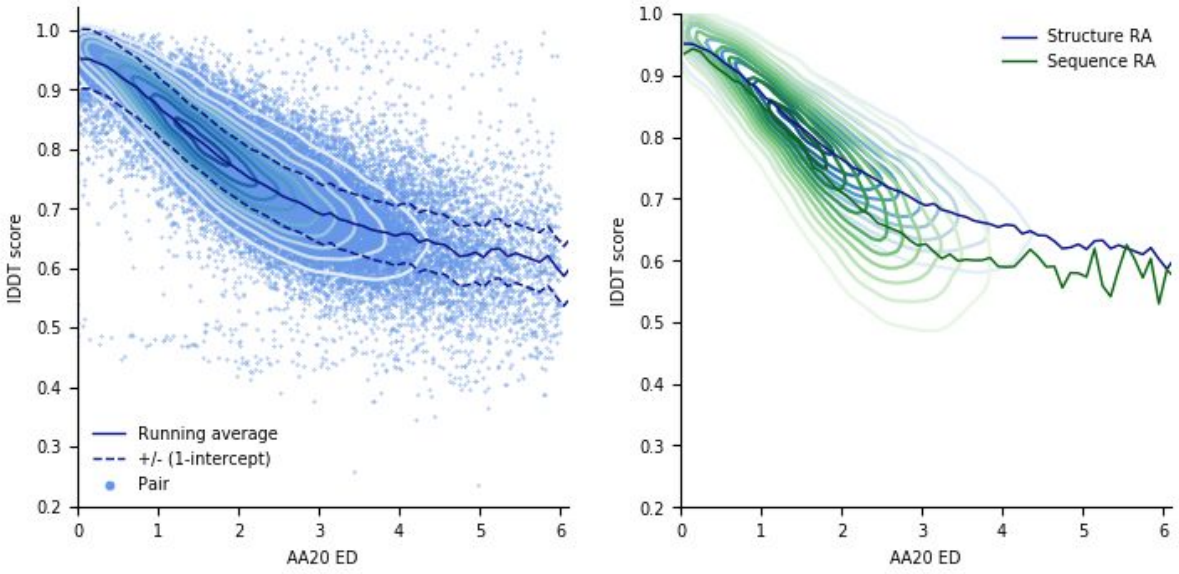

Figure 1 (Left) Results (49042 pairs, 16729 domains, 967 folds) from the structural workflow between AA20 ED 0-6. The dark blue line is the total running average for the broad dataset. The intercept at the y-axis, the minimum expected structural difference by nearly identical structures (1-0.95 = 0.05 lDDT score) is plotted as a dashed blue line. (Right) A comparison between the structural and sequence workflows. The structure running average (RA) and point density (kde) is plotted in blue, and the sequence in green. The Pearson correlation coefficient between the two workflows is 0.89 for the lDDT scores and 0.88 for the AA20 ED.

Comparing structural scores and sequence cardinalities

When superimposing the homology and fold datasets, the sequence-structure relationship appears continuous for AA20 ED and lDDT score. Irrespective of structural score or cardinality used, the structural response decelerates as sequence evolution proceeds (see Figure S2), eventually reaching a plateau. The observed relationship is thus non-linear, opposite to the previously reported linear relationship using the same sequence and structural measures, but only analyzing protein cores [17]. The results indicate that the analyzed sequences are allowed to change beyond the point of their maximal structural dissimilarity in the full structures, while the linear relationship found in cores does not. This is evident when comparing the full structures from the core analysis, which yields an almost identical sequence-structure relationship to ours (see Figure 2).

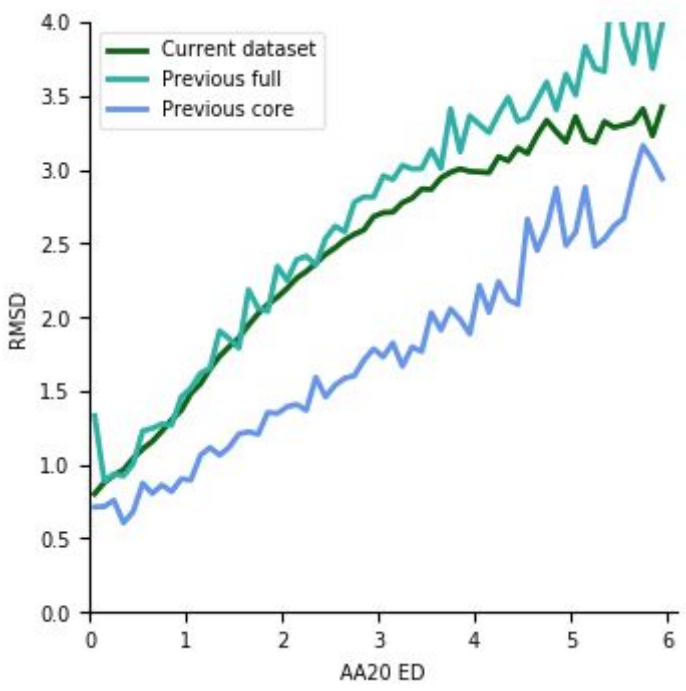

Figure 2 Comparison between running averages of the current dataset (Broad dataset) using the full domain structures and the dataset from a previous study [17] using only domain cores. Analysing the full structures dataset from the previous study yields an almost identical relationship to ours.

Extreme outliers

The two outlier sets, one with lDDT scores >0.9 and AA20 ED >2 (Set 1) and one with lDDT scores <0.45 and AA20 ED <2 (Set 2), contain 142 pairs and 6 pairs respectively (see Figure S4 and Table S1). Set 1 has 117 domains from the class Mainly Alpha, and contains mostly long straight alpha-helical structures. Set 1 has only 7 pairs from class Mainly Beta, 6 from Alpha Beta and 12 from few secondary structures. Set 2 contains only pairs from the class Mainly Beta. These are beta propellers and similar beta structures with central holes.

Since there are only 6 pairs in Set 2, it is hard to attribute feature importance to their deviation. When examining structural superpositions in this set (see Figure S3), one finds many variable loops. These structures thus appear similar globally but differ in local structure, hence the low lDDT scores. The sequences in Set 1 are very short, with very low contact densities and contact orders (see Figure 2 and 3). The most deviating CATH topology with at least 10 pairs (“Single alpha-helices involved in coiled-coils or other helix-helix interfaces”, 1.20.5 [7], 80 pairs) is part of these structures. This topology has an average standard deviation of around 3.

Explaining the variance

In the structural workflow, the sequence-structure relationship is remarkably stable across domain pairs, with 69 % (34069/49042) of all pairs being within the minimal expected deviation (0.05 lDDT score) from the total running average. Yet, substantial deviations are observed, with some points deviating as much as 0.4 in lDDT score, i.e. some set of mutation cause some domains to change their structure significantly more (or less) than others. Next, we set out to try to identify the causes of these variations.

Correlation analysis

Figures 3 and 4 show the relationship between the deviation from the total running average for each pair in the broad dataset, against different features. Each pair was assessed individually to avoid grouping biases that may arise due to the inherent CATH structure. The strongest correlation is found for the query domain length, having a Pearson correlation coefficient of -0.2. Analyzing the figure representing domain length, it is evident that shorter domain sequences have more variation in their sequence-structure relationship. This suggests that smaller domains have both weaker and stronger structural responses than larger ones due to a similar amount of mutations.

The fraction aligned of the shortest sequence in each pair, the length of the subject sequence and then percentage of gaps in the subject sequence all show correlations close to 0.2 as well. The fraction aligned and gap percentage are very similar features, both assessing sequence similarity, similarly, the query and subject lengths both represent the size of domains.

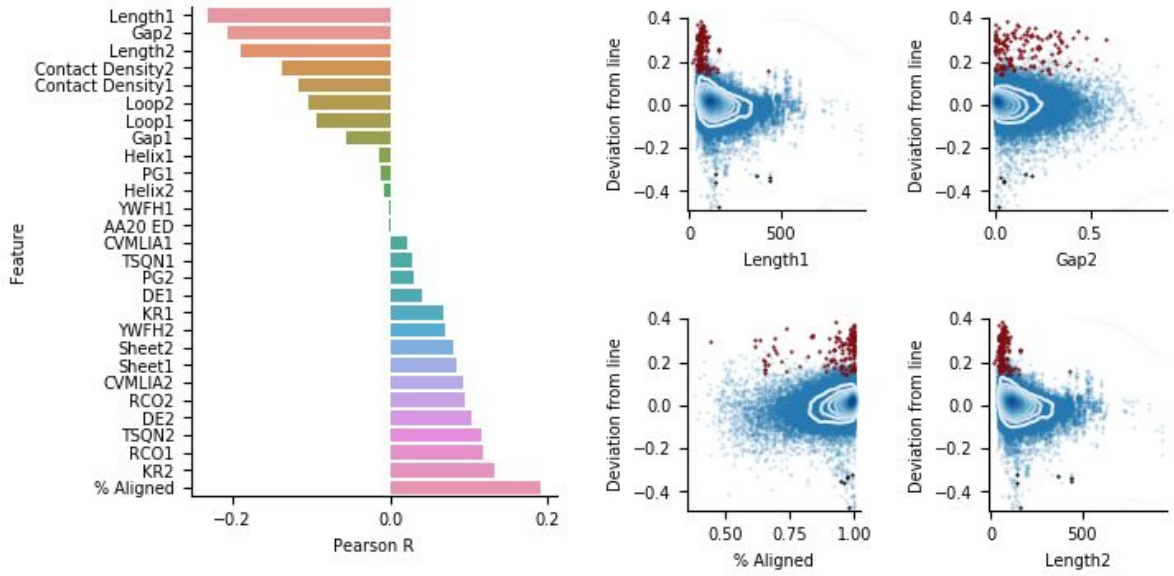

Figure 3. Deviation from the total running average in IDDT score for each pair in the broad dataset, against different features displayed by query domain (1) and subject (2). (Left) The Pearson correlation coefficients for different features are shown as a barplot, with the length of the query domain having the strongest correlation. (Right) All points together with kde for the features with the strongest correlations, the red data points represent the positive outliers (Set 1), the black the negative (Set 2) and the blue all other pairs. The relationships for the remaining features are shown in Figure 3.

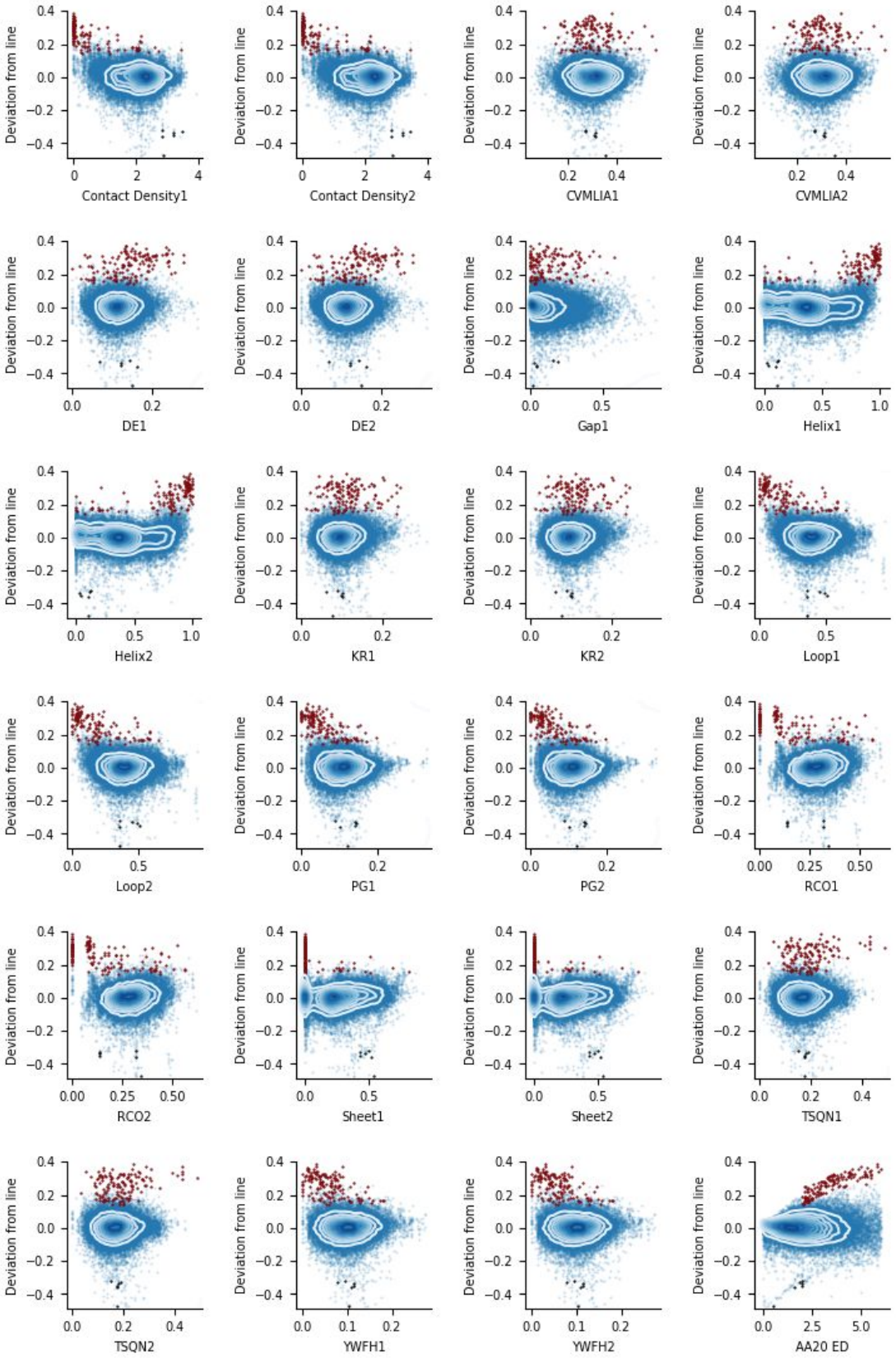

Figure 4. Deviation from the total running average in lDDT score for each pair in the broad dataset, against different features displayed by query domain (1) and subject (2). All points together with kde, the red data points represent the positive outliers (Set 1), the black the negative (Set 2) and the blue all other pairs.

Decomposition of feature importance with machine learning

Since it may be the case that different combinations of features explain the deviation and not one feature alone, we performed machine learning in the form of random forest regression. Our model predicts the deviation of pairs with an average accuracy of 0.029 (± 0.00026) lDDT score in a five-fold cross-validation, and Pearson correlation coefficient of 0.721 (± 0.007). The correlation coefficient for the lDDT scores is 0.923 (± 0.003). Using only the average lDDT score for each ED of 0.1 yields an average accuracy of 0.042 (± 0.00035) and Pearson correlation coefficient for the lDDT scores of 0.838 (± 0.006). The correlation coefficient for the deviation is zero, since no deviations from the average relationship can be mapped. The minimal expected deviation between almost identical pairs is 0.05 lDDT. The accuracy in prediction does not seem to be affected by the ED, as similar spread in prediction error is observed across EDs (see Figure 5).

Decomposing the feature importance (permutation importance [21]) shows the length of the query domain being the most important feature for both training and testing, responsible for 0.56 and 0.32 decrease in model score respectively (see Figure 5). The percent of the shortest sequence in each pair that has been aligned follows in importance and the ED thereafter, for the training. In the test set, these two features show almost identical importances. These two features should represent quite similar descriptors, as both are related to the similarity between the domain sequences in each pair. The percent of beta sheet in the query domain is also indicated to be an important characteristic for predicting structural change, although not in comparison with the aforementioned ones. All other features, such as amino acids grouped due to their characteristics (see Cardinality, AA6), the secondary structure elements (Helix, Loop) and even the length of the subject sequence are of relatively negligible importance for the predictor.

Removing all structurally related features and training a new random forest regressor with the same parameters as previously, assesses how well the sequence alone can estimate sequence-structure relationships. Using a model with only sequence features predicts the deviation of pairs with an average accuracy of 0.032 +/- 0.00032 in a stratified five-fold, and Pearson correlation coefficient of 0.631 ± 0.014. The correlation coefficient for the lDDT scores is 0.904 (± 0.005). The average prediction error is very similar to using structural features, but the correlation between the true and predicted scores worse, suggesting predictions of more exaggerated deviations. The highest feature importances appear in the same order as when using all features (see Figure S3).

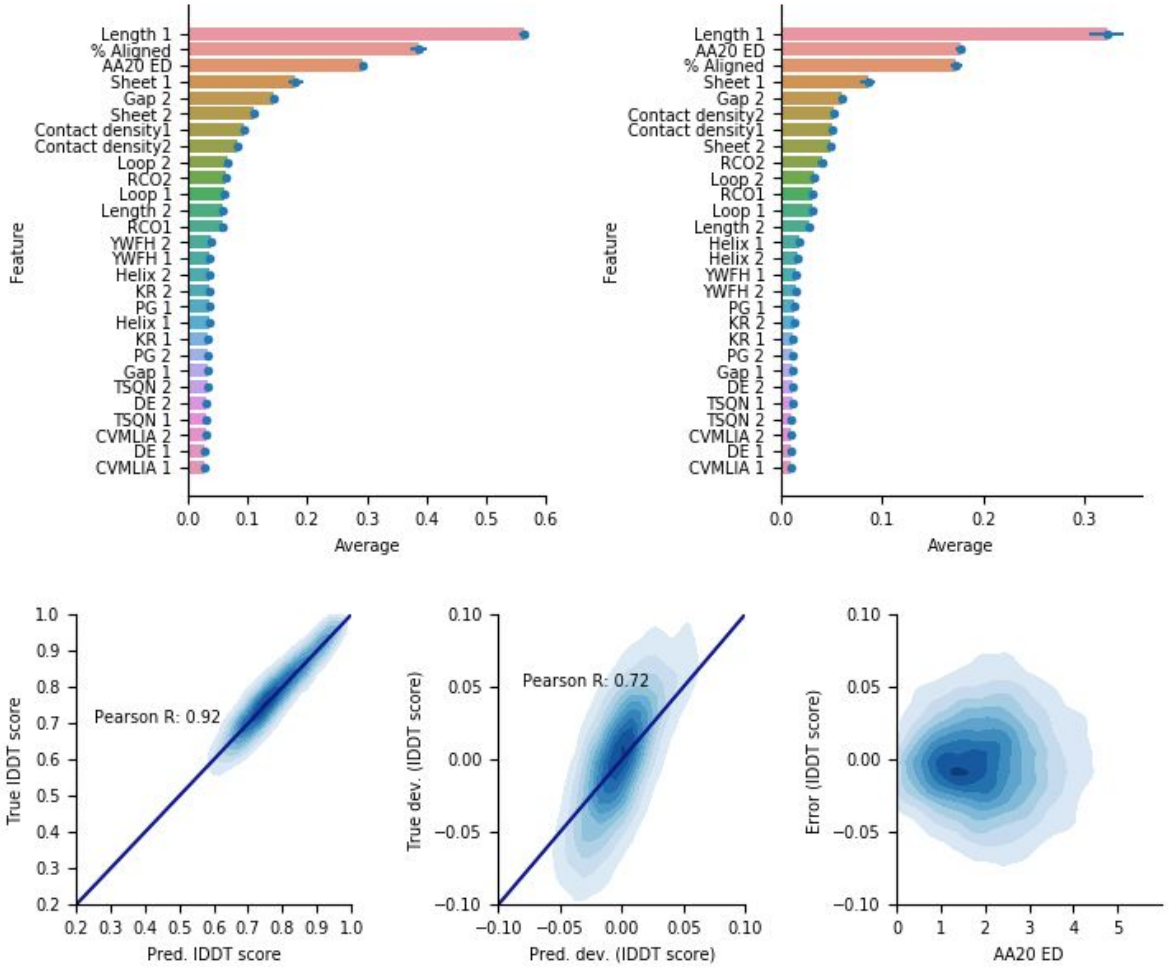

Figure 5. Feature importance by query domain (1) and subject (2) with standard deviations for predicting deviating pairs across a five-fold cross-validation. The left feature importances are calculated on the train set and the right on the test set, both are permutation importances (the decrease in predicted score when a single feature value is randomly shuffled). The true scores are compared with the predicted ones, so is the true deviation and the predicted deviation. The error in predicted score is also assessed for different EDs.

Discussion

The high correlation between initial sequence and structural alignments show that both options can be used to analyze sequence-structure relationships. The structural workflow was chosen for the main analyses due to the ability of structural alignments to capture more distant similarities.

To compare sequence and structural changes is not straightforward, as the amino acid alphabet consists of 20 discrete states, while structural measures are continuous. It is difficult to assess which structural score and cardinality are the most informative when evaluating sequence-structure relationships. Length dependent scores such as RMSD have an obvious bias, but may function well at very low discrepancies. Length independent scores such as DIFFSS and DIFFC may, in turn, enable comparisons between very different domain pairs, but have trouble representing small differences. Some scores are normalized, such as lDDT, running from 0-1. Others, such as rmsd and ED, can in theory reach infinite values. The normalized scores will have natural plateaus, while this is not necessarily true for the unlimited ones. The score lDDT was chosen as the primary score to assess structural similarity here, as this score is not length-dependent and is able to capture both global and local structural differences[22]. Similarly, to obtain a realistic sequence representation, AA20 ED was used. A cutoff of 6 AA20 ED was chosen due to the large uncertainty and lack of pairs beyond this point (see Figure S2).

The outlier sets exemplify extreme deviations from the average sequence-structure relationship. The deviation from the running average in outlier Set 1 occurs due to the flexibility and irregularity in protein domains with low contact orders and contact densities. Such domains lack structure in a manner, which may explain their high structural similarity but varying sequence similarity. When examining structural superpositions in outlier Set 2, one finds that the variable loops seem to be the reason for the low lDDT scores, even though the global structure looks very similar. The reasons behind these outliers are therefore not structural differences, but rather that the structure of these domains is more dynamic.

Why some pairs deviate more than others from the average relationship assesses why domains change with varying rates in evolution. Using statistical methods and machine learning to assign importance to protein features, decompose the underlying explanations to the observed deviations. Both statistical analysis and machine learning reveal the query domain length, or analogously, size being by far the most important feature for describing the relative structural response due to sequence changes in protein domains. The correlation analysis and feature importance differ in their ranking of the remaining features, although they both regard measures of sequence similarity to be of greatest importance.

Our model enables accurate prediction of relative structural response due to sequence changes in protein domains. How much domain structures will change due to mutations in the sequence can thus be assessed by assuming certain structural features such as contact densities remain similar across evolutionary distances, a property that can be observed in the strong means and relatively small spread in contact densities in figure 4. Compared to only using the average sequence-structure relationship, our model has (0.923-0.838)/0.838 = 10 % higher correlation and (0.042-0.029)/0.029 = 45 % better average error, relatively. The model overpredicts deviations both positively and negatively, still, the correlation coefficient is quite high, especially compared to using the average alone (0.72 and 0.00 respectively). Using only sequence features the improvements are (0.904-0.838)/0.838 = 8% and (0.042-0.032)/0.032 = 31 % in correlation and average error respectively. Our model thereby displays significant improvements in both correlation and average error.

Irrespective of structural score or cardinality used, the structural response decelerates as sequence evolution proceeds, suggesting sequences can continue to change beyond the point of their maximal structural dissimilarity. Another explanation for this observation can be that it is not possible to measure structural differences at a certain point, while more long-range sequence differences can be. Since we analyze evolutionarily related domains from both the same and different homologies and both datasets display a saturation in structural response at similar points, CATH groupings should not be the reason for these observations.

In the structural workflow, 69 % of all pairs are within the minimal expected deviation (0.05 lDDT) from the total running average. We find this uniformity remarkable, yet there are substantial deviations. The finding that shorter domain sequences have more variation in their sequence-structure relationship than longer is not surprising for the case of increased structural response. The relative change in a small structure due to a similar number of mutations should naturally be greater than in a larger one, due to a small structural disturbance having a larger relative effect the smaller a structure is. The opposite case is surprising though, and can perhaps be attributed to smaller structures having less defined global structure than larger ones, why many sequences could encode such ill-defined structures.

Conclusions

In this paper, we investigate the relationship between changes in protein sequence and structure. As reported before we note that there is, on average, a strong correlation between sequence and structure divergence. We find that irrespective of structural score or cardinality used, the average structural response decelerates as sequence evolution proceeds in full domain structures. For individual domain pairs however, there exist considerable variation between sequence and structure divergence. Investigating differences in deviation between individual pairs using common protein descriptors, the strongest correlation is found for domain query length, having a Pearson correlation coefficient of -0.2. Furthermore, shorter domain sequences are found to have more variation in their sequence-structure relationship than longer. This suggests that smaller domains have both weaker and stronger structural responses than larger ones when subject to a similar amount of mutations. Using random forest regression to assess feature combinations, we predict the structural similarity of pairs with an average accuracy of 0.029 lDDT score and Pearson correlation coefficient of 0.92 in a five-fold cross-validation. Decomposing the feature importance in our model reveals again domain query length being the most important feature. Since both statistical analysis and machine learning reveal the domain length, or analogously, size being the most important feature, we emphasize that further studies should focus on sizes of proteins when assessing evolutionary rates. The statistical analysis does not make the prediction of evolutionary trajectories possible though. Our random forest regressor enables prediction of relative structural response due to sequence changes in protein domains. This is useful to assess possible trajectories of domains and their allowed relative structural change in evolution.

Materials and Methods

Workflow and datasets

To investigate the relationship between changes in structure and sequence, the database CATH[7] was used. CATH consists of domain groupings on mainly four different levels: Class, Architecture, Topology and Homology. The homology groupings (H-groups) and folds (topology groupings) were studied here (see Figure S1 for a visual guide to the workflow).

Homology Dataset

To create a homology dataset, all entries in CATH were reduced on 95 % sequence identity using CD-HIT[23]. All domain structures which original PDB structure had been solved with X-ray crystallography and had a resolution of less than 2.6 Å were then chosen. From each family (H-group), 2-15 random entries were selected, taking the maximum possible number. These were then compared pairwise, resulting in 1-105 pairs per H-group.

For each pair, both sequence and structural alignments were performed, using HHalign (with HMMs created from HHblits)[24] and TM-align[25] respectively. This created the starting points for two workflows, one with initial sequence alignments and one with structural ones. To ensure the alignments are of sufficient quality, a reduction on the initial sequence alignments was made. The pairs with at least 80 % of the shortest sequence aligned were selected. The corresponding pairs from the structural workflow were selected, and all data were merged for comparison of sequence- and structure alignments as starting points.

In the case of initial sequence alignments, corresponding structural alignments were made by extracting aligned alpha carbons from the PDB-files. These were then superposed with TM-score[26]. DSSP[27,28] was run on the original PDB-files and secondary structure annotations and surface accessibilities were extracted, following both initial sequence- and structure alignments. From both the TM-score and TM-align superpositions, lDDT[22], rmsd, DIFFSS (see Discrete secondary structure and surface accessibility measures) and DIFFC (see Contact measures) were calculated. To calculate sequence distances (ED), tree-puzzle[29] was run on phylip files[30] created from the sequence alignments from HHalign, and sequence alignments extracted from the TM-align result[17].

Fold Dataset

In order to assess more distant evolutionary relationships, a fold dataset was created. Continuing after the PDB-filter step from the homology dataset (see Figure S1), all pairs from within the same CATH fold were selected. Using only one pair per H-group and aligning with another randomly chosen pair, a fold dataset was created. The same sequence- and structural alignment procedure as in the homology dataset then followed. Since as distant evolutionary distances as possible were sought here, no reduction on the sequence alignments, as described above, was made.

Broad Dataset
To assess a broad range of evolutionary distances, the homology dataset and the topology dataset were concatenated into the broad dataset.

Table 1. The number of pairs, groups and domains for the different datasets. Note that in the homology dataset, the groups refer to homology groupings. For the fold and broad datasets, the groups refer to topology groupings.

Dataset	Pairs	Groups	Domains
Homology	49103	2426	14636
Fold	1799	442	3598
Broad	50902	967	16729

Metrics describing the construction of these datasets, such as the percent passed in each step of the workflow, are displayed in tables S2 and S3.

Cardinality

Different amino acid groupings were made to obtain similar amounts of states in sequence and structural measures. This enables the assessment of the effect of varying cardinality. In total, four amino acid alphabets were used, including the regular 20 state one[17].

AA2: LIVFMWCPA are classified as hydrophobic (H), all other as polar (P)
AA3: LIVFMWCPA are classified as hydrophobic (H), one class consisting of P and G (helix breakers), and the remaining in one class.
AA6: the following groupings were used [KR],[DE],[YWFH],[TSQN],[CVMLIA], [PG]
AA20: the 20 standard amino acids

Tree-puzzle was run on the different cardinality versions of the aligned sequences to obtain ML distances.

Discrete secondary structure and surface accessibility measures

DSSP annotations of secondary structure were grouped accordingly:
(H)elix = G,H,I | (S)trand = E, B | (L)oop = T, S, C
The fractional identity (ID_{SS}) between aligned pairs in the ungapped parts were calculated by matching aligned secondary structure annotations:

$$ID_{SS} = \frac{(HH+SS+LL)}{(HH+SS+LL+HS+SH+HL+LH+SL+LS)}$$

where HH, SS and LL are the number of matched helix, strand and loop states and the rest unmatched states. The differing fraction of aligned secondary structure states is thus:

$DIFF_{SS} = 1\text{-}ID_{SS}$

The surface accessibility of each residue obtained from DSSP was normalized according to empirical measurements[31], with a maximum allowed value of 100 %. These relative surface areas (RSAs) were then assessed for all residues, annotating the residues with RSA> 25 % as exposed (E) and all others as buried (B). The mismatching RSA states in each aligned pair were counted and divided by the number of aligned residues, obtaining the measure $DIFF_{RSA}$.

Contact measures

All beta carbons (alpha carbons for glycine) within each domain that were more than 5 residues apart in sequence and less than 8 Å in root mean square deviation (RMSD) were selected. For each alignment, the number of matching contacts (C) between the domain structures was compared with the total number of contacts (M, N) in each of the domain structures. This represents the fraction of conserved contacts, and the fraction of unconserved contacts is thus:

$$DIFF_C = 1 - \frac{C}{M+N-C}$$

The relative contact order (RCO) of each domain was also calculated. Again classifying all beta carbons within each domain that were more than 5 residues apart in sequence, and less than 8 Å in RMSD as contacts.

$$RCO = \frac{1}{L \cdot N} \sum^{N} \Delta S_{i,j} ,$$

where L is the total number of residues in the domain, N is the total number of contacts and $\Delta S_{i,j}$ is the sequence separation in residues between each contact. The domains which have no contacts more than five residues apart will thus have an RCO of 0.

The contact density (CD) is the number of contacts per residue, which is calculated by taking the total amount of contacts N and dividing by the peptide length L.

$$CD = \frac{N}{L}$$

Running average and deviation

For the interval of 0.1 maximum likelihood amino acid distance (ED), using different cardinalities, the running average was computed for different structural scores. The same interval and procedure were used to calculate all mentioned running averages. All sequence structure-relationships are displayed using the midpoint of each ED interval, e.g. 0.05 for interval 0-0.1.

To assess the deviation from the total running average, all pairs with AA20 ED 0-6, 49042/50902 = 96 % out of all possible pairs, were taken. The deviation from the total running average for these were calculated to compare their correlation with different protein features. The features AA20 ED, % of shortest sequence aligned, Contact density, RCO, Domain Length, % AA6 groupings: [KR],[DE],[YWFH],[TSQN],[CVMLIA], [PG], % Gap, % secondary structure: Helix, Loop, Sheet were used. All features except for AA20 ED and % of shortest sequence aligned have a value for both query and subject sequences in each pair. The minimal expected deviation is computed by subtracting a perfect score (e.g. 1.0 for the lDDT score) , with the average in the first interval (ED 0-0.1).

Machine learning

A random forest regressor was trained using scikit-learn(Pedregosa et al., 2011) to predict deviation from the total running average. The same features as for the correlation analysis were used. The meaningful optimized parameters are min_samples_split = 2 and n_estimators = 500, all other parameters were set to default values. The accuracy in prediction and Pearson correlation coefficient are assessed in a five-fold cross-validation. The feature importance (permutation importance, [21]) was decomposed in five repeats on all five cross-validation splits for both the training and test-sets, to assess the most important features for the deviation prediction.

Extreme outliers

Some pairs have high structural similarity, but low sequence similarity and some have low structural similarity, but high sequence similarity. These pairs deviate from the average sequence-structure relationship. The pairs from the structural alignment workflow in the broad dataset were analyzed by creating two outlier sets, one with lDDT scores >0.9 and AA20 ED >2 (Set 1) and one with lDDT scores <0.45 and AA20 ED <2 (Set 2).

Code and software versions

All code used in this study is freely available from
https://github.com/ElofssonLab/evolutionary_rates

Acknowledgements

This work was supported by grants **VR 2016-06301 to AE.**

We want to thank Petras Kundrotas, David Menéndez Hurtado, Sudha Govindarajan, Claudio Bassot, Gabriele Pozzati and Wensi Zhu for remarks and discussions during the process of this work.

Conflicts of interest

We declare no conflicts of interest.

Supplementary Material

https://docs.google.com/document/d/1Tdfk2ULYxRmMz2rwlI--e2327FNd_org3SULKtZOCTU/edit?usp=sharing

References

1. Anfinsen CB. Principles that Govern the Folding of Protein Chains. Science. 1973. pp. 223–230. doi:10.1126/science.181.4096.223

2. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. The EMBO Journal. 1986. pp. 823–826. doi:10.1002/j.1460-2075.1986.tb04288.x

3. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991;9: 56–68.

4. Sonnhammer EL, Kahn D. Modular arrangement of proteins as inferred from analysis of homology. Protein Sci. 1994;3: 482–492.

5. Levitt M. Nature of the protein universe. Proceedings of the National Academy of Sciences. 2009. pp. 11079–11084. doi:10.1073/pnas.0905029106

6. Ekman D, Björklund AK, Frey-Skött J, Elofsson A. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. J Mol Biol. 2005;348: 231–243.

7. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2017;45: D289–D295.

8. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014;42: D304–9.

9. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014;10: e1003926.

10. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, et al. The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res. 2009;37: D310–4.

11. Burke S, Elber R. Super folds, networks, and barriers. Proteins: Structure, Function, and Bioinformatics. 2012. pp. 463–470. doi:10.1002/prot.23212

12. Gilson AI, Marshall-Christensen A, Choi J-M, Shakhnovich EI. The role of evolutionary selection in the dynamics of protein structure evolution. doi:10.1101/059741

13. Krishna SS, Grishin NV. Structural drift: a possible path to protein fold change. Bioinformatics. 2005. pp. 1308–1310. doi:10.1093/bioinformatics/bti227

14. Cuff A, Redfern OC, Greene L, Sillitoe I, Lewis T, Dibley M, et al. The CATH Hierarchy Revisited—Structural Divergence in Domain Superfamilies and the Continuity of Fold Space. Structure. 2009;17: 1051.

15. McGuffin LJ, Bryson K, Jones DT. What are the baselines for protein fold recognition? Bioinformatics. 2001;17: 63–72.

16. Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J, et al. Recognizing the fold of

a protein structure. Bioinformatics. 2003;19: 1748–1759.

17. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence-A study of structural response in protein cores. Proteins: Structure, Function, and Bioinformatics. 2009. pp. 499–508. doi:10.1002/prot.22458

18. Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. The relationship of protein conservation and sequence length. BMC Evol Biol. 2002;2: 20.

19. Bloom JD, Drummond DA, Arnold FH, Wilke CO. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol. 2006;23: 1751–1761.

20. Zhou T, Drummond DA, Wilke CO. Contact density affects protein evolutionary rate from bacteria to animals. J Mol Evol. 2008;66: 395–404.

21. Breiman L. Machine Learning. 2001. pp. 5–32. doi:10.1023/a:1010933404324

22. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29: 2722–2728.

23. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152.

24. Steinegger M, Meier M, Mirdita M, Voehringer H, Haunsberger SJ, Soeding J. HH-suite3 for fast remote homology detection and deep protein annotation. doi:10.1101/560029

25. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33: 2302–2309.

26. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004;57: 702–710.

27. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011;39: D411–9.

28. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22: 2577–2637.

29. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 2002. pp. 502–504. doi:10.1093/bioinformatics/18.3.502

30. Baum BR. PHYLIP: Phylogeny Inference Package. Version 3.2. Joel Felsenstein. The Quarterly Review of Biology. 1989. pp. 539–541. doi:10.1086/416571

31. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. PLoS One. 2013;8: e80635.