THE PROTEIN SOCIETY

Homology modeling in a dynamical world

Alexander Miguel Monzon [1], Diego Javier Zea,[2] Cristina Marino-Buslje,[2]* and Gustavo Parisi[1]*

[1]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, B1876BXD Bernal, Argentina
[2]Structural Bioinformatics Unit, Fundación Instituto Leloir, CONICET, C1405BWE Ciudad Autónoma de Buenos Aires, Argentina

Received 6 June 2017; Accepted 9 August 2017
DOI: 10.1002/pro.3274
Published online 17 August 2017 proteinscience.org

Abstract: A key concept in template-based modeling (TBM) is the high correlation between sequence and structural divergence, with the practical consequence that homologous proteins that are similar at the sequence level will also be similar at the structural level. However, conformational diversity of the native state will reduce the correlation between structural and sequence divergence, because structural variation can appear without sequence diversity. In this work, we explore the impact that conformational diversity has on the relationship between structural and sequence divergence. We find that the extent of conformational diversity can be as high as the maximum structural divergence among families. Also, as expected, conformational diversity impairs the well-established correlation between sequence and structural divergence, which is nosier than previously suggested. However, we found that this noise can be resolved using *a priori* information coming from the structure-function relationship. We show that protein families with low conformational diversity show a well-correlated relationship between sequence and structural divergence, which is severely reduced in proteins with larger conformational diversity. This lack of correlation could impair TBM results in highly dynamical proteins. Finally, we also find that the presence of order/disorder can provide useful beforehand information for better TBM performance.

Keywords: protein structure; homology modeling; protein dynamics; protein sequence; conformational diversity

Introduction

Template-based modeling (TBM) is the most reliable, accurate, and fastest approach for protein structure prediction.[1–3] TBM includes both the threading techniques and comparative modeling.[4] The accumulation of experimental structures in the Protein Data Bank (PDB) has increased the fold-space coverage,[5] which in combination with the steady enhancement of template-detection techniques over the last several years,[2] allows prediction of three-dimensional (3D)

Abbreviations: TBM, template-based modeling; CD, maximum conformational diversity of a given protein; SD, structural divergence; MSD, maximum structural divergence between two homologous proteins; PDB, Protein Data Bank; RMSD, C-alpha root mean squared deviation.

Additional Supporting Information may be found in the online version of this article.

Gustavo Parisi and Cristina Marino-Buslje contributed equally to this work.

Importance/impact: Template-based modeling (TBM) relies on the high correlation between sequence and structural divergence, and template selection is a key step to obtain a reliable 3D model. We show that the extent of conformational diversity can be as high as the maximum structural differences reached by homologous proteins. In proteins with higher flexibility, the well-established correlation between sequence and structural divergence is nosier than previously suggested due to the presence of structural change without sequence variation.

Grant sponsor: Agencia de Ciencia y Tecnología; Grant number: PICT-2014-3430; Grant sponsor: Universidad Nacional de Quilmes; Grant number: 1402/15.

*Correspondence to: Gustavo Parisi, Roque Sáez Peña 182, B1876BXD Bernal, Argentina. E-mail: gusparisi@gmail.com Cristina Marino-Buslje, Av. Patricias Argentinas 435, C1405BWE Ciudad Autónoma de Buenos Aires, Argentina. E-mail: cmb@leloir.org.ar

structures in at least 50% of the human proteome, and almost 70% for some prokaryotic proteomes using current TBM methods.[6,7] TBM relies on the fact that homologous proteins, with detectable sequence similarity, possess similar 3D structures. Pioneering work by Chothia and Lesk found that structural divergence (SD) increases with evolutionary distance, measured as percent identity, following a non-linear relationship.[8] Very similar sequences show modest structural differences, which suddenly increase when sequence identity drops below 30%. This trend was later confirmed by others;[9–11] however, in recent studies a linear correlation between sequence and SD has been found.[12–14]

Using the relationship between sequence identity and structural distance, the first step in TBM involves the search for an adequate template.[3] In view of the above-mentioned studies on structure-sequence relationships, the better template will certainly be one with the maximum sequence similarity to the target sequence in the known structural database. Evolutionary distances between target and template, the presence of ligands, and resolution are also useful guidelines for template selection.[4] This step is followed by alignment between the template and target sequence to detect conserved and variable regions. The final step of TBM is refinement through a combination of methods to render a 3D model of the target sequence. Using sequence identity as a measure of the distance between target and template sequences, it was found that structural models differ 1–2 Å C-alpha root mean squared deviation (RMSD) from a selected native structure for templates with more than 50% sequence identity. In the case of templates between 30 and 50% identity, the distance between a model and a native structure is about 4 Å RMSD, while for templates below 30% identity, template-free methods outperform TBM techniques.[1]

In spite of the outstanding contributions of TBM approaches to a great variety of fields,[15] it is still difficult to obtain high quality 3D models. Errors derived from target and template alignments,[16] along with refinement of the initial model to obtain more native-like models,[4] are among the major problems to solve in order to improve 3D model quality.[3] However, there is still a conceptual issue to face in order to improve TBM predictions. This issue is related to the nature of the "native state" of proteins, which are composed of different conformers in equilibrium, a key concept for understanding protein function.[17] In this sense, TBM techniques should progress toward a new step in its development to predict the "native state of proteins," and not simply to "predict the structure" (in terms of the alpha carbon scaffold) of a target sequence. Several authors have previously pointed out the impact of conformational diversity on TBM approaches,[18] primarily because a given template (with a determined distance to the target sequence) can have different conformers sampling a large conformational space.[13,19] A wide range of structural differences among conformers can be observed by comparing structures of the same protein obtained under different crystallization conditions. These differences result from the relative movements of large domains,[20] secondary and tertiary element rearrangements,[21] and loop movements,[22] which overall can produce a conformational diversity up to 4–5 Å of RMSD.[23–26] Even up to 15–20 Å can be observed, depending on the structural alignment algorithm used to calculate the RMSD.[18] Taking into account this extent of conformational diversity, performance of TBM methods should be re-evaluated. Blind evaluation protocols use only one conformation of the selected templates, and the performance of the resulting model depends highly on that selection.[27]

Underneath the effect of conformational diversity in TBM techniques, the more complex problem of solving how structural information is codified in the protein primary structure remains unattended.[14] The so-called local model maintains that a few positions in the protein define the global structural arrangement. Non-linear behavior in the structure-sequence relationship supports this hypothesis due to the observation that a large amount of sequence variation is required in order to dramatically change the structure (mostly below 20–25% sequence identity). On the contrary, the global model supports the idea that several positions spread along the protein define the structural arrangement. A linear relationship between structural change and sequence divergence will support this model by showing proportional change between those variables. However, considering that a single sequence can adopt several conformations, makes it even more complicated to predict how non-synonymous substitutions correlate with SD.

As a key concept to be taken into account in TBM approaches, here we explore the impact of conformational diversity on the relationship between structural and sequence divergence. To this end, a curated data set of 2024 proteins with experimentally known conformational diversity, clustered in 524 homologous families (>30% local identity and 90% coverage). These proteins cover the four main classes of the CATH database[28] with 17% mainly alpha structures, 25% mainly beta, 57% alpha beta, and 1% proteins with few secondary structures. These homologous families were analyzed to derive structural and sequence similarities between their members. We found that the use of a highly redundant sequence dataset (i.e., considering the conformational diversity) blurs the well-established relationship between sequence and structure divergence more than shown in previous studies. However, we also found that this trend could be solved using *a priori* information from the structure-function relationship. We show that families containing proteins with low conformational

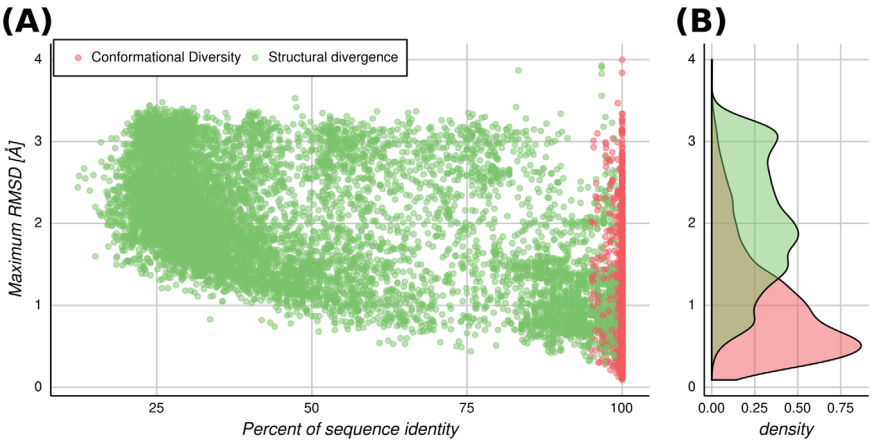

Figure 1. Maximum RMSD (MSD and CD) versus sequence percent identity. Points refer to the maximum RMSD obtained from an all versus all comparison between structures from two homologous proteins (MSD), or from the same protein (CD). (A) Green dots: comparisons between homologous protein pairs. Red dots: comparison between conformers of the same protein. (B) Distributions of the maximum RMSD between two homologous proteins (green) and between conformers of the same protein (red).

diversity, which we call "rigid" proteins, show a well-correlated behavior of sequence and SD; on the contrary, this correlation is severely reduced in protein families with larger conformational diversity. This lack of correlation could impair TBM results in highly dynamic proteins. Finally, we found that the presence of order/disorder regions could be useful prior information resulting in a better TBM performance.

Results

Protein conformational diversity can be as high as the SD in family evolution

We performed an "all against all" structural alignment within each family using MAMMOTH,[29] against our dataset of 524 protein families containing 2024 proteins with known conformational diversity, totalling 37,755 structures (at least five conformers of each protein, with approximately 19 conformers in average) extracted from the CoDNaS database.[23] For each pair of homologous proteins within each family, the percent sequence identity was calculated, aligning each pair of sequences with the Needleman-Wunch algorithm.[30] Since each protein is represented by an ensemble of conformers, the maximum RMSD derived from an "all versus all" comparison of conformers, belonging to the homologous proteins being compared, is called maximum SD (hereafter MSD for simplicity). When this procedure is repeated between the pairs of conformers (different structures of the same protein, see Methods section), the maximum conformational diversity (hereafter CD for simplicity) is obtained, measured as the maximum RMSD of a given protein (see schematic protocol in Fig. S1, Supporting Information). In Figure S2, the plot contains all comparisons between conformers for each pair of homologous protein being studied, containing approximately 3.5×10^6 dots. Figure 1(A) shows the relationship between the MSD (green points) versus percent sequence identity. Figure 1(A) also shows the CD values (red dots). Green dots show the typical behavior previously found between sequence and structure divergence[8,14] with a steep decrease of structural similarity at low identity percent (below approximately 30%). However, our dataset also shows high SD at very high sequence similarities, as a consequence of proteins with conformational diversity as high as the SD of the family.

The distribution of the CD shows [Fig. 1(B)] mostly moderated RMSD values, with an average of 1 Å. Nevertheless, it also shows high positive skewness, toward larger RMSD values. This is in concordance with previous work.[18] The 90th percentile of the analyzed proteins show a CD below 2 Å of RMSD, and then 10% of the proteins can exist in a conformational space as large as the MSD, coming from comparison of remote homologous proteins (approximately 3 Å). This is an interesting result that indicates that a given sequence can potentially exist in a conformational space as big as the SD that arose from the accumulation of substitutions, namely the evolutionary process. In the light of the conformational diversity, it is easy to understand that closely homologous proteins (suppose above 80% sequence identity) can have either high or low RMSD values when superposing their structures, depending on the particular conformers being compared. So, conformational diversity can lead to large RMSD values between proteins over short evolutionary time periods, instead of reaching these RMSD values through the long process of accumulation of sequence mutations. Besides RMSD, two other measures of SD were estimated. The fraction of unconserved secondary structure (SS) and the relative solvent accessibility category (RSA) (see Methods section) are shown in Figures S3 and S4, respectively. It is possible to see that both parameters follow the general trend observed in Figure 1(A) for the RMSD.

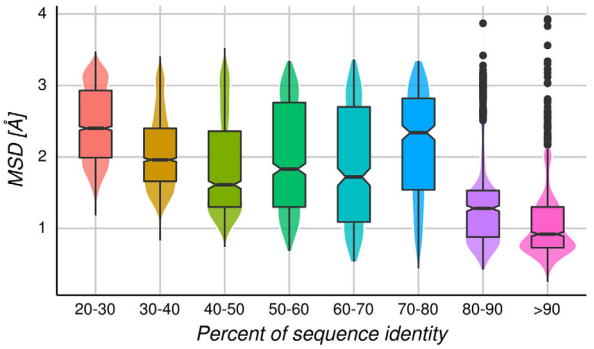

Figure 2. MSD distributions over all homologous protein pairs by bins of 10% sequence identity. Above and below the median (horizontal line inside each box) are the first and third quartiles, respectively. The notches display the median absolute deviation.

Again, it is interesting to observe the high dispersion of RSA and secondary structural variation at high values of sequence identity.

The other important consequence of this finding is that the correlation between sequence and SD is weaker than stated in previous works.[12–14] In other words, due to the CD, a given sequence can adopt different conformations, so the structural change due to non-synonymous substitutions in a divergent evolution process will make the relationship between sequence and structure noisier. In fact, the distribution of the MSD [green dots in Fig. 1(A)] and sequence identity, have a Spearman's rank correlation rho of −0.52. The relationship between sequence and structure will be visible in light of the conformational diversity, as explained below.

Template selection and structural diversity

As mentioned before, TMB approaches require the use of a protein with known structure as the template. This identification can be performed using a broad variety of techniques with different sensitivities.[3,31,32] A key point in this step is selection of the best template, which is based upon the commonly used relationship between structural and sequence divergence,[8] and the one maximizing both coverage and percent sequence identity against the target.[33] However, as we can see from Figure 1, this criterion is not as simple as previously stablished. In Figure 2, we show the MSD distribution among bins of 10% sequence identity between pairs of homologous proteins. It is possible to observe the great variation in RMSD for each particular bin. More importantly, the maximum RMSD value is almost equal for all the considered bins of percent identity (mean = 3.54 Å and standard deviation = 0.23 Å). Therefore, template selection is not as straightforward as just selecting a structure in a given identity bin, because it is not known how those structures belong to the conformational ensemble of the sequence to be modeled.[27]

However, the distributions per bin shown in Figure 2 could be influenced by the presence of a given protein family with exceptionally large or small structural diversity. For that reason, in Figure 3 we show the average MSD per protein family in a bin of sequence identity. The averaged RMSD values are between 1.34 and 2.10 Å for the different bins, with standard deviations between 0.19 and 0.29, showing that the dispersion is not related to the sequence identity. In Figure S5, we show that the average MSD does not depend on how populated (amount of homologous proteins) the corresponding family is.

Taking these results into account, the selection of an adequate template will depend heavily on the target protein, whether it has a high or a small conformational diversity. Which then will be the general recommendations for selecting a good template? The relationship between SD and CD will give us a clue.

How does SD correlate with conformational diversity?

In a simple evolutionary scenario of the evolution of a protein family, we can consider that CD is a conserved trait among its members and that the common ancestor of the family shows low conformational diversity at the backbone level. If all the conformers belonging to this common ancestor are structurally aligned, the resulting RMSD values would be about 0.5 Å (a value equivalent to the crystallographic error), meaning that the conformer population is almost identical. It is important to say that these conformers are structurally equivalent at the backbone level, as the RMSD is measured using the alpha carbons, but conformational differences at the residue level cannot be discarded.[34,35] Considering that this family has a selective pressure to maintain its conformational diversity in most of their proteins (i.e., due to functional restrictions[36]), most of the SD of this family would have originated by the accumulation of nonsynonymous substitutions. On the

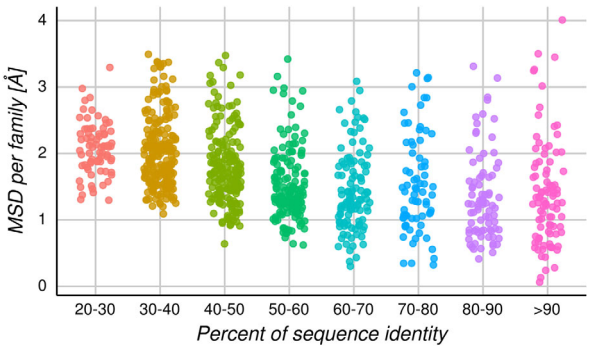

Figure 3. MSD distributions over bins of 10% sequence identity per family. Each dot represents the average MSD for all homologous protein pairs per family in a given bin of percent identity. It is possible to see a great dispersion of SD even at low percent identity and that the different families spread approximately 2.9 Å RMSD in average per bin.

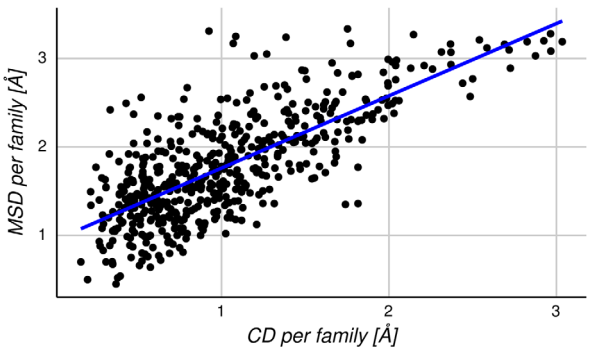

Figure 4. Relationship between MSD and CD. Each dot represents the average RMSD values for the MSD and the CD in a specific family. The data show a Pearson's correlation coefficient of 0.75.

contrary, in the case that a family originated with a common ancestor showing large conformational diversity (i.e., a RMSD of approximately 2 Å), the process of divergence due to the accumulation of nonsynonymous substitutions will certainly increase the available repertory of conformations and eventually increase the SD. We have found that the dispersion of the CD extent is rather low, possibly indicating that the CD can be a trait conserved within families (see Fig. S6). However, more work needs to be done to address the question of how the conservation of the conformational diversity inside a protein family is.

It is our central hypothesis that the maximum conformational diversity of a protein will correlate with the MSD that can be reached by that family. To probe this hypothesis, the average of the CD and the MSD per family was calculated, and a Pearson's correlation coefficient of 0.75 (P value <0.01) was observed (see Fig. 4).

As MSD could change proportionally with the sequence divergence of each family, in Figure S7 we show that the association is independent of the sequence divergence found in each family.

In the light of the results shown in Figure 4, we study the relationship between structural and sequence divergence by splitting the dataset into homologous protein pairs with large and small conformational diversity ($\geq$0.5 and $<$0.5 Å RMSD, respectively, obtained as the average of CD between each pair of homologous proteins). This threshold is near the crystallographic error (approximately 0.5 Å[37]). It is interesting to note that the Spearman's rank correlation rho between structure (MSD) and sequence divergence (sequence identity percent) is -0.83 (with a significant P value <0.01) in the subset of protein families with small conformational diversity per homologous protein pair, and -0.51 (P value <0.01) in the subset of protein families with large conformational diversity (see Fig. 5). These results indicate that the known correlation between sequence and structure[8] is strong in the subset of protein families with low conformational diversity. In these families, the biological function can be achieved with conformers almost identical at the backbone level. That will make the relationship between sequence and structure straightforward, namely, the change in structure is proportional to the observed change at the sequence level. In the opposite cases, for the subset of families with high conformational diversity, two scenarios are possible: either the biological function is less tight with a single conformation, or inversely, the function requires high plasticity of the structure. In this sense, the subset with conformational diversity below 0.5 Å RMSD, sequence versus structure divergence has a linear and exponential fit with R^2 values of 0.54 ± 0.15 and 0.66 ± 0.12, respectively. While for the subset with conformational diversity above 0.5 Å, RMSD has a linear and exponential fit with R^2 values of 0.23 ± 0.21 and 0.29 ± 0.18, respectively. These R^2 values are the mean obtained for testing datasets in a fivefold cross validation. We also found that splitting the distribution in bins of CD, these correlation coefficients change monotonically as the CD increases (Table S1). RMSD seems to be more sensitive to CD change than other parameters showed above (unconserved relative solvent accessibility category (RSA) and secondary structure) (Table S1).

Accordingly to these results, TBM approaches will be much more reliable in protein families with low conformational diversity because the expected change in structure is proportional to the sequence divergence. In these families, where selecting the template as the one showing the highest sequence similarity and coverage, will increase the reliability of TBM. As we can see in Figure 5(A), both linear and exponential regressions give a RMSD approximately 0.45 for 100% sequence identity. On the contrary, in families with a larger CD, that relationship loses predictability due to the observed structural variability for the same sequence (approximately 1.3 RMSD for both linear and exponential regressions at 100% sequence identity). Differences between linear and exponential fitness for the sets below and above 0.5 Å RMSD are meaningless because they could be associated with different causes (such as intrafamily variations or RMSD non-linear dependence with protein length among others). However, regression intercepts are informative about structural similarity dispersion at high sequence identities.

How can we turn these findings into practical advice for use in TBM methods? It is very difficult to know the conformational diversity of the target sequence to be modeled by TBM protocols before starting. However, our previous work shows that proteins with disordered regions have larger conformational diversity compared with ordered proteins, on average.[38] In the section "How does protein disorder correlate with SD?," we address this question: How is protein disorder related to SD?

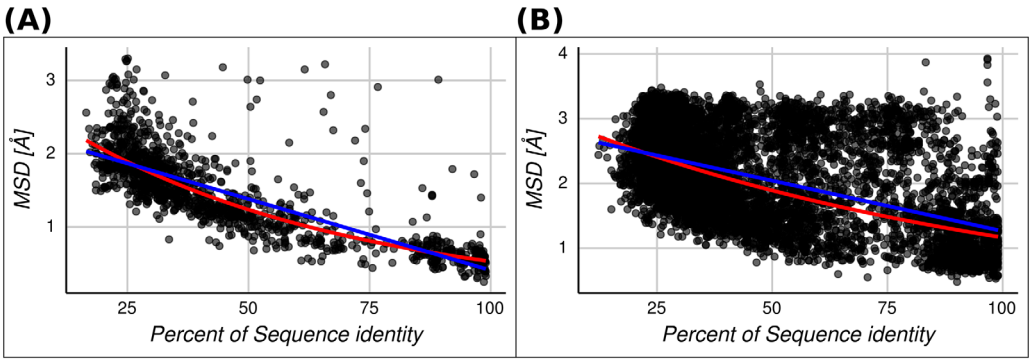

Figure 5. MSD versus percent of sequence identity for each homologous protein pair. The lineal (blue line) and exponential (red line) regressions are shown for two sets. (A) Homologous protein pairs with an average conformational diversity less or equal than 0.5 Å. The linear and exponential fitted expressions are RMSD= 2.354 − 0.019 SEQID and RMSD= $e^{1.06}e^{-0.017\ SEQID}$, respectively. (B) Homologous protein pairs with an average conformational diversity greater than 0.5 Å. The linear and exponential fitted expressions are RMSD= 2.823 − 0.015 SEQID and RMSD= $e^{1.121}e^{-0.010\ SEQID}$, respectively.

How does protein disorder correlate with SD?

Disordered regions in proteins are known to be involved in several important biological functions.[39,40] Intrinsically disordered regions (IDRs) or proteins (IDPs) are characterized by their high flexibility and mobility, displayed as missing regions in crystallographic structures.[41] It is difficult to estimate the extent of flexibility in the disordered regions, but it is possible to measure conformational diversity in the ordered regions of these proteins using the RMSD between different conformers, for example.[42] We found that proteins containing IDRs have larger conformational diversity than those with full ordered structures, when disorder-order transitions take place between protein conformations.[38] Furthermore, we recently found that proteins with IDRs can be split into two groups with different structure-function relationships, depending on how structure-based features change among the available conformer population for each protein.[43] Therefore, it is interesting to ask if the structure-sequence relationship could be also separated into two groups, mainly proteins with and without IDRs, following the above-mentioned results using CD. We find that pairs of homologous proteins containing at least one disordered region (in any of their available conformers) show higher values of MSD than the population of ordered homologous proteins (see Fig. 6). These distributions were found to be statistically different using the Wilcoxon and Kolmogorov Smirnov test with P value <0.01.

Since homologous pairs containing disordered regions have higher MSDs, we expect that the correlation between structure and identity percent would be worse, as shown above (Fig. 5). We found that the Spearman's rank correlation rho is −0.36 and −0.58 for disordered and ordered pairs of homologous proteins, respectively. These results show that the presence of disordered regions in the template and/or in the target sequence could predict a large CD and could make the relationship between sequence and structure less predictable. Although this high correlation value between full ordered proteins and MSD, using both linear and exponential fits still produces high MSD dispersions at high percent identity (Fig. S8). However, we know that there are few full-ordered protein families with very large CDs. These proteins have been extensively studied thanks to the pioneering work of Chothia and coworkers[20,21,44] and represent less than approximately 20% of our dataset.[43] Removing these highly dynamic proteins with non-disordered regions, we obtain the relationship shown in Figure 7 where the Spearman's rank correlation rho is −0.71 for the ordered set of pairs.

Based on these correlation coefficients, we can say that the presence of disorder regions alone has a moderate capacity for predicting the presence of noise between structure and sequence variables. Moreover, these values were obtained considering that most ordered proteins do not have CD, or at

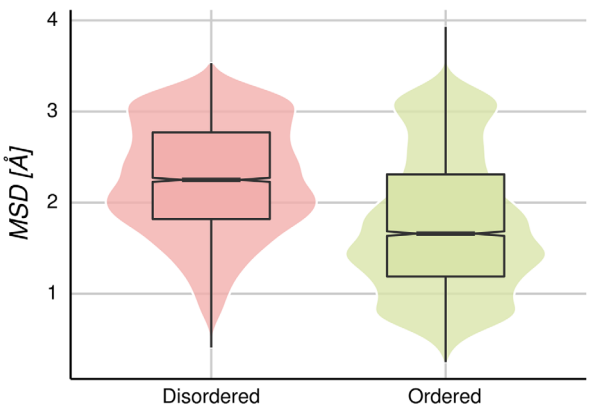

Figure 6. MSD distributions in homologous protein pairs. The disordered set (4439 homologous protein pairs) has proteins with at least one conformer with IDRs, while the ordered set (5034 homologous protein pairs) has no proteins with IDRs, in any of the conformers.

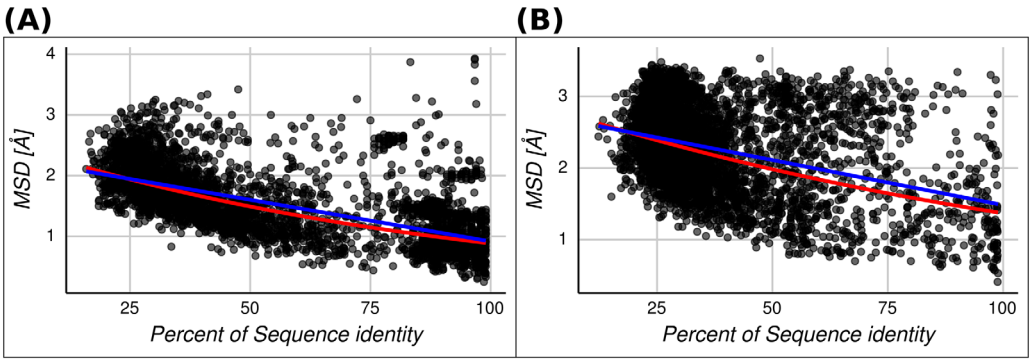

Figure 7. MSD versus percent of sequence identity for each homologous protein pair. The lineal (blue line) and exponential (red line) regressions are shown for two sets. (A) Homologous protein pairs containing just ordered conformers. The linear and exponential fitted expressions are RMSD= 2.288 − 0.014 SEQID and RMSD= $e^{0.092}e^{-0.01\ SEQID}$, respectively. (B) Homologous protein pairs with at least one of the conformers containing disordered regions. The linear and exponential fitted expressions are RMSD= 2.741 − 0.012 SEQID and RMSD= $e^{1.055}e^{-0.007\ SEQID}$, respectively.

least show moderate CD, as we have previously described.[43] Taking into account these considerations, and the easy and reliable prediction capacity for disordered regions in proteins, this information could be still used as guidance in TBM approaches.

Discussion

The study of structure-sequence relationships embraces a foundational concept for several areas focused on the study of proteins. Biological function prediction,[45] protein evolution,[46] structural proteomics,[47] and homology modeling[3] are just a few examples of the broad and active research areas that take advantage of that relationship. Among all the situations that structure-sequence could adopt, we have focused on how the presence of conformational diversity in proteins could influence the relationship between sequence and structural change and therefore affect TBM approaches. A central point was derived early by Chothia and Lesk, showing that the success of 3D prediction will depend on the extent of the target sequence identity with the corresponding template.[8] Basically, the behavior between RMSD and percent identity established a relationship in which structural similarity increased as sequence similarity increased. Their results, and conclusions, have been verified by numerous studies.[9,10,12,13,48–50] These studies found moderate-to-high correlation coefficients between different parameters that were proportional to structural and sequence similarity, that is, RMSD and percent identity,[10,50] evolutionary distance,[13] and statistical significance of RMSD.[14] They also found linear and nonlinear behavior, and an invariably low structural variation, at 100% identity (approximately 0.5 Å). Departure from linear fitness has been explained as being derived from errors in the alignments (at high and low sequence similarity), use of redundant datasets, or by ignoring the multiple substitutions per site during evolution.[13,14,48]

In this work, we found that the extent of the CD is related to the MSD of the family, and that the structure-sequence relationship is more complex than previously thought. First, we found that the extent of the CD could be as large as the MSD (Fig. 1). Conformational diversity is a key concept to understand many processes and mechanisms in protein function, such as enzyme catalysis,[51] promiscuity in protein interactions,[52] protein-protein recognition,[53] signal transduction,[54] mechanisms of disease-related mutations,[55] immune escape,[56] the origin of neurodegenerative diseases,[57] protein evolutionary rates,[58] conformer-specific substitution patterns,[59] the origins of new biological functions,[60] molecular motors,[61] and co-evolutionary measurements between residues[24,62] (for a recent review please see[17]). Furthermore, we have recently shown that the distribution of CD in a large dataset of proteins, with experimentally determined CD (approximately 5000 proteins), results in three main groups of proteins with different structure-function relationships.[43] More recently, we found that the dynamical behavior in a given family could change with minor sequence variations making difficult to predict CD by homology.[63]

Distributions in Figure 1 show that CD could be as large as the MSD, but it is also evident that most of the proteins in our dataset have modest to low CD, meaning that they could function with very low or absent backbone movements.[43,64,65]

Several studies drew attention to the importance of CD in TBM methods,[18,19,27] but the consideration of CD in the study of structure-sequence relationships, as an essential ingredient in TBM methods, was often considered a source of bias or "noise."[13,14] On one side, avoiding the noise introduced by the use of redundant data (i.e., considering CD) would allow us to assume that structural changes would be proportional to sequences changes.

The existence of a single sequence with multiple conformations defines a "degeneracy" of the structural information coded in a given sequence, introducing a nonlinear behavior in the protein space. It was also found that this nonlinearity could possibly impair the performance of knowledge-based methods in bioinformatics.[66]

However, taking into account the remarkable importance of CD for explaining biological processes and protein behavior, it appears impossible to ignore. As derived from Figure 2, when expressed as an average between all families, there is a large dispersion in RMSD even at high sequence identities but interestingly, this dispersion remains at the family level (Fig. 3). These figures show the large uncertainty in template selection even at high sequence similarities. We have also found that CD is proportional to the MSD reached in a family (Fig. 4). Indeed, CD is independent of the maximum sequence divergence of the family (Fig. S5). It is at this point that the total set used in this work could be split in two sets of reduced and large CD impact. In proteins evolving under selective pressure to maintain a reduced CD, we find that the correlation between sequence and structure variation is high (Spearman's rank correlation rho = −0.83). We have previously characterized this group of proteins.[43] Briefly, rigid proteins have low CD (approximately 0.8 Å RMSD in average), they are mainly proteins without disordered regions and have important tunnels and cavities. When conformers of rigid proteins are compared, for example, in their bound and unbound states, we find that they mainly differ in backbone positions associated with their tunnels and cavities. This indicates that the minimum movements for rigid proteins are associated with the movement of functional structures to allow the transit of substrates and/or products between the inside and the surface of the protein.[67–69] It is for this group of proteins that sequence-structure relationships show a high correlation between variables, and for whom it would be possible to reliably predict 3D models using TMB techniques [see Fig. 5(A)].

According to our results, proteins with higher CD also have larger SDs (Fig. 4), and a larger variability of accessible structures, even at high percentages of sequence identity [Fig. 5(B)]. We found that correlations and proportionalities between variables in this group of proteins are low, blurring the common idea in TBM approaches that similar sequences show similar structures. For example, the expected value at 100% identity using the linear regression gives approximately 1.3 Å.

Our results indicate that the relative success of the 3D model using TBM approaches will be strongly associated with the CD of the corresponding target protein. Because it is difficult to predict the extent of CD in a given protein, we used the presence of disordered regions as an indicator of higher values of CD, based upon previous results.[38] We found that the presence of highly ordered proteins (without any disordered regions in any conformer) in pairs of homologous proteins have a Spearman's correlation rho of −0.58 for RMSD and a sequence identity relationship. In this way, we found that presence of disorder/order is not a strong indicator of a well-correlated sequence and structural change (Fig. 7), compared with the knowledge of the extent of CD (Spearman's correlation rho of −0.83, Fig. 5). Removing ordered and highly dynamic proteins, we found a better Spearman's correlation rho of −0.71. This increment in reliability again confirms the higher correlation between sequence and SD for rigid proteins with low CD. Based on the many and reliable predictor methods for detecting disordered regions in proteins,[70] and the above mentioned considerations, order/disorder could be an easy way for evaluating the expected dispersion of RMSD for a given sequence similarity between template and target sequences. Alternatively, since the CD of some proteins is well correlated with the MSD of the family (Fig. 4), comparing all the known structures of the family can predict the expected flexibility of our target. However, further studies and experimental data are required in order to address the question of how well CD is conserved through evolution.

In summary, sequence and structure divergence is a more complex process than previously thought. Protein conformational diversity challenges the ordered and well-accepted relationship between sequence and structural similarity, a cornerstone of TBM techniques, as well as our understanding of the nature of the protein folding code. Further work is needed to deepen our current knowledge in such a basic topic for many areas associated with the study of proteins, as well as to encourage a reappraisal of current methods for obtaining and evaluating 3D protein models.

Materials and Methods

Protein families with conformational diversity selection

The CoDNaS database,[23] containing a redundant collection of three dimensional structures for the same protein (at least 95% of sequence identity among structures to include putative sequence variations), was used to recruit proteins exhibiting conformational diversity. All structures belonging to this dataset were obtained by x-ray diffraction at a resolution equal or less than 2.5 Å. The CD is the maximum C-alpha RMSD derived from all conformer pairwise comparisons. With the aim to obtain a reliable and comparable estimation of conformational diversity of each protein, our dataset only contains proteins with a minimum of

five conformers (average approximately 19 conformers per protein) as was previously suggested.[71]

In order to identify homologous proteins, we ran BLASTClust[72] to obtain all available clusters at least 30% of local sequence identity with a minimum coverage of 90% between all sequences in the cluster. The PDB SEQRES records were used to extract the sequences and to perform the clustering. We only considered those clusters with at least two different proteins using the UniProt ID for identifying each protein. The sequences in our dataset are close homologous proteins (at least 30% local sequence identity and 90% coverage). Considering these restrictions in the dataset building we do not expect to have a large influence by unaligned loop regions or gaps (the average coverage of structurally aligned residues is 89% with 14% standard deviation).

The final dataset contains 2024 different protein chains with a total of approximately 38,000 conformers (which represents about 25% of the total proteins included in the version used of the CoDNaS database). These proteins are grouped in 524 families with an average of approximately 4 proteins per family (with a minimum of 2 and a maximum of 61) and according to the classification of the CATH database represent 250 different folds. The average length of the proteins are 283 residues with a standard deviation of 141, and the extension of percent identity ranges from 20 to 98% with a median of approximately 47%.

Sequence and structure comparisons

To estimate the SD for each homologous protein pairs in a cluster, we calculated the C-alpha RMSD using MAMMOTH[29] for all possible pairs of conformers belonging to the proteins being compared. MAMMOTH is a sequence-independent structural alignment program, which not only has a very good accuracy aligning proteins with different folds, but also provided the statistical reliability of the resulted structural alignment. Also, the RMSD values calculated by MAMMOTH neither show dependence with protein size nor length.

The MSD for a pair of homologous proteins is the pair with maximum RMSD value among all versus all conformers comparisons between them. Additionally, we calculated the percent sequence identity for each homologous protein pairs using a global sequence alignment obtained with the Needleman-Wunch algorithm.[30] Furthermore, we defined disordered regions for each conformer when it has five or more consecutive missing residues that were not in the amino or carboxyl terminal of the protein sequence (the first or least twenty residues). If a residue has missing electron density coordinates in a structure obtained with x-ray crystallography, it assumed to be disordered.[73] If a protein has at least one conformer with disordered regions we classified it as disordered.

The total comparisons among all versus all conformers for each homologous protein pairs and structures of the same protein give an amount of approximately 3.5×10^6 of pairs.

In addition, we calculated the fraction of secondary structure variation between each homologous protein pairs of MSD and conformer pairs of CD. We assigned the secondary structure elements to each structure using DSSP.[74] The eight states in DSSP were grouped into three states: alpha-helices (H and G), strand (E) and loops (T, S, B, I, ′ ′) as in previous works.[75] We compared the secondary structure of the corresponding aligned positions in the pair of structures and calculated the fractional identity of secondary states as:[13]

$$\text{Identity}_{SS} = \frac{(HH+SS+LL)}{(HH+SS+LL+\ SH+HS+LH+HL+LS+SL)},$$

where HH, SS, and LL are the number of aligned residues in both structures with helix (H), strand (S), and loop (L) states, respectively. HS and SH are residues in helix aligned with residues in strand, the same is for the other SS combinations. In order to estimate the local differences between the structures of the pair, we calculated the fraction of unconserved aligned secondary structure elements as follows: $\text{diff}_{SS} = 1 - \text{Identity}_{SS}$. Additionally, we used Naccess 2.1.1[76] with a probe size of 1.4 Å to obtain the relative solvent accessible area (RSA) for each pair of aligned residues in the pairs of MSD and DC. According to RSA values we classified each residue into two categories: buried (B) for RSA $\leq$25% and exposed (E). Using the same methodology explained before, we compared the RSA classification of the corresponding aligned positions in the pair of structures and calculated the fractional identity of RSA. Then, we used the difference to obtain the fraction of residues that changed between buried and exposed.

Statistical analysis

The correlations coefficients showed in this work were obtained using the function *cor.test* of R package,[77] with the corresponding two-way test (null hypothesis is that the correlation coefficient is equal to 0). We used Spearman rank correlations coefficients since it does not assume linearity of the data, only searches for a monotonic relationship.

Cross-validated function fitting were performed using the Julia language libraries LsqFit and MLBase. All the regressions were weighted, and each point (MSD vs. sequence identity percent) has a weight of 1 over the number of protein pairs in the protein family. It was done in order to avoid that populated families predominate the results.

However, the R^2 informed was calculated as the mean of the R^2 in each testing subset from a fivefold cross-validation without using weights. In that way, we use the amount of unweighted variation explained by the model as a measure of goodness of fit.

Acknowledgments

GP and CMB are researchers of CONICET and AMM and DZ are PhD and Postdoctoral fellows of the same institution.

Additional Information

Competing financial interests

The author(s) declare no competing financial interests.

Author contributions

Conceived and designed the experiments: AMM, CMB, and GP. Performed the experiments: AMM. Analyzed the data: AMM, GP, CMB, and DZ. Contributed reagents/materials/analysis tools: AMM and DZ. Wrote the paper: GP, CMB, AMM, and DZ.

References

1. Baker D (2001) Protein structure prediction and structural genomics. Science 294:93–96.
2. Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18:342–348.
3. Qu X, Swanson R, Day R, Tsai J (2009) A guide to template based structure prediction. Curr Protein Pept Sci 10:270–285.
4. Fiser A (2010) Template-based protein structure modeling. Methods Mol Biol 673:73–94.
5. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci USA 111:3733–3738.
6. Schwede T (2013) Protein modeling: what happened to the "protein structure gap"? Structure 21:1531–1540.
7. Anand P, Sankaran S, Mukherjee S, Yeturu K, Laskowski R, Bhardwaj A, Bhagavat R, Consortium OSDD, Brahmachari SK, Chandra N (2011) Structural annotation of Mycobacterium tuberculosis proteome. PLoS One 6:e27044.
8. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826.
9. Flores TP, Orengo CA, Moss DS, Thornton JM (1993) Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 2:1811–1826.
10. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 269:423–439.
11. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297:233–249.
12. Panchenko AR, Wolf YI, Panchenko LA, Madej T (2005) Evolutionary plasticity of protein families: coupling between sequence and structure variation. Proteins 61:535–544.
13. Illergård K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins 77:499–508.
14. Wood TC, Pearson WR (1999) Evolution of protein sequences and structures. J Mol Biol 291:977–995.
15. Zhang Y (2009) Protein structure prediction: when is it useful?. Curr Opin Struct Biol 19:145–155.
16. Kopp J, Schwede T (2004) Automated protein structure homology modeling: a progress report. Pharmacogenomics 5:405–416.
17. Wei G, Xi W, Nussinov R, Ma B (2016) Protein ensembles: How does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell. Chem Rev 116:6516–6551.
18. Burra PV, Zhang Y, Godzik A, Stec B (2009) Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. Proc Natl Acad Sci USA 106: 10505–10510.
19. Kosloff M, Kolodny R (2008) Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins 71:891–902.
20. Gerstein M, Lesk AM, Chothia C (1994) Structural mechanisms for domain movements in proteins. Biochemistry 33:6739–6749.
21. Gerstein M, Krebs W (1998) A database of macromolecular motions. Nucleic Acids Res 26:4280–4290.
22. Gu Y, Li D-W, Brüschweiler R (2015) Decoding the mobility and time scales of protein loops. J Chem Theory Comput 11:1308–1314.
23. Monzon AM, Rohr CO, Fornasari MS, Parisi G (2016) CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. Database [Internet]. https://doi.org/10.1093/database/baw038. [Accessed 5 July 2016].
24. Parisi G, Zea DJ, Monzon AM, Marino-Buslje C (2015) Conformational diversity and the emergence of sequence signatures during evolution. Curr Opin Struct Biol 32:58–65.
25. Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A (2016) PDBFlex: exploring flexibility in protein structures. Nucleic Acids Res 44:D423–D428.
26. Monzon AM, Juritz E, Fornasari MS, Parisi G (2013) CoDNaS: a database of conformational diversity in the native state of proteins. Bioinformatics 29:2512–2514.
27. Palopoli N, Monzon AM, Parisi G, Fornasari MS (2016) Addressing the role of conformational diversity in protein structure prediction. PLoS One 11:e0154923.
28. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, Lehtinen S, Studer RA, Thornton J, Orengo CA. (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43:D376–D381.
29. Ortiz AR, Strauss CEM, Olmea O (2009) MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Sci 11:2606–2621.
30. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453.
31. Yan R, Xu D, Yang J, Walker S, Zhang Y (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep 3:2619.
32. Xiang Z (2006) Advances in homology protein structure modeling. Curr Protein Pept Sci 7:217–227.
33. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94.

34. Eyal E, Gerzon S, Potapov V, Edelman M, Sobolev V (2005) The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J Mol Biol 351:431–442.
35. Daily MD, Gray JJ (2007) Local motions in a benchmark of allosteric proteins. Proteins 67:385–399.
36. Ma B, Nussinov R (2016) Protein dynamics: Conformational footprints. Nat Chem Biol 12:890–891.
37. Kuriyan J, Karplus M, Petsko GA (1987) Estimation of uncertainties in X-ray refinement results by use of perturbed structures. Proteins 2:1–12.
38. Zea DJ, Monzon AM, Gonzalez C, Fornasari MS, Tosatto SCE, Parisi G (2016) Disorder transitions and conformational diversity cooperatively modulate biological function in proteins. Protein Sci 25:1138–1146.
39. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res 6:1882–1898.
40. Oldfield CJ, Keith Dunker A (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem 83:553–584.
41. Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27:527–533.
42. DeForte S, Uversky VN (2016) Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. Protein Sci 25:676–688.
43. Monzon AM, Zea DJ, Fornasari MS, Saldaño TE, Fernandez-Alberti S, Tosatto SCE, Parisi G (2017) Conformational diversity analysis reveals three functional mechanisms in proteins. PLoS Comput Biol 13: e1005398.
44. Lesk AM, Chothia C (1984) Mechanisms of domain closure in proteins. J Mol Biol 174:175–191.
45. Sadowski MI, Jones DT (2009) The sequence-structure relationship and protein function prediction. Curr Opin Struct Biol 19:357–362.
46. Xia Y, Levitt M (2004) Simulating protein evolution in sequence and structure space. Curr Opin Struct Biol 14:202–207.
47. Drew K, Winters P, Butterfoss GL, Berstis V, Uplinger K, Armstrong J, Riffle M, Schweighofer E, Bovermann B, Goodlett DR, Davis TN, Shasha D, Malmström L, Bonneau R (2011) The Proteome Folding Project: proteome-scale prediction of structure and function. Genome Res 21:1981–1994.
48. Koehl P, Levitt M (2002) Sequence variations within protein families are linearly related to structural variations. J Mol Biol 323:551–562.
49. Hubbard TJ, Blundell TL (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Eng 1: 159–171.
50. Russell RB, Barton GJ (1994) Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. J Mol Biol 244: 332–350.
51. Henzler-Wildman KA, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, Pozharski E, Wilson MA, Petsko GA, Karplus M, Hübner CG, Kern D (2007) Intrinsic motions along an enzymatic reaction trajectory. Nature 450:838–844.
52. Khersonsky O, Roodveldt C, Tawfik DS (2006) Enzyme promiscuity: evolutionary and mechanistic aspects, Current Opinion in Chemical Biology, pp 498–508. http://www.sciencedirect.com/science/article/pii/S1367593106001189
53. Yogurtcu ON, Erdemli SB, Nussinov R, Turkay M, Keskin O (2008) Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations. Biophys J 94:3475–3485.
54. Nussinov R, Ma B (2012) Protein dynamics and conformational selection in bidirectional signal transduction. BMC Biol 10:2.
55. Juritz E, Fornasari MS, Martelli PL, Fariselli P, Casadio R, Parisi G (2012) On the effect of protein conformation diversity in discriminating among neutral and disease related single amino acid substitutions. BMC Genomics 13: S5.
56. Sethi A, Tian J, Derdeyn CA, Korber B, Gnanakaran S (2013) A mechanistic understanding of allosteric immune escape pathways in the HIV-1 envelope glycoprotein. PLoS Comput Biol 9:e1003046.
57. Kopito RR, Ron D (2000) Conformational disease. Nat Cell Biol 2:E207–E209.
58. Javier Zea D, Miguel Monzon A, Fornasari MS, Marino-Buslje C, Parisi G (2013) Protein conformational diversity correlates with evolutionary rate. Mol Biol Evol 30:1500–1503.
59. Juritz E, Palopoli N, Fornasari MS, Fernandez-Alberti S, Parisi G (2013) Protein conformational diversity modulates sequence divergence. Mol Biol Evol 30:79–87.
60. Tokuriki N, Tawfik DS (2009) Protein dynamism and evolvability. Science 324:203–207.
61. Dos Santos HG, Abia D, Janowski R, Mortuza G, Bertero MG, Boutin M, Guarín N, Méndez-Giraldez R, Nuñez A, Pedrero JG, Redondo P, Sanz M, Speroni S, Teichert F, Bruix M, Carazo JM, Gonzalez C, Reina J, Valpuesta JM, Vernos I, Zabala JC, Montoya G, Coll M, Bastolla U, Serrano L (2013) Structure and non-structure of centrosomal proteins. PLoS One 8: e62633.
62. Sfriso P, Duran-Frigola M, Mosca R, Emperador A, Aloy P, Orozco M (2016) Residues coevolution guides the systematic identification of alternative functional conformations in proteins. Structure 24:116–126.
63. Marino-Buslje C, Monzon AM, Zea DJ, Fornasari MS, Parisi G (2017) On the dynamical incompleteness of the Protein Data Bank. Brief. Bioinformatics [Internet]. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx084. [Accessed 15 August 2017]
64. Gutteridge A, Thornton J (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. J Mol Biol 346:21–28.
65. Mesecar AD, Stoddard BL, Koshland DE, Jr (1997) Orbital steering in the catalytic power of enzymes: Small structural changes with large catalytic consequences. Science 277:202–206[PAGE #S].
66. Rackovsky S (2015) Nonlinearities in protein space limit the utility of informatics in protein biophysics. Proteins 83:1923–1928.
67. Chovancova E, Pavelka A, Benes P, Strnad O, Brezovsky J, Kozlikova B, Gora A, Sustr V, Klvana M, Medek P, Biedermannova L, Sochor J, Damborsky J (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput Biol 8:e1002708.
68. Zhou HX, Wlodek ST, McCammon JA (1998) Conformation gating as a mechanism for enzyme specificity. Proc Natl Acad Sci USA 95:9280–9283.
69. Biedermannová L, Prokop Z, Gora A, Chovancová E, Kovács M, Damborsky J, Wade RC (2012) A single mutation in a tunnel to the active site changes the mechanism and kinetics of product release in

haloalkane dehalogenase LinB. J Biol Chem 287: 29062–29074.
70. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK (2009) Predicting intrinsic disorder in proteins: an overview. Cell Res 19:929–949.
71. Best RB, Lindorff-Larsen K, DePristo MA, Vendruscolo M (2006) Relation between native ensembles and experimental structures of proteins. Proc Natl Acad Sci USA 103:10901–10906.
72. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410.
73. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, et al. (2001) Intrinsically disordered protein. J Mol Graphics Modell 19:26–59.
74. Kabsch W, Sander C (1983) DSSP: definition of secondary structure of proteins given a set of 3D coordinates. Biopolymers 22:2577–2637.
75. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003) Protein disorder prediction: implications for structural proteomics. Structure 11:1453–1459.
76. Hubbard SJ, Thornton JM (1993) Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London 2.
77. Team RC (2014) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.