BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 3 2013, pages 331–337
doi:10.1093/bioinformatics/bts694

Structural bioinformatics

Advance Access publication November 30, 2012

Non-local residue–residue contacts in proteins are more conserved than local ones

Orly Noivirt-Brik[1], Gershon Hazan[2], Ron Unger[2,†] and Yanay Ofran[2,†,*]

[1]Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel and [2]The Goodman Faculty of Life Sciences, Bar Ilan University, Ramat Gan 52900, Israel

Associate Editor: Burkhard Rost

ABSTRACT

Non-covalent residue–residue contacts drive the folding of proteins and stabilize them. They may be local—i.e. involve residues that are close in sequence, or non-local. It has been suggested that, in most proteins, local contacts drive protein folding by providing crucial constraints of the conformational space, thus allowing proteins to fold. We compared residues that are involved in local contacts to residues that are involved in non-local contacts and found that, in most proteins, residues in non-local contacts are significantly more conserved evolutionarily than residues in local contacts. Moreover, non-local contacts are more structurally conserved: a contact between positions that are distant in sequence is more likely to exist in many structural homologues compared with a contact between positions that are close in sequence. These results provide new insights into the mechanisms of protein folding and may allow for better prediction of critical intra-chain contacts.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: yanay@ofranlab.org

Received on June 3, 2012; revised on October 25, 2012; accepted on November 23, 2012

1 INTRODUCTION

The importance of local versus non-local contacts is debated: The folding of a peptide chain into a biologically active 3D structure is driven primarily by the non-covalent interactions between atoms from different amino acids (Creighton, 1990; Dill *et al.*, 2008). Not all the non-covalent interactions in a protein are equally consequential for folding and stabilization of the structure. Identifying the interactions that drive folding and distinguishing them from non-critical interactions is a key to understanding protein folding. A common distinction that is often mentioned in this context is between local interactions, namely those that occur between residues that are adjacent in sequence, and non-local ones. The relative importance of local versus non-local interactions for protein folding and stability is bitterly debated for >35 years. Experimental and theoretical analyses of their respective importance give conflicting results (Ionescu and Matthews, 1999). An earlier study that was based on computer simulation of models of proteins on a two dimensional lattice concluded that non-local interactions are critical for folding (Taketomi *et al.*, 1975). Around the same time, however, other studies relied on theoretical and experimental results to propose that folding is determined primarily by local interactions (Anfinsen and Scheraga, 1975). This view was later supported by several studies that suggested that local interactions dominate folding (Harrison and Durbin, 1985; Karplus and Weaver, 1976; Rooman *et al.*, 1992; Wright *et al.*, 1988). Other studies used statistical mechanical models to suggest that the importance of non-covalent interactions is defined more by their physical nature (e.g. hydrophobic versus electrostatic) than by the sequence separation of the participating residues (Dill, 1990; Kaya and Chan, 2003). Notwithstanding, numerous studies have continued to argue that the sequence separation is a relevant feature for distinguishing between more and less important interactions. With the growth of protein structure databases and the advent of computational methods, more sophisticated analyses were applied to larger datasets. One computational analysis determined that the contribution of local interactions to the stability of the native state is small (Govindarajan and Goldstein, 1995), while another study concluded that foldability of a sequence is determined primarily by local interactions (Unger and Moult, 1996). In several large scale analyses, the notion of interaction was replaced by that of a contact, namely analysing all pairs of residues that are close enough in space to allow for an interaction between them.

Plaxco, Simons and Baker (Plaxco *et al.*, 1998) linked the distribution of local and non-local contacts to folding rate. They used the relative contact order (CO), which is the average sequence distance between all pairs of contacting residues in a chain. Higher CO indicates more non-local contacts, while lower CO indicates that more of the contacts are local. They analysed the CO of few dozen single domain proteins and concluded that proteins with lower CO fold more rapidly than ones with higher CO (Ivankov *et al.*, 2003; Plaxco *et al.*, 1998). A common interpretation of these results is that the ratio of local to non-local contacts in a protein determines the kinetics of folding. A recent analysis suggested that local and non-local contacts have significantly different effect on the folding rate (Zou and Ozkan, 2011).

Most of the above studies are based on an attempt to directly assess the contributions of different interactions to folding based on various energetic models, and in some cases on incidental experimental analysis of a small set of proteins. The different results may be ascribed, at least to some extent, to the difference in the models and techniques used in the assessment. Here, we propose to use the wealth of sequence and structure data to tackle this question in a different way. Rather than attempting

*To whom correspondence should be addressed.
†These authors contributed equally to this work.

to explicitly assess the energetic contribution of each type of intermolecular interaction to the folding and stability of the chain, we propose to indirectly assess their importance through their evolutionary and structural conservation. We assume that positions that form contacts that are more critical for folding and stability of the proteins are likely to be more conserved. This approach enables large scale analysis of all known structures and does not require theoretical models and approximations that may or may not faithfully represent the physical reality of protein folding. We checked whether there are significant differences in evolutionary conservation between local and non-local contacts in the native structure of proteins. We found that in the vast majority of proteins with known 3D structure, non-local contacts are more conserved than local ones. Furthermore, the average conservation of residues that are involved in non-local interactions is significantly higher than that of proteins involved in local interactions. Analysis of the structural similarity between homologous structures allowed us to assess the structure conservation of individual residues. This was done by checking how often the position of the residue is structurally aligned in structural homologues. Using a proper background, we were able to show that residues that are involved in non-local contacts are more likely to be structurally conserved compared with residues that are involved in local contacts. Furthermore, the non-local contacts themselves are more conserved than their local counterparts. These differences indicate that residues that are involved in non-local contacts are less tolerant to substitutions and structural changes than residues that are involved in local contacts. Together, they offer a new perspective in the discussion about the importance of local and non-local contacts to protein folding and stability.

2 METHODS

Construction of datasets: The analysis was based on a non-redundant set of PDB (Berman *et al.*, 2000). The PDBselect-25% list (Griep and Hobohm, 2010) was downloaded from http://bioinfo.tg.fh-giessen.de/pdbselect/ (April 2009 version). This list contained 4423 different single protein chains, each sharing no more than 25% sequence identity with any of the other sequences. Only the X-ray structures were further used for our analysis. The final list contained 2983 chains.

Sequence alignment: For the analysis of evolutionary conservation, we used the HSSP database (Dodge *et al.*, 1998). HSSP files were downloaded from the HSSP database (ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/hssp) (Schneider and Sander, 1996), and a list containing 1912 PDB entries of all HSSP alignments with minimal length of 100 amino acids and minimal size of 50 sequences was generated. Within these alignments, the average pairwise sequence identity was 37% (±13%). We will refer to this list as PDB-List1. This list was used to construct single-chain sequence alignments based on their full HSSP alignments. The HSSP alignments may contain more than one peptide chain for each PDB entry, since it includes all the chains in the entry. Thus, the size and length of the new single-chain alignments were re-evaluated. The final set of the HSSP-based alignments contained 1396 different alignments of single-chain proteins. The length distribution of these alignments (using the length of the 'master' in each alignment) is shown in Supplementary Figure S1. Within these alignments, the average pairwise sequence identity was 42% (±12%).

Structure alignment: DALI database (Holm and Sander, 1998) was taken from http://ekhidna.biocenter.helsinki.fi/dali/downloads/download.html (updated for 5.11.09). Proteins from the corresponding PDBselect-90% list that was used to build this version of DALI were identified. The DALI-based structure alignments were generated for all PDB entries included in PDB-List1, by excluding from the original DALI alignment structures corresponding to PDB entries not included in the PDBselect-90% list. This filtering restricted the level of redundancy in our DALI-based structure alignments. Since structural data is scarcer than sequence data, we did not require here that each alignment will contain 50 structures, but >85% of these alignments contained >20 structures.

The final set of DALI-based alignments contained only 1776 different alignments owing to missing PDB structures or missing DALI information. Note that these 1776 are not necessarily the same as the ones in the HSSP-based set. For control purposes, additional alignment programs like Mammoth (Ortiz *et al.*, 2002) and CE (Shindyalov and Bourne, 1998) were used for a subset of the alignments.

Generation of local and non-local contact sets: All PDB entries that were used to construct the sequence alignment and structure alignment were analysed to determine the spatial distance and the sequence separation of all pairs of residues. For the sequence alignments, residues that corresponded to positions that had >10% gaps in the HSSP sequence alignments were not included in any of the contact sets. Two residues were defined to be in contact if the spatial distance between their respective C_β atoms (or C_α in the case of glycine) was ≤6Å. Contacting residues with sequence separation of 5–10 residues were defined as local. Contacting residues with sequence separation ≥20 residues were defined as non-local. Other contacts were not included in any set. To confirm the robustness of the results to the choice of sequence separation cut-offs, we also constructed datasets where the local contacts were defined as sequence separation of 4–8, 6–10 or 2–10 and non-local were defined as ≥15. We also varied the contact cut-off from 6 Å to 6.5 Å and 7 Å. In all cases results remained highly significant (see Supplemantary Material). We removed alignments in which there were no contacts (either local or non-local) for any of the sequence separation of distance cut-offs. Thus, for example, Table 1 contains data about 1355 alignments.

Equal accessibility sets: Since the core of the protein is, on average, more conserved than its surface (Bustamante *et al.*, 2000; Choi *et al.*, 2006; Conant and Stadler, 2009; Franzosa and Xia, 2009; Goldman *et al.*, 1998; Overington *et al.*, 1992), we generated local and non-local contact sets with equal accessibility distributions, to avoid any biased due to such effect. This was done both with absolute accessibility (in terms of $Å^2$) and for relative accessibility (% of theoretical maximal accessibility). First, we obtained the solvent accessibility value of each residue in the dataset [for the DALI-based dataset, these values were calculated using the DSSP program (Kabsch and Sander, 1983), and for the HSSP-based dataset, they were taken from the HSSP files]. Then, the accessible surface area was binned in intervals of $10Å^2$. E.g. all residues with solvent accessible area that is <10 $Å^2$ were assigned an index accessibility of 1,

Table 1. Evolutionary conservation of local versus non-local contacts in 1355 alignments of protein families

$N_{\text{total}} = 1355$	Entropy	Conservation weight
Local > non-local	849 (63%)	563 (41%)
Non-local > local	494 (36%)	771 (57%)
Local = non-local	12 (1%)	21 (2%)
P-value	4.47×10^{-22}	1.45×10^{-8}

The number of proteins in which the average conservation score/Entropy of residues in local contacts is larger than, smaller than or equal to the average score of residues in non-local contacts is presented in the first, second and third rows, respectively. Note that smaller values of entropy represent stronger conservation while for conservation weight higher values represent stronger conservation. *P*-values were obtained using paired sign-test.

residues with accessibility that was $\geq$10 Å^2 and <20 Å^2 were assigned an index accessibility of 2, and so forth. Then, all contacts were grouped according to the pair of indices that describe the accessibility of the residues that form them. That is, all contacts that are formed between a residue with accessible surface of 30–40 Å^2 and a residue with accessible surface of 60–70 Å^2, belong to the same group. Finally, a set of local contacts and a set of non-local contacts were equally sampled from all groups by selecting N contacts from each such accessibility-defined group, with N defined as:

$$N_i = \min(N^i_{local}, N^i_{non-local}) \tag{1}$$

where i designates a specific accessibility group and N^i_{local} and $N^i_{non-local}$ designate the number of local and non-local contacts in bin i, respectively. Thus, we generated two sets of contacts with equal accessibility distribution. We also created as set of *Equal accessibility–Equal neighbours* in which in each bin, contacts have both the same accessibility and the same number of neighbours.

Calculation of sequence conservation scores: We used two evolutionary conservation scores: (i) The Shannon-entropy score, E, and (ii) The conservation weight, CW (Sander and Schneider, 1991). The entropy E is defined as:

$$E(i) = -\sum_{x=1}^{20} p(x)\log_{20} p(x) \tag{2}$$

where i is a given position in the HSSP multiple sequence alignment (MSA), x is an amino acid type and $p(x)$ is the frequency of that amino acid at position i. This score can range from 0 for fully conserved position to 1 for maximally variable ones.

The conservation weight CW is defined as:

$$CW(i) = \frac{\sum_{k,l}^{N_{pairs}} w_{kl}\,sim(R_{ik}, R_{il})}{\sum_{k,l}^{N_{pairs}} w_{kl}} \tag{3}$$

where i is a given position in the MSA, the sum runs over all the possible pairs of sequences in the MSA, k and l ($k \neq l$). $sim(R_{ik},R_{il})$ is the similarity between the residues located in the i^{th} position in sequence k and l. The similarity values are the normalized values of PAM250 matrix (Dayhoff *et al.*, 1978). w_{kl} is the weight of a sequence pair which is defined here as the fraction of amino acid mismatches over the alignment length L:

$$w_{kl} = 1 - \frac{1}{L}\sum_{i}^{L} \delta(R_{ik}, R_{il}) \tag{4}$$

where $\delta(R_{ik}, R_{il}) = \begin{cases} 1 & R_{ik} = R_{il} \\ 0 & otherwise \end{cases}$

This score ranges between 0 for residues that are not conserved and 1 for fully conserved residues.

For each protein in the dataset, we calculated the average of each of the conservation scores (E and CW) for residues that are involved in local contacts. Then we calculated these averages also for residues that are involved in non-local contacts. One residue can be involved in more than one contact, hence we reconsidered a residue for every additional contact it has with another residue. Note that by this scheme, a residue could be included both in the average for local interaction and in the average for non-local interactions. Average conservation was calculated separately for local contacts and for non-local contacts for each protein.

Structure conservation scores: Structure conservation was evaluated using two scores. The first, f_{aligned}, defined for a given pair of interacting residues as:

$$f_{\text{aligned}} = \frac{N_{\text{aligned}}}{N_{\text{total}}} \tag{5}$$

where N_{aligned} is the number of structures in the DALI structure alignment in which both residues are structurally aligned (i.e. there is no gap in the corresponding position), and N_{total} is the number of all structures in the DALI structure alignment for the protein under discussion. The second score, $f_{\text{contacts,}}$ is defined as:

$$f_{\text{contacts}} = \frac{N_{\text{contacts}}}{N_{\text{total}}} \tag{6}$$

where N_{contacts} is the number of structures in the DALI alignment in which these two residues are in contact (i.e. their C_β–C_β are $\leq$6 Å apart).

For each protein in the dataset, the average f_{aligned} and f_{contacts} values were calculated separately for local and non-local contacts, and these average values were compared. We defined the Non-local Preference measure (NLP_{aligned}) as the percentage of proteins in a dataset in which the average f_{aligned} for non-local contacts is greater than for local contacts, and in a similar way we defined the NLP_{contacts} measure, which is the percentage of proteins in a dataset in which the average f_{contacts} for non-local contacts is greater than for local contacts.

Control set for structural conservation: Structural conservation of a pair of residues can be the result of the conservation of an interaction between them, but it can also be the result of the conservation of a structural neighbourhood independent of their interaction. To distinguish between these two possibilities, we defined a set of pseudo-contacts that included all pairs of residues whose spatial distance is 10–12 Å. The width of the range was chosen to be 2 Å similar to the range of real contacts, which are in distance between 4 Å and 6 Å. Arguably, residues that are 10–12 Å apart are not in any type of direct physical interaction, and this distance has no particular significance. Thus, this set can be used as a control for real contacts. We will refer to this set as $S_{10\text{–}12}$. The number of local and non-local pairs that were sampled from the control $S_{10\text{–}12}$ set was equal to the number of those used in the respective contact set of the same structure, and these samples were used to calculate both f_{contacts} scores for the control set. It is important to note that the averaged scores calculated for the local versus non-local sets or for contacts versus $S_{10\text{–}12}$ are comparable owing to the fact that they were derived from the same sequence or structure alignment and, thus, they share the same evolutionary background. Such comparison filters out the evolutionary noise and, therefore, its results are more reliable.

Comparing pseudo-contacts and real contacts: To compare pseudo-contacts and real contacts for local and non-local contacts, we defined the following measures:

$$\Delta_{\text{aligned}} = NLP^{c}_{aligned} - NLP^{pc}_{aligned} \tag{7}$$

Where $NLP^{c}_{aligned}$ is the percentage of proteins in a dataset in which the average f_{aligned} is greater for non-local contacts than for local contacts, and $NLP^{pc}_{aligned}$ is the same measure for pseudo-contacts. Thus, if in most proteins, non-local contacts are more structurally conserved than local contacts, but this preference exists also in pseudo-contacts, then Δ_{aligned} will be close to 0. Similarly, if in most proteins, local contacts are more structurally conserved than non-local contacts, but this preference exists also in pseudo-contacts, then Δ_{aligned} will also be close to 0. If, however, in one of the groups (either local or non-local), there is a large discrepancy between the structural conservation of real contacts and pseudo-contacts, Δ_{aligned} will be largely positive (if non-local contacts are more conserved) or largely negative (if local contacts are more conserved).

Similarly, in order to assess the conservation of the actual contacts, we define:

$$\Delta_{\text{contacts}} = NLP^{c}_{contacts} - NLP^{pc}_{contacts} \tag{8}$$

3 RESULTS

Two residues were defined to be in contact if their C_β atoms were <6 Å apart (C_α for glycine). We defined local contacts as contact between amino acids that are 5–10 residues apart. Non-local contacts were defined between amino acids that are 20 and more residues apart. This range was selected to avoid most interactions that occur within secondary structure elements. All data presented in the main text are based on these definitions. The supplementary material contains data for other cut-off values that we have tried. Although the strength of the signal we observed may slightly vary for different cut-off values, the trend was robust to the choice of cut-off for all measures.

Figure 1 shows the mean entropy of residues plotted against the sequence separation of the contacts in which they are involved. The mean entropies of contacts at very low sequence separation are much higher than the rest of the graph. The trend remains not only when comparing local and non-local contacts but also when comparing different non-local contacts. For instance, looking at contacting residues that are between 20 and 40 positions apart, virtually all of them have a mean entropy that is >0.39. However, 93% of the contacting residues that are >40 positions apart have mean entropies that are <0.39. Supplementary Figure S2 presents a similar plot for CW.

We also investigated the balance of local and non-local conservation within each protein. Table 1 compares the number (and percentage) of proteins in which residues involved in local contacts were, on average, more or less conversed than non-local ones. For both metrics of evolutionary conservation (E and CW), in a significant majority of proteins, residues in non-local contacts are more conserved than those involved in local ones. These results have very significant *P*-values (paired sign-test) demonstrating the clear preference of conservation of residues involved in non-local contacts compared with residues involved in local contacts. Supplementary Figure S3 presents histograms of the difference between local and non-local values for Entropy and CW, showing, again, higher conservation of non-local contacts. Supplementary Table S3b shows that this is true for different structural classes and is independent of secondary structure composition. We found no significant differences between the residue–residue preferences in local contacts and in non-local contacts (see Supplementary Fig. S4).

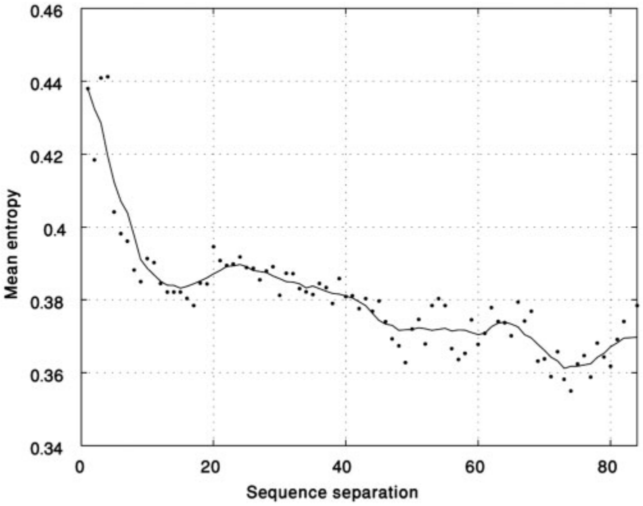

Fig. 1. Mean entropy versus sequence separation: The mean entropy (y-axis) of all the residues that are in contact (C_β distance of <6 Å) for a given sequence separation (x axis). The graph is based on 1 238 642 contacts. Dots represent mean entropy of all the contacts with the same sequence separation. The smoothed line is based on a running window of length 9. Note that the decrease in entropy continues throughout the range of sequence separations: the entropy for sequence separation >20 is lower than for sequence separation <20, and the entropy for sequence separation >40 is lower than sequence separation <40. The analysis included all the consecutive distances from 1 to 84, for which we have >1500 data points

Next, we wanted to check whether the contacts themselves, rather than the residues involved in the contacts, are more conserved in non-local versus local contacts. f_{align} measures the percentage of structures in the alignment in which the two contacting residues are structurally aligned (i.e. there is no gap in the structural alignment in either of these positions). $f_{contacts}$ reflects the fraction of structures in a structural alignment in which the considered contact itself exists (i.e. the two corresponding residues in the aligned structures are in contact with each other). We found that non-local contacts are more structurally conserved than local contacts (*P*-value of 10^{-36} and 10^{-20} for f_{align} and $f_{contacts}$, respectively, Supplementary Figs S5 and S6). However, since errors in structural alignments may build up as sequence separation grows, it is not simple to compare the structural conservation of local contacts with that of non-local contacts. Moreover, conservation of a contact may be the result of the conservation of a structural neighbourhood independent of their interaction. To control for these possible biases, we used $\Delta_{aligned}$ and $\Delta_{contact}$ (see Section 2), which measure to what extent real non-local contacts are more conserved than real local contacts, and compares this difference to the respective difference between pseudo-contacts. If local contacts are indeed more conserved, these measures should be negative. If there is no difference, they should be close to 0, and if non-local contacts are more conserved, it should be positive. Table 2 shows, for different definitions of local and non-local contacts, that $\Delta_{aligned}$ and $\Delta_{contact}$ have large positive value indicating that non-local contacts are more structurally conserved (using the baseline of pseudo-contacts) than local contacts. These results are based on the Dali alignments; similar results were obtained for Mammoth and CE alignments (data not shown).

Another way of looking at the data can be seen in Figure 2 where the $f_{contacts}$ parameter of local (Fig. 2A) and non-local

Table 2. Structural conservation of contacts

Contact definition	$\Delta_{aligned}$	$\Delta_{contact}$
Local: 5–10, Non-local: ≥20	10.65	35.35
Local: 6–10, Non-local: ≥15	9.24	33.41
Local: 4–8, Non-local: ≥15	9.63	51.91
Local: 2–10, Non-local: ≥20	6.09	47.83

$\Delta_{aligned}$ and $\Delta_{contact}$ assess the tendency of contacts to exist in different structurally similar proteins (see Section 2). If local contacts are always more conserved than non-local contacts then both $\Delta_{aligned}$ and $\Delta_{contact}$ should be equal to (−100). If there is no difference between local and non-local, then they should be 0. As seen in the table, non-local contacts are more conserved regardless of the choice of cut-off for local and non-local contacts.

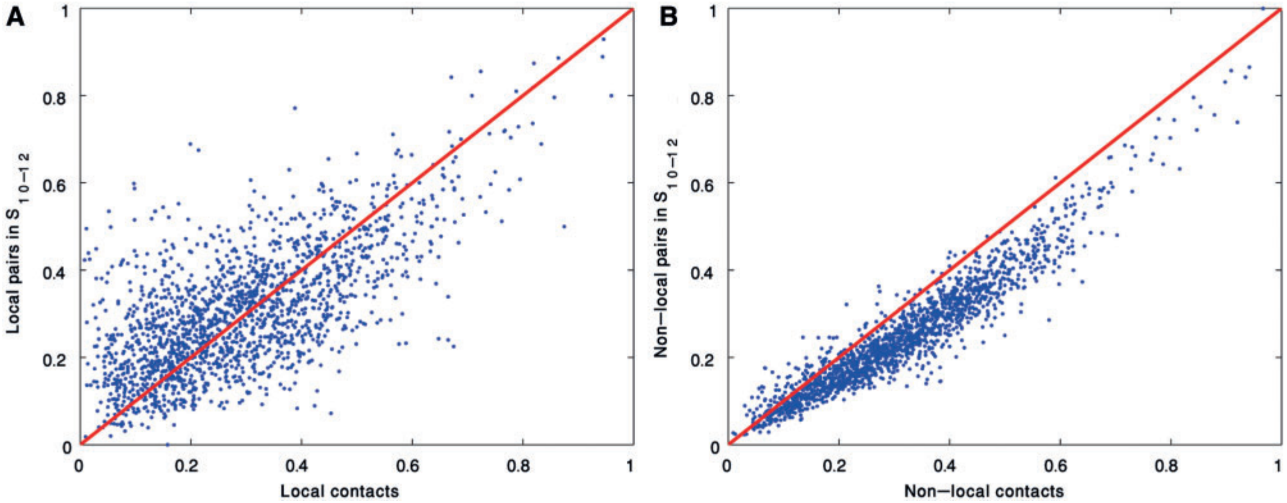

Fig. 2. Structural conservation of real versus pseudo-contacts. Each blue dot represents one protein chain. On the X axis is the average structural conservation score of real contacts ($f_{contacts}$) and on the Y axis is the average structural conservation score of pseudo-contacts ($S_{10–12}$). If there is no structural conservation of real contacts, proteins should fall on or around the red line, which represents equal structural conservation of real and pseudo-contacts. Structural conservation of contacts should result in more dots below the line (i.e. proteins in which real contacts are more conserved than pseudo-contacts). The values were calculated for local (**A**) and non-local (**B**) pairwise interactions for all the 1774 DALI-based structure alignments. The difference between real and pseudo-contacts was extremely significant for non-local contacts (P-values $< 1.0 \times 10^{-275}$ paired sign test)

(Fig. 2B) pairwise interactions in $S_{10–12}$ is plotted against the respective parameter in the contacts set using the Dali set. For local contacts, the structural conservation of real contacts is virtually the same as that of pseudo-contacts. However, for non-local contacts, real contacts are more conserved than pseudo-contacts in almost all structures (P-value $< 1.0 \times 10^{-275}$ paired sign test). These results indicate that for local contacts, spatial distance between the residues tends to be maintained, possibly reflecting the conservation of structural neighbourhood and not necessarily the importance of the specific contact itself. In a striking contrast, distances between residues that are far apart in the sequence do not tend to be conserved unless there are in real contact.

4 DISCUSSION

Conceivably, the higher conservation of non-local contacts compared with local ones may simply be the result of other, already known, characteristics of protein structure. For example, it is well established that buried residues tend to be more conserved than exposed ones (Bustamante *et al.*, 2000; Choi *et al.*, 2006; Conant and Stadler, 2009; Franzosa and Xia, 2009; Goldman *et al.*, 1998; Overington *et al.*, 1992). Possibly, non-local contacts may involve more buried residues, while local contacts may occur more frequently between residues that are accessible to solvent. If true, this may be reflected in the average conservation of each of these groups of contacts.

However, when we calculated the correlation between the accessible surface area of the residues involved in contacts and their entropy, we found a rather weak correlation coefficient of 0.34. In a more direct analysis, we created equal accessibility sets (see Section 2) such that the local and non-local interactions set have the same distributions of solvent accessibilities. In this comparison, the difference remained highly significant, though smaller in magnitude (P-values $< 1.7 \times 10^{-5}$ for entropy, see Supplementary Table S1B). Using relative accessibility gave similar results (Supplementary Table S1C).

Another possible source for the difference may be the number of spatial neighbours of each residue. Residues involved in non-local contacts may be involved, on average, in more contacts than other residues, and such 'hubs' may also be more conserved than other residues. To control for this possible effect, we created an *Equal accessibility–Equal neighbours* set (see Section 2). Again, non-local contacts were significantly more conserved even when we controlled for number of spatial neighbours (see Supplementary Table S2). We also checked whether the differences we observed between local and non-local contacts may stem from using multi-domain proteins, from using the DALI structural alignment algorithm or from bias toward a specific structural class of proteins. Although in some cases smaller number of alignments reduced the statistical significance, the trend of greater structural conservation of non-local contacts remained clear in all cases (see Supplementary Tables S1–S3).

As an illustration we bring the example of non-fluorescent flavoprotein (PDB 1nfp). Figure 3 highlights in red residues that participate in contacts that are highly structurally conserved (i.e. high $f_{contacts}$) in this protein. The average $f_{contacts}$ of all local contact (i.e. between residues that are 5–12 positions apart) was 0.11, which is less than a half of the average for non-local contacts (between residues that were >20 positions apart), which was 0.25. As can be clearly seen in Figure 3, most of the highly conserved contacts are between residues that are distant in sequence (in this case, >20 residues apart), while only a few may be considered local (<12 residues apart). It seems that the non-local contacts stabilize interactions between helices and within sheets. In both categories, some of the residues are buried and others are exposed.

Thus, using a non-redundant version of all proteins with known 3D structure, we showed that non-local contacts are, on average,

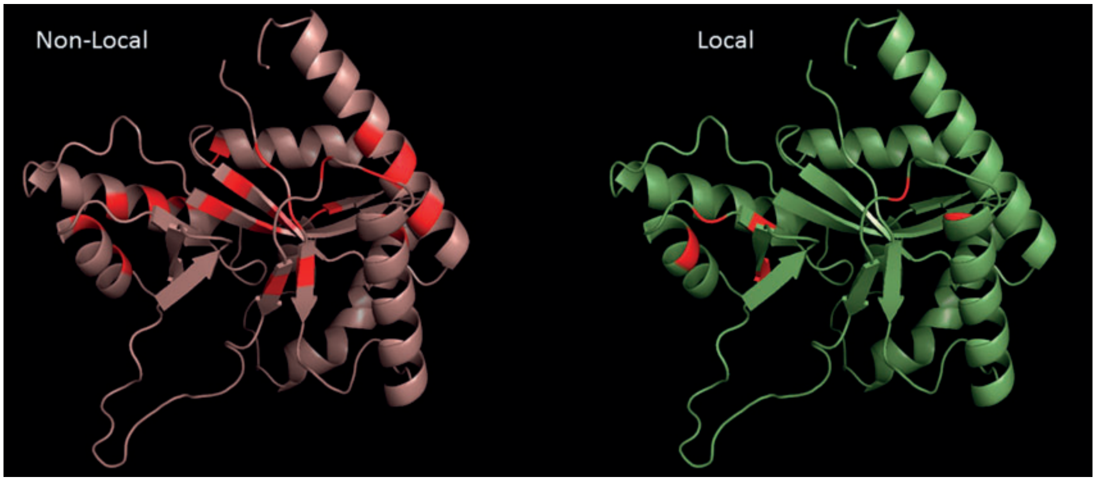

Fig. 3. Residues in structurally highly conserved local and non-local contacts. We identified the residues that participate in the contacts with the highest f_{contacts} in the non-fluorescent flavoprotein (PDB 1nfp). These residues form the most structurally conserved contacts in the protein. On the right-hand side we highlighted in red residues in local contacts and on the left-hand side residues in non-local contacts

more conserved both evolutionarily and structurally. To the best of our knowledge, this is the first time that such differences have been observed and statistically assessed. The difference between these two classes of contacts can stem from a variety of structural or evolutionary factors. One possible explanation may be that the contribution of non-local contacts to the folding or to the stabilization of the native structure is greater than that of local contacts. This claim is consistent with the predictions made based on polymer statistics that a longer polymer chain is less likely to have contacts between its ends than shorter chains (Doi and Edwards, 1988). Therefore, in order to ensure the existence of a non-local contact, there might be a higher need to conserve the residues providing such a contact. This hypothesis is supported by several recent experimental observations that found that non-local native interactions are being formed in very early stages of the folding process (Felitsky *et al.*, 2008; Orevi *et al.*, 2009). They are also consistent with the conjecture that formation of such interactions in the early stages of folding reduce the configurational entropy of the chain and consequently accelerates folding. If true, these results may be useful in devising structure prediction tools and in protein engineering.

ACKNOWLEDGEMENT

Thanks to Elisha Haas (BIU) for useful discussion.

Funding: This research was supported in part by Israel Science Foundation grant 1339/08 to RU. Y.O. is supported in part by the Israeli Science Foundation, grants No. 511/10(www.isf.org.il). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest: None declared.

REFERENCES

Anfinsen,C.B. and Scheraga,H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, **29**, 205–300.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bustamante,C.D. *et al.* (2000) Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Mol. Biol. Evol.*, **17**, 301–308.

Choi,S.S. *et al.* (2006) Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. *Mol. Biol. Evol.*, **23**, 2131–2133.

Conant,G.C. and Stadler,P.F. (2009) Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol. Biol. Evol.*, **26**, 1155–1161.

Creighton,T.E. (1990) Protein folding. *Biochem. J.*, **270**, 1–16.

Dayhoff,M.O. *et al.* (1978) *Atlas of Protein Sequence and Structure*, Dayhoff,M.O. ed., Vol. 5, (**Suppl. 3**). National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Dill,K.A. (1990) Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.

Dill,K.A. *et al.* (2008) The protein folding problem. *Annu. Rev. Biophys.*, **37**, 289–316.

Dodge,C. *et al.* (1998) The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.*, **26**, 313–315.

Doi,M. and Edwards,S.F. (1988) *The Theory of Polymer Dynamics*. Oxford University Press, USA.

Felitsky,D.J. *et al.* (2008) Modeling transient collapsed states of an unfolded protein to provide insights into early folding events. *Proc. Natl Acad. Sci. USA*, **105**, 6278–6283.

Franzosa,E.A. and Xia,Y. (2009) Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.*, **26**, 2387–2395.

Goldman,N. *et al.* (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458.

Govindarajan,S. and Goldstein,R.A. (1995) Optimal local propensities for model proteins. *Proteins*, **22**, 413–418.

Griep,S. and Hobohm,U. (2010) PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Res.*, **38**, D318–D319.

Harrison,S.C. and Durbin,R. (1985) Is there a single pathway for the folding of a polypeptide chain? *Proc. Natl Acad. Sci. USA*, **82**, 4028–4030.

Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.

Ionescu,R.M. and Matthews,C.R. (1999) Folding under the influence. *Nat. Struct. Biol.*, **6**, 304–307.

Ivankov,D.N. *et al.* (2003) Contact order revisited: influence of protein size on the folding rate. *Protein Sci.*, **12**, 2057–2062.

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

Karplus,M. and Weaver,D.L. (1976) Protein-folding dynamics. *Nature*, **260**, 404–406.

Kaya,H. and Chan,H.S. (2003) Contact order dependent protein folding rates: kinetic consequences of a cooperative interplay between favorable nonlocal interactions and local conformational preferences. *Proteins*, **52**, 524–533.

Orevi,T. *et al.* (2009) Early closure of a long loop in the refolding of adenylate kinase: a possible key role of non-local interactions in the initial folding steps. *J. Mol. Biol.*, **385**, 1230–1242.

Ortiz,A.R. *et al.* (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Overington,J. *et al.* (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*, **1**, 216–226.

Plaxco,K.W. *et al.* (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.

Rooman,M.J. *et al.* (1992) Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry*, **31**, 10226–10238.

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

Schneider,R. and Sander,C. (1996) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **24**, 201–205.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Taketomi,H. *et al.* (1975) Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.*, **7**, 445–459.

Unger,R. and Moult,J. (1996) Local interactions dominate folding in a simple protein model. *J. Mol. Biol.*, **259**, 988–994.

Wright,P.E. *et al.* (1988) Conformation of peptide fragments of proteins in aqueous solution: implications for initiation of protein folding. *Biochemistry*, **27**, 7167–7175.

Zou,T. and Ozkan,S.B. (2011) Local and non-local native topologies reveal the underlying folding landscape of proteins. *Phys. Biol.*, **8**, 066011.