On the amino acid composition and mechanical stability of protein structures

Antonio Deiana[1,2]*, Kana Shimizu[3,4]*, Andrea Giansanti[1,2,5]*§

[1] Department of Physics, La Sapienza University of Rome, P.le A. Moro 5, 00185, Rome, Italy
[2]Interdepartmental Research Centre for Models and Information Analysis in Biomedical Systems (CISB), La Sapienza University of Rome, C.so Vittorio Emanuele II 244, 00186, Rome, Italy
[3]Department of Computer Science, Graduate School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinijuku-ku, Tokyo, 169-8555, Japan
[4]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan
[5] INFN, Sezione di Roma, P.le A. Moro 2, 00185, Rome, Italy

*These authors contributed equally to this work
[§]Corresponding author

Email addresses:
AD: Antonio.Deiana@roma1.infn.it
KS: shimizu-kana@aist.go.jp
AG: Andrea.Giansanti@roma1.infn.it

Abstract

Amino acid composition is an important determinant of protein structures. In this paper we investigate the relationship between amino acid composition and mechanical stability of protein sequences. We divide the protein structures deposited in the Protein Data Bank (PDB) as ordered, disordered and in the twilight zone, depending on their amino acid composition. We use a consensus score S_{SU} among three predictors of global disorder, Poodle-W, gVSL2 and mean pairwise energy. Mechanical stability is evaluated through Miyazawa-Jernigan potential. We find that the three groups of protein sequences have different contact energy, disordered sequences being the most unstable and ordered ones being the most stable. Secondary structure energy and global mechanical stability, on the other hand, are about the same in the three groups of proteins, pointing to a fundamental role of backbone interactions in the stabilization of the tertiary structure. Proteins with relative high contact energy tend to remain short in length and they do not enrich in disorder-promoting amino acids. Moreover, several short proteins in the twilight zone compensate their relative instability through disulfide bridges. Our results support the hypothesis that backbone interactions play a fundamental role in the stabilization of protein structures. However, the role of long-range interactions and its relation with protein length must be further investigated. It is necessary to develop a more fundamental theory to understand the exact relation between amino acid composition and the mechanical stability of protein sequences.

Background

Tertiary structure is important to protein function. However, particularly in the last decade, several proteins have been discovered having important biological functions but lacking a well-defined three-dimensional (3-D) structure. It is known that these *intrinsically disordered* proteins (IDP), which functionally display an ensemble of flexible conformations are involved in DNA binding, signalling, targeting and other important cellular functions, as well as in cancer development and amyloidotic diseases. These proteins have been denoted variably, not only as IDP, but also as *intrinsically unstructured* or as *natively unfolded* proteins. Despite the term disorder might be misleading (an intrinsically structured protein is clearly not a fragment of an ordered, periodic solid) we follow here what seems to be the prevailing choice. Nevertheless, we believe that a terminological revision is needed, based on empirical classifications. A comprehensive introduction to the field is given in the recent book by Peter Tompa [1], several significant papers are listed in the bibliography [2-13] and a cured database is available on the net [14].

From a physical point of view, there is consensus in considering the tertiary structure of a protein as an equilibrium configuration corresponding, in the protein conformational space, to a minimum of the free energy landscape. Protein folding can be viewed then as a pathway in which the protein sequence rolls down on the energy landscape [15-19]. Since there is a high energy barrier between the folded and the unfolded configuration, a protein in the folded status cannot leave this conformation unless it is forces by denaturant agents. In this equilibrium configuration, the protein is able to fulfil its biological function. Natively unfolded proteins, on the other hand, are characterized by free energy landscapes with many minima separated by low energy barriers, of the order of k_BT. Therefore, they cannot have a stable tertiary structure, but they are characterized by an ensemble of high-flexible interchangeable extended three-dimensional conformations [20, 21].

It has been observed that ordered and disordered protein sequences have a different amino acid composition. Romero *et al.* report that well-structured segments of polypeptide chains have a higher frequency of T, C, F, I, Y, V and L than disordered ones. They indicate these amino acids as order-promoting. Analogously, unstructured segments of polypeptide chains are enriched in disorder-promoting amino acids (M, A, R, Q, S, P, E and K). [22] In many papers, amino acid composition is the main physical-chemical properties used to infer, from an *ab-initio* analysis, whether a protein sequences is structured or not. This observation leads to the conclusion that amino acid composition is important for a protein to reach a free energy minimum. Sequences enriched in disorder-promoting amino acid could have high free energies than those enriched in order-promoting ones.

Interestingly, however, several recent studies have indicated that amino acid composition is not sufficient to determine the fold of a protein sequence. Szilagyi *et al.* [23] observe that ordered and disordered proteins overlap when plotted on a hydrophobicity-charge plane, pointing to the existence of a *twilight zone* between order and disorder, an overlap volume in the vector space of amino acidic compositions, occupied by both ordered and disordered sequences. In their work Szilagyi *et al.* report that the twilight zone is wider for short proteins, and by means of lattice models, they infer that amino acidic composition alone is not sufficient to determine whether a protein sequence of short length folds into a tertiary structure. They also observe that longer proteins, on the other hand, can sample a comparatively higher number of conformations and then they have a higher probability to reach a

stable one. In these cases, the order of the amino acids in the sequence is less crucial and amino acidic composition only is sufficient to determine whether the sequence folds or not. The main merit of [23] is to have clarified the interplay of composition and length in determining the tendency of a protein to fold.

In this paper we investigate about the relation between amino acid composition and mechanical stability of protein sequences. Mechanical stability is highly related to thermodynamic stability, as we discuss in the Discussion section. We take into consideration protein structures deposited in the Protein Data Bank (PDB). Searching the PDB [24] for disordered amino acids can be useful to evaluate the occurrence of disorder in structured proteins, as a starting point toward understanding how disorder influences their biological function. Le Gall *et al.* [25] report that about 10% of proteins in the PDB have disordered fragments longer than 30 amino acids, and the percentage raises to 40% if protein fragments shorter than 30 amino acids are considered. This indicates that in the PDB there is a non-negligible fraction of putatively intrinsically disordered proteins. Let us note that the connection between the global stability of the fold of a protein and the presence of highly flexible segments (possibly related to the occurrence of non-observable, disordered residues) is still controversial, and it is one of the matters of the present paper to contribute a critical assessment of the problem. In this paper we scan the proteins in the PDB by means of sequence-only global predictors of folding, dichotomic classifiers which, in general, evaluate the amino acidic composition of a protein and compare it with the compositions of typical folded and unfolded proteins in a training set. In particular, we consider three IDP predictors and we combine them into a strictly unanimous consensus score S_{SU} [26]. We use mean pairwise energy, *gVSL2* and Poodle-W [26]. Mean pairwise energy has been introduced by Dosztanyi *et al.* [27] to discriminate ordered and disordered residues in proteins, and is implemented in the IUPred program [28]. *gVSL2* [26] is the arithmetic mean of the disorder score returned by the VSL2 predictor [29, 30]. Poodle-W is a predictor of natively unfolded proteins developed by Shimizu *et al.* [31] and is based on a spectral graph transducer [32]. Operationally, the strictly unanimous consensus score S_{SU} [26] classifies a protein as ordered if all three indexes agree in classifying it as folded and classifies a protein as disordered if all three indexes agree in classifying it as disordered. If at least one index disagrees, S_{SU} classifies the protein in the twilight zone. We have shown that S_{SU} is effective in selecting out proteins in the twilight zone between order and disorder from a generic set of proteins [26]. By using S_{SU} we classify about 3% of proteins in the PDB as disordered and 12% are assigned to the twilight zone between order and disorder. Generally, a protein predicted as disordered by a global predictor is considered as lacking a well-defined three-dimensional conformation. Following this assumption one should consider the predictions of natively unfolded proteins in the PDB as false predictions. In this paper however, we adopt the view of considering unstructured proteins as *structured proteins with peculiar amino acidic compositions*, similar to those of unfolded sequences, and we investigate their distribution in the PDB. We also start here to investigate the physical meaning behind the peculiar amino acidic compositions. More precisely, we investigate the relation between amino acidic composition, length of the proteins and their mechanical and thermodynamic stability.

Thermodynamic stability is an important property, since it is related to the propensity of a protein to reach a minimum of the free energy landscape, and therefore to fold into a tertiary structure, to crystallize, to bind or aggregate with other biological macromolecules and to other important biological functions [33]. As stated above, we

take into consideration mechanical stability, a property strictly related to thermodynamic stability. To estimate mechanical stability, we use the Miyazawa-Jernigan (MJ) potential [34-37]. We separately evaluate, in the MJ potential, *secondary structure energy* and *contact energy*; the first term is the interaction, in a specific protein conformation, between a tripeptide and the residue located at its centre, and therefore it considers short-range, local, interactions; the second term is the interactions among residues in contact, i.e. whose geometric centres are closer, in the structure, than 6.5 Å, and therefore it considers long-range interactions between residues which are far apart, along the backbone. Let us anticipate here what we have found: i) MJ energy correlates with protein length, consistently with the observation that longer proteins are more stable than shorter ones [33]; on the contrary, ii) MJ energy *per residue* does not depend on protein length; iii) secondary structure energy does not discriminate ordered from disordered proteins, as classified by S_{SU}, indicating that it does not depend on amino acidic compositions; whereas iv) contact energy is different in ordered and disordered proteins, the latter in fact, have a higher energy per residue than the former. Moreover, proteins in the twilight zone have a contact energy per residue intermediate between that of ordered and disordered proteins, with the exception of a certain number of short proteins with a quite low contact energy per residue, that, as we discuss in the paper, make their structure mechanically more stable by means of disulfide bridges, a property that it is not exhibited by disordered proteins. In a nutshell, S_{SU} acts as a filter of the MJ contact energy per residue.

The analysis of MJ contact energy distribution in the PDB shows that ordered and disordered proteins differ in their long-range interactions, which strongly depend on amino acidic composition, whereas the contribution due to secondary structure is comparable in both groups. Interestingly, the sum of secondary structure and contact energy does not distinguish between ordered or disordered proteins, suggesting that secondary structure energy gives an important contribution to the overall mechanical stability of proteins, quite independently of amino acidic composition.

Disordered and twilight zone proteins result to be depleted in order-promoting and rich in disorder-promoting amino acids, and this corresponds to higher contact energies and lower mechanical stability. These proteins tend also to have short polypeptide chains, due to the fact that they become more and more unstable and hard to be found in the PDB as their length increases. It appears therefore, also from the analysis here presented, that a strict relation between long-range conformational energy and protein length exists, and worth to be further investigated.

Results

Analysis of the amino acidic compositions of the proteins in the PDB

As mentioned above and detailed in the Methods, we combined three global IDP predictors into a strictly unanimous consensus score S_{SU} [26], which allows classification of protein sequences into three classes. In this work, we investigate the occurrence of these three classes of structural order in the PDB, seeking for a physical, energetic meaning behind this classification. From now on, *ordered*, *disordered*, or *twilight zone*, in denoting a protein of the PDB, refers to this classification.

We selected from the PDB 70684 sequences, excluding complexes. As expected, the majority of proteins are classified by S_{SU} as ordered; of the rest, about 3% are classified as disordered and 12% in the twilight zone.
Since the predictors of intrinsically disordered proteins combined in S_{SU} are functionals of amino acidic compositions, then disordered and twilight zone proteins should have measurably different residue frequencies with respect to ordered ones. This is shown in figure 1, where we show histograms of residue frequencies for the three classes. Ordered proteins are rich in W, C, F, I, Y, V, L, M, A, N, D, H and T, a list consistent with the list of order-promoting residues of Romero *et al.* [22]. Disordered proteins are rich in R, Q, S, P, E, K, and G. This list also coincides with the list of disorder-promoting amino acids proposed in [22]. Proteins assigned to the twilight zone have frequencies of W, F, I, Y, V, L, R, Q, P, E, K, G and N intermediate between those of ordered and disordered proteins. The amino acids C, A, S, D and T, on the other hand, are more frequent in proteins in the twilight zone than in those belonging to the other classes. To investigate the distribution of amino acidic compositions in the PDB, we computed the log-odds ratio S of the likelihoods, for a protein sequence, of being composed by either order-promoting or disorder-promoting amino acids [26, 38] (see Methods). The S distribution in the PDB is shown in figure 2. Ordered proteins are characterized by positive values of S (mean value 18.72 ± 0.06, median 15.73), and are mainly composed by order-promoting amino acids, while disordered proteins tend to display negative values of S (mean value -2.64 ± 0.18, median -1.68), being composed mainly by disorder-promoting amino acids. Proteins in the twilight zone have intermediate S-values, centred around zero (mean value 3.73 ± 0.05, median 2.79) with a slight shift of the distribution towards positive values.

Estimate of the conformational energies of proteins through Miyazawa-Jernigan contact potential

In this section we investigate whether ordered, disordered and twilight zone proteins in the PDB differ in the conformational energy of their tertiary structure. It has been reported that conformational energy can be related to the thermodynamic stability of proteins, as we discuss below (see Discussion) [23, 39]. To estimate the conformational energy, we used the MJ potential [34-37], sum of the local *secondary structure energy*, and of the long-range *contact energy* (see Methods for details).
We checked that the MJ potential is strongly correlated with protein lengths (Pearson's correlation coefficient -0.93): longer proteins have a lower conformational energy. Increasing the number of residues increases the number of interactions inside the protein structure making the fold mechanically more stable. This result is in line with the observation by Ghosh *et al.* [33] that thermodynamic stability increases with the number of residues in a polypeptide chain. To understand whether the conformational energy of proteins depends either on the mere number of amino acids (the protein length) or on their physical-chemical properties, we consider the MJ energy per residue. It does not significantly correlate with protein length (Pearson's correlation coefficient is equal to -0.20). Moreover, secondary structure energy per residue does not discriminate ordered, disordered and twilight zone proteins; contact energy per residue, on the other hand, is higher in disordered proteins and lower in ordered ones (see figure 3). In the supplemental figure 1 we report the distributions of contact MJ energy per residue E_c in ordered, disordered and twilight zone proteins; average values are: (-0.2511 ± 0.0003) arbitrary energy unit (a.e.u) for ordered

proteins and (-0.058 ± 0.003) a.e.u. for disordered ones, so disordered proteins have a lower $<E_c>$ than ordered ones. Proteins in the twilight zone have an average contact energy per residue $<E_c>$ equal to (-0.165 ± 0.001) a.e.u., intermediate between that of ordered and disordered proteins. We observe, however, a certain number of short proteins in the twilight zone with especially low contact energies per residue; we discuss these proteins, potentially rich in disulfide bridges, below.
Let us now express the MJ contact energy of a protein sequence as:

$$E_C \approx n_{OO} < \varepsilon_{OO} > + n_{OD} < \varepsilon_{OD} > + n_{DD} < \varepsilon_{DD} > \quad (1)$$

where n_{OO}, n_{DD} and n_{OD} are the number of contacts between order-promoting, between disorder-promoting and between order-promoting and disorder-promoting residues, respectively; $<\varepsilon_{OO}>$, $<\varepsilon_{DD}>$ and $<\varepsilon_{OD}>$ are the corresponding mean contact energies (see Methods for definitions). Then, contact energy per residue can be expressed as:

$$\frac{E_C}{N} \approx \frac{n_{OO}}{N} < \varepsilon_{OO} > + \frac{n_{OD}}{N} < \varepsilon_{OD} > + \frac{n_{DD}}{N} < \varepsilon_{DD} > \quad (2)$$

where N is the number of residues in the protein sequence. We checked that $<\varepsilon_{OO}>$ is equal to -0.284 ± 0.025 a.e.u., $<\varepsilon_{DD}>$ is equal to 0.200 ± 0.030 a.e.u. and $<\varepsilon_{OD}>$ is equal to 0.036 ± 0.018 a.e.u.; it is evident, then, that a global reduction in contact energy per residue is mainly due to the contacts among order-promoting residues, and only marginally to the contacts among disorder-promoting ones. Moreover and interestingly, if one considers the sum of the secondary structure (short-range) energy per residue with the contact energy (long-range) per residue, then the distinction among ordered, disordered and twilight zone proteins is lost, as shown in supplemental figure 2. There the distributions of E_{MJ}, the MJ energy per residue, are shown for ordered, disordered and twilight zone proteins of the PDB, and it is clear that they tend to overlap, particularly with respect to what is shown for the contact energy per residue, in contrast, in supplemental figure 1. This indicates that the distinction between ordered, disordered and twilight zone proteins is mainly due to the contact Miyazawa-Jernigan energies, and then, that the consensus predictor S_{SU} is an effective filter of MJ contact energy per residue, in the space of protein sequences.

Protein lengths and amino acidic compositions

In figure 4 we report the logarithmic plot of the number of ordered, disordered and twilight zone proteins, versus the length of their sequences. The number of disordered proteins is about constant until sequences are shorter than 90 amino acids, then it tends to decrease as lengths increase. A similar trend is shown by proteins in the twilight zone; these proteins however tend to be slightly longer than disordered ones. In both groups the number of proteins longer than 200 amino acids is negligible. On the contrary, the number of ordered proteins reaches a maximum for sequence lengths comprised between 120 and 400 residues, and a considerable number of longer proteins is present. We conclude that disordered and twilight zone proteins are generally shorter than ordered ones. As said above, ordered proteins are rich in order-promoting amino acids. The previous result therefore indicates that proteins enriched in order-promoting amino acids are generally longer than the other proteins. We evaluated the correlation between the log-odds ratio S and the length of proteins in the

PDB. Pearson's correlation coefficient is equal to 0.82 for ordered proteins, 0.52 for those in the twilight zone and -0.10 for the disordered ones. Regression lines are shown in figure 5. The strong positive correlation between the S score and chain lengths, observed for ordered proteins, indicates that these proteins become enriched in order-promoting amino acids as their chain length increases. On the other hand, S is substantially uncorrelated with protein length for disordered and twilight zone proteins, indicating that the balance of order-promoting and disorder-promoting amino acids in these proteins is independent from chain length. The dependence of amino acidic composition on protein length has been the subject of recent studies. Bastolla *et al.* [39] note that the frequency of positively charged residues decreases with chain length, while the frequency of alanine and glycine increases, moreover they also observe that the frequency of V, I, L, M, F, T, W, D and E does not depend on protein length. We checked that these observations are confirmed in the PDB.

Our key observations, based on figure 5, are: i) ordered proteins are longer than disordered proteins, and tend to be enriched in order-promoting amino acids as their chain length increases; ii) twilight zone proteins appear to have also a slight tendency to increase the fraction of disorder promoting amino acids with length; iii) in disordered proteins there is a weak tendency to increase the fraction of disorder-promoting amino acids with length.

Since amino acidic composition modulates contact energy per residue, we investigate now the relation between contact energy per residue and length of protein sequences. From relations (1) and (2) in the previous section, we observe that contact energy per residue is decreased mainly by contacts among order-promoting amino acids. We checked that the number of contacts among order-promoting amino acids and the number of these amino acids are strongly correlated (Pearson's correlation coefficient is equal to 0.98). We found a similar correlation between the number of disorder-promoting amino acids and the number of contacts made among them and also between the number of order-promoting amino acids and the number of their contacts made with disorder-promoting amino acids. If a protein sequence contains a high number of order-promoting (disorder-promoting) amino acids, it contains also a high number of contacts among them. If we write $n_{OO} \sim n_O$ and $n_{DD} \sim n_D$, relation (2) becomes:

$$\frac{E_C}{N} \approx \frac{n_O}{N} < \varepsilon_{OO} > + \frac{n_O}{N} < \varepsilon_{OD} > + \frac{n_D}{N} < \varepsilon_{DD} > = f_O (< \varepsilon_{OO} > + < \varepsilon_{OD} >) + f_D < \varepsilon_{DD} > \tag{3}$$

where f_O is the frequency of order-promoting amino acids and f_D is the frequency of disorder-promoting amino acids. We can say therefore that a protein enriched in order-promoting amino acids has a low contact energy and, conversely, a protein enriched in disorder-promoting amino acids has a high contact energy (compare the values of $<\varepsilon_{OO}>$, $<\varepsilon_{DD}>$ and $<\varepsilon_{OD}>$ given above).

In figure 6 we graph cumulative probabilities *P*(l)* for a PDB protein to be longer than a fixed length, versus the length of the protein. The cumulative probability is evaluated for different ranges of contact energy per residue (see the last section of Methods). From this figure it is evident that only if the contact energy per residue is sufficiently low, it is possible to observe in the PDB stable long proteins having that energy content. In contrast, considering the cumulative probabilities corresponding to the energy ranges typical of disordered and twilight zone proteins (black, red and blue lines) is evident that there is a cut-off length.

To be more quantitative, in table 1 we collect, for various contact energy ranges the values of L_{20}, a threshold corresponding to a cumulative probability of 20%. From the table it is evident that proteins with contact energy per residue below -0.24 a.e.u. have a L_{20} greater than 200 while in proteins with higher energies this characteristic length tends to decrease, indicating that long proteins with high conformational energy per residue are rarely observed in the PDB. Since proteins enriched in disorder-promoting amino acids tend to have high conformational energies per residue, this can explain why the examples of disordered proteins in the PDB are, in general, quite short.

Disulfide bridges stabilize short proteins in the twilight zone

We have previously mentioned that some proteins belonging to the twilight zone between order and disorder are characterized by a short chain length (less than 100 residues) and a quite low conformational energy per residue (see figure 3). To understand the low energy of these proteins, we estimated their amino acidic frequencies and compared them with those of the other proteins in the PDB, of the same length. The result is reported in figure 7. We can see that twilight zone proteins with conformational energy lower than -0.4 a.e.u. are quite rich in cysteines and glycines. Then, these are the amino acids responsible for the decreasing of the conformational energy of these proteins. Being cysteine-rich suggests that these proteins reduce their potential energy through disulfide bridges. To validate this hypothesis, we checked that the correlation between the number of cysteines and the number of disulfide bridges in the proteins of the twilight zone is 0.96, against a correlation of 0.88 exhibited by ordered proteins. This points out that about all cysteines in the twilight zone proteins form disulfide bridges, differently from ordered proteins where we find a large number of cysteine that are not saturated in this covalent bond. Our result is in line with that reported by Bastolla *et al.* [39] who also point out that the frequency of cysteines decreases with chain length and that in short proteins cysteines often link themselves in disulfide bridges. It is then reasonable to conclude that twilight zone proteins are rich in disulfide bridges that compensate their intrinsic tendency to remain unstable.

Discussion

Let us now synthesize and discuss the observations in this paper, devoted to the distribution of disorder in the PDB and aiming at discriminating the relative importance of amino acidic compositions and structural energetic for the folding propensity of proteins. Through the S_{SU} consensus score we classify uncomplexed PDB proteins into three groups: ordered, disordered and twilight zone. These three groups differ in their amino acidic composition, as it is evident from the histograms of amino acidic frequencies and from the distributions of the log-odds ratio S. Proteins classified as ordered have an amino acidic composition typical of folded proteins, while proteins classified as disordered have an amino acidic composition typical of unfolded ones (e. g. proteins in the DisProt [14]). Nevertheless, both groups of proteins have a tertiary structure experimentally characterized and deposited in the PDB. Amino acidic composition, therefore, is not sufficient to determine whether a protein acquires a tertiary structure or not, but, as shown in the Results and as we further discuss below, it modulates the mechanical and thermodynamic stability of protein structures.

The MJ potential is here used to estimate the conformational energy of proteins at a residue level [34-37]. This choice is sound, since this potential is commonly used in several bioinformatics tasks (e.g., in threading algorithms and in discriminating native from non-native tertiary structures ab-initio inferred from sequences). Moreover, it is well known that conformational energies are related to the thermodynamic stability of proteins [23, 39]. Let us now discuss this point, with specific reference to the MJ potential. Consider the unfolding Gibbs free energy difference of a generic protein as:

$$\Delta G = \Delta E - T \Delta S \quad (4)$$

where $\Delta E = E_{folded} - E_{unfolded}$, and $\Delta S = S_{folded} - S_{unfolded}$. We can assume that the conformational energy of the unfolded state is zero, then $\Delta E = E_{folded}$. Entropy change is due to two contributions: entropy change due to transfer of amino acids into the protein core $\Delta S_{transfer}$ and entropy change due to re-ordering and packing of amino acids ΔS_{conf}. As pointed out by Ghosh *et al.* [33], both terms scale with the number N of residues in the polypeptide chain. Therefore, we can write:

$$\Delta G \approx E_{ground} - NT\Delta S \quad (5)$$

From relation (5) we see that higher conformational energies E_{folded} imply higher Gibbs free energies and then a lower thermodynamic stability.

We verified that the MJ potential can be assumed as a coarse-grained estimator of Gibbs free energy. We considered 67 proteins selected by Robertson *et al.* [40], for which we have enthalpy and entropy of unfolding, determined by differential calorimetry or optical spectroscopy, and we found a significant correlation coefficient equal to -0.71 between MJ energy and both enthalpy and Gibbs free energy of unfolding. As recalled above, MJ potential is the sum of two terms: secondary structure energy and contact energy. We have shown that secondary structure energy per residue does not discriminate ordered from disordered proteins, whereas contact energy per residue is generally higher in disordered proteins than in ordered ones. We conclude that ordered and disordered proteins differ in their long-range interactions, and these are modulated by amino acidic compositions: ordered proteins have lower contact energy per residue than the others. Since conformational energy correlates with Gibbs free energy, we can deduce that ordered proteins are more stable than the other proteins. Secondary structure energy, on the other hand, does not depend on amino acidic composition. Interestingly, the sum of secondary structure and contact energy per residue does not discriminate between ordered proteins and disordered ones. This result points to the important role of short-range interactions in stabilizing protein structures.

Actually, the relative importance, for the stability of protein structures, of the hydrophobic effect and of backbone hydrogen bonds is matter of debate. Hydrophobic effect is commonly considered as the driving force of the folding process [15-19]. In this view, the protein free energy landscape is considered as funnel-shaped, and in the folding process a protein reduces its free energy descending the gradient of the landscape toward the folded ground state. The existence of amino acids with different hydrophobicities implies a rugged energy landscape with several local metastable states; these states are thought to be circumvented by evolution through the selection of minimally frustrated amino acidic sequences and thus smoothing the free energy landscape [16, 19]. From this point of view, evolution in the space of amino acidic composition plays a fundamental role in the stabilization of protein structures. A

different point of view consider the protein backbone as fundamental to build a stable tertiary structure, while amino acid side-chains have a minor role, limited to a local selection between alpha-helices or beta-sheets. In this scheme hydrogen bonds play a fundamental role, as shown by Rose and collaborators [41-44]. These authors give several arguments to support their hypothesis. In first instance, they have been able to construct topology of several proteins from secondary structure alone, without considering long-range interactions [43, 44]. Second, they report that the hydrophobic effect has been usually over-estimated [42]. Moreover, they point the attention on the fact that a polypeptide chain can be driven toward the folded or the unfolded state by the presence of denaturing or protecting osmolytes. [41, 42]. Denaturing osmolytes stabilize the unfolded state favouring residue-solvent interactions, whereas protecting osmolytes stabilize the folded state favouring intramolecular interactions. This effect is due mainly to the backbone, so osmolytes modulate backbone hydrogen bonds formation: the polypeptide chain folds if it gains a sufficient number of backbone hydrogen bonds, and this can be favoured or disfavoured by osmolytes [39, 40]. Our results on the MJ stabilization of structures indicate that short-range secondary structure energy plays the major role in the formation of a stable fold, supporting the hypothesis that backbone hydrogen bonds are indeed fundamental. Secondary structure energy must be considered to decide whether a sequence has a structure. Amino acidic composition modulates only long-range interactions; therefore the analysis of the amino acidic composition of a protein is not sufficient to determine whether it gets a tertiary structure since it does not give us information about its secondary structure energy. Disordered proteins in the PDB have an amino acidic composition that, as an effect of long-range interactions, makes them less stable than ordered ones. But they can nonetheless have a fold, since their global mechanical energy can be stabilized by secondary structure energy, mainly due to backbone hydrogen bonds.

As mentioned in the introduction, Szilagyi *et al.* investigated the relation between conformational energy and protein length by means of lattice models [23]. In a nutshell, they suggest that short proteins may reach only a limited number of conformations, so they gain a stable conformation with more difficulty than longer proteins, which can exploit a higher number of available conformations. This means that short proteins have a lower probability to reach a contact map compatible with a fold. The number of residues therefore plays an important role to make a fold stable; in long proteins, amino acidic composition is sufficient to determine the conformational energy of a protein, while it is not sufficient in short proteins, where the order of the residues and the interactions among them are more important. We have shown that amino acidic composition modulates only the contact energy. On the other hand, we have shown the important role of secondary structure energy in reaching a stable fold. It is reasonable that secondary structure energy is highly dependent from the order of residues in the polypeptide chain, since the order of residues modulates dihedral angles among adjacent residues and these angles, in turn, modulate secondary structure energy.

Our results, however, suggest that there is a subtle relation among amino acidic compositions and protein length that must be further investigated. We have shown that the probability to have long proteins with high conformational energy is low; proteins enriched in disorder-promoting amino acids have high conformational energy, since disorder-promoting amino acids give a low contribution to the decrease of contact energy; this explain can explain why disordered proteins, rich of disorder-promoting amino acids, generally have short polypeptide sequences. It will be

interesting to further investigate the relation among amino acidic composition and protein length. We have observed also that a certain number of short proteins in the twilight zone tend to compensate their relative instability by means of disulfide bridges (see last sub-section of the Results). Conformational energy is reduced mainly by order-promoting amino acids, so they largely contribute to make the fold of a protein stable. Disordered proteins and those in the twilight zone tend to remain more unstable than ordered proteins due to their lower content of order-promoting amino acids. These proteins tend also to have short polypeptide chains. We have shown that the probability to have long chains decreases rapidly in proteins with high conformational energy per residue; since proteins depleted in order-promoting amino acids have higher conformational energy per residue, they are generally of short length. Moreover, there exists a family of short proteins, that we classify in the twilight zone between order and disorder, that use disulfide bridges to enhance their mechanical stability. The reason of this behaviour must be further investigated.

This paper is one of the first explorations of the distribution of structural disorder in the PDB. It is clear that the final goal of this research would be to understand how many different flavours of intrinsically disordered structures are distributed among different structural and functional classes and how disorder is evolutionarily conjugated with the need for functional stability. Let us just recall that we have screened 70684 structures classifying 75% of them as ordered and the remaining 15% as twilight zone and disordered. The problem then raises: is this distribution peculiar of the set of proteins sampled in the PDB or this percentages are representative of the general probabilities for a protein to be folded or unfolded? Just as a preliminary observation, we checked on TrEMBL database (release 2010_06) that of 1908802, 146774 entries (roughly 7%) have a negative S value and whereas 1762028 entries (about 93%) have a positive S value. On the other hand Orengo and co-workers [45] using Hidden Markov models have estimated that a fraction between 10% and 20% of protein sequences in a genome are singletons, homologically unrelated to other sequences within their own genome or in other genomes. Moreover, we have checked that a significant fraction of these singletons appears to have low secondary structure and therefore are candidates to be intrinsically disordered. This just to conclude that the percentage of disorder found in the PDB, seems to be consistent with the general expectation for proteins in general.

Conclusions

This work has been focussed on the relative role of amino acidic composition, protein lengths and mechanical and thermodynamic stability in determining whether a protein gets a stable fold or remains partially or intrinsically unstructured. We think to have gained some insight.

Through the analysis of the two terms of the MJ potential, we have shown that amino acidic composition modulates long-range interactions. Proteins enriched in disorder-promoting amino acids have a lower mechanical stability than proteins enriched in order-promoting amino acids, since their long-range interactions are weaker. Nonetheless, our study has shown that secondary structure energy gives an important contribution in making a fold stable. As previously discussed, this is in line with recent works stressing the importance of backbone hydrogen bonds for the stability of protein tertiary structures. Since the analysis of amino acidic composition does not give us information about this secondary structure energy, it is intrinsically insufficient to determine whether an amino acidic segment has a well-defined three-

dimensional conformation or not. When a disorder predictor classifies a protein segment as disordered, this segment can have a fold if it succeeds in making a sufficient number of backbone hydrogen bonds, and this information cannot be inferred from amino acidic composition; however, we can nonetheless say that if the tertiary structure exists, it has a low mechanical stability, since it is characterized by weak long-range interactions. This can explain why we find disordered proteins in the PDB, where we know that proteins have a tertiary structure experimentally characterized. Further studies are necessary to establish the relation about mechanical stability due to long-range interactions and dynamics of the protein.
It is important to point out that the previous observations do not mean that amino acidic composition has no role for a protein to have a stable structure. Our analysis has evidenced that proteins enriched in disorder-promoting amino acids, then with low mechanical stability, tend to be quite short.

Methods

Selection of proteins in the PDB

We extensively analysed all non-complexed proteins deposited in the Protein Data Bank (PDB) database [24]. To determine whether a protein is complexed with other macromolecules, we parsed protein annotation in the PDB searching for "COMPLEX" or "COMPLEXED" keywords. All proteins reporting these keywords were excluded from out dataset. To parse the file, we used the Molecules To Go web-server application (http://molbio.info.nih.gov/cgi-bin/pdb). In the total we selected 70684 proteins.

Strictly unanimous consensus score among predictors of natively unfolded proteins

The strictly unanimous consensus score S_{SU} is a consensus index among different predictors of natively unfolded proteins. It combines three predictors: mean pairwise energy [27, 28], *gVSL2* [26, 29, 30], and Poodle-W [31]. The protocol used to compute these indexes is described elsewhere [26]. It should be noted that pairwise energy between residues does not take into account whether the two residues are in contact in the protein structure [27] or not. Mean pairwise energy is simply the arithmetic mean of the global pairwise energy of the protein sequence, and therefore it is an estimate of the pairwise energy per residue. *gVSL2* is the arithmetic mean of the disorder scores obtained through *VSL2* [29, 30], a good performing disorder predictor. Poodle-W [31] evaluates whether a protein sequence is disordered through a spectral graph transducer [32].
The strictly unanimous consensus score S_{SU} considers the three predictions by mean pairwise energy, *gVSL2* and Poodle-W. If the three predictors agree in classifying a protein as folded, S_{SU} classifies the protein as ordered. If the three predictors agree in classifying a protein as unfolded, S_{SU} classifies the protein as disordered. If at least two predictors disagree in classifying a protein as folded or unfolded, S_{SU} assigns the protein to the twilight zone.

Log-odds ratio of the likelihoods that a sequence has amino acidic composition typical of folded and unfolded proteins

Referring to a simple probabilistic model, one assumes to have reliable estimates of the probability of occurrence of each amino acid *a* in folded and unfolded proteins $\{\pi_a^{(F)}\}$ and $\{\pi_a^{(U)}\}$, respectively where *a* runs over all amino acid labels. We estimated these probabilities on the set of folded and natively unfolded proteins selected by Shimizu *et al.* to test Poodle-W [31]. Then, a folded protein can be thought of as if its sequence were sampled from $\{\pi_a^{(F)}\}$ through a sequence of independent extractions. The likelihood that a sequence has amino acidic composition typical of a folded protein is:

$$L_F = \prod_{a=1}^{20} \left(\pi_a^{(F)}\right)^{n_a}$$

where $\pi_a^{(F)}$ is the probability of amino acid *a* and n_a is the occurrence of amino acid *a* in the sequence. The probabilistic model implicit in the above definition is a 0-order Markov chain. Similarly we can define L_U by using $\pi_a^{(U)}$. L_F/L_U is the ratio of the likelihoods, for a given sequence and through its amino acidic composition $\{n_a\}$, to have been generated from $\{\pi_a^{(F)}\}$ and $\{\pi_a^{(U)}\}$, respectively. The log-odds ratio of a given sequence is then defined as:

$$S = \sum_{a=1}^{20} n_a \cdot \ln\left(\frac{\pi_a^{(F)}}{\pi_a^{(U)}}\right)$$

Order-promoting amino (disorder-promoting) acids contribute with positive (negative) terms to S, since their ratios $\pi_a^{(F)}/\pi_a^{(U)}$ are bigger (lower) than one. Therefore, S is positive (negative) if the protein is composed mainly by order-promoting (disorder-promoting) amino acids. When a protein is composed by approximately the same number of order-promoting and disorder-promoting amino acids, its S score has a value close to zero.

Miyazawa-Jernigan potential

MJ potential is used in this paper to estimate mechanical stability of proteins. We used the definition of the potential following ref. [34].

Miyazawa-Jernigan potential can be expressed as the sum of two terms: the secondary structure energy and the contact.

Secondary structure energy is an estimate of the interaction between a tripeptide in a given conformation and a residue located at its centre. We indicate with (s_{p-1}, s_p, s_{p+1}) the conformation of a residue at position *p* in the sequence, where *s* indicates one of five possible conformations: α, β, proline β, left-handed α, left-handed β [34]. The conformation adopted by the residue depends on its peptide dihedral angles Φ and Ψ. We estimate secondary structure energy as [34]:

$$E_S = \sum_p \left[e\left(s_{p-1}, s_p, s_{p+1}\right) + \delta e\left(s_{p-1}, s_p, s_{p+1}, i_p\right)\right]$$

where $e(s_{p-1}, s_p, s_{p+1})$ is the energy of the backbone and $\delta e(s_{p-1}, s_p, s_{p+1}, i_p)$ is the interaction between the tripeptide and the side-chain of the residue located at its centred. The sum is over all amino acids in the protein sequence. Note that, differently from [26], we chose to include also the backbone-backbone interactions since we expect that they contribute to the mechanical stability of the protein.
Contact energy is an estimate of the interactions among residues in contact in the protein structure. Two residues are considered in contact whether their geometric centres are less distant than 6.5 Å.
We estimate contact energy as [34]:

$$E_C = \frac{1}{2}\sum_p \sum_j n_{i_p j} \cdot \left(e_{i_p j} - e_{rr}\right)$$

where $n_{i_p j}$ is the number of residues of type *j* in contact with residue of type *i* at position *p*, $e_{i_p j}$ is the pairwise energy between residue of type *i* and type *j* and e_{rr} is a reference collapse energy.

Finally, the total conformational energy is the sum of the secondary structure and contact energy, as in the following:

$$E = E_S + E_C$$

Conformational energy per residue is simply the conformational energy divided by the number of residues in the protein sequence.

Probability for a protein to be longer than a fixed length

Let $p(l)$ be the probability distribution, for proteins, to have length *l*. This distribution can be evaluated operatively as the ratio between the number $n(l)$ of observed proteins with length between *l* and $l+\Delta l$ and the number of all proteins considered.

$$p(l) = P(l < \lambda < l + dl) \approx \frac{n(l)}{N}$$

The probability, for a protein, to be shorter than *l* is distributed following the cumulative probability distribution $P(l)$ defined as:

$$P(l) = P(\lambda < l) = \int_{-\infty}^{l} p(\lambda) d\lambda$$

This probability can be estimated as:

$$P(l) \approx \frac{1}{N}\sum_{\lambda=0}^{l} n(\lambda).$$

Conversely, the probability for a protein to be longer than *l* can be expressed as:

$$P^*(l) = P(\lambda \geq l) = 1 - P(l) = 1 - \int_{-\infty}^{l} p(\lambda)d\lambda = \int_{l}^{+\infty} p(\lambda)d\lambda$$

and this is the quantity estimated with the proteins in the PDB, which is plotted in figure 6. In table 1 are reported, for several ranges of contact energy per residue, values of L_{20} a length which is defined by the following equation:

$$P^*(L_{20}) = 0.2$$

Authors' contributions

All authors participated in study design, research and manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

References

1.Tompa P: *Structure and Function of intrinsically disordered proteins*. CRC Press 2010.

2.Wright P, Dyson HJ: **Intrinsically unstructured proteins: re-assessign the protein structure-function paradigm.** *J. Mol. Biol.* 1999, 293: 321-331.

3.Dyson HJ, Wright P: **Intrinsically unstructured proteins and their functions**. *Nat. Rev. Mol. Cell. Biol*. 2005, 6: 197-208

4.Dunker A, Lawson J, Brown C, Romero P, Oh J, Oldfield C, Campen A, Ratliffl C, Hipps K, Ausio J, Nissen M, Reeves R, Kang C, Kissinger C, Bailey R, Griswold M, Chin W, Garner E, Obradovic Z: **Intrinsically disordered proteins.** *J. Mol. Graph. Model* 2001, 19: 26-59

5.Demchenko AP: **Recognition between flexible protein molecules: induced and assisted folding.** *J. Mol. Recognit.* 2001, 14: 42-61

6.Uversky VN: **Natively unfolded proteins: a point where biology waits for physics.** *Protein Sci.* 2002, 11: 739-756

7.Tompa P: **Intrinsically unstructured proteins.** *TRENDS Biochem. Sci.* 2002, 27: 527-533

8.Fink AL: **Natively unfolded proteins.** *Curr. Opin. Struct. Biol.* 2005, 15: 35-41

9.Uversky VN, Oldfield CJ, Dunker AK: **Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signalling.** *J. Mol. Recognit.* 2005, 18: 343-384

10.Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z., Dunker AK: **Intrinsic disorder in cell-signalling and cancer-associated proteins.** *J. Mol. Biol.* 2002, 323: 573-584

11.Uversky VN, Oldfield CJ, Dunker AK: **Intrinsically disordered proteins in human diseases: introducing the D^2 concepts.** *Annu. Rev. Biophys.* 2008, 37: 215-246

12.Uversky VN: **Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?** *Cell and Mol. Life Sciences* 2003, 60: 1852-1871

13.Uversky VN, Fink AL: **Conformational constraints for amyloid fibrillation: the importance of being unfolded.** *Biochimica and Biophysica Acta* 2004, 1698: 131-153
14.Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. 2006: **DisProt: the database of disordered proteins.** *Nucleic Acids Res*. 2007, 35:D786-93.
15.Bryngelson JD, Wolynes PG: **Spin glasses and the statistical mechanics of protein folding.** *Proc. Natl. Acad. Sci. USA* 1987, 84: 7524-7528
16.Dill KA: **Dominant forces in protein folding.** *Biochemistry* 1990: 29: 7133-7155
17.Wolynes PG: **Folding funnels and energy landscapes of larger proteins within the capillarity approximation.** *Proc. Natl. Acad. Sci. USA* 1997, 94: 6170-6175
18.Dill K.A.: **Polymer principles and protein folding.** *Protein Sci.* 1999, 8: 1166-1180
19.Finkelstein AV, Galzitskaya OV: **Physics of protein folding.** *Physics of Life Review* 2004, 1: 23-56
20.Rose G. (Ed): **Unfolded proteins.** In *Advances in protein chemistry* 2002, 62: 1-398
21.Banavar JR, Hoang TX, Maritan A, Seno F, Trovato A: **Unified perspective on proteins: a physics approach.** *Phys. Rev. E* 2004, E70:041905
22.Romero P, Obradovic Z, Xiaohong L, Gamer EC, Brown CJ, Dunker AK: **Sequence complexity of disorder proteins.** *Proteins* 2001, 42: 38-48
23.Szilagyi A, Gyorffy D, Zavodszky P: **The twilight zone between protein order and disorder.** *Biophysical J.* 2008, 95: 1612-1626
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov JN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res.* 2000, 28: 235-242
25.Le Gall T, Romero PR, Cortese MS, Uversky VN, Dunker AK: **Intrinsic disorder in the Protein Data Bank.** *J. Mol. Struct. Dyn.* 2007, 24: 325-341
26.Deiana A, Giansanti A: **Predictors of natively unfolded proteins: unanimous consensus score to detect a twilight zone between order and disorder in generic datasets.** *BMC Bioinformatics* 2010, 11: 198
27.Dosztanyi Z, Csimok V, Tompa P, Simon I: **The pairwise energy content estimated from amino acid composition discriminate between folded and intrinsically unstructured proteins.** *J. Mol. Biol.* 2005, 347: 627-639
28.Dosztanyi Z, Csizmok V, Tompa P, Simon I. **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**. *Biinformatics 2005, 21: 3433-3434*
29.Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK: **Exploiting heterogeneous sequence properties improves prediction of protein disorder.** *Proteins* 2005, 7: 176-182
30.Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z: **Length-dependent prediction of protein intrinsic disorder.** *BMC Bioinformatics 2006, 7: 208-225*
31.Shimizu K, Muraoka Y, Hirose S, Tomii K, Noguchi T: **Predicting mostly disordered proteins by using structure-unknown protein data.** *BMC Bioinformatics* 2007, 8: 78-92
32.Joachims T: **Transductive learning via Spectral Graph Transducer.** *Proceeding of International Conference on Machine Learning* 2003: 143-151
33.Ghosh K, Dill K.: **Computing protein stabilities from their chain length.** *Proc. Natl. Acad. Sci. USA* 2009, 106: 10649-10654

34.Miyazawa S, Jernigan RL: **Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space.** *Proteins* 2003, 50: 35-43
35.Miyazawa S, Jernigan RL: **Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues.** *Proteins* 1999, 34: 49-68
36.Miyazawa S, Jernigan RL: **Evaluation of short-range interactions as secondary structure energies for protein Fold and sequence recognition.** *Proteins* 1999, 36: 347-356
37.Miyazawa S, Jernigan RL: **An empirical energy potential with a reference state for protein fold and sequence recognition.** *Proteins* 1999, 36: 357-369
38.Higgs PG, Attwood TK.: *Bioinformatics and molecular evolution.* Blackwell Publishing 2006
39.Bastolla U, Demetrius L: **Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds.** *Protein Eng. Des. Sel.* 2005, 18: 405-415
40.Robertson AD, Murphy KP: **Protein structures and the energetics of protein stability.** *Chem. Rev.* 1997, 97: 1251-1267
41.Rose GD, Fleming PJ, Banavar JR, Maritan A: **A backbone-based theory of protein folding.** *Proc. Natl. Acad. Sci. USA* 2006, 103: 16623-16633
42.Bolen DW, Rose GD: **Structure and energetics of the hydrogen-bonded backbone in protein folding**. *Annu. Rev. Biochem.* 2008, 77: 339-362
43.Gong H, Fleming PJ, Rose GD: **Building native protein conformation from highly approximate backbone torsion angles.** *Proc. Natl. Acad. Sci. USA* 2005, 102: 16227-16232
44.Fleming PJ, Gong H, Rose GD: **Secondary structure determines protein topology**. *Protein Sci.* 2006, 15: 1829-1834
45.Orengo CA, Thornton JM: **Protein families and their evolution . a structural perspective.** *Annu. Rev. Biochem.* 2005, 74: 867-900

Tables

Table 1 - Threshold lengths L_{20} as a function of the range contact energy per residue

E_c, Contact energy per residue	L_{20}*
<-0.80	34
[-0.80 , -0.69]	351
[-0.68 , – 0.58]]	365
[-0.57, -0.47]	248
[-0.46, -0.36]	323
[-0.35, -0.25]	362
[-0.24, -0.14]	208
[-0.13, -0.03]	54
[-0.02, 0.08]	90
[0.09, 0.19]	66
[0.20, 0.30]	6

* For the equation defining L_{20} see the last section of Methods.

Figures

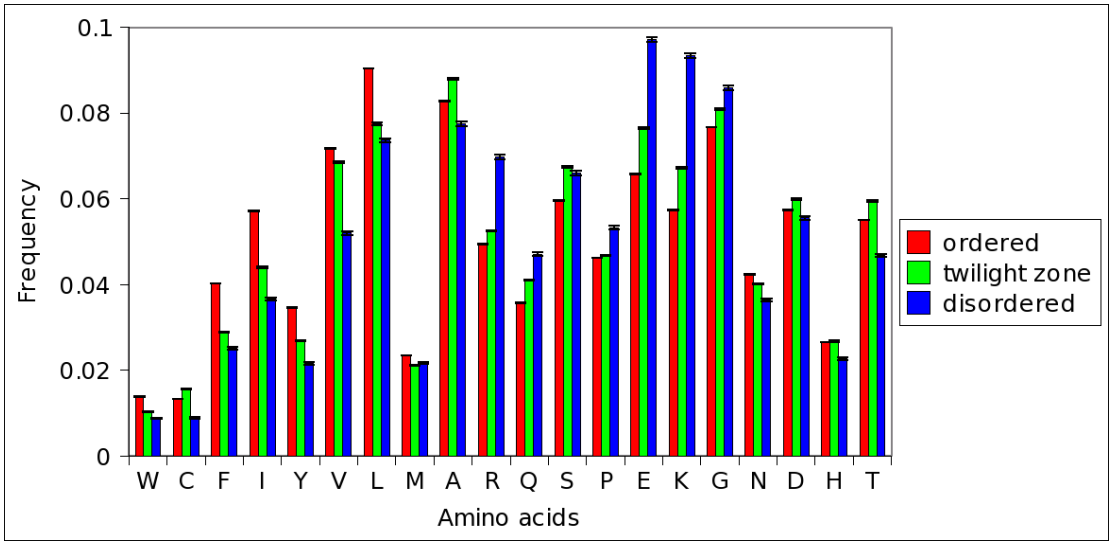


Figure 1 - Frequencies of amino acids in the three flavours of disorder in the PDB

Amino acidic frequencies of ordered, disordered and twilight zone proteins in the PDB, as selected by the strictly unanimous consensus score S_{SU} (see text). Red bars refer to ordered proteins, blue to disordered and green to proteins in the twilight zone. Ordered proteins are enriched in order-promoting amino acids whereas disordered proteins are enriched in disorder-promoting amino acids. Proteins in the twilight zone have amino acidic composition intermediate between those of ordered and disordered proteins, with the exceptions of residues C, A, S, D and T that are more present in this class of proteins. The error bars are standard deviations from the mean.

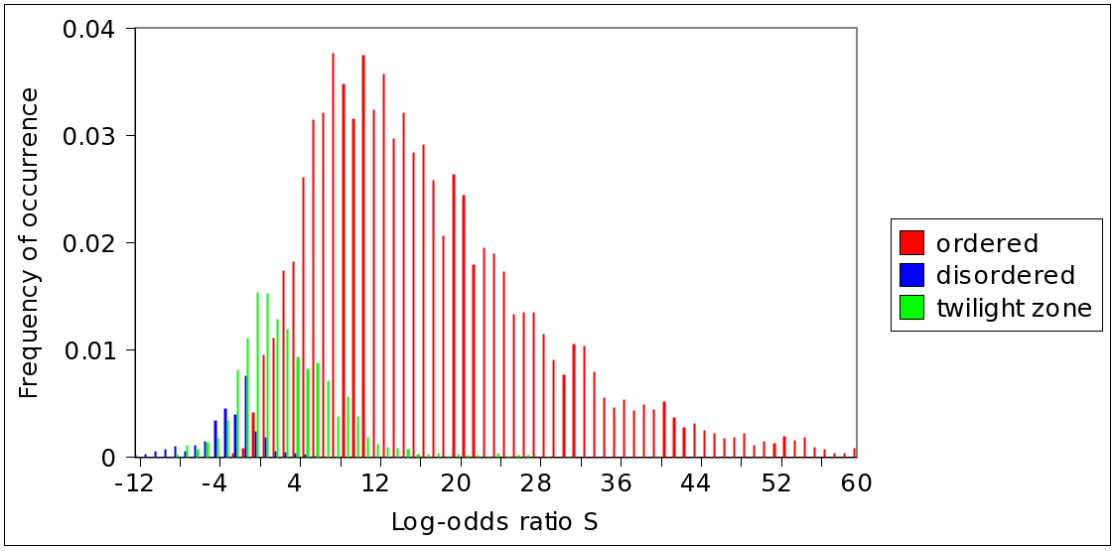


Figure 2 - Distribution of the log-odds ratio S of ordered, disordered and twilight zone proteins; as selected in the PDB by the strictly unanimous consensus score S_{SU}

Ordered proteins have mainly positive values of S, disordered proteins mainly negative, and twilight zone proteins have S-values intermediate between those of the other two classes, centred around 0. Twilight zone proteins are then characterized by a balanced composition of order-promoting and disordered-promoting amino acids.

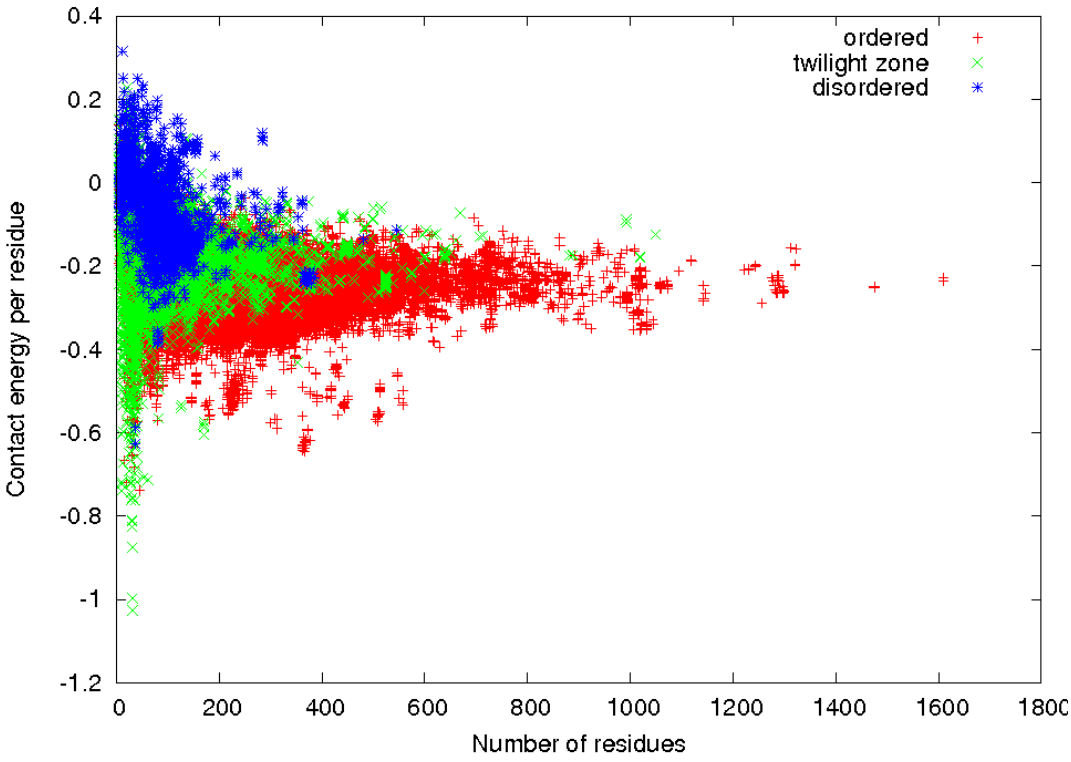


Figure 3 - Miyazawa-Jernigan contact energy (long-range) per residue of the ordered, disordered and twilight zone proteins in the PDB, as a function of their length

Red points refer to ordered proteins, blue points to disordered proteins and green to proteins in the twilight zone. Ordered proteins are the most stable ($\langle E_c \rangle$ = -0.2511 ± 0.0003), disordered proteins are the least stable ($\langle E_c \rangle$ = -0.058 ± 0.003). Proteins in the twilight zone have values of contact energy per residue distributed between those of ordered and disordered proteins ($\langle E_c \rangle$ = -0.165 ± 0.001). Note the presence of proteins in the twilight zone with particularly low contact energies per residue.

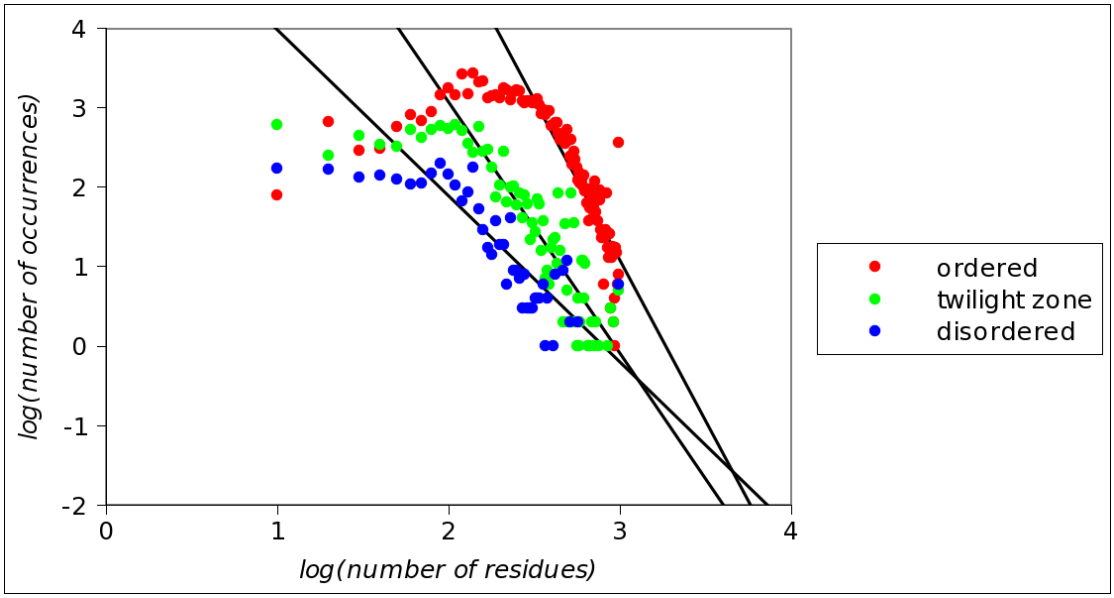


Figure 4 – Double logarithmic plot of the number of ordered, disordered and twilight zone proteins in the PDB, versus their lengths

The number of disordered proteins and of those in the twilight zone decreases starting from chain length of 90 – 140 amino acids. The number of ordered proteins exhibits a maximum for chain length between 120 and 400 amino acids. Ordered proteins are generally longer than proteins in the twilight zone and disordered ones.

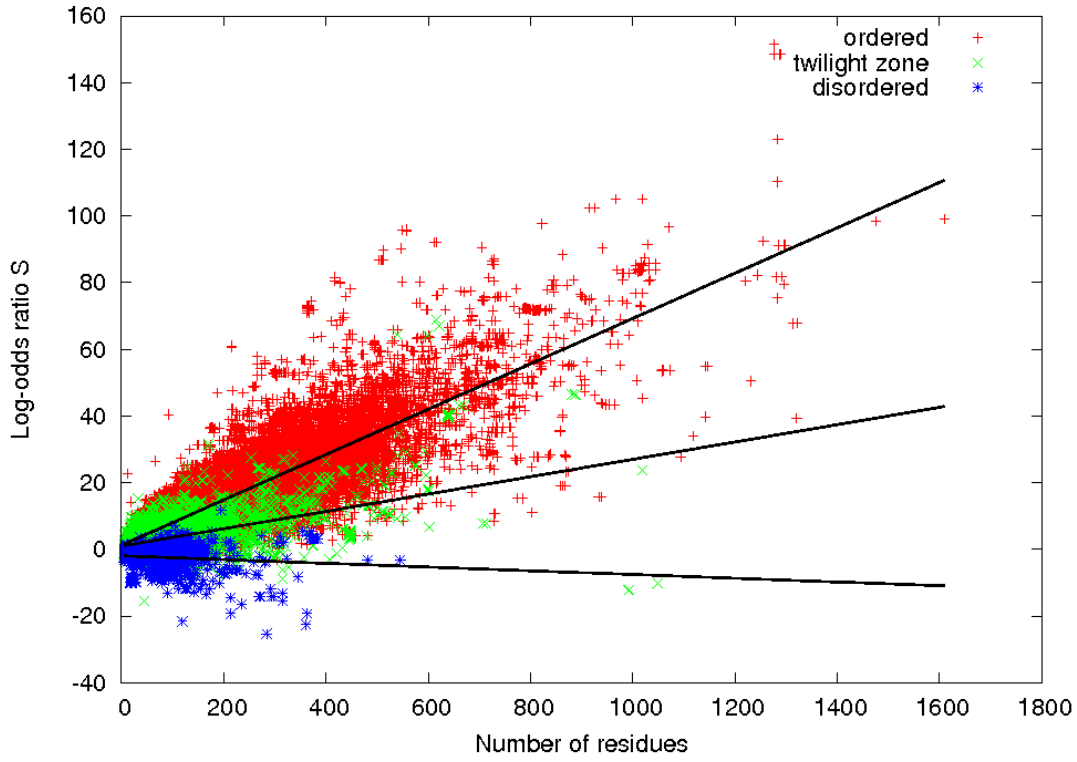


Figure 5 - Log-odds ratios S of ordered, disordered and twilight zone proteins in the PDB as a function of the respective protein lengths

Regression lines between S and protein length are reported as solid lines in black. Pearson's correlation coefficients are: 0.82 for ordered, -0.10 for disordered and , 0.52 for proteins of the twilight zone. This observation clearly shows that structured proteins, rich in order-promoting amino acids, can be longer than proteins with higher relative content of disorder-promoting amino acids.

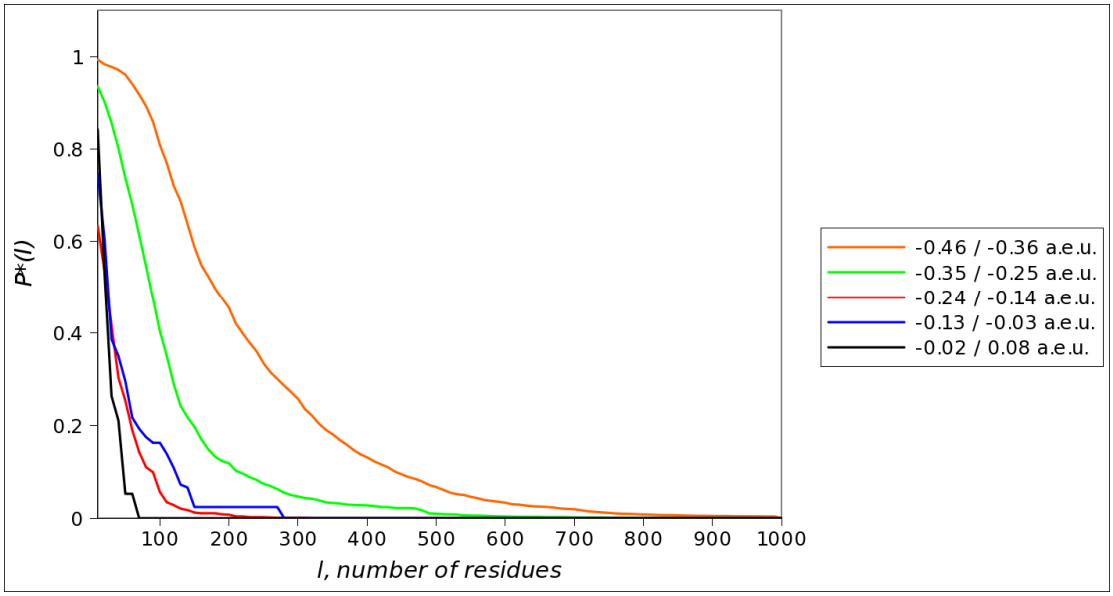


Figure 6 - Cumulative probability *P*(l)* for an uncomplexed protein in the PDB of being longer than a fixed length *l* (see Methods), for several ranges of MJ contact energies per residue.

For the sake of illustration, note that the probability to have proteins longer than 100 residues is close to 10% in the range of contact energy per residue [-0.25, 0.14] and it is above 60% for proteins in the remaining lowest energy ranges.

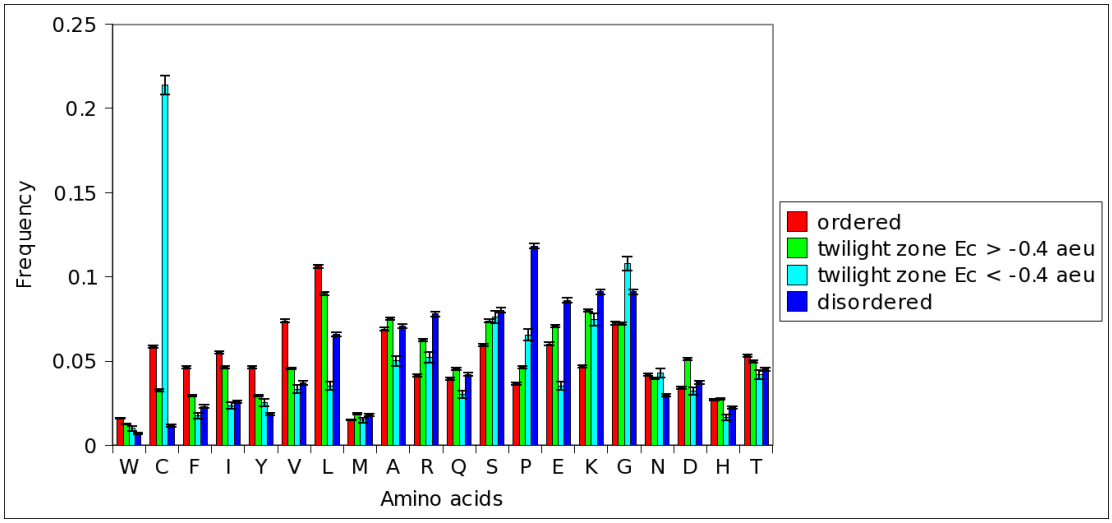
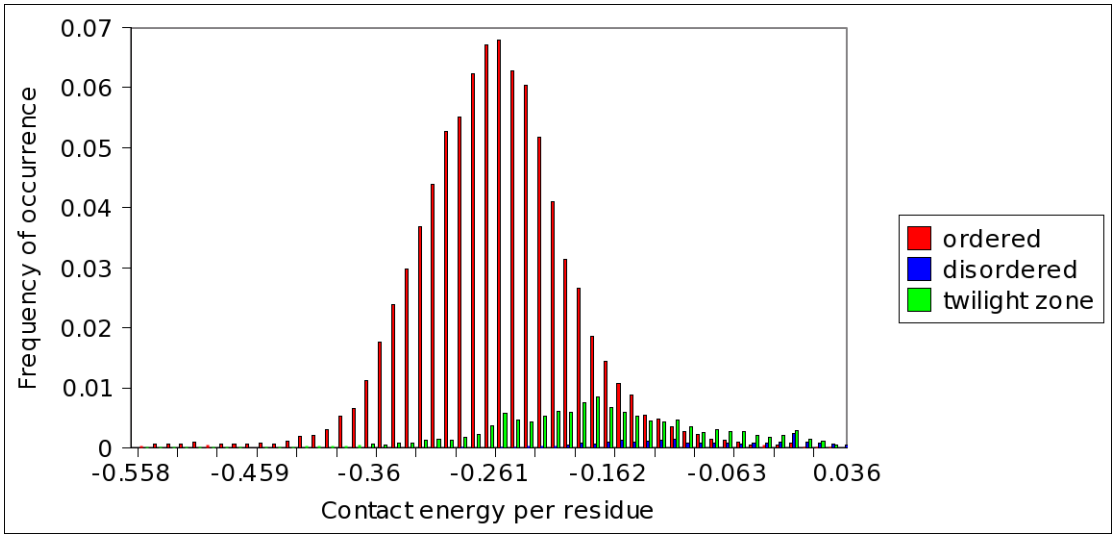


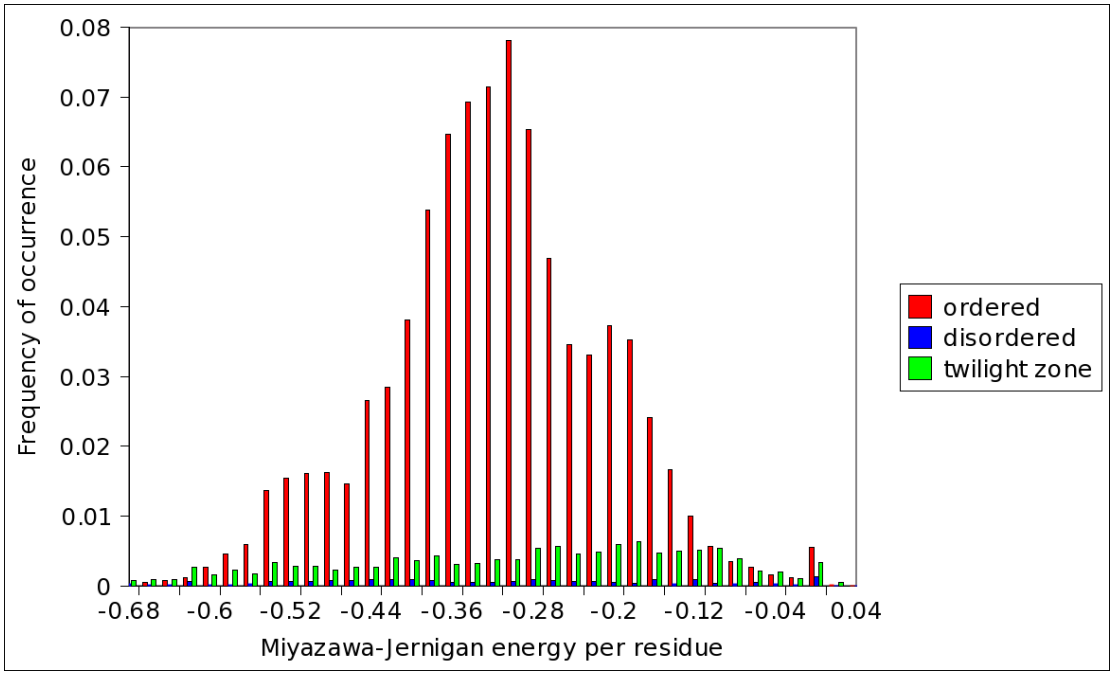
Figure 7 - Frequencies of amino acids in the ordered, disordered and twilight zone proteins shorter than 100 residues.

Red bars refer to ordered proteins, blue bars refer to disordered proteins, green bars refer to proteins of the twilight zone with a contact energy per residue higher than -0.4 arbitrary energy units (a.e.u), light blue bars refer to proteins of the twilight zone with a contact energy per residue lower than -0.4 a.e.u. Note that the latter proteins are quite rich in cysteines and glicines.



Supplemental figure 1 - Distribution of Miyazawa-Jernigan *contact energy* per residue for ordered, disordered and twilight zone proteins in the PDB

Red bars refer to ordered proteins, blue to disordered and green to proteins in the twilight zone. Ordered proteins are the most stable with respect to E_c ($\langle E_c \rangle$ = -0.2511 ± 0.0003), disordered proteins are the least stable ($\langle E_c \rangle$ = -0.058 ± 0.003). Proteins in the twilight zone have values of contact energy per residue distributed between those of ordered and disordered proteins ($\langle E_c \rangle$ = -0.165 ± 0.001).



Supplemental figure 2 - Distribution of the Miyazawa-Jernigan energy ($E_{MJ}=E_S+E_C$) per residue for ordered, disordered and twilight zone proteins in the PDB

Red bars refer to ordered proteins, blue bars to disordered proteins, green bars to proteins in the twilight zone. Ordered proteins are the most stable ($<E_{MJ}>$ = -0.3072± 0.0005), disordered proteins are the least stable ($<E_{MJ}>$ =-0.320±0.007). Proteins in the twilight zone have values of E_{MJ} distributed between those of ordered and disordered proteins ($<E_{MJ}>$ = -0.293± 0.002).