Functional and Early Folding Residues are separated in proteins to increase evolvability and robustness

Sebastian Bittrich[1,2*], Michael Schroeder[2], Dirk Labudde[1]

1 University of Applied Sciences Mittweida, Technikumplatz 17, 09648, Mittweida, Germany
2 Biotechnology Center (BIOTEC), TU Dresden, Tatzberg 47-49, 01307, Dresden, Germany

* sebastian.bittrich@hs-mittweida.de

Abstract

The three-dimensional structure of proteins captures evolutionary ancestry, and serves as starting point to understand the origin of diseases. Proteins adopt their structure autonomously by the process of protein folding. Over the last decades, the folding process of several proteins has been studied with temporal and spatial resolution which allowed the identification of so-called Early Folding Residues (EFR) in the folding process. These structurally relevant residues become affected early in the folding process and initiate the formation of secondary structure elements and guide their assembly.

Using a dataset of 30 proteins and 3,337 residues provided by the Start2Fold database, discriminative features of EFR were identified by a systematical characterization. Therefore, proteins were represented as graphs in order to analyze topological descriptors of EFR. They constitute crucial connectors of protein regions which are distant at sequence level. Especially, these residues exhibit a high number of non-covalent contacts such as hydrogen bonds and hydrophobic interactions. This tendency also manifest as energetically stable local regions in a knowledge-based potential. Conclusively, these features are not only characteristic for EFR but also differ significantly with respect to functional residues. This unveils a split between structurally and functionally relevant residues in proteins which can drastically improve their evolvability and robustness.

The characteristics of EFR cannot be attributed to trivial features such as the accessible surface area. Thus, the presented features are novel descriptors for EFR of the folding process. Potentially, these features can be used to design classifiers to predict EFR from structure or to implement structure quality assessment programs. The shown division of labor between functional and EFR has implications for the prediction of mutation effects as well as protein design and can provide insights into the evolution of proteins. Finally, EFR allow to further the understanding of the protein folding process due to their pivotal role.

Author summary

Proteins are chains of amino acids which adopt a three-dimensional structure and are then able to catalyze chemical reactions or propagate signals in organisms. Without external influence, most proteins fold into their correct structure, and a small number of Early Folding Residues (EFR) have been shown to become affected at the very start of

the process. We demonstrated that these residues are located in energetically stable local conformations. EFR are in contact to many other residues of a protein and act as hubs between sequentially distant regions of a proteins. These distinct characteristics can give insights into what causes certain residues to initiate and guide the folding process. Furthermore, it can help our understanding regarding diseases such as Alzheimer's or amyotrophic lateral sclerosis which are the result of protein folding gone wrong. We further found that the structurally relevant EFR are almost exclusively non-functional. Proteins separate structure and function, which increases evolvability and robustness and gives guidance for the artificial design of proteins.

Introduction

Most proteins adopt their three-dimensional conformation autonomously during the process of protein folding [1, 2] which is strongly connected to protein design as well as the quality assessment of structures and *in silico* models [1, 3]. Various diseases are caused by misfolding or aggregation of proteins [4–7]. During the protein folding process, the denatured chain of amino acids passes a energetic barrier, called transition state, to form a compact and functional native structure [2].

How proteins fold is an open question [1]. There is a lack of experimental data describing which events or residues guide the folding process [8–10]. The protein sequence resembles the starting point and the three-dimensional structure captures the result of the protein folding process for a wide range of proteins, yet how they connect via the transition state is unclear. The unstable nature of the transition state hinders its experimental determination [11, 12]. Another hindrance for the understanding of the sequence-structure relation is that some proteins depend on chaperons to fold correctly [7].

The defined pathways model

Whether general folding patterns exist [13] and whether folding is stochastic or deterministic [14] remains to be answered – even within some protein families the process differs [15]. The defined pathways model proposes that small fragments fold first and then guide a step-wise assembly of further parts of the protein until the native structure is formed [14, 16, 17]. The process is believed to be deterministic and fragments folding first do so autonomously from other parts of the protein – no other region directly supports or hinders their formation [14, 17]. Which parts of the protein initiate the formation of local, ordered structures, e.g. secondary structure elements, is encoded in their sequence [18–23]. Consequently, these regions decrease in energy as well as entropy and stabilize the protein during the folding process [23, 24]. This also supports the observation that proteins fold cotranslationally as they are being synthesized by a ribosome and stabilizing long-range contacts cannot be formed yet [25]. These local structures form long-range contacts and assemble the global structure [14, 18, 22, 26, 27]. The formation of a native structure causes a further decrease in free energy [3, 17, 28]. Long-range contacts are especially important for the stability of the hydrophobic core of the native structure [29].

Identifying Early Folding Residues during protein folding

In recent years, various experimental strategies [30–33] were established which can identify residues crucial for the folding process. Probably the most elegant approach to track the protein folding process with spatial and temporal resolution is pulse labeling hydrogen-deuterium exchange (HDX) [14, 29, 34–39]. The state of a protein can be

controlled e.g. by denaturants or temperature [35]. Starting from a denatured protein, folding conditions are gradually established until the protein refolded completely. The resulting folding trajectory can be studied by HDX. Depending on the state of the folding process, individual amino acids will be susceptible to or protected from an exchange of the hydrogen atom of their amide group. Residues become protected when their amide group is isolated from the solvent as the effect of other residues surrounding them. When the folding process affects a residues, its spatial neighborhood is altered. Where and when these exchanges occur is tracked by a downstream mass spectroscopy or nuclear magnetic resonance spectroscopy. Residues which are protected from the exchange at the earliest stages [14, 37–39] are called Early Folding Residues (EFR). Residues which became protected at later stages or not at all are referred to as Late Folding Residues (LFR). EFR were shown to initiate the folding process and the formation of secondary structure elements [39] or even larger autonomously folding units [14]. They tend to be conserved, but non-functional residues [40]. In contrast, LFR may be relevant during later stages of the folding process, implement protein function, or be mere spacers between protein regions.

The data obtained by HDX experiments is still difficult to interpret [41] and results of other experiments or techniques are tedious to compare [29, 39]. The Start2Fold database [39] provides an invaluable annotation of EFR which became protected early in a standardized manner [29]. In a previous study [38], EFR have been shown to exhibit lower disorder scores and higher backbone rigidity. Regions with relatively high backbone rigidity are likely to constitute ordered secondary structure elements and this tendency is manifested in local sequence fragments [19, 20, 38, 39, 42]. Especially aromatic and hydrophobic amino acids were linked to ordered regions of proteins [38]. Subsequently, it was shown that EFR are likely buried according their relative accessible surface area (RASA) and proposed that they are also the residues which form the greatest number of contacts in a structure [39]. EFoldMine [10] is a classifier that predicts EFR from sequence. Due to the nature of the trained models [10, 38], it is still unclear what the relation between sequence and structure is and if EFR cause their surroundings to fold first or vice versa [23].

Representing proteins by Energy Profiling and graphs

A protein's native structure exhibits minimal free energy [14]. Thus, knowledge-based potentials are a potent tool to describe the process of protein folding [28] and have been previously employed for the quality assessment of protein structures [3]. In an approach called Energy Profiling the surroundings of each residue are expressed as energy value. Low energy values occur for hydrophobic amino acids which are stabilized by many contacts. Thus, this approach is a valuable feature to assess the stability of individual residues as well as their interactions with their spatial neighborhood.

Individual residues can also be characterized in the context of protein structures by topological features derived from network analysis. Protein structures are represented as graphs: amino acids constitute the nodes and contacts between residues are represented as edges [12, 43–49]. There is a plethora of contact definitions and most are based on distance cutoffs between certain atoms of amino acids [50]. Graph representations of proteins were previously employed to describe residue flexibility [51] as well as residue fluctuation [43], protein folding [12, 46], structural motifs [52], and evolvability [49]. Furthermore, protein graphs were shown to exhibit the character of small world networks [12, 43–46] whereby a small number of residues has high connectivity and the average path length in the graph is small. Hydrophilic and aromatic amino acids were found to be crucial connectors in the graph – so-called hubs – which underlines their importance in the context of protein folding [53].

Graph representations of proteins also allow to assess whether proteins feature a

modular design [54, 55]. Reinvention is rarely observed in nature and whenever possible existing, established, and safe strategies are reused [56]. This can explain why the conceivable sequence and structure space is explored so little: by evolving established sequences, misfolding sequences or those prone to aggregation are avoided [56, 57]. This behavior is likely the result of a separation of residues relevant for folding and those relevant for protein function [40] such as ligand binding sites or active sites. Functional residues also were shown to exhibit distinct topological features [45]. A division of labor between fold and function increases robustness and evolvability of protein sequences [40, 49, 54, 55, 58] because functional residues can be changed without any impairment of the protein's stability and the fold can be improved without compromising function.

Motivation

The Start2Fold database [39] constitute a dataset of EFR [10, 14, 17, 23]. Previous studies considered a small number of proteins, whereas the 30 proteins of the Start2Fold database [39] allow a more robust characterization of EFR. Because the annotation of EFR is standardized, a workflow can be established to analyze also future results of HDX experiments added to the database.

It is unknown what sequence features causes particular residues to fold early and how these residues contribute to the formation of the native structure (Fig 1A). EFR are strongly connected to the defined pathways model and provide an opportunity to understand the driving forces behind the assembly of stabilizing local structures as well as the formation of tertiary contacts [14, 23].

Fig 1. Graphical abstract. (**A**) During the folding process, an extended protein chain passes the transition state and forms a native structure [2]. (**B**) Protein structures are represented as graphs to derive topological descriptors of residues. Amino acids constitute nodes, whereas residue contacts are represented as edges. EFR are structurally relevant residues which participate early in the folding process by forming local contacts to other residues. They are separated from functional residues which are primarily ligand binding sites and active sites as derived from UniProt [59]. EFR show a great number of long-range contacts which furnish the spatial arrangement of protein parts which are far apart on sequence level.

In this study, several novel structural features are employed for the characterization of EFR. Especially, the Energy Profiling approach, topological descriptors of protein graphs, and the explicit consideration of non-covalent contacts types provides a new level of information in order to describe the folding process. EFR exhibit lower, more stable energy values in their Energy Profile [3, 28]. A network analysis reveals that EFR are more connected to other residues and that they are located at crucial positions in the protein graph (Fig 1B). This distinct wiring to the rest of the protein is especially furnished by hydrophobic interactions. EFR are likely structurally relevant for the correct protein fold [10]. This information is used to demonstrate that proteins separate structurally relevant residues from functional residues (Fig 1B).

Results and Discussion

A previously described dataset [23] of 30 proteins and 3,377 residues is the basis of this study and summarized in S1 Table. 482 (14.3%) of the residues are labeled as EFR, the remaining residues are considered LFR.

To characterize EFR in more detail, various features were defined and compared to the values of LFR. EFR form a significantly greater number of contacts than their LFR counterparts (Fig 2A). The loop fraction is defined as the ratio of unordered secondary structure elements in a window centered on a particular residue [60]. Fewer unordered secondary structure elements can be found around EFR (Fig 2B), whereas LFR exhibit a higher propensity to occur in coil regions. EFR are on average closer to the centroid of a protein structure and are likely embedded in the hydrophobic core (Fig 2C). Analogously, they also tend to be more distant to the N- or C-terminus of the sequence than other residues and are likely buried regarding their RASA as per S2 Table.

Fig 2. General properties discriminative between EFR and LFR. (**A**) EFR form more contacts to their surroundings than LFR. (**B**) The loop fraction [60] is the ratio of unordered secondary structure elements which are observed in a windows of nine amino acids around a residue. EFR are more commonly surrounded by ordered secondary structure elements. (**C**) EFR are located significantly closer to the centroid of the protein than LFR.

The propensity of EFR to participate in more contacts and to occur in the core of a protein are in agreement with previous studies [14, 23, 38, 46]. The shift in loop fraction can also be attributed to these findings and is further substantiated by the fact that long ordered secondary structure elements tend to contain more EFR [23]. It has been reported that buried residues are more likely to be EFR [23, 29] which also explains why they are closer to the spatial centroid of a protein and more separated from sequence termini (S2 Table). Yet, all these factors cannot explain why some residues become EFR and others do not.

Early Folding Residues constitute stable local conformations

To assess the energetic contribution of EFR to the native structure, the proteins of the dataset were transformed by the Energy Profiling approach [3, 28]. Computed energy values of EFR are significantly lower than the values of LFR. A more detailed investigation of the computed energy values (Fig 3) shows that this trend can be observed for individual amino acids, but the change is insignificant for aspartic acid and isoleucine. Hydrophilic amino acids commonly feature high energy values, whereas the values for hydrophobic and aromatic amino acids are low. The changes in energy for amino acids with hydroxyl groups in their side chain such as serine and threonine are remarkable. This trend also manifests in sequence; thus, energy values predicted by sequence using the eGOR method [28] are also lower for these residues (see S2 Table). Regarding the average absolute contact frequencies, a EFR participates in 3.87 hydrogen bonds and forms 1.30 hydrophobic interactions to other residues. This constitutes a significant increase compared to LFR (see S2 Table).

Fig 3. Computed energy by amino acid. A knowledge-based potential [3, 28] was used to characterize the surrounding of each residue. Hydrophobic and aromatic amino acids have a high tendency to be located in the buried core of a protein. Hydrophilic and polar amino acids prefer to be exposed to the polar solvent. This tendency is reflected by low and high average energy values respectively. The distribution of energy values of EFR always exhibits a lower median than LFR. Significance in change is indicated by asterisks (*). EFR observations of serine and threonine exhibit relatively low energy values. The side chains of both amino acids can form hydrogen bonds. The decrease in energy is insignificant for aspartic acid and isoleucine. No annotation of EFR is available for proline.

EFR exhibit significantly lower values in computed Energy Profiles as well as those predicted from sequence. This indicates that they occur in parts of proteins which are more stable and contain an increased number of hydrophobic amino acids in their spatial surroundings. Especially amino acids such as serine or threonine, which can form hydrogen bonds via their side chains, feature relatively low energy values even though they have an overall tendency to be exposed to the solvent due to their hydrophilic nature. The energy contribution of hydrogen bonds has been shown to be context specific [61], but also crucial for the correct formation of protein structure [53]. Especially amino acids capable of forming side chain hydrogen bonds contribute to the protein stability [1, 61]. Hydrophilic and aromatic amino acids like arginine, histidine, and methionine are considered strong hubs in protein structures, which is substantiated by a significant change in computed energy values for EFR. Hydrophobic amino acids occur in the core of a protein and are stabilized by an increased number of hydrophobic interactions. Thus, they have an intrinsic propensity to form stable, low energy conformations which is also reflected by the computed energy values. EFR might be the mediators between the formation of local structure elements and their assembly in the context of the three-dimensional structure. Secondary structure elements such as helices interact e.g. by hydrophobic interactions [62], however, it seems that single contacts are neither strong nor specific enough to guide their assembly [17, 63, 64]. A future, fine-grained distinction of contact types including π-stacking and hydrophobic interactions is needed to assess the role of EFR as potential driving force behind the correct of arrangement of secondary structure elements.

Network analysis shows a unique wiring of Early Folding Residues

The way residues interact with their spatial surrounding was assessed by network analysis based on protein graphs. Regarding the topological properties of residues derived from network analysis, EFR show a higher interconnectedness than LFR. They exhibit higher betweenness (Fig 4A) and closeness (Fig 4B) values. High betweenness values are observed for well-connected nodes which are passed by many of the shortest paths in a graph. High closeness values occur for nodes which can be reached by relatively short paths from arbitrary nodes. The distinct neighborhood count expresses to how many sequentially separated regions of a protein a residue is connected. Again a significant increase can be observed for EFR (Fig 4C). Residues are considered separated when they are more than five sequence positions apart. This threshold was also used to distinguish local contacts (i.e. less than six residues apart) and long-range contacts. Interestingly, the clustering coefficient features a significant decrease when EFR are considered. The clustering coefficient of a node is the number of edges between its adjacent nodes divided by the theoretical maximum of edges these nodes could form. However, EFR are biased to be in the core of the protein [39], thus, it was assessed if this change is also significant when only buried [65] residues are considered. The differences are insignificant in that case (see S2 Table).

Fig 4. Topological properties of EFR and LFR. Proteins were represented as graphs and a network analysis was performed. (**A**) EFR have higher betweenness values implying that shortest paths in the graph tend to pass through these nodes more often. (**B**) They also exhibit higher closeness values because their average path length to other nodes is lower on average. (**C**) The distinct neighborhood count of a residues describes to how many separated regions it is connected. Residues are considered separated when their separation on sequence level is greater than five. EFR connect significantly more regions of a protein than LFR.

By topological terms, EFR are more connected to the rest of the protein as expressed by betweenness, closeness, and the distinct neighborhood count. The betweenness property is closely related to the small-world characteristics of networks and can be observed in this case due to the ratio of protein surface and volume [46]. Residues relevant for the folding process have been shown to exhibit high betweenness values in the transition state and to be crucial for the formation of the folding nucleus [46]. Interestingly, the clustering coefficient shows no difference between EFR and LFR when only buried residues are considered. Also, the value is higher for LFR, which is probably an effect of EFR being hubs which connect several separated regions of a protein (as shown by the distinct neighborhood count). These regions themselves are not well-connected, which results in a lower clustering coefficient for EFR. The performed network analysis aids the understanding on the idiosyncratic properties of EFR in the context of the whole protein and is in agreement with previous studies [11, 46, 53]. EFR are hubs between sequentially distant protein regions which underlines their importance for the correct assembly of the tertiary structure of a protein. The distinction between local and long-range contacts provides new insights into the structural relation of residues with their respective neighborhood. Nevertheless, the increased number of local and long-range contacts of EFR point to their importance for the whole protein folding process as described by the defined pathways model [14, 17]. The existence of disordered proteins [7, 38], chaperons [7], cotranslational folding [25], and the peculiarities of membrane proteins [62] conceal important properties and EFR may be a welcome simplification to advance the understanding of the protein folding process.

Early Folding Residues are non-functional residues

Division of labor is one of the most successful strategies of evolution [40, 54, 55, 66–69]. The separation of residues crucial for folding and those furnishing function may allow reuse of established protein folds [32, 40, 54–56, 58]. The sequence and structure space ascertained over the course of evolutions seems small for a truly random exploration. Reusing established folds could also avoid slow-folding sequences or those prone to aggregation [31, 56, 70, 71]. There seem to be a delicate balance in proteins between robustness and evolvability [55, 58]. Thus, functional residues [72] can be mutated and new functions can be adopted without compromising the fold of the protein [32]. In consequence, a clear division should be observable between EFR – which initiate and guide the folding process – and the functional ones implementing protein function.

To address this question, residues in the dataset were labeled as either EFR or LFR as well as being either functional or non-functional. Active sites and ligand binding regions were considered to be the functional parts of proteins. The distribution of both binary variables (Table 1) shows that the majority of residues in the dataset are neither EFR (87.2%) nor functional (95.4%) residues. Only 0.5% share both classes, resulting in a Cramér's V of 0.01. The distribution of both variables separated by individual proteins is presented in S1 Table. For most proteins, no residues are both EFR and functional (Fig 5A). Furthermore, EFR tend to be located in the core of proteins, whereas functional residues are exposed towards the solvent in order to realize their respective function (Fig 5). Acyl-coenzyme A binding protein (STF0001) [33, 73, 74] features five residues which are both EFR and functional (Fig 5B).

For the majority of the dataset, a clear separation of EFR and functional residues can be observed. The acyl-coenzyme A binding protein may exhibit five residues which are both EFR and functional because its a rather small protein of 86 residues which binds ligands with large aliphatic regions. Intuitively, the residues furnishing the bowl-like shape of the protein are also those which participate in the function of ligand binding [33, 73, 74]. For acyl-coenzyme A binding protein, roughly half of its residues are marked as EFR which further accentuates why the division of labor is less strict in this

Table 1. Contingency table of folding characteristics and functional relevance.

	functional	non-functional
early	14	345
late	116	2332

Out of 2807 observations, 0.5% are EFR and functional at the same time. Cramér's V amounts to 0.01 – this minimal association between both categories implies that EFR are not functional and vice versa.

Fig 5. Rendered structures of 2 dataset entries. EFR are rendered in blue, functional residues are rendered in orange. (**A**) In the case of lysozyme (PDB:2eql_A) the intersection is empty. For most proteins in the dataset, there is a clear distinction between both classes and structurally relevant residues have a propensity to be located in the core, while functional residues are exposed on the protein's surface. (**B**) Five residues are both EFR and functional in the acyl-coenzyme A binding protein (PDB:2abd_A) which is one of the exceptions in the dataset where some residues are both EFR as well as functional.

case. Another case in which natures avoids limitations imposed by a defined structural fold can be found in aminoacyl tRNA synthetases [68, 69, 75]. Ancient enzymes may have existed as functional molten globules [76, 77] in their earliest implementations [68, 69] in order to not restrict evolution prematurely by ensuring integrity of the protein's fold [78]. Disordered proteins are another example of proteins without structural integrity which achieve a high robustness of function [49]. In structural biology, structure is commonly considered to be equal to function [49, 79]. However ultimately, it is most important that proteins are functional [79, 80]. This potential unimportance of a particular fold underlines that structurally and functionally relevant residues are detached entities in proteins and that their separation is advantageous for evolvability. Another interpretation with respect to the defined pathways model [14] is that EFR initiate and guide the folding process. By assigning this responsibility to a small number of residues, the remaining residues are available to carry other responsibility such as constituting active sites.

Early Folding and functional residues exhibit distinct features

The previously described features were employed to substantiate the identified separation of structure and function on residue level (S3 Table). EFR show significantly lower computed energy values when compared to LFR or functional residues (Fig 6A). Functional residues exhibit higher energy values than their non-functional counterparts. Most residues form only a small number of hydrophobic interactions, however, the number for EFR is significantly increased (Fig 6B). 97.7% of EFR form hydrogen bonds and 64.3% participate in hydrophobic interactions. Functional residues participate to 93.1% in hydrogen bonds and to 43.8% in hydrophobic interactions. On the contrary, the change between the hydrogen bond count of EFR and functional residues in a buried state is insignificant (S3 Table). The clustering coefficient of a node captures how many edges can be observed between the adjacent nodes and, thus, describes how well-connected the direct surroundings of a node are. Functional residues show an insignificant change regarding this property (S3 Table). In contrast, the clustering coefficient significantly decreases when EFR are compared to LFR or functional residues (Fig 6C). In summary, EFR exhibit distinct properties compared to functional residues. Their surrounding secondary structure elements, values in Energy Profiles, and the

number of hydrophobic interactions are especially discriminative.

Fig 6. Characteristics of EFR and functional residues. EFR and LFR are compared to functional and non-functional residues. (**A**) EFR show lower energy values as they are in contact with many residues and tend to be embedded in the hydrophobic core. In contrast, functional residues are exposed to the solvent in order to constitute e.g. binding sites. (**B**) Hydrophobic interactions occur especially in the core of a protein. Therefore, most residues do not form any. EFR however show an significant increase compared to LFR. (**C**) The clustering coefficient of a node describes how well-connected its adjacent nodes are. EFR connect regions of a protein which are separated on sequence and, thus, not well-connected on their own. Functional residues exhibit higher values.

Due to their purpose, EFR are located in the hydrophobic core and functional residues are primarily exposed to the solvent. This distinct requirements manifest in the computed energy values. Furthermore, protein function can commonly be broken down to amino acids which feature hydrophilic, chemically functional groups [72]. Hydroxyl groups are a prominent examples for functional groups contributing to catalysis [72]. Thus, functional residues are likely to exhibit above average energy values because of their higher propensity to contain hydrophilic side chains. Analogously, fewer hydrophobic amino acids constitute the functional residues of binding sites and they form fewer hydrophobic interactions. Most of the hydrophobic interactions are accumulated in the hydrophobic core of a protein [1, 28, 81]. EFR tend to be crucial connectors in proteins, however, their clustering coefficient is low. This can be attributed to the fact that EFR connect many distinct neighborhoods. Furthermore, functional residues feature above average closeness values: they are well-connected to other parts of the protein, even though they are unaffected during the early folding process. It was shown that functional residues have special requirements on how they are wired to the rest of a protein [45]: Surrounding residues ensure the correct placement of functional residues [45, 82, 83], modulate their chemical properties such as pK_a values [45, 72, 84], or propagate signals to other parts of a protein [45].

Modularity in proteins is also present in domains [54], secondary structure elements, and autonomous folding units of the defined pathways model [17, 27]. Particularized knowledge of EFR may improve synthetic biology and could allow the design of proteins combining existing functional domains without influencing one another negatively [2, 54, 55, 85]. Furthermore, understanding the differences of structurally relevant residues and those implementing function could help in predicting mutation effects and provide a new level of detail by allowing whether a mutation disrupts the protein's fold or its function [86, 87].

Conclusion

A dataset of EFR for the protein folding process was studied. They were found to be highly connected nodes in protein graphs and were observed to be located in energetically favorable conformations as pointed out by the approach of Energy Profiling [3, 28]. These structurally relevant residues have distinct properties e.g. regarding the number of hydrophobic interactions compared to functional residues.

Future HDX data can substantiate the presented trends regarding the nature of EFR. Potentially, the arsenal of experimental techniques to study the folding process of proteins will expand and become more refined and standardized, so that the underlying dataset of studies like this one will become more robust. EFR are an excellent tool to gain insights into the folding process with spatial and temporal resolution. Future

studies may link them to characteristics on sequence level to understand the sequence composition which causes particular regions of a protein to initiate the folding process. Features presented in this study were shown to be highly discriminative for EFR. Insights into topological properties of residues can also improve structure quality assessment programs [3]. Classifiers for EFR based on sequence [23] or structure may annotate residues crucial for protein folding. Trained classifiers can also report as well as visualize the most discriminative features [88, 89] which may further delineate EFR. This information is also invaluable for mutation studies, ϕ-value analysis, or protein design and can serve as basis for the prediction of mutation effects [86]. Understanding the protein folding problem may also give insights into the cause of diseases such as amyotrophic lateral sclerosis [4, 5], Alzheimer's [7], and Parkinson's disease [7]. The same is true for the observed division of structurally relevant and functional residues in proteins. Understanding these topological differences provides insights into the way they interact with the rest of the protein and to what degree they tolerate or compensate manipulation. For decades, scientists longed for a glimpse into the folding process [8–10] and the dataset of EFR [39] provides just that. It is stunning that not more studies are focused on this resource.

Methods

Dataset creation

Folding characteristics of residues were obtained from the Start2Fold database [39]. Therein, the authors adopted the definition of EFR from Li et al. [29] and presented a refined dataset which ignores possible back-unfolding and aggregation events [90].

This procedure resulted in a dataset for EFR characteristics encompassing 30 proteins and 3,377 residues – 482 of the EFR class and 2,895 of the LFR class. Due to the nature of the HDX experiments no data can be obtained for proline residues [37], rendering them LFR in any case. Annotation of functional residues was performed using the SIFTS [91] and UniProt [59] resources. For 23 proteins an annotation of binding sites or regions existed, totaling in 2,807 residues – 130 classified as functional and 2,677 as non-functional. A detailed summary of the dataset is provided in S1 Table. Information used from the Start2Fold database can be found in S1 File. Residues annotated as functional are summarized in S2 File.

Graph representation and analysis

Protein structures are commonly represented as graphs. This allows a scale-invariant characterization of the neighborhood relation of individual amino acids in the context of the whole protein [47].

In this study, amino acids constitute the nodes of a graph, whereas covalent bonds and residue contacts are represented as edges. Residues were considered in contact when their C_β atoms were less than 8 Å apart – if no C_β atom was present the C_α position was used as fallback. Furthermore, contacts were labeled as either local (i.e. the separation in sequence is less than six) or long-range (i.e. sequence separation greater than five) [92]. This distinguishes contacts stabilizing secondary structure elements and those which represent contacts between secondary structure elements. The set of distinct neighborhoods of a node is defined as all adjacent nodes which do not share any local edge to any element of the set. Betweenness is defined the number of shortest paths on the graph passing through a specific node, normalized by the number of node pairs [46, 93]. Closeness of a node is defined as the inverse of the average path length to any other node [45]. The clustering coefficient of node is the number of edges

between its n_k adjacent nodes divided by the maximal number of edges between n_k nodes: $0.5 \cdot n_k \cdot (n_k - 1)$ [46].

Feature computation

Energy Profiles were calculated from structure and predicted from sequence according to the methodology used in the eQuant web server [3, 28]. Energy Profiles represent a protein's complex three-dimensional structure as one-dimensional vector of energy values. Thereby, the surroundings of each residue are characterized by one energy value. RASA values were computed by the algorithm of Shrake and Rupley [94]. Buried residues are defined as those with RASA values less than 0.16 [65]. Non-covalent residue-residue contacts were detected by PLIP [95]. Secondary structure elements were annotated using DSSP [96]. The loop fraction is defined as fraction of unordered secondary structure in a window of nine residues around the evaluated amino acid [60]. This yields a fraction, where high values are tied to regions of high disorder, whereas amino acids embedded in α-helices or β-sheets result in scores close to zero. The centroid distance of a residue is the spatial distance of its centroid to that of all atoms. The terminus distance is lower of the sequence separation to either terminus divided by the number of residues.

Data integration was performed by a Java library publicly available at `https://github.com/JonStargaryen/jstructure`.

Statistical analysis

Association between distributions of nominal variables was quantified with Cramér's V. Dependence of distributions of real-valued variables was tested by the Mann-Whitney U test. Dependence of distributions of count variables was tested using the Dunn test with Bonferroni correction. * corresponds to significant p-values <0.05 for the Mann-Whitney U and p-values <0.025 for the Dunn test.

Acknowledgments

The authors thank Florian Kaiser, Christoph Leberecht, Sebastian Salentin, and Alexander Eisold for scientific discussions and/or proofreading the manuscript.

References

1. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Biophys. 2008;37:289–316.

2. Haglund E, Danielsson J, Kadhirvel S, Lindberg MO, Logan DT, Oliveberg M. Trimming down a protein structure to its bare foldons: spatial organization of the cooperative unit. J Biol Chem. 2012;287(4):2731–2738.

3. Bittrich S, Heinke F, Labudde D. eQuant-A Server for Fast Protein Model Quality Assessment by Integrating High-Dimensional Data and Machine Learning. In: Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery. Springer; 2015. p. 419–433.

4. Bystrom R, Andersen PM, Grobner G, Oliveberg M. SOD1 mutations targeting surface hydrogen bonds promote amyotrophic lateral sclerosis without reducing apo-state stability. J Biol Chem. 2010;285(25):19544–19552.

5. Shaw BF, Valentine JS. How do ALS-associated mutations in superoxide dismutase 1 promote aggregation of the protein? Trends Biochem Sci. 2007;32(2):78–85.

6. Jahn TR, Radford SE. Folding versus aggregation: polypeptide conformations on competing pathways. Arch Biochem Biophys. 2008;469(1):100–117.

7. Balchin D, Hayer-Hartl M, Hartl FU. In vivo aspects of protein folding and quality control. Science. 2016;353(6294):aac4354.

8. Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. Trends in biochemical sciences. 1999;24(1):26–33.

9. Baldwin RL, Rose GD. Is protein folding hierarchic? II. Folding intermediates and transition states. Trends in biochemical sciences. 1999;24(2):77–83.

10. Raimondi D, Orlando G, Pancsa R, Khan T, Vranken WF. Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins. Scientific reports. 2017;7(1):8826.

11. Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. Nature. 2001;409(6820):641–645.

12. Dokholyan NV, Li L, Ding F, Shakhnovich EI. Topological determinants of protein folding. Proceedings of the National Academy of Sciences. 2002;99(13):8637–8641.

13. Daggett V, Fersht AR. Is there a unifying mechanism for protein folding? Trends in biochemical sciences. 2003;28(1):18–25.

14. Englander SW, Mayne L. The nature of protein folding pathways. Proceedings of the National Academy of Sciences. 2014;111(45):15873–15880.

15. Nickson AA, Wensley BG, Clarke J. Take home lessons from studies of related proteins. Current opinion in structural biology. 2013;23(1):66–74.

16. Panchenko AR, Luthey-Schulten Z, Wolynes PG. Foldons, protein structural modules, and exons. Proc Natl Acad Sci USA. 1996;93(5):2008–2013.

17. Englander SW, Mayne L. The case for defined protein folding pathways. Proceedings of the National Academy of Sciences. 2017;114(31):8253–8258.

18. Lesk AM, Rose GD. Folding units in globular proteins. Proceedings of the National Academy of Sciences. 1981;78(7):4304–4308.

19. Rooman MJ, Kocher JP, Wodak SJ. Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. Biochemistry. 1992;31(42):10226–10238.

20. Rooman MJ, Wodak SJ. Extracting information on folding from the amino acid sequence: consensus regions with preferred conformation in homologous proteins. Biochemistry. 1992;31(42):10239–10249.

21. Myers JK, Oas TG. Preorganized secondary structure as an important determinant of fast protein folding. Nat Struct Biol. 2001;8(6):552–558.

22. Krishnan A, Giuliani A, Zbilut JP, Tomita M. Network scaling invariants help to elucidate basic topological principles of proteins. J Proteome Res. 2007;6(10):3924–3934.

23. Pancsa R, Raimondi D, Cilia E, Vranken WF. Early folding events, local interactions, and conservation of protein backbone rigidity. Biophysical journal. 2016;110(3):572–583.

24. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nature Structural & Molecular Biology. 1997;4(1):10–19.

25. Deane CM, Dong M, Huard FP, Lance BK, Wood GR. Cotranslational protein folding–fact or fiction? Bioinformatics. 2007;23(13):i142–148.

26. Karplus M, Weaver DL. Protein folding dynamics: The diffusion-collision model and experimental data. Protein Science. 1994;3(4):650–668.

27. Maity H, Maity M, Krishna MM, Mayne L, Englander SW. Protein folding: the stepwise assembly of foldon units. Proc Natl Acad Sci USA. 2005;102(13):4741–4746.

28. Heinke F, Schildbach S, Stockmann D, Labudde D. eProS—a database and toolbox for investigating protein sequence–structure–function relationships through energy profiles. Nucleic acids research. 2012;41(D1):D320–D326.

29. Li R, Woodward C. The hydrogen exchange core and protein folding. Protein Science. 1999;8(8):1571–1590. doi:10.1110/ps.8.8.1571.

30. Fersht AR, Sato S. Phi-value analysis and the nature of protein-folding transition states. Proc Natl Acad Sci USA. 2004;101(21):7976–7981.

31. Oliveberg M, Wolynes PG. The experimental survey of protein-folding energy landscapes. Q Rev Biophys. 2005;38(3):245–288.

32. Nishimura C, Prytulla S, Dyson HJ, Wright PE. Conservation of folding pathways in evolutionarily distant globin sequences. Nature Structural & Molecular Biology. 2000;7(8):679–686.

33. Teilum K, Kragelund BB, Knudsen J, Poulsen FM. Formation of hydrogen bonds precedes the rate-limiting formation of persistent structure in the folding of ACBP. Journal of molecular biology. 2000;301(5):1307–1314.

34. Roder H, Elove GA, Englander SW. Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. Nature. 1988;335(6192):700–704.

35. Bai Y, Sosnick TR, Mayne L, Englander SW. Protein folding intermediates: native-state hydrogen exchange. Science. 1995;269(5221):192–197.

36. Chu R, Pei W, Takei J, Bai Y. Relationship between the native-state hydrogen exchange and folding pathways of a four-helix bundle protein. Biochemistry. 2002;41(25):7998–8003.

37. Englander SW, Mayne L, Krishna MM. Protein folding and misfolding: mechanism and principles. Q Rev Biophys. 2007;40(4):287–326.

38. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. From protein sequence to dynamics and disorder with DynaMine. Nat Commun. 2013;4:2741.

39. Pancsa R, Varadi M, Tompa P, Vranken WF. Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability. Nucleic acids research. 2016;44(D1):D429–D434.

40. Ptitsyn OB, Ting KLH. Non-functional conserved residues in globins and their possible role as a folding nucleus. Journal of molecular biology. 1999;291(3):671–682.

41. Bedard S, Mayne LC, Peterson RW, Wand AJ, Englander SW. The foldon substructure of staphylococcal nuclease. J Mol Biol. 2008;376(4):1142–1154.

42. Warshel A, Levitt M. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol. 1976;103(2):227–249.

43. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. Biophysical journal. 2004;86(1):85–91.

44. Bagler G, Sinha S. Network properties of protein structures. Physica A: Statistical Mechanics and its Applications. 2005;346(1):27–33.

45. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, et al. Network analysis of protein structures identifies functional residues. Journal of molecular biology. 2004;344(4):1135–1146.

46. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Physical Review E. 2002;65(6):061910.

47. Royer L, Reimann M, Andreopoulos B, Schroeder M. Unraveling protein networks with power graph analysis. PLoS computational biology. 2008;4(7):e1000108.

48. Kayikci M, Venkatakrishnan A, Scott-Brown J, Ravarani CN, Flock T, Babu MM. Visualization and analysis of non-covalent contacts using the Protein Contacts Atlas. Nature Publishing Group; 2018.

49. Rorick MM, Wagner GP. Protein Structural Modularity and Robustness Are Associated with Evolvability. Genome Biology and Evolution. 2011;3:456–475. doi:10.1093/gbe/evr046.

50. Duarte JM, Sathyapriya R, Stehr H, Filippis I, Lappe M. Optimal contact definition for reconstruction of contact maps. BMC bioinformatics. 2010;11(1):283.

51. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins: Structure, Function, and Bioinformatics. 2001;44(2):150–165.

52. Wangikar PP, Tendulkar AV, Ramya S, Mali DN, Sarawagi S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. Journal of molecular biology. 2003;326(3):955–978.

53. Brinda K, Vishveshwara S. A network representation of protein structures: implications for protein stability. Biophysical journal. 2005;89(6):4159–4170.

54. Bhattacharyya RP, Remenyi A, Yeh BJ, Lim WA. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev Biochem. 2006;75:655–680.

55. Rorick M. Quantifying protein modularity and evolvability: a comparison of different techniques. BioSystems. 2012;110(1):22–33.

56. Levy Y. Protein Assembly and Building Blocks: Beyond the Limits of the LEGO Brick Metaphor. Biochemistry. 2017;.

57. Khan T, Ghosh I. Modularity in protein structures: study on all-alpha proteins. J Biomol Struct Dyn. 2015;33(12):2667–2681.

58. Hleap JS, Susko E, Blouin C. Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture. BMC Struct Biol. 2013;13:20.

59. Consortium U, et al. UniProt: a hub for protein information. Nucleic acids research. 2014; p. gku989.

60. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic acids research. 2009;37(suppl_2):W510–W514.

61. Pace CN, Fu H, Fryar K, Landua J, Trevino SR, Schell D, et al. Contribution of hydrogen bonds to protein stability. Protein Science. 2014;23(5):652–661.

62. DeGrado WF, Gratkowski H, Lear JD. How do helix–helix interactions help determine the folds of membrane proteins? Perspectives from the study of homo-oligomeric helical bundles. Protein Science. 2003;12(4):647–665.

63. Dyrka W, Nebel JC, Kotulska M. Probabilistic grammatical model for helix-helix contact site classification. Algorithms for molecular biology: AMB. 2013;8(1):31–31.

64. Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proceedings of the National Academy of Sciences. 1992;89(1):20–22.

65. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins: Structure, Function, and Bioinformatics. 1994;20(3):216–226.

66. Nicolini C, Bifone A. Modular structure of brain functional networks: breaking the resolution limit by Surprise. Sci Rep. 2016;6:19250.

67. Guimera R, Nunes Amaral LA. Functional cartography of complex metabolic networks. Nature. 2005;433(7028):895–900.

68. Carter CW. Coding of Class I and II aminoacyl-tRNA synthetases. In: Protein Reviews. Springer; 2017. p. 103–148.

69. Martinez-Rodriguez L, Erdogan O, Jimenez-Rodriguez M, Gonzalez-Rivera K, Williams T, Li L, et al. Functional class I and II amino acid-activating enzymes can be coded by opposite strands of the same gene. Journal of Biological Chemistry. 2015;290(32):19710–19725.

70. Baldwin RL. The nature of protein folding pathways: the classical versus the new view. Journal of biomolecular NMR. 1995;5(2):103–109.

71. Wolynes PG. Three paradoxes of protein folding. Protein folds: A Distances Based Approach. 1996; p. 3–17.

72. Gutteridge A, Thornton JM. Understanding nature's catalytic toolkit. Trends Biochem Sci. 2005;30(11):622–629.

73. Kragelund BB, Andersen KV, Madsen JC, Knudsen J, Poulsen FM. Three-dimensional structure of the complex between acyl-coenzyme A binding protein and palmitoyl-coenzyme A. Journal of molecular biology. 1993;230(4):1260–1277.

74. Burton M, Rose TM, Færgeman NJ, Knudsen J. Evolution of the acyl-CoA binding protein (ACBP). Biochemical Journal. 2005;392(2):299–307.

75. Kaiser F, Bittrich S, Salentin S, Leberecht C, Haupt VJ, Krautwurst S, et al. Backbone brackets and arginine tweezers delineate class I and class II aminoacyl tRNA synthetases. bioRxiv. 2017;doi:10.1101/198846.

76. Pervushin K, Vamvaca K, Vögeli B, Hilvert D. Structure and dynamics of a molten globular enzyme. Nature structural & molecular biology. 2007;14(12):1202–1206.

77. Hu H. Wild-type and molten globular chorismate mutase achieve comparable catalytic rates using very different enthalpy/entropy compensations. Science China Chemistry. 2014;57(1):156–164.

78. Mirny LA, Abkevich VI, Shakhnovich EI. How evolution makes proteins fold quickly. Proc Natl Acad Sci USA. 1998;95(9):4976–4981.

79. Najmanovich RJ. Evolutionary studies of ligand binding sites in proteins. Current opinion in structural biology. 2017;45:85–90.

80. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proceedings of the National Academy of Sciences. 2006;103(15):5869–5874.

81. Dobson CM, Karplus M. The fundamentals of protein folding: bringing together theory and experiment. Curr Opin Struct Biol. 1999;9(1):92–101.

82. Kaiser F, Eisold A, Bittrich S, Labudde D. Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data. Bioinformatics. 2015;32(5):792–794.

83. Kaiser F, Labudde D. Unsupervised Discovery of Geometrically Common Structural Motifs and Long-Range Contacts in Protein 3D Structures. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2017;.

84. Brodkin HR, DeLateur NA, Somarowthu S, Mills CL, Novak WR, Beuning PJ, et al. Prediction of distal residue participation in enzyme catalysis. Protein Science. 2015;24(5):762–778.

85. Jacobs TM, Kuhlman B. Using anchoring motifs for the computational design of protein–protein interactions; 2013.

86. Hecht M, Bromberg Y, Rost B. News from the protein mutability landscape. Journal of molecular biology. 2013;425(21):3937–3948.

87. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nature biotechnology. 2017;35(2):128–135.

88. Hammer B, Villmann T. Generalized relevance learning vector quantization. Neural Networks. 2002;15(8):1059–1068.

89. Kästner M, Hammer B, Biehl M, Villmann T. Functional relevance learning in generalized learning vector quantization. Neurocomputing. 2012;90:85–95. doi:10.1016/j.neucom.2011.11.029.

90. Silow M, Oliveberg M. Transient aggregates in protein folding are easily mistaken for folding intermediates. Proceedings of the National Academy of Sciences. 1997;94(12):6084–6086.

91. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, et al. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic acids research. 2012;41(D1):D483–D489.

92. Adhikari B, Cheng J. Improved protein structure reconstruction using secondary structures, contacts at higher distance thresholds, and non-contacts. BMC bioinformatics. 2017;18(1):380.

93. Freeman LC. A set of measures of centrality based on betweenness. Sociometry. 1977; p. 35–41.

94. Shrake A, Rupley J. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. Journal of molecular biology. 1973;79(2):351IN15365–364371.

95. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein–ligand interaction profiler. Nucleic acids research. 2015;43(W1):W443–W447.

96. Kabsch W, Sander C. DSSP: definition of secondary structure of proteins given a set of 3D coordinates. Biopolymers. 1983;22:2577–2637.

Supporting information

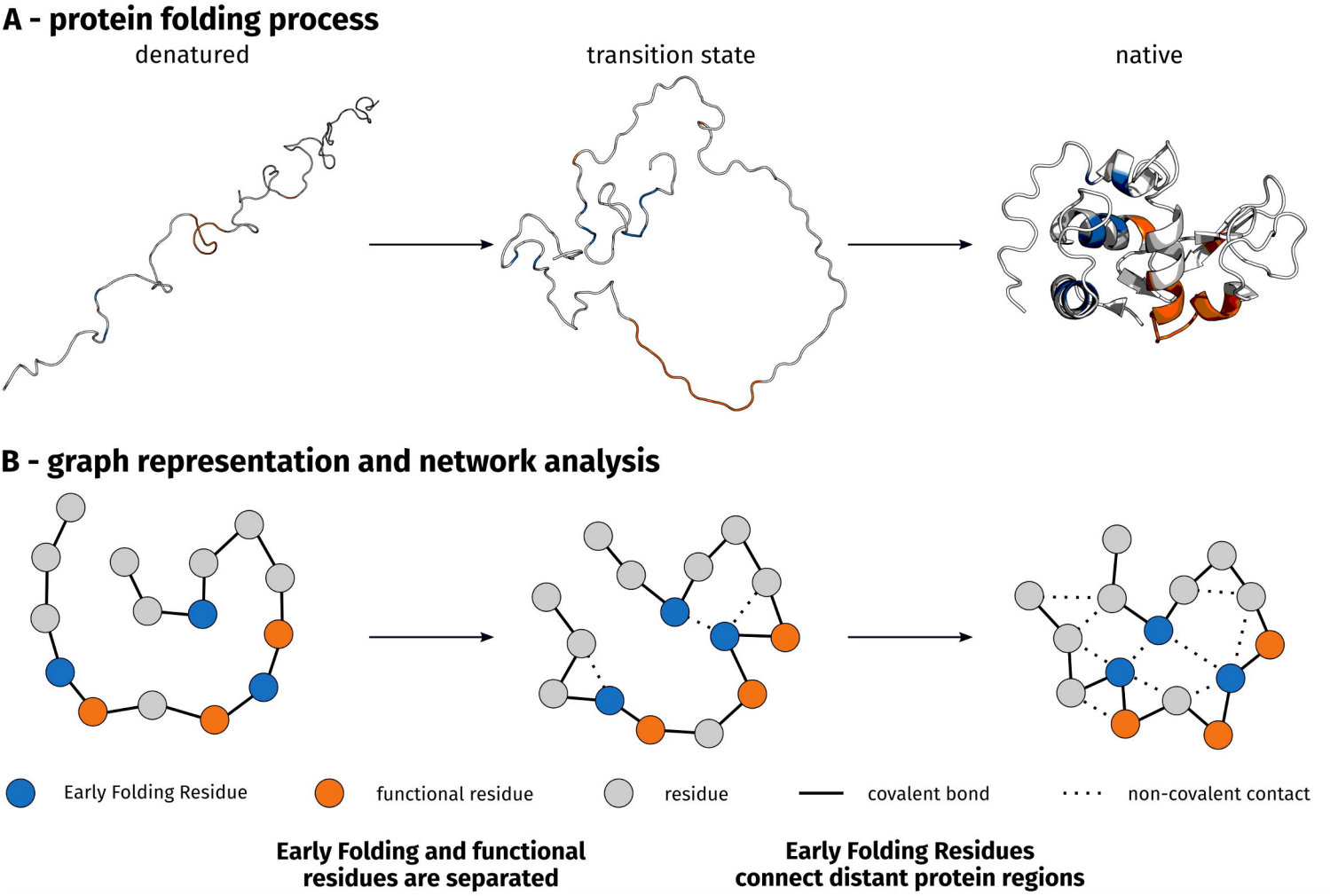
S1 Table. EFR dataset summary. Summarizes identifiers [23] of each entry as well as the number of residues in the corresponding protein chain, the number of EFR and functional residues as well as the cardinality of the intersection of both sets. Proteins not containing any functional residues according to UniProt [59] are marked with dashes.

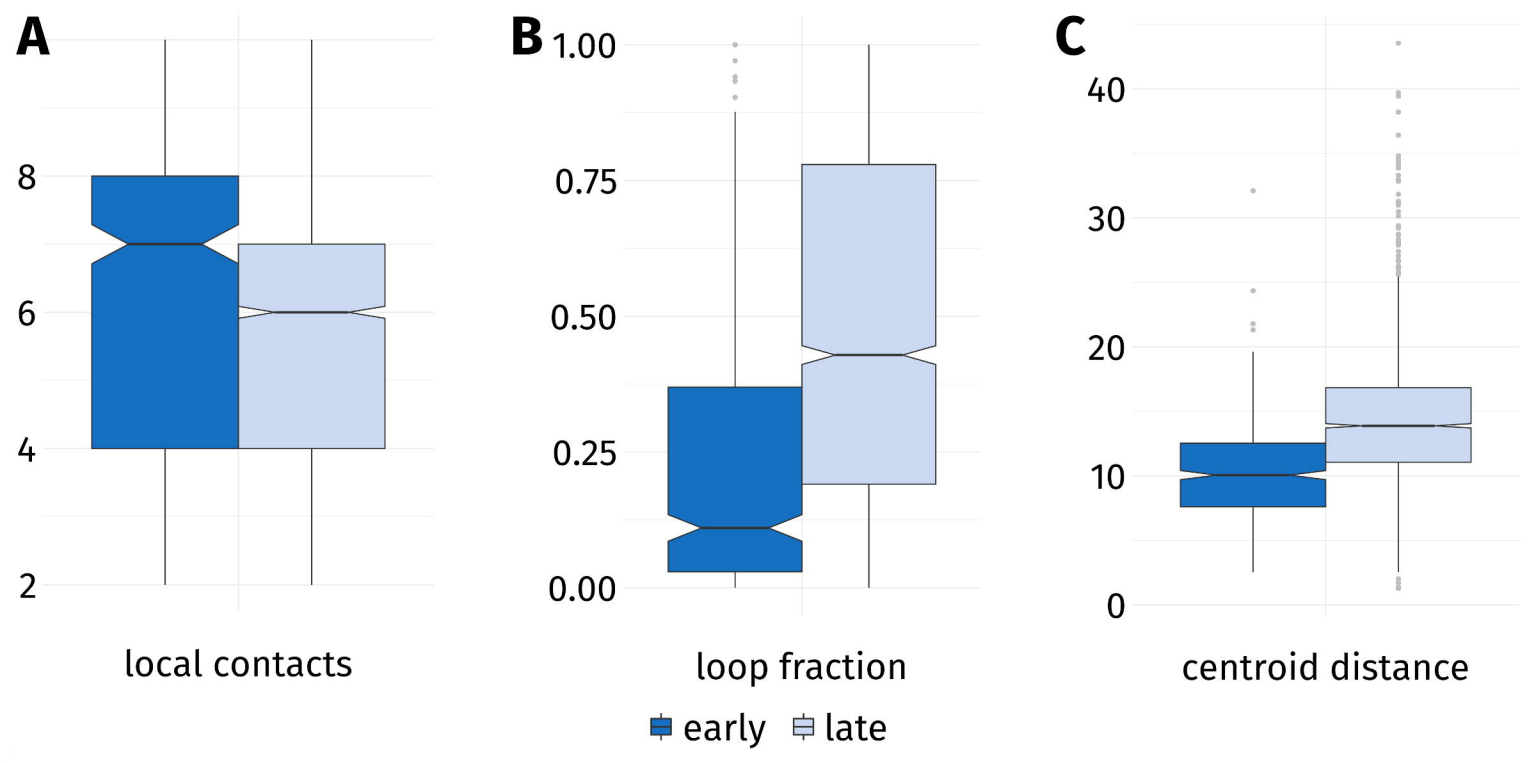
S2 Table. Statistical characterization of EFR. For each presented feature the mean (μ) and standard deviation (σ) of both the EFR and LFR category is reported. p_{buried} refers to the p-value of the test on residues buried according their RASA value, this was done because EFR have a tendency to be located in the core of a protein and without filtering all differences are significant. Features and employed tests are described in the Methods section.

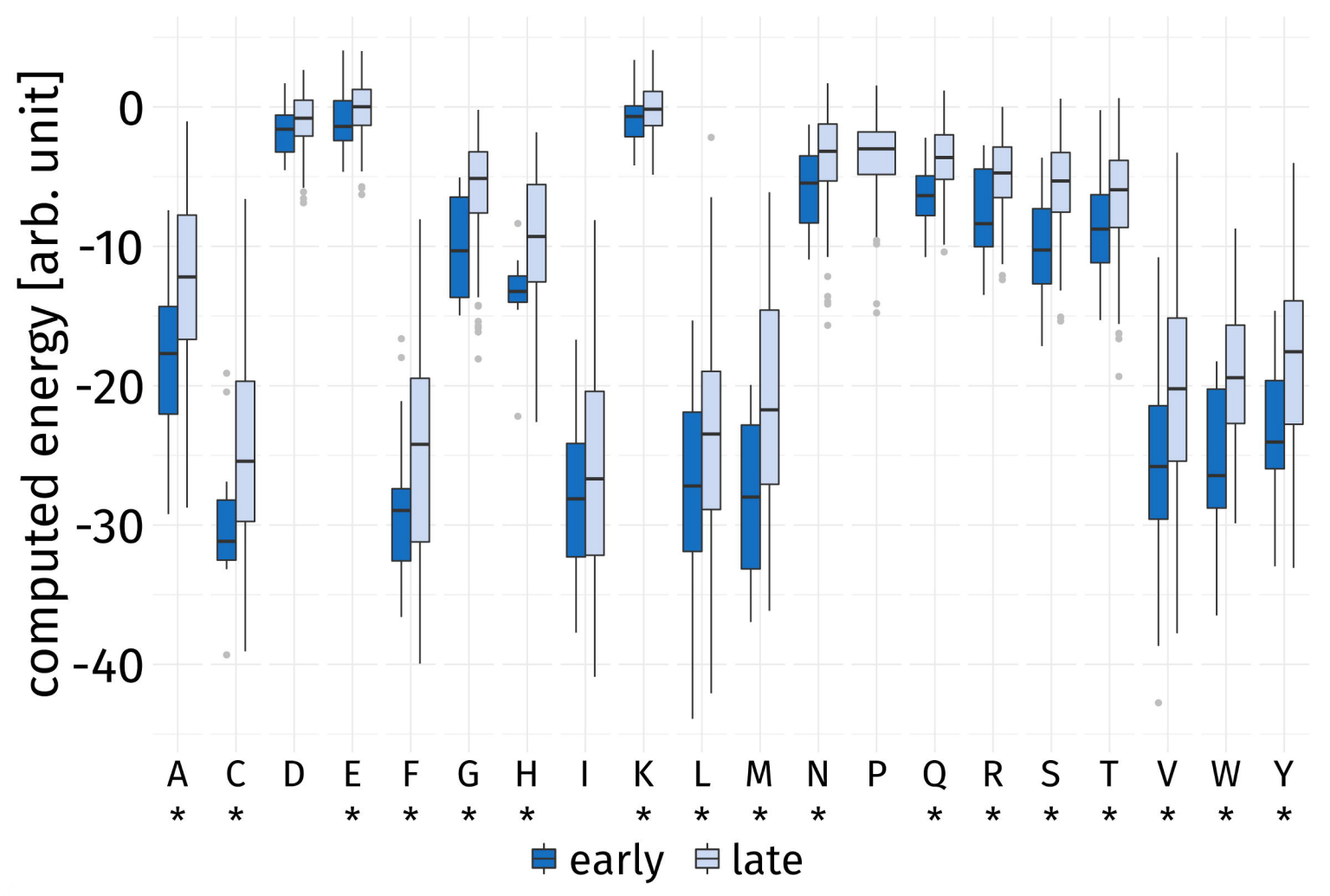
S3 Table. Comparison of EFR and functional residues. For each presented feature the distribution of values is compared between functional and non-functional residues as well as EFR and functional residues. The corresponding p-values and significance level are stated for buried residues. Mean values are shown for EFR (μ_{early}) and functional residues (μ_{func}). Features and employed tests are described in the Methods section.

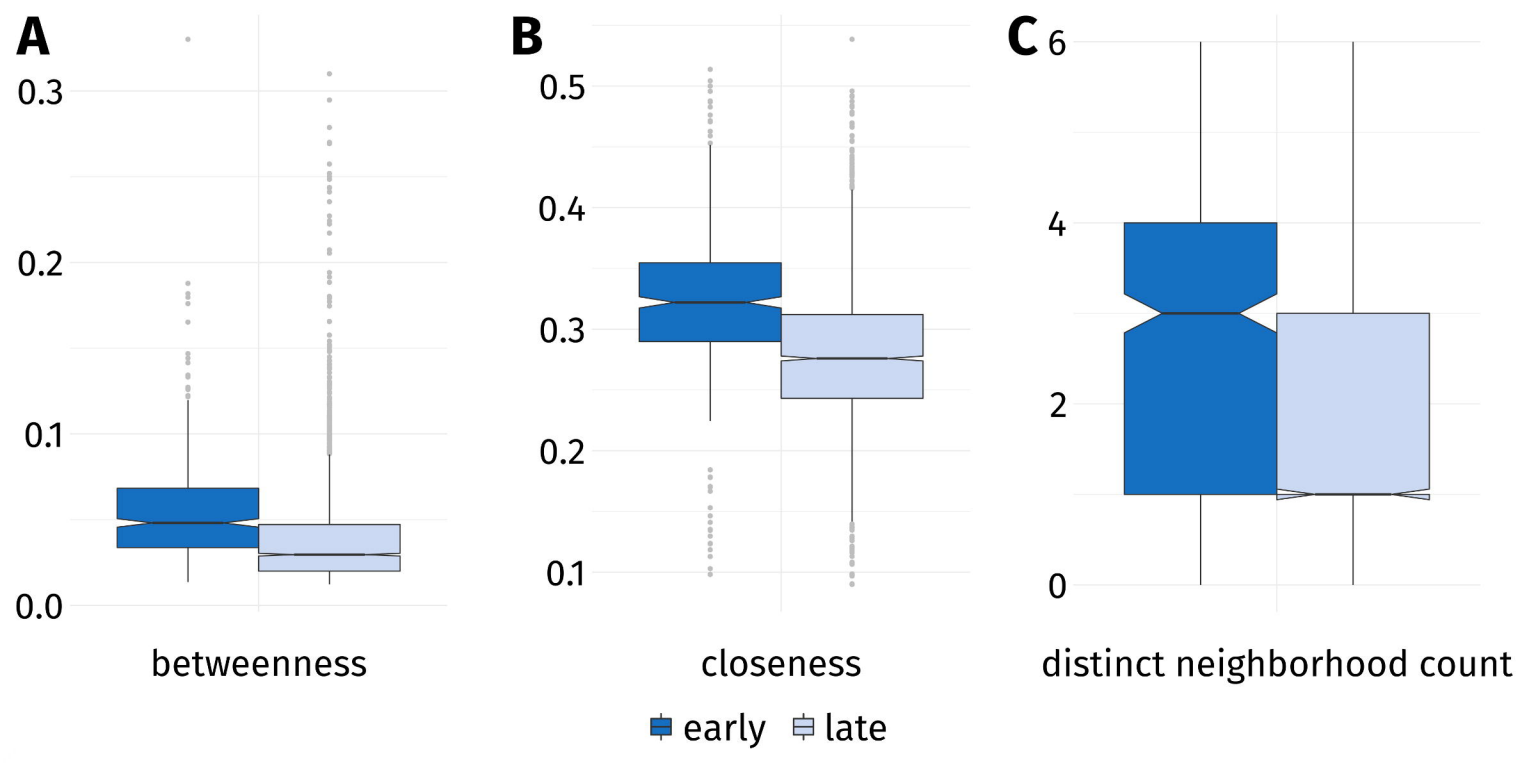
S1 File. Dataset as JSON file. Machine-readable JSON version of the dataset. Provides protein name, Start2Fold identifier, PDB identifier, UniProt identifier, number of EFR, range of residues numbers, and the secondary structure element composition for each dataset entry.

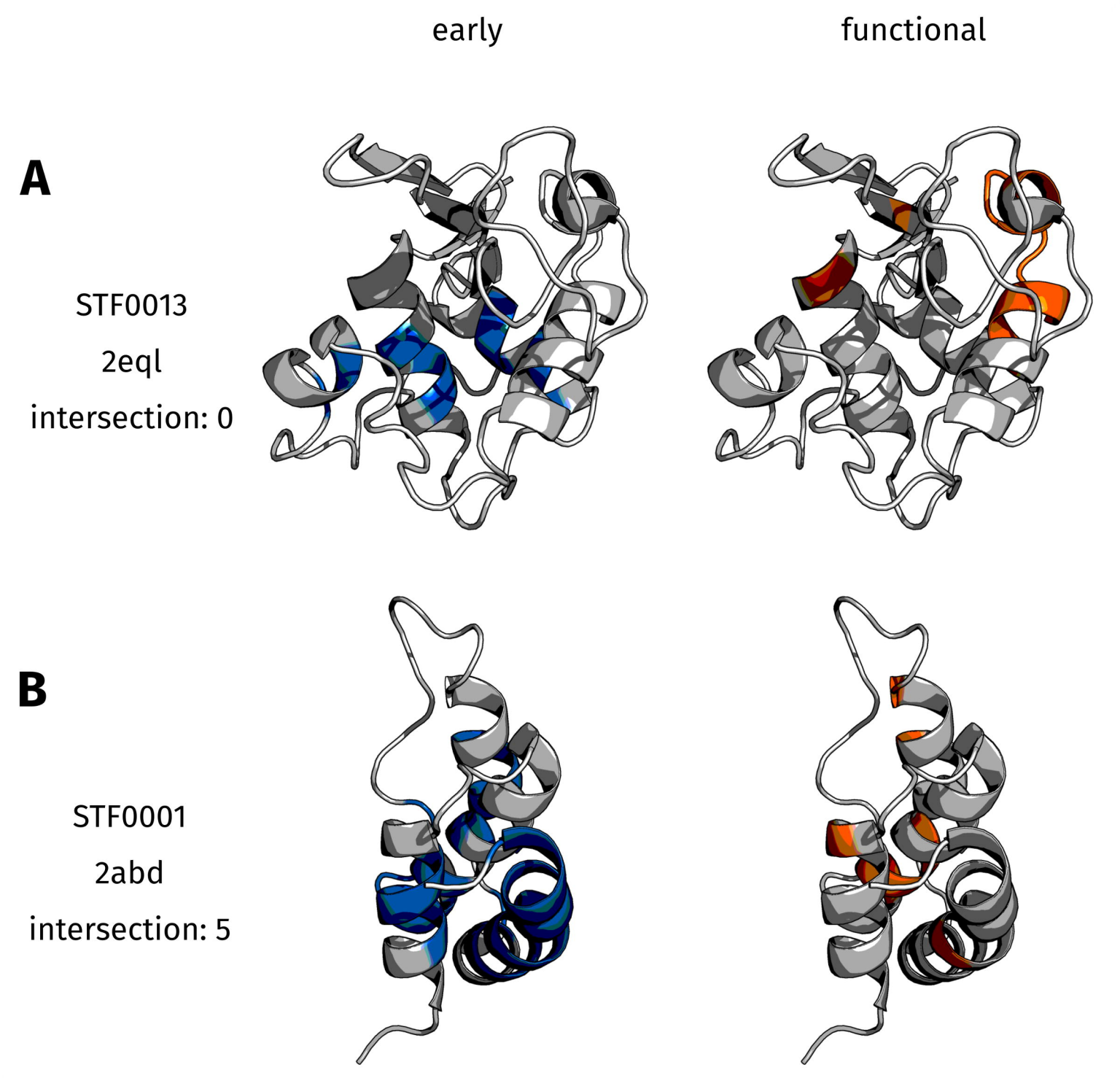
S2 File. Dataset as table. Summary table of all protein chains used for the analysis. Provides Start2Fold identifier, PDB identifier, evaluated experiments, number of EFR, UniProt identifier, and identifiers of functional residues derived from UniProt.

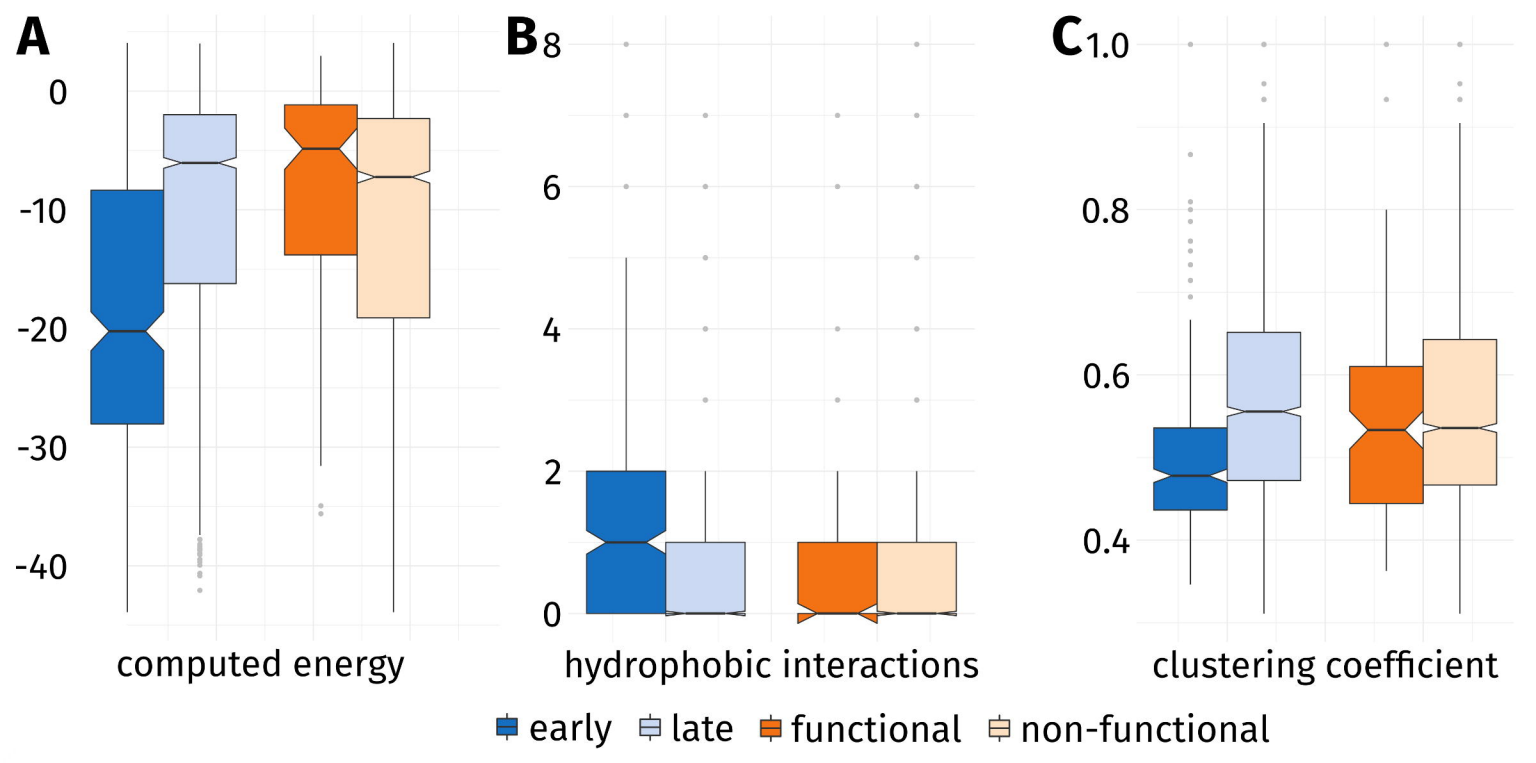

A - protein folding process
denatured
transition state
native
B - graph representation and network analysis
Early Folding Residue
functional residue
residue
covalent bond
non-covalent contact
Early Folding and functional residues are separated
Early Folding Residues connect distant protein regions


A
8
6
4
2
local contacts
B
1.00
0.75
0.50
0.25
0.00
loop fraction
C
40
30
20
10
0
centroid distance
early
late


computed energy [arb. unit]
0
-10
-20
-30
-40
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
*
*
*
*
*
*
*
*
*
*
*
*
*
*
*
*
*
early
late


A
0.3
0.2
0.1
0.0
betweenness
B
0.5
0.4
0.3
0.2
0.1
closeness
C
6
4
2
0
distinct neighborhood count
early
late


early
functional
A
STF0013
2eql
intersection: 0
B
STF0001
2abd
intersection: 5


A
0
-10
-20
-30
-40
computed energy
B
8
6
4
2
0
hydrophobic interactions
C
1.0
0.8
0.6
0.4
clustering coefficient
early
late
functional
non-functional