PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Reassessing the exon–foldon correspondence using frustration analysis

Ezequiel A. Galpern[a,b], Hana Jaafari[c,d], Carlos Bueno[c], Peter G. Wolynes[c,e,f,1], and Diego U. Ferreiro[a,b,1]

Affiliations are included on p. 7.

Edited by Victor Muñoz, University of California Merced School of Engineering, Merced, CA; received January 3, 2024; accepted May 31, 2024 by Editorial Board Member J. A. McCammon

Protein folding and evolution are intimately linked phenomena. Here, we revisit the concept of exons as potential protein folding modules across a set of 38 abundant and conserved protein families. Taking advantage of genomic exon–intron organization and extensive protein sequence data, we explore exon boundary conservation and assess the foldon-like behavior of exons using energy landscape theoretic measurements. We found deviations in the exon size distribution from exponential decay indicating selection in evolution. We show that when taken together there is a pronounced tendency to independent foldability for segments corresponding to the more conserved exons, supporting the idea of exon–foldon correspondence. While 45% of the families follow this general trend when analyzed individually, there are some families for which other stronger functional determinants, such as preserving frustrated active sites, may be acting. We further develop a systematic partitioning of protein domains using exon boundary hotspots, showing that minimal common exons correspond with uninterrupted alpha and/or beta elements for the majority of the families but not for all of them.

exon | protein folding | energy landscape | foldon

Significance

If globular protein domains consist of smaller units, folding and evolution would be facilitated. The fact that natural eukaryotic proteins are genetically partitioned in exons suggests that these may correspond with foldable regions. Here, we revisit the correspondence between exons and foldons, quasi-independent folding units, using concepts derived from energy landscape theory. We find that conserved exons are more foldable than other partitions of the primary structure. We describe that exon boundaries rarely interrupt the continuous secondary structures in the folded domain in most but not all of the protein families analyzed.

Protein evolution and folding are two intertwined aspects of a complex problem. Over the past decades, a reasonable shortcut to simplify this problem has been to try to break down protein structures into distinct modules. In 1973, Wetlaufer proposed that the initial stages of folding nucleation may occur independently in separate regions (1). Addressing Levinthal's paradox, he claimed that if there were individual modules that fold in parallel, the searching time for folding the entire molecule can be exponentially reduced and would be comparable to the isolated segments' folding time. With the discovery of silent DNA interrupting coding regions in Eukarya, Gilbert (2) and Blake (3) posited that if genes resemble a mosaic divided into pieces, then the coding pieces—christened by Gilbert "exons"—can reasonably be expected to translate into integrally folded protein pieces, such as domains or supersecondary structures. These fragments could then shuffle and combine over evolutionary timescales, giving rise to novel functional proteins. Indeed, exons of several proteins were early characterized as structural units, including hemoglobin (4). Among various theories, it has been argued that exon-shuffling may have played a significant role in metazoan evolution, coinciding with a burst of evolutionary creativity during the emergence of multicellularity (5).

Energy landscape theory explains how proteins fold within relevant timescales using parallel paths without explicitly dividing the molecule into parts. When a polymer is minimally frustrated, parallel search can be done in a delocalized manner as native contacts can guide the polymer folding (6). Of course, some paths may be modestly favored over others, and these variations have been successfully predicted by perfectly funneled models (7). Different protein regions may fold at different times quasi-independently if the sufficiently strong native interactions largely contained within them can overcome their entropy loss. These units then may fold in a single cooperative step which have been called foldons by Panchenko et al (8). Using a simple energy field model and a searching algorithm, they assigned foldons to many proteins and they compared them with exons. They found only a weak correlation between the evolutionary units and the folding regions (8, 9).

Exons have also been compared with secondary structure elements, with negative results (10, 11). Evidence of a co-occurrence between exon boundaries and protein domain border positions has been found by others and used to support the exon-shuffling

Author contributions: E.A.G., P.G.W., and D.U.F. designed research; E.A.G., H.J., C.B., P.G.W., and D.U.F. performed research; E.A.G. and H.J. contributed new reagents/analytic tools; E.A.G., H.J., C.B., P.G.W., and D.U.F. analyzed data; and E.A.G., P.G.W., and D.U.F. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. V.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: wolynes@rice.edu or ferreiro@qb.fcen.uba.ar.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2400151121/-/DCSupplemental.

Published July 2, 2024.

theory (12, 13). At least for some genomes, it has been shown that this correlation of domains and exons can not be explained with a neutral null model (14).

The search for folding elemental units on some particular proteins has been pursued directly with various experimental methods and models. Foldons have been identified for Cytochrome C through Hydrogen exchange experiments (15, 16). These agreed with those found computationally with a perfectly funneled energy model (17). Dihydrofolate reductase (DHFR) has been analyzed by molecular dissection (18), circular permutation (19), systematic Alanine insertion (20) and overlapped contact volume (21), leading to potential modular decompositions.

Folding units are not necessarily continuous in sequence. Secondary structure motifs have been grouped into overlapping foldons (22) and physically connected amino acids in the tertiary structure have been correlated into "protein sectors" (23).

In the case of repeat-proteins, their structural symmetry allows a way to naturally define folding units for an entire protein family (24, 25). Remarkably, by modeling the interactions between these minimal common foldons, different groups of elements that fold at the same time emerge naturally for each protein, defining domains that coincide with those described experimentally (26). Interestingly, it has been seen that repeat-proteins are made of exons that encode one or two complete repeats, exhibiting a striking conservation of intron position and phase (27).

In this work, we revisit the concept of exon regions as potential protein folding modules. By leveraging gene annotation and protein sequence databases, we explore exon conservation across 38 protein families to assess whether exons exhibit foldon-like behavior through energy landscape measures. To accomplish this, we use the coarse-grained forcefield Associative memory, Water Mediated, Structure and Energy Model (AWSEM) (28) to establish a quantitative score that assesses the independence of foldability for sequence fragments. Furthermore, we investigate a systematic partitioning of proteins into nonoverlapping units using exon boundary hotspots.

Results

Exons As Protein Segments. We mapped exon positions to the amino acid sequence in the multiple sequence alignments (MSA) for 38 protein domain families. Details about the data for each family are summarized in *SI Appendix*, Table S1, with curation specifics available in *Materials and Methods*. We divided the protein sequences into the segments that are encoded by each exon. It is noteworthy that the distribution of exons per protein in this set follows an exponential pattern (*SI Appendix*, Fig. S1). The distribution of exon sizes for the entire set also exhibits an exponential decay, a result expected under the assumption that intron positions are the result of independent trials of a neutral stochastic process (*SI Appendix*, Fig. S1).

However, when we focus our analysis on individual protein families, we observe deviations from the general trends. We present specific results for the DHFR family in Fig. 1 for illustrating this phenomenon. In the DHFR family, certain exceptionally large exon sizes are overrepresented (*B*), suggesting that natural selection may influence exon lengths. Results for other families are presented in *SI Appendix*, Fig. S3. Interestingly, none of the families individually shows a clear exponential decay trend. Instead, some preferred exon sizes stand out. For structurally symmetric domains like the Cristall or MHC-I family, a characteristic exon length emerges and the exponential decay is not present at all.

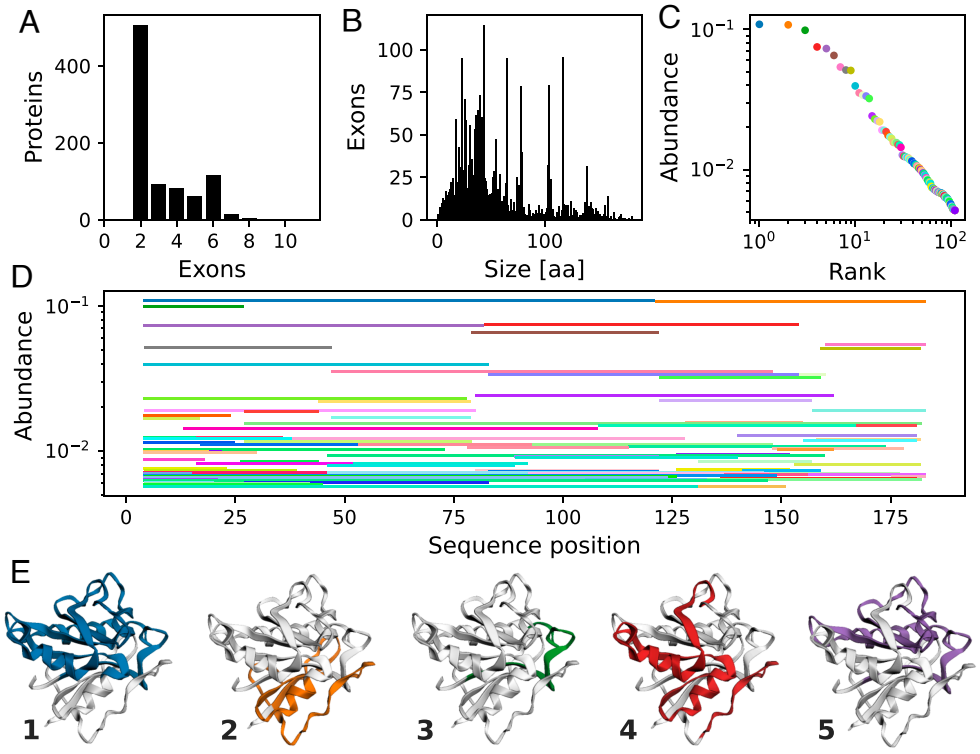

Fig. 1. Exon characterization for DHFR family. (*A*) Number of exons per protein. (*B*) Exon size distribution, measured in the amino acids of the corresponding protein segment. (*C*) Abundance-rank plot, including exons present in at least 0.5% of the effective sequences. (*D*) Abundance as a function of the aligned sequence position for the exons in panel (*C*). (*E*) Projection on the 3D family reference structure (PDB: 8dfr) for the most abundant exon (blue), the second one (orange), the third one (green), the fourth one (red), and fifth one (purple). Color assignment to exon is shared between panels (*C*–*E*).

Along the MSA, exon positions are sometimes exactly conserved allowing us to measure exon relative abundance. Abundance-rank plots present power-law trends, which can be a consequence of spreading phylogenetic diversity represented by exons. The DHFR case is shown in Fig. 1 *C*–*E*. We see that the two most abundant exons (blue and orange) are present in 10% of the sequences, causing a division of the domain at residue 119 into two consecutive fragments. The fourth (red) and the fifth (purple) most frequent exons define an almost completely alternative partition of the structure. In contrast, some other exons while abundant do not always come along with a specific exon in the complementary part of the chain. An example of this is the third-most abundant exon (green). This pattern suggests the existence of multiple alternative options that can complete the open reading frame.

Exon Foldability. Do natural exons behave as foldons? Foldons have been defined as quasi-independent foldable protein segments (8). A foldon then should be at least as minimally frustrated by itself as in the context of the whole protein that contains it. We therefore examine exons comparing their frustration using two schemes (Fig. 2*A*). In one scheme, the protein segment encoded in an exon is treated as a totally independent polymer folding to its final three-dimensional structure. In the other scenario, the folding of the same segment is treated in the context, still interacting with the rest of the protein that contains it. Using both the independent scheme (I) and the context scheme (C) we compute the correspondent total frustration index, a Z score defined as $f = \Delta E/\delta E$, where ΔE is the energy gap between the native configuration and the molten globule state, represented by a set of decoys, and δE^2 is the energy variance of those decoys (29). The quantities are related to the characteristic transition temperatures of the chain segments through the configurational entropy loss upon folding from a compact molten globule S. For the protein to be foldable on a relevant timescale, the folding temperature ($T_f \propto \Delta E/S$) should exceed its glass transition temperature ($T_g \propto \delta E/S^{1/2}$). Protein foldability, which has been used to search foldons (8, 9), can be written as $\Theta = f\,S^{1/2} \propto T_f/T_g$.

Here, we have employed the mutational frustration index, where decoys are scrambled versions of the original sequence. The energies are computed using the coarse-grained AWSEM potential (28). The exon energy is averaged over all the sequences in the alignment that share that same exon. A single Protein Data Bank (PDB) structure is used as reference for each family, threading the corresponding sequence each time. We take the independent segment to retain the structure that it has in context. Details of the implementation are provided in *Materials and Methods*.

We introduce δf, the relative change in total frustration of a protein segment in the transition from the independent (I) to the in context (C) scheme

$$\delta f = \frac{f_C - f_I}{-|f_I|}. \qquad [1]$$

If the configurational entropy loss S is the same in the two scenarios, the relative change in the total frustration can be seen also as the relative change in the foldability Θ and in the ratio T_f/T_g.

In Fig. 2*A* we present two examples. On the one hand, DHFR exon 1—the most conserved exon in Fig. 1 *E*1—shows a small change in total frustration $\delta f_1 = 8\%$. It's a segment that in isolation is almost as minimally frustrated in the context of the

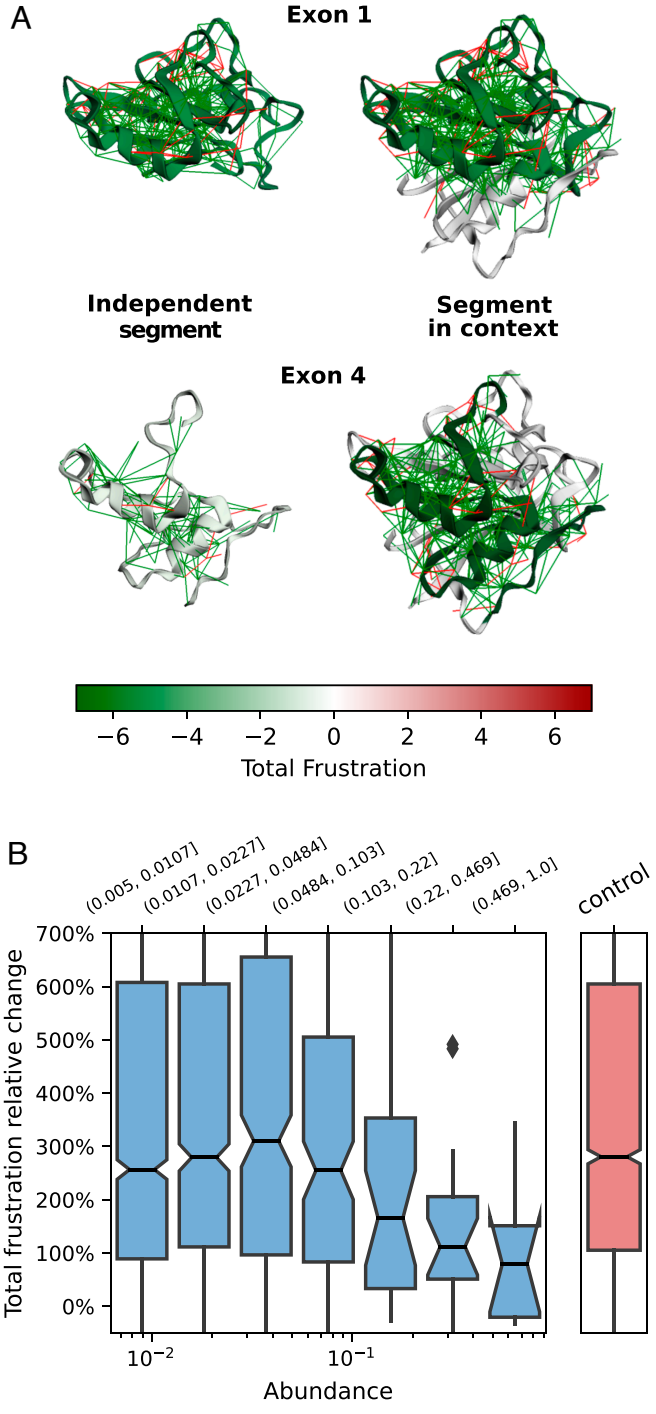

Fig. 2. Relative change in total frustration. (*A*) Definition, using as examples DHFR exons 1 and 4 (blue and red in Fig. 1*E*). The relative change in total frustration of a protein segment δf compares it in two schemes, as independent segment (*Left*) and the same segment in the context of the whole protein (*Right*). We color the segments in the reference structure (PDB: 8dfr) according to the total frustration. For exon 1, the segment in context ($f_{C1} = -5.47$) is as frustrated as the independent scheme ($f_{I1} = -5.05$), the relative change $\delta f_1 = 8\%$ is small, the exon as minimally frustrated as it can be without the rest of the protein. But for exon 4, $f_{C4} = -6.59$ while $f_{I4} = -0.43$, therefore the relative change $\delta f_4 = 1{,}446\%$ is huge. Contact frustration is added for indicating the contacts that stabilize (green) or destabilize (red) the segment in each scheme. (*B*) Total frustration relative change as a function of exon abundance for all the families together, in a box plot in logarithmic scale. Blue boxes contain the central 50% of data, with a black line in the median and a notch indicating its CI. Abundance interval for each box is indicated on *Top*. The red box on the *Right* represents the distribution of a control group of alternative exons, sampled from each family size distribution. Below an abundance of 5%, natural exons distribution is indistinguishable from the control one. Over that frequency, the frustration relative change smoothly decrease.

whole structure, as the foldon definition requires. On the other hand, exon 4—which is slightly shorter and less conserved than exon 1—has a huge relative change $\delta f_4 = 1446\%$. In this case, the exon present in the context numerous stabilizing contacts between the fragment and the rest of the protein that minimize the total frustration. Those stabilizing native interactions are absent for the independent segment. By itself, the segment is notably less foldable, making it difficult to characterize it as a foldon.

To see whether there is a systematic relationship between the foldability independence and the exon conservation, for each protein family we compute δf for all the exons having an abundance greater than 0.5% and for a control group made of exon alternatives sampled from the family size distribution. Considering all families together, the relative change in total frustration median decreases with exon abundance, as Fig. 2*B* box plot shows. Below an abundance of 5%, δf distribution for natural exons is not distinguishable from the control group distribution. But for the more abundant exons, the frustration relative change starts a descending trend. This effect does not directly result from exon length, which does not significantly change with abundance (*SI Appendix*, Fig. S4). Most conserved exons are more likely to behave as foldons than do the less abundant exons.

Minimal Common Exons. Exon boundaries are not evenly distributed along the sequences. We present a histogram of exon boundary positions for the DHFR family as a case study in Fig. 3*A* (black bars). We note that there are no absolutely prohibited positions for the exon boundaries when one considers the entire sequence alignment. In addition, high-frequency hotspots appear every 20 to 40 residues. Taking into account the exon size distribution for DHFR (Fig. 1*B*), the hotspots are too close to each other to be explained just by repeating some very abundant exons. Instead, they reveal an overlap of different sequence partitions. The hotspots can be interpreted as alternative breakup points in the exon–intron structure, conserved through the family. A similar pattern is seen for the other studied families (*SI Appendix*, Figs. S6 and S7).

The local maxima in the histogram of Fig. 3*A* (red stars) can be used to divide each protein domain (and the corresponding MSA) into a set of segments, that we call *minimal common exons* (MCE). We identify the minimal common exons with different colors along the secondary structure description on *Top* of the histogram of Fig. 3*A* and the PDB structure in Fig. 3*B*.

The relationship between the MCE and secondary structure stands out in this case. With a single exception (position 121) the hotspots do not break alpha helices or beta strands; instead, the breaks occur in coil-like regions. Equivalently, one can describe the MCE as being complete secondary structure elements or combinations of them.

We compare this picture with a neutral model, where alternative exons are generated by sampling the exponential size distribution of MCE from each and every family (*Materials and Methods*). A Z score comparing the natural MCE and those alternative pieces reveals that for the majority of the families that we studied (including DHFR) the actual boundaries occur more than expected in coil-like regions and rarely occur in alpha or beta elements (Fig. 4).

A local smoothed frustration signal can be defined computing the total mutational frustration for a segment of five residues on a sliding window (Fig. 3*A*, blue line). This signal shows some correlation with secondary structure categories. The beta regions generally have lower frustration than the rest of the structure on average (*SI Appendix*, Fig. S5). A Z score comparing natural MCE and pieces sampled from a neutral model shows that frustration is higher than expected on boundaries for the majority of families (Fig. 4), but there are some exceptions to the pattern.

We compute δf, the relative change in total frustration for the MCE. A comparison of the MCE δf distribution with that generated by the neutral model yields heterogeneous results. Only around one third of the families that we studied have significantly more independently foldable MCE than the neutral model would give (Fig. 4).

We see that MCE are not as independently foldable as actual exons. They seem to be too short to be independent from the rest of the protein. Instead, they work as fundamental units that

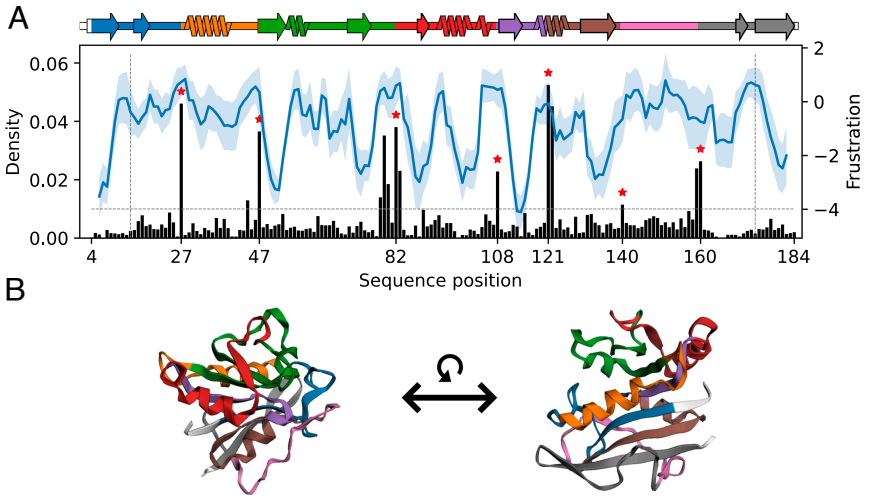

Fig. 3. Exon boundary analysis for DHFR. (*A*) Histogram of exon boundaries (black bars). Boundary hotspots, histogram local maxima, are marked with red stars. Below the horizontal dashed gray (density = 0.01) line we ignore the peaks, considering them background noise. We also ignore peaks closer than 10 residues from each other or to alignment limits (vertical dashed gray lines). Over the histogram, the local smoothed frustration signal (blue), its average over the sequences as a solid line and SD as a shadow. On *Top*, the secondary structure representation of the reference structure of the family (PDB: 8dfr). Colors represent the minimal common exons (MCE), the sequence partition given by boundary histogram hotspots. Almost all the MCE are made of uninterrupted secondary structure elements or combinations of them. (*B*) Minimal common exons projected on the reference 3D structure with the same colors used in panel (*A*), with two different orientations of the structure.

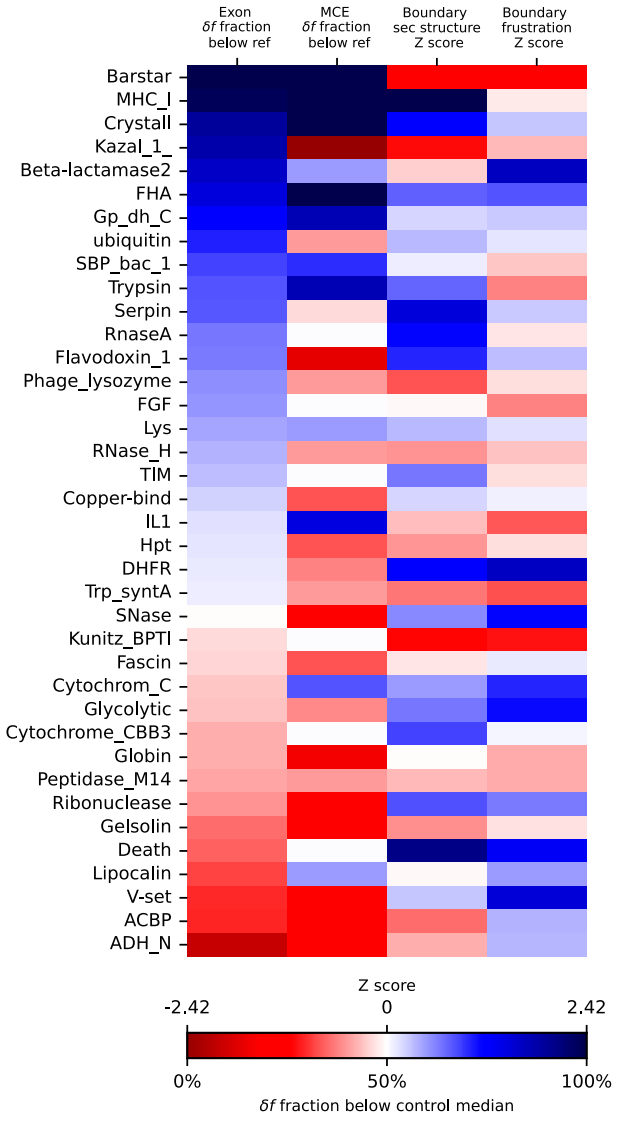

Fig. 4. Summary for each family. We include two groups of results in heat maps, each one with the corresponding scale at the *Bottom*. First and second columns on the *Left* represent the fraction of δf distribution below the family control group median for the actual exons (first) and the minimal common exons (second). These scores go from 0% (red), where all the exons are less independently foldable than the reference, to 100% (blue) where all the exons are more independently foldable than it. The families are sorted using the first column score. The last two columns on the *Right* represent Z scores comparing boundary hot spot positions (MCE boundaries) with alternative ones generated with a neutral model. In the third column, the score is positive (blue) when boundaries occur more than expected in coil-like regions. In the last column, the score is positive (blue) when boundaries occur in regions that are more frustrated than expected.

can alternatively combine into different bigger segments (the real exons). Within these possibilities, the most frequent ones stand out as more independently foldable than random segments.

Family Specific Characteristics. We summarize the results obtained from four different approaches for the actual exons of the 38 studied families in a heat map (Fig. 4). The first heat map column on the *Left* represents the fraction of the δf distribution having values below the median of the alternative exons control group, which is family-specific. This score runs from 0% (red), where all the exons are less independently foldable than the alternative exon sampled reference, to 100% (blue) where all the exons are more independently foldable than in the reference. The families are sorted according to this statistic. We performed a two-tailed Mann–Whitney U statistical test for each family and found that the trend for each family individually is significant for 17 out of all 38 studied families (*SI Appendix*, Fig. S9 and Table S2).

For the majority but not all of the families studied the natural exons are more independently foldable than would be expected by chance. Nevertheless, the results are somewhat diverse, and for some families, the alternative pieces generated at random are more independently foldable than the naturally occurring segments. This may reflect distinct evolutionary histories of the different families, highlighting that there may be no common mechanism involved in the exonic partitioning for every individual protein family. It is likely that the selection pressure for foldability is not the only controlling driving force for the location of exon boundaries. This is not unexpected because functional sites are often required to be frustrated but must be conserved for function (30). Those protein families for which nonnatural chance partitions are more independently foldable than the actual exons (red in the first column of Fig. 4) may also hint that other strong biological determinants are at play, for example at the RNA level.

We compare the δf for the MCE with the correspondent control group. The fraction of the δf distribution below the control group median for MCE is shown in the second column of the heat map (Fig. 4). With a few exceptions (IL1, Cytochrome C, Lipocalin), the actual exons show higher scores than do the MCE of the same family.

The third column represents the Z score that compares MCE boundaries to those of a neutral control group. A positive Z score (blue) indicates there are more boundaries in coil-like regions than what would be expected based on a neutral model, while a negative score (red) would indicate that there are more boundaries in the stable secondary structure regions. Finally, the last column shows whether the MCE boundaries are more frustrated than expected, measuring frustration using a five-residues sliding window.

We find that protein families display several patterns when we take together the analysis of exon foldability and boundary occurrence regions. For some families, MHC-I, Crystall, Beta-lactamase, forkhead associated domain (FHA), Gpdh-C, and Trypsin, exons are independently foldable and codify mostly uninterrupted stable secondary structure regions. In another set of families, the most common boundaries are clearly not random, but exon folding does not seem to be the most relevant signature for their evolutionary selection. This is the case for Ribonuclease, Death, V-set, Glycolytic, and SNase. There are some other examples however, Barstar, IL1, Kazal, and Phage lysozyme, where the actual exons are more independently foldable than expected, but their boundaries do not particularly occur in highly frustrated or coil-like regions.

In the case study presented previously in this work, DHFR, as in ubiquitin, Serpin, and Flavodoxin, we find that the actual exons are more independently foldable than expected, but this is not the case for the minimal common exons. Interestingly, these minimal common exons are mostly contiguous secondary structure elements that may not fold by themselves but can combine into larger segments—the actual exons—that are less frustrated.

Concluding Remarks

We have revisited the correspondence between exons and protein folding modules. By mapping the exon–intron boundaries to multiple sequence alignments, we identified conserved exon

partitions. For each protein family, the size distribution of exons deviates from exponential decay due to particular and very common instances. A neutral model, where intron positions are chosen through sequential independent trials of a stochastic process, cannot explain these patterns. Through frustration analysis, we found that protein segments corresponding to the most common exons are clearly more independently foldable than others. On average, the size of the foldable fragments does not change with exon abundance. Presumably, natural selection acting on exons is influencing the size distribution by taking into account the folding of the corresponding protein fragment. If exons have been shuffled during evolution, the foldability independence of the protein region encoded by an exon becomes an advantageous feature, allowing it to be inserted in a different topology or copied in tandem and maximizing the chances of giving rise to a foldable polymer.

Unfortunately, we are still lacking a unique and consistent way to experimentally define foldons in natural proteins, thus there is no universal ground truth in the laboratory to directly evaluate the foldon-exon correspondence. We have proposed in this paper an alternative computational evaluation of the independent foldability of exonic regions as compared to alternative partitions of the primary structure.

The most common exons can function as folding units. Nevertheless, it's important to note that these exons may not always span the entire protein domain; instances exist where they overlap with each other. We define a systematic way of partitioning a multiple sequence alignment into nonoverlapping segments using the exon boundary histogram hot spot positions along the sequence. These selected hot spots divide the protein into MCE. For the majority of the studied families, the MCE consists of uninterrupted alpha and/or beta elements, and the boundaries between them occur in highly frustrated or coil regions. This co-occurrence has been previously studied in earlier works, but no significant tendency to co-occur was reported (10, 11).

While it has been observed that domain boundaries may match exon boundaries (12–14), our results show that there is an internal structure within the protein domains. The most conserved intron positions define possible splitting points for the actual modules of a protein domain. The diversity of exons within a protein family arises from the alternative usage of these breaking points, forming the actual exons. It should be noted that each family may have a different evolutionary history, where the exon boundaries may be seen as scars of that history and may be maximizing the chances of giving rise to a new fold. We have shown that in certain families, conserved exon boundaries clearly delineate secondary structure elements, whereas in other families, exon frustration is remarkably minimal. Folding in vivo may require the assistance of external factors that may interfere with autonomous protein folding, and this effect may be acting not on a family but on specific family members (31). We propose that both aspects must be taken into account when analyzing the relations between protein folding and evolution of particular protein domains.

Materials and Methods

Data Curation. A total of 38 well-behaved protein families with distinct topologies (all alpha, all-beta, alpha+beta, alpha/beta) were used, including the instances studied in the first paper that analyzed exon–foldon correspondence (8) along with an additional 26 protein families of the Start2fold database (32), consulted in January 2023. Protein MSA for each family were obtained in December 2022 from Pfam (33), now hosted by InterPro database (34). For minimizing phylogenetic bias within each MSA, we clustered by full sequence similarity using CD-hit (35) at 90% cutoff and we assigned a weight to each sequence defined as $1/n_i$, being n_i the number of sequences in the ith cluster. All the statistics were made taking into account these sequence weights. We used a target 3D structure selected from the PDB (36) for each family (*SI Appendix*, Table S1) and we aligned the MSA to its corresponding sequence, keeping only the positions of the MSA that are present in the target sequence. To summarize, MSA positions are Pfam domain positions in the target PDB structure. All the calculations that involve the protein tertiary structure were made using the target PDB structure selected for the family. Secondary structure data were obtained using the Define Secondary Structure of Proteins (DSSP) algorithm (37) on the target structures. Exon data were obtained from GenBank database (38). We downloaded all the gene files corresponding to the Uniprot IDs (39) in our MSAs, excluding Bacteria. Single-exon sequences were excluded from the analysis. We parsed the gene files to get the exon positions and we mapped them to the corresponding MSA, obtaining the amino acid sequence segment corresponding to each exon. Every exon starting and ending position was referenced to the MSA. We calculated the exon relative abundance as the sum of the sequence weights of all the sequences that have an exon in the same position of the MSA, normalized by the sum of the sequence weights of the protein family. Data download and curation were carried out using python scripts. The code is available at GitHub: https://github.com/eagalpern/exon-foldon.

Total Frustration Relative Change. To determine the energy of a protein segment we used the AWSEM coarse-grained forcefield, including only the burial and the contact terms (28). We used a single 3D target structure for each family and we threaded it with the particular sequence we wanted to evaluate. The total energy of a segment is calculated according to two different scenarios. The independent (I) scheme energy includes the contact terms of all the pairs within the segment, while the in-context scheme (C) considers also all the contacts between segment residues and other protein positions outside it,

$$H_I = \sum_{i=a}^{b} H_i^{\text{burial}} + \sum_{i=a}^{b}\sum_{j=a}^{b} H_{ij}^{\text{contact}} \quad \textbf{[2a]}$$

$$H_C = \sum_{i=a}^{b} H_i^{\text{burial}} + \sum_{i=a}^{b}\sum_{j=1}^{L} H_{ij}^{\text{contact}}, \quad \textbf{[2b]}$$

where the segment goes from position a to b, within a sequence of L residues. For each exon, segment energy is a weighted average over all the sequences in the alignment that have the exon. We exclude from the average exons where gaps represent more than 50% of their sequence. The frustration f was calculated using decoy sets, constructed by randomizing the identity of the amino acids of the complete sequences. Decoy sequences were also threaded through the same tertiary structure selected for the family. Energy calculations were made using a python implementation of the protein frustratometer (40), available at Github: https://github.com/HanaJaafari/DCA_Frustratometer. Total frustration calculation scripts are available at GitHub: https://github.com/eagalpern/exon-foldon.

Local Frustration. The local frustration of position x was calculated evaluating the total frustration (as defined previously) of a five-residues segment centered on x. The signal was obtained by sliding this five-residue window along the sequence. We consider each segment in the context of the whole protein.

Boundary Local Maxima Searching Criteria. We searched for relative maxima in the exon boundary histograms comparing positions with another 10 to each side. We discarded any maximum closer than 10 positions to the beginning or to the end of the sequence, and also positions with histogram density below 0.01. The exon assignment is robust to small changes in these parameters for most of the families, as the effective exons count is large.

Visualization Tools. Secondary structure linear visualizations were made adapting SSDraw python library (41). Tertiary structure visualizations were made using py3Dmol python library (42).

Exon Control Groups. To compare the properties of actual exons and MCE, we generated specific sets of exon alternatives as control groups. Each exon alternative represents a fragment defined by its initial sequence position and length within a protein family. For the actual exons, we sampled each family's exon size distribution and a random initial position to generate exon alternatives, resulting in family-specific exon control groups. To measure the energy of an exon alternative, we assigned 100 sequences randomly selected from the corresponding positions of the family's multiple sequence alignment to each generated fragment. These sequences were then threaded along the reference tertiary structure, and we then calculated the energy average over the 100 sequences. As MCEs are limited in number per family, MCE alternatives were obtained by generating consecutive segments, sampling their sizes from a geometrical distribution fitted from the sizes of all MCEs across families. This approximation follows a neutral model for the size distribution, leading us to designate this set of MCE alternatives as the neutral control group. MCE alternatives shorter than 10 residues were eliminated, as they are smaller than the minimum distance we imposed between histogram maxima (10 residues).

Fraction Scores. The distribution of the total frustration relative change for exons δf_{exon} was compared for each family with the corresponding size-wise control group distribution. We used as a score the weighted fraction of δf_{exon} below the median of the distribution $\delta f_{\text{control}}$, given by

$$frac.score^{\text{exon}} = \sum_i w_i \, \delta_{i,\, \delta f_i < median(\delta f_{\text{control}})}, \qquad [3]$$

where w_i is the abundance of the exon i and $\delta_{i,x}$ is the Kronecker symbol, taking value one if the condition x is True for the exon i and zero otherwise.

For the MCE, the reference is given by the median of $\delta f_{\text{control}}$ for the MCE control group

$$frac.score^{\text{MCE}} = \sum_i \delta_{i,\, \delta f_i < median(\delta f_{\text{control}})} / N, \qquad [4]$$

where i are the MCE and N is the number of MCE for the family.

Boundary Secondary Structure Z Score. The occurrence of exon boundary local maxima (or MCE boundaries) on coil-like regions (not alpha or beta) for a family was compared to the occurrence on the alternative partitions that define the MCE control using a Z score defined as

$$Z\,score^{\text{coil}} = \frac{\bar{\delta}_{\text{MCE}} - \langle \bar{\delta}_{\text{control}} \rangle}{\sigma_{\delta_{\text{control}}}}, \qquad [5]$$

where δ is the Kronecker symbol, $\bar{*}$ represents the average over the partition, $\langle * \rangle$ the average over all the decoy partitions and σ the SD. We consider the boundary as being the ending position of each segment (MCE or control) and the first position of the next one. If at least one of them is not an alpha or beta region, we take that boundary i as a positive case $\delta^i = 1$, while if the two of them are beta and/or alpha, $\delta^i = 0$.

Boundary Frustration Z Score. Local frustration on exon boundary local maxima (or MCE boundaries) for a family was compared with the frustration on the alternative partitions that define the MCE control using a Z score defined as

$$Z\,score^{f} = \frac{\bar{f}_{\text{MCE}} - \langle \bar{f}_{\text{control}} \rangle}{\sigma_{f_{\text{control}}}}, \qquad [6]$$

where f is the local frustration, $\bar{*}$ represents the average over the boundary positions of a partition, $\langle * \rangle$ the average over all the control partitions and σ the SD. We consider as boundary the ending position of each segment (MCE or control) and the first position of the next one.

Data, Materials, and Software Availability. All input data needed to reproduce the main results, including Figs. 1 *D* and *E* and 3 for the 38 protein families that we studied is available at GitHub: https://github.com/eagalpern/exon-foldon (43), along with a Jupyter notebook for visualization.

ACKNOWLEDGMENTS. This work was supported by the Consejo de Investigaciones Científicas y Técnicas (CONICET) (D.U.F. is CONICET researchers and E.A.G. is a postdoctoral fellow); CONICET Grant PIP2022-2024–11220210100704CO. Universidad de Buenos Aires UBACyT 20020220200106BA. Additional support from NASA Astrobiology Institute (NAI) and Grant Number 80NSSC18M0093 Proposal ENIGMA: EVOLUTION OF NANOMACHINES IN GEOSPHERES AND MICROBIAL ANCESTORS (NASA ASTROBIOLOGY INSTITUTE CYCLE 8). PGW was supported both by the Bullard–Welch Chair at Rice University, grant C-0016, and by the Center for Theoretical Biological Physics sponsored by NSF grant PHY-2019745. We call the attention of the international scientific community to the potential erosion of Argentina's strong scientific tradition due to current funding constraints and the sudden termination of long-term policies.

Author affiliations: [a]Protein Physiology Lab, Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina; [b]Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales, Consejo Nacional de Investigaciones Cientificas y Tecnicas - Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina; [c]Center for Theoretical Biological Physics, Rice University, Houston, TX 77005; [d]Applied Physics Graduate Program, Smalley-Curl Institute, Rice University, Houston, TX 77005; [e]Department of Chemistry, Rice University, Houston, TX 77005; and [f]Department of Physics, Rice University, Houston, TX 77005

1. D. B. Wetlaufer, Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **70**, 697–701 (1973).
2. W. Gilbert, Why genes in pieces? *Nature* **271**, 501–501 (1978).
3. C. C. Blake, Do genes-in-pieces imply proteins-in-pieces? *Nature* **273**, 267–267 (1978).
4. M. Gō, Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**, 90–92 (1981).
5. L. Patthy, Genome evolution and the evolution of exon-shuffling–a review. *Gene* **238**, 103–114 (1999).
6. J. D. Bryngelson, P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7524–7528 (1987).
7. P. G. Wolynes, Energy landscapes and solved protein-folding problems. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **363**, 453–467 (2005).
8. A. R. Panchenko, Z. Luthey-Schulten, P. G. Wolynes, Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 2008–2013 (1996).
9. A. R. Panchenko, Z. Luthey-Schulten, R. Cole, P. G. Wolynes, The foldon universe: A survey of structural similarity and self-recognition of independently folding units. *J. Mol. Biol.* **272**, 95–105 (1997).
10. A. Stoltzfus, D. F. Spencer, M. Zuker, J. M. Logsdon Jr, W. F. Doolittle, Testing the exon theory of genes: The evidence from protein structure. *Science* **265**, 202–207 (1994).
11. K. Weber, W. Kabsch, Intron positions in actin genes seem unrelated to the secondary structure of the protein. *EMBO J.* **13**, 1280–1286 (1994).
12. M. Liu, A. Grigoriev, Protein domains correlate strongly with exons in multiple eukaryotic genomes-evidence of exon shuffling? *Trend. Genet.* **20**, 399–403 (2004).
13. B. Smithers, M. Oates, J. Gough, 'why genes in pieces?'–revisited *Nucleic Acids Res.* **47**, 4970–4973 (2019).
14. X. Cui, M. Stolzer, D. Durand, Evidence for exon shuffling is sensitive to model choice. *J. Bioinform. Comput. Biol.* **19**, 2140013 (2021).
15. P. Weinkam, C. Zong, P. G. Wolynes, A funneled energy landscape for cytochrome c directly predicts the sequential folding route inferred from hydrogen exchange experiments. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12401–12406 (2005).
16. H. Maity, M. Maity, M. M. Krishna, L. Mayne, S. W. Englander, Protein folding: The stepwise assembly of foldon units. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 4741–4746 (2005).
17. P. Weinkam, J. Zimmermann, F. E. Romesberg, P. G. Wolynes, The folding energy landscape and free energy excitations of cytochrome c. *Acc. Chem. Res.* **43**, 652–660 (2010).
18. C. V. Gegg, K. E. Bowers, C. R. Matthews, Probing minimal independent folding units in dihydrofolate reductase by molecular dissection. *Protein Sci.* **6**, 1885–1892 (1997).
19. M. Iwakura, T. Nakamura, C. Yamane, K. Maki, Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.* **7**, 580–585 (2000).
20. R. Shiba *et al*., Systematic alanine insertion reveals the essential regions that encode structure formation and activity of dihydrofolate reductase. *Biophysics* **7**, 1–10 (2011).
21. Y. Takase, Y. Yamazaki, Y. Hayashi, S. Toma-Fukai, H. Kamikubo, Structure elements can be predicted using the contact volume among protein residues. *Biophys. Physicobiol.* **18**, 50–59 (2021).
22. M. O. Lindberg, M. Oliveberg, Malleability of protein folding pathways: A simple reason for complex behaviour. *Curr. Opin. Struct. Biol.* **17**, 21–29 (2007).
23. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein sectors: Evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
24. N. P. Schafer *et al*., Discrete kinetic models from funneled energy landscape simulations. *PLoS One* **7**, e50635 (2012).

25. D. U. Ferreiro, A. M. Walczak, E. A. Komives, P. G. Wolynes, The energy landscapes of repeat-containing proteins: Topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLoS Comput. Biol.* **4**, e1000070 (2008).
26. E. A. Galpern, J. Marchi, T. Mora, A. M. Walczak, D. U. Ferreiro, Evolution and folding of repeat proteins. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2204131119 (2022).
27. T. O. Street, G. D. Rose, D. Barrick, The role of introns in repeat protein gene formation. *J. Mol. Biol.* **360**, 258–266 (2006).
28. A. Davtyan *et al*., AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **116**, 8494–8503 (2012).
29. D. U. Ferreiro, J. A. Hegler, E. A. Komives, P. G. Wolynes, Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19819–19824 (2007).
30. M. I. Freiberger *et al*., Local energetic frustration conservation in protein families and superfamilies. *Nat. Commun.* **14**, 8379 (2023).
31. P. To, B. Whitehead, H. E. Tarbox, S. D. Fried, Nonrefoldability is pervasive across the *E. coli* proteome. *J. Am. Chem. Soc.* **143**, 11435–11448 (2021).
32. R. Pancsa, M. Varadi, P. Tompa, W. F. Vranken, Start2Fold: A database of hydrogen/deuterium exchange data on protein folding and stability. *Nucleic Acids Res.* **44**, D429–D434 (2016).
33. R. D. Finn *et al*., The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
34. T. Paysan-Lafosse *et al*., Interpro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
35. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
36. H. M. Berman *et al*., The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
37. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopol. Orig. Res. Biomol.* **22**, 2577–2637 (1983).
38. D. A. Benson *et al*., Genbank. *Nucleic Acids Res.* **41**, D36–D42 (2012).
39. E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, A. Bairoch, *Uniprotkb/Swiss-Prot in Plant Bioinformatics* (Springer, 2007), pp. 89–112.
40. A. O. Rausch *et al*., Frustratometer: An R-package to compute local frustration in protein structures, point mutants and MD simulations. *Bioinformatics* **37**, 3038–3040 (2021).
41. E. A. Chen, L. L. Porter, SDDraw: Software for generating comparative protein secondary structure diagrams. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.08.25.554905 (Accessed 1 December 2023).
42. N. Rego, D. Koes, 3Dmol.js: Molecular visualization with WebGL. *Bioinformatics* **31**, 1322–1324 (2015).
43. E. A. Galpern, H. Jaafari, C. Bueno, P. G. Wolynes, D. U. Ferreiro, Data from 'Reassessing the Exon-Foldon correspondence using Frustration Analysis'. Github. https://github.com/eagalpern/exon-foldon/tree/main/data. Deposited 26 December 2023.