Where Natural Protein Sequences Stand out From Randomness

Laura Weidmann[1], **Tjeerd Dijkstra**[2], **Oliver Kohlbacher**[2,3,4,5,6], **Andrei Lupas**[1]

***For correspondence:**
lweidmann@tuebingen.mpg.de

[1]Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany; [2]Biomolecular Interactions, Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany.; [3]Applied Bioinformatics, Department for Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany; [4]Institute for Biomedical Informatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany; [5]Center for Quantitative Biology, University of Tübingen, Sand 14, 72076 Tübingen, Germany; [6]Translational Bioinformatics, University Hospital Tübingen, Hoppe-Seyler-Str. 9, 72076 Tübingen

---

Abstract Biological sequences are the product of natural selection, raising the expectation that they differ substantially from random sequences. We test this expectation by analyzing all fragments of a given length derived from either a natural dataset or different random models. For this, we compile all distances in sequence space between fragments within each dataset and compare the resulting distance distributions between sets. Even for 100mers, 95.4% of all distances between natural fragments are in accordance with those of a random model incorporating the natural residue composition. Hence, natural sequences are distributed almost randomly in global sequence space. When further accounting for the specific residue composition of domain-sized fragments, 99.2% of all distances between natural fragments can be modeled. Local residue composition, which might reflect biophysical constraints on protein structure, is thus the predominant feature characterizing distances between natural sequences globally, whereas homologous effects are only barely detectable.

Introduction

Natural proteins form the backbone of the complicated biochemical network that has given rise to the great variety of life on Earth. This highly interwoven framework of reactions seems impossible to have arisen by chance, simply because the great majority of random protein sequences fails to form a specific structure, let alone possess chemical activity. Features that distinguish naturally evolved from random sequences are therefore of great interest, both in order to understand protein evolution ***Shah et al.*** (***2015***) ***Luigi Luisi*** (***2003***) and to guide the design of new proteins ***Woolfson et al.*** (***2015***) ***Pande et al.*** (***1994***).

Searches for such differences have hitherto focused on the exhaustive enumeration of short peptides and their statistical analysis by exact occurrence ***Poznański et al.*** (***2018***) ***Lavelle and Pearson*** (***2009***). These studies showed that the natural frequency of most peptides is similar to that expected from random sequences with the same composition. Nevertheless, the frequency of some peptides was found to deviate substantially from random occurrence, an observation which was variously discussed in terms of homologous descent and convergence due to structural and

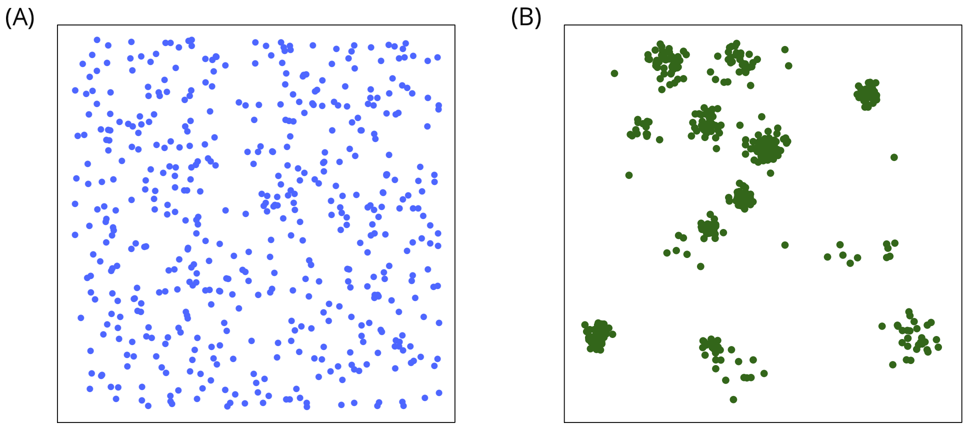

Figure 1. Sketches of sequence space occupation. (A) A random distribution has no structure. Purely random sequences are thus distributed homogeneously over the possible sequence space. With a blue circle, we represent a sequence in an abstract representation of space. (B) Natural sequences are known to frequently arise by replication and diversification. Recently duplicated sequences that have not diverged beyond obvious recognition form cluster in local areas of sequence space, indicated by green circles. Through significant similarity among multiple sequences, even very distant relatives can be assigned to a cluster of homologous sequences. This concept of sequence clusters, like islands in a gigantic ocean of possibilities, is commonly used.

---

functional constraints. This enumeration approach quickly reaches its limits at sequence lengths above 5, due to the fact that there are simply not enough natural sequences to populate the exponentially growing sequence space. Furthermore, pentapeptides are far from having a relevant length for understanding protein sequences. Even if proteins are dissected into their constituent domains, relevant sequence lengths still mostly range above 80 residues. At a complexity of 20^{80}, it is clear that this sequence space cannot be analyzed by an enumeration approach.

Although the sequence space of domain-sized fragments appears intractable due to its size, we have nevertheless developed expectations about its occupation by natural sequences through decades of bioinformatic research. This is because most proteins have arisen by descent and differentiation from a set of domain prototypes, and can thus be classified into a hierarchy of domain families and superfamilies. This points to the fact that the sequence space around domains is substantially populated by their homologs, resulting in an image of local islands of natural sequences within a global sea of virtual, unrealized possibilities (***Figure 1***). The extent to which this image is adequate to describe the global sequence distribution is however unclear.

A first step to extend from local sequence islands to a more global view has been taken with searches for variants close to existing proteins ***Bershtein et al.*** (***2017***) ***Starr et al.*** (***2017***) ***Harms and Thornton*** (***2014***) ***Urlinger et al.*** (***2000***). By testing exhaustively all mutations at certain sites, these studies bypass intermediate mutants that would not have viable in evolution. Contrasting the abundance of possible functional variants to the small number of natural sequences demonstrates how sparsely nature has explored sequence space, even locally. The high energy barriers, epistatic effects, and functional dependencies prevent the establishment of random mutations and seem to entrench already existing and functional forms ***Starr and Thornton*** (***2016***) ***Shah et al.*** (***2015***).

Modern techniques of protein design allow to reach out further into the global sequence space to find possible exemplars in unknown territory ***Huang et al.*** (***2016***) ***Woolfson et al.*** (***2015***). Scaling these scans up to the currently highest practicable level for a given structure or function has uncovered viable solutions far from existing proteins ***Stiffler et al.*** (***2019***) ***Chevalier et al.*** (***2017***) ***Larson et al.*** (***2002***), showing that sequence similarity to existing proteins is not required for functionality. This leads to the hypothesis that the usable part of sequence space is mostly randomly structured, which has been proposed for unrelated natural sequences before ***Lavelle and Pearson*** (***2009***).

Apart from the seemingly random global structure, there are nevertheless biophysical requirements for all usable protein sequences, natural as well as designed, such as foldability, hydrophobic core formation and solubility. This indicates that these proteins may share some convergent features, which would restrict a random drift away into unstructured space. Natural sequence space could thus be characterized globally by sequences with the potential to fold, i.e. by convergent features.

In this paper, we analyze the global structure of natural sequence space, aiming to identify general features that characterize natural sequences and to evaluate the relative contributions of convergence and homology to this space. We do this by contrasting natural data with a variety of random models, in order to extract sequence features arising from different natural mechanisms.

Results and Discussion

Natural sequence data and random sequence models

Choice of a natural dataset

For an adequate dataset that reflects the natural protein sequence space, we aimed to achieve a reasonable coverage of deep phylogenetic branches with complete and well-annotated proteomes. Given that the genome coverage for the archaeal and eukaryotic lineages is still sparser than for bacteria and that particularly eukaryotic genomes are affected by issues of assembly, gene detection, and intron-exon boundaries, we built our database from the derived bacterial proteomes collected in UniProt ***Apweiler*** (***2009***). To control for redundancy, we selected only one genome per genus and filtered each for identical open reading frames and low-complexity regions. In total our dataset comprises 1,307 genomes, $4.7 \cdot 10^6$ proteins, and $1.2 \cdot 10^9$ residues. We simplified complexities arising from the use of modified versions of the 20 proteinogenic amino acids, which occurred in a few hundred cases, by converting these to their unmodified precursors, thus maintaining an alphabet of 20 characters throughout. Further details on the generation of our dataset and its specific content are provided in the Methods section.

In order to evaluate where our natural dataset differs from randomness, we developed a series of increasingly specific random models that account for compositional effects.

How random is random?

Our most basal model considers completely random sequences of the 20 proteinogenic amino acids, in which each occurs with an equal probability of 5% (E-model). This model is known to approximate natural sequences only poorly ***de Lucrezia et al.*** (***2012***) ***Munteanu et al.*** (***2008***). This is hardly surprising as natural amino acid frequencies in fact range between 1% and 10%, a bias which is associated with metabolic pathways, bio-availability, and codon frequency. We therefore built models that factor in this compositional bias at increasingly local levels. The first model incorporates the global amino acid composition of our natural dataset, which we refer to as the A-model.

More specific models consider increasingly local fluctuations in composition. The composition of different genomes, for example, varies with GC-content and environmental influences ***Fukuchi and Nishikawa*** (***2001***) ***Fukuchi et al.*** (***2003***). This effect can be factored in using the individual genome composition (G-model). With an increasingly local focus, compositional bias can be accounted for at the level of proteins (P-model) ***Chou*** (***2001***) ***Cedano et al.*** (***1997***), domains (D-model) ***Lavelle and Pearson*** (***2009***) and even sub-domain-sized fragments ***Poznański et al.*** (***2018***).

Having accounted for compositional effects resulting from environment, metabolism, and the need to form a hydrophobic core, the remaining differences between natural and random sequences must be attributed to sequence effects, due either to *divergence* from a common ancestor ***Alva et al.*** (***2015***) or *convergence* as a result of secondary structure formation ***Pande et al.*** (***1994***).

Table 1. Random sequence models based on amino acid composition.

model	natural feature	class of feature
E	natural amino acid alphabet, equal propensity for each letter	single, overall
A	overall amino acid composition	descriptor
T	overall dipeptide frequency	
G	composition of individual genomes	context-specific
P	composition of proteins	composition
D	composition of domain-sized fragments	
D1	D-model + homology sequence bias	mixed models that
D2	D-model + analogy sequence bias	incorporate
D3	D-model + homology and analogy sequence bias	sequence bias

Table 1–source data 1. Random sequence models. Completely random sequences, where each amino acid occurs with the same probability of 5%, are represented by the E-model. The natural frequency of specific amino acids deviates remarkably form such an equal distribution, thus, random sequence models are usually based on the overall amino acid composition, represented by the A-model. The overall dipeptide frequency is considered by the T-model. The diversity of amino acid composition across genomes, is accounted for by the G-model. On a more specific level, the composition occuring in natural proteins or even domain-sized fragments can be used to generate random sequence models, here referred to as P- and D-models. In order to estimate the contribution of analogous and homologous relationships to the global occupation of sequence space, we generated models D1, D2 and D3 that include sequence bias in addition to the composition bias od the D-model. (These models will be explained in detail in the last section of the Results.) We compare our natural dataset to all of these models and illustrate to what extent they differ from the natural sequence space occupation. Our implementation of the models are described in the Methods section.

Representing sequence space occupation based on pairwise distances

Sequence space has frequently been analyzed with a direct approach based on the exhaustive enumeration of natural kmers, and the comparison of their frequencies to those derived from a random model ***Poznański et al.*** (***2018***) ***Lavelle and Pearson*** (***2009***). This approach is restricted to kmers of length 5 or smaller, due to sequence space complexity and the data sparsity caused thereby. It also does not represent the relative position among kmers within the global sequence space.

We use an indirect approach to circumvent these problems. Our approach is built on the probability mass function of pairwise distances between sequences of the same length, in the following referred to as *distance distribution*. A distance distribution illustrates how often sequences are positioned at a certain distance to each other and we use it to study the way sequences are spread across the possible space. We built distance distributions for the natural dataset and for each dataset of random sequences derived from specific models. By using lengths of up to 100 residues, our sequences thus reach domain size ***Wheelan et al.*** (***2000***).

As a metric for distance, we use the *normalized local alignment score* of a Smith-Waterman alignment, since this metric is commonly used to capture similarities between natural sequences ***Rost*** (***1999***) ***Schneider et al.*** (***1997***). We note that the choice of distance metric is not of great relevance for the main implications of our study; relative to each other, the distance distributions of the random models deviate similarly from that of natural sequences irrespective of the chosen metric, as outlined in the following Results sections. Details on the derivation of distance distributions and the used distance metrics are provided in the Methods section. In this context, it is important to note that our method differs from common approaches, as it only considers the pairwise similarity between two sequences and thus their actual distance in sequence space. In contrast, most bioinformatic methods that compare sequences to each other scale distances according to their statistical significance and in many cases iterate comparisons in order to extract patterns of conserved residues, as indicators of homologous relationships. These approaches result in distances that reflect evolutionary relationships, visualized as islands of higher density in sequence cluster maps ***Alva et al.*** (***2009***).

Studying the layout of space through pairwise distances is common in other fields, such as protein structure determination ***Wüthrich*** (***1986***), spatial statistics ***Diggle*** (***2014***) and economics ***Duranton and Overman*** (***2005***), but has not, to our knowledge, been applied to investigate protein sequence space. Such distance-based approaches do not preserve information about specific positions of data points in space, but rather characterize their global distribution, which includes *global clustering and dispersion*. A corollary of this is that distinct datasets become comparable through their distance distribution, even if they do not share any specific data points.

Comparing distance distributions

For the comparison of the natural to a random distance distribution, we first subtract the fraction of distances observed in the random dataset from that observed in the natural dataset for each alignment score. We refer to this difference as the *residual*. Over all alignment scores, residuals sum up to zero and may have values that are either positive (more natural distances) or negative (more random distances). In order to obtain an overall measure of how different two distance distributions are, we derive the *total residual*, which is the variational distance between two probability mass functions. More precisely, the total residual is the sum over the absolute residuals, normalized to a range between 0% and 100%.

If the two distance distributions are completely non-overlapping, the total residual assumes the maximal value of 100%, indicating that no distance between natural fragments can be modeled with the underlying random sequences. If they are identical, the total residual assumes a value of 0%, indicating that 100% of all distances in the natural distribution have a corresponding distance in the random distribution. Thereby, the total residual represents the fraction of natural distances that are not accounted for by the distance distribution of a random model.

Global amino acid composition (A-model)

We start our analysis by assessing to what extent the global amino acid composition, as captured in the A-model, can describe natural sequences. We compare the distance distributions of the two datasets for fragment lengths up to 100 residues, in increments of 10. At all fragment lengths, the results are closely comparable. We show the results for 100mers as representative for domain-sized sequences in ***Figure 2*** and provide the others in the supplementary figures.

The distance distributions of natural and A-model data overlap extensively (***Figure 2***: A). Both are uni-modal with a peak at a low alignment score of 11%. Their minor differences only become apparent, when their residuals are considered (***Figure 2***: B). These take the shape of a wave, with two crests at alignment scores of 9% and 15% (reflecting an over-representation of the corresponding distances in the natural dataset), and a trough at 11% (reflecting an under-representation). The over-representation of distances both longer and shorter than expected from the random model, suggests that natural sequences are less homogeneously distributed in space. We rationalize this effect with the observation that natural sequences are enriched in certain parts of sequence space, leading to an increase in shorter distances. This may occur both in regions with rare amino acids (such as Cys, Trp and His in small proteins dominated by zinc-coordination and disulfide bonds ***Vallee and Auld*** (***1990***)) and in regions with abundant amino acids (such as Leu, Ala and Glu in all-alpha proteins, most extremely in coiled coils ***Lupas et al.*** (***1991***)). The compositional differences in these enriched regions mean that their distance in sequence space will be larger than expected from the A-model, and thus lead to a complementary increase in longer distances. Since residuals add up to zero, the number of intermediate distances is correspondingly decreased.

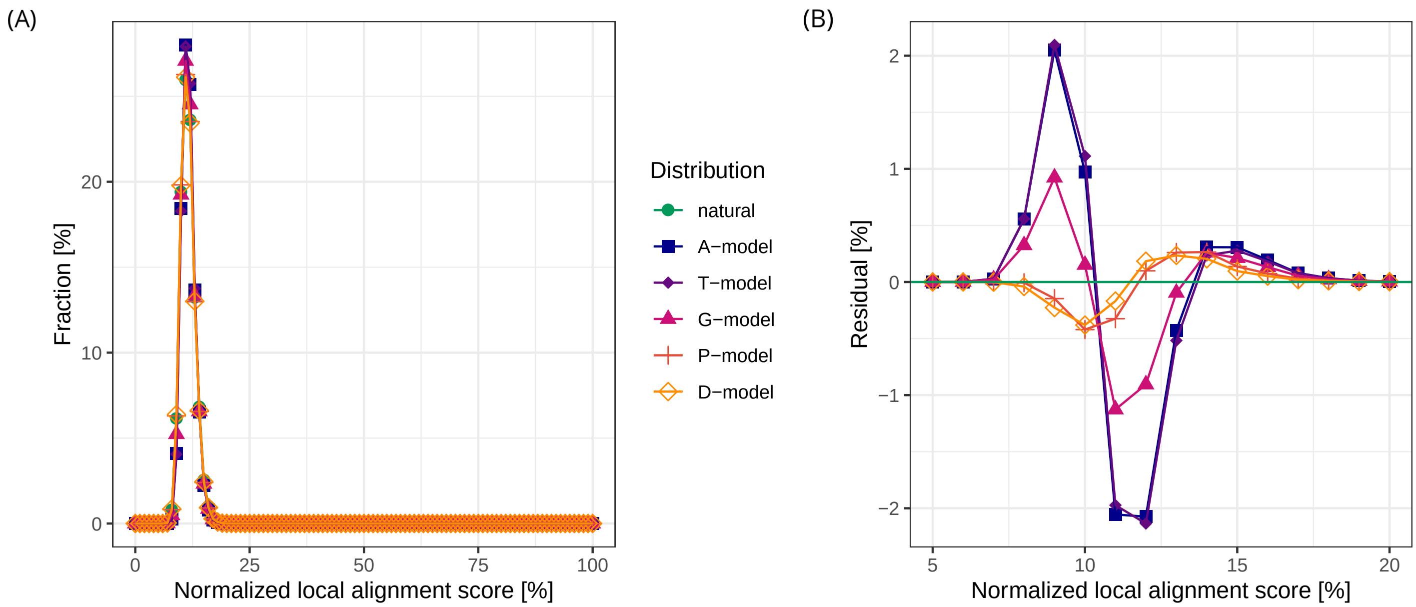

Figure 2. Comparing the sequence space occupation of random protein sequence models and natural sequence data. (A) Distance distributions are a descriptor of sequence space occupation. The distance between sequence fragments of the same length, defined as the sequence identity score obtained from a Smith-Waterman alignment, are plotted against the fraction of fragment pairs with the respective distance. We sampled 500 Million distances between fragments of length 100 for each model as well as for the natural sequence data. All distance distributions spike in the area of long-range distances with a mean sequence identity score around 11%. Both natural and random distance distributions are almost entirely overlapping. (B) Residuals represent the difference in sequence space occupation of random models compared to the natural sequences. We extract the distance-specific difference by subtracting the random from the natural distance distribution. The resulting residuals for each model indicate distances between natural fragments that are unaccounted for by the respective model (crests above zero). The A-, T- and G-model display a 2-peak behavior, associated with more long-range and short-range distances between natural fragments than modeled, reflecting an increased amount of both diversity and clustering in natural sequence space. The residuals of the P- and D-model possess only one peak for more short-range distances between natural sequences, hence an unexpected amount of clustering.

We note, however, that this discrepancy between natural sequences and the A-model is not very pronounced, as the total residual has a value of only 4.6% for 100mers (***Figure 3***: A). It is even less pronounced at smaller fragment lengths, reaching 0.4% for 10mers. We conclude that the A-model becomes less accurate in describing the sequence space occupation of natural sequences at lengths that are biologically relevant, but that it already achieves considerably higher accuracy than the completely random model (E-model), which has a total residual of 30.4% for 100mers (data shown in Methods).

We evaluated whether adding sequence information to the unified compositional bias of the A-model could further improve it. Since nature favors certain amino acid combinations as neighboring residues, a model that reflects the natural dipeptide frequency (T-model), has been proposed to represent natural sequences better than the A-model ***Lavelle and Pearson*** (***2009***). We implemented the T-model by extracting the dipeptide frequencies from our natural dataset and using them to generate random sequences with a Markov Chain Model. For all fragment lengths, we derived the distance distribution of the T-model (***Figure 2***: A), its residuals (***Figure 2***: B) and the total residual (***Figure 3***: A). By all these measures the T- and the A-model yielded essentially identical results in modeling the natural distance distribution. This outcome was somewhat surprising, as the addition of dipeptide frequencies to the A-model did produce a measurable improvement in the enumeration study of 5mers ***Lavelle and Pearson*** (***2009***). This may be due to the different methodology in that study, which collated exact 5mer frequencies, corresponding to a position-wise Hamming distance of zero, and thus being close to a global, not to a local alignment as used in our study. In fact, when using the Hamming distance as metric, the T-model achieves a slightly better accuracy over the A-model for sequences of 50 or less residues (***Figure 3***: D). From the results we obtained with the A- and T-models, we conclude that global measures of composition and sequence bias already approximate natural sequences fairly accurately, but that this accuracy decreases with sequence length. Especially for longer fragments, we expect further improvement by including local compositional biases as outlined in the previous section.

Context-specific composition

In order to capture context-dependent features, we investigated the effects of naturally occurring local amino acid compositions. As a first step we considered a model that accounts for genome diversity (G-model). Therein, the random dataset is produced by shuffling residues of the natural dataset within the boundaries of each genome. Given that our natural dataset holds 1,307 genomes, the derived sequences are thus sampled from 1,307 distinct compositions. Further locality is achieved by accounting for the composition of individual proteins (P-model). Here, the random dataset is produced by shuffling residues within each natural protein, corresponding to $4.7 \cdot 10^6$ compositions.

Since proteins are generally composed of domains, which are usually autonomous in structure and also often in function, the next level of locality would be achieved by accounting for the compositional biases of individual domains. Producing such a D-model is however not straightforward, as determining domain boundaries for proteins of unknown structure is fraught with errors and many residues in our dataset cannot be assigned to a domain family. As a proxy for domains we therefore derived all possible fragments of length 100 from our natural dataset and generated the D-model by shuffling residues within each fragment (see Methods). Correspondingly, we considered all natural sequences, whether or not they are part of a structured domain and thus included linker sequences and intrinsically unstructured regions. The extent to which this model is an accurate approximation of natural domains will be discussed below.

Comparing the G-model to the A- and T-models over the bacterial dataset shows a dampened wave for the residuals, with the same shape, but a decreased amplitude (***Figure 2***: B). The total residual is correspondingly smaller by a factor of about 2 for all fragment lengths (***Figure 3***: A), implying that

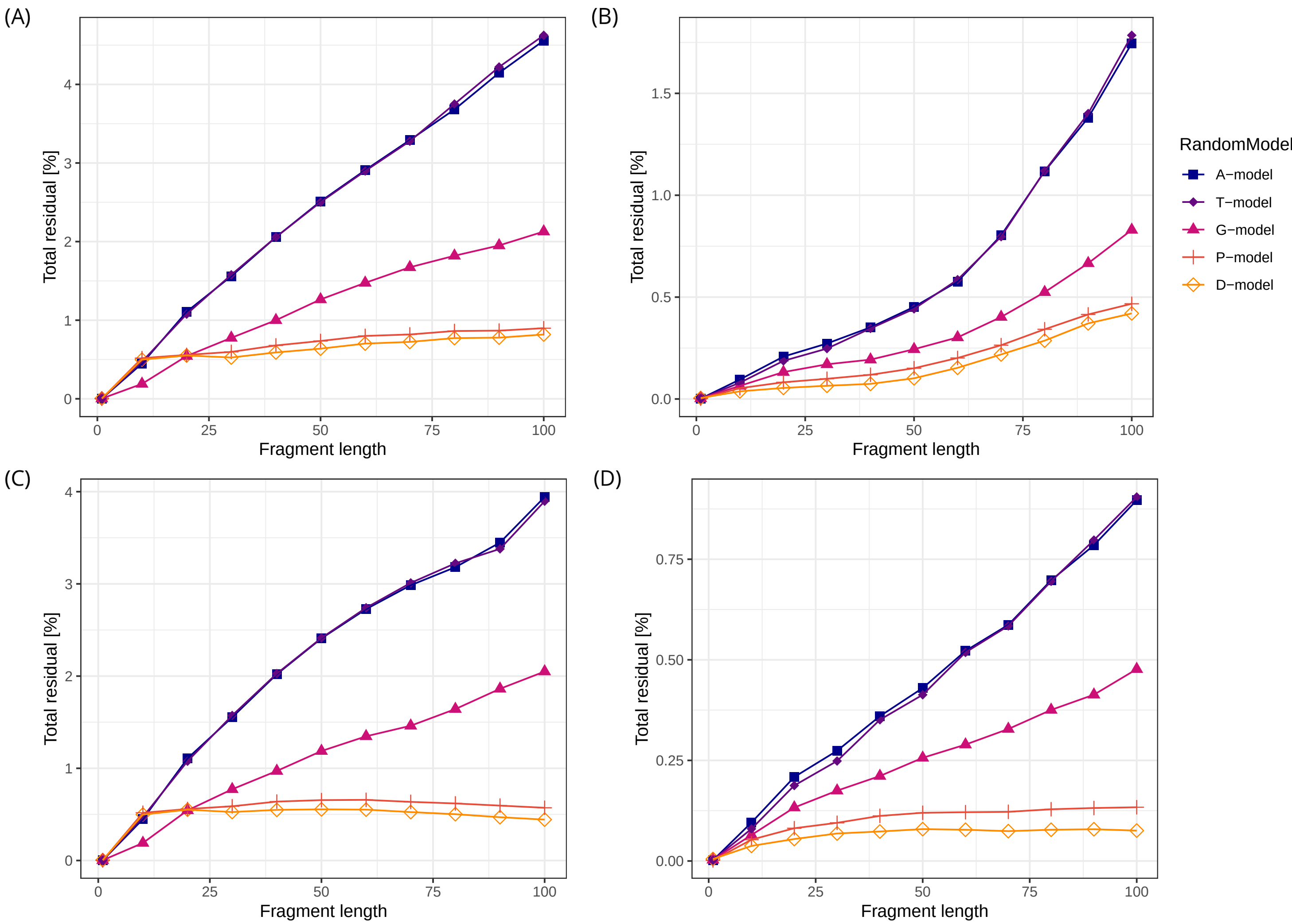

Figure 3. Deviation of random sequence models from natural sequences as a function of fragment length. (A) Total residuals when using a local Smith-Waterman alignment The total residual indicates the extent to which the distance distribution of random sequence models deviates from the natural. It reflects the fraction of distances between natural fragments that are unaccounted for in the random model. With increasing fragment length, the total residual of all models increases, implying that for longer fragments all models become worse in approximating similarities between natural fragments. The A-model (overall amino acid composition) and the T-model (overall dipeptide frequency) deviate furthest followed by the G-model (residue composition of genomes), the P-model (residue composition of proteins) and the D-model (residue composition of domain-sized fragments of length 100), which deviates the least. The intercept of the total residuals of the T- and D-model with the other models at fragment length around 10 is associated with edge effects of natural sequences and the usage of a local alignment as distance metric. (B) Total residuals when using a global Needleman-Wunsch alignment. The inconsistent continuation of the total residuals at sequence length 10 when using a local alignment has disappeared. Generally, the total residuals are reduced by 2.5-fold compared to the local alignment, reflecting that a global alignment captures less effects of natural sequences than a local alignment. (C) Total residuals when using a local Shift alignment. A Shift alignment does not penalize beginning and end gaps and prohibits internal gaps. Similar to the Smith-Waterman alignment, the Shift alignment displays an inconsistency at fragment length around 10. (D) Total residuals when using a Hamming distance without alignment. It reflects the most stringent interpretation of similarity in sequence space, as the n-th position of one sequence is always compared with the n-th position of another sequence. It corresponds to a metric that considers the number of dimensions (positions in sequence) that are identical.

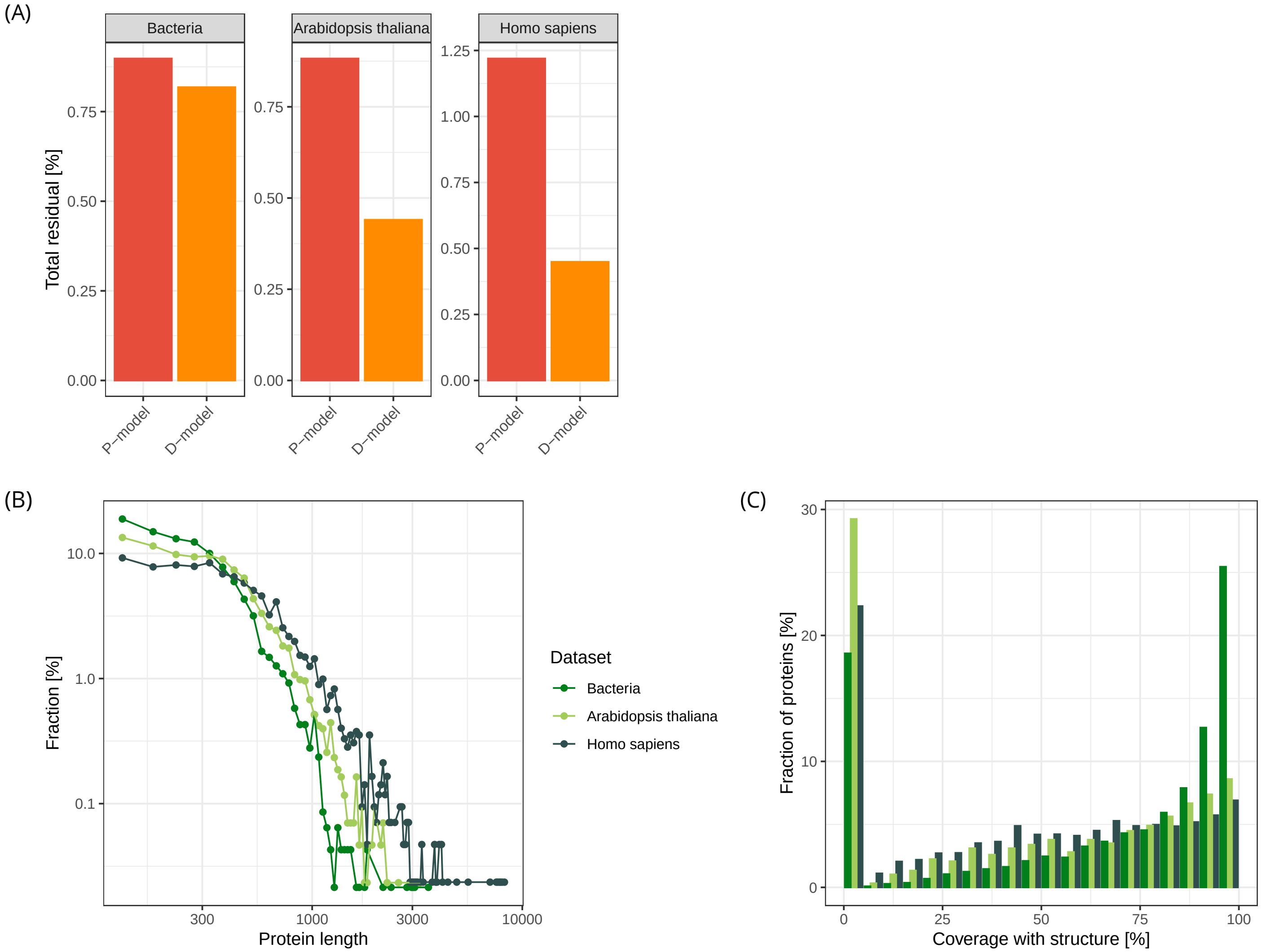

Figure 4. Contrasting the results of our bacterial dataset with those from two eukaryotic proteomes. (A) Total residuals of random models for bacterial dataset, the proteome of *Arabidopsis thalia* and *Homo sapiens* of the P- and D-models. Relative to the total residual of the P-model, the total residuals of the D-models differ in the three presented datasets. In bacteria, both are almost identical, whereas for the eukaryotic datasets the D-models have a more than 2-fold increase in accuracy over the P-models. (B) Distribution of protein length. The median protein length is smallest for bacteria with 315 residues, 400 residues in the *Arabidopsis thaliana* dataset and 550 residues in the *Homo sapiens* dataset. The increase of median protein length correlates with the decrease in the total residual of the D-model relative to the P-model. (C) Coverage of proteins by structured domains. For each protein in the three datasets, an estimate of the coverage by structured domains was obtained by assigning ECOD families to regions in the protein. The fraction of residues within assigned domains compared to the protein length was obtained and plotted as a histogram over all sampled proteins. In bacteria 40% of the sampled proteins are almost completely structured (coverage of >90%), a fraction that is greater compared to that in *Arabidopsis thaliana* (15%) and *Homo sapiens* (13%).

controlling for genome composition provides a substantial improvement in modeling the natural distance distribution. A further improvement is clearly achieved with the P-model, even though, at sequence lengths below 20 residues, it produces minor inconsistencies in its total residuals relative to the A-, T-, and G-models (***Figure 3***: A). We suspect that this is an artifact of using local alignments (***Figure 3***: A, C) and, indeed, the effect disappears when using a global alignment as distance metric over the same dataset (***Figure 3***: B, D). As for the A-, T-, and G-models, the residuals of the P-model also have a wave shape, which is however qualitatively different from the shapes for the less local random models, as it has only one crest at an alignment score of 13%. The crest for the unexplained long-range distances is gone, which we attribute to the fact that accounting for composition at the level of individual proteins has introduced the heterogeneity of natural sequences into the random model. For 100mers the total residual of the P-model is 0.9% (***Figure 3***: A), a value that is not improved remarkably by an even greater locality: The residuals of the D-model have the same wave shape as those of the P-model and a comparable amplitude, providing only a minor improvement with a total residual of 0.8%. This was somewhat surprising, as it is well established that many proteins are composed of disparate parts such as domains of distinct fold classes, intrinsically unstructured regions or fibrous parts, which are known to be characterized by different residue compositions ***Dosztányi et al.*** (***2005***). The composition of proteins that are composed of heterogeneous parts should thus be scrambled in the P-model and preserved in the D-model. We therefore expected that the D-model would provide a clearer improvement over the P-model.

Similar results of D- and P-models are associated to the dataset

We see two reasons why the total residuals of the D- and the P-models are almost identical. One is a technical reason, namely that there is no room for fluctuation of local residue composition in our bacterial dataset, as it may comprise a large number of short and single-domain proteins. The other is a potential qualitative characteristic of our dataset, namely that in long bacterial proteins the local residue composition does not fluctuate remarkably. In order to distinguish how these two reasons contribute to the comparable total residuals of the D- and P-models, we added two eukaryotic datasets for comparison to the following analysis. We retrieved the highly curated proteomes of *Homo Sapiens* and *Arabidopsis thaliana* from UniRef ***Apweiler*** (***2009***) and pruned them according to the procedure used for our bacterial dataset. Comparisons of total residuals between the bacterial and eukaryotic datasets show that, whereas the P- and D-models for the bacterial dataset are essentially equivalent, the D-models for the eukaryotic datasets are roughly 2-fold smaller than those of the P-models (***Figure 4***: A), and thus closer to our expectation.

In order to evaluate the technical reason, we analyzed sequence lengths in all three datasets and estimated the number of single- and multi-domain proteins. The bacterial dataset has the shortest proteins with a median length of 315 residues, the *Arabidopsis thaliana* dataset a median length of 400 residues and the *Homo sapiens* dataset the longest proteins with a median length of 550 residues (***Figure 4***: B). To estimate the number of single and multi-domain proteins, we randomly sampled each of the three datasets and used HHpred ***Remmert et al.*** (***2012***) for their domain annotation against the ECOD database ***Cheng et al.*** (***2014***), which represents the most recent and comprehensive classification of domains of known structure (see Methods). ECOD is the current "gold standard" in domain assignments and, at more than 13,000 families, provides a structural basis for most known domains (as captured in databases such as Pfam ***Punta et al.*** (***2012***), SMART ***Schultz et al.*** (***1998***) or COGs ***Tatusov et al.*** (***2000***)). We considered proteins multi-domain if they had at least 2 domains assigned to them, otherwise we considered them as single-domain proteins. The predicted fraction of multi-domain proteins in our bacterial dataset is 30%, which is smaller in *Arabidopsis thaliana* (25%) and greater in *Homo Sapiens* (35%). The overall length distribution thus indeed correlates with the ratio between the total residuals of the P- and D-models, and potentially contributes to the observed effect, whereas the number of domains per protein does not.

In order to evaluate the qualitative reason, namely that sequences of distinct composition are combined within proteins, we assessed the fraction of structured and unstructured regions in the used proteins. To that end, we estimated the fraction of structured regions for each protein with HHpred against the ECOD database (***Figure 4***: C). For the bacterial dataset, 40% of all sampled proteins are predicted to be structured over >90% of their sequence, a fraction that is smaller in *Arabidopsis thaliana* (15%) and *Homo Sapiens* (13%). The structure content of proteins thus also correlates with the ratio between the total residuals of the P- and D-models (***Figure 4***: A), possibly because scrambling between structured and unstructured regions leads to greater compositional disturbance than scrambling within these regions.

We conclude that the D-model approximates the natural distance distribution better than the P-model in all cases, however in a more pronounced way for datasets containing heterogeneous mixtures of long sequences combining structured with unstructured regions. In our analysis, these effects were more pronounced in eukayotic than in bacterial proteins.

Sequence bias caused by homology

Having accounted for compositional effects at increasingly local level, the remaining discrepancy between the distance distributions of the D-model and the natural dataset must be related to the actual sequence of amino acids. This discrepancy can arise either through divergence from a common ancestor (homology) or convergence as a result of structural constraints, particularly secondary structure formation (analogy). In order to evaluate the relative contribution of these mechanisms to the natural distances between sequence fragments we aimed to identify what proportion of distances could be assigned confidently to either homologous or analogous relationships and evaluated their contribution to the natural distance distribution.

The detection of homologous relationships requires advanced approaches, which are computationally much more expensive than the simple sequence alignments used to determine distances in sequence space. We therefore only considered a small subset of our sequences and their relationships within this subset, which could be derived computationally in a reasonable amount of time. For this, we randomly sampled our natural dataset to form 10 unbiased groups of 100mers, containing approximately 650 sequences each. We used HHblits to generate profile Hidden Markov Models (HMMs) for all individual sequences within these groups, then derived a set of relationships by aligning the retrieved HMMs from one set of sequences to those of another. This we repeated for arbitrary sets of 100mers, resulting in multiple unbiased samples of relationships. The likelihood of homology between two HHMs was derived using the tool HHalign and required a strict threshold of minimally 90% probability (see Methods). This process identified 0.11% of pairwise relationships as homologous (***Figure 5***: A, yellow), with a standard error of the mean (SEM) of 0.0033% (***Figure 5***: A).

For the remaining sequence pairs, we evaluated the likelihood of analogy by comparing their HMMs to those of the ECOD database ***Cheng et al.*** (***2014***). By virtue of containing only domains of known structure, ECOD is the currently best resource for distinguishing between homology and analogy in protein domains. For our analysis, we scored pairs of sequences as analogous if they matched distinct X-groups in the ECOD hierarchy using the same probability cutoff of 90% as for the homology assignment. In most cases, the X-level is the highest level at which homology still needs to be considered as a possibility; requiring fragments to match different X-groups within this level thus provided a conservative estimate of analogous relationships. This process identified 52.22% of pairwise relationships as analogous (***Figure 5***: A, purple), with a SEM of 0.84%. We conclude from this that the number of confident analogous pairs exceeds the number of confident homologous pairs by more than 2 orders of magnitude. This already indicates that the influence of homology on the global distance distribution in natural sequences will be dwarfed by analogy. All sequence pairs that could not be confidently assigned to either group were considered to be of unknown

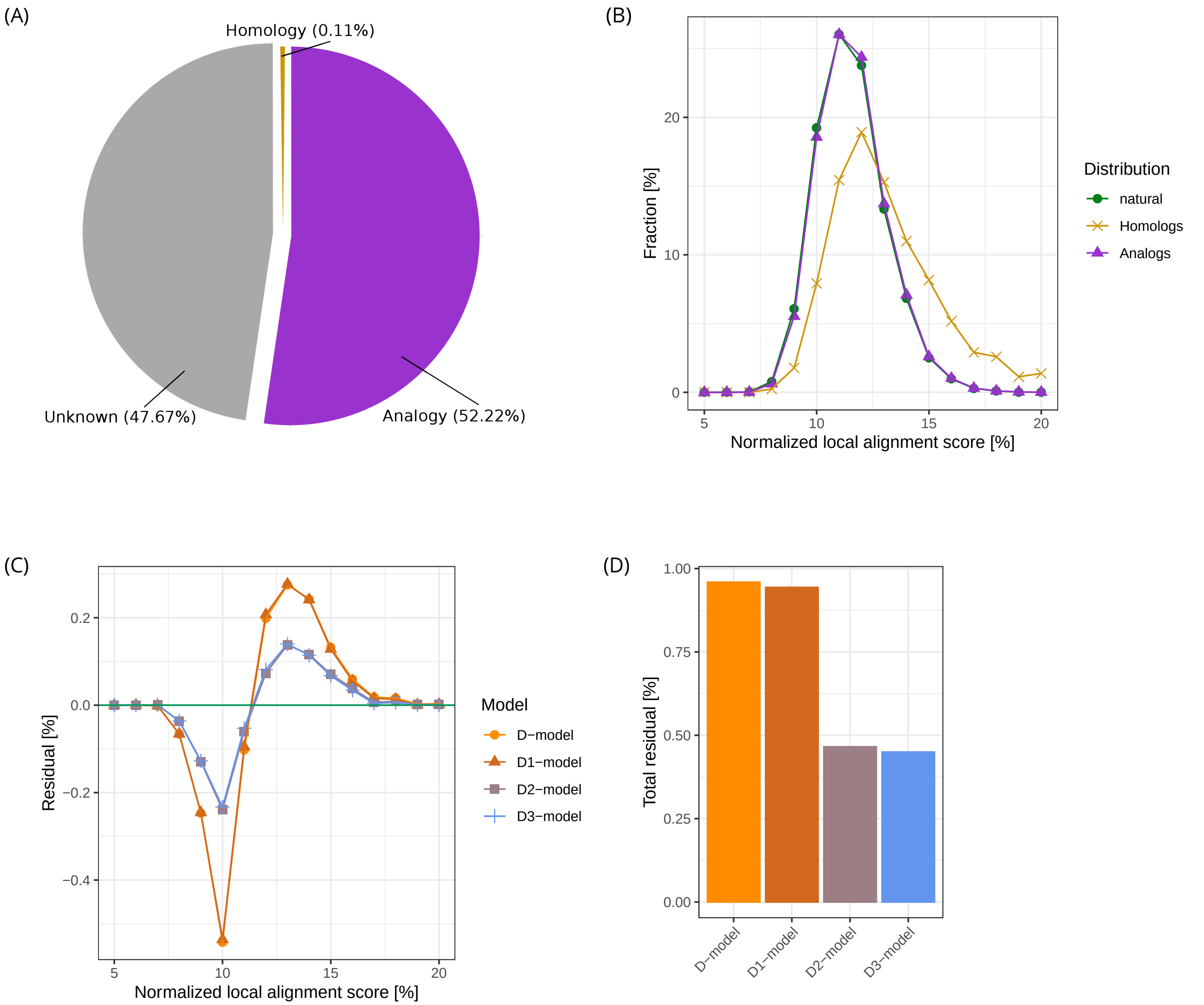

Figure 5. The contribution of homology and analogy to the global occupation of sequence space. (A) Decomposition of fragment pairs into their origins. We sampled 2 Million fragments pairs and analyzed if their relationship is confidently homologous or analogous. The fraction of analogous relationships was determined to be 52.22%, homologous relationships only 0.11% and the remaining fraction is labeled of unknown origin. Thus, the majority of relationships is generally analogous. (B) Distance distribution between homologs and analogs contrasted with the natural distance distribution. The qualitative difference between the distance distribution of analogs and that of all fragments is relatively small. Compared to this, the distance distribution of homologs displays a strong tendency towards a higher sequence identity score; it nevertheless has a major overlap with the natural distribution. (C) Residuals of the models incorporating the sequence bias of homology and analogy. We generated mixed models, that include the sequence bias of homology (D1-model), analogy (D2-model) and both (D3-model) into the D-model, which is only based on the composition of natural 100mers. The D1-model, which includes homologous sequence bias, displays almost the same residuals as the purely composition-based D-model. The residuals of the D2-model, which includes analogous sequence bias, deviate severely from that of the D-model. The D3-model yields similar results as the D2-model. (D) Total residuals of mixed models. The total residuals behave accordingly to the residuals. The D1-model has displays an only improvement in the total residual of 0.016% compared to that of the D-model. The D2-model reaches a total residual of 0.46% and is more than 2-fold more accurate than the D-model (0.96%). Adding the homology bias to the D2-model to obtain the D3-model has almost no effect.

relationship, amounting to 47.6% of the total with a SEM of 0.84% (***Figure 5***: A, grey).

Having decomposed sequence pairs into confident homologous and analogous relationships, we analyzed to what extent the remaining total residual (0.8%) can be explained by incorporating corresponding sequence biases into our D-model. Therefore, we generated three new hybrid models in the following way: we omitted either homologous pairs, or analogous pairs, or both from our set of assigned relationships, generated a D-model for the remaining fragment pairs through the same shuffling procedure as used previously, and then added back the omitted pairs without shuffling. In the following we refer to the hybrid model that adds the sequence bias of homologs to the domain composition as the D1-model, the one that adds the sequence bias of analogs as the D2-model, and the one that adds both biases as the D3-model.

The residuals of these three models are compared to that of the D-model in (***Figure 5***: C). Due to the reduced sampling over only 2 million fragment pairs, instead of 500 million, the total residual of the D-model in this analysis deviates slightly from that obtained over the entire dataset and has a value of 0.96% instead of 0.8% (***Figure 5***: D).

Relative to this total residual of the purely compositional D-model, the D1-model, which includes homologous sequence effects, is only minimally better (total residual reduced by 0.016%) at approximating the natural distance distribution (***Figure 5***: D). We assume that two reasons are mainly responsible for this only minor improvement: First, the proportion of homologous relationships is only 0.11%, giving them little leverage. Second, the distance distribution of homologs (***Figure 5***: B, yellow) differs only to a small extent from the distance distribution of the natural dataset. This is not entirely unexpected, given how difficult it is to distinguish distant homology from random fluctuation in sequence comparisons. In fact, it has been recognized previously that most homologous sequences share no significant similarity ***Rost*** (***1997***).

In contrast, the total residual of the D2-model (0.46%), which includes analogous sequence effects, is decreased about 2-fold relative to the D-model. Thus, although analogs have a distance distribution that is very similar to the natural (***Figure 5***: B, purple and green), their leverage is 2 orders of magnitude higher than that of homologs, causing these small differences to improve substantially the fit of the D2-model to the natural distance distribution. This is again not entirely unexpected, as most sequences in our natural dataset share the ability to form secondary structures (***Figure 4***: C), resulting in a sequence bias that is not fully captured by residue composition ***Pande et al.*** (***1994***) ***Lavelle and Pearson*** (***2009***). As expected from the D1-model, adding the homologous sequence bias to the D2-model did not really improve its ability to approximate the natural distance distribution. We conclude that the sequence space of natural proteins is almost entirely shaped by compositional effects and that the remaining sequence bias is almost entirely due to analogy, which we interpret to result from secondary structure formation.

Conclusion

In this article we have undertaken a study of natural protein sequence space, using an approach built on the probability mass function of pairwise distances between sequence fragments. With this approach we were able to analyze the occupation of sequence space by fragments up to 100 residues in length, substantially exceeding previous efforts and for the first time characterizing globally the relative position of sequences in space. Our results show that the global compositional bias of natural proteins is already sufficient to approximate the distance distribution of natural sequences by 95.4% and that accounting for local compositional bias down to the level of individual 100mers further improves this to 99.2%. The remaining 0.8% of unaccounted distances between natural 100mers are almost entirely contributed by sequence effects arising from analogous relationships, leaving only a negligible contribution to homology in the global characterization of sequence space occupation.

This surprised us, as decades of bioinformatic work have mapped out an increasingly comprehensive description of sequence space around protein families, based on the detection of ever more remote homology. We therefore expected to find that homology also has a substantial role in shaping the global structure of sequence space occupation, corresponding to the image of islands formed by natural sequences within a global sea of possibilities. This expectation was not borne out and in retrospect this might not seem as surprising, given that even within protein families, the influence of homology is smaller than generally perceived. This is substantially due to the way in which family relationships are represented, strongly emphasizing common features (such as the generally few conserved residues) and omitting variable ones. This focus on biological significance over raw sequence similarity leads to a perception of sequence space that is distorted by *evolutionary distance* and does not reflect actual distances. Evidence for this can be seen for example in the progressively more complex statistical methods needed to substantiate homology across increasingly large evolutionary distances, the resulting difficulties to classify the detected relationships into a hierarchy of protein families and superfamilies, and the remaining inability in many cases to judge on the homologous or analogous nature of similarities, even in the presence of extensive sequence and structure information ***Rost*** (***1997***). These considerations show that even at the level of protein families, many sequence relationships comprise a large random element, substantially indistinguishable from random fluctuation and sequence convergence. This random element not only results from our inability to detect homologs that have diverged strongly due to low selective pressure, but also from the fact that in many families, a conserved core has been elaborated in different ways with analogous sequences.

We find a much larger influence of analogous sequence biases on the global shape of naturally occupied sequence space. The main common feature of proteins in our natural dataset is the ability to fold, which translates into a propensity to assume secondary structure locally. We see this as the main reason for the sequence bias that we observe between analogous sequences. Nevertheless, the sequence biases of homology and analogy together account for only 0.8% of all distances between natural sequences. We conclude that natural sequences stand out from randomness primarily through their biased use of the 20 amino acids. Accounting for this bias at increasingly local levels is largely sufficient to model the global structure of sequence space occupation. This major relevance of composition has been acknowledged as it has been implemented into BLAST ***Schaffer*** (***2002***) and been demonstrated to be key for the aggregation of intrinsically unstructured proteins ***Vymětal et al.*** (***2019***).

There seems to be no other striking feature of the primary structure in natural protein sequences and in consequence there are also no other obvious features that distinguish natural from random sequences. We conclude that viable proteins could be located anywhere in the sequence space defined by natural residue compositions. The main reason why the proteome of nature currently only comprises some 10^{12} proteins ***Lupas and Koretke*** (***2008***) and that these mainly fall into only about 10^4 families ***Punta et al.*** (***2012***) is therefore not due to the limited availability of useful sequence space, but rather to their evolutionary history. There is treasure everywhere.

Methods

Natural data

Genome selection

With the aim to achieve a reasonable coverage of deep phylogenetic branches with complete and well-annotated proteomes, we selected the majority of bacterial genomes provided by UniRef on 22.09.2017 ***Apweiler*** (***2009***). Some genomes stood out as they possessed multiple replicas of the same protein and were excluded, leaving 4,098 to remain. For each of the 1,307 genera we randomly chose one representative for our natural data set. The genus was derived from the full-length genome name via string matching.

We are aware of the general ambiguity of the definition of a genus ***Parks et al.*** (***2018***). However, with the genus selection we only aimed to reduce redundancy caused by some species that have been sequenced many times. Lastly, we note that the bias towards bacteria that are easy to cultivate prohibits a sampling of the true diversity among bacterial genomes.

Genome curation

Apart from redundancy at the genome level, we control for recent gene duplication events. For each genome, we cluster its proteins using cd-hit (version 4.6 with 99% sequence identity and 90% coverage). A representative protein sequence, as defined by cd-hit, was then selected for each cluster; all other proteins were discarded.

Low complexity filtering

Low-complexity regions (LCRs) are a well-known features of natural sequences, that do not occur as frequently in random sequences. We first analyzed our data including LCRs and found that they majorly contribute to the total residual between natural sequences and our models (data not shown). Therefore, we pruned LCRs of our dataset using segmasker ***Wootton and Federhen*** (***1996***) (version 2.3.0+ with the standard settings), to obtain differences between natural and random sequences that are not due to this well-known feature. This pruning of LCRs leads to sequences of slightly higher complexity than expected for short peptides (data not shown). The pruning bias plays an insignificant role, especially for longer sequences, which are of most interest in our study. Since, N-terminal methionines were sometimes included, we stripped them to standardize our sequences.

Sequence adjustments

To simplify our analysis we changed a couple of hundred cases of uncommon amino acids to their most similar proteinogenic amino acid. In order to use the exact same dataset for all sequence lengths, we pruned our data set of sequences shorter than 100. Additionally, we removed the invalid amino acid X by replacing it with an end-of-line-character, effectively dividing a protein sequence into multiple parts. However, since some of our random models depend on shuffling intact genomes or proteins, we performed this division into multiple parts after the shuffling (more detail below).

Complete statistics and data availability

Taken together our dataset holds $1.2 \cdot 10^9$ valid amino acids of 1,307 genomes comprising $4.7 \cdot 10^6$ proteins. In the supplements we provide:

- fasta-file of original genomes
- fasta-file of adjusted genomes
- overall amino acid composition

Fragment pair selection and random sequence models

Fragment selection

We selected random fragments such that each character (amino acids and end-of-line-character) in the dataset had the same probability of being chosen and that the same fragment pair would never be chosen twice. We ensured this by implementing two linear congruential generator ***Press et al.*** (***2007***) to enumerate all possible pairs of fragments. In detail, one linear congruential generator was used for each member of the pair with multiplier $a = 1$ and moduli $m_1 = 223$ and $m_2 = 34,211$, where both moduli are prime numbers relative to the total number of characters $1,168,754,000$. Depending on the starting points of the two generators, a different subset of index pairs can be selected. This enabled us to calculate disjunctive fragment pairs in parallel. We selected $5 \cdot 10^8$ valid pairs of fragments to accurately estimate the distance distributions and rejected fragments that straddled protein boundaries or invalid regions, indicated by the end-of-line-character.

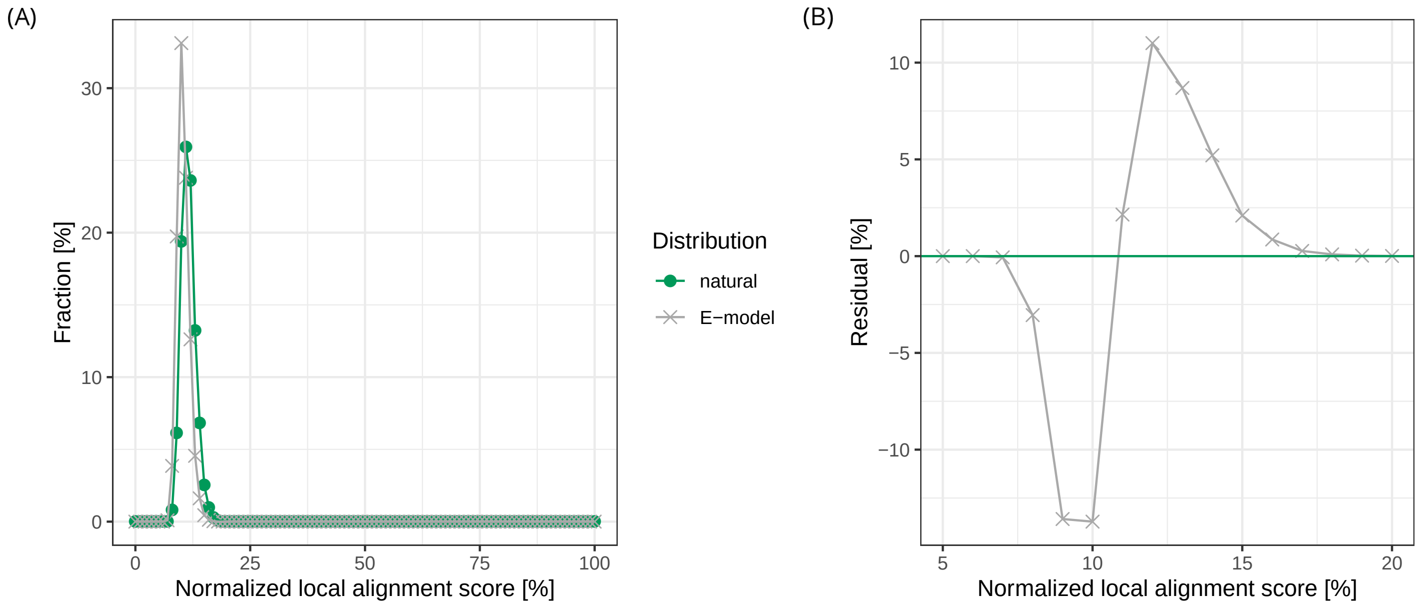

Figure 6. The E-model deviates severely from the natural dataset. (A) Distance distributions of natural dataset and E-model. In contrast to the primarily used models (A-, T-, G-, P-, and D-models) in this paper, the distance distribution of the E-model has an obvious deviation from that of the natural data. (B) Residuals of the E-model. Compared to the distances between fragments derived from the E-model, the distances between natural sequences have a strong tendency towards being shorter.

Models incorporating overall amino acid composition

The most standard random sequence model is based on the underlying amino acid composition of a given dataset. We obtained randomized data for this A-model by randomly shuffling all amino acids of the natural data. Thereby, protein length is maintained and the number of amino acids stays exactly the same. As all our random models are based on random permutations, we used the Mersenne Twister algorithm mt19337 of the C++ 14 std library with the standard seed value of 19650218. This algorithm is considered one of the best pseudo-random number generators and in a test with a smaller dataset we found that our results did not depend on the type or seeding of the random number generator.

For the E-model, we proceeded the same way as for the A-model. The only difference is that we replaced the natural dataset, by writing over all valid amino acids with the 20 possible amino acids in lexicographical order. When reaching the character Y for tryptophan, we started over with A for alanine. The distance distribution of the E-model deviates severely from that of the natural dataset (***Figure 6***) with a total residual of 30.4% for 100mers.

Models based on the amino acid composition of genomes or proteins

To account for genome or protein composition, we shuffled amino acids within the context of genomes or proteins. For the G-model, we shuffled valid amino acids within each of the 1,307 genomes. For the P-model we shuffled valid amino acids within each protein. We used one instance for genome and protein composition bias and stored them to generate the distance distribution for the corresponding models. After shuffling, we divided proteins containing the invalid amino acid X by replacing it with an end-of-line-character.

Model based on the amino acid composition of domain-sized fragments

For the D-model, we randomly shuffled natural fragments of length 100. In contrast to the previous random models, generating a single randomly shuffled dataset is not computationally feasible since storing an instance of all shuffled 100mers would increase the data size approximately 100-fold. We therefore shuffled 100mers on the fly during the calculation of the distance distributions. In detail, we select pairs of natural fragments as described above and consider the target fragment of length N to be located in the middle of the domain. If the domain straddles any protein boundaries, we adjust the domain boundaries such that the domain fits into the protein boundaries by shifting to the right (starting sequences) or to the left (terminal sequences). Note that because of this adjustment, the selection probability of amino acids into domains is not uniform but the selection probability of amino acids in fragments is. The alternative would be a rejection procedure, where we would reject fragments that are so close to protein boundaries that the domain of length of 100 would not fit. The downside of such a rejection procedure is that fragments close to protein boundaries are not selected and hence the selection probability of fragments is not uniform anymore, which differs from the selection of natural fragments or fragments for the A-, G-, and P-models. The D1, D2, and D3-models, which incorporate the sequence bias of homologous and analogous fragments, are presented further down.

Pairwise distances as descriptor for sequence space occupation

Distance metric

We define the distance between two fragments of the same length N as the normalized rounded score s from a Smith-Waterman alignment. In the alignment, an amino acid match is scored with 1, a mismatch with 0, gap opening penalty is equal to 3 and gap extension penalty is 0.1, which are the same parameters for gaps as used in ***Rost*** (***1999***) ***Schneider et al.*** (***1997***). Due to gaps, the alignment scores p can rank between 0 and N in 0.1 steps; to obtain integer distances, we round scores to the closest integer number. Distances exactly between two integers (such as 1.5) are assigned to the smaller one. To compare the score p across different fragment lengths N, we transform it into the normalized score s, scaling between 0-100%, as follows:

$$s = \frac{round(p)}{N}$$

This score s thereby reflects the number of dimensions in sequence space (positions in sequence), which differ between two fragments, while allowing for gaps and insertions. In some cases, we use the normalized global alignment score from a Needleman-Wunsch alignment with identical scoring parameters, the Hamming distance or a Shift metric that allows for terminating and starting gap without penalties, to illustrate differences to the used Smith-Waterman metric.

We also diversified gap penalties, leading to comparable results (data not shown). For all alignments, we used the SeqAn C++ library, version 2.4 ***Rahn et al.*** (***2018***), which enables many sequence comparisons in parallel.

Comparing distance distributions

The residual corresponds to the variational distance at each possible sequence identity between two distance distributions. We use it to demonstrate the qualitative difference between the distance distribution of a random model and that natural sequences. Denoting the residual by r, the random model distance distribution by D_{rand} and the natural one by D_{nat} we have:

$$r(s) = D_{nat}(s) - D_{rand}(s)$$

where s is the alignment score. For residuals $r(s)$ exceeding zero, there is a higher frequency of these alignment scores in natural fragments relative to random fragments.

To summarize the difference between natural and random model distance distributions in a single metric, we sum the absolute residuals over all sequence identities and normalize it to a range between 0 and 100%:

$$R = \sum_{0 \le s \le N} \frac{|r(s)|}{2}$$

We call R the total residual, which is variously called the variational distance, total variation distance or Kolmogorov distance ***Deza and Deza*** (***2014***).

Decomposition into homologous and analogous relationships

Homology

We derived the fraction of sequence pairs that are confidently homologous using the tools of HH-suite (version number 3.0.3) ***Remmert et al.*** (***2012***). To derive this fraction, we systematically sampled our dataset and extracted 10 sets of natural 100mers that are equally distributed over our dataset, each containing approximately 650 fragments. With HHblits, we generated HMMs with the standard settings for each of these fragments with two iterations, using uniclust30 as underlying database (version August 2018).

Then, we pairwise aligned the generated HMMs with HHalign, in order to estimate whether two fragments are homologous. We did this by aligning all fragments in one set to all of those in another set, resulting in 90 possible directed combinations of which we chose 10 as representative sets of pairwise relationships. Each set of fragments was considered twice in this comparison, once as the set of query sequences and once as the set of target sequences in the alignment. This resulted in 2 Million pairwise fragment comparisons divided into 10 disjunctive sets. Pairs of fragments were considered to be homologous, if HHalign predicted them to be homologous with a probability above 90%. In total 0.11% of the fragment pairs were found to be homologous; the standard error of the mean (SEM) derived from the 10 sets of 0.0033%.

Analogy

We derived the fraction of sequence pairs that are confidently analogous using a similar procedure as used for the homology detection. We first assigned structured domains to each 100mer. We then assumed a pair of 100mers to be of analogous origin, if the two 100mers matched only distinct domains that are confidently not related to each other.

For the assignment of structured domains, we used the ECOD classification ***Cheng et al.*** (***2014***), which is the currently best resource for distinguishing between homology and analogy in protein domains. The HMMs of each 100mer (same as in the homology detection) were thereby compared against all ECOD entries (retrieved on 9.4.2019) with HHsearch. We used HHsearch with the standard parameter and assigned the best-scoring non-overlapping hits with a probability above 90% to the corresponding fragment. Of all 100mers 70% could be assigned to a single domain and less than 1% to multiple domains, of which we considered each. Other 100mers were not assigned to any domain, which we directly excluded to be analogous to any other sequence, since we are uncertain about their origin.

For the assignment of analogous relationships, we considered only pairs of 100mers that were assigned to at least one domain. If their domains matched only distinct X-groups in the ECOD hierarchy, the pair was assumed to have an anologous relationship. The X-group is the highest level at which homology still needs to be considered as a possibility. All pairs of fragments that were assigned to domains of only distinct X-levels were considered to be confidently analogous.

With this procedure 52.22% of the fragment pairs were found to be analogous; the standard error of the mean derived from the 10 sets is 0.84%. The remaining 47.6% of the fragment pairs is of unknown relationship.

Mixed models containing sequence bias of homology or analogy

In order to estimate the influence of homology and analogy to the natural distance distribution, we generated mixed models that that account for their sequence bias. The D1-model includes the homologous sequence bias by including the distances between all confidently homologous fragment pairs without shuffling. We applied the D-model to the remaining fragment pairs and shuffled the fragments of the corresponding pairs that are not homologous with the Unix command shuf followed by deriving their distance. All distances combined resulted into the distance distribution of the D1-model. The sequence bias between homologous fragments is therein preserved while for other fragment pairs only their composition is accounted for. We proceeded the same way for the D2-model by including the distances of unshuffled fragments that are confidently analogous, and distances of the remaining pairs after shuffling the residues within each fragment. For the D3-model we included both sequence bias of homologous and analogous natural fragments.

References

Alva V, Remmert M, Biegert A, Lupas AN, Söding J. A galaxy of folds. Protein Science. 2009 jan; 19(1). http://www.ncbi.nlm.nih.gov/pubmed/19937658http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2817847http://doi.wiley.com/10.1002/pro.297, doi: 10.1002/pro.297.

Alva V, Söding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. eLife. 2015; 4. doi: 10.7554/eLife.09410.001.

Apweiler R. The universal protein resource (UniProt) in 2010. Nucleic Acids Research. 2009 jan; 38:D190–5. http://www.ncbi.nlm.nih.gov/pubmed/18045787http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2238893, doi: 10.1093/nar/gkp846.

Bershtein S, Serohijos AW, Shakhnovich EI. Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. Curr Opin Struct Biol. 2017; 42:31–40. http://dx.doi.org/10.1016/j.sbi.2016.10.013, doi: 10.1016/j.sbi.2016.10.013.

Cedano J, Aloy P, Perez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. Journal of Molecular Biology. 1997 feb; 266(3):594–600. https://www.sciencedirect.com/science/article/pii/S0022283696908049, doi: 10.1006/jmbi.1996.0804.

Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV. ECOD: An Evolutionary Classification of Protein Domains. PLoS Computational Biology. 2014 dec; 10(12):e1003926. https://dx.plos.org/10.1371/journal.pcbi.1003926, doi: 10.1371/journal.pcbi.1003926.

Chevalier A, Silva DA, Rocklin GJ, Hicks DR, Vergara R, Murapa P, Bernard SM, Zhang L, Lam KH, Yao G, Bahl CD, Miyashita SI, Goreshnik I, Fuller JT, Koday MT, Jenkins CM, Colvin T, Carter L, Bohn A, Bryan CM, et al. Massively parallel de novo protein design for targeted therapeutics. Nature. 2017 oct; 550(7674):74–79. http://www.nature.com/articles/nature23912, doi: 10.1038/nature23912.

Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS Struct Funct Genet (Erratum ibid, 2001, Vol44, 60). 2001; 43(3):246–255. doi: 10.1017/CBO9781107415324.004.

Deza MM, Deza E. Encyclopedia of Distances. Springer Berlin Heidelberg; 2014. https://books.google.de/books?id=q_7FBAAAQBAJ.

Diggle PJ. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. CRC press; 2014.

Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. Journal of Molecular Biology. 2005; 347(4):827–839. doi: 10.1016/j.jmb.2005.01.071.

Duranton G, Overman HG. Testing for Localization Using Micro-Geographic Data. The Review of Economic Studies. 2005 oct; 72(4):1077–1106. https://academic.oup.com/restud/article-lookup/doi/10.1111/0034-6527.00362, doi: 10.1111/0034-6527.00362.

Fukuchi S, Nishikawa K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol. 2001 jun; 309(4):835–843. https://www.sciencedirect.com/science/article/pii/S0022283601947187?via{%}3Dihub, doi: 10.1006/jmbi.2001.4718.

Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K. Unique amino acid composition of proteins in halophilic bacteria. J Mol Biol. 2003; 327(2):347–357. doi: 10.1016/S0022-2836(03)00150-5.

Harms MJ, Thornton JW. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. Nature. 2014 aug; 512(7513):203–207. http://www.ncbi.nlm.nih.gov/pubmed/24930765http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4447330, doi: 10.1038/nature13410.

Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016; 537(7620):320–7. http://www.nature.com/doifinder/10.1038/nature19946{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27629638, doi: 10.1038/nature19946.

Larson SM, England JL, Desjarlais JR, Pande VS. Thoroughly sampling sequence space: large-scale protein design of structural ensembles. Protein Sci. 2002; 11(12):2804–13. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2373757{&}tool=pmcentrez{&}rendertype=abstract, doi: 10.1110/ps.0203902.

Lavelle DT, Pearson WR. Globally, unrelated protein sequences appear random. Bioinformatics. 2009; 26(3):310–318. doi: 10.1093/bioinformatics/btp660.

de Lucrezia D, Slanzi D, Poli I, Polticelli F, Minervini G. Do natural proteins differ from random sequences polypeptides? natural vs. random proteins classification using an evolutionary neural network. PLoS ONE. 2012 may; 7(5):e36634. https://dx.plos.org/10.1371/journal.pone.0036634, doi: 10.1371/journal.pone.0036634.

Luigi Luisi P. Contingency and determinism. Philos Trans R Soc A Math Phys Eng Sci. 2003; 361(1807):1141–1147. http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.2003.1189, doi: 10.1098/rsta.2003.1189.

Lupas A, Koretke K. In: Evolution of Protein Folds; 2008. p. 131–151. doi: 10.1142/9789812778789_0006.

Lupas AN, Dyke MV, Stock J. Predicting Coiled Coils from Protein Sequences. Science. 1991; 252:1162–1164.

Munteanu CR, González-Díaz H, Borges F, de Magalhães AL. Natural/random protein classification models based on star network topological indices. Journal of Theoretical Biology. 2008; 254(4):775–783. doi: 10.1016/j.jtbi.2008.07.018.

Pande VS, Grosberg aY, Tanaka T. Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? Proc Natl Acad Sci U S A. 1994; 91(26):12972–12975. doi: 10.1073/pnas.91.26.12972.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018; 36(10):996. doi: 10.1038/nbt.4229.

Poznański J, Topiński J, Muszewska A, Dębski KJ, Hoffman-Sommer M, Pawłowski K, Grynberg M. Global pentapeptide statistics are far away from expected distributions. Sci Rep. 2018 dec; 8(1):15178. http://www.nature.com/articles/s41598-018-33433-8, doi: 10.1038/s41598-018-33433-8.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes 3rd Edition: The Art of Scientific Computing. 3 ed. New York, NY, USA: Cambridge University Press; 2007.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. Nucleic Acids Research. 2012 jan; 40(D1):D290–D301. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr1065, doi: 10.1093/nar/gkr1065.

Rahn R, Budach S, Costanza P, Ehrhardt M, Hancox J, Reinert K. Generic accelerated sequence alignment in SeqAn using vectorization and multi-threading. Bioinformatics. 2018; p. 1–9. https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty380/4992147, doi: 10.1093/bioinformatics/bty380.

Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods. 2012 feb; 9(2):173–175. http://www.nature.com/articles/nmeth.1818, doi: 10.1038/nmeth.1818.

Rost B. Twilight zone of protein sequence alignments. Protein Engineering Design and Selection. 1999; 12(2):85–94. doi: 10.1093/protein/12.2.85.

Rost B. Protein structures sustain evolutionary drift. Folding and Design. 1997 jun; 2(3):S19–S24. https://www.sciencedirect.com/science/article/pii/S135902789700059X, doi: 10.1016/S1359-0278(97)00059-X.

Schaffer AA. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Research. 2002; 29(14):2994–3005. doi: 10.1093/nar/29.14.2994.

Schneider R, de Daruvar A, Sander C. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res. 1997 jan; 25(1):226–230. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/25.1.226, doi: 10.1093/nar/25.1.226.

Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: Identification of signaling domains. PNAS. 1998; 95(May):5857–5864. doi: 10.1073/pnas.95.11.5857.

Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. . 2015; 2015. http://arxiv.org/abs/1404.4005{%}0Ahttp://dx.doi.org/10.1073/pnas.1412933112, doi: 10.1073/pnas.1412933112.

Starr TN, Picton LK, Thornton JW. Alternative evolutionary histories in the sequence space of an ancient protein. Nat Publ Gr. 2017; 549. https://www.nature.com/articles/nature23902.pdf, doi: 10.1038/nature23902.

Starr TN, Thornton JW. Epistasis in protein evolution. . 2016; 25(7):1204–1218. http://www.ncbi.nlm.nih.gov/pubmed/26833806http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4918427, doi: 10.1002/pro.2897.

Stiffler MA, Poelwijk FJ, Brock K, Stein RR, Teyra J, Sidhu S, Marks DS, Gauthier NP, Sander C. Protein structure from experimental evolution. bioRxiv. 2019; https://www.biorxiv.org/content/early/2019/06/13/667790, doi: 10.1101/667790.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research. 2000 jan; 28(1):33–36. http://www.ncbi.nlm.nih.gov/pubmed/10592175http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC102395, doi: 10.1093/nar/28.1.33.

Urlinger S, Baron U, Thellmann M, Hasan MT, Bujard H, Hillen W. Exploring the sequence space for tetracycline-dependent transcriptional activators: Novel mutations yield expanded range and sensitivity. Proc Natl Acad Sci. 2000 jul; 97(14):7963–7968. doi: 10.1073/PNAS.130192197.

Vallee BL, Auld DS. Zinc Coordination, Function, and Structure of Zinc Enzymes and Other Proteins. Biochemistry. 1990; 29(24):5647–5659. doi: 10.1021/bi00476a001.

Vymětal J, Vondrášek J, Hlouchová K. Sequence Versus Composition: What Prescribes IDP Biophysical Properties? Entropy. 2019 jul; 21(7):654. https://www.mdpi.com/1099-4300/21/7/654, doi: 10.3390/e21070654.

Wheelan SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. Bioinformatics. 2000; 16(7):613–618. doi: 10.1093/bioinformatics/16.7.613.

Woolfson DN, Bartlett GJ, Burton AJ, Heal JW, Niitsu A, Thomson AR, Wood CW. De novo protein design: How do we expand into the universe of possible protein structures? . 2015; 33:16–26. http://dx.doi.org/10.1016/j.sbi.2015.05.0090959-440X/{#}, doi: 10.1016/j.sbi.2015.05.009.

Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods in Enzymology. 1996 jan; 266:554–571. https://www.sciencedirect.com/science/article/pii/S0076687996660352, doi: 10.1016/S0076-6879(96)66035-2.

Wüthrich K. NMR of Proteins and Nucleic Acids; 1986.