The intrinsic dimension of protein sequence evolution

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

The intrinsic dimension of protein sequence evolution

Elena Facco[1], **Andrea Pagnani**[2,3,4]*, **Elena Tea Russo**[1], **Alessandro Laio**[1,5]

1 SISSA, Trieste, Italy, **2** DISAT, Politecnico di Torino, Torino, Italy, **3** IIGM, Italian Institute for Genomic Medicine, Torino, Italy, **4** INFN, Sezione di Torino, Torino, Italy, **5** ICTP, International Centre for Theoretical Physics, Trieste, Italy

* andrea.pagnani@polito.it

Abstract

It is well known that, in order to preserve its structure and function, a protein cannot change its sequence at random, but only by mutations occurring preferentially at specific locations. We here investigate quantitatively the amount of variability that is allowed in protein sequence evolution, by computing the intrinsic dimension (ID) of the sequences belonging to a selection of protein families. The ID is a measure of the number of independent directions that evolution can take starting from a given sequence. We find that the ID is practically constant for sequences belonging to the same family, and moreover it is very similar in different families, with values ranging between 6 and 12. These values are significantly smaller than the raw number of amino acids, confirming the importance of correlations between mutations in different sites. However, we demonstrate that correlations are not sufficient to explain the small value of the ID we observe in protein families. Indeed, we show that the ID of a set of protein sequences generated by maximum entropy models, an approach in which correlations are accounted for, is typically significantly larger than the value observed in natural protein families. We further prove that a critical factor to reproduce the natural ID is to take into consideration the phylogeny of sequences.

Author summary

Protein sequence evolution is an extremely complex process, whose roles are ultimately determined by the necessity of living organisms to adapt to changes in the environment. We here address a fundamental question related with this process: in how many independent directions can a sequence evolve, without compromising the protein capability of folding and of performing its function? We find that the number of these directions is surprisingly small, of 10 or less in most of the families we considered. This property is not correctly accounted for by most of the theoretical model we considered, which predict that sequence evolution can take place in 30-40 independent directions. The only way to accomplish the task of generating low-dimensional sequences is to take into consideration sequence phylogeny.

OPEN ACCESS

Citation: Facco E, Pagnani A, Russo ET, Laio A (2019) The intrinsic dimension of protein sequence evolution. PLoS Comput Biol 15(4): e1006767. https://doi.org/10.1371/journal.pcbi.1006767

Editor: David Liberles, Temple University, UNITED STATES

Received: April 27, 2018

Accepted: December 25, 2018

Published: April 8, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: AP acknowledges financial support from Marie Skl odowska-Curie, grant agreement No. 734439 (INFERNET). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Protein sequence evolution is an extremely important process in living organisms. During evolution, due to insertions, deletions, substitutions, a sequence can significantly change. Still, and most importantly, three-dimensional structures and functions are conserved, so that protein domains descending from a common ancestor share fundamental common traits; such protein domains form the so-called families. The birth of a new family is a rare event, while existing families are conserved by evolution. Despite the fact that the sequence similarity between members of the same family can be extremely low, by looking at the multiple sequence alignment (MSA) of a protein family one immediately notices patterns. Amino acids in specific columns of the MSA are often conserved, and mutations in different columns are in many cases correlated. This observation is at the very basis of statistical models for assessing the probability that a protein sequence belongs to a family [1] or for predicting the three-dimensional structure of the protein from the MSA [2, 3]. Frequent occurrences of the same amino acid in a column of the MSA together with covariation between different columns suggest that evolution modifies the sequences along a number of directions that is much lower than the bare dimension of the space sampled by randomly substituting amino acids.

We here address a specific question: how many independent directions are explored during evolution in a protein family? This issue can be rephrased in the conceptual framework of Intrinsic Dimension [4]. The Intrinsic dimension (ID) of a data set is defined as the minimum number of parameters needed to describe the data without information loss. Several methods are available to estimate the ID [4]; projective methods aim at representing points onto a lower dimensional space by minimizing an error function [5–7], while fractal methods measure the scaling of the number of points within a certain radius as such radius grows larger [8]. Nearest neighbors-based methods assume that close points are drawn from a uniform distribution and extract models for the statistical distribution of the first distances [4, 9, 10]. We here estimate the ID by TWO-NN [11], since the method is able to deal with density variations in the dataset. This is a fundamental requirement, as the protein sequences belonging to a family are not sampled uniformly: the sequence identity between a sequence and its closest neighbor can vary significantly, even within the same family. Moreover, and possibly more importantly, the TWO-NN estimator provides a criterion of reliability on the ID measure based on the quality of a linear regression.

Key to all the statistical properties of distances, is a notion of metrics on the set of protein sequences. This is a delicate point, since the metric is entangled with the ID estimate. The features that such distance should possess to be suitable for our purposes are essentially two: (i) being a metric (many dissimilarity measures used in bioinformatics do not satisfy the triangular inequality), (ii) to depend only on the pair of sequences it is computed on (thus excluding distances deriving from multiple sequence alignments). Taking into consideration these issues, we develop two notions of distances that we call Modified Hamming Distance, and BLOSUM distance and relate them to the well established Edit distance [12]. By using such distances we are able to compute the dimensions of several Pfam protein families [13], discovering that their IDs range approximately from 6 to 12. These values are robust to the choice of the metric and are correlated, to a certain extent, with the average length of the sequences as well as with the number of different architectures present in the family.

Finally, we study the reliability of artificial generative models for protein sequences from the point of view of the intrinsic dimension. The ID is a complex function of the data, but a meaningful one, and we suggest that it can be employed to assess the goodness of artificial models. We benchmark the capability of reproducing the correct ID of sets of protein sequences generated by HMMER [1], Direct Coupling Analysis [14, 15] and ProteinEvolver

[16], an approach which allows simulating protein sequence evolution taking into account phylogenetic history.

Methods

In this section we describe in detail the procedure to compute the intrinsic dimension of protein families. In particular, we address the following issues:

- Definition of a distance between sequences
- ID computation

The datasets we analyze are obtained by downloading the FASTA sequences of full families from the Pfam website [17]. Since we are interested in estimating the intrinsic dimension of protein sequence evolution at intermediate evolutionary distances, as a preliminary step we filter out correlated entries by means of CD-HIT [18], with a threshold of 80% of sequence similarity.

Definition of a distance between sequences

Defining a "good" distance between points is a crucial step to compute the intrinsic dimension of a dataset. In general under different metrics the ID can change. If the dataset is not representable in terms of coordinates (as in the case of protein sequences) the space itself is fully described by a set of pairwise distances. It is therefore of fundamental importance to analyze the relationship between the notion of distance employed and the resulting intrinsic dimension. In a formal context two metrics d and $\tilde{d}$ are said to be *equivalent* [19] if and only if there exists a finite positive constant C such that $\frac{1}{C}d(x,y) \leq \tilde{d}(x,y) \leq Cd(x,y)$. In practice, when comparing two notions of metrics it is possible to look at their correlation to infer their equivalence. Indeed, the intrinsic dimension is unchanged when the metric is altered to an equivalent metric [19]. Thus, even if we have at our disposal only the finite set of pair distances defined on a set of sequences, we expect that in the case of a good correlation between the distances obtained with different metrics the ID will be unchanged; this means that the intrinsic dimension is not only an attribute of a notion of distance, but rather of a class of distances associated to each other in terms of correlation.

Several methods are available in the literature to estimate pairwise sequence distances and similarities [20]. A definition of distance has to fulfill some fundamental requirements in order to describe a set of relationships between sequences where the ID estimation is well-posed. First of all we want our definition of dissimilarity D to resemble as much as possible a metric, meaning it should be non-negative, equal to zero only for identical sequences, symmetric, and it should satisfy the Triangular Inequality (TI). Another requisite we prescribe is that the notion of dissimilarity between two sequences depends only on the sequences themselves. Some of the well-known notions of distances, for example the Hamming distance [21], are based on a Multiple Sequence Alignment (MSA); in a MSA the match between two residues in two different sequences does not depend only on the two sequences, but on all the sequences used to derive the MSA: in this way the distance between two entries builds upon all the other sequences in the set, and a simple operation as adding new entries to the dataset could change the overall distribution of distances. For this reason our definitions of dissimilarity will rely only on pairwise sequence alignments. In the following we describe three notions of distances that fulfill the requirements enumerated above.

Modified Hamming distance. This measure of distance is based on pairwise alignment between sequences by means of BLAST [22]. Since the TWO-NN intrinsic dimension

estimator requires finding only the closest and the second closest nearest neighbors, only entries with an E-value lower than 10 are retained. If for sequences s_1 and s_2 two (or more) relevant MSPs, or alignments, are found, the one with lowest E-value is retained. Note that, due to its heuristic nature, BLAST is not symmetric, meaning that in principle, it could align differently an ordered pair of sequences (A, B) from the reversed (B, A). In this case the best (in terms of E-value) of the two alignments is retained. We define the Modified Hamming distance as:

$$d_{MH}(s_1, s_2) = \begin{cases} \frac{m - L \times \frac{P}{100}}{m} & \text{if } \mathrm{E-value}(s_1, s_2) < 10. \\ 10 & \text{otherwise} \end{cases} \quad (1)$$

Here by P we denote the percent identity of the best alignment between s_1 and s_2, by E-value (s_1, s_2) we indicate its E-value, while $m = \max\{m_1, m_2\}$ is the maximum between the two sequences lengths and L is the alignment length. In words, this corresponds to scoring matches as 0, mismatches as 1, counting the not aligned amino acids as mismatches and dividing by the maximum length of the two sequences. Note that by definition the distances lower than 10, that is to say those actually deriving from an alignment, are bounded above by 1.

We analyzed all significant triplets of sequences in a number of different PfamA families to count the percent of violations of the Triangular Inequality. The results obtained confirm that Modified Hamming is susceptible to TI violations but in such a small amount that it can be considered a metric. For instance in the case of PfamA family DnaJ the number of violations is 645 over a number of proper triplets of 69×10^7, thus only $\sim 9 \times 10^{-5}$% of the entries are involved.

BLOSUM distance. We also defined a variation of the Modified Hamming distance that instead of scoring the mismatches as 1 assigns them a score according to a BLOSUM matrix. The computation employs the bitscore of pairwise alignments (again part of the BLAST output), that is based on the score matrix BLOSUM62. Given two sequences s_1 and s_2 we define the BLOSUM distance d_{BL} as:

$$d_{BL}(s_1, s_2) = \begin{cases} \frac{M - S}{M} & \text{if } \mathrm{E-value}(s_1, s_2) < 10. \\ 10 & \text{otherwise} \end{cases} \quad (2)$$

Here M is the maximum bit score between s_1 against itself and s_2 against itself, and S is the bitscore of the best alignment. We empirically verified that d_{BL} is, to a good approximation, a metric.

In the same framework, we also considered a distance d_{SD} capable of capturing sequence divergence for short evolutionary times. At this scope, we performed sequence alignments using the substitution matrix from [23], which takes into account the evolutionary likelihood of a substitution from a codon model. We then followed the procedure in [24] to obtain an amino acid substitution matrix, at a reference sequence identity of 96%. We estimated the distance by Eq 2, aligning the sequences and computing the the scores M and S with this matrix. As shown in S1 Fig, the modified Hamming distance d_{MH}, the BLOSUM distance d_{BL} and the distance d_{SD} are well correlated at short distances.

Normalized edit/Levinstein distance. In information theory the Levenshtein (or Edit) distance is a string metric for measuring the difference between two sequences (see [25]). Informally, the Levenshtein distance between two words is the minimum number of single character edits (insertions, deletions or substitutions) required to change one word into the other. The Edit distance in its basic formulation is a true metric, and the TI can be formally

proved. Instead of using plain Edit we normalize it by the average length of the sequences, to apply a correction to the fact that it is easier for short sequences to present fewer mismatches. The normalized Edit can be also considered a metric, and is well correlated with d_{MH} at short distances, as displayed in S1 Fig). In the following we show that, as expected, these three notions of distance lead to consistent ID measures.

Finally we considered other definitions of dissimilarity between sequences, namely the p distance, the Kimura distance and the Jukes Cantor distance [20, 26]. We verified that none of these distances is well correlated with d_{MH}, mainly for the reason that they are computed on sequences aligned in an MSA. These dissimilarity measures do not depend only on pairs of sequences, and therefore, based on the considerations above, are not considered in our analysis.

ID computation

In this section we describe a procedure to estimate the intrinsic dimension of protein families where distances between sequences are defined according to d_{MH}. Even if d_{MH} is to a good approximation a metric, and thus a basic requirement of TWO-NN is fulfilled, it has an upper bound $u = 1$ that may induce artificial inhomogeneities in the space. TWO-NN method [11] is rooted on the computation, for each point $\mathbf{x}$ in the dataset, of its first and second nearest neighbors distances r_1 and r_2; the ratios $\mu(\mathbf{x}) \doteq \frac{r_2(\mathbf{x})}{r_1(\mathbf{x})}$ are collected to provide a measure of ID by a fitting procedure; if the hypothesis of the method are fulfilled, the set S given by $S = \{(\log(\mu_i), -\log(1 - F^{emp}(\mu_i))) \mid i = 1, \ldots, N\}$, where F^{emp} defines the empirical cumulate, is well fitted by a straight line whose slope corresponds to the intrinsic dimension of the dataset.

If we blindly apply TWO-NN to a dataset of d_{MH} distances obtained on Pfam family DnaJ we obtain a curved S set, that cannot be fitted by a straight line, indicating that TWO-NN cannot be applied in a straightforward fashion. We verified that this effect is a consequence of the upper bound interference. So if we could restrict our measure to neighborhoods whose radius r is relatively smaller than 1., we should be able to wash away the effects of the upper bound.

To implement this idea, we proceed as follows: let A be the set of μ values obtained on the whole dataset of d_{MH} distances. For different values of $\bar{r}$ we retain sets of μ values $A_{\bar{r}} \subset A$ such that a value $\mu = \frac{r_2}{r_1}$ belongs to $A_{\bar{r}}$ if and only if $r_2 \leq \bar{r}$. In symbols:

$$A_{\bar{r}} \doteq \{ \quad \mu \in A \quad \| \quad r_2 < \bar{r} \quad \}. \tag{3}$$

We then estimate the ID using TWO-NN only for the μ belonging to $A_{\bar{r}}$. To find out the optimal value of $\bar{r}$, for each value of $\bar{r} = 0.25, 0.26, \ldots, 0.6$ we compute the Root-Mean-Square Deviation (RMSD) of the fit of $S_{\bar{r}}$ to a straight line. The best choice for $\bar{r}$ is ideally the one minimizing the RMSD. The problem is that the RMSD around the minimum value fluctuates, therefore we set up a procedure to find such minimum together with an estimate of the error on the ID measure. First, we fit the points of the RMSD by a fourth-degree polynomial (cfr. Fig 1) in order to roughly locate the minimum. By performing the fit, we also compute the standard deviation (SD) of the data from the fitted curve. We then consider as putative minima all the data points falling within SD from the minimum of the polynomial. The optimal $\bar{r}$ is obtained by fitting a quadratic curve to this points, and finding its minimum argument, thus refining the previous computation. The error ϵ_d on the ID measure is estimated by the difference between the highest and the lowest value of the ID among the ones obtained for all the putative minima. In Fig 1, upper Panel, we plot the RMSD as a function of $\bar{r}$ for the PfamA family PF00226. In the bottom panels we plot the corresponding sets $S_{\bar{r}}$. It is clear that as $\bar{r}$ decreases towards the value $\bar{r} = 0.3$ the pronounced curvature that is visible in the set $S_{0.6}$ starts to diminish, and $S_{0.3}$ can be well fitted by a straight line. For lower values of $\bar{r}$ the

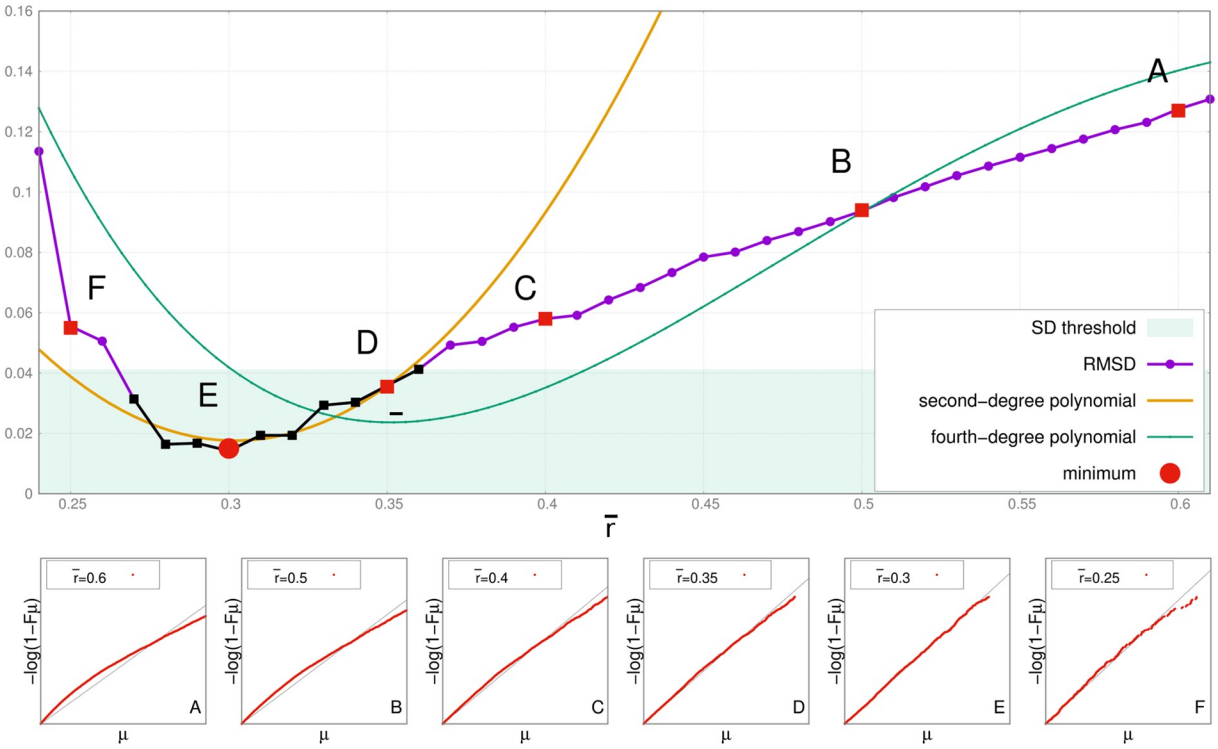

Fig 1. Technical procedure to locate the threshold value minimizing the RMSD. Upper Panel: RMSD of the sets $S_{\bar{r}}$ for different values of $\bar{r}$. The minimum of the RMSD is obtained for $\bar{r} = 0.3$ (point E). In green we show the four-degree polynomial fitting the RMSD, with standard deviation SD. The RMSD values falling below the minimum of such polynomial plus SD (green area) are plausible RMSD minima (highlighted in black). We fit these plausible RMSD minima by a second-degree polynomial (yellow curve). The argument of the minimum of such polynomial is considered as the putative value minimizing the RMSD. Panel A: sets $S_{\bar{r}}$ for $\bar{r} = 0.6$. Panel B: same as A for $\bar{r} = 0.5$. Panel C: same as A for $\bar{r} = 0.4$. Panel D: same as A for $\bar{r} = 0.35$. Panel E: same as A for $\bar{r} = 0.3$. Panel F: same as A for $\bar{r} = 0.25$. The S_{set} here is well fitted by a straight line, while the others show a curvature.

https://doi.org/10.1371/journal.pcbi.1006767.g001

curvature becomes visible again as we begin to see the effects of a new boundary, this time represented by $\bar{r}$ itself. We see that the hypotheses underlying TWO-NN are fulfilled within a range of $\bar{r}$ in which r_2 is far enough from the boundary, yet $\bar{r}$ is large enough not to influence the μ distribution. In the case of the example, the minimum of the RMSD is well defined and located at a value of 0.3. the ID corresponding to $\bar{r} = 0.3$ is ~ 9.

Wrapping up, the procedure to compute the intrinsic dimension of protein families is the following:

- Cluster sequences by sequence similarity through CD-HIT in order to obtain a reduced dataset where sequences that are too similar are excluded. This allows estimating the ID on an intermediate scale of sequence identity.
- Use BLAST to perform pairwise alignments and compute the Modified Hamming distance between sequences. The result is a sparse matrix where very high distances are not measured, but set to a default value. This choice speeds up the computation of distances, but introduces a boundary effect.
- Cure the boundary effects by computing the ID on a reduced dataset whose second nearest neighbors are "far enough" from the boundary, that is to say they are below a threshold $\bar{r}$. To find the optimal upper bound $\bar{r}$ try different values and choose the value corresponding to a set $S_{\bar{r}}$ that minimizes the RMSD of the fit to a line.

We verified that, in accordance with the considerations about correlation and equivalence we made above, the ID obtained by this procedure is consistent in the class of equivalent metrics encompassing the Edit distance, the Modified Hamming distance and the BLOSUM distance. In the case of Pfam family PUA the ID computed with the three different metrics is 11.2 for the Modified Hamming distance, 9.9 for the BLOSUM distance and 8.8 for the EDIT distance. This variation is of the order of magnitude of the estimated error on the ID. We also considered the distance obtained by using the substitution matrix from [23], which takes into account the evolutionary likelihood of a substitution from a codon model (see Methods). The ID estimated with this distance is $\sim$ 4.5, somehow lower. This discrepancy is relatively small, taking into account that this distance is built on a completely different principle with respect to the other three distances: indeed, it is designed to capture the quantified sequence divergence for short evolutionary times, while the other are more appropriate for describing divergence at intermediate or large times.

In the next section we apply the methods discussed above to the analysis of the ID of a set of Pfam families.

Results and discussion

Computing the intrinsic dimension of Pfam protein families

We analyzed several Pfam families belonging to different Pfam clans in order to explore a wide range of cases. The procedure explained in the previous section was first applied on a selection of families of Pfam release 31.0 enumerated in Table 1. The families are extracted from clans that are very different from each other: clan CL0489 for instance includes antifreeze proteins, while clan CL0378 consists of enzymes including luciferase. In Fig 2 we summarize the results obtained on some of these families.

First of all we note that in all the cases the reduced S-sets are very well fitted by a straight line, meaning that the procedure introduced in this work leads to a well defined intrinsic dimension. More precisely, this property implies that the ID is practically constant for sequences belonging to the same family [11]. The slopes of these straight lines, corresponding to the dimensions of the families, span from a value of around 6 to around 12; these values are quite similar and low relatively to the dimension of the embedding space: in fact, representing all the sequences of a family in a vector space would require in principle L coordinates, where L is the maximum sequence length in the family; L is normally of the order of at least 300 in all the cases displayed in Table 1. If we look at the sequence similarity within a family, we find entries sharing only 20% of the amino acids so that the number of mutations observed in a family is enormous. The low ID of the manifold containing the sequences can be interpreted in terms of allowance for mutations: the evolutionary pressure results in a lack of variations at specific positions and in correlated variations across different positions, both restricting the number of degrees of freedom. These results are consistent with the low dimension of 4 found in [27] for 6651 encoding for voltage sensor domains. In this case, though, the ID is computed by means of the Hamming distance on sequences aligned in a MSA. As we discussed above, this measure of similarity is not, strictly speaking, a metric, since its value for a pair of sequences depends on all the other sequences in the MSA. The intrinsic dimension is in principle different in different families, and this is what we observe in our measures. The natural question arising at this point is what determines a specific value of the intrinsic dimension.

In order to address this question, we first tested the dependence of the ID on the length of the Hidden Markov Model (HMM) of the family. A possible guess is that the longer is the length of the HMM, the larger the ID, as a longer HMM could allow for larger variability across the family. We study the correlation between the HMM length and the ID of the Pfam

Table 1. Intrinsic dimension, lenght of the seed alignment, architecture entropy and error on the ID measure ϵ_d for 27 Pfam familes in Pfam notation.

Family	ID	Hmm length	Entropy	ϵ_d
PF04266	7.2	107	0.17	0.16
PF01472	9.7	74	1.65	1.93
PF14306	11.	159	1.39	0.87
PF07786	8.3	223	0.38	1.41
PF02009	10.2	321	0.48	1.72
PF16976	8.4	116	0.78	1.67
PF02886	7.4	238	1.11	4.7
PF08666	11.6	63	1.61	3.57
PF13144	7.6	196	0.12	1.72
PF16177	10.9	55	0.61	2.87
PF04326	9.1	122	1.53	4.83
PF01918	8.3	67	0.16	1.12
PF01177	10.3	224	0.10	1.80
PF00471	7.3	47	0.01	0.27
PF01020	5.6	50	0.99	0.98
PF05170	8.	608	1.34	4.71
PF00185	10.5	157	0.43	2.42
PF13502	6.8	229	1.41	3.91
PF06827	9.9	30	0.89	1.59
PF13116	6.6	289	0.78	1.80
PF00226	9.1	63	2.64	0.75
PF08613	8.4	161	0.44	1.48
PF04357	7.1	383	0.44	1.35
PF00382	9.1	71	1.74	2.70
PF04080	8.8	255	0.39	0.67
PF00049	5.7	85	0.66	1.53
PF05426	10.1	272	0.92	0.79

https://doi.org/10.1371/journal.pcbi.1006767.t001

families listed in Table 1. This analysis is to be carried out carefully, since the measure on the ID is affected by an error that could, in principle, hide the correlation signal. To deal with such error, we retain only the 10 families with smaller error on the ID measure. On this subset of families the correlation between the ID and the seed length is rather small, with a Pearson coefficient of ~ 0.2. This indicates that the ID value is only slightly correlated with the typical length of the sequences belonging to a family.

We also investigated the possibility that the ID is correlated with the number of domain architectures in the family. Some families show a great variability of architectures while others are well represented by a single one. It is plausible that families encompassing a wider number of different architectures have a greater variability and thus a larger ID. To test this hypothesis, we examined the Pearson correlation between the intrinsic dimension and the entropy of the distribution of domain architectures across the family, defined as:

$$S = -\sum_a \frac{N_a}{N} \log\left(\frac{N_a}{N}\right).$$

Here N is the total number of sequences in the family and N_a is the number of sequences in the family associated with a given architecture a. Again, we compute the correlation on the

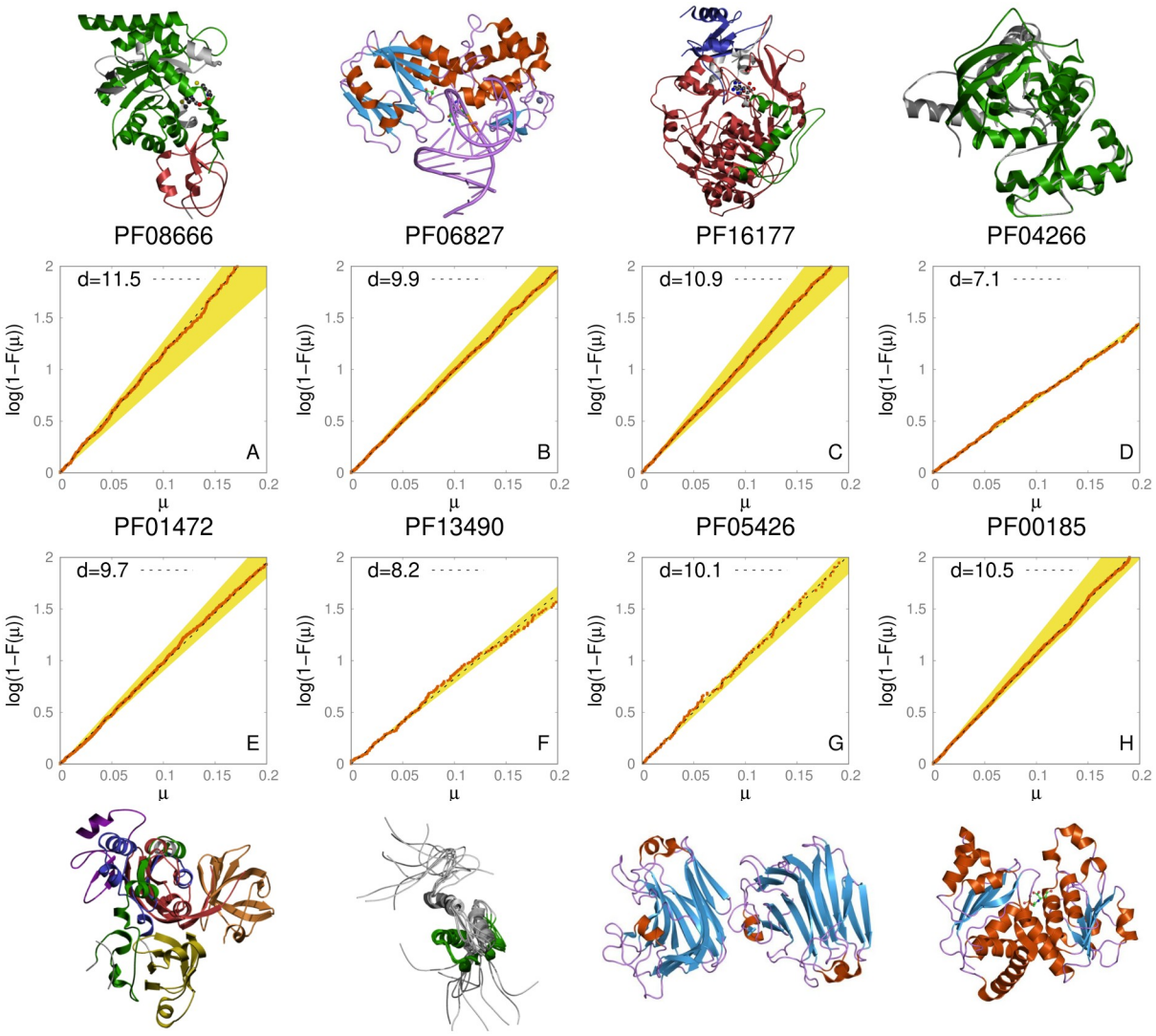

Fig 2. S set and estimated ID for a selection of Pfam A families. Panel A: we display the S set and the intrinsic dimension for PFAM domains PF08666 and the relative pdb structure. Panel B: same as A for PFAM domain PF06827. Panel C: same as A for PFAM domain PF16177. Panel D: same as A for PFAM domain PF04266. Panel E: same as A for PFAM domain PF01472. Panel F: same as A for PFAM domain PF13490. Panel G: same as A for PFAM domain PF05426. Panel H: same as A for PFAM domain PF00185. The pdb structures show that the results do not seem to depend on the structural characteristics that, among the chosen families, turn out to be highly heterogeneous.

https://doi.org/10.1371/journal.pcbi.1006767.g002

first 10 families in which the ID is computed with smaller error. In this case the computed Pearson coefficient is ~ 0.3. The analysis suggests that part of the variability of the ID in different families can be ascribed to the differences in the length of their typical sequence and to the number of architectures included in the family.

Estimating the ID of artificially generated protein sequences

We finally studied the reliability of artificial generative models for protein sequences from the point of view of the intrinsic dimension; we have seen that the dimension of protein families is relatively low; this is likely to be due to the constraints imposed by the three dimensional structure, that narrow to a great extent the allowed mutations. The aim of statistical models of protein sequence evolution is to unveil the statistical constraints underlying the variability of

sequences within a family, as well as to link this information to the conservation of biological structures and functions [14]. Once the statistical model has been set up, it could in principle be used in an inverse fashion, to generate sequences. A virtually perfect generative model should be able to reproduce the salient constraints acting on a family, and in particular the low ID. Over the last decade a number of global statistical inference approaches has been developed [1–3, 28–35]. In all the cases, the starting point is a Multiple Sequence Alignment (MSA) on a protein family. MSAs are rectangular matrices $A = \{a_i^v | i = 1, \ldots, L, v = 1, \ldots, M\}$, where M is the number of sequences, L the length of the alignment and a_i^v is either one of the 20 amino acids or a gap "-" standing for insertions or deletions (we stress that in this way gaps are considered on a par with any of the 20 amino acids, in accordance with standard practice). Thus, each line in a MSA corresponds to a protein sequence $(a_1, \ldots, a_L)$. A viable assumption for modeling the MSA is that it constitutes a sample of a Boltzmann distribution in the space of sequences:

$$P(a_1, \ldots, a_L) = \frac{1}{\mathcal{Z}} \exp\{-\mathcal{H}(a_1, \ldots, a_L)\} \quad , \tag{4}$$

where $\mathcal{Z}$ is a normalization constant. The key point of model inference is the reconstruction of the form, together with its specific coefficients, of the Hamiltonian $\mathcal{H}$ in the exponent of Eq 4. For instance, if the purpose is to reproduce exactly the first empirical moments computed from the MSA, as in the case of [1], $\mathcal{H}$ will take the shape:

$$\mathcal{H}(a_1, \ldots, a_L) = -\sum_{i=1}^{L} h_i(a_i), \tag{5}$$

where $h_i(a_i) = \log f_i(a_i)$ and $f_i(a_i)$ is the empirical frequency count computed from the reference MSA for the occurrence of amino-acid a_i at alignment position i. Increasing the level of complexity of the generative model, we considered artificial protein families generated by `hmmemit` (from the HMMER suite [1]), where the emission probabilities are those of the hidden Markov model which produced the MSA. Beyond reproducing the local frequency counts as in (5), this strategy allows taking into account the gap-gap non-local correlation structure generated by the frequent appearance of repeated gap stretches in specific parts of the alignment. Finally, as the most accurate generative model, we use Direct Coupling Analysis (DCA), that aims at reproducing also the empirical distribution of second moments, i.e. the empirical pair frequency counts $f_{ij}(a_i, a_j)$. These are obtained by counting the co-occurrence of amino acids a_i, a_j at position i, j in the MSA. Maximum entropy modeling dictates for $\mathcal{H}$ the following functional form:

$$\mathcal{H}(a_1, \ldots, a_L) = -\sum_{1 \le i < j \le L} J_{ij}(a_i, a_j) - \sum_{i=1}^{L} h_i(a_i). \tag{6}$$

The model parameters encoded in the interaction matrix $J_{ij}(a_i, a_j)$, and in the biases $h_i(a_i)$ can be estimated, for instance, by means of Boltzmann learning (BL) [36, 37]. The gaps are taken care of following the standard procedure of DCA, i.e. considering gaps as the 21st amino acid. Their frequencies and length are therefore by construction consistent with the corresponding distributions observed in the MSA.

The interest of the latter model lies in the fact that strongest pairwise couplings turn out to provide accurate prediction of contacts between residues, thus enabling protein-structure prediction. It is then natural to expect that artificial sequences generated from 6 can mirror other characteristics of natural sequences, which are not explicitly fitted by the model. To test the validity of models 5 and 6 with respect to the ID we study the case of Pfam family PF00076. In

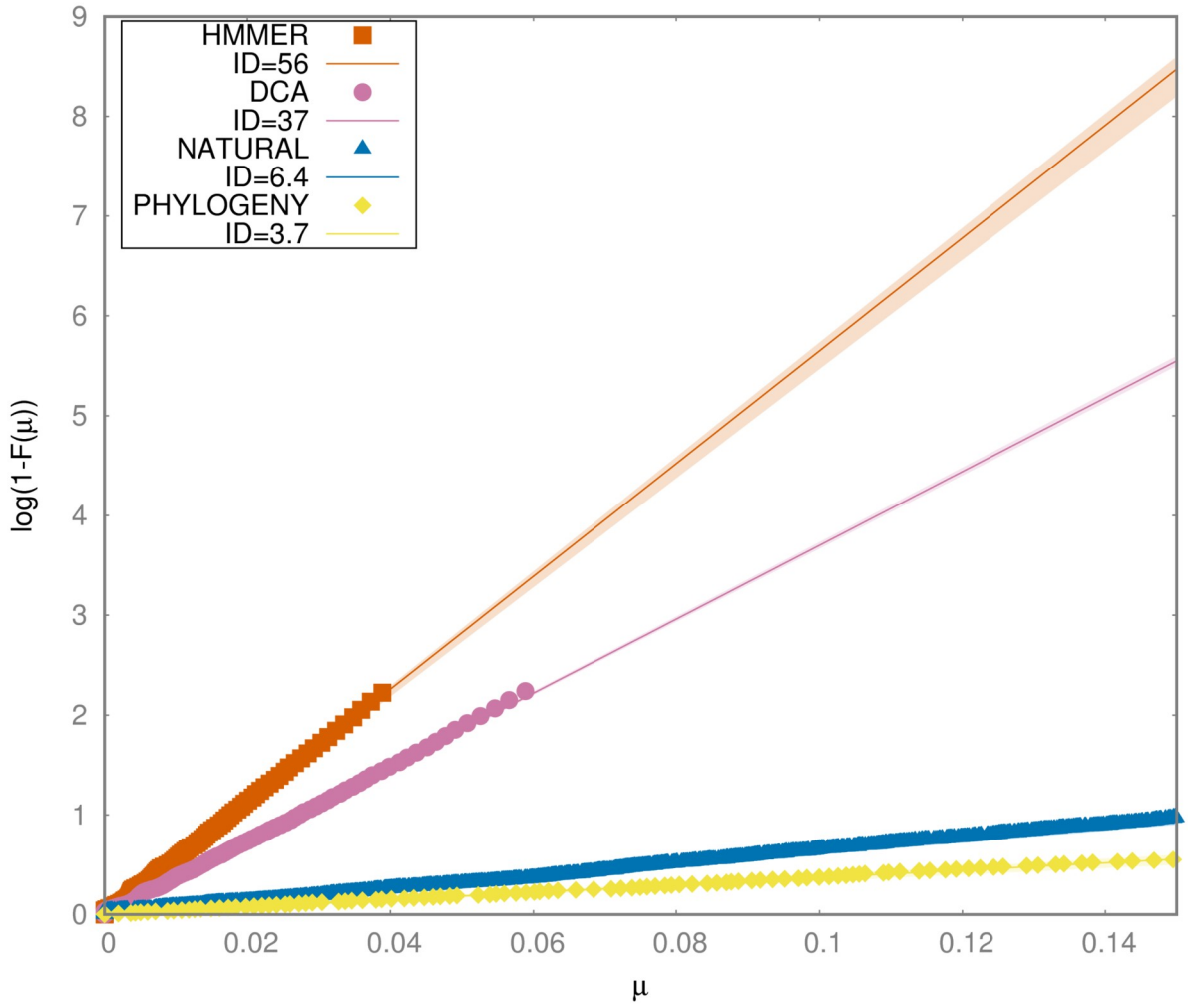

Fig 3. S set and fitting for generated dataset. S set and fitting lines together with the corresponding error on the fit (filled area) for sequences generated with HMMER [1], ACE [37], with the model by Arenas *et al* [16] introducing the phylogeny factor and for natural ones, in the case of Pfam family PF00076.

https://doi.org/10.1371/journal.pcbi.1006767.g003

Fig 3 we analyze the ID of sequences generated with HMMER [1] and DCA, and compare it to the ID of natural ones. The intrinsic dimension of sequences generated with HMMER [1] is the highest, with a value of $\sim$ 56 (with lowest ID within the error bars equal to 52), since the constraints related to covariation are not taken into account by construction. Sequences generated by means of DCA have a lower ID, as a consequence of the couplings, but the dimension is $\sim$37 (with lowest ID within the error bars equal to 27), still high with respect to the natural ones; in fact, according to the other Pfam families we analyzed, natural sequences from family PF00076 lie on a manifold with ID $\sim$ 6.

Taken together, these observations suggest that even if DCA is able to provide accurate predictions of contacts between residues, and thus to give insight into protein structure, yet it is not able to reproduce the ID of natural protein families. To do so, probably, pairwise couplings and local fields are not sufficient, and a more complex model has to be considered. From this point of view, the ID indicator turns out to be a severe and stringent touchstone for artificial sequence generation. To test this hypothesis, we simulated protein sequence evolution by the model introduced Arenas *et al* [16]. This approach allows taking into account the structure of

the native state, and phylogenetic history at the same time. This problem was recently raised in an interesting paper by Qin *et. al* [38], where the naive use of of DCA from MSAs was criticized due to the neglected confounding effect induced by phylogenetic branching.

We first performed the dynamics using as a reference the chain A of the structure 1g2e from the Protein Data Bank. This chain is one of the representative structures of Pfam family PF00076. We then selected at random 1000 sequences from the same family, and we built a phylogenetic tree using `FastTree v2.1` [39], an approximately-maximum-likelihood phylogenetic trees reconstruction from alignments of nucleotide or protein sequences. We ran ProteinEvolver [16], generating 5 replicates with the default settings provided in the distribution (the input files are produced as shown in S1 File). In this manner, we were able to generate a set of 5000 sequences that we analyzed with the standard procedure.

The result is shown in Fig 3: the ID of the set of sequences is ~ 4, slightly lower than the true value for this family ~ 6 but of the same order of magnitude. We verified that this result is robust with respect to changes of the parameters of the model (size of the phylogenetic tree, substitution model, sequence of the seed, etc.). In particular, we repeated the procedure starting from ten different sequences from the same family PF00076: A0A0P7BDN8_9HYPO/316-385, A0YL37_LYNSP/3-73, A0A0K8LDY3_9EURO/366-436, A0A0K3AU17_CAEEL/271-341, A4RRZ6_OSTLU/23-94, Q86BL5_DROME/654-724, A0A0V1HP59_9BILA/14-84, F1QWP0_DANRE/26-96, E2RGA8_CANLF/456-526, A0A0D1X4N1_9EURO/250-320. These sequence differ significantly in their sequence. For each of these initial seeds we repeated the whole procedure. The ID in each case differs by maximum 0.3 from the one generated starting from the sequence of 1g2e. This analysis demonstrates that taking into account phylogeny drastically reduces the dataset ID, thus reproducing the characteristic low intrinsic dimension peculiar to natural sequences.

Concluding remarks

In this work we study the intrinsic dimension of samples of protein sequences belonging to the same families. This value is a measure of the number of independent directions that evolution can take starting from any given sequence. A key point to properly address the problem is defining an appropriate distance between data points, namely the protein sequences. Many measures of sequence similarity introduced in the literature cannot be used in this context, either because they do not define a metric (not even approximately) or because their value does not depend only on pairs of sequences. Remarkably, the three measures of sequence similarity that can be considered appropriate distance measures, the Modified Hamming distance, the BLOSUM distance, and the Edit distance, are (roughly) equivalent [19], and lead to approximately the same value of intrinsic dimension. While we cannot exclude that other viable distance measures exist, which are not equivalent to the ones we consider here, this result implies that the intrinsic dimension we find is rather robust with respect to the exact choice of the metric, within the same class of equivalence.

By exploiting the properties of the TWO-NN estimator [11], we find that the intrinsic dimension is approximately constant within a family. This claim is non-trivial: in principle, one could expect the ID to vary in different branches of the evolutionary tree, but our results suggest that this is not the case. Empirically, we find that the intrinsic dimension varies within 6 and 12 among the families we considered, and we find hints that this variation can be at least partially ascribed to the average length of the protein sequence, and to the number of architectures that are present in a family. These values are remarkably small, especially because we considered only sequences differing by at least 20% of residues. As a consequence, if one observes ~ 60 mutations in a sequence of 300 amino acids, typically these mutations are

strongly correlated, and can be accurately described by a "basis" of dimension between 6 and 12. The low ID can be at least partially interpreted in terms of allowance for mutations: the evolutionary pressure results in a lack of variations at specific positions and in correlated variations across different positions, both restricting the number of degrees of freedom of the sequences. However, we find that the value of the ID is not reproduced by Direct Coupling Analysis, an approach that models these effects by a pairwise Potts-like Hamiltonian defined on the sequence space. Indeed, the value of the ID measured on sets of artificial sequences generated by DCA is several times larger than the value observed in natural sequences, even though DCA reproduces exactly the pair probability of observing two amino acids in two different sites (see also S2 Fig). Given the recent interest in using maximum entropy models to generate *in silico* functional protein sequences [40–44], we believe that ID analysis provides a very stringent test to assess the accuracy of the generative model to reproduce natural sequences belonging to a given protein family. We finally demonstrate that taking into consideration phylogeny in protein sequence evolution [16] implies a drastic reduction in the ID, to values that are close to those observed in the natural sequences. This indicates that the phylogenetic structure of the mutation history is essential for generating ensembles of structures with an amount of correlation consistent with observations.

Supporting information

S1 Fig. Correlation plots in the case of pfamA family PUA. (A) Correlation plot between d_{BL} and normalized Edit distances (B) Correlation plot between d_{BL} and d_{MH} distances. (C) Correlation plot between d_{SD} and d_{MH} distances. The four distances are correlated, especially at low values.
(TIF)

S2 Fig. Histograms of Hamming distances. Artificial sequences should in principle be indistinguishable from natural ones. One of the characteristics they are supposed to share, tested in [14], is the set of Hamming distances of sequences to the consensus sequence $(a_1^*, \ldots, a_L^*)$, defined by the most frequent amino acids $a_i^* = \mathrm{argmax}_a f_i(a)$ in the MSA. Here, we show the histograms of Hamming distances from the consensus for natural sequences and artificial ones in the case of Pfam family PF00076; here the DCA method employed to generate artificial sequences is Adaptive Cluster Expansion (ACE) [37, 45], that accurately reproduces the sampled and correlation at the cost of a high computational demand. Hamming distances of natural and model-generated sequences from Pfam family PF00076. The two histograms show that, from the point of view of the Hamming distance, natural and artificial sequences are in fact indistinguishable.
(TIF)

S1 File. ID Pipeline. The archive contains the following files we used to run the ID analysis: (i) a readme file, (ii) a shell script file, (iii) a fasta file containing protein sequences, (iv) the main c++ file for the analysis.
(ZIP)

Author Contributions

Investigation: Elena Facco, Andrea Pagnani, Elena Tea Russo, Alessandro Laio.

Methodology: Elena Facco, Andrea Pagnani, Alessandro Laio.

Software: Elena Facco.

Supervision: Andrea Pagnani, Alessandro Laio.

Visualization: Elena Facco.

Writing – original draft: Elena Facco, Andrea Pagnani, Alessandro Laio.

Writing – review & editing: Elena Facco, Andrea Pagnani, Elena Tea Russo, Alessandro Laio.

References

1. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011; 39(suppl_2):W29–W37. https://doi.org/10.1093/nar/gkr367 PMID: 21593126
2. Jaynes ET. Information Theory and Statistical Mechanics. Physical Review Series II. 1957; 106:620–630.
3. Jaynes ET. Information Theory and Statistical Mechanics II. Physical Review Series II. 1957; 108:171–190.
4. Campadelli P, Casiraghi E, Ceruti C, Rozza A. Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. Mathematical Problems in Engineering. 2015; 2015. https://doi.org/10.1155/2015/759567
5. Kruskal JB, Wish M. Multidimensional scaling. vol. 11. Sage; 1978.
6. Cox TF, Cox MA. Multidimensional scaling. CRC press; 2000.
7. Jolliffe I. Principal component analysis. Wiley Online Library; 2002.
8. Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. In: The Theory of Chaotic Attractors. Springer; 2004. p. 170–189.
9. Levina E, Bickel PJ. Maximum likelihood estimation of intrinsic dimension. In: Advances in neural information processing systems; 2004. p. 777–784.
10. Rozza A, Lombardi G, Ceruti C, Casiraghi E, Campadelli P. Novel high intrinsic dimensionality estimators. Machine learning. 2012; 89(1-2):37–65. https://doi.org/10.1007/s10994-012-5294-7
11. Facco E, d'Errico M, Rodriguez A, Laio A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific Reports. 2017; 7. https://doi.org/10.1038/s41598-017-11873-y PMID: 28939866
12. Ristad ES, Yianilos PN. Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20(5):522–532. https://doi.org/10.1109/34.682181
13. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research. 2016; 44(D1):D279–D285. https://doi.org/10.1093/nar/gkv1344 PMID: 26673716
14. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse Statistical Physics of Protein Sequences: A Key Issues Review. arXiv preprint arXiv:170301222. 2017;.
15. De Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nature Reviews Genetics. 2013; 14(4):249–261. https://doi.org/10.1038/nrg3414 PMID: 23458856
16. Arenas Miguel and Dos Santos Helena G. and Posada David and Bastolla Ugo, Protein evolution along phylogenetic histories under structurally constrained substitution models Bioinformatics, 2013; 29 (23);3020–3028. https://doi.org/10.1093/bioinformatics/btt530 PMID: 24037213
17. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic acids research. 2016; 44(D1):D279–D285. https://doi.org/10.1093/nar/gkv1344 PMID: 26673716
18. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658. https://doi.org/10.1093/bioinformatics/btl158 PMID: 16731699
19. Barnsley MF. Superfractals. Cambridge University Press; 2006.
20. Hollich V, Milchert L, Arvestad L, Sonnhammer ELL. Assessment of Protein Distance Measures and Tree-Building Methods for Phylogenetic Tree Reconstruction. Molecular Biology and Evolution. 2005; 22(11):2257–2264. https://doi.org/10.1093/molbev/msi224 PMID: 16049194
21. Robinson DJ. An introduction to abstract algebra. Walter de Gruyter; 2003.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

23. Kosiol Carolin, Holmes Ian, Goldman Nick. An Empirical Codon Model for Protein Sequence Evolution. Molecular Biology and Evolution. 2007; 24(7):1464–1479. https://doi.org/10.1093/molbev/msm064 PMID: 17400572

24. Rizzato F., Rodriguez A., and Laio A. Non-Markovian effects on protein sequence evolution due to site dependent substitution rates BMC Bioinformatics, 1t:258.

25. Mantaci S, Restivo A, Sciortino M. Distance measures for biological sequences: Some recent approaches. International Journal of Approximate Reasoning. 2008; 47(1):109–124. https://doi.org/10.1016/j.ijar.2007.03.011

26. Nei M, Zhang J. Evolutionary distance: estimation. eLS. 2005;.

27. Granata D, Carnevale V. Accurate Estimation of the Intrinsic Dimension Using Graph Distances: Unraveling the Geometric Complexity of Datasets. Scientific Reports. 2016; 6. https://doi.org/10.1038/srep31377 PMID: 27510265

28. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proceedings of the National Academy of Sciences. 2009; 106 (1):67–72. https://doi.org/10.1073/pnas.0805923106

29. Burger L, van Nimwegen E. Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. PLoS Comput Biol. 2010; 6(1):e1000633. https://doi.org/10.1371/journal.pcbi.1000633 PMID: 20052271

30. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011; 108(49):E1293–E1301. https://doi.org/10.1073/pnas.1111471108

31. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLoS ONE. 2011; 6(12):e28766. https://doi.org/10.1371/journal.pone.0028766 PMID: 22163331

32. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28:184. https://doi.org/10.1093/bioinformatics/btr638 PMID: 22101153

33. Sreekumar J, ter Braak C, van Ham R, van Dijk A. Correlated mutations via regularized multinomial regression. BMC Bioinformatics. 2011; 12(1):444. https://doi.org/10.1186/1471-2105-12-444 PMID: 22082126

34. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Physical Review E. 2013; 87(1):012707. https://doi.org/10.1103/PhysRevE.87.012707

35. Feinauer C, Skwark MJ, Pagnani A, Aurell E. Improving contact prediction along three dimensions. PLoS Computational Biology. 2014; 10(10):e1003847. https://doi.org/10.1371/journal.pcbi.1003847 PMID: 25299132

36. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Molecular Biology and Evolution. 2016; 33(1):268–280. https://doi.org/10.1093/molbev/msv211 PMID: 26446903

37. Cocco S, Monasson R. Adaptive cluster expansion for inferring Boltzmann machines with noisy data. Physical review letters. 2011; 106(9):090601. https://doi.org/10.1103/PhysRevLett.106.090601 PMID: 21405611

38. Qin C., and Colwell L. J. Power law tails in phylogenetic systems Proceedings of the National Academy of Sciences, Jan 2018, 201711913. https://doi.org/10.1073/pnas.1711913115

39. Price M.N., Dehal P.S., and Arkin A.P FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments PLoS ONE, 5(3):e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

40. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in artificial WW domains. Nature. 2005; 437(7058):579–583. https://doi.org/10.1038/nature03990 PMID: 16177795

41. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature. 2005; 437(7058):512–518. https://doi.org/10.1038/nature03991 PMID: 16177782

42. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Molecular Biology and Evolution. 2016; 33(1):268–280. https://doi.org/10.1093/molbev/msv211 PMID: 26446903

43. Jacquin H, Gilson A, Shakhnovich E, Cocco S, Monasson R. Benchmarking Inverse Statistical Approaches for Protein Structure and Design with Exactly Solvable Models. PLOS Computational Biology. 2016; 12(5):1–18. https://doi.org/10.1371/journal.pcbi.1004889

44. Asti L, Uguzzoni G, Marcatili P, Pagnani A. Maximum-Entropy Models of Sequenced Immune Repertoires Predict Antigen-Antibody Affinity. PLOS Computational Biology. 2016; 12(4):1–20. https://doi.org/10.1371/journal.pcbi.1004870

45. Barton J. P., De Leonardis E., Coucke A., and Cocco S., Ace: adaptive cluster expansion for maximum entropy graphical model inference Bioinformatics, 2016; 32(20);3089–3097. https://doi.org/10.1093/bioinformatics/btw328 PMID: 27329863