OPEN ACCESS Freely available online

PLoS one

Markov Models of Amino Acid Substitution to Study Proteins with Intrinsically Disordered Regions

Adam M. Szalkowski[1,2]*, **Maria Anisimova**[1,2]

1 Swiss Institute of Bioinformatics, Lausanne, Switzerland, **2** Computational Biochemistry Research Group, Computer Science Department, ETH Zurich, Zurich, Switzerland

Abstract

Background: Intrinsically disordered proteins (IDPs) or proteins with disordered regions (IDRs) do not have a well-defined tertiary structure, but perform a multitude of functions, often relying on their native disorder to achieve the binding flexibility through changing to alternative conformations. Intrinsic disorder is frequently found in all three kingdoms of life, and may occur in short stretches or span whole proteins. To date most studies contrasting the differences between ordered and disordered proteins focused on simple summary statistics. Here, we propose an evolutionary approach to study IDPs, and contrast patterns specific to ordered protein regions and the corresponding IDRs.

Results: Two empirical Markov models of amino acid substitutions were estimated, based on a large set of multiple sequence alignments with experimentally verified annotations of disordered regions from the DisProt database of IDPs. We applied new methods to detect differences in Markovian evolution and evolutionary rates between IDRs and the corresponding ordered protein regions. Further, we investigated the distribution of IDPs among functional categories, biochemical pathways and their preponderance to contain tandem repeats.

Conclusions: We find significant differences in the evolution between ordered and disordered regions of proteins. Most importantly we find that disorder promoting amino acids are more conserved in IDRs, indicating that in some cases not only amino acid composition but the specific sequence is important for function. This conjecture is also reinforced by the observation that for 27% of our data set IDRs evolve more slowly than the ordered parts of the proteins, while we still support the common view that IDRs in general evolve more quickly. The improvement in model fit indicates a possible improvement for various types of analyses e.g. *de novo* disorder prediction using a phylogenetic Hidden Markov Model based on our matrices showed a performance similar to other disorder predictors.

Citation: Szalkowski AM, Anisimova M (2011) Markov Models of Amino Acid Substitution to Study Proteins with Intrinsically Disordered Regions. PLoS ONE 6(5): e20488. doi:10.1371/journal.pone.0020488

Editor: Vladimir N. Uversky, University of South Florida College of Medicine, United States of America

Received March 22, 2011; **Accepted** April 27, 2011; **Published** May 27, 2011

Funding: This work was partially supported by the Swiss National Science Foundation grant to M.A. (ref. 31003A/127325). A.S. and M.A. are also funded by the Eidgenössische Technische Hochschule (ETH) Zürich. 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.

* E-mail: adam.szalkowski@inf.ethz.ch

Introduction

Contrary to the traditional sequence-structure-function paradigm, the function of a protein is not determined solely by its stable 3D structure. Today it is known that naturally unfolded or so called intrinsically disordered proteins (IDPs) fulfill a multitude of functions, such as signaling and regulation. While some proteins are completely unstructured, others may contain only short disordered regions. Current estimates suggest that more than 30% of eukaryotic proteins contain long intrinsically disordered regions (IDRs), but IDRs are also frequently found in prokaryotes [1]. According to estimates from Ward *et al.* [2], on average 10% of proteins are fully unstructured, while half of all proteins contain at least one long IDR. IDPs (or proteins with IDRs) often depend on structure instability for their function [3–5]. The absence of a stable 3D or secondary structure makes IDPs more flexible when binding and forming protein complexes, providing important advantages over ordered proteins [6]. Compared to ordered proteins, IDPs often participate in molecular recognition, signaling processes, cell-cycle regulation and modulating gene expression or chaperone activity [7]. Due to their flexibility, IDPs are more resistant to perturbations in the molecular interactions environment and tend to act as hubs in molecular interaction networks [8]. Proteins with IDRs are increasingly associated with diseases such as cancer and neurodegeneration [7]. For example, the CREB transcription factor is crucial in neuronal plasticity and long-term memory formation in the brain; malfunctions of CREB may contribute to the development of Huntington's disease and some types of cancers. Other famous examples include prion protein and tumor suppressor proteins p53 and BRCA1 [7,9].

To date most studies contrasting the differences between ordered and disordered proteins focused on simple summary statistics, such as sequence complexity and amino acid composition [10,11]. For example, regions of low sequence complexity are likely to be disordered [10]. IDRs usually have few large hydrophobic residues but favor polar and charged amino acids. Such sequence composition properties are often used by computational methods of disorder prediction (see [12]).

Brown *et al.* [11] estimated separate Markov amino acid substitution models for ordered and (wholly) intrinsically disordered

proteins at three levels of sequence similarity. These models were used to compare amino acid frequencies and average rates of evolution. The authors concluded disordered proteins having a generally higher rate of evolution than ordered. Midic *et al.* [13] published a scoring matrix for the alignment of protein sequences with disordered regions. This study also confirms a higher rate of evolution in IDRs and shows differences in amino acid substitution patterns between ordered and disordered parts of proteins.

Here, we take an evolutionary approach and study multiple sequence alignments of homologous proteins with IDRs using Markov amino acid substitution models in the maximum likelihood (ML) framework. Based on a large set of homologous groups with experimentally annotated IDRs, we estimate two empirical amino acid substitution models, each describing the evolution either in ordered or disordered regions. An expectation-maximization (EM) algorithm is used to obtain ML estimates of model parameters [14]. A new method is suggested to evaluate whether components of two inferred substitution models are significantly different. This test shows that our models are indeed significantly different and capture the essential features of ordered and disordered regions. As using the new model with *a priori* known IDRs significantly improves the fit to data, the new model may be recommended for other downstream evolutionary analyses. For example, using the new two component order-disorder model we define a phylogenetic Hidden Markov Model (phylo-HMM) and apply it as a *de novo* predictor of intrinsic disorder in multiple sequence alignments of homologous proteins. Our predictor demonstrates the potential of achieving a competitive accuracy-power balance compared to other disorder prediction methods.

Further, the estimated empirical models were used to contrast patterns and rates of the evolution in disordered and the corresponding ordered protein regions. It is typically thought that disordered regions are in general faster evolving than structured regions of the same protein. While we confirm that IDRs tend to have a higher rate of evolution compared to the rest of the protein, we find a significant number of protein groups where the reverse is the case. We present examples of proteins where the evolutionary rates at IDRs are significantly slower (or faster) compared to the corresponding ordered regions. Finally, we discuss other properties of IDPs such as their distribution among functional categories and biochemical pathways and their preponderance to contain repeats in tandem (another important property correlating with enhanced protein binding [15]).

Materials and Methods

Assembling homologous protein groups with IDRs

Previous analyses [16] relied on the computational prediction of intrinsic disorder [13,17–21]. Here, we decided not to use computational prediction methods. One one hand this drastically reduces the amount of data available for model estimation. On the other hand we avoid introducing unforeseeable biases due to prediction inaccuracies.

Instead our analyses were based on the DisProt database [22] as a starting point of data acquisition. This database comprises about 500 proteins annotated with a total of about 1000 experimentally verified intrinsically disordered regions.

DisProt was scanned for the presence of homologous proteins using the BLASTCLUST program, which finds pairs of sequences with statistically significant matches (using the BLAST algorithm) and groups them based on single-linkage clustering. This program is part of the BLAST suite [23]. We found that the DisProt database contains very few homologs. Consequently, we expanded the set of IDPs through further searches for homologous proteins in SwissProt and Pfam-based PANDIT databases. The similarity threshold was set to be sufficiently stringent to assume structural homology so that disorder annotations could be propagated to all homologous positions. A more detailed description of this procedure is provided below.

PANDIT data set

Each DisProt sequence entry was mapped to a representative homologous group in the PANDIT database [24] based on pairwise local alignment [25] score with BLOSUM62 [26]. The score threshold was set to 100 which corresponds to an E-value of 10^{-6}. PANDIT consists of Pfam protein families [27] together with multiple sequence alignments and inferred phylogenetic trees based on protein-coding DNA and amino acid sequences.

When multiple DisProt sequences mapped onto a single homologous group in PANDIT, the group was successively bisected by its longest branch until the mapping became injective, so that there was only one disorder annotation per group. Groups with no mapping or with <3 taxa were discarded. The corresponding multiple sequence alignments were restricted to the homologous sites as determined by the pairwise alignments to the reference sequence from DisProt. To avoid noise in the matrix estimation, distant sequences were filtered out based on the alignment score. The final set contained 223 homologous groups with a total of 1805 sequences with 54233 disordered and 254308 ordered residues.

The PANDIT data comprises a set of reliable alignments but due to its limited size it imposes considerable uncertainty on model estimation. Thus we use this data set mainly for verification of our results and for functional analyses.

SwissProt data set

To improve the reliability of our model estimation, we constructed a second larger data set of homologous protein groups with IDRs based on the SwissProt database [28]. For each DisProt entry an initial homologous group from SwissProt sequences was built from pairwise alignments. Multiple sequence alignments were constructed from pairwise homologies and trimmed to sites present in the reference sequence. Further, the groups were refined by removing distant sequences so that each sequence had a distance <100 PAM to the reference sequence. The resulting data set included 373 homologous protein groups with a total of 15490 sequences with 1043845 disordered and 3986493 ordered residues and was used as the main source for the estimation of our Markov model of evolution. To overcome potential biases due to errors in the group-wise multiple sequence alignments and estimated phylogenies, we compare the separate model estimates for both data sets.

The estimation of Markov amino acid substitution models

The evolution of amino acids was described by a Markov process with the generator matrix $Q=(q_{ij})$ defining the instantaneous rates of changes from amino acid i to j. As usual, the substitution process was assumed to be reversible so that $\pi_i q_{ij} = \pi_j q_{ji}$, where π_i are the *equilibrium amino-acid frequencies*. For a reversible process the instantaneous rates of change from i to j can be expressed as $q_{ij} = s_{ij}\pi_j$, a product of equilibrium (or stationary) amino acid frequency π_j and the exchangeability s_{ij} between residues i and j. We further refer to the matrix $S=(s_{ij})$ as the amino acid exchangeability matrix. For a multiple sequence alignment the substitution process flows along a phylogeny relating

the sequences in a sample. The transition probability matrix over time t is computed as $P(t)=e^{Qt}$. On this basis a likelihood function can be constructed for each site for a given tree. The total likelihood of the alignment is calculated as a product of site likelihoods based on the site-independence assumption (for computational reasons).

We estimated separate amino acid substitution models for ordered and disordered regions, each described with instantaneous substitution matrices D and O, respectively. Overall, the mixed *DO* model describes the evolution of *a priori* annotated IDRs using matrix *D*, while structured regions are described using matrix *O*.

Model parameters were estimated using an EM algorithm [14] on our two assembled training sets. The EM approach finds the ML estimates of substitution model parameters, with the substitution histories and counts being unobserved latent variables. The EM iteratively estimates parameters and latent variables in an alternating manner until convergence. Each model (both for ordered and disordered regions) required estimating 190 exchangeability and 19 amino acid frequency parameters.

For the training set based on PANDIT groups we used phylogenies provided by the PANDIT database. For the SwissProt data set phylogenies were built using PhyML3.0 [29] with $LG+\Gamma_4+I$ [30,31], thereby estimating evolutionary rates per site.

We followed the procedure described by Le *et al.* [30] and separated the alignment columns by their most likely rate class, as estimated by PhyML, to normalize for among-site heterogeneity of evolutionary rates.

Evaluating the significance of differences between models estimated for ordered and disordered regions

The significance of differences in estimated amino acid frequencies was evaluated using two likelihood-ratio tests on the estimated amino acid counts computed by XRate [14]: Pearson's χ^2-test and the G-test. Both tests compare the null hypothesis that the two count vectors arose from a common distribution against the alternative hypothesis where each vector originates from a distinct distribution. Similarly, exchangeability rates were compared using the estimated substitution counts.

In addition to Pearson's χ^2 and the G tests, confidence intervals for model estimates were computed by a bootstrapping technique. For each homologous group replicate data sets were generated by bootstrap on alignment columns and by jackknife on rows. For each replicate, substitution models for ordered and disordered regions were re-estimated using the EM-based procedure identical to that applied to the original data. The resulting distributions of model estimates were used to estimate empirical variances for exchangeabilities and amino acid frequencies (Figure 1).

In particular, we investigated whether the IDRs may be characterized only by the bias in amino acid composition, or if a bias in exchangeability between different classes of amino acids (order and disorder promoting) may also be observed. To achieve this we computed the substitution rates between order and disorder promoting residues for ordered and disordered regions separately:

$$\sigma'_{\mathcal{S}_\mathcal{O}} = \sum_{i\in S_O}\sum_{j\in S_O, i\neq j} \pi_i q_{ij}$$

$$\sigma'_{\mathcal{S}_\mathcal{D}} = \sum_{i\in S_D}\sum_{j\in S_D, i\neq j} \pi_i q_{ij}$$

where $\mathcal{S}_\mathcal{O}=\{C,F,I,L,V,W,Y\}$ and $\mathcal{S}_\mathcal{D}=\{A,E,G,K,P,Q,R,S\}$ are the sets of order promoting and disorder promoting amino acids, respectively. In order to compare these terms between ordered and disordered regions we normalized the terms by frequencies of occurrences of amino acids in sets S_O and S_D:

$$\sigma_{\mathcal{S}_\mathcal{O}} = \frac{\sigma'_{\mathcal{S}_\mathcal{O}}}{(\sum_{i\in\mathcal{O}}\pi_i)^2} \text{ and } \sigma_{\mathcal{S}_\mathcal{D}} = \frac{\sigma'_{\mathcal{S}_\mathcal{D}}}{(\sum_{i\in\mathcal{S}_\mathcal{D}}\pi_i)^2}$$

which rendered these terms independent of the target and source amino acid frequencies. To detect bias in exchangeabilities with regard to order and disorder promoting residues, we compared the ratios $\sigma_{\mathcal{S}_\mathcal{D}}/\sigma_{\mathcal{S}_\mathcal{O}}$ between ordered and disordered regions.

Comparison of evolutionary rates in ordered and disordered regions

To compare average rates of evolution we computed the group-wise total tree lengths (sum of branch lengths) for the SwissProt data set from pairwise distances and least-squares distance trees estimated with Darwin [32], because for a given set of taxa tree lengths are expected to be proportional to the average rates of evolution. We will refer to the estimated evolutionary rates as ρ_D for IDRs and ρ_O for ordered regions.

The ordered and disordered portions of multiple alignments were bootstrapped separately, and the significance was computed with the Mann-Whitney-U-Test.

Prediction of IDRs using phylogenetic Hidden Markov Models (phylo-HMMs)

The estimated empirical models for ordered and disordered regions may be used to define a phylo-HMM for predicting IDRs. We applied XRate [14] in annotation mode to obtain a prediction of order/disorder for each alignment column in the testing set compiled from PANDIT. This was done using the model estimates for order and disorder trained on either the PANDIT or the SwissProt sets. The HMM consists of 4 hidden states: start, end, and states for emitting ordered and disordered alignment columns. The emission probabilities are defined by the estimated evolutionary model and the transition probabilities were trained simultaneously from data. To correct for the differences in group size we divided the error statistics by the corresponding number of sequences in the homologous group.

The quality of prediction of intrinsic disorder for our phylo-HMM was compared with the quality of two sequence-based disorder predictors: VSL2 [20] and iupred [21]. VSL2 was used with two different parameter sets. One version of VSL2 uses auxiliary information from PSI-Blast PSSM and PSI-Pred secondary structure prediction, while the "fast" version is executed without this additional data. Iupred was used with its "long" and "short" presets.

Results

The new *DO* model requires twice as many parameters to be estimated from data compared to a standard empirical amino acid model that does not distinguish between order and disorder. Despite this, the model significantly improved the model fit to data with *a priori* annotated IDRs. For example, for the SwissProt data set the AIC decreased by 1916 (with an increase in log-likelihood of $\Delta l=1167$). Consequently, we used the DO model to analyze differences between ordered and disordered regions in terms of amino acid composition and exchangeabilities, evolutionary rates, and content of tandem repeats. We also tested whether the two

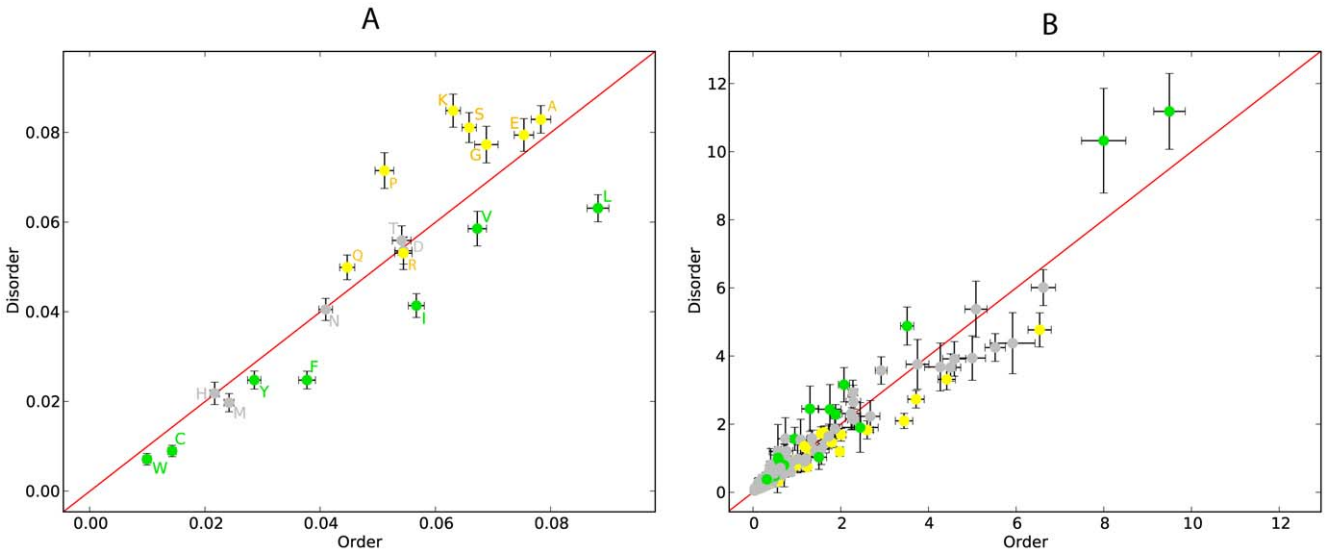

Figure 1. Scatter plot for amino acid frequencies (A) and exchangeablilities (B) in SwissProt data set. Error bars are 1.96 standard deviations. Order promoting amino acids are green, disorder promoting ones yellow. Exchangeabilities between order and disorder promoting residues are gray.
doi:10.1371/journal.pone.0020488.g001

components of *DO* may be used for disorder prediction from multiple alignments of homologous protein sequences.

Comparison of model estimates

We compared the ML estimates for the disorder and order components of the *DO* model in the PANDIT and SwissProt data sets. The majority of the ML estimates were similar between the two data sets. Only for a few amino acids the estimated equilibrium frequencies differed significantly between the data sets, based on variances estimated with bootstrap and jackknife resampling (Supplementary Figure S1). This may reflect a heterogeneity of gene and lineage composition. On the other hand, no significant differences were observed between exchangeability estimates in the two training sets. Such stability of our estimates is reassuring.

The uncertainty in model estimates for the SwissProt training set was lower compared to that for the PANDIT set. Moreover, we observed a lower variance in the estimates for the ordered regions compared to the disordered model. This may be explained by the amount of data available for each estimation, since the SwissProt set is larger than the PANDIT set, and since we have more residues in ordered regions compared with IDRs. Consistent with this explanation, we observe high variance in exchangeabilities between rare amino acids.

Next, we contrasted model estimates for ordered and disordered regions in the SwissProt data set. The estimates of model parameters for IDRs are shown in figure 2 and can be downloaded as supplementary datasets S3 and S4 in a format compatible with PAML [33]. The amino acid equilibrium frequencies and exchangeabilities are displayed and compared separately. The components of *O* and *D* matrices for ordered and disordered models were found to be significantly different based on Pearson's χ^2-test and the G-test ($p<0.01$ for both tests) applied to estimated substitution counts.

Based on the estimates of amino acid frequencies for ordered and disordered regions (Figure 2), we observed that order-promoting amino acids I, L, V (large and hydrophobic) and W, Y, F (aromatic) appeared in IDRs at a lower frequency. In addition, IDRs contained a low frequency of the non-polar amino acid C. On the other hand, IDRs contained high frequencies of disorder-promoting amino acids: positively charged R and K, polar E and Q, and small A, G, S and P. Our estimates of amino acid frequencies were largely in agreement with other empirical observations [10,11]. Our observations held for both data sets with only minor differences.

We clearly observed that IDRs are enriched with disorder-promoting amino acids while ordered regions are enriched with order-promoting amino acids. Further, significant differences in the amino acid exchangeability patterns between the models inferred for ordered and disordered regions were found (Figure 1).

In IDRs we observed relatively fewer substitutions between disorder promoting residues compared to ordered regions ($\sigma_{\mathcal{S}_\mathcal{D}}/\sigma_{\mathcal{S}_\mathcal{O}}=0.41$ in IDRs and 0.66 in ordered regions). In addition, in IDRs the exchangeability rates are higher between order-promoting residues, whereas in the ordered regions the exchangeability rates tend to be higher between disorder-promoting residues and between residues from the two classes (order or disorder promoting). Thus, it may be concluded that IDRs are characterized not only by the compositional bias but also by exchangeability biases between the classes of order and disorder-promoting residues.

Performance of HMMs for *de novo* disorder prediction

Using a test data set compiled from PANDIT, we compared the performance of our phylo-HMM based disorder predictor with two well established sequence-based predictors. Table 1 summarizes the numbers of correctly and incorrectly annotated sites with different methods tested. In our tests, VSL2 exposed the best performance in marking sites as disordered, while Iupred was too conservative, annotating too many sites as ordered. According to precision and recall values (Table 1), our phylo-HMMs outperformed the simple sequence based Hidden Markov Models based only on amino acid frequencies and yield results comparable to iupred and VSL2. It should be noted that VSL2 and iupred preformed similar or even better on the test set compared to predictions on DISPROT (results not shown).

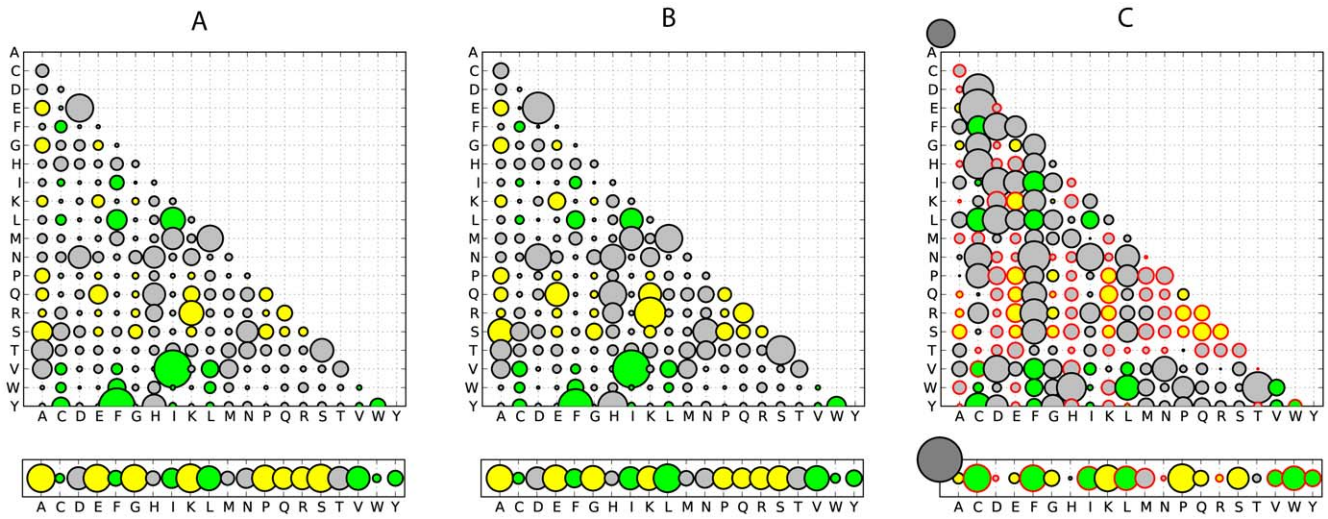

Figure 2. Amino acid exchangeability matrices and amino acid frequencies for ordered and disordered regions derived from the SwissProt data set: here, the area of each bubble represents the rate of a substitution or the amino acid frequency. (A) Model estimates for IDRs. (B) Model estimates for ordered regions. (C) Relative difference ($\frac{\rho_D-\rho_O}{\rho_O}$) between the corresponding values for disordered and ordered models (plots A and B). ρ_O and ρ_D stand for the relative evolutionary rates in ordered and disordered regions, respectively. Order promoting amino acids are green, disorder promoting ones yellow. Exchangeabilities between order and disorder promoting residues are gray. Bubbles with red border correspond to negative values, i.e. have a lower frequency in IDRs.
doi:10.1371/journal.pone.0020488.g002

Comparison of evolutionary rates in ordered and disordered protein regions

It is typically thought that IDRs evolve at a higher rate compared to proteins with stable 3D structure [11,34]. Here, we tested this consensus view using the rate estimates from our inferred models.

For about 60% of the PANDIT data set the evolutionary rates were significantly higher in IDRs compared to rates in ordered regions (Table 2). For example, the tumor suppressor protein p53 (PF00870) was found to evolve significantly faster in its IDR ($\rho_D/\rho_O = 4.2$). Another example from this class is discussed below in more detail. However, for 25% of our homologous groups the estimated rate of evolution in IDRs was significantly lower than in respective ordered regions ($p < 0.05$). The full results are available in supplementary dataset S1 ($\rho_D > \rho_O$) and S2 ($\rho_D < \rho_O$). This may indicate that the contribution of IDRs to the overall function of the protein may vary significantly, and which is confounded with a multitude of other factors including the properties of the primary sequence.

Further we explored the distribution of functional categories among groups with either significantly higher or lower evolutionary rates between ordered and disordered residues. For this task we used the PANDIT dataset since the information on functional categories (GO [35] terms) and biochemical pathways (KEGG [36]) was already available from PanditPlus [37]. For each protein group we parsed GO terms from the highest hierarchical level down to collect all relevant ancestral terms. The class with higher rate in IDRs ($\rho_D > \rho_O$) was enriched with proteins from the functional categories *'nucleotide binding'* ($p = 0.0162$, p-values before multiple testing correction), and especially *'adenyl nucleotide binding'* ($p = 0.0423$) and *'ATP binding'* ($p = 0.0569$).

In the other class with $\rho_D < \rho_O$ the cellular component *'membrane part'* ($p = 0.0241$) and the biological process *'regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process'* ($p = 0.0671$) were overrepresented. Due to the small number of homologous groups we had, these were the only terms close to a significant level. However, functional categories such as binding and regulation

Table 1. Comparison of disorder prediction.

Tool	TP	TN	FP	FN	accuracy	recall
VSL2 fast	3003	9209	3309	1237	0.2459	0.7083
VSL2	3268	8822	3696	973	0.2703	0.7706
iupred long	2175	11287	1230	2065	0.1616	0.5130
iupred short	2074	11092	1426	2166	0.1575	0.4892
phyHMM SwissProt	2728	9430	3217	1598	0.2244	0.6306
phyHMM PANDIT	3123	10136	2511	1203	0.2355	0.7219
HMM SwissProt	2313.1	11803.6	715.346	1928.08	0.1639	0.5454
HMM PANDIT	2113.29	11667.1	851.875	2127.89	0.1534	0.4983

Comparison of phylo-HMM based disorder prediction using the models estimated from the PANDIT or the SwissProt data set with other sequence based predictors. Shown are true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) for different parameter configurations of each method.
doi:10.1371/journal.pone.0020488.t001

Table 2. Comparison of evolutionary rates.

	$\rho_D > \rho_O$	$\rho_D < \rho_O$	$\rho_D \approx \rho_O$
LG	188	85	47
DO	186	82	52

Comparison of evolutionary rates between ordered and disordered columns in the SwissProt data set. Each cell contains the number of homologous groups which pass a test of significance at $p < 0.05$ or the number of those with indistinguishable rates.

doi:10.1371/journal.pone.0020488.t002

were reported to be enriched in other studies, which used disorder predictors for such analyses [38,39].

For KEGG pathways we found that proteins with less conserved IDRs ($\rho_D > \rho_O$) tend to be involved in *'proteasome'* ($p = 0.0370$),*'apoptosis'* ($p = 0.0569$) and *'colorectal cancer'* ($p = 0.0906$) pathways. Proteins with conserved IDRs ($\rho_D < \rho_O$) were found to be highly significantly overrepresented in *'tyrosine metabolism'* ($p = 0.00765$) as well as *'pyruvate metabolism'* ($p = 0.0242$), *'valine, leucine and isoleucine degradation* ($p = 0.0188$), *'urea cycle and metabolism of amino groups'* ($p = 0.0188$), *'1- and 2-methylnaphthalene degradation'* ($p = 0.0290$), *'fatty acid biosynthesis'* ($p = 0.0290$), and *'3-chloroacrylic acid degradation'* ($p = 0.0241$) pathways. Interestingly, according to KEGG most of these pathways fall into the same larger category or/and are related.

Note that the estimates of tree lengths were robust to the model choice (estimates for new model *DO* and LG differed by only $\pm 5\%$), and thus had little influence on our conclusions regarding the comparisons of the evolutionary rates ρ_D and ρ_O (Table 3).

Relationship between intrinsic disorder and protein repeats in tandem

It has been suggested that about 50% of the protein regions with tandem repeats may be intrinsically disordered [40–42], implying a higher incidence of IDRs in proteins with tandem repeats compared to their average frequency among all proteins. Here, we examined whether the reverse observation may be made, i.e. proteins with IDRs are more likely to contain tandem repeats. To assess whether IDPs are enriched with tandem repeats, we examined the frequency of tandem repeats in our homologous groups. For each group in our SwissProt data set and for each DisProt sequence, protein repeats were detected using a recent algorithm based on a k-means clustering approach [43]. We found that 69% (362/522) of the sequences in DisProt contained predicted repeats and that 76% (285/373) of our SwissProt groups contained at least one sequence with tandem repeats. This is significantly higher than what is typically observed among all proteins (reportedly 9% in SwissProt and 14% in a GenBank-based protein census [44]).

This analysis demonstrated that tandem repeats tend to occur more frequently in intrinsically disordered regions ($p < 0.05$; Table 4). Furthermore, in our data proteins with tandem repeats tended to have higher rates of evolution in IDRs ($\rho_D > \rho_O$) more frequently compared to proteins without tandem repeats (Table 4).

Table 3. Overlaps between rate estimate classes.

↓ LG/DO →	$\rho_D > \rho_O$	$\rho_D < \rho_O$	$\rho_D \approx \rho_O$
$\rho_D > \rho_O$	176	5	7
$\rho_D < \rho_O$	5	75	5
$\rho_D \approx \rho_O$	5	2	40

Overlaps between different rate estimate classes based on the LG and the *DO* models. Especially the overlaps between the opposing classes are within the targeted level of confidence ($p < 0.05$). The estimates based on the LG and DO models are not significantly different. Thus, using a single model for rate estimation is considered sufficient.

doi:10.1371/journal.pone.0020488.t003

Examples of proteins with IDRs

$\rho_D > \rho_O$ in mouse SOCS3. Significantly higher rate of evolution in the IDR compared to the ordered portion of the protein was found in the homologous protein group constructed for the DISPROT sequence DP00446. This protein is a suppressor of cytokine signaling (SOCS3) in mouse. The IDR between the SH2 domain and the C-terminal SOCS box (Figure 3) is believed to be a PEST-like sequence and is not required for primary function (phosphothyrosine binding) [45]. Instead it is likely to have an enhancing effect in protein degradation.

SOCS3 is involved in the following GO biological processes: '*branching involved in embryonic placenta morphogenesis*', '*negative regulation of insulin receptor signaling pathway*', '*negative regulation of signal transduction*', '*placenta blood vessel development*', '*positive regulation of cell differentiation*', '*regulation of growth*', '*regulation of protein phosphorylation*', '*spongiotrophoblast differentiation*', '*trophoblast giant cell differentiation*'. Further it is annotated with the molecular function '*protein binding*'. The protein is part of the following KEGG pathways: '*Ubiquitin mediated proteolysis*', '*Osteoclast differentiation*', '*Jak-STAT signaling pathway*', '*Insulin signaling pathway*', '*Adipocytokine signaling pathway*', '*Type II diabetes mellitus*', '*Hepatitis C*'.

We conducted a more thorough analysis of this protein and assembled a superset of this homologous group from the OMA project [46]. By doing so we obtained a group of 27 sequences and a total alignment length of 225 amino acids with 36 disordered columns. Babon *et al.* [45] note that the IDR of this protein is highly conserved in mammals. Despite this, our analysis confirmed a significantly higher substitution rate in the IDR compared to the rest of the protein. For this data set the estimated average tree length measured in expected substitutions per site was 1.2 in ordered regions but 8.4 for the IDR, with highly significant Mann-Whitney-U-Test. Further, protein-coding DNA sequences were analyzed using codon models with variable selection pressure over sites (models M0, M1, M2, M3, M7 and M7 in PAML [33]). No positive selection was detected on this protein, but the purifying selection pressure was less stringent in the IDR compared to the ordered part of the protein - the trend consistent with our observation of $\rho_D > \rho_O$.

$\rho_D < \rho_O$ in rat GNMP. The Glycine N-methyltransferase (GNMP) is an example of a protein where the rate of evolution is significantly lower in the IDR compared with the ordered regions of the protein - contrary to the predominant view. This protein creates a tetrameric complex shaping a molecular basket. The 40 unstructured N-terminal residues of each subunit regulate access to the active site by filling the core of this basket (Figure 4). In presence of AdoHcy these IDRs unclog the core and give access to the active site [47].

GNMP is involved in the following GO biological processes: '*adenosylhomocysteine metabolic process*', '*S-adenosylmethionine metabolic process*', '*folic acid metabolic process*', '*protein homotetramerization*'. Further it is annotated with the molecular functions '*folic acid binding*', '*glycine N-methyltransferase activity*', and '*glycine binding*'. The protein is part of the KEGG pathway '*Glycine, serine and threonine metabolism*'.

Similar to SOCS3, we expanded the original homologous group containing GNMP (around the DisProt sequence DP00031) with additional sequences from OMA. Thus this group was extended

Table 4. Intrinsic disorder and tandem repeats.

	# residues in order	# residues in disorder	$\rho_D > \rho_O$	$\rho_D < \rho_O$	$\rho_D \approx \rho_O$
TR	5217	5945	44	5	9
noTR	81207	47759	62	26	16

The first 2 columns contain numbers of ordered or disordered characters in DisProt which are predicted to be inside or outside of tandem repeats. Tandem repeats are significantly more frequent in disordered regions ($p < 0.001$).
The last 3 columns represent a comparison of evolutionary rates between ordered and disordered columns in the SwissProt data set restricted to groups with or without tandem repeats, respectively. Each cell contains the number of homologous groups which pass a test of significance at $p < 0.05$ or the number of those with indistinguishable rates.
doi:10.1371/journal.pone.0020488.t004

from 5 (originally) to 43 taxa and a total alignment length of 292 amino acids with 40 disordered columns.

The analysis of the extended dataset resulted in the estimated average tree length of 4.7 for the ordered regions versus 2.0 for the IDR, with highly significant Mann-Whitney-U-Test. The analyses with codon models (as for SOCS3) reported higher purifying selective pressure in the IDR. This again confirms our previous result suggesting that the IDR in GNMP protein is more conserved although it does not contain the active site.

Discussion

Here we estimated an empirical Markov amino acid substitution model for IDRs and ordered regions of proteins, which provided a significant improvement in model fit to data (as measured by AIC).

Figure 3. In murine SOCS3 the IDR (yellow) between the SH2 domain and the SOCS box is little conserved. It presumably just has an effect in the degradation of the protein. This structure is available as PDB identifier 2BBU.
doi:10.1371/journal.pone.0020488.g003

Based on the *a priori* annotated alignments, the mixed *DO* model succeeded at detecting several significant distinctions between evolutionary patterns in IDRs and the corresponding structured parts of the protein. First, the stationary amino acid distribution was found to be significantly skewed towards disorder promoting amino acids, which confirmed previous empirical observations [10,11]. Moreover, the exchangeability rates in IDPs were also biased, with significantly higher rates between order promoting residues. At the same time, the exchangeability rates for other types of changes were lower compared to what was observed in ordered regions. Probably, in IDRs disorder promoting amino acids are under higher functional constraints than order promoting residues. As a result, the *DO* model may better reflect the biological reality for IDPs and therefore may improve the accuracy of inferences for various types of analyses, such as maximum likelihood phylogeny inference with mixture models [48], ancestral reconstruction, and sequence alignment. As an example, we used our model to construct a phylo-HMM to predict intrinsic disorder from a multiple sequence alignment of IDPs based on the difference in evolutionary patterns. The phylo-HMM based on the estimated models was shown to be competitive compared with other sequence-based predictors. Combining this approach with the use of summary statistics, such as energy calculations or the inclusion into a meta-predictor may improve

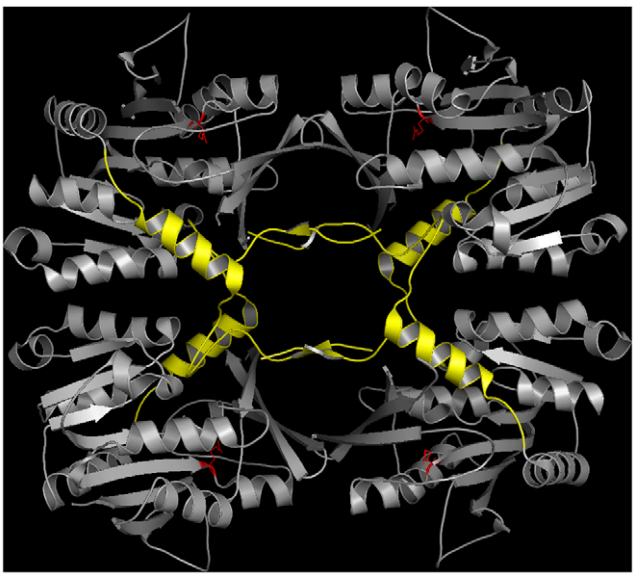

Figure 4. In rat GNMP the N-terminal IDRs (yellow) are strongly conserved. They give access to the active sites (red) in the presence of AdoHcy. This structure is available as PDB identifier 2IDJ.
doi:10.1371/journal.pone.0020488.g004

the prediction even further. One limitation of our study was the relatively small sample of IDPs that are currently known to contain structural disorder based on experimental work. Recently however, the natively unfolded proteins have been under spotlight [49]. The increased attention to IDPs is likely to increase the amount of structural information available for this kind of proteins. Larger sets of homologous alignments of IDPs may be used in the future to re-estimate the two empirical components of our *DO* model. Such new estimates should be more accurate and have smaller variances. However, the composition of IDRs may depend on their relevance to the function of the protein and the specifics of performed function. Given sufficient data, different *DO* models may be estimated for different classes of IDPs, where IDRs play different functional roles. At the moment this is not foreseeable due to lack of both structural and functional data.

Our analyses suggest that for the majority (59%) of IDPs the unstructured regions are indeed less conserved than the rest of the protein, as is typically thought. However, this is not a general rule and many exceptions exist. For 15% of IDPs the rates of evolution in IDRs and ordered regions were not significantly different. Moreover, a large proportion of IDPs in our set (27%) had higher conservation in their disordered parts, contradicting the common view (one such example, the GNMP protein, was presented above). Our functional enrichment analyses of this protein class showed that IDPs with $\rho_D < \rho_O$ tend to be involved in pathways responsible for amino acid and carbohydrate metabolism or related. In particular, the amino acid metabolism function involved the metabolism of order promoting residues (but not disorder promoting) - namely Tyrosine, Valine, Leucine and Isoleucine. We hypothesize that this may be related to our observation of higher exchangeability rates between order-promoting residues in IDRs compared to ordered regions. Overall, IDPs with slower evolving IDRs (compared to their structured parts) seem to exhibit a preferential involvement in certain biochemical pathways. Indeed, proteins whose IDRs are directly involved in function, or are crucially important for function, may be expected to evolve more slowly due to additional functional constraints. In addition, IDRs abundantly found in alternatively spliced regions [50,51] may evolve slower with respect to other regions due to additional constraints for functional proteins in different alternative frames.

A recent study [42] found that tandem protein repeats are enriched with IDRs. Here, we found that the reverse statement also may be made, i.e., proteins with IDRs are enriched in tandem repeats. So the presence of tandem repeats in a protein should have strong correlation with the presence of IDRs. This supports the theory that at least some of IDRs originate via repeat expansion [40]. This evolutionary mechanism provides a means of interactome scaling, where certain nodes in the interaction network increase their fitness by incorporating intrinsic disorder and repeats [9]. Sandhu [52] is also supportive of this view in his study of chromatin remodeling proteins that frequently contain IDRs. The IDRs resulting from repeat expansion may enable reversible binding to different interacting partners, which overall contributes to functional diversity and specialization of chromatin remodeling complexes. Moreover, Jorda *et al.* [42] found that the level of repeat perfection correlates with the amount of intrinsic disorder. If the repeat perfection is representative of recent evolutionary origin (rather than due to functional importance), then this finding is in a perfect agreement with the hypothesis that repeat expansion drives the origin of new IDRs. With time the repeat perfection should be decreased, especially that in our study we found that most IDPs with repeats evolve significantly faster in their IDRs compared to the structured regions. This may be also indicative that IDPs with tandem repeats fall into particular functional classes, a premise that should be studied when more structural and functional data (especially on IDPs) becomes available.

Supporting Information

Figure S1 Scatter plot of Pandit vs. SwissProt amino acid frequencies (A) and exchangeabilities (B) for the disordered model. Error bars are 1.96 standard deviations. Order promoting amino acids are green, disorder promoting ones yellow. Exchangeabilities between order and disorder promoting residues are gray.
(TIF)

Dataset S1
(CSV)

Dataset S2
(CSV)

Dataset S3
(QMAT)

Dataset S4
(QMAT)

Author Contributions

Conceived and designed the experiments: AMS MA. Performed the experiments: AMS. Analyzed the data: AMS MA. Contributed reagents/materials/analysis tools: AMS. Wrote the paper: AMS MA.

References

1. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, et al. (2001) Intrinsically disordered protein. Journal of Molecular Graphics & Modelling 19: 26–59.
2. Ward J, Sodhi J, McGuffin L, Buxton B, Jones D (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. Journal of Molecular Biology 337: 635–645.
3. Tompa P (2002) Intrinsically unstructured proteins. Trends in Biochemical Sciences 27: 527–533.
4. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6: 197–208.
5. Tompa P, Kalmar L (2010) Power law distribution defines structural disorder as a structural element directly linked with function. Journal of Molecular Biology 403: 346–350.
6. Uversky VN (2010) Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes. Chemical Society Reviews.
7. Iakoucheva LM, Brown CJ, Lawson J, Obradovic Z, Dunker A (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. Journal of Molecular Biology 323: 573–584.
8. Huang Y, Liu Z (2010) Smoothing molecular interactions: The kinetic buffer effect of intrinsically disordered proteins. Proteins: Structure, Function, and Bioinformatics 78: 3251–3259.
9. Dosztnyi Z, Chen J, Dunker AK, Simon I, Tompa P (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. Journal of Proteome Research 5: 2985–2995.
10. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, et al. (2001) Sequence complexity of disordered protein. PROTEINS-NEW YORK- 42: 3848.
11. Brown CJ, Johnson AK, Daughdrill GW (2010) Comparing models of evolution for ordered and disordered proteins. Molecular Biology and Evolution 27: 609–621.
12. Ferron F, Longhi S, Canard B, Karlin D (2006) A practical overview of protein disorder prediction methods. Proteins: Structure, Function, and Bioinformatics 65: 1–14.
13. Midic U, Dunker AK, Obradovic Z (2009) Protein sequence alignment and structural disorder: a substitution matrix for an extended alphabet. In: Proceedings of the KDD-09 Workshop on Statistical and Relational Learning in Bioinformatics. New York, NY, USA: ACM, StReBio '09, 2731 p. doi:10.1145/1562090.1562096. ACM ID: 1562096.

14. Klosterman P, Uzilov A, Bendana Y, Bradley R, Chao S, et al. (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics 7: 428.
15. Andrade MA, Perez-Iratxeta C, Ponting CP (2001) Protein repeats: Structures, functions, and evolution. Journal of Structural Biology 134: 117–131.
16. Dunker AK, Silman I, Uversky VN, Sussman JL (2008) Function and structure of inherently disordered proteins. Curr Opin Struct Biol 18: 75664.
17. Jones DT, Ward JJ (2003) Prediction of disordered regions in proteins from position specific score matrices. Proteins 53 Suppl 6: 5738.
18. Hecker J, Yang J, Cheng J (2008) Protein disorder prediction at multiple levels of sensitivity and specificity. BMC genomics 9: S9.
19. Schlessinger A, Liu J, Rost B (2007) Natively unstructured loops differ from other loops. PLoS Comput Biol 3: e140.
20. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7: 208.
21. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21: 3433–3434.
22. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, et al. (2005) DisProt: a database of protein disorder. Bioinformatics 21: 13740.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.
24. Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N (2006) PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Research 34: 327331.
25. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. Journal of Molecular Biology 147: 195–197.
26. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America 89: 10915–10919.
27. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins: Structure, Function, and Genetics 28: 405–420.
28. Bairoch A, Boeckmann B (1992) The SWISS-PROT protein sequence data bank. Nucleic Acids Research 20: 2019–2022.
29. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696704.
30. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25: 13071320.
31. Guindon S, Dufayard J, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate Maximum-Likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol 59: 307–321.
32. Gonnet GH, Hallett MT, Korostensky C, Bernardin L (2000) Darwin v. 2.0: an interpreted com-puter language for the biosciences. Bioinformatics 16: 101–103.
33. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586–1591.
34. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, et al. (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. Journal of Molecular Evolution 55: 104–110.
35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nat Genet 25: 25–29.
36. Kanehisa M (2002) The KEGG database. Novartis Foundation Symposium 247: 91–101; discussion 101–103, 119–128, 244–252.
37. Dimitrieva S, Anisimova M (2010) PANDITplus: toward better integration of evolutionary view on molecular sequences with supplementary bioinformatics resources. Trends in Evolutionary Biology 2: e1.
38. Lobley A, Swindells MB, Orengo CA, Jones DT (2007) Inferring function using patterns of native disorder in proteins. PLoS Comput Biol 3: e162.
39. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, et al. (2007) Functional anthology of intrinsic disorder. 3. ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J Proteome Res 6: 191732.
40. Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. BioEssays 25: 847–855.
41. Simon M, Hancock JM (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biology 10: R59.
42. Jorda J, Xue B, Uversky VN, Kajava AV (2010) Protein tandem repeats - the more perfect, the less structured. The FEBS Journal 277: 2673–2682.
43. Jorda J, Kajava AV (2009) T-REKS: identification of tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics 25: 2632–2638.
44. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D (1999) A census of protein repeats. Journal of Molecular Biology 293: 151–160.
45. Babon JJ, Yao S, DeSouza DP, Harrison CF, Fabri LJ, et al. (2005) Secondary structure assignment of mouse SOCS3 by NMR deffnes the domain boundaries and identifies an unstructured insertion in the SH2 domain. FEBS Journal 272: 6120–6130.
46. Schneider A, Dessimoz C, Gonnet GH (2007) OMA browser exploring orthologous relations across 352 complete genomes. Bioinformatics 23: 2180–2182.
47. Huang Y, Komoto J, Konishi K, Takata Y, Ogawa H, et al. (2000) Mechanisms for auto-inhibition and forced product release in glycine n-methyltransferase: crystal structures of wild-type, mutant R175K and s-adenosylhomocysteine-bound R175K enzymes. Journal of Molecular Biology 298: 149–162.
48. Le SQ, Lartillot N, Gascuel O (2008) Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B Biol Sci.
49. Dunker AK, Oldfield C, Meng J, Romero P, Yang J, et al. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 9: S1.
50. Kovacs E, Tompa P, Liliom K, Kalmar L (2010) Dual coding in alternative reading frames correlates with intrinsic protein disorder. Proceedings of the National Academy of Sciences 107: 5429–5434.
51. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, et al. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proceedings of the National Academy of Sciences 103: 8390–8395.
52. Sandhu KS (2009) Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. Journal of Molecular Recognition 22: 1–8.