Disordered Patterns in Clustered Protein Data Bank and in Eukaryotic and Bacterial Proteomes

Michail Yu. Lobanov, Oxana V. Galzitskaya*

Group of Bioinformatics, Institute of Protein Research Russian Academy of Sciences, Pushchino, Moscow Region, Russia

Abstract

We have constructed the clustered Protein Data Bank and obtained clusters of chains of different identity inside each cluster, http://bioinfo.protres.ru/st_pdb/. We have compiled the largest database of disordered patterns (141) from the clustered PDB where identity between chains inside of a cluster is larger or equal to 75% (version of 28 June 2010) by using simple rules of selection. The results of these analyses would help to further our understanding of the physicochemical and structural determinants of intrinsically disordered regions that serve as molecular recognition elements. We have analyzed the occurrence of the selected patterns in 97 eukaryotic and in 26 bacterial proteomes. The disordered patterns appear more often in eukaryotic than in bacterial proteomes. The matrix of correlation coefficients between numbers of proteins where a disordered pattern from the library of 141 disordered patterns appears at least once in 9 kingdoms of eukaryota and 5 phyla of bacteria have been calculated. As a rule, the correlation coefficients are higher inside of the considered kingdom than between them. The patterns with the frequent occurrence in proteomes have low complexity (PPPPP, GGGGG, EEEED, HHHH, KKKKK, SSTSS, QQQQQ), and the type of patterns vary across different proteomes, http://bioinfo.protres.ru/fp/search_new_pattern.html.

Introduction

Intrinsically disordered regions serve as molecular recognition elements, and play an important role in the control of many cellular processes and signaling pathways [1–6]. It is useful to be able to predict positions of disordered regions in protein chains. Prediction methods are aimed at identifying disordered regions through the analysis of amino acid sequences using mainly the physicochemical properties of amino acids [7–16] or evolutionary conservation [17–20].

Many examples of proteins with intrinsically disordered regions which exhibit coupling between folding and binding have been described in the literature [4–6,21–23]. Nevertheless, the universality of this phenomenon and functional importance of many disordered regions remain unclear.

A database of continuous protein fragments (Molecular Recognition Features or MORFs) was compiled from the Protein Data Bank which includes short protein chains (with fewer than 70 residues) bound to larger proteins [24,25]. It has been argued that MORFs participate in the coupling of binding and folding, a hypothesis that was supported by the analysis of the composition and predicted disorder of MORF segments. As a result of studying the subtle structural differences of the same proteins in bound (Complex) and unbound (Single) states in relation to their intrinsic disorder the database of protein structures (ComSin) has been constructed [26].

Recently several computational tools for identifying Linear motifs [27] and minimotifs in protein-protein interactions [28] have been published. Linear motifs are short segments of multidomain proteins that provide regulatory functions independently of protein tertiary structure [27] but minimotifs are short functional peptide sequences obtained after analysis of known protein-protein interactions [28].

Low-complexity regions attract our attention since they are regions of a protein in which a particular amino acid, or a small number of different amino acids, are enriched. Single amino acid repeats (homorepeats) belong to these regions. It turned out that homorepeats play important roles in some biological process [29] and may play a more important role in human diseases than it was previously recognized.

In the current study we search for sequence patterns consisting of a number of consecutive residues along the polypeptide chain that are nearly always associated with disordered segments. It has been found that two types of patterns appear to be recurrent: a proline-rich pattern and a positively or negatively charged pattern [30]. It should be noted that the old and new versions of our libraries include patterns enriched by proline and charged residues [31].

The statistical analysis of disordered residues was done considering 34 464 unique protein chains taken from the PDB database. In this database, 4.95% of residues are disordered (i.e. invisible in X-ray structures) [31]. The statistics was obtained separately for the N- and C-termini as well as for the central part of the protein chain. It has been shown that frequencies of occurrence of disordered residues of 20 types at the termini of protein chains differ from the ones in the middle part of the protein chain [31,32].
It is necessary to construct a clustered PDB because this simplifies the filtering process of protein structures under their analysis and searches general structural characteristics among non-identical proteins. It is necessary to construct a clustered PDB which is important for the analysis of actualized data.

In this work we constructed a clustered PDB and used clusters of protein chains where identity between chains inside of the cluster exceeds 75% (version of 28 June 2010). Combining the motif discovery and disorder protein segment identification in the clustered PDB allows us to create the largest library of disordered patterns [31]. At present the library includes 141 disordered patterns. Such an approach is new and promising for further studying and understanding the functional role of the obtained patterns in different proteomes. Taking into consideration the library of disordered patterns will help one improve accuracies of predictions for residues to be structured or unstructured inside the given region. The previous version of the library includes 109 disordered patterns and has restrictions on the minimal length of the patterns. Using more simple rules without restriction on the pattern length and clustered PDB of the same version we constructed the largest library of disordered patterns.

The patterns occur more often as short fragments. Patterns of four-six residues long occur more frequently (105 out of 141) among the disordered patterns of the library. It should be noted that six residue patches affect the folding/aggregation features of proteins, and they are important “words” for the understanding of protein dynamics [33]. Moreover, nucleation sites are constrained by patches of approximately six residues [34,35]. There is evidence that the minimum length necessary for a peptide to elicit an allergic response and molecular mimicry (a patch of a protein eliciting an immune response equivalent to the entire protein) is about six [36]. All these facts suggest the existence of a fragment of biologically meaningful information located along approximately six residues [33].

With the library of disordered patterns taken into account, it would be easier to improve accuracy of prediction of ordered/disordered residues inside the given region.

Proteome-wide calculations are a great way to place our work in a larger, evolutionary frame. In this paper of interest is the occurrence of disordered residues inside the cluster. Further we consider the identity of 75% because the general grouping has occurred below 90% identity. Considering this level of identity, we have created the Clustered Disordered Residues Data Bank (CDRDB), its elements are 18,775 clusters of protein chains. Among 116,997 chains, approximately 4.5% of their residues are disordered, i.e. are not resolved by X-ray analysis. To reveal such residues, we compared (for each protein chain) records SEQRES and records ATOM in the corresponding PDB-file. Residues which were present in record SEQRES, but their coordinates were absent in record ATOM (namely, the coordinates of the Cα-atom were absent in record ATOM), were considered as unstructured ones. We considered the residues as disordered if there were not coordinates of Cα atoms.

Below we consider only clusters with ≥75% identities between any pair of chains inside each cluster because the general grouping has occurred below 90% identity. Considering this level of identity, we have created the Clustered Disordered Residues Data Bank (CDRDB), its elements are 18,775 clusters of protein chains.

Construction of the library of disordered patterns

Among 116,997 chains, approximately 4.5% of their residues are disordered, i.e. are not resolved by X-ray analysis. To reveal such residues, we compared (for each protein chain) records SEQRES and records ATOM in the corresponding PDB-file. Residues which were present in record SEQRES, but their coordinates were absent in record ATOM (namely, the coordinates of the Cα-atom were absent in record ATOM), were considered as unstructured ones. We considered the residues as disordered if there were not coordinates of Cα atoms.

Below we consider only clusters with ≥75% identities between any pair of chains inside each cluster because the general grouping has occurred below 90% identity. Considering this level of identity, we have created the Clustered Disordered Residues Data Bank (CDRDB), its elements are 18,775 clusters of protein chains. Figure 2 illustrates two clusters with 100% identity combined in one cluster with 75% identity. One can see that the sequences from two clusters are different in one position 110, serin is changed for cysteine, and the weight of the chain from the first cluster is

\[
W_{\text{3kn2A}} = W_{\text{2a4C}} = \frac{1}{N_{\text{C100} \times M_{\text{C75}}}} = \frac{1}{8 \times 2^6}.
\]

and the weight of the chain from the second cluster is

\[
W_{\text{2a4kA}} = \frac{1}{2 \times 2^5}.
\]

respectively. Analogously the weight of each chain from any cluster is calculated by using equation:

\[
W_{\text{chain}} = \frac{1}{N_{\text{C100} \times M_{\text{C75}}}}.
\]
where $NC_{100}$ is the number of chains in the cluster with 100% identity and $MC_{75}$ is the number of clusters with 75% identity. It should be noted that the sum of weights inside one cluster with 75% identity will be equal to one. The weight of residue we consider to be the same as the weight of chain so at the level of 75% identity a cluster may include protein chains of different lengths.

Our goal is to create a database of disordered patterns i.e. amino acid sequences that are likely to be found in disordered parts of protein chains using CDRDB by applying simpler rules for the creation of the library of disordered patterns than in our previous work [31]. Let $P$ be a protein chain and $A$ be a pattern of length $L$. The database was compiled using a two-stage procedure.

At the first stage, we created a list of candidate patterns. To be a candidate in the patterns the considered pattern should be disordered in half cases among the chains from the cluster with 100% identity. Then the desired disordered patterns were selected into the candidate list. 855,775 candidates in the disordered patterns were gathered.

We say that pattern $A$ matches chain $P$ at position $s$ if the following conditions are valid:

1) two residues from each end should coincide:

\[
\begin{align*}
A[1] &= \text{chain}_s[s+1], \quad A[2] = \text{chain}_s[s+2], \\
A[L-1] &= \text{chain}_s[s+L-1], \quad A[L] = \text{chain}_s[s+L];
\end{align*}
\]

2) there could be done substitutions at most $L/5$ positions $r$ in the middle of pattern in which

\[
A[r] \neq \text{chain}_s[s+r].
\]

This means that for patterns with a length of $L \leq 5$ no change is possible, for $5 < L \leq 10$ – only 1 change, for $10 < L \leq 15$ – 2 changes, etc. The occurrence is terminal if it belongs to the first 40 residues (“N-terminal”) or last 40 residues (“C-terminal”) of the chain. The other occurrences are called internal ones.

If the distance between the edges of the pattern and the chain is less than 40 residues the pattern is considered to match these patterns.

Figure 1. Dependence of the number of clusters on identity between protein chains. Inside each cluster at the given identity between chains the identity is larger than the considered identity between clusters.
doi:10.1371/journal.pone.0027142.g001

Figure 2. Example of two clusters with 100% identity combined in one cluster with 75% identity. The sequences from two clusters are different only in one position (110), serin is changed for cystein. U denotes disordered residues in the chain and dash denotes ordered residues, respectively.
doi:10.1371/journal.pone.0027142.g002
residues. The pattern length is not limited in this paper. Further we
consider the following terminology: \(N_f\) is the sum of weights \(\langle w_{\text{chain}} \rangle\)
of disordered residues matched by the pattern; \(N_i\) is the sum of
weights \(\langle w_{\text{chain}} \rangle\) of ordered residues matched by the pattern; \(C_u\) is
the number of clusters with identity 75% \((C75)\) in which \(N_u\); 
\(C_f\) is the number of clusters with identity 75% \((C75)\) in which
\(N_f \leq N_i\). Protein \(P\) has an occurrence of pattern \(A\) if \(A\) matches \(P\) at
position \(s\).

Fragment \(A = P_{i+j-1} \ldots P_{i+L_j}\) of chain \(P_j\) is considered as a
candidate disordered pattern if it meets the following conditions:

\[ C1)\; C_u \geq 5; \; C2)\; C_u > C_f; \; C3)\; N_u > N_f. \]

There are 16,918 patterns meeting conditions C1, C2, and C3.
The longest pattern has the length of 45 amino acid residues
(\(\text{HHHHHHSSGLVPRGS} \ldots \text{DDDKAMA}\)), and the shortest pattern has 2 residues (\(\text{HH}\)). In
the next step we selected disordered patterns from the candidate list
using the following iterative greedy procedure. From 16,918
patterns we chose the pattern with the maximal value \(D = N_u - N_f\).
Then for the rest patterns the values of \(N_u, N_i, C_u, C_f\) were recalculated not taking into account the residues matched by
the first pattern. Again all the rest patterns were checked to meet
conditions C1, C2, and C3. Among the rest patterns meeting
conditions C1, C2, and C3 the pattern with a maximal \(D\) value
was chosen. If there were no patterns meeting conditions C1, C2, and
C3, then the procedure was stopped. The iterative procedure was
stopped when 390 patterns were selected (\(D > 0\)). Finally, we were
interested in the patterns for which \(D \geq 10\) and \(D \geq 25\) (the value 25
coincides, for example \(i = 39\) for \(n = 6\)).

Significance of disordered occurrences

We have studied the statistical significance of the selected patterns from two points of view. First, we were interested whether
the disordered fragments are overrepresented among the occurrences
of each pattern, and, second, whether the patterns are
overrepresented in the database. The features are described with
the proper \(Z\)-scores, called \(Z_{\text{disorder}}\) and \(Z_{\text{occur}}\), respectively. To
estimate the significance of the number of disordered occurrences of
pattern \(P\) we have implemented the following procedure. First,
we determined the fraction of disordered fragments among all
fragments with the given length taking into account the weight of
the disordered residues in each case:

\[
p(n) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{L_i-n+1} w_{i} \delta_{ik}}{\sum_{i=1}^{N} \sum_{k=1}^{L_i-n+1} w_{i}}, \quad (4)
\]

where \(N\) is the number of chains in the CDRDB, \(L_i\) is the length of
the considered chain, \(n\) is the fragment length, \(w_i\) is the chain
weight, \(\delta_{ik}\) is equal to 1 if the fragment with adjoining regions is
disordered more than in half positions, and 0 in the opposite case.
For each pattern we know the number of clusters \(C_u\) where this
pattern in more than half cases is disordered, and also the number of
clusters \(C_f\) where this pattern is folded in more than half cases
(see Dataset S1, columns J and K). We should calculate the
probability \(P(Y)\) that the number of successes will be larger or
equal to \(C_u\) at the given number of trials \(Y = C_u + C_f\).

In other words, this is the probability that at the given or larger
number of trials:

\[
P = \sum_{i=0}^{Y} \binom{C_u}{i} \left( \frac{p}{C_u + C_f} \right)^i \left( \frac{1-p}{C_u + C_f} \right)^{C_u - i}, \quad (5)
\]

where \(p\) is the probability of success of one trial (see above). The
significance of disordered occurrences is estimated with the \(Z\)-
score:

\[
Z_{\text{disorder}} = \frac{C_u - Y \cdot p}{\sqrt{C_u \cdot (1-p)}}. \quad (6)
\]

Statistical significance of the observed number of
occurrences of pattern \(X\) in proteomes

The probability of finding patterns with possible changes is
equal to the summation of probabilities over all sequences compatible with the given pattern.

\[
p(X) = \prod_{i=1}^{n} p_i, \quad (7)
\]

\(X^i\) is the sequence compatible with the given one (see the rules of
coincidence, for example \(i = 39\) for \(n = 6\)).

\[
p(X, N) = \frac{1}{\sqrt{2\pi N}} \exp \left(-\frac{1}{2} \left(\frac{p(X)}{N} - 1\right)^2\right). \quad (8)
\]

where the probability \(p(X)\) that pattern \(X\) occurs in a sequence and
\(p_i\) is the probabilities of occurrence of amino acids in the
considered proteome. We calculated the probability \(p(X, N)\) that
pattern \(X\) with \(n\) amino acid residues occurs in a sequence of
length \(N\):

\[
p(X, N) = \frac{p(X)}{N} \cdot (N - n + 1). \quad (9)
\]

The probability distribution on protein sequences is assumed to be
binomial.

The statistical significance of pattern \(X\) is estimated with the \(Z\)-
score

\[
Z_{\text{occur}} = Z(X, N) = \frac{S - R \cdot p(X, N)}{\sqrt{R \cdot p(X, N) \cdot (1 - p(X, N))}}, \quad (10)
\]

where \(S\) is the number of sequences containing at least one
occurrence of homorepeat \(X\). \(R\) is the number of proteins in the
considered proteome, \(N\) is the average length of proteins in the
considered proteome.

Statistical significance of the observed number of
occurrences of pattern \(X\) in two different proteomes

Let \(n_i\) and \(n_j\) be the numbers of proteins with the given pattern \(X\)
in proteomes \(i\) and \(j\). \(N_i\) and \(N_j\) are the whole numbers of proteins
in both proteomes, and \(p = n_j/N_j\) is the frequency of proteins with
the given pattern, \( s = \frac{\sqrt{r}}{C_{ij}} \) is the standard deviation. \( L_i \) and \( L_j \) are the average length of proteins in the considered proteomes \( i \) and \( j \). The scoring function is:

\[
Z_{\text{diff.a}} = \frac{p_i/L_i - p_j/L_j}{\sqrt{S_i/L_i^2 + S_j/L_j^2}}
\]

(11)

We consider that the difference is significant if its Z-score exceeds the given pattern. \( s \approx \sqrt{r} \) is standard deviations for variables \( x \) and \( y \).

\[
r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

\[
S_x, S_y
\]

(12)

\[
Z_{\text{diff.a}} = \frac{p_i/L_i - p_j/L_j}{\sqrt{S_i/L_i^2 + S_j/L_j^2}}
\]

(11)

We consider that the difference is significant if its Z-score exceeds the proper value with absolute meaning 3 and 5. These values correspond to the probabilities 3*10^{-3} and 6*10^{-5}, respectively.

The correlation coefficient \( r \) was calculated using the equation:

\[
\text{Database of proteomes}
\]

We considered 3279 proteomes from the EBI site (ftp://ftp.ebi.ac.uk/pub/databases/SPProteomes/uniprot/proteomes/). Since the patterns with the frequent occurrence in proteomes have low complexity we did a preliminary analysis. The analysis showed that the number of proteins with at least one occurrence of homorepeats of 6 residues long is less than 500 for proteomes with an overall number of residues below 2500000. Even so, only 22 proteomes out of 3156 have more than 100 proteins with at least one occurrence of 6-residue homorepeats. The data gave grounds for our research involving only proteomes with an overall number of residues exceeding 2500000.

We obtained 123 proteomes taking into account the length of proteomes representing 9 kingdoms of eukaryotes and 5 phyla of bacteria (see Table 1 and Dataset S2). Unfortunately, only three kingdoms of eukaryotes (Metazoa, Viridiplantae, and Fungi) are given at http://www.ncbi.nlm.nih.gov/Taxonomy/. In other cases, the rank of kingdom is missing. In such situations, we chose the highest taxonomic category proceeding from the subkingdom of eukaryotes instead of the kingdom. We chose 97 out of 120 eukaryotic proteomes, and a small number of bacterial proteomes. The smallest eukaryotic proteome belongs to *Hemiselmis andersenii*, class Cryptophyta. It is evident that 498 proteins with an overall number of 167452 of amino acid residues are not sufficient for reliable statistics. Historically, the superkingdom of bacteria is divided into phyla but not kingdoms. We preferred to consider such phyla separately.

Among 97 eukaryotic proteomes, 17 belong to the kingdom of Metazoa or animals: *Homo sapiens* (51778 protein sequences), *Bos taurus* (18405), *Mus musculus* (42120), *Rattus norvegicus* (29166), *Gallus gallus* (12954), *Danio rerio* (21576), and *Tetraodon nigroviridis* (27836) belong to Chordata phylum, *Drosophila melanogaster* (15101), *Drosophila pseudobscura* (16000), *Aedes aegypti* (16042), *Anopheles darlingi* (11437), and *Anopheles gambiae* (12455) to arthropods, and *Caenorhabditis briggsae* (18531), *Caenorhabditis elegans* (23817), *Loa loa* (16271), and *Trichinella spiralis* (16040) belong to nematodes, *Nematostella vectensis* (24435) belongs to cnidaria phylum.

**Results and Discussion**

**Library of disordered patterns**

Following the procedure described in the Materials and Methods section, we constructed the clustered PDB (CDRDB) at the level identity of 75% (http://bioinfo.protres.ru/stpdb/) and obtained a library of disordered patterns. The dataset includes 141 patterns (see Dataset S4). Figure 3 demonstrates the distribution of the patterns according to their lengths. The patterns occur more frequently as short fragments (105 out of 141 are patterns of 4-6 residues long). The largest pattern with condition D=25 consists of 17 amino acid residues [HHHHHSGSGLYVFQGP]. It is interesting that the strong pattern is HHHH, but not HHHHHH as in the last version of the library [31]. We suggest that the residues matched by these patterns will be disordered in new protein chains because more than half of residues in these patterns are disordered (see conditions C2 and C3 in the Materials and Methods section).

The statistical significance of disordered occurrences in the selected patterns was estimated with the \( Z \)-score (see Materials and Methods). We calculated the probability that the number of successes will be larger or equal to \( N_i \) at the given number of \( C_i = C_f \) for each pattern we know the number of clusters \( C_f \) where this pattern in more than half cases is disordered, and also the number of clusters \( C_f \) where this pattern is ordered in more than half cases. This probability for all 141 disordered patterns is less than \( 7 \times 10^{-5} \).

All 141 patterns have \( Z_{\text{disorder}} \geq 6.4 \) that corresponds to the P-value of \( 7 \times 10^{-5} \), which is in good agreement with the procedure of the disordered patterns determination. The worst variant is \( C_i = 5 \), \( C_f = 4 \), and the length of patterns is 6. We have four such cases: SVAESS, ASIGQA, PPCGSP, and DSDVSL (see Dataset S1, columns O and P).

**Comparison of the new and the previous libraries of disordered patterns**

After construction of the new library the question about similarity of two databases (previous and new) arises. For this purpose the previous patterns matched the clustered pdb (CDRDB) and the sum of weights was calculated analogously to the new patterns. Then we calculated the sum of weights for residues matching both the previous and the new patterns (intersections, \( I_{12} \)). The number of clusters with identity of 75% in which there were new and previous patterns was calculated, as well as the number of intersections. The degree of coincidence was calculated using equations (13) and (14):

\[
F_1 = I_{12}/(N_1 + N_2 - I_{12})
\]

(13)

\[
F_2 = I_{12}/\min(N_1, N_2)
\]

(14)

where \( I_{12} \) is the sum of weights for intersections (coincidences), and \( N \) is the weight of a single pattern. We considered only pairs where \( F_2 > 0.1, F_2(C75) > 0.1, I_{12} \geq 3 \), and the number of clusters where two patterns appear together, \( C_{12} \geq 3 \) (see Dataset S3). The measure \( F_1 \) points to the coincidence between two considered patterns. At the same time the measure \( F_2 \) demonstrates a level of inclusion of the pattern with smaller \( N \) into a larger one. Large difference between \( N_1 \) and \( N_2 \) results in a wide difference between \( F_1 \) and \( F_2 \).

For example, the sequence GSISHHHHHHSSGGLYVFQGP occurs in 393 clusters on the N-termini, where it is disordered more than half in 387 clusters. This sequence is matched by pattern GSHM, and its beginning is matched by the HHHH pattern. If we have a test database with one protein where there is such a sequence at the N-end, then \( N_{\text{GSHM}} = 20 \). \( N \) is the weight of a pure pattern with the neighboring part, in this case this is the length of the whole N-terminal fragment, \( N_{\text{HHHH}} = 9, I_{12} = 9 \).
Table 1. Names of 97 eukaryotic and 26 bacterial proteomes.

| Eukaryota     | Eukaryota (Fungi) | Bacteria*** |
|---------------|-------------------|------------|
| Metazoa       | 25.H_sapiens      | Acidobacteria | 25797.S_usitatus |
|               | 22974.B_taurus    | Actinobacteria | 37022.A_mediterranei |
| 59.M_musculus | 34495.A_ermia     |             |
| 122.R_norvegicus | 34498.A_ermia    |             |
| 21457.G_gallus | 35919.A_benhamiae |             |
| 20721.D_reio   | 29154.A_clavatus  |             |
| 22388.T_nigroviridis | 33020.A_flavus  |             |
| 17.D_melanogaster | 23118.A_fumigatus FGSC_A1100 |             |
| 25396.D_pseudoobscura | 31018.A_fumigatus CEA10 |             |
| 31436.A_aegypti | 29130.A_niger    |             |
| 78607.A_darlingi | 23077.A_oryza    |             |
| 22426.A_gambiae | 28239.A_terreus  |             |
| 21633.C_briggsae | 30100.B_fuckeliana |             |
| 9.C_elegans   | 22024.C_albicans_CS314 |             |
| 64800.L_loa   | 32738.C_dubliniens |             |
| 79720.T_spiralis | 19665.C_glabrata | Proteobacteria | 112.B_japonicum |
| 30565.N_vectensis | 34491.C_tropicalis |             |
| Viridiplantae  | 23214.O_sativa    |             |
| 3.A_thaliana  | 34493.C_lusatiae  |             |
| 33157.Micromonas_sp | 34218.C_posadasi |             |
| 29351.O_lucimarinus | 79902.C_graminicola |             |
| 25972.O_tauri | 20018.D_hansenii |             |
| Stramenopiles* | 35109.E_siliculosus |             |
| Choanoflagellida** | 30562.M_brevicollis |             |
| Euglenozoa*   | 83400.L_braziliensis |             |
| Alveolata*    | 71330.T_brucii    |             |
| Diplomonadida* | 33600.G_intestinalis_ATCC_50803 |             |
| Amoebozoa*    | 31998.P_chabaudi  |             |
| Diatomaceae   | 34387.P_marina   |             |
| 31342.P_knowlesi | 34392.P_brasiilensis_ATCC_MYA-826 |             |
| 31632.P_vivax | 31898.P_chrysogenum |             |
| Diplomonadida* | 33000.G_intestinalis_ATCC_50803 |             |
| Amoebozoa*    | 35301.P_pallidum  |             |
| Diplomonadida* | 35295.G_intestinalis_ATCC_50581 |             |
| 65115.G_intestinalis | 79905.P_teres |             |
| 30091.S_cerevisiae_YJM789 |             |             |
| 31651.S_cerevisiae_RN11-1a |             |             |
| 35062.S_cerevisiae_Lalvin_EC1118 |             |             |
| 31632.P_vivax | 31898.P_chrysogenum |             |
| Diplomonadida* | 32999.P_marneffei |             |
| 31342.P_knowlesi | 34392.P_brasiilensis_ATCC_MYA-826 |             |
| 31632.P_vivax | 31898.P_chrysogenum |             |
| Diplomonadida* | 32999.P_marneffei |             |
| 31342.P_knowlesi | 34392.P_brasiilensis_ATCC_MYA-826 |             |
| 31632.P_vivax | 31898.P_chrysogenum |             |
| Diplomonadida* | 32999.P_marneffei |             |
F₁ = 9/20 = 0.45, F₂ = 9/9 = 1. In a real situation, in the whole CDRDB N_HHHH = 29,560.4, N_GSHM = 8,452.0, I₁₂ = 3,163.1, F₁ = 0.09, F₂ = 0.37. It should be noted that real F₂ is less than test F₂. This occurs because GSHM appears usually in sequence GSSHHHHHSSGLVPRGSHM or in similar sequences. Yet sometimes GSHM appears alone.

The result of intersections of the two libraries (the previous library includes 109 patterns and the new one includes 390 patterns if D > 0) is presented in Fig. 4. One can see that there are 16 precise coinciding patterns: ENLYFQ, ASMTGGQQMGR, GSSHHH, WSHPQFEK, EGSHHHHHH, RRGKKK, PTTENLYFQGAM, PTTENLYFQGAM, SHHHHHHSQDP, HHHHHMA, SMGGQQMQRGS, KKGKST, SRSSHHH, ENLYFGGS, GGRHHH, HHHGSM, GSHMSQ, and 8 with not precise coincidence, for example HHHHHH and HHHH (Dataset S3).

It is interesting that some patterns appear in a protein together with other patterns (57 out of 141). Such pairs can be seen in Dataset S4. Also we calculated the number of patterns which appear in proteins together with the considered pattern (see Fig. 5). Pattern HHHH occurs more often with other patterns in proteins. It should be noted that there are several patterns which appear alone in the CDRDB (see Fig. 5, Dataset S4). We used the same criteria as for the intersections of the two libraries.

**Occurrence of disordered patterns in 97 eukaryotic and 26 bacterial proteomes**

After creating the library of disordered patterns taken from the CDRDB, another interesting question arises: how often the obtained patterns could occur in some proteomes. Since eukaryotic proteomes include more disordered regions than other proteomes [17,37,38] we compared 97 eukaryotic proteomes and 26 bacterial ones (see Table 1, Dataset S1, and Materials and Methods).

We considered two cases for coincidence. In the first case we calculated the number of proteins where the patterns match with precise coincidence a polypeptide chain fragment. In the second case we analyzed the coincidence according to the definition

---

**Table 1. Cont.**

| Eukaryota | Eukaryota (Fungi) | Bacteria*** |
|-----------|-------------------|-------------|
| 35359.V_albo-atrum | 20011.Y_lipolytica | |
| 31020.C_cinerea | 20846.C_neoformans_JEC21 | |
| 21380.C_neoformans_B-3501A | 31023.L_bicolor | |
| 33031.P_placenta | 22029.U_maydis | |

*Category without rank is given.
**The name of order is given because the highest ranks are missing in the taxonomic description.
***The superkingdom of bacteria is divided in phyla rather than kingdoms.

doi:10.1371/journal.pone.0027142.t001

---

**Figure 3. Dependence of the number of patterns on the length (number of amino acid residues).**

[Graph showing the relationship between pattern length and number of patterns]

doi:10.1371/journal.pone.0027142.g003
suggested here and in the paper [31]. According to the rule mentioned in the Materials and Methods section for patterns with a length of \(L\leq5\) no change may occur, for \(5<L\leq10\) – only 1 change may take place, for \(10<L\leq15\) – 2 changes, etc.

Among 141 disordered patterns 17 occur (with precise coincidence) only in the PDB but are very sparse in 123 proteomes (see Dataset S5). Such patterns as RASQPELAPEDPED, SMTGGQQMGRGS, SHHHHHHSQDP, PTENLYFQGAM, HHHHHHSSGLEVLFQGP, EQKLISEEDLN, and ASMTGGQMGR do not appear in the analyzed proteomes even in two cases (precise coincidence and exact coincidence of two terminal residues and no coincidence in \(L/5\) positions) (see Figure 6). This suggests that such patterns are an artificial addition to proteins from the CDRDB for their better purifications.

From Figure 6 it is evident that the homorepeats occur very often in eukaryotic proteomes. The patterns with the most frequent occurrence in the eukaryotic proteomes have low complexity: PPPPP, GGGGG, EEEEED, HHHH, KKKKK, SSTSS, and QQQQQP. From Tables 2 and 3 it is evident that the disordered patterns with the most frequent occurrence in the eukaryotic and even in bacterial proteomes are patterns with low complexity GGGGG, PPPPP, TTTPTT, GGGGSGG, KKKKK, etc.

According to work [31] we suggest that these patterns will be disordered in most cases. It should be noted that low-complexity regions can additionally include ordered structural proteins or proteins with strong structural propensity, like collagens, coiled-coils or fibrous proteins [40]. Recently, it has been demonstrated that an increased number of perfect tandem repeats correlates with their stronger tendency to be unstructured [41]. Moreover, strong association between homorepeats and unstructured regions was shown elsewhere [42]. Such patterns as GGGGSGG, EEEEVEE, EDEREE, APIPAP, and PSRSPS (see Table 2) often occur in the considered 17 animal proteomes.

It should be noted that poly H fragments are artificial parts of proteins in the PDB which have been added for better purification of proteins, but in eukaryotic proteomes such a repeat is likely to have a biological function. The locations of poly-H fragments can
be found in different proteomes from our site, http://bioinfo.protres.ru/fp/search_new_pattern.html.

We calculated the statistical significance of the observed patterns in 123 proteomes by using equation (10) (see Materials and Methods). It should be noted that the average length of proteins in considered proteomes is larger (about 400 residues) than the average length of the protein in the PDB database (about 260 residues). On the one hand, $Z_{\text{occur}} \leq 0$, varies from 40 patterns for the human proteome to 91 ones for the bacterial proteome $B. xenovorans$. On the other hand, $Z_{\text{occur}} > 5$ varies from 65 patterns in the rice proteome ($O. sativa$) to 8 patterns in the bacterial proteome $B. xenovorans$. Several examples deserve our attention. For instance, the appearance of pattern GGSGGGGSGGG varies from 7 cases in $T. spiralis$ (the expected occurrence is 0.0004) to 149 cases in humans (the expected occurrence is 0.013), but the $Z_{\text{occur}}$ value is 353 and 1291, respectively. Such patterns as MSLN and SNAM appear more sparsely in comparison with the expected value ($Z_{\text{occur}}$,$0$) for all considered 17 animal proteomes. Although the first pattern occurs 100 times (that is not rare) in the human proteome, and the second pattern appears 61 times, correspondingly. At the same time pattern HHHH appears more often than expected (from 10 for the human to 4 for the actinia ($N. vectensis$), but $Z$ is 68 and 12, respectively.

We calculated the frequencies of occurrence of 141 disordered patterns in 123 proteomes. To make a statement that the given pattern X occurs more often in the i proteome than in the j one we introduced the scoring function for such difference between occurrences of the pattern in two proteomes (by using equation (11), see Materials and Methods). This scoring function should have a normal distribution according to the central limit theorem. We considered the difference occurrence of 141 patterns in some pairs of proteomes (see Dataset S5) and illustrated here the example for eukaryota and bacteria superkingdoms. It turns out that the appearances of 55 patterns in the two superkingdoms do not differ significantly at the level of $10^{-7}$. The negative value of the scoring function points out that the frequency of appearance of the given pattern is higher in bacteria than in eukaryota superkingdoms. For example pattern APIPAP occurs 1.5 times more frequently in 26 bacterial proteomes than in 97 eukaryotic proteomes ($Z_{\text{diff}} = -20.4$). It should be added, that HHHH and QQQQQP patterns occur in Arthropoda’s proteomes more often than in the Chordata proteomes ($Z_{\text{diff}} = -38.4$ and $-34.7$, correspondingly) (see Table 2 and Dataset S5).

For each proteome we calculated a set of 141 values reflecting the number of proteins containing at least one disordered pattern for each of the 141 patterns from the library. Then considering all possible pairs of proteomes, the correlation coefficients between the 141 values have been calculated resulting in the matrix of correlation coefficients. The correlation coefficient was calculated for each pair of proteomes separately (see Table 4), and then averaging has been done inside each kingdom and phylum (see Table 5). As a rule, the correlation coefficients are higher inside the studied kingdom and phylum than between them.

From Table 4 four clusters can be selected with a high correlation coefficient between the numbers of proteins where all considered patterns appear for all pairs between 17 animal proteomes. The first cluster corresponds to phylum Chordata (7 proteomes), the second corresponds to Arthropoda (5 proteomes), the third to Nematoda (4 proteomes), and the fourth to Cnidaria (only 1 proteome). In Tables 4 and 5, bold formatting is used to show a correlation higher than 75%, normal size of numbers to show the correlation from 50% to 75%, and smaller size of numbers to show the correlation smaller than 50%. From Table 4 it is evident that the number of proteins from the human proteome

![Figure 6. Occurrence of disordered patterns in four proteomes.](doi:10.1371/journal.pone.0027142.g006)
Table 2. Number of proteins with the most frequent occurrence of disordered patterns in 17 considered animal proteomes in the case of incomplete coincidence.

| Number of proteins | 51778 | 18405 | 42120 | 28166 | 15101 | 16000 | 16042 | 11437 | 12455 | 18531 | 23817 | 16271 | 16040 | 24435 |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Proteome           | H. sapiens | B. taurus | M. musculus | R. norvegicus | G. gallus | D. rerio | T. nigroviridis | D. melanogaster | D. pseudoobscura | A. aegypti | A. darlingi | A. gambiae | C. briggsae | C. elegans | L. loa,16271 | T. spiralis | N. vectensis |
| PPPPPP             | 1260       | 467       | 1002       | 739       | 282       | 352       | 581       | 560       | 526       | 276       | 385       | 326       | 318       | 478       | 127       | 130       | 153       |
| GGGGG              | 779        | 339       | 606       | 435       | 198       | 270       | 522       | 718       | 885       | 391       | 1025      | 782       | 201       | 385       | 64        | 115       | 97        |
| GGGGGGG            | 478        | 203       | 364       | 266       | 128       | 148       | 276       | 529       | 579       | 225       | 698       | 513       | 103       | 212       | 46        | 59        | 56        |
| EEEED              | 800        | 320       | 604       | 524       | 235       | 340       | 350       | 159       | 170       | 140       | 182       | 186       | 215       | 182       | 41        | 67        | 71        |
| EEEEEE             | 804        | 279       | 650       | 581       | 189       | 295       | 346       | 114       | 112       | 68        | 130       | 125       | 148       | 144       | 49        | 54        | 55        |
| HHHHH              | 224        | 71        | 201       | 138       | 67        | 164       | 192       | 383       | 416       | 325       | 817       | 409       | 116       | 165       | 62        | 107       | 29        |
| TTTTTT             | 178        | 56        | 160       | 110       | 45        | 102       | 76        | 373       | 412       | 256       | 579       | 283       | 271       | 420       | 232       | 199       | 58        |
| XXXXXX             | 503        | 330       | 423       | 353       | 179       | 564       | 196       | 102       | 88        | 139       | 84        | 91        | 169       | 181       | 78        | 79        | 118       |
| SSTSS              | 353        | 115       | 310       | 218       | 125       | 181       | 237       | 295       | 277       | 218       | 301       | 168       | 238       | 333       | 116       | 82        | 69        |
| QQQQQQ             | 240        | 60        | 195       | 145       | 57        | 78        | 107       | 437       | 528       | 230       | 501       | 239       | 109       | 176       | 56        | 82        | 23        |
| EDEREE             | 467        | 185       | 360       | 354       | 129       | 235       | 242       | 143       | 135       | 97        | 140       | 109       | 138       | 179       | 37        | 39        | 55        |
| NSSSSS             | 182        | 74        | 179       | 136       | 82        | 116       | 102       | 272       | 293       | 181       | 287       | 143       | 73        | 114       | 58        | 99        | 31        |
| PAPP               | 385        | 146       | 298       | 230       | 74        | 112       | 181       | 148       | 144       | 74        | 108       | 81        | 100       | 113       | 23        | 56        | 40        |
| PPAPP              | 397        | 161       | 327       | 260       | 103       | 104       | 176       | 111       | 115       | 56        | 81        | 88        | 91        | 90        | 23        | 22        | 37        |
| APIPAP             | 388        | 150       | 282       | 242       | 59        | 95        | 131       | 122       | 211       | 63        | 82        | 71        | 81        | 127       | 12        | 13        | 26        |
| PSRSPS             | 310        | 113       | 251       | 225       | 79        | 129       | 204       | 98        | 111       | 66        | 72        | 54        | 59        | 79        | 31        | 18        | 45        |
| KGKEKK             | 210        | 131       | 197       | 177       | 100       | 170       | 121       | 57        | 47        | 76        | 50        | 56        | 151       | 209       | 35        | 40        | 97        |
| DODDDED            | 108        | 41        | 120       | 104       | 30        | 144       | 73        | 105       | 172       | 104       | 311       | 102       | 60        | 85        | 131       | 61        | 77        |
| PSPP               | 306        | 113       | 269       | 183       | 74        | 109       | 170       | 95        | 82        | 62        | 80        | 59        | 48        | 61        | 22        | 18        | 65        |
| KKEKK              | 225        | 74        | 189       | 145       | 76        | 147       | 73        | 48        | 52        | 66        | 39        | 46        | 117       | 143       | 35        | 24        | 47        |
Table 3. Average number of proteins with the most frequent occurrence of disordered patterns in 123 considered proteomes in the case of incomplete coincidence.

|                | Metazoa (17) | Viridiplantae (5) | Stramensipiles (1) | Choano-flagellida (1) | Euglenozoa (4) | Alveolata (6) | Amoebozoa (2) | Diplomonadida (3) | Fungi (58) | Acidobacteria (14) | Proteobacteria (8) | Bacteroidetes (2) | Chloroflexi (1) |
|----------------|--------------|------------------|-------------------|----------------------|---------------|--------------|--------------|------------------|-------------|-------------------|-----------------|-----------------|-----------------|
| GGGGG          | 460          | 1176             | 2691              | 75                   | 277           | 41           | 409          | 7                | 121         | 75                | 67              | 73              | 21              | 26              |
| PPPPP          | 468          | 679              | 618               | 133                  | 297           | 18           | 249          | 25               | 171         | 54                | 38              | 70              | 14              | 24              |
| TTTPTT         | 224          | 147              | 115               | 196                  | 275           | 18           | 1826         | 17               | 112         | 10                | 23              | 10              | 8               | 19              |
| GGGGGGG        | 287          | 562              | 1941              | 47                   | 105           | 30           | 233          | 4                | 72          | 49                | 40              | 53              | 9               | 16              |
| KKKKK          | 216          | 152              | 119               | 73                   | 57            | 741          | 196          | 9                | 68          | 0                 | 0               | 4               | 17              | 4               |
| EEEED          | 270          | 233              | 348               | 71                   | 107           | 103          | 280          | 7                | 98          | 1                 | 1               | 3               | 6               | 10              |
| SSTSS          | 214          | 118              | 135               | 96                   | 104           | 28           | 455          | 53               | 106         | 2                 | 7               | 9               | 5               | 18              |
| EEEVEE         | 244          | 187              | 470               | 53                   | 125           | 94           | 299          | 8                | 83          | 0                 | 1               | 4               | 3               | 10              |
| QQQQQP         | 192          | 81               | 142               | 67                   | 101           | 5            | 693          | 11               | 81          | 7                 | 6               | 4               | 3               | 8               |
| EDEREE         | 179          | 140              | 163               | 78                   | 74            | 82           | 167          | 9                | 82          | 1                 | 6               | 7               | 8               | 10              |
| DDDDEDD        | 108          | 206              | 203               | 154                  | 58            | 82           | 308          | 3                | 83          | 0                 | 2               | 10              | 3               | 3               |
| HHHH           | 229          | 168              | 99                | 44                   | 58            | 17           | 292          | 1                | 62          | 4                 | 4               | 5               | 5               | 10              |
| APIPAP         | 127          | 174              | 179               | 116                  | 43            | 3            | 30           | 19               | 63          | 42                | 84              | 88              | 5               | 22              |
| PSRSPS         | 114          | 135              | 90                | 52                   | 80            | 8            | 134          | 21               | 81          | 7                 | 26              | 13              | 1               | 6               |
| NSSSS          | 142          | 90               | 73                | 30                   | 51            | 54           | 534          | 18               | 53          | 2                 | 2               | 4               | 3               | 8               |
| PAPPP          | 136          | 155              | 89                | 59                   | 55            | 4            | 27           | 6                | 53          | 39                | 33              | 44              | 5               | 8               |
| PSPPPP         | 107          | 236              | 133               | 48                   | 85            | 5            | 60           | 8                | 56          | 8                 | 10              | 15              | 0               | 6               |
| PPAPP          | 132          | 134              | 102               | 50                   | 87            | 2            | 21           | 4                | 41          | 17                | 35              | 44              | 4               | 10              |
| KKGEKK         | 113          | 77               | 64                | 26                   | 36            | 238          | 106          | 10               | 39          | 2                 | 1               | 3               | 23              | 7               |
| RGPRGRG        | 89           | 161              | 132               | 55                   | 41            | 11           | 17           | 3                | 44          | 4                 | 21              | 18              | 2               | 10              |

doi:10.1371/journal.pone.0027142.t003
| Phylum     | Proteome | H. sapiens | B. taurus | M. musculus | R. norvegicus | G. gallus | D. rerio | T. nigroviridis | D. melanogaster | D. pseudoobscura | A. aegypti | A. darlingi | A. gambiae | C. briggsaeaelegans | C. elegans | L. loa | T. spiralis | N. vectensis |
|------------|----------|------------|-----------|-------------|---------------|-----------|---------|----------------|----------------|----------------|------------|-------------|------------|----------------------|-----------|-------|-------------|-------------|
| Chordata   | H. sapiens | 98 | 99 | 99 | 97 | 86 | 97 | 73 | 69 | 71 | 57 | 68 | 84 | 80 | 56 | 66 | 83 |
|            | B. taurus | 98 | 98 | 98 | 97 | 91 | 95 | 71 | 67 | 70 | 56 | 67 | 83 | 79 | 54 | 65 | 86 |
|            | M. musculus | 99 | 98 | 99 | 97 | 88 | 97 | 74 | 70 | 73 | 59 | 68 | 86 | 82 | 59 | 69 | 85 |
|            | R. norvegicus | 99 | 98 | 99 | 97 | 89 | 96 | 69 | 65 | 70 | 54 | 64 | 85 | 79 | 57 | 66 | 84 |
|            | G. gallus | 97 | 97 | 97 | 97 | 92 | 95 | 72 | 68 | 75 | 58 | 68 | 88 | 83 | 60 | 70 | 88 |
|            | D. rerio | 86 | 91 | 88 | 89 | 92 | 83 | 62 | 58 | 70 | 53 | 59 | 84 | 76 | 63 | 70 | 88 |
|            | T. nigroviridis | 97 | 95 | 97 | 96 | 95 | 83 | 80 | 77 | 78 | 67 | 77 | 84 | 82 | 57 | 69 | 83 |
| Anthropoda | D. melanogaster | 73 | 71 | 74 | 69 | 72 | 62 | 80 | 99 | 95 | 94 | 96 | 79 | 87 | 69 | 83 | 67 |
|            | D. pseudoobscura | 69 | 67 | 70 | 65 | 68 | 58 | 77 | 99 | 94 | 96 | 96 | 74 | 83 | 66 | 80 | 62 |
|            | A. aegypti | 71 | 70 | 73 | 70 | 75 | 70 | 78 | 95 | 94 | 94 | 92 | 85 | 90 | 78 | 90 | 73 |
|            | A. darlingi | 57 | 56 | 59 | 54 | 58 | 53 | 67 | 94 | 96 | 94 | 97 | 68 | 77 | 67 | 80 | 54 |
|            | A. gambiae | 68 | 67 | 68 | 64 | 68 | 59 | 77 | 96 | 96 | 92 | 97 | 71 | 79 | 60 | 75 | 61 |
| Nematoda   | C. briggsae | 84 | 83 | 86 | 85 | 88 | 84 | 84 | 79 | 74 | 85 | 68 | 71 | 97 | 84 | 88 | 86 |
|            | C. elegans | 80 | 79 | 82 | 79 | 83 | 76 | 82 | 87 | 83 | 90 | 77 | 79 | 97 | 85 | 90 | 84 |
|            | L. loa | 56 | 54 | 59 | 57 | 60 | 63 | 57 | 69 | 66 | 78 | 67 | 60 | 84 | 85 | 91 | 74 |
| Cnidaria   | N. vectensis | 83 | 86 | 85 | 84 | 88 | 88 | 83 | 67 | 62 | 73 | 54 | 61 | 86 | 84 | 74 | 74 |

Table 4. Correlation coefficients (in percent) between 17 animal proteomes (kingdom Metazoa).

doi:10.1371/journal.pone.0027142.t004
Table 5. Averaged correlation coefficients (in percent) between numbers of proteins where at least once a disordered pattern for each of 141 patterns appears in 9 kingdoms of eukaryota and 5 phyla of bacteria.

|                  | Metazoa (17) | Viridiplantae (5) | Stramenopiles (1) | Choano- | Eugleno- | Alveola- | Amoebo- | Diplomonadida (3) | Fungi (58) | Acidobacteria (1) | Actinobacteria (14) | Proteobacteria (8) | Bacteroidetes (2) | Chloroflexi (1) |
|------------------|--------------|------------------|-------------------|----------|----------|----------|----------|-------------------|------------|-------------------|--------------------|-------------------|-----------------|---------------|
| Metazoa (17)     | 78           | 71               | 59                | 70       | 75       | 32       | 55       | 47                | 75         | 55                | 43                 | 48                | 52              | 69            |
| Viridiplantae (5)| 71           | 77               | 73                | 67       | 73       | 20       | 43       | 32                | 66         | 63                | 51                 | 55                | 49              | 64            |
| Stramenopiles (1)| 59           | 73               | —                 | 34       | 60       | 7        | 25       | 12                | 42         | 73                | 52                 | 54                | 43              | 55            |
| Choano-flagellida (1) | 70       | 67               | —                 | —        | 74       | 22       | 71       | 57                | 78         | 48                | 53                 | 53                | 42              | 72            |
| Euglenozoa (4)   | 75           | 73               | 60                | 74       | 79       | 14       | 55       | 53                | 73         | 65                | 54                 | 57                | 47              | 73            |
| Alveolata (6)    | 32           | 20               | 7                 | 22       | 14       | 96       | 15       | 12                | 29         | —                 | —                  | —                | —               | 4              |
| Amoebozoa (2)    | 55           | 43               | 25                | 71       | 55       | 15       | 99       | 39                | 56         | 17                | 18                 | 12                | 28              | 48            |
| Diplomonadida (3)| 47           | 32               | 12                | 57       | 53       | 12       | 39       | 94                | 58         | 28                | 31                 | 34                | 35              | 60            |
| Fungi (58)       | 75           | 66               | 42                | 78       | 73       | 29       | 56       | 58                | 80         | 44                | 39                 | 43                | 46              | 66            |
| Acidobacteria (1)| 55           | 63               | 73                | 48       | 66       | —        | 17       | 28                | 44         | —                 | 83                 | 84                | 44              | 70            |
| Actinobacteria (14)| 43        | 51               | 52                | 53       | 54       | —        | 18       | 31                | 39         | 83                | 90                 | 85                | 32              | 68            |
| Proteobacteria (8)| 48          | 55               | 54                | 53       | 57       | —        | 12       | 34                | 43         | 84                | 85                 | 84                | 44              | 70            |
| Bacteroidetes (2)| 52           | 49               | 43                | 42       | 47       | 45       | 28       | 35                | 46         | 44                | 32                 | 39                | 64              | 54            |
| Chloroflexi (1)  | 69           | 64               | 55                | 72       | 73       | 6        | 48       | 60                | 66         | 70                | 68                 | 70                | 54              | —             |

doi:10.1371/journal.pone.0027142.t005
correlates with that from chicken and fish lesser than with bovine, rat, and mouse proteomes. At the same time, the correlation between the number of proteins from proteomes from the Chordata phylum is high for such proteomes as C. briggsae and C. elegans. High correlation coefficients also are observed for such pairs as T. spirals for the Arthropoda proteomes, and N. truncus for the Chordata proteomes.

Combining the motif discovery and disorder protein segment identification in the clustered PDB allows us to create the largest library of the disordered patterns. At present the library includes 141 disordered patterns. Such an approach is promising for further studying and understanding the functional role of the obtained patterns in different proteomes. We came to some general conclusions after analysis of 123 proteomes. The disordered patterns appear more often in eukaryotic than in bacterial proteomes. We can conclude that the occurrence of disordered patterns is more monotonous within the same kingdom (phylum) than between kingdoms (phyla). One can suggest that such short similar motifs are responsible for common functions for nonhomologous, unrelated proteins from different organisms.

**Supporting Information**

**Dataset S1** List of 141 disordered patterns. (XLS)

**References**

1. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38(Database issue): D211–222.
2. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37(Database issue): D212–215.
3. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38(Database issue): D161–166.
4. Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27: 527–533.
5. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293: 321–331.
6. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6: 197–206.
7. Linding R, Russell RB, Nedvova V, Gibson TJ (2003) GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 31: 3701–3708.
8. Lobanov MY, Galzhitskaya OV (2011) The Ising model for prediction of disordered residues from protein sequence alone. Phys Biol 8: 035004.
9. 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.
10. Coeytaux K, Popoun A (2005) Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 21: 1891–1900.
11. Galzhitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. Bioinformatics 22: 2949–2950.
12. Galzhitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) Prediction of amyloidogenic and disordered regions in protein chains. PLoS Comput Biol 2: e177.
13. Schlessinger A, Punta M, Rost B (2007) Natively unstructured regions in proteins identified from contact predictions. Bioinformatics 23: 2376–2384.
14. Peng K, Radivoyevich P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7: 208.
15. Obradovic Z, Peng K, Vucetic S, Radivoyevich P, Dunker AK (2005) Employing heterogeneous sequence properties improves prediction of protein disorder. Proteins 61 Suppl 7: 176–182.
16. Obradovic Z, Peng K, Vucetic S, Radivoyevich P, Brown CJ, et al. (2003) Predicting intrinsic disorder from amino acid sequence. Proteins 53 Suppl 6: S66–S72.
17. Ward JJ, Sudhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337: 635–645.
18. Hecker J, Yang Y, Cheng J (2006) Protein disorder prediction at multiple levels of sensitivity and specificity. BMC Genomics 9 Suppl 1: S9.
19. Su CT, Chen CY, Ou YY (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinformatics 7: 319.
20. Yang ZR, Thomson R, McNeil P, Kanouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21: 3369–3376.
21. Sugase K, Dyson HJ, Wright PE (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447: 1021–1025.
22. Bordelon T, Montegudo SK, Pakhomova S, Oldham ML, Newcomer ME (2004) A disorder to order transition accompanies catalysis in retinaldehyde dehydrogenase type II. J Biol Chem 279: 43085–43091.
23. Shoemaker BA, Portman JJ, Wolynes PG (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc Natl Acad Sci U S A 97: 8966–8973.
24. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, et al. (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry 46: 13468–13477.
25. Mohan A, Oldfield CJ, Radivoyevich P, Vacic V, Cortese MS, et al. (2006) Analysis of molecular recognition features (MoRFS). J Mol Biol 362: 1043–1059.
26. Lobanov MYu, Shoemaker BA, Garbuzynskiy SO, Feng GH, Panchenko AR, et al. (2010) ComSini: Database of protein structures in bound (Complex) and unbound (Single) states in relation to their intrinsic disorder. Nucleic Acids Res 38: D293–D307.
27. Rajasekaran S, Merlin JC, Kundi V, Mi T, Oomen M, et al. (2011) A computational tool for identifying mimotifs in protein–protein interactions and improving the accuracy of mimotif predictions. Proteins 79: 153–164.
28. Gould CM, Dietz F, Via A, Puntervoll P, Gemmell G, et al. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res. D167–D180.
29. Karlin S, Burge C (1996) Trinucleotide repeats and long homopolymeres in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci U S A 93: 1560–1565.
30. Lise S, Jones DT (2005) Sequence patterns associated with disordered regions in proteins. Proteins 58: 144–150.
31. Lobanov MY, Furletova EI, Bogatyreva NS, Roytberg MA, Galzhitskaya OV (2010) Library of disordered patterns in 3D protein structures. PLoS Comput Biol 6: e1000958.
32. Lobanov MY, Garbuzynskiy SO, Galzhitskaya OV (2010) Statistical analysis of unstructured amino-acid residues in protein structures. Biochemistry (Moscow) 75: 236–246.
33. Zhitluk JP, Chua GH, Krishnan A, Bossa C, Colafranceschi M, et al. (2006) Entropic criteria for protein folding derived from recurrences: six residues patch as the basic protein word. FEBS Lett 580: 4861–4864.
34. Galzhitskaya OV (2008) Search for folding induction sites from amino acid sequence. J Bioinform Comput Biol 6: 681–691.
35. Nikiforovich GV, Frieden C (2002) The search for local native-like nucleation centers in the unfolded state of beta-sheet proteins. Proc Natl Acad Sci USA 99: 10239–10239.
36. Hemmer B, Kondo T, Gran B, Pinilla C, Cortese I, et al. (2000) Minimal peptide length requirements for CD4(+) T cell clones—implications for molecular mimicry and T cell survival. Int. Immunol 12: 375–383.
37. Bogatyreva NS, Finkelstein AV, Galzitskaya OV (2006) Trend of amino acid composition of proteins of different taxa. J Bioinform Comput Biol 4: 597–608.
38. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ (2000) Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform 11: 161–171.
39. Abschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
40. Dosaínyi Z, Mézáros B, Simon I (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform 11: 225–243.
41. Jordà J, Xue R, Uversky VN, Kajava AV (2010) Protein tandem repeats - the more perfect, the less structured. FEBS J 277: 2673–2682.
42. Simon M, Hancock JM (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol 10: R59.