Published online 22 October 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D273–D278***
doi:10.1093/nar/gkt927

HRaP: database of occurrence of HomoRepeats and patterns in proteomes

Mikhail Yu. Lobanov, Igor V. Sokolovskiy and Oxana V. Galzitskaya*

Group of Bioinformatics, Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russia

Received August 15, 2013; Revised and Accepted September 20, 2013

ABSTRACT

We focus our attention on multiple repeats of one amino acid (homorepeats) and create a new database (named HRaP, at http://bioinfo.protres.ru/hrap/) of occurrence of homorepeats and disordered patterns in different proteomes. HRaP is aimed at understanding the amino acid tandem repeat function in different proteomes. Therefore, the database includes 122 proteomes, 97 eukaryotic and 25 bacterial ones that can be divided into 9 kingdoms and 5 phyla of bacteria. The database includes 1 449 561 protein sequences and 771 786 sequences of proteins with GO annotations. We have determined homorepeats and patterns that are associated with some function. Through our web server, the user can do the following: (i) search for proteins with the given homorepeat in 122 proteomes, including GO annotation for these proteins; (ii) search for proteins with the given disordered pattern from the library of disordered patterns constructed on the clustered Protein Data Bank in 122 proteomes, including GO annotations for these proteins; (iii) analyze lengths of homorepeats in different proteomes; (iv) investigate disordered regions in the chosen proteins in 122 proteomes; (v) study the coupling of different homorepeats in one protein; (vi) determine longest runs for each amino acid inside each proteome; and (vii) download the full list of proteins with the given length of a homorepeat.

INTRODUCTION

It was found that motifs with low complexity occurred in eukaryotic proteomes (including the human one) more frequently than other protein motifs (1–3). One such motif is a homorepeat, which is the region with repeating of a single amino acid. It turned out that homorepeats play important roles in some biological processes (1,2,4,5). Homorepeats of some amino acids occur more frequently than homorepeats of other amino acids, and the type of homorepeats varies in different proteomes (3). For example, EEEEEE appears to be most frequent for Chordata, QQQQQQ for Arthropoda and SSSSSS for Nematoda (3). One can suggest that such homorepeats may be molecular recognition elements for proteins. A growing number of studies suggest that homorepeats may have a broader role in human diseases than was previously recognized (6). It should be stressed that expansion of homorepeats is a molecular basis for at least 18 human neurological diseases. For example, expansion of poly-A in polyadenine-binding protein 2 is associated with oculopharyngeal muscular dystrophy (7). Long poly-A tracts cause several human developmental diseases (5,8,9). Expansion of poly-Q (larger than 36 residues long) in the Huntington gene results in Huntington's disease. Moreover, poly-Q tracts are associated with several ataxias (8,10). Therefore, perceiving the functional role of these patterns, homorepeats in particular, in the proteomes is a formidable challenge.

With active studying of disordered regions and their functioning, we focus our attention on multiple long repeats of one amino acid (homorepeats) (see Figure 1). The longest uninterrupted runs in the *Dictyostelium discoideum* proteome are of 306 residues for serine, 79 for glutamine, 90 for asparagine and 55 for glutamic acid. The longest uninterrupted runs in the human proteome are of 58 residues for serine, 74 for glutamine, 58 for aspartic acid and 53 for lysine. It is just the time to make a more careful analysis of occurrence, evolution and conservation of these repeats to find their functions. It is still unknown why genetically unstable homorepeats have been preserved throughout evolution, but now it is very important to perform evolution searching of occurrences of homorepeats in different classes. Recently the functional determination of some such motifs has been done. For example, histidine repeats in the protein kinase DYRK1A (length of 13) and in the protein FAM76B (length of 10) mediate nuclear speckle trafficking (5,11,12). Poly-A tract in the HOXD13 protein (length of 15) is important in limb development (5). It has been

*To whom correspondence should be addressed. Tel: +7 4967 318275; Fax: +7 4967 318435; Email: ogalzit@vega.protres.ru

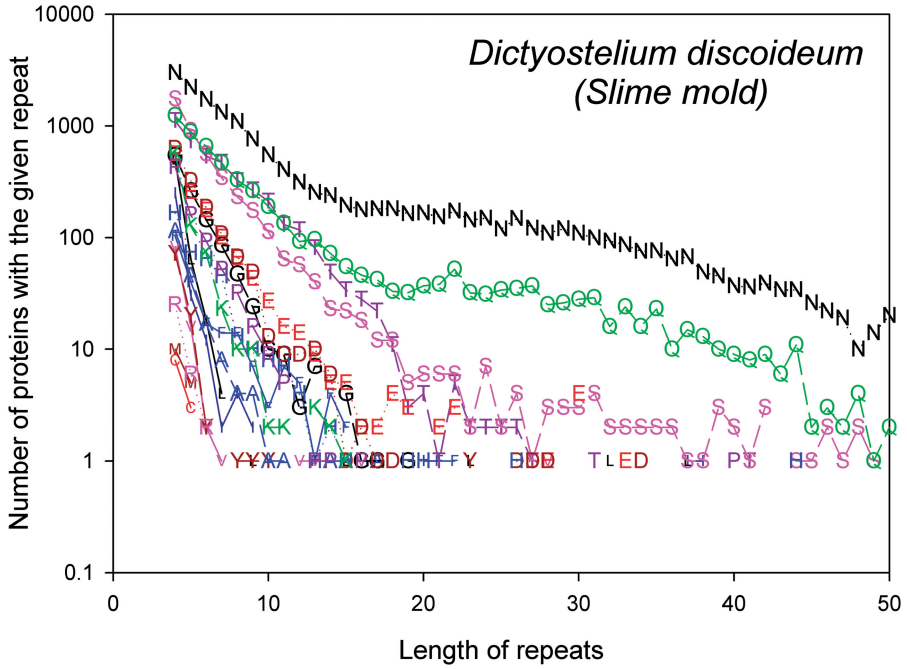


Figure 1. Dependence of the number of proteins that contain homorepeats of different lengths for 20 amino acids in *D. discoideum* proteome.

predicted that the most parts of the homorepeats are disordered (3,13). It should be noted that homorepeats such as KKKKK, PPPPP and HHHH are included in the library of disordered patterns (14). It is worth mentioning that in living organisms, homopeptides can be of non-ribosomal origin as well (2). Comparative analysis of amino acid repeats in some proteomes has been done (2,9,15,16). To gain a clear insight into the abundance of homorepeats and disordered patterns, we have created a database of occurrence of homorepeats with different lengths and disordered patterns (HRaP) in 122 eukaryotic and bacterial proteomes. Our database includes 1 449 561 protein sequences from 122 proteomes, 771 786 sequences of proteins with GO annotations (17) and all homorepeats and 412 disordered patterns from three sets (14,18,19).

DESCRIPTION OF THE DATABASE

We considered 3617 proteomes from the European Bioinformatics Institute site (ftp://ftp.ebi.ac.uk/pub/databases/SPproteomes/uniprot/proteomes/). Because the disordered patterns with the frequent occurrence in proteomes have low complexity (homorepeats), we performed a preliminary analysis. Figure 2 shows the dependence of the number of proteins with at least one occurrence of homorepeats of 6 and more residues long on the size of proteomes. One can see the weak dependence of the occurrence of homorepeats on the size of proteomes. The general result following from this analysis is that the homorepeats appear more often in the eukaryotic proteome than in other proteomes (bacterial, archaeal and viral ones). From Figure 2, one can also see that the number of proteins with at least one occurrence of homorepeats of 6 residues long is <100 for proteomes with an overall number of residues <2 500 000. The data gave grounds for our research involving only proteomes with an overall number of residues exceeding 2 500 000 (19). We obtained 122 proteomes taking into account the length of proteomes representing nine kingdoms of eukaryotes and five phyla of bacteria (see Table in HRaP: proteomes). In view of these proteomes, we have 1 449 561 protein sequences. It should be mentioned that the possible use of this database (named HRaP) is not restricted only to the tasks connected with investigations of disordered regions in proteins and proteomes. Disordered regions can be calculated by using our programs IsUnstruct (14,20) and FoldUnfold (21). It should be noted that recently the new published methods for the prediction of disordered regions are usually meta-servers that combine multiple disorder predictors, e.g. MD (22), PONDR-FIT (23) and MFDp (24). There are separate methods for predicting short [≤15 residues in the program PONDR VSL2 (25)] and long disordered residues [≥30 residues PONDR VSL1 (26)]. Our method IsUnstruct demonstrates a high accuracy in predicting both short and long disordered regions.

HRaP can be used to analyze evolution differences between proteins from different proteomes and connections of these regions with some definite functions. The database includes 771 786 of proteins with GO annotations. It has been found that leucine repeats were especially abundant in the 'Receptor and/or Membrane' group, glutamine and alanine repeats in 'Transcription Factor and/or Development' group, and lysine repeats in the 'Metabolism' group (2,5).

To see the occurrence of a homorepeat, at the first step the user should choose a proteome among 122 considered ones, and then at the second step choose the investigated

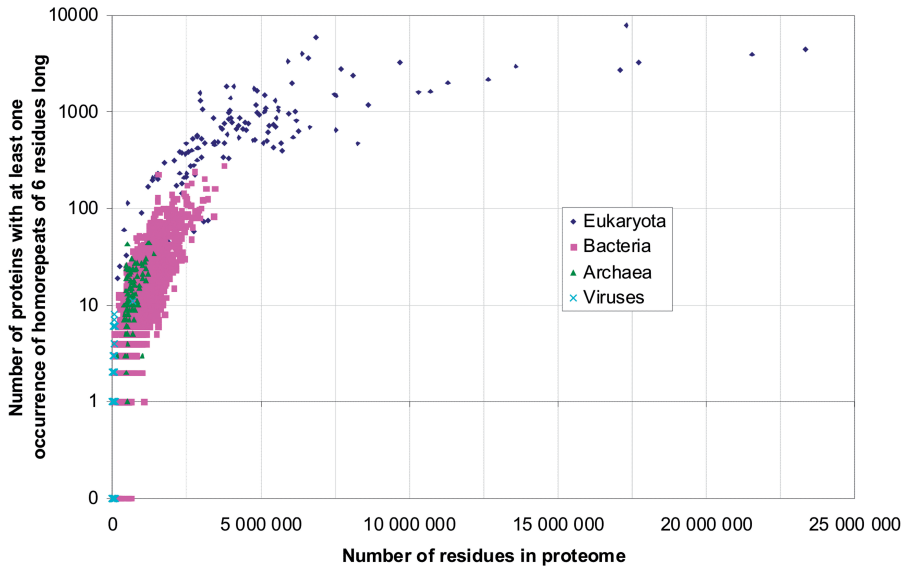


Figure 2. Dependence the number of proteins with at least one occurrence of homorepeats of ≥6 residues long in 3617 proteomes on the size of proteomes.

homorepeat with the given length or pattern (see Figure 3). It should be noted that the order of amino acids and patterns is not random. As concerns the first, the order groups similar amino acids together. From the table of occurrence of homorepeats for different lengths, one can see that long homorepeats appear more often for polar and charge amino acids. The patterns have been ordered according to their significance for prediction of disordered regions. These numbers have been assigned in the corresponding articles (14,18,19). After that, the list of proteins with the given homorepeat or pattern appears with GO annotations (if such is determined). Usually, long proteins contain a homorepeat or several different homorepeats. If several homorepeats and patterns exist in a protein, then all these regions will be marked by different colors in the sequence. In the section HomoRepeats and Patterns, the user can find the occurrence of homorepeats with different lengths and disordered patterns for all 122 proteomes. The largest fraction of homorepeats of six and more residues long belongs to Amoebozoa proteomes (*D. discoideum*), 46% (see Figure 4). The longest uninterrupted runs in *D. discoideum* proteome are of 306 residues for serine, 79 for glutamine, 90 for asparagine and 55 for glutamic acid. The most frequent amino acid runs in the 122 proteomes occur for poly-Q ($6 \leq$ the length of tract ≤ 15), poly-S, poly-A, poly-G, poly-N, poly-P and poly-E (in decreasing order). The acidic runs poly-E and poly-D exceed the runs poly-Q and poly-N for tracts with a short length until 5. The relationship is changed for the long tracts. The occurrence of basic runs poly-K exceeds the runs poly-R, and poly-S exceeds the runs poly-T for all lengths of homorepeats.

Homorepeats and patterns associated with the function

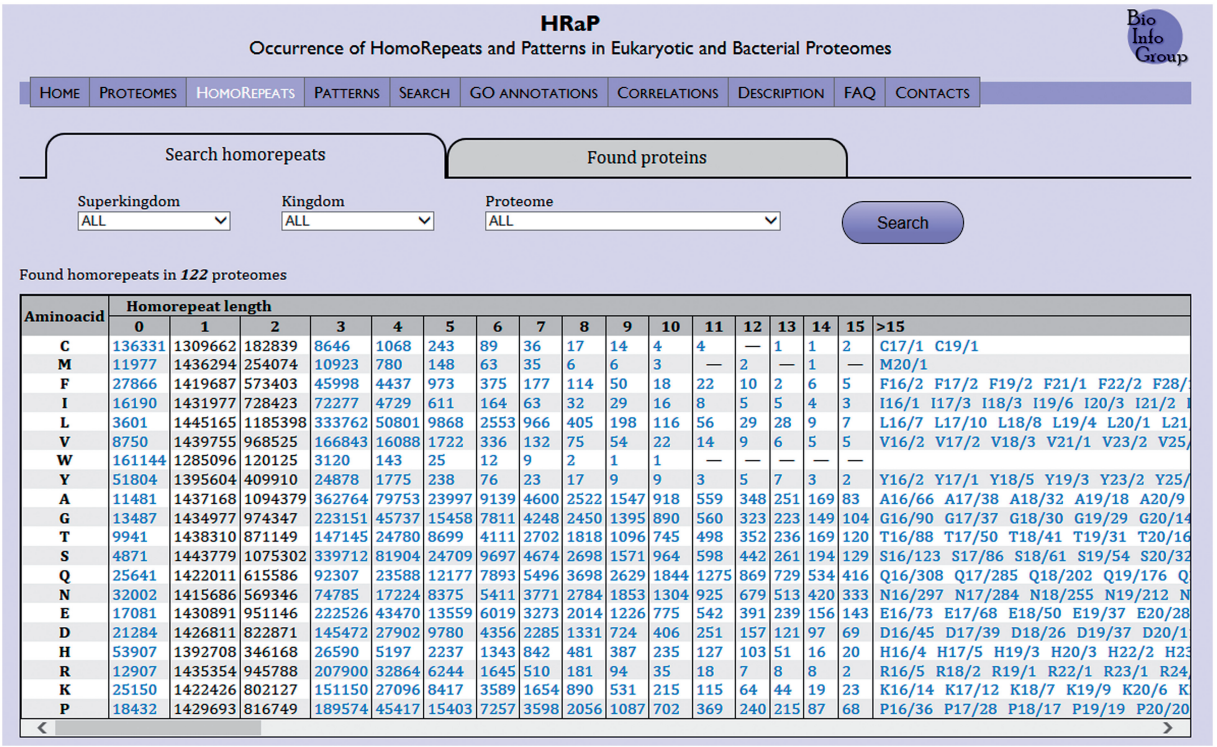
We can suggest that homorepeats and patterns are responsible for common functions of nonhomologous, unrelated proteins from different organisms. To confirm this, we have done the following analysis. All possible GO annotations for proteins were taken for the set of 122 proteomes. The number of different kinds of all annotations is 11 313. Proteins without annotations were combined into the class «absent annotation». The number of proteins including at least one pattern from the last version of the library [171 patterns, set 2012 (14)] was calculated, «N_{pt}». The number of proteins including homorepeats of length ≥6 residues long was calculated, «N_{hm}» as well. The number of proteins with the given annotation was also calculated and indicated in the column «N_{go}». For example, we found 60 proteins with GO annotations of functional kind (F) as ATP binding and including the pattern IKSHHNVGGLP. The same pattern is associated also with the guanosine monophosphate (GMP) synthase (glutamine-hydrolyzing) activity and GMP biosynthetic process. For each pattern or homorepeat, we can calculate the frequency of occurrence in all proteins:

$$w = \frac{N_{pt}}{N_{all}}$$

where N_{pt} is the number of proteins with the given homorepeat or with the given pattern, $N_{all} = 1\,449\,561$ is the full number of proteins in 122 proteomes and the total number of GO-annotated proteins is 771 786. The number of proteins with the given homorepeat (or pattern) and annotation ($N_{pt,go}$) is given in the Table (section GO annotations). The probability to find the number of proteins $N_{pt,go}$ and larger among all proteins with the given annotation was calculated as:

$$p_z = \sum_{i=N_{pt,go}}^{N_{go}} \frac{N_{go}!}{i! \cdot (N_{go} - i)!} \cdot w^i \cdot (1 - w)^{N_{go} - i}$$

Taking into account 171 patterns, 20 homorepeats and 11 313 kinds of GO annotations, we have



HRaP

Occurrence of HomoRepeats and Patterns in Eukaryotic and Bacterial Proteomes

Bio Info Group

HOME | PROTEOMES | HOMOREPEATS | PATTERNS | SEARCH | GO ANNOTATIONS | CORRELATIONS | DESCRIPTION | FAQ | CONTACTS

Search homorepeats | Found proteins

Superkingdom: ALL | Kingdom: ALL | Proteome: ALL | Search

Found homorepeats in *122* proteomes

Aminoacid	Homorepeat length																
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	>15
C	136331	1309662	182839	8646	1068	243	89	36	17	14	4	4	—	1	1	2	C17/1 C19/1
M	11977	1436294	254074	10923	780	148	63	35	6	6	3	—	2	—	1	—	M20/1
F	27866	1419687	573403	45998	4437	973	375	177	114	50	18	22	10	2	6	5	F16/2 F17/2 F19/2 F21/1 F22/2 F28/
I	16190	1431977	728423	72277	4729	611	164	63	32	29	16	8	5	5	4	3	I16/1 I17/3 I18/3 I19/6 I20/3 I21/2 I
L	3601	1445165	1185398	333762	50801	9868	2553	966	405	198	116	56	29	28	9	7	L16/7 L17/10 L18/8 L19/4 L20/1 L21
V	8750	1439755	968525	166843	16088	1722	336	132	75	54	22	14	9	6	5	5	V16/2 V17/2 V18/3 V21/1 V23/2 V25/
W	161144	1285096	120125	3120	143	25	12	9	2	1	1	—	—	—	—	—	
Y	51804	1395604	409910	24878	1775	238	76	23	17	9	9	3	5	7	3	2	Y16/2 Y17/1 Y18/5 Y19/3 Y23/2 Y25/
A	11481	1437168	1094379	362764	79753	23997	9139	4600	2522	1547	918	559	348	251	169	83	A16/66 A17/38 A18/32 A19/18 A20/9
G	13487	1434977	974347	223151	45737	15458	7811	4248	2450	1395	890	560	323	223	149	104	G16/90 G17/37 G18/30 G19/29 G20/14
T	9941	1438310	871149	147145	24780	8699	4111	2702	1818	1096	745	498	352	236	169	120	T16/88 T17/50 T18/41 T19/31 T20/16
S	4871	1443779	1075302	339712	81904	24709	9697	4674	2698	1571	964	598	442	261	194	129	S16/123 S17/86 S18/61 S19/54 S20/32
Q	25641	1422011	615586	92307	23588	12177	7893	5496	3698	2629	1844	1275	869	729	534	416	Q16/308 Q17/285 Q18/202 Q19/176 Q
N	32002	1415686	569346	74785	17224	8375	5411	3771	2784	1853	1304	925	679	513	420	333	N16/297 N17/284 N18/255 N19/212 N
E	17081	1430891	951146	222526	43470	13559	6019	3273	2014	1226	775	542	391	239	156	143	E16/73 E17/68 E18/50 E19/37 E20/28
D	21284	1426811	822871	145472	27902	9780	4356	2285	1331	724	406	251	157	121	97	69	D16/45 D17/39 D18/26 D19/37 D20/1
H	53907	1392708	346168	26590	5197	2237	1343	842	481	387	235	127	103	51	16	20	H16/4 H17/5 H19/3 H20/3 H22/2 H23
R	12907	1435354	945788	207900	32864	6244	1645	510	181	94	35	18	7	8	8	2	R16/5 R18/2 R19/1 R22/1 R23/1 R24/
K	25150	1422426	802127	151150	27096	8417	3589	1654	890	531	215	115	64	44	19	23	K16/14 K17/12 K18/7 K19/9 K20/6 K
P	18432	1429693	816749	189574	45417	15403	7257	3598	2056	1087	702	369	240	215	87	68	P16/36 P17/28 P18/17 P19/19 P20/20

Figure 3. A screenshot of HRaP results filtered for HomoRepeats of the all 122 proteomes.

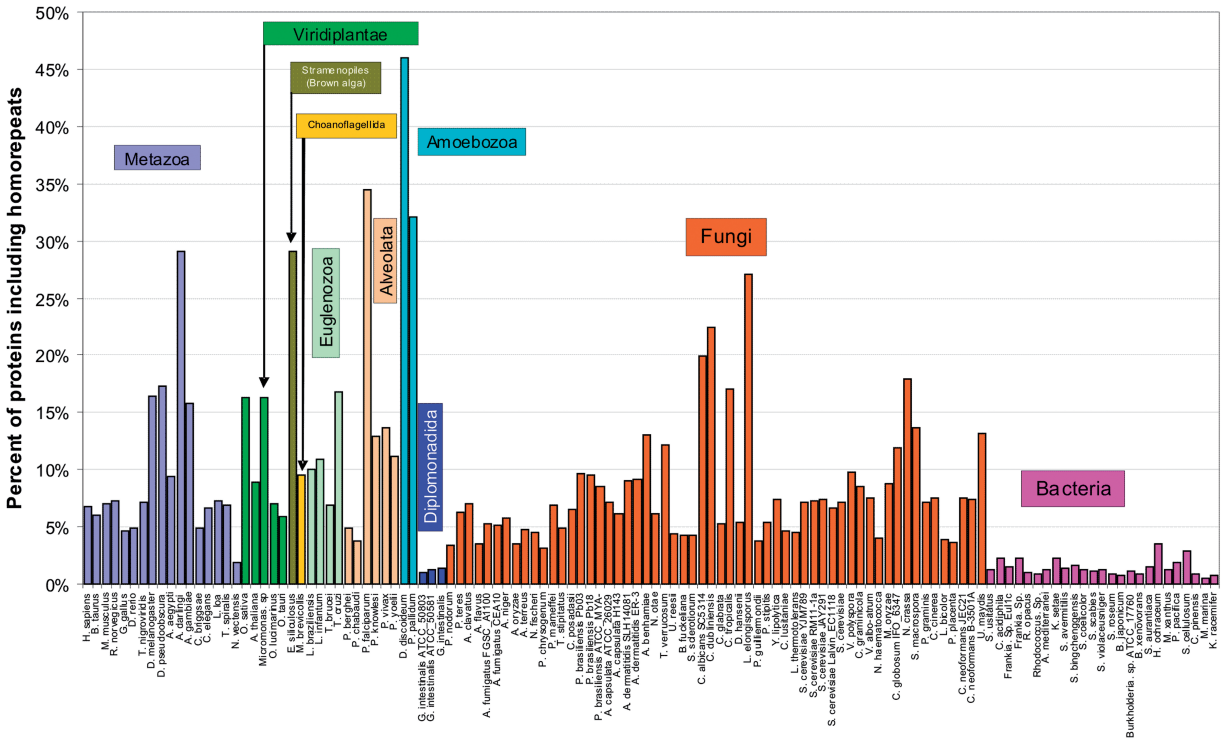


Figure 4. The percentage of proteins with at least one occurrence of homorepeats of ≥6 residues long in 122 proteomes.

$11\,313*(171+20) = 2\,160\,783 \approx 2{\cdot}10^6$ possible combinations. Therefore, we should not pay attention to the events in which the probability is higher than 10^{-7}. Taking this into account, the probabilities p_z were colored according to the following conditions: green corresponds to $p_z < 10^{-15}$, light green corresponds to $10^{-15} \le p_z < 10^{-10}$ and yellow corresponds to $10^{-10} \le p_z < 10^{-7}$.

We also calculated the probabilities:

$$p_1 = \frac{N_{hm,go}}{N_{hm}} \text{ and } p_2 = \frac{N_{hm,go}}{N_{go}}.$$

The patterns and homorepeats are sorted by p_1 and p_2 using the following colors: green–$p_1 > 0.5$, light green–$0.3 < p_1 < 0.5$ and light yellow–$0.1 < p_1 < 0.3$. The patterns and homorepeats associated with the functions are presented in section GO annotations. It is interesting to note that histidine, alanine, glutamine and glutamine acid repeats are connected with GO annotation 'C: nucleus'. As has been mentioned in the Introduction, histidine repeats mediate nuclear speckle trafficking in several transcription factors (5,11,12). The methionine repeat is connected with the voltage-gated calcium channel activity. Proline homorepeats are associated with many GO annotations: dendrite self-avoidance, central nervous system morphogenesis, bacterial cell surface binding, axon guidance receptor activity, axon extension involved in axon guidance, actin polymerization or depolymerization, Rho GTPase binding, mushroom body development, actin cortical patch, axonal fasciculation, actin cytoskeleton organization, peripheral nervous system development, cell morphogenesis, tropomyosin binding and stereocilium. Also, it should be noted that not all amino acid repeats are associated with some functions.

Among 109 disordered patterns (set 2010), 8 occur (with precise coincidence) only in the Protein Data Bank but are absent in 122 proteomes. Among 141 patterns (set 2011), there are only 6 such patterns, and 8 among 171 patterns (set 2012). Such patterns as TTTATT and NNNNN (from set 2012) occur > 17 000 times in the considered 122 proteomes. The leader is QQQQQQQ, which occurs >20 000 times. Moreover, the pattern NNNNN is connected with such process as symbiosis, encompassing mutualism through parasitism. This pattern occurs very seldom in the human proteome, only in 21 proteins.

We have created the list of human proteins with homorepeats that are associated with disease. The list can be found in the frequently asked questions section. Also, the list of proteins with homorepeats of 6 and more residues long from the clustered Protein Data Bank (14) can be found in the frequently asked questions section.

Correlations between number of proteins with homorepeats or patterns in any proteome

For each proteome, we calculated a set of 109 values reflecting the number of proteins containing at least one disordered pattern for each of the 109 patterns from the library. Then considering all possible pairs of proteomes, the correlation coefficients between the 109 values have been calculated resulting in the matrix of correlation coefficients. The correlation coefficient was calculated for each pair of proteomes separately, and then averaging has been done inside each kingdom and phylum (see Correlations section). Similar values have been calculated for a set of 141 disordered patterns, 171 disordered patterns and 20 homorepeats. A comparative analysis of the number of proteins containing homorepeats of 6 and more residues long in 122 proteomes has demonstrated that the correlation coefficients between numbers of proteins, where at least once a homorepeat of 6 and more residues long for each of the 20 types of amino acid residues appears in 9 kingdoms of eukaryota and 5 phyla of bacteria, are higher inside the considered kingdom than between them (3). The same result is valid for the 109 disordered selected patterns (set 2010) (18), the 141 disordered selected patterns (set 2011) (19) and the 171 disordered selected patterns (set 2012) (14).

CONCLUSIONS AND FUTURE DIRECTIONS

We have collected an exhaustive database of occurrence of homorepeats and patterns in 122 proteomes with the number of residues larger than 2 500 000 in each proteome. The found patterns and homorepeats associated with the function point to the tremendous importance of homorepeats in a large variety of cellular processes and merit further studying. In future work, we are planning to include the analysis of coupling between occurrences of different homorepeats in one protein and to make clusterization of proteins to escape the influence of homologous proteins for determination of homorepeat functions. We will be grateful for any contribution to the database from the community.

ACKNOWLEDGEMENTS

The authors thank O.I. Sokolovskaya for assistance in programming.

FUNDING

Russian Foundation for Basic Research [11-04-00763]; Russian Academy of Sciences (programs 'Molecular and Cell Biology' [01201353567] and 'Fundamental Sciences to Medicine'). Funding for open access charge: Russian Academy of Sciences program 'Molecular and Cell Biology' [01201353567].

Conflict of interest statement. None declared.

REFERENCES

1. Tompa,P. (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays*, **25**, 847–855.
2. Jorda,J. and Kajava,A.V. (2010) Protein homorepeats sequences, structures, evolution, and functions. *Adv. Protein Chem. Struct. Biol.*, **79**, 59–88.
3. Lobanov,M.Y. and Galzitskaya,O.V. (2012) Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. *Mol. BioSyst.*, **8**, 327–337.
4. Karlin,S. and Burge,C. (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl Acad. Sci. USA*, **93**, 1560–1565.
5. Mularoni,L., Ledda,A., Toll-Riera,M. and Albà,M.M. (2010) Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res.*, **20**, 745–754.
6. Siwach,P. and Ganesh,S. (2008) Tandem repeats in human disorders: mechanisms and evolution. *Front. Biosci.*, **13**, 4467–4484.

7. Brais,B., Bouchard,J.P., Xie,Y.G., Rochefort,D.L., Chretien,N., Tome,F.M., Lafrenière,R.G., Rommens,J.M., Uyama,E., Nohira,O. *et al.* (1998) Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.*, **18**, 164–167.
8. Brown,L.Y. and Brown,S.A. (2004) Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet.*, **20**, 51–58.
9. Mularoni,L., Veitia,R.A. and Alba',M.M. (2007) Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics*, **89**, 316–325.
10. Gatchel,J.R. and Zoghbi,H.Y. (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.*, **6**, 743–755.
11. Alvarez,M., Estivill,X. and de la Luna,S. (2003) DYRK1A accumulates in splicing speckles through a novel targeting signal and induces speckle disassembly. *J. Cell Sci.*, **116**, 3099–3107.
12. Salichs,E., Ledda,A., Mularoni,L., Alba',M.M. and de la Luna,S. (2009) Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet.*, **5**, e1000397.
13. Jorda,J., Xue,B., Uversky,V.N. and Kajava,A.V. (2010) Protein tandem repeats - the more perfect, the less structured. *FEBS J.*, **277**, 2673–2682.
14. Lobanov,M.Y., Sokolovskiy,I.V. and Galzitskaya,O.V. (2013) IsUnstruct: prediction of the residue status to be ordered or disordered in the protein chain by a method based on the Ising model. *J. Biomol. Struct. Dyn.*, **31**, 1034–1043.
15. Alba',M.M. and Guigo,R. (2004) Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.*, **14**, 549–554.
16. Dalby,A.R. (2009) A comparative proteomic analysis of the simple amino acid repeat distributions in Plasmodia reveals lineage specific amino acid selection. *PLoS One*, **4**, e6231.
17. Gene Ontology Consortium. (2013) Gene Ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.
18. Lobanov,M.Y., Furletova,E.I., Bogatyreva,N.S., Roytberg,M.A. and Galzitskaya,O.V. (2010) Library of disordered patterns in 3D protein structures. *PLoS Comput. Biol.*, **6**, e1000958.
19. Lobanov,M.Y. and Galzitskaya,O.V. (2011) Disordered patterns in clustered Protein Data Bank and in eukaryotic and bacterial proteomes. *PLoS One*, **6**, e27142.
20. Lobanov,M.Y. and Galzitskaya,O.V. (2011) The Ising model for prediction of disordered residues from protein sequence alone. *Phys. Biol.*, **8**, 035004.
21. Galzitskaya,O.V., Garbuzynskiy,S.O. and Lobanov,M.Y. (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, **22**, 2948–2949.
22. Schlessinger,A., Punta,M., Yachdav,G., Kajan,L. and Rost,B. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **4**, e4433.
23. Xue,B., Dunbrack,R.L., Williams,R.W., Dunker,A.K. and Uversky,V.N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*, **1804**, 996–1010.
24. Mizianty,M.J., Stach,W., Chen,K., Kedarisetti,K.D., Disfani,F.M. and Kurgan,L. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.
25. Peng,K., Radivojac,P., Vucetic,S., Dunker,A.K. and Obradovic,Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.
26. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P. and Dunker,A.K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61(Suppl. 7)**, 176–182.