Protein tandem repeats - the more perfect, the less structured.

Julien Jorda, Bin Xue, Vladimir N Uversky, Andrey V Kajava

Protein tandem repeats: the more perfect the less structured

Julien Jorda[1], Bin Xue[2,3], Vladimir N. Uversky[2-5], Andrey V. Kajava[1,*]

[1]Centre de Recherches de Biochimie Macromoléculaire, CNRS UMR-5237, University of Montpellier 1 and 2, Montpellier, France; [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [3]Institute for Intrinsically Disordered Protein Research, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [4]Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia; [5]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Running title: Structural state of perfect protein repeats

Keywords: bioinformatics, disordered conformation, evolution, sequence analysis, protein structure.

*To whom correspondence should be addressed:

Dr. Andrey V. Kajava

Centre de Recherches de Biochimie Macromoléculaire, CNRS,

1919 Route de Mende, 34293 Montpellier, Cedex 5, France

FAX: +33 4 67 521559 ; Phone number: +33 4 67 61 3364

e-mail : andrey.kajava@crbm.cnrs.fr

ABSTRACT

We analysed structural properties of protein regions containing arrays of perfect and nearly perfect tandem repeats. Naturally occurring proteins with perfect repeats are practically absent among the proteins with known 3D structures. The great majority of such regions in the Protein DataBank (PDB) are found in the *de novo* designed proteins. The abundance of natural structured proteins with tandem repeats is inversely correlated with the repeat perfection: the chance to find natural structured proteins in PDB increases with a decrease in the level of repeat perfection. Prediction of intrinsic disorder within the tandem repeats in the SwissProt proteins supports the conclusion that the level of repeat perfection correlates with their tendency to be unstructured. This correlation is valid across the various species and subcellular localizations, although the level of disordered tandem repeats varies significantly between these datasets. On average, in prokaryotes, tandem repeats of cytoplasmic proteins were predicted to be the most structured, whereas in eukaryotes, the most structured portion of the repeats was found in the membrane proteins. Our study supports the hypothesis that in general, the repeat perfection is a sign of recent evolutionary events rather than of exceptional structural and (or) functional importance of the repeat residues.

INTRODUCTION

Genome sequencing projects are producing knowledge about a large number of protein sequences. The understanding of the biological role of many of these proteins requires information about their 3D structure as well as their evolutionary and functional relationships. At least 14% of all proteins and more than one third of human proteins carrying out fundamental functions contain arrays of Tandem Repeats (TR) [1]. The 3D structures of many of these proteins have already been determined by X-ray crystallography and NMR methods. Fibrous proteins with repeats of 2 to 7 residues (collagen, silk fibroin, keratin, tropomyosin) were the first objects studied by methods of structural biology [2]. Proteins with the repeat length from 5 to 50 residues gained special interests in 90s, when several unusual structural folds, including β-helices [3], β-rolls [4], horse-shoe shaped structure of Leucine-Rich-Repeat proteins [5], β-propellers [6], and α-helical solenoids [7] were resolved by X-ray crystallography. Many proteins with repeats longer than 30 residues have a "beads-on-a-string" organization with each repeat being folded into a globular domain, e.g. Zn-finger domains [8], Ig-domains[9] and the human matrix metalloproteinase [10]. It was noticed that frequently proteins with repeats do not have unique stable 3D structures [11]. Rough estimates propose that half of the regions with TRs may be naturally unfolded [12, 13]. Low-complexity regions of eukaryotic proteins that are enriched in repetitive motifs are rare among the known 3D structures from the Protein Data Bank (PDB) [14]. Common structural features, functions and evolution of proteins with TRs have been summarized in several reviews [7, 11, 15-18].

Perfect TRs occupy a special place among protein repeats which are usually imperfect due to mutations (substitutions, insertions, deletions) accumulated during evolution. The high level of perfection of repeats can mean high structural and functional importance of each

residue in the repeat as it was observed in collagen molecules or some β-roll structures [2, 19]. It can also mean recent evolutionary events that, for example, in pathogens, can allow a rapid response to environmental changes and can thus lead to emerging infection threats and, in higher organisms, can lead to rapid morphological effects [20].

Perfect and nearly perfect repeats occur in a significant portion of proteins. Recently, by using a newly developed algorithm for *ab initio* identification of TRs, we detected this type of repeats in 9% of proteins of SwissProt database [21]. To estimate the level of perfection of the TRs we used a parameter called "P_{sim}" which is based on calculation of Hamming distances between the consensus sequence and aligned repeats of TR region (see Materials and Methods). In this work we analysed perfect and nearly perfect TRs with P_{sim} >=0.7.

Specific structural and evolutionary properties of the perfect repeats pose challenges for annotation of genomic data. First, in contrast to the aperiodic globular proteins, prediction of structure-function by sequence similarity can not be directly applied to the perfect or nearly perfect repeats due to their different evolutionary mechanisms. Second, although *ab initio* structural prediction of proteins with TRs generally yields reliable results [11], very high fidelity of sequence periodicity decreases the accuracy and reliability of the information obtained from the sequence alignment of the repeats. Each position of the perfect repeats is conserved and this hampers distinguishing between residues that form the interior of the structure and those that face the solvent.

TRs are often found in proteins associated with various human diseases. For example, expansion of homorepeats is the molecular cause of at least 18 human neurological diseases, including myotonic dystrophy 1 (DM1), Huntington's disease (HD), Kennedy disease (also known as spinal and bulbar muscular atrophy, SBMA), dentatorubral-pallidoluysian atrophy (DRPLA), and a number of spinocerebellar ataxias (SCAs), such as spinocerebellar ataxia

type 1 (SCA1), spinocerebellar ataxia type 2 (SCA2), Machado-Joseph disease (MJD/SCA3), SCA6, SCA7 and SCA17 [22, 23]. A number of clinical disorders, including prostate cancer, benign prostatic hyperplasia, male infertility and rheumatoid arthritis are associated with polymorphisms in the length of the polyglutamine and polyglycine repeats of the androgen receptor [24].

Thus, proteins with perfect or nearly perfect TRs play important functional roles, are abundant in genomes, are related to major health threats and, at the same time, represent a challenge for *in silico* identification of their structures and functions. Along this line, the objective of our work was a systematic bioinformatics analysis of arrays of perfect or nearly perfect TRs to obtain a global view on their structural properties.

RESULTS AND DISCUSSION

The 3D structures of naturally occurring proteins with perfect repeats are practically absent in PDB

Our analysis shows that among 20800 sequences of non-redundant PDB (95% identity) only 9 naturally occurring proteins (0.04%) have perfect TRs with P_{sim} =1 (Table 1). Furthermore, these arrays of TRs are short (less than 19 residues) and they are missing from the determined structures representing regions with blurred electron density. A common reason for missing electron density is that the unobserved atom, side chain, residue, or region fails to scatter X-rays coherently due to variation in position from one protein to the next, e.g. the unobserved atoms are flexible or disordered. Exceptions are two proteins: (1) an antibody molecule where Gly-rich TR region represents a crosslink between two domains (PDB code 1F3R) [25], and, (2) a substrate with $(Arg\text{-}Ser)_8$ tract that was co-crystallised with protein

kinase (PDB code 3BEG) [26]. This Arg-rich peptide being alone in solution, most probably, will be unstructured due to the absence of non-polar residues and the presence of eight Arg residues carrying charge of the same sign. Thus, this analysis suggested that regions of natural proteins with perfect repeats have tendency to be unstructured.

To retrace this tendency, we analysed further the regions with less perfect TRs. The TRs with $0.9 \leq P_{sim} < 1.0$ are also rare among natural proteins of the PDB. Furthermore, the conformation of almost all of them is not resolved by the X-ray crystallography because they are located in regions with missing electron density. Only one of them, human CD3-e/d dimer (PDB 1XIW) [27] has a short region of two 9 residue repeats corresponding to a loop followed by β-strand. We also analysed TRs with $0.8 \leq P_{sim} < 0.9$ and found already 17 TRs of natural proteins with the 3D structures (Table 1). In addition to relatively short regions of less than 20 residues, corresponding to the α-helical elements, we also found longer regions which form immunoglobulin-like structures (1D2P) [28], β-roll (1GO7) [29]; α-solenoid (2AJA) [30] and unusual long β-hairpin (1JHN) [31] (Fig. 1). Three of these four structures are formed by bacterial proteins.

De novo design proteins with perfect repeats fold into stable 3D structures

In the PDB, majority (80%) of the proteins having perfect TRs are *de novo* designed proteins (Table 1). TR regions of a large portion of these proteins fold into the well-defined repetitive 3D structures such as collagen triple-helices, α-helical coiled coils and α-helical solenoids [2, 17]. The fact that the designed perfect TRs can form the stable 3D structures indicates that the absence of such structures in natural proteins is due to evolutionary reasons and not due to the problems with their folding propensities *per se*.

Prediction of intrinsically disordered regions in SwissProt database supports tendency of TRs to be unfolded

The ability of TRs to be structured or disordered was further tested by using a larger datasets extracted from SwissProt. The analysed dataset of TRs from PRDB (http://bioinfo.montp.cnrs.fr/?r=repeatDB) were filled in by T-REKS program [21]. The TRs with P_{sim} range from 0.7 to 1 are consisted of 51,685 repeats found in 33,151 proteins which represent 9.1% of all proteins in the SwissProt release of January, 2009 (364,403 sequences). The level of intrinsic disorder in these repeats and repeat-containing proteins was evaluated by using several computational tools.

Compositional profiling

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are known to be different from structured globular proteins and domains with regard to many attributes, including amino acid composition, sequence complexity, hydrophobicity, charge, flexibility, and type and rate of amino acid substitutions over evolutionary time. For example, IDPs/IDRs are significantly depleted in a number of so-called order-promoting residues, including bulky hydrophobic (Ile, Leu, and Val) and aromatic amino acid residues (Trp, Tyr, and Phe), which would normally form the hydrophobic core of a folded globular protein, and also possess low content of Cys and Asn residues. On the other hands, IDPs/IDRs were shown to be substantially enriched in so called disorder-promoting amino acids: Ala, Arg, Gly, Gln, Ser, Pro, Glu, and Lys [32-36]. These biases in the amino acid compositions of IDPs and IDRs can be visualized using a normalization procedure known as compositional profiling [32, 33, 37]. In brief, compositional profiling is based on the evaluation of the $(C_{s1} - C_{s2})/C_{s2}$ values, where C_{s1} is a content of a given residue in a set of interest (regions and proteins with TRs), whereas C_{s2} is the corresponding value for the reference dataset (set of ordered proteins

or set of well-characterized IDPs). Negative values of the profiling correspond to residues which are depleted in a given dataset in comparison with a reference dataset, and the positive values correspond to the residues which are over-represented in the set of interest.

Figure 2 compares the amino acid compositions of (i) all TR regions analyzed in this study, (ii) proteins containing these TRs, and (iii) a dataset of IDPs, with the composition of ordered proteins. The datasets of IDPs and fully structured proteins were taken from our previous analysis [38, 39]. It shows that the compositions of proteins containing TRs and TRs themselves are different from compositions of ordered proteins. They follow the trend for IDPs, being generally depleted in major order-promoting residues. This tendency towards disorder is stronger for the TR regions indicating that they contribute to this trend. At the same time, amino acid composition of the TRs has a bias when compared with the composition of "typical" disordered proteins (Fig. 2). TRs have especially low occurrence of order-promoting Met and disorder-promoting charged residues Asp, Glu and Lys. On the other hand, TRs are highly enriched in Cys and disorder-promoting Pro, Gly, Ser and His.

To test the tendency of TRs towards disorder as a function of their level of perfection the TRs were subdivided into four subsets accordingly to their P_{sim} values ($0.7 < P_{sim} \leq 0.8$ (32691 TRs), $0.8 < P_{sim} \leq 0.9$ (8322 TRs), $0.9 < P_{sim} \leq 1.0$ (1471 TRs), and homorepeats with $P_{sim} = 1.0$ (5259 TRs). Homorepeats were analyzed separately from the other TRs because they significantly outnumber the other types of repeats and being in the same group would obscure the effect related to the other repeats. The amino acid compositions of these subsets were compared with the compositions of fully structured proteins. Figure 3 represents the results of compositional profiling for TRs with different level of perfection. Both homorepeats and the other TRs show the same trend. With the increase in the perfection of the repeated segment, the amount of order-promoting residues is gradually reduced, whereas the relative

contents of disorder-promoting polar residues are gradually increased. The content of Gly and Pro residues does not significantly change.

Prediction of intrinsic disorder

Since the compositional profiling showed that TRs and repeat-containing proteins have a noticeable increase in the number of disorder-promoting residues, we further analyzed the abundance of predicted intrinsic disorder in these sequences by several computational tools, including PONDR® VLXT [34, 40] and VSL2 [41, 42] algorithms, as well as predictors such as IUpred [43, 44], FoldIndex [45], and TopIDP [37]. Results of this analysis are summarized in Table 2, which clearly shows that both TRs and repeat-containing proteins are highly disordered. Furthermore, TRs have higher percentage of disordered residues when compared to the entire TR-containing sequences. Prediction of intrinsic disorder also confirmed an observation that the amount of disorder in both datasets increases with an increase in the repeat perfection (Table 2).

This observation is further illustrated by the distributions of values representing the number of predicted disorder residues divided by the number of residues in the considered region (Fig. 4). These distributions are generated for TR regions of different levels of perfection (Fig. 4A) and for the corresponding repeat-containing proteins (Fig. 4B). Figure 4A shows that all analysed TRs are highly disordered irrespective of the level of their perfection. At the same time, as the perfection of TR increases, the relative content of disorder also increases. For example, at least 70% of TRs with $0.7 < P_{sim} \leq 0.8$ are predicted to have disorder ratio of more than 0.95. For TR regions with $0.8 < P_{sim} \leq 0.9$ this percentage increases to 85%, for segments with $0.9 < P_{sim} \leq 1.0$ it is 86% and for perfect homorepeats it reaches 97% (Fig. 4A). Fig. 4B shows that only 6% of the whole sequences of proteins containing perfect repeats are well-structured (disordered ratio less than 0.2). The rest of these

sequences is wide spread within the disordered ratio ranging from 0.25 to 1. Proteins containing the least perfect repeats ($0.7 < P_{sim} \leq 0.8$) are almost evenly distributed among various disorder ratios at the level of about 5%. Thus, perfect repeats preferentially occur in proteins which have the disorder ratio of more than 0.2 and are poorly represented in more structured proteins, while less perfect repeats are equally probable in sequences with different disorder ratios.

Intrinsic disorder of tandem repeats across species and subcellular localizations

PONDR® VLXT predictor and TopIDP index were used to establish variation of the disorder level among TRs of viral, eukaryotic and prokaryotic proteins. The tested dataset included TRs with $P_{sim} \geq 0.9$ identified in SwissProt. The homorepeats were excluded and analyzed separately from the other TRs because their predominant occurrence in eukaryotic proteins would obscure the results. Prior to the analysis, the redundancy of the dataset related to existence of protein sequences from different strains of the same species (especially for bacteria and viruses) had been filtered out by using the species name, consensus motif, number and location of repeats. As a result the dataset contained 245 repeats from prokaryotic proteins, 1059 repeats from eukaryotic proteins and 70 repeats from viral proteins. Our analysis shows that TRs from all species have the tendency to be unstructured (Table 3). At the same time, TRs from eukaryotic proteins have ratio of disordered proteins slightly higher than TRs from viral or prokaryotic proteins.

The ratio of disordered repeats was also investigated as a function of the subcellular localization of corresponding repeat-containing proteins. We performed this analysis separately for homorepeats and the other TRs of SwissProt with $P_{sim} \geq 0.8$. The obtained distribution among cellular compartments were similar in these two datasets, therefore, Table 4 represent the combined results for both types of repeats. The lowest portion of disordered

repeats (54.3%) was found in the cytoplasmic proteins of prokaryotes (Table 4). The ratio increases from cytoplasm to the cellular exterior, being equal to 72.3% and 83.6% in membrane and secreted proteins, respectively. Survey of amino acid sequences of the bacterial cytoplasmic repeats which were predicted to be structured revealed a large number (90 TRs) of $(GGM)_n$ repeats. These repeats are located at the C-terminal extremity of GroEL chaperone and plays important role in refolding of proteins [46]. In the crystal structure of GroEL complex, these C-terminal tails are not resolved and located inside the complex chamber. This suggests that inside of the GroEL complex they are disordered. Such repeats are also found in mitochondria of eukaryotes in HSP60, a eukaryotic homolog of GroEL. The cytoplasmic TRs of prokaryotes with excluded GGM repeats still have the highest percentage of predicted structured regions among the cellular compartments.

In eukaryotes, the ratio of disorder varies differently depending on the cellular localization. The lowest level of TR disorder is found in membrane proteins, followed by secreted and nuclear proteins. The cytoplasmic TRs are the most disordered in eukaryotes (82%). The high percentage of ordered TRs in membrane proteins suggests that they may be a part of transmembrane regions. However, our analysis revealed that only 12% of them predicted to be within the TM regions.

MATERIAL AND METHODS

Detection of protein tandem repeats

A program T-REKS was used for *ab initio* identification of the TRs in protein sequences (http://bioinfo.montp.cnrs.fr/?r=t-reks) [21]. It is based on clustering of lengths between identical short strings by using a K-means algorithm. Benchmarks on several sequence datasets showed that TREKS detects the TRs in protein sequences better than the other tested

software. Several parameters of the program can be defined by users. Among them are: Δl - allowed percentage of length variability (default value of Δl used in this analysis is equal to 20% of the repeat length). It was chosen based on the analysis of known repeats of biological importance. The program also evaluates the level of sequence similarity between the identified repeats of each run by using the following approach. Based on the Multiple Sequence Alignment (MSA) of the repeats constituting a given tandem array, T-REKS deduces a consensus sequence and uses it as a reference for similarity calculation. In this alignment an indel is considered as an additional 21st type of amino acid residue. We calculate a Hamming distance D_i [47] between the consensus sequence and a repeat R_i with $1 \leq i \leq m$, where m is a number of repeats in one run. Then, we define a similarity coefficient for the whole alignment as $P_{sim} = (N - \Sigma^{m}_{i=1} D_i)/N$ with $N = m \times l$ (l is the repeat length). The P_{sim} value can be used to estimate the level of perfection of the TR region. The maximal value $P_{sim} = 1$ corresponds to the run of the perfect repeats. In this work we analysed TRs with $P_{sim} >= 0.70$. The minimal length of TR regions was determined by estimation of the expected number of perfect TRs found by chance in a random sequence dataset (of the SwissProt database size) which follows a binomial distribution approximated by a Poisson Distribution [21]. The lengths for which the expected number of perfect TRs is equal or close to zero correspond respectively to 9 residues for homorepeats regions and 14 residues for the other repeats.

Two databases were analyzed: (i) a non-redundant databank of sequences (with less than 95% identity) from the July, 2008 release of PDB [48] and (ii) SwissProt, release of January 2009 [49]. During analysis of PDB, artificial His-tags attached to proteins were not taken into consideration. Short peptides of less than 20 residues which represent ligands bound to proteins were also not taken into consideration. Several errors of PDB sequence annotations were found and excluded from the analysis. The 3D structures of the remaining

164 repeats divided into three groups by the level of perfection (P_{sim}=1, 1>P_{sim}≥0.9, 0.9>P_{sim}≥0.8) were analysed manually (Table 1). The identified TRs were stored in Protein Repeat DataBase (PRDB) (http://bioinfo.montp.cnrs.fr/?r=repeatDB).

Compositional profiling

Biases in the amino acid compositions of intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) can be visualized using a normalization procedure known as compositional profiling [32, 33, 37]. Compositional profiling is based on the evaluation of the $(C_{s1} - C_{s2})/C_{s2}$ values, where C_{s1} is a content of a given residue in a set of interest (regions and proteins with TRs), whereas C_{s2} is the corresponding value for the reference dataset (set of ordered proteins or set of well-characterized IDPs). Datasets of fully disordered and structured proteins were taken from DisProt and PDB databases [38, 39].

Prediction of disordered regions

Two disorder predictors from PONDR® family, VLXT [34, 40] and VLS2 [41, 42], as well as a set of orthogonal predictors such as IUPred [43, 44], FoldIndex [45], and TopIDP [37], were used to analyze the differences between the above-described datasets. PONDR® VLXT is an integration of three artificial neural networks which were designed for each of the termini and the internal part of the sequences, respectively. Each individual predictor was trained in a dataset containing only the corresponding part of sequences. The inputs of the neural networks were amino acid composition, hydropathy, net charge, flexibility, and coordination number. The final prediction result was an average over the overlapping regions of three independent predictors [34, 40].

PONDR® VSL2 utilized support vector machines to train on long sequences with length ≥ 30 and on short sequences of length < 30, separately. The inputs included

hydropathy, net charge, flexibility, coordination number, PSSM from PSI-blast [50], predicted secondary structures from PHDsec [51] and PSIPRED [52]. The final output was a weighted average with the weights determined by a meta-predictor [41, 42]. VSL2 is accurate in detecting both short and long disordered sequences.

IUPred assumes that globular proteins have larger inter-residue interactions than disordered proteins [43, 44]. Hence, it is possible to derive a sequence-based pair-wise interaction matrix from globular proteins of known structures. The averaged energy based on this pair-wise interaction matrix for globular proteins should be different from that of disordered proteins.

FoldIndex is a method developed from charge-hydrophobicity plot [35] by adding the technique of sliding windows [45]. The charge-hydrophobicity plot was designed to determine if a protein is disordered or not as a whole [35]. By applying a sliding window of 21 amino acids centered at a specific residue, the position of this segment on the charge-hydrophobicity plot can be calculated, and the distance of this position away from the boundary line is taken as an indication whether the central residue is disordered or not [45].

TopIDP index is an amino acid scale that discriminates between order and disorder [37]. It is based on a set of general intrinsic properties of amino acid residues that are responsible for the absence of ordered structure in intrinsically disordered proteins. The corresponding TopIDP score for each amino acid along the sequence is an average over a sliding-window of 21 residues. It reflects the conditional possibility of disordered status for the central amino acid in the sliding-window [37].

All these predictors calculate a prediction score for each residue in the sequence. By setting up the threshold value of the prediction score, all the residues whose prediction scores were higher than the threshold value were assigned to be disordered, and the lower-score residues were assigned to be structured.

CONCLUSION

TRs of proteins with known 3D structures are generally imperfect. They have consensus sequences with both conserved and variable amino acid residues. Analysis of these 3D structures reveals that each sequence repeat corresponds to a repetitive structural unit and their tandem arrangement yields elongated regular structures [11]. The conserved residues of repeats are frequently located inside of the structure, because they are important for its stability, whereas variable residues are exposed on the protein surface. This may lead one to expect that all residues of highly perfect TRs are conserved due to their important structural roles. However, our present study shows that this rule does not work for perfect or almost perfect repeats. We showed that increasing repeat perfection correlates to a stronger tendency to be unstructured. This result is in agreement with the previous conclusion about a strong association between homorepeats and unstructured regions [13]. Coding for protein disorder is more permissive and does not require exact sequence motifs in contrast to the coding for the 3D structures. It allows a higher variability in amino acid sequences. Therefore, TR perfection cannot be explained by the necessity to encode disordered conformations. The other reason of high conservation of amino acid residues may be their functional importance, such as involvement of all or almost all residues of the repeat in interaction with the other molecule. This scenario is also unlikely due to the fact that only some residues of the repeat motif can be in contact with the other molecule and, therefore, will be conserved due to the specific functional interactions. Thus, the TR's structural role and functional interactions, even when they are considered together, cannot explain repeat perfection. This consideration favours explanations based on evolutionary reasons. For example, the perfection of TRs may reflect their recent appearance during evolution. It is known that the repetitive regions evolve more rapidly than the other parts of genes [53], such as in microsatellites where the mutation

rate is 10^6 higher than in other regions (10^{-3} to 10^{-4} per locus per generation) [54]. This generic instability of TRs together with the structurally permissive nature of their disordered state may provide a higher chance for newly emerged repeats to be fixed during evolution and allow a rapid response to the environmental changes [12, 55, 56]. The explanation of the repeat perfection by the evolutionary reason is in line with previously suggested hypothesis that IUP may evolve by repeat extension [12]. Functional constraints such as the ability of TRs to bind to the repetitive surfaces of other molecules or to provide a spacer that can vary in length in rapid response to the environmental treats may play a role in their selection during evolution.

Our results suggest that until a certain level of repeat perfection, conservation of amino acid residues has structural reasons and these types of residues may stabilize the unique 3D structure. However, when a certain threshold of the conserved residues in the repeat is exceeded, the repetitive regions of proteins are predominantly disordered and the main reason of residue conservation in TRs **may** change from a structural to an evolutionary one. **This hypothesis can be tested by further evolutionary analysis**. The results of our analysis also lead to a practical recommendation for prediction of structure and function of proteins. If one sees a perfect TR in a protein of interest, this region is most probably unstructured by itself but still may adopt 3D structures upon binding to the other molecular partners.

ACKNOWLEDGEMENT

This work was supported in part by the grants R01 LM007688-01A1 (to A.K.D and V.N.U.) and GM071714-01A2 (to A.K.D and V.N.U.) from the National Institute of Health, the grant EF 0849803 (to A.K.D and V.N.U.) from the National Science Foundation, and the Program of the Russian Academy of Sciences for the “Molecular and Cellular Biology” (to V.N.U.). We gratefully acknowledge the support of the IUPUI Signature Centers Initiative. This work

was also supported by Ministère de l'Education Nationale, de la Recherche et de la Technologie (MENRT) grant to J.J. . We thank A. Ahmed for critical reading of the manuscript and suggestions.

REFERENCES

1. Pellegrini M, Marcotte EM & Yeates TO (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* **35**, 440-446.

2. Fraser RDB & MacRae TP (1973) *Conformation in fibrous proteins and related synthetic polypeptides*. Academic Press, London and New York.

3. Yoder MD, Lietzke SE & Jurnak F (1993) Unusual structural features in the parallel beta-helix in pectate lyases. *Structure* **1**, 241-251.

4. Baumann U, Wu S, Flaherty KM & McKay DB (1993) Three-dimensional structure of the alkaline protease of *Pseudomonas aeruginosa*: a two-domain protein with a calcium binding parallel beta roll motif. *Embo J* **12**, 3357-3364.

5. Kobe B & Kajava AV (2001) The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* **11**, 725-732.

6. Fulop V & Jones DT (1999) Beta propellers: structural rigidity and functional diversity. *Curr Opin Struct Biol* **9**, 715-721.

7. Groves MR & Barford D (1999) Topological characteristics of helical repeat proteins. *Curr Opin Struct Biol* **9**, 383-389.

8. Lee MS, Gippert GP, Soman KV, Case DA & Wright PE (1989) Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* **245**, 635-637.

9. Sawaya MR, Wojtowicz WM, Andre I, Qian B, Wu W, Baker D, Eisenberg D & Zipursky SL (2008) A double S shape provides the structural basis for the extraordinary binding specificity of Dscam isoforms. *Cell* **134**, 1007-1018.

10. Elkins PA, Ho YS, Smith WW, Janson CA, D'Alessio KJ, McQueney MS, Cummings MD & Romanic AM (2002) Structure of the C-terminally truncated human ProMMP9, a gelatin-binding matrix metalloproteinase. *Acta Crystallogr D Biol Crystallogr* **58**, 1182-1192.

11. Kajava AV (2001) Review: proteins with repeated sequence--structural prediction and modeling. *J Struct Biol* **134**, 132-144.

12. Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* **25**, 847-855.

13. Simon M & Hancock JM (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol* **10**, R59.

14. Huntley MA & Golding GB (2002) Simple sequences are rare in the Protein Data Bank. *Proteins* **48**, 134-140.

15. Andrade MA & Bork P (1995) HEAT repeats in the Huntington's disease protein. *Nat Genet* **11**, 115-116.

16. Heringa J (1998) Detection of internal repeats: how common are they? *Curr Opin Struct Biol* **8**, 338-345.

17. Kobe B & Kajava AV (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem Sci* **25**, 509-515.

18. Matsushima N, Yoshida H, Kumaki Y, Kamiya M, Tanaka T, Izumi Y & Kretsinger RH (2008) Flexible structures and ligand interactions of tandem repeats consisting of proline, glycine, asparagine, serine, and/or threonine rich oligopeptides in proteins. *Curr Protein Pept Sci* **9**, 591-610.

19. Aachmann FL, Svanem BI, Guntert P, Petersen SB, Valla S & Wimmer R (2006) NMR structure of the R-module: a parallel beta-roll subunit from an *Azotobacter vinelandii* mannuronan C-5 epimerase. *J Biol Chem* **281**, 7350-7356.

20. Fondon JW, 3rd & Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* **101**, 18058-18063.

21. Jorda J & Kajava AV (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* **25**, 2632-2638.

22. Cummings CJ & Zoghbi HY (2000) Trinucleotide repeats: mechanisms and pathophysiology. *Annu Rev Genomics Hum Genet* **1**, 281-328.

23. Cummings CJ & Zoghbi HY (2000) Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum Mol Genet* **9**, 909-916.

24. McEwan IJ (2001) Structural and functional alterations in the androgen receptor in spinal bulbar muscular atrophy. *Biochem Soc Trans* **29**, 222-227.

25. Kleinjung J, Petit MC, Orlewski P, Mamalaki A, Tzartos SJ, Tsikaris V, Sakarellos-Daitsiotis M, Sakarellos C, Marraud M & Cung MT (2000) The third-dimensional structure of the complex between an Fv antibody fragment and an analogue of the main immunogenic region of the acetylcholine receptor: a combined two-dimensional NMR, homology, and molecular modeling approach. *Biopolymers* **53**, 113-128.

26. Ngo JC, Giang K, Chakrabarti S, Ma CT, Huynh N, Hagopian JC, Dorrestein PC, Fu XD, Adams JA & Ghosh G (2008) A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1. *Mol Cell* **29**, 563-576.

27. Arnett KL, Harrison SC & Wiley DC (2004) Crystal structure of a human CD3-epsilon/delta dimer in complex with a UCHT1 single-chain antibody fragment. *Proc Natl Acad Sci U S A* **101**, 16268-16273.

28. Deivanayagam CC, Rich RL, Carson M, Owens RT, Danthuluri S, Bice T, Hook M & Narayana SV (2000) Novel fold and assembly of the repetitive B region of the *Staphylococcus aureus* collagen-binding surface protein. *Structure* **8**, 67-78.

29. Hege T, Feltzer RE, Gray RD & Baumann U (2001) Crystal structure of a complex between *Pseudomonas aeruginosa* alkaline protease and its cognate inhibitor: inhibition by a zinc-NH2 coordinative bond. *J Biol Chem* **276**, 35087-35092.

30. Kuzin AP, Chen Y, Acton T, Xiao R, Conover KMC, Kellie R, Montelione GT, Tong L & Hunt JF (To be published) X-Ray structure of an ankyrin repeat family protein Q5ZSV0 from *Legionella pneumophila*.

31. Schrag JD, Bergeron JJ, Li Y, Borisova S, Hahn M, Thomas DY & Cygler M (2001) The Structure of calnexin, an ER chaperone involved in quality control of protein folding. *Mol Cell* **8**, 633-644.

32. Vacic V, Uversky VN, Dunker AK & Lonardi S (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* **8**, 211.

33. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, et al. (2001) Intrinsically disordered protein. *J Mol Graph Model* **19**, 26-59.

34. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ & Dunker AK (2001) Sequence complexity of disordered protein. *Proteins* **42**, 38-48.

35. Uversky VN, Gillespie JR & Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**, 415-427.

36. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN & Dunker AK (2007) Intrinsic disorder and functional proteomics. *Biophys J* **92**, 1439-1456.

37. Campen A, Williams RM, Brown CJ, Meng J, Uversky VN & Dunker AK (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett* **15**, 956-963.

38. Xue B, Li L, Meroueh SO, Uversky VN & Dunker AK (2009) Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Mol Biosyst*.

39. Xue B, Oldfield CJ, Dunker AK & Uversky VN (2009) CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett* **583**, 1469-1474.

40. Romero P, Obradovic Z, Kissinger C, Villafranca J & Dunker A (1997) Identifying disordered regions in proteins from amino acid sequence. *Proc IEEE Int Conf Neural Networks* **1**, 90–95.

41. Peng K, Radivojac P, Vucetic S, Dunker AK & Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208.

42. Obradovic Z, Peng K, Vucetic S, Radivojac P & Dunker AK (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61 Suppl 7**, 176-182.

43. 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.

44. Dosztanyi Z, Csizmok V, Tompa P & Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**, 827-839.

45. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I & Sussman JL (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435-3438.

46. Tang YC, Chang HC, Roeben A, Wischnewski D, Wischnewski N, Kerner MJ, Hartl FU & Hayer-Hartl M (2006) Structural features of the GroEL-GroES nano-cage required for rapid folding of encapsulated protein. *Cell* **125**, 903-914.

47. Hamming R (1950) Error Detecting and Error Correcting Codes. *AT&T TECH J* **29**, 147-160.

48. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242.

49. Bairoch A & Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45-48.

50. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.

51. Rost B, Sander C & Schneider R (1994) PHD--an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci* **10**, 53-60.

52. McGuffin LJ, Bryson K & Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.

53. Buard J & Vergnaud G (1994) Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *Embo J* **13**, 3203-3210.

54. Weber JL & Wong C (1993) Mutation of human short tandem repeats. *Hum Mol Genet* **2**, 1123-1128.

55. Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet* **16**, 551-558.

56. Williamson MP (1994) The structure and function of proline-rich regions in proteins. *Biochem J* **297 (Pt 2)**, 249-260.

TABLE 1

Number of structured and unstructured regions found for each range of P_{sim} in the PDB TRs dataset

P_{sim} ranges	**Sn**	Ln	Un	**Sd**	Ld	Ud
$P_{sim}=1.0$	0	2	7	16	4	14
$0.9 \leq P_{sim} < 1.0$	1	2	8	20	2	5
$0.8 \leq P_{sim} < 0.9$	17	8	31	24	1	12

The following tags were assigned to each analysed region with TRs: Sn and Sd - fragments containing secondary structures from natural and designed proteins correspondingly; Ln and Ld – fragments connecting secondary structures; Un and Ud – fragments whose structure was not determined

TABLE 2

Analysis of intrinsic disorder distribution in TRs and TR-containing proteins

		P_{sim} 0.7-0.8	P_{sim} 0.8-0.9	P_{sim} 0.9-1	Homorepeats
TRs*	Total No.	34286	5519	1382	5259
	Avg. Length	25.5	41.0	59.1	13.8
	ID Ratio – VSL2	80.4%	88.6%	88.9%	98.4%
	ID Ratio – IUpred	56.0%	62.7%	67.2%	86.5%
	ID Ratio – FoldIndex	62.4%	68.6%	70.3%	79.9%
	ID Ratio – TopIDP	85.6%	88.8%	91.1%	74.4%
Sequences**	Total No.	25649	4915	1295	3663
	Avg. Length	643.4	752.0	840.2	790.4
	ID Ratio – VSL2	49.3%	58.6%	57.0%	61.6%
	ID Ratio – IUpred	32.1%	41.7%	41.7%	45.4%
	ID Ratio – FoldIndex	46.6%	52.3%	52.3%	52.7%
	ID Ratio – TopIDP	71.2%	75.3%	72.2%	74.9%

* Tandem repeat segments; ** Whole proteins containing these tandem repeats.

TABLE 3

Variation of the disorder level among TRs of viral, eukaryotic and prokaryotic proteins

	Prokaryotes	Viruses	Eukaryotes
PONDR® VLXT *	84%	85.0%	88.4%
TOP-IDP**	71.4%	72.4%	77.8%

*Protein regions with VLXT CDF distance less than 0 are identified as disordered. P_{sim} range for this dataset is 0.9-1. Disorder level is estimated as percentage of residues predicted to be disordered.

** Protein regions with TOP-IDP less than 0 are identified as disordered. P_{sim} range for this dataset is 0.9-1. Disorder level is estimated as percentage of TRs with negative TOP-IDP values.

TABLE 4

Abundance of disordered repeats as a function of the subcellular localization of corresponding repeat-containing proteins

	Prokaryotes			Eukaryotes			
	cytoplasm	membrane	secreted	nucleus	cytoplasm	membrane*	secreted
Ratio of TOP-IDP	54.3%	72.3%	83.6%	74%	81.2%	60.2%	72.7%
Number of TRs	459	264	140	3650 (1898 hr**)	1181 (476 hr)	1436 (637 hr)	782 (178 hr)

* Membrane localization for eukaryotes combines “membrane” and “cell membrane” terms from SwissProt. ** Number of homorepeats (hr) among TRs.

LEGENDS TO FIGURES

FIGURE 1. The 3D structures of proteins with almost perfect tandem repeats. Repeat regions are shown in color.

FIGURE 2. Compositional profiling of tandem repeats, entire sequences of proteins containing these tandem repeats, and a set of fully disordered proteins from DisProt in comparison with the composition of fully structured proteins from PDB. C^{Struct}_{AA} is the content of a given amino acid in the set of structured proteins; $C^{Dataset}_{AA}$ is the content of this amino acid in the dataset of interest. Amino acids are denoted by one letter code and arranged in order of decreasing structure-promoting property suggested by TOP-IDP scale [37].

FIGURE 3. (**A**) Difference of amino acid compositions between tandem repeat segments subdivided into groups with different level of the repeat perfection and fully structured proteins. The homorepeats are analyzed separately (**B**) due to their unusually high occurrence in comparison to the other tandem repeats. For this purpose, a dataset of perfect and cryptic homorepeats was created and subdivided in three groups depending on the P_{sim} values. C^{tr}_{AA} and C^{hr}_{AA} are the contents of a given amino acid in the set of tandem repeats (excluding homorepeats) and only homorepeats, correspondingly. Amino acids residues are arranged in four sets: order-promoting aromatic and aliphatic amino acids (W, F, Y, I, M, L, V, and A) which are denoted as non-polar; glycine, as order-neutral and, at the same time, specific residue, disorder promoting polar residues (N, C, T, Q, S, R, D, H, E, and K) and disorder-promoting proline.

FIGURE 4. Length distribution of predicted disordered segments. (**A**) Length distribution of predicted disorder for 4 groups of tandem repeats. (**B**) Length distribution of predicted disorder for whole protein sequences containing the tandem repeats in 4 groups.

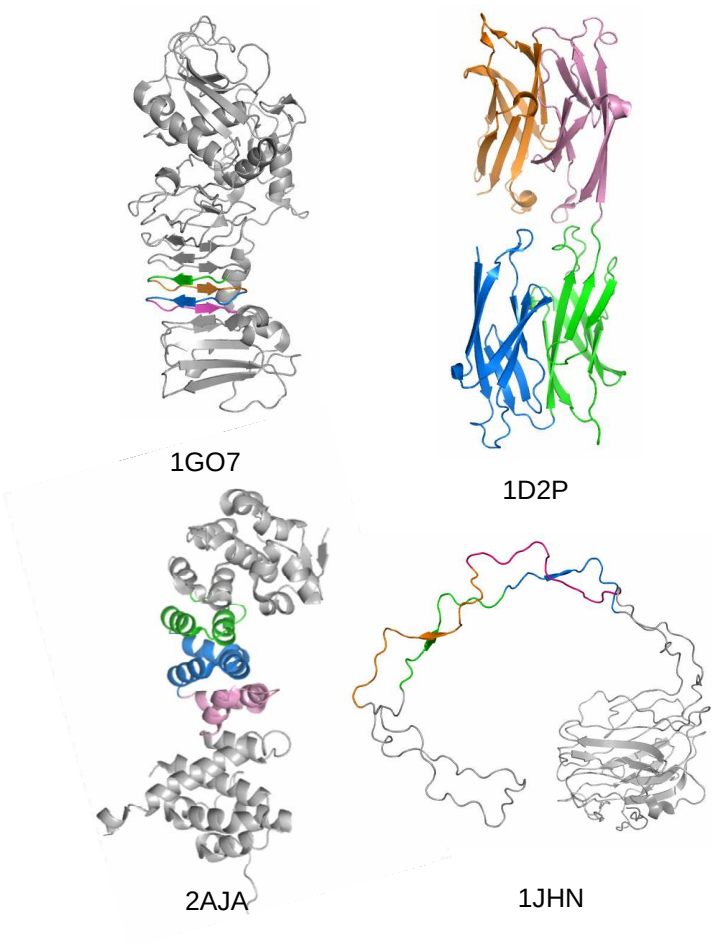


Fig. 1

Jorda et al.

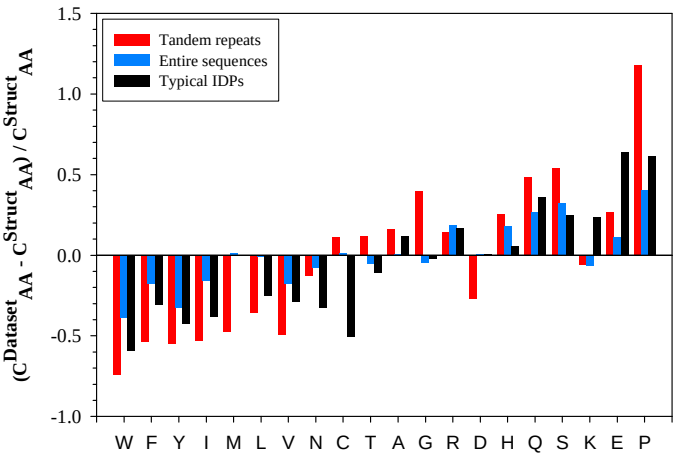


Fig. 2

Jorda et al.

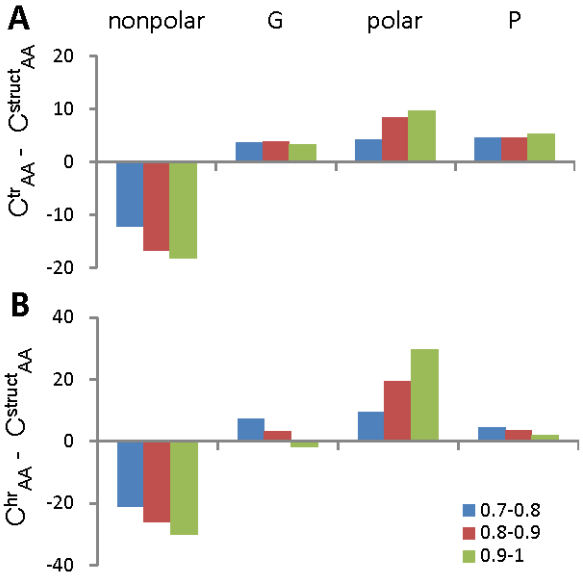
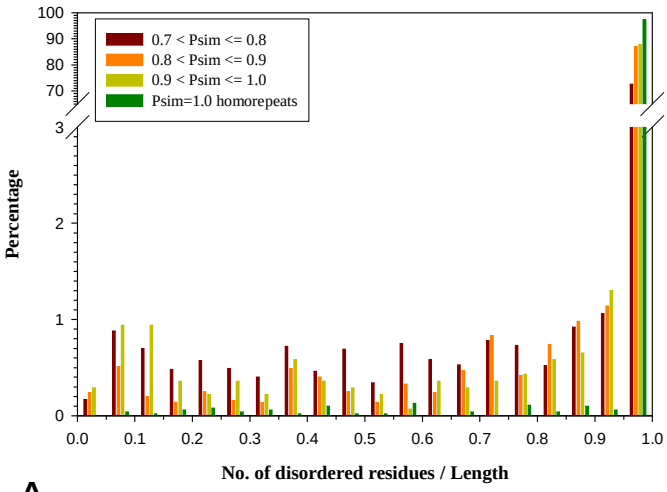
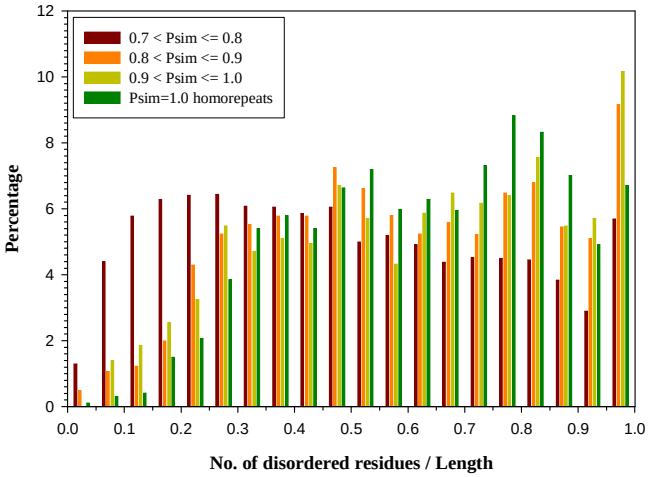


Fig. 3

Jorda et al.



A



B

Fig. 4

Jorda et al.