**BRIEFINGS IN BIOINFORMATICS.** VOL 15. NO 4. 582–591
Advance Access published on 14 February 2013
doi:10.1093/bib/bbt003

# Understanding and identifying amino acid repeats

*Hong Luo and Harm Nijveen*

Submitted: 22nd November 2012; Received (in revised form): 17th January 2013

## Abstract

Amino acid repeats (AARs) are abundant in protein sequences. They have particular roles in protein function and evolution. Simple repeat patterns generated by DNA slippage tend to introduce length variations and point mutations in repeat regions. Loss of normal and gain of abnormal function owing to their variable length are potential risks leading to diseases. Repeats with complex patterns mostly refer to the functional domain repeats, such as the well-known leucine-rich repeat and WD repeat, which are frequently involved in protein–protein interaction. They are mainly derived from internal gene duplication events and stabilized by 'gate-keeper' residues, which play crucial roles in preventing inter-domain aggregation. AARs are widely distributed in different proteomes across a variety of taxonomic ranges, and especially abundant in eukaryotic proteins. However, their specific evolutionary and functional scenarios are still poorly understood. Identifying AARs in protein sequences is the first step for the further investigation of their biological function and evolutionary mechanism. In principle, this is an NP-hard problem, as most of the repeat fragments are shaped by a series of sophisticated evolutionary events and become latent periodical patterns. It is not possible to define a uniform criterion for detecting and verifying various repeat patterns. Instead, different algorithms based on different strategies have been developed to cope with different repeat patterns. In this review, we attempt to describe the amino acid repeat-detection algorithms currently available and compare their strategies based on an in-depth analysis of the biological significance of protein repeats.

***Keywords:*** *amino acid repeat; detection algorithm; low complexity sequence; repeat containing protein; protein domain repeats*

## INTRODUCTION

Amino acid repeats (AARs) are abundant in protein sequences either as periodic elements in structural proteins such as collagens, keratins, silk and cell wall proteins, or as structural modules in functional proteins such as transcription factors, receptors, ion channels, histones, ubiquitins and calcium storage proteins. Table 1 shows some well-known examples of human repeat-containing proteins (RCPs) gathered in the UniProt/Swiss-Prot Knowledgebase (http://www.uniprot.org/). For example, the major prion protein (PRIO_HUMAN) contains an N-terminal repeat region with several octamers (PHGGGWGQ); the extra-embryonic spermatogenesis homeobox 1 protein (ESX1_HUMAN) has a sequence motif PPxxPxPPx repeated nine times and the alpha-1 type I collagen protein contains a repeat of various lengths of the periodic tri-amino acid GPP. The giant muscle protein Titin composed of 34 350 amino acid residues (TITIN_HUMAN) contains several types of repeating domains. Single amino acid repeats (SAARs) are also common, such as the polyQ repeats in the Forkhead box protein P2 (FOXP2_HUMAN), the androgen receptor (ANDR_HUMAN) and the Huntington's disease (HD) protein (HD_HUMAN). Other SAARs including polyL, polyA and polyH can also be found in many other proteins. RCPs are distributed in all life kingdoms, and especially abundant in eukaryotes [1].

Corresponding author. Hong Luo, Wageningen University Laboratory of Bioinformatics, Radix, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands. Tel: +86-10-62836682; Fax: +86-10-62836682; E-mail: hong.luo@ibcas.ac.cn

**Hong Luo** is a PhD student at Wageningen University and Research Center. His research interests include data mining and comparative genomics of repeat patterns in protein sequences.

**Harm Nijveen** is a scientific programmer at Wageningen University and Research Center. His research interests are in bioinformatics and molecular biology.

**Table 1:** Some examples of AARs in human proteins

| UniProt ID | Protein | AA | Repeat pattern |
|---|---|---|---|
| SECR_HUMAN | Secretin | 121 | polyL |
| PRIO_HUMAN | Major prion protein | 253 | $(PHGGGWGQ)_4$ |
| ANKRI_HUMAN | Ankyrin repeat domain-containing protein 1 | 319 | Ankyrin repeat |
| CASQ2_HUMAN | Calsequestrin-2 | 399 | D/E-Rich |
| ESXI_HUMAN | Homeobox protein ESX1 | 406 | $(PPxxPxPPx)_9$ |
| WDRI_HUMAN | WD repeat-containing protein 1 | 606 | WD repeat |
| UBC_HUMAN | Polyubiquitin-C | 685 | Ubiquitin |
| FOXP2_HUMAN | Forkhead box protein P2 | 715 | polyQ |
| LRRNI_HUMAN | Leucine-rich repeat neuronal protein 1 | 716 | Leucine Rich Repeat |
| ANDR_HUMAN | Androgen receptor | 919 | polyQ, polyG, polyP |
| SRBP2_HUMAN | Sterol regulatory element-binding protein 2 | 1141 | polyS, $(PQ)_4$, $(SGSS)_2$ |
| BRD4_HUMAN | Bromodomain-containing protein 4 | 1362 | polyP, polyH, polyQ, K-Rich, S-Rich |
| COIAI_HUMAN | Collagen alpha-1(I) chain | 1464 | $(GPP)_n$ |
| CACIA_HUMAN | Brain calcium channel 1 | 2505 | polyQ, polyH, polyG |
| HD_HUMAN | Huntington disease protein | 3142 | polyQ, polyP, polyT, polyE, HEAT domain |
| MLL2_HUMAN | Histone-lysine N-methyltransferase MLL2 | 5537 | $(S/P\text{-}P\text{-}P\text{-}E/P\text{-}E/A)_{15}$ |
| TITIN_HUMAN | Titin | 34350 | Several types of repeating domains: TPR WD RCCI PEVK Kelch Z Ig repeats |

It is known that some AARs such as the leucine-rich repeats (LRRs) form the structural framework for protein–protein interaction, and the repeat fragment in zinc finger transcription factors binds to *cis*-elements of DNA promoters. AARs can also cause problems such as the mis-folding of prion proteins [2]. Furthermore, modification of repeat length may introduce abnormal function. A typical case is the expansion of polyQ, resulting in several neurological disorders such as mental retardation, HD, inherited ataxias and muscular dystrophy.

## Classification of amino acid repeat patterns at sequence level

Mathematical and statistical methodologies can be applied to study the particular functional and evolutionary background of an AAR. Several approaches have been proposed to classify AARs into different categories depending on the characteristics of repeat units, including the sequence similarity among repeat units, the distance between adjacent repeat units and the complexity of the sequence pattern of the repeat units.

The first approach is to classify AARs according to the similarity among the repeat units. Based on this approach, AARs can be classified into two main groups: perfect repeats and imperfect repeats. The repeat units in perfect repeat fragments are identical, e.g. AAAAAAA and PQPQPQPQ, whereas the repeat units in imperfect repeat fragments are not exactly the same, e.g. AAWAAAA and QQQMLQQQFL. Imperfect repeats with highly variable, but still recognizable, repeat units are also called divergent repeats.

The second approach for repeat classification is based on the distance between adjacent units. AARs can be classified as tandem repeats (TRs) or non-tandem repeats (NTRs). The units in TRs are continuously distributed in the repeat sequence, whereas the units in NTRs are sequentially interspersed.

The third approach takes the complexity of the sequence pattern of the repeat units into consideration. Based on this approach, AARs can be roughly classified as simple repeats or complex repeats. Simple repeats generally refer to the continuous or interrupted runs of single amino acid residues or short peptides. The regions in a protein sequence containing simple repeats are often called simple sequences (SSs) or low complexity regions (LCRs). On the other hand, most of the complex repeats usually have sophisticated patterns of repeat units with variable lengths ranging from 10 to >100 residues, and these complex repeats patterns are frequently recognized as repeated protein domains [3].

In practice, it is rather difficult to strictly distinguish the different classes owing to the complicated patterns of AARs. For example, some domain repeats also contain SSs, such as the abundant leucine residues found in an LRR domain. And in the case of point mutations or insertions/deletions (INDELs), the original perfectly repeated units in proteins could gradually evolve into non-perfect tandem repeats (NPTRs).

The above approaches used to classify AARs are all based on the protein sequence. However, they are

insufficient to reveal the biological significance of AARs, as proteins play their functional roles by folding into particular secondary and tertiary structures, which are difficult to deduce through amino acid patterns at sequence level. Data from several experiments show that proteins with similar tertiary structures may share low sequence identity [4, 5]. And similar functional domains of proteins do not necessarily correspond to recognizable sequence repeat patterns [3, 6–8]. Therefore, in-depth study of protein repeats requires better understanding of the correspondence of repeat sequences with their structures and functions. In addition, the acquisition of such biological knowledge is more sophisticated than simply classifying sequential repeat data.

## Biological significance of different patterns of AARs

Biologically, different amino acid repeat patterns imply different functional and evolutionary backgrounds. Repeats with simple patterns, such as single AARs, mainly exist in intrinsically unstructured regions (IURs) of proteins [9, 10]. Such protein regions that do not fold into a 3D structure commonly have functions related to molecular recognition and molecular assembly [11, 12]. Single amino acid or trinucleotide repeats like polyQ are involved in neurodegenerative diseases such as HD [13], where their length variations often result in either loss of normal or gain of abnormal function[14, 15].

Most SAARs are presumed to be originally derived from replicative DNA slippage [16] in the coding region. Expansion of some SAARs might also result from unequal chromosomal crossover, such as the polyA in the human HOX13 gene [17]. In general, perfect amino acid runs are inherently mutable and are frequently interrupted by point mutations [18] to become simple sequences [19].

In addition to SAARs, sequential tandem repeats (PTRs and NPTRs) with highly similar units are prevalent in protein sequences. We have found that ~13% of all proteins deposited in the public protein databases contain at least one tandem repeat fragment. And >40% of the tandem repeats are PTRs, while ~60% PTRs are single amino acid runs [1]. Errors in sequencing and automatic annotation procedures might have introduced some false-positive PTRs into the public protein knowledgebase. However, this cannot undermine the biological significance of frequently occurring PTRs in protein sequences, especially considering the fact that functional PTRs are being continuously experimentally identified, and most of them are conserved among orthologous proteins [20–22].

Consistent with this scenario, conservation of amino acid tandem repeats is a strong indication for biological relevance. The phylogenetically conserved repeat fragments among orthologous proteins should have a conserved function, such as the conserved polyQ regions in primate FOXP2 proteins [23]. In contrast, however, variable repeat unit length in corresponding regions of orthologous proteins indicates a different scenario. These repeats are probably going through a rapid change driven by selection [24]. More interestingly, tandem repeats have been shown to play an important role in micro-evolution by catalysing the rapid production of genetic and phenotypic variation among organisms [25–28].

Repeats with complex patterns have comparatively stable structures and conserved functions, which are generally called domain repeats. Domain repeats are among the most common protein motifs in the Pfam database [29], such as LRRs, Zinc finger repeats, Ankyrin repeats and Tetratricopeptide repeats (TPRs) [30]. These domain repeats are mostly involved in transcription regulation, cell-cycle control and signal transduction [31–34] and widely spread in the proteomes of different species across different life kingdoms [35]. Many genes containing these domain repeats in the coding region are significant in certain diseases [36], as sequence identity increases the chance of protein aggregation [37] and mis-folding. Domain repeats are thought to have evolved through internal gene duplications arising from recombination events [3, 38], such as unequal crossing over [39] and exon shuffling [40]. The duplications may involve several domains at a time [3, 41]. In addition, a number of specific sequence-based signals such as the 'gate-keeper' residues [41] play a crucial role in preventing inter-domain aggregation. Therefore, these repeat patterns are generally obscure at sequence level, and a sophisticated search is required to detect them.

# REPEAT DETECTION STRATEGIES

During the past decade, several strategies for the identification of AARs from protein sequences have been reported. Among these approaches, the three major ones are self-comparison, pattern recognition and complexity measurement. Table 2 shows

**Table 2:** Repeat detection algorithms

| Method | Repeat type[a] | Ref | Availability |
|---|---|---|---|
| Self-comparison | | | |
| REP | Domain | [42] | http://www.embl.de/~andrade/papers/rep/search.html |
| COACH | Domain | [43] | http://www.drive5.com/lobster/ |
| TPRpred | Domain | [44] | http://tprpred.tuebingen.mpg.de/ |
| REPRO | Domain | [45] | http://www.ibi.vu.nl/programs/reprowww/ |
| TRUST | Divergent | [46] | http://www.ibi.vu.nl/programs/trustwww/ |
| Internal Repeat Finder | Divergent | [47] | http://nihserver.mbi.ucla.edu/Repeats/ |
| HHrep | Divergent | [48] | http://hhrep.tuebingen.mpg.de/hhrep/ |
| RADAR | Divergent | [49] | http://www.ebi.ac.uk/Tools/Radar/ |
| HHrepID | Divergent | [50] | http://toolkit.tuebingen.mpg.de/hhrepid/ |
| Pattern recognition | | | |
| REPETITA | Solenoid | [51] | http://protein.bio.unipd.it/repetita/ |
| LSTM | Domain | [52] | http://www.bioinf.jku.at/software/LSTM_protein/ |
| ARD | Alpha-Rod | [53] | http://www.ogic.ca/projects/ard/ |
| Complexity measurement | | | |
| SIMPLE | Simple | [19] | http://www.biochem.ucl.ac.uk/bsm/SIMPLE/ |
| GBA | Simple | [54] | xli@cise.ufl.edu |
| Others | | | |
| XSTREAM | NPTR | [55] | http://jimcooperlab.mcdb.ucsb.edu/xstream/ |
| Apriod | PPP | [56] | hwan@mindgen.org |
| LocRepeat | PPP | [57] | http://www.cs.cityu.edu.hk/~lwang/software/LocRepeat/ |
| REPfind | NPTR | [58] | adebiyi@informatik.uni-tuebingen.de |
| Reptile | Perfect | [59] | http://reptile.unibe.ch/ |
| SUFFIX | Perfect | [60] | http://www.cs.ucdavis.edu/~gusfield/strmat.html |

[a]NPTR = non-perfect tandem repeat; PPP = pseudo-periodic partitions.

the algorithms and publicly available tools including online resources that can be used to detect AARs of various types.

In the following section, we will give a brief introduction to the amino acid repeat-detection strategies focusing on the general principles behind these strategies.

## The self-comparison strategy

One of the most intuitive strategies to detect repeat patterns in protein sequences is the self-comparison method. The idea of this approach is rather simple, i.e. comparing a protein sequence to itself. Sequence comparison is a fundamental bioinformatics method that has been extensively used to search similar regions among biological sequences. The global sequence alignment method was first proposed in the 1970s [61] and focuses on finding the optimal alignment of two entire biological sequences using dynamic programming. Soon after, the Smith–Waterman local alignment algorithm [62] was developed to recognize the better aligned sub-regions between two sequences in order to show meaningful biological relevance.

On aligning a sequence with itself for the purpose of identifying repeat patterns, the sub-optimal alignments become obscured by the best (and most obvious) alignment. This optimal alignment should be excluded from the initial search. The reliability of identifying sub-optimal alignments of protein sequences using the dynamic programming method has been evaluated [62]. A very distinguishing feature of this method is the use of a scoring system that gives scores to paired amino acids and penalties to unmatched gaps. Substitution matrices such as PAM [63] and BLOSUM [64] are the basis of the scoring system and represent the specific evolutionary relevance among different amino acids. More specifically tuned scoring matrices have also been proposed. These matrices take special features of amino acids such as polarity, electrostatic charge, structure, molecular volume and codon bias [65] into account. One of the greatest advantages of using a scoring system for identifying sub-optimal alignments is that statistical models can be applied to define reliable criteria [66, 67].

In principle, the self-alignment repeat-detection methods are the extension of an alignment-based homology-detection approach. Thus, they have inherited characteristics that are more suitable for detecting divergent internal repeats in protein sequences. The units of these repeats generally have

low identities and ambiguous boundaries, but share evolutionarily conserved sites or motifs, which are presumed to have crucial functions. As such, the accurate definition of repeat length and repeat number according to substantial biological significance is a sophisticated problem. And this is especially true for detecting repeat patterns without prior knowledge, also called '*de novo*' repeat detection. On the other hand, the algorithms depending on prior knowledge, such as REP, COACH and TPRpred [42–44], generally search repeat patterns from sequence databases by profiles constructed with known repeat families using hidden Markov models (HMMs) [68]. Therefore, the repeat patterns identified by these programs are usually well-known, and some of them are experimentally studied functional protein domain repeats.

It is generally believed that detecting repeat patterns with a self-alignment-based method is a feasible strategy. However, it also has some flaws and limitations. First, the computational complexity of performing self-alignment is high. The general complexity for a sequence with n amino acids is O ($n^2$) for both time and space, which will increase exponentially with the increase of the sequence length. Fortunately, this problem is not too serious for protein sequences, as their average length is around 320 AAs [69]. And the computational capacity of current computer hardware is powerful enough to handle this problem within acceptable time and space. In addition, several optimization strategies have been recently applied to sequence alignments, such as the implementation of the Smith–Waterman algorithm with the new technology of graphics processing units (GPUs) [70], and the parallel computing version of the REPRO [71] algorithm [72] can handle much longer sequences within a reasonable time.

One of the main purposes for detecting AARs is to find novel repeat patterns and infer their functional and evolutionary roles. As the majority of repeat patterns in protein sequences have not been well studied, *de novo* repeat-detection algorithms are more widely used, such as PEPRO, Internal Repeat Finder, RADAR, TRUST, HHrep and HHrepID [45–50, 56, 57]. All of them identify repeats using the self-comparison strategy, but differ in some aspects. For example, Internal Repeat Finder assumes that the statistically significant sub-optimal alignment scores should have a Poisson distribution [47]. TRUST uses the particular strategy on sub-optimal alignments, which could increase the chance and reliability to identify divergent repeats [46]. HHrep [48] and its optimized version HHrepID [50] compares a sequence with itself by the HMM–HMM [73] strategy, which looks for the sub-optimal alignments using a profile HMM constructed by iterations of PSI-BLAST [74].

## The pattern recognition strategy

The second strategy to detect AARs from protein sequences uses the conventional method of pattern recognition. The two main algorithms of this strategy are the discrete Fourier transform (DFT) and neural networks.

DFT has been widely applied in the research area of signal processing. Generally, it can decompose signals into constituent frequencies, so that the cryptic patterns hidden in the signals could be analysed intuitively. Early studies showed that DFT can be used to detect periodic patterns in collagen protein [75], but also has some fundamental difficulties which limit its usage [45]. The accuracy of DFT-based methods is easily biased by the length variation of the repeat units caused by mutations or INDELs, as this will weaken the periodical pattern of the transformed Fourier spectral amplitudes.

Some recent algorithms make efforts to provide better discrimination on Fourier spectral amplitudes using newly developed methods. For example, REPETITA yields better accuracy than self-alignment methods on detecting solenoid repeats by introducing several optimized strategies of the DFT-based method [51]. In addition, the stationary wavelet packet transform has been widely used in bioinformatics and computational biology in recent years [76]. As a state of the art optimization DFT algorithm [77], it has been shown to have good quality on detecting protein repeat patterns [78].

The neural network-based method is another well-studied pattern-recognition strategy, which is also capable of identifying similar patterns in protein sequences [79]. A well-established neural network is able to associate homologous patterns in the protein sequence with the input patterns and can be trained to adapt the patterns. Several neural network algorithms show good accuracy and time efficiency on protein homologue detection. LSTM is able to combine amino acid properties with patterns and does not rely on pre-defined scoring matrices for similarity measurements [52]. The ARD neural network is designed to identify specific alpha-rod repeat patterns

and has been applied to the analysis of Huntingtin protein sequences [53].

## The complexity measurement strategy

The third approach of identifying AARs takes complexity measurement into consideration. LCRs are widely distributed in protein sequences. LCRs commonly contain particular repeat patterns that have continuous repetitions of very short units, such as the SAARs and cryptically simple sequences [19]. Apparently, these repeats have special functional and evolutionary properties that differ from the repeats with more complex patterns and longer units. Their typical short unit length makes both the self-comparison- and the pattern recognition-based strategies less well suited to identify LCR repeats efficiently.

Fortunately, several algorithms have been introduced to detect repeats involved in LCRs, most of them using a strategy to measure the complexity of sequences within a sliding window. As for complexity measuring, SIMPLE [19] awards simplicity score to the central amino acid of each window, and is most suitable for detecting short unit cryptic repeats. SEG [80], DSR [81] and CARD [82] are based on Shannon entropy [83], which displays several limitations when decoding complex protein sequences (43).

The main drawback of sliding windows-based algorithms is that they all require a pre-specified window size, and repeats that are longer or shorter than the window are not detectable. On the other hand, non-sliding window algorithms show more flexibility on detecting repeats in LCRs. GBA [54] constructs a graph for each protein sequence, and finds short subsequences as LCR candidates through traversing. Coronado [84] introduces the composition-modified scoring matrices to identify LCRs within cell wall proteins of fungi. These algorithms are an important complement to the sliding window-based algorithms.

## Other strategies

As described above, the self-comparison strategy and the pattern recognition strategy are mostly suitable for detecting divergent repeats, whereas the complexity measurement strategy is mostly suitable for detecting simple unit repeats. In addition, exclusive and optimized strategies for sequential tandem repeats are also particularly useful. Sequential tandem repeats implicated in the amino acid fragments with tandem repeat patterns are comparatively more explicit than divergent repeats. They are widely spread in many proteomes across wide taxonomic ranges, but are still insufficiently studied.

Hamming distance [85] and edit distance, also called Levenshtein distance [86], are widely used for measuring the similarity of sequential tandem repeats [87–90]. Differing from hamming distance, which only accounts for point mutations, edit distance-measuring algorithms also consider insertions and deletions. In addition, Apriod [56] and LocRepeat [57] focus on finding the 'pseudo-periodic partitions', which are gradually evolved patterns among repeat units. Given that NPTRs are originally evolved from PTRs, Xstream [55] and REPfind [58] detect NPTRs based on the extension of exact repeats seeds, which could decrease the computational complexity of both time and space.

Most of the repeat-detection algorithms can identify PTRs together with other repeat patterns incidentally. But as some of the PTRs are nested in larger NPTR fragments, which can hardly be distinguished by the common strategies, an exclusive algorithm for detecting PTRs is also necessary. For example, the suffix tree-based strategy is supportive to identify all PTRs in a protein sequence with linear time complexity [60]. Reptile uses a 'brute-force' strategy to detect PTRs from the proteins of parasite antigens [59]. Following the definition of statistically significant repeat runs in protein sequences [91], the cut-off sizes of five, four, three and two of the repeat unit repetitions are common criteria for identifying mono-amino, di-amino, tri-amino and all other repeats, respectively.

# SUMMARY AND PERSPECTIVE

Identifying repeat patterns in proteins is the first step towards the understanding of their physiological function and evolutionary mechanism. During the evolution process, these patterns become so intricate that no single algorithm is adequate to identify all of them. There is no doubt that an in-depth investigation of their biological background is required to choose proper algorithms for the identification of specific patterns. In general, self-comparison algorithms are suitable to detect *de novo* repeats with complex patterns. Pattern recognition-based algorithms are suitable to detect repeats with low sequence identities but high intrinsic biological similarities. Complexity measurement-based algorithms can be

applied to detect repeats with simple patterns involved in LCRs. For the tandem repeats that have more sequentially repetitive patterns, one should consider the strategies that measure the similarity of repeat units by edit or hamming distance.

The biological significance of protein repeats has been discussed for years. Internal duplication in genomes is one of the most important evolutionary mechanisms for species to adapt the environment [92–94]. As a result, repetitive patterns at the DNA level such as interspersed microsatellites and tandem tri-nucleotide repeats are prevalent. Intragenic repeats are presumed to have potential roles on generating functional variability [95, 96]. And the repeats in coding regions corresponding to AARs are more likely to go through adaptive competition [24, 97, 98]. Therefore, large amount of repeats in proteins is less likely to be regarded as 'junk proteins' [99], which merely have non-essential roles. At the same time, their variable characters and vulnerabilities to disorder and diseases has been a scientific puzzle for a long time. Frequently asked questions are: Is the characteristics of similar repeat patterns coherent in different proteomes across different life kingdoms? Could the functional and evolutionary roles of certain repeats correspond to their particular characters, such as position bias, GC content constrains and codon usage? How could the conserved functions of particular repeats have been evolved by selection? And what are the structure and sequence-based strategies to prevent repeats from aggregation?

The insufficient understanding of protein repeats is not only due to the difficulty of identification, but also because of the lack of integrated repository for large-scale investigation and comparison of repeats among a variety of proteomes across different kingdoms. To that end, we developed ProRepeat (http://prorepeat.bioinformatics.nl), which integrates non-redundant tandem repeats detected by several algorithms from the UniProt [69] and RefSeq [100] protein databases and offers powerful analysis tools for finding biologically interesting properties of query results. In addition, we also integrated ProRepeat with ProGMap—a tool we developed for the integration of annotation resources for protein orthology [101]. With this set-up, we will be making large-scale orthologous comparisons on protein repeats over a broad taxonomy range especially eukaryotes in the near future.

**Key Points**

- Amino acid repeats are abundant in protein sequences.
- They can be classified into different categories depending on the characters of the repeat units.
- Different amino acid repeat patterns imply different functional and evolutionary backgrounds.
- The three major approaches for detection of amino acid repeats are the self-comparison strategy, the pattern recognition strategy and the complexity measurement strategy.

**Acknowledgements**

This article is dedicated to the memory of Jack Leunissen, a former Professor of Bioinformatics at Wageningen University and Research Center in the Netherlands, and a former Editorial Board member of Briefings in Bioinformatics, who passed away on 14 May 2012. He conceived and supervised the Protein Repeat project, which was supported partly by the BioRange project of the Netherlands Bioinformatics Centre (NBIC). Thanks to Dr Lisa Mullan for critically reading and language editing of this manuscript. Thanks to Professor Lettie Lubsen for critically reading and informative comments of the manuscript.

## FUNDING

Funding for open access charge: BioRange Project of the Netherlands Bioinformatics Centre (financial support to H.L.).

## *References*

1. Luo H, Lin K, David A, *et al*. ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins. *Nucleic Acids Res* 2011;**40**:D394–9.
2. Cordeiro Y, Kraineva J, Gomes MPB, *et al*. The amino-terminal PrP domain is crucial to modulate prion misfolding and aggregation. *BiophysJ* 2005;**89**:2667–76.
3. Bjorklund AK, Ekman D, Elofsson A. Expansion of protein domain repeats. *PLoS Comput Biol* 2006;**2**:e114.
4. Sussman JL, Lin D, Jiang J, *et al*. Protein data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 1998;**54**:1078–84.
5. Chikenji G, Fujitsuka Y, Takada S. Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study. *Proc Natl Acad Sci USA* 2006;**103**:3141–6.
6. Ferreiro DU, Walczak AM, Komives EA, *et al*. The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLoS Comput Biol* 2008;**4**:e1000070.
7. Main ER, Lowe AR, Mochrie SG, *et al*. A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr Opin Struct Biol* 2005;**15**:464–71.
8. Ferreiro DU, Komives EA. The plastic landscape of repeat proteins. *Proc Natl Acad Sci USA* 2007;**104**:7735–6.

9. Simon M, Hancock JM. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol* 2009;**10**:R59.
10. Dunker AK, Brown CJ, Lawson JD, *et al*. Intrinsic disorder and protein function. *Biochemistry* 2002;**41**:6573–82.
11. Dunker AK, Silman I, Uversky VN, *et al*. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* 2008;**18**:756–64.
12. Dunker AK, Uversky VN. Drugs for 'protein clouds': targeting intrinsically disordered transcription factors. *Curr Opin Pharmacol* 2010;**10**:782–8.
13. Orr HT, Zoghbi HY. Trinucleotide repeat disorders. *Annu Rev Neurosci* 2007;**30**:575–621.
14. Buchanan G, Yang M, Cheong A, *et al*. Structural and functional consequences of glutamine tract variation in the androgen receptor. *Hum Mol Genet* 2004;**13**:1677–92.
15. Brown L, Paraso M, Arkell R, *et al*. *In vitro* analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. *Hum Mol Genet* 2005;**14**:411–20.
16. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 1987;**4**:203–21.
17. Warren ST. Polyalanine expansion in synpolydactyly might result from unequal crossing-over of HOXD13. *Science* 1997;**275**:408–9.
18. Hancock JM, Vogler AP. How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: implications for phylogeny reconstruction. *Mol Phylogenet Evol* 2000;**14**:366–74.
19. Alba MM, Laskowski RA, Hancock JM. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* 2002;**18**:672–8.
20. Salichs E, Ledda A, Mularoni L, *et al*. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet* 2009;**5**:e1000397.
21. Anan K, Yoshida N, Kataoka Y, *et al*. Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. *Mol Biol Evol* 2007;**24**:281–7.
22. Wu HT, Su YN, Hung CC, *et al*. Interaction between PHOX2B and CREBBP mediates synergistic activation: mechanistic implications of PHOX2B mutants. *Hum Mutat* 2009;**30**:655–60.
23. Enard W, Przeworski M, Fisher SE, *et al*. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 2002;**418**:869–72.
24. Mularoni L, Ledda A, Toll-Riera M, *et al*. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res* 2010;**20**:745–54.
25. Huntley MA, Clark AG. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. *Mol Biol Evol* 2007;**24**:2598–609.
26. Fondon JW3rd, Garner HR. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci USA* 2004;**101**:18058–63.
27. Caburet S, Vaiman D, Veitia RA. A genomic basis for the evolution of vertebrate transcription factors containing amino acid runs. *Genetics* 2004;**167**:1813–20.
28. Caburet S, Cocquet J, Vaiman D, *et al*. Coding repeats and evolutionary "agility". *Bioessays* 2005;**27**:581–7.
29. Finn RD, Mistry J, Tate J, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2010;**38**:D211–22.
30. Kajander T, Cortajarena AL, Main ER, *et al*. A new folding paradigm for repeat proteins. *J Am Chem Soc* 2005;**127**:10188–90.
31. Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* 2001;**11**:725–32.
32. D'Andrea LD, Regan L. TPR proteins: the versatile helix. *Trends Biochem Sci* 2003;**28**:655–62.
33. Gamsjaeger R, Liew CK, Loughlin FE, *et al*. Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends Biochem Sci* 2007;**32**:63–70.
34. Li J, Mahajan A, Tsai MD. Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry* 2006;**45**:15168–78.
35. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 2001;**310**:311–25.
36. Ponting CP, Mott R, Bork P, *et al*. Novel protein domains and repeats in *Drosophila melanogaster*: insights into structure, function, and evolution. *Genome Res* 2001;**11**:1996–2008.
37. Wright CF, Teichmann SA, Clarke J, *et al*. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* 2005;**438**:878–81.
38. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. *J Struct Biol* 2001;**134**:117–31.
39. Weatherall DJ, Clegg JB. Recent developments in the molecular genetics of human hemoglobin. *Cell* 1979;**16**:467–79.
40. Patthy L. Genome evolution and the evolution of exon-shuffling–a review. *Gene* 1999;**238**:103–14.
41. Han JH, Batey S, Nickson AA, *et al*. The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol* 2007;**8**:319–30.
42. Andrade MA, Ponting CP, Gibson TJ, *et al*. Homology-based method for identification of protein repeats using statistical significance estimates. *J Mol Biol* 2000;**298**:521–37.
43. Edgar RC, Sjolander K. COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* 2004;**20**:1309–18.
44. Karpenahalli MR, Lupas AN, Soding J. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC Bioinformatics* 2007;**8**:2.
45. Heringa J, Argos P. A method to recognize distant repeats in protein sequences. *Proteins* 1993;**17**:391–411.
46. Szklarczyk R, Heringa J. Tracking repeats using significance and transitivity. *Bioinformatics* 2004;**20(Suppl 1)**:i311–7.
47. Pellegrini M, Marcotte EM, Yeates TO. A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* 1999;**35**:440–6.
48. Soding J, Remmert M, Biegert A. HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res* 2006;**34**:W137–42.
49. Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* 2000;**41**:224–37.
50. Biegert A, Soding J. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* 2008;**24**:807–14.

51. Marsella L, Sirocco F, Trovato A, *et al*. REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics* 2009;**25**:i289–95.
52. Hochreiter S, Heusel M, Obermayer K. Fast model-based protein homology detection without alignment. *Bioinformatics* 2007;**23**:1728–36.
53. Palidwor GA, Shcherbinin S, Huska MR, *et al*. Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comput Biol* 2009;**5**:e1000304.
54. Li X, Kahveci T. A Novel algorithm for identifying low-complexity regions in a protein sequence. *Bioinformatics* 2006;**22**:2980–7.
55. Newman AM, Cooper JB. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* 2007;**8**:382.
56. Li L, Jin R, Kok PL, *et al*. Pseudo-periodic partitions of biological sequences. *Bioinformatics* 2004;**20**:295–306.
57. Liu X, Wang L. Finding the region of pseudo-periodic tandem repeats in biological sequences. *Algorithms Mol Biol* 2006;**1**:2.
58. Adebiyi EF, Jiang T, Kaufmann M. An efficient algorithm for finding short approximate non-tandem repeats. *Bioinformatics* 2001;**17(Suppl 1)**:S5–12.
59. Fankhauser N, Nguyen-Ha TM, Adler J, *et al*. Surface antigens and potential virulence factors from parasites detected by comparative genomics of perfect amino acid repeats. *Proteome Sci* 2007;**5**:20.
60. Gusfield D, Stoye J. Linear time algorithms for finding and representing all the tandem repeats in a string. *J Comput Syst Sci* 2004;**69**:525–46.
61. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;**48**:443–53.
62. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
63. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. *Atlas Protein Seq Struct* 1978;**5**: 345–51.
64. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;**89**: 10915–9.
65. Atchley WR, Zhao J, Fernandes AD, *et al*. Solving the protein sequence metric problem. *Proc Natl Acad Sci USA* 2005; **102**:6395–400.
66. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* 1990;**87**:2264–8.
67. Karlin S, Altschul SF. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci USA* 1993;**90**:5873–7.
68. Rabiner LR. A Tutorial on hidden markov-models and selected applications in speech recognition. *Proc IEEE* 1989;**77**:257–86.
69. Consortium TU. Ongoing and future developments at the universal protein resource. *Nucleic Acids Res* 2010;**39**: D214–9.
70. Alexandrov V, van Albada G, Sloot P, *et al*. GPU Accelerated Smith-Waterman. *Computational Science – ICCS 2006*. Berlin/Heidelberg: Springer, 2006;188–95.
71. George RA, Heringa J. The REPRO server: finding protein internal sequence repeats through the Web. *Trends Biochem Sci* 2000;**25**:515–7.
72. Romein JW, Heringa J, Bal HE. 'A Million-Fold Speed Improvement in Genomic Repeats Detection'. *Proceedings of the 2003 ACM/IEEE conference on Supercomputing*. Phoenix, AZ, USA: IEEE Computer Society, Phoenix, AZ, USA, 2003;20.
73. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 2005;**21**:951–60.
74. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
75. McLachlan AD. Analysis of periodic patterns in amino acid sequences: collagen. *Biopolymers* 1977;**16**:1271–97.
76. Lio P. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics* 2003;**19**:2–9.
77. Sweldens W. Wavelets: What next? *Proceedings of the IEEE* 1996;**84**:680–5.
78. Vo A, Nguyen N, Huang H. Solenoid and non-solenoid protein recognition using stationary wavelet packet transform. *Bioinformatics* 2010;**26**:i467–73.
79. Bishop CM. *Neural Networks for Pattern Recognition*. Oxford, New York: Clarendon Press, Oxford University Press, 1995.
80. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* 1996;**266**:554–71.
81. Wan H, Li L, Federhen S, *et al*. Discovering simple regions in biological sequences associated with scoring schemes. *J Comput Biol* 2003;**10**:171–85.
82. Shin SW, Kim SM. A new algorithm for detecting low-complexity regions in protein sequences. *Bioinformatics* 2005;**21**:160–70.
83. Shannon CE. The mathematical theory of communication. *1963*. MD Comput 1997;**14**:306–17.
84. Coronado JE, Attie O, Epstein SL, *et al*. Composition-modified matrices improve identification of homologs of saccharomyces cerevisiae low-complexity glycoproteins. *Eukaryot Cell* 2006;**5**:628–37.
85. Hamming RW. Error detecting and error correcting codes. *Bell Syst Tech J* 1950;**29**:147–60.
86. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 1966; **10**:707–10.
87. Groult R, Lonard M, Mouchard L. Speeding up the detection of evolutive tandem repeats. *Theor Comput Sci* 2004; **310**:309–28.
88. Hammock EA, Young LJ. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* 2005;**308**:1630–4.
89. Katti MV, Sami-Subbu R, Ranjekar PK, *et al*. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci* 2000;**9**:1203–9.
90. Bannen RM, Bingman CA, Phillips GN, Jr. Effect of low-complexity regions on protein structure determination. *J Struct Funct Genomics* 2007;**8**:217–26.
91. Karlin S. Statistical significance of sequence patterns in proteins. *Curr Opin Struct Biol* 1995;**5**:360–71.
92. Long M, Betran E, Thornton K, *et al*. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 2003;**4**:865–75.

93. Ohno S. *Evolution by Gene Duplication*. Berlin, New York: Springer-Verlag, 1970.
94. Crow KD, Wagner GP. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol* 2006;**23**:887–92.
95. Verstrepen KJ, Jansen A, Lewitter F, *et al*. Intragenic tandem repeats generate functional variability. *Nat Genet* 2005;**37**: 986–90.
96. Gibbons JG, Rokas A. Comparative and functional characterization of intragenic tandem repeats in 10 Aspergillus genomes. *Mol Biol Evol* 2009;**26**:591–602.
97. Rorick MM, Wagner GP. The origin of conserved protein domains and amino acid repeats via adaptive competition for control over amino acid residues. *J Mol Evol* 2010;**70**:29–43.
98. Haerty W, Golding GB. Genome-wide evidence for selection acting on single amino acid repeats. *Genome Res* 2010;**20**:755–60.
99. Haerty W, Golding GB. Low-complexity sequences and single amino acid repeats: not just "junk" peptide sequences. *Genome* 2010;**53**:753–62.
100. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;**35**:D61–5.
101. Kuzniar A, Lin K, He Y, *et al*. ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Res* 2009;**37**:W428–34.