Bioinformatics Advance Access published June 22, 2016

Bioinformatics, 2016, 1–7
doi: 10.1093/bioinformatics/btw319
Advance Access Publication Date: 9 June 2016
Original Paper

Sequence analysis

Revealing aperiodic aspects of solenoid proteins from sequence information

Thomas Hrabe, Lukasz Jaroszewski and Adam Godzik*

Department of Bioinformatics and Systems Biology, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 27, 2015; revised on April 28, 2016; accepted on May 13, 2016

Abstract

Motivation: Repeat proteins, which contain multiple repeats of short sequence motifs, form a large but seldom-studied group of proteins. Methods focusing on the analysis of 3D structures of such proteins identified many subtle effects in length distribution of individual motifs that are important for their functions. However, similar analysis was yet not applied to the vast majority of repeat proteins with unknown 3D structures, mostly because of the extreme diversity of the underlying motifs and the resulting difficulty to detect those.

Results: We developed FAIT, a sequence-based algorithm for the precise assignment of individual repeats in repeat proteins and introduced a framework to classify and compare aperiodicity patterns for large protein families. FAIT extracts repeat positions by post-processing FFAS alignment matrices with image processing methods. On examples of proteins with Leucine Rich Repeat (LRR) domains and other solenoids like proteins, we show that the automated analysis with FAIT correctly identifies exact lengths of individual repeats based entirely on sequence information.

Availability and Implementation: https://github.com/GodzikLab/FAIT.

Contact: adam@godziklab.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Almost 20%, of all plant and animal, including human, proteins are entirely built from or contain large domains that consist of multiple repeats of short sequence motifs (Andrade *et al.*, 2001). These repeated motifs typically contain between 3 and 40 amino acids and cannot form independent structures by themselves (Kajava, 2012). A subset of repeat proteins that folds into solenoid-like structures have motif lengths typically in the range of 20–40 amino acids. Since all repeats in any given protein are homologous, it is often assumed that their lengths would be constant and, even if not, the small differences in the length of individual motifs are usually treated as noise. However, when a few available 3D structures of solenoid proteins were analyzed, it was shown that small length variations of 1–5 amino acids between individual motifs can modify structure in subtle but functionally important ways. For instance, structures of the Leucine Rich Repeat (LRR) family are all known to fold into curved solenoid (Kobe and Kajava, 2001; Kajava, 1998), horseshoe-like structures and small motif length variations affect their local curvature (Matsushima *et al.*, 2005) and, hence, the specific shape of the binding cavity and their binding specificity. This feature of LRR proteins is being explored in *in silico* design of artificial solenoid-like proteins with desired shapes and curvatures (Bazan and Kajava, 2015; Park *et al.*, 2015). The interest in detecting and analyzing repeat proteins grew over the last decade as more functionally important groups of such proteins were discovered. This led to the development of many methods for their recognition and analysis based, among others, on self-alignment—RADAR (Heger and Holm, 2000), HMMs—HHRepID (Biegert and Söding, 2008), self-comparison approaches based on tiling (Parra *et al.*, 2013) or predefined databases and Fourier transformation—Repetita (Marsella *et al.*, 2009). Results of such structure analysis now available in dedicated resources such as the RepeatDB database

(Di Domenico *et al.*, 2013), as analyzed with Raphael (Walsh *et al.*, 2012). However, nearly all methods focused on the first step of the analysis—detecting the motifs—and not on identifying exact lengths and variations thereof.

At the same time, detecting variations in periodic signals in other fields, such as cardiographic signals or intense sunspot activities that oscillate with an approximate period of 11 years, but with significant variations, created a large repertoire of mathematical frameworks that can be applied to detect length variability in solenoid-like proteins (Jacobson, 2001; Scholkmann *et al.*, 2012). Here, we present a novel approach called *F*FAS-based *A*periodicity detection using *I*mage-processing *T*echniques (FAIT) for the in detail analysis of solenoid like proteins. FAIT is an adaptation of image-processing algorithms to the post-processing the graphic representation of results of the sensitive profile–profile alignment method FFAS developed previously in our group (Jaroszewski *et al.*, 2011). We analyze FAIT's performance comparing FAIT to annotations from our structure-based tool, ConSole (Hrabe and Godzik, 2014), since no standard benchmark exists that contains information about positions and lengths of individual repeat units for sufficiently large sets of proteins. Finally, we use FAIT to analyze variability of several large families of solenoid proteins.

1.1 Aperiodicity profiles

In our previous publication, we introduced a concept of aperiodicity profiles of solenoid-like protein structures to describe their precise variation in solenoid motif lengths (Hrabe and Godzik, 2014). These are plots of motif lengths, indicating the variability between individual solenoid units. Properties of repeat structures include the mean unit length λ_μ and the lengths of individual units λ_i, which can also be described as differences from the mean length. Differences from λ_μ are specific for individual repeats, describing how they vary from the average repeat motif length. They can be indicative for presence of binding sites or local curvature of the protein, distinguishing specific individual repeats from other repeats of the same family. For instance, aperiodicity in the TLR4 profile (2Z64) indicates the binding region to MD2 with variable electrostatic potential in several species (Anwar *et al.*, 2015). Analysis of these parameters allowed us to automatically highlight aperiodic regions in solenoid structures and identify their specific aperiodicity signatures.

Here, we first benchmark FAIT's sequence-derived aperiodicity profiles by comparing them to more-accurate structure-based methods. We also introduce a novel aperiodicity score, the area *A* under the aperiodicity profile curve, to use in a large-scale analysis of the aperiodicity in LRR proteins (see Section 2 for details). Flat aperiodicity profiles and hence low *A* scores indicate periodic structures, while profiles with many peaks and hence high *A* scores are an indicator for highly aperiodic structures. For instance, *A* allows us to automatically sort structures by their aperiodicity (Fig. 1). We use aperiodicity profiles and *A* scores as descriptors to benchmark FAIT and to analyze large protein families.

2 Methods

Detection of aperiodicity patterns with FAIT is based on a customized signal-processing pipeline where the FFAS-generated profile–profile scoring matrix *M* (Wilson, 1996) is enhanced with image-processing techniques to reveal positions of repeat units in the query sequence. The FFAS program (Jaroszewski *et al.*, 2011), similarly to other sequence alignment algorithms, returns the sequence alignment as the main output, but the full scoring matrix used to generate the alignment can also be exported. Visualization of such matrices was used in many early protein and nucleic acid alignment programs to manually identify the alignment in a method referred to as a dot-plot analysis (Vingron and Argos, 1991), but its popularity waned with improvement of the automated methods for alignment identification. Here, we identify and enhance repetitive patterns in a profile–profile scoring matrix using image-processing tools.

A scoring matrix *M* is an *n*-by-*m* matrix, with *n* being the length of the reference protein sequence, and *m* being the length of the query protein sequence. In a classical alignment problem, values at position *i,j* in *M* are the result of comparing query sequence residue *Q* at position *i* and the reference sequence residue R at position *j*:

$$M_{i,j} = \boldsymbol{B}(Qj, Rj) \tag{1}$$

where *B* is an amino acid substitution matrix. An optimal path through matrix *M*, i.e. the alignment, is typically found by dynamic programming (Wilson, 1996). This approach can be generalized to a profile–profile scoring matrix in which positions in the two profiles being compared are described by vectors rather than amino acids (Rychlewski *et al.*, 2000; Xu *et al.*, 2014). Values at position *i,j* in the scoring matrix are now determined using a matrix–vector product defined as:

$$M_{i,j} = \vec{v_{2,j}} \cdot \boldsymbol{B} \cdot \vec{v_{1,i}}, \tag{2}$$

where $\vec{v_{1,i}}$ is a vector describing a sequence variation at position *i*th in the reference protein sequence, ***B*** is the amino acid substitution matrix [BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992) in FFAS], and $\vec{v_{2,j}}$ is a vector describing a sequence variation at the *j*th position in the query sequence (Xu *et al.*, 2014).

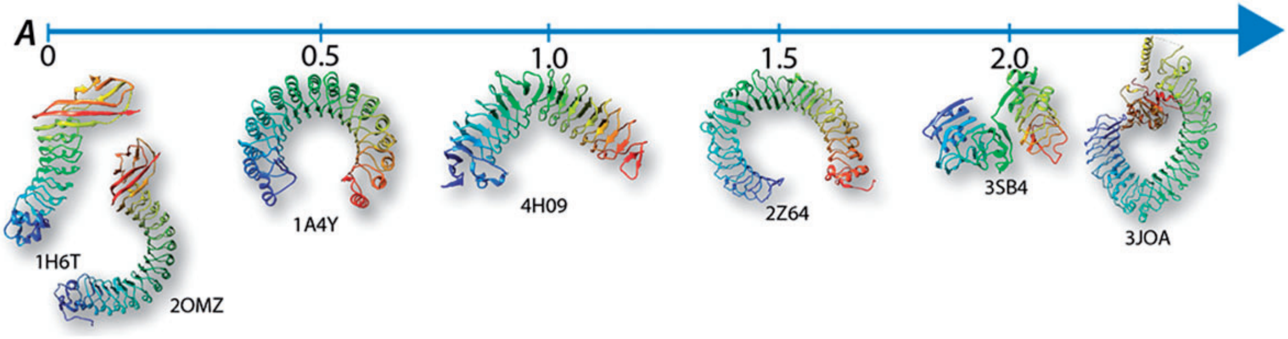

Fig. 1. Several LRR structures sorted by their respective profile area *A* score. Ideally periodic internalin structures (1H6T $A = 0$, 2OMZ $A = 0$) have a low *A* score. Ribonuclease inhibitors (1A4Y $A = 0.5$) show a larger degree of aperiodicity with their λ_i fluctuating between 28 aa and 29 aa. Structures with higher aperiodicity such as the TLR4 (2Z6A $A = 1.5$) or TLR5 (3J0A $A = 2.1$) have larger *A* scores

2.1 Reference sequence

The matrix analysis described below relies on the fact that starts and ends of individual solenoid units in the reference protein are known from structure analysis. For LRRs, we selected Listeria internalin (PDB id: 2OMZ, chain A) as the reference protein. The LRR domain in this structure is relatively long and consists of $n = 15$ LRR units. Internalins are ideally periodic LRR structures, so that:

$$\forall \lambda_i : \lambda_i = \lambda_\mu ; i \ \epsilon \ [1; N] \tag{3}$$

Hence, 2OMZ provides a robust reference for detecting aperiodicity in LRRs. The whole protein was manually truncated to the LRR domain, which was extracted from the 2OMZ structure, and all LRR unit start positions were labeled in the sequence based on the ConSole results. For Ankyrin and Armadillo sequences, we similarly extracted positions from the structures 1AWC-B and 3TJ3-A, respectively.

2.2 Matrix processing

Elucidating LRR aperiodicity from the profile–profile scoring matrix M (Fig. 2.1) involves several steps modeled after image processing. In the following description, we will refer to each matrix value at position i,j as pixel $p_{i,j}$.

1. We apply an averaging filter to amplify local similarity at each position (Fig. 2.2). The averaging filter calculates the mean value of 10 pixels along the diagonal fragment $[p_{i,j};\ p_{i+10,j+10}]$. High values detected for such diagonals indicate regions of high local similarity between corresponding regions of two compared sequences (or profiles).
2. Because repeat unit positions in the reference sequence are known, we now split the whole matrix into several submatrices SM with the size of $\lambda_{\text{reference}} \times m$, where $\lambda_{\text{reference}}$ is the length of the repeat unit in 2OMZ and m is the length of the analyzed protein.
3. We calculate the average matrix AM of all SM: $\text{AM} = \Sigma_i^N \text{SM}_i$ (Fig. 2.3). Strong similarities, now visible on some diagonals, indicate high local sequence similarities. This approach to reveal strong features in the matrix is similar to noise-reduction in signal or image processing as it is done by averaging images acquired by cryo electron microscopes for instance (Hrabe and Förster, 2011).
4. We use the discrete Laplace operator filter to amplify diagonals (Fig. 2.4). The Laplace filter is a standard image-processing technique to detect edges in an image (Forsyth and Ponce, 2003). The outcome of the convolution is matrix L (Fig. 2.4), which is formally defined as $L(x,y) = \frac{\partial^2 \text{AM}}{\partial x^2} + \frac{\partial^2 \text{AM}}{\partial y^2}$. The filter is applied by convoluting each position in AM with the kernel K, which is defined as

$$K = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{4}$$

5. Finally, we average over the first five rows of L to analyze similarity between the query sequence and the static LxLxx pattern of the internalin LRR unit (Fig. 2.5).

2.3 Signal processing

As a result of the steps described in the previous paragraph, we obtain a 1D signal of identical length as the query sequence length (Fig. 2e). High values of the signal indicate high similarity to starts of the LRR units in the reference protein. Hence, by extracting positions of these peaks, one can find start positions of units in the query sequence. We use the signal statistics to detect peaks in the

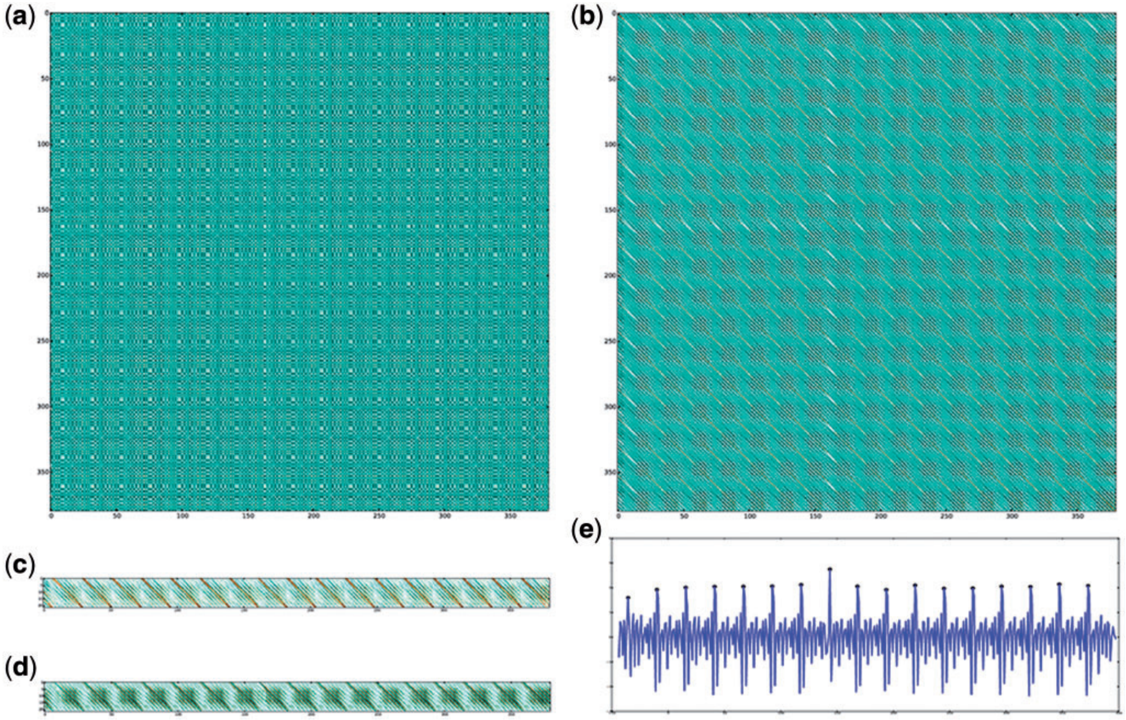

Fig. 2. Processing the profile–profile scoring matrix from its original state (1) to the final signal (5) where LRR units can be detected. The respective steps (1–5) are described in more detail in the main text. (5) Peaks in the final signal are indicators for the starting positions of the LRR units in the query sequence. The highest peak in the profile identifies positions of two engineered residues in the 2OMZ sequence

signal. We define significant peaks as values larger than 2σ of the signal over the query sequence length. The average distance between the detected peaks corresponds to the mean unit length λ_μ. The method accepts λ_μ as a valid unit length if it falls within the known unit length intervals [18;32] for LRRs, [23;43] for Ankyrin and [31;51] for Armadillo (Kajava, 2012; Li *et al.*, 2006; Tewari *et al.*, 2010). If λ_μ is not within that interval, the σ_{signal} factor used in unit detection is gradually decreased. If a valid value of λ_μ is still not found at 1σ, the query sequence is classified as a non-LRR.

Once λ_μ is known, we iteratively identify LRR units from peaks in the signal. Here, a numeric optimization method starts detection at the first peak and optimizes Equation (5) to detect individual units:

$$u_i = \operatorname{argmax}_{j\in[\lambda_\mu-\delta;\lambda_\mu+\delta]}(v_i + v_{i+j}) \tag{5}$$

Starting at the first peak with position i and a signal value v_i, the unit length j is sampled within the $[\lambda_\mu - \delta;\quad \lambda_\mu + \delta]$ interval, where δ specifies the maximum length of a unit. The only predefined parameter in FAIT and is set to $\delta = \lambda_\mu/2$. Once a highest scored unit is assigned to residue i, the method continues at position $i+j+1$.

2.4 Similarity of aperiodicity profiles

In order to compare two aperiodicity profiles, we calculate the profile similarity s with a sliding L^2 window. Here, we slide the shorter profile p_1 along the longer profile p_2 and detect the minimal Euclidean distance for all overlapping regions [Equation (6)], where N is the number of elements in the shorter sequence p_1 and M the number of elements in the longer sequence p_2:

$$s = \min_i^{M-N}\sqrt{\sum_{j=0}^{N}(p_{1,j} - p_{2,i+j})^2} \tag{6}$$

We experimented with a similar approach by calculating the correlation of p_1 and p_2, but the L_2 norm yields more stable results. The problem with normalized correlation is that the similarity of two vectors where one has all identical elements is not defined. Correlation becomes not meaningful for extremely periodic proteins, and hence L_2 is a more stable score in this application. Finally, each element in each profile is the respective difference λ_{u_i}–λ_μ measured in amino acids (aa); then, the unit of s is *aa*, too.

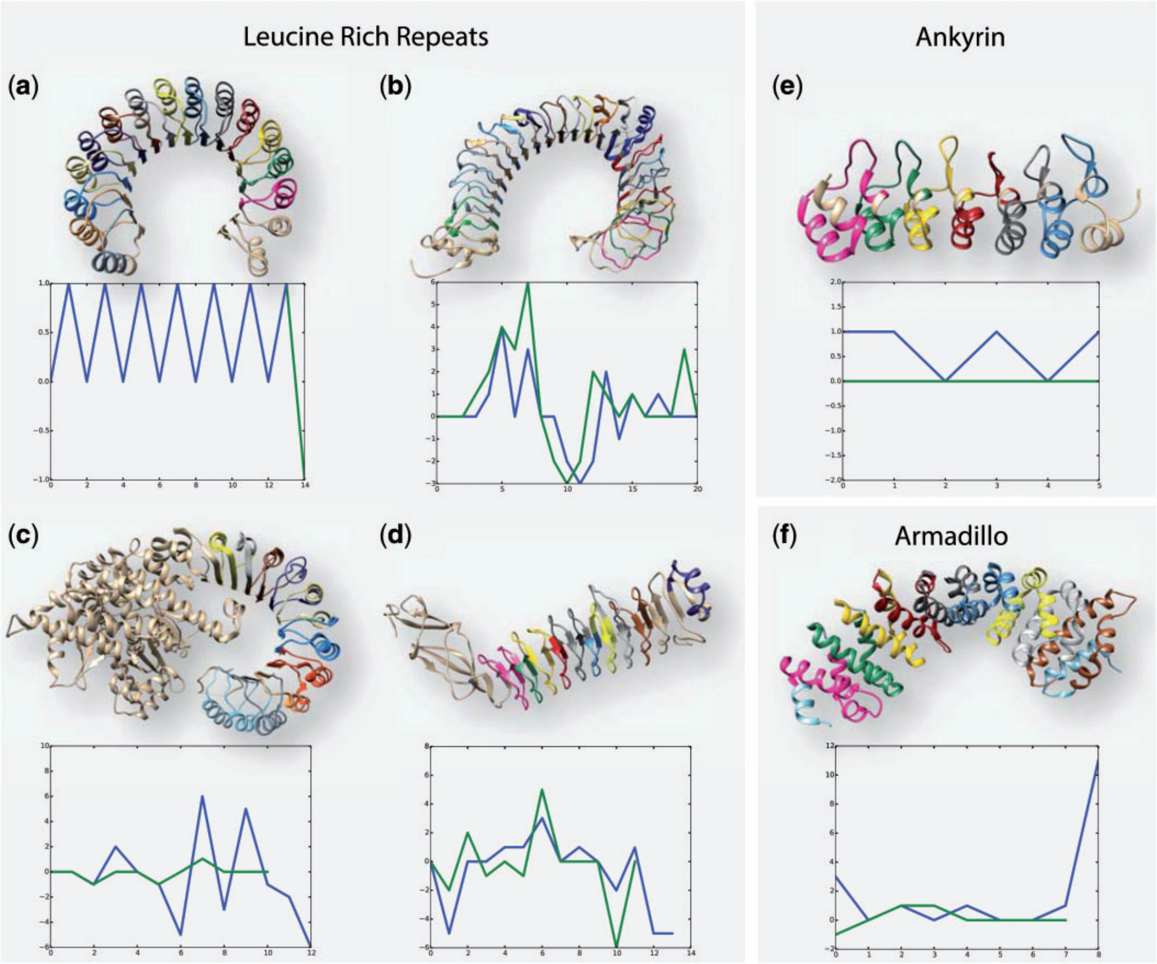

Fig. 3. Examples of aperiodicity profiles detected for solenoid structures with FAIT (blue curve) or ConSole (green curve). *X*-axis is the solenoid unit index (starting at 0) and *Y*-axis is the difference $\lambda_{u_i} - \lambda_\mu$ measured in amino acids (aa), solenoid units were colored as detected by FAIT. LRR box: (**a**) The aperiodicity profiles of the ribonuclease inhibitor (1A4Y-A) match with a profile similarity of 0.06, and both show the characteristic, sawtooth-like pattern. (**b**) Profiles of the mouse Toll-like receptor 4 (2Z64-A) shows a high aperiodicity, and both indicate aperiodicity at the same LRR units. Profile similarity equals 0.17. (**c**) Mouse Nod-like receptor 4 (4KXF-B) structure with its N-terminal LRR domain. Profile similarity equals 0.39. (**d**) Structure of bacterial LRR human gut symbionts with unknown function (4F0D-A). Unusually for LRRs, the structure is not curved and haves rather linear LRR domains with varying LRR unit lengths. The similarity between FAIT and ConSole profiles are 0.32. Ankyrin box: (**e**) 4UUC-A structure with a profile similarity of 0.33. Armadillo box: (**f**) 3TJ3-A structure with a profile similarity of 0.54. Highlighted repeat units in sequence of all six structures are presented in the Supplementary Material

2.5 Aperiodicity score *A*

The aperiodicity score A used for sorting structures in Figure 1 is defined as the area under the aperiodicity profile:

$$A = \frac{\sum_{i=0}^{N} |\lambda_\mu - \lambda_i|}{N} \tag{7}$$

where N is the length of the aperiodicity profile.

3 Results

3.1 Benchmark based on aperiodicity profiles from structure

In order to benchmark FAIT, we compared its sequence-based aperiodicity profiles against structure-based profiles from our structure-based method, ConSole. A total of 54 LRR, 148 Ankyrin and 80 Armadillo proteins were used to compare aperiodicity profiles derived directly from structure (Console) to analogous profiles predicted from sequence (FAIT). For all structures, we could identify the error distribution for the profiles as an average profile difference of 0.6 aa and a standard deviation of 0.41 aa, as detected by Equation (6). Moreover, profile lengths differed by 2 *aa* on average. Figure 3 depicts structures and their compared aperiodicity profiles. Our results indicate that aperiodicity profiles detected with FAIT are in good agreement with the ones detected from structures with ConSole.

Table 1. Benchmark of HHRepID and FAIT aperiodicity profiles against structure-based ConSole results

	Avg. L^2 distance	L^2 deviation
LRR		
FAIT	0.6 aa	0.41 aa
HHRepID	3.82 aa	14.57 aa
Ankyrin		
FAIT	1.53 aa	1.29 aa
HHRepID	1.56 aa	1.38 aa
Armadillo		
FAIT	0.92 aa	0.19 aa
HHRepID	4.9 aa	4.15 aa

3.2 Comparing FAIT and HHRepID

We compared FAIT to HHRepID repeat detector (Biegert and Söding, 2008). HHRepID is probably the conceptually most similar method from all the repeat detection algorithms, as the HMM–HMM comparison not only detects any repeat motifs in a sequence, but returns positions of repeat units as well. Hence, for direct comparison, HHRepID was applied to generate aperiodicity profiles for the identical LRR dataset used to benchmark FAIT.

Here, we show that FAIT clearly outperforms HHRepID (Table 1). For instance, for the ribonuclease inhibitor (PDB ID 1A4Y-A), HHRepID returns unit some lengths as long as 57 aa. This result is not satisfactory as it ignores the fact that these 57 aa contains two LRR units that vary in length between 28 aa and 29 aa, as shown by FAIT (Fig. 3). Similar effects could be observed for Armadillo repeats. Results for Ankyrin repeats, however, were in good agreement for both methods. This must be attributed to the low, intrinsic aperiodicity of the whole protein family, which we also observed in our previously study (Hrabe and Godzik, 2014).

While HHRepID performs well as a generalized repeat detector from sequence, it underperforms in detecting individual repeat units in sequence. HHRepID with other sequence-based methods such as RADAR, Repettita or Tiling were developed and benchmarked for the general purpose of detecting whole repeat domains in sequence, and were not specifically tailored to detect individual repeat units and their lengths (Luo and Nijveen, 2014).

3.3 Model bias analysis

In order to analyze FAIT's sensitivity to the type of reference used for analysis, we varied the reference sequence for detecting LRR units. Specifically, we used the Ankyrin (PDB: 1AWC-B) sequence as reference for FAIT analysis. Ankyrin proteins also form solenoid-like structures, but their sequence does not resemble any LRR pattern and their structures significantly differ from LRR structures. The difference between the Ankyrin-based FAIT and ConSole profiles detected by the sliding Euclidean score was significantly larger than using the LRR 2OMZ-A as reference. The average difference between aperiodicity profiles was 1.62 aa with a standard deviation of 2.67 aa.

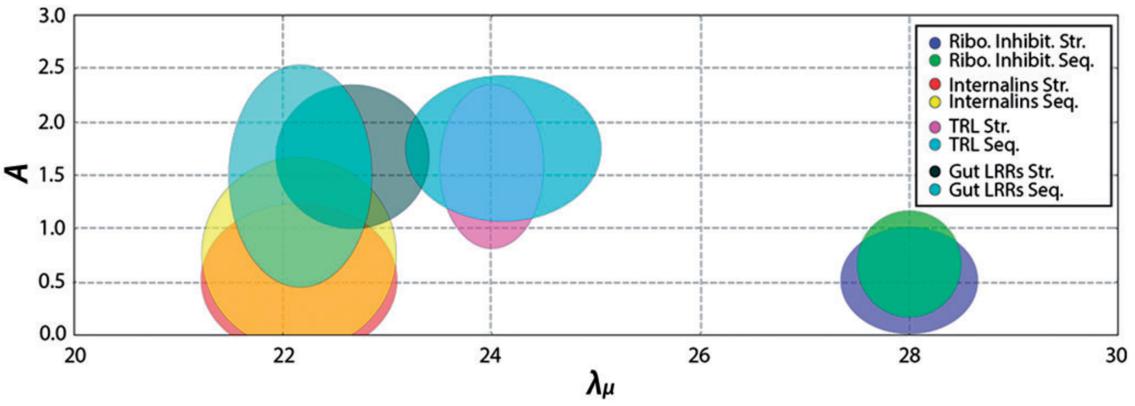

Fig. 4. Comparison of structure aperiodicities detected by ConSole (Str.) and FAIT (Seq.). The distribution of λ_μ and A score determines the position of each subfamily cluster in the plot. The cluster centers are determined by the mean values of λ_μ and A score and the cluster width by corresponding standard deviations in the cluster. Structures of the bacterial internalin family and the ribonuclease inhibitor family form clusters in the lower A-score regions, with the ribonuclease inhibitor cluster overlapping the NLR subfamily cluster (not shown). TLR structures generally have a higher aperiodicity (they form clusters in higher A-score regions). LRR structures from human gut show the largest aperiodicity of all subfamilies. The positions and widths of FAIT-based clusters are highly correlated with structure-based results

3.4 Pan-family LRR analysis

The intended application of FAIT was to analyze large protein families, with most proteins known only on the sequence level. Here, we analyzed various LRR subfamilies ranging from ideally periodic internalin proteins to highly aperiodic TLR proteins and generated aperiodicity profiles in each subfamily based entirely on sequence information. Aperiodicity profiles were generated for subfamily specific structures and sequences. Figure 4 shows a comparison of structure- and sequence-based aperiodicity analysis. The plot confirms our previous observations about existence of several subfamilies of LRR proteins and hence corroborates the robustness of FAIT: (i) human NLRs and ribonuclease inhibitors have a repeat-unit length of *28 aa* and generally a low aperiodicity, (ii) internalins are known to have short unit lengths of typically *22 aa*, and (iii) human TLRs cluster around a unit length of *24 aa* and have a larger aperiodicity than all previously discussed subfamilies. In addition, our results show that bacterial LRRs from human gut symbionts have the largest aperiodicity we observed in our analysis and are systematically different both from TLR and NLR-like proteins. Again, as shown in Figure 4, FAIT analysis of sequences is highly correlated with structure-based results from ConSole and validates the robustness of our method.

4 Conclusion

FAIT is a novel method developed to analyze variations in length of individual repeats in solenoid-like proteins, and it is the first method developed specifically to detect it only from sequences. It can, therefore, be used to analyze proteins from newly sequenced genomes and/or large families that were never structurally characterized. It uses information from the profile–profile alignment scoring matrix calculated with FFAS and extracts subtle variations of individual repeats with image-processing algorithms.

FAIT was benchmarked against an established structure-based method and shown to give results closely correlated with those obtained using structural data. This allows it to be used on much larger groups of proteins, as 3D structures of only a very small percentage of all repeat proteins are known. HHRepID is probably the sequence-based method with the most similar annotation detail, but as nearly all existing methods it was developed to detect repeat domains and not to specifically provide exact positions of individual units as FAIT. Hence, the major difference between existing methods and FAIT is the analysis of individual units, not the detection of repetitive domains in sequence.

Using a previously annotated protein as reference in FFAS comparison is the key feature of our approach, as it allows us to precisely detect individual repeats. Our results indicate that while varying references for cross-family experiments yields valuable results, to get the best results possible it is preferable to use a family-specific reference. Hence, using annotated reference protein seems to be a required step to increase the resolution in analyzing repeat sequences. We mainly demonstrate the proof of concept of our algorithm on the large LRR protein family, using a previously annotated, long and extremely periodic sequence. However, results on other solenoid-like protein families such as Ankyrin and Armadillo repeats indicate that the FAIT procedure can be readily adapted to other protein families with one previously annotated sequence and known length of repeat motifs.

In FAIT, we also introduce a novel framework using the *A* score and aperiodicity profiles to analyze large groups of solenoid-like proteins and group them by their variability, allowing us to investigate structural and sequence flexibility in more detail than currently possible on the PDBFlex.org webserver (Hrabe *et al.*, 2015). We performed such analysis to visualize aperiodicity in subfamilies of LRRs. These results also provided additional validation of our method since we compared clustering according to FAIT with structure-based clustering and showed that they are highly correlated. We are currently conducting a larger study in which we use FAIT to analyze the evolution of aperiodicity in NLR and TLR families.

As all of our previous methods, FAIT is currently available as an open-source package, and will be extended in a future web-server implementation that automatically distinguishes between multiple solenoid families.

Acknowledgements

We would like to thank Andreas Biegert for HHRepID installation support and Cindy Cook for carefully reading the manuscript.

Funding

This work has been supported by the National Institute for Health RO1 grant GM101457

Conflict of Interest: none declared.

References

Andrade,M.A. *et al.* (2001) Protein repeats: structures, functions, and evolution. *J. Struct. Biol.*, **134**, 117–131.

Anwar,M.A. *et al.* (2015) Insights into the species-specific TLR4 signaling mechanism in response to Rhodobacter sphaeroides lipid A detection. *Sci. Rep.*, **5**, 7657.

Bazan,J.F. and Kajava,A.V. (2015) Designs on a curve. *Nat. Publ. Gr*, **22**, 103–105.

Biegert,A. and Söding,J. (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, **24**, 807–814.

Di Domenico,T. *et al.* (2013) RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Res*, **42**, D352–D357.

Forsyth,D.A. and Ponce,J. (2003) *Computer Vision*. 1st edn. Pearson, New York.

Heger,A. and Holm,L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41**, 224–237.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 10915–10919.

Hrabe,T. *et al.* (2015) PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res.*, **44**, D423–D428.

Hrabe,T. and Förster,F. (2011) Structure determination by single particle tomography. *Encycl. Life Sci.* doi:10.1002/9780470015902.a0023175

Hrabe,T. and Godzik,A. (2014) ConSole: using modularity of contact maps to locate Solenoid domains in protein structures. *BMC Bioinformatics*, **15**, 119.

Jacobson,M.L. (2001) Auto-threshold peak detection in physiological signals. *Eng. Med. Biol. Soc. 2001. Proc. 23rd Annu. Int. Conf. IEEE*, **3**, 2194–2195.

Jaroszewski,L. *et al.* (2011) FFAS server: novel features and applications. *Nucleic Acids Res.*, **39**, W38–W44.

Kajava,A.V. (1998) Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol.*, **277**, 519–527.

Kajava,A.V. (2012) Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.*, **179**, 279–288.

Kobe,B. and Kajava,A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.*, **11**, 725–732.

Li,J. *et al.* (2006) Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry*, **45**, 15168–15178.

Luo,H. and Nijveen,H. (2014) Understanding and identifying amino acid repeats. *Brief. Bioinf.*, **15**, 582–591.

Marsella,L. *et al.* (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*, **25**, i289–i295.

Matsushima,N. *et al.* (2005) Structural analysis of leucine-rich-repeat variants in proteins associated with human diseases. *Cell. Mol. Life Sci.*, **62**, 2771–2791.

Park,K. *et al.* (2015) Control of repeat-protein curvature by computational protein design. *Nat. Struct. Mol. Biol. Mol. Biol.*, **22**, 167–174.

Parra,R.G. *et al.* (2013) Detecting repetitions and periodicities in proteins by tiling the structural space. *J. Phys. Chem. B*, **117**, 12887–12897.

Rychlewski,L. *et al.* (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.

Scholkmann,F. *et al.* (2012) An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals. *Algorithms*, **5**, 588–603.

Tewari,R. *et al.* (2010) Armadillo-repeat protein functions: questions for little creatures. *Trends Cell Biol.*, **20**, 470–481.

Vingron,M. and Argos,P. (1991) Motif recognition and alignment for many sequences by comparison of dot-matrices. *J. Mol. Biol.*, **218**, 33–43.

Walsh,I. *et al.* (2012) RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics*, **28**, 3257–3264.

Wilson,S.R. (1996) Introduction to computational biology: maps, sequences and genomes. *Stat. Med.*, **15**, 2264.

Xu,D. *et al.* (2014) FFAS-3D: Improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics*, **30**, 660–667.