This tutorial provides an introduction to Tandem Repeats (TR) in proteins. This is an advanced course which requires familiarity with the Pfam [4] database and profile Hidden Markov Models, described here [5]. TRs are a distinct class of protein modules that are found in many natural proteins. This course offers a description of TRs and their importance in biology. Furthermore, we describe one method of TR detection which is based on sequence homology, as well as discussing the challenges in characterizing them.

**Learning objectives:**

- List features of tandem repeats
- Describe the importance of repeats in biology
- List the differences between repeats and domains
- Describe the process used to create repeat entries in Pfam

**What are repeats?**

Repeats are functionally versatile structural units that occur in all domains of life, with a higher level of prevalence in eukaryotic sequences [1] [6]. These repetitive units usually occur adjacent to one another, and hence are called tandem repeats (TR). Figure 1 (below) shows an example of a TR, specifically the WD40 repeat unit appearing tandemly in one protein sequence [2] [6]. TRs possess regular secondary structures which collectively form three-dimensional assemblies of varied size and function. The interactions between the individual repeat units help to stabilise the fold. Multiplicity is one of the defining features of TRs and as a consequence they constitute a separate class of protein element, distinct from domains and motifs which unlike repeats, occur stably in isolation [3] [6].

**Figure 1:** Example of a WD40 tandem repeat unit appearing multiple times in one protein sequence.
Repeats in biology

An estimated one third of human proteins contain tandem repeats, and TR- proteins are involved in a variety of cellular activities and human diseases including neurodevelopmental disorders and cancer. The WD40 repeat shown above is one of the most commonly detected repeats in human proteins. It is composed of ~ 40 amino acids with well-conserved Trp (W) and Asp (D) residues, hence the name. The WD40 beta-fold is shown to act as a recruitment platform for proteins, DNA, and RNA, as such it is involved in a diverse array of cellular pathways. Notably, several WD40 repeat-containing proteins are involved in disease-associated pathways such as CDC20 protein, which is linked to the ability of human glioblastoma stem cells to generate brain tumours [4 [6]].

Another widespread example of TRs are leucine rich repeats (LRR). LRRs occur in proteins that are involved in a huge range of cellular activities including protein-protein interactions, cell adhesion, cell signalling, cellular trafficking, platelet aggregation, RNA processing, cellular polarization, neuronal development, apoptotic signalling, bacterial adhesion and invasion, as well as immunological response to pathogens and pathogen recognition. NOD-like receptors sense molecular determinants from a wide range of pathogens through their LRRs. Three polymorphisms in the LRR region of NOD2 have been found to directly associate with Crohn’s disease which might contribute to defective sensing of microbial components, leading to the inhibition of NOD2 dimerization and proper activation of the NF-kB in monocytes necessary for pathogen clearance [5 [6]].

Repeats vs domains

As stated, repeats are a distinct class of protein element. In comparison with domains, individual repeats are typically shorter and are unlikely to fold in isolation. While the overall tertiary structures are conserved, the sequence conservation of the repeat can be extremely variable, with some copies of the repeat hardly recognisable compared to the consensus.

The presence of short repeat units in proteins confers greater opportunity for evolution compared to domains. For example, repeat elements can expand and contract frequently in terms of unit numbers, especially in "open" structures where both termini are distant from each other, see Figure 2. Theoretically there is no limit to how many repeat units can be added on either termini in "open" structures, since there is no steric impediment [4] [6].

Figure 2: Armadillo repeat structure in Homo sapiens, adapted from PDBe 4wv6 [7].
Problems detecting repeats

Detection of repeat units is a challenge for a number of reasons:

1. It is difficult to detect individual repeat units because on average they are relatively short (< 60 aa)
2. It is problematic to define the boundaries of repeats
3. There is considerable sequence divergence among units of the same repeat (i.e. repeats within a single protein or members of a protein family [8] might degenerate at sequence level)
4. Repeats often correspond with areas of low compositional complexity such as disordered regions. (i.e. separating the two different signals can be challenging)

Despite of the aforementioned problems with detecting repeats, there is a continuous effort in identifying and classifying them. One approach used for identifying repeats is based on sequence homology detection, because sequence is conserved among evolutionary related proteins, thus it is possible to use profile hidden Markov models (HMMs) to identify repeats.

Profile HMMs are described in detail in our Pfam Database: Creating Protein Families tutorial [5].

Creating a Repeat entry

Creating a Repeat type entry in Pfam [9] broadly follows the same process as creating other entry types, with the exception of threshold adjustment and boundaries determination. These two steps pose a more arduous task in cases of repeat units compared to domains. This is due to the sequence divergence and short length of repeat sequences.

Detecting repeats by profile HMMs involves careful, manual curation [10] and fine tuning of different parameters (allowance of inserts, deletes and amino acid [11] mismatches in the seed alignments, E-values and bit-scores for profile HMMs) in order to account for the intricacies of repeats. The process is explained in more detail in the next sessions (Threshold adjustment and Boundaries determination).

Threshold adjustment

In order to capture all examples of a repeating unit in a sequence, the inclusion threshold cut-offs (i.e. those sequence regions deemed to be part of the family) have to be carefully defined.

As described in our tutorial Pfam Database: Creating Protein Families [12], there are two different values in the HMMER output that we consider, E-values and bit scores. For each of these, there are two scores reported for every match, one for the individual hit and one for the sequence. The sequence bit score is the sum of all hit bit scores found between the sequence and the profile HMM. For most Pfam entries of type domain, the sequence and hit thresholds are very similar, around 25.0 bit, as you do not typically find domains repeated within a sequence.

In the case of repeats, we often define very different sequence and hit bit-scores. Consider a sequence with 6 identical repeats. Each individual repeat may each score 15 bit, meaning the sequence has 90 bit score. However, we have already highlighted that repeat sequence motifs degenerate, so in reality we usually have some repeats that match the consensus very well and others that do not. Therefore, it is more likely that our 6 repeat sequence contains imperfect repeats and we might find that the bit scores vary between 4 and 35 (see example in Figure)
3). Setting a bit sequence of 50 and bit hit threshold of 4 would allow us to capture all instances of the repeat. Having a stringent sequence threshold requires the repeat to be observed multiple times, which prevents false positives from single matches that would happen at a threshold as low as 4 bit.

![Repeat units and Bit Score](image)

**Figure 3:** To detect all repeat units in one sequence, we set a stringent bit sequence score and a low bit hit score.

### Boundaries determination

While we have talked about defining the inclusion thresholds in Pfam, we have not talked about defining the actual repeating motif. In the case of repeat units, the hit boundaries can be ambiguous due to their short length and sequence divergence. Furthermore, the repeats near the termini of a sequence are less like those in the middle, as such defining the correct repeating unit can be challenging. However, in the case of our perfect 6 repeat protein, this is actually very easy.

Plotting the sequence against itself using software such as dotter [13] will reveal the start and end of each repeat. However, in reality it is often significantly more difficult to see where one repeat stops and another begins due to the divergence in sequence.

In Pfam, we adopt a number of different approaches to define the correct boundaries:

1. Using known structures to determine the repeating unit
2. Using dot plots with different sequences plotted against each other
Building profiles and shifting the boundaries until an optimal solution is found

In the perfect scenario, all Pfam repeat entries would model a single repeat unit (left in Figure 4), and for some repeats, such as the TAL effector [14], this is the case. However, single repeats have the drawback that despite our best optimisation of inclusion thresholds, we cannot detect every instance of the repeat, in every sequence. As a result, you sometimes detect a mosaic pattern of hits across the sequence.

In some cases the repeat single units are impossible to significantly distinguish from random noise (a problem exacerbated by the entropy weighting scheme employed by HMMER3), and so a sequence containing multiple repeats (usually 2 or 3 single copies) is used to construct the profile HMM (centre in Figure 4). While this has the advantage of typically covering more of the tandem repeat in the sequence, it results in a series of overlapping matches, that may still only reflect a partial match to the profile HMM. In such cases, the graphical representations of the matches on the Pfam website may appear as strange patterns of matching lengths.

Finally, some repeats are modelled as the whole set of tandem repeats (right in Figure 4). While it may seem that this should capture the entire repetitive region, the local-local matching search algorithm [15] means that multiple partial matches may occur, separated by instances of very poorly conserved repeats.

As stated previously, it can be useful to compare the profile HMM to known structures, to help determine the repeating unit. However, structure identification of proteins with tandem repeats can itself be difficult due to the nature of the repeat region being intrinsically unstructured.

Figure 4: Modelling of repeat units. In reality, the N and C termini repeat units can have less sequence similarity with the central units. This is due to fewer steric restrictions at the termini and also because they may have a role in capping the repeat structure. Therefore, the Pfam family model often represents multiple copies of the repeat units which offers an increased selectivity of the search model compared to a single repeat unit.

Clustering of entries

Structural properties of proteins are often more conserved than sequence. Therefore, a single profile HMM is often insufficient to model an entire, diverse, superfamily [16] of structurally related proteins. In Pfam [9] there is a hierarchal level of classification which groups evolutionary related entries in to sets, termed Clans.

The relationship between entries in a Clan may be defined by:
sequence similarity (whilst still originating from a common ancestor [17]

- similarity of known three-dimensional structures
- functional similarity
- and/or similarity between their profile HMMs, as determined by algorithms such as HHsearch

While the majority of Pfam Clans are groupings of domains and families, they have also been used to group Repeat type entries. Clans like the Pentapeptide [18] repeats include profile HMMs (Pfam families) composed of different numbers of repeat units which are known to share a common structural topology. In other cases Clans can include repeat units as diverse as Tetratricopeptide-like repeats (TPR [19]) which differ not only in sequence length but also in overall structure (whilst maintaining the same fold). Consequently, Pfam entries representing repeats within a Clan may be very different as shown in Figure 5.

<table>
<thead>
<tr>
<th>Pfam Accession</th>
<th>Pfam ID</th>
<th>UniProt ID</th>
<th>Domain architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF00515</td>
<td>TPR_1</td>
<td>W7AJR4</td>
<td></td>
</tr>
<tr>
<td>PF13371</td>
<td>TPR_9</td>
<td>U6MH8</td>
<td></td>
</tr>
<tr>
<td>PF14559</td>
<td>TPR_19</td>
<td>T2G7R6</td>
<td></td>
</tr>
<tr>
<td>PF14853</td>
<td>Fis1_TPR</td>
<td>W2TCN0</td>
<td></td>
</tr>
<tr>
<td>PF09976</td>
<td>TPR_21</td>
<td>Z5XQH3</td>
<td></td>
</tr>
<tr>
<td>PF12688</td>
<td>TPR_5</td>
<td>U5W915</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: TPR Clan (Pfam CL0020) contains different entry types to model TPR repeat units.

Summary

Tandem repeats (TR) in protein sequence and structure, are a distinct class compared to domains and motifs that may be present in a single or multiple fashion in each protein. Their detection is challenging due to the relatively short size and considerable sequence divergence between their units. Due to their high prevalence in human proteins and potential roles in a myriad of human diseases, there are considerable efforts in identifying and classifying them. One approach is based on protein sequence analysis which uses detection algorithms based on homology such as profile Hidden Markov Models (HMM). Profile HMMs offer an improvement in sequence detection methods compared to sequence-sequence comparison by utilising sequence-profile comparison. This allows greater sensitivity for detection of highly divergent repeat units in protein sequences. For more information on profile HMMs please refer to our tutorial Pfam Database: Creating Protein Families [5].

Get help and support on Pfam

Support

You can reach us via the Pfam helpdesk, click pfam-help [at] ebi.ac.uk (subject:
Modelling Protein Repeats) (here)

The Pfam Consortium
For a list of the current Pfam team members, see our team page [20].

References


Contributors
Sara El-Gebali [1]
EMBL-EBI
Scientific Database Curator

Dr. Sara El-Gebali is a Curator for Pfam Database at the European Bioinformatics Institute (EMBL-EBI). She did her undergraduate studies in Molecular Biology at the University of Lund, Sweden before moving to Queen Mary School of Medicine and Dentistry to gain an MSc degree in Molecular Pathology and Genomics. After working in the Medical Research Institute (MRI) laboratories at the University of Dundee and at the Institute of Oncology Research in Southern Switzerland (IOSI), she joined the University of Bern, Switzerland where she completed her PhD in Biochemistry and Molecular Medicine. She worked as a Database Curator for EMBO in Heidelberg, Germany before joining the Protein Families team at EMBL-EBI in 2016.

Lorna Richardson [2]
EMBL-EBI
Scientific Curator

Lorna is a scientific curator in the sequence families team at EMBL-EBI. She is responsible for InterPro, Pfam and Genome Properties curation, as well as outreach and training. Prior to joining the sequence families team at the EBI, she worked with the eMouseAtlas project at the University of Edinburgh, curating 3D anatomy and gene expression data. She has been working in the field of biological database curation since 2001.
Rob Finn [3]

EMBL-EBI
Team Leader, Sequence Families

Dr Rob Finn leads EMBL-EBI’s Sequence Families team, which is responsible for the InterPro, Pfam and Rfam data resources as well as the HMMER web searches. Rob joined EMBL-EBI in 2014 from the HHMI Janelia Research Campus in the US, where he led a group that designed fast, web-based, interactive protein sequence searches and annotations. Between 2001-2010, he was the project leader for Pfam at the Wellcome Trust Sanger Institute in the UK. Rob’s academic background is in microbiology and he holds a PhD in biochemistry from Imperial College, London.

Your feedback

Please tell us what you thought about this Quick tour. Your feedback is invaluable and helps us to improve our courses and enhance your learning experience.

Source URL: https://www.ebi.ac.uk/training/online/course/repeats-pfam

Links
[1] https://www.ebi.ac.uk/training/online/trainers/selgebali_9606
[2] https://www.ebi.ac.uk/training/online/trainers/lornar_8729
[3] https://www.ebi.ac.uk/training/online/trainers/rdf
[5] https://www.ebi.ac.uk/training/online/course/pfam-database-behind-scenes
[7] https://www.ebi.ac.uk/pdbe/entry/pdb/4vw6
[8] https://www.ebi.ac.uk/training/online/glossary/protein-family
[9] https://www.ebi.ac.uk/training/online/glossary/pfam
[10] https://www.ebi.ac.uk/training/online/glossary/curation
[11] https://www.ebi.ac.uk/training/online/glossary/amino-acid
[12] https://www.ebi.ac.uk/training/online/course/pfam-database-creating-protein-families
[14] https://pfam.xfam.org/family/PF03377
[15] https://www.ebi.ac.uk/training/online/glossary/algorithm
[16] https://www.ebi.ac.uk/training/online/glossary/superfamily
[17] https://www.ebi.ac.uk/training/online/glossary/ancestor
[21] https://europepmc.org/abstract/MED/10512723
[22] https://europepmc.org/abstract/MED/11551174
[23] https://europepmc.org/articles/PMC3402919
[25] https://europepmc.org/abstract/MED/29483781
[26] https://europepmc.org/articles/PMC4918410
[27] https://europepmc.org/abstract/MED/26442257
[28] https://europepmc.org/abstract/MED/9918945
[29] https://europepmc.org/abstract/MED/26673716