Phylogenetics of Tandem Repeats with Circular HMMs: A Case Study on Armadillo Repeat Proteins

Spencer Bliven[1,2,*] **Maria Anisimova**[1,2]

[1] *Institute for Applied Simulation, Zurich University of Applied Sciences, Wädenswil, Switzerland* [2] *Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland*

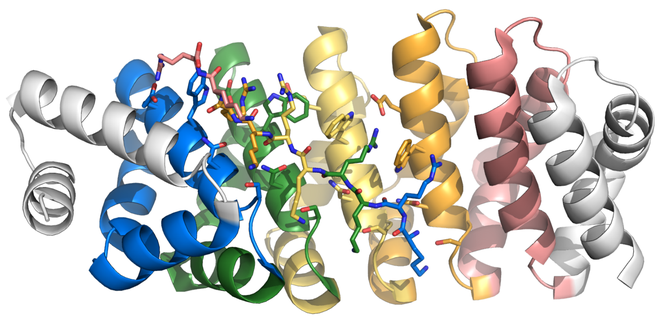
Abstract

Tandem repeat proteins are characterized by multiple sequential copies of repeats with significant structural or sequence similarity. Tandem repeats evolve via repeat expansion, duplication and loss, and many protein families exhibit very diverse repeat counts. Identifying the complex relationships between homologous proteins and between individual repeats is a challenging task. Using tools developed in our group, we present a detailed phylogenetic analysis of the repeats in the Armadillo Repeat Protein (ArmRP) family.

The ArmRP family is very diverse, appearing throughout the eukaryotes and having a wide range of functions. They are well characterized structurally, with ~42 amino acid repeats forming three alpha-helices which assemble into a solenoid structure. ArmRP are exciting candidates for protein design, as they have been shown to bind peptides in a modular manner (Reichen 2016).

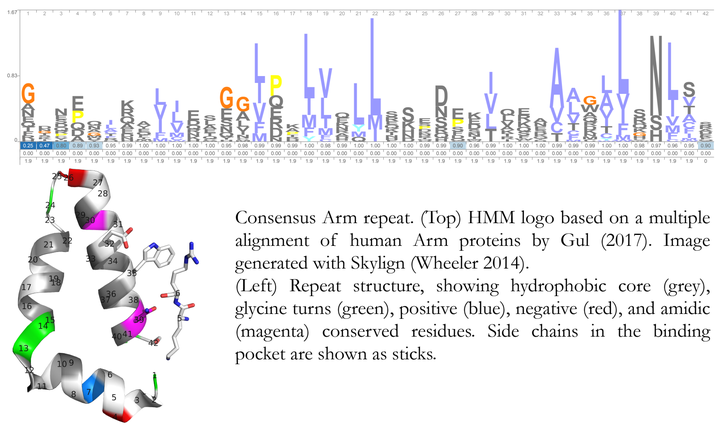
Phylogenetic analysis of tandem repeats has several unique challenges. Identifying homologous regions is complicated by the repetitive nature of the sequence, which can cause register shifts when applying standard alignment tools. We surmount this problem using the Tandem Repeat Annotation Library (TRAL), a tool for accurately identifying repeats using circular profile hidden Markov models (Schaper 2015). After constructing a multiple alignment of the repeats of ArmRP representatives, we infer a phylogenetic tree relating the different ArmRP and use it to analyze the conservation and diversification patterns through evolution, based on the information about tandem repeat number, order and their distribution on phylogenies.

Armadillo Repeat Proteins



Designed ArmRP $Y_{III}M_5A_{II}$ bound to a $(KR)_5$ peptide [5AEI]

Armadillo repeat proteins (ArmRP) are composed of ~40 amino acid repeats forming a alpha solenoid structure. Each repeat contains three alpha helices, with a hydrophobic core and conserved binding residues along the third helix.



Consensus Arm repeat. (Top) HMM logo based on a multiple alignment of human Arm proteins by Gul (2017). Image generated with Skylign (Wheeler 2014).
(Left) Repeat structure, showing hydrophobic core (grey), glycine turns (green), positive (blue), negative (red), and amidic (magenta) conserved residues. Side chains in the binding pocket are shown as sticks.

ArmRP bind extended peptides, which often are disordered prior to binding. Two amino acids binding each repeat. The conserved Asn at position 39 of the repeat typically forms a hydrogen bond to the peptide backbone. Specificity is conferred by a shallow binding pocket between helix 3 of adjacent repeats, which binds the peptide side chain. However, many ArmRP show weak specificity and engage multiple partners.

References

Gul, I. S., Hulpiau, P., Saeys, Y., & van Roy, F. (2017). Metazoan evolution of the armadillo repeat superfamily. Cellular and Molecular Life Sciences : CMLS, 74(3), 525–541. http://doi.org/10.1007/s00018-016-2319-6

Nguyen, L.-T., Schmidt, H. A., Haeseler, von, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution, 32(1), 268–274. http://doi.org/10.1093/molbev/msu300

Reichen, C., Hansen, S., Forzani, C., Honegger, A., Fleishman, S. J., Zhou, T., et al. (2016). Computationally Designed Armadillo Repeat Proteins for Modular Peptide Recognition. Journal of Molecular Biology, 428(22), 4467–4489. http://doi.org/10.1016/j.jmb.2016.09.012

Schaper, E., Korsunsky, A., Pečerska, J., Messina, A., Murri, R., Stockinger, H., et al. (2015). TRAL: tandem repeat annotation library. Bioinformatics, 31(18), 3051–3053. http://doi.org/10.1093/bioinformatics/btv306

Wheeler, T. J., Clements, J., & Finn, R. D. (2014). Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics, 15(1), 7. http://doi.org/10.1186/1471-2105-15-7

Poster presented at SIB Days 2018 (Biel, CH)

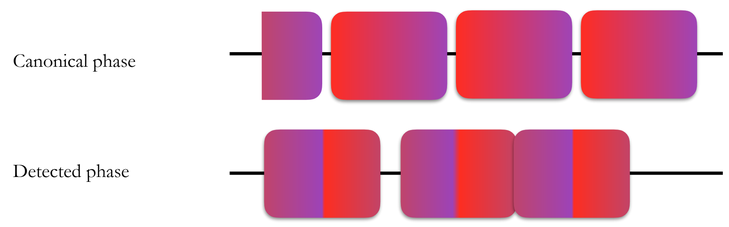
This research was supported by COST Action BM1405 and COST Switzerland SEFRI project IZCNZ0-174836.

TRAL availability

http://elkeschaper.github.io/tral/

Tandem Repeat Annotation

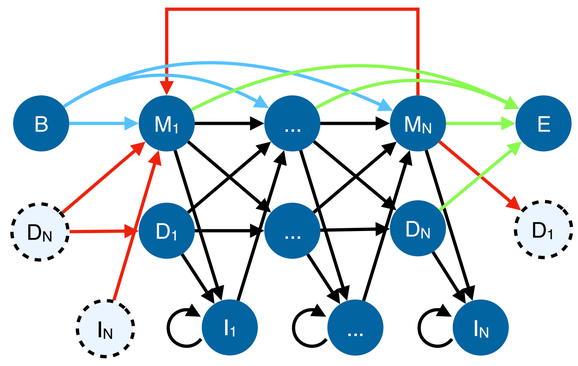
One issue with identifying tandem repeats is that the first and last repeats are often truncated or atypical. With standard sequence alignment methods, this can lead to missing partial repeats and incorrect phase identification:



(Top) Protein with 3.5 repeats with repeat boundaries assigned from literature. (Bottom) Incorrect tandem repeat identification, with non-standard boundaries, a missing partial repeat, and overlapping repeats.

Additionally, misalignments near the repeat boundaries can cause overlapping repeats which must be reconciled.

These issues are resolved in the **Tandem Repeat Annotation Library (TRAL)** (Schaper 2015). The library makes use of circular profile hidden Markov models (cpHMM) to capture TR profiles.

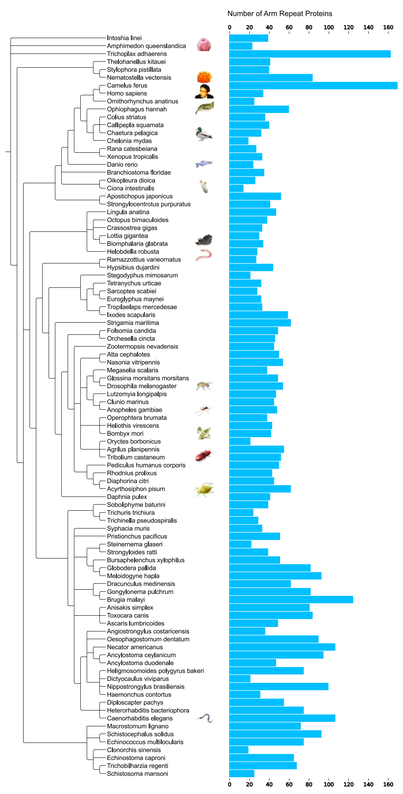


Circular TR sequence profile HMM (cpHMM). The alignment starts in any position within the TR with equal probability (blue). States representing alignment matches, insertions and deletions are included as in the HMMER model (black). Additional transitions are added from the end of a repeat back to the beginning (red) to align multiple repeats. The alignment may end at any position with equal probability (green).

The cpHMM allows the TR region to start and end at any point in the repeat, making the phase inconsequential. Additionally, it prevents overlaps and provides a rigorous statistical model for dealing with insertions between and within repeats.

TRAL contains functionality for constructing a cpHMM from a set of possibly inconsistent or overlapping seed TR hits.

ArmRP Prevalence

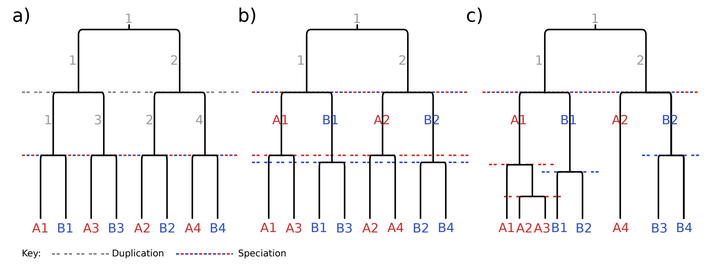


Prevalence of ArmRP members across metazoan species. ArmRP proteins were detected using TRAL across a 94 metazoan species. Between 14 and 170 proteins were detected containing at least two consecutive repeats.

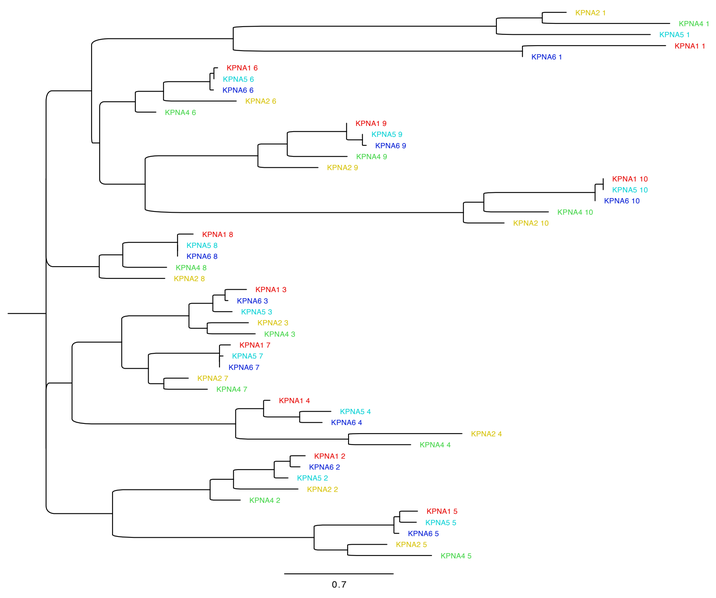
Species Icons curtesy of the TogoDB (http://togodb.biosciencedbc.jp)

Evolution

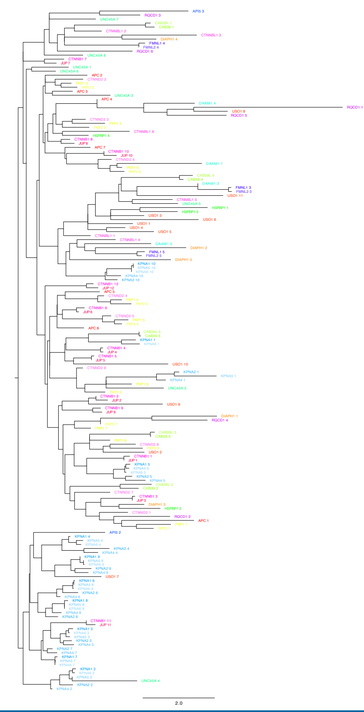
Determining the evolution of tandem repeat proteins is a challenge. The repeats typically have significant homology (for instance, Asn39 is conserved in 51% of human ArmRP; Leu22 in 67% (Gul, 2017)). This leads to the hypothesis that most tandem repeats evolved through duplication and fusion events. A number of mechanisms exist for the duplication/fusion of repeats, both individually and collectively through gene duplication/fusion. Combining this with gene-level speciation and duplications creating paralogs can explain the complex relationships we see among TR families.



Possible evolutionary histories for two 4-repeat proteins. a) TR expansion occurs before speciation, giving identical TR orders in both species. b) Both species undergo independent whole-gene duplications, giving similar subtree patterns in both species. c) Both species undergo independent TR duplications showing different patterns.



Phylogenetic tree for human karyopherin (KPNA) proteins. Each protein has a 10 Arm repeats. Each repeat clusters together, indicating that the order of repeats has not changed since the paralogous expansion. The support for this tree is too weak to draw conclusions about the relationships between repeats. Calculated using IQ-Tree (Nguyen 2015).



Preliminary tree of human ArmRP proteins with structural information in the Protein Data Bank. The shown tree does not explicitly model duplication events. Leaves are colored by gene. Calculated using IQ−Tree.

Conclusions & Outlook

We have shown that using cpHMM in TRAL we are able to improve our identification of ArmRP repeats. It is able to capture the correct phase of the repeats, as well as handle partial terminal repeats.

Using the ArmRP alignment we construct phylogenetic trees of the repeats. Work continues to reconstruct detailed evolutionary histories for individual repeats. This will allow the identification of repeat duplication and loss across this complex family.

Work is also ongoing to incorporate duplication and fusion events into the model for phylogenetic inference. This would allow reconstruction of the evolutionary events leading to tandem repeats.