Data in Brief 11 (2017) 510–516

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# Comparative analysis data of SF1 and SF2 helicases from three domains of life

Wafi Chaar [a], Hiba Ibrahim [b], Juliana Kozah [c], Hala Chamieh [d,e,*]

[a] Hloul Business Analytics, Omar Daouk Street, Beirut, Lebanon
[b] Beirut Arab University, Faculty of Science, Tripoli, Lebanon
[c] Université Saint Esprit, Jounieh, Lebanon
[d] Azm Center for Research in Biotechnology and its Applications, Lebanese University, Lebanon
[e] Department of Biology, Lebanese University, Faculty of Science, Tripoli, Lebanon

ARTICLE INFO

*Article history:*
Received 10 June 2016
Received in revised form
23 January 2017
Accepted 27 February 2017
Available online 3 March 2017

*Keywords:*
Helicase
Archaea
SF1
SF2
Phylogenetics

ABSTRACT

SF1 and SF2 helicases are important molecular motors that use the energy of ATP to unwind nucleic acids or nucleic-acid protein complexes. They are ubiquitous enzymes and found in almost all organisms sequenced to date. This article provides a comparative analysis for SF1 and SF2 helicase families from three domains of life archaea, human, bacteria. Seven families are conserved in these three representatives and includes Upf1-like, UvrD-like, Rad3-like, DEAD-box, RecQ-like. Snf2 and Ski2-like. The data highlight conservation of the helicase core motifs for each of these families. Phylogenetic analysis presented on certain protein families are essential for further studies tracing the evolutionary history of helicase families. The data supplied in this article support publication "Genome-wide identification of SF1 and SF2 helicases from archaea" (Chamieh et al., 2016) [1].



DOI of original article: http://dx.doi.org/10.1016/j.gene.2015.10.007
* Corresponding author at: Lebanese University, Faculty of Science, Department of Biology, Tripoli, Lebanon.
*E-mail address:* hala.chamieh@ul.edu.lb (H. Chamieh).

http://dx.doi.org/10.1016/j.dib.2017.02.047

**Specifications Table**

| | |
|---|---|
| Subject area | *Biology* |
| More specific subject area | *Genomics, Phylogenetics, helicase, archaea* |
| Type of data | *Figures* |
| How data was acquired | *Computational analysis* |
| Data format | *Analyzed* |
| Experimental factors | *Protein sequences were retrieved from online databases and used for detection of protein domain conservation and Phylogenetic analysis.* |
| Experimental features | *Human, E.coli and archaea protein helicase sequences were aligned using TCOFFEE or PROMALS3D..Conserved motifs were detected from multiple sequence alignments using WebLOGO software. Phylogenetic analysis were performed using Maximum Likelihood Methods or Bayesian Methods after protein alignment trimming by TrimAl.* |
| Data source location | *Lebanese University* |
| Data accessibility | Data is available within this article |

**Value of the data**

- The presented data on highly conserved amino acids in each of the seven conserved families across the three domains of life is important to design mutagenic studies and therefore determine functional conservation required for helicase function.
- Protein sequence comparison between SF1 and SF2 helicase families will allow establishing key experiments for genetic and biochemical analysis of helicase action.
- Phylogenetic tree data of Upf1-like, ski2-like and rad3-like shed light on the phylogenic relationship between these helicases in archaea, human and *E.coli*. The data offers valuable information on the complex evolutionary history within a helicase family and is a starting point for more detailed evolutionary studies on helicase subfamilies.

## 1. Data

Four figure files are presented. Fig. 1 denotes a comparative analysis of helicase core motifs in conserved families from archaea, bacteria and human. Figs. 2–4 are phylogenetic trees obtained after Maximum Likelihood analysis for Upf1-like and Rad3-like families, and Bayesian analysis for ski2-like helicase family.

## 2. Experimental design, materials and methods

All protein sequences were retrieved from existing protein databases and were used with their UniProt accession numbers and were classified into different families as shown in Chamieh et al. [1,2]. Multiple protein sequence alignment was performed using T-COFFEE EXPRESSO program for small sequence numbers ($<150$ sequences) [3] or PromalS3D for large sequence numbers ($>150$ sequences) [4]. Fig. 1 was obtained from the multiple sequence alignment files for protein sequences within the same family using the WEBLOGO software [5]. Sequences were inspected for their correct alignment within the helicase core domain. Multiple sequence alignment was trimmed using TrimAl v1.3 method set to automated [6]. The best evolutionary fit model was identified using ProtTest [7].

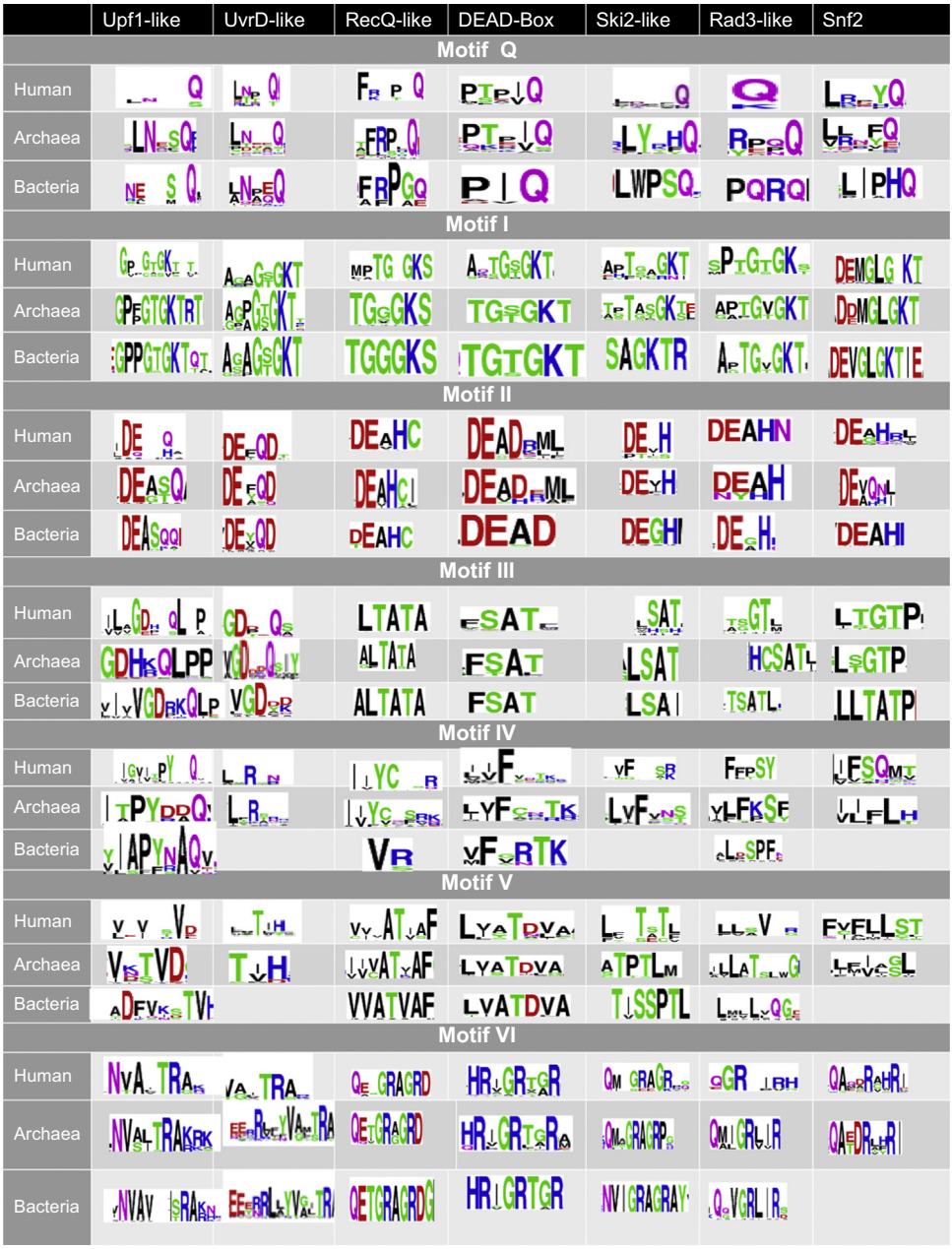


**Fig. 1.** Conserved motifs of the helicase core domain for SF1 and SF2 families across the three domains. All protein sequences were retrieved from existing protein databases. Multiple protein sequence alignment was performed using T-COFFEE EXPRESSO program for small sequence numbers ( < 150 sequences) (2) or PromalS3D for large sequence numbers ( > 150 sequences). Conserved motifs were generated from the multiple sequence alignment files for protein sequences within the same family using the WEBLOGO software.

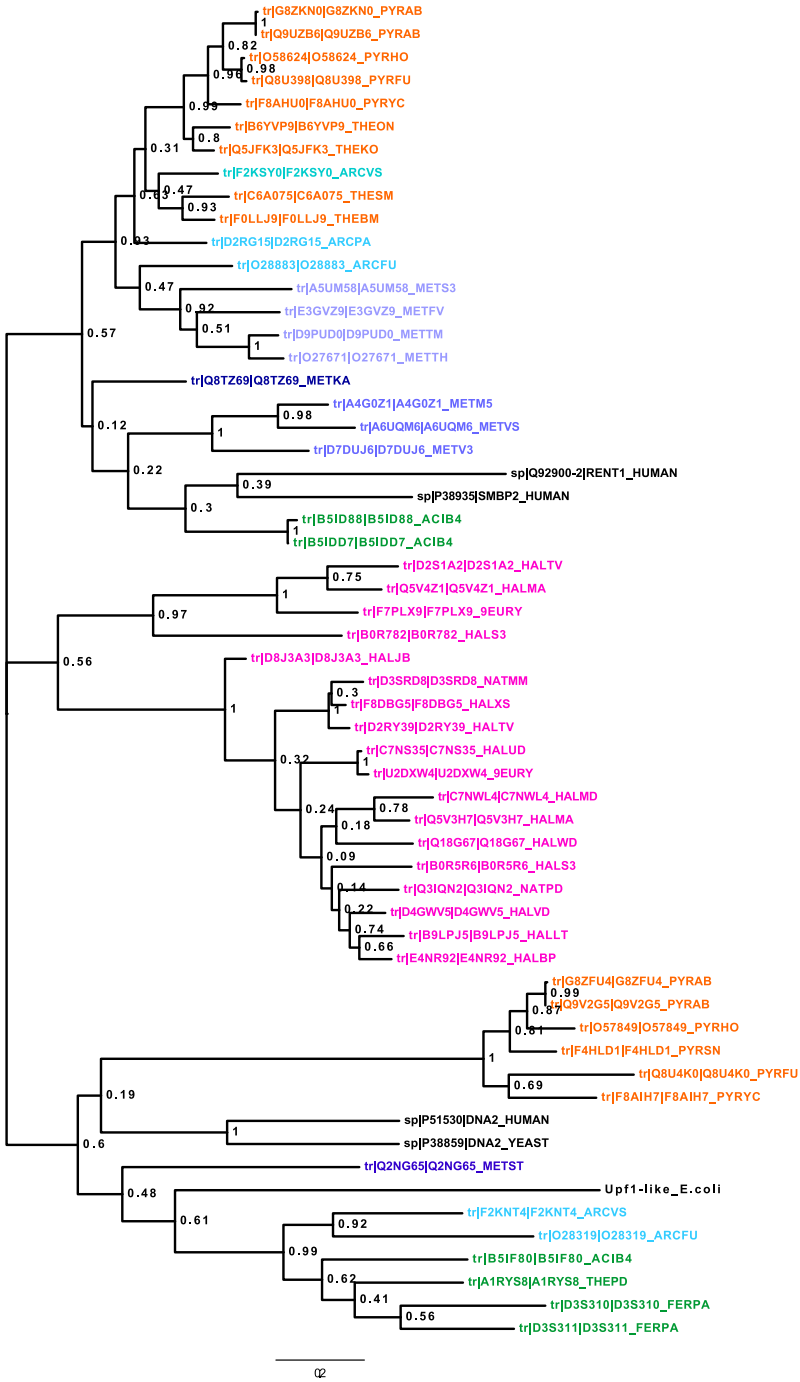


**Fig. 2.** Molecular Phylogenetic analysis of the Upf1-like family by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan And Goldman+Freq. model (WAG+F). The percentage of trees in which the associated taxa clustered together is shown next to the branches. The analysis involved 58 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 230 positions in the final dataset. Evolutionary analyses were conducted in MEGA7.


**Fig. 3.** Molecular Phylogenetic analysis of Ski2-like family by Bayesian Method. The evolutionary history was inferred by using the Bayesian method based on the MTMam model. The analysis involved 178 amino acid sequences. Evolutionary analyses were conducted in MrBayes. Two runs of 750,000 generations were conducted. Burn-in was set to 25%. Robustness of nodes was assessed with Bayesian posterior probabilities.

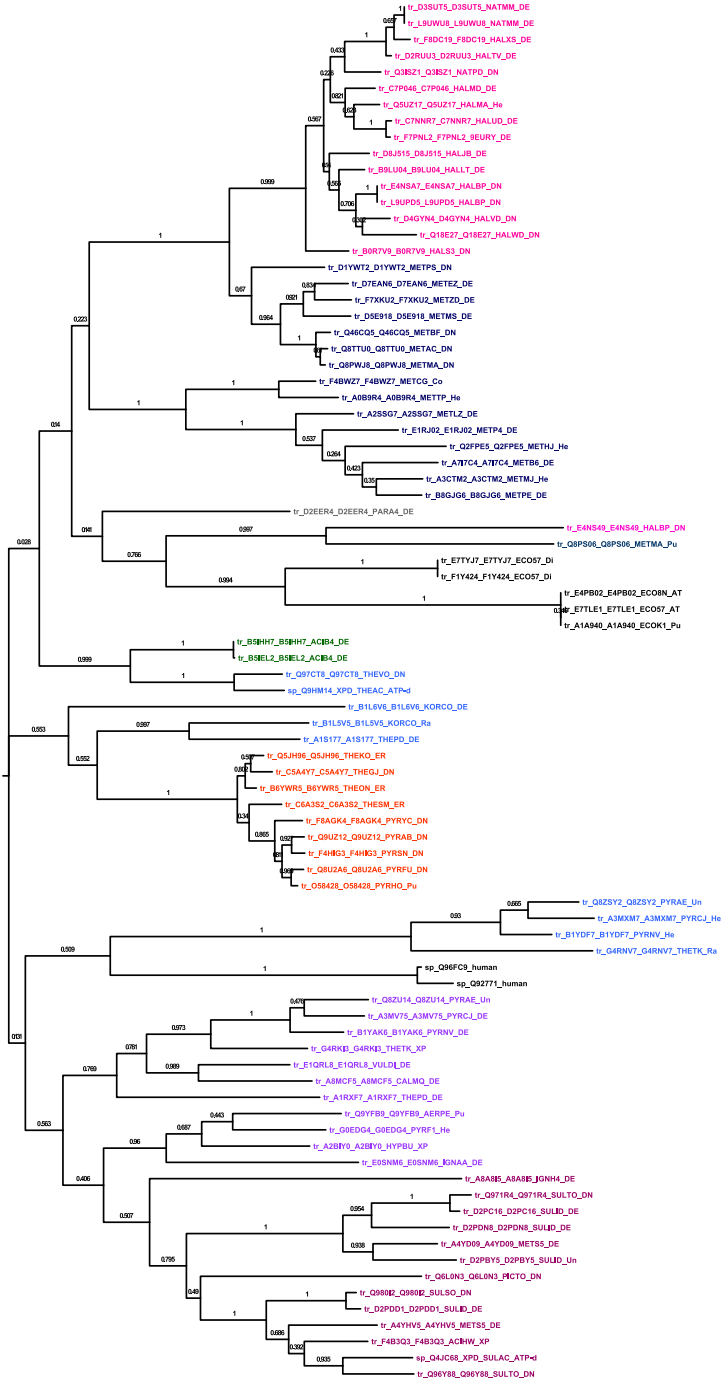


**Fig. 4.** Molecular Phylogenetic analysis of rad3-like family by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method based on the WAG+F model. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The analysis involved 85 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 268 positions in the final dataset. Evolutionary analyses were conducted in MEGA7.

Phylogenetic analysis was performed using Maximum Likelihood analysis from MEGA7 software [8] or MrBayes with the TOPALI platform [9,10].

## Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.02.047.

## References

[1] H. Chamieh, H. Ibrahim, J. Kozah, Genome-wide identification of SF1 and SF2 helicases from archaea, Gene 576 (1 Pt 2) (2016) 214–228.
[2] R. Apweiler, UniProt: the Universal Protein knowledgebase, Nucleic Acids Res. 32 (90001) (2004) 115D–119D.
[3] J.-F. Taly, C. Magis, G. Bussotti, J.-M. Chang, P. Di Tommaso, I. Erb, et al., Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures, Nat. Protoc. 6 (11) (2011) 1669–1682.
[4] J. Pei, N.V. Grishin, PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information, Methods Mol. Biol. 1079 (2014) 263–271.
[5] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (6) (2004) 1188–1190.
[6] S. Capella-Gutierrez, J.M. Silla-Martinez, T. Gabaldon, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, Bioinformatics 25 (15) (2009) 1972–1973.
[7] D. Darriba, G.L. Taboada, R. Doallo, D. Posada, ProtTest 3: fast selection of best-fit models of protein evolution, Bioinformatics 27 (8) (2011) 1164–1165.
[8] S. Kumar, G. Stecher, K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets, Mol. Biol. Evol. 33 (7) (2016) 1870–1874.
[9] F. Ronquist, M. Teslenko, P. van der Mark, D.L. Ayres, A. Darling, S. Höhna, et al., MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space, Syst. Biol. 61 (3) (2012) 539–542.
[10] I. Milne, D. Lindner, M. Bayer, D. Husmeier, G. McGuire, D.F. Marshall, et al., TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops, Bioinformatics 25 (1) (2009) 126–127.