Attempting an annotation on the gene *Resuscitation-promoting factor (RpfB)* of *Mycobacterium Tuberculosis* (*MTb*) using a Sequence Analysis approach.

Ioannis Valasakis[1]

[1]Birkbeck, University of London

January 19, 2018

Abstract

This coursework will provide computational research information on *RpfB*, a gene in the *Mycobacterium Tuberculosis (MTb)*. It is also known as *RV1009*. The approach is going to be methodological, as discussed in Sequence Analysis lectures. Computational Software, Online Databases and Web Services like Protein BLAST, HHBlits and InterPro are utilised to retrieve the required information to generate further insights.

Introduction

RpfB is a gene responsible for the creation of the protein *Resuscitation-promoting factor (RpfB)* in the *Mycobacterium Tuberculosis (MTb)*. Resuscitation of *MTb* is crucial to the aetiology of tuberculosis, not only because latent tuberculosis is estimated to affect one-third of the world population(Ruggiero et al., 2009). Kapoor et all showed that the resuscitation-promoting factor *RpfB* is mainly responsible for *MTb* resuscitation from dormancy(Kapoor et al., 2013). Given the impact of latent Tuberculosis, *RpfB* represents an interesting target for tuberculosis drug discovery. Currently, no molecular models of substrate binding and catalysis are hitherto available for this enzyme.

As described in the concept of the Central Dogma of Molecular Biology by Sir Francis Crick(CRICK, 1970), this is the one-way process where each gene in the DNA molecule carries the information needed to construct one protein, which, enzymatically, controls one chemical reaction in the cell. The subject of this coursework is the annotation of the gene *RpfB*, logically organised according to the dogma mentioned above.

KEYWORDS

rpfb; tuberculosis; sequence analysis; cell wall

Results

General Genome Features

The gene *RpfB* is located in the forward strand of the MTb chromosome in the location 1,128,091-1,129,179, with a length of 1089 base pairs (bp). It consists of 362 amino acids, and it has the transcript ID CCP43759 (also known as Rv1009), a protein ID 362aa and a UniProt ID P9WG29. This reference is taken from the MTb strain H37Rv with the genome assembly ID ASM19595_v2 from Sanger Institute. MTb has a genome size of 4,411,532 bp.

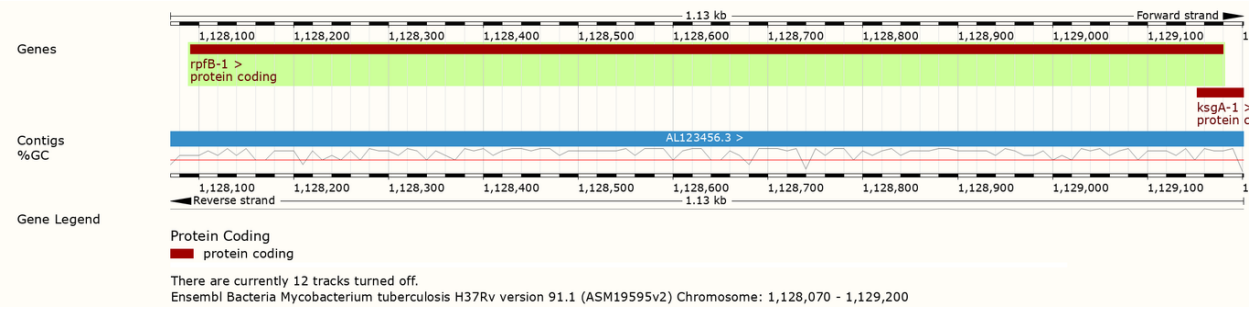

Figure 1: EnsemblBacteria view of the gene *RpfB*. (Zerbino et al., 2017)

Transcriptional Regulation

Protein synthesis is a process that happens in two steps. First, the DNA is transcribed to mRNA by an RNA polymerase complex, and second, the mRNA is translated to protein by a ribosome, which is a complex of proteins and rRNA. Having as a starting point the sequence of the gene *RpfB,* I was able to analyse it using the *Open Reading Frame* (*ORF*) Finder from the NCBI website, to explore further ideas. An ORF is a part of a reading frame with the potential to be translated to mRNA. Specifically, the ORF2 seems to match the *RpfB* (as expected) but the ORF13 is part of ksgA (an adjacent gene overlapping *RpfB*). Interestingly enough, ORF1 (length 393bp) seems to also create the protein deoxyribonuclease.

Given the fact that a match with -10 and -35 promoter sequences, while lacking an ORF may encode a functional RNA (tRNA, rRNA, etc), it is valuable to try and locate those promoters. From the Berkeley Drosophila Genome Project, which has a promoter prediction algorithm, it seems that a quite possible promoter lies in the position 524-569 upstream

> AGGTGGTTGAGTGTTGCCGAGGTCGGGGATATAGCGCGTTGACTCTACTT

The ORF, including this promoter, can also be part of an operon, meaning that it will be regulated along with other genes and may be functionally related to them as well. Finally, as those sequences are used under different conditions in the cell, there may be binding sites for repressor or activator proteins that are involved in regulating many different genes in the bacterial cell. In fact, considering that the promoter lies in the position between 524-569 upstream and there's an ORF at the same position with a -10 promoter 'TATA', there is really possible that the annotation of the protein is misplaced: the beginning of the coding

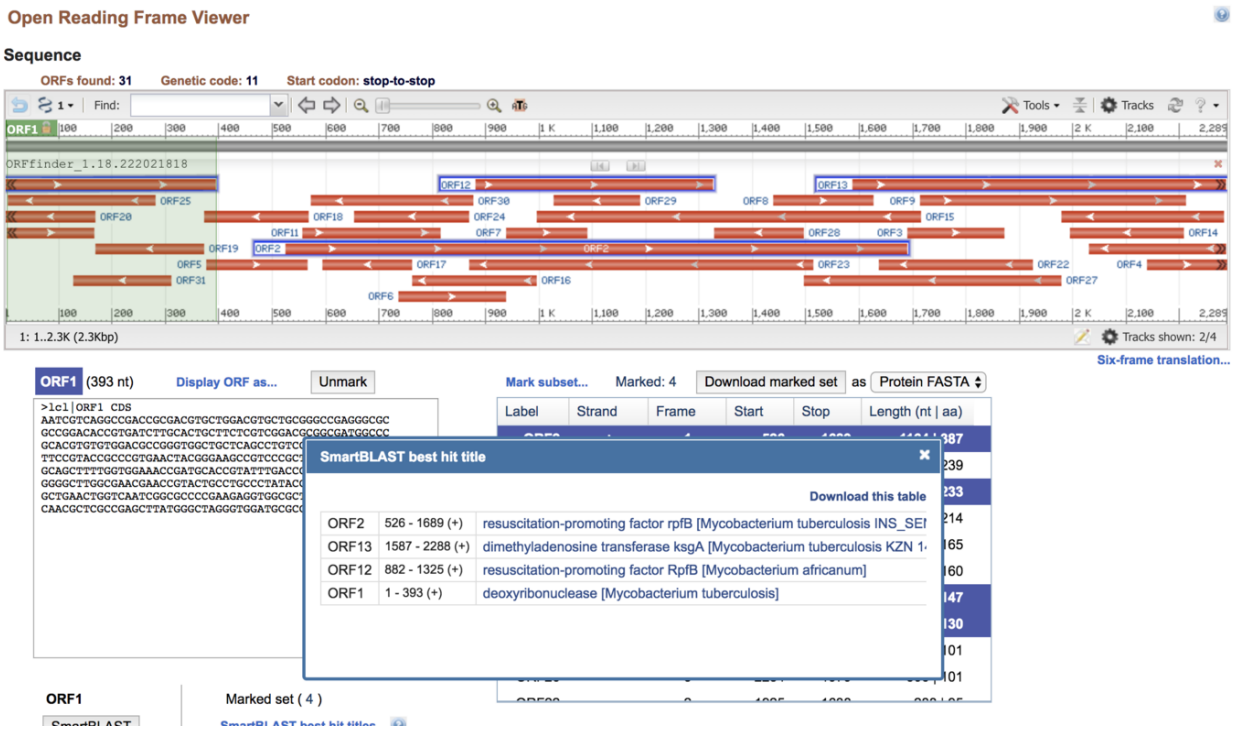

Figure 2: Using the ORF Finder from NCBI, combined with Smart Blast on the longest operons some hypothesis can be made.

sequence, in that case, includes 75 nucleotides upstream the originally annotated start codon. The starting codon would be a 'GTG'. It is indeed more usual the starting codon to be 'ATG' but there has been about 15% of the cases in E. Coli bacterium that this is a 'GTG'.

There is a really high probability of the gene *RpfB* and the *ksgA* to share a common operon, namely the aop0185 (Arkin Laboratory naming system. (Price, 2005)

The cellular location of the gene after getting the following motif result via Expasy,

| MLRL————VVGALLLVLAFAGGYAVAAC

can be assumed that it lies in the prokaryotic membrane and the lipoprotein lipid attachment site. (Gasteiger, 2003)

Protein

To discover the conserved protein domains of the protein, I used NCBI's Conserved Domain Database (*CDD*). The protein characterised as a resuscitation-promoting factor, which is a cell-wall glycosidase that cleaves cell-wall peptidoglycan; it stimulates resuscitation of dormant cells. Three domains are conserved: YabE, DUF348, Transglycosylase and G5. Similar results can be assumed by using the InterPro protein database.

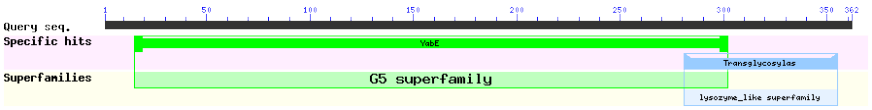

Figure 3: Conserved protein domains of *RpfB*.

In the literature, the protein is described as a factor that stimulates resuscitation of dormant cells. Has peptidoglycan (*PG*) hydrolytic activity. Active in the *pM* concentration range. Has little to no effect on actively-growing cells. *PG* fragments could either directly activate the resuscitation pathway of dormant bacteria or serve as a substrate for endogenous Rpf, resulting in low molecular weight products with resuscitation activity. A fragment (residues 194-362) hydrolyzes an artificial lysozyme substrate, by itself has little activity on the cell wall, in combination with *RipA* is active against cell wall extracts from a number of Actinobacteria; this activity is inhibited by *PBP1A* (*ponA1*). Sequential gene disruption indicates *RpfB* and RpfE are higher than *RpfD* and *RpfC* in the functional hierarchy. (Lee et al., 2014; Chauviac et al., 2014; Bhuwan et al., 2016; Ruggiero et al., 2013)

There is substantial evidence that the protein is strongly correlated to its adjacent proteins, forming one group of inter-related behaviour. See 4.

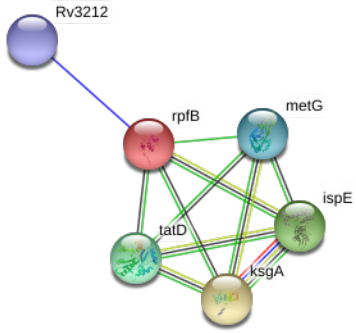

Figure 4: Statistical analysis of the interaction combined with the randomly observed probability between *RpfB* and its adjacent proteins using the String database.(Hubbard, 2004b)

Another relation that is useful phylogenetically is the one between the gene families whose occurrence patterns across genomes show similarities. Hidden Markov Model (HMM, PSI-BLAST) analysis combined with observed experimental data and predicted probability, created the conserved domains as shown in 5.

Finally, using the CATH database, I acquired the 11 domains that are most relevant to the protein. (Sillitoe et al., 2015)

Some of the domains belong to the cluster 2.20.230.10.1.1.1

> 4fuoA01
> 4funA01
> 4fum01
> 4fupA01
> 4fuoB01

as well as the cluster 2.20.230.10.3.1.1.1

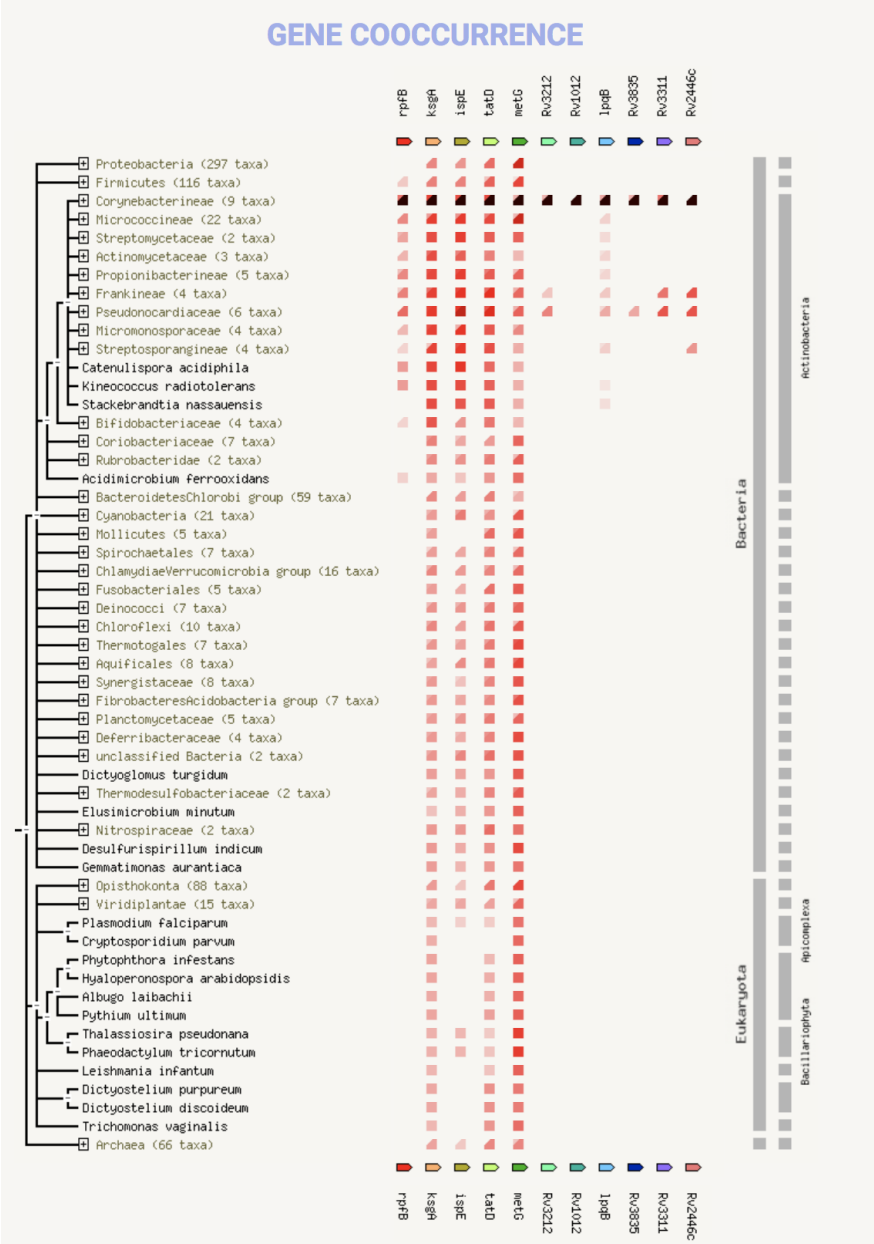

Figure 5: Phylogenetic representation of gene family occurances between different species as a relation of conserved topological domains. (Hubbard, 2004a)

> 4fzqD00
> 4fzqF00
> 4fzqB00
> 4fzqA00
> 4fzqE00
> 4fzCD00

With limited space and time, it makes more sense to concentrate on the few that have the highest sequence

similarity, i.e. 3eo5, 4fupA01, 4emn. Those two superfamilies adopt the fold members of Lyases(48%) and Hydrolases (98.4%) respectively. The architecture descriptions are "Orthogonal Bundle" and "Single Sheet". The functional family of this CATH domain is defined as *Accumulation associated protein* and *Hydrophilic* (grouped as a membrane protein). This information is available online on the PDBSum link provided by the CATH entry (PDB id: 4fup).(Conrady et al., 2012) Similar protein features exist on the *Escherichia coli* and *Staphylococcus epidermidis* (92%) match respectively. There is also another domain, identical to 3eo5 (Lysozyme) namely the 1xsfA00, which is experimentally verified in the CATH database and verifies the Lysozyme prediction made above.

Concluding, some Gene Ontology (GO) annotations sequenced by UniProt-GO. Summarizing:(Huntley et al., 2014)

> protein binding
> extracellular region
> positive regulation of gene expression
> negative regulation of gene expression
> positive regulation of growth rate
> dormancy exit of symbiont in host

Further analysis could prove valuable on getting more in-depth information about the protein *RpfB* using the CATH domains, gene ontologies, Gene3D results analysis and HMM predictions. At the same time, an investigation about the sub-processes the interactors of this protein are involved.

License

References

Manish Bhuwan, Naresh Arora, Ashish Sharma, Mohd Khubaib, Saurabh Pandey, Tapan Kumar Chaudhuri, Seyed Ehtesham Hasnain, and Nasreen Zafar Ehtesham. Interaction of MTb Virulence Factor RipA with Chaperone MoxR1 Is Required for Transport through the TAT Secretion System. *mBio*, 7(2):e02259–15, mar 2016. doi: 10.1128/mbio.02259-15. URL https://doi.org/10.1128%2Fmbio.02259-15.

Francois-Xavier Chauviac, Giles Robertson, Doris H. X. Quay, Claire Bagnéris, Christian Dumas, Brian Henderson, John Ward, Nicholas H. Keep, and Martin Cohen-Gonsaud. The RpfC (Rv1884) atomic structure shows high structural conservation within the resuscitation-promoting factor catalytic domain. *Acta Crystallographica Section F Structural Biology Communications*, 70(8):1022–1026, jul 2014. doi: 10.1107/s2053230x1401317x. URL https://doi.org/10.1107%2Fs2053230x1401317x.

D. G. Conrady, J. J. Wilson, and A. B. Herr. Structural basis for Zn2 +- dependent intercellular adhesion in staphylococcal biofilms. *Proceedings of the National Academy of Sciences*, 110(3):E202–E211, dec 2012. doi: 10.1073/pnas.1208134110. URL https://doi.org/10.1073%2Fpnas.1208134110.

FRANCIS CRICK. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, aug 1970. doi: 10.1038/227561a0. URL https://doi.org/10.1038%2F227561a0.

E. Gasteiger. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, 31(13):3784–3788, jul 2003. doi: 10.1093/nar/gkg563. URL https://doi.org/10.1093%2Fnar%2Fgkg563.

T. Hubbard. Ensembl 2005. *Nucleic Acids Research*, 33(Database issue):D447–D453, dec 2004a. doi: 10.1093/nar/gki138. URL https://doi.org/10.1093%2Fnar%2Fgki138.

T. Hubbard. Ensembl 2005. *Nucleic Acids Research*, 33(Database issue):D447–D453, dec 2004b. doi: 10.1093/nar/gki138. URL https://doi.org/10.1093%2Fnar%2Fgki138.

Rachael P Huntley, Tony Sawford, Maria J Martin, and Claire O'Donovan. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *GigaScience*, 3(1), mar 2014. doi: 10.1186/2047-217x-3-4. URL https://doi.org/10.1186%2F2047-217x-3-4.

Nidhi Kapoor, Santosh Pawar, Tatiana D. Sirakova, Chirajyoti Deb, William L. Warren, and Pappachan E. Kolattukudy. Human Granuloma In Vitro Model for TB Dormancy and Resuscitation. *PLoS ONE*, 8 (1):e53657, jan 2013. doi: 10.1371/journal.pone.0053657. URL https://doi.org/10.1371%2Fjournal.pone.0053657.

Jino Lee, Jihye Kim, Jeewon Lee, Sung Jae Shin, and Eui-Cheol Shin. DNA immunization of MTb, Cell Responses. *Clinical and Experimental Vaccine Research*, 3(2):235, 2014. doi: 10.7774/cevr.2014.3.2.235. URL https://doi.org/10.7774%2Fcevr.2014.3.2.235.

M. N. Price. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research*, 33(3):880–892, feb 2005. doi: 10.1093/nar/gki232. URL https://doi.org/10.1093%2Fnar%2Fgki232.

Alessia Ruggiero, Barbara Tizzano, Emilia Pedone, Carlo Pedone, Matthias Wilmanns, and Rita Berisio. Crystal Structure of the Resuscitation-Promoting Factor (DeltaDUF) RpfB from M. tuberculosis. *Journal of Molecular Biology*, 385(1):153–162, jan 2009. doi: 10.1016/j.jmb.2008.10.042. URL https://doi.org/10.1016%2Fj.jmb.2008.10.042.

Alessia Ruggiero, Jan Marchant, Flavia Squeglia, Vadim Makarov, Alfonso De Simone, and Rita Berisio. Molecular determinants of inactivation of the resuscitation promoting factor B from Mycobacterium tuberculosis. *Journal of Biomolecular Structure and Dynamics*, 31(2):195–205, feb 2013. doi: 10.1080/07391102.2012.698243. URL https://doi.org/10.1080%2F07391102.2012.698243.

Ian Sillitoe, Natalie Dawson, Janet Thornton, and Christine Orengo. The history of the CATH structural classification of protein domains. *Biochimie*, 119:209–217, dec 2015. doi: 10.1016/j.biochi.2015.08.004. URL https://doi.org/10.1016%2Fj.biochi.2015.08.004.

Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M Staines, Stephen J Trevanion, Bronwen L Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1): D754–D761, nov 2017. doi: 10.1093/nar/gkx1098. URL https://doi.org/10.1093%2Fnar%2Fgkx1098.