FEBS Letters 445 (1999) 6–8 FEBS 21549

Identification of a ribonuclease H gene in both *Mycoplasma genitalium* and *Mycoplasma pneumoniae* by a new method for exhaustive identification of ORFs in the complete genome sequences

Matthew I. Bellgard[a,b,*], Takashi Gojobori[a]

[a] *Center for Information Biology, National Institute of Genetics, Mishima, Shizuoka-ken 411-8540, Japan*
[b] *School of Information Technology, Murdoch University, Murdoch, WA, Australia*

Received 25 December 1998

Abstract Exhaustive identification of open reading frames in complete genome sequences is a difficult task. It is possible that important genes are missed. In our efforts to reanalyze the intergenic regions of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*, we have newly identified a number of new open reading frames (ORFs) in both *M. genitalium* and *M. pneumoniae*. The most significant identification was that of a ribonuclease H enzyme in both species which until now has not been identified or assumed absent and interpreted as such. In this paper we discuss the biological importance of RNase H and its evolutionary implication. We also stress the usefulness of our method for identifying new ORFs by reanalyzing intergenic regions of existing ORFs in complete genome sequences.

Key words: Complete genome sequence; *Mycoplasma genitalium*; *Mycoplasma pneumoniae*; Ribonuclease H; Open reading frame

1. Introduction

Complete genome sequence data are accumulating at a staggering pace. In order to annotate these data, identification of ORFs is usually conducted by several computer programs along with some manual refinement and intervention. However, this strategy is not infallible and it is still possible that the identification of very important genes are often missed. These missed genes can result in errors in functional and evolutionary interpretations. Until they are identified or confirmed to be absent by detailed analysis, there is a potential for error propagation. Hence, there is an urgent need for careful reanalyses of genome sequences. Compounding this problem, typical homology searches are usually conducted at the protein level because the sequenced whole genomes currently available are distantly related and thus have low DNA sequence similarity. This means that untranslated regions between predicted ORFs (intergenic regions) of the sequenced genomes are not subject to typical protein query searches. In order to make an attempt to exhaustively identify all ORFs, we show that it is useful to conduct a homology search by using intergenic regions as query sequences.

By our analysis of intergenic regions in *Mycoplasma genitalium*, we have identified nine new ORFs in *M. genitalium* [1] and seven in *Mycoplasma pneumoniae* [2] (Table 1). They have not been identified in previous studies and they are supported by the observation that they share high sequence homology and are in orthologous positions, in relation to flanking annotated ORFs. Five of these ORFs currently have no other significant database matches.

Most significant in our findings is the identification of a ribonuclease H (RNase H) enzyme in both species which, until now, has not been identified [1–3] and has been interpreted as absent [4,5]. It is significant because the biological role and evolutionary origins of the RNase H enzyme are not clearly known and are fast becoming an intriguing puzzle to solve [6,7]. It is known that the RNase H enzyme cleaves the RNA strand in hybrid molecules containing paired RNA and DNA strands [8]. It has also been suggested that multiple genes of RNase H per genome are a general feature of a wide variety of organisms [9]. For example, *Escherichia coli* possesses two enzymes with RNase H activity: RNase HI [10] and RNase HII [9]. These two proteins are encoded by *rnhA* and *rnhB* genes, respectively. While extensive studies have been conducted on *rnhA*, little is known about *rnhB*. There is a general consensus that RNase H is ubiquitously found in all living organisms [8,12]. Frank et al. [12] show that yeast RNase H(35) is the counterpart to mammalian RNase HI and is evolutionarily related to prokaryotic RNase HII. To date, a RNase H protein has not been identified in either *M. genitalium* or *M. pneumoniae*. As a possible explanation for the absence of this gene in *M. genitalium*, it has been proposed, by a theory of non-orthologous displacement [4,5], that another gene within *M. genitalium* performs the required function. In this case, the ORF MG262 displays high homology to the 5′-3′ exonuclease domain and it is proposed to displace RNase H in *M. genitalium*. However, our new identification of an RNase H gene in *M. genitalium* and *M. pneumoniae* show that this theory is not needed to explain its apparent absence. Thus, we would like to propose that it is useful to apply our new method for conducting homology search by using intergenic regions as query sequences to identify new ORFs.

2. Materials and methods

The complete annotated *M. genitalium* and *M. pneumoniae* genome sequences were obtained from the internet sites http://www.tigr.org/tdb/tdb.html and http://mail.zmbh.uni-heidelberg.de/M_pneumoniae, and a program was written to extract all intergenic sequences. All sequences were subjected to a BLAST search [13] against the non-redundant databases. Matches with a score ($E < 0.2$) were tentatively identified as potential ORFs. Intergenic regions larger than 60 nucleotides were also translated in all six frames and those with a conceptual translation greater than 20 amino acids were compared against orthologous intergenic regions in *M. pneumoniae*. The GeneMark.hmm [14] was used for further support of potential ORFs.

The PSI BLAST program [13] was used to extract all RNase HII

*Corresponding author. Fax: (61) (8) 9360 2941.
E-mail: m.bellgard@murdoch.edu.au

PII: S0014-5793(99)00075-7

Table 1
Newly identified genes within *Mg* and *Mp* by reanalyzing the intergenic regions of *Mg* and *Mp*

Intergenic region	Corrected annotation	Length (aa)/% identity	Comment
MG103-MG104	New ORF in *Mg* and *Mp*	77 *Mg* 76 *Mp*/88	Weak homology to *Bs* ORF yvaL (76 aa)
MG114-MG115	Longer N-terminus of *Mg*115 (extra 84 aa) New *Mp* ORF	163/50	Unknown function
MG131-MG132	New ORF in *Mg* and *Mp*	96/35	Unknown function
MG141-MG142	New ORF in *Mg* and *Mp*	88 *Mg* 92 *Mp*/55	Unknown function
MG149-MG150	New ORF in *Mg*. It has an ortholog in *Mp*, ORF VXpSPT7_orf112	154/68	
MG199-MG200	Identified R RNase HII gene in both *Mg* and *Mp*. Intergenic region+ORF MG199 (C09_orf143b in *Mp*) constitute the RNase HII gene	235 *Mg* 244 *Mp*/49	31% aa with *Sp* RNase H not identified in *Mg* (1) and *Mp* (3) Koonin et al. predict RNase H's function is substituted (10) BLAST search reveals weak homology to a hypothetical ORF in both Bs and Aa
MG269-MG270	New *Mg* and *Mp*. A frame shift error (run of 5 adenines) in *Mg*	132/46	Unknown function
MG291-MG292	New ORF in *Mg* and *Mp*	139 *Mg* 141 *Mp*/59	Match to hyp *Bs* ORF (yrrK, 138 aa)
MG357-MG358	New ORF in *Mg*. Ortholog in *Mp*, previously reported to be unique to *Mp* (G12_orf140b)	142/54	
MG395-MG396	New ORF in *Mg*. Ortholog in *Mp*, previously reported to be unique to *Mp* (D02_orf122a)	98/34	Unknown function
MG456-MG457	New ORF in *Mg* and *Mp*	193/43	Unknown function

genes from DDBJ/EMBL/GenBank and ClustalW [15] Ver 1.74 was used to create a multiple sequence alignment. Gapped regions in this alignment were removed and the resultant sequences were aligned again. Subsequent gapped regions were removed and the process was repeated until there was no further improvement to the alignment. The resulting alignment (152 amino acids) was used to construct the phylogenetic tree. The 22 species included in this analysis (along with accession numbers) are: *Pyrococcus kodakaraensis* (*Pk*) (AB012613), *Pyrococcus horikoshii* (*Ph*) (AP000006), *Haemophilus influenzae* (*Hi*) (P43808), *Mycobacterium tuberculosis* (*Mt*) (Q10793), *Mycobacterium leprae* (*Ml*) (Z97369), *Chlamydia trachomatis* (*Ct*) (AE001277 (*Ct1*), AE001275 (*Ct2*)), *Streptomyces coelicolor* (*Sso*) (AL022374), *Streptococcus pneumoniae* (*Sp*) (U93576), *Bacillus subtilis* (*Bs*) (Z99112 (*Bs1*), Z75208 (*Bs2*)), *Aquifex aeolicus* (*Aa*) (AE000765 (*Aa1*), AE000755 (*Aa2*)), *Archaeoglobus fulgidus* (*Af*) (AE001062), *Borrelia burgdorferi* (*Bb*) (AE001118), *Brucella melitensis* (*Bm*) (AF054610), *Methanococcus jannaschii* (*Mj*) (Q57599), *Helicobacter pylori* (*Hp*) (P56121), *Escherichia coli* (*Ec*) (P10442), *Magnetospirillum* sp. (*Ma*) (D32253), *Synechocystis* sp. (*Sy*) (D90899), *Caulobacter crescentus* (*Cc*) (p52975), *Methanobacterium thermoautotrophicum* (*Mth*) (AE000875), *Homo sapiens* (*Hs*) (Z97029), and *Saccharomyces cerevisiae* (*Sc*) (P53942). Note that the genes for *Mj*, *Aa2*, and *Bs2* are currently annotated as hypothetical proteins. Phylogenetic analysis was conducted using PHYLIP [16].

3. Results and discussion

A particular intergenic sequence in *M. genitalium* was highly homologous to the N-terminal region of *Sp* RNase HII gene [9]. Further analysis revealed that when concatenated with the adjacent annotated ORF (MG199), this sequence fragment was a RNase HII homolog. Between MG199 and the intergenic region, there is a stop codon due to a frame shift which prevents continuous conceptual translation. We also identified an RNase HII ORF homolog in *M. pneumoniae*. Surprisingly, however, there was no frameshift error between the orthologous intergenic region and the adjacent *M. pneumoniae* ORF (C09_143b).

It is likely that the frameshift in the RNase HII ORF in *M. genitalium* may be due to a sequencing error as there is a run of eight consecutive adenines at the frame shift error site, and the removal of a single adenine from the sequence brings the translation back into frame. (In the relatively high AT-rich *M. genitalium* genome, there are a total of 604 runs of at least a length of seven consecutive adenines distributed fairly uniformly.) Other possible explanations for the frameshift are that the ORF contains a translational frameshift, or that the RNase HII in this *M. genitalium* strain is non-functional (a pseudogene). When this sequence is submitted to the gene prediction program GeneMark.hmm, the program predicts, with a high probability, two ORFs on two consecutive frames (on the same strand) corresponding to the two ORFs separated by the stop codon. In addition, there is a 49% (116/235 amino acids) amino acid identity between *M. genitalium* and *M. pneumoniae* when the extra adenine in *M. genitalium* is removed. From our analysis we suspect that the frameshift is due to a sequencing error.

A BLAST homology search using the *rnhB* gene from *M. genitalium* as a query sequence reveals weak but significant homologies ($E < 4^{-10}$) to a hypothetical protein in both *Bacillus subtilis* [17] and *Aquifex aeolicus* [18]. This sequence match implies that there is a new *rnhB* gene in both species in addition to the one already annotated in each. We shall refer to these new *rnhB* genes as *Bs2* and *Aa2*, respectively. A multiple sequence alignment of the *rnhB* genes in *M. genitalium*, *M. pneumoniae*, *Bs2*, *Aa2*, and *Sp* revealed that there were conserved motifs which are specific to the *rnhB* gene [6,11] (data not shown). Our work confirms the result of an early study in which two RNase H proteins exist in *Bs*, although the sequence data was not shown [9]. The identification of the second RNase HII in both *Bs* and *Aa* is of evolutionary interest as Zhang et al. [11] indicated that *Sp* was the first genome in which only RNase HII is present (in the absence of a RNase HI enzyme). They also speculated that RNase HI may be the more dispensable of the two RNase H enzymes. Our findings support this proposal because the minimal genomes of *M. genitalium* and *M. pneumoniae* contain only *rnhB* and both *Bs* and *Aa* each contain two copies of this

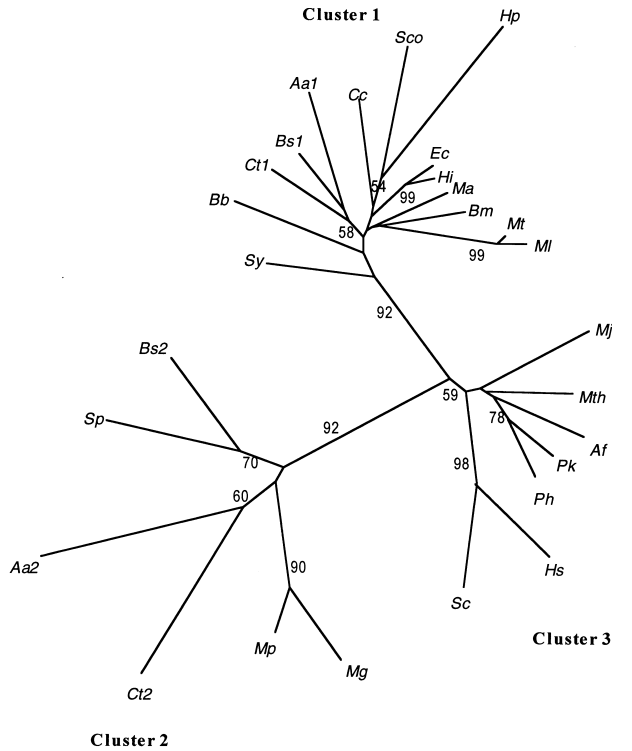

Fig. 1. Phylogenetic tree of *rnhB* genes. The two eubacterial clusters are labelled cluster 1 and cluster 2. *Mg* and *Mp* are contained within cluster 2. *Mj*, *Bs2* and *Aa2* are currently annotated as hypothetical ORFs in the public databases. See the text for abbreviations and accession numbers. Only bootstraps larger than 50 are shown in the figure.

gene within their genome. As further evidence for the dispensability of RNase HI, the recently sequenced genome of *Chlamydia trachomatis* (*Ct*) [19] also contains two copies of *rnhB*. For our phylogenetic analysis, we shall refer to these as *Ct1* and *Ct2*, respectively.

Fig. 1 shows the phylogenetic tree of the *rnhB* genes for 22 species representing Eubacteria, Archaebacteria and Eukaryotes. As can be seen in this figure, there are two bacterial clusters. Cluster 1 includes the most available species such as: Proteobacteria, Spirochaetales, Cyanobacteria, Firmicutes, Aquificales and Chlamydiales. Cluster 2 contains only a limited number of species: Firmicutes, Aquificales and Chlamydiales. Note that for the species with two copies of *rnhB*, *Ct*, *Aa* and *Bs*, one copy belongs in cluster 1 and the other in cluster 2. A phylogenetic tree was also constructed by the maximum likelihood method and a similar tree was produced (data not shown). A branch separating cluster 1 from cluster 2 has a high bootstrap value. This would imply that the *rnhB* gene has been involved in an early duplication event in the bacterial lineage as is evidenced by two copies of genes shared among the species *Bs*, *Aa* and *Ct*. Subsequent to the duplication, most bacterial species must have lost one or the other gene. *M. genitalium*, *M. pneumoniae* and *Sp* have a single copy of the *rnhB* gene belonging to cluster 2 and lost the duplicated gene belonging to cluster 1. On the other hand, the reverse has occurred for the other bacterial species. It should be noted that all the species examined in the present study have at least one copy of *rnhB*. Although the recently sequenced genome of *Treponema pallidum* (*Tp*) [20], which is a spirochete, does not have a *rnhB* gene, it does possess a *rnhA* gene. Thus, it is reasonable to conclude that either an *rnhA* or at least one *rnhB* gene is essential for every organism. Therefore, it is not conceivable that they are all absent from *M. genitalium* and *M. pneumoniae*.

Our finding is important for many reasons. First, a RNase H gene does exist in both *M. genitalium* and *M. pneumoniae* despite the acceptance by many of its absence in these two species. Second, the non-orthologous gene displacement theory is not necessary in this case. Third, there is now further support that RNase H, in the form of either *rnhA* or at least one *rnhB*, exists in all living organisms including the smallest known free living organisms. Fourth, while experimental studies are crucial for elucidating functional roles for the variety of RNase H enzymes found in organisms, our phylogenetic analysis reveals an interesting duplication event of the *rnhB* gene within the bacterial lineages. Thus, not only is there now a distinction between *rnhA* and *rnhB* genes, but there is also a distinction among the *rnhB* genes of various species. Finally, our method of using intergenic regions as query searches has been extremely useful. It will be even more useful for genome comparisons of closely related species as more become completely sequenced. Thus, there is need for detailed reanalysis and for a more robust, systematic approach for ORF identification. Further up to date information of the ORFs described in this paper, including sequence alignments and ORF identification strategies, may be obtained from http://arginine.it.murdoch.edu.au/research.html.

Acknowledgements: We thank Dr. Naoko Takezaki for her advice on the phylogenetic analysis. This research was made possible by a postdoctoral fellowship from the Japan Society for the Promotion of Science in Australia.

References

[1] Fraser, C.M. et al. (1995) Science 270, 397–403.
[2] Himmelreich, R. et al. (1997) Nucleic Acids Res. 24, 4420–4449.
[3] Himmelreich, R. et al. (1997) Nucleic Acids Res. 25, 701–712.
[4] Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) Trends Genet. 12, 334–336.
[5] Koonin, E.V., Mushegian, A.R. and Rudd, K.E. (1996) Curr. Biol. 6, 404–416.
[6] Frank, P. et al. (1998) Proc. Natl. Acad. Sci. USA 95, 12872–12877.
[7] Cerritelli, S.M. and Crouch, R.J. (1998) Genomics 53, 300–307.
[8] Stein, H. and Hausen, P. (1969) Science 166, 393–395.
[9] Itaya, M. (1990) Proc. Natl. Acad. Sci. USA 87, 8587–8591.
[10] Kanaya, S. and Crouch, R.J. (1983) J. Biol. Chem. 258, 1276–1281.
[11] Zhang, Y.-B., Ayalew, S. and Lacks, S.A. (1997) J. Bacteriol. 179, 3828–3836.
[12] Frank, P., Braunshofer-Reiter, C. and Wintersberger, U. (1998) FEBS Lett. 421, 23–26.
[13] Altschul, S.F. et al. (1997) Nucleic Acids Res. 25, 3389–3402.
[14] Lukashin, A.V. and Borodovsky, M. (1998) Nucleic Acids Res. 26, 1107–1115.
[15] Higgins, D.G. and Sharp, P.M. (1988) Gene 73, 237–244.
[16] Felsenstein, J. (1993) Phylogeny Inference Package, version 3.5c, distributed by the author, Department of Genetics, University of Washington, Seattle, WA.
[17] Kunst, F. et al. (1997) Nature 390, 249–256.
[18] Deckert, G. et al. (1998) Nature 392, 353–358.
[19] Stephens, R.S. et al. (1998) Science 282, 754–759.
[20] Fraser, C.M. et al. (1998) Science 281, 375–387.