Published online 31 July 2009

Nucleic Acids Research, 2009, Vol. 37, No. 12 ***1–5***
doi:10.1093/nar/gkn000

Helicobacter pylori Homology Database: the case of tryptophan

Markus Joppich [1,*], Cindy? [2] Luisa Jimenez [3] and Ralf Zimmer [1*]

[1]Affiliation of Corresponding Author and [2]Affiliation of Both Co-Authors

Received January 1, 2018; Revised February 1, 2018; Accepted March 1, 2018

ABSTRACT

To study microorganisms it is necessary to evaluate phenotypes, which include biochemical and visual characteristic, and relation with its environment. The advance of sequencing technologies allow access to information about the organism's genome and how it is transcribed from multiple exemplars. The combination of genotype and phenotype result in a integral picture of a microorganism. However, large quantities of genomes are being submitted every day at a pace challenging the capacities of analysis for those studying an organism. To date most searches for homology genes / proteins are done with the objective of find unknown function based on homology to already known genes or proteins. This searches are important to estimate the function of unknown proteins or genes. However studies involving the analysis of homologous proteins across several clones from the same species are not being fully exploited at the moment; and if the researcher wants to take evaluate homlogous genes, it needs to rely on BLASTs or in a unanimous annotation of the genes of interest. The last is hardly achievable, leaving an tenuous and limited search through alignment. In order to show the advantages of genome / proteome information from several strains , we have developed a database using as model organism Helicobacter pylori. With the use of this database we have found that strains from H. pylori use tryptophan in proteins in a strain-specific way with potential changes in the function of membrane related and cation-binding proteins. We extended this analysis to other bacterial species adapting this database to their genome information, showing that it can be adapted to any microrganism of interest of which several genomes are complete.

INTRODUCTION

The idea that similar protein or gene sequences have a higher probability of fulfilling the same or similar function has been the foundation for searches based on alignments, being BLAST the most popular. Although similarity could mean analogous function, the lack of similitude does not exclude it. Alignments can help to discover the function of unknown genes or proteins under the premise that the function of a homologous sequence is already known, as variations of in sequences are considered part of the evolutionary process (reviewed by Pearson WR 2013).

Until 2015 there were genome sequences from 50 bacterial and 11 archaeal phyla available with a total of around 14000 (February 2015, NCBI) (Land M 2015). Today, over three years later, this number has increase to a total of 133148 (March 15, 2018. NCBI Genome). The rate at which genomes are being published increases the possibility of finding a homologous gene or protein. At the same time, analysis of multiple genomes from different specimens (strains) belonging to same species allow to estimate normal variations of the organisms in ecosystems. However, this data growth rate challenges the ability to analyze it in a coherent manner based on already available information.

$$\mathrm{Ascorbate} + \mathrm{EDTA}\cdot\mathrm{Fe}^{3+} \rightarrow \text{Oxidized ascorbate}$$

$$\mathrm{EDTA}\cdot\mathrm{Fe}^{2+} + \mathrm{H_2O_2} \rightarrow \mathrm{EDTA}\cdot\mathrm{Fe}^{3+} + \cdot\mathrm{OH} + \mathrm{OH}^{-}$$

One of the many difficulties in the analysis of homologous proteins within one species is the variation of annotations in databases. Each submitted genome uses a different annotation for the genes/ proteins found in their data. Some are achieved through automated alignment and homology assignment. Other annotations are based on historic references made by the researchers at submission. One case are proteins components of the Type IV secretion System (T4SS). Bacterial species that present components for a T4SS, like Bordetella pertussis, Helicobacter pylori and Legionella pneumophila, have had their components described in different ways depending on the researcher submitting genomes, or as a result of an automated homology search made. One example of the consequences for these variation in annotation of genomes is the L. pneumophila's component DotL (From the Dot/Icm T4SS). It can be found as across literature and genomes as DotL, IcmO, or referred as VirD4 homologue based on its resemblance to the first T4SS defined in Agrobacterium tumefasciens . However, if the text search is done using VirD4 in L. pneumophila, none of these proteins will appear. Instead the results will include LvhD4, VirD4 component, conjugal transfer protein TraG, hypothetical protein or Type IV secretory system conjugative DNA transfer family protein.

*To whom correspondence should be addressed. Tel: +49 89 2180 4045; Email: joppich@bio.ifi.lmu.de

This kind of variations in annotations makes the work with bacterial proteins names extremely difficult and frustrating. This case is not isolated: all other bacteria named above present the same challenges while working with T4SS, and although the variations in annotations of the T4SS could have been an exception, sadly it is more the rule for many of the annotations of genes in their genome.

In order to work with all homologous genes from one organism, the annotation’s variations forces the use of sequence BLAST against the taxID of the species of interest, resulting in a list of proteins/genes with homologies to the query. This alignments can then be exported and used for analysis. Although functional is not the most efficient way to obtain several homologous genes from all strains sequenced.

The database presented here is designed to eliminate the need of BLAST of each protein against its own database, and facilitate the comparison of genomic and proteomic data available in the EMBL genome database from different strains from the same organism.

To evaluate the effectiveness and usefulness of the database, we have chosen as model organism *Helicobacter pylori*, an epsilon proteobacteria present in the human stomach and associated with gastric pathologies(ref). H. pylori is a natural competent bacterium, able to capture and integrate extracellular DNA into its genome. At the same time it present a high genetic variability between strains. The first strain sequenced from this organism was the strain 26695 (ref). Its open-reading frames (ORFs or Locus Tags) have been used for the definition of many *H. pylori* characteristic features, like the Cag Pathogenicity Island (Ref), the Outer Membrane Protein families (ref) and recently to describe the transcriptome of *H. pylori*(Sharma et al). The strain J99 sequence followed and its locus tags have been used as synonym to describe proteins or genes. While these descriptions were sufficient to analyze single genes in the following years, the exponential growth of new genomes available and variations in annotations between strains makes it difficult to find the known homologue genes. During experiments requiring semiquantitative analysis of western blot signals, the normalization of sample load for each bacterial strain in a polyacrylamide gel electrophoresis required the detection of proteins in a way compatible with further immunological analysis. The use of 2,2,2, trichloroethanol (TCE) with ultraviolet activation (ref) allowed the estimation of protein present in the gel and the use of the same samples for western blot analysis. The TCE allows fluorescent detection of tryptophan present in proteins at 305 nm without creating a crosslink of proteins with the acrylamide matrix. The imaging of tryptophan revealed that different strains presented variation in their patterns showing changes in tryptophan content in their proteins (Rojas et al, submitted) In order to find the homologous genes in two of the standard strains we were confronted with the need to identify the right homologous proteins in both genomes to quantify their tryptophan residues. This proved tedious and not really feasible if more strains would be analyzed. Therefore we created a database that allows the identification of homologous groups of proteins across different genomes independently of their annotations, and allows the introduction of transcriptome information if available for the organism. To show the biological reference using the database we show here that not only H. pylori uses tryptophan in an specific way, but ..... and ... as well, while other bacterial species like *Escherichia coli* and *Campylobacter pylori* restrict the variations in tryptophan residues. Text (2, 3).

Text (4).

MATERIALS AND METHODS

Materials subsection one

Text. Text.

Materials subsubsection one. Text. Text:

$$\mathrm{LD}^r = \frac{\mathrm{LD}}{A_{\mathrm{iso}}} = 1.5S\left(3\cos^2\alpha_i - 1\right) \quad \mathbf{(1)}$$

Text. Text.

Materials subsection two

Text. Text (see Figure 1).

Text. Text.

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

$$\mathrm{LD}(t) = \sum_i a_i \exp\left(\frac{-t}{\tau_i}\right)$$

Text. Text.

RESULTS

Results subsection one

Text. Text.

Results subsection two

Text. Text (see Table 1).

Text. Text (see Figure 2a).

Text. Text.

Table 1. This is a table caption

Col. head 1	Col. head 2 (%)	Col. head 3 (s^{-1})	Col. head 4 (%)	Col. head 5 (s^{-1})
Row 1	Row 1	Row 1	–	–
Row 2	Row 2	Row 2	Row 2	Row 2

This is a table footnote

Results subsection three

Text. Text.

DISCUSSION

Discussion subsection one

Text. Text.

Discussion subsection two

Text. Text.

Text. Text.

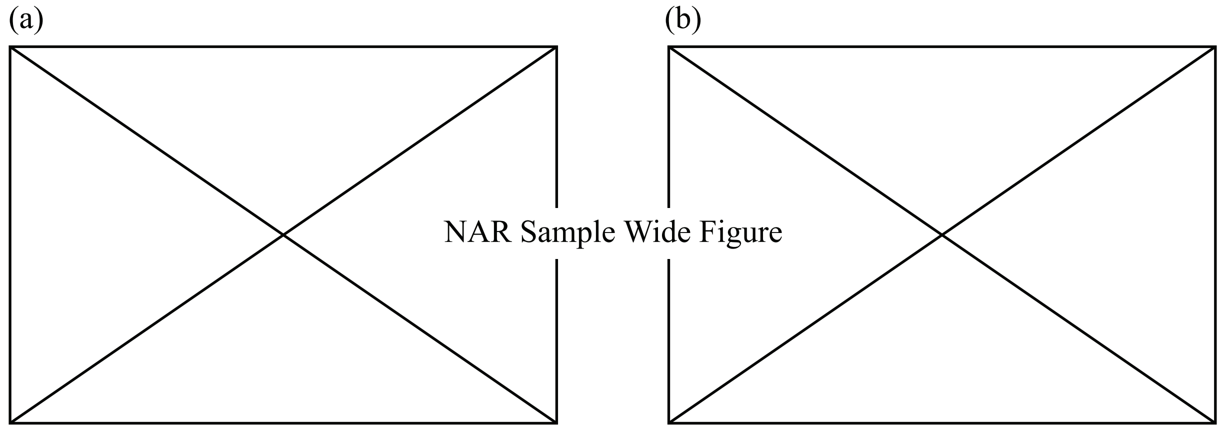

Figure 2. Caption for wide figure over two columns. **(a)** Left figure. **(b)** Right figure (see (a)).

Text. Text.

Discussion subsection three

Text. Text.

Text. Text.

Text. Text.

CONCLUSION

Text. Text.

ACKNOWLEDGEMENTS

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

Conflict of interest statement. None declared.

REFERENCES

1. Author,A.B. and Author,C. (1992) Article title. *Abbreviated Journal Name*, **5**, 300–330.
2. Author,D., Author,E.F. and Author,G. (1995) *Book Title*. Publisher Name, Publisher Address.
3. Author,H. and Author,I. (2005) Chapter title. In Editor,A. and Editor,B. (eds), *Book Title*, Publisher Name, Publisher Address, pp. 60–80.
4. Author,Y. and Author,Z. (2002) Article title. *Abbreviated Journal Name*, **53**, 500–520.