

# Helicobacter pylori Homology Database: the case of tryptophan

Markus Joppich<sup>1,\*</sup>, Luisa Jimenez<sup>3</sup> and Ralf Zimmer<sup>1,\*</sup>

<sup>1</sup>Affiliation of Corresponding Author and <sup>2</sup>Affiliation of Both Co-Authors

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

## ABSTRACT

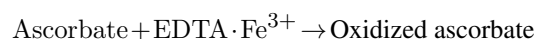
To study microorganisms it is necessary to evaluate phenotypes, which include biochemical and visual characteristic, and relation with its environment. The advance of sequencing technologies allow access to information about the organism's genome and how it is transcribed from multiple exemplars. The 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 homologous 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 microorganism 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.



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 tumefaciens* . 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

[illegible]

## Materials subsection two

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

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

[illegible]

## RESULTS

### Results subsection one

[illegible]

## Results subsection two

[illegible]

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.  
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.  
Text. Text. Text. Text. Text. Text (see Figure 2a).

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

**Table 1.** This is a table caption

Col. head 1	Col. head 2 (%)	Col. head 3 (s <sup>-1</sup> )	Col. head 4 (%)	Col. head 5 (s <sup>-1</sup> )
Row 1	Row 1	Row 1	–	–
Row 2	Row 2	Row 2	Row 2	Row 2

---

This is a table footnote

[illegible]

### Results subsection three

[illegible]

## DISCUSSION

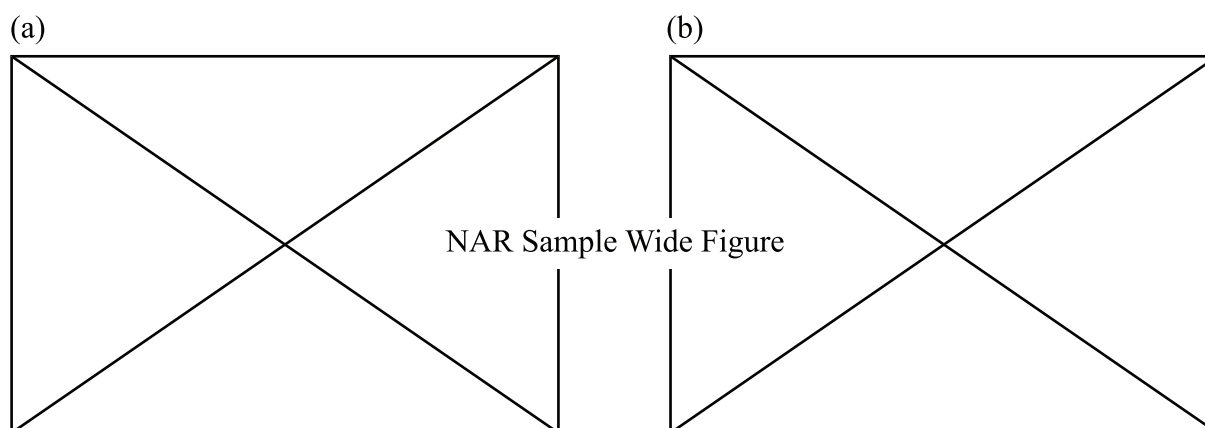
### Discussion subsection one

[illegible]

### Discussion subsection two

[illegible]

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



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

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

## CONCLUSION

Natural variations at the genome and protein level within a species can show us the flexibility of the genetic code in each organism. Although the search for homology across species is the most used method for search of function, the variations within a species can tell us about the intrinsic activity and tolerance levels of mutations for a single protein inside their species. Our Database allows to overcome the irregularities in gene annotations present in genome databases, and allows to adapt the emergent information about an microorganism like transcriptome and small RNAs, making it unique compared to other homologue search databases, where the primary function is to find the homology across different species.

## Discussion subsection three

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

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

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

