Assignment of Novel Functions to *Helicobacter pylori* 26695's genome

Tiago Resende[1], Daniela M. Correia[1] and Isabel Rocha[1]

[1] IBB – Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho, Portugal
tiagoresende@ceb.uminho.pt;{danielacorreia;irocha}@deb.uminho.pt

Abstract. *Helicobacter pylori* is a pathogenic bacterium that colonizes the human epithelia, causing duodenal and gastric ulcers as well as gastric cancer. The genome of *H. pylori* 26695 has been sequenced and annotated. In addition, two genome-scale metabolic models have been developed. In order to maintain accurate and relevant information on coding sequences (CDS) and to retrieve new information, the assignment of new functions to *Helicobacter pylori* 26695's genes was performed. The use of software tools, on-line databases and an annotation pipeline for inspecting each gene allowed the attribution of validated E.C. numbers to metabolic genes, and the assignment of 177 new functions to the CDS of this bacterium. This information provides relevant biological information for the scientific community dealing with this organism and can be used as the basis for a new metabolic model reconstruction.

Keywords: *Helicobacter pylori*, Genome annotation, Metabolic functions, Genome-Scale Reconstruction.

1 Introduction

Helicobacter pylori, first cultivated in 1982 [1], is a gram-negative, spiral-shaped bacterium that belongs to the Proteobacteria [2, 3]. It is well known that this bacterium colonizes the stomach of more than 50% of the human population worldwide, reaching 80% of infection rate in developing countries [1, 3]. When in the gastric mucosa, this bacterium induces a chronic inflammation causing an increase in the risk of developing a disease such as duodenal and gastric ulcer, gastric cancer and mucosa associated lymphoid tissue (MALT) lymphoma [1]. However, only few individuals develop any *H. pylori* related gastric disease [3]. This may be due to fact that this bacterium possesses mechanisms to increase genomic diversity yielding multiple and diverse strains [4]. At the present time, there are 43 completely sequenced genomes of different *H. pylori* strains on NCBI, which highlights this bacterium genetic variability. *H. pylori* 26695, a highly pathogenic strain, was originally isolated from a patient in the United Kingdom with gastritis and had its complete genome sequenced and published in 1997 using whole-genome random sequencing [5]. This organism presents a small size genome of

around 1.67 Mbp, with approximately 1590 coding sequences (CDS) identified [5].

The genome functional annotation can be seen as the process of allocating functional information to the genes of a sequenced genome. The majority of gene functions are assigned by homology search from characterized sequences, found in several online databases; and, if a given gene product is unknown, it is labeled as hypothetical protein [6]. The re-annotation can be viewed as the process of updating the functional information of a genome. Databases and computational methods are constantly evolving and over time new information is also being published, making possible to assign new gene functions [7]. The last re-annotation of *H. pylori* 26695 was published in 2003. This re-annotation generated a specific database for *H. pylori* (PyloriGene) [8] and allowed the reduction of the percentage of hypothetical proteins from approximately 40% to 33%, allowing also the reassignment of functions to 108 CDS [8]. Unfortunately, this re-annotation does not contemplate the allocation of E.C. numbers to the annotated metabolic genes and therefore it compromises some of the applications of the annotation. A very important application of gene functional annotation is the reconstruction of the metabolic network of a sequenced organism. This reconstruction allows the development of a genome-scale metabolic model based on the well-known stoichiometry of biochemical reactions catalyzed by the enzymes encoded in the annotated genes of an organism [9, 10]. These models can then be used for simulating *in silico* the phenotypic behavior of a microorganism under different environmental and genetic conditions, thus representing an important tool in metabolic engineering design and the identification of novel drug targets for pathogens [10].

To date, two metabolic models of *H. pylori* 26695 were published. The model *i*CS291 was published in 2002 and contains 291 genes and 388 reactions [11]; in 2005, based on the previous model, a new model was reconstructed, the *i*IT341 GSM/GPR with 341 genes and 476 reactions, including also 355 gene-protein reaction associations [12]. Most of improvements made in the latter model were a result of the increase of available literature and the revised annotation of the *H. pylori* genome [12].

Here we present a new re-annotation of the *H. pylori* 26695 genome. The function of each gene previously annotated was reevaluated, new functions were identified and EC numbers were assigned to genes with metabolic functions, thus presenting the combined results of updated databases and new annotation methodologies. This re-annotation will be used as the basis for reconstructing an updated genome-scale metabolic model for *H. pylori* 26695.

2 Methods

H. pylori 26695's genome was retrieved, in the amino acid fasta format, from the GenBank repository at ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Helicobacter_pylori_26695_uid57787/.

merlin

merlin (MEtabolic model Reconstruction using genome scaLe INformation) is a software tool created in our group to assist on the processes of (re) annotation and reconstruction of genome-scale metabolic models. *merlin* is available for download at http://www.merlin-sysbio.org. It performs automatic genome-wide functional (re)annotations and provides a numeric confidence score for each automatic assignment, taking into account the frequency and taxonomy within the annotation of all similar sequences [13]. In the present work the confidence score was kept with the default parameters, with a set threshold of 0.7. To perform homology searches, *merlin* uses both BLAST (Basic Local Alignment Search Tool) (from NCBI) and profile HMM (Hidden Markov Models) (from HMMER [14]) algorithms. *merlin*'s interface was used throughout the re-annotation process to assign functions to each, gene based on the highest confidence scores [13].

Annotation pipeline

After the automatic re-annotation performed by *merlin*, each candidate function was manually inspected by following several confirmation steps as described in Fig.1. For this, three on-line databases were used: UniProt [15] which contains up-to-date information in many *H. pylori* protein coding genes; BRENDA [16] which is an enzyme curated information database, used to confirm gene product names of a certain E.C. number; and PyloriGene [8] the specific *H. pylori* annotation database released in January 2003 upon the last re-annotation of the strain 26695 and last updated, through blastp homology search, in March, 2011[8].

The manual curation of *merlin* results began with the correspondence between each candidate and the information on different databases, giving priority to Uniprot reviewed information, followed by Uniprot unreviewed and finally the information in PyloriGene. When a match with reviewed information occurred, *merlin* candidates were annotated with a very high confidence level. On the other hand, when there was a match with Uniprot unreviewed data, the candidates were annotated with high or medium confidence levels, according to the type of information present, such as E.C. numbers, for example. If there was no information on Uniprot for candidates, *merlin* homology data, PyloriGene annotation and relevant bibliographic references (if existent on PyloriGene) were analyzed. Results with the best scores were selected and annotated with high, medium or low confidence levels, according to bibliographic evidence. When mismatches occurred between *merlin* results and Uniprot, *merlin* homology results were analyzed to search for matching information, or this was manually added. Each of the potential enzyme encoding candidates was revised in BRENDA to verify its function and confirm E.C. number assignment. Some of the enzymes were assigned with incomplete EC numbers; thus, this database was also used to identify complete EC numbers when available, by searching for enzyme's product name. As previous annotation lacked E.C. number information, and due to the importance of this kind of information, for instance in the reconstruction of metabolic models, an effort was made to try to retrieve every possible E.C. number belonging to candidates encoding enzymes.

Therefore, each of the potential enzyme encoding candidates' E.C. number was sought in the different sources of information, including Uniprot, *merlin* results and BRENDA. Nevertheless, despite following the annotation pipeline, genes with no metabolic function, naturally, were not assigned with an E.C. number and therefore were annotated according to the source of information, whether it was Uniprot reviewed, unreviewed, PyloriGene or *merlin* homology data.

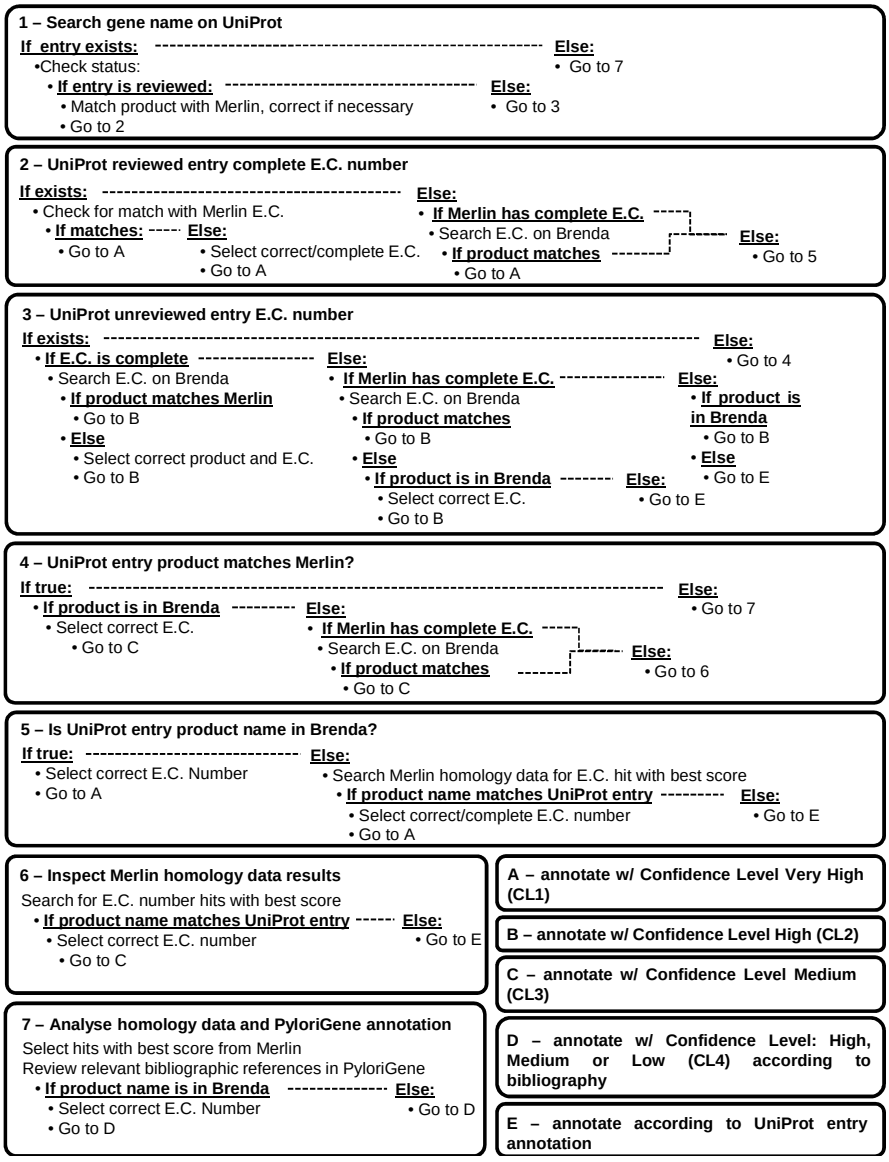

Fig. 1 Re-annotation pipeline for manual inspection of each gene candidate.

3 Results and Discussion

All protein encoding genes present in *H. pylori* 26695 genome were annotated according to the proposed methodology and reviewed by the developed annotation pipeline. The number of genes inspected was different from the last re-annotation because, in the genome retrieved from NCBI, the number of genes has been updated, having now 1573 genes, instead of the previous 1590.

Function assignment

Analyzing the results obtained from homology search with *merlin*, it was noticed that new assigned functions were based in homology with other *H. pylori* strains. This might be due to the exponential amount of *H. pylori* strains being sequenced in recent years, increasing the volume of available information on their genome.

As depicted in Table 1, the total number of coding sequences (CDS) annotated with a function was 1203, divided into 581 metabolic CDS (528 with complete E.C. numbers and 53 with incomplete E.C. numbers) and 622 non-metabolic CDS. The number of hypothetical CDS was 370, representing a total of 24% of the CDS in the genome, a lower number than the previous annotation which contained 510 hypothetical proteins (32%).

Comparing the present annotation with the previous one, it is possible to observe that both annotations are in agreement in the function of 1026 CDS. The number of CDS with differently assigned functions is 177, of which 137 correspond to the allocation of new functions to previous hypothetical CDS and 40 to the assignment of new functions to CDS previously annotated with another function. The 177 new functions assigned are divided in 71 metabolic CDS, 77 non-metabolic CDS and 29 CDS with a generic function, which has a lower level of specificity, such as, for example: HP1234, a membrane transport protein. In more than half of cases the new assignment of a function is related to increasing specificity of the function previously assigned and not necessarily to a modification in the function. For instance, the protein encoding gene HP1450, which had been annotated as an “inner membrane protein”, is now assigned as a “Membrane integrase YidC”.

Table 1 Distribution of CDS according to functional category

		This work	***PyloriGene***
Total CDS		1573	1590
Metabolic CDS	Complete E.C.	528	1080
	Incomplete E.C.	53	
Non-metabolic CDS		622	
Hypothetical CDS		370	510

Annotation confidence level

As a result of inspecting CDS according to the annotation pipeline, an annotation confidence level has been attributed to each protein coding sequence, according to the robustness of the information generated. Table 2 presents the confidence level for the total CDS and CDS with new functions.

Table 2 Confidence levels of function assignments to total CDS and CDS with new functions

Confidence Level	Total CDS (1573)	New functions (177)
Very high	529	39
High	93	4
Medium	65	6
Low	886	128

For a total of 1573 CDS annotated, 529 (33.6%) were classified with a very high confidence level, which is the highest classification, indicating that these genes are reviewed on Uniprot, and, therefore well characterized and curated manually by experts. This is also true for the 39 new functions (22%) classified in the same way. “High”, is the classification level of 93 (6%) of total CDS and 4 (2%) of new functions, which, along with the medium confidence level (65 (4%) of total CDS and 6 (3%) of new functions) also indicates a good/average confidence in the results, although in a lesser extent. This classification was assigned to genes with high similarity with other genes well characterized. The majority of the total CDS, 886 (56%), and of new functions, 128 (72%) were assigned with a low confidence level, indicating that these genes are not well characterized, lacking reviewed information and validation, being the result of pure homology search data and inference methodologies. This outcome was, somewhat, expected for new functions, once new homology information is more rapidly generated than direct biological/biochemical experimental data and also because the revision, by experts, of all existing information is a laborious and time consuming task.

Enzyme class distribution

More than 88% (513) of the CDS assigned with metabolic activities were classified with only one complete E.C. number (monofunctional). Nevertheless, two other groups appeared, depending on the number and class of assigned E.C. number. As depicted in table 3, most of complete monofunctional E.C. numbers are classified as transferases, 163 (28%) CDS. On the other hand, most of the CDS encoding incomplete E.C. numbers are hydrolases, 23 (4%). Nevertheless, only 9% (53) of enzymes have an incomplete E.C. number. Oxireductases, transferases and hydrolases represent more than 75% of the identified enzymes. Multifunctional genes encode for more than one enzymatic function within the same class, but with different functions. They catalyze similar reactions using substrates with small differences.

Table 3 Enzyme encoding genes classification

	Complete E.C.			Incomplete E.C.		
	A[1]	B[2]	C[3]	A[1]	B[2]	C[3]
Oxidoreductases	94	1	1	2	0	0
Transferases	163	5	2	19	0	0
Hydrolases	128	2	1	23	0	0
Lyases	45	1	1	1	0	1
Isomerases	32	0	1	4	0	0
Ligases	51	0	0	3	0	0

1- A = Monofunctional; 2- B = Multifunctional; 3- C = Multiclass.

For instance, HP0683, a bifunctional N-acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase (2.3.1.157, 2.7.7.23) catalyzes a reaction where the product of the first is the substrate of the second. Multiclass genes encode for more than one enzymatic activity whose E.C. numbers belong to different classes, meaning they have dissimilar catalytic functions, as for example, HP0326 which encodes for pseudaminic acid cytidylyltransferase and UDP-2,4-diacetamido-2,4,6-trideoxy-beta-L-altropyranose hydrolase (5.3.1.24, 4.1.1.48) that are classified as a transferase and a hydrolase, respectively. For constructing table 3, when classifying a protein coding sequence with more than one E.C. number, such CDS was assigned to the subgroup of first enzyme annotated, because such function was assumed as the main function.

4 Conclusions

In the present work, the assignment of new functional activities to the CDS of *H. pylori* 26695 genome was performed. Using a software tool for re-annotation and an annotation pipeline, all gene functions were inspected and updated, when necessary, being assigned with a confidence level for their function. The E.C. numbers for all metabolic CDS were searched, validated and attributed when found.
A total of 177 new functions were assigned, 137 of which were attributed to CDS previously classified as “hypothetical proteins”. 40 new functions were assigned to CDS already annotated; many of them had been classified with only generic annotations. From the new functions assigned, 71 were metabolic, 77 non-metabolic and 29 had generic descriptions, indicating, for instance the localization of the protein. A total of 581 E.C. numbers were assigned to CDS, being 528 of them complete E.C. numbers. These results bring new and more comprehensive information to the *H. pylori* 26695 genome, increasing and improving the existing knowledge on this human pathogen, with special relevance for the attributed metabolic functions. The assignment of E.C. numbers is a fundamental task, since these data can be used for the reconstruction of a new genome-scale metabolic model for this organism.

Acknowledgements. This work was supported by the project FCOMP-01-0124-FEDER-009707, entitled "HeliSysBio – molecular Systems Biology in *Helicobacter pylori*" (Ref.: FCT PTDC/EBB-EBI/104235/2008). Daniela Correia is grateful for financial support from the FCT (PhD grant: SFRH/BD/47596/2008).

References

1. Marshall, B.: *Helicobacter pylori* : 20 years on. Clinical Medicine. 2, 147–152 (2002).
2. Ge, Z., Taylor, D.E.: Contributions of genome sequencing to understanding the biology of *Helicobacter pylori*. Annu. Rev. Microbiol. 53, 353–387 (1999).
3. Kusters, J.G., Van Vliet, A.H.M., Kuipers, E.J.: Pathogenesis of *Helicobacter pylori* infection. Clinical microbiology reviews. 19, 449–90 (2006).
4. Costa, A.C., Figueiredo, C., Touati, E.: Pathogenesis of *Helicobacter pylori* Infection. Helicobacter. 14, 15–20 (2009).
5. Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., et al.: The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature. 388, 539–547 (1997).
6. Dias, O., Gombert, A.K., Ferreira, E.C., Rocha, I.: Genome-wide metabolic (re-) annotation of Kluyveromyces lactis. BMC genomics. 13, 517 (2012).
7. Médigue, C., Moszer, I.: Annotation, comparison and databases for hundreds of bacterial genomes. Research in microbiology. 158, 724–36 (2007).
8. Boneca, I.G., Reuse, H. De, Epinat, J.C., Pupin, M., Moszer, I., De Reuse, H., Labigne, A.: A revised annotation and comparative analysis of *Helicobacter pylori* genomes. Nucleic Acids Research. 31, 1704–1714 (2003).
9. Durot, M., Bourguignon, P.-Y., Schachter, V.: Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS microbiology reviews. 33, 164–90 (2009).
10. Rocha, I., Förster, J., Nielsen, J.: Design and Application of Genome-Scale Reconstructed Metabolic Models. In: Gerdes, S.Y. and Ostermnan, A.L. (eds.) Methods in Molecular Biology, vol. 416: Gene Essentiality. pp. 409–433. Humana Press Inc. (2007).
11. Schilling, C.H., Covert, M.W., Famili, I., Church, G.M., Edwards, J.S., Palsson, B.Ø.O.: Genome-scale metabolic model of *Helicobacter pylori* 26695. Journal of bacteriology. 184, 4582–4593 (2002).
12. Thiele, I., Vo, T.D., Price, N.D., Palsson, B.Ø.: Expanded Metabolic Reconstruction of *Helicobacter pylori* (iIT341 GSM / GPR): an In Silico Genome-Scale Characterization of Single- and Double-Deletion Mutants. Society. 187, 5818–5830 (2005).
13. Dias, O., Rocha, M., Ferreira, E.C., Rocha, I.: Merlin : Metabolic Models Reconstruction using Genome-Scale Information. In: Banga, J.R., Bagaerts, P., Impe, J. Van, Dochain, D., and Smets, I. (eds.) Proceedings of the 11th International Symposium on Computer Applications in Biotechnology (CAB 2010). pp. 120–125. , Leuven, Belgium (2010).
14. Finn,R.D., Clements J., Eddy S.R.: HMMER web server: interactive sequence similarity searching.. Nucleic Acids Research. Web Server Issue 39, W29-W37 (2011).
15. The UniProt Consortium: Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research. 40, D71-D75 (2012).
16. Scheer M., Grote A., Chang A., Schomburg I., Munaretto C., Rother M., Söhngen C., Stelzer M., Thiele J., Schomburg D.:BRENDA, the enzyme information system in 2011. Nucleic Acids Research. 3, 670-676 (2011).