Re-annotation of the genome sequence of Helicobacter pylori 26695

Tiago Resende[1], Daniela M. Correia[1], Miguel Rocha[2] and Isabel Rocha[1,*]

[1]IBB – Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho, Portugal

[2]CCTC, School of Engineering, University of Minho, Portugal

Summary

Helicobacter pylori is a pathogenic bacterium that colonizes the human epithelia, causing duodenal and gastric ulcers, and gastric cancer. The genome of *H. pylori* 26695 has been previously 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* 26695s genes was performed in this work. The use of software tools, on-line databases and an annotation pipeline for inspecting each gene allowed the attribution of validated EC numbers and TC numbers to metabolic genes encoding enzymes and transport proteins, respectively. 1212 genes encoding proteins were identified in this annotation, being 712 metabolic genes and 500 non-metabolic, while 191 new functions were assignment 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.

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 human population worldwide, reaching 80% of infection rate in developing countries, mainly due to unsanitary conditions [1, 3]. *H. pylori* is thought to be transmitted by the faecal-oral route, since it is associated and has been detected in contaminated water, but also by the oral-oral route through the spreading of contaminated secretions [1, 3]. When in the gastric mucosa, this bacterium induces a chronic inflammation causing an increase of 4-6 fold in the risk of developing a disease such as duodenal and gastric ulcer, and gastric cancer. Mucosa associated lymphoid tissue (MALT) lymphoma, the most common gastric lymphoma, is deeply associated with *H. pylori* infection and normally enters into complete remission once eradicated the infection. The treatment for duodenal and gastric ulcer is also the eradication of *H. pylori* [1]. Despite inducing chronic inflammation, 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, such as endogenous mutation and recombination, yielding multiple and diverse strains. This means that *H. pylori* strains are not equally pathogenic nor

[*] To whom correspondence should be addressed. Email: irocha@deb.uminho.pt

doi:10.2390/biecoll-jib-2013-233

virulent [4]. For instance, persons carrying strains containing full cag Pathogenicity Island (cagPAI) and active vacuolating cytotoxin (VacA), have higher risk for duodenal ulceration, compared to those without cagPAI [5]. 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 [6]. This organism presents a small size genome of around 1.67Mbp, with approximately 1590 coding sequences (CDS) identified [6].

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 on-line databases; if a given gene product is unknown, it is labelled as hypothetical protein [7]. 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 [8]. The last re-annotation of *H. pylori* 26695 was published in 2003. This re-annotation generated a specific database for *H. pylori* (PyloriGene) [9] and allowed the reduction of the percentage of hypothetical proteins from approximately 40% to 33%, allowing also the reassignment of functions to 108 CDS [9]. However, this re-annotation does not contemplate the allocation of EC numbers or TC 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 catalysed by the enzymes encoded in the annotated genes of an organism [10, 11]. These models can then be used for simulating *in silico* the phenotypic behaviour 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 [11].

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 [12]; 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 [13]. 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 [13].

Here, we present a new re-annotation of the *H. pylori* 26695 genome. The function of each gene previously annotated was re-evaluated and new functions were identified. EC numbers were assigned to genes involved in metabolites conversion, and TC numbers were assigned to genes involved in the metabolite carriage throughout the cell, thus presenting the combined results of updated databases and new annotation methodologies. This re-annotation will be used in the future as the basis for reconstructing an updated genome-scale metabolic model for *H. pylori* 26695 and will provide relevant biological information for the scientific community dealing with this organism.

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.

2.1 *merlin*

merlin (MEtabolic model Reconstruction using genome scaLe INformation) is a software tool created to assist in 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 [14]. In the present work, the confidence score was kept with the default parameters, with a threshold of 0.7. To perform homology searches, *merlin* uses both BLAST (Basic Local Alignment Search Tool) (from NCBI) and profile HMMs (Hidden Markov Models) (from HMMER [15]) algorithms. *merlin*'s interface was used throughout the re-annotation process to assign functions to each gene based on the highest confidence scores [14].

2.2 Annotation pipeline

After the automatic re-annotation performed by *merlin*, each candidate was manually inspected by following several confirmation steps as described in Figure 1. For that purpose, three online databases were used: UniProt [16] which contains up-to-date information in many *H. pylori* protein coding genes; BRENDA [17] which is an enzyme curated information database, used to confirm gene product names of a certain EC number; and PyloriGene [9] the specific *H. pylori* annotation database released in January 2003 upon the last re-annotation of strain 26695 and updated, through BLASTp homology search, in March 2011.

The manual curation of *merli*n 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 corresponding to unreviewed data, the candidates were annotated with high or medium confidence levels, according to the type of information present, such as EC numbers, for example. If there were no information on Uniprot for candidates, *merlin* homology data, PyloriGene annotation and relevant bibliographic references (if existing on PyloriGene) were analysed. Results with the best score were selected and annotated with high, medium or low confidence levels, according to bibliography. When mismatches occurred between *merlin* results and Uniprot, *merlin* homology results were analysed in search of matching information, or manually added. Each of the potential enzyme encoding candidates was revised in BRENDA to verify its function and confirm EC number assignment. Some of the enzymes were assigned with incomplete EC numbers; therefore, this database was also used to identify complete EC numbers for such cases, by searching for enzymes product names.

As previous annotation lacked EC number information, and due to the importance of this kind of

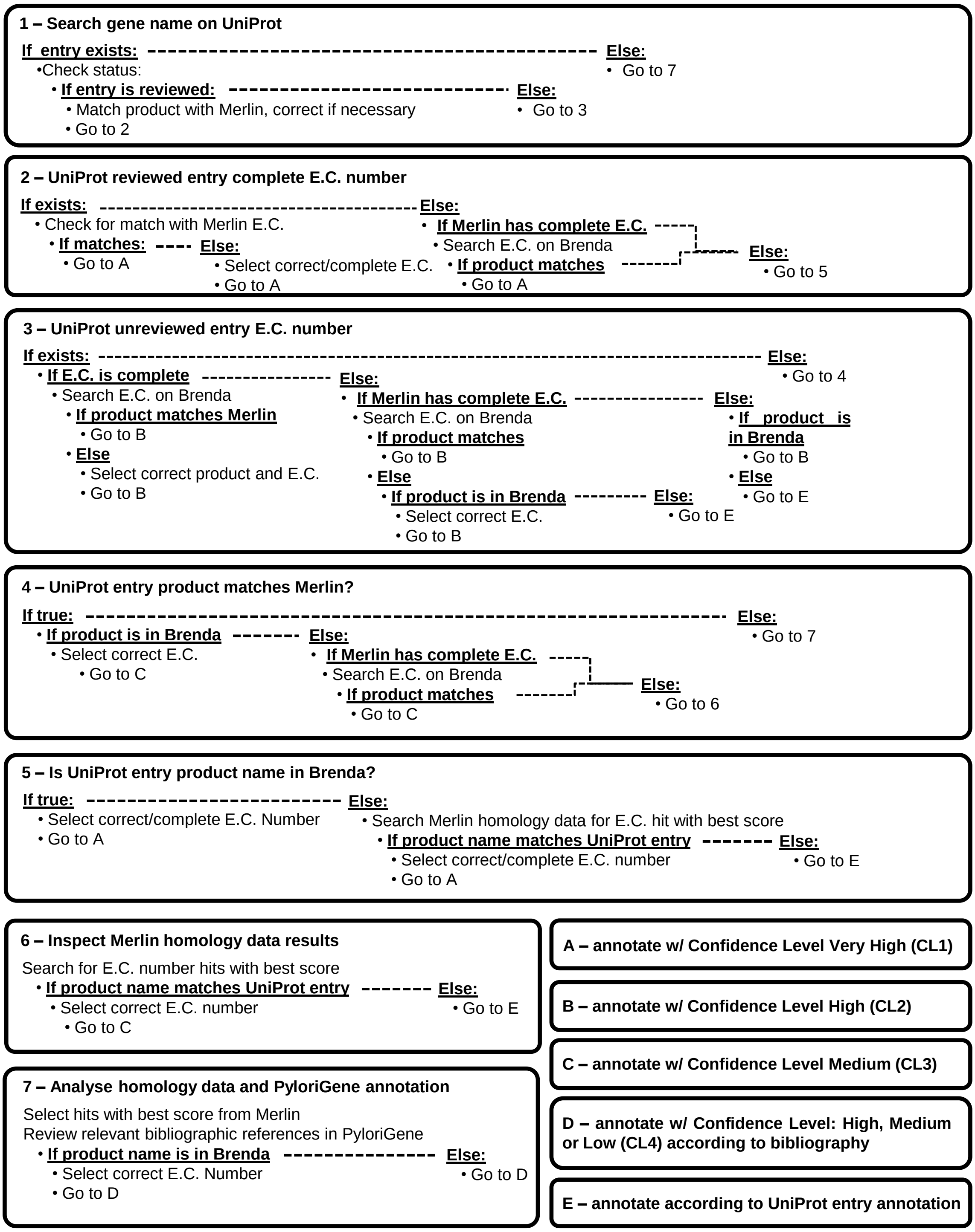
1 – Search gene name on UniProt
If entry exists:
•Check status:
• If entry is reviewed:
• Match product with Merlin, correct if necessary
• Go to 2
Else:
• Go to 3
Else:
• Go to 7
2 – UniProt reviewed entry complete E.C. number
If exists:
• Check for match with Merlin E.C.
• If matches:
• Go to A
Else:
• Select correct/complete E.C.
• Go to A
Else:
• If Merlin has complete E.C.
• Search E.C. on Brenda
• If product matches
• Go to A
Else:
• Go to 5
3 – UniProt unreviewed entry E.C. number
If exists:
• If E.C. is complete
• Search E.C. on Brenda
• If product matches Merlin
• Go to B
• Else
• Select correct product and E.C.
• Go to B
Else:
• If Merlin has complete E.C.
• Search E.C. on Brenda
• If product matches
• Go to B
• Else
• If product is in Brenda
• Select correct E.C.
• Go to B
Else:
• Go to E
Else:
• If product is in Brenda
• Go to B
• Else
• Go to E
Else:
• Go to 4
4 – UniProt entry product matches Merlin?
If true:
• If product is in Brenda
• Select correct E.C.
• Go to C
Else:
• If Merlin has complete E.C.
• Search E.C. on Brenda
• If product matches
• Go to C
Else:
• Go to 6
Else:
• Go to 7
5 – Is UniProt entry product name in Brenda?
If true:
• Select correct/complete E.C. Number
• Go to A
Else:
• Search Merlin homology data for E.C. hit with best score
• If product name matches UniProt entry
• Select correct/complete E.C. number
• Go to A
Else:
• Go to E
6 – Inspect Merlin homology data results
Search for E.C. number hits with best score
• If product name matches UniProt entry
• Select correct E.C. number
• Go to C
Else:
• Go to E
7 – Analyse homology data and PyloriGene annotation
Select hits with best score from Merlin
Review relevant bibliographic references in PyloriGene
• If product name is in Brenda
• Select correct E.C. Number
• Go to D
Else:
• Go to D
A – annotate w/ Confidence Level Very High (CL1)
B – annotate w/ Confidence Level High (CL2)
C – annotate w/ Confidence Level Medium (CL3)
D – annotate w/ Confidence Level: High, Medium or Low (CL4) according to bibliography
E – annotate according to UniProt entry annotation

Figure 1: Re-annotation pipeline for manual inspection of each gene candidate

information, for example in the reconstruction of metabolic models, an effort was made to try to retrieve every possible EC number belonging to candidates encoding enzymes. Therefore, each of the potential enzyme encoding candidates EC 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 EC number and therefore were annotated according to the source of information, whether it was Uniprot reviewed, unreviewed, PyloriGene or *merlin* homology data.

2.3 Identification of genes encoding transport proteins

Transport proteins are located on cell membranes and have a key role in the flow of nutrients and other metabolites within different compartments of the cell and between the cell and its environment. Thus, information on transport proteins is imperative for the reconstruction of a robust genome-scale metabolic model.

The Transport Classification Database (TCDB - `http://www.tcdb.org`) is a curated database that details a comprehensive classification system for membrane transport proteins known as Transporter Classification (TC) system. The TC system is analogous to the Enzyme Commission (EC) system for the classification of enzymes, except that it incorporates both functional and phylogenetic information [18].

There are only six genes of *H. pylori* 26695's genome included in TCDB as transporter proteins (see Additional file 1: Table S1 of the supplemental material available in the URL `http://darwin.di.uminho.pt/hpylori`). Thus, being this information so scarce, it was necessary to identify transport proteins through similarity search in TCDB. This homology analysis was performed with *merlin*, which uses the Smith-Waterman (SW) dynamic programming algorithm [19] to perform local similarity searches with the TCDB, to identify the TCS (Transporter Classification Superfamily) number of the genes that encode transporter proteins [7].

Since the SW algorithm is exact, its alignments are very accurate but computationally very demanding. To solve this problem, and because the transport of metabolites is usually performed by proteins located on membranes [20], only genes encoding proteins with transmembrane helices were searched against TCDB. For that, we used the TransMembrane prediction using Hidden Markov Models (TMHMM) software [21] that predicts the number of transmembrane helices in a protein, using Hidden Markov Models.

Only genes encoding proteins with one or more transmembrane domain were aligned in the TCDB, reducing the search time. *merlin* uses a similarity threshold of 10% when performing SW searches and has internal heuristics to lower the threshold inversely to the number of domains [7]. In the end of this process, each gene identified as a potential transport protein encoding gene was annotated and the TCS number, as well as the metabolites transported by such protein, were assigned. In order to prevent being too restrictive, TCS numbers were chosen over TC family numbers [7].

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.

3.1 Function assignment

Analysing 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.

The developed annotation pipeline successfully reviewed all the 1573 coding sequences (CDS) from *H. pylori* 26695's genome, finding functions to 1212 of its genes and assigning EC numbers to 581 of them. Also, the transporter annotation function from *merlin* was able to find similarities with 155 CDS from the genome, although 24 of the genes identified as transporter proteins and assigned with TC numbers were also annotated with EC numbers, being, therefore, classified with both EC and TC numbers.

As depicted in Figure 2, the total number of coding sequences annotated with a function was 1212, divided into 712 genes (45.26%) with a metabolic function and 500 (31.79%) with non-metabolic functions. The number of hypothetical CDS was 361, representing a total of 22.95% of the CDS in the genome, a lower number than the previous annotation which contained 510 hypothetical proteins (32%). From the 712 genes annotated with metabolic functions, 581 were annotated with at least one EC number, making up about 37% of the genome, and being 557 of them assigned only with EC numbers. From the 155 genes annotated as transport proteins, 131 CDS were annotated as exclusively transporters, being only assigned with TC or TCS numbers, and 24 were assigned with both enzymatic and transport functions.

The final annotation of *H. pylori* 26695 is available in Table S2 of the Additional file 1 (`http://darwin.di.uminho.pt/hpylori`).

3.2 Comparison with the last re-annotation

H. pylori 26695's last re-annotation was performed in 2003 [9]. However, the annotated genes were not assigned with EC or TC numbers, making it challenging to use the re-annotation in some of its applications, namely in the reconstruction of genome-scale metabolic models.

Table 1 displays the distribution of annotated CDS according to function category from this work and the last re-annotation. As seen before, the number of CDS annotated with a function is 1212. The decrease in the number of hypothetical CDS, from 510 in last re-annotation to 361 in this work, represents an increase of approximately 9.5 percentage points in the attribution of functions to CDS.

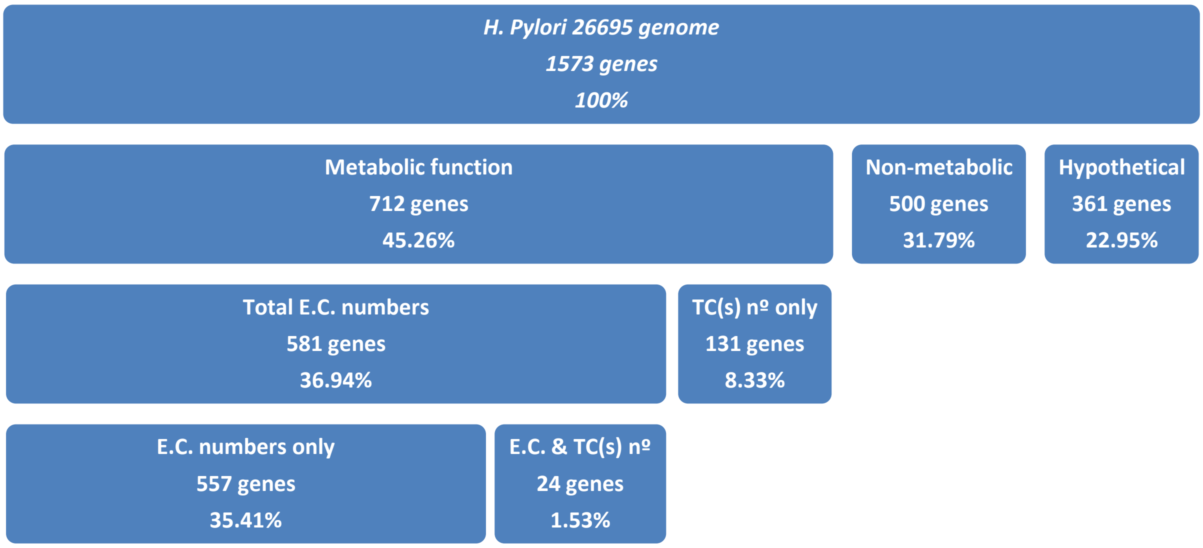

Figure 2: ***H. Pylori*** **26695 re-annotation results and statistics**

As depicted in Additional file 1: Table S3 of supplemental material, when comparing the present annotation with the previous one, it is possible to observe that both annotations are in agreement with the functions of 1021 CDS. The number of CDS assigned with new functions is 191, of which 149 correspond to the allocation of functions to hypothetical CDS (as already referred) and 42 to the assignment of new functions to CDS previously annotated with a function. The 191 new functions assigned are divided in 103 metabolic CDS, 64 non-metabolic CDS and 24 CDS with a generic function which have a low level of specificity, such as, for example: HP0114, a motility accessory protein. In more than half of the 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	*Last re-annotation* [9]
Total CDS		1573	1590
Metabolic CDS	Complete EC	527	1080
	Incomplete EC	54	
	Only TC(s)	131	
Non-metabolic CDS		500	
hypothetical CDS		361	510

3.3 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 partition of total CDS and new functions CDS by confidence level.

Table 2: Confidence levels of function assignment to total and new functions CDS

Confidence Level	Total CDS (1573)	New functions (191)
Very high	529	40
High	93	4
Medium	65	5
Low	886	142

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 40 new functions (21%) classified in the same way. The classification level of *high* is assigned to 93 CDS (6% of total) and to 4 (2%) of the new functions, which along with the *medium* confidence level assigned to 65 CDS (or 4% of the total) with 5 (3%) of new functions, also indicates a good confidence in results, although in a lesser extent. This classification was assigned to genes well/middling characterized but with no direct reviewed information. The majority of total CDS, 886 (56%), and new functions, 142 (74%) were assigned with *low* confidence level, indicating that these genes are not well characterized, lacking reviewed information and validation, being the assigned function a 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 the revision, by experts, of all existing information is a laborious and time consuming task.

3.4 Enzyme class distribution

More than 88% (512) of the CDS assigned with metabolic activities were classified with only one complete EC number (monofunctional). Nevertheless, two other groups appeared, depending on the number and class of assigned EC number. As depicted in table 3, most of complete monofunctional EC numbers are classified as transferases, 162 (28%) CDS. On the other hand, most of the CDS encoding incomplete EC numbers are hydrolases (24 or 4%). Nevertheless, only 9% (54) of enzymes have an incomplete EC number. Oxireductases, transferases and hydrolases represent more than 75% of the identified enzymes. Multifunctional genes encode for more than one enzyme within the same class, but with different functions. They catalyse similar reactions using substrates with small differences.

For instance, HP0683, a bifunctional N-acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase (2.3.1.157, 2.7.7.23) catalyses two reactions where the product of the first is the substrate of the second. Multiclass genes encode for enzymes whose EC numbers are attributed 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 (2.7.7.81, 3.6.1.57) that are classified as a transferase and a hydrolase, respectively. For constructing Table 3, when classifying a protein coding sequence with more than one EC number, such CDS was assigned to the subgroup of first enzyme annotated, because such function was assumed as the main function.

Table 3: Enzyme encoding genes classification

	Complete EC			**Incomplete EC**		
	A	B	C	A	B	C
Oxidoreductases	96	1	1	2	0	0
Transferases	162	5	2	19	0	0
Hydrolases	127	2	1	24	0	0
Lyases	45	1	1	1	0	1
Isomerases	31	0	1	4	0	0
Ligases	51	0	0	3	0	0
(A- monofunctional; B- multifunctional; C- multiclass)						

3.5 Transporter classification

In the process of re-annotation, during the classification of genes encoding transport proteins, it was noticed that some genes were being identified and classified with both EC and TC numbers. This occurs because some proteins can be classified by both systems with the same function. For example, HP0145 encodes the cbb3-type cytochrome c oxidase subunit II and was assigned with the EC number 1.9.3.1 and TC number 3.D.4.3.#, having both the same function: the proton-translocating cytochrome oxidase. Another example is HP1270, encoding NADH dehydrogenase subunit K, with the EC 1.6.99.5 and the TC number 3.D.1.1.# which belongs to the H+ or Na+-translocating NADH Dehydrogenase (NDH) Family.

In this re-annotation, 131 genes were assigned only with TC numbers, meaning that these genes encode exclusively transport proteins. Table 4 shows the distribution of these encoded transport proteins in the classes of the transporter classification database.

Table 4: Distribution of TC numbers in the TCDB class

Class	**TC numbers**
1: Channels/Pores	18
2: Electrochemical Potential-driven Transporters	47
3: Primary Active Transporters	57
4: Group Translocators	1
5: Transmembrane Electron Carriers	0
8: Accessory Factors Involved in Transport	0
9: Incompletely Characterized Transport Systems	8

Table 4 highlights that 104 (79.4%) of the 131 genes encoding transport proteins assigned only with TC numbers are divided into two classes: 2: Electrochemical Potential-driven Transporters, with 47 genes; and 3: Primary Active Transporters, with 57 genes. The electrochemical potential-driven transporters class gathers several protein families, including sugar porters, monocarboxylate porters, drugs antiporters, amino acid transporters, iron permeases, zinc permeases and phosphate symporters, among others [20]. The Primary Active Transporters use a primary source of energy to drive active transport of a solute against a concentration gradient. Transport systems are included if they, for example, hydrolyze the diphosphate bond of inorganic pyrophosphate, ATP, or another nucleoside triphosphate, to drive the active uptake and/or extrusion of a solute or solutes [20]. In Additional file 1: Table S4 of the supplemental material, the annotation of the genes encoding transporter proteins is displayed, including TCS number and description and TC family number and description. In there it is possible to notice that most of the genes assigned as electrochemical potential-driven transporters belong to the subclass: 2.A. Porters, the subclass of uniporters, symporters, antiporters, including 2.A.1 - The Major Facilitator Superfamily (MFS) and 2.A.6 - The Resistance-Nodulation-Cell Division (RND) Superfamily. On the other hand, the genes assigned as primary active transporters belong to the subclass 3.A - P-P-bond-hydrolysis-driven transporters, in which transporter proteins hydrolyze the diphosphate bond of inorganic pyrophosphate, ATP, or another nucleoside triphosphate, to drive the active uptake and/or extrusion of a solute or solutes [20].

3.6 Association of annotated enzymes with KEGG pathways

Table 5 depicts the association of assigned EC numbers from the annotated genes encoding enzymes, with global pathways from the KEGG database (`http://www.genome.jp/kegg`). Global pathways are universal, and include enzymes from several other pathways. In the table, 3 global pathways are identified with the total amount of EC numbers that compose each, and the number of EC numbers assigned to *H. pylori* 26695 present in each global pathway. We can observe that, as expected, the global metabolic pathway encompasses the largest amount of EC numbers from *H. pylori* 26695, and that the amount of EC numbers present in each global pathway decreases proportionally with the total amount of EC numbers in that pathway.

In Additional file 1: Table S5 of the supplemental material all the pathways in which enzymes have been identified in the new annotation are provided. There, we can see that almost every pathway includes at least one EC number from *H. pylori* 26695. The pathway with the most enzymes present from *H. pylori* 26695 is the carbohydrate metabolism, followed by the amino acid metabolism. Added, they represent more than 37% of all enzymes present in the different pathways. The single pathways with higher representation of *H. pylori* 26695 enzymes are amino sugar and nucleotide sugar pathway, and purine metabolism pathway, both with more than 100 EC numbers present, indicating their importance in the metabolism. Note that encoding enzymes belonging to specific pathways do not confirm that such pathway exists in an organism. *H. pylori* 26695 encodes enzymes from several pathways, nevertheless we cannot establish that such pathways are present in its metabolism, especially when only few genes are present in it.

Table 5: EC number distribution in global pathways

Global pathway	Total amount of EC	*H. pylori* 26695 EC
Metabolic pathways	1382	219
Biosynthesis of secondary metabolites	645	98
Microbial metabolism in diverse environments	489	57

This annotation offers new insights on *H. pylori* 26695 metabolic reactions and gathers information on the distribution of EC numbers in KEGG pathways, identifying pathways that contain EC numbers assigned to *H. pylori* 26695 enzyme encoding genes.

4 Conclusion

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 EC numbers from all metabolic CDS were searched, validated and attributed when found. A total of 191 new functions were assigned, 149 of which attributed to CDS previously classified as hypothetical proteins. 42 new functions were assigned to CDS already annotated; many of them had been classified with only generic annotations. From the new functions assigned, 103 were metabolic, 64 non-metabolic and 24 had generic descriptions, indicating, for instance the localization of the protein. A total of 581 EC numbers were assigned to CDS, being 527 of them complete EC numbers. The total number of genes assigned as hypothetical proteins decreased from 510, from last re-annotation, to 361 in this work.

The new annotation also includes novelties, such as the assignment of transporter superfamily numbers to genes identified as transporter proteins, being identified and assigned TC numbers to 155 transporter proteins, 131 of them only assigned with TC numbers. 24 genes were assigned with both EC and TC numbers, indicating that some functions can be assigned by both classifications systems. For a total of 1573 CDS annotated, 529 (33.6%) were classified with a very high confidence level, which is the highest classification. Moreover, it was demonstrated that Oxidoreductases, Transferases and Hydrolases represent more than 75% of the identified enzymes. Electrochemical potential-driven transporters and primary active transporters represent almost 80% of all the genes classified as transporter proteins and assigned with only TC numbers. Lastly, *H. pylori* 26695 annotated enzymes were associated with KEGG pathways and EC number distribution in global and specific pathways discussed.

These results bring new and more comprehensive information to *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 EC numbers and TC numbers are a fundamental task, since this 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] B. Marshall. Helicobacter pylori: 20 years on. *Clinical Medicine*, 2(2):147–152, 2002.

[2] Z. Ge and D. Taylor. Contributions of genome sequencing to understanding the biology of Helicobacter pylori. *Annual Reviews in Microbiology*, 53:353–387, 1999.

[3] J. G. Kusters, A. H. M. van Vliet and E. J. Kuipers. Pathogenesis of Helicobacter pylori infection. *Clinical Microbiology Reviews*, 19(3):449–490, 2006.

[4] A. C. Costa, C. Figueiredo and E. Touati. Pathogenesis of Helicobacter pylori infection. *Helicobacter*, 14(Supplement s1):15–20, 2009.

[5] A. Marais, G. L. Mendz, S. L. Hazell and F. Mégraud. Metabolism and genetics of Helicobacter pylori: the genome era. *Microbiology and Molecular Biology Reviews*, 63(3):642–674, 1999.

[6] J. F. Tomb, O. White, a. R. Kerlavage et al. The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature*, 388(6642):539–547, 1997.

[7] O. Dias, A. K. Gombert, E. C. Ferreira and I. Rocha. Genome-wide metabolic (re-) annotation of Kluyveromyces lactis. *BMC Genomics*, 13(1):517, 2012.

[8] C. Médigue and I. Moszer. Annotation, comparison and databases for hundreds of bacterial genomes. *Research in Microbiology*, 158(10):724–736, 2007.

[9] I. G. Boneca. A revised annotation and comparative analysis of Helicobacter pylori genomes. *Nucleic Acids Research*, 31(6):1704–1714, 2003.

[10] M. Durot, P.-Y. Bourguignon and V. Schachter. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Reviews*, 33(1):164–190, 2009.

[11] I. Rocha, J. Förster and J. Nielsen. Design and Application of Genome-Scale Reconstructed Metabolic Models. In A. L. Osterman and S. Y. Gerdes (editors), *Microbial Gene Essentiality: Protocols and Bioinformatics*, volume 416 of *Methods in Molecular Biology*, pages 409–431. Humana Press, 2008.

[12] C. H. Schilling, M. W. Covert, I. Famili, G. M. Church, J. S. Edwards and B. O. Palsson. Genome-Scale Metabolic Model of Helicobacter pylori 26695. *Journal of Bacteriology*, 184(16):4582–4593, 2002.

[13] I. Thiele, T. D. Vo, N. D. Price and B. O. Palsson. Expanded Metabolic Reconstruction of Helicobacter pylori (iIT341 GSM / GPR): an In Silico Genome-Scale Characterization of Single- and Double-Deletion Mutants. *Journal of Bacteriology*, 187(16):5818–5830, 2005.

[14] O. Dias, M. Rocha, E. Ferreira and I. Rocha. Merlin: Metabolic models reconstruction using genome-scale information. In J. R. Banga, P. Bogaerts, J. V. Impe, D. Dochain and I. Smets (editors), *Proceedings of the 11th International Symposium on Computer Applications in Biotechnology (CAB 2010), Leuven, Belgium, July 7-9, 2010*, pages 120–125. 2010.

[15] R. Finn, J. Clements and S. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl 2):W29–W37, 2011.

[16] The UniProt Consortium. Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Research*, 40(D1):D71–D75, 2012.

[17] M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Söhngen, M. Stelzer, J. Thiele and D. Schomburg. Brenda, the enzyme information system in 2011. *Nucleic Acids Research*, 39(suppl 1):D670–D676, 2011.

[18] M. H. Saier, C. V. Tran and R. D. Barabote. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Research*, 34(suppl 1):D181–D186, 2006.

[19] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[20] M. H. Saier. A Functional-Phylogenetic Classification System for Transmembrane Solute Transporters. *Microbiology and Molecular Biology Reviews*, 64(2):354–411, 2000.

[21] A. Krogh, B. Larsson, G. von Heijne and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, 2001.