The *secE* Gene of *Helicobacter pylori*

Claudine Médigue,^{1,2} Benjamin Chun-Yu Wong,³ Marie Chia-Mi Lin,⁴ Stéphanie Bocs,² and Antoine Danchin^{5*}

Génétique des Génomes Bactériens, Institut Pasteur, Paris,¹ and Génomique Comparative des Micro-organismes Pathogènes, Génopole, Evry,² France, and Department of Medicine³ and Institute of Molecular Biology,⁴ University of Hong Kong, and HKU-Pasteur Research Center,⁵ Hong Kong, China

Received 28 December 2001/Accepted 6 February 2002

Despite extensive annotation by two independent teams, the *Helicobacter pylori* genome appeared to lack a complete secretion machinery. The use of clinical isolates to substantiate *in silico* annotation is used here to identify the missing *secE* component of the major secretion machinery of *Helicobacter pylori*.

Two independent sequences of the *Helicobacter pylori* genome, with many clinical isolates from hospital laboratories, have been annotated by two independent consortia. It is therefore expected that the identification errors have been kept to a minimum. Naturally, some genes of unknown function may have escaped attention while spurious sequences have been taken for bona fide genes. It is, however, important for the community to be sure that no essential gene has been missed or misannotated since most scientists now rely on data library searches to substantiate their experiments. How would the discovery of an important gene sequence stand in a much explored (and patented) sequence? By combining and chaining a series of independent tasks meant to identify coding se-

quences (CDSs) in bacterial genomes (2), we explored *in silico* the genome of *H. pylori* to see whether important genes have escaped notice. To predict CDSs, this strategy combined periodical Markov chain analysis (the original GeneMark program, which works by discrimination of relevant protein coding sequences from the background [5]) and now popular derivatives (such as Glimmer, which works by assimilation from previously known sequences [10]) together with BlastX computation and identification of tRNAs, terminators, and putative ribosome binding sites by using the platform Imagen (6). In contrast with the usual approaches, which mostly rest on one single method for gene identification, this allows one to discriminate with fair certainty between spurious genes and bona fide genes.

FIG. 1. The *nusG* region in *H. pylori*. Four frames with putative CDSs are displayed in the *tufB* *nusG* region. The upper line is the prediction by GeneMark, and the lower line is the prediction by Glimmer. While Glimmer predicts a long CDS in the opposite strand, GeneMark predicts an excellent CDS between *mmpG* and *nusG*. The present work shows that this corresponds to the *secE* gene.

* Corresponding author. Mailing address: HKU Pasteur Research Centre, Dexter HC Man Building, 8, Sassoon Road, Pokfulam, Hong Kong, China. Phone: 852 2816 8403. Fax: 852 2168 4427. E-mail: adanchin@hkucc.hku.hk.

posed a longer sequence in the opposite strand (and a poor indication of the former putative CDS) (Fig. 1). In many instances GeneMark is preferred over Glimmer because it discriminates between coding and noncoding regions while Glimmer assimilates putative coding regions to known ones. This results in the carrying over of features from one strand to its complement as soon as they are palindromic in nature (e.g.,

FIG. 3. The SecE protein in context. Phylogenetically conserved residues are in grey. Inverted colors correspond to the transmembrane segment of the protein. Stars correspond to residues encoded by synonymous codons in the various *H. pylori* isolates. The six variable residues are shown as indicated (e.g., K > E at the beginning of the protein).

the RNY coding rule is true both in the coding strand and in its complement [9]). A Blast search revealed that the latter sequence did not display similarity with known sequences, whereas the former was similar to the *secE* gene present in a variety of organisms. The neighboring gene order was consistent with this, since *secE* is often part of an operon with genes involved in translation, as shown by Pohlschroder et al. at a time when the first genome sequences appeared (8).

The sequence of the two known *H. pylori* genomes (<http://genolist.pasteur.fr/PyloriGene>) suggested that *secE* was a bona fide gene, but the data were too scarce and contradictory to warrant this hypothesis. (Only eight codons differed in the two reference sequences, including one modified in its second base position, therefore arguing against the hypothesis. One expects that most sequences display synonymous mutations; therefore, mutations would appear in the third codon position.) To substantiate this interpretation, we sequenced the homologous region in seven *H. pylori* isolates collected at the University of Hong Kong. The new sequences differed from those of the model strains at 10 more significant positions. Twelve positions corresponded to synonymous replacements while six others yielded conservative replacements in the *secE* frame (Fig. 2). The only position which was not strictly conservative (AAA to GAA) is located immediately after the start of the protein at a nonconservative position (a gap in some SecE proteins). These same mutations would yield several non-

conservative replacements in the complementary putative coding sequence (in particular an AAA (lysine) → ATA (isoleucine) replacement). Interestingly, the sequences indicate that the Asian isolates are from a common group that differs from those of the rest of the world (1).

The structure of the SecE protein is comprised of a cytoplasmic domain, a transmembrane helix, and a periplasmic domain. The bacterial translocase consists of the SecEYG membrane protein complex and the peripheral-membrane-associated SecA dimer (for a review, see reference 3). In the present sequence, residues in both the cytoplasm and the periplasm are conserved (in particular, the residues making contact with SecY) in comparison with other bacterial counterparts (this is in complete agreement with the identifications made by Hartmann et al. [4] and Murphy and Beckwith [7], thus further substantiating our identification of the *secE* gene) while the extremities of the transmembrane helix are also conserved, with the hydrophobic core preserving not the residue but its hydrophobic nature (11) (Fig. 3). SecE, which in most bacteria is comprised of a single transmembrane domain (as predicted by the present study), is an essential component of the highly conserved general secretion machinery. In addition to placing *H. pylori* in the normal class of secreting bacteria and providing a new example of the SecE structure, this work emphasizes the need for continuous reannotation of genome sequences, including those regions which have been previously annotated thoroughly, and the associated experimental substantiation.

ADDENDUM IN PROOF

We have recently noted that Doig et al. (P. Doig, B. L. de Jonge, R. A. Alm, E. D. Brown, M. Uria-Nickelsen, B. Noonan, S. D. Mills, P. Tummino, G. Carmel, B. C. Guild, D. T. Moir, G. F. Vovis, and T. J. Trust, *Microbiol. Mol. Biol. Rev.* **63**:675–707, 1999) in their in silico prediction of the *H. pylori* genes have suggested in passing the existence of a *secE* counterpart. However, this prediction has not been included in the corresponding databases, perhaps for want of experimental substantiation. This is now done in the present work.

REFERENCES

1. Achtman, M., T. Azuma, D. E. Berg, Y. Ito, G. Morelli, Z. J. Pan, S. Suerbaum, S. A. Thompson, A. van der Ende, and L. J. van Doorn. 1999. Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol. Microbiol.* **32**:459–470.
2. Boes, S., A. Danchin, and C. Médigue. 2002. Re-annotation of genome microbial coding sequences: finding new genes and incorrectly annotated genes. *BMC Bioinformatics* **3**:5. [Online.] <http://www.biomedcentral.com/1471-2105/3/5>.
3. Driessen, A. J., E. H. Manting, and C. van der Does. 2001. The structural basis of protein targeting and translocation in bacteria. *Nat. Struct. Biol.* **8**:492–498.
4. Hartmann, E., T. Sommer, S. Prehn, D. Gorlich, S. Jentsch, and T. A. Rapoport. 1994. Evolutionary conservation of components of the protein translocation complex. *Nature* **367**:654–657.
5. McIninch, J. D., W. S. Hayes, and M. Borodovsky. 1996. Applications of GeneMark in multispecies environments. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4**:165–175.
6. Médigue, C., F. Rechenmann, A. Danchin, and A. Viari. 1999. Imagen: an integrated computer environment for sequence annotation and analysis. *Bioinformatics* **15**:2–15.
7. Murphy, C. K., and J. Beckwith. 1994. Residues essential for the function of SecE, a membrane component of the *Escherichia coli* secretion apparatus, are located in a conserved cytoplasmic region. *Proc. Natl. Acad. Sci. USA* **91**:2557–2561.
8. Pohlschroder, M., W. A. Prinz, E. Hartmann, and J. Beckwith. 1997. Protein

- translocation in the three domains of life: variations on a theme. *Cell* **91**: 563–566.
9. **Rother, K. I., O. K. Clay, J. P. Bourquin, J. Silke, and W. Schaffner.** 1997. Long non-stop reading frames on the antisense strand of heat shock protein 70 genes and prion protein (PrP) genes are conserved between species. *Biol. Chem.* **378**:1521–1530.
  10. **Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White.** 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**:544–548.
  11. **Veenendaal, A. K., C. van Der Does, and A. J. Driessen.** 2001. Mapping the sites of interaction between SecY and SecE by cysteine scanning mutagenesis. *J. Biol. Chem.* **276**:32559–32566.