ELSEVIER

FEMS Microbiology Letters 206 (2002) 253–258

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Evolution of functional polymorphism in the gene coding for the *Helicobacter pylori* cytotoxin

Xuhuai Ji [a,b], Francesco Frati [c], Silvia Barone [a], Cristina Pagliaccia [a], Daniela Burroni [a], Guoming Xu [b], Rino Rappuoli [a], Jean-Marc Reyrat [a,1], John L. Telford [a,*]

[a] *IRIS, Chiron SpA, Via Fiorentina 1, 53100 Siena, Italy*

[b] *Department of Gastroenterology, Changhai Hospital, Second Military Medical University, 200433 Shanghai, PR China*

[c] *Department of Evolutionary Biology, University of Siena, Via Mattioli 4, 53100 Siena, Italy*

Received 23 October 2001; accepted 23 November 2001

First published online 11 December 2001

Abstract

There are two functionally different alleles of the *Helicobacter pylori vacA* gene, which code for proteins with different in target cell specificity. The alleles (m1 and m2) differ by approximately 50% in amino acid sequence in a 300 amino acid region, the m-region, which determines specificity. An analysis of partial likelihood anomalies in a set of eight Chinese and six Western *vacA* genes revealed highly significant phylogenetic deviation of a region of the gene including the m-region. Phylogenetic analysis of the conserved regions of these genes failed to reveal any distinction between m1 alleles and m2 alleles, however clear cut geographic variation was observed. In the m-region, the m1 alleles also show separate clustering of Chinese and Western isolates, however the m-region of the m2 alleles has a phylogenetic structure markedly different from the rest of the gene. The data indicate that the m2 m-region was acquired and spread through the population by horizontal transfer of DNA.

Keywords: Toxin; VacA; Evolution; Geographic variation; *Helicobacter pylori*

1. Introduction

The human gastric pathogen, *Helicobacter pylori*, causes chronic gastritis and peptic ulcer and is associated with increased risk of gastric cancer [1,2]. A major virulence factor produced by the bacteria is the vacuolating cytotoxin, VacA, which causes vacuolar degeneration of target cells in vitro and gastric epithelial erosion in vivo [3]. VacA is produced as a 140 kDa precursor polypeptide [4]. The carboxy-terminal 45 kDa of the precursor has autotransporter activity and is cleaved from the protein after translocation across the outer-membrane [4,5]. The toxin is released from the bacteria as a high molecular mass oligomeric protein consisting of several copies of an approximately 95 kDa polypeptide [6,7]. Each monomer is further structured in two distinct subunits of approximately 37 kDa and 58 kDa (Fig. 1). The 37 kDa subunit contains the functional vacuolating activity [8] and the 58 kDa subunit is responsible for receptor binding and interaction with the target cells [9].

There are two forms of VacA, called m1 and m2 [6,10], which differ in their capacity to bind target cells [11]. The m2 form fails to bind and hence to intoxicate HeLa cells but is fully active if expressed intracellularly by DNA transfection. Both forms bind and intoxicate a rabbit kidney cell line RK13 and primary cells from human gastric biopsies. The region of the protein which defines target cell specificity is within a region of about 300 amino acids in

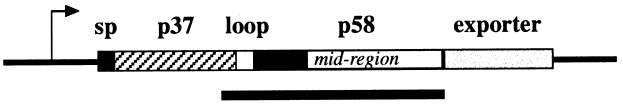

Fig. 1. Schematic representation of the *vacA* gene. sp, signal peptide; p37 and p58, 37 kDa and 58 kDa subunits; loop, flexible loop containing cleavage site between subunits; exporter, outer-membrane autotransporter. Bar below the gene shows the anomalous region detected by the program Plato.

* Corresponding author. Tel.: +39 (577) 243470; Fax: +39 (577) 243564.

E-mail address: john_telford@biocine.it (J.L. Telford).

[1] Present address: Mycobacterial Genetics Unit, Institute Pasteur, 25 Rue du Dr. Roux, 75015 Paris, France.

0378-1097/02/$22.00
PII: S0378-1097(01)00543-2

the 58 kDa subunit which differs by about 50% in amino acid sequence between the two forms of the protein [12]. In contrast, the two forms share greater than 90% amino acid sequence identity in the rest of the protein. An understanding of the origin of this functional polymorphism might shed light on the evolution of virulence in *H. pylori*.

Previous studies of several genes from *H. pylori* have failed to reveal any strong phylogenetic structure [13] although week clonality was detected in some Asian strains [14] and differences were detected between European and Asian populations in The Netherlands [15]. It has been suggested that the lack of detectable structure is due to frequent recombination in *H. pylori* [13].

2. Materials and methods

2.1. H. pylori isolates

H. pylori single colony isolates were obtained from dyspeptic patients undergoing upper gastrointestinal endoscopy at the Changhai Hospital, Shanghai, China. Cultures were prepared as previously described [16]. Western isolates with previously published *vacA* gene sequence were obtained from the respective authors.

2.2. Genomic DNA isolation and DNA sequencing

Genomic DNA extraction was as described previously [17]. On the basis of published NCTC 11638 and 95-54 gene sequences [4,11], oligonucleotides were synthesized to amplify the *vacA* genes in three overlapping segments. The PCR products were directly sequenced using an ABI PRISM dye terminator cycle sequencing ready reaction kit with Amplitaq DNA polymerase in an automatic DNA sequencer model 373A, Applied Biosystems, USA. GenBank accession numbers for the determined sequences are given in Table 1.

Table 1
GenBank accession numbers of sequences analyzed

Strain	Acc. no.	Reference
11638	S72494	[4]
60190	U05676	[6]
95-54	U95971	[11]
185-44	Z26883	[5]
26695	AE000511	Tomb et al., 1997
Tx30a	U29401	[10]
43526	AF001358	Ogura et al., unpublished
5060d	AF050328	this work
5038c	submitted	this work
3554a	submitted	this work
4611a	submitted	this work
1811a	AF050326	this work
5114a	AF50327	this work
3295b	AF050319	this work
5147c	AF050320	this work

2.3. Phylogenetic analysis

Phylogenetic analysis was conducted using the program PAUP* [18]. The maximum likelihood (ML) approach was used following the procedure proposed in Swofford et al. [19], as outlined in Frati et al. [20] and Sullivan et al. [21]. The best model of evolution was determined by using the likelihood ratio test [22] on a set of initial trees. Heuristic searches were then run, optimizing model parameters during the run. Statistical confidence of the nodes was assessed with bootstrap analysis by running 100 replicates with the parameters fixed at those estimated during the search. The ratio between the number of non-synonymous (d_N) and synonymous (d_S) substitutions was estimated using the Nei and Gojobori [23] method as implemented in MEGA version 2.0 [24]. The program Plato [25] was used to detect recombination by assessing regions of likelihood anomalies across the whole gene.

3. Results

H. pylori was cultured from gastric biopsies from 20 patients from the Shanghai region in China and DNA was prepared from single colonies. Partial sequence analysis of the m-region revealed that 15 of the patients were infected with m2 bearing strains and five with m1 bearing strains. The nucleotide sequence of the entire *vacA* gene from eight representative strains from China (four m1 and four m2) was determined. The deduced amino acid sequences were compared with four previously published sequences of *vacA* m1 alleles and two m2 alleles from Western isolates. The sequence comparison permitted a more precise localization of the m-region. The first difference, consistent in all strains, between m1 and m2 forms is found at amino acid 501 of the reference strain NCTC 11638 which is Arg in all m1 alleles and Lys in all m2 alleles. The last consistent difference between m1 and m2 forms is at amino acid 834 of NTCT 11638 which is Gly in the m1 forms and Asn in the m2 forms. This latter position is about 30 amino acids upstream of the cleavage point between the mature toxin and the outer-membrane exporter domain [26]. Within this region, 45% (151/334) of the amino acid positions vary in at least one of the genes sequenced and of these, 59 positions (39%) are identical within each m-type. It is hence likely that the functional difference between m1 and m2 forms resides in these consistent differences in amino acid sequence.

In contrast, in the conserved regions coding for the p37 subunit and the putative outer-membrane exporter, no consistent differences were observed between m1 and m2 alleles, however, Chinese isolates were more similar to each other than to isolates from Western countries. In the exporter region, of a total of 67 positions of variability in the deduced amino acid sequences, 15 are consistently different between the Chinese and Western isolates.

The above observations indicate different phylogenetic structure between the conserved regions of the gene and the allelic m-region. To assess this, partial likelihood anomalies were determined within the DNA sequence using the program Plato [25] which uses a sliding window to identify regions within an alignment which do not fit with a global phylogenetic hypothesis. The initial phylogeny was determined using maximum likelihood and the general time-reversible model [27] with Γ rate heterogeneity (GTR+Γ). Plato identified only a single anomalous region extending from position 881 to 2769 in the sequence lineup (z-value = 15.337, significant if > 4.057). This region codes for a part of the protein extending from about 50 amino acids upstream of the start of the 58 kDa subunit to the end of the mature secreted protein. No anomalies larger than 4 bp were detected when the m1 sequences were analyzed independently of the m2 sequences suggesting that the anomalies arise due to phylogenetic differences between m1 and m2 forms.

To confirm this, ML analysis was performed on the three regions 1–180, 881–2769 and 2770–4032 of the gene independently. Fig. 2 shows the topologies obtained and the bootstrap support for the major nodes. Both the non-allelic 1–880 region (Fig. 2A) and the 2769–4032 region (Fig. 2C) show clear division into Chinese and Western clades with no evidence of separation of m1 and m2 types. In marked contrast, the major distinction in the topology of the 881–2769 region is a clear separation between m1 and m2 types (Fig. 2B). However, there is reasonably strong bootstrap support for independent clustering of Chinese and Western sequences within both the m1 and m2 clades. Hence, the topology of this central region, which essentially encompasses the p58 subunit, indicates that there are four independent clusters of genes in the

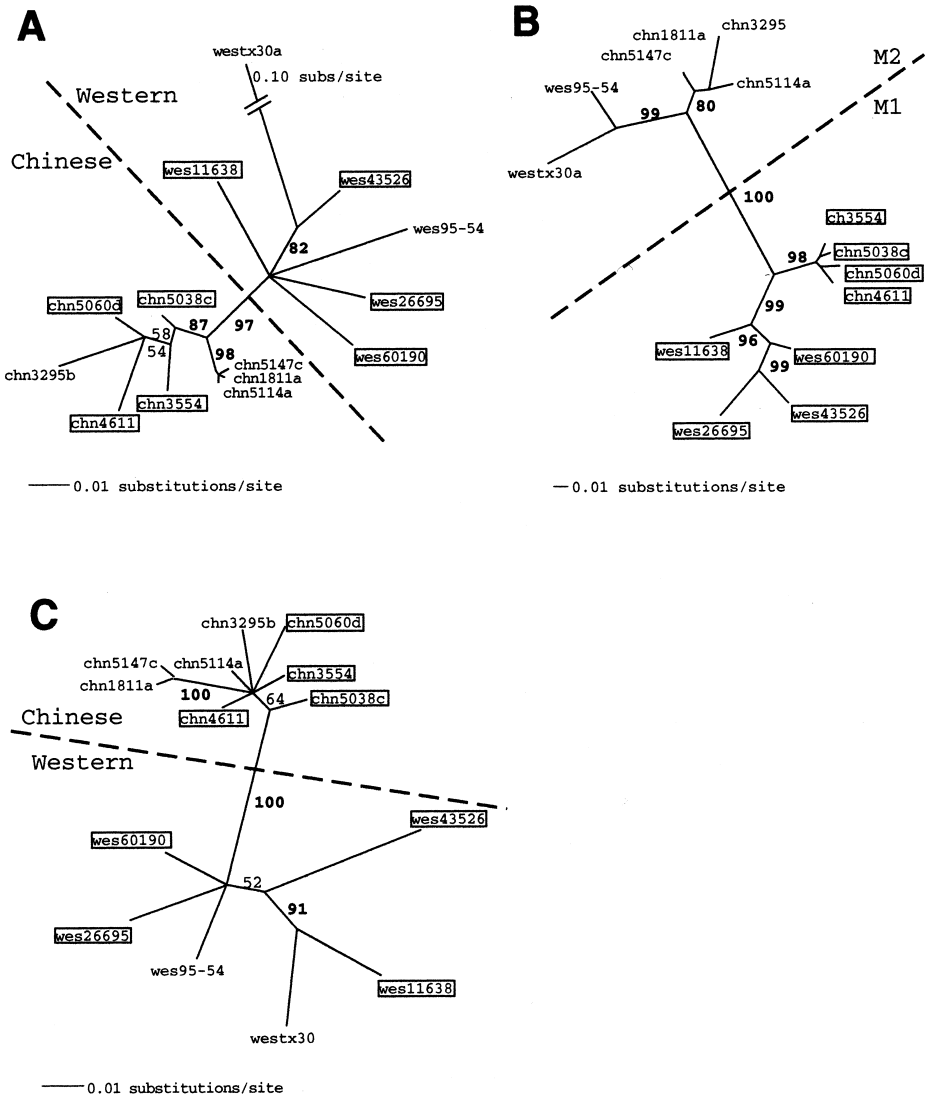

Fig. 2. Phylogenetic analysis of the evolutionary relationship between Chinese and Western *vacA* genes based on maximum likelihood analysis (using the GTR+Γ model) of the regions 1–880 (A), 881–2769 (B) and 2770–4032 (C). The designations of m1 strains are boxed. Bootstrap values are shown at the nodes. Nodes with bootstrap values below 50% have been collapsed.

sample (Chinese m1, Western m1, Chinese m2 and Western m2).

We were surprised that the anomalous region identified by Plato is larger than the m-region in the protein which can be clearly identified on the basis of the amino acid sequence comparison (see above). In fact, the anomalous region ends very close to end of the m-region but begins approximately 180 amino acids upstream of the first consistent amino acid difference between m1 and m2 forms. To analyze the anomalous region in more detail, the ML topology for the region 881–1539 corresponding to the region upstream of the start of the m-region was compared with topologies for the regions 1540–2058 and 2059–2769 corresponding to the m-region in two parts. The reason for this latter split was on the one hand to create regions of approximately similar length and on the other because some strains have been identified which appear to have hybrid m1/m2 genes [28,12] with breakpoints around a short conserved region near the center of the m-region. As can be seen from Fig. 3A, the topology of the first region, like that of the conserved p37 and export regions, shows a major split between Chinese and Western genes (bootstrap support, 100%). In contrast, both halves of the m-region (Fig. 3B and C) split clearly into m1 and

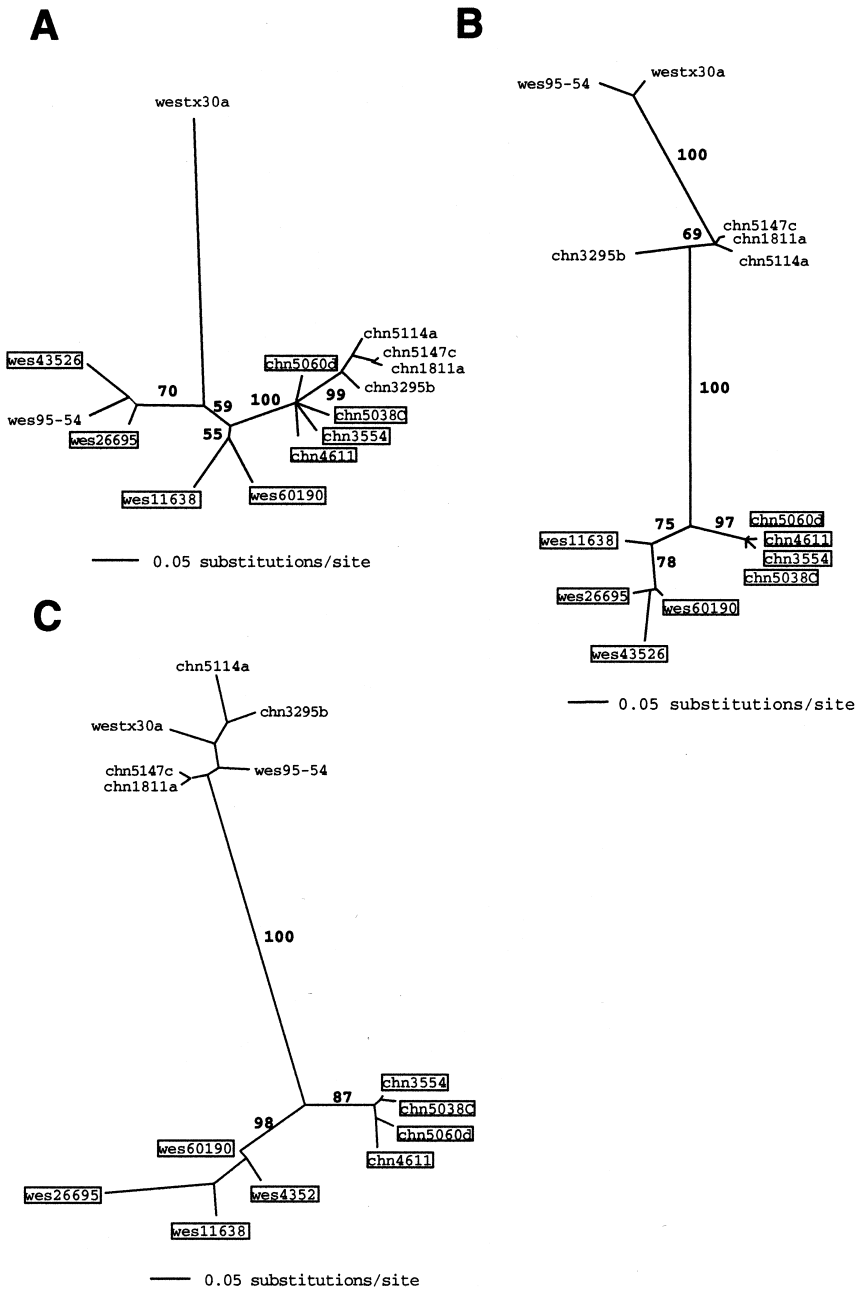

Fig. 3. Phylogenetic analysis showing the different evolutionary relationships obtained for the regions 881–1539 (A), 1540–2058 (B) and 2059–2769 (C). Topologies are derived using the maximum likelihood method using the GTR+Γ model. Bootstrap values are shown at the nodes. Nodes with bootstrap values below 50% have been collapsed.

m2 forms (bootstrap support, 100%). Interestingly, the m1 genes split into Chinese and Western clusters in both halves of the m-region. Thus the topologies of all regions tested in the m1 genes showed the same clustering of Chinese genes separately from Western genes. These data indicate that the m1 form of the gene has existed since before the separation of the Chinese and Western populations of *H. pylori*.

In contrast, the topologies obtained for the two halves of the m-region of the m2 alleles were strikingly different. The 3′ half of the m-region of the m2 alleles is highly conserved and showed no geographic separation (Fig. 3C). ML analysis of a short 282 bp (2064–2346) sequence within this region from a total of 15 Chinese isolates confirmed the lack of any geographic clustering (data not shown). In the 5′ half of the m-region the Western m2 alleles clustered together with 100% bootstrap support. Clustering of the Western genes separate from the majority of the Chinese genes was confirmed by ML analysis of a short region of 174 bp (1540–1714) in this region from 15 Chinese isolates (data not shown). However this analysis revealed three Chinese genes which clustered with the Western sequences (bootstrap support 71%). Thus, although the 3′ half of the m2 m-region is highly conserved and shows no geographic clustering, two distinct 5′ groups can be detected which are not strictly clustered along geographic lines in that one group contains both Western and Chinese genes and the other Chinese genes only. These two groups code for substantially different amino acid sequences in this region. There are in fact about 17% positions which are identical within each group but different between the groups.

In order to test whether positive selection on the different regions of the gene contributed to the difference in phylogenetic histories, we estimated the ratio between non-synonymous and synonymous substitutions (d_N/d_S) in the whole gene and in the selected subregions of the gene. In the case of positive selection we would expect this ratio to be >1.0, while if only purifying selection is acting, the ratio should be much less than 1.0 as purifying selection will constrain non-synonymous sites at a much greater extent than the neutral or nearly neutral synonymous sites. The d_N/d_S ratio averaged across all pairwise comparisons of the complete sequences was 0.336 ± 0.111. Similar average values were calculated for all the subsets examined (1–880, 880–2769, 2770–4032, 880–1539, 1540–2058, 2059–2769), the highest value (0.509 ± 0.213) being found in the 1540–2058 region. These data suggest that positive selection does not contribute significantly to the evolution of this gene, nor can it explain the apparent discrepancies in the phylogenetic history of different gene regions.

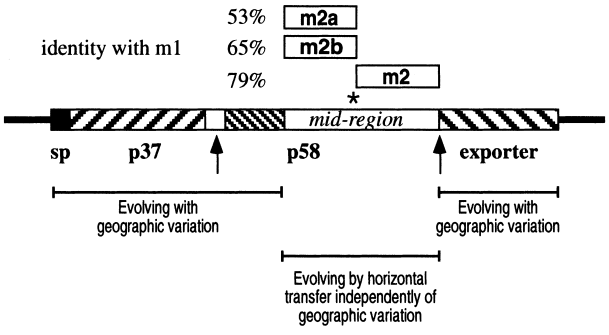

Fig. 4. Schematic representation of the VacA protein showing the position of the m-region and the relative distance of the m2a and m2b alleles from the m1 allele. The asterisk shows the approximate position of the junction between m1 and m2 in the chimeric *vacA* gene reported by Ji et al. [12].

4. Discussion

The sequence variation in the conserved regions of the *vacA* gene indicates that the Chinese isolates form a consistent evolutionary group quite separate from Western isolates, although there is little or no support for any topology within the groups. Given the free recombination in *H. pylori* demonstrated by homoplasmy analysis [13], it is perhaps not surprising that a strong phylogenetic signal is only detectable between populations which have been geographically separate for a considerable time. The data presented indicate that while the m-region of m1 alleles has an evolutionary topology similar to the conserved regions of the genes, the m-region of m2 alleles has an evolutionary history independent of the rest of the gene and strongly suggests that the m2 m-region has spread after the separation of Chinese and Western strains. Furthermore, there is a clear contrast between the differences in the 5′ m-region and the lack of divergence of the 3′ m-region of all m2 alleles which suggests that these two regions of the m2 m-region have evolved independently (shown schematically in Fig. 4). In support of this, Pan et al. [28] have recently described clinical isolates with chimeric m1/m2 *vacA* genes in which the breakpoint between the two forms is very close to the point at which the two m2 alleles converge. Furthermore, Ji et al. [12] have demonstrated that the sequences which determine cell specific toxicity are limited to the 5′ half of the m-region. Whereas different rates of evolution in different regions of the same gene might be expected, dramatically different patterns of evolution between different regions of the same set of genes are best explained by recombination. Hence it is likely that the mosaicism in the *vacA* gene is maintained in the population by genetic recombination and since there is a single copy of the *vacA* gene in *H. pylori* this implies horizontal transfer of DNA.

The high incidence of the m2 allele in the Chinese population may indicate that it was first acquired in this population and subsequently spread to Western populations. Alternatively, the m2 allele may have a selective advantage in the Chinese population due to human polymorphism. The two isoforms of the protein have distinct functional

differences in cell binding indicating differences in cell surface receptors for the two forms (Pagliaccia et al. [11]). It is tempting to speculate that there may be polymorphism in the expression of these receptors in different populations. In this context, the higher variability observed in the allelic m-region may be related to the ability to provide the cytotoxin with different variants to respond to polymorphism of the receptors of the host.

Acknowledgements

We wish to thank A. Muzzi for synthesis of oligonucleotides and S. Guidotti for automatic DNA sequencing. We thank G. Corsi for artwork, A. Covacci, V. Scarlato, and G. Grandi for helpful discussions. This work was supported by European Commission grants IC18CT95-0024 and ICA4CT1999-10010 and National Natural Science Foundation of China grant No. 396706481.

References

[1] Cover, T.L. and Blaser, M.J. (1992) *Helicobacter pylori* and gastroduodenal disease. Annu. Rev. Med. 43, 135–145.
[2] Covacci, A., Telford, J.L., Del Giudice, G., Parsonnet, J. and Rappuoli, R. (1999) *Helicobacter pylori* virulence and genetic geography. Science 284, 1328–1333.
[3] Reyrat, J.-M., Pelecic, V., Papini, E., Montecucco, C., Rappuoli, R. and Telford, J.L. (1999) Towards deciphering the *Helicobacter pylori* cytotoxin. Mol. Microbiol. 34, 197–204.
[4] Telford, J.L., Ghiara, P., Dell Orco, M., Burroni, D., Bugnoli, M., Tecce, M., Censini, S., Covacci, A., Xiang, Z., Pappini, E. and Rappuoli, R. (1994) Gene structure of the *Helicobacter pylori* cytotoxin and evidence of its key role in gastric disease. J. Exp. Med. 179, 420–460.
[5] Schmitt, W. and Haas, R. (1994) Genetic analysis of the Helicobacter pylori vacuolating cytotoxin: structural similarities with the IgA protease type of exported protein. Mol. Microbiol. 12, 307–319.
[6] Cover, T.L., Tummuru, M.K.R., Cao, P., Thompson, S.A. and Blaser, M.J. (1994) Divergence of genetic sequences for the vacuolating cytotoxin among *Helicobacter pylori* strains. J. Biol. Chem. 269, 10566–10573.
[7] Lupetti, P., Heuser, J.E., Manetti, R., Massari, P., Lanzavecchia, S., Bellon, P.L., Dallai, R., Rappuoli, R. and Telford, J.L. (1996) Oligomeric and subunit structure of the *Helicobacter pylori* vacuolating cytotoxin. J. Cell Biol. 133, 801–807.
[8] de Bernard, M., Burroni, D., Papini, E., Rappuoli, R., Telford, J.L. and Montecucco, C. (1998) Identification of the Helicobacter pylori VacA toxin domain active in the cytosol. Infect. Immun. 66, 6014–6016.
[9] Reyrat, J.-M., Lanzavecchia, S., Lupetti, P., de Bernard, M., Pagliaccia, C., Pelecic, V., Charrel, M., Ulivieri, C., Norais, N., Ji, X., Cabiaux, V., Papini, E., Rappuoli, R. and Telford, J.L. (1999) 3D imaging of the 58 kDa cell binding subunit of the *Helicobacter pylori* cytotoxin. J. Mol. Biol. 290, 459–470.
[10] Atherton, J.C., Cao, P., Reek, R.M., Tummuru, M.K.R. and Blaser, M.J. (1995) Mosaicism in vacuolating cytotoxin alleles of *Helicobacter pylori*. J. Biol. Chem. 270, 17771–17777.
[11] Pagliaccia, C., de Bernard, M., Lupetti, P., Ji, X., Cover, T.L., Papini, E., Rappuoli, R., Telford, J.L. and Reyrat, J.M. (1998) The m2 form of the *Helicobacter pylori* cytotoxin has cell type-specific vacuolating activity. Proc. Natl. Acad. Sci. USA 95, 10212–10217.
[12] Ji, X., Fernandez, T., Burroni, D., Atherton, J.C., Reyrat, J.M., Rappuoli, R. and Telford, J.L. (2000) Cell specificity of the Helicobacter pylori cytotoxin is determined by a short region in the polymorphic mid-region. Infect. Immun. 68, 3754–3757.
[13] Suerbaum, S., Maynard Smith, J., Bapumia, K., Morelli, G., Smith, N.H., Kunstmann, E., Dryeck, I. and Achtman, M. (1998) Free recombination in *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA 95, 12619–12624.
[14] Achtman, M., Azuma, T., Berg, D.E., Ito, Y., Morelli, G., Pan, Z-J., Suerbaum, S., Thompson, S.A., van der Ende, A. and van Doorn, L-J. (1999) Recombination and clonal groupings within Helicobacter pylori from different geographical regions. Mol. Microbiol. 32, 459–470.
[15] van der Ende, A., Pan, Z.J., Bart, A., van der Hulst, R.W., Feller, M., Xiao, S.D., Tytgat, G.N. and Dankert, J. (1998) cagA-positive Helicobacter pylori populations in China and The Netherlands are distinct. Infect. Immun. 66, 1822–1826.
[16] Figura, N., Guglielmetti, P., Rossilini, A., Barberi, A., Cusi, G., Musmanno, R.A., Russi, M. and Quarranta, S. (1989) Cytotoxin production by *Campylobacter pylori* strains isolated from patients with peptic ulcers and from patients with chronic gastritis only. J. Clin. Microbiol. 27, 225–226.
[17] Xiang, Z., Censini, S., Bayeli, M., Telford, J.L., Figureura, N., Rappuoli, R. and Covacci, A. (1995) Analysis of expression of CagA and VacA virulence factors in 43 strains of *Helicobacter pylori* reveals that clinical isolates can be divided into two major types and that *cagA* is not necessary for expression of the vacuolating cytotoxin. Infect. Immun. 63, 94–98.
[18] Swofford, D.L. (1998) PAUP*. Phylogenetic Analysis Using Parsimony * and other methods, Version 4. Sinauer Associates, Sunderland, MA.
[19] Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996) Phylogenetic inference. In: Molecular Systematics, 2nd Edn. (Hillis, D.M., Moritz, C. and Mable, B.K., Eds.), pp. 407–514. Sinauer Associates, Sunderland, MA.
[20] Frati, F., Simon, C., Sullivan, J. and Swofford, D.L. (1997) Evolution of the mitochondrial cytochrome oxidase II gene in Collembola. J. Mol. Evol. 44, 145–158.
[21] Sullivan, J., Markert, J.A. and Kilpatrick, C.W. (1997) Phylogeography and molecular systematics of the *Peromyscus atzecus* species group Rodentia: Muridae inferred using parsimony and likelihood. Syst. Biol. 46, 426–440.
[22] Yang, Z., Goldman, N. and Friday, A. (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. 44, 384–439.
[23] Nei, M. and Gojobori, T. (1986) Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.
[24] Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. (2001) *MEGA2:* Molecular Evolutionary Genetics Analysis software. Bioinformatics, in press.
[25] Grassly, N.C. and Holmes, E.C. (1997) A likelihood method for the detection of selection and recombination using sequence data. Mol. Biol. Evol. 14, 239–247.
[26] Nguyen, V.Q., Caprioli, R.M. and Cover, T.L. (2001) Carboxy-terminal proteolytic processing of Helicobacter pylori vacuolating toxin. Infect. Immun. 69, 543–546.
[27] Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.
[28] Pan, Z.J., Berg, D.E., van der Hulst, R.W.M., Su, W.W., Raudonikiene, A., Xiao, S.D., Dankert, J., Tytgat, G.N.J. and van der Ende, A. (1998) Prevalence of vacuolating cytotoxin production and distribution of distinct vacA alleles in *Helicobacter pylori* from China. J. Infect. Dis. 178, 220–226.