Sequence anomalies in the Cag7 gene of the Helicobacter pylori pathogenicity island

GUOYING LIU†, TIMOTHY K. MCDANIEL‡, STANLEY FALKOW‡, AND SAMUEL KARLIN†§

†Department of Mathematics, Stanford University, Stanford, CA 94305-2123; and ‡Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305-5124

Contributed by Samuel Karlin, April 5, 1999

ABSTRACT The severity of Helicobacter pylori-related disease is correlated with a pathogenicity island (the Cag region of about 26 genes) whose presence is associated with the up-regulation of an IL-8 cytokine inflammatory response in gastric epithelial cells. Statistical analysis of the Cag gene sequences calculated from the complete genome of strain 26695 revealed several unusual features. The Cag7 sequence (1,927 aa) has two repeat regions. Repeat region I runs 317 aa in a form of baaaba proximal to the protein N terminal; repeat region II extends 907 aa in the middle of the protein sequence consisting of 74 contiguous segments composed from selections among six consensus sequences and includes 58 regularly distributed cysteine residues with consecutive cysteines mostly 12, 18, or 24 aa apart. This "regular" cysteine arrangement may provide a scaffolding of linker elements stabilized by disulfide bridges. When Cag7 homologues from different strains are compared, differences were found almost exclusively in the repeat regions, resulting from deletion and/or insertion of repeating units. These observations suggest that the anomalous repetitive structure of the sequence plays an important role in the conformation of Cag7 gene product and potentially in the function of the pathogenicity island. Other facets of the Cag7 sequence show significant charge clusters, high multiplet count, and extremes of amino acid usage.

Helicobacter pylori (HP) is a Gram-negative spiral-shaped bacterium that colonizes the human stomach. About 50% of humans are infected by HP but only 10% exhibit clinical disease, including chronic gastritis, gastric carcinoma, and peptic ulcer (1). The more severe forms of disease are associated with infection by specific strains called type I. Two type I HP strains have been sequenced in their entirety [strains 26695 (2) and J99 (3)]. Virulent HPs differ from less virulent strains (type II) by the presence of a 40-kb block of genes called the Cag pathogenicity island (abbreviated Cag PAI or CagA region; ref. 4). No specific function is established for any gene from the Cag island. However, Cag-positive, but not Cag-negative, strains cause cultured gastric epithelial cells to secrete the proinflammatory cytokine IL-8 (4,5), and this ability is abolished by specific mutation of many of the 26 ORFs found in the Cag island (4–6). Several of these genes are modestly similar to genes of other pathogens that encode subunits of specialized type IV secretory systems that directly deliver bacterial virulence factors to the surface and possibly into host cells. Control of bacterial virulence often is mediated by changes at the DNA sequence level that affect gene regulation or expression (7). Three Cag PAI now have been sequenced from the complete genomes of strains 26695 and J99 and the sequenced cosmid 36 from strain NCTC11638. All three contain an unusual ORF (annotated Cag7 or HP527 in strain 26695), which is significantly variable among HP pathogenic strains, but no mechanisms for this variation have been proposed and no features of the Cag7 sequence have been noted to account for the origin of this variation.

We present here a rigorous statistical analysis of the Cag7 protein (1,927 aa) from strain 26695. Of particular interest, we underscore several sequence features of this protein, including distinctive repeat patterns, a remarkable cysteine residue distribution, a statistically significantly high multiplet count (defined below), a pronounced charge residue cluster (8,9), extremes of lysine and glutamate amino acid usage, and identification of hydrophobic potential transmembrane segments. Expansion or contraction of the repeats could account for the size variations seen in the ORF of Cag7.

RESULTS

Unusual Sequence Features of Cag7. The SAPS (Statistical Analysis of Protein Sequences) program (8) was applied to all the putative proteins encoded from the CagA region of strain 26695. This analysis reveals several unusual sequence features especially for the Cag7 protein, which was found to contain two impressive regions composed of contiguous repeated amino acid sequences.

Repeat I. Repeat I (Fig. 1), covering amino acid positions 9-325 inclusive, in the pattern baaaba** (130 aa), has baaaba (130 aa) aligned with baaaba (130 aa), showing only three mismatches and baaaba (57 aa), a truncated copy of baaaba*, which matches perfectly over their common 57 aa and, more impressively, in perfect DNA agreement. Remarkably the baaaba and baaaba** differ at only three DNA positions, which all occur in codon site 1. There are no synonymous (silent site) substitutions. The almost perfect DNA identities comparing baaaba to baaaba* or baaaba** strongly suggest a recent origin to these repeats.

Repeat II. Repeat II (Figs. 2 and 3) consists of 74 contiguous segments composed from selections among six different consensus sequences, which we call α, β, λ, μ, δ, ε, stretching over amino acid positions 477-1383. The underline signifies perfect conservation of the amino acids at that position among the ensemble of sequences of α, of β, etc.

Some α have one or two appended aa, generally E, EE, or QE: β = L K D L E K D L Q K K V L (14 aa length).

ε = C E K L L T P E A (K/R) K L L E (14 aa length).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked ‘advertisement’ in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Abbreviation: HP, Helicobacter pylori.

†To whom reprint requests should be addressed. e-mail: fd.zgg@forsythe.stanford.edu.

7011
The explicit order of the subsequences of repeat II is displayed next:

\[
\begin{align*}
\lambda &= (\tau_1 \cdot \epsilon \cdot \lambda) \cdot (\tau_2 \cdot \epsilon \cdot \lambda) \\
\alpha &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\beta \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\beta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\delta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\mu &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\end{align*}
\]

The explicit order of the subsequences of repeat II is displayed next:

\[
\begin{align*}
\lambda &= (\tau_1 \cdot \epsilon \cdot \lambda) \cdot (\tau_2 \cdot \epsilon \cdot \lambda) \\
\alpha &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\beta \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\beta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\delta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\mu &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\end{align*}
\]

The explicit order of the subsequences of repeat II is displayed next:

\[
\begin{align*}
\lambda &= (\tau_1 \cdot \epsilon \cdot \lambda) \cdot (\tau_2 \cdot \epsilon \cdot \lambda) \\
\alpha &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\beta \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\beta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\delta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\mu &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\end{align*}
\]

The explicit order of the subsequences of repeat II is displayed next:

\[
\begin{align*}
\lambda &= (\tau_1 \cdot \epsilon \cdot \lambda) \cdot (\tau_2 \cdot \epsilon \cdot \lambda) \\
\alpha &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\beta \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\beta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\delta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\mu &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\end{align*}
\]

The explicit order of the subsequences of repeat II is displayed next:

\[
\begin{align*}
\lambda &= (\tau_1 \cdot \epsilon \cdot \lambda) \cdot (\tau_2 \cdot \epsilon \cdot \lambda) \\
\alpha &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\beta \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\beta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\delta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\mu &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\end{align*}
\]

The explicit order of the subsequences of repeat II is displayed next:

\[
\begin{align*}
\lambda &= (\tau_1 \cdot \epsilon \cdot \lambda) \cdot (\tau_2 \cdot \epsilon \cdot \lambda) \\
\alpha &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\beta \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\beta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\delta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\mu &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\end{align*}
\]

The explicit order of the subsequences of repeat II is displayed next:

\[
\begin{align*}
\lambda &= (\tau_1 \cdot \epsilon \cdot \lambda) \cdot (\tau_2 \cdot \epsilon \cdot \lambda) \\
\alpha &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\beta \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\beta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\delta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\mu &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\end{align*}
\]

The explicit order of the subsequences of repeat II is displayed next:

\[
\begin{align*}
\lambda &= (\tau_1 \cdot \epsilon \cdot \lambda) \cdot (\tau_2 \cdot \epsilon \cdot \lambda) \\
\alpha &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\beta \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\beta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\delta &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\mu &= (\alpha \cdot \epsilon \cdot \alpha) \cdot (\alpha \cdot \delta \cdot \alpha) \\
\end{align*}
\]
dramatic DNA identity within the repeat structures putatively generated through recombination or replication strand slippage allows opportunities for changes in the repeat length. Different lengths may produce alternative protein conformations or serve to switch the protein's expression on and off, thus affording the HP bacterium a means to confound host immune system surveillance.

**Significantly High Multiplet Count in the Cag7 Sequence.** A measure of the homopeptide density of a protein sequence is provided by the multiplet count, i.e., the number of distinct homooligopeptide runs of two or more residues. Specifically, multiplet counts refer to the number of homopeptides in protein sequences counting all homodipeptides $XX(=X_2)$, homotripeptides $YYY(=Y_3)$, homotetrapeptides $ZZZ(=Z_4)$, etc., where $X, Y, Z$ denotes any amino acid. A statistical assessment of the counts and locations of these multiplets compares the observed multiplet set to the multiplet distribution in a random (shuffled) reconstruction of the protein sequence. A significance test of high multiplet counts would take account of the amino acid composition of the protein sequence under study and is described in Karlin et al. (12). The scarce occurrence of proteins in possession of an abundance of amino acid multiplets stands out in *Escherichia coli* and in most prokaryotes. The percentage of human proteins with significantly high multiplet counts is about 1.5% with similar percentages observed in mouse and yeast. A greater number of proteins with significantly many multiplets is detected in *Drosophila* (about 10%), usually associated with developmental regulatory genes (13). Strikingly, Cag7 in *HP* (strain 26695) and *HP* (strain J99) is the only protein sequence of *HP* that carries a significantly high multiplet count. In the case at hand, the bulk of the multiplets concentrate in the two repeat regions (see Fig. 5).

**Fig. 2.** Repeat II in the Cag7 protein extends continuously from amino acids 477-1383. The sequences of $\alpha, \beta, \delta, \epsilon, \lambda$, and $\mu$ are aligned, and the consensus sequences are displayed at the top. Residues that appear the same number of times at one position both are displayed in the consensus sequence indicated by a colon. Note that the sequences of $\epsilon, \lambda$, and $\mu$ start with a cysteine. Lowercase letters represent nonaligned residues. The $*$ underneath the K locates the terminal point of ORF14 in cosmid 36 and the @ underneath the m locates the start point of ORF13 in cosmid 36. The conservation index (defined below) among the sequences of $\alpha$ is 0.82; among the sequences of $\beta$, 0.79; among the sequences of $\delta$, 0.81; among the sequences of $\epsilon$, 0.60; among the sequences of $\lambda$, 0.68; and among the sequences of $\mu$, 0.78. The conservation index (10) provides a means to quantitate similarity among aligned sequences. A similarity score between a pair of amino acids is determined according to a similarity substitution matrix, say BLOSUM 62 (11). Normalized scores for an amino acid pair ($a$ and $b$) are calculated by the formula

$$S(a, b) = \frac{S(a, b)}{\sqrt{S(a, a) \times S(b, b)}},$$

where $S(a, b), S(a, a), S(b, b)$ are similarity values given by the BLOSUM 62 matrix. For each position (column) of these sequences, the conservation index is calculated by taking the average normalized score from all residue pairs at that position.
and other long amino acid runs also are correlated with high multiplet counts (13).

**Potential Transmembrane Segments in Cag7.** The Cag7 distinguishes two statistically significant long predominantly hydrophobic uncharged runs, traversing coordinates 343-370 proximal downstream to repeat I and 1836-1870 near the carboxyl end.

**DISCUSSION**

In the CagA region of the HP genome strain 26695, the Cag7 (HP527) gene (1927 aa) is replete with unusual sequence features. This gene has been noted by other researchers because of its marginal sequence similarity to the virB10 family (percent similarity about 30%) of type IV secretory genes and its necessity for HP’s induction of IL-8 secretion in gastric epithelial cells and the strain-to-strain variation in size. We have found that this variation occurs within two repeat regions in the Cag7 protein. The amino end of Cag7 is distinguished by the long tandem repeat **(14 aa)** – **(15 aa)** (total length 317 aa). The middle part of Cag7, repeat II, covering amino acid positions 477-1383 consists of 74 subsequences selected from six consensus sequences α (generally 14 aa), β (14 aa), λ (13 aa), μ (13 aa), δ (10 aa), and ε (5 aa) (see previous text or Fig. 2 for the explicit sequences). The DNA identity among different representations of the consensus sequences is very high. Other strains of HP maintain polymorphic versions of repeats I and II associated strictly with variation in repeat subunits. The published sequences of Cag7 (HP527) with homologues from two other strains suggest that the strain-to-strain variation could be explained by recombination within the gene mediated by repeat subunits or can, in part, result from replication strand slippage. The three strains also may reflect on Cag7 variation among separate population sources. Thus, strain 26695 comes from a United Kingdom individual, strain J99 comes from a United States individual, and cosmid 36 (strain NCTC11638) was sequenced from an Australian individual. The extent over time of Cag7 variation from a single strain has not been adequately ascertained.

---

**Fig. 3.** Aligned DNA sequences corresponding to the amino acid sequences of group **1** are displayed. 1 indicates a position strictly conserved among these sequences, and • indicates a position with average conservation index (CI, defined in Fig. 2 legend) exceeding 0.6. The scores for nucleotide comparisons are as follows: identity has average conservation index (CI, defined in Fig. 2 legend) exceeding 0.8, and 87% of the columns show a CI value above 0.6 emphasizing a high level of conservation.

Where a preponderance of lysine and glutamate doublets KK and EE (or EEE) appear (see Discussion for possible implications).

It is interesting that significantly high multiplet counts are also present in the genes containing the PGRS repeats of *Mycobacterium tuberculosis* contemplated also as pathogenicity islands (14, 15). The human neurological disease genes associated with long trinucleotide CAG (glutamine) iterations

---

**Fig. 4.** The Cag7 protein sequence is aligned with the translated protein in cosmid 36 combining ORF14, ORF13, and the intervening part requiring a single base (+1) frame shift after amino acid 682 (counting from the N terminus of ORF14). When introducing the frame shift the DNA sequence encodes a protein that, apart from two gaps, aligns more than 90% with Cag7. The first gap corresponds to amino acids 9–138 of Cag7, consisting of unit **a**, the second gap corresponds to amino acids 1114–1182, consisting of two consecutive repeat triplet groups, namely (α-ε-λ)-(β-δ-μ) (see text). The Cag7 ortholog jhp0476 in strain J99 is displayed below Cag7. The jhp0476 sequence is missing a segment equivalent to **b** of Cag7, and the unit corresponding to **a** is 16 aa longer. The same two consecutive triplet groups missing from cosmid 36 also are missing from jhp0476, whereas the repeat **II** in jhp0476 extends longer by 78 aa augmented by the two successive triplet groups (δ-μ-α)-(δ-μ-α). The Cag7 and jhp0476 can be divided into three parts corresponding to ORF14, ORF13, and the intervening piece in cosmid 36. The *loc* locate two significantly long uncharged (potential transmembrane) segments, the first traversing amino acid positions 343–370 downstream proximal to repeat I and the second segment of positions 1836-1870 is near the C terminus. + corresponds to a concentrated charge region. The arrows indicate the extent and orientation of the ORFs.

---

**Fig. 4.** The Cag7 protein sequence is aligned with the translated protein in cosmid 36 combining ORF14, ORF13, and the intervening part requiring a single base (+1) frame shift after amino acid 682 (counting from the N terminus of ORF14). When introducing the frame shift the DNA sequence encodes a protein that, apart from two gaps, aligns more than 90% with Cag7. The first gap corresponds to amino acids 9–138 of Cag7, consisting of unit **a**, the second gap corresponds to amino acids 1114–1182, consisting of two consecutive repeat triplet groups, namely (α-ε-λ)-(β-δ-μ) (see text). The Cag7 ortholog jhp0476 in strain J99 is displayed below Cag7. The jhp0476 sequence is missing a segment equivalent to **b** of Cag7, and the unit corresponding to **a** is 16 aa longer. The same two consecutive triplet groups missing from cosmid 36 also are missing from jhp0476, whereas the repeat **II** in jhp0476 extends longer by 78 aa augmented by the two successive triplet groups (δ-μ-α)-(δ-μ-α). The Cag7 and jhp0476 can be divided into three parts corresponding to ORF14, ORF13, and the intervening piece in cosmid 36. The *loc* locate two significantly long uncharged (potential transmembrane) segments, the first traversing amino acid positions 343–370 downstream proximal to repeat I and the second segment of positions 1836-1870 is near the C terminus. + corresponds to a concentrated charge region. The arrows indicate the extent and orientation of the ORFs.
Apart from the striking repeat patterns, Cag7 is extraordinary in other sequence attributes, including high multiplet count, significant charge clusters (not shown), several extreme amino acid usages, and potential of transmembrane segments. Issues and potential experiments to be considered are:

(i) What part of Cag7 is necessary for virulence? The frame shift in the intervening region between ORF14 and ORF13 of cosmid 36 converts the ORFs into an almost complete homologue of Cag7. The polymorphism resulting from variations of the repeat numbers and lengths may enhance or curtail interactions with the host and serve as a means of shielding the bacterium from an immune system attack. The changes in the repeat numbers may affect how the Cag7 protein surface looks to the immune system and thereby may avoid recognition by antibodies made during previous infections. The almost perfect DNA identities within repeats I and II strongly argues for rather recent changes in repeat numbers. These repeat patterns may indicate a facility of HP for allowing rapid changes prompted by some host immune attacks.

(ii) The regular distribution of cysteine residues in repeat II provides a possible scaffolding involving disulfide bridges cross-linking secondary structures and/or domains of the protein structure. It would be informative to synthesize a triplet unit of the repeat II, say α-ε-λ and/or β-δ-μ, and evaluate its secondary structure in an aqueous medium.

(iii) The role of repeats in protein sequences is generally unclear. They may be benign, arising through replication strand slippage or recombination. They may provide flexibility and variation to protein conformation and function in response to environmental stress or host surveillance. They may contribute a regulatory role in gene transcription, translation, and expression. They may facilitate binding capacities in protein–protein and protein–DNA interactions. The pattern of repeat II, coupled to the regular cysteine distribution and an abundance of KK and EE diresidues, may contribute to several of these activities.

(iv) The high multiplet count of Cag7 is dominated by lysine and glutamate doublets, that are especially rife in repeat region II. These conceivably provide opportunities for multiple salt bridges, facilitating conformational stability and contributing to protein–protein interactions and quaternary structure formations (16).

(v) The main question about Cag7 is what is its function? Its similarity to secretory genes (virB10 family) of other species and its necessity for IL-8 secretion would support the idea that it is a component of a secretory apparatus that delivers a product or products that induces the IL-8 response. The virB10 gene has been proposed to play a regulatory function of the type IV secretory system of Agrobacterium tumefaciens (17). It is noteworthy that the portion containing the repeat regions is absent from members of the virB10 family in all other sequenced species. These observations suggest that the repeat regions and their contractions and expansions play key regul-
latory roles in the function of the putative HP secretory apparatus.

We are happy to acknowledge valuable discussions with Drs. B. E. Blaisdell, L. Brocchieri, A. M. Campbell, J. Mrázek, and N. Salama. This work was supported by National Institutes of Health Grants 5R01GM10452-34 (G.L. and S.K.), 5R01HG00335-11 (G.L. and S.K.), and AI38459 (T.M. and S.F.), National Science Foundation Grant DMS9704552 (G.L. and S.K.), and the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation, DRG-1456 (T.M.).