Published online December 2, 2004

*Nucleic Acids Research, 2004, Vol. 32, No. 21* ***6347–6357***
*doi:10.1093/nar/gkh935*

# Identification of genes with fast-evolving regions in microbial genomes

**Yu Zheng[1], Richard J. Roberts[3] and Simon Kasif[1,2,*]**

[1]Bioinformatics Graduate Program and [2]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA and [3]New England Biolabs, Beverly, MA 01915, USA

Received August 30, 2004; Revised and Accepted October 25, 2004

## ABSTRACT

**Complete sequences of multiple strains of the same microbial species provide an invaluable source for studying the evolutionary dynamics between orthologous genes over a relatively short time scale. Usually the intensity of the selection pressure is inferred from a comparison between the nonsynonymous substitution rate and the synonymous substitution rate. In this paper, we propose an alternative method for detecting genes with one or more fast-evolving regions from pairwise comparisons of orthologous genes. Our method looks for regions with overrepresented nonsynonymous mutations along the alignment, and requires a higher nonsynonymous evolution rate in those regions than the neutral evolution rate. It identifies gene targets under intensive selection pressure that are not detected from the conventional rate comparison analysis. For those identified genes with known annotations, most of them have a clear role in processes such as bacterial defense and host–pathogen interactions. Gene sets reported from our method provide a measure of the phenotypic divergence between two closely related genomes.**

## INTRODUCTION

The increasing availability of microbial genomes has stimulated comparative studies on genome sequence variation and its evolutionary implications (1,2). Each individual gene and each individual amino acid in its protein product are selected differently and as a result, various levels of sequence variation are observed between orthologs, even from phylogenetically close strains of the same species. Comparative sequence analysis has suggested that variation in evolutionarily close genes is usually not uniformly distributed along the sequences but is mostly found in specific regions (3). What is even more interesting is that the heterogeneous pattern of evolution rates along the sequence is often correlated with differential selection pressure shaping the slow-evolving and fast-evolving parts of the protein for functional purposes (4–6). In a number of examples, function is crucially associated with a high degree of regional sequence diversity. For instance, a higher evolution rate has been observed at the extracellular antigen recognition sites (ARS) of the MHC class I α chain gene than in the other domains of the same gene (3). It would be very useful to identify all the possible genes with regional fast evolution, and to analyze the functional implication of those fast-evolving regions.

Conventionally, the intensity of selection pressure is estimated by the ratio between the nonsynonymous substitution rate and the synonymous substitution rate, which are calculated from a set of closely related sequences (7). A coding sequence is considered to be positively selected if the nonsynonymous substitution rate is significantly higher than the synonymous substitution rate. It has been known that many genes involved in host defense display this property. However, a large-scale search in the database for positively selected genes based on this criterion found very few genes (8). It is known that the statistical test involving the evolution rates is usually too conservative (9), and more often, there are no sufficient instances of closely related sequences to get statistically accurate estimations.

An alternative approach pioneered by Tang and Lewontin (TL method hereafter) looks for the spatial clustering of variable sites along an alignment (4). The clustering of the mutations highlights the regions where divergence starts to occur. In our previous work, we have proposed the concept of segmentally variable genes (SVGs) (10). However, SVGs are inferred from the comparison across species, and there is no inherent assumption on the evolution rate of the variable regions in SVGs. In this paper, we propose a two-step procedure to detect genes with fast-evolving regions from comparison of closely related bacterial strains: first, regions with overrepresented nonsynonymous mutations are identified in an alignment, based on a modified TL method; second, the nonsynonymous evolution rate in the regions detected in the first step is required to be higher than the estimated overall neutral evolution rate. The new procedure is an alternative way to detect genes under intensive selection pressure in microbial genomes.

## METHODS

### Nucleotide alignment of orthologous gene pairs

Whole-genome sequences of different strains of the same microbial species were retrieved from GenBank. Putative

*To whom correspondence should be addressed. Tel: +1 617 358 1845; Fax: +1 617 353 6766; Email: kasif@bu.edu
Present address:
Yu Zheng, New England Biolabs, 32 Tozer Road, Beverly, MA 01915, USA

orthologous gene pairs were identified as reciprocal best BLAST (11) hits using translated protein sequences. The amino acid sequences of each orthologous gene pair were aligned using the Needleman–Wunsch global alignment algorithm (12). Orthologous gene pairs with a low percent identity are highly divergent so that mutation saturation may significantly influence the detection of the clustering of the variable sites and the evolution rate estimation (4). For simplicity, we use 50% as a cutoff, and all pairs with <50% percent identity were excluded from further analysis. Among 1347 orthologous gene pairs between *Helicobacter pylori* strain 26695 and strain J99, there are 27 gene pairs with <50% percent identity in amino acid sequences (see Table S.1 in the Supplementary Material). The amino acid alignments from the global alignment were then converted to nucleotide alignments with the translation frame retained. Gaps in the amino acid alignments were excluded in the nucleotide alignments.

### Detection of regions with significantly higher variability

A variant of the TL method (4) is used to detect regions with significantly higher variability within an alignment. At each codon position in the alignment, the number of nonsynonymous mutations is estimated. For example, AAC (Asn) → AAA (Lys) indicates one nonsynonymous mutation. In cases where multiple parsimonious pathways exist, we assume equal weighting among these pathways (7). For example, the nonsynonymous mutation number between TTT and GTA is 1.5 (two pathways with equal probability: TTT (Phe) → GTT (Val) → GTA (Val), 1 nonsynonymous mutation; TTT (Phe) → TTA (Leu) → GTA (Val), 2 nonsynonymous mutations). The alignment is then converted to a numerical string where each non-zero entry gives the number of nonsynonymous mutations that might have happened at that site. Compared with the original TL method (4), in which the mismatch in the alignment is converted to 1 and the match to 0, this modified approach is expected to give a more sensitive measure in sequence divergence. Naturally, other measures on the distance between amino acids can be used, for instance, a distance metric based on their physicochemical properties.

The TL method uses the ECDF statistics (13) to calculate the difference between the observed cumulative distribution ($F_n$) of events and the expected cumulative density function ($F$). Let us use the following notation: the total number of codons in the alignment is $N$; the positions of all non-zero elements in the alignment string are $X_1, X_2, \ldots, X_v$; the numerical values of each non-zero elements are $x_1, x_2, \ldots, x_v$ ($x_1$ is the nonsynonymous substitution number in position $X_1$ and so on); the total number of nonsynonymous mutation events in the alignment is $n$ ($n = \sum_{i=1}^{v} x_i$). We then calculate $F_n$ and $F$ at position $X_i$ as follows:

$$F_n(X_i) = \frac{\sum_{j=1}^{i} x_j}{n},$$

$$F(X_i) = \frac{X_i}{N}.$$

The TL method defines a function $G$ to record the difference between the above two values at each non-zero position $X_i$:

$$G(X_i) = F_n(X_i) - F(X_i), \quad i = 1, \ldots, v.$$

In a region bounded by two non-zero elements, let

$$\Delta G_{i,j} = G(X_j) - G(X_i), \quad 1 \le i \le j \le v.$$

Regions with significantly overrepresented nonsynonymous mutations will have a higher positive $\Delta G$ value. The test statistic $T$ is $\max(\Delta G)$, where the maximum is taken over all intervals in which the function G increases monotonically.

The critical value of $T$ is estimated by bootstrapping, where we randomly distribute $n$ nonsynonymous mutation events over $N$ sites with replacement. This process is repeated 10 000 times to get a distribution of $T$ under the null hypothesis (Figure S.2 in the Supplementary Material). The critical value of $T$ is then derived from this distribution by setting the type I error rate at 1%. Any orthologous gene pairs with a calculated $T$ value bigger than the critical $T$ value are considered to be candidates for genes with fast-evolving regions. In the next step, they will be filtered by another criterion.

### Estimates of the overall neutral evolutionary rates

When the selection pressure is absent or very weak in a particular region of the protein, it may accumulate neutral substitutions freely so that it may appear that this region has a significantly higher nonsynonymous substitution rate than other regions. Since we are interested in fast-evolving regions that are selected for function, we should eliminate those cases in the previous list from the TL method. For this purpose, we further require that the nonsynonymous mutation rate in the regions identified by the TL method should be higher than the overall neutral evolution rate in that gene. The neutral rate for each orthologous gene pair is estimated separately from all the 4-fold degenerate sites (4D sites) in the original alignment (14). The 4D site encodes the same amino acid no matter what nucleotide is in the third codon position, e.g. CGx → Glycine (x can be A, T, G, C). It is assumed that there is minimal selection pressure operating on these sites (14). The mean ($m_{4d}$) and the standard deviation ($sd_{4d}$) of the neutral rate are calculated by the modified Nei and Gojobori method (15). The transition/transversion ratio (R) is estimated by $R = P/Q$, where $P$ and $Q$ are the observed frequencies of transition-type substitutions and transversion-type substitutions in the alignment (7). The $Z$-score is then calculated as $Z = (p - m_{4d})/sd_{4d}$, where $p$ is the nonsynonymous substitution rate in the regions identified by the first step. The $Z$-score cutoff is set at 1 in this paper. Orthologous gene pairs that pass this test are reported as genes with fast-evolving regions.

### Test of overrepresentation of radical substitutions in fast-evolving regions

We carried out an additional $\chi^2$ test to test overrepresentation of the radical substitutions in the fast-evolving regions. The 20 amino acids are classified into three groups according to their charges: the positively charged group (Arg, His, Lys), the negatively charged group (Asp, Glu) and the neutral group (other amino acids). Amino acid substitutions within groups are considered as conservative while those between groups are

considered as radical (16,17). The $\chi^2$ statistic is calculated as follows:

$$\chi^2 = \frac{(N_R(F) - E[N_R(F)])^2}{E[N_R(F)]} + \frac{(N_R(NF) - E[N_R(NF)])^2}{E[N_R(NF)]},$$

where $N_R(F)$ and $N_R(NF)$ are the number of radical substitutions in the fast-evolving regions and non-fast-evolving regions, respectively. $E[N_R(F)]$ and $E[N_R(NF)]$ are the expected value of $N_R(F)$ and $N_R(NF)$, respectively. $E[N_R(F)]$ is calculated as follows:

$$E[N_R(F)] = (N_R(F) + N_R(NF))\frac{L(F)}{L(F) + L(NF)},$$

where $L(F)$ is the length of the fast-evolving region and $L(NF)$ is that for non-fast-evolving region. $E[N_R(NF)]$ is calculated in a similar way. Any orthologous gene pairs with a $P$-value <0.001 are significant and are reported.

### $K_A/K_S$ ratio test using maximum likelihood method

Positive selection is detected using the *codeml* program in the PAML package (18,19). A window of 90 nt (30 codons) is slid along the nucleotide alignment with a step size of 3 nt. Inside each window, the *codeml* program is run twice: first with the $K_A/K_S$ ratio (ω) fixed at 1 and second with this ratio as a free parameter. The difference in the maximum likelihood ratio is calculated as $LR = 2(\ln ML(\omega = 1) - \ln ML(\omega))$. $LR$ is then compared against the $\chi^2$ distribution with one degree of freedom to test whether the $K_A/K_S$ ratio is significantly >1. The significance level is set at 0.05. Genes that pass this test are reported as positively selected.

## RESULTS

A two-step procedure was developed to identify genes with fast-evolving regions by pairwise alignment of orthologous genes. The first step looks for regions with overrepresented nonsynonymous mutations in the alignment. In the second step, regional divergences probably caused by random drifting are filtered out by requiring a higher regional nonsynonymous evolution rate than the neutral evolution rate. Compared with the method of estimating the ratio between the nonsynonymous substitution rate and the synonymous substitution rate, our proposed method usually identifies more gene targets, some of which are not even suitable for the rate comparison analysis.

We applied our method to several sets of complete genomic sequences, including *Helicobacter pylori* strains 26695 (20) and J99 (21), *Neisseria meningitidis* strains MC58 (22) and Z2491 (23), *Mycobacterium tuberculosis* strains H37Rv (24) and CDC1551 (25), *Streptococcus pneumoniae* strains R6 (26) and TIGR4 (27), and three *Escherichia coli* strains [K12 (28), O157:H7 (29), CFT073 (30)]. Full lists of genes reported from our method can be found in the Supplementary Material (also see our website at http://geneva.bu.edu/fasta.html).

### Genes having fast-evolving regions at the genome level

In *H.pylori*, 1320 orthologous gene pairs are included in the analysis, among which 111 are detected as having one or more fast-evolving regions. We then classified them into different functional categories according to the shared keywords in their current Genbank annotation, as shown in Figure 1.

Among the largest classified groups are the hypothetical genes, the outer membrane protein family, the restriction–modification family and the transporter gene family. Hypothetical genes, most of which are unique to *H.pylori* species, form the largest group. This is in agreement with the observation that the conserved gene set is more stable than the unique genes (2,31), etc. Among the hypothetical genes, several of them can actually be annotated using simple similarity searches. For instance, HP0373 (Genbank id = 2313477) and HP0170 (Genbank id = 2313837) both encode proteins homologous to the existing outer membrane protein (OMP) family (see Figure S.1 in Supplementary Material). They represent two new members that should be added to this family. Members of the identified OMP family are believed to play important roles in adhesion and adaptive antigenic variation (20,32). Figure 2a shows a graphical representation

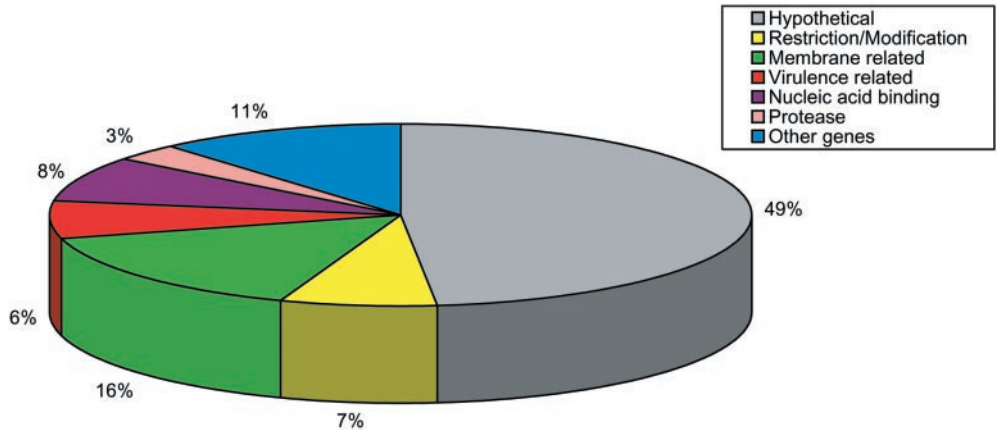


**Figure 1.** Functional classification of genes with fast-evolving regions in *H.pylori*. Classification is based on the shared keywords in the current annotations of the genes.

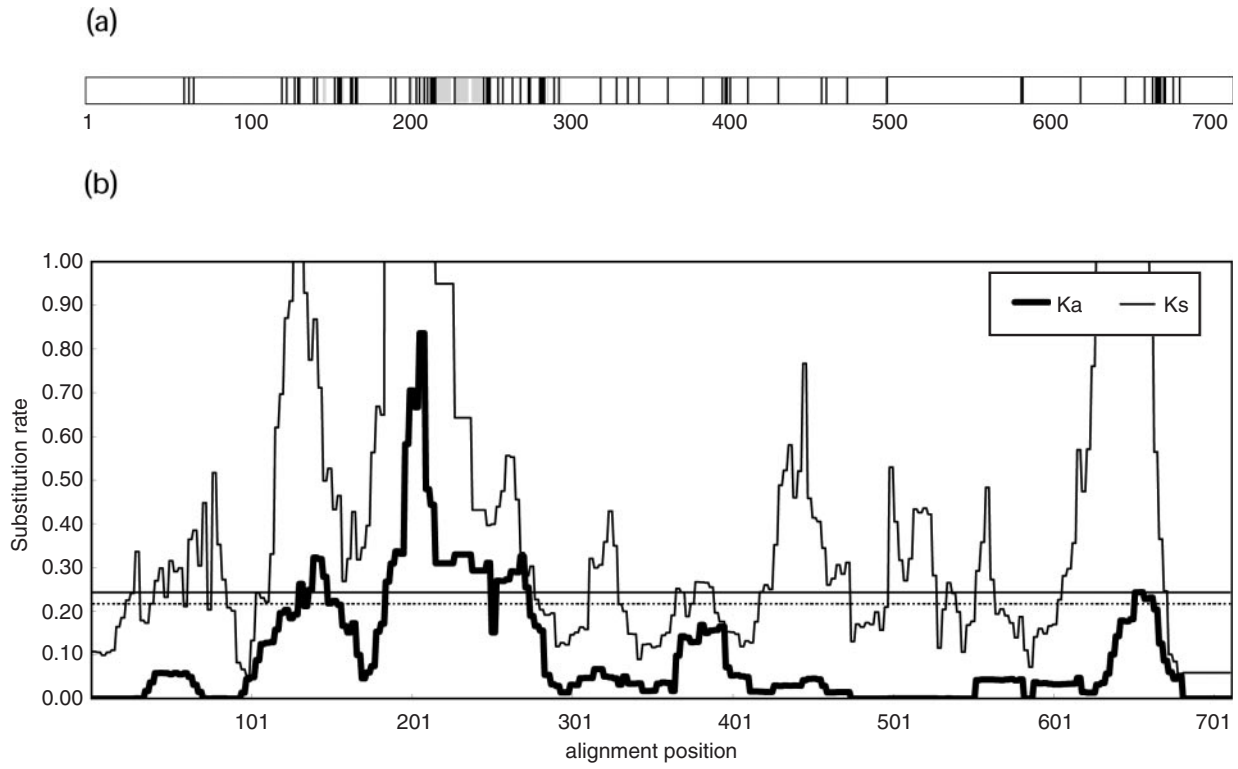


**Figure 2.** (**a**) Alignment of the HP0025 (OMP2) orthologous gene pair. The black lines represent mismatches and gray lines represent gaps. Black boxes over the alignment represent potential transmembrane helices reported by Tmpred (33). (**b**) $K_A$ (solid black line) and $K_S$ (gray line) along the OMP2 alignment. The dashed horizontal line is the mean of the neutral rate ($m_{4d}$). The solid horizontal line is the cutoff of the neutral rate constraint ($m_{4d} + sd_{4d}$).

of the pairwise alignment between orthologs for one of the OMP genes (omp2) (33). The variation sites clearly cluster after the second transmembrane segment in the N-terminus (Figure 2a). This region may correspond to the extracellular portion of the gene and the variation inside it could reflect the intense selection pressure operating on it, supporting its active role in host–pathogen interaction.

Figure 3 displays the relative position of the fast-evolving regions (normalized by alignment length) and the length distribution in *H.pylori* genes. As shown in Figure 3, most of the fast-evolving regions are 10–80 amino acids long, and they do not show any preference of location along the sequences. Although one might reason that the extreme N- or C-terminal segment is less functionally important and thus free to mutate, we do not observe this trend in our results. This is partly due to our requirement that the evolution rate in the fast-evolving region should be larger than the neutral evolution rate, which should exclude most of the functionally unimportant divergent regions.

Statistics of results from several other strain–strain pairs are shown in Figure 4a. In Figure 4b, we show the phylogenetic tree constructed from 16S ribosomal RNA sequences of each strain included (34). The percentage of genes with fast-evolving regions in the whole genome appears to be correlated with the phylogenetic distance between the two strains. For instance, the two *N.meningitidis* strains are more distant from each other than any other strain pairs, and they have the highest percentage of genes reported in the genome. The genome size, which corresponds to overall complexity of the organism, does not seem to correlate with this percentage judged from this figure. Interestingly, although the two *M.tuberculosis* strains are very close to each other (there are no substitutions in their 16S ribosomal RNA sequences), there are still a small percentage of genes (48 genes) reported by our method. Among these genes, there are ones from the glycine-rich PE and PPE family with possible roles in pathogenicity (24), a number of transporter genes and hypothetical genes (see our website).

### Comparison with a positively selected gene set using the $K_A/K_S$ ratio test

We applied the sliding window $K_A/K_S$ test for detecting positive selection between orthologous gene pairs of two strains. A coding sequence is considered as being positively selected if $K_A$, the nonsynonymous substitution per nonsynonymous site, is significantly larger than $K_S$, the synonymous substitution per synonymous site. A window (length = 30 codons) is slid along each pairwise nucleotide alignment of orthologous genes in two strains of *H.pylori*. Inside each window, the $K_A/K_S$ ratio is estimated by the maximum likelihood method (see Materials and Methods). Gene pairs that have windows with this ratio significantly >1 are reported. Among the 1320 orthologous pairs, there are 42 pairs reported to have positively selected regions. Only 14 pairs (33%) within this set are shared with the results from our method. Many factors contribute to this marked difference: first, the two approaches start from different hypotheses, one is based on the $K_A/K_S$ ratio and the other is based on the spatial clustering of variable sites. Second, the comparison here is limited. The sliding window procedure of

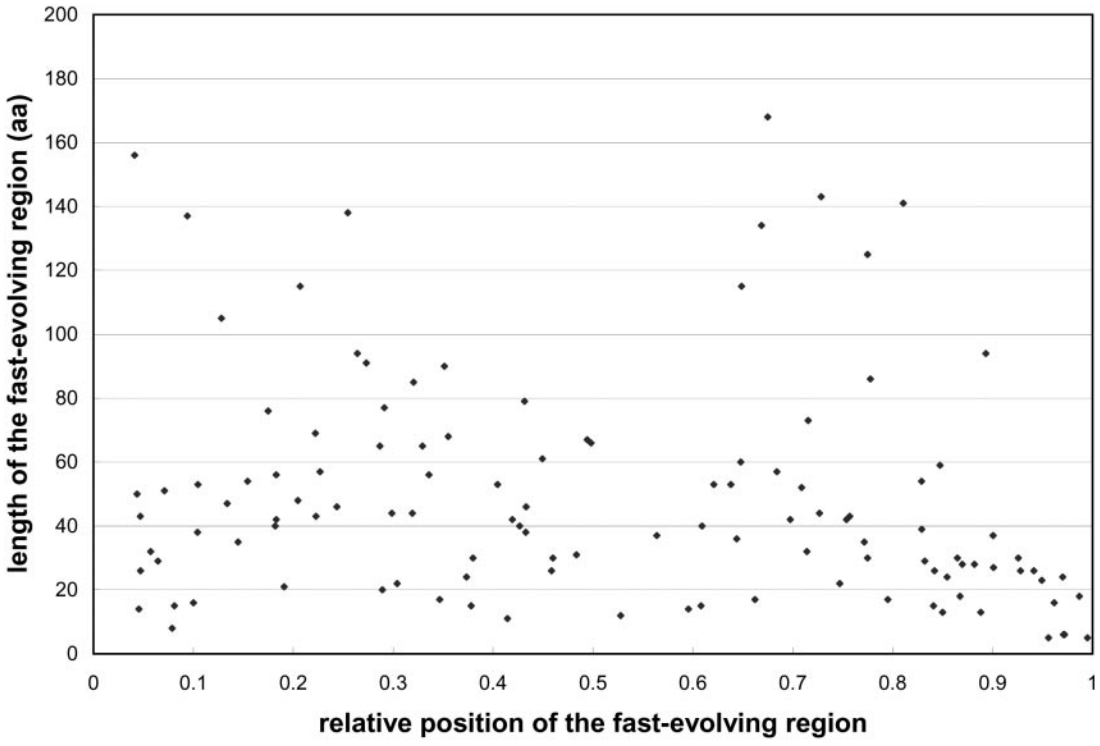


**Figure 3.** Relative positions versus their lengths of the fast-evolving regions in *H.pylori* genes. The *x*-axis is the relative position of the fast-evolving regions, calculated as the midpoint position of the region normalized by the alignment length; the *y*-axis the length of the fast-evolving regions in the alignment.

the $K_A/K_S$ ratio test used averages of the selection pressure over sites inside each window, as a result, the overall statistical power is low, especially when there are only limited number of sequences (two in this paper). When the number of sequenced strains, or a particular gene is large, it is possible to use the $K_A/K_S$ analyses to calculate the site-dependent selection pressure more accurately (35). Third, when the sequences become divergent, the estimation of $K_A/K_S$ becomes less reliable. In particular, as shown in Figure 2b, when the sequences inside windows are divergent, the estimation of $K_A$ and $K_S$ tends to give inaccurate numbers. This is especially the case for the $K_S$ estimation (Figure 2b). Last, we assume the neutral substitution rate does not vary inside a gene. For most genes this assumption may hold. However, when there is a regional clustering of mutational hot-spots, our method may still report them although the region may be under neutral selection. Nonetheless, in cases where the estimations of $K_A$ and $K_S$ become problematic, the proposed method in this paper provides an effective alternative.

## Genes with regional abundance in radical substitutions

Radical substitutions here are defined as those substitutions between groups of amino acids with different charges (e.g. Histidine (+) → Aspartate (−), see Materials and Methods section). Given the importance of charged residues in protein–protein, protein–small molecule and protein–nucleic acid interactions (17), fast-evolving regions with overabundant radical substitutions strongly suggests possible functional divergence in binding activity. Among the 111 genes in *H.pylori*, a majority of them (72 or 65%) are found to have significantly overrepresented radical substitutions in fast-evolving regions. In Table 1, we list those genes and sort them by the number of radical substitutions they have in the fast-evolving regions. In the following section, we will discuss several interesting case studies.

*HP1517: a putative restriction enzyme*. This gene has four fast-evolving regions, as shown in Figure 5. It shows significant similarity (*E*-value < 1E−10 by BLAST, percent identity ∼20% by Smith–Waterman alignment) to a type IIG restriction endonuclease Eco57I, which possesses both DNA cleavage activity and methylation activity (36). The N-terminal region of Eco57I was shown to have DNA cleavage activity by mutation of its $PDX_nEXK$ motif, which is the likely site for $Mg^{2+}$ binding and catalysis (36,37). The N-terminal region of HP1517 (region I in Figure 5) can be aligned with that of Eco57I and a variant of this motif is present ($GDX_nERK$). It is possible that the N-terminal region of HP1517 may have DNA cleavage activity. However, this motif is absent in the orthologous gene of HP1517 in the *H.pylori* strain J99 genome. The C-terminus of HP1517 (region IV in Figure 5) probably contains the target recognition domain for DNA methylation (38). Fast evolution in the C-terminus suggests possible changes in the methylation specificity. Since the cleavage and methylation activities are expected to have the same specificity for the same enzyme, this example may be explained by the co-evolution between the N-terminal cleavage domain and the C-terminal target recognition domain. It is also possible that one or both genes serve no function *in vivo* and are decaying (39). However, if that is the case, we would expect to see a more uniformly distributed pattern of nonsynonymous mutations, than in specific regions. The other two fast-evolving regions (region II and III

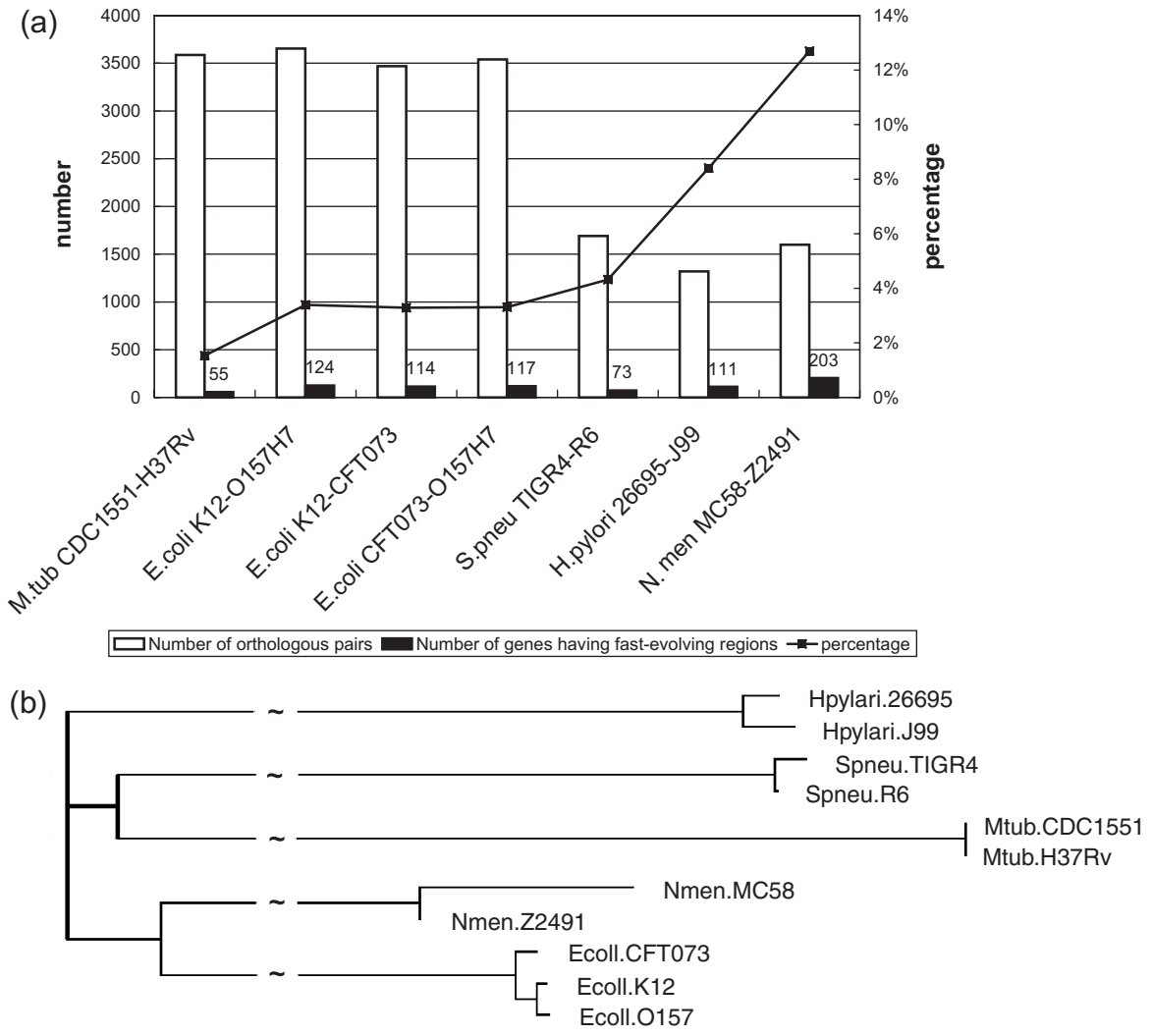


**Figure 4.** (a) Statistics of results reported from several strain–strain pairs. The white bars show the number of orthologous pairs between two strains; the black bars show the number of genes with fast-evolving regions; the line plot shows the percentage of genes with fast-evolving regions among all orthologous pairs (secondary *y*-axis). The strain–strain pairs are ordered by their phylogenetic distances (from left to right). (**b**) Phylogenetic tree constructed from 16 S RNA sequences. The Kimura 2 parameter model is used to compute the distance matrix (implemented in DNADIST). Neighbor-joining algorithm (implemented in NEIGHBOR) is used to construct the tree. DNADIST and NEIGHBOR are from the Phylip package (34).

in Figure 5) located between N-terminus and C-terminus cannot be assigned a clear function now (Region III appears to be an inserted segment which is absent in Eco57I.).

*HP0462: type I restriction enzyme specificity subunit (hsdS).* This gene encodes the subunit of a type I restriction endonuclease and determines the specificity of recognition. As shown in Figure S.1(a), the two orthologous genes (HP0462 and JHP0414) share their N-terminus, but are quite divergent in their C-terminus. Previous work (40) has shown that N- and C-termini of the hsdS gene in *E.coli* K12 each recognize separate half of the sequence motif (AAC$(N)_6$GTGC). Re-assortment of individual domains generates new specificities (40), e.g. by phage transduction (41). As a result, it is a reasonable assumption that these two enzymes share their N-terminal recognition motif, but have different C-terminal recognition motifs. Further searches against REBASE (42) were carried out using each domain looking for matches to an S subunit with known specificity. The shared N-terminal domain is slightly similar (*E*-value <1E−10, percent identity ∼26%) to S.HindI. The C-terminal domain of HP0462 is similar (*E*-value <1E−5, percent identity∼26%) to the N-terminal domain of S.EcoDI with a specificity of TTA$N_7$GTCY. The C-terminal domain of JHP0414 does not show significant similarity to enzymes with known specificity.

*ATP-dependent proteases: Lon and ClpA.* Energy-dependent proteases are indispensable for the elimination of defective or invading proteins in the cell (43). A key feature of these enzymes is the precise selection specificity for protein degradation. In the Lon protease, the fast-evolving region of about 30 amino acid resides in its substrate-recognition domain (SSD) (43), where the Lon of *H.pylori* 26695 has four repeated copies of 'KSEDQ' and the Lon of *H.pylori* J99 does not (Figure S.3 in Supplementary Material).

For ClpA, the fast-evolving region resides in the N-terminal domain. [see Figure S.1(b)]. The percent identity of amino

**Table 1.** List of genes in *H.pylori* with overrepresented radical substitutions in fast-evolving regions

| Gene name (26695) | GenBank ID | Gene name (J99) | GenBank ID | Fast-evolving regions number | Radical substitutions number | Current function in GenBank |
|---|---|---|---|---|---|---|
| HP1517 | 2314695 | jhp1409 | 4156028 | 1 | 65 | type IIS restriction enzyme R and M protein (ECO57IR) |
| HP0790 | 2313919 | jhp0726 | 4155286 | 2 | 62 | anti-codon nuclease masking agent (prrB) |
| HP0462 | 2313566 | jhp0414 | 4154940 | 1 | 47 | type I restriction enzyme S protein (hsdS) |
| HP1116 | 2314275 | jhp1044 | 4155633 | 1 | 29 | *H.pylori* predicted coding region HP1116 |
| HP1105 | 2314258 | jhp1032 | 4155618 | 1 | 28 | LPS biosynthesis protein |
| HP1471 | 2314648 | jhp1364 | 4155976 | 2 | 28 | type IIS restriction enzyme R protein (BCGIB) |
| HP0373 | 2313477 | jhp1008 | 4155594 | 2 | 26 | conserved hypothetical protein |
| HP1433 | 2314608 | jhp1326 | 4155937 | 1 | 22 | *H.pylori* predicted coding region HP1433 |
| HP0063 | 2313149 | jhp0058 | 4154566 | 1 | 22 | *H.pylori* predicted coding region HP0063 |
| HP0922 | 2314061 | jhp0856 | 4155426 | 1 | 20 | toxin-like outer membrane protein |
| HP0059 | 2313145 | jhp0052 | 4154552 | 1 | 19 | *H.pylori* predicted coding region HP0059 |
| HP0322 | 2313421 | jhp0305 | 4154836 | 2 | 16 | poly E-rich protein |
| HP0025 | 2313112 | jhp0021 | 4154526 | 2 | 16 | outer membrane protein (omp2) |
| HP1354 | 2314522 | jhp1272 | 4155873 | 1 | 15 | putative adenine specific DNA methyltransferase |
| HP0527 | 2313642 | jhp0476 | 4155007 | 1 | 15 | cag pathogenicity island protein (cag7) |
| HP1437 | 2314619 | jhp1330 | 4155940 | 1 | 15 | *H.pylori* predicted coding region HP1437 |
| HP0995 | 2314140 | jhp0941 | 4155519 | 1 | 14 | Integrase/recombinase (xerD) |
| HP0033 | 2313108 | jhp0029 | 4154534 | 1 | 14 | ATP-dependent C1p protease (clpA) |
| HP0289 | 2313383 | jhp0274 | 4154798 | 1 | 13 | toxin-like outer membrane protein |
| HP0896 | 2314032 | jhp1164 | 4155775 | 1 | 12 | outer membrane protein (omp19) |
| HP0508 | 2313624 | jhp0458 | 4154998 | 1 | 11 | *H.pylori* predicted coding region HP0508 |
| HP1396 | 2314570 | jhp1431 | 4156055 | 1 | 11 | *H.pylori* predicted coding region HP1396 |
| HP0850 | 2313984 | jhp0786 | 4155366 | 1 | 11 | type I restriction enzyme M protein (hsdM) |
| HP1142 | 2314298 | jhp1070 | 4155663 | 1 | 10 | *H.pylori* predicted coding region HP1142 |
| HP1379 | 2314549 | jhp1293 | 4155905 | 1 | 10 | ATP-dependent protease (lon) |
| HP1499 | 2314683 | jhp1392 | 4156012 | 1 | 9 | *H.pylori* predicted coding region HP1499 |
| HP0338 | 2313442 | jhp0320 | 4154857 | 1 | 9 | *H.pylori* predicted coding region HP0338 |
| HP1053 | 2314208 | jhp0372 | 4154904 | 1 | 9 | *H.pylori* predicted coding region HP1053 |
| HP1472 | 2314649 | jhp1365 | 4155977 | 1 | 8 | type IIS restriction enzyme M protein (mod) |
| HP0186 | 2313284 | jhp0174 | 4154702 | 1 | 8 | *H.pylori* predicted coding region HP0186 |
| HP0651 | 2313769 | jhp0596 | 4155139 | 1 | 8 | fucosyltransferase |
| HP1079 | 2314229 | jhp0346 | 4154870 | 1 | 8 | *H.pylori* predicted coding region HP1079 |
| HP0710 | 2313837 | jhp0649 | 4155201 | 1 | 8 | conserved hypothetical protein |
| HP0669 | 2313798 | jhp0613 | 4155166 | 1 | 8 | *H.pylori* predicted coding region HP0669 |
| HP0392 | 2313493 | jhp0989 | 4155570 | 1 | 7 | histidine kinase (cheA) |
| HP0583 | 2313700 | jhp0530 | 4155060 | 1 | 7 | *H.pylori* predicted coding region HP0583 |
| HP0747 | 2313873 | jhp0684 | 4155251 | 1 | 7 | conserved hypothetical protein |
| HP1363 | 2314529 | jhp1281 | 4155896 | 1 | 7 | conserved hypothetical integral membrane protein |
| HP1521 | 2314701 | jhp1410 | 4156029 | 1 | 6 | type III restriction enzyme R protein (res) |
| HP0203 | 2313295 | jhp0189 | 4154708 | 1 | 6 | *H.pylori* predicted coding region HP0203 |
| HP0417 | 2313521 | jhp0967 | 4155547 | 1 | 6 | methionyl-tRNA synthetase (metS) |
| HP0717 | 2313841 | jhp0655 | 4155207 | 1 | 6 | DNA polymerase III gamma and tau subunits (dnaX) |
| HP0805 | 2313935 | jhp0741 | 4155297 | 1 | 6 | lipooligosaccharide 5G8 epitope biosynthesis-associated protein (lex2B) |

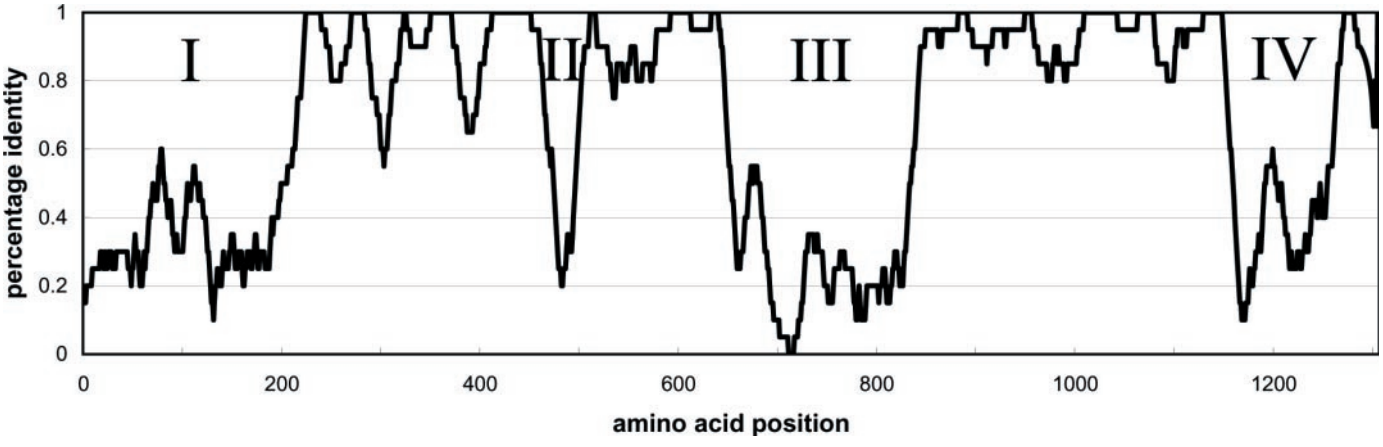


**Figure 5.** Percent identity plot using the sliding-window (width = 20 amino acids) along the amino acid alignment of the HP1517 orthologous pair. The fast-evolving regions are labeled as (I,II,III and IV) in the figure.

acids for the N-terminal domain (1–140 amino acids) is 69% while that for the rest of the protein is 95%. The N-terminal domain is implicated in substrate binding via recognition of a ClpA-specific adaptor protein, ClpS (44). ClpS is a relatively small protein of 90 amino acids. The percent identity for the ClpS orthologous pair between the two *H.pylori* strains is 69%, which is significantly lower than the average identity among other orthologous pairs (92 ± 11%). It is likely that fast evolution in the ClpS sequence then triggers fast evolution in the N-terminal domain of ClpA.

### Genes with fast-evolving regions from multi-strain comparison

When the genome sequences of more than two strains of the same species are available, it is of interest to look for orthologous genes with fast-evolving regions in all the strains or in most of the strains. One simple way is to intersect lists of identified genes from each pairwise strain–strain comparison. In doing this, we do not require that the fast-evolving regions should overlap with each other among all the orthologous genes. Intuitively, genes that appear in most of the pairwise comparison suggest that they are undergoing constant selection pressure present in most of the strains, while genes that only appear in a single pairwise comparison may suggest that they are probably the result of strain-specific evolution. Table 2 lists the genes that are detected in each pairwise strain–strain comparison among three *E.coli* strains. There are only 7 genes reported among the 3397 orthologous genes shared by all the three strains. In website, we also list the genes that appear only in two of the three pairwise comparisons. In the following, we choose a few very interesting examples to illustrate the functional implication of the regional fast evolution.

The first example is the *LamB* protein (Table 2), which functions as an outer membrane transporter of maltose and maltodextrins and at the same time serves as a receptor for several types of phages (45). Phage-resistant sites have been located mostly in four regions along the *LamB* protein sequence: area I (around position 152), area II (around 247), area III (around 382), area IV (around 401) (45). All these sites are located in the outer-membrane loops between the consecutive transmembrane segments. The crystal structure of *LamB* shows that only part of the sites that confer phage resistance are exposed to the outside surface (site 155, 164, 259, 386, 387, 394 and 401) (46). The fast-evolving region (380–400 amino acids) falls between area III and area IV, and is exposed to the surface of the molecule (Figure S.4 in Supplementary Material). In other regions, such as I, II, the multiple alignment shows 100% amino acid sequence identity among the three orthologs. Why is it that only the region between areas III and IV evolves rapidly, and diverges among strains, while other regions are completely conserved? Maltose transport assays on mutant *LamB* shows that only mutations at certain positions (154, 155, 259, 382, 401) do not impair transport activity (47). As a result, fast evolution between residue 380 and 400 is very likely to be the result of a balancing strategy between the arms race against hostile phages and the requirement to preserve transport ability. Fast evolution in this region is almost certain to be driven by the defense against phage attack.

The second example is the fhuA gene, which is detected between *E.coli* strain K12 and CFT073, O157 and CFT073, but not between K12 and O157 (Figure 6). The FhuA protein is the outer-membrane receptor for ferrichrome-iron, the related antibiotic albomycin and several bacteriophages (T1, T5, UC-1, Φ80) (48). The structure of FhuA is mainly composed of antiparallel transmembrane β strands. Loops connect adjacent strands and are either exposed to the periplasmic region or to the outside. The surface-exposed loops (labeled L3 through L10 in Figure 6) may interact directly with the substrates or serve as binding sites for bacteriophages (48). As shown in Figure 6, most variable sites are located inside these loops (L4, L5, L6, L7). At most of these variable sites, FhuA in strain CFT073 differs from its orthologs in both K12 and O157 while the latter two are kept constant. For instance, surface loop L4 has been proposed to form a gating loop to be opened upon ferrichrome binding (49). The same loop also confers phage sensitivity, e.g. binding of T5 triggers the opening of FhuA (49). Most of the sites have been altered in the L4 loop of strain CFT073 (Figure 6), so we predict that fast evolution in this loop strongly suggests that FhuA in CFT073 has a different phage sensitivity profile from its orthologs in K12 and O157. On the other hand, several other surface-exposed loops are almost completely conserved, which may suggest either important roles in normal functioning of FhuA or they do not provide phage binding sites. For instance, deletion of part of L3 loop results in the loss of ferrichrome-iron uptake (49).

**Table 2.** Genes in *E.coli* K12 with fast-evolving regions from all three pairwise comparisons

| GenBank id | Name | Annotation |
|---|---|---|
| 1788436 | yehI | putative regulator |
| 1789421 | b3042 | orf, hypothetical protein |
| 1790051 | rfaC | heptosyl transferase I; lipopolysaccharide core biosynthesis |
| 1789805 | yhgE | putative transport |
| 1790469 | lamB | phage lambda receptor;maltose high-affinity receptor |
| 1786335 | folK | 7,8-dihydro-6-hydroxymethylpterin pyrophosphokinase |
| 1788309 | flu | outer membrane fluffing protein, similar to adhesin |

## DISCUSSION

In this paper, we studied a unique set of genes that appear to possess one or more fast-evolving regions from pairwise comparative genomics. From our results, we show cases where the regional fast evolution is present under different biological scenarios: an arms race with invaders like hostile bacteriophages, antigenic variation to evade the host immune system, co-evolution in the molecular recognition, etc. Conventional gene content analysis focuses on gene sets that are present or absent between genomes and attribute unique physiological characteristics to these genes. Here, we find that functional divergence occurs even in a small portion of putative orthologous genes that are present in both genomes. Although such divergence is usually restricted to specific regions inside genes, they could contribute to the marked phenotypic differences between closely related organisms, together with the strain-specific genes.

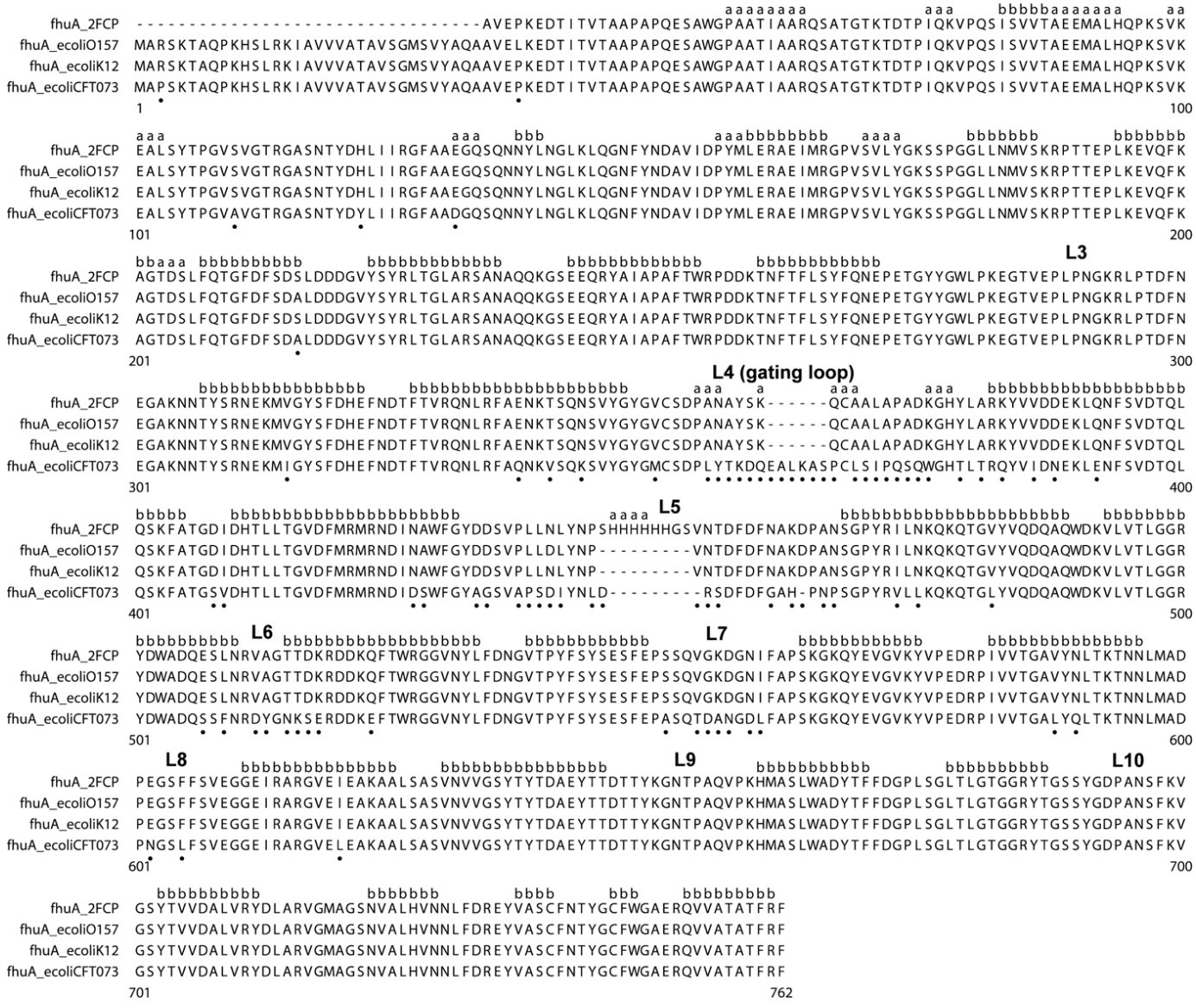


**Figure 6.** Multiple alignment of FhuA from three strains of *E.coli* and crystal data (fhuA_2FCP). Alpha helices and beta strands are shown on top of the alignment labeled as 'a' and 'b'. Sites where the residues from the three *E.coli* orthologs differ from each other are marked with a dot below the alignment. Large surface-exposed loops, L3–L10, are labeled over the alignment.

Choices of strain pairs with different phylogenetic distance influence the results. As we have shown, it seems natural that strains that are very close to each other usually have fewer genes identified as having fast-evolving regions, since there are few fixed mutations in most of the genes. The choice of the proper phylogenetic distance between the strains is expected to maximize the power of the method. Moreover, choices of strains from distinct biological niches give different results. For instance, the results from all three pairs of the three *E.coli* strains do not completely overlap. Genes that appear to diverge between strain A and B do not necessarily diverge in strain B and C. Only a limited number of genes that face common environmental challenges are subject to regional divergence among all strains. For instance, the LamB protein shows variation in phage-resistant sites, which may bring advantage against phage attacks.

Compared with the methods of estimating the $K_A/K_S$ ratio, the method we describe provides an alternative way of identifying genes under intensive selection pressure in microbial genomes. It is based on the observation of the clustering pattern of the variable sites along coding sequences, and is less dependent on the exact evolutionary model as is the case for most rate estimation methods. Between closely related strains, mutation saturation has not been reached for most genes so that confounding factors such as severe multiple substitutions at many sites may be ignored (50). Interesting questions, such as 'what genes are among the first group to diverge during strain evolution?', 'where does the divergence start to occur inside an individual gene?', 'what genes among different strains are always changing?', etc., can be approached from this angle. The results even help to find genes that are involved in host–pathogen interactions. For example, 4 of 8 genes that

have been used in vaccine development in *N.meningitidis* (51) are also detected by our method. Along with genes known to be related with pathogenesis in pathogenic bacterial genomes, many genes with hypothetical annotation are found to have fast-evolving regions. These genes should be given a priority for experimental determination of function and their role in the cell.

## SOFTWARE AVAILABILITY

A software package Faster is freely available at http://geneva.bu.edu/faster.html for all users.

## SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

## ACKNOWLEDGEMENTS

Y.Z. thanks Dr R. C. Lewontin for helpful discussion. This work was supported in part by NSF grant DBI-0239435.

## REFERENCES

1. Elena,S.F. and Lenski,R.E. (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Rev. Genet.*, **4**, 457–469.
2. Jordan,I.K., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Microevolutionary genomics of bacteria. *Theor. Popul. Biol.*, **61**, 435–447.
3. Hughes,A.L. and Nei,M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.
4. Tang,H. and Lewontin,R.C. (1999) Locating regions of differential variability in DNA and protein sequences. *Genetics*, **153**, 485–495.
5. Fares,M.A., Elena,S.F., Ortiz,J., Moya,A. and Barrio,E. (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J. Mol. Evol.*, **55**, 509–521.
6. Zhang,J. and Rosenberg,H.F. (2002) Diversifying selection of the tumor-growth promoter angiogenin in primate evolution. *Mol. Biol. Evol.*, **19**, 438–445.
7. Nei,M. and Kumar,S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, UK.
8. Endo,T., Ikeo,K. and Gojobori,T. (1996) Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.*, **13**, 685–690.
9. Yang,Z. and Bielawski,J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.*, **15**, 496–503.
10. Zheng,Y., Roberts,R.J. and Kasif,S. (2004) Segmentally variable genes: a new perspective on adaptation. *PLoS Biol.*, **2**, E81.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
13. Stephens,M.A. (1986) *Tests for the Uniform Distribution*. Marcel Dekker, London.
14. Hardison,R.C., Roskin,K.M., Yang,S., Diekhans,M., Kent,W.J., Weber,R., Elnitski,L., Li,J., O'Connor,M., Kolbe,D. *et al.* (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.*, **13**, 13–26.
15. Zhang,J., Rosenberg,H.F. and Nei,M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA*, **95**, 3708–3713.
16. Zhang,J. (2000) Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.*, **50**, 56–68.
17. Hughes,A.L., Ota,T. and Nei,M. (1990) Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol. Biol. Evol.*, **7**, 515–524.
18. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
19. Nekrutenko,A., Makova,K.D. and Li,W.H. (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.*, **12**, 198–202.
20. Tomb,J.F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A. *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.
21. Alm,R.A., Ling,L.S., Moir,D.T., King,B.L., Brown,E.D., Doig,P.C., Smith,D.R., Noonan,B., Guild,B.C., deJonge,B.L. *et al.* (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–180.
22. Tettelin,H., Saunders,N.J., Heidelberg,J., Jeffries,A.C., Nelson,K.E., Eisen,J.A., Ketchum,K.A., Hood,D.W., Peden,J.F., Dodson,R.J. *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, **287**, 1809–1815.
23. Parkhill,J., Achtman,M., James,K.D., Bentley,S.D., Churcher,C., Klee,S.R., Morelli,G., Basham,D., Brown,D., Chillingworth,T. *et al.* (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*, **404**, 502–506.
24. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E., III *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
25. Fleischmann,R.D., Alland,D., Eisen,J.A., Carpenter,L., White,O., Peterson,J., DeBoy,R., Dodson,R., Gwinn,M., Haft,D. *et al.* (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.*, **184**, 5479–5490.
26. Hoskins,J., Alborn,W.E., Jr., Arnold,J., Blaszczak,L.C., Burgett,S., DeHoff,B.S., Estrem,S.T., Fritz,L., Fu,D.J., Fuller,W. *et al.* (2001) Genome of the bacterium *Streptococcus pneumoniae* strain R6. *J. Bacteriol.*, **183**, 5709–5717.
27. Tettelin,H., Nelson,K.E., Paulsen,I.T., Eisen,J.A., Read,T.D., Peterson,S., Heidelberg,J., DeBoy,R.T., Haft,D.H., Dodson,R.J. *et al.* (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, **293**, 498–506.
28. Blattner,F.R., Plunkett,G., III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
29. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
30. Welch,R.A., Burland,V., Plunkett,G., III, Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
31. Makarova,K.S., Aravind,L., Galperin,M.Y., Grishin,N.V., Tatusov,R.L., Wolf,Y.I. and Koonin,E.V. (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.*, **9**, 608–628.
32. Aspholm-Hurtig,M., Dailide,G., Lahmann,M., Kalia,A., Ilver,D., Roche,N., Vikstrom,S., Sjostrom,R., Linden,S., Backstrom,A. *et al.* (2004) Functional adaptation of BabA, the *H. pylori* ABO blood group antigen binding adhesin. *Science*, **305**, 519–522.
33. K. Hofmann,W.S. (1993) TMbase—a database of membrane spanning proteins segments. *Biol. Chem.*, **374**, 166.
34. Felsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
35. Clark,A.G., Glanowski,S., Nielsen,R., Thomas,P.D., Kejariwal,A., Todd,M.A., Tanenbaum,D.M., Civello,D., Lu,F., Murphy,B. *et al.* (2003)

Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–1963.
36. Rimseliene,R. and Janulaitis,A. (2001) Mutational analysis of two putative catalytic motifs of the type IV restriction endonuclease Eco57I. *J. Biol. Chem.*, **276**, 10492–10497.
37. Pingoud,A. and Jeltsch,A. (1997) Recognition and cleavage of DNA by type-II restriction endonucleases. *Eur. J. Biochem.*, **246**, 1–22.
38. Malone,T., Blumenthal,R.M. and Cheng,X. (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.*, **253**, 618–632.
39. Lin,L.F., Posfai,J., Roberts,R.J. and Kong,H. (2001) Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. *Proc. Natl Acad. Sci. USA*, **98**, 2740–2745.
40. Fuller-Pace,F.V. and Murray,N.E. (1986) Two DNA recognition domains of the specificity polypeptides of a family of type I restriction enzymes. *Proc. Natl Acad. Sci. USA*, **83**, 9368–9372.
41. Bullas,L.R., Colson,C. and Van Pel,A. (1976) DNA restriction and modification systems in Salmonella. SQ, a new system derived by recombination between the SB system of *Salmonella typhimurium* and the SP system of *Salmonella potsdam*. *J. Gen. Microbiol.*, **95**, 166–172.
42. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2003) REBASE: restriction enzymes and methyltransferases. *Nucleic Acids Res.*, **31**, 418–420.
43. Smith,C.K., Baker,T.A. and Sauer,R.T. (1999) Lon and Clp family proteases and chaperones share homologous substrate-recognition domains. *Proc. Natl Acad. Sci. USA*, **96**, 6678–6682.
44. Zeth,K., Ravelli,R.B., Paal,K., Cusack,S., Bukau,B. and Dougan,D.A. (2002) Structural analysis of the adaptor protein ClpS in complex with the N-terminal domain of ClpA. *Nature Struct. Biol.*, **9**, 906–911.
45. Gehring,K., Charbit,A., Brissaud,E. and Hofnung,M. (1987) Bacteriophage lambda receptor site on the *Escherichia coli* K-12 LamB protein. *J. Bacteriol.*, **169**, 2103–2106.
46. Schirmer,T., Keller,T.A., Wang,Y.F. and Rosenbusch,J.P. (1995) Structural basis for sugar translocation through maltoporin channels at 3.1 A resolution. *Science*, **267**, 512–514.
47. Charbit,A., Gehring,K., Nikaido,H., Ferenci,T. and Hofnung,M. (1988) Maltose transport and starch binding in phage-resistant point mutants of maltoporin. Functional and topological implications. *J. Mol. Biol.*, **201**, 487–496.
48. Ferguson,A.D., Hofmann,E., Coulton,J.W., Diederichs,K. and Welte,W. (1998) Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide. *Science*, **282**, 2215–2220.
49. Killmann,H., Herrmann,C., Wolff,H. and Braun,V. (1998) Identification of a new site for ferrichrome transport by comparison of the FhuA proteins of *Escherichia coli*, *Salmonella paratyphi B*, *Salmonella typhimurium*, and *Pantoea agglomerans*. *J. Bacteriol.*, **180**, 3845–3852.
50. Lenski,R.E., Winkworth,C.L. and Riley,M.A. (2003) Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20 000 generations. *J. Mol. Evol*, **56**, 498–508.
51. Pizza,M., Scarlato,V., Masignani,V., Giuliani,M.M., Arico,B., Comanducci,M., Jennings,G.T., Baldi,L., Bartolini,E., Capecchi,B. *et al.* (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*, **287**, 1816–1820.