J Mol Evol (2001) 52:193–204
DOI: 10.1007/s002390010148

JOURNAL OF MOLECULAR EVOLUTION

The Interaction of Protein Structure, Selection, and Recombination on the Evolution of the Type-1 Fimbrial Major Subunit (*fimA*) from *Escherichia coli*

Andrew S. Peek,[1] Valeria Souza,[2] Luis E. Eguiarte,[2] Brandon S. Gaut[1]

[1] Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, Irvine, CA 92697-2525, USA
[2] Departmento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado Postal 70-275, México D.F. 04510, México

Received: 13 April 2000 / Accepted: 28 October 2000

Abstract. Fimbrial adhesins allow bacteria to interact with and attach to their environment. The bacteria possibly benefit from these interactions, but all external structures including adhesins also allow bacteria to be identified by other organisms. Thus adhesion molecules might be under multiple forms of selection including selection to constrain functional interactions or evolve novel epitopes to avoid recognition. We address these issues by studying genetic diversity in the *Escherichia coli* type-1 fimbrial major subunit, *fimA.* Overall, sequence diversity in *fimA* is high ($\pi = 0.07$) relative to that in other *E. coli* genes. High diversity is a function of positive diversifying selection, as detected by d_N/d_S ratios higher than 1.0, and amino acid residuces subject to diversifying selection are nonrandomly clustered on the exterior surface of the peptide. In addition, McDonald and Kreitman tests suggest that there has been historical but not current directional selection at *fimA* between *E. coli* and *Salmonella.* Finally, some regions of the *fimA* peptide appear to be under strong structural constraint within *E. coli,* particularly the interior regions of the molecule that is involved in subunit to subunit interaction. Recombination also plays a major role contributing to *E. coli fimA* allelic variation and estimates of recombination ($2N_ec$) and mutation ($2N_e\mu$) are about the same. Recombination may act to separate the diverse evolutionary forces in different regions of the *fimA* peptide.

Key words: Diversifying selection — Surface protein — *Escherichia coli* genetic variation

Correspondence to: Andrew S. Peek, Cytoclonal Pharmaceutics Inc., 9000 Harry Hines Boulevard, Dallas, TX 75235, USA; *e-mail:* apeek@cytoclonal.com

Introduction

Type-1 fimbria are adhesive protein fibers produced by *Escherichia coli* (Klemm 1984) and other members of the Enterobacteriaceae (Eisenstein 1987). Fimbria mediate bacterial attachment to host tissue and play a central role in bacterial colonization and pathogenesis (Elliott and Kaper 1997; Hultgren et al. 1996). In addition, fimbria probably act as targets for the immunological recognition of bacteria (Tewari et al. 1993). Thus, it is likely that adhesion molecules are under multiple forms of selection, including selection to maintain function and selection to evolve novel epitopes to avoid recognition.

The molecular structure of type-1 fimbrial subunits has been determined recently by X-ray crystallography (Choudhury et al. 1999). The structural model proposes that type-1 fimbria are complex hair-like structures composed of many protein subunits. The subunits form a hollow right-handed helix, with three subunits per turn. Along the hollow interior of the fimbrial structure, the subunits interact by donor strand complementation, where one subunit complements the incomplete immunoglobulin-like β-barrel of an adjacent subunit. This mo-

lecular structure ensures that fimbrial proteins have distinct external and internal domains.

Four proteins make up type-1 fimbria: *fimA, fimC, fimG,* and *fimH.* Of these, *fimA* is predominant, representing 98% of type-1 fimbria by weight. *E. coli* fimbriae have long been known to be polymorphic and the gene encoding *fimA* is also known to be highly polymorphic, possibly as a result of immunological selection (Boyd and Hartl 1998; Klemm et al. 1982; Marc and Dho-Moulin 1996; Salit et al. 1983). However, these studies of *fimA* relied on a small sample of sequences and the molecular structure was inferred from primary sequence, not by direct methods. Given the important function of type-1 fimbria, the availability of structural information, and potentially interesting selection regimes, additional study of the molecular evolution of *fimA* is merited.

Here we revisit the issue of selection on *fimA,* with the goal of better understanding the evolutionary forces acting on *fimA* in the context of its molecular structure. It is well known that functional and structural regions can be under different selective constraints (Kimura 1983). As one of several possible examples, a classic study of MHC genes demonstrated that antigen-recognition sites are subject to diversifying selection, whereas other regions of MHC molecules are subject to structural constraint (Hughes and Nei 1988; Hughes et al. 1990). Diversifying selection can be inferred directly by measuring the ratio of nonsynonymous-to-synonymous substitutions, where a ratio higher than 1.0 provides evidence of positive selection. Other types of selection can be inferred from tests of neutrality that compare genetic polymorphism within and between closely related species (e.g., Akashi 1999; Fu 1996; Hudson et al. 1987; McDonald and Kreitman 1991; Tajima 1989).

To connect structural and evolutionary information, we sequenced *fimA* from 21 *E. coli* that were isolated from a broad range of natural hosts. We combined these data with seven published *fimA* sequences and used the sequence data to address four questions: (i) Is there evidence for selection acting on the *fimA* gene of *E. coli,* as suggested previously? (ii) If so, what is the history of selection in different regions of the *fimA* protein? (iii) Is it necessary to have a structural model to make accurate inferences about the evolutionary history of protein regions, or is structural prediction from primary sequence sufficient to study the relationship between structure and evolution? and (iv) Finally, what is the interplay between genetic recombination and natural selection at the *fimA* locus?

Materials and Methods

Strains

A sample of 21 *Escherichia coli* isolates was chosen from the Souza collection (Souza et al. 1999) (and unpublished) to represent a range of hosts and geographic origins (Table 1). The 21 strains represent a diverse sampling of *E. coli,* including both human and nonhuman hosts as well as pathogenic and nonpathogenic strains. We amplified and sequenced the complete *fimA* gene from each of the 21 isolates.

PCR and Sequencing

Genomic DNAs from bacterial strains were purified for each isolate by CsCl gradient centrifugation. PCR primer pairs were designed from the *E. coli* K-12 MG1655 (Blattner et al. 1997) genome sequence and used to amplify the *fimA* locus (position −97 from start, forward, 5′-ACTGTGCAGTGTTGGCAG-3′; position +63 from stop, reverse 5′-GTTATTTTATCGCACAAGG). PCR amplifications included 2.5 μl of 10× polymerase buffer, 2.5 μl of 25 m*M* $MgCl_2$, 2.5 μl of 2 m*M* dNTPs, 1.0 μl of a 10 μ*M* concentration of each primer, 0.1 μl of DNA template, 0.1 μl of 10 U/μl Taq DNA polymerase (Promega, Madison, WI), and ddH_2O to 25 μl. The PCR profile (95°C for 1 min, 60°C for 1 min, 72°C for 1 min) continued for 30 cycles, followed by a final extension at 72°C for 10 min. Five microliters of reactions was checked by agarose gel and the remaining 20 μl was precipitated at 4°C with 10 μl 8 *M* NH_4OAc and 40 μl 95% EtOH for direct sequencing. Each PCR product was sequenced with both amplification primers separately, as well as an additional reverse primer (position +4 from stop, reverse 5′-TAGGTTATTGATACTGAACC-3′) by cycle-sequencing according to the manufacturer's instructions (ABI, Foster City, CA).

Sequence Analyses

The 21 *fimA* sequences generated for this study (AF206639–AF206659) were combined with 7 previously published *E. coli fimA* sequences (GenBank accession Nos. D13186, M27603, U14003, U20815, X00981, Y10902, and Z37500; Table 1) and a *fimA* sequence from *Salmonella enterica* (serovar Typhimurium LT2; GenBank accession No. L19338). The sequences were aligned by inferring amino acids and aligning corresponding codons with GDE 2.2a (Smith et al. 1994). The *fimA* gene contained two insertion/deletion (indel) polymorphisms of one or two codons in length within *E. coli,* and there were four indels fixed between *E. coli* and *S. enterica* (serovar Typhimurium LT2; GenBank accession No. L19338) (Figs. 1 and 4). We did not use indels in analyses.

Nucleotide sequence diversity was measured by the number of segregating sites with Watterson's (1975) θ and by π, the average pairwise difference (*k*) per site ($\pi = k$/sites) (Kimura 1969), as implemented in Sites 1.1a (Hey and Wakeley 1997). Standard deviations (SD) of estimates were calculated assuming minimal recombination, θ SD estimates were derived from Eq. 8, and π SD estimates were derived from $V_{\min}[k]$ Eq. 16 of Tajima (1993). θ and π were estimated for all sites and for nonsynonymous and synonymous sites separately. Tajima's *D* was calculated with the program Sites 1.1a (Hey and Wakeley 1997).

McDonald–Kreitman (MK) (1991) tests, as implemented in DNASP 3.0 (Rozas and Rozas 1997), were applied to the data, using the single *S. enterica* sequence to measure divergence. It is important to note that the two programs used in analyses, Sites and DNASP, count synonymous sites and nonsynonymous sites differently, and this contributes to the differences in Tables 2 and 3. More specifically, complex codons with histories of multiple substitutions were excluded from MK tests and logistic regression but included in polymorphism estimates. Excluded positions correspond with amino acids 20, 66, 70, 92, 105, 106, 148, and 168 in Fig. 1 and nucleotide sites 58–60, 196–198, 208–210, 274–276, 313–318, 439–441, and 503–504 in Fig. 4. The important feature is that exclusion of these codons is conservative with respect to MK test results. Logistic regression was based on JMP v. 4.0 (SAS Institute), using sites counted by DNASP (Table 3).

The likelihood method of Nielsen and Yang (1998) was used to test

Table 1. GenBank accession numbers of *fimA* and *mdh* sequences and their *E. coli* strains and hosts of origin

fimA GenBank No.	Strain ID code	Strain host/information	Host genus species
D13186	PDI-386	Avian (chicken), pathogenic O78	*Gallus gallus*
M27603	J96	Human, uropathogenic,	*Homo sapiens*
U14003	K-12 MG1655	Human, OR:H?	*Homo sapiens*
U20815	—	Human, O157:H7	*Homo sapiens*
X00981	—	Laboratory strain K-12	Laboratory K-12
Y10902	536	Human, uropathogenic,	*Homo sapiens*
Z37500	MT78	Avian (chicken), pathogenic O2:K1	*Gallus gallus*
AF206642	2[a]	Domestic dog, O8:H49, Mexico	*Canis familiaris*
AF206639	55[a]	Golden eagle, Mexico	*Aquila chrysaethus*
AF206643	78[a]	Painted spiny pocket mouse, O112:H2, Mexico	*Lyomis pictus*
AF206644	161[a]	Flycatcher (bird), Mexico	*Empidonax* sp.
AF206645	268[a]	Puma, O11:H15, Mexico	*Felis concolor*
AF206646	270[a]	Jaguar,[b] O159:H46, EPEC, Mexico	*Panthera onca*
AF206647	271	Jaguar,[b] O159:H46, EPEC, Mexico	*Panthera onca*
AF206648	287[a]	Merriam's kangaroo rat,[c] O138:H28, Mexico	*Dipodomys merriami*
AF206640	288[a]	Merriam's kangaroo rat,[c] O138:H28, Mexico	*Dipodomys merriami*
AF206649	298[a]	Pack rat, O19:H?, Mexico	*Neotoma albigula*
AF206650	820[a]	Desert pocket mouse, O88:H9, Mexico	*Perognathus penicillatus*
AF206651	1684	Nectar feeding bat, Mexico	*Choeronycteris mexicana*
AF206641	1698	Grey four-eyed opossum, O?:H4, Mexico	*Philander opossum*
AF206652	2165	Human, DEC 4E O157:H7, EHEC, Denmark	*Homo sapiens*
AF206653	2274 (B-427)[a]	Savanna baboon, O?:H?, Tanzania, Africa	*Papio cynocephalus*
AF206654	2314 (TA 120)[a]	Narrow footed marsupial mouse, O?:H?, Australia	*Sminthopsis delichura*
AF206655	2339 (TA 237)[a]	Broad footed marsupial mouse, OR:H6, Australia	*Antechinus flavipes*
AF206656	4981	Human, O111ab:H12, EPEC, Cairo Egypt	*Homo sapiens*
AF206657	4999	Human, O55:H7, EPEC, England	*Homo sapiens*
AF206658	5026	Human, O157:H7, EHEC, Thailand	*Homo sapiens*
AF206659	5028	Human, O157:H7, EHEC, Mexico	*Homo sapiens*

[a] Described by Souza et al. (1999).
[b] Strains 270 and 271 isolated from different jaguars.
[c] Strains 287 and 288 isolated from the same kangaroo rat.

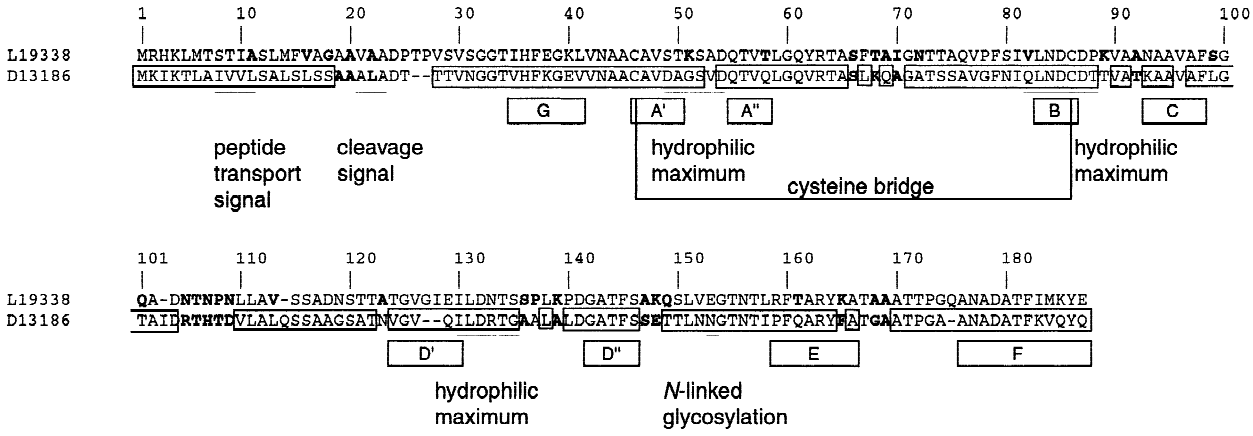

Fig. 1. Aligned amino acid sequence of the type-1 fimbria major subunit, *fimA,* from *Salmonella enterica* (**top;** L19338) and one haplotype from *Escherichia coli* (**bottom;** D13186). *Boxed areas* indicate identical amino acid regions within the present *E. coli* sample. Amino acids with d_N/d_S higher than 1.0 are identified in *boldface* on the *Salmonella* sequence for the analysis including *Salmonella* or the *E. coli* sequence for the analysis without *Salmonella.* Regions of known function are indicated below the aligned sequence, and *underlined*amino acid positions correspond to the text below. For example, three regions of maximal amino acid hydrophilicity are underlined. Secondary structure β-strands of the *fimA* peptide are designated with a *rectangular box* below the alignment containing the strand letter designation, and a single intramolecular cysteine bridge is indicated between position 47 and position 86.

for d_N/d_S ratios higher than 1.0 as implemented in PAML 2.0g (Yang 1997). To perform d_N/d_S ratio analyses we reconstructed allelic genealogies with the neighbor-joining (NJ) method based on Kimura (1980) two-parameter distances as implemented in Phylip 3.57 (Felsenstein 1989). Maximum-likelihood reconstructions were performed with PAUP* (Swofford 1999) and topologies were visualized with Treeview 1.5.2 (Page 1996).

We studied recombination among sequences with four methods. The first method was Sawyer's (1989) test for recombination. Sawyer's test is based on 10,000 permutations for the sum of the squares of

Table 2. Genetic diversity estimates at the *fimA* locus in natural isolates of *Escherichia coli*

fimA partition	All sites	Nonsynonymous	Synonymous
Complete molecule (546 sites)			
No. polymorphic sites	119	54	70
k (SD)	40.6 (3.81)	19.4 (2.64)	23.8 (2.92)
θ per site (SD)	0.05606 (0.00513)	0.02544 (0.00346)	0.03297 (0.00394)
π (SD)	0.07436 (0.00698)	0.03560 (0.00483)	0.04367 (0.00535)
Tajima's *D*	0.4347	0.6198	0.3487
β-strand (174 sites)			
No. polymorphic sites	18	2	16
k (SD)	4.4 (1.26)	0.8 (0.53)	3.6 (1.14)
θ per site (SD)	0.02753 (0.00627)	0.00306 (0.00209)	0.02447 (0.00591)
π (SD)	0.02551 (0.00724)	0.00454 (0.00305)	0.02098 (0.00657)
Tajima's *D*	−0.5867	1.0094	−0.8318
Non-β-strand (303 sites)			
No. polymorphic sites	79	36	48
k (SD)	32.0 (3.38)	16.1 (2.40)	18.6 (2.58)
θ per site (SD)	0.06705 (0.00754)	0.03055 (0.00509)	0.04074 (0.00587)
π (SD)	0.10577 (0.01118)	0.05303 (0.00791)	0.06157 (0.00853)
Tajima's *D*	0.9516	1.2394	0.7202
Interior (147 sites)			
No. polymorphic sites	13	1	12
k (SD)	3.6 (1.13)	0.07 (0.16)	3.5 (1.12)
θ per site (SD)	0.02273 (0.00630)	0.00175 (0.00175)	0.02098 (0.00606)
π (SD)	0.02433 (0.00770)	0.00049 (0.00109)	0.02385 (0.00762)
Tajima's *D*	−0.2463	−1.1514	−0.0869
Exterior (330 sites)			
No. polymorphic sites	84	37	52
k (SD)	33.4 (3.46)	17.1 (2.47)	19.0 (2.61)
θ per site (SD)	0.06667 (0.00714)	0.02937 (0.00474)	0.04127 (0.00561)
π (SD)	0.10110 (0.01047)	0.05173 (0.00749)	0.05763 (0.00791)
Tajima's *D*	0.8248	1.3472	0.4471

Table 3. McDonald and Kreitman tests for *fimA*[a]

	Total	F_{ns}	F_s	P_{ns}	P_s	*p*
All sites	187	113	83	21	81	$<10^{-7}$
Mature peptide	159	85	78	18	73	$<10^{-7}$
β-strand	58	25	30	2	18	5.7×10^{-3}
Non-β-strand	101	60	48	16	55	1.3×10^{-5}
Interior	50	24	31	1	14	7.3×10^{-3}
Exterior	109	61	47	17	59	4.0×10^{-6}

[a] Total is the number of amino acid positions in aligned sequences. F_{ns} is the number of fixed nonsynonymous differences between *E. coli* and *S. enterica*, F_s is the number of fixed synonymous differences between *E. coli* and *S. enterica*, P_{ns} is the number of polymorphic nonsynonymous sites within *E. coli*, and P_s is the number of polymorphic synonymous sites within *E. coli*. The probability *p* was calculated by a two-tailed Fisher's exact test.

condensed fragment lengths (SSCF), the maximum condensed fragment length (MCF), and the analogous metrics with uncondensed fragment lengths (SSUF and MUF). The second method was Stephens' test for deviations from random distributions of polymorphic sites. These tests were applied to all congruent nucleotide partitions according to Stephens' (1985). Eqs. 5, 9, and 10. The third method was Maynard Smith's (1992) maximum χ^2 (max-χ^2). This method was used to detect significant recombination boundaries, and results were based on 10,000 nucleotide sequence permutations of the allelic pair under consideration. Sawyer's, Stephens', and Maynard Smith's methods were implemented in programs written by the authors. Finally, the methods of Hey and Wakeley (1997) (γ) and Hudson (1987) (*C*) were used to estimate the population recombination parameter, $2N_e c$.

Structural Analyses and Partitions

Many of the analyses depended on partitioning the *fimA* amino acid residues into different structural categories based on homology to the pilin domain of *fimH*. The categories were based on the X-ray crystallography data for the pilin domain of *fimH* in the *fimH–fimC* complex (PDB accession No. 1QUN) and followed the alignment of *fimH* to position *fimA* peptide α carbons (Choudhury et al. 1999). Peptide structures were analyzed and visualized with RasMol and MolScript. Inferential methods from primary sequence for secondary structure prediction were based on Predator (Frishman and Argos 1996) and tertiary structure prediction was based on Hopp and Woods' (1981) amino acid hydrophilicity index.

From X-ray crystallography data, we partitioned *fimA* on the basis of two functional and structural categories: secondary structure and tertiary structure. We presume that positions involved in secondary and tertiary structure formation may have different functional constraints than other positions. Regions of secondary structure were defined as those involved in the formation of the β-barrel, and individual β-strands are depicted as rectangles below the aligned sequences in Fig. 1. Briefly, positions 36–41, 46–50, 55–58, 83–86, 93–98, 124–130, 142–146, 159–166, and 176–187 of the mature peptide were in β-strands, while all others were within the "non-β-strand" category.

The *fimA* tertiary structure can be summarized as follows. The N-terminal region (labeled G in Fig. 1) of one *fimA* molecule completes an immunoglobulin-like β-barrel between strand A and strand F of a second adjacent molecule along the fimbria's hollow interior (Choudhury et al. 1999). Regions G, A′, A″, and F are exposed to the 25-Å-diameter hollow center of the fimbria and are involved in the function of intersubunit interaction and chaperone binding, while strands B, C, D′, D″, and E face to the 70-Å-diameter exterior but are

still involved in β-barrel formation. Based on the tertiary structure, we defined interior and exterior regions. Interior regions correspond to positions 28–61 and 172–187 in the mature peptide, while exterior correspond to positions 62–171 (Fig. 1). Note that secondary and tertiary structure positions overlap slightly but are not identical (i.e., about half of the β-strand positions are interior and half exterior). Finally, we mention that the mature peptide starts at amino acid position 24, but this region of *fimA* contains some amino-terminal heterogeneity, and excluding indel polymorphism we defined amino acid positions 28–187 as the mature peptide (Fig. 1).

Results

fimA *Genetic Variation*

The nucleotide sequence of the *fimA* locus had been determined from seven *E. coli* isolates (Blattner et al. 1997; Klemm 1984; Marc and Dho-Moulin 1996; Orndorff and Falkow 1985; Sekizaki et al. 1993). To increase the number of isolates and to increase statistical power for analyses, we determined the *fimA* sequence for an additional 21 *E. coli* isolates. The total sample of 28 *fimA* sequences represents 20 unique nucleotide sequence haplotypes. Overall, there is a large amount of polymorphism segregating at the *fimA* locus, and increasing sequences from 7 to 28 raised π from 0.061 to 0.074 (Table 2). Also, π at replacement sites over the entire molecule (0.035) was slightly lower than π at synonymous sites (0.043) (Table 2). Tajima's (1989) D for all sites was positive ($D = 0.043$; Table 2) but not significantly different from zero. Despite the large amount of nonsynonymous polymorphism at *fimA*, amino acid positions proposed to be involved in the transport signal, cysteine disulfide bond, and a glycosylation site are invariant (Fig. 1).

Nonneutral Evolution of fimA

We tested neutrality with the MK test (McDonald and Kreitman 1991) and found significant deviations from neutrality for the mature peptide [$p < 10^{-7}$, Fisher's exact test (FET)] (Table 3). In addition, we divided the peptide into partitions defined by either secondary or tertiary structure based on the homology of *fimA* to the pilin domain of *fimH* (Choudhury et al. 1999). With respect to secondary structure, amino acid residues were defined as either β-strand or non-β-strand. This definition resulted in 58 β-strand amino acid residues and 101 non-β-strand amino acid residues. We applied MK tests to β-strand and non-β-strand regions separately, and tests of neutrality also showed significant deviations from neutrality for both β-strand ($p = 0.005$) and non-β strand sites ($p < 10^{-4}$) (Table 3). With respect to tertiary structure, we defined amino acid residues as either interior or exterior (see Materials and Methods). Both interior ($p = 0.007$) and exterior ($p < 10^{-5}$) regions rejected neutrality by the MK criterion (Table 3).

These tests all indicate that *fimA* is evolving nonneutrally, but the MK tests indicate both similarities and differences among regions of the molecule. One similarity is the direction of deviation from neutrality—i.e., in each molecular region there is a high ratio of nonsynonymous-to-synonymous fixed differences relative to nonsynonymous-to-synonymous polymorphisms. Differences among regions are more subtle. For example, both the interior and the β-strand regions segregate low amounts of nonsynonymous polymorphism within *E. coli* and contain fewer nonsynonymous than synonymous fixed differences. In contrast, exterior and non-β-strand regions segregate relatively high amounts of nonsynonymous polymorphism within *E. coli* and also contain more nonsynonymous than synonymous fixed differences between *E. coli* and *S. enterica* (Table 3). Altogether, the MK tests suggest that all regions of the molecule are evolving nonneutrally but that different regions may have slightly different evolutionary histories.

To examine these dynamics more closely and to control for nonindependence among MK tests, we recoded Table 3 data in a multifactorial design and applied logistic regression. The independent variables for logistic regression were tertiary structure (internal or external), secondary structure (non-β-strand and β-strand), and change type (nonsynonymous or synonymous). The dependent variable was the variable state (fixed or polymorphic). Log linear models including tertiary structure and change type were significantly better at predicting the dependent variable than models without tertiary structure and change type, at probability values of $p = 0.02$ and $p = 0.001$, respectively. Conversely, models that included secondary structure ($p = 0.14$) and second-order interactions among independent variables (data not shown) were not a significant improvement. (It was not possible to test third-order interactions among independent variables because of cells with low counts in the multifactorial table.) Biologically, these results imply two things. The first is that the distribution of fixed and polymorphic sites varies among nonsynonymous and synonymous sites; in short, this result corroborates the MK test (Table 3). The second is that the distribution of fixed and polymorphic sites varies with tertiary structure, implying that evolutionary dynamics differ between interior and exterior sites of the molecule.

To characterize further differences among molecular regions, we estimated polymorphism metrics for different structural partitions (Table 2). Considering only within-species variation, nonsynonymous and synonymous polymorphism was nonrandomly distributed between interior (1 nonsynonymous, 12 synonymous) and exterior (37 nonsynonymous, 52 synonymous) regions of the molecule ($p = 0.0282$, FET), with a higher ratio of nonsynonymous-to-synonymous polymorphism in exterior regions. These differences also extend to β-strand and non-β-strand regions ($p = 0.0143$). Furthermore, the interior and β-strand regions contained much lower lev-

els of variation (β-strand $\pi = 0.02$ and interior $\pi = 0.02$) than their complements ($\pi = 0.06$). Finally, none of Tajima's tests deviated significantly from zero, but interior and β-regions of the fimbrial protein exhibit a negative value rather than positive value for Tajima's *D* (Table 2). Overall, MK tests, logistic regression, and polymorphism metrics suggest that different structural regions are segregating different amounts and patterns of polymorphism.

Sites of High d_N/d_S

One objective of this study was to identify sites of positive diversifying selection within *E. coli* and test whether these were nonrandomly distributed across the structural model of the *fimA* peptide. We tested the hypothesis of positive diversifying selection by comparing two models of amino acid evolution (Nielsen and Yang 1998). The first model was a neutral model that had two categories of codon sites: a neutral site where the d_N-to-d_S ratio was 1.0 and a purifying selection site where d_N/d_S was 0.0. The second model was a positive selection model that contains both types of codon sites in the neutral model and also included positively selected codon sites where d_N/d_S was allowed to be higher than 1.0. The models were applied to two data sets. One data set contained the 28 *E. coli* alleles and the *Salmonella* outgroup (Fig. 2), and the second data set included only the 28 alleles from within *E. coli.*

Comparisons between the neutral and the positive selection models for the data set including *Salmonella* yielded a significantly better fit of the data to the positive selection model relative to the neutral model by a likelihood-ratio (LR) test (LR = 22.9; $p = 1.06 \times 10^{-5}$, df = 2), and the positive selection model yielded a d_N/d_S estimate of 2.64. Comparisons between models for the data set including only *E. coli* sequences also showed a significantly better fit of the data to the positive selection model compared to the neutral model (LR = 29.88; $p = 3.25 \times 10^{-7}$), with a corresponding d_N/d_S estimate of 2.74. These results are not dependent on the neighbor-joining tree topology since several candidate trees from heuristic maximum-likelihood searches also yielded significant differences between the two models (not shown). These results suggest that positive diversifying selection has acted on some amino acid residues within *E. coli.*

The location of codons where the ratio of nonsynonymous-to-synonymous substitution exceeds 1.0 may help reveal the nature of positive selection occurring at the *fimA* gene both within and between *E. coli* and *S. enterica.* For the data set including *S. enterica*, 34 of 179 amino acid residues were inferred to evolve under positive selection (Fig. 1). Five of the 34 positively selected amino acid sites were in the leader peptide region in the *fimA* molecule and 29 were in the mature subunit. The 29

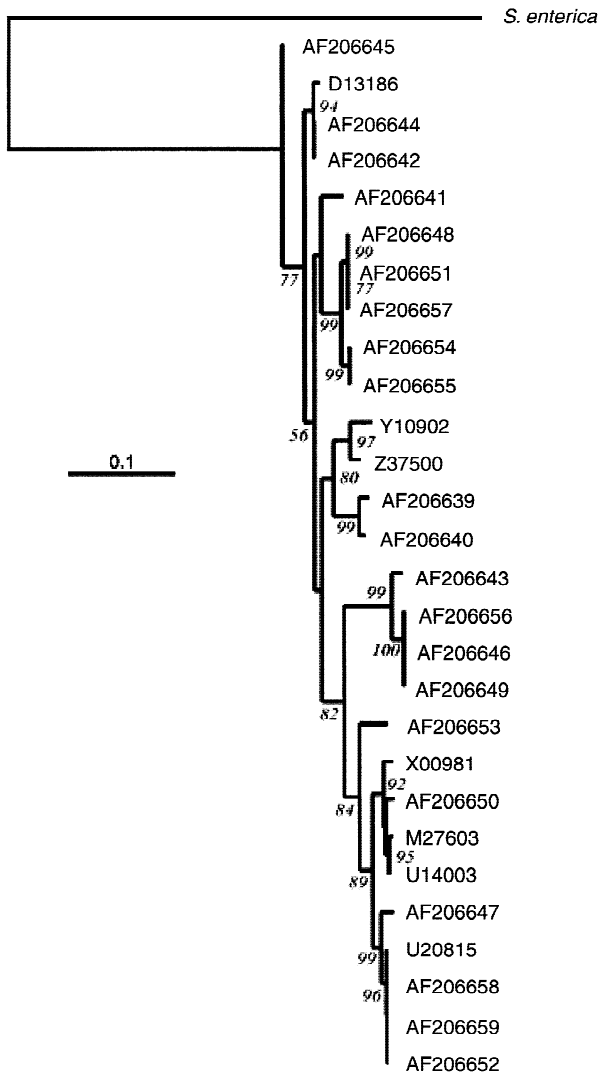

Fig. 2. The neighbor-joining phylogeny of *E. coli fimA* sequences, with an *S. enterica fimA* sequence as outgroup. OTU labels for *E. coli* are Genbank accession numbers given in Table 1. Bootstrap percentages, based on 1000 bootstrapped data sets, are given in *italics* either below or to the right of nodes; percentages greater than 50 are shown. The scale bar indicates the level of per-site sequence divergence.

residues in the mature subunit were highly nonrandom in distribution, with 27 of the 29 in exterior exposed regions (27 positive exterior, 82 nonpositive exterior, 2 positive interior, and 46 nonpositive interior; $p = 0.0015$, FET) and 25 of the 29 in non-β-strand structures ($p < 10^{-4}$). Nineteen amino acids were inferred to be under positive selection in the *E. coli* data set. Eighteen of these 19 amino acids were also identified as positively selected in the data set that included *S. enterica* (Fig. 1). Sixteen of the 19 positively selected amino acids are in the mature peptide, and all of these 16 occurred in exterior exposed regions ($p = 0.0031$). Fifteen of the 16 were in non-β-strand structures ($p = 0.0059$).

The physical location of positively selected amino acid positions are shown in Fig. 3. Amino acid positions

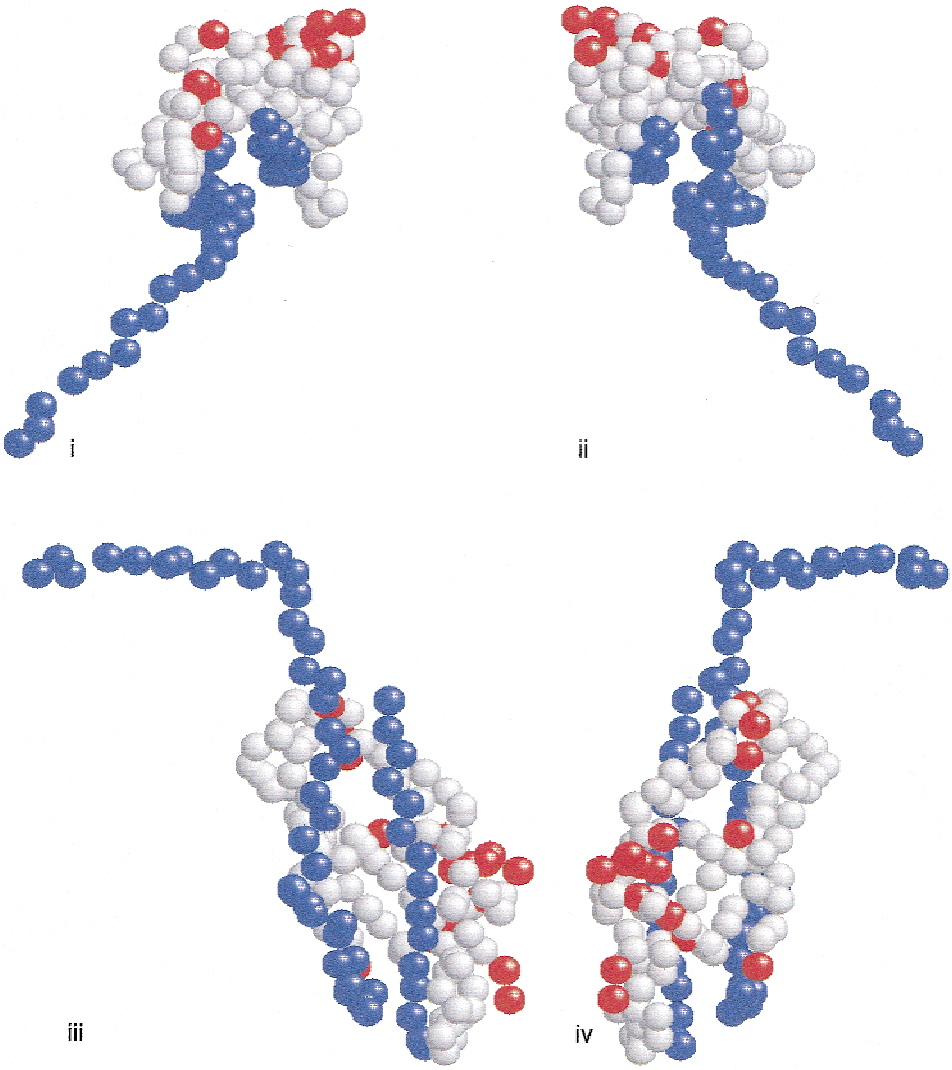

Fig. 3. Location of *fimA* peptide α carbons. *Blue* sites are on the interior of the fimbrial structure and *gray* sites are on the exterior. A subset of the exterior sites under positive diversifying selection within *E. coli* is shown in *red.* Four perspectives of the *fimA* peptide are **(i)** exterior up, G-strand (involved in donor-strand complimentation) to the front; **(ii)** exterior up, G-strand to the rear; **(iii)** exterior to the rear, G-strand to the left; and **(iv)** exterior to the front, G-strand to the right. Side views (i and ii) reveal the hollow interior channel involved in donor-strand complimentation in blue and the exterior surface of red positively selected sites. Interior view iii also reveals the A-strand and F-strand regions involved in donor-strand complimentation in blue, and exterior view iv shows the exterior surface of red positively selected sites.

under positive diversifying selection in *E. coli* are exclusively located on the exterior regions of *fimA*, the fimbrial tertiary structure, and tend not to be involved in the secondary structure formation of the immunoglobulin-like β-barrel.

Comparisons Between Direct and Indirect Estimates of Fimbrial Structure

An additional question we addressed was whether the secondary and tertiary structural information from X-ray crystallography was necessary for effectively assigning the molecular location of sites under positive selection. We estimated secondary structure from primary *fimA* sequences (Frishman and Argos 1996) and found few sites accurately inferred according to structural information. For tertiary structure prediction, hydrophilic amino acid regions tend to be exposed to exterior regions (antigenic), while hydrophobic regions tend to be associated with the interior of globular proteins (Hopp and Woods 1981). We examined the *fimA* gene for overall correlations between hydrophilicity and genetic variation, expecting a positive correlation if hydrophilic regions are more variable. A series of sliding-window analyses was performed to compare amino acid hydrophilicity to variation, but no significant correlations were observed and most correlations were negative. Finally, amino acids under positive selection do not have significantly different hydrophilicities than other amino acid postions. Overall, secondary and tertiary structure predictions

from primary sequence would not be useful for identifying sites of different constraint catagories.

Recombination

Positive diversifying selection appears to be responsible for some of the variation observed at the *fimA* locus within *E. coli.* To examine whether recombination also contributes to allelic variation, we applied Sawyer's (1989) tests on the *E. coli* nucleotide sequences. All four of Sawyer's metrics were highly significant ($p < 10^{-3}$). Stephens' (1985) tests identified 19 phylogenetically congruent partitions, where 14 of these contained significantly ($p < 0.05$) nonrandom distributions of sites and 10 of the 14 were highly significant ($p < 10^{-4}$). Lengths of the significant partitions identified by Stephens' tests ranged from 2 to 130 nucleotide sites. Overall, these tests give the impression that recombination is a common occurrence at the *fimA* locus.

Close examination of polymorphic sites (Fig. 4) also suggests prevalent recombination. For example, AF206653 is identical to the K-12 allele U14003 from position 1 to position 302 ($p < 10^{-3}$, max-χ^2) (Maynard Smith 1992) and identical to the AF206648 allele from position 304 to position 521 ($p < 10^{-3}$, max-χ^2). A second example of recombination is the AF206641 allele, which contains regionally distinct similarity to both allele Z37500 and allele AF206651 ($p < 10^{-3}$). Other alleles such as Y10902 appear to contain a region in the leader peptide between nucleotide site 59 and nucleotide site 65 that does not occur elsewhere in our sample but may have resulted from recombination events with an unsampled allele.

Estimates of $2N_ec$ per site recombination also suggest that recombination is a powerful force at the *fimA* locus. Estimates of $2N_ec$ per site by (Hey and Wakeley 1997) or (Hudson 1987) method were 0.087632 or 0.067641, while estimates of $2N_e\mu$ per site mutation by π or θ were 0.07436 or 0.05606, respectively. Ratios of $2N_ec/2N_e\mu$ compare relative amounts of recombination to mutation per nucleotide site and are about 1, suggesting that recombination and mutation occur at roughly the same rate in the *fimA* gene.

Discussion

We have made the following observations about genetic variation in the *fimA* locus of *E. coli.* First, the gene is highly variable, even more variable than the unusually variable *gnd* locus ($\pi = 0.04$) (Nelson and Selander 1994), which has been suggested to segregate high polymorphism due to selection acting at the nearby O surface antigen locus *rfb.* Second, significant MK tests suggest nonneutral evolution between *E. coli* and *S. enterica* or within *E. coli.* Third, logistic regression indicates that the distribution of fixed and polymorphic sites differs between interior and exterior regions of the molecule. Differences among molecular regions are also supported by other polymorphism metrics. Fourth, d_N/d_S analyses detect amino acid residues with ratios higher than 1.0, and these are clustered nonrandomly in the exterior and non-β-strand regions of the *fimA* peptide. Finally, there is much evidence for recombination within *fimA;* estimates suggest that recombination and mutation occur at similar rates.

Diversifying Selection

These observations contribute to a comprehensive view of *fimA* evolution. First consider high variability in the *fimA* locus, which could be caused by numerous factors. One concern is that high polymorphism at *fimA* reflects our sampling strategy more than it reflects a property of *fimA* evolution. Two points argue against this notion. First, the seven previously published *fimA* sequences also exhibited extremely high nucleotide polymorphism (Boyd and Hartl 1998). Second, we sequenced the housekeeping gene malate dehydrogenase (*mdh*) from the same 21 wild isolates (Genbank accession Nos. AF230568, AF230569, AF230571–AF230589) and found that *mdh* polymorphism ($\pi = 0.01$) is very similar to polymorphism for *mdh* sequences sampled from the ECOR collection (Boyd et al. 1994; Pupo et al. 1997). Thus, high polymorphism in our *fimA* sample cannot be attributed solely to our sampling strategy.

A more likely reason for high variability is positive selection for diversity, and evidence from this and a previous study (Boyd and Hartl 1998) strongly suggests that diversifying selection is acting on *fimA.* Evidence for diversifying selection comes not only from high genetic variability but also from analyses that detect d_N/d_S ratios higher than 1.0, particularly on the exterior of the *fimA* molecule. Moreover, it is biologically reasonable to presume that diversifying selection acts on *fimA,* because diversifying selection can be fueled by frequency-dependent avoidance of immune responses. There are approximately 1×10^3 *fimA* subunits per fimbriae, 5×10^2 fimbriae per *E. coli* cell, and 1×10^7 *E. coli* cells per ml within the intestine, for a total of roughly 5×10^{12} *fimA* subunits per ml in the intestine (Eisenstein 1987; Selander et al. 1987). With these large numbers, it is not hard to imagine the potential for immune system mediated selection on the *fimA* peptide.

One can consider additional alternative interpretations to diversifying selection, but none of the alternatives appear likely in this case. For example, the interpretation of d_N/d_S values implicitly assumes that synonymous substitutions are neutral, and d_N/d_S values can be inflated if synonymous substitutions are highly constrained. However, there is no strong evidence of synonymous constraint in *fimA.* Furthermore, if there were synonymous

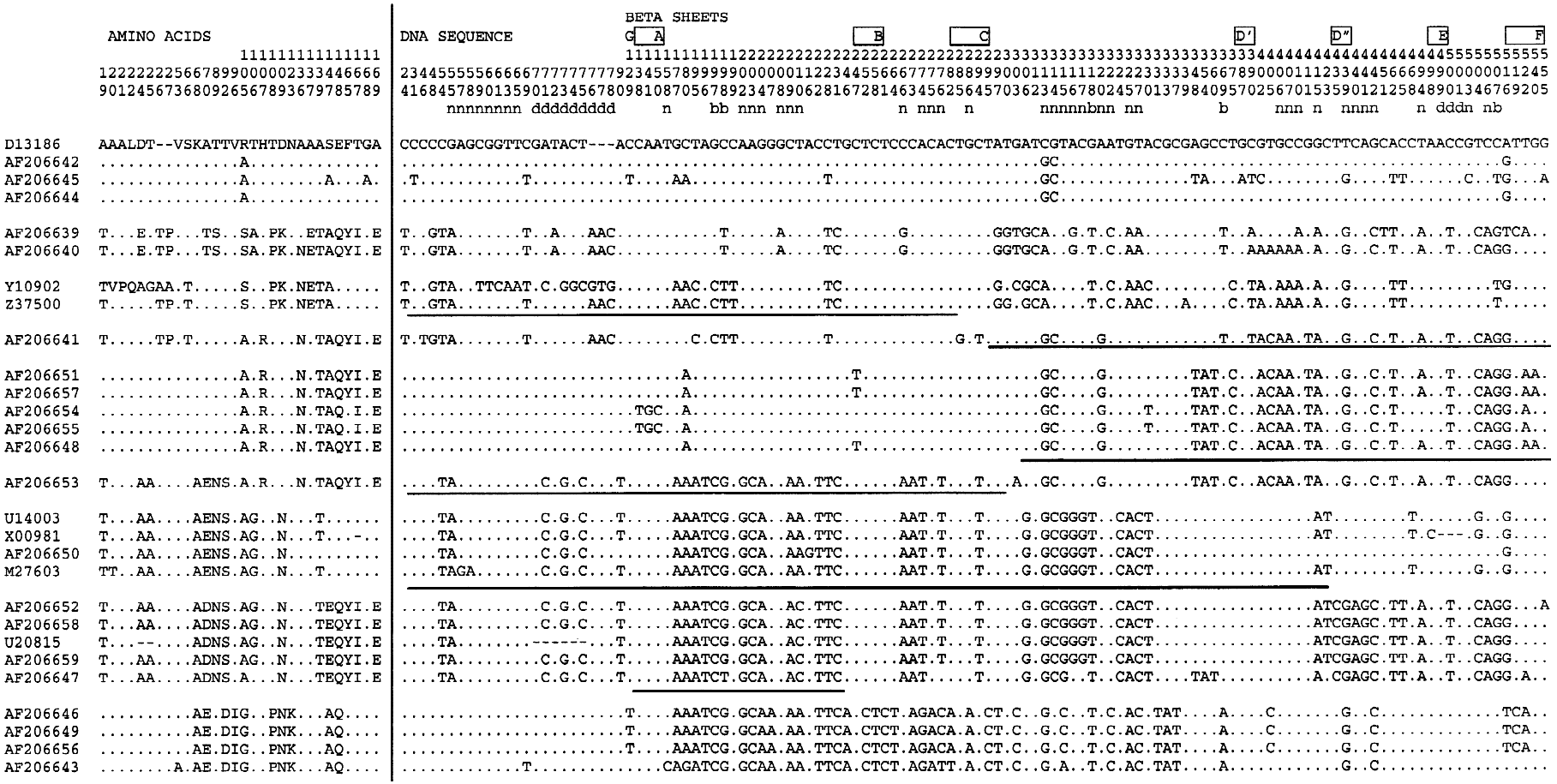

Fig. 4. All polymorphic positions in the amino acid and nucleotide sequences of *Escherichia coli* are shown in 28 *fimA* sequences. The uppermost haplotype sequence is complete, whereas each sequence below contains a *period* to represent an identical character. To the left of the vertical bar are the variable amino acid positions, labeled across the top with position numbers corresponding to Fig. 1. Sequences were separated in groups, spaced top to bottom, to minimize amino acid sequence variation within groups. To the right of the vertical bar are the variable nucleotide positions, labeled by position across the top. Above the nucleotide site, numbers correspond to letter designations of sites in β sheets (if appropriate), and below are letter designations for polymorphic nucleotide site classes: nonsynonymous (n), synonymous (blank), both nonsynonymous and synonymous (b), and insertion/deletion (d). *Horizontal lines* between grouped haplotypes represent regions of significantly nonrandom similarities between nucleotide sequences above and below the line by Maynard Smith's max-χ^2 method (i.e., recombination). Not all significant recombination events identified by the maximum-χ^2 method are shown.

constraint, the *fimA* gene would have either low total amounts of synonymous variation or particularly low variation in regions where high d_N/d_S ratios are detected (Smith et al. 1995). In addition, this would require purifying selection to act differently on synonymous substitutions on the exterior of the molecule. If this were the case, one might expect codon usage indices [e.g., codon adaptation index (CAI), third-position G+C content] to differ between regions. We observe no significant differences, suggesting elevated synonymous constraint at locations with higher d_N/d_S ratios for CAI (Sharp and Li 1987) or third-position G + C content in either interior (CAI = 0.408, G + C = 0.414) versus exterior (0.416, 0.447) or β-strand (0.473, 0.422) versus non-β-strand (0.384; 0.445) regions.

Other Kinds of Selection on fimA

The evidence for diversifying selection is strong, but diversifying selection alone does not explain all of our observations. For example, MK tests suggest that the processes contributing to low nonsynonymous-to-synonymous polymorphism do not represent the processes that acted during the divergence of *E. coli* and *S. enterica.* MK tests that deviate toward a high ratio of nonsynonymous:synonymous fixed differences are often interpreted as indicative of directional selection—i.e., adaptive selection after divergence (Long and Langley 1993; McDonald and Kreitman 1991). Perhaps, then, the *fimA* locus has had a history that features episodes of directional selection in addition to diversifying selection. Another possible cause of significant MK tests could be the result of weak selection during the divergence between *E. coli* and *S. enterica* at either synonymous or nonsynonymous positions (Akashi 1995). If weak selection were occurring at synonymous positions, the relative contributions of codon usage selection and mutational bias would be difficult to know, but the *fimA* locus has an overall low CAI, thus the majority of synonymous polymorphism may be slightly deleterious. In the end, we are uncertain about the cause of rejection of MK tests, but the data are consistent with diversifying selection within *E. coli* and directional selection between *E. coli* and *S. enterica* at nonsynonymous positions.

Another observation that is not easily explained solely by diversifying selection is the differences in patterns of polymorphism among molecular regions. More specifically, our analyses detect few d_N/d_S values higher than 1.0 within the interior or β-strand regions of the molecule, and logistic regression indicates that the distribution of fixed and polymorphic sites varies between interior and exterior regions. It is not unreasonable to presume that the interior and β-strand regions are subject to strong selective constraint, because these regions are important for the structural integrity of individual subunits, intersubunit interactions, and interactions with chaperone transport peptides (Choudhury et al. 1999). This view is consistent with the low ratio of nonsynonymous-to-synonymous polymorphisms in this region and also consistent with the slightly negative Tajima's *D* for the interior region. Also, the single variable amino acid position in the interior region (position 53; Fig. 1) occurs in only a single *fimA* allele, and this pattern is consistent with a slightly deleterious model for this polymorphism (Akashi 1999). Thus, we favor the view that the interior region—and probably β-strand regions as well—is evolving under strong selective constraint within *E. coli.*

One surprising aspect of our analyses is that the regions presumably dominated by purifying selection in *E. coli* (i.e., interior and β-strand regions) also reject the neutrality hypothesis of the MK test, suggesting that they may not have evolved neutrally since the divergence of *E. coli* from *S. enterica.* It is difficult to determine the forces acting on these regions of molecule, but there are several possibilities. These include (i) recent adaptive selection in β-strand and interior regions that have removed amino acid polymorphism, (ii) historically reduced selective constraint in these regions during the divergence of *E. coli* and *S. enterica,* and (iii) some combination of linkage and selection that has skewed historical or current ratios of nonsynonymous-to-synonymous polymorphism. We cannot discriminate among these possibilities, but the first seems unlikely since an equal reduction of synonymous polymorphism is expected and not observed.

Recombination, Structural Inference, and the Evolution of fimA

A previous study of bacterial adhesins included seven alleles from the *fimA* locus and suggested that positive diversifying selection is predominantly responsible for *fimA* allelic variation within *E. coli* (Boyd and Hartl 1998). With an increased number of haplotypes, we observe extensive evidence for recombination and this result is consistent with other evidence for recombination in this region of the *E. coli* genome (Louarn et al. 1994). In *fimA,* we calculate that the ratio of recombination rate to mutation rate is roughly equal to one. This ratio has also been estimated for other *E. coli* genes and found to range from about 1 to 10 (Guttman 1997; Whittam 1996). The *fimA* locus is on the low end of this range, but it must be noted that this ratio is estimated with neutral models. Under these models, diversifying selection likely decreases the *fimA* ratio estimate. In any event, recombination plays a large role in the evolution of *fimA* and probably uncouples the forces of diversifying and purifying selection that act in different regions of the molecule.

Finally, we incorporated X-ray structural information of the type-1 fimbria major subunit *fimA* to partition the molecule into regions of known function. Structural in-

formation is not available for most gene products, and we were interested in whether inferential techniques were useful in identifying structural regions for genetic diversity analyses. Inferring secondary structure from the primary sequence is generally about 65% accurate (Frishman and Argos 1996), and this is close for *fimA* since 15 of the 22 (68%) predicted in β-strands were correctly identified. However, a larger problem in secondary structure inference was the incorrect assignment of most sites that are actually involved in β-strand formation. Assignment of tertiary structure by amino acid hydrophilicity is also ineffective for the highly hydrophobic *fimA* peptide (Eisenstein 1987; Klemm 1984; Klemm and Hedegaard 1990). Anomalous tertiary structure estimates may be due partly to the hollow interior of type-1 fimbria containing hydrophilic amino acids (see Fig. 1) and the interior region is likely subject to purifying not diversifying selection. Overall for *fimA,* direct methods for structure assignment appear to be necessary for the accurate designation of specific amino acid positions to functional and structural regions. Without accurate designation of amino acid positions, it would have been difficult to appreciate fully the disparate evolutionary forces acting on exterior and interior regions of the *fimA* peptide.

Acknowledgments. We thank Valerie Bouchet, Aldo Valera, and Rebecca Gaut for technical assistance, Armando Navarro and Dr. Alejandro Cravioto for serotypes and human strains, and two anonymous reviewers for insightful comments, especially the idea to apply logistic regression. This work was supported in part by Grants IN-218698 from DGAPA-UNAM and 27557-M from CONACyT to V.S., a sabbatical scholarship from DGAPA-UNAM and CONACyT to V.S. and L.E., a Sloan Young Investigator Award to B.G., and NSF Biological Informatics Postdoctoral Fellowship DBI-9974237 to A.P.

References

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics 139:1067–1076

Akashi H (1999) Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151:221–238

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474

Boyd EF, Hartl DL (1998) Diversifying selection governs sequence polymorphism in the major adhesion proteins *fimaA, papA,* and *sfaA* of *Escherichia coli.* J Mol Evol 47:258–267

Boyd EF, Nelson K, Wang F-S, Whittam TS, Selander RK (1994) Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica.* Proc Natl Acad Sci USA 91:1280–1284

Choudhury D, Thompson A, Stojanoff V, Langermann S, Pinkner J, Hultgren SJ, Knight SD (1999) X-ray structure of the *fimC-fimH* chaperone-adhesin complex from uropathogenic *Escherichia coli.* Science 285:1061–1066

Eisenstein BI (1987) Fimbriae. In: Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE (eds) *Escherichia coli* and *Salmonella typhiurium* cellular and molecular biology. American Society for Microbiology, Washington, DC, pp 84–90

Elliott SJ, Kaper JB (1997) Role of type I fimbriae in EPEC infections. Microbial Pathogen 23:113–118

Felsenstein J (1989) PHYLIP—Phylogeny inference package (version 3.2). Cladistics 5:164–166

Frishman D, Argos P (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng 9:133–142

Fu Y-X (1996) New statistical tests of neutrality for DNA samples from a population. Genetics 143:557–570

Guttman DS (1997) Recombination and clonality in natural populations of *Escherichia coli.* Trends Ecol Evol 12:16–22

Hey J, Wakeley J (1997) A coalescent estimator of the population recombination rate. Genetics 145:833–846

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78: 3824–3828

Hudson RR (1987) Estimating the recombination parameter of a finite population model without selection. Genet Res 50:245–250

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 89:167–170

Hughes AL, Ota T, 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

Hultgren SJ, Jones CH, Normark S (1996) Bacterial adhesins and their assembly. In: Neidhardt FC, Curtis RR, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umbarger HE (eds) *Escherichia coli* and *Salmonella typhiurium* cellular and molecular biology. ASM Press, Washington, DC, pp 2730–2756

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903

Kimura M (1980) A simple method for estimating the evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Klemm P (1984) The *fimA* gene encoding the type-1 fimbrial subunit of *Escherichia coli.* Eur J Biochem 143:395–399

Klemm P, Hedegaard L (1990) Fimbriae of *Escherichia coli* as carriers of heterologous antigenic sequences. Res Microbiol 141:1013–1017

Klemm P, Ørskov I, Ørskov F (1982) F7 and type I-like fimbriae from three *Escherichia coli* strains isolated from urinary tract infections: Protein chemical and immunological aspects. Infect Immun 36: 462–468

Long M, Langley CH (1993) Natural selection and the origin of *jingwei,* a chimeric processed functional gene in *Drosophila.* Science 260:91–95

Louarn J, Cornet F, Francois V, Patte J, Louarn JM (1994) Hyperrecombination in the terminus region of the *Escherichia coli* chromosome: Possible relation to nucleoid organization. J Bacteriol 176:7524–7531

Marc D, Dho-Moulin M (1996) Analysis of the fim cluster of an avian O2 strain of *Escherichia coli:* Serogroup-specific sites within *fimA* and nucleotide sequence of *fimI.* J Med Microbiol 44:444–452

Maynard Smith J (1992) Analyzing the mosaic structure of genes. J Mol Evol 34:126–129

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila.* Nature 351:652–654

Nelson K, Selander RK (1994) Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene *(gnd)* in enteric bacteria. Proc Natl Acad Sci USA 91:10227–10231

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Orndorff PE, Falkow S (1985) Nucleotide sequence of *pilA*, the gene encoding the structural component of type 1 pili in *Escherichia coli*. J Bacteriol 162:454–457

Page RDM (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. Comp Appl Biol Sci 12:357–358

Pupo GM, Karaolis DKR, Lan R, Reeves PR (1997) Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. Infect Immun 65:2685–2692

Rozas J, Rozas R (1997) DnaSP version 2.0: A novel software package for estensive molecular population genetics analysis. Comp Appl Biol Sci 13:307–311

Salit IE, Vavougios J, Hofmann T (1983) Isolation and characterization of *Escherichia coli* pili from diverse clinical sources. Infect Immun 42:755–762

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Sekizaki T, Ito H, Asawa T, Nonomura I (1993) DNA sequence of type 1 fimbrin, *Fpul1*, gene from a chicken pathogenic *Escherichia coli* serotype O78. J Vet Med Sci 55:395–400

Selander RK, Caugant DA, Whittam TS (1987) Genetic Structure and variation in natural populations of *Escherichia coli.* In: Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE (eds) *Escherichia coli* and *Salmonella typhiurium* cellular and molecular biology. American Society for Microbiology, Washington, DC, pp 1625–1648

Sharp PM, Li W-H (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Smith NH, Maynard Smith J, Spratt BG (1995) Sequence evolution of the *porB* gene in *Neisseria gonorrhoeae* and *Neisseria meningitidis:* Evidence of positive Darwinian selection. Mol Biol Evol 12: 363–370

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM (1994) The genetic data environment and expandable GUI for multiple sequence analysis. Comp Appl Biol Sci 10:671–675

Souza V, Rocha M, Valera A, Eguiarte LE (1999) Genetics structure of natural populations of *Escherichia coli* in wild hosts on different continents. Appl Environ Microbiol 65:3373–3385

Stephens JC (1985) Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. Mol Biol Evol 2:539–556

Swofford DL (1999) PAUP*. Phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland, MA

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tajima F (1993) Measurement of DNA polymorphism. In: Takahata N, Clark AG (eds) Mechanisms of molecular evolution. Sinauer Associates, Sunderland, MA pp 37–59

Tewari R, MacGregor JI, Ikeda T, Little JR, Hultgren SJ, Abraham SN (1993) Neutrophil activation by nascent *fimH* subunits of type 1 fimbriae purified from the periplasm of *Escherichia coli.* J Biol Chem 268:3009–3015

Watterson GA (1975) On the number of segregating sites in genetic models without recombination. Theor Popul Biol 7:256–276

Whittam TS (1996) Genetics variation and evolutionary processes in natural populations of *Escherichia coli.* In: Neidhardt FC (ed) *Escherichia coli* and *Salmonella typhiurium* cellular and molecular biology. American Society for Microbiology, Washington, DC, pp 2708–2720

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comp Appl Biol Sci 13:555–556