The Molecular Evolution of Bacterial Alkaline Phosphatase: Correlating Variation among Enteric Bacteria to Experimental Manipulations of the Protein

Robert F. DuBose and Daniel L. Hartl

Department of Genetics, Washington University School of Medicine

The phylogenetic distribution of the gene coding for bacterial alkaline phosphatase (phoA) was examined in nine species of enteric bacteria closely related to Escherichia coli. The nucleotide and protein sequences from the E. fergusonii and Serratia marcescens genes are presented. The spatial distribution of replaced amino acid residues in the aligned sequences is shown to be highly nonrandom and can be correlated with specific regions within the tertiary structure of the protein. There is an avoidance of replacements within the beta sheet of the protein, and there is an excess of replacements elsewhere, particularly in solvent-exposed residues. In addition, all positions across alpha helices do not accept replacements with equal frequency; there is a bias toward acceptance of replacements in the carboxyl ends of helices. To examine this further, mutations within the E. coli phoA gene were created using site-directed mutagenesis. The patterns seen from the sequence comparisons were verified in the laboratory-created mutants. The average activity of mutations within or near the beta sheet was approximately one-third of that within or near alpha helices, and multiple mutations within the carboxyl ends of alpha helices always possessed greater activity than did multiple mutations within the corresponding amino ends. The results indicate that identifiable regions within the protein are under different selective pressures and are therefore evolving at different rates.

Introduction

Comparative studies have been the foundation of molecular evolution over the past 25 years in providing hypotheses concerning phylogenetic reconstruction, population structure, and molecular processes of variation. If they have been limited in any way, it is in that they do not generally suggest an experimental method to be used in testing the theories which they generate, except by the further gathering of comparative data. For example, the extensive compilation of cytochrome and globin sequences has allowed workers such as Fitch (1976), Holmquist et al. (1983), and Kunisawa et al. (1987) to provide evidence for unusual patterns of evolution within proteins across lineages. Examination of more sequences has refined the distribution parameters but has not shed any light on the original question of what leads certain residues to be so variable or so inflexible. This aspect of the problem, reflecting a need for a functional morphology for proteins, can be resolved in part by using proteins whose tertiary structures are known, proteins such as the globins (Perutz et al. 1965; Zuckerlandl and Pauling 1965; Lcsk and Chothia 1980; Perutz 1983). While it is

1. Key words: protein evolution, site-directed mutagenesis, phoA, Enterobacteriaceae.
2. Present address: Institute of Molecular Biology, University of Oregon.

Address for correspondence and reprints: Daniel L. Hartl, Department of Genetics, Washington University School of Medicine, 4566 Scott Avenue, St. Louis, Missouri 63110-1095.
It is in this respect that a direct experimental approach can be useful. By means of molecular techniques, it is now possible to create in vitro mutations that are not found in nature and to examine their effects on the structure and activity of the protein. One technique in particular, site-directed mutagenesis (Zoller and Smith 1983; Smith 1985), has revolutionized the field of protein structure and function. By enabling one to make specific changes in a DNA sequence, it gives the investigator the ability to manipulate a protein in ways not previously possible. Its use in biochemistry has been widespread, ranging from investigation of the roles of one or a few key residues in enzymatic function to exploration of the patterns of mutational flexibility that may ultimately lead to the prediction of protein structure (Butler-Ransohoff et al. 1988; Chaidaroglou et al. 1988; Reidhaar-Olson and Sauer 1988; Bowie and Sauer 1989; Lim and Sauer 1989). In general, however, directed mutagenesis has not been used in molecular evolution, where the ability to make specific changes at the DNA level makes it ideal for use in testing the hypotheses generated from comparative studies.

In the present study, we have combined both comparative surveys and site-directed mutagenesis. As a model system we have chosen a bacterial protein, alkaline phosphatase, for which the *E. coli* DNA sequence and tertiary structure, as well as an extensive background of biochemical and genetic information, are available (Torriani-Gorini et al. 1987). We have obtained the DNA sequence for the gene coding for bacterial alkaline phosphatase (*phoA*) from related species of enteric bacteria and have investigated the patterns of accumulation of amino acid replacements to develop hypotheses concerning the selective forces shaping the evolution of the protein. To test these predictions, we have used site-directed mutagenesis to introduce changes in the *E. coli* *phoA* gene, on the basis of the hypotheses provided by the comparative study. We have created alterations in the gene that mimic those seen within the various sequences and have also tested the effects of replacements in regions that show a high degree of constraint across the various lineages. In this manner, we show that there is a clear functional interpretation of the patterns of amino acid replacements seen in the comparative analysis.

**Material and Methods**

**Bacterial Strains**

Strains identified by an American Type Culture Collection (ATCC) number were obtained directly from the ATCC unless otherwise noted. Strains used were *Escherichia coli* K-12 (strain W3110), *Shigella dysenteriae* (ATCC 13313), *S. flexneri* (ATCC 29508), *S. sonnet* (ATCC 29930), *E. hermanii* (ATCC 33650), *E. vulneris* (ATCC 33821), *E. fergusonii* (ATCC 35469 was from Don Brenner, and ATCC 35470, 35471, 35472, and 35473 were from the ATCC), *E. blattae* (ATCC 29907), *Salmonella typhimurium* LT-2, and *Serratia marcescens* (ATCC 13880).
DNA Manipulation

Protocols to extract bacterial chromosomal DNA follow those described elsewhere (Sawyer et al. 1987; DuBose et al. 1988). Southern blotting, isolation of DNA fragments from agarose, and subcloning were performed according to a method described by Sawyer et al. (1987). The phoA DNA used as a probe in the Southern blotting experiments has been described by DuBose et al. (1988). Probe DNAs from trpR and gap were provided by Jeff Lawrence and Howard Ochman, respectively. For sequencing, the following constructs were created: (1) a 2.4-kb SalI-PstI chromosomal DNA fragment from *E. fergusonii* (ATCC 35469) containing the 5' half of phoA and associated flanking region subcloned into SalI-PstI-digested M13mp18, (2) a 2.0-kb PstI chromosomal DNA fragment from *E. fergusonii* containing the 3' half of phoA and associated flanking region subcloned into PstI-digested M13mp19, and (3) a 2.7-kb SalI-EcoRI chromosomal DNA fragment from *Serratia marcescens* containing the entire phoA gene subcloned into SalI-EcoRI-digested M13mp18. The polymerase chain reaction (PCR) was used to amplify the region from bp 2429 to bp 3017 in the *E. fergusonii* sequence from the four additional strains described above (ATCC 35470–35473), according to the method of Saiki et al. (1988). The primers used were (+2429) CGCCTCACCCGAAATATC and (−3017) CACTCAAAGGGGCAACTG. The plus sign (+) denotes the forward primer, and the minus sign (−) denotes the reverse primer. Single-stranded DNA preparation and sequencing was performed according to a method described by DuBose et al. (1988). PCR products were purified for double-stranded sequencing by using Sepharose CL-6B spun columns and were sequenced according to a method described by DuBose and Hartl (1990). Where necessary, PCR primers and additional sequencing primers were synthesized on an Applied Biosystems oligonucleotide synthesizer.

Site-directed Mutagenesis

Oligonucleotide-directed mutagenesis was performed according to a method described by DuBose and Hartl (1989), by using the oligonucleotide primers shown in table 1. Mutants were verified by DNA sequencing to ensure that no additional secondary changes had occurred. Sequences used in structural replacement experiments were chosen so as to keep the total length of the altered structure the same as that of the wild-type sequence.

Protein Preparation and Enzymatic Assays

Alkaline phosphatase was prepared according to a method described by Reid and Wilson (1971) and Dykhuizen et al. (1985). The Michaelis constant ($K_m$) and turnover rate ($K_{cat}$) were obtained by monitoring the accumulation of paranitrophenol at 37°C in 0.6 M Tris, pH 8.2, by using a Beckman DU-7 spectrophotometer. The $K_m$ was estimated by a regression of $v$ against $[S]$, and the $K_{cat}$ was estimated by a regression of $[S]/v$ against $[S]$ (Segel 1976). Protein concentrations were determined by the method of Bradford (1976).

Computer Analysis

DNA sequences were compiled into a computer data base by using a Science Accessories Corporation GP-8 sonic digitizer (Southport, Conn.) and Microgenie software (Beckman, Palo Alto, Calif.) and were collated by using the Staden DNA sequence
Table 1
Oligonucleotide Primers Used in Site-directed Mutagenesis

<table>
<thead>
<tr>
<th>Mutation</th>
<th>5' Sequence 3*</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 → 1.A</td>
<td>TCGGCGAAAATGCCTCACGTCGCGCCATGGAATGTTCGCGGTCATGAGACGCTCAGG</td>
</tr>
<tr>
<td>7 → 1.B</td>
<td>TGCTACGCCTCGAGGCGCAATCTTGGAATTCCTCAGGTTACGGCTGGAAGG</td>
</tr>
<tr>
<td>7 → 1.AB</td>
<td>TGCTACGCCTCGAGGCGCAATCTTGGAATTCCTCAGGTTACGGCTGGAAGG</td>
</tr>
<tr>
<td>7 → T4.A</td>
<td>TCGGCGAAAATGCCTACGTCGCGCCATGGAATGTTCGCGGTCATGAGACGCTCAGG</td>
</tr>
<tr>
<td>7 → T4.B</td>
<td>TACGGCTCCGAGGCGCAATCTTGGAATTCCTCAGGTTACGGCTGGAAGG</td>
</tr>
<tr>
<td>7 → T4.AB</td>
<td>TACGGCTCCGAGGCGCAATCTTGGAATTCCTCAGGTTACGGCTGGAAGG</td>
</tr>
<tr>
<td>4 → 7.A</td>
<td>TACGGCTCCGAGGCGCAATCTTGGAATTCCTCAGGTTACGGCTGGAAGG</td>
</tr>
<tr>
<td>4 → 7.B</td>
<td>TACGGCTCCGAGGCGCAATCTTGGAATTCCTCAGGTTACGGCTGGAAGG</td>
</tr>
<tr>
<td>4 → 7.AB</td>
<td>TACGGCTCCGAGGCGCAATCTTGGAATTCCTCAGGTTACGGCTGGAAGG</td>
</tr>
<tr>
<td>2 → a1.A</td>
<td>GCCGATGGATGGGGAAATTGGTTAAATAATCGACGGATTTATGCC</td>
</tr>
<tr>
<td>2 → a1.B</td>
<td>GCCGATGGATGGGGAAATTGGTTAAATAATCGACGGATTTATGCC</td>
</tr>
<tr>
<td>2 → a1.AB</td>
<td>GCCGATGGATGGGGAAATTGGTTAAATAATCGACGGATTTATGCC</td>
</tr>
<tr>
<td>EcB → Ps1.A</td>
<td>TATAGATGCTCCACCGAGGCGCCGATACATCGACTAAG</td>
</tr>
<tr>
<td>EcB → Ps1.B</td>
<td>TATAGATGCTCCACCGAGGCGCCGATACATCGACTAAG</td>
</tr>
<tr>
<td>EcB → Ps1.AB</td>
<td>TATAGATGCTCCACCGAGGCGCCGATACATCGACTAAG</td>
</tr>
<tr>
<td>EcH → Ps5.A</td>
<td>GCTAAAAAGGAGGTACCTCGCGCATGTAGTTAGTACCCCT</td>
</tr>
<tr>
<td>EcH → Ps5.B</td>
<td>GCTAAAAAGGAGGTACCTCGCGCATGTAGTTAGTACCCCT</td>
</tr>
<tr>
<td>EcH → Ps5.AB</td>
<td>GCTAAAAAGGAGGTACCTCGCGCATGTAGTTAGTACCCCT</td>
</tr>
<tr>
<td>A61G</td>
<td>GGAAATTACTGCACTGGAGCAAGGTGCGAGGG</td>
</tr>
<tr>
<td>A61G+A65G</td>
<td>GGAAATTACTGCACTGGAGCAAGGTGCGAGGG</td>
</tr>
<tr>
<td>Q852</td>
<td>ACGGCGCAATACACTCATAC</td>
</tr>
<tr>
<td>T107S+A108G</td>
<td>GCATTACACGCGCCGTCGCAACC</td>
</tr>
<tr>
<td>T174S</td>
<td>CGCACTCACGCGCCGTCGCAACC</td>
</tr>
<tr>
<td>L196V</td>
<td>GGAAATTACTGCACTGGAGCAAGGTGCGAGGG</td>
</tr>
<tr>
<td>A200P</td>
<td>AGCCTCGTCGCGCCGTCGCAACC</td>
</tr>
<tr>
<td>T203K</td>
<td>CGCACTCACGCGCCGTCGCAACC</td>
</tr>
<tr>
<td>A260P</td>
<td>GCCGCGATGGCTCCGAGGCGCAAT</td>
</tr>
<tr>
<td>H331P</td>
<td>AACGAGATGCTCCGAGGCGCAAT</td>
</tr>
<tr>
<td>N334D</td>
<td>ATGCTCGGCGATCTTGTGGG</td>
</tr>
<tr>
<td>A348P</td>
<td>TCGATGAATCGGTAACACG</td>
</tr>
</tbody>
</table>

* Underscored boldface letters denote altered bases. The nomenclature is explained in detail in table 7.

analysis package (Staden 1984a, 1984b, 1984c, 1984d, 1986) and programs available from the Wisconsin Genetics Computer Group. Computer analysis was performed on an IBM PC/XT, a Digital Equipment Corporation VAX 6220, and a Sun Microsystems SUN 3/60 workstation. To determine the significance of the alignment between the Serratia marcescens 5' and 3' phoA flanking sequences and those of E. coli, a modification of the method of Lipman et al. (1984) was used (also see Wilbur and Lipman 1983; Lipman and Pearson 1985). Sequences were aligned initially, and a score was obtained that describes the percentage of mismatch in the alignment. The E. coli sequence was then permuted, and a new alignment score was generated. This process was repeated 100,000 times to generate a distribution of alignment scores expected at random, and the observed score was compared with this distribution to determine significance. Since the random sequences would align usually from their first matching bases, it was necessary to retain unshuffled a small number of bases at the 3' or 5' end of the E. coli test sequence to force anchoring of the two sequences at their respective start or termination codons. Fifteen bases is sufficient to force alignment but is small enough that it does not artificially inflate the alignment scores.

To determine the significance of the distribution of protein polymorphisms among E. coli, E. fergusonii, and Serratia marcescens, a modification of the method described by DuBose et al. (1988) and Sawyer (1989) was used. Protein sequences were aligned,
and an observed score equal to the sum of the squares of the lengths of monomorphic residues between consecutive polymorphic sites was determined. The empirical significance of the observed value was expressed as the frequency of the number of randomly generated scores which exceeded the observed value (see fig. 8A for methodological details).

Estimates of the severity of amino acid replacements seen among the sequences were determined using the values of Grantham (1974) to describe the physicochemical distance between amino acid replacements. Histograms were produced by performing pairwise comparisons between the three species and recording the frequency of polymorphisms within distance classes grouped by 10s (e.g., 0–10, 11–20). Divergence at silent and replacement sites was corrected for multiple hits by using the DIVERGE program available in the Wisconsin Genetics Computer Group software package (also see Kimura 1980; Perler et al. 1980; Li et al. 1985). G+C-content graphs were produced by sliding a window of length 99 bp across the sequence and displaying the average value for the window as a point positioned at the center of the window.

**Results**

**Distribution of phoA within Enteric Bacteria**

As can be seen from figure 1, phoA does not hybridize to all of the bacterial species examined. In particular, it is absent from *Escherichia hermanii* and *E. vulneris*, and the probe only weakly hybridizes to *Salmonella typhimurium*. Although the signal from *E. blattae* appears strong, the *E. coli* probe also contains DNA from a gene flanking phoA. Other hybridization experiments show that the signal from phoA alone from *E. blattae* is no stronger than that from *Salmonella* (data not shown). To control for the possibility of extreme divergence in the two *Escherichia* species in general, duplicate filters were prepared and hybridized with labeled DNA from the genes encoding the tryptophan operon aporepressor (*trpR*) and glyceraldehyde 3-phosphate dehydrogenase (*gap*), also shown in figure 2. Since the divergence between the *Escherichia* species and *Salmonella* has been estimated by DNA-DNA hybridization as being on the order of 50%–80% identity (Brenner et al. 1969), and since the divergence

![Fig. 1.—Southern hybridizations. From left to right, the panels show the hybridization signal for phoA, trpR, and gap, respectively. In each gel, lanes 1–10 correspond to *Escherichia coli*, Shigella dysenteriae, *S. flexneri*, *S. sonnei*, *E. fergusonii*, *Salmonella typhimurium*, *E. hermanii*, *E. vulneris*, *E. blattae*, and *Serratia marcescens.*](image-url)
among the various *Escherichia* species has been estimated as being 70%–80% (Brenner et al. 1972; Farmer et al. 1985a, 1985b), we can conclude that the lack of signal is not due to overall extreme divergence of the two *Escherichia* species, as *trp* and *gap* give strong hybridization signals.

DNA Sequence Comparisons of Alkaline Phosphatase within Enteric Bacteria

The DNA-aligned sequence of the *phoA* region from *E. coli*, *E. fergusonii*, and *Serratia marcescens* is shown in figure 2 (the sequences from the *Shigella* strains are to be presented elsewhere). In comparison with figure 1, the hybridization signal seen for *Serratia* corresponds to $\sim 30\%$ overall divergence at the DNA level, while that for *E. fergusonii* corresponds to $20\%$. The pairwise percentage divergences for the three sequences are shown in table 2. The values in table 2 represent divergence within *phoA* only and do not reflect the divergence of the flanking regions, as it was not possible to accurately align the *Serratia* sequence with the others, except for certain landmark features (see below). *Escherichia coli* and *E. fergusonii* are more closely related to each other both by sequence divergence and nucleotide composition than either is to *Serratia*; however, they are not equally divergent from *Serratia* (see table 2).

1. "Silent" Sites as Saturated and Replacement Sites as Unsaturated

From the values in table 2 it can be seen that divergence at synonymous sites is essentially saturated (corrected values are all >100%), while that at replacement sites is not. In addition, when either synonymous or replacement substitutions are considered, *E. fergusonii* is more divergent from *Serratia marcescens* than is *E. coli*, suggesting the possibility of unequal rates of evolution along the *Escherichia* lineage. The values range from a deficiency of 10%–20% with respect to replacement sites to >200% for corrected synonymous sites.

2. Flanking Sequences as More Divergent than Coding Regions

In the comparisons involving *Serratia marcescens* the similarity (and possibly homology) ends abruptly at the 5' and 3' ends of the coding sequence. The divergence of the intergenic region is less between *E. coli* and *E. fergusonii*; however, two large insertion/deletion events have occurred in the region downstream of *phoA*. Also, there are two open reading frames (ORFs) downstream of *phoA* in *E. coli*. Neither sequence is present close to *phoA* in *Serratia*, and only the first ORF (proposed to encode the phosphate-inducible gene *psiF*; see Chang et al. 1986) is present intact in *E. fergusonii*. The second ORF (only a fragment of which has been sequenced in *E. coli*) has become a pseudogene in *E. fergusonii* and requires a number of insertion/deletion events to align optimally in the two *Escherichia*. To verify that this situation was not unique to the particular strain of *E. fergusonii* used, four other isolates of the bacterium were obtained, and the region from bp 2429 to bp 3017 (encompassing the second ORF) was amplified by PCR and sequenced (DuBose and Hartl 1990). All five strains possessed identical sequences in the region, indicating that they are all isolates of one clonal group (*E. fergusonii* was isolated as a pathogen; see Farmer et al. 1985b).

Although the 5' flanking region cannot be accurately aligned with respect to *Serratia marcescens*, certain regulatory elements of *phoA* that are common to all three sequences serve as landmarks. The promoter elements at the Shine-Dalgarno sequence and the $-10$ region (shown in boldface in fig. 2) are conserved in all three species, as
is the "pho box," a regulatory consensus sequence seen in many genes involved in phosphate metabolism in *E. coli* (Shinigawa et al. 1987). Although the overall divergence in the 5' region of *Serratia* is extreme with respect to *E. coli*, the *Serratia* gene is still capable of being regulated by the *E. coli* phosphate regulon (Signer et al. 1961; Bhatti 1973, 1975; Bhatti and Done 1973), presumably because of the presence of the conserved pho box. An additional structural feature conserved among the three species is the transcription termination region beyond the 3' end of the gene. In *Serratia* the A+T-rich stem/loop structure seen in *E. coli* and *E. fergusonii* has been replaced with a similar stem/loop structure; however, the composition is now almost entirely G+C (reflecting the overall higher G+C content of the *Serratia* genome).

3. *G+C Content of Gene as Varying between Species*

The percentage G+C content of *phoA* from each of the three species is shown in table 3. The elevated level for *Serratia* (~60% on average) reflects the higher G+C content (57%-60%; see Grimont et al. 1977; Grimont and Grimont 1978, 1984) of the species in general. However, the G+C content is not uniform across the entire 2.7-kb sequenced region. In particular, the G+C content of the *Serratia* sequence in the regions 5' and 3' to the gene is much lower than that in the coding region. Figure 3 shows a graph of the local G+C content across the sequence. The G+C content of the 3' flanking sequence in *Serratia* (the second dashed line) is much lower than that in the coding region (45% for the flanking sequence vs. 65% for the coding sequence), as has been observed with other *Serratia* genes sequenced (Sharp, accepted). The promoter region and pho box for each species is very A+T rich (see section marked by the asterisk in fig. 3).

Protein Sequence Comparisons of Alkaline Phosphatase within Enteric Bacteria

The aligned amino acid sequence for each of the three species is shown in figure 4, along with the secondary-structure assignments of Sowadski et al. (1985). The *E. coli* signal sequence for export of the protein into the periplasm is also indicated, along with the cleavage site and the start of the mature peptide. Although the *Serratia* protein sequence is 24% divergent from the *E. coli* enzyme, it is still capable of forming a functional hybrid dimer with the *E. coli* protein of intermediate electrophoretic mobility (Signer et al. 1961).

1. *Divergence within Signal Sequence as Higher than That in the Remainder of the Protein*

The pairwise percentage divergences for each of the three enzymes are shown in table 4. These data represent only those residues from pro-19 to the end of the protein, as it is not possible to accurately align the amino terminal portion and signal sequence from *Serratia marcescens* (see fig. 4). The primary sequence of the signal region in *phoA* shows the high divergence characteristic of such regions in general (von Heijne 1985). However, while the primary sequence is not conserved, the overall hydrophobic makeup is. In the signal sequence of all three species, both the general structure and the splice-site junction between the nascent form and the mature peptide are preserved (fig. 5). Kendall et al. (1986) have shown that the entire signal sequence can be "idealized" by replacement with highly hydrophobic residues (e.g., leucine) yet still direct export of the protein to the periplasm. Using *phoA* fusion proteins, other workers have demonstrated that the amino terminal portion of the mature peptide can tolerate
FIG. 2.—Aligned phoA DNA sequence. Boldface nucleotides in the 5' end of the sequence denote the conserved promoter elements (the pho box and the Shine-Dalgarno sequence). The AUG start codon (GUG in the *Escherichia coli* and *E. fergusonii* sequence) and the UAA termination codon are in boldface type and denoted by an asterisk.

In the 5' end of the sequence, gaps have been inserted only to align the pho box and the other regulatory features. Otherwise, the alignment between the *Serratia* sequence and the two *Escherichia* species is no better than that expected at random. The alignment in the signal sequence region is that alignment produced by multiple iterations of the alignment algorithm and does not necessarily reflect the evolutionary history of the sequences. The boldface bases at the 3' end of the sequence denote the inverted repeats of the *E. coli* transcription terminator. As with the 5' end, the 3' alignment between the *Serratia* sequence and the *Escherichia* species is no better than that expected at random. Gaps have been inserted in the 3' end to improve the alignment between *E. coli* and *E. fergusonii*. 

---

**EF 1092**

---

**EC 829**

---

**SM 1201**

---

**EF 1212**

---

**EC 949**

---

**SM 1321**

---

**EF 1332**

---

**EC 1389**

---

**SM 1681**

---

**EF 1452**

---

**EC 1549**

---

**SM 1801**

---

**EF 1572**

---

**EC 1651**

---

**SM 1861**

---

**EF 1692**

---

**EC 1699**

---

**SM 1871**

---

**EF 1812**

---

**EC 1932**

---

**SM 2041**

---
### Table 2
Nucleotide Sequence Differences (%)

<table>
<thead>
<tr>
<th>A. Total&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Escherichia fergusonii</th>
<th>Serratia marcescens</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli . . . . . . .</td>
<td>20.9</td>
<td>27.3</td>
</tr>
<tr>
<td>E. fergusonii . . . .</td>
<td>31.9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. Silent and Replacement&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Escherichia coli</th>
<th>Escherichia fergusonii</th>
<th>Serratia marcescens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent:</td>
<td>1   2    3</td>
<td>1   2    3</td>
<td>1   2    3</td>
</tr>
<tr>
<td>E. coli . . . . . . .</td>
<td>32.2 36.4 64.0</td>
<td>38.6 62.1 55.8</td>
<td></td>
</tr>
<tr>
<td>E. fergusonii . . . .</td>
<td>155.0 78.8 143.9</td>
<td>46.9 46.7 69.5</td>
<td></td>
</tr>
<tr>
<td>S. marcescens . . . .</td>
<td>221.2 267.4 102.4</td>
<td>418.5 120.4 196.5</td>
<td></td>
</tr>
<tr>
<td>Replacement:</td>
<td>1   2    3</td>
<td>1   2    3</td>
<td>1   2    3</td>
</tr>
<tr>
<td>E. coli . . . . . . .</td>
<td>6.1   5.8    4.3</td>
<td>17.2 11.4 12.9</td>
<td></td>
</tr>
<tr>
<td>E. fergusonii . . . .</td>
<td>19.4   9.1    4.4</td>
<td>20.0 13.8 13.7</td>
<td></td>
</tr>
<tr>
<td>S. marcescens . . . .</td>
<td>63.4   18.8   14.1</td>
<td>76.6 23.1 15.1</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Uncorrected pairwise percentage differences between *phoA* among the three bacterial species. The values represent data from amino acid residue pro-19 to the end of the gene only and do not reflect the divergence of the flanking regions.  
<sup>b</sup> Values above the diagonal denote the actual percentage difference, and those below the diagonal denote the percentage difference as corrected by the DIVERGE program of the Wisconsin Genetics Computer Group program (see Perler et al. 1980). The numbers 1, 2, and 3 refer to the number of potential base substitutions for silent or replacement changes.

Large insertions (150 amino acids) and still produce a functional enzyme (Hoffman and Wright 1985). From the aligned sequences in figure 4, it would appear that the essential region of the protein begins with pro-19, as the similarity increases dramatically at that point.

#### 2. Absence of Nonrandom Distribution of Amino Acid Replacements across the Protein

Figure 6 shows the location of residues polymorphic among the three species. The alignment shown in figure 6 runs from leu-7 to lys-449 (the carboxyl terminus), although, because of the high divergence in the signal sequence and amino terminal residues, only those residues from pro-19 to lys-449 will be used in the subsequent analysis. The other residues are included here to show the increased divergence in the amino terminal region of the protein. Residues polymorphic between each of the three pairs of species are denoted by different symbols (E, F, S, and X). The secondary-structure assignment for each residue according to Sowadski et al. (1985) is also indicated. We will consider first the spatial distribution of substitutions in all of the three species and will consider the individual pairwise comparisons later. Two statistical tests were employed to judge the significance of the observed distributions of polymorphisms. The first is a modification of the occupancy test of Feller (1968; modified in Stephens 1985). The relevant values are $n$, the total length of the sequence; $s$, the number of polymorphic positions; $d_o$, the size of the largest span end-to-end of residues which are polymorphic; and $g_o$, the size of the largest gap between successive polymorphisms. The test statistics are $P_d$, the probability that an observed value $d_o$ is less...
Table 3
G+C Content (%)  

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>5'</th>
<th>Coding</th>
<th>3'</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli</em></td>
<td>51.6</td>
<td>42.5</td>
<td>53.7</td>
<td>51.5</td>
</tr>
<tr>
<td><em>Escherichia fergusonii</em></td>
<td>47.4</td>
<td>39.5</td>
<td>49.2</td>
<td>48.8</td>
</tr>
<tr>
<td><em>Serratia marcescens</em></td>
<td>59.6</td>
<td>61.7</td>
<td>64.8</td>
<td>44.5</td>
</tr>
</tbody>
</table>

Note.—Total, 5', coding, and 3' refer to the entire sequenced region, noncoding DNA upstream from the initiation codon of *phoA*, the coding region of *phoA*, and noncoding DNA downstream from the termination codon of *phoA*, respectively.

than or equal to that expected by chance (given the length of the sequence and the number of polymorphic sites), and $P_g$, the probability that the largest gap is greater than that expected at random. For the protein sequences here, $n = 431$, $s = 96$, $d_o = 422$, and $g_o = 35$. From these numbers, $P_d = 0.36$, and $P_g = 0.0089$. This indicates that, although the polymorphisms are not clustered in one region of the protein in general ($P_d$ is not significant), there is one large gap (marked by an asterisk in fig. 6) that is statistically significant. The next largest gap length ($g_o = 25$) is not significant ($P_g = 0.13$). Therefore, by one of Stephens’s two criteria the polymorphisms within one region of the protein are significantly clustered; the gap marked by the asterisk in figure 6 is longer than that expected at random. In this context, however, this test suffers the limitation that it yields only the probability that any individual gap is significantly longer than that expected by chance. It is possible to have a large number of gaps in the sequence, each of which is nonsignificant, yet the total combination of so many such gaps could still be quite improbable. To address this question, a second statistical test is used, modified from that of Sawyer (1989; also see DuBose et al. 1988). The test is based on the method in figure 7A. For the total polymorphisms across the three species, $P = 0.00072$ ($72/100,000$ had a score greater than that expected). Thus, in addition to having one gap which is significantly long in and of itself, the total distribution of gap lengths is also highly significant (see fig. 7B), indicating either that there are a large number of regions of the protein that accept an unusually high number of replacements or that there are regions constrained against accepting substitutions.

3. Beta Sheets, Alpha Helices, and Unassigned Structures as Accepting Replacements at Differing Frequencies

Figure 6 shows that amino acid replacements in the protein occur primarily within alpha helices or unassigned structural regions and are generally deficient in beta sheets. To quantify this, a $\chi^2$ test of the observed distribution of polymorphic residues with respect to secondary structure was performed (again by using only those residues from pro-19 to the end). For the three species together, 30 replacements are observed to occur in alpha helix, three in beta sheet, and 63 in neither. For each of these classes the expected values (calculated from the proportion of residues constituting each class) are 27.6, 13.4, and 55.0, respectively. The $\chi^2$ value for this is 9.4, corresponding to $P < 0.01$, with 2 degrees of freedom. Similar calculations were performed by classifying the residues either as beta sheet or non-beta sheet. Here the observed values are 93 and 3, and the expected values are 82.6 and 13.4. The $\chi^2$ (with 1 degree of freedom) value equals 9.3, which is also significant ($P < 0.005$). If the beta-sheet
regions are excluded from the analysis and only the alpha-helix and unassigned regions are considered, the resulting $\chi^2$ value for replacements between the two is nonsignificant, indicating that it is the beta sheets that are deficient; by this test, polymorphisms are occurring between alpha-helix and unassigned regions essentially at random (however, see below).

The relative lack of replacements within beta sheets accounts for a large proportion of the clustering but not for all of it. If those residues corresponding to the beta sheets are removed, and if the simulations are performed again, the result still provides significant evidence for clustering (for 370 remaining residues with 93 polymorphisms, $P = 0.008$). Note, however, that there are now about 10-fold more simulated scores exceeding the observed value than were previously obtained when the beta sheets were included (803 vs. 72 simulations of 100,000 exceeded the observed score). If those residues which correspond to alpha helices are also removed, the sum-of-squares score is no longer significant ($P = 0.18$, for 246 residues with 63 polymorphisms).

Finally, figure 6 shows that there are three areas of unassigned structure which
have a number of consecutive variable residues, in particular the regions ile-124 to pro-130, thr-210 to ala-217, and thr-287 to val-296. When these are compared with the tertiary-structure diagram, two of them (ile-124 and thr-210) correspond to regions that are exposed to solvent. In particular, the run of seven consecutive polymorphic residues beginning at ile-124 forms a loop extending away from the main body of the enzyme.

Fig. 4.—Aligned alkaline phosphatase protein sequence. Dots denote residues identical to the *Escherichia coli* sequence, while dashes denote gaps introduced during alignment. Alpha helices and beta sheets are shaded. Residues constituting the signal sequence are also shaded. The start of the mature peptide (N-terminal threonine) is denoted by a vertical arrow above the sequence. The dagger denotes residue pro-19, the cutoff point for the clustering analysis described in the text. Residues are numbered according to the protocol of Sowadski et al. (1985).
Table 4
Divergence and Diversity of Amino Acid Replacements

<table>
<thead>
<tr>
<th></th>
<th>Escherichia fergusonii</th>
<th>Serratia marcescens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein sequence divergence (%):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>7.4</td>
<td>19.7</td>
</tr>
<tr>
<td>E. fergusonii</td>
<td>20.4</td>
<td></td>
</tr>
<tr>
<td>Average severity of amino acid replacement:*</td>
<td>63.0</td>
<td>66.4</td>
</tr>
<tr>
<td>E. coli</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. fergusonii</td>
<td>67.2</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—All data are from residue pro-19 to the end of the protein.
* Data were tabulated by using the values of Grantham (1974) to express pairwise amino acid distances. Average severity of replacement for two random sequences is 98.9, with a standard deviation of 46.9.

4. Alpha Helices as Not Showing Equal Acceptance of Replacements across Their Lengths

Although the helical regions of alkaline phosphatase are less constrained against accepting replacements than are beta sheets, they do not uniformly do so across their lengths. To demonstrate this, the 14 helices were aligned at their respective amino (N) and carboxyl (C) termini, as shown in table 5. By means of the methods of Presta and Rose (1988) and Richardson and Richardson (1988), residues were then classified as being upstream or downstream from each terminus (upstream N', the N terminus itself, and downstream N1 and N2—and comparably for C2, C1, C, and C') and as “middle” (M). As the lengths of the helices varied, the M class represents residues from N3, . . . , C3 grouped together. For each class, the number of helices that contain a polymorphism in any of the three species is totaled, and this total is compared with that expected at random, given the total number of polymorphisms and the frequencies of residues belonging to each class. The histogram constructed from the data in table 5 is shown in figure 8. The excess of polymorphisms at the carboxyl end was verified by a $\chi^2$ test. When positions N', N, and N1 are classified as “amino,” N2, M, and C2 as M, and C1, C, and C' as “carboxyl,” there are six amino, 11 M, and 16 carboxyl positions polymorphic, compared with the 9.24, 14.52, and 9.24 values expected at random. The $\chi^2$ value (with 2 degrees of freedom) equals 6.94, which is significant at the 5% level ($P < 0.05$). If the classification is considered as carboxyl versus the rest (i.e., amino and M grouped together), the values become 16 and 17 versus 9.24 and 23.76, for the observed and expected values, respectively. The $\chi^2$ value (with 1 degree of freedom) equals 6.87, which is significant at the 1% level ($P < 0.01$). Thus, in addition to a deficiency of replacements within beta sheets, across alpha helices there is an excess of replacements at the carboxyl end, compared with the rest of the helix.

5. Pairwise Sequence Comparisons

The results to this point have concerned the distribution of replacements across all three sequences. If the sequences are broken down into pairwise comparisons, the significant results still hold, except with respect to the E. coli/E. fergusonii comparison. This is due primarily to the lower level of divergence between the two species. As the percentage divergence decreases, the sensitivity of each test also decreases. Despite the fact that they are no longer statistically significant, the same general pattern of having
Fig. 5.—Signal sequences. The panels depict the alkaline phosphatase signal sequence in the form of an alpha helix, bottom→top corresponding to amino→carboxyl. Strongly hydrophobic residues are shaded. The panel on the left is from *Escherichia coli*, that in the middle is from *E. fergusonii*, and that on the right is from *Serratia marcescens*. The sequences have been rotated to bring the hydrophobic core to the same relative position in each.

fewer replacements within beta sheets as compared with the value expected at random holds. For example, the alpha-helix-versus-beta-sheet comparison for *E. coli* versus *E. fergusonii* generates a $\chi^2$ value which is borderline significant at the 10% level with 1 degree of freedom. Because the percentage divergence of the two *Escherichia* species is low (32 replacements in the region compared, as compared with >90 for the comparisons with *Serratia*), the deviations would have to be extreme to be detectable. The values for the gap permutation test are also nonsignificant within *Escherichia* ($P = 0.26$). However, as with the $\chi^2$ tests, the observed value does lie in the same direction as the tests which were significant (see fig. 7B).

6. Amino Acid Replacements within the Three Species as Being Primarily Conservative

Grantham (1974) has presented a table of pairwise distance values for amino acid replacements. The values were determined by three objective criteria comparing only the composition, volume, and hydrophobicity (polarity) of each amino acid residue, without respect to observed replacement frequencies in proteins. Grantham was able to show that the distance values correlated well with amino acid replacement frequencies (the RSF value of McLachlan 1971a, 1971b). Miyata et al. (1979) used two of the three criteria (volume and polarity) to show that, for proteins in which the likelihood of multiple replacements at a residue is small (i.e., replacements do not involve amino acids that are two mutational steps away from each other by the genetic code), the PAM value (Dayhoff et al. 1972, 1978) also correlates quite well with the amino acid distance measures. That is, physically similar residues interchange more
FIG. 6.—Spatial distribution of amino acid replacements. Dashes denote monomorphic residues; E, F, and S denote residues where differences are unique to *Escherichia coli*, *E. fergusonii*, and *Serratia marcescens*, respectively. X denotes those residues different in all three sequences. Only those residues from pro-19 (denoted by the arrow) were used in the clustering analysis. The gap denoted by the asterisk is significantly long by Stephens's (1985) gap test. Alpha helices are shaded lightly (a), and beta sheets are shaded heavily (■).

frequently than dissimilar ones. Figure 9 shows a histogram of the frequency of replacement by amino acid distance for alkaline phosphatase for each of the three pairwise comparisons, and table 4 shows the mean distance values for each comparison. As with the substitution distribution data, only residues from pro-19 to the end were used, to avoid artificially inflating the scores. The mean substitution value does increase slightly as the divergence of the proteins compared increases, indicating that with increased evolutionary distance more nonconservative replacements are accepted (Lesk and Chothia 1980, 1982; Chothia and Lesk 1982, 1985, 1987). As shown in figure 10, the histograms show a skewed distribution, with peaks near 30, 60, and 90. A distance score ≈30 on Grantham's scale corresponds to relatively conservative replacements (e.g., arginine for lysine).

Site-directed Mutagenesis of *E. coli* Alkaline Phosphatase

To correlate sequence variation with possible functional effects, we have employed site-directed mutagenesis to introduce single and multiple mutations in the *phoA* gene of *E. coli*. Using the strategy shown in figure 10, we have targeted specific structures (alpha helices and beta sheets) for replacement. In brief, for each of three separate helices and two beta sheets, we have made multiple mutations that replace the wild-type sequences with amino acid residues that encode a similar structure from another source. For each helix or beta sheet, we have created three replacement mutations: the first (A) replaces the amino portion of the structure, the second (B) replaces the carboxyl portion of the structure, and the last (AB) creates a double mutant with both halves replaced. The replacement sequences and their associated activities are shown in tables 6 and 7, along with selected point mutations also created in *phoA*. A detailed analysis of the alpha-helix replacements has been described elsewhere (DuBose and Hartl 1989). However, there are several results which should be noted: (1) *There is no tendency to recover only mutations with either near-wild-type or near-zero activity.* In addition to the mutations described in table 7, we have created >20 additional point mutations at random locations scattered across the gene (data not shown). Ranking these with respect to their activities yields a continuous distribution. (2) *Replacements involving beta sheets are much more severe (in terms of activity) than those involving alpha helices.* Of the multiple mutations in helices, only those involving helix 2 do not yield an active enzyme. In contrast, all but one of the beta-sheet replacements are inactive (*EcB*→*Ps1.B*; see fig. 10 and tables 6 and 7). Table 7 also
A

1. Align sequences
   _________________X______________
   _________________X______________
   _________________X______________

2. Identify polymorphic positions

3. Record gap lengths, calculate sum-of-squares
   \[ 3^2 + 4^2 + 3^2 + 9^2 + 3^2 + 2^2 + 7^2 = 162 \]

4. Scramble columns in sequences
   _________________X______________
   _________________X______________
   _________________X______________

5. Calculate a new sum-of-squares score
   \[ 10^2 + 0^2 + 2^2 + 8^2 + 4^2 + 1^2 + 3^2 = 194 \]

Fig. 7.—Computer simulation strategy. The diagram shows the steps involved in the permutation method. Steps 4 and 5 are repeated 100,000 times, and the \( P \) value for the observed distribution is expressed as the frequency of the simulations in which the randomized sum-of-squares score exceeded the observed value. B. Simulated distributions of monomorphic gap lengths. Both panels show the distribution of the sum of the gap lengths squared from scrambling the polymorphic sites in alkaline phosphatase 100,000 times. The top panel represents the distribution from the combined polymorphisms within all three species, and the bottom panel represents that from the \( \text{Escherichia coli/E. fergusonii} \) comparison only. The vertical line denotes the observed sum-of-squares value.

shows selected point mutations introduced into \( \text{phoA} \), classified as being within or near helices or sheets. With one exception, the activities of point mutations near alpha helices are greater than those near beta sheets. The activities of the helix point mutations are 56%–108% (average 86%) of those of the wild type, whereas those near beta sheets are 0%–69% (average 29%) of those of the wild type. This distinction is further emphasized by the fact that two of the point mutants are actually combinations of two mutated residues (A61G+A65G and T107S+A108G), which nonetheless retain a significant fraction of enzymatic activity. (3) For each helix mutational class (A, B, and AB), there is a consistent order of activities: 7→T4 > 7→T4 > 2→7 > 2→8.A. This reflects in part the assumption that the different structures should be more or less constrained by their relative positions within the tertiary structure. Recall that helix 7 is on the periphery of the enzyme, that helix 4 is adjacent to the active site, and that helix 2 lies on the monomer/monomer interface. (4) Within each helix, the replacement mutation of the amino half (A) possesses less activity than that of the carboxyl half (B). This pattern is maintained despite the fact that the multiple replacements in different helices are not equally conservative. For example, the 4→7.B mutant is a much more drastic replacement in the helix (two charged residues replace a neutral alanine and a bulky aromatic tryptophan which is buried within the protein) than is the 4→7.A mutation, yet it still retains a much higher activity. In addition, the one beta-sheet replacement that retained activity was a “B” mutant as well. (5) There is a nonadditive relationship between the activities of the amino, carboxyl, and double mutants. This also affects the \( K_m \) and \( K_{cat} \) of the enzyme differently. When the \( K_m \) is examined, the effect of combining the amino and carboxyl mutations is seen to be dominant: the \( K_m \) of the double mutant is not significantly different from that of the least active single mutant. The \( K_{cat} \) does not behave so nicely, however. Here the activity of the double mutant is neither dominant nor the product of the activities of
FIG. 7 (Continued)

the respective single mutants. Only with helix 4 does it appear that the single mutants
may be combining in an additive fashion.

Discussion

Sequence analysis indicates that several levels of constraint affect the evolution
of alkaline phosphatase. We have used site-directed mutagenesis to test the inferred
constraints directly. The apparent paucity of replacements within beta sheets relative
to other areas of the protein suggests that those regions should be less tolerant of amino
acid replacement than are alpha helices or unassigned structures. This was verified in
two ways. First, the introduction of consecutive structure-preserving mutations dem-
strates that such changes within alpha helices are accepted much more readily than
similar mutations within beta sheets: nine of the 12 helical replacement mutations
retained activity, whereas only one of the six beta-sheet mutations did (see table 7).
Second, point mutations near alpha helices have higher activities than do similar
mutations near beta sheets. This high-versus-low activity difference does not arise
merely from an all-or-none response of the enzyme to replacements. Ranking all the
mutations created (table 7 and data not shown) with respect to their activities yields
a continuous distribution—there is no tendency to recover only mutations with either
low (e.g., <1%) or high wild-type activity. The fact that residues in one strand
which compose part of a beta sheet must interact with residues in other such strands
provides one physical explanation for the differences seen between alpha helices and
beta sheets. To restore stability in the structure, a mutation in any position in a beta
sheet would need to be accompanied by a complementary mutation elsewhere (Lesk
and Chothia 1980; also see Wyckoff 1987 for an example of such an occurrence within
alkaline phosphatase). In contrast, while such a complementary mutation could relieve
the stress created by a replacement within an alpha helix (as has been observed with
the cytochromes and hemoglobin; see Perutz 1983; Chothia and Lesk 1985), not all
replacements within alpha helices would require such changes. The stress created by
the new side chain may be negligible if it is on that side of the helix that faces the
solvent, or it may be accommodated by local conformational change of the residue
in question. Although our changes were designed to be structure preserving, the re-
placement of many consecutive residues within beta sheets would nonetheless require
the concomitant replacement of their complementary partners.
Although an examination of the locations of polymorphisms indicated that beta-sheet regions are generally deficient in polymorphisms, this deficiency does not totally account for the clustering scores observed in the simulation studies. In particular, if those residues corresponding to beta-sheet structure are removed, leaving only alpha-helix and unassigned structures, the score is still significant, but with a $P$ value 10-fold larger than that with the beta sheets included. It is not until the alpha-helix residues are also removed that the score is nonsignificant. The distinction between alpha helix and beta sheet is a convenient one (as they represent clearly discernible structures), but it is not necessarily the only functionally important one. In particular, most of the surface of the monomer/monomer interface is rigidly conserved. Helices 1 (gln-29), 2 (asp-55), 3 (ile-74), 13 (his-425), and 14 (thr-436)—along with beta strands J, B, and I—lie along the interface. Comparing this with the spatial distribution of polymorphisms in figure 6 shows that these regions have generally low replacement rates. In particular, the region encompassing helices 2 and 3 and sheet B contains only one replacement, and that from sheet I to the middle of helix 14 (encompassing helix 13 and sheet J) shows only two replacements. With regard to mutagenesis, the one helical replacement that did not function was helix 2, which is in this constrained region. In contrast, the analysis indicated polymorphic runs elsewhere, in particular the regions between residues ile-124 and pro-130 (adjacent to helix 5), between thr-210 and ala-217 (between sheet D and helix 9), and between thr-287 and val-296 (adjacent to helix 11). Although no mutagenesis data are available for these regions, the latter do correspond to areas of the protein that are relatively exposed to solvent.
In the protein sequence comparison shown in figures 4 and 8, it was noted that, compared with the carboxyl or M regions, the amino ends of alpha helices accepted fewer replacements than expected. In particular, >50% of the helices were polymorphic at their C-terminus position (see table 5); this was the highest value for any of the nine classes. Site-directed mutagenesis verified this, in that in each of the helical replacement mutations the activity of the amino-end mutant was at least sevenfold less than its corresponding carboxyl-end counterpart (see table 7). From a structural standpoint, perturbations at the amino end of a helix might be propagated down the length of the helix, potentially having a greater effect than those that begin midway.

Our observations of increased divergence in the signal sequence and in the first several amino acids of alkaline phosphatase supports the results of experiments by others. Signal sequences in general can be highly variable (von Heijne 1985); the principle structural feature required is that of a hydrophobic core flanked by a charged N-terminus and a polar C-terminus region. Kendall et al. (1986) created "idealized" alkaline phosphatase signal sequences, showing that the entire core region is replaceable or extendable as long as the replacement residues are highly hydrophobic. Indeed, if the Escherichia coli, E. fergusonii, and Serratia marcescens signal sequences are compared, it is evident that, while the primary sequences are divergent in both length and composition, the structural arrangement of a hydrophobic core flanked by charged residues is conserved (see fig. 5). With respect to the increased divergence within the first 20 amino acids of the mature peptide, a similar result was seen when phoA fusion proteins were used to study protein secretion (Hoffman and Wright 1985). In these experiments, phoA (minus the residues coding for its signal sequence) was fused to the promoter or amino terminus of another gene, and the resulting fusion protein was
Severity of amino acid replacement in *phoA*

Fig. 9.—Severity of amino acid replacements. The severity of amino acid replacement for each of the three pairwise comparisons—eco-fer (■ ■), eco-sma (▲ ▲), and fer-sma (● ●)—is shown. In each pairwise comparison, each amino acid replacement is assigned a replacement severity score according to the method of Grantham (1974). Severity scores were then grouped, by 10’s, into classes (plotted on the abscissa), and the frequency of scores in each severity class was plotted on the ordinant. Also plotted is the expected curve (● ●), which denotes the frequency of each severity class when calculated from all possible pairwise replacements of the 20 amino acids.
FIG. 10.—Site-directed mutagenesis strategy. The heavy line in the middle-left panel denotes the amino acid sequence of alkaline phosphatase. Alpha helices and beta sheets are shown as filled and open boxes, respectively. The two disulfide bridges are shown as lines connecting regions of the protein below the heavy line. The beta-sheet structures replaced by site-directed mutagenesis (EcB→Ps1 and EcH→Ps5) and the point mutations created are shown above the heavy line, and the alpha helices replaced (7→1, 7→T46, 4→7, and 2→α1) are shown below the heavy line. The arrows indicate the direction of the replacement. The inset depicts the mutagenesis strategy used to create the structural replacement mutations. The heavy line denotes the wild-type phoA DNA sequence, and the helical structure below it denotes the alpha helix that it encodes (the same strategy was employed for the beta sheets). In two separate rounds of mutagenesis (A and B), the amino and carboxyl ends of the structure (shown in boldface) were replaced by using oligonucleotides complementary to the region of interest (denoted by sawtooth lines above the heavy line). The double mutant (AB) is created by using the sequence of the altered amino mutant (A) as the template DNA in a third round of mutagenesis. References to bacteriophage T4 lysozyme, Pseudomonas aeruginosa azurin, and the synthetic peptide αpha1 can be found in reports by Matthews and Remington (1974), Canters (1987), and Regan and DeGrado (1988), respectively.
Table 6
Multiple Amino Acid Replacements Created by Mutagenesis

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Helix 7 wild type:</strong></td>
</tr>
<tr>
<td>→ phoA1.A</td>
</tr>
<tr>
<td>→ 3 phoA1.B</td>
</tr>
<tr>
<td>→ phoA1.AB</td>
</tr>
<tr>
<td>→ T4L6.A</td>
</tr>
<tr>
<td>→ T4L6.B</td>
</tr>
<tr>
<td>→ T4L6.AB</td>
</tr>
<tr>
<td><strong>Helix 4 wild type:</strong></td>
</tr>
<tr>
<td>→ phoA7.A</td>
</tr>
<tr>
<td>→ phoA7.B</td>
</tr>
<tr>
<td>→ phoA7.AB</td>
</tr>
<tr>
<td><strong>Helix 2 wild type:</strong></td>
</tr>
<tr>
<td>→ al.A</td>
</tr>
<tr>
<td>→ al.B</td>
</tr>
<tr>
<td>→ al.AB</td>
</tr>
<tr>
<td><strong>Sheet B wild type:</strong></td>
</tr>
<tr>
<td>→ Ps1.A</td>
</tr>
<tr>
<td>→ Ps1.B</td>
</tr>
<tr>
<td>→ Ps1.AB</td>
</tr>
<tr>
<td><strong>Sheet H wild type:</strong></td>
</tr>
<tr>
<td>→ Ps5.A</td>
</tr>
<tr>
<td>→ Ps5.B</td>
</tr>
<tr>
<td>→ Ps5.AB</td>
</tr>
</tbody>
</table>

**Note.**—Boldface residues denote the alpha helix or beta sheet replaced. Oligonucleotide primers used are described in Table 1. T4L6 refers to the sixth helix of T4 lysozyme (Matthews and Remington 1974), al refers to the synthetic peptide alpha1 (Regan and DeGrado 1988), and Ps1 and Ps5 refer to Pseudomonas aeruginosa azurin sheets 1 and 5, respectively (Canters 1987).

Examined for activity. Alkaline phosphatase is active only when localized in the periplasmic space (the area between the inner bacterial cell membrane and the cell wall of *E. coli*) and not when in the cytoplasm. That the enzyme is functional, then, implies that it was secreted into the periplasm, and therefore the foreign residues must contain a signal sequence for protein secretion. One such fusion resulted in the deletion of the first 13 amino acids from the mature peptide, another resulted in the deletion of the first 39 amino acids of the mature peptide, and a third resulted in the addition of 150 residues from β-lactamase. Both the fusion protein with the 13-residue deletion and that with the 150-residue insertion functioned at near-wild-type levels; however, the mutation with the large deletion was inactive. Deleting the first 39 residues would remove the first alpha helix, which comprises part of the monomer/monomer interface, and thus could be essential to dimerization and hence activity. On the basis of our sequence analysis, we would predict that the cutoff point should be at about 20 amino acids, near pro-19.

Finally, with respect to the DNA sequences, two observations are worth mentioning. The first is the conservation of an upstream regulatory sequence. A number of sequences for phosphate-regulated genes in *E. coli* have been obtained, and, on the basis of these, a common regulatory consensus sequence, the pho box, has been proposed (Shinigawa et al. 1987). The two additional sequences presented here also possess this consensus sequence, in approximately the same positions relative to the
### Table 7
Enzymatic Activity of Replacement and Point Mutations

#### A. Wild-Type *Escherichia coli*

<table>
<thead>
<tr>
<th>Mutant</th>
<th>( K_m ) (( \mu M ))^b</th>
<th>( K_{cat} ) (min(^{-1} ))</th>
<th>( K_{cat}/K_m )</th>
<th>Relative ( K_{cat}/K_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> K-12</td>
<td>36.2 ± 2.50</td>
<td>103. ± 3.00</td>
<td>2.85</td>
<td>1.00</td>
</tr>
</tbody>
</table>

#### B. Replacement Mutations

<table>
<thead>
<tr>
<th>Mutant</th>
<th>( K_m ) (( \mu M ))^b</th>
<th>( K_{cat} ) (min(^{-1} ))</th>
<th>( K_{cat}/K_m )</th>
<th>Relative ( K_{cat}/K_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 → 1A</td>
<td>14.1 ± 0.92</td>
<td>8.32 ± 0.10</td>
<td>0.59</td>
<td>0.21</td>
</tr>
<tr>
<td>7 → 1B</td>
<td>22.2 ± 0.58</td>
<td>90.1 ± 1.05</td>
<td>4.05</td>
<td>1.42</td>
</tr>
<tr>
<td>7 → 1AB</td>
<td>21.2 ± 1.71</td>
<td>73.5 ± 3.20</td>
<td>3.46</td>
<td>1.21</td>
</tr>
<tr>
<td>7 → T46A</td>
<td>40.9 ± 5.62</td>
<td>2.65 ± 0.12</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>7 → T46B</td>
<td>18.5 ± 0.96</td>
<td>65.2 ± 1.42</td>
<td>3.51</td>
<td>1.23</td>
</tr>
<tr>
<td>7 → T4AB</td>
<td>42.0 ± 3.74</td>
<td>8.90 ± 0.45</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>4 → 7A</td>
<td>29.0 ± 2.54</td>
<td>1.25 ± 0.06</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>4 → 7B</td>
<td>9.42 ± 0.59</td>
<td>26.1 ± 0.42</td>
<td>2.77</td>
<td>0.97</td>
</tr>
<tr>
<td>4 → 7AB</td>
<td>26.2 ± 1.08</td>
<td>0.42 ± 0.01</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>2 → a1A</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
<tr>
<td>2 → a1B</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
<tr>
<td>2 → a1AB</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
<tr>
<td>EcB → Ps1.A</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
<tr>
<td>EcB → Ps1.B</td>
<td>11.7 ± 1.18</td>
<td>6.53 ± 0.19</td>
<td>0.56</td>
<td>0.20</td>
</tr>
<tr>
<td>EcB → Ps1.AB</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
<tr>
<td>EcH → Ps5.A</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
<tr>
<td>EcH → Ps5.B</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
<tr>
<td>EcH → Ps5.AB</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
</tbody>
</table>

#### C. Point Mutations

<table>
<thead>
<tr>
<th>Mutant</th>
<th>( K_m ) (( \mu M ))^b</th>
<th>( K_{cat} ) (min(^{-1} ))</th>
<th>( K_{cat}/K_m )</th>
<th>Relative ( K_{cat}/K_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A61G a</td>
<td>19.5 ± 0.42</td>
<td>46.5 ± 0.51</td>
<td>2.38</td>
<td>0.84</td>
</tr>
<tr>
<td>A61G+ As5G a</td>
<td>12.5 ± 0.25</td>
<td>19.9 ± 0.17</td>
<td>1.59</td>
<td>0.56</td>
</tr>
<tr>
<td>T107S → A108G a</td>
<td>17.1 ± 1.12</td>
<td>36.2 ± 1.13</td>
<td>2.11</td>
<td>0.74</td>
</tr>
<tr>
<td>T174S a</td>
<td>23.5 ± 4.57</td>
<td>70.2 ± 6.23</td>
<td>2.99</td>
<td>1.05</td>
</tr>
<tr>
<td>L196V a</td>
<td>22.8 ± 0.72</td>
<td>64.2 ± 0.95</td>
<td>2.81</td>
<td>0.99</td>
</tr>
<tr>
<td>N334D (a)</td>
<td>42.6 ± 2.70</td>
<td>90.2 ± 1.93</td>
<td>2.12</td>
<td>0.74</td>
</tr>
<tr>
<td>A348P a</td>
<td>26.9 ± 0.65</td>
<td>82.7 ± 0.76</td>
<td>3.07</td>
<td>1.08</td>
</tr>
<tr>
<td>Q82P b</td>
<td>26.5 ± 0.45</td>
<td>47.8 ± 0.26</td>
<td>1.80</td>
<td>0.63</td>
</tr>
<tr>
<td>A200P (β)</td>
<td>36.7 ± 2.50</td>
<td>72.7 ± 1.82</td>
<td>1.98</td>
<td>0.69</td>
</tr>
<tr>
<td>T203K β</td>
<td>45.9 ± 0.83</td>
<td>12.4 ± 0.14</td>
<td>0.27</td>
<td>0.10</td>
</tr>
<tr>
<td>A260P (β)</td>
<td>92.0 ± 3.04</td>
<td>3.36 ± 0.07</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>H331P (β)</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
<td>~0</td>
</tr>
</tbody>
</table>

* Sequences of the replacement mutations are given in detail in table 6. For each, the wild-type structure is designated before the arrow, the mutant after. For example, EcB → Ps5 refers to the mutations in which the sequences coding for *E. coli* beta-sheet B are replaced with those coding for *Pseudomonas aeruginosa* beta-sheet 5. For each of the point mutations, the wild-type residue and position (see Sowadski et al. 1985) is followed by the mutant residue. Point mutations are classified as being within alpha helices or beta sheets by the symbols a and p, respectively, following the mutant designation; residues near rather than within the structures in question are enclosed in parentheses.

* ~0 indicates that enzyme did not possess detectable activity above background hydrolysis of the substrate. All values are given ± 1 standard error, calculated from at least three independent measurements.

* To the K-12 value.
start of the gene as compared with *E. coli*, matching in 16 of 18 bp. To date, these represent the only genes for *phoA* sequenced from bacteria other than *E. coli*. Multiple structural genes for alkaline phosphatase have been cloned from *Bacillus licheniformis* (Hulet 1987), but it is as yet unknown whether this regulatory feature is conserved in that species as well. The second observation concerns the codon bias within *phoA*. By either Ikemura's (1985) optimization parameter (*F*<sub>op</sub>) or Sharp and Li’s “codon adaptation index” (CAI; Sharp and Li 1986, 1987a, 1987b; Sharp et al. 1988), *phoA* shows low-to-moderate codon bias, usually corresponding to lower levels of expression.

In the absence of inorganic phosphate, however, *phoA* is highly expressed, to the point where 5%-10% of the total cellular protein content can be alkaline phosphatase (Reid and Wilson 1971). Taken together, these indicate that, while *phoA* may be highly expressed, it is not expressed often. To put the matter another way, *E. coli* does not experience conditions of phosphate starvation regularly. A similar situation is seen with β-galactosidase—also a gene whose levels of expression can be high but which yet shows low codon bias (Sharp and Li 1986, 1987a, 1987b; Sharp et al. 1988). As is the case with β-galactosidase, *phoA* is apparently a dispensable gene in some species.

Of the 10 bacterial species screened, it is not present in four of them, having either been lost independently at least twice (once in the lineage leading to *Salmonella* and at least a second time in the lineage(s) leading to *E. hermanii*, *E. vulneris*, and *E. blattae*, whose precise phylogenetic relationships are uncertain; see Brenner et al. 1969, 1972; Farmer et al. 1985a, 1985b) or introduced into the *E. coli/E. fergusonii* ancestor by horizontal transfer. With only three sequences, it is not possible to distinguish between the two possibilities. However, *phoA* is loosely linked to *lac*, which is known to occur on plasmids in *Klebsiella*. The area between *lac* and *phoA* (especially the area surrounding *lac*) is extremely rich in insertion elements (Deonier 1987), which could mediate recombination from an incoming plasmid and thus facilitate horizontal transfer. In addition, Southern hybridizations have detected plasmid-borne sequences hybridizing to *phoA* in *Serratia grimsii* and in two strains in the ECOR (*E. coli* standard reference) strain collection (data not shown).

The results presented here, in conjunction with those of other studies (Perutz et al. 1965; Lesk and Chothia 1980, 1982; Chothia and Lesk 1982, 1985, 1986, 1987; Perutz 1983; Stuart and Wilson 1987; Prager and Wilson 1988), present interesting implications with regard to the possible pathways of protein evolution. In a detailed study of globin structural evolution, Lesk and Chothia (1980) proposed that protein evolution in the buried cores of proteins occurs by the fixation of replacements that retain similar polarity but not necessarily similar size or identity. They suggest that canalization of protein structure will limit the potential pathways that further evolution may take and that a major rearrangement of secondary structure is unlikely to occur during evolution, although minor shifts of structural elements relative to each other may occur (this has been seen in a variety of different types of structural proteins; see Chothia and Lesk 1982, 1985; Lesk and Chothia 1982). Alkaline phosphatase appears to be governed by the same general principles: the lack of replacements within the beta-sheet core and in the monomer/monomer interface, where replacements should have the greatest structural effect, supports this view. Similarly, a comparison between human placental and bacterial alkaline phosphatase (Kam et al. 1985) indicates that the major design of the *E. coli* enzyme has been conserved. While the primary sequence shows only 30% similarity to *E. coli*, the active site and metal ion–binding ligands have been maintained. Finally, although the secondary and tertiary structures have not been determined for the placental enzyme, the amino acid sequence indicates that
some of the E. coli structures have been moved relative to each other and that still others are apparently missing. Helix 7 in particular is absent; this is the helix which, of those altered via mutagenesis in the present study, was the most tolerant of replacements (see Wyckoff 1987).

The question of the extent to which individual amino acid residues coevolve during evolution is still unresolved. Further experimentation in two directions should enable this problem to be addressed as well. First, the effect of multiple replacements can be investigated by combining individual mutations, as was done in the helical and beta-sheet replacement experiments. The helical experiments have already indicated that the activity of double mutants is not readily predictable from that of their respective single mutants (DuBose and Hartl 1989). Both dominance and interactive effects are seen in the $K_m$ and $K_{cat}$ of the mutants. By combining mutations whose activities diverge from wild type in opposite directions, it is possible to address this question experimentally (Dunn et al. 1988). Second, these mutations can serve as a baseline for improvement by artificial selection. By selecting revertants with increased activity, it is possible to examine the relative contribution of secondary mutations either within the local structures or elsewhere in the protein, as has been done with $ebg$ and other cryptic genes (Hall 1983) in E. coli. By combining inferential approaches of sequence comparisons with experimental ones using techniques such as site-directed mutagenesis, we can gain more insight into the processes of DNA and protein evolution than would be possible by using either approach alone.

Acknowledgments

R.F.D. thanks Howard Ochman and Jeff Lawrence for providing DNA probes for gap and trpR. We would like to thank Dan Dykhuizen, Kyoko Maruyama, Howard Ochman, and two anonymous reviewers for comments on previous versions of the manuscript. This work was supported by National Institutes of Health grants GM 30201 and GM 08036.

LITERATURE CITED


CANTERS, G. W. 1987. Cloning and sequencing of the azurin gene from *Pseudomonas aeruginosa*. 
Protein Eng. 1:243.

Function of arginine-166 in the active site of *Escherichia coli* alkaline phosphatase. Biochemistry 
27:8338–8343.

CHANG, C. N., W. KUANG, and E. Y. CHEN. 1986. Nucleotide sequence of the alkaline phosphatase 


—. 1985. Helix movements and the reconstruction of the haem pocket during the evolution 

—. 1986. The relation between the divergence of sequence and structure in proteins. EMBO 
J. 5:823–826.

399–405.

DAYHOFF, M. O., R. V. ECK, and C. M. PARK. 1972. A model of evolutionary change in 
National Biomedical Research Foundation, Washington, D.C.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change 
Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

American Society for Microbiology, Washington, D.C.


DUBOSE, R. F., and D. L. HARTL. 1989. An experimental approach to testing modular evolution: 
directed replacement of alpha helices in a bacterial protein. Proc. Natl. Acad. Sci. USA 86: 
9966–9970.

—. 1990. Rapid purification of PCR products for DNA sequencing using Sepharose CL- 

DUNN, I. S., R. COWAN, and P. A. JENNINGS. 1988. Improved peptide function from random 


FARMER, J. J., B. R. DAVIS, F. W. HICKMAN-BRENNER, A. MCWHORTER, G. P. HUNTLEY- 
CARTER, M. A. ASBURY, C. RIDDLE, H. G. WATHEN-GRADY, C. ELIAS, G. R. FANNING, 
A. G. STEIGERWALT, C. M. O’HARA, G. R. MORRIS, P. B. SMITH, and D. J. BRENNER. 
1985a. Biochemical identification of new species and biogroups of Enterobacteriaceae isolated 

FARMER, J. J., G. R. FANNING, B. R. DAVIS, C. M. O’HARA, C. RIDDLE, F. W. HICKMAN- 
BRENNER, M. A. ASBURY, V. A. LOWERY, and D. J. BRENNER. 1985b. *Escherichia fergusonii* 
and *Enterobacter taylorae*, two new species of Enterobacteriaceae isolated from clinical speci- 


FITCH, W. M. 1976. The molecular evolution of cytochrome-c in eukaryotes. J. Mol. Evol. 8: 
13–40.

185:862–864.

248.


PERUTZ, M. F., J. C. KENDREW, and H. C. WATSON. 1965. Structure and function of hae-


Sharp, P. M. Processes of genome evolution reflected by base frequency differences among *Serratia marcescens* genes. Mol. Microbiol. (accepted).


BARRY G. HALL, reviewing editor

Received January 4, 1990; revision received May 8, 1990

Accepted May 22, 1990