Temporal and Topological Clustering of Diverged Residues among Enterobacterial Dihydrofolate Reductases

Lorrie Daggett Garvin and Stephen C. Hardies

Department of Biochemistry, University of Texas Health Science Center at San Antonio

The complete nucleotide and encoded amino acid sequences were determined for the dihydrofolate reductase (DHFR) from the bacteria Enterobacter aerogenes and Citrobacter freundii. These were compared with the closely related Escherichia coli DHFR sequence. The ancestral DHFR sequence common to these three species was reconstructed. Since that ancestor there have been seven, nine, and one amino acid replacements in E. coli, E. aerogenes, and C. freundii, respectively. In E. coli, five of its seven replacements were located in the beta-sheet portion of the protein, and all seven were located in a single restricted region of the protein. In E. aerogenes, all nine of its replacements were located within surface residues, with five clustered in a region topologically distinct from the E. coli cluster. The replaced side chains are sometimes in direct contact but more often are separated by an intervening side chain. It is argued that the temporal clustering of replacements is typical for the evolution of most proteins and that the associated topological clustering gives a picture of how evolutionary change is accommodated by protein structure.

Introduction

Temporal clustering occurs when groups of replacements fall more closely in time than expected by chance and is a well-studied phenomenon affecting the long-term divergence of proteins. Numerous investigations of the rate at which proteins diverge have indicated a variance 2-2.5 times that of a Poisson process (for a review, see Wu et al. 1974; Zuckerkandl 1987). The extra variance means that replacements accumulate only half as steadily as would be expected if they occurred independently of each other with a constant average probability. This clustering in time requires that replacements be interdependent in some way, either (1) by having some replacements dependent on the prior occurrence of others or (2) by having multiple replacements triggered by some common influence external to the protein. In the present paper we measure temporal clustering as a variation in the total number of replacements occurring in different bacterial species over the same period of time.

Topological clustering occurs when replacements fall closer together in space than expected by chance and has been recognized between widely diverged proteins. A model called the “covarion model” has been developed, which seeks to describe the portion of residues in a protein that are free to vary (Fitch and Markowitz 1970; Fitch 1971, 1976; Shoemaker and Fitch 1989). The covarion model, when applied to cytochrome c from a wide variety of species, indicated that only about 12 residues are free to vary in any given sequence but that this concomitantly variable set of

1. Key words: dihydrofolate reductase (DHFR), protein structure, sequence analysis, rate of protein evolution, molecular clock.

Address for correspondence and reprints: Stephen C. Hardies, Department of Biochemistry, University of Texas Health Science Center, 7703 Floyd Curl Drive, San Antonio, Texas 78284-7760.

© 1991 by The University of Chicago. All rights reserved.
0737-4038/91/0805-0007$02.00

654
residues shifts to different sites over time. Thus, replacements are eventually found at most positions in the protein. In support of this model, those areas of cytochrome c that are most densely populated by replacements are not the same for different phylogenetic groupings (Margoliash et al. 1972). A similar observation was made for myoglobin (Castillo et al. 1977). A different kind of analysis of snake venom toxins revealed topological clustering in the sense that certain pairs of replacements repeatedly occurred together (Breckenridge and Dufton 1987). In the present paper we measure topological clustering as a spatial grouping of amino acid replacements in a region where there are no replacements in the homologous proteins from closely related bacterial species.

Any circumstance that causes one replacement to influence the acceptability of subsequent replacements may cause temporal and topological clustering. We will use the term “context effects” to describe such interactions among replacements. Context effects have been observed in the following kinds of studies: (1) Genetic studies often reveal the suppression of deleterious effects of some mutations by second-site intragenic suppressor mutations (Yanoisky et al. 1964; Hecht and Sauer 1985; Hampsey et al. 1986; Klig et al. 1988; Minor et al. 1989; Nagata et al. 1989). (2) In evolutionary studies, the shifting of the context effect was proposed to drive the movement of the covariable sets in the covarion theory (Fitch and Markowitz 1970). (3) Context effects on packing interactions have been experimentally explored by saturation mutagenesis of the lambda repressor (Reidhaar-Olson and Sauer 1988; Lim and Sauer 1989). Reconstruction of ancestral lysozyme sequences by in vitro mutagenesis has similarly been used to explore the interactions among packed residues (Malcolm et al. 1990). (4) A theoretical study of the context effect on packing of hydrophobic cores has been carried out by computational methods (Ponder and Richards 1987).

Both theories of evolution and theories of protein structure frequently invoke context effects in the form of compensation by one replacement for the effects of another. Lesk and Chothia (1980, 1982; Chothia and Lesk 1987) have conducted detailed structural comparisons among families of homologous proteins, concluding that changes are usually coupled so as to maintain the active-site configuration but that complementary and adjacent changes are not the rule. In light of these studies, the topological clusters indicated by other methods should not be expected to consist of pairs of packed side chains. Rather they should reflect replacements scattered within a cooperatively folded local region.

Temporal and topological clustering are closely related phenomena. Both can be attributed to the interactions among side chains, as discussed above, and both are natural consequences of the covarion model. We keep them conceptually separate because they correspond to two different kinds of measurements. Temporal clusters are observed by measuring the variance in branch lengths of an evolutionary tree. Unfortunately, only the sizes of temporal clusters are measured in this way, not the identities of those replacements that are concentrated within a short time interval. Topological clusters are measured by observing the positions of replacements in the three-dimensional structure. Unfortunately, the order and timing of the replacements within the topological cluster are usually not known. In the Discussion we will consider further the correspondence between topological and temporal clustering.

When widely diverged proteins are compared, as in most cases cited above, the resultant picture is blurred because many different clusters of replacements are superimposed. At lower divergence, however, sequences of homologous genes may show discrete topological clusters that are related to single sets of causally interrelated re-
placements. Similarly, a single temporal cluster might stand out in one lineage that is heavily diverged compared with other lineages of similar duration.

We have examined the divergence in the sequence for dihydrofolate reductase (DHFR) from three closely related species of bacteria. DHFR catalyzes the reduction of dihydrofolate to tetrahydrofolate, which is essential for the biosynthesis of DNA. The crystal structure of *Escherichia coli* DHFR is known to 1.7-Å resolution and features a central parallel beta sheet buried by alpha helices and surface loops (Bolin et al. 1982). In vitro mutagenesis has been carried out at several positions of the DHFR molecule (Villafranca et al. 1983; Chen et al. 1985; Howell et al. 1986; Benkovic et al. 1988). Garvey and Matthews (1989) have combined in vitro mutagenesis with investigation of folding kinetics to identify positions participating in key folding steps. On the basis of the extensive existing characterization, DHFR is an excellent system in which to follow up sequence observations with molecular biological experimentation.

Comparison of DHFR sequences and/or X-ray crystal structures have been made among *E. coli*, *Lactobacillus casei*, *Streptococcus faecium*, chicken, and mammals (for review, see Freisheim and Matthews 1984). We have added the sequences of *Enterobacter aerogenes* and *Citrobacter freundii*, which are gram-negative bacteria closely related to *E. coli*. These two proteins have 92% and 96% similarity with *E. coli* DHFR, respectively. We observed both temporal and topological clustering, including both a cluster of diverged residues in the beta sheet of *E. coli* DHFR and a cluster of diverged residues on the surface of *E. aerogenes* DHFR.

**Experimental Procedures**

**Bacterial Strains**

*Enterobacter aerogenes*, strain 13048, and *C. freundii*, strain 8090, were obtained from the American Type Culture Collection.

**Cloning of the *E. aerogenes* and *C. freundii* DHFR Genes**

Plasmid pCV37 containing the *E. coli* DHFR gene between the BamHI and SalI sites of pBR322 (Smith and Calvo 1980) was provided by J. Calvo. Flanking regions of the DHFR gene were removed by excision of an internal 400-bp SalI fragment and a 3.6-kb EcoRI fragment. The resulting plasmid, named “pCVc,” has a 500-bp RI/Sal fragment containing most of the DHFR coding region inserted between the EcoRI and SalI sites of pBR322. pCVc was radiolabeled and used for subsequent detection of the *C. freundii* and *E. aerogenes* DHFR genes on Southern blots and in genomic libraries prepared in the vectors M13mp18 or M13mp19.

A 10-kb EcoRI fragment from *E. aerogenes* was cloned into M13mp19. A 3.8-kb spontaneous deletion occurred during the cloning of this fragment; however, sequencing revealed that the whole DHFR gene was present, including 442 bp 5' of the coding region. A 6.0-kb EcoRI fragment of *C. freundii* was cloned into M13mp18. This fragment was missing the last 20 codons, so we also cloned into M13mp18 a 2.0-kb PstI fragment which overlapped the 3' end of the gene. Southern blot hybridization of both genomic *E. aerogenes* and *C. freundii* DNA by the *E. coli* probe revealed that these fragments were the only hybridizing fragments in the respective genomes.

**Subcloning and Sequencing**

Shotgun cloning was by the method of Bankier et al. (1987). The DHFR genes were fragmented for subcloning into M13mp18 either by digestion with HaeIII, AluI, HhaI, or HinfI or by sonication. DNA sequencing was by the dideoxy chain-termi-
Clustering of Diverged Residues in DHFR

Cloning method as modified by Bankier et al. (1987), using the universal M13 primer. Single-strand gaps were filled by priming with internal oligonucleotides. Figure 1 shows the sequencing strategy for both E. aerogenes and C. freundii. Sequence data were compiled with programs written by R. Staden (1982). The structures displayed in figure 2 are derived from the crystal structure of E. coli DHFR (Bolin et al. 1982) displayed on an Evans and Sutherland model PS350 running Frodo 6.6.

Assignment of Replacements to Individual Species

A replacement was assigned to one species at positions in the aligned sequences where it differed and where the other two agreed. For example, at codon 23, where C. freundii and E. coli have AAC (Asn) and where E. aerogenes has GAC (Asp), a replacement was assigned to E. aerogenes. At some codons, two replacements had to be assigned to interchange the codons (see legend to table 1). At codon 88 all three species have a different residue. It was assumed that two were the result of replacements and that one was still the same as the ancestral residue; however, there is no way to decide which of the three species still has the ancestral residue. Therefore the minimally required two replacements at codon 88 were distributed two-thirds to each of the three species.

Reliability of the Assigned Replacements

The assigned replacements will differ from the true history of this gene system in instances where two replacements have occurred at the same site such that the second replacement masks the first. For example, we can estimate that how often a single assigned replacement in E. aerogenes really represents a series of two successive replacements at the same position is equal to 1.3 [i.e., 9 (number of assigned replacements) × 27/180 (fraction of synonymous sites that have been mutated)]. We chose the fraction of altered synonymous sites, rather than the fraction of altered replacement sites, to represent the probability of the second replacement, because a residue already in the covariable set is more likely to be replaced than a residue chosen at random. Summing over all three species, we should expect two or three replacements to be hidden by this effect. Empirically, the number of hidden replacements should be about the same as the number of visible double replacements (i.e., changes at two different positions in the same codon). This happens twice in our data (see legend to table 1).

![Diagram](https://example.com/diagram.png)

**Fig. 1.**—DNA sequencing strategy. The double-strand sequencing strategy for the DHFR gene in (A) Enterobacter aerogenes and (B) Citrobacter freundii is shown. The boxes indicate the reported sequences.
Also, a replacement may be assigned to the wrong species if, in truth, an identical change had coincidently occurred in both of the other two species (parallel replacements). When parameters from table 1 are used, the probability of misassigning a replacement to \textit{E. coli} in this way is 0.5 \[\left(\frac{8.7 \times 27.7}{180} + 0.7 \times 27.2/180\right) \frac{1}{3}\] = 0.24. The \(\frac{1}{3}\) factor was used because both replacements would have to result from the same base change. There is a similar chance of misassigning a replacement to \textit{E. aerogenes}, but there is no chance of misassigning a replacement to \textit{C. freundii}, since there are no replacements uniquely assigned to this species. Empirically, the number of misassignments should be less than the number of instances where the second replacement is not identical to the first, causing three different residues to appear at the same position in the three different species. This happens only once in our data (see legend to table 1). Synonymous changes were assigned to individual species by the same principle used to assign replacements.

Finally, our ancestral sequence represents the node of a three-member unrooted tree. If the tree were rooted, then the root would have to fall in one of the three species, separating its replacements into two separate groups.

Note that each potential error affecting replacement assignments is in the direction of obscuring clustering. Hidden replacements would preferentially suppress the count of replacements in the more heavily diverged sequences. Parallel replacements would be misassigned from the most heavily diverged to the least heavily diverged sequences. Both of these effects would tend to obscure temporal clustering. The superimposition of two groups of replacements that might be separable by adding an outgroup would tend to obscure the topological distinctiveness of each group. None of these sources of error has the potential to artifactually create the appearance of clustering. The possibility of unequal opportunity for change among the three species is addressed in the Results section.

Results

The complete nucleotide and encoded amino acid sequences for the DHFR genes of \textit{Citrobacter freundii} and \textit{Enterobacter aerogenes} are shown, in comparison with the \textit{Escherichia coli} sequence, in figure 3. The \textit{E. coli} DHFR gene was reported to be completely sequenced on both strands (Smith and Calvo 1980). Since our analysis depends on the differences among the sequences, we have also confirmed the accuracy of the two new sequences by completely sequencing both strands.

Temporal Clustering of Diverged Residues

Inspection of figure 3 reveals that the number of residues unique to each bacterial species varies from one to eight. The reason for this variation is further explored in table 1. The three genes were assumed to have simultaneously diverged from a common ancestral DHFR gene which is defined as having the basepair shared by two out of the three sequences at any divergent position. This allows most of the replacements to be assigned to one of the bacterial species. The handling of multiply changed codons is further described in the Experimental Procedures section and causes the number of replacements attributed to \textit{E. coli} and \textit{E. aerogenes} to increase to seven and nine, respectively. An assessment of factors affecting the reliability of the assignments (see the Experimental Procedures section) indicated that they should be reliable to within three of four total replacements and that the identifiable sources of error would tend to obscure, rather than artifactually enhance, the appearance of clustering.

The divergence at synonymous sites is tabulated (table 1) as a control for unequal
FIG. 2.—Distribution of replacements in tertiary structure of DHFR. Positions of altered residues are superimposed on the crystal structure of *Escherichia coli* DHFR (Bolin et al. 1982). The view is into the active-site cleft. NADPH binds in the upper part of the cleft laying along the beta E strand with the nicotinamide moiety directed downward (Filman et al. 1982). Methotrexate binds in the lower rear part of the cleft, making contacts with the alpha helix which runs behind the beta E strand (Bolin et al. 1982). Replacements incurred during the recent descent of (A) *E. coli* DHFR and (B) *Enterobacter aerogenes* DHFR from a common ancestor are shown. The space-filling representations of the replaced residues are drawn without hydrogens. The side chains for the residues peculiar to *E. aerogenes* were placed by analogy to the *E. coli* structure. C, Differences between *E. coli* and *E. aerogenes* DHFR. The dots represent amino acid differences between the two bacteria placed at the backbone hydrocarbon position. D, Differences between *E. coli* and *Lactobacillus casei* DHFR. Triangles represent deletions or insertions.

opportunity for each bacterial species to have accumulated replacements in the gene for DHFR. If one of the bacterial species were more diverged than the others (e.g., because it had really diverged from the common ancestor at a much earlier time) then the number of synonymous base changes would be unequally distributed among the three sequences. Since the number of synonymous base changes is not distributed in proportion to the replacements, the disparity in replacements among the three sequences should not be attributed to unequal opportunity.

From table 1 the mean number of replacements is 5.3, and the variance is 17.3, giving a variance about three times the mean. The situation does not change appreciably if the values are weighted by the numbers of synonymous changes (not shown). The
The general rule for evolutionary divergence is for the variance to be 2–2.5 times the mean. If the replacements were entirely independent, the variance should equal the mean, so the extra variance indicates an interdependence within groups of two or three replacements. The observed variation among the three DHFR genes is therefore consistent with the rate variation generally observed in other gene systems.

Topological Clustering of Diverged Residues

Figure 2a–c illustrates the positions of the *E. coli* and *E. aerogenes* replacements superimposed on the *E. coli* crystal structure. The *E. coli* replacements (fig. 2a) appear in an obvious cluster on one side of the protein. For the most part, the altered side chains are not in direct contact but instead are separated by one intervening unchanged side chain or elements of the backbone. This forms our operational definition of a topological cluster as several changes with no more than one intervening unchanged moiety in a region which is unaltered in other species with equal opportunity for change.

The replacements in *E. aerogenes* are at surface residues (fig. 2b). Six of nine replacements, including a double replacement at residue 89, line the top of the protein. As in *E. coli*, most of the replaced residues are not in direct contact but have one
Table 1
Number of Synonymous and Replacement Changes Having Occurred During Descent from Common Ancestor

<table>
<thead>
<tr>
<th></th>
<th>No. of Synonymous Changes</th>
<th>No. of Replacement Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Citrobacter freundii</em></td>
<td>27.67</td>
<td>0.67</td>
</tr>
<tr>
<td><em>Enterobacter aerogenes</em></td>
<td>27.17</td>
<td>8.67</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>39.17</td>
<td>6.67</td>
</tr>
</tbody>
</table>

a The fractional values for synonymous changes result from 11 positions where all three genes have a different basepair. These were split 1/3, 1/3, and 1/3— or sometimes 1, 1/2, and 1/2 if necessary to avoid introducing an extra replacement at the codon.

b The 2/3 fractional values for replacement changes are due to residue 88 being different in all three genes.

c At residue 89, two replacements are necessary to convert the ancestral codon, CCA, to the *Enterobacter aerogenes* codon, GAA (Pro→Ala→Glu or Pro→Gln→Glu).

d At residue 73, two replacements are necessary to convert the ancestral codon, CAG, to the *Escherichia coli* codon, ACG (Gln→Lys→Thr or Gln→Pro→Thr).

unchanged intervening residue. The residue separating His 108 and Glu 89 is the N-terminal methionine which is not free to change. Since, in *E. coli* or *C. freundii*, there are no replacements to the left of position 88, this group (a double replacement Glu 89, His 108, Asp 129, and Glu 132) satisfies our operational definition of a topological cluster.

The long-term trends in bacterial DHFR divergence are represented by the comparison of sequences from *E. coli* and *Lactobacillus casei* (fig. 2d). At this much greater level of divergence, which is typical of many evolutionary comparisons that have been made, the spatial distribution of individual clusters is totally obscured. Figure 2c shows the effects of leaving out the *C. freundii* data. Without the third sequence, the number of comparisons is reduced from three to one, eliminating the possibility of seeing the temporal aspect of clustering in each local region. Also, the divergence time separating *E. coli* from *E. aerogenes* is twice that separating each from its common ancestor. Even this small increase in divergence is enough to partially obscure the spatial grouping. Therefore, individual topological clusters are only apparent in a relatively narrow range of divergence.

Changes to the Hydrophobic Core

Perhaps the most striking feature of the *E. coli* cluster is that five of the seven replacements (including the double replacement at 73) reside in the beta sheet forming the floor of the right domain of the protein. Of these, three of the side chains are buried in the hydrophobic core. Since replacements in proteins are generally found concentrated at surface residues (Fitch 1976), the cluster of internal replacements in *E. coli* would be easier to understand if they were causally related.

Figure 4 shows a detailed view of the beta sheet containing the four altered residues in *E. coli*. Since these are the amino acids that were actually determined in the *E. coli* crystal structure, their side chains are well localized. Ile 41, Leu 62, and Thr 73 project up out of the page, and Ile 61 projects down into the page. Therefore, side-chain contact among Leu 62, Ile 61, and Thr 73 is precluded, although Leu 62 and Ile 41 are in direct contact. Residues 61 and 62 are in contact through the backbone, and 61 and 73 are in contact through the hydrogen-bonding network of the beta sheet.
FIG. 3.—Comparison of sequences from Escherichia coli (EC), Enterobacter aerogenes (EA), and Citrobacter freundii (CF) DHFR genes. Only differences from C. freundii are shown for E. coli and E. aerogenes. The E. coli sequence is that reported by Smith and Calvo (1980). The base compositions are 112 A, 116 C, 143 G, and 109 T for C. freundii and 112 A, 114 C, 142 G, and 112 T for E. aerogenes.

Ile 61 is completely buried and could be considered the center of the hydrophobic core. Ile 41 is also completely buried, being wedged between the hydrophobic cores of the two domains of the protein. Leu 62 is mainly buried in hydrophobic contacts, although the tip of the side chain has slight surface exposure. Thr 73 is positioned like a plate, with one face lying on the hydrophobic core and with the other face exposed to solvent. The changes affecting the three buried residues (Val 41→Ile, Val 61→Ile, and Ile 62→Leu) are chemically conservative and of the type customarily found within hydrophobic cores (Chothia and Lesk 1987; Bowie et al. 1990). The total volume increase contributed by these three replacements is 36.5 Å³, calculated according to the method of Richards (1974).

Is E. coli DHFR Strange?

On a per-site basis, E. coli DHFR has actually diverged faster in the hydrophobic core than on the outside. We count 40 residues in DHFR that are on beta strands not including the ends. Eight percent of these (3 of 40) have been replaced in E. coli. Only 3% (4/119) of the remaining residues have been replaced. So, the buried beta strands have diverged 2.5 times faster than the remainder of the protein. This is not characteristic of the usual behavior of proteins (Fitch 1976). The long-term-average behavior of DHFR can be obtained from a comparison of E. coli with L. casei. Counting over the same two sets of residues (table 2) yielded 68% (27/40) differences within the beta sheet and 75% (69/119) differences elsewhere. It is clear that DHFR is not generally more variable within the beta sheet than elsewhere.

If the E. coli enzyme is not representative of the long-term-average behavior of DHFR, then it is equally true that the E. aerogenes and C. freundii enzymes are not
Clustering of Diverged Residues in DHFR

**FIG. 4.** Close-up of replacements in beta strands B, C, and D in *Escherichia coli* DHFR. The orientation has been rotated 90° forward from that of fig. 2, resulting in a view from the top of the protein.

representative either. In aggregate, *E. aerogenes* and *C. freundii* DHFR have experienced 8% (9/119) replacements outside the beta sheet but none inside of the beta sheet (table 2). This is a far greater preference for surface replacements than is exhibited in the long term, as discussed above. However, the aggregate of all three genes has 8% (3/40) replacements in the beta sheet and 13% (15/119) elsewhere, in good agreement with the long-term average. We conclude that none of these individual bacterial DHFRs are strange but that they represent an assortment of the short-term behaviors that are randomly mixed to yield the customary patterns of long-term divergence.

Relation to Some Earlier In Vitro Mutagenesis Studies

His 45 of *E. coli* lines the active-site cleft. It was previously altered to Glu by in vitro mutagenesis, with the finding that activity and stability of the protein were not significantly altered (Chen et al. 1985). *Enterobacter aerogenes* has a Leu at that same position.

Valine 75 has been subjected to an extensive series of substitutions in a study of folding in DHFR (Garvey and Matthews 1989). Mutations at this residue were observed to alter folding at the rate-limiting step which was proposed to be a reorganization of the beta sheet. Valine 75 is in intimate association with the cluster of replacements that have occurred within that beta sheet in the *E. coli* enzyme.

Relation to Other DHFR Strain Differences

Several investigators have reported sequence differences in other *E. coli* isolates (Stone et al. 1977; Bennett et al. 1978; Baccanari et al. 1981; Smith and Calvo 1982;
Table 2
Distribution of Replacements in Interior of Beta Sheet of DHFR

<table>
<thead>
<tr>
<th>Sequence under Comparison</th>
<th>Replacements in Beta Sheet</th>
<th>Replacements Outside Beta Sheet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citrobacter freundii</td>
<td>0/40</td>
<td>0.67/119</td>
</tr>
<tr>
<td>Enterobacter aerogenes</td>
<td>0/40</td>
<td>8.67/119</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>3/40</td>
<td>3.67/119</td>
</tr>
<tr>
<td>Escherichia coli/Lactobacillus casei</td>
<td>27/39</td>
<td>89/120</td>
</tr>
</tbody>
</table>

* Sequences other than C. freundii and E. aerogenes are reviewed by Freisheim and Matthews (1984).
  * The first and last residue of each beta strand were excluded, in that such positions are generally less well conserved than the interior of beta strands (Lesk and Chothia 1982).
  * All remaining residues, including ends of beta strands.
  * The double replacement at residue 73 in E. coli has been counted in the outside class.

Smith et al. 1982; and Flensburg and Sköld 1987). These strains were each resistant to trimethoprim or, in one case, were selected for altered interaction between DHFR and another gene product. The replacements in these strains do not fall near the cluster of replacements we have indicated for E. coli DHFR but, rather, tend to fall in the alpha helix which runs behind beta-strand E in figure 2. This region forms part of the binding site of the drug. Consequently, replacements in this area, forced by selection against the action of the drug, are not surprising. Some of these strains have multiple replacements that do not fall either in our cluster or near the trimethoprim binding site. Since it is unclear whether these replacements are also related to the drug-resistant phenotype or are strain differences that existed before conversion to drug resistance, we are not prepared to draw any interpretation about their distribution.

Discussion

In the comparison of closely related bacterial DHFR proteins, we observed the replacements to be both temporally and topologically clustered into two distinct groups of about six replacements each. We have considered these clustering effects to be reflections of the fundamental process by which proteins accommodate evolutionary change. We propose that, during evolution, most replacements occur in small spurts, each confined to a limited spatial region within the tertiary structure. The distribution of replacements between divergent homologues would then consist of a random sampling of these small topological clusters. We offer the following observations to argue that our model—and the data supporting it—are consistent with accepted theory of protein structure and evolution.

1. Studies on the variance of the molecular clock have consistently indicated that, in proteins in general, the density of replacements during short time intervals varies more than is expected purely by chance (Wu et al. 1974; Zuckerkandl 1987). The variation in the density of replacements among the three enterobacterial DHFR sequences is quantitatively consistent with the traditionally observed increased variance in replacement rates. The average of these fluctuations over longer times converges on a roughly constant divergence rate, which is the more familiar representation of the molecular clock.

2. Studies on the crystal structures of a number of homologous proteins have indicated that, although structural changes of the internally packed residues are often coupled to maintain the geometry of the active site, complementary changes affecting
adjacent side chains are not the rule (Lesk and Chothia 1980, 1982; Chothia and Lesk 1987). The distribution of replaced side chains within the internal *Escherichia coli* DHFR cluster adheres to this generalization. They are most often not in direct contact but are usually separated by an intervening side chain or by elements of the backbone. The chemically conservative nature of the replacements and the resulting volume change are in line with established precedent (Chothia and Lesk 1987; Bowie et al. 1990).

3. Comparisons of divergent proteins have indicated that the surface is more susceptible to change than is the interior (Fitch 1976). The long-term divergence pattern of DHFR is much like that of other proteins built around beta sheets (Lesk and Chothia 1982). Although *E. coli* DHFR appears to be unusual when considered in isolation, averaging of the three *enterobacterial* species together results in a distribution of replacements, between the interior and exterior, much like the long-term average behavior of DHFR. So the behavior of the interior follows a topological equivalent of the temporal variance discussed above. During most short intervals there are no replacements in the interior, but, when they occur, they occur as a cluster. The familiar divergence patterns are expected to appear after a number of short intervals are averaged together.

Surface versus Interior Changes

Most published commentary on coupled changes has been in reference to the packed interior (Lesk and Chothia 1980, 1982; Chothia and Lesk 1987; Ponder and Richards 1987; Reidhaar-Olson and Sauer 1988; Lim and Sauer 1989; Bowie et al. 1990). The replacement distribution in *Enterobacter aerogenes* suggests that topological clustering may also affect noncore residues. This is backed up by the observation that the *E. aerogenes* replacements are also part of a temporal clustering pattern. The distribution of noncore replacements is nine, four, and one among the three species, indicating that they too appear in spurts. Therefore, the rules governing patterns of replacements in the interior versus on the surface may be similar.

Mechanisms Enforcing Clustering

If changes at one residue influence the acceptability of changes at other residues (Fitch 1971), then there would be temporal clustering. Such context effects would reduce the independence among individual replacements and increase the variance in the divergence rate, therefore causing temporal clustering. We emphasize that by a temporal cluster we mean a series of replacements in evolutionary time, not a saltatory burst of mutations in the same bacterium. If the context effect is stronger between residues that are close together, then there would be topological clustering. That the affected side chains do not have to be in direct contact is a reflection of the cooperative nature of protein structure. We do not believe that either of these arguments is difficult to reconcile with traditional theories of protein structure.

Two different approaches can be taken in envisioning how clusters might be triggered. A chance independent replacement or a replacement from a succeeding cluster could act as the trigger for a new cluster. This mechanism would be self-sustaining and would fit nicely with recent versions of neutral theory (Ohta 1987; Takahata 1987). Alternatively, each cluster could be in response to a change in the selective environment, in agreement with recent versions of selective theory (Gillespie 1986; Malcolm et al. 1990). Therefore, we think that our observations are consistent with either neutral or selective theory.
Relationship between Temporal and Topological Clustering

Our model agrees with the covarion theory in invoking the shifting of positions of covariable codons such that topological clusters of replacements occur one after the other. Clearly, the development of a topological cluster implies temporal clustering of replacements in that local area of the protein. Perhaps the simplest way to view the relationship between topological and temporal clustering would be to equate the two by assuming that all replacements that fell close together in space were also compressed together in time. This view would hold that, between times when clusters are developing, there are periods when only a few isolated replacements are occurring. This is the easiest way to reconcile the lack of replacements in *Citrobacter freundii* vis-à-vis the clusters in the other two species.

More complicated relationships between temporal and topological clusters are possible, wherein some of the replacements within a topological cluster are more closely spaced in time than others. As an extreme example, the shifting of the covariable codons to include new residues may allow replacements to occur which improve fitness but which were previously excluded by steric considerations. These replacements would be fixed rapidly by positive selection. In this model, both the last replacement that made the selective change possible and the selective replacement itself would be strongly clustered in time, whereas further replacements to reoptimize the disturbed area might occur at a slower pace. Other considerations include the question of how the number of covarions associated with a topological cluster expands or contracts during its development. In any case, the exact relationship between temporal and topological clustering depends on the mechanism by which the covariable sets shift, which is a subject for further investigation.

Sequence Availability

The nucleotide sequences reported in the present paper have been submitted to the GenBank™/EMBL Data Bank under accession numbers M26022 and M26023.

Acknowledgments

We would like to acknowledge the fine technical assistance of Estella Acosta and Barry Korinek. We thank Peggy Rifleman (UTHSCA Biopolymer Sequencing and Synthesis Facility) for the synthesis of the oligonucleotides. This work was supported by grant AQ-1107 from the Robert A. Welch Foundation and by grant HG00190 from the NIH.

LITERATURE CITED


Clustering of Diverged Residues in DHFR


NAGATA, S., C. C. HYDE, and E. W. MILES. 1989. The alpha subunit of tryptophan synthase: evidence that aspartic acid 60 is a catalytic residue and that the double alteration of residues 175 and 211 in a second-site revertant restores the proper geometry of the substrate binding site. J. Biol. Chem. 264:6288–6293.


SHOEMAKER, J. S., and W. M. FITCH. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270–289.


WALTER M. FITCH, reviewing editor

Received September 10, 1990; revision received February 21, 1991

Accepted February 22, 1991