Selection Intensity for Codon Bias and the Effective Population Size of *Escherichia coli*

Otto G. Berg

Department of Molecular Biology, University of Uppsala Biomedical Center, S-75124 Uppsala, Sweden

Manuscript received July 10, 1995
Accepted for publication January 8, 1996

ABSTRACT

The selection intensity for codon bias and the synonymous diversity have been used in the recent literature to estimate the effective population size of *Escherichia coli*. The results have varied between 10^5 and 10^8. It is suggested here that most of this disparity can be explained by a model that accounts for the population structure of the species. Thus it is assumed that weakly selected characters, like synonymous substitutions, are selectively fixed within individual lines or colonies but spread throughout the population in an essentially neutral way when colonies replace one another. In this way, the effective population size that enters expressions for the codon bias will be that of an individual colony, which, if hitchhiking effects are considered, can be a very small number. The effective population size that appears together with the mutation rate in expressions for the synonymous diversity, on the other hand, will be related to the total number of colonies that make up the species and can be a very large number.

THE effective population size of a species is determined in practice by comparing the mathematical expressions for diversity based on some ideal population genetic model with the actual genetic diversity observed. In a recent study on the diversity of synonymous codon usage in homologuous genes from different strains of *Escherichia coli*, HARTL *et al.* (1994) deduced that the effective population size of this species is $\sim 2 \times 10^8$. This estimate is close to a commonly accepted value of $\sim 10^9$ (*e.g.*, OCHMAN and WILSON 1987) and leads to a selection coefficient for a preferred codon of $\sim 7 \times 10^{-9}$.

The preferential use of some synonymous codons over others, the codon bias, in some species is thought to be the result of selection for translational efficiency and/or accuracy (*e.g.*, IKEMURA 1985). Based on a translation model, BULMER (1991) calculated that the growth advantage and the selection coefficient for a preferred or major codon would be $\sim 10^{-5}$ (depending on gene expression) and deduced that the effective population size of *E. coli* is of the order of 10^5. This low, some would say ridiculously low, number agrees with an independent estimate based on the sequence diversity of binding sites for gene regulatory proteins (BERG 1992).

There is a major difference between these disparate estimates of the effective population size: the large number from HARTL *et al.* (1994) is based on the *intergenomic* diversity (virtual heterozygosity) where the same gene is compared in different strains; the small numbers (BULMER 1991; BERG 1992), on the other hand, are based on the *intragenomic* diversity, *i.e.*, the nucleotide selection bias within the same genome. Below it will be suggested that the heterozygosity and the codon bias are determined by different kinds of processes at different levels in the population structure, and that they are therefore connected to very different estimates of the effective population size. The codon bias is established as a local equilibrium of synonymous substitutions within each line and determined by the fixation probabilities of a small population size. The synonymous substitutions are spread throughout the species as different lines replace each other via extinction-recolonization events. The synonymous heterozygosity, therefore, is determined by the genealogy of lines and their separation times, which are related to a large population size determined primarily from the total number of lines. The distinction between the population sizes for bias and heterozygosity becomes particularly simple if it can be assumed that the colonization events are neutral with respect to the synonymous substitutions.

MODEL

Many model calculations, like those by HARTL *et al.* (1994) and BULMER (1991) for instance, assume that recombination is sufficiently frequent so that each mutation can spread and become fixed independently of all others. Thus, all of *E. coli* is considered as a single population with random mixing of alleles. However, *E. coli* has a distinct population structure and may be largely a clonal species (HARTL and DYKHUIZEN 1984; SELANDER *et al.* 1987; LENSKI 1993), as is indicated, for instance, by the strong linkage disequilibrium and by the fact that it is possible to construct genealogies for its strains. Furthermore, in comparisons of homologous genes from *E. coli* and Salmonella, codon bias is virtually

Corresponding author: Otto G. Berg, Department of Molecular Biology, BMC, Box 590, S-75124 Uppsala, Sweden.
E-mail: otto@xray.bmc.uu.se

the same in both species, while differences in heterozygosity suggest different effective population sizes (GUTTMAN and DYKHUIZEN 1994). Thus, codon bias cannot be directly related to the heterozygosity via the same effective population size.

Consider a group of sites in the genome where there are two nucleotide choices, one selected (weakly) over the other with selection coefficient s. This group could consist of, for instance, all sites for a twofold degenerate amino acid where one codon is preferred over the other. The expected distribution of such synonymous choices will be the focus of the following discussion. The central idea is that the population genetics of *E. coli* can be considered at two levels. On the species level, the population is subdvided into a large number of colonies (lines) (MARUYAMA and KIMURA 1980). These could either be in the guts of individual animals, or possibly in larger family groups of animals. It is on the level of an individual colony that mutations arise and first become fixed before spreading to others via invasion or recolonization events. The basic assumption, to be justified further below, is that selection within a colony is very different from selection between colonies.

Fixation within a colony: In contrast to the picture of complete mixing, below it is assumed that mutations are tightly linked such that neutral and weakly selected or counterselected mutations, like the synonymous codon changes considered here, are fixed mostly by hitchhiking (MAYNARD SMITH and HAIGH 1974) with a new strongly selected variant. A colony in a chemostat continuosly undergoes periodic selection events (selective sweeps) when new strongly selected mutations appear and take over the population (*e.g.*, HELLING *et al.* 1987). This seems to be mostly a consequence of the large size of the colony, and it can be argued that colonies in "the wild" will behave in a similar manner (MARUYAMA and KIMURA 1980; LEVIN 1981; KIMURA 1983; BERG 1995). The substitution rates, α_1 and α_2, whereby the selected codon (with selection coefficient s) replaces the counterselected one throughout the colony and vice versa, can be calculated in a hitchhiking model (BARTON 1994; BERG 1995). The results are determined primarily by the parameter combination $s\bar{T}$, where $\bar{T}$ is the average time between periodic selection events. If hitchhiking is the dominant process, one finds (BERG 1995)

$$\alpha_1 = u\frac{\exp(s\bar{T}) - 1}{s\bar{T}}, \tag{1a}$$

where u is the mutation rate constant. The corresponding result from a random-mixing haploid two-allele model for a stable population of size N_{col} (*e.g.*, BULMER 1991) is

$$\alpha_1 = \frac{2uN_{col}s}{1 - \exp(-2N_{col}s)}. \tag{1b}$$

When $2N_{col}s$ and $s\bar{T}$ are not larger than 1, these two expressions, (1a) and (1b), give virtually the same result if $2N_{col}s$ is replaced by $s\bar{T}$ (BERG and MARTELIUS 1995). Similarly, the codon bias is found to be

$$B = \alpha_1/\alpha_2 = \exp(s\bar{T}) \tag{2a}$$

if hitchhiking dominates, and

$$B = \alpha_1/\alpha_2 = \exp(2N_{col}s) \tag{2b}$$

in the random mixing model. Thus, the main effect of hitchhiking is to introduce $2N_{col}$ as the time between selective sweeps in a colony. This can be a very small number compared to the real population size of the colony.

Fixation throughout the species: On the species level, let us assume with MARUYAMA and KIMURA (1980) that the total population of *E. coli* can be subdivided into n distinct local subpopulations (colonies, lines). A colony becomes extinct with the probability λ per generation and is replaced by members from another colony (recolonization). MARUYAMA and KIMURA (1980) showed that the effective population size for neutral heterozygosity in this situation is dominated by the number of colonies:

$$N_{nh} \cong n/2\lambda \quad \text{if } n \gg 1. \tag{3}$$

In this model, N_{nh} corresponds to the average time to the most recent common ancestor of two randomly chosen colonies. Furthermore, the probability $p(t)$ per unit time that two randomly chosen colonies split off from their most recent common ancestral colony in a recolonization event at time t before the present is exponentially distributed with density (HUDSON 1990)

$$p(t) = \frac{2\lambda}{n-1}\exp\left(\frac{-2\lambda t}{n-1}\right) = \frac{1}{N_{nh}}\exp\left(\frac{-t}{N_{nh}}\right). \tag{4}$$

During the separation time t, each colony (line) has been accumulating mutations. One finds that the expected fraction of sites that are different after time t of separation is (BERG and MARTELIUS 1995)

$$f(t) = \frac{2B}{(1+B)^2}[1 - \exp(-2(\alpha_1 + \alpha_2)t)], \tag{5}$$

where the bias is $B = \alpha_1/\alpha_2$. In the case considered here, α_1 and α_2 refer to the rates of synonymous replacement within a colony as discussed above (Equations 1 and 2). Once a mutation is fixed in a colony, it can spread to others via colonization events. In the model of MARUYAMA and KIMURA (1980) the invading line in a recolonization is randomly chosen and the replacements of colonies therefore nonselective. In this picture, the spreading of the synonymous changes through colony replacements will be neutral and expressed by (3) and (4). As discussed further below, it is not unlikely that this spreading will be essentially neutral also in a more realistic picture where colony replacements are due to invasion-takeover events in which case λ corre-

sponds to a rate of invasion times the probability of successful takeover.

Heterozygosity: A measure for the heterozygosity is the average fraction of differences between two randomly chosen individuals. If the individuals are from different lines, the expectation value for the fraction of synonymous differences can be calculated as

$$\langle f \rangle = \int_0^\infty f(t)\,p(t)\,dt = \frac{2}{1+B}\cdot\frac{2N_{nh}\alpha_1}{1+2N_{nh}(\alpha_1+\alpha_2)}. \quad (6)$$

For a truly neutral mutation, $s = 0$ and $\alpha_1 = \alpha_2 = u$, where u is the mutation rate constant (for simplicity of discussion, u is assumed to be the same in both directions), one finds from (6)

$$\langle f \rangle = \frac{2uN_{nh}}{1+4uN_{nh}}, \quad (7)$$

where N_{nh} from (3) is the effective population size for neutral heterozygosity. Equation 7 has the term $4uN_{nh}$ in the denominator rather than $2uN_{nh}$, as given by MARUYAMA and KIMURA (1980), since this is a two-allele model rather than the infinite-allele model considered by them.

Rewriting the expression for the expected fraction of differences, (6), using (1b) and (2b), one finds

$$\langle f \rangle = \frac{2}{1+B}\cdot\frac{2uN_{nh}\cdot 2N_{col}s}{2uN_{nh}\cdot 2N_{col}s\cdot(1+1/B)+1-1/B}. \quad (8)$$

The structure of this expression remains the same if one uses the hitchhiking model within the colony, (1a) and (2a), instead. Thus, the effective population size enters the result in two places: in combination with the mutation rate constant u, one finds the effective population size for neutral heterozygosity N_{nh} from (3); together with the selection coefficient s one finds the local effective population size of a colony, N_{col}.

DISCUSSION

The calculations above show how one can get very different estimates for the effective population size associated with codon bias and with synonymous heterozygosity. The crucial assumption in the present model is that synonymous differences are spread throughout the population in an essentially neutral way as different lines replace each other. In a picture where lines must compete with each other in a colony replacement, this requires that they have picked up different kinds of strongly favorable mutations during their different recent histories. If so, a few synonymous differences will not be decisive and will therefore be neutral. The periodic selection experiments (HELLING *et al.* 1987; LENSKI *et al.* 1991) suggest that the organism readily picks up favorable mutations even after a long time in the same environment. The selection coefficients for these favorable mutations are found to be of the order of 10^{-1}, while the synonymous codon changes are estimated to have very small selection coefficients, anywhere from 10^{-5} (BULMER 1991) to 10^{-8} (HARTL *et al.* 1994). Natural populations of *E. coli* probably cycle through many different kinds of environments and may pick up different favorable mutations in the process. Thus *E. coli* may be well adapted not to a stable environment but to an unpredictable one that is continuously changing. The alternative view that all strains of *E. coli* are equally and optimally adapted to all its natural environments so that very few strongly favorable mutations ever occur seems less likely to me; however, this is a conjecture that remains to be tested.

In a chemostat, selective sweeps take place about once every hundred generations or so (HELLING *et al.* 1987). This corresponds to a "local effective population size" on the order of $N_{col} = 10^2$. For colonies in the wild, selective sweeps may be less frequent and N_{col} correspondingly larger. Replacements of resident strains in the intestinal flora may take place with a comparable rate (SELANDER *et al.* 1987) and may therefore interfere with periodic selection. Such invasion-recolonization events probably involve a very small number of cells and will therefore lead to a random sampling of the genetic variants in the invading line. If the invader were a single cell, each invasion-takeover would have the same effect for hitchhiking as a periodic selection event. However, since an invasion is not likely to start with a single cell, the random sampling of the invading line will be a more complicated process than a periodic selection event and the corresponding effective population size would be increased. Thus, N_{col}, which appears together with the selection coefficient s, both in the codon bias (Equation 2) and in the synonymous diversity (Equation 6), can be a very small number, though probably not as small as 10^2. Hitchhiking is not an essential part of the present model calculations, but it provides an obvious mechanism by which the effective population size of a colony can be very small.

The effective population size for the process of spreading the weak mutation across the whole *E. coli* population is determined by N_{nh}, which, from (3), probably is larger than the number of colonies. Some genetic material in *E. coli* has spread faster through the population and therefore has a much lower divergence than others (GUTTMAN and DYKHUIZEN 1994). Such material is thought to be linked to some strongly selected mutation and spread also via recombination. Effectively such parts of the genome will have a faster rate of spreading (larger λ) and therefore smaller N_{nh}. Clearly, the random-replacement model of MARUYAMA and KIMURA (1980) used above is only a first approximation that needs to be extended to include such more complex processes.

HARTL *et al.* (1994) identify two parameter combinations in their analysis of the intergenomic diversity of *E. coli:* $uN_e \approx 9 \times 10^{-2}$ and $sN_e \approx 1.3$. Using DRAKE's

(1991) number for the mutation rate $u = 5 \times 10^{-10}$, they find the effective population size $N_e \approx 2 \times 10^8$. Inserting this estimate for N_e in the result for sN_e, they deduce that the selection intensity for a preferred synonymous codon is extremely small, $s \approx 7 \times 10^{-9}$. As support for their results, HARTL *et al.* (1994) also look at the intragenomic diversity as determined by the codon bias, $sN_e = 0.5 \ln(B)$ from (2), and find that this is essentially the same as that determined from the intergenomic diversity. In spite of the uncertainty in many of the numbers, this indeed provides strong independent support for the usefulness of their theory. However, based on the analysis presented above, I would argue that if one uses a random-mixing model for a species with a strongly structured population, like *E. coli,* it is necessary to consider (at least) two very different numbers for the effective population size: the local size $N_e = N_{col}$, which appears in combination with the selection coefficient s, could well be 10^5 or smaller; the other and much larger number, $N_e = N_{nh}$ from (3), appears together with the mutation rate u and could well be 10^9 or larger. Thus BULMER's (1991) estimate based on translational efficiency for the selection coefficient of $\sim 10^{-5}$ for a preferred codon may be reasonable and the findings of HARTL *et al.* (1994) could be reconciled with the well supported notion (*e.g.,* IKEMURA 1985) that codon bias in *E. coli* is determined in some way by translational efficiency.

The primary purpose of the present model is to show how a simple allowance for the population structure in a straightforward way can account for the very disparate estimates of the effective population size. Some of the assumptions are clearly oversimplified, as discussed, and will require more extensive work before real quantitative predictions can be made. On the other hand, this situation is no different from that of existing random-mixing models where the single estimate of N_e is used as a fitting parameter: the effective population size is simply that number that is required to make the diversity of the real population look like that of an ideal random-mixing population of size N_e. In the present model, two such fitting parameters are required and it is also shown how they may arise in the dynamics of a structured population.

I thank PEDRO SILVA for critical reading of an earlier version of the manuscript and for useful discussions. This work has been supported by The Swedish Natural Science Research Council.

LITERATURE CITED

BARTON, N. H., 1994 The reduction in fixation probability caused by substitutions at linked loci. Genet. Res. **64:** 199–208.

BERG, O. G., 1992 The evolutionary selection of DNA base pairs in gene-regulatory binding sites. Proc. Natl. Acad. Sci. USA **89:** 7501–7505.

BERG, O. G., 1995 Periodic selection and hitchhiking in a bacterial population. J. Theor. Biol. **173:** 307–320.

BERG, O. G., and M. MARTELIUS, 1995 Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. J. Mol. Evol. **41:** 449–456.

BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

DRAKE, J. W., 1991 Spontaneous mutation. Annu. Rev. Genet. **25:** 125–146.

GUTTMAN, D. S., and D. E. DYKHUIZEN, 1994 Detecting selective sweeps in naturally occurring *Escerichia coli.* Genetics **138:** 993–1003.

HARTL, D. L., and D. E. DYKHUIZEN, 1984 The population genetics of *Escherichia coli.* Annu. Rev. Genet. **18:** 31–68.

HARTL, D. L., E. N. MORIYAMA and S. A. SAWYER, 1994 Selection intensity for codon bias. Genetics **138:** 227–234.

HELLING, R. B., C. N. VARGAS and J. ADAMS, 1987 Evolution of *Escherichia coli* during growth in a constant environment. Genetics **116:** 349–358.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. **7:** 1–44.

IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:** 13–34.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, England.

LENSKI, R. E., 1993 Assessing the genetic structure of microbial populations. Proc. Natl. Acad. Sci. USA **90:** 4334–4336.

LENSKI, R. E., M. R. ROSE, S. C. SIMPSON and S. C. TADLER, 1991 Long-term experimental evolution in *Escherichia coli.* I. Adaptation and divergence during 2000 generations. Am. Nat. **138:** 1315–1341.

LEVIN, B. R., 1981 Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics **99:** 1–23.

MARUYAMA, T., and M. KIMURA, 1980 Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc. Natl. Acad. Sci. USA **77:** 6710–6714.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitchhiking effect of a favourable gene. Genet. Res. **23:** 23–35.

OCHMAN, H., and A. C. WILSON, 1987 Evolutionary history of enteric bacteria, pp. 1649–1654 in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology,* edited by F. C. NEIDHARDT. ASM Press, Washington DC.

SELANDER, R. K., D. A. CAUGANT and T. S. WHITTAM, 1987 Genetic structure and variation in natural populations of *Escherichia coli,* pp. 1625–1648 in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology,* edited by F. C. NEIDHARDT. ASM Press, Washington DC.

Communicating editor: M. KIRKPATRICK