Recombination processes and non-linear Markov chains

S.A. Pirogov¹, A.N. Rybko¹, A.S. Kalinina¹, M.S. Gelfand¹,²

¹Kharkevich Institute for Information Transmission Problems, RAS, Moscow, Russia
²Lomonosov Moscow State University, Department of Bioengineering and Bioinformatics, Moscow, Russia

Abstract

Bacteria are known to exchange genetic information by horizontal gene transfer. Since the frequency of homologous recombination depends on the similarity of recombining segments, several studies examined whether this could lead to the emergence of subspecies. Most of them used simulations of fixed size Wright-Fisher populations, in which the genetic drift should be taken into account. Here, we use non-linear Markov chains to describe a bacterial population evolving under mutation and recombination processes. We consider a population as a space of genomes and a probability measure on it. Thus, the genetic drift is not assumed. We prove that under these conditions the emergence of subspecies is impossible, so the genetic drift is a necessary driving force of bacterial speciation.

Introduction

Despite the absence of sexual reproduction in prokaryotes, they are able to obtain genetic information from sources other than their maternal cells. In 1928, Griffith showed that after adding heat-killed cells of a pneumococcus strain virulent for mice (Streptococcus pneumoniae) to alive cells of another, non-virulent, strain, the latter had became virulent [1].

As it was shown later by Avery, MacLeod and McCarty, in Griffith’s experiment bacteria changed their genotype (and hence phenotype) by uptake of DNA from environment [2]. This process is called natural transformation. Approximately 1% of bacteria are known to have this
ability, i.e. are competent \[3, 4\]. Many of these species are not permanently competent, their ability to uptake DNA being induced by many factors such as stress and starvation. Restriction-modification systems can form a barrier for the natural transformation \[5\], for example, in *Neisseria meningitidis* the length of homologous recombination segments significantly depends on whether the recombinating strains have the same restriction-modification systems \[6\]. Other mechanisms for horizontal DNA transfer are conjugation and transduction. Non-competent species, such as *Escherichia coli*, acquire DNA from other bacteria via conjugative plasmids (conjugation) or phages (transduction) \[7, 8\].

Following uptake, DNA can be used by a cell as food or integrated in the genome by homologous. The probability of successful homologous recombination depends, firstly, on similarity of recombining segments, and, secondly, on their length \[9, 10, 11\].

The process of homologous recombination is believed to be more intensive within bacterial species than between them due to higher similarity of genomes and common environment. Thus, bacterial species should be homogeneous, but, in fact, they often form stable subspecies or phylogenetic groups \[12\]. This can be explained by a variety of reasons. Phylogroups may differ by restriction-modification systems or can occupy different local optima on the fitness landscape.

Simulations for some neutral models of speciation in a Wright-Fisher population lead to the conclusion that under certain parameters clusters can emerge \[13, 14\]. Further, it has been analytically shown that distinct populations may be maintained by the mutation and homologous recombination processes, without other factors \[15\]. However, in that study the distance between two populations was defined as the mean distance between all pairs of genomes, so if two similar populations with high dispersion form one cluster, they will still have non-zero distance between them.

Here we consider the possibility of phylogroup emergence due to solely mutations and recombination. We define a bacterial population as a set of genomes that continuously exchange genetic information via homologous recombination. For simplicity, we assume that
the genomes can be aligned throughout their entire length, so that coordinates in a genome completely define the homologous region in another genome. Below, after giving formal definitions, we write a differential equation that describes a population under mutation and recombination processes in terms of probability measures on the space of genomes, and examine its fixed points. Non-trivial fixed points of this equation correspond to population structures that can emerge in this model.

**Results**

Let $K$ be a finite alphabet (a set of nucleotides) and let a genome $x$ be a word of length $n$ over it. We consider two transformations of a genome:

1) mutations, when one letter changes to another $x_i \to y_i$, $i \in \Lambda = \{1, \ldots, n\}$;

2) homologous recombination, when a subword $x_I$ changes with a certain probability to subword $y_I$ with the same coordinates from another genome $y$.

The fundamental difference between these two transformations is that mutations occur in a genome independently of other genomes. Formally, for any position $i$ in a genome, for any nucleotides $a, b \in K, a \neq b$, there exists a transition probability $a \to b$ denoted by $\alpha_i(a, b)$. This means that for a small period of time $dt$ the probability of mutation of nucleotide $a$ to nucleotide $b$ approximately equals $\alpha_i(a, b)dt$.

Homologous recombination results from interaction of genomes in the space of genomes $X$. The recombination probability depends on the distribution of genomes in the space $X$ and a function $\varphi(x_I, y_I)$ which defines similarity between genomes $x$ and $y$ on subword $I$. This function is symmetric and non-negative. The distribution of genomes in the space of genomes $X$ is characterized by the probability distribution $\mu(x)$. Thus, the probability of substitution of a subword $x_I$ in genome $x$ to subword $y_I$ from genome $y$ equals $\kappa \varphi(x_I, y_I) \mu_I(y_I)dt$.
up to terms of order \((dt)^2\), \(\kappa\) is a constant and \(\mu_i(y_I)\) is the marginal distribution, i.e. the probability distribution of subword \(y_I\).

Importantly, the probability of recombination on a segment \(I\) in a genome depends on the probability distribution of all genomes in \(X\). Such processes are called continuous-time non-linear Markov chains, because the dependence of the probability distribution \(\mu(x)\) on time is described by a non-linear differential equation

\[
\frac{d\mu_\Lambda(x_\Lambda)}{dt} = \sum_i \sum_y \left( \alpha_i(y_i, x_i)\mu_\Lambda(x_{\Lambda\setminus i}, y_i) - \alpha_i(x_i, y_i)\mu_\Lambda(x_\Lambda) \right) + \\
\kappa \sum_I \sum_{y_I} \left( \varphi_I(y_I, x_I)\mu_I(x_I)\mu_\Lambda(x_{\Lambda\setminus I}, y_I) - \varphi_I(x_I, y_I)\mu_I(y_I)\mu_\Lambda(x_\Lambda) \right),
\]

(1)

(1) (unlike the linear Kolmogorov forward equation for usual Markov processes). The right-hand side of this equation consist of the sum of the following terms:

1) linear terms for mutations;
2) non-linear terms for segments \(I \subset \Lambda\), where recombination is possible.

Here we prove that if only mutation and recombination processes are considered and the probability of recombination \(\varphi(x_I, y_I)\) is symmetric, then for all values of other parameters, such as the ratio of the intensity of mutations and recombinations or an initial distribution of genomes, there is a unique fixed point \(q\). This fixed point, as we show below, is the stationary distribution of the Markov chain for the mutation process.

**Theorem.** Equation (1) has a unique fixed point \(q\) and all trajectories of (1) \(\mu(t) \to q\) as \(t \to \infty\).

Consider the mutation and recombination processes separately. If \(dt\) is small, the recombination process on the segment \(I\) can be described as a non-linear discrete time Markov chain on the space \(X\)
with transition probabilities
\[ P_\mu^{(I)}(x \rightarrow y) = \mu \delta(x_{\Lambda \setminus I}, y_{\Lambda \setminus I}) \varphi(x_I, y_I) \mu_I(y_I) dt \]
for \( y \neq x \) and \( P_\mu^{(I)}(x \rightarrow x) = 1 - \sum_{y \neq x} P_\mu^{(I)}(x \rightarrow y) \). Here \( \delta \) is the Kronecker delta.

Obviously, for this Markov chain the probability distribution
\[ \hat{\mu}_\Lambda(x_\Lambda) = \mu_{\Lambda \setminus I}(x_{\Lambda \setminus I}) \mu_I(x_I) \]
is an invariant measure (here it is important, that the similarity function \( \varphi(x_I, y_I) = \varphi(y_I, x_I) \) is symmetric). Moreover, any measure \( \nu_\Lambda(x_\Lambda) \) on the space \( X \), that has marginal distributions \( \mu_{\Lambda \setminus I}(x_{\Lambda \setminus I}) \) and \( \mu_I(x_I) \), turns to a measure of the same type, so the measure \( \nu P_\mu^{(I)}(x) = \sum_{y \in X} \nu_\Lambda(y) P_\mu^{(I)}(y \rightarrow x) \) has the same marginal distributions on \( I \) and \( \Lambda \setminus I \).

We formulate an inequality for finite Markov chains, although it is more general [16].

**Lemma.** Let \( P \) be a stochastic matrix, i.e. matrix \( P_{xy} \) such that \( P_{xy} \geq 0 \) and \( \sum_y P_{xy} = 1 \), and let \( \hat{\mu} \) be an invariant probability measure, \( \hat{\mu} = \mu P \). Suppose \( \hat{\mu}(x) > 0 \) for any \( x \). Then, for any probability measure \( \mu \),
\[
\sum_x \left( \ln \left( \frac{\mu P(x)}{\hat{\mu}(x)} \right) \right) (\mu P)(x) \leq \sum_x \left( \ln \left( \frac{\mu(x)}{\hat{\mu}(x)} \right) \right) \mu(x) \tag{2}
\]

In our case \( \hat{\mu}_\Lambda(x_\Lambda) = \mu_{\Lambda \setminus I}(x_{\Lambda \setminus I}) \mu_I(x_I) \), so \( \ln \hat{\mu} = \ln \mu_I(x_I) + \ln \mu_{\Lambda \setminus I}(x_{\Lambda \setminus I}) \) is a sum of functions depending only on \( x_I \) and \( x_{\Lambda \setminus I} \). Since \( P_\mu^{(I)} \), acting on measure \( \mu \), retains marginal distributions of \( x_I \) and \( x_{\Lambda \setminus I} \), it follows that
\[
\sum_x (\ln \hat{\mu}(x)) (\mu P)(x) = \sum_x (\ln \hat{\mu}(x)) \mu(x) \tag{3}
\]
Finally, we obtain the entropic inequality
\[
\sum_x (\ln (\mu P)(x)) (\mu P)(x) \leq \sum_x (\ln (\mu(x))) \mu(x) \tag{4}
\]
Now consider the mutations. The transition intensities \( \alpha_i(a,b) \) form a connected continuous-time Markov chain on alphabet \( K \), so it is possible to pass from any \( a \) to any \( b \) in several steps. By definition, \( \alpha_i(a,a) = -\sum_{b \neq a} \alpha_i(a,b) \). Matrix \( A_i = (\alpha_i(a,b), a,b \in K) \) is called the infinitesimal matrix of a time-continuous Markov chain. For this chain there exists a unique invariant distribution \( q_i(a), a \in K \) and \( q_i(a) > 0 \). In terms of matrix \( A_i \) this means that \( q_iA_i = 0 \) (by definition \( (q_iA_i)(x) = \sum_y q_i(y)\alpha_i(y,x) \)). To describe mutations in any position in the genome consider the following continuous-time Markov chain. Let \( A_\Lambda = (a_\Lambda(x,y), x,y \in X) \) be the infinitesimal matrix, \( a_\Lambda(x,y) = \sum_i \delta(x_\Lambda \setminus i, y_\Lambda \setminus i) \alpha_i(x_i, y_i) \). The invariant distribution of the chain, defined by matrix \( A_\Lambda \), is

\[
q_\Lambda(x_\Lambda) = \prod_i q_i(x_i)
\]

Obviously, this chain is connected on the space \( X \).

Finally, we use a general theorem about the entropy monotonicity (this theorem is well known from folklore and from the results of \([17]\) as a special case).

**Theorem.** Let \( \alpha_{xy} \) be the transition intensities of a connected finite continuous-time Markov chain and let \( q_x \) be its stationary distribution. Then the relative entropy \( D(p|q) = \sum_x p(x) \ln \frac{p(x)}{q(x)} \) is strictly decreasing (and, furthermore, its derivative is strictly negative) along the trajectory of the Kolmogorov forward equation \( \dot{p} = pA, A \) is the infinitesimal matrix of the considered Markov chain.

**Proof.** (for the reader’s convenience)

Let \( p(t) \) be a solution of the Kolmogorov forward equation and denote \( \frac{p_x}{q_x} \) by \( f_x \), then the derivative \( \frac{d}{dt} D(p(t)|q) \) can be written as

\[
\frac{dD}{dt} = -\sum_{x,y} \left( \frac{f_x}{f_y} \ln \frac{f_x}{f_y} - \frac{f_x}{f_y} + 1 \right) q_x \alpha_{xy} f_y
\]

Obviously, after removing parentheses the two last terms in this formula cancel out, but they are needed to prove monotonicity. The
expressions in parentheses are non-negative and, as the Markov chain is connected, simultaneously equal to 0, only if $f_x = f_y$ for all $x, y$, i.e. if the distributions $p$ and $q$ are the same.

Now collect the properties of the mutation and homologous recombination processes described above.

1) For the recombination process on segment $I$

$$H(\mu) = \sum_x \mu(x) \ln \mu(x)$$

monotonically (maybe, non-strictly) decreases, so its time derivative is non-positive.

2) For the same process, the value $\sum_x (\ln q_\Lambda) \mu(x)$ does not change, because this logarithm is the sum of functions of $x_I$ and $x_{\Lambda \setminus I}$, and as shown above, the means of such functions remain constant.

3) Hence, the relative entropy

$$D(\mu|q) = \sum_x \mu(x) \ln \frac{\mu(x)}{q(x)}$$

also has a non-positive derivative.

4) For the mutation process, the relative entropy $D(\mu|q)$ has a strictly negative derivative.

The right part of equation (1) consist of terms for the recombination process on all segments $I$ and the mutation process. Since the relative entropy $D(p|q)$ has a non-positive derivative by equations for the recombination process and a strictly negative derivative for the mutation process, the derivative of $D(p|q)$ by equation (1) is strictly negative, if $p \neq q$. This means, that $D(p|q)$ strictly decreases along the trajectory of equation (1) and this equation has a unique fixed point $q$. As we noted above, fixed points of equation (1) correspond to different population structures. A unique fixed point $q$ depends only on the substitution matrix for the mutation process, so it gives us a population without any structure – if $q_i$ do not depend on $i$, then the
probability of a genome depends only on its nucleotide frequencies. Note that if the similarity function $\varphi(x_I, y_I)$ and the constant $\kappa$ do depend on time, it would not affect the above calculations.

**Conclusions**

Our results are consistent with simulations in [13] with one difference. When the recombination rate is low, mutations lead to an increase of variance in a mostly clonal population, otherwise clusters are mixed by recombination. However, in that study genetic drift may cause speciation by chance as in [14], if the recombination rates vary appreciably between members of the population.

Here we do not examine the behaviour of a system in time, so we can not claim that clusters may not emerge temporarily, but there is no force, which could maintain them. The model is general, in particular, it accommodates different types of dependence of recombination rate on sequence similarity, e.g. log-linear [11]. However, the symmetry of the function $\varphi$ is a strong restriction and it seems to be weakly related to natural populations. For example, in the case of conjugative plasmids, the probabilities of DNA transfer between $F^+$ and $F^-$ cells in different directions are not equal [7], and hence $F^+$ genomes may form clusters.

**Acknowledgements**

This study was partially supported by grants from RAS (programs ”Molecular and Cellular Biology” and ”Living Nature and Biodiversity”) and RFBR (12-04-01105).

**References**

[1] Griffith F. The Significance of Pneumococcal Types. J Hyg (Lond). Jan;27(2):113-59 (1928).
[2] Avery O.T., MacLeod C.M., McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med. Feb 1;79(2):137-58 (1944).

[3] Thomas C.M., Nielsen K.M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. Sep;3(9):711-21 (2005).

[4] Jonas D.A. et al. Safety considerations of DNA in food. Ann. Nutr. Metab. 45, 235254 (2001).

[5] Berndt C., Meier P., Wackernagel W. DNA restriction is a barrier to natural transformation in Pseudomonas stutzeri JM300. Microbiology 149, 895901 (2003).

[6] Budroni S. et al. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci U S A. Mar 15;108(11):4494-9 (2011).

[7] Arutyunov D., Frost L.S. F conjugation: back to the beginning. Plasmid. Jul;70(1):18-32 (2013).

[8] Weinbauer M.G, Rassoulzadegan F. Are viruses driving microbial diversification and diversity? Environ Microbiol. Jan;6(1):1-11 (2004).

[9] Shen P., Huang H.V. Homologous recombination in Escherichia coli: dependence on substrate length and homology. Genetics. Mar;112(3):441-57 (1986).

[10] Majewski J., Cohan F.M. DNA sequence similarity requirements for interspecific recombination in Bacillus. Genetics. Dec;153(4):1525-33 (1999).
[11] Vuli M., Dionisio F., Taddei F., Radman M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci U S A. Sep 2;94(18):9763-7 (1997).

[12] Chaudhuri R.R., Henderson I.R. The evolution of the *Escherichia coli* phylogeny. Infect Genet Evol. Mar;12(2):214-26 (2012).

[13] Fraser C., Hanage W.P., Spratt B.G. Recombination and the nature of bacterial speciation. Science. Jan 26;315(5811):476-80 (2007).

[14] Falush D. et al. Mismatch induced speciation in Salmonella: model and data. Philos Trans R Soc Lond B Biol Sci. Nov 29;361(1475):2045-53 (2006).

[15] Doroghazi J.R., Buckley D.H. A model for the effect of homologous recombination on microbial diversification. Genome Biol Evol. 3:1349-56 (2011).

[16] Yosida K. Functional analysis (Russian). Moscow, Mir 1967 (transl. from English edition).

[17] Batishcheva Ya.G., Vedenyapin V.V. The second law of thermodynamics for chemical kinetics Mat. Model. 17 (8) 106-110 (2005).