BMC Evolutionary Biology

Research article

A model of evolution with constant selective pressure for regulatory DNA sites

Farida N Enikeeva*[1], Ekaterina A Kotelnikova[2,4], Mikhail S Gelfand[1,3] and Vsevolod J Makeev[2,5]

Address: [1]Institute for Information Transmission Problems (the Kharkevich Institute) of RAS, Bolshoi Karetny pereulok, 19, GSP-4, Moscow, 127994, Russia, [2]State Research Institute of Genetics and Selection of Industrial Microorganisms, 1st Dorozhnyj proezd, 1, Moscow, 113535, Russia, [3]Faculty of Bioengineering and Bioinformatics, Moscow State University, Vorobyevy Gory 1-73, Moscow, 119992, Russia, [4]Ariadne Genomics Inc. 9700 Great Seneca Highway, Suite 113, Rockville, MD 20850, USA and [5]Engelgardt Institute of Molecular Biology of RAS, Vavilova 32, Moscow, 119991, Russia

Email: Farida N Enikeeva* - enikeeva@iitp.ru; Ekaterina A Kotelnikova - ekotelnikova@gmail.com; Mikhail S Gelfand - gelfand@iitp.ru; Vsevolod J Makeev - makeev@genetika.ru

* Corresponding author

Published: 27 July 2007

BMC Evolutionary Biology 2007, **7**:125 doi:10.1186/1471-2148-7-125

Received: 18 January 2007
Accepted: 27 July 2007

This article is available from: http://www.biomedcentral.com/1471-2148/7/125

Abstract

Background: Molecular evolution is usually described assuming a neutral or weakly non-neutral substitution model. Recently, new data have become available on evolution of sequence regions under a selective pressure, e.g. transcription factor binding sites. To reconstruct the evolutionary history of such sequences, one needs evolutionary models that take into account a substantial constant selective pressure.

Results: We present a simple evolutionary model with a single preferred (consensus) nucleotide and the neutral substitution model adopted for all other nucleotides. This evolutionary model has a rate matrix in which all substitutions that do not involve the consensus nucleotide occur with the same rate. The model has two time scales for achieving a stationary distribution; in the general case only one of the two rate parameters can be evaluated from the stationary distribution. In the middle-time zone, a counterintuitive behavior was observed for some parameter values, with a probability of conservation for a non-consensus nucleotide greater than that for the consensus nucleotide. Such an effect can be observed only in the case of weak preference for the consensus nucleotide, when the probability to observe the consensus nucleotide in the stationary distribution is less than 1/2. If the substitution rate is represented as a product of mutation and fixation, only the fixation can be calculated from the stationary distribution. The exhibited conservation of non-consensus nucleotides does not take place if the elements of mutation matrix are identical, and can be related to the reduced mutation rate between the non-consensus nucleotides. This bias can have no effect on the stationary distribution of nucleotide frequencies calculated over the ensemble of multiple alignments, e.g. transcription factor binding sites upstream of different sets of co-regulated orthologous genes.

Conclusion: The derived model can be used as a null model when analyzing the evolution of orthologous transcription factor binding sites. In particular, our findings show that a nucleotide preferred at some position of a multiple alignment of binding sites for some transcription factor in the same genome is not necessarily the most conserved nucleotide in an alignment of orthologous sites from different species. However, this effect can take place only in the case of a mutation matrix whose elements are not identical.

Background

The controlled expression of genes is the main mechanism responsible for the life cycle and biodiversity [1]. Transcription, which is a crucial process defining the level of gene expression, is regulated by interaction of transcription factor proteins (TFs) with transcription factor binding sites (TFBSs) in a DNA molecule [2]. Thus, adaptable interaction between TFs and TFBSs is one of the main driving forces of biological evolution [3].

Both TFs and TFBSs are subject to mutations and selection affecting their interaction [4]. In this study we focus on mutations in TFBSs. New experimental [5-7] and computational [8] methods of TFBS identification produce an increasing amount of data about TFBS sequences, which creates a possibility to study evolutionary events in these regions.

Modelling evolution of regulatory sequences can be useful for understanding both the general mechanisms of gene expression control and the regulatory history of particular genes.

The evolution of regulatory regions has a complex pattern [3,9-11] and is still unclear in many aspects. DNA segments can gain and lose the TFBS function [12], and that can bring new genes under regulation by a particular TF [13-15], or divert regulation from other genes [16,17]. One particular type of events is emergence of new or changed sites following displacement of a transcription factor by horizontal transfer [18]. Sometimes this leads to considerable changes in the regulon content [19,20] or even partial or complete rewiring of regulatory cascades [21-26], reviewed in [27].

Evolution of functional TFBS sequences is strongly non-neutral [11,28] and under a positive selection [29], which makes it difficult to calculate the rate of TFBS evolution. This rate varies between TFBS positions [30,31]. Moreover, co-evolution of TFBS and TF [16,20] can make the selective pressure vary in different lineages. Indeed, although in some cases the DNA motif bound by orthologous factors may be conserved at surprisingly large evolutionary distances [14,32-34], in other cases not only the motifs themselves may be different [35,36], but even the symmetry of the motif (e.g. palindrome or direct repeat) may change [37,38].

The existing evolutionary models, which were successful in reconstruction of phylogenetic relations, can be applied to evolution of regulatory sequences only with a caution. Such models are historically related to the Jukes–Cantor [39] and Kimura [40] models of molecular evolution. Existing modifications of these models take into account various global characteristics like transition/transversion rate or local GC composition [41-44]. They are not applicable to the case of strong selection for a specific nucleotide at a particular position.

On the other hand, models developed specifically for the evolution of TFBS are needed to reconstruct the evolutionary origin of a particular TFBS and to evaluate the position-specific mutation rate and selective pressure.

Because of the position-specific variation in the rate of TFBS evolution [30], the rate matrix must also be position-specific. The data produced by mass experiments on TFBS identification or comparative genomic studies produce tens of TFBS for each TF (more exactly, a group of orthologous TFs). That might be sufficient to evaluate the evolutionary rate at each TFBS position.

Here we consider the simplest model of position-specific evolution with one preferred (consensus) nucleotide and three other (minor) nucleotides, the latter considered in a symmetric setting, without any selection or rate preferences [45]. Such a model can be deduced from physical requirements of the TF/TFBS interaction [46] and can explain the observed TFBS fuzziness.

We build a rate matrix, which enhances the model of [45]. We calculate the substitution probability for each finite time and show that the nucleotide conservation in phylogenetic lineages can be non-trivial for some parameter values. Particularly, a non-consensus nucleotide may appear more conserved than the consensus nucleotide, although the latter has a selective preference. This happens when the rate of mutations between non-consensus nucleotides is lower than the rate of mutation into the consensus, or there is selection against any mutation in non-consensus nucleotides.

Results and Discussion

Model

We start with definitions. We consider an alignment of several sequences; all positions in this alignment are assumed to be independent, and thus may be modelled independently. *Consensus nucleotide* (or simply *consensus*) is the most frequent nucleotide in an alignment column (*position*). Other nucleotides are called *non-consensus*. The frequency of the consensus nucleotide N_c is a fraction of the number of consensus nucleotides in a particular position. Obviously, $1/4 < N_c \leq 1$. The consensus is called *weak*, if $1/4 < N_c < 1/2$; the consensus is *strong*, if $1/2 \leq N_c < 1$.

Consider the model of nucleotide substitutions given by a Markov process $X(t)$ with four states $\{g_1, g_2, g_3, g_4\}$. Without loss of generality, assume that the state g_1 is the consensus state and the states g_2, g_3, g_4 are equiprobable non-consensus states. Suppose that the transition rate matrix $A = (q_{ij})$ is given by

$$A = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \beta & -\beta - 2\delta & \delta & \delta \\ \beta & \delta & -\beta - 2\delta & \delta \\ \beta & \delta & \delta & -\beta - 2\delta \end{pmatrix},$$

where q_{ij} is the transition rate from the state g_i to g_j and α, β, and δ are positive unknown parameters.

Transitional probabilities

Let $\mathbf{P}_{ij}(t)$ be the transitional probability from the state g_i to the state g_j for the time t. From the theory of Markov chains we have for $h \to 0$

$$\begin{aligned} 1 - \mathbf{P}_{ii}(h) &= -q_{ii}h + o(1), \\ \mathbf{P}_{ij}(h) &= q_{ij}h + o(1), \quad i \neq j \end{aligned}$$

For brevity we denote by 'c' the subscript '1' corresponding to the consensus state g_1 and by 'n' and 'm' the subscripts 2, 3, and 4 corresponding to the non-consensus states g_2, g_3, g_4. Thus we have five transitional probabilities $\mathbf{P}_{cc}$, $\mathbf{P}_{cn}$, $\mathbf{P}_{nc}$, $\mathbf{P}_{nn}$, and $\mathbf{P}_{nm}$, where n and m stand for two distinct non-consensus states. The formulas for $\mathbf{P}_{ij}$ are derived from simple calculation:

$$\begin{aligned} \mathbf{P}_{cc}(t) &= \frac{\beta}{3\alpha+\beta} + \frac{3\alpha}{3\alpha+\beta} e^{-(3\alpha+\beta)t}, \\ \mathbf{P}_{cn}(t) &= \frac{\alpha}{3\alpha+\beta} - \frac{\alpha}{3\alpha+\beta} e^{-(3\alpha+\beta)t}, \\ \mathbf{P}_{nc}(t) &= \frac{\beta}{3\alpha+\beta} - \frac{\beta}{3\alpha+\beta} e^{-(3\alpha+\beta)t}, \\ \mathbf{P}_{nn}(t) &= \frac{\alpha}{3\alpha+\beta} + \frac{\beta}{3(3\alpha+\beta)} e^{-(3\alpha+\beta)t} + \frac{2}{3} e^{-(3\delta+\beta)t}, \\ \mathbf{P}_{nm}(t) &= \frac{\alpha}{3\alpha+\beta} + \frac{\beta}{3(3\alpha+\beta)} e^{-(3\alpha+\beta)t} - \frac{1}{3} e^{-(3\delta+\beta)t}. \end{aligned}$$

For simplicity, introduce a new time-scale, $u = t\beta$, and denote $\lambda = \alpha/\beta$, $\mu = \delta/\beta$. Then

$$\begin{aligned} \mathbf{P}_{cc}(u) &= \frac{1}{3\lambda+1} + \frac{3\lambda}{3\lambda+1} e^{-(3\lambda+1)u}, \\ \mathbf{P}_{cn}(u) &= \frac{\lambda}{3\lambda+1} - \frac{\lambda}{3\lambda+1} e^{-(3\lambda+1)u}, \\ \mathbf{P}_{nc}(u) &= \frac{1}{3\lambda+1} - \frac{1}{3\lambda+1} e^{-(3\lambda+1)u}, \\ \mathbf{P}_{nn}(u) &= \frac{\lambda}{3\lambda+1} + \frac{1}{3(3\lambda+1)} e^{-(3\lambda+1)u} + \frac{2}{3} e^{-(3\mu+1)u}, \\ \mathbf{P}_{nm}(u) &= \frac{\lambda}{3\lambda+1} + \frac{1}{3(3\lambda+1)} e^{-(3\lambda+1)u} - \frac{1}{3} e^{-(3\mu+1)u} \end{aligned}$$

and

$$A = \beta \cdot \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ 1 & -1-2\mu & \mu & \mu \\ 1 & \mu & -1-2\mu & \mu \\ 1 & \mu & \mu & -1-2\mu \end{pmatrix}.$$

Note that the parameter μ is related to the conservation within the set of non-consensus states $\{g_2, g_3, g_4\}$. Simply, λ is a rate of transition from the consensus nucleotide and μ is a rate of transition between non-consensus states up to the time-scale parameter β.

The stationary distribution $\pi = (\pi_c, \pi_n, \pi_n, \pi_n)$ of the process $X(t)$ is given by

$$\pi = \left(\frac{1}{3\lambda+1}, \frac{\lambda}{3\lambda+1}, \frac{\lambda}{3\lambda+1}, \frac{\lambda}{3\lambda+1} \right).$$

According to our definition of the consensus as the most frequent nucleotide, from $\pi_c = 1/(3\lambda + 1) > 1/4$ it follows that $0 < \lambda < 1$. The parameter λ is responsible for transitions from the consensus state to the non-consensus ones. At the same time, the transition from a non-consensus state to the consensus state occurs with the rate 1 (up to the time-scale parameter β). Intuitively, in the model with one selected consensus state the probability of transition to a non-consensus state should be smaller than the probability of transition to the consensus state; thus, λ should be less than 1. Indeed, if $\lambda \geq 1$, then $\pi_c \leq 1/4$, $\pi_n \geq 1/4$ and we have three equiprobable consensus states.

If the consensus is strong, then $\pi_c \geq 1/2$, and, consequently, $0 < \lambda \leq 1/3$. Note that we can not impose any condition on μ other than $\mu > 0$.

Conservation and transition of nucleotides

The goal of this section is to compare the transitional probabilities between different states in our model. We will see that the relations we get have clear biological interpretation.

Simple cases

Let a unique consensus exist in our model, i.e. $0 < \lambda < 1$. It is not difficult to obtain the following relations between the transitional probabilities, which hold for any $\lambda \in (0, 1)$, $u > 0$.

1. $\mathbf{P}_{cc}(u) > \mathbf{P}_{cn}(u)$;

2. $\mathbf{P}_{cc}(u) > \mathbf{P}_{nc}(u)$;

3. $\mathbf{P}_{cc}(u) > \mathbf{P}_{nm}(u)$ for any $\mu > 0$;

4. $\mathbf{P}_{nn}(u) > \mathbf{P}_{cn}(u)$ for any positive λ and μ;

5. $\mathbf{P}_{nc}(u) > \mathbf{P}_{cn}(u)$;

6. $\mathbf{P}_{nn}(u) > \mathbf{P}_{nm}(u)$ for any $\mu > 0$.

The interpretation of these results is straightforward. Figures 1, 2, 3, 4 show the graphs of the transitional probabilities for different values of λ and μ. The relations 1–6 between the transitional probabilities are clearly shown in the figures. The first three inequalities show that the probability of conservation of the consensus state is always higher than the probability of transition between the

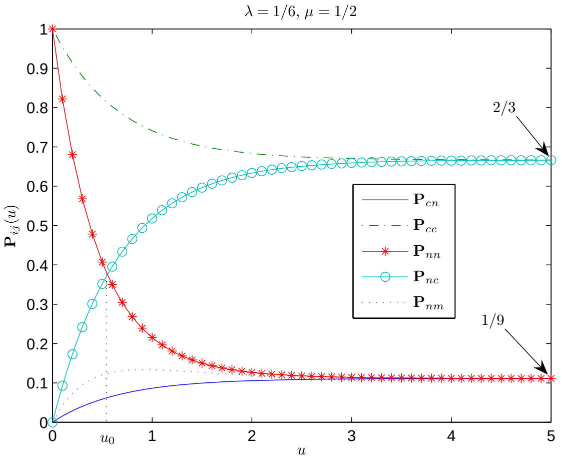

Figure 1

Transitional probabilities $\mathbf{P}_{cn}$, $\mathbf{P}_{cc}$, $\mathbf{P}_{nn}$, $\mathbf{P}_{nc}$, $\mathbf{P}_{nm}$ for $\lambda = 1/6$, $\mu = 1/2$. This figure describes the case $\mathbf{P}_{cc}(u) > \mathbf{P}_{nn}(u)$ for all $u > 0$. The consensus is strong, since $\pi_c = 2/3$. Since the rate of transition between non-consensus states μ is greater than the rate of transition from the consensus state λ, we have $\mathbf{P}_{cn}(u) < \mathbf{P}_{nm}(u)$. The relation between the probability of conservation of a non-consensus nucleotide and the probability of transition from the non-consensus state to the consensus one is shown, $\mathbf{P}_{nn}(u) > \mathbf{P}_{nc}(u)$ for $u \in (0, u_0)$ and, vice versa, $\mathbf{P}_{nn}(u) < \mathbf{P}_{nc}(u)$ for $u > u_0$. This figure exemplifies all simple cases that hold for any $\lambda \in (0, 1)$, $\mu > 0$, $u > 0$ (see the corresponding paragraph in the main text).

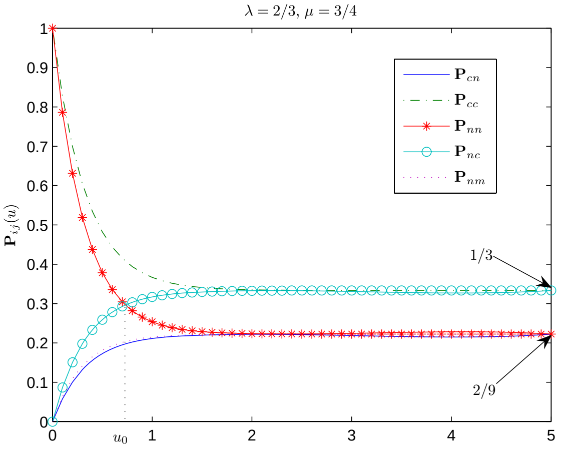

Figure 2

Transitional probabilities $\mathbf{P}_{cn}$, $\mathbf{P}_{cc}$, $\mathbf{P}_{nn}$, $\mathbf{P}_{nc}$, $\mathbf{P}_{nm}$ for $\lambda = 2/3$, $\mu = 3/4$. Here we have $\mathbf{P}_{nn}(u) > \mathbf{P}_{nc}(u)$ for $u \in (0, u_0)$ and, vice versa, $\mathbf{P}_{nn}(u) < \mathbf{P}_{nc}(u)$ for $u > u_0$. In this figure $\mathbf{P}_{cc}(u) > \mathbf{P}_{nn}(u)$ for all $u > 0$, as $3\lambda \le 2\mu + 1$. The consensus is weak, since $\pi_c = 1/3$. Since $\mu > \lambda$, we have $\mathbf{P}_{cn}(u) < \mathbf{P}_{nm}(u)$. This figure shows all simple cases that hold for any $\lambda \in (0, 1)$, $\mu > 0$, $u > 0$.

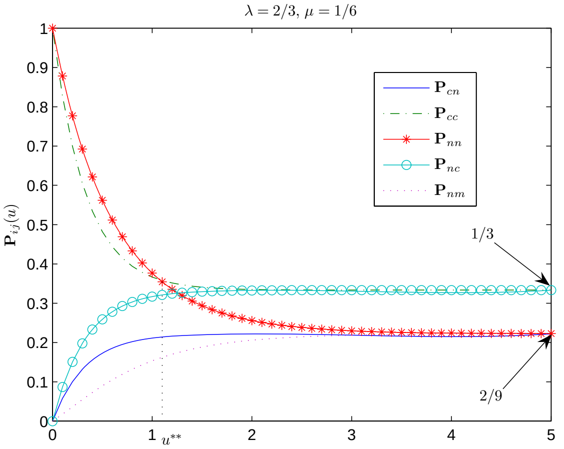

Figure 3

Transitional probabilities $\mathbf{P}_{cn}$, $\mathbf{P}_{cc}$, $\mathbf{P}_{nn}$, $\mathbf{P}_{nc}$, $\mathbf{P}_{nm}$ for $\lambda = 2/3$, $\mu = 1/6$. In this case the probability of conservation of the consensus nucleotide is less than the probability of conservation of a non-consensus nucleotide, $\mathbf{P}_{cc}(u) < \mathbf{P}_{nn}(u)$ for $u \in (0, u^{**})$, since $3\lambda > 2\mu + 1$ and $\lambda > \mu$. The stationary distribution of the consensus state is $\pi_c = 1/3$. Thus, the consensus is weak. Since $\mu < \lambda$, we have $\mathbf{P}_{cn}(u) > \mathbf{P}_{nm}(u)$ for all $u > 0$.

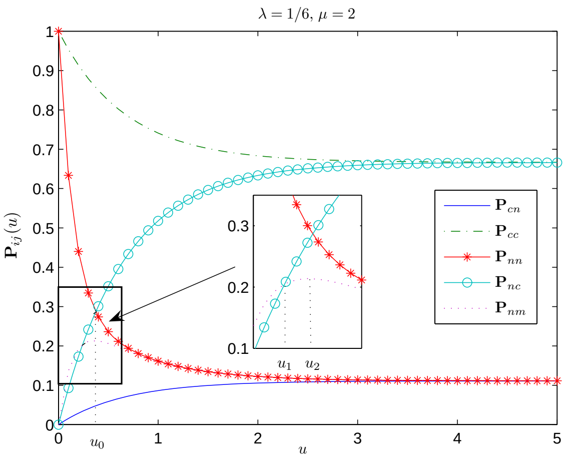

Figure 4
Transitional probabilities $\mathbf{P}_{cn}$, $\mathbf{P}_{cc}$, $\mathbf{P}_{nn}$, $\mathbf{P}_{nc}$, $\mathbf{P}_{nm}$ for $\lambda = 1/6$, $\mu = 2$. Here we have $\mathbf{P}_{cc}(u) > \mathbf{P}_{nn}(u)$ for all $u > 0$, since $3\lambda \leq 2\mu + 1$. The consensus is strong and the stationary distribution of the consensus state is $\pi_c = 2/3$. Since $\mu > \lambda$, we have $\mathbf{P}_{cn}(u) < \mathbf{P}_{nm}(u)$. The probability of substitutions between non-consensus nucleotides $\mathbf{P}_{nm}(u)$ is not monotone with a maximum point at $u_2 = 2 \log 7/11 \approx 0.354$. Since $\mu > 1$, transitions between non-consensus nucleotides occur more frequently than transitions from a non-consensus to the consensus nucleotide, $\mathbf{P}_{nc}(u) < \mathbf{P}_{nm}(u)$ on the time interval $u \in (0, u_1)$, whereas $\mathbf{P}_{nc}(u) > \mathbf{P}_{nm}(u)$ for $u > u_1$. This figure also shows the relation between $\mathbf{P}_{nn}(u)$ and $\mathbf{P}_{nc}(u)$.

states of different type. Inequalities 1, 4, and 5 imply that the probability of transition from the consensus state to a non-consensus one is always less than the probability of conservation of a nucleotide or the probability of transition from a non-consensus state to the consensus one. Inequalities 3 and 6 show that the probability of transition between two different non-consensus states is always less than the probability of state conservation. All these results correlate well with our intuitive ideas about an evolutionary model with selective pressure. Note that relations 1–6 hold for any positive μ, $\lambda \in (0, 1)$ and for any time period.

Interesting cases

The most interesting case is the conservation of the consensus or a non-consensus nucleotide. Recall that $\mathbf{P}_{\mathrm{cc}}(\mathrm{h}) = 1 - 3\lambda\beta\mathrm{h} + \mathrm{o}(1)$, $\mathbf{P}_{\mathrm{nn}}(\mathrm{h}) = 1 - (2\mu + 1)\beta\mathrm{h} + \mathrm{o}(1)$ for $\mathrm{h} \to 0$. Thus, the relation between the probabilities of conservation of the consensus state and conservation of a non-consensus state during the time u depends on the relation between 3λ and $2\mu + 1$. We obtained the following result.

1. If $3\lambda \leq 2\mu + 1$, then

$$\mathbf{P}_{cc}(u) > \mathbf{P}_{nn}(u) \text{ for } u > 0.$$

2. If $3\lambda > 2\mu + 1$, $\lambda > \mu$, then

$$\begin{aligned} \mathbf{P}_{cc}(u) &< \mathbf{P}_{nn}(u) \quad \text{for } u \in (0, u^{**}), \\ \mathbf{P}_{cc}(u) &> \mathbf{P}_{nn}(u) \quad \text{for } u > u^{**}. \end{aligned}$$

Here $u^{**} = u^{**}(\lambda, \mu) > 0$ is a time moment depending on λ and μ. More precisely, u^{**} is a non-zero solution of the equation $\mathbf{P}_{cc}(u) = \mathbf{P}_{nn}(u)$. Let $F(u) = \mathbf{P}_{cc}(u) - \mathbf{P}_{nn}(u)$. It can be shown that for $\lambda > (2\mu + 1)/3$ and $\lambda > \mu$

$$u^{**} > u^{*} = \frac{1}{3(\lambda - \mu)} \log\left[\frac{9\lambda - 1}{2(3\mu + 1)}\right] > 0.$$

Indeed,

$$F(u) = \frac{1-\lambda}{3\lambda + 1} + \frac{9\lambda - 1}{3(3\lambda + 1)} e^{-(3\lambda+1)u} - \frac{2}{3} e^{-(3\mu+1)u},$$

and $F'(u) = 0$ if and only if

$$2(3\mu + 1)\, e^{-(3\mu+1)u} = (9\lambda - 1)e^{-(3\lambda+1)u}.$$

The solution of the latter equation is u^*. If $3\lambda > 2\mu + 1$ and $\lambda > \mu$, then $u^* > 0$ and F increases for $u > u^*$ and decreases for $u < u^*$. At the same time, $F(0) = 0$. Thus, in this case $F(u) < 0$ for $u \in (0, u^{**})$ and $F(u) > 0$ for $u > u^{**}$. This implies the second statement.

Further, if $3\lambda \leq 2\mu + 1$ and $\lambda > \mu$, then $u^* < 0$ and $F'(u) > 0$ for all $u > 0$. Therefore, $F(u)$ increases for $u > 0$ and $F(u) > 0$, since $F(0) = 0$. Next, for $\lambda = \mu$ we obtain that

$$F(u) \geq 3(1 - \lambda)(1 - e^{-(3\lambda+1)u}) > 0$$

for any $u > 0$. Thus, $F(u) > 0$ if $3\lambda \leq 2\mu + 1$, and the first inequality follows.

The second relation partly describes the case of the weak consensus (Fig. 3). Since $3\lambda + 1 > 2(\mu + 1)$, the stationary distribution of the consensus state is estimated from above as

$$\pi_c = \frac{1}{3\lambda + 1} < \frac{1}{2(\mu + 1)} < \frac{1}{2}.$$

At the same time, the first relation holds both in the case of strong consensus $\pi_c \geq 1/2$ (Fig. 1) and in the case of weak consensus $1/4 < \pi_c < 1/2$ (Fig. 2).

The second interesting result is a relation between the probability of conservation of the non-consensus nucleotide and the probability of transition from a non-consensus state to the consensus state. We have

$$\mathbf{P}_{nn}(u) > \mathbf{P}_{nc}(u), \quad u \in (0, u_0),$$
$$\mathbf{P}_{nn}(u) < \mathbf{P}_{nc}(u), \quad u > u_0,$$

where $u_0 \equiv u_0(\lambda, \mu)$ is the solution of equation $\mathbf{P}_{nn}(u) = \mathbf{P}_{nc}(u)$. It can be shown that u_0 is always positive. This relation shows that the probability of conservation of a non-consensus nucleotide is less than the probability of transition to the consensus nucleotide from a non-consensus one for sufficiently long time $u > u_0$. However, on the interval $(0, u_0)$ the opposite inequality holds (see Fig. 1, 2).

Less interesting cases

It remains to study the relations between $\mathbf{P}_{nm}$ and $\mathbf{P}_{cn}$, $\mathbf{P}_{nc}$. From the practical point of view, these cases are not very interesting, since the events of transition between different non-consensus states can be hardly observed on practice. However, we consider them for the sake of completeness.

The following relations imply that the transitional probability from the consensus to a non-consensus state could be less or greater than the transitional probability between two different non-consensus states depending on the relation between μ and λ (see Fig. 2 and 3):

$$\mathbf{P}_{cn}(u) < \mathbf{P}_{nm}(u), \quad \mu > \lambda,$$
$$\mathbf{P}_{cn}(u) = \mathbf{P}_{nm}(u), \quad \mu = \lambda,$$
$$\mathbf{P}_{cn}(u) > \mathbf{P}_{nm}(u), \quad \mu < \lambda,$$

Indeed, if $\mu > \lambda$, then $\delta > \alpha$ and the transition rate d between the non-consensus states is greater than the transition rate from the consensus state to non-consensus. Clearly, in this case $\mathbf{P}_{nm}$ has to be greater than $\mathbf{P}_{cn}$ (see Fig. 2 and Fig. 3).

The next case concerns the relation $\mathbf{P}_{nm}$ and $\mathbf{P}_{nc}$ for $\lambda \in (0, 1)$ (see Fig. 1 and 4):

$$\mathbf{P}_{nc}(u) > \mathbf{P}_{nm}(u), \quad \mu \in (0,1), \quad u > 0$$
$$\mathbf{P}_{nc}(u) < \mathbf{P}_{nm}(u), \quad \mu \geq 1, \quad u \in (0, u_1)$$
$$\mathbf{P}_{nc}(u) > \mathbf{P}_{nm}(u), \quad \mu \geq 1, \quad u > u_1,$$

where $u_1 = u_1(\lambda, \mu)$ is a solution of the equation $\mathbf{P}_{nc}(u) = \mathbf{P}_{nm}(u)$.

An interesting fact is that the probability $\mathbf{P}_{nm}(u)$ is not monotone in u for $\mu > \lambda$ (see Fig. 4). If $\mu > \lambda$, then $\mathbf{P}_{nm}$ increases on $(0, u_2)$ and decreases for $u > u_2$, where

$$u_2 = \frac{1}{3(\mu - \lambda)} \log(3\mu + 1).$$

Finally, consider the case $\mu = \lambda$. As it has been shown above, $\mathbf{P}_{cc}(u) > \mathbf{P}_{nn}(u)$ for $u > 0$ (see Fig. 5). It is easy to see that in this case $\mathbf{P}_{nm} \equiv \mathbf{P}_{cn}$. The function $\mathbf{P}_{nm}(u)$ is monotonically increasing for all u. Note that in the degenerate case $\lambda = \mu = 1$ all states are equiprobable, the stationary distribution of the process $\pi = (1/4, 1/4, 1/4, 1/4)$ and $\mathbf{P}_{nn} = \mathbf{P}_{cc}$, $\mathbf{P}_{nc} = \mathbf{P}_{cn} = \mathbf{P}_{nm}$ (see Fig. 6).

Generalization

The obtained results can be generalized on the following model. Consider a Markov chain $\tilde{X}(t)$ with $M + N$ states $\{g_1, ..., g_M, g_{M+1}, ..., g_{M+N}\}$, where the states $g_1, ..., g_M$ are consensus states, and the states $g_{M+1}, ..., g_{M+N}$ are non-consensus ones. Define the transition rate matrix $\tilde{A} = (\tilde{q}_{ij})$ for this process by

$$\tilde{q}_{ij} = \begin{cases} \alpha, & i, j = 1, \dots, M, \quad i \neq j \\ \gamma, & i = 1, \dots, M, \quad j = M+1, \dots, M+N \\ \beta, & i = M+1, \dots, M+N, \quad j = 1, \dots, M \\ \delta, & i, j = M+1, \dots, M+N, \quad i \neq j, \end{cases}$$

$$\tilde{q}_{ii} = \begin{cases} -(M-1)\alpha - N\gamma, & i = 1, \dots, M \\ -M\beta - (N-1)\delta, & i = M+1, \dots, M+N, \end{cases}$$

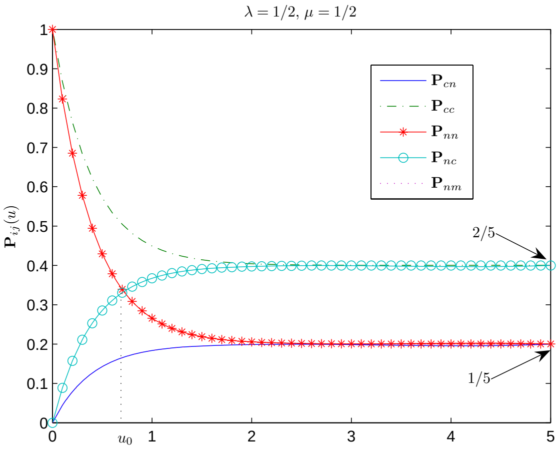

Figure 5

Transitional probabilities for $\lambda = \mu = 1/2$. This case is similar to the one of Fig. 1, since here $3\lambda \leq 2\mu + 1$ and, consequently, the probability of consensus conservation is always greater than the probability of conservation of a non-consensus nucleotide, $\mathbf{P}_{cc}(u) > \mathbf{P}_{nn}(u)$ for all $u > 0$. As $\lambda = \mu$, we have $\mathbf{P}_{nm} = \mathbf{P}_{cn}$.

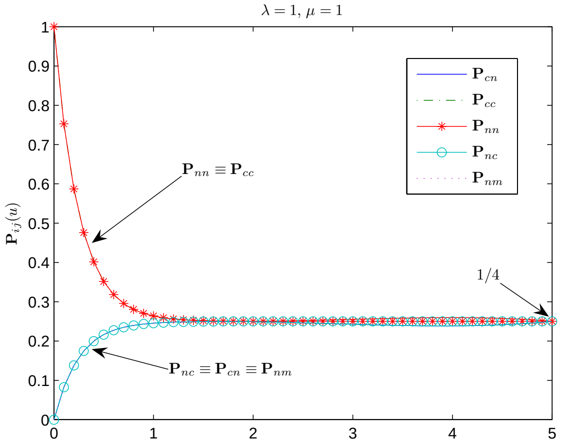

Figure 6
Degenerate case λ = μ = 1. In this case all states are equiprobable, the stationary distribution of the process π = (1/4, 1/4, 1/4, 1/4) and $\mathbf{P}_{nn} \equiv \mathbf{P}_{cc}$, $\mathbf{P}_{nc} \equiv \mathbf{P}_{cn} \equiv \mathbf{P}_{nm}$.

where α, β, γ, δ, are positive unknown parameters.

In this case we have two groups of M consensus and N non-consensus states. We use the same notation for subscripts of transitional probabilities $\tilde{\mathrm{P}}_{ij}$. If c and d denote different consensus states, n and m stand for different non-consensus states as before, we have the following transitional probabilities:

$$\mathrm{P}_{cc}(t) = \frac{\beta}{\beta M + \gamma N} + \frac{\gamma N}{M(\beta M + \gamma N)} e^{-(\beta M + \gamma N)t} + \frac{M-1}{M} e^{-(\alpha M + \gamma N)t},$$

$$\mathrm{P}_{cd}(t) = \frac{\beta}{\beta M + \gamma N} + \frac{\gamma N}{M(\beta M + \gamma N)} e^{-(\beta M + \gamma N)t} - \frac{1}{M} e^{-(\alpha M + \gamma N)t},$$

$$\mathrm{P}_{nc}(t) = \frac{\beta}{\beta M + \gamma N} - \frac{\beta}{\beta M + \gamma N} e^{-(\beta M + \gamma N)t},$$

$$\mathrm{P}_{cn}(t) = \frac{\gamma}{\beta M + \gamma N} - \frac{\gamma}{\beta M + \gamma N} e^{-(\beta M + \gamma N)t},$$

$$\mathrm{P}_{nn}(t) = \frac{\gamma}{\beta M + \gamma N} + \frac{\beta M}{N(\beta M + \gamma N)} e^{-(\beta M + \gamma N)t} + \frac{N-1}{N} e^{-(\beta M + \delta N)t},$$

$$\mathrm{P}_{nm}(t) = \frac{\gamma}{\beta M + \gamma N} + \frac{\beta N}{N(\beta M + \gamma N)} e^{-(\beta M + \gamma N)t} - \frac{1}{N} e^{-(\beta M + \delta N)t}.$$

Thus, this process has the stationary distribution $\tilde{\pi}$ = (π_c, ..., π_c, π_n, ..., π_n) with

$$\pi_c = \frac{\beta}{\beta M + \gamma N}, \quad \pi_n = \frac{\gamma}{\beta M + \gamma N}.$$

Clearly, $\beta > \gamma$, since the first M states are the consensus ones. Next, compare the probabilities of conservation of the consensus and non-consensus states $\mathbf{P}_{cc}$ and $\mathbf{P}_{nn}$. In a similar way, we obtain that there exists $t^* > 0$ such that $\mathbf{P}_{cc}(t) < \mathbf{P}_{nn}(t)$ for $t \in (0, t^*)$. Analyzing the probabilities $\mathbf{P}_{cc}$ and $\mathbf{P}_{nn}$ we can show that this is possible only for $\delta(N - 1) - \alpha(M - 1) + \beta M - \gamma N < 0$. Then the frequency of the consensus state (stationary distribution of consensus) is estimated from above as

$$\pi_c = \frac{\beta}{\beta M + \gamma N} < \frac{\beta}{2\beta M + \delta(N-1) - \alpha(M-1)}.$$

If M = 1, N = 3, then

$$\pi_c < \frac{\beta}{2\beta + 2\delta} < \frac{1}{2}.$$

This condition coincides with the condition on weak consensus obtained for the case of four states with one consensus (M = 1, N = 3) considered above.

Comparison with the Molecular Evolution Theory

In the framework of the molecular evolution theory, the element a_{ij} of the transition rate matrix is considered to be proportional to the product of the mutation rate p_{ij} and the probability of fixation of a mutation f_{ij}, $a_{ij} = kp_{ij}f_{ij}$, where k is an arbitrary scaling constant [47]. As in [45] we start from the simplest Jukes–Cantor (1 - p)-scheme, to which we introduce selection. Thus, we ignore the difference between transitions and transversions in p_{ij}.

TFBS regulating different genes in the same genome most likely evolve independently and thus the nucleotide composition π_i at the respective positions of different TFBS occurrences approximates the equilibrium frequencies. With these equilibrium frequencies at hand it is possible to relate p_{ij} and f_{ij} with the equation (see [47])

$$f_{ij} \propto \frac{\log\left(\dfrac{\pi_j p_{ji}}{\pi_i p_{ij}}\right)}{1 - \dfrac{\pi_i p_{ij}}{\pi_j p_{ji}}}.$$

In our case, for the substitution rate between three non-consensus positions we obtain $\pi_i = \pi_j = \pi_n$ and $p_{ij} = p_{ji} = p_{nm}$, which yields $f_{nm} \propto 1$ by the l'Hôpital rule as in [47]. Thus, $\delta = r_{nm} = k_{pnm}$.

For the substitutions between non-consensus and consensus positions, r_{nc}, both the selection preferences and mutation asymmetry come into consideration. In this case the "asymmetry constant" λ is crucial, which satisfies the inequality $r_{cn}/r_{nc} = p_n/p_c = \lambda < 1$. The following expression is valid:

$$f_{cn} \propto \frac{\log\left(\lambda \frac{p_{nc}}{p_{cn}}\right)}{1-\lambda^{-1}\frac{p_{cn}}{p_{nc}}}.$$

This fixation rate is linked with the Kimura selection constant [48] s by the relation $f_{cn} = (1 - e^{-2s})/(1 - e^{-2Ns})$, where N is the population size.

If the mutation rate is symmetric, $p_{nc} = p_{cn}$, then $f_{cn} \propto \lambda \log(1/\lambda)/(1 - \lambda)$. Conversely, for the non-consensus to consensus substitution $f_{nc} \propto \log(1/\lambda)/(1 - \lambda) = f_{cn}/\lambda$, the greater flux from a non-consensus state to the consensus maintains a greater consensus frequency. Note that in alignments of sites in a single genome we can observe only the equilibrium constants π_n, π_c (which actually are rather rough approximations), thus the assumption $p_{nc} = p_{cn}$ may be too strong, and the above general formula for f_{cn} might be more relevant.

The coefficients in the matrix A for the symmetric mutation rates are given by

$$\alpha = r_{cn} = kp_{cn}\frac{\lambda\log(1/\lambda)}{1-\lambda},$$
$$\beta = r_{nc} = kp_{nc}\frac{\log(1/\lambda)}{1-\lambda}.$$

If the background mutation rate is identical for all consensus and non-consensus nucleotides, we obtain $p_{cn} = p_{nc} = p_{nm} = p$ and p may be merged with the constant k. In this most simple case we obtain

$$\alpha = k\frac{\lambda\log(1/\lambda)}{1-\lambda}, \quad \beta = k\frac{\log(1/\lambda)}{1-\lambda}, \quad \delta = k;$$

and, consequently, $\beta > \delta > \alpha$. This is the simplest generalization of the Jukes–Cantor model for the case with introduced selection.

It should be noted that in this case $\mu = \delta/\beta = (1 - \lambda)/\log(1/\lambda)$, which implies $\mu > \lambda$. Thus, $3\lambda \le 2\mu + 1$ and we are in Case 1 of "Interesting Cases" for which $\mathbf{P}_{cc}(u) > \mathbf{P}_{nn}(u)$ for $u > 0$.

Conservation of non-consensus nucleotides at the reduced mutation rate

Previously [8] we have observed that non-consensus nucleotides may be highly conserved in the alignments of orthologous TFBS in bacterial genomes. On the other hand, as shown above, the non-consensus nucleotide in the alignment of orthologous sites from different species cannot be more conserved than the consensus nucleotide if we adopt the Kimura model with the identical mutation rate for all pairs.

One way to explain the observation made in [8] is to drop the equivalence of non-consensus nucleotides and assume that different non-consensus nucleotides are under different selection in the sites regulating different rows of orthologous genes. Interestingly, it is possible to observe such conservation pattern in the model with an identical probabilities of mutation and fixation for all non-consensus nucleotides in all orthologous site rows, with a single preferred nucleotide, the consensus, the same in all sites. The only necessary relaxation of the model is to drop the condition $p_{cn} = p_{nc} = p_{nm} = p$ for the condition $p_{nm} < p_{nc} = p_{cn}$. Doing this it is possible to satisfy the inequalities determining Case 2 of "Interesting Cases": $\mu < \lambda$, $3\lambda > 2\mu + 1$.

The condition $p_{nm} < p_{nc} = p_{cn}$ means that the rate of direct mutations from one non-consensus nucleotide to another one is lower than the mutation rate in pairs involving the consensus nucleotide. A possible example of such specific reduction of the mutation rate comes from correlations of nucleotides occupying different positions of the same binding site. Assume that two positions within the site are not independent and must be occupied by correlated nucleotides. These may be, e.g., adjacent positions in a DNA site or base-paired positions in an RNA structure. Assume also that if any of the two positions is occupied by the consensus nucleotide, the correlating nucleotide may be arbitrary. Conversely, if one position is occupied by a non-consensus nucleotide, the other position should be occupied with some specific nucleotide, e.g. the complementary one in the case of an RNA structure.

In this case, the preferred pathway from a non-consensus nucleotide to another non-consensus nucleotide would become not via a direct mutation, but via an intermediate mutation into the consensus nucleotide. For example, if "C" is both cytosine and consensus, and for some case this position is correlated with another one, so that only A-A, T-T, and G-G pairs involving the non-consensus nucleotides at the first position are allowed, then only mutation A>C is valid, whereas mutations A>T and A>G are forbidden, and may occur via mutation into C and then the compensating mutation in the second position of the site. This simple model to some extent agrees with recent studies demonstrating that protein-DNA interactions are rather complex and probably may not be described by a simple position-independent model such as a positional weight matrix [49].

At the same time, this effect is not in contradiction with the uniform distribution of non-consensus nucleotides obtained from alignments of multiple TFBS regulating different genes in the same genome. It also allows for high conservation of some non-consensus nucleotides in the alignment of orthologous transcription factor binding sites from different species. Indeed, if the context-dependent pattern of conservation is specific to a particular position in a particular set of orthologous TFBSs, exactly this type of behavior may be expected.

Conclusion

The evolutionary model derived here can be generalized to the case of M consensus and N non-consensus nucleotides that are equiprobable, respectively. Thus, for $M = 2$, $N = 2$ and $\alpha = \beta$, $\gamma = \delta$. we get the case of neutral evolution with different rates of transitions and transversions [40]. If $M = 1$, $N = 3$, then we have the evolution model under constant selective pressure considered in this paper.

One somewhat non-obvious feature of this model is the existence of a combination of the rate of transition between non-consensus states μ and the rate of transition from the consensus state λ, for which the conservation of the consensus nucleotide in a sequence alignment can be lower than the conservation of a neutral (non-consensus) nucleotide. However, this can be observed only for the case of a weak consensus $1/4 < N_c < 1/2$ and for a relatively short time interval $u \in (0, u^{**})$, and a mutation matrix with different elements, e.g. context-dependent.

Models of this type can be applied not only for the analysis of regulatory sites, but in other situations, e.g. for the analysis of functional sites in proteins [50] or analysis of evolution in the case of nucleotide biases [41-44,51,52].

In future work, we intend to estimate the parameters of the model from the data based on the method of maximum likelihood trees. Consider a Markov process $X(t)$ describing the evolution at some fixed position. We assume that we observe only the endpoints of k different paths of $X(t)$ depending on the evolution branch. In other words, our data is a set of k nucleotides at a fixed position in k genomes. The parameter λ can be easily estimated by the 'naive' estimator $\hat{\lambda} = (1/N_c - 1)/3$, where N_c is the frequency of the consensus nucleotide at the fixed position within k genomes. This estimator is obtained from the formula for the stationary distribution of the consensus state $\pi_c = 1/(3\lambda + 1)$. However, it is a good approximation of the true value of λ. The preliminary analysis of numerically simulated data shows that the performance of the maximum likelihood estimator is also rather good. On the other hand, it is not clear whether it is possible to construct an explicit estimator for the rate of transition between non-consensus states μ from the data, although μ participates in the expression for transitional probabilities.

The inspiration for the constructed model was the example of a set of orthologous transcription factors interacting with the cognate regulatory regions. However, it appears that evolution of sequences under a low selection pressure is a more widespread phenomenon. Indeed, a recent study [53] demonstrated that the force causing conservation of some non-coding genome regions of human, mouse and chimpanzee can be explained by a rather small selective pressure at the genomic level. A similar problem appears in the context of the CG composition of genomic regions [54]. Again, in this case the selective pressure appears to be low, although unlike the previous examples, here two nucleotides become selected for rather than one consensus nucleotide. The appropriate model in this case includes differences in the transition and the transversion rates, which makes the model more complicated, and probably would result in more complex time behavior of the substitution probabilities.

Anyhow, the emerging huge amount of data on orthologous non-coding regions, which has became available recently, brings forward a problem of modelling evolution with a selection pressure at finite times.

Methods

The numerical simulations and figures were produced using Matlab.

Authors' contributions

MSG and VJM conceived the study. FNE and EAK developed the model. FNE performed numerical simulations. FNE, VJM and MSG wrote the paper. All authors have read and approved the final version.

Acknowledgements

This study was partially supported by the Russian Federation Agency in Science and Innovation State Contract # 02.531.11.9003 (VJM, EAK), by the Russian Foundation for Basic Research, projects no. 07-04-01584 (VJM), 07-04-01623 (EAK) and by grants from the Howard Hughes Medical Institute (55005610, MSG), INTAS (05-1000008-8028, VJM and MSG), and the Russian Academy of Sciences (program "Molecular and Cellular Biology", MSG). FNE was supported by the Russian Science Support Fund.

References

1. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424(6945):**147-151.
2. Pabo CO, Sauer RT: **Transcription factors: structural families and principles of DNA recognition.** *Annu Rev Biochem* 1992, **61:**1053-1095.
3. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20(9):**1377-1419.
4. Mirny LA, Gelfand MS: **Structural analysis of conserved base pairs in protein-DNA complexes.** *Nucleic Acids Res* 2002, **30(7):**1704-1711.

5. Wells J, Farnham PJ: **Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation.** *Methods* 2002, **26**:48-56.
6. Weinmann AS, Farnham PJ: **Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation.** *Methods* 2002, **26**:37-47.
7. Hube F, Myal Y, Leygue E: **The promoter competition assay (PCA): a new approach to identify motifs involved in the transcriptional activity of reporter genes.** *Front Biosci* 2006, **11**:1577-1584.
8. Kotelnikova EA, Makeev VJ, Gelfand MS: **Evolution of transcription factor DNA binding sites.** *Gene* 2005, **347**:255-263.
9. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover.** *Mol Biol Evol* 2002, **19(7)**:1114-1121.
10. Dermitzakis ET, Bergman CM, Clark AG: **Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites.** *Mol Biol Evol* 2003, **20(5)**:703-714.
11. Rajewsky N, Socci ND, Zapotocky M, Siggia ED: **The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons.** *Genome Res* 2002, **12(2)**:298-308.
12. Mustonen V, Lassig M: **Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies.** *Proc Natl Acad Sci USA* 2005, **102(44)**:15936-41.
13. Gerasimova AV, Gelfand MS: **Evolution of the NadR regulon in Enterobacteriaceae.** *J Bioinform Comput Biol* 2005, **3**:1007-1019.
14. Rodionov DA, Gelfand MS: **Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling.** *Trends Genet* 2005, **21**:385-389.
15. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Comparative genomics of the regulation of methionine metabolism in Gram-positive bacteria.** *Nucleic Acids Res* 2004, **32**:3340-3353.
16. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB: **Conservation and evolution of cis-regulatory systems in ascomycete fungi.** *PLoS Biol* 2004, **2(12)**:e398.
17. Costas J, Casares F, Vieira J: **Turnover of binding sites for transcription factors involved in early Drosophila development.** *Gene* 2003, **310**:215-20.
18. Mazon G, Campoy S, Erill I, Barbe J: **Identification of the Acidobacterium capsulatum LexA box reveals a lateral acquisition of the Alphaproteobacteria lexA gene.** *Microbiology* 2006, **152(Pt 4)**:1109-18.
19. Erill I, Jara M, Salvador N, Escribano M, Campoy S, Barbe J: **Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics.** *Nucleic Acids Res* 2004, **32(22)**:6617-26.
20. Rodionov DA, Dubchak IL, Arkin AP, Alm EJ, Gelfand MS: **Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks.** *PLoS Computational Biology* 2005, **1**:e55.
21. Tsong AE, Tuch BB, Li H, Johnson AD: **Evolution of alternative transcriptional circuits with identical logic.** *Nature* 2006, **443(7110)**:415-20.
22. Tanay A, Regev A, Shamir R: **Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast.** *Proc Natl Acad Sci USA* 2005, **102(20)**:7203-8.
23. Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, Berman J, Barkai N: **Rewiring of the yeast transcriptional network through the evolution of motif usage.** *Science* 2005, **309(5736)**:938-40.
24. Rodionov DA, Gelfand MS, Todd JD, Curson ARJ, Johnston AWB: **Computational reconstruction of iron- and manganese-responsive transcriptional networks in alpha-proteobacteria.** *PLoS Comput Biol* 2006, **2(12)**:3163.
25. Permina EA, Gelfand MS: **Heat Shock (sigma32 and HrcA/CIRCE) regulons in beta-, gamma and epsilon-proteobacteria.** *J Mol Microbiol Biotechnol* 2004, **6(3-4)**:174-181.
26. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS: **Regulation of biosynthesis and transport of aromatic amino acid in low-GC Gram-positive bacteria.** *FEMS Microbiol Lett* 2003, **222**:211-220.
27. Gelfand MS: **Evolution of transcriptional regulation networks in microbial genomes.** *Curr Opin Struct Biol* 2006, **16**:420-429.
28. Nei M: **Selectionism and neutralism in molecular evolution.** *Mol Biol Evol* 2005, **22(12)**:2318-2342.
29. Rockman MV, Hahn MW, Soranzo N, Goldstein DB, Wray GA: **Positive selection on a human-specific transcription factor binding site regulating IL4 expression.** *Curr Biol* 2003, **13(23)**:2118-23.
30. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB: **Position specific variation in the rate of evolution in transcription factor binding sites.** *BMC Evol Biol* 2003, **3**:19.
31. Kechris KJ, van Zwet E, Bickel PJ, Eisen MB: **Detecting DNA regulatory motifs by incorporating positional trends in information content.** *Genome Biol* 2004, **5(7)**:R50.
32. Mazon G, Lucena JM, Campoy S, Fernandez de Henestrosa AR, Candau P, Barbe J: **LexA-binding sequences in Gram-positive and cyanobacteria are closely related.** *Mol Genet Genomics* 2004, **271**:40-9.
33. Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek M, Cebrat S: **Where does bacterial replication start? Rules for predicting the oriC region.** *Nucleic Acids Res* 2004, **32(13)**:3781-3791.
34. Rodionov DA, Mironov AA, Gelfand MS: **Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea.** *Genome Research* 2002, **12**:1507-1516.
35. Campoy S, Mazon G, Fernandez de Henestrosa AR, Llagostera M, Monteiro PB, Barbe J: **A new regulatory DNA motif of the gamma subclass Proteobacteria: identification of the LexA protein binding site of the plant pathogen Xylella fastidiosa.** *Microbiology* 2002, **148(Pt 11)**:3583-97.
36. Zverlov VV, Schwarz WH: **Organization of the chromosomal region containing the genes lexA and topA in Thermotoga neapolitana. Primary structure of LexA reveals phylogenetic relevance.** *Syst Appl Microbiol* 1999, **22(2)**:174-8.
37. Mazon G, Erill I, Campoy S, Cortes P, Forano E, Barbe J: **Reconstruction of the evolutionary history of the LexA-binding sequence.** *Microbiology* 2004, **150(Pt 11)**:3783-95.
38. Danilova LV, Gel'fand MS, Liubetski VA, Lakova ON: **Computer analysis of regulating metabolism of glycerol-3-phosphate in proteobacteria genome.** *Mol Biol (Mosk)* 2003, **37(5)**:843-9.
39. Jukes T, Cantor C: **Evolution of protein molecules.** In *Mammalian Protein Metabolism* Academic Press; 1969:21-132.
40. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotides sequences.** *J Mol Evol* 1980, **16**:111-120.
41. Galtier N, Gouy M: **Inferring phylogenies from DNA sequences of unequal base compositions.** *Proc Natl Acad Sci USA* 1995, **92(24)**:11317-21.
42. Galtier N, Gouy M: **Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis.** *Mol Biol Evol* 1998, **15(7)**:871-9.
43. Goncalves I, Robinson M, Perriere G, Mouchiroud D: **JaDis: computing distances between nucleic acid sequences.** *Bioinformatics* 1999, **15(5)**:424-5.
44. Tamura K, Kumar S: **Evolutionary distance estimation under heterogeneous substitution pattern among lineages.** *Mol Biol Evol* 2002, **19(10)**:1727-36.
45. Gerland U, Hwa T: **On the selection and evolution of regulatory DNA motifs.** *J Mol Evol* 2002, **55(4)**:386-400.
46. Gerland U, Moroz J, Hwa T: **Physical constraints and functional characteristics of transcription factor-DNA interaction.** *Proc Natl Acad Sci USA* 2002, **99(19)**:12015-20.
47. Halpern A, Bruno W: **Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies.** *Mol Biol Evol* 1998, **15(7)**:910-7.
48. Kimura M: **On the probability of fixation of mutant genes in a population.** *Genetics* 1962, **47**:713-9.
49. Kinney J, Tkacik G, Callan CJ: **Precise physical models of protein-DNA interaction from high-throughput data.** *Proc Natl Acad Sci U S A* 2007, **104(2)**:501-6.
50. Pollock DD, Bruno WJ: **Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition.** *Mol Biol Evol* 2000, **17**:1854-1858.
51. Perna NT, Kocher TD: **Unequal base frequencies and estimation of substitution rates.** *Mol Biol Evol* 1995, **12**:359-361.
52. Collins TM, Wimberger PH, Naylor GJP: **Compositional bias, character-state bias, and character-state reconstruction using parsimony.** *Syst Biol* 1994, **47**:482-496.

53. Keightley P, Kryukov G, Sunyaev S, Halligan D, Gaffney D: **Evolutionary constraints in conserved nongenic sequences of mammals.** *Genome Res* 2005, **15(10):**1373-8.
54. Lercher M, Smith N, Eyre-Walker A, Hurst L: **The evolution of isochores: evidence from SNP frequency distributions.** *Genetics* 2002, **162(4):**1805-10.