Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 347 (2005) 255–263

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Evolution of transcription factor DNA binding sites

Ekaterina A. Kotelnikova[a,*], Vsevolod J. Makeev[a,b], Mikhail S. Gelfand[a,c]

[a]State Research Institute of Genetics and Selection of Industrial Microorganisms, 1st Dorozhnyj proezd, 1, Moscow, 113535 Russia
[b]Engelhardt Institute of Molecular Biology, Russian Academy of Science, Moscow, Russia
[c]Institute for Information Transmission Prolems, Russian Academy of Science, Moscow, Russia

Received 14 September 2004; received in revised form 12 November 2004; accepted 2 December 2004

Received by M. Vendruscolo

Abstract

In bioinformatics, binding of transcription regulatory factors to the cognate binding sites is usually described by sequence-specific binding energy, which is estimated from a training sample of sites. This model implies that all binding sites with binding energy above some threshold are functional and site sequence variations should be considered neutral until they do not reduce this energy below the threshold. To quantify this energy, the binding profile (positional weight matrix, PWM) model or consensus-based model is usually applied. Here we show that in many cases available data are not sufficient to construct a relevant PWM, and modified consensus-based model could be more effective to describe binding properties.

Further, using the data about binding sites of several transcription factors, we demonstrate that some non-consensus nucleotides in "orthologous sites" (that is, binding sites of the same factor upstream of orthologous genes), which have been believed to be irrelevant or even hindering the regulation, are evolutionary very stable and specific for the regulated gene. For each two considered genomes, the number of substitutions between non-consensus nucleotides is far less than the expected number of neutral substitutions. Moreover, in several positions of binding sites regulating different genes, there are non-consensus nucleotides conserved in distant genomes. It means that there exists a selection pressure, which results in the stability of non-consensus nucleotides.

Keywords: Genomics; Transcription; Regulation; Consensus; Binding site; Evolution

1. Introduction

Predictions of transcription regulation rely on the fact that transcription factor (TF) binding sites upstream of regulated genes, although not identical, are similar to each other (Stormo and Fields, 1998). The similarity of a site to sites from a training set is measured by a positional weight matrix (PWM, or profile) that assigns a score to each nucleotide at each signal position. The specificity of such predictions is often rather low. One way of obtaining more reliable predictions is to require that candidate sites are present upstream of orthologous genes in several genomes (Gelfand, 1999).

A common theory of the TF binding specificity (Berg and von Hippel, 1987) implies equiprobability for all sites with the same TF affinity to be selected and independence of DNA base-pair contributions to this affinity. In the case when all contributions from different non-consensus nucleotides are assumed to be the same, so called "two-state model" can be implemented. It considers the TF binding energy as a sum of contributions from consensus nucleotides of the binding site, so that the contribution to the energy from non-consensus nucleotides is believed to be zero. An evolutionary theory based on this model (Peliti, 2002; Lassig, in press) explains the appearance and

Abbreviations: PWM, positional weight matrix; TF, transcription factor; COG, cluster of orthologous genes; WCM, weighted consensus matrix; ANOVA, analysis of variance; DNA, deoxyribonucleic acid.

* Corresponding author.
E-mail address: Ekotelnikova@gmail.com (E.A. Kotelnikova).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.12.013

evolution of binding sites as well as selection on the factor's binding affinity to sites upstream of different genes, which should fall within some fixed range in order to provide the right level of regulation.

These models with assumption of equiprobability imply the absence of selectivity pressure on non-consensus nucleotides, which is not necessarily true. The aim of this work was to study the evolution of prokaryotic transcription factors binding sites, and in particular the behavior of non-consensus positions.

2. Materials and methods

The following terms will be used: *orthologous sites* are transcription factor binding sites upstream of orthologous genes in different genomes. *Consensus nucleotide* or *consensus* of a signal position in a genome is the most frequent nucleotide in this genome (in all binding sites for the corresponding factor).

The following genomes were downloaded from: *Escherichia coli* [AE005174, U00096], *Salmonella typhi* [AE014613, AL513382], *Salmonella typhimurium* [AE006468], *Yersinia pestis* [AL590842], *Yersinia enterocolitica* [NC_003222], *Erwinia amylovora*, *Haemophilus influenzae* [L42023], *Haemophilus somnus* [NZ_AACJ00000000], *Pasteurella multocida* [AE004439], *Vibrio cholerae* [AE003852, AE003853], *Vibrio fischeri*, *Vibrio parahaemolyticus* [BA000031, BA000032], *Vibrio vulnificus* [BA000038], *Pectobacterium carotovorum* (*Erwinia carotovora*) [NC_004547], *Actinobacillus actinomycetemcomitans* [NC_002924].

Binding sites for 16 transcriptional regulators were used to analyze the uniformity of frequencies of non-consensus nucleotides (Table 1). Further, sites for 6 regulators (PurR, FruR, LexA, GalRS, Sigma-32, ScrR) with strong signals and well-conserved regulons were used to analyze conservation of non-consensus nucleotides. At that, only orthologous sites were taken into account (Table 2). To demonstrate that the non-consensus nucleotides tend to be conserved in particular rows of orthologous sites for some factor it is necessary to have such sites in many genomes and a sufficient number of them in each genome. Only three such examples, Sigma-32 (Sections 3.1, 3.3), LexA(Sections 3.3, 3.4) and PurR (Section 3.3) were available (Table 1, Table 2). The other three factors studied, FruR, GalRS, and ScrR have a small number of sites in most considered genomes.

The rates of the neutral evolution and the intergenomic distances were taken from (Novichkov et al., 2004). For a given set of genomes, a set of COGs (Tatusov et al., 2001) was identified, in which each genome was represented by exactly one gene (i.e., all species present, no paralogs). Additionally, the constituent genes were required to have a sufficient alignable length (at least 60 codons in conserved blocks; see below). This search resulted in 563 COGs, which were used to measure the degree of conservation at

Table 1
The number of sites used in testing the consensus and the PWM models

Transcriptional factors	HrcA	BirA	MalR	GalRS	ZuR	GntR	FNR	NifA	UxuR	Sig32	LexA	FruR	KdgR	PurR	Crp	CcpA
Number of searched genomes	10	25	13	9	7	8	6	10	7	10	6	5	7	6	1	10
Minimum number of binding sites in genome	1	1	1	2	1	2	1	5	4	11	9	3	8	13	48	42
Maximum number of binding sites in genome	7	5	5	8	12	8	23	23	24	21	22	22	36	36	48	132
Average number of binding sites in genome	3	3	3	4	4	5	10	12	12	13	15	16	19	24	48	70
References	Permina and Gelfand, 2004	Rodionov et al., 2002	Laikova[a]	Laikova[a]	Panina et al., 2003	Rodionov et al., 2000	Gerasimova et al., 2001	Rodionov[a]	Rodionov et al., 2000	Permina and Gelfand, 2004	Permina[a]	Laikova[a]	Rodionov[a]	Ravcheev et al., 2002	Rodionov et al., 2000	Rodionov et al., 2001

[a] O.N. Laikova, E.A. Permina and D.A. Rodionov personal communications.

Table 2
Transcription factors and their regulons used for detailed evolutionary studies

Transcription regulator	Description	Genomes	Regulated genes	References
PurR	Purine metabolism	*E.coli, S. typhi, Y. pestis, H. influenzae, P. multocida, V. cholerae*	*cvpA, purB, purE, purH, purL, purM, upp, yjcD, gltS, purR, prsA, purC, yhhQ, yicE, yieG, glyA, guaB, purT, pyrC, rnt, serA, speA, codB, folD, gcvT, pyrD, ydiJ, ydiK, purA, tsx, glnB, rpiA, uraA*	Ravcheev et al., 2002
FruR	Fructose catabolism	*E. coli, K. pneumoniae, S. typhi, Y. pestis, V. cholerae*	*aceB, adhE, adk, crp, edd, epd, fbp, fruB, frvA, glk, icdA, mtlA, pckA, pdhR, pfkA, ppc, ppsA, ptsH, pykF, tpiA, yibO*	Laikova[a]
LexA	SOS response	*E.coli, S. typhi, P. multocida, Y. pestis, V. cholerae, H. influenzae*	*dinD, dinG, dinI1, dinI2, dinJ, dinP, lexA1, lexA2, lexA3, mutH, recA, recN1, recN2, recN3, ruvA1, ruvA2, ssb, sulA, TFOX, umuD, uvrA, uvrB, uvrD, yebG*	Permina[a]
GalRS	Galactose catabolism	*A. actinomycetemcomitans , E.coli, H. somnus, K. pneumoniae, P. multocida, P. carotovorum, S. typhi, V. fischeri, V. parahaemolyticus, V. vulnificus, Y. pestis*	*galR, mglB, galT, galE1, galE2, galP, galS1, galE2, galP3, sglT1, sglT2, sglT3*	Laikova[a]
Sigma 32	Heat shock response	*E.coli, S. typhi, Y. pestis, Y. enterocolitica, H. influenzae, P. carotovorum, K. pneumoniae, P. multocida, V. cholerae, V. fischeri*	*Var1, b0492, b4140, clpB, clpP, dnaK, ftsJ, gapA, grpE, hslT, hslV, htpG, htpX, lon, mopB, rpsL*	Permina and Gelfand, 2004
ScrR	Sucrose catabolism	*E. amylovora, E.coli, K. pneumoniae, P. carotovorum , S. typhimurium, Y. enterocolitica*	*scrK, scrY1, scrY2*	Laikova[a]

If there were more than one orthologous sites upstream the same gene, the corresponding sites are numbered.
[a] O. Laikova, E. Permina and D. Rodionov, personal communications.

synonymous codon positions. The distance between two genomes was defined as the median of the distribution of the distances between orthologs from these genomes (Wolf et al., 2002).

MS Excel and Statistica software package were used for statistical tests. ANOVA tests were used as implemented in the Statistica software package. The purpose of *analysis of variance* (ANOVA) is to test the differences in means (for groups or variables) for statistical significance (Neter et al., 1996). This test is based on a comparison of the variance due to the between-groups variability (called *Mean Square Effect*, MS effect) with the within-group variability (called *Mean Square Error*, or MS error). Under the null hypothesis (that there are no mean differences between groups in the population), some minor random fluctuation in the means for the two groups is still expected. Therefore, under the null hypothesis, the variance estimated based on the within-group variability should be about the same as the variance due to the between-groups variability. Those two estimates of the variance can be compared via the *F*-test.

3. Results and discussion

3.1. Binding site model (consensus and non-consensus nucleotides)

Two main models of transcription regulatory signals are the consensus (Day and McMorris, 1992) and the position weight matrix (Schneider, 1986). When the consensus is used to search for new candidate sites, the relevance of a site is measured by the number of matches. In the case of PWM, the score of a candidate site is the sum of positional weights. The most popular way to define positional nucleotide weights is to set them proportional to the logarithms of positional frequencies of each nucleotide in a multiple alignment of a training set of known sites. Thus a consensus can be considered as a particular case of a PWM with positional weights of one and zero. The relevance of the PWM model was discussed many times, and it is believed to be the optimal pattern model (Schneider, 1986). It has been also shown that the PWM score is correlated with the factor affinity to the site (Berg and von Hippel, 1987).

One can see that the criteria for defining consensus positions are often arbitrary (formally, in each alignment position there is a most frequent nucleotide(s), whereas only positions with strong leaders should be accepted, Day and McMorris, 1992). On the other hand, the PWM weights for each signal position are usually evaluated from nucleotide counts in a multiple alignment of binding sites. Because the total of all counts must be equal to the number of sites in the alignment, three independent parameters for each alignment column need to be evaluated from the training data set. However, in most cases the number of known sites in a genome (Table 1) is not (and will never be) sufficient to evaluate three independent parameters for each site alignment column and construct a relevant weight matrix. It means that in these cases the PWM model could be overfitted, leading to low sensitivity.

To find out whether the PWM model could be simplified without loss of selectivity, two tests were performed for each genome and each set of known DNA regulatory sites for several transcriptional factors. The first one was uniformity (chi-square 5%) test for frequencies of all types of nucleotides at each position of the regulatory signal. This test revealed whether there is a preferred "consensus nucleotide" in the given position. The second uniformity (chi-square 5%) test was performed only for non-consensus nucleotides, and showed whether there is a preferred nucleotide among non-consensus ones. As an example, the analysis of sigma-32 binding sites is given in Table 3.

The consensus of this signal, CTTGAAA-$N_{11–16}$-CCCCAT, is well conserved in the studied genomes. The results of the chi-square test also are consistent: in position 1 of the first half-site, G is a significantly preferred non-consensus nucleotide; in positions 1, 3 and 4 of the second half-site there also are preferred non-consensus nucleotides. Thus, the consensus (with degenerate positions) may be re-written as (C/G)TTGAAA-$N_{11–16}$-(C/a/g)C(C/t)(C/T)AT. Similar observations were made for other considered transcription factors (data not shown). The summary of test results for binding sites of 16 transcriptional regulators is presented in Table 4. It can be seen that in many cases one (the consensus) nucleotide prevails and differences in frequencies of all non-consensus nucleotides are statistically insignificant. Other possibilities are the uniform distribution (a non-informative position) and the cases of one consensus and one preferred non-consensus nucleotide (two-letter choice). Below both latter cases will be referred to as *positions with poor consensus*.

Table 3
Chi-square test results for sigma-32 transcriptional binding sites

Genome	Number of sigma-32 binding sites	1	2	3	4	5	6	7		1	2	3	4	5	6
		C	**T**	**T**	**G**	**A**	**A**	**A**	**- (12-14) -**	**C**	**C**	**C**	**C**	**A**	**T**
Pectobacterium carotovoru	chi-square	0.01	0	0	0	0	0	0		0.02	0	0.01	0	0	0
	chi-square without max	0.17	N/A	N/A	N/A	0.14	0.37	0.37		0.07	0.37	0.17	0.05	0.37	0.37
		C	**T**	**T**	**G**	**A**	**A**	**A**	**- (11-15) -**	**C**	**C**	**C**	**C**	**A**	**T**
Escherichia coli	chi-square	0.01	0	0	0	0	0	0		0	0	0	0	0	0
	chi-square without max	0.02	0.37	N/A	0.37	0.05	0.22	0.14		0.37	0.14	0.05	0.1	N/A	N/A
		C	**T**	**T**	**G**	**A**	**A**	**A**	**- (12-16) -**	**C**	**C**	**C**	**C**	**A**	**T**
Haemophilus influenzae	chi-square	0	0	0	0	0	0	0.02		0.03	0.04	0.15	0.15	0	0.03
	chi-square without max	0.14	N/A	N/A	0.61	0.37	0.14	0.17		0.78	0.07	0.14	0.14	N/A	0.78
		G	**T**	**T**	**G**	**A**	**A**	**A**	**- (11-15) -**	**C**	**C**	**C**	**C**	**A**	**T**
Klebsiella pneumoniae	chi-square	0.34	0	0	0	0	0	0.03		0	0	0	0.01	0	0
	chi-square without max	0.37	N/A	0.37	0.61	0.37	0.14	0.25		0.61	N/A	0.05	0.01	N/A	0.37
		C	**T**	**T**	**G**	**A**	**A**	**A**	**- (12-16) -**	**C**	**C**	**C**	**T**	**A**	**T**
Pasteurella multocida	chi-square	0	0	0	0	0	0	0		0.08	0.02	0.01	0.11	0	0
	chi-square without max	1	0.37	N/A	0.37	0.37	0.61	0.37		0.25	0.37	0.01	0.45	N/A	0.37
		C	**T**	**T**	**G**	**A**	**A**	**A**	**- (11-16) -**	**C**	**C**	**C**	**C**	**A**	**T**
Salmonella typhi	chi-square	0.02	0	0	0	0	0	0		0	0	0	0	0	0
	chi-square without max	0.05	0.37	N/A	0.37	0.61	0.25	0.07		0.78	N/A	0.02	0.01	N/A	N/A
		C	**T**	**T**	**G**	**A**	**A**	**A**	**- (11-15) -**	**C**	**C**	**C**	**T**	**A**	**T**
Vibrio cholerae	chi-square	0.03	0	0	0	0	0	0.03		0.01	0	0.02	0.01	0	0.01
	chi-square without max	0.03	0.37	N/A	0.37	0.14	0.37	0.25		0.17	0.14	0.07	0.01	0.37	0.17
		C	**T**	**T**	**G**	**A**	**A**	**A**	**- (12-15) -**	**C**	**C**	**C**	**C**	**A**	**T**
Vibrio fischeri	chi-square	0.01	0	0	0	0	0	0.04		0.01	0.15	0	0.02	0	0
	chi-square without max	0.02	N/A	0.37	1	0.37	0.61	0.07		0.01	0.82	0.05	0.17	0.14	0.37
		C	**T**	**T**	**G**	**A**	**A**	**A**	**- (11-15) -**	**C**	**C**	**C**	**C**	**A**	**T**
Yersinia enterocolitica	chi-square	0.04	0	0	0	0	0.02	0.08		0	0	0.02	0	0	0
	chi-square without max	0.07	0.37	N/A	0.37	0.37	0.17	0.25		0.37	N/A	0.17	0.37	0.37	N/A
		G	**T**	**T**	**G**	**A**	**A**	**A**	**- (11-15) -**	**C**	**C**	**C**	**C**	**A**	**T**
Yersinia pestis	chi-square	0.48	0	0	0	0	0	0.03		0	0	0	0	0	0
	chi-square without max	0.37	0.61	N/A	0.14	N/A	0.61	0.78		0.05	N/A	0.37	0.05	N/A	N/A

The values that pass the first test (existence of a preferred "consensus" nucleotide) are marked in light grey. The values that pass the second test (non-uniformity of non-consensus nucleotide frequencies) are marked in dark grey. Absolutely conserved positions are marked as N/A.

Table 4
The summary of the uniformity test results for binding sites of 16 transcriptional regulators in available genomes

Transcriptional factors	HrcA	BirA	MalR	GalRS	ZUR	GntR	FNR	NifA	UxuR	Sig32	LexA	FruR	KdgR	PurR	Crp	CcpA
Number of counted positions	486	450	260	180	161	160	96	160	126	130	120	80	147	96	22	160
Number of consensus nucleotides	272	127	114	70	53	87	69	109	100	122	101	71	140	91	16	160
Number of second meaningful nucleotides	7	2	1	1	3	8	5	15	23	17	17	11	59	17	3	132
Percentage of two meaningful nucleotides	3	2	1	1	6	9	14	14	23	14	17	15	42	19	19	83

Number of counted positions is number of genomes multiplied by the number of positions in the signal.

The fact that the first situation is frequent indicates that instead of a PWM, a model with a variable number of parameters appears to be more relevant. For positions with statistically uniform non-consensus letters only two parameters can be used: the weight of the consensus nucleotide and the weight of a non-consensus nucleotide. We call this model *weighted consensus matrix* (WCM).

Below we study the evolution of nucleotides occupying such positions in orthologous binding sites; in this case the position occupied by the consensus and the non-consensus nucleotides will be referred to as the *consensus* and the *non-consensus* positions, respectively.

3.2. Conservation of non-consensus nucleotides in orthologous sites

Consensus, WCM and PWM models are based on aligned binding sites from one genome and expose their averaged characteristics. However, a look at aligned orthologous binding sites reveals unexpected conservation of non-consensus nucleotides. To check the relevance of this observation and to determine whether the choice of non-consensus nucleotide depends on the genome and gene we performed the Multi-Factor ANOVA test with the following parameters for each transcriptional regulator (see Table 2): the independent variables were “Genome” (names of available genomes), “Gene” (names of regulated genes), “Position” (number of nucleotide position in the binding signal), and “Nucleotide” (A, T, G, C, a, t, g, c), whereas the dependent variable was the “Count” (yes/no). For a given position each nucleotide in a particular site could be consensus (uppercase) or non-consensus (lowercase). Only non-consensus nucleotides from a set of orthologous sites were considered.

Here we considered the variability within the “Position”*”Nucleotide” group as a within-group variability, and computed between-groups variability for two different three-way interactions: “Genome”*”Position”*”Nucleotide” and “Gene”*”Position”*”Nucleotide”.

The within-group variability of “Position”*”Nucleotide” interaction is the same for both of these cases, and reflects the importance of nucleotide arrangement in the signal. The three-way interaction “Genome”*”Position”*”Nucleotide” measures whether signals taken from the different genomes are significantly different (taking together all binding sites from each genome). Similarly, the three-way “Gene”*”Position”*”Nucleotide” interaction measures whether sets of orthologous sites are significantly different (taking together sites from all available genomes for each orthologous set). Thus, the former test measures the dependence of the nucleotide choice at a given position on the genome, and the latter test measures the dependence of this choice on the gene (Table 5).

Significant “Genome”*”Position”*”Nucleotide” interaction reflects the intuitively obvious fact that the binding preferences (here only for non-consensus nucleotides) of the

Table 5
Variance analysis of "Nucleotide" occurrences

Regulator	Effect	SS	Degrees of freedom	MS	*F*	*p*	+/−
PurR	Gene*Position*Nucleotide	111.8918	960	0.11655	1.8667	0.000000	+
	Genome*Position*Nucleotide	14.0194	150	0.09346	1.4969	0.000103	+
	Error	331.7285	5313	0.06244			
FruR	Gene*Position*Nucleotide	129.1413	900	0.14349	4.4530	0.00	+
	Genome*Position*Nucleotide	17.8102	180	0.09895	3.0706	0.00	+
	Error	130.1496	4039	0.03222			
LexA	Gene*Position*Nucleotide	135.9799	1311	0.10372	2.0271	0.00	+
	Genome*Position*Nucleotide	19.1335	285	0.06714	1.1061	0.110593	−
	Error	293.0938	5728	0.05117			
GalRS	Gene*Position*Nucleotide	68.16715	627	0.10872	2.1633	0.000000	+
	Genome*Position*Nucleotide	31.15729	570	0.05466	1.0876	0.102605	−
	Error	96.59463	1922	0.05026			
ScrR	Gene*Position*Nucleotide	8.08262	114	0.070900	1.35449	0.011447	+
	Genome*Position*Nucleotide	11.19467	285	0.039280	0.75040	0.998054	−
	Error	46.06318	880	0.052345			
Sigma-32	Gene*Position*Nucleotide	50.4434	504	0.10009	3.1041	0.000000	+
	Genome*Position*Nucleotide	10.9415	324	0.03377	1.0473	0.273235	−
	Error	213.0328	6607	0.03224			

Other factors are "Gene", "Genome" and "Position".
Three-way between interactions. For each regulator, the spreadsheet shows whether "Position"*"Nucleotide" effect is qualified by the third independent factor. Columns: SS is the Sum of (deviation) squares; MS is the mean squares for the effect with given degrees of freedom (Degrees of freedom); *F* is the Fischer test result; *p* is the *p*-value; the +/− column is the outcome of the significance test.

regulator can be different in different genomes, that is, the binding signal can change in course of molecular evolution.

Significant "Gene"*"Position"*"Nucleotide" interaction confirms the novel observation that non-consensus nucleotides in certain positions of the binding signal are evolutionary stable and depend on the downstream regulated gene or operon.

It can be seen that the "Gene"*"Position"*"Nucleotide" interaction in our data is always more significant than "Genome"*"Position"*"Nucleotide" one. This is the result of considering closely related genomes and very well-conserved transcriptional regulators that means slow changes of the binding signal in course of evolution.

3.3. Neutral evolution and evolutionary stability of non-consensus nucleotides

At the first glance, an obvious explanation for the observed conservation of non-consensus nucleotides is the small evolutionary distances between genomes, i.e. high conservation of all non-coding regions. If this were the case, the frequency of "non-consensus mutations" (i.e. mutations of non-consensus nucleotides resulting in another non-consensus nucleotide) should be comparable with the frequency of mutations observed in nearly neutral positions, for instance, synonymous codon positions in orthologous genes from available genomes.

To account for this possibility, for each pair of genomes (G,H), all aligned nucleotide pairs in all orthologous sites were considered. All such pairs were divided into two groups, those with both non-consensus nucleotides, and those with consensus nucleotide in at least one genome. At each position i, the value $c_i(G,H)$ was defined as the fraction of genome pairs (G,H) with identical non-consensus nucleotides among the former (both non-consensus) type.

Several types of aligned positions in a pair of genomes (G,H) could be distinguished: *conserved consensus positions* (with a negligible number of non-consensus pairs, so that for genome pairs (G,H), $c_i(G,H)$ was not defined); *positions with a poor consensus* or different consensi in different genomes ($c_i(G,H)$ was not well defined); *standard positions* (some non-consensus nucleotides, but $c_i(G,H)$ differs for different (G,H)); and *conserved non-consensus positions* (with stable high conservation $c_i(G,H)>1/2$ for almost all genome pairs (G,H)). The summary of position types is given in Table 6. It can be seen, that for each of the reviewed transcriptional factors there is about 25% of the aligned binding site positions, which are conserved non-consensus positions.

If the effect of non-consensus conservation is due to neutral evolution and there is no selection for non-consensus nucleotides, the value of conservation for all positions should be almost the same. To show the average effect in a time-scale, for each transcription factor and each (G,H) genome pair the *average non-consensus conservation* $C(G,H)$ was defined as the fraction of conserved nucleotide pairs among the *non-consensus* nucleotide pairs in the pre-selected set of *conservative non-consensus positions*.

In a standard approximation of neutral evolution, mutations in third positions of synonymous codons (coding for the same amino acid) with different degree of degeneracy are usually chosen (though there is some evolutional pressure on these positions, in our case considering this fact could only increase the observed

Table 6
Different types of positions in binding signals of several transcriptional regulators

Transcriptional factor	Number of conservative consensus positions	Number of positions with poor consensus	Number of standard positions	Number of conservative non-consensus positions	Total number of positions (signal length)	Average number of available binding sites in a genome
PurR	5	3	4	**4**	16	30
LexA	6	5	5	**4**	20	24
Sigma-32	4	3	2	**4**	14	18

Different types of positions are described in text. The column with the number of conservative non-consensus positions for each transcriptional factor is highlighted.

effect). In the neutral case, $C(G,H)$ should be equal to the average conservation observed for codon with three-fold degeneracy in the set of aligned clusters of orthologous genes (COGs, see Materials and methods) from the same genomes. Because there is only one amino acid (isoleucine) with degeneracy 3, other cases (degeneracy 2 and 4) were also considered as providing an upper and lower bounds respectively. To measure the value of neutral conservation, aligned coding sequences representing relevant COGs in each pair of genomes (G,H) were taken. The conservation was defined as a number of conserved third codon positions from all conserved amino acids with the given codon degeneracy divided by the total number of these amino acids in all aligned COGs.

As an example, comparison of $C(G,H)$ and the respective *neutral conservation* in synonymous codon positions for one factor, LexA is given in Fig. 1. As a measure of distances between two genomes, the median of the distribution of the distances between orthologs from the given pair of genomes was taken (Novichkov et al., 2004; Wolf et al., 2002).

It can be seen that the value of non-consensus conservation, averaged over four relevant positions, is always higher than the neutral (residual) conservation even for the two-fold degeneracy. Under the null hypothesis, the average neutral conservation and the non-consensus conservation should be the same. Because there is no information about the distribution of these variables, several nonparametric methods, including the Wald-Wolfowitz, Mann-Whitney and Kolmogorov-Smirnov tests were applied. For each considered transcriptional regulator these tests were not passed with p-value <0.001. It means that the conservation of some non-consensus positions in orthologous binding sites is significantly higher than expected solely based on the distance between the genomes.

3.4. Comparison of consensus and non-consensus conservation

To demonstrate the observed phenomena in another way, the conservation of consensus and non-consensus positions

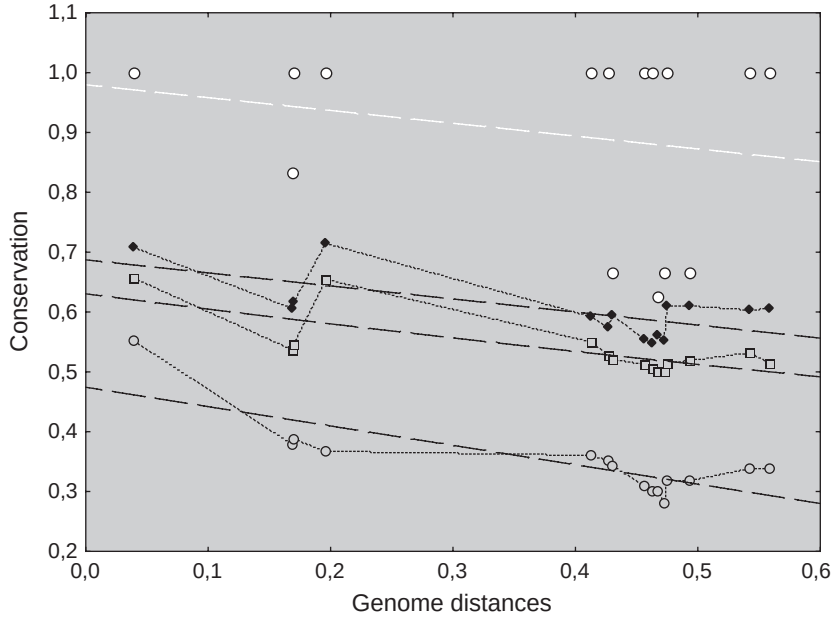

Fig. 1. Conservation of non-consensus positions in orthologous LexA binding sites in comparison to the quasi-neutral conservation of codon positions. ◆ Neutral conservation in synonymous codon positions with degeneracy 2; □ Neutral conservation in synonymous codon positions with degeneracy 3; ○ Neutral conservation in synonymous codon positions with degeneracy 4; ○ Average non-consensus conservation $C(G,H)$.

relative to the average conservation of all signal positions was computed.

For each pair of genomes (G,H) and all aligned nucleotide pairs in all orthologous sites, we divided all such pairs into two groups, those with both non-consensus nucleotides, and those with a consensus nucleotide in at least one genome. Denote by $W(G,H)$ the fraction of pairs with identical nucleotides.

At each position i, let $c_i(G,H)$ be the fraction of pairs with identical non-consensus nucleotides among the non-consensus type, defined in the previous paragraph, and let $d_i(G,H)$ be the fraction of pairs with both consensus nucleotides among the pairs of the latter, consensus type. To take into account the evolutionary distances between genomes, we normalize $c_i(G,H)$ and $d_i(G,H)$ by dividing them by $W(G,H)$. After that, these normalized values are averaged over different (G,H) genome pairs for each position i.

A typical example (the LexA signal) is shown in Fig. 2. The histogram reflects the conservation of positions with respect to the "average" conservation of a binding site, so that values exceeding 1 correspond to strong conservation, which demonstrates the considerable effects for non-consensus positions. It can be seen that there are several positions with strong non-consensus conservation, which is even higher than the consensus conservation. This once again confirms the unexpected conservation of non-consensus nucleotides in orthologous binding sites. Another observation is that strongly conserved consensus positions are clustered, in agreement with a recent work of Kechris et al., 2004, where the theory for the same effect in yeast and *E. coli* binding sites was discussed. Similar observations were made for several other considered factors (data not shown).

4. Conclusion

We have demonstrated that non-consensus nucleotides in many bacterial signals of transcriptional regulation are evolving much slower than could be expected assuming that these nucleotides are random deviations from the consensus. Thus there exists an evolutionary pressure stabilizing these positions. The nature of this pressure is unclear and requires further investigation.

One possible cause of conservation of a particular non-consensus letters in some ortolog row, could be overlapping transcription factor binding sites. Indeed, an additional binding site of a yet unknown transcription regulatory factor can be present only upstream of one particular gene, subject to regulation by this hypothetical factor. The latter may not regulate other genes. At the present stage of our knowledge about the architecture of regulatory regions we cannot neither confirm nor reject this hypothesis, but the situation will be clarified as more regulatory circuits in bacteria become characterized. However, we do not think that this could be a universal explanation, as existence of overlapping sites would yield additional conserved positions outside of studied sites, which has not been observed.

The other idea is less obvious, and comes from the requirement to maintain the right level of regulation for any given gene. Some genes should be under highly sensitive regulation, others by weaker regulation, and the efficiency of regulation is determined by binding affinity

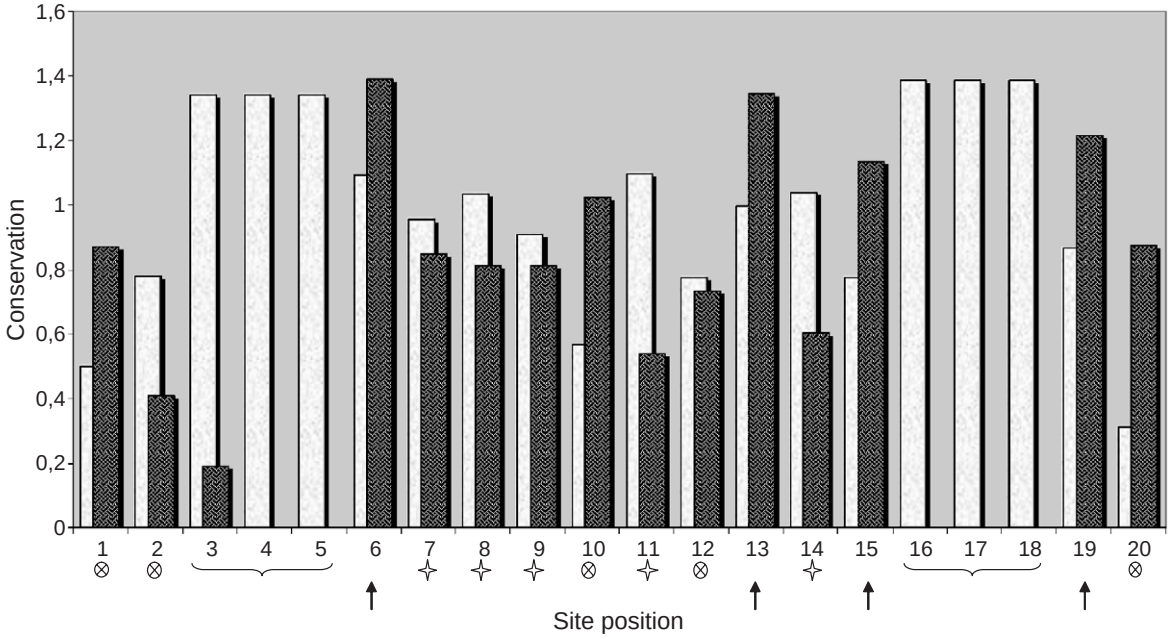

Fig. 2. Conservation of consensus and non-consensus nucleotides in orthologous LexA binding sites. □ Consensus coservation d_i; ⌣ Conserved consensus positions; ✧ Standard positions; ▩ Non-consensus conservation c_i; ⊗ Positions with poor consensus; ↑ Conserverd non-consensus positions.

of the regulatory factor (Berg and von Hippel, 1987). For instance, recently it has been shown, that the closer the enzyme is to the beginning of the pathway, the shorter is the response time of the activation of its promoter and the higher is its maximal promoter activity (Zaslaver et al., 2004). In this model, the consensus nucleotide provides the maximal binding site affinity, which is required, however, only for a subset of all genes. For the function of other genes, the affinity of this factor should be smaller, which is maintained by conservation of a particular non-consensus letter. The substitution of this letter either to a consensus letter or to another non-consensus letter would change the activity, which would reduce the fitness. This means that neither the evolution of non-consensus nucleotides is neutral, nor mutations towards consensus nucleotides are always preferred.

Acknowledgements

We are grateful to Olga Laikova, Elizabeth Permina and Dmitry Rodionov for data about binding sites, to Pavel Novichkov for assistance with determining the intergenomic distances, and to Andrei Mironov and Alexander Favorov for valuable discussions.

This work was supported by the Russian Fund of Basic Research (grant 04-04-49601 to V.J.M.), the Howard Hughes Medical Institute (grant 55000309 to M.S.G.), and Russian Academy of Sciences Programs “Molecular and Cellular Biology” (V.J.M. and M.S.G.) and “Origin and Evolution of the Biosphere” (M.S.G.).

References

Berg, O.G., von Hippel, P.H., 1987. Selection of DNA binding sites by regulatory proteins. J. Mol. Biol. 193, 723–750.

Day, W.H., McMorris, F.R., 1992. Threshold consensus methods for molecular sequences. J. Theor. Biol. 159, 481–489.

Gelfand, M.S., 1999. Recognition of regulatory sites by genomic comparison. Res. Microbiol. 150, 755–771.

Gerasimova, A.V., Rodionov, D.A., Mironov, A.A., Gel’fand, M.S., 2001. Computer analysis of regulatory signals in bacterial genomes. Fnr binding segments. Mol. Biol. (Mosk.) 35, 1001–1009.

Kechris, K.J., van Zwet, E., Bickel, P.J., Eisen, M.B., 2004. Detecting DNA regulatory motifs by incorporating positional trends in information content. Genome Biol. 5, R50.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied Linear Statistical Models. Irwin, Chicago.

Novichkov, P.S., Omelchenko, M.V., Gelfand, M.S., Mironov, A.A., Wolf, Y.I., Koonin, E.V., 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J. Bacteriol. 186 (19), 6575–6585.

Panina, E.M., Mironov, A.A., Gelfand, M.S., 2003. Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. Proc. Natl. Acad Sci. U. S. A. 100, 9912–9917.

Peliti, L., 2002. Quasispecies evolution in general mean-field landscapes. Europhys. Lett. 57, 745–751.

Permina, E.A., Gelfand, M.S., 2004. Heat shock (sigma32 and HrcA/CIRCE) regulons in beta-, gamma- and epsilon-proteobacteria. J. Mol. Microbiol. Biotechnol. 6, 174–181.

Ravcheev, D.A., Gel’fand, M.S., Mironov, A.A., Rakhmaninova, A.B., 2002. Purine regulon of gamma-proteobacteria: a detailed description. Genetika 38, 1203–1214.

Rodionov, D.A., Mironov, A.A., Rakhmaninova, A.B., Gelfand, M.S., 2000. Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria. Mol. Microbiol. 38, 673–683.

Rodionov, D.A., Mironov, A.A., Gelfand, M.S., 2001. Transcriptional regulation of pentose utilisation systems in the Bacillus/Clostridium group of bacteria. FEMS Microbiol. Lett. 205, 305–314.

Rodionov, D.A., Mironov, A.A., Gelfand, M.S., 2002. Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. Genome Res. 12, 1507–1516.

Schneider, T.D., 1986. Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188, 415–431.

Stormo, G.D., Fields, D.S., 1998. Specificity, energy and information in DNA-protein interactions. Trends Biochem. Sci. 23, 109–113.

Tatusov, R.L., et al., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29, 22–28.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Koonin, E.V., 2002. Genome trees and the tree of life. Trends Genet. 18, 472–479.

Zaslaver, A., et al., 2004. Just-in-time transcription program in metabolic pathways. Nat. Genet. 36, 486–491.