Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy

Vitali A. Likhoshvai*, Yuri G. Matushkin

Laboratory of Molecular Evolution, Institute of Cytology and Genetics, Prospekt Lavrentieva 10, Novosibirsk 630090, Russia

Received 19 December 2001; revised 20 February 2002; accepted 20 February 2002

First published online 6 March 2002

Edited by Takashi Gojobori

Abstract We analyzed the interrelation between the efficiency of a gene expression and the nucleotide composition of all protein-coding sequences in 38 unicellular organisms whose complete genomic sequences are known. These organisms comprise 37 prokaryotic (29 eubacteria and eight archaebacteria) and one eukaryotic (yeast) species. We demonstrated that frequency analysis of gene codon composition fails to reflect adequately the gene expression efficiency of all these organisms. We constructed a measure, the elongation efficiency index, that considers simultaneously the information on codon frequencies and the degree of mRNA local self-complementarity. This measure recognizes the ribosome-coding genes as highly expressed in all the unicellular organisms studied. According to our analysis, these species fall into five groups differentiated by the process that makes the key contribution to the elongation rate. © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Gene expression; Local complementarity; Codon frequency; Mathematical model; Computer analysis

1. Introduction

The efficiency of protein synthesis of several organisms has been shown to correlate well with the codon frequencies of the corresponding genes [1–5]. The speed of codon reading has been correlated with the concentration of the corresponding tRNA fractions [6,7]. Analysis of the dynamic models of mRNA translation that take into account template-dependent protein synthesis, degeneracy of the genetic code, and availability of the adapter tRNAs has significantly advanced our understanding of relationship between the codon frequencies in a gene and the efficiency of translation of the corresponding mRNA [8–13], which in turn appears to be higher in the highly expressed genes. The correlation between the codon frequency bias and the efficiency of translation appears sufficient to predict gene expression levels in *Escherichia coli* and *Saccharomyces cerevisiae* [14], it is, however, not a good predictor for a number of other organisms [15–17].

In this work, we analyzed protein-coding genes of 38 organisms whose complete genomes are already known.

We suggest that more highly expressed genes display, on average, smaller average time required for one amino acid residue to be attached to the growing polypeptide chain (that is, they display faster elongation).

We defined an *elongation efficiency index* that considers simultaneously the codon composition profiles and those local secondary structures in mRNA being translated that can slow down the translation process. As we show below, gene expression efficiency and codon composition are not necessary correlated: in some cases, selection against mRNA self-complementarity appears to be stronger than selection for optimum codon composition.

In this context, we studied the relative contributions to elongation efficiency of the codon frequency bias and the local mRNA self-complementarities for the 38 completely sequenced organisms. Basing on the computed values of the elongation efficiency index, we divided the set of 38 organisms into five major groups that display markedly different patterns in terms of prevalence of selection for codon frequency bias or selection against local self-complementarity. Therefore, we may be able to use the elongation efficiency index as a better predictor of efficiency of gene expression than a codon frequency bias index alone.

2. Materials

We analyzed genes from the following bacteria: *Aquifex aeolicus*, *B. halodurans* C-125, *Bacillus subtilis*, *Borrelia burgdorferi*, *Buchera* sp. APS, *Campylobacter jejuni*, *C. muridarum*, *C. pneumoniae*, *C. pneumoniae* AR39, *C. pneumoniae* J138, *Chlamydia trachomatis*, *D. radiodurans* R1, *E. coli* K-12 MG1655, *Haemophilus influenzae* Rd, *Helicobacter pylori* 26695, *H. pylori* J99, *Mycobacterium tuberculosis* H367RV, *Mycoplasma genitalium* G37, *Mycoplasma pneumoniae* M129, *Neisseria meningitidis* strain MC58, *N. meningitidis* serogroup A strain Z2491, *Pseudomonas aeruginosa* PA01, *Rickettsia prowazekii* strain Madrid E, *Synechocystis* PCC6803, *Thermotoga maritima*, *Treponema pallidum*, *Ureaplasma urealyticum*, *Vibrio cholerae*, and *X. fastidiousa*; archebacteria *Archaeoglobus fulgidus*, *Aeropyrum pernix*, *Halobacterium* sp. NRC-1, *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum* delta H, *Pyrodictium abyssi*, *Pyrococcus horikoshii* OT3, and *Thermoplasma acidophilum*, as well as from the eukaryote *S. cerevisiae*. The nucleotide sequences and layout of the genes were extracted from EMBL (European Molecular Biology Lab database), release 58, and from GenBank.

The methods, implemented in Fortran, are described in Appendix A.

3. Results and discussion

Translation of mRNA in unicellular organisms is one of the most energy-consuming stages in the path from a gene to the corresponding protein. For example, an *E. coli* cell growing in

*Corresponding author. Fax: (7)-3832-33 12 78.
E-mail address: likho@bionet.nsc.ru (V.A. Likhoshvai).

0014-5793/02/$22.00 © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
PII: S0014-5793(02)02507-3

a nutrient-rich environment may expend, for translation, up to 50% of its material and energy resources. Consequently, the efficiency of the translation machinery may be a long-term factor in natural selection. The elongation stage is one of the stages that is potentially affected by this evolutionary optimization. The elementary act of elongation attachment of an amino acid residue to a growing polypeptide chain comprises three successive steps: (1) placement of a charged isoacceptor tRNA in the ribosomal A site, (2) transpeptidation, and (3) translocation. Low efficiency at any of these three stages can pose a selective disadvantage.

The concentrations of charged tRNAs determine primarily the mean tRNA placement time; therefore, the codon composition of translated mRNA does influence the elongation efficiency. Another possible way to improve the speed of translation is to speed up the transpeptidation or translocation stages.

Our current knowledge does not permit us to relate directly the mRNA sequence to the speed of transpeptidation, but we can use mRNA sequence to estimate the speed of ribosome translocation. Indeed, since the translocation stage of translation involves a physical advance of a ribosome along the mRNA at a distance of three nucleotides, it can be sped up through reduction of the number of local mRNA hairpin structures. We assumed that the time required for translocating a codon depends inversely on the amount of local mRNA self-complementarity in 15–65 nucleotide window sliding along the protein-coding part of the mRNA.

Thus, we related the time required for peptide-bond formation to the two following characteristics of nucleotide context: characteristics of codon usage frequency and characteristics of mRNA local secondary structures. We formalized these characteristics as a set of three indices (see Appendix A for formal mathematical definitions): a qualitative index T_a that estimates the mean time required for placement of an amino acyl tRNA; an index LCI ζ that summarizes the average number of self-complementary regions within mRNA window of a specified length, disregarding the energies of the secondary structures; and an index LCI ψ that reflects the average energy of the secondary structures over the mRNA region of a specified length. Using the method described in Appendix A and the latter two indices, we can determine the mean time required for a ribosome to translocate. We use a weighted sum of these three indices to define an elongation index, EEI.

To analyze the interplay among conflicting selective constraints acting on protein-coding gene regions, we ordered all genes in each of the 38 organisms according to the value of one index. We then calculated the position of genes encoding ribosome proteins (GRPs) in the ordered gene list (Table 1).

The GRPs are known to have a universally high expression level; therefore, we expected that the GRPs would tend to have index values smaller than those of the other genes (thus, the GRPs would be placed at the beginning of the list of genes, which is given in ascending index value order). To assess the significance of the position of the GRPs in the ordered list of genes, we calculated the probability that the observed distribution would occur in a random reordering of genes. If the probability was below 10^{-3}, we considered that the index evaluated the GRPs adequately as highly expressed genes and, consequently, that the rest of genes probably also were ranked correctly.

The larger the positive shift of the ribosome protein genes from the center of the ordered list, the more significant the relation between the characteristic observed (accounted for by the index calculated) and the expression efficiency. Note separately that we constructed the indices of several organisms by analyzing the frequency characteristics of incomplete codon sets (Table 2). A small fraction of codons was omitted from the analysis due to methodical considerations: The algorithm for ordering genes according to EEI values occasionally got caught in an endless loop because certain codons in these organisms occurred with unusual rarity in the genes that had the lowest EEI values (we interpret these genes as highly expressed).

We calculated EEI values in five different ways according to the factors that we took into account in the calculations (see Table 1):

1. In column A, the genes are ordered according to EEI($u_1=1,u_2=0$), taking into account the codon composition and disregarding local complementarity.
2. In column ζ, the genes are ordered according to EEI($u_1=0,u_2=1$), disregarding the codon composition and taking into account local complementarity ignoring the energies of potential secondary structures (LCI ζ form).
3. In column ψ, the genes are ordered according to EEI($u_1=0,u_2=1$), disregarding the codon composition and calculating local complementarity taking into account the energies of potential secondary structures (LCI ψ form).
4. In column Aζ, the genes are ordered according to EEI($u_1=1,u_2=1$), taking into account both the codon composition and local complementarity ignoring the energies of potential secondary structures (LCI ζ form).
5. In column Aψ, the genes are ordered according to EEI($u_1=1,u_2=1$), taking into account both the codon composition and local complementarity calculated taking into account the energies of potential secondary structures (LCI Aψ form).

It is evident from Table 1 that of the five calculated values of GRP shift from the center of the ordered list, there is at least one positive value of d that exceeds 40 for all the organisms except for *M. pneumoniae* M129. Thus, the observed regularity is significant ($P<10^{-7}$).

These data suggest that in highly expressed genes at least one of the two stages of elongation – tRNA attachment to ribosome or ribosome translocation – tends to be highly optimized for speed. The genome of *M. pneumoniae* M129 is a relatively rare exception, with a maximal shift amounting to $d=28$, which value corresponds to a relatively high but still highly significant P value of 4×10^{-4}. Presumably, the elongation efficiency of this particular organism depends weakly on both the codon composition and the mRNA secondary structure through mechanisms as yet unknown to us.

These results allow us to divide the organisms studied into five groups (see column Σ of Table 1). In the organisms belonging to group A, the efficiency of translocation is increased through optimization of codon composition and through optimization of the local self-complementarity (a weak selection against self-complementarity is conceivable for *Synechocystis* PCC6803 and *P. abbysi*).

Table 1
The shift from the center of the genes encoding ribosomal proteins in the total gene sample ordered according to values of various EEI modifications

Organism	*N*[a]	*K*[b]	A[c,d]	ζ[c,e]	ψ[c,f]	Aζ[c,g]	Aψ[c,h]	Σ[i]
B. halodurans	4066	55	**78**	−22	−33	−16	−32	A
B. subtilis	4178	53	**85**	0	2	0	15	A
C. muridarum	818	48	**44**	0	−19	40	13	A
C. pneumoniae	1052	52	**65**	−1	20	23	45	A
C. pneumoniae AR39	997	48	**67**	−7	20	62	53	A
C. pneumoniae J138	1087	52	**65**	−7	22	37	51	A
C. trachomatis	894	51	**45**	−12	−15	24	12	A
E. coli K12 MG1655	4289	56	**90**	12	8	38	21	A
H. influenzae	1709	57	**74**	27	−57	65	−40	A
P. abyssi	1765	64	**66**	42	−22	63	9	A/Aζ[j]
Synechocystis PCC6803	3168	54	**46**	18	−35	44	−40	A/Aζ[j]
S. cerevisiae	6337	78	**98**	−4	−44	29	−30	A
V. cholerae	3828	59	**95**	0	2	11	17	A
B. burgdorferi	850	54	−19	**49**	−4	42	−30	ζ
Buchnera sp. ASP	569	54	−47	**63**	0	45	−25	ζ
C. jejuni	1654	53	−38	**67**	−39	55	−53	ζ
H. pylori J99	1491	53	−38	**54**	−18	48	−26	ζ
H. pylori 26695	1566	53	−36	**50**	−23	43	−31	ζ
M. genitalium	480	51	−62	**56**	−48	0	−62	ζ
T. acidophilum	1478	54	22	**51**	−8	48	0	ζ/Aζ[j]
U. urealyticum	611	51	−15	**77**	−60	67	−58	ζ
P. aeruginosa PA01	5638	57	−68	79	**82**	58	78	ψ
A. aeolicus	1522	55	32	63	−14	**66**	−10	Aζ
D. radiodurans	2937	56	62	45	38	**73**	72	Aζ/Aψ[k]
M. tuberculosis	3920	55	25	20	9	**43**	30	Aζ
M. pneumoniae M129	677	51	−4	25	−37	**28**	−34	Aζ
R. prowazekii Madrid E	834	54	20	20	−13	**52**	−4	Aζ
T. maritima	1846	51	44	62	24	**69**	37	Aζ
T. pallidum	1031	52	27	47	31	**50**	37	Aζ
A. fulgidus	2407	61	50	40	0	**61**	29	Aζ
A. pernix K1	2694	57	54	34	21	**73**	56	Aζ
Halobacterium	2058	56	−27	31	24	**37**	28	Aζ
M. jannaschii	1715	65	−6	68	−18	**77**	−20	Aζ
M. thermoautotrophicum delta H	1869	61	26	56	25	**66**	27	Aζ
P. horikoshii OT3	2095	52	42	54	−5	**69**	14	Aζ
N. meningitidis MC58	1989	56	11	34	47	12	**79**	Aψ
N. meningitidis Z2491	2121	55	11	27	47	5	**77**	Aψ
X. fastidiosa	2766	55	−38	11	42	37	**64**	Aψ

[a] The total number of genes analyzed in each organism.
[b] The total number of genes annotated as genes encoding ribosomal proteins (GRPs).
[c] The index of GRP shift from the center of the genes sampled, ranked in ascending order of EEI value.
[d] > Genes ranked by EEI($u_1 = 1$,$u_2 = 0$).
[e] Genes ranked by EEI($u_1 = 0$,$u_2 = 1$) and LCI ζ.
[f] Genes ranked by EEI($u_1 = 0$,$u_2 = 1$) and LCI ψ.
[g] Genes ranked by EEI($u_1 = 1$,$u_2 = 1$) and LCI ζ.
[h] Genes ranked by EEI($u_1 = 1$,$u_2 = 1$) and LCI ψ.
[i] The stage that reflects maximally the expression efficiency.
[j] Complementarity presumably affects elongation efficiency.
[k] Secondary structure energy presumably affects translocation efficiency.

In contrast, for the organisms of group ζ, only the degree of self-complementarity seems to be optimized.

Group ψ contains only one organism, *P. aeruginosa* PA01. In this case, the elongation stage is optimized through decreases in the number and the energy of mRNA local secondary structures, rather than through optimization of the codon frequencies.

Group Aζ is most common, comprising 13 organisms; however, *D. radiodurans* may fall into group Aψ, which has three members. Typical of the organisms belonging to these groups is the optimization of elongation efficiency through both the codon compositions of genes and the degree of self-complementarity venues. Moreover, this dependence is cooperative: Taking into account both characteristics of gene nucleotide composition increases the GRP positive shift compared to the shifts calculated using individual indices. However, the energy of the secondary structures is unimportant for elongation in group Aζ organisms, whereas it is an essential factor for group Aψ organisms. Presumably, this distinction stems from differences in the mechanisms that these gene ribosomes use to overcome various hindrances at the stage of elongation, including those presented by local secondary structures. Our hypothesis is that, in the organisms belonging to groups ζ and Aζ, a hindrance (a hairpin) encountered by a ribosome triggers an unknown mechanism that removes the hindrance, regardless of the hairpin energy. In contrast, in group ψ and Aψ organisms, the effort towards removing hairpins is proportional to hairpin number and energy. There are at least two possible explanations for the independence of the elongation stage of the codon composition

Table 2
List of the organisms with codons that were discarded from the analysis

Organism	Codons discarded from analysis
A. aeolicus	cga, cgg, cgc, cgt, tgc, tgt
A. fulgidus	cga, cgg, cgc, cgt
C. jejuni	cgg
D. radiodurans R1	cga, agg, ata, tgt, tta, cta, tca
Halobacterium	aga, cta, tta, ata, cct, aat
H. influenzae Rd	cgg, agg
M. jannaschii	cga, cgg, cgc, cgt, tcg, ccg, ccc
P. aeruginosa PA01	aga, tta, tca, tct, aca, tgt, ata
P. abyssi	cga, cgg, cgc, cgt, tgc, tgt
P. horikoshii	cga, cgg, cgc, cgt, tgc, tgt
V. cholerae	cgg, agg
S. cerevisiae	cga, cgg

of genes, which is characteristic of groups ζ and ψ. First, the speed of amino acyl tRNA placement in the ribosomal site A is approximately similar for all the codons. Second, the tRNA placement in site A proceeds in parallel with certain processes providing a normal course of further elongation. The rate of the former process is considerably higher that those of the latter; thus, the latter process shadows the former one.

The fact that related organisms refer to the same group according to the way of optimization of the primary structure of highly expressed genes is quite natural exactly due to their relatedness. However, this distribution is not an artefact, because in a single group one could see hardly closely related organisms. For example, to the A group, we classify: *S. cerevisiae*, *V. cholerae*, *H. influenzae*, *C. trachomatis*, *E. coli*. To the Aζ group: *M. tuberculosis*, *R. prowazekii* Madrid E, *T. pallidum* eubacterium, whereas the rest organisms in the Aζ group refer to archaebacteria.

Acknowledgements: We are grateful to Galina Orlova and Galina Chirikova for translation of the manuscript into English. We express our great thanks to Dr. Andrey Rzhetsky for reviewing the manuscript and providing helpful suggestions and to Lyn Dupre for valuable comments on an earlier version of this article. This work was supported by the Interdisciplinary Integration Projects of the Siberian Branch of the Russian Academy of Sciences No. 65 and No. 66, and by the Russian Foundation for Basic Research (grants No. 00-04-49229 and 02-04-48802).

Appendix A

A.1. Elongation efficiency index (EEI)

Let G be a subset of 64 standard codons. We divide the codons from G into 21 synonymous groups corresponding to the 20 amino acids and translation termination codons; we will refer to each group of synonymous codons as a generalized codon. Let C be the total number of generalized codons, labeled by consecutive integer numbers. Since some codons may be missing in a particular gene, we may need to consider only a subset of all generalized codons; we will refer to such codon subset as *accountable* codons. While defining the EEI, we consider only generalized codons. Let N be the total number of genes in the sample analyzed; n_i, the total number of accountable codons in the ith gene ($i = 1,...,N$), and $\delta(i,j)$, the number of codons belonging to the jth accountable codon of the ith gene ($1 \le \delta(i,j) \le C$). We introduce an index EEI(i) that characterizes the nucleotide sequence of gene i and has a physical meaning of the mean elongation time per codon. We define EEI(i) as

$$\mathrm{EEI}(i) = u_1 T_{\mathrm{a}}(i) + u_2 T_{\mathrm{e}}(i)$$

where $u_1 = 0.1$ and $u_2 = 0.1$ are weight coefficients, which determine contribution of each term to this index. In total, only three non-trivial combinations of weight coefficients are meaningful in the context of our study: (1) if $u_1 = 1$ and $u_2 = 0$, only the term $T_{\mathrm{a}}(i)$ is taken into account; (2) if $u_1 = 0$ and $u_2 = 1$, only the term $T_{\mathrm{e}}(i)$ is taken into account; and (3) if $u_1 = 1$ and $u_2 = 1$, both terms $T_{\mathrm{a}}(i)$ and $T_{\mathrm{e}}(i)$ are taken into account. The first term, T_{a}, reflects the mean time required for isoacceptor amino acyl tRNA to attach to the ribosomal A site as determined by codon frequencies alone. The corresponding value is calculated according to the following equation:

$$T_{\mathrm{a}}(i) = \sum_{j=1}^{n_i} \beta_{\delta(i,j)}/n_i,\ \beta_{\delta(i,j)} = \frac{\sum_{m=1}^{C} \sqrt{\alpha_m}}{\sqrt{\alpha_{\delta(i,j)}}}$$

where the variable $1/\beta_{\delta(i,j)}$ is interpreted as the optimal relative concentration of amino acyl tRNA complementary to the jth accountable codon, while $\alpha_{\delta(i,j)}$ and α_m have the meaning of mean frequencies of the codons $\delta(i,j)$ and m in a gene [12].

The second term, $T_{\mathrm{e}}(i)$, estimates the mean time required for a ribosome to translocate when only local self-complementarity of the mRNA is taken into account:

$$T_{\mathrm{e}}(i) = t_{\min}\cdot(1-p(i)) + t_{\max}\cdot p(i)$$

where $t_{\min}$ is the minimal conventional time required for translocation; $t_{\max}$, the maximal conventional time required for translocation; and $p(i)$, the probability of realization of the maximal conventional time required for translocation calculated according to the following equation:

$$p(i) = \int_0^{\mathrm{LCI}(i)} \frac{k^{n+1}x^n}{G(n+1)} \mathrm{e}^{-kx}\mathrm{d}x,\ k = m/\sigma^2,\ n = (m/\sigma)^2$$

where m and σ^2 are the mean and variance, respectively, of the positive random variable with a distribution density of

$$p(i) = \frac{k^{n+1}x^n}{G(n+1)} \mathrm{e}^{-kx}$$

$G(n+1)$ is a gamma function, and LCI(i), the local complementarity index.

We used the two following forms of the local complementarity index:

A.1.1. LCI ζ form

The average number of self-complementary regions in a window of m_i nucleotides is calculated according to the following equation:

$$\mathrm{LCI}\ \zeta = \frac{\sum_{m=1}^{m_i-s_{\max}-l_{\max}} \left\{ \sum_{s=s_{\min}}^{s_{\max}} \left[\sum_{l=l_{\min}}^{l_{\max}} \zeta(\mathrm{con}(m,m+s-1), \overline{\mathrm{con}(m+s+l-1,2m+2s+l-2)}) \right] \right\}}{m_i-s_{\max}-l_{\max}} \quad (1)$$

In this computation we disregard information about the energy of hairpins formed in mRNA. In the latter formula m_i equals the triple number of codons contained in the ith gene plus 53 nucleotides from its 3′ end, con(i,j) is just the sequence of nucleotides between ith and jth nucleotides; $\overline{\mathrm{con}(i,j)}$ denotes sequence that is complementary to nucleotide sequence between ith and jth nucleotides; and ζ(conext1,conext2) has value 1, if the words conext1 and conext2 are identical, otherwise ζ(conext1,conext2) = 0. The length of accountable inverted repeat falls between $s_{\min}$ and $s_{\max}$; the distance between accountable inverted repeats falls between $l_{\min}$ and $l_{\max}$ (in this paper, $s_{\min} = s_{\max} = 3$, $l_{\min} = 3$, and $l_{\max} = 50$).

A.1.2. LCI ψ form

The average energy of a random hairpin formed in the window of m_i nucleotides (here, $s_{\min} = 3$, $s_{\max} = 6$, $l_{\min} = 3$, and $l_{\max} = 50$) is calculated with the following equation:

$$\mathrm{LCI}\ \psi(i) = \frac{\sum_{m=1}^{m_i-s_{\max}-l_{\max}} \left\{ \sum_{s=s_{\min}}^{s_{\max}} \left[\sum_{l=l_{\min}}^{l_{\max}} \psi(\mathrm{con}(m,m+s-1), \overline{\mathrm{con}(m+s+l-1,2m+2s+l-2)}) \right] \right\}}{m_i-s_{\max}-l_{\max}} \quad (2)$$

where ψ is the energy of a hairpin, calculated conventionally [18].

The time required for a ribosome to perform the transpeptidation stage was assumed equal for all the codons and genes and, hence, was disregarded while constructing EEI(i).

A.2. Algorithm for automated gene ordering

This algorithm ranks the genes iteratively in an ascending order of EEI value. To run the algorithm one first needs to specify M highly expressed genes ($0 < M \leq N$) and choose one of the three forms of EEI index. In this algorithm, the M genes that display the minimal index values at the previous iteration are used to calculate the $\beta_{\delta(i,j)}$ values, which are in turn used to calculate the indexes of the genes at the next iteration. See [13] for detailed description of this algorithm.

All the accountable codons were used in analysis in the case of organisms not indicated in Table 2. The order of genes for these organisms was set after 10–15 iteration cycles. However, this algorithm got caught in an endless loop when applied to the organisms listed in Table 2. This infinite looping manifested itself as a cyclic change in the gene order after several iterations, so that the orders at two successive iteration steps were different. The infinite looping is explained by the presence of codons with abnormally low frequencies in sequences of the genes displaying minimal EEI values calculated by the algorithm. We inferred that such codons may be involved in a specific regulation and discarded them from the analysis. This allowed us to stabilize the algorithm operation for all the organisms involved.

We excluded the theoretical possibility that the algorithm proposed assigns low values to GRPs because they fall into the sample of M genes with the lowest values (the artifact of self-recognition) through parallel computations performed for all the organisms analyzed when the codons contained in GRPs were excluded from the statistics of codon usage. The results obtained were principally similar to the data listed in Table 1.

A.3. Specification of values of the parameters M, u_1, u_2, $t_{\min}$, $t_{\max}$, m, and σ

It is necessary to specify the parameters M, u_1, u_2, $t_{\min}$, $t_{\max}$, m, and σ for the algorithm of automated gene ordering to run correctly.

A.3.1. Parameter M

We used the same value of this parameter ($M = 300$) for all the organisms studied. However, the results remained principally the same when M was varied from 50 to 1000 while analyzing the genes of *V. cholerae*.

A.3.2. Parameters u_1, u_2

All the non-trivial combinations of the parameters were used, namely $u_1 = 1$, $u_2 = 0$; $u_1 = 0$, $u_2 = 1$; and $u_1 = u_2 = 1$.

A.3.3. Parameters t_{min}, t_{max}, m, and σ

Specification of these parameters has meaning only when $u_1 = 1$ and $u_2 = 1$. Then, different weights may be assigned to the terms T_{a} and T_{e} varying $t_{\min}$ and $t_{\max}$, while m and σ specify the degree of LCI (local complementarity index) impact on the T_{e} value. We included the terms T_{a} and T_{e} into EEI with equal weights. For this purpose, we specified $t_{\min} = 0$, $t_{\max} = 0$ for the first iteration step, so that this first step calculated the gene indices disregarding the complementarities; and then after each iteration step, we calculated the minimal and maximal values of the term T_{a} over all the genes and used them as $t_{\min}$ and $t_{\max}$, respectively, for each subsequent iteration step. The arithmetic mean of LCI values over all the genes was used as m. A double value of the mean deviation between LCI and m was used as σ for all the organisms except for those listed below: triple value for *M. tuberculosis*, and four-fold for *M. pneumoniae*, *A. fulgidus*, *A. pernix*, *M. thermoautotrophicum*, *P. abyssi*. The multiplicity was selected to maximize the shift of GRPs from the center (see the precise definition of the shift below).

A.4. Method for estimating the EEI compliance with expression efficiency

Assume that the GRPs are highly expressed in all the organisms analyzed. Consequently, if the EEI values for these genes are low a low EEI index appears to be a good predictor of high gene expression.

To evaluate the significance of this criterion, we calculated P values of the GRP distribution under a null model (a random permutation of genes). Assume that the GRP set consists of K genes ($K < N$) and designate this set as $\Im$. Let $n(g)$ be the rank of gene g in the sample ranked according to ascending order of EEI values. The measure of sample $\Im$ deviation from the mean is calculated according to the following equation:

$$d(\Im) = 100\cdot\left(2\sum_{g\in\Im} n(g)/K - N - 1\right)/(N-K)$$

The requirement that $-100 \leq \mathrm{d}(\Im) \leq 100$ is tested directly. If the location of the gene sample $\Im$ in the determined gene order is random, we can calculate the probability $p = p(d \geq d(\Im))$ of a random sample of K genes to have the deviation of not less than $d(\Im)$. If the obtained value of p is small, we interpret this as a confirmation of the non-random location of GRPs in the fixed order of all the genes. We took $\varepsilon = 10^{-3}$ as a significance threshold.

References

[1] Sharp, P.M. and Li, W.-H. (1986) Nucleic Acids Res. 14, 7737–7749.

[2] Sharp, P.M. and Devine, K.M. (1989) Nucleic Acids Res. 17, 5029–5039.

[3] Shields, D.C. et al. (1988) Mol. Biol. Evol. 5, 704–716.

[4] Shields, D.C. and Sharp, P.M. (1987) Nucleic Acids Res. 15, 8023–8040.

[5] Lloyd, A.T. and Sharp, P.M. (1993) Yeast 9, 1219–1228.

[6] Ikemura, T. (1985) Mol. Biol. Evol. 2, 13–24.

[7] Yamato, F. et al. (1991) Nucleic Acids Res. 19, 7737–7749.

[8] Bulmer, M. (1991) Genetics 129, 897–907.

[9] Bulmer, M. (1987) Nature 325, 728–730.

[10] Shields, D.C. (1990) J. Mol. Evol. 31, 71–80.

[11] Baglioni, F. and Lio, P. (1995) J. Theor. Biol. 173, 271–281.

[12] Likhoshvai, V.A. (1992) in: Modeling and Computer Methods in Molecular Biology and Genetics (Ratner, V.A. and Kolchanov, N.A., Eds.), pp. 463–469, Nova Sc. Publishers, New York.

[13] Likhoshvai, V.A. and Matushkin, J.-G. (2000) Comput. Technol. 5 (N2), 57–63.

[14] Li, W.-H. and Luo, L.J. (1996) J. Theor. Biol. 181, 111–124.

[15] Andersson, S.G. and Sharp, P.M. (1996) J. Mol. Evol. 42, 525–536.

[16] Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M. and Wolfe, K.H. (1999) Nucleic Acids Res. 27, 1642–1649.

[17] Likhoshvai, V.A. and Matushkin, Y.-G. (2000) Mol. Biol. (Mosk.) 34, 345–350.

[18] Turner, D.H. and Sugimoto, N. (1988) Annu. Rev. Biophys. Chem. 17, 167–192.