J Mol Evol (2004) 59:520–527
DOI: 10.1007/s00239-004-2644-5

JOURNAL OF MOLECULAR EVOLUTION

© Springer Science+Business Media, Inc. 2004

Overlapping Messages and Survivability

Ofer Peleg, Valery Kirzhner, Edward Trifonov, Alexander Bolshoy

Genome Diversity Center, Institute of Evolution, Haifa University Mt. Carmel, Haifa 31905, Israel

Received: 13 October 2003 / Accepted: 10 May 2004 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. The phenomenon of overlapping of various sequence messages in genomes is a puzzle for evolutionary theoreticians, geneticists, and sequence researchers. The overlapping is possible due to degeneracy of the messages, in particular, degeneracy of codons. It is often observed in organisms with a limited size of genome, possessing polymerases of low fidelity. The most accepted view considers the overlapping as a mechanism to increase the amount of information per unit length. Here we present a model that suggests direct evolutionary advantage of the message overlapping. Two opposing drives are considered: (a) reduction in the amount of vulnerable points when the overlapping of two messages involves common critical points and (b) cumulative compromising cost of coexistence of messages at the same site. Over a broad range of conditions the reduction of the target size prevails, thus making the overlapping of messages advantageous.

Key words: Mutation rate — Lethal mutation — Survival — Fitting landscape

Introduction

DNA sequences are often compared with the texts written in a language, either a natural human language or an artificial one. The fundamental difference is that all usual human texts are organized linearly so that if there is more than one message, the messages do not overlap but follow one another. This does not apply to natural genetic DNA sequences. The same fragment of a sequence typically carries more than one functional message, and these messages overlap. Holliday (1968) predicted this phenomenon considering signals responsible for recombination, which may well reside within the protein-coding sequences. This would be allowed, perhaps, due to degeneracy of the triplet code. The idea was developed further in works by Schaap (1971), Zuckerkandl (1976), Eigen and Schuster (1979), and Trifonov (1981) and confirmed by numerous experimental data, as reviewed by Normark et al. (1983). In a generalized form (Trifonov 1989) any sequence pattern that corresponds to a certain function is considered a code. Due to the degeneracy of the codes, "corresponding messages are not only interspersed, but actually overlapped, so that some nucleotides belong to several messages simultaneously." The idea of the multiplicity of the codes of nucleotide sequences and their overlapping also appeared in studies by Caporale (1984), Staden (1984), Kypr (1986), and Konings et al. (1987) (reviewed by Trifonov 1996).

One example of such overlapping is when a biologically significant RNA secondary structure overlaps sequencewise with the protein-coding region. Several well-known RNA secondary structures along the HIV-1 genome play various functional roles during the virus life cycles, while residing in protein-coding regions. One of the best known is the RNA structure of the Rev Responsive Element (RRE), which interacts with the Rev Trans-Activator protein (Dayton et al. 1989; Kjems et al. 1991; Malim et al. 1989, 1990). Biologically functional RNA secondary

Correspondence to: Ofer Peleg; *email:* peleg@research.haifa.ac.il

structure motif may stimulate ribosomal frameshifting in the gag–pol overlapping region in retroviruses. The exact nature of this signal is still controversial. Various researchers define it as either a pseudoknot (Le et al. 1991; Morikawa and Bishop 1992), a stem–loop structure (Parkin et al. 1992; Vickers and Ecker 1992), or a two-stem structure (Dulude et al. 2002). The translational frameshifting message is also suggested to reside in (G–non-G–N)$_n$ mRNA periodicity (Trifonov 1987, 1992; Lagunez-Otero and Trifonov 1992), which may be responsible for monitoring the correct reading frame during translation by forming transient complementary complexes with the C-periodical structure of the ribosomal RNA. Staple and Butcher (2003) fortified the frameshifting theory by NMR analysis of the RNA of the gag–pol region. This analysis indicated a frameshift-inducing stem–loop element made of an A form helix capped by a structured ACAA tetraloop.

Superimposed messages can be revealed even without knowing the biological meaning of the hidden message. A simple statistical approach based on processing data from a combination of multiple alignments and RNA secondary structure predictions, allowed to deduce distinct RNA folds in the first conserved (C1) protein-coding region of the *env* gene (Peleg et al. 2002) and in the β3/β4 region of the *nef* gene of HIV-1 (Peleg et al. 2003). It is important to distinguish between overlapping messages that play a role in the expression of the gene they overlap with and those that are functionally unrelated to the gene. The first group includes ribosomal frameshift signals, attenuators, antiterminators, and nucleosome positioning codes. An example of a message from the second group is the RRE RNA secondary structure encoded by the sequence of the *env* gene in retroviruses, which does not play any role in expression of the protein gp160 encoded by the same sequence.

Viruses often have very compact genomes, and overlapping of messages in this case occurs frequently. Konings (1992) studied a specific type of multiple-coding problem in lentiviruses, taking the RRE structure as a unique biological example. TAR is another well-known RNA secondary structure of retroviruses, which is not located inside any coding sequence. Comparison of the mutation rate between the TAR and the RRE regions revealed that in the case of RRE overlapping with the protein-coding message could be viewed as a factor constraining the evolutionary divergence of the element by increasing its selective value. Wagner and Stadler (1999) studied the genomes of single-stranded RNA viruses (dengue virus, hepatitis C virus, and HIV-1), focusing on the mutational stability of conserved and nonconserved viral secondary structure elements. Using this comparison, they concluded that the mutation robustness and monomorphism are associated with the RNA secondary structure message that overlaps with the protein coding.

The overlapping messages occur frequently in prokaryotic genomes and prokaryote-derived organelles such as mitochondria (Normark et al. 1983). Fukuda et al. (1999) identified 160 overlapping gene pairs in the genome of *M. pneumonia* and 155 overlapping gene pairs in the genome of *M. genitalium*.

A number of reports indicate the existence of overlapping messages in nuclear genomes of eukaryotes. Coelho et al. (2002) found that coding information for protein (TAR1) and structural RNAs (rRNA) in *Saccharomyces cerevisiae* can overlap, raising issues regarding the coevolution of such complex genes. Shintani et al. (1999) reported the overlapping of genes ACAT2 and TCP1 in their 3′ untranslated regions (UTRs). They appear in a tail-to-tail orientation, while their coding sequences are located on the opposite strands. Zhou and Blumberg (2003) found that the genes VLCAD and DLG4 are arranged in a head-to-head orientation on human chromosome 17p13 and share a 245-bp overlapping region that contains part of DLG4 exon 1 and the entire exon 1 of VLCAD including 62 bp of protein-coding sequence. Edgar (2003) found that the genes ABHD1 and Sec12 overlap. These genes, located on human chromosome 2p23.3, share 42 bp of the 3′-UTR in an antisense manner.

Species with a low quality of replication can maintain only a short genome (Eigen and Schuster 1979), which may be too small to store all the necessary information in a sequential manner. To increase the amount of information that can be stored, "the quantity of information per length unit has to be increased; i.e., part of the genome has to code for multiple functions" (Huynen et al. 1993). Hogeweg and Hesper (1992) studied the evolutionary dynamics leading to multiple coding. They showed that a high mutation rate and crossing-over lead to "multiple coding." However, they concluded that "multiple coding often does not increase the fitness of the population; nevertheless, it is selected." Huynen et al. (1993) evaluated the transition from RNA primary sequences to RNA secondary structure in the RRE region in the *env* gene of lentiviruses and Visna virus. The results of this study indicated a variation in the initial ruggedness of fitness landscapes that plays an essential role in the evolution and optimization of RNA secondary structure encoded in the translated region. On the other hand, simulation of an evolutionary search process for a specific secondary structure shows a reduction of allowable point mutations and a reduction of the possibility for small-scale adaptation. Apparently, this decreases the final fitness of the region of overlapping. High fitness as a prerequisite for multiple coding is discussed by Pavesti et al. (1997). Studying the informational content of overlapping genes in prokaryotic and eukaryotic viruses, the authors revealed an increased frequency of amino acid residues with high levels of degeneracy in proteins encoded by overlapping genes. Krakauer (2000) proposed a mathematical model to estimate the stability and evolution of overlapping genes in various orientations in terms of information cost. Krakauer assumed that the superposition increases coupling between functionally related genes and concluded that overlapping at the 3′ end decays more slowly than that at the 5′ end.

If the cost of the superposition is so high, is it at all advantageous? In this article, we propose two versions of a model in which multiple coding directly enhances the fitness under conditions of a high mutation rate.

Models of Advantageous Overlapping of Biological Messages

A Simple Model

Let us consider an example with two adjacent messages in a given genome. Each message contains several crucial residues (uppercase) along its base sequence.

```
Seq 1:
ActggtGttaTCtttaCcgATAggaTGgccttActC
Seq 2:
CaGggaaggAaaCagtTAgCcaGtcaAtcgGtagT
```

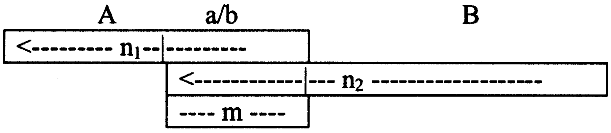
Here the total number of crucial residues in both messages is 23 ($N = 23$). For simplicity, other residues are here considered neutral, replaceable by any other residue. Consider now the overlapping as below, such that the matching residues (boldface letters) are identical in both sequences:

```
Seq 1:          ActggtGttaTCtttaCcgATAggaTGgccttActC---
                      |  |     || | |||   | |  |
Seq 2:          --- CaGggaaggAaaCagtTAgCcaGtcaAtcgGtagT
Seq_overlap:    ActgCaGgtaTCtAtaCcgATAgCaTGtccAtAcGCagT
```

The superposition ($\mathrm{Seq}_{\mathrm{overlap}}$) contains both messages, but the total number of crucial residues (uppercase) including common ones is now smaller (18). A reduction in the number of crucial residues (target size) by 5, thus, would increase the fitness of the organism. In other words, for fitness λ and number of crucial positions N we have $\lambda \downarrow N$. In the case of overlapping messages, $N_{\mathrm{overlap}} < N_{\mathrm{non}}$. As a result, $\lambda_{\mathrm{overlap}} > \lambda_{\mathrm{non}}$.

The Model Based on Probabilities of Lethal Mutations per Nucleotide in A, B, and A/B

The above model can be described in detail by considering two sequence messages sliding toward their superposition forming a common region. Let us have two messages, A and B. Message A has length n_1; message B has length n_2. m is the length of the overlapping region a/b in the A/B merge.



Let μ_1 be the probability of a (lethal) mutation per one nucleotide for message A. (Further on, we consider only lethal mutations.) Let μ_2 be the probability of a mutation per one nucleotide of message B. Therefore, the probability that there will be no mutation in the interval n_1–m is equal to $(1-\mu_1)^{n_1-m}$, and by the same token the probability of no mutation in the interval n_2–m is $(1-\mu_2)^{(n_2-m)}$. The probability that there will be no mutation in the overlapping interval m is equal to $(1-\nu)^m$. Consequently, the probability P_m of the total absence of any mutation in the whole interval is equal to the product of these probabilities,

$$P_m = (1-\mu_1)^{n_1-m} \cdot (1-\mu_2)^{n_2-m} \cdot (1-\nu)^m \qquad (1)$$

Obviously, $0 \le \mu_1, \mu_2, \nu \le 1$. Let us rewrite P_m

$$P_m = C\left(\frac{1-\nu}{(1-\mu_1)(1-\mu_2)}\right)^m \qquad (2)$$

where the constant C

$$C = (1-\mu_1)^{(n_1)}(1-\mu_2)^{(n_2)} \qquad (2)$$

From (2), it follows that the extreme value of P_m is reached at one of the limits of the parameter range. Namely, if $(1-\nu) < (1-\mu_1)(1-\mu_2)$, then the value P_m reaches a maximum when $m = 0$ ($P_0 > P_i$ for $0 < i \le m$). If $(1-\nu) > (1-\mu_1)(1-\mu_2)$, then the value P_m reaches the maximum when m is maximal, that is, $m = \min(n_1, n_2)$ ($P_m > P_i$ for $0 \le i < m$). Note that parameter ν, in principle, may take any value between zero and one in a species-dependent or environment-dependent manner. One reasonable possibility is that ν is equal to the larger of the probabilities μ_1 and μ_2: $\nu = \max(\mu_1, \mu_2)$. In this case, the inequality $(1-\nu) > (1-\mu_1)(1-\mu_2)$ is always true; consequently, the optimum corresponds to the maximal possible overlap. In the case of $\nu = \mu_1\mu_2$, the probability that there would not be a mutation in the interval m is equal to $(1-\mu_1\mu_2)^m$, and again, $(1-\nu) > (1-\mu_1)(1-\mu_2)$.

To illustrate this equation graphically we assumed that the probability of a lethal mutation per one nucleotide μ_1 is equal to mutation rate α multiplied by "index of lethality" λ and $\nu = \mu_1 = \mu_2$. Then the survival index σ for estimating the difference between the nonoverlapping situation and overlapping of m bases ($\sigma = P_0 – P_m$) is

$$\nu = \mu_1 = \mu_2 = \alpha\lambda, \sigma = P_0 - P_m$$

$$\sigma = C\left(\frac{1-\nu}{(1-\mu_1)(1-\mu_2)}\right)^0 - C\left(\frac{1-\nu}{(1-\mu_1)(1-\mu_2)}\right)^m = C\left(1 - \left[\frac{1}{1-\alpha \cdot \lambda}\right]^m\right) \tag{3}$$

Figure 1A–C illustrate how the survival index is influenced by the index of lethality, mutation rate, and length of the superposition. All figures correspond to mutation rate values between 0 and 0.0001 and superposition lengths between 0 and 100 bases. Figure 1A corresponds to lethality index $\lambda = 0.3$ (30% of critical points). Figure 1B and C correspond to lethality index $\lambda = 0.6$ and $\lambda = 0.9$, respectively. We would like to emphasize that the actual values of the survival index are not important. For the purposes of this analysis what matters is its behavior as a function of the parameters.

Figure 1A–C demonstrate that the survival index increases with degree of overlapping, in particular, under conditions of a high mutation rate. This is a simple formal illustration of the general idea of the advantage of overlapping. Formally the survival index as defined and parameterized in this model would increase indefinitely and cause the whole genome to overlap. This situation is clearly unrealistic. The penalty should be introduced to take into account a consequence of the compromise, which decreases the ability of the overlapping (degenerate) messages to vary. The compromise of the overlapping may lead, for example, to a nonoptimal choice of an amino acid at a particular location in the protein. It might cause a mismatch of a specific nucleotide in an RNA secondary structure to its counterpart in the opposite strand of the stem. Both would cause a reduction of the phenotypic repertoire. This justifies the assumption that the fitness also depends on the length of the overlap. Let us introduce the penalty for the overlapping as a function f(m) that decreases with an increase in m. Then a general fitness λ_m becomes

$$\lambda_m = P_m * \mathrm{f}(m) \tag{4}$$

Examples with different possible types of penalty function f(m) are considered below.

(1) f(m) = $(1–\beta)^m$. In this case a penalty for each overlapping nucleotide (letter) equals a constant value β. Then the probability that no letter would be penalized is $(1–\beta)^m$.

$$\lambda_m = C\left(\frac{1-\nu}{(1-\mu_1)(1-\mu_2)}\right)^m (1-\beta)^m = C\left(\frac{(1-\nu)(1-\beta)}{(1-\mu_1)(1-\mu_2)}\right)^m \tag{5}$$

Obviously, the maximum of λ_m is reached either when $m = 0$ or when $m = \min(n_1, n_2)$. As in the previous case, if $(1–\nu)(1–\beta) < (1–\mu_1)(1–\mu_2)$, then P_m is maximal when $m = 0$. If $(1–\nu)(1–\beta) > (1–\mu_1)(1–\mu_2)$, then P_m is maximal when $m = \min(n_1, n_2)$.

(2) f(m) = $1/(1+m)$. In this case

$$\lambda_m = C\left(\frac{1-\nu}{(1-\mu_1)(1-\mu_2)}\right)^m \frac{1}{1+m} \tag{6}$$

If $(1–\nu) < (1–\mu_1)(1–\mu_2)$ then λ_m decreases with an increase in m. This means that $m = 0$ (no overlap) has the best fitness. If $(1–\nu) > (1–\mu_1)(1–\mu_2)$ then λ_m is not a monotonous function of m. An extreme value would be when

$$m_0 = \frac{1}{\ln\left(\frac{(1-\nu)}{[(1-\mu_1)(1-\mu_2)]}\right)} - 1 \tag{7}$$

However, it is easy to show that λ_{m0} is the minimum of the function. Thus, as before, the maximum of λ_m is reached either when $m = 0$ or when $m = \min(n_1, n_2)$.

(3) $\mathrm{f}(m) = \left[\frac{1}{(1+\omega m)}\right]^m$, where ω is a relatively small value. This penalty function has a weaker length dependence than in case 2. Consequently,

$$\lambda_m = C\left(\frac{1-\nu}{(1-\mu_1)(1-\mu_2)}\right)^m \left(\frac{1}{1+\omega m}\right)^m \tag{8}$$

An extreme value would be reached when

$$m = \left[\frac{1-\nu}{(1-\mu_1)(1-\mu_2)} - 1\right] \Big/ \omega \tag{9}$$

$0 \leq m_0 \leq \min(n_1, n_2)$ and, in principle, may be any value in this range. If $\nu = \max(\mu_1, \mu_2)$, assuming $\mu_1 > \mu_2$, then

$$m = \left[\frac{\mu_2}{(1-\mu_2)}\right] \Big/ \omega \tag{10}$$

That is, approximately

$$m = \left(\frac{\mu_2}{\omega}\right) \tag{11}$$

Thus, the optimal m, which is the optimal length of the overlapping region, depends on the ratio between the mutation rate μ and the coefficient ω of the cumulative influence of an overlapping length m in the penalty function f(m).

From formulae (10) and (11) it is apparent that the optimal m does not depend upon the lengths of the messages. It should obey the inequality $0 \leq m \leq$ min

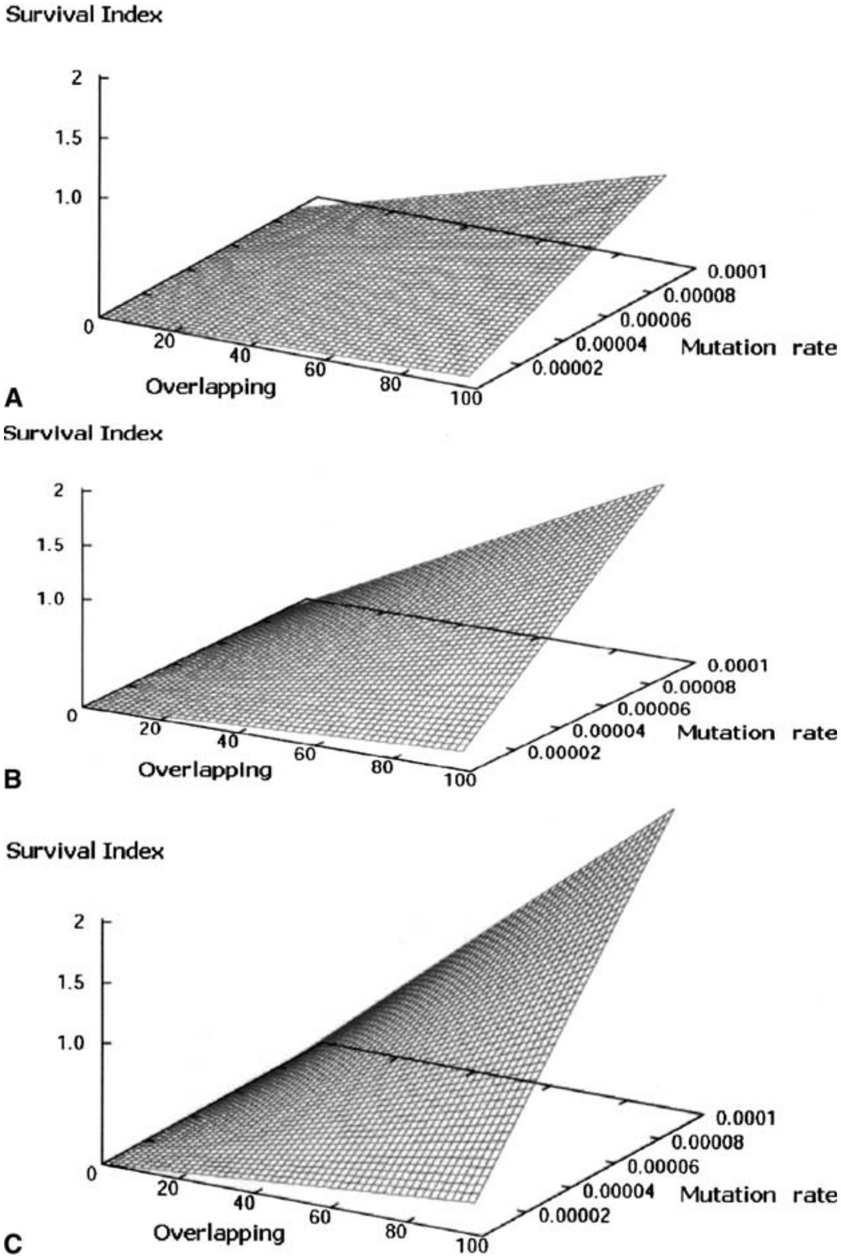


Fig. 1. Survival index as a function of overlap length and mutation rates. **A** λ = 0.3 (30% of critical points). **B** λ = 0.6. **C** λ = 0.9.

(n_1,n_2), and it may happen that m would be less than required by (10) and equal to min (n_1,n_2), which is the case of the maximal possible overlap.

Figure 2A–C correspond, respectively, to Figs. 1A–C modified by the addition of the penalty function. The penalty function is $\mathrm{f}(m) = \left[\frac{1}{(1+\omega m)}\right]^m$. The survival index $\sigma = \lambda_0 - \lambda_m$ is

$$\begin{aligned}\sigma &= C\left(\frac{1-\nu}{(1-\mu_1)(1-\mu_2)}\right)^0 - C\left(\frac{1-\nu}{(1-\mu_1)(1-\mu_2)}\right)^m \\ &\quad \cdot \left(\frac{1}{1+\omega m}\right)^m \\ &= C\left(1 - \left[\frac{1}{(1-\nu)(1+\omega m)}\right]^m\right) \qquad (12)\end{aligned}$$

The survival index indicates an evolutionary advantage for overlapping as long as it is higher than zero. Figure 2 shows that since the penalty depends only on the extent of the overlapping, its increase leads to accumulation of the compromises. The information contained in at least one of the messages decreases, which leads to the loss of the evolutionary advantage of the overlapping. Values of the parameters in the formulae above (overlapping length m, mutation rate μ, lethality index λ, parameter ω) are only of illustrative significance. Moreover, each sequence position has, actually, its own penalty for the message overlapping. However, the plots above illustrate in a semiquantitative way how the accumulation of compromises that

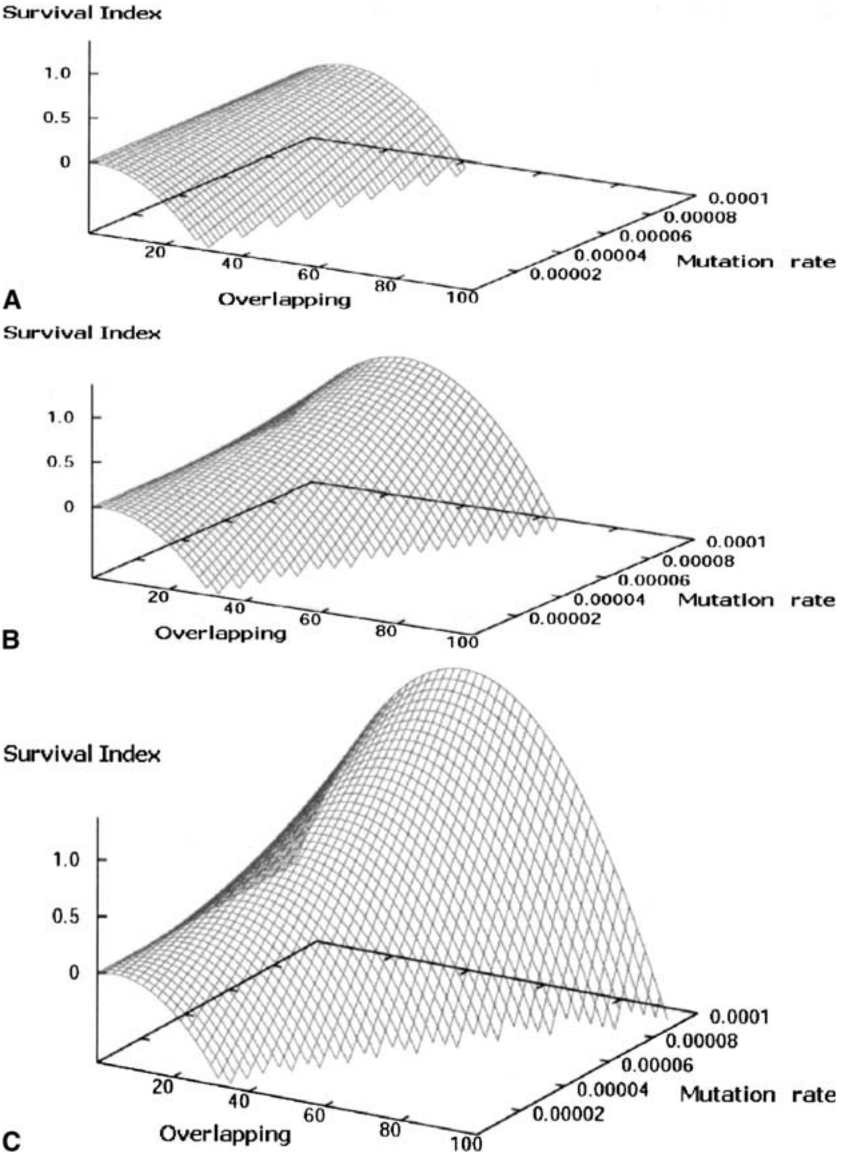


Fig. 2. Same as Fig. 1, with penalty introduced. **A** λ = 0.3. **B** λ = 0.6. **C** λ = 0.9 (90% of critical mutations).

accompanies the overlapping restricts the length of the overlapping region. This occurs even under favorable conditions for overlapping such as a high mutation rate and reduction of the number of crucial points.

The Model Based on Probabilities of Lethal Mutations per Interval

In this model we assume that a probability of one (or more) lethal mutation within the interval is proportional to the length of the interval. The definitions follow.

Q_m is the probability that a lethal mutation happens inside the overlapping interval,

$$Q_m = \frac{n_1 + n_2 - m}{n_1 + n_2} = 1 - \frac{m}{n_1 + n_2} \tag{13}$$

and the probability P_m,

$$P_m = \frac{m}{n_1 + n_2} \tag{14}$$

to compare with (1).

From (14) it follows that the optimal m corresponds to the maximal overlap $m = \min(n_1, n_2)$. Let us introduce a penalty for each overlapping nucleotide (letter) equal to a constant value β. Then the probability that no letter would be penalized is $(1-\beta)^m$. Accordingly

$$\lambda_m = \frac{m}{n_1 + n_2}(1 - \beta)^m \tag{15}$$

and the extreme value of a general fitness λ_m would be reached when

$$m = \frac{1}{\ln(1 - \beta)} \approx \frac{1}{\beta} \tag{16}$$

A very simple and meaningful conclusion follows from this formula: The higher the penalty for the overlapping, the lower the advantage of the overlapping.

Discussion

A genome region with two overlapping messages is considered, such that each message contains several crucial residues, while other residues are assumed to be neutral. The models are proposed to demonstrate how the superposition of critical points and mutation rate may influence the survival index. The interplay between an advantage of the superposition and a penalty for impairment of an advantage of the variability of the (degenerate) message is analyzed. The conclusion is that under reasonable assumptions the degree of the optimal overlapping depends on the ratio between the mutation rate and the cumulative penalty function. A low mutation rate and incompatibility of the messages (too high a cost for the compromise in the overlapping region) make overlapping disadvantageous. Importantly, a high mutation rate and high tolerance of the juxtaposed sequence messages make the superposition beneficial.

Usually, the phenomenon of overlapping is explained by the need to store a large quantity of information in a small genome (Eigen and Schuster 1979). The model that we present here is in general accord with previous models associating high mutation rate with message overlapping. However, unlike the previous models (Hogeweg and Hesper 1992; Huynen et al. 1993), our analysis suggests a direct evolutionary advantage for message overlapping under conditions of a high mutation rate. Moreover, our model proposes a particular mechanism to realize this evolutionary advantage through superposition of critical points, thus reducing their amount in the genome. This simple model involves two opposing forces that balance the degree of overlapping. The first force is the reduction of the number of vulnerable points and the second, opposing factor is a penalty for deterioration of the messages, which gradually reach the point of zero gain. This penalty is quite similar to the information cost described by Krakauer (2000). He described the information cost in terms of the tendency of a population to become monomorphic, which restricts the ability of polypeptides to be fine-tuned. Although our model describes the overlapping messages as sliding toward each other, we do not mean that this is the only possible mechanism. The sliding presentation is only a simplification that helps to illustrate the idea of two opposing forces optimizing the effect of overlapping.

Acknowledgments. We thank all members of the Genome Diversity Center for fruitful discussions and critical comments on the paper. We thank Ms. Naava Rubinstein for the editing. A.B. is grateful to Professors T. Ratiu and J.H. Maddocks, Bernoulli Institute, Swiss Federal Institute of Technology, for the kind invitation to visit Centre Bernoulli for two months.

References

Caporale LH (1984) Is there a higher level genetic code that directs evolution? Mol Cell Biochem 64:5–13

Coelho PS, Bryan AC, Kumar A, Shadel GS, Snyder MA (2002) Novel mitochondrial protein, Tar1p, is encoded on the antisense strand of the nuclear 25S rDNA. Genes Dev 16:2755–2760

Edgar AJ (2003) The gene structure and expression of human ABHD1: Overlapping polyadenylation signal sequence with Sec12. BMC Genomics 4:18

Dayton E, Powell D, Dayton A (1989) Functional analysis of CAR, the target sequence for the Rev protein of HIV-1. Science 246:1625–1629

Dulude D, Baril M, Brakier-Gingras L (2002) Characterization of the frameshift stimulatory signal controlling a programmed -1 ribosomal frameshift in the human immunodeficiency virus type 1. Nucleic Acids Res 30:5094–5102

Eigen M, Schuster P (1979) The hypercycle: A principle of natural self-organization. Springer-Verlag, Berlin

Fukuda Y, Washio T, Tomita M (1999) Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium, Mycoplasma pneumoniae*. Nucleic Acids Res 27:1847–1853

Hogeweg P, Hesper B (1992) Evolutionary dynamics and the coding structure of sequences: Multiple coding as a consequence of crossover and high mutation rates. Computers Chem 4:300–314

Holliday R (1968) Genetic recombination in fungi. In: Peacock WJ, Brock RD (eds) Replication and recombination of genetic material. Australian Academy of Science, Canberra, pp 157–174

Huynen MA, Konings DA, Hogeweg P (1993) Multiple coding and the evolutionary properties of RNA secondary structure. J Theor Biol 165:251–267

Kjems J, Brown M, Chang D, Sharp S (1991) Structural analysis of the interaction between the human immunodeficiency virus Rev protein and the Rev response element. Proc Natl Acad Sci USA 88:683–687

Konings DA (1992) Coexistence of multiple codes in messenger RNA molecules. Comput Chem 16:153–163

Konings DA, Hogeweg P, Hesper B (1987) Evolution of the primary and secondary structures of the E1a mRNAs of the adenovirus. Mol Biol Evol 4:300–314

Krakauer DC (2000) Stability and evolution of overlapping genes. Int J Org Evol 54:731–739

Kypr J (1986) A part of codon bias in genes protects protein spatial structures from destabilization by random single point mutations. Biochem Biophys Res 139:1094–1097

Lagunez-Otero J, Trifonov EN (1992) mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. J Biomol Struct Dyn 10:455–464

Le S-Y, Shapiro BA, Chen JH, Nussinov R, Maizel JV (1991) RNA pseudoknots downstream of the frameshift sites of retroviruses. Genet Anal Tech Appl 8:191–205

Malim M, Hauber J, Le S-Y, Maizel J, Cullen B (1989) The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. Nature 338:254–257

Malim M, Tiley L, McCarn D, Rusche J, Hauber J, Cullen B (1990) HIV-1 structural gene expression requires binding of the Rev trans-activator to its RNA target sequence. Cell 60:675–683

Morikawa S, Bishop DH (1992) Identification and analysis of the gag-pol ribosomal frameshift site of feline immunodeficiency virus. Virology 186:389–397

Normark S, Bergstrom S, Edlund T, Grundstrom T, Jurin B, Lindberg FP, Olsson O (1983) Overlapping genes. Annu Rev Genet 17:499–525

Parkin NT, Chamorro M, Varmus HE (1992) Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: Demonstration by expression in vivo. J Virol 66:5147–5115

Pavesi A, De Iaco B, Granero MI, Porati A (1997) On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. J Mol Evol 44:625–631

Peleg O, Brunak S, Trifonov EN, Nevo E, Bolshoy A (2002) RNA secondary structure and sequence conservation in C1 region of human immunodeficiency virus type 1 *env* gene. AIDS. Res Hum Retro 18:867–878

Peleg O, Trifonov EN, Bolshoy A (2003) Hidden messages in the *nef* gene of human immunodeficiency virus type 1 suggest a novel RNA secondary structure. Nucleic Acids Res 31:4192–4200

Schaap T (1971) Dual information in DNA and the evolution of the genetic code. J Theor Biol 32:293–298

Shintani S, O'hUigin C, Toyosawa S, Michalova V, Kelin J (1999) Origin of gene overlap: The case of TCP1 and ACAT2. Genetics 152:743–754

Staden R (1984) Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res 12(1,Pt 2):505–519

Staple DW, Butcher SE (2003) Solution structure of the HIV-1 frameshift inducing stem-loop RNA. Nucleic Acids Res 31:4326–4331

Trifonov EN (1981) Structure of DNA in chromatin. In: Schweiger H (ed) International cell biology 1980–1981. Springer-Verlag, Berlin, pp 128–138

Trifonov EN (1987) Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. J Mol Biol 194:643–652

Trifonov EN (1989) The multiple codes of nucleotide sequences. Bull Math Biol 51:417–432

Trifonov EN (1992) Recognition of correct reading frame by the ribosome. Biochimie 74:357–362

Trifonov EN (1996) Interfering contexts of regulatory sequence elements. Comput Appl Biosci 12:423–429

Vickers TA, Ecker DJ (1992) Enhancement of ribosomal frameshifting by oligonucleotides targeted to the HIV gag-pol region. Nucleic Acids Res 20:3945–3953

Wagner A, Stadler PF (1999) Viral RNA and evolved mutational robustness. J Exp Zool 285:119–127

Zhou C, Blumberg B (2003) Overlapping gene structure of human VLCAD and DLG4. Gene 305:161–166

Zuckerkandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. II. A selectionist model for random fixations in proteins. J Mol Evol 7:269–311