DNA SEQUENCE SPECIFICITY OF RecA-FILAMENT CORRELATES WITH GENETIC CODE

Mikhail PONOMARENKO, Igor TITOV, and Nikolay KOLCHANOV[#]

Institute of Cytology and Genetics,
Russian Academy of Science, Siberian Branch,
10 Lavrentyev Ave., Novosibirsk 630090 Russia

Telephones: +7(3832)355335, +7(3832)356406, +7(3832)351263
FAXs: +7(3832)355336, +7(3832)356558

E-mails: pon@bionet.nsc.ru, titov@bionet.nsc.ru, kol@bionet.nsc.ru

Alexander MAZIN, and Stephen KOWALCZYKOWSKI

Division of Biological Sciences, Sections of Microbiology and
of Molecular and Cellular Biology, University of California at Davis

Hutchison Hall, Davis, CA 95616-8665, USA

Telephone: +1(916)7525938
FAX: +1(916)7525939

E-mails: avmazin@ucdavis.edu, sckowalczykowski@ucdavis.edu

Keywords: RecA, filament, ssDNA[$] , affinity, genetic code, fuzzy logic, decision making

[#] Corresponding author.

[$] Abbreviations: **ssDNA**, single-stranded DNA; **nls**, nucleotides (unit of ssDNA length)

ABSTRACT

Experimental data on the affinity of the RecA-filament to ssDNA have been analyzed using the computer system SITEVIDEO. The affinity is maximal when the ssDNA sequence is devoid of the trinucleotides DRV={AAA, AAC, AGA, AGC, GAA, GAC, GGA, GGC, TAA, TGA, AAG, AGG, TAG, TGG, GAG, GGG, TAC, TGC} and decreases as the concentration of the trinucleotides increases in the neighborhood of the ssDNA 5'-end. This contextual feature is consistent with the experimental evidence that '5→3' direction is preferred by the main functions of RecA, namely RecA-filament formation and strand exchange. The trinucleotides and the genetic code have been shown to correlate. This finding fits in with the well-known fact that about 90% of the *E. coli* genome encodes proteins. The DRV-codons include all stop-codons, some codons for the residues forming protein surface, and none of the codons for the residues forming protein globular nuclei. The RecA-filament can therefore recognize the gene regions encoding protein functional sites and prevent damaging while recombination is under way. As is known, the sequences of protein functional sites are much more conserved than those of protein secondary structure, which provides further support to the last conclusion.

INTRODUCTION

The RecA protein plays a key role in both homologous recombination and DNA repair, and also because RecA-promoted homologous recombination is widely used in constructing artificial strains of *E. coli* with predefined properties (for instance, superproducers of given polypeptides) by genetic engineering techniques [1-4]. That is why an investigation of the role of DNA sequence-specificity in RecA-mediated homologous recombination is of biological, bioengineering and genetical interests.

A huge body of experimental evidence has now been gathered that many key events of the *E. coli* life cycle would not proceed normally unless RecA is functioning. Most importantly, this protein protects the cell against DNA damaging agents: *E. coli* cells that lack RecA are one million times more sensitive to UV-irradiation than wild-type cells. The high evolutionary conservation of the RecA protein family among bacteria, archaea and eukaryotes, including mammals, suggests that the role RecA-like protein plays in human cells should be pivotal as good [5].

RecA binds to single-stranded DNA to form a multimeric nucleoprotein filament of which unique helical structure is essential for all biologically important reactions mediated by RecA, such as DNA strand exchange and SOS induction [1-5]. Up to now it was common accepted, that vital importance of these functions of RecA-filament for the whole *E. coli* genome do not permit the RecA-filament to have preference to any DNA sequences. [1-5].

Thus the recent discovery that the RecA protein binds preferentially to certain characteristic DNA sequences *in vitro* [6, 7] appeared unexpected. This finding parallels with the well-known, although not so well understood observations that the frequency of homologous recombination varies significantly at different genetic loci, resulting in some extreme cases in the so-called "hot-spots" of recombination. Hence, understanding the relationship between DNA binding affinity and specificity for the RecA protein, and the efficiency of RecA-promoted homologous recombination is a key problem. Solving this would provide an insight into the mechanisms of homologous recombination and would also have an impact on applied research dealing with gene targeting.

To elucidate the role of DNA sequence context in homologous recombination promoted by RecA protein, the experimental data on the affinity of the RecA-filament to ssDNA [6] were analyzed by the computer system SITEVIDEO [8] The relevant version of this system is presented in this paper. It was determined, that the affinity of the RecA-filament to ssDNA depends on the ssDNA sequence. This affinity is maximal if the ssDNA has not trinucleotides DRV={AAA, AAC, AGA, AGC, GAA, GAC, GGA, GGC, TAA, TGA, AAG, AGG, TAG, TGG, GAG, GGG, TAC, TGC} and decreases with the increase of their concentration in the neighborhood of the 5'end of this ssDNA. It fits the experimental data on preference of 5'→3' direction as for RecA-filament formation [9], as for strange exchange [10-12]. It was also shown, that these DRV trinucleotides are significant for genetic code in *E. coli* [13]. It corresponds to the well-known fact that about 90% of the *E. coli* genome consist of protein genes. The trinucleotides DRV were found to be stop-codons and codons for residues for protein surface and out for globular nucleus. Hence, RecA-filament can recognize the gene regions encoding protein functional sites to protect them in recombination. It matchs by current knowledge on higher conservatively of protein sequences for functional sites in comparison with those for secondary structure.

MATERIALS

The experimental data on affinity of the RecA-filament to ssDNA [6] presented in the Table 1 were analyzed. These data were derived by the following way [6].

The $ssDNA_0$ was synthesized by sequence S_0 (Table 1).

The RecA protein was added with ratio 1 monomer per 3 nucleotides of $ssDNA_0$. This ratio is a character of RecA-filament [6]. As result of binding between $ssDNA_0$ и RecA the RecA-filament was formed.

Another $ssDNA_n$ with sequence S_n (Table 1, n-th line) was synthesized using ^{32}P-labeled nucleotides.

It was added to RecA-filament in the ratio 1:1. As the result of binding the ^{32}P-labeled $ssDNA_n$ and the RecA-filament, their ^{32}P-labeled complex was formed.

The ^{32}P-labeled complex was separated from ^{32}P-labeled $ssDNA_n$ by gel-electrophoresis.

The ^{32}P-labeled complex concentration, $Comp_n$, was measured by a phosphoroimager.

These measurements, $Comp_n$, were done for each of sequences from Table 1 ($0 \leq n \leq 15$).

It was noted [6], the concentrations differed from each other in the range of two orders.

By this way, the fact that the affinity of the RecA-filament to ssDNA depends on the ssDNA sequence was discovered [6].

Thus, the question raised what features of an arbitrary ssDNA sequence are responsible for the affinity of the RecA-filament to the ssDNA.

To diminish the heterogeneity of experimental data [6] the above concentrations $Conp_n$ were normalized by the concentration $Comp_0$ and expressed in logarithmic scale:

Table 1. Affinity of the RecA-filament to ssDNA [6]

Data set	n	Mark	ssDNA sequence, S_n	Affinity, F_n
Training	**0**	**IDENT**	**CCATCCGCAAAAATGACCTCTTATCAAAAGGA**	**0.00**
	1	dC	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	0.54
	2	#40	ACCACCACACACGCGCACACCACCACACACGC	0.48
	3	htr#3	TTCACAAACGAATGGATCCTCATTAAAGCCAG	0.34
	4	#39	GCGTGTGTGGTGGTGTGCGCGTGTGTGGTGGT	0.33
	5	dT	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	0.09
	6	htr#4	CATGGAGCAGGTCGCGGATTTCGACACAATTT	-0.02
	7	#7	GGCGGGCGGCGCGGCCGGGCGGCGGGCGCGCG	-1.99
	8	htr#2	AATTCTTCGAAGCTAGCCCTCAGGCCTAGGCA	-2.42
	9	dA	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	-5.01
Control	10	A>T	CCTTCCGCTTTTTTGTCCTCTTTTCTTTTGGT	1.20
	11	G>C	CCATCCCCAAAAATCACCTCTTATCAAAACCA	0.03
	12	G>T	CCATCCTCAAAAATTACCTCTTATCAAAATTA	-0.40
	13	C>G	GGATGGGGAAAAATGAGGTGTTATGAAAAGGA	-1.00
	14	C>T	TTATTTGTAAAAATGATTTTTTATTAAAAGGA	-1.20
	15	C>A	AAATAAGAAAAAATGAAATATTATAAAAAGGA	-3.40

$$\mathbf{F_n = - \ln \{ Comp_n / Comp_0 \}.} \qquad (1)$$

The value F_n characterizes the affinity of the RecA-filament to the $ssDNA_n$. (Table 1).

Data were divided into training and control sets (Table 1). The training set contains pairs (S_n,F_n) with n between 0 and 9. This set was used to reveal the ssDNA contextual features responsible for the affinity of the RecA-filament to ssDNA. Control set consisted of the all other pairs (S_n,F_n) with n between 10 and 15 (Table 1) and was used for independent testing of the result obtained.

METHOD

The above experimental data (Table 1) were investigated using SITEVIDEO computer system [8], modified as following.

In the sequence $S=s_1...s_i...s_L$ at length L nls with the known affinity F_S the short subsequences $Z=z_1...z_j...z_k$ at length k between 1 and 4 nls were taken (k<<L). The nucleotides codes z_j were from widely used set of all 15 possible codes of nucleotide combinations {A, T, G, C, W=A/T, R=A/G, M=A/C, K=T/G, Y=T/C, S=G/C, B=T/G/C, V=A/G/C, H=A/T/C, D=A/T/G, N=A/T/G/C}. The concentration of the subsequences Z weighted by the weight function w(i) were calculated for the sequence S at length L, such as:

$$\mathbf{X_{Zw}(S) = \Sigma_{1\le i\le L-k+1} I_{ZS}(i)\times w(i);} \qquad \mathbf{(2)}$$

where $I_{ZS}(i)=1$ if $\{s_{i+j-1}=z_j\}$and $I_{ZS}(i)=0$ if otherwise; w(i) is weight of i-th position.

In Formula (2), the indicator $I_{ZS}(i)$ equals 1 in the case of the subsequence Z starts at i-th position of sequence S. The weight function w(i) simulates the influence of the subsequence Z at i-th position of a sequence S to the affinity value F_S. According to Zadeh's fuzzy logic [14], the simple rule «the greater weight, the greater influence on the affinity» was used.

Fig 1 exemplifies one of such weight function w(i) simulating the greatest influence on affinity for the subsequences Z in the neighborhood of the 5'end of any sequence S of length 32 nls. Totally 180 these weight functions w(i) simulating different DNA regions having greatest influence on protein-binding affinity were used.

Concentrations $X_{Zw}(S_n)$ at the fixed Z and w for the whole number of sequences S_n with known affinity F_n from the set $\{(S_1,F_1),...(S_n,F_n)...(S_N,F_N)\}$ were calculated by Formula (2). It resulted in the set of pairs «concentration→affinity», $\{X_{Zw}(S_n)\to F_n\}$.

The simple regression was optimized for the set $\{X_{Zw}(S_n)\to F_n\}$ by standard method [15]:

$$\mathbf{F_{Zw}(S_n)=a_0 + a_1\times X_{Zw}(S_n);} \qquad \mathbf{(3)}$$

where: a_0 and a_1 are the standard coefficients of simple linear regression [15].

The set of pairs $[F_{Zw}(S_n); F_n]$ of predicted and experimental affinity were made up by Formula (3) at the fixed combination (Z, w). It allows to estimate the predictability of the subsequence Z and the weight function w(i) fixed.

The predicting capability of the combination (Z,w) was evaluated using the theory of decision making [16]. That is why a great number of different types of matches between

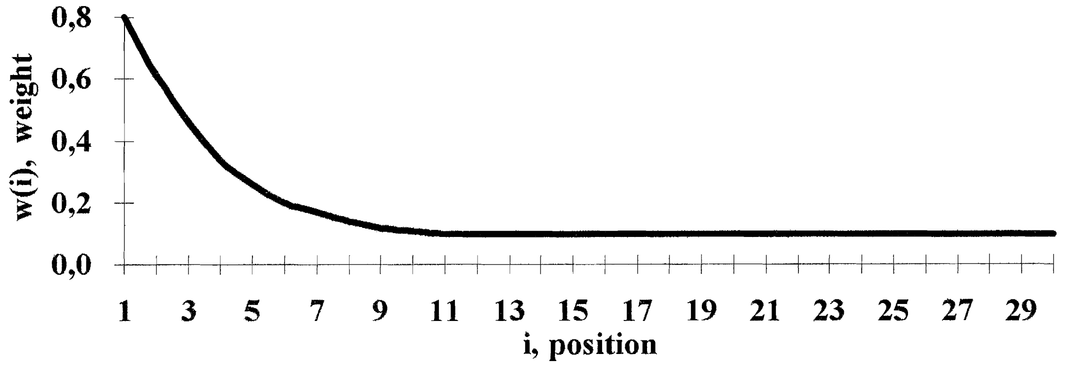

Figure 1. The example of a weight function w(i) which simulates a high influence on DNA/protein-affinity for trinucleotides Z near the 5'end of sequence S at length 32 nls.

predicted and experimental affinity, $F_{Zw}(S_n)$ and F_n, were tested, namely linear, rank and sign correlation's [17], means and variances [18], the Gaussian distribution for the deviations $\{\Delta_n=F_n-F_{Zw}(S_n)\}$ and also the uniform distributions for $\{F_{Zw}(S_n)\}$ and $\{F_n\}$ values [19]. The total number of statistical tests used is equal eleven.

To diminish the heterogeneity of experimental data analyzed, each criterion was tested on a large number of different data sets, such as

1) the complete set $\{[F_{Zw}(S_1); F_1], \ldots[F_{Zw}(S_n); F_n], \ldots[F_{Zw}(S_N); F_N]\}$;
2) half the set with low F_n;
3) half the set in the vicinity of mean F_n;
4) half the set with high F_n;
5) half the set with low $F_{Zw}(S_n)$;
6) half the set in the vicinity of mean $F_{Zw}(S_n)$;
7) half the set with high $F_{Zw}(S_n)$.

Totally, 11×7=77 different matches between predicted and experimental affinities, $F_{Zw}(S_n)$ and F_n, were tested in significance. Each significant match was assigned by positive mark between 0 and 1, and also each dissignificant match was assigned by negative mark between -1 and 0. According to [16], the average value of these marks, U(Z,w), was used as an estimation of predicting capability of the subsequence Z and weight function w(i) fixed.

This value U(Z,w) is so called «Utility» [16]. The utility has two important properties:

U(Z,w)<0 means "the prediction $F_{Zw}(S_n)$ is groundless for F_n"; (4)

U(Z,w)>U(Q,v) means "the prediction $F_{Zw}(S_n)$ is better than $F_{Qv}(S_n)$". (5)

Properties (4, 5) mean «**the highest U(Z,w) pinpoints the best prediction $F_{Zw}(S_n)$**».

As soon as neither subsequences Z nor weight functions w(i) significant for affinity $F_{Zw}(S)$ prediction by an arbitrary sequences S were known, all possible concentrations X_{Zw} were tested. The used code of 15 possible combinations of nucleotides allowed to compose the total number of all possible mono-, di-, tri-, and tetranucleotides which was equal to 14+14×14 + 14×15×14 + 14×15×15×14 = 14 + 196 + 2940 + 44100 = 47250. By combining of each of these 47250 short subsequences with each of 180 weight functions used gave the total number 47250×180=7938000 of different concentration's X_{Zw}.

For each of these 7938000 concentrations X_{Zw} the utility U(Z,w) was calculated. On the basis of Property (4) all concentrations with negative utility were discarded. By using the Property (5) the concentration X_{Zw} with highest utility U(Z,w) observed was selected.

That is the result of SITEVIDEO for the entered set of sequences with known affinity.

RESULTS AND DISCUSSION

Using method described above the training set of sequences S_n with known affinity F_n (Table 1) was analyzed. All possible 7938000 concentrations $X_{Z,w}$ calculated by formula 2 were tested. The concentration $X_{DRV,w}$ was found to have the highest utility U(DRV,w)=0.27 observed. This concentration $X_{DRV,w}$ was specified by the trinucleotide DRV={AAA, AGA, TAA, TGA, GAA, GGA, AAG, AGG, TAG, TGG, GAG, GGG, AAC, AGC, TAC, TGC, GAC, GGC} and the weight function w(i) with maximum at the ssDNA 5'end (Fig. 1).

On the training set (Table 1) the simple regression for predicting the affinity was made:

$$\mathbf{F_{DRV,w}(S_n) = 0.54 - 1.03 \times X_{DRV,w}(S_n).} \qquad \mathbf{(6)}$$

Fig.2a presents the correlation between affinity experimental, F_n, and predicted by Formula (6), $F_{DRV,w}(S_n)$, for the training set from Table 1. The linear correlation coefficient between these values reached r=0.888 (α<0.01 [19]).

The formula 6 was tested in significance for control set from Tabl.1. The test results are shown in Fig.2b. In this case the linear regression coefficient between experimental and predicted affinity equaled r=0.812 (α<0.05). It means that the $F_{DRV,w}(S)$ value calculated by Formula (6) is a reliable prediction of affinity F_S between RecA-filament and ssDNA by the ssDNA sequence S.

The interpretation of the revealed ssDNA contextual feature, namely weight concentration $X_{DRV,w}$, resulted in the following. Firstly, from all possible mono-, di-, tri- and tetranucletides the most significant appeared to be indeed a trinucletide, namely the trinucleotide DRV. It is in agreement with the character ratio of RecA-filament, 1 monomer RecA per 3 nucleotides of ssDNA, which has been determined experimentally [6].

As shown in Fig.1, the damaging effect of the trinucleotides DRV on the affinity of the RecA-filament to ssDNA decreased in the 5'→3' direction. It fits the experimental data on preference of this 5'→3' direction as for the RecA-filament formation [9] as for strange exchange promoted by RecA-filament [10-12].

Table 2 presents the comparison between trinucleotide DRV=={AAA, AGA, TAA, TGA,

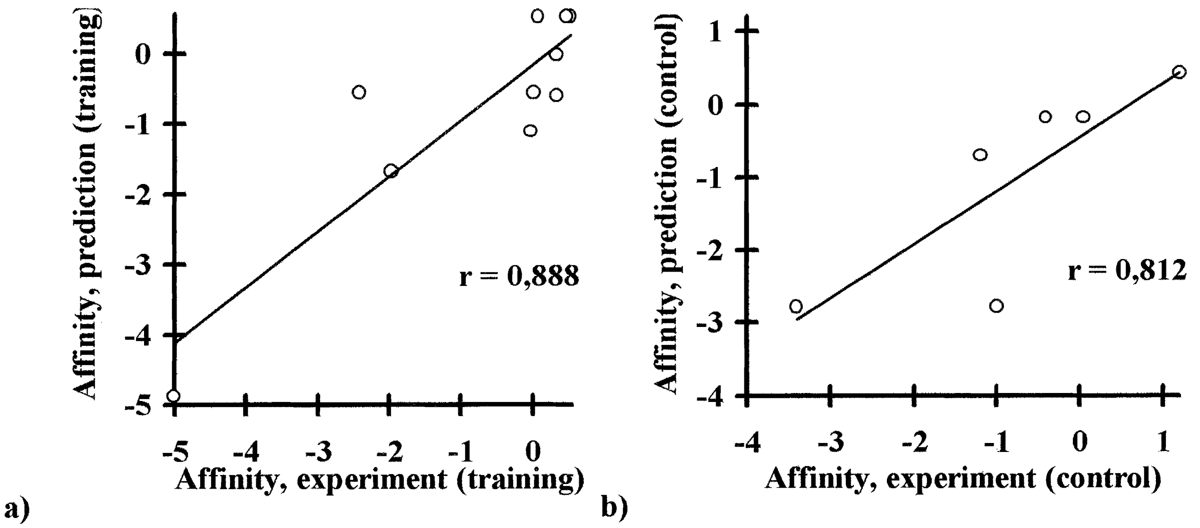

Figure 2. Correlation between predicted and experimental affinity on the Table 1 training (a) and control (b) sets. Solid line is the optimal linear regression.

Table 2. Comparison between the trinucleotide DRV and the *E. coli* genetic code [13]

Residues		Genetic code **(DRV is bold-face)**	Residues		Genetic code
R	Arg	**AGG AGA** CGG CGA CGT CGC	A	Ala	GCG GCA GCT GCC
N	Asn	AAT **AAC**	Q	Gln	CAG CAA
D	Asp	GAT **GAC**	H	His	CAT CAC
C	Cys	TGT **TGC**	I	Ile	ATA ATT ATC
E	Glu	**GAG GAA**	L	Leu	TTG TTA CTG CTA CTT CTC
G	Gly	**GGG GGA** GGT **GGC**	M	Met	ATG
K	Lys	**AAG AAA**	F	Phe	TTT TTC
S	Ser	AGT **AGC** TCG TCA TCT TCC	P	Pro	CCG CCA CCT CCC
W	Trp	**TGG**	T	Thr	ACG ACA ACT ACC
Y	Tyr	TAT **TAC**	V	Val	GTG GTA GTT GTC
Stop		**TGA TAG TAA**			

GAA, GGA, AAG, AGG, TAG, TGG, GAG, GGG, AAC, AGC, TAC, TGC, GAC, GGC} and the *E. coli* genetic code [13]. It is shown, that the trinucleotides DRV contained all three Stop-codons; three codons for Glycine (G); and two codons for each of the following residues Arginine (R), Glutamic acid (E), Lysine (K); and one codon for each of the following Asparagine (N), Aspartic acid (D), Cystein (C), Serine (S), Tryptophan (W), Tyrosine (Y). Thus, the trinucleotides DRV are the Stop-codons and some codons of residues.

To test the significance of the trinucleotides DRV for the *E.coli* genetic code [13] a large number of different physico-chemical and statistical properties were taken into consideration. It

Table 3. Physico-chemical, statistical and genetical properties significant for the DRV-codons

Property	Reference	Residues		codon numbers$^{\&}$				Significance$^{\$}$
		All possible	**of DRV-codon**	N_{++}	N_{-+}	N_{+-}	N_{--}	
Stop-codon$^{\#}$	[13]	**TGA, TAG, TAA**	**TGA, TAG, TAA**	3	15	0	46	0.025
Coil	[20]	A,V,**Y,D,N,E,K,G**	**Y,D,N,E,K,G**	10	8	12	34	0.05
Surface	[21]	H,Q,**D,E,K,N,R**	**D,E,K,N,R**	8	10	8	38	0.05
Charge	[20]	H,**D,E,K,R**	**D,E,K,R**	7	11	7	39	0.05
H-/SS-bonds	[20]	H,T,**C,Y,D,E,S,K,R**	**C,Y,D,E,S,K,R**	$7^{@}$	$3^{@}$	$2^{@}$	$8^{@}$	0.05
Nucleus	[21]	L,I,M,F,V	**none**	0	18	16	30	0.0025
Izostructural	[20]	V,L,I,M	**none**	0	18	14	32	0.01
Aliphatic	[20]	A,V,L,I,**C**	**C**	1	15	18	28	0.01
β-structure	[20]	A,V,L,I,F,**C,S,G**	**C,S,G**	5	13	26	20	0.05
Degeneration	[13]	A,V,L,P,T,**S,R,G**	**S,R,G**	6	12	32	14	0.01

Notes: &) N_{++} - the number of the DRV-codons with the Property; N_{-+} - the number of the DRV-codons without the Property; N_{+-} - the number of the other codons with the Property; N_{--} - the number of the DRV-codons without the Property; #) stop-codons; @) the number of residues; $) significance level α determined by precise Fisher's criterion [18].

was done as the following. One of the properties was fixed. The numbers of the DRV-codons with ($\mathbf{N}_{++}$) and without ($\mathbf{N}_{-+}$) this property were accounted. The same amounts, $\mathbf{N}_{+-}$ and $\mathbf{N}_{--}$, were calculated for the other codons. The values $\mathbf{N}_{++}$, $\mathbf{N}_{-+}$, $\mathbf{N}_{+-}$ и $\mathbf{N}_{--}$, were derived which characterized the coding capacity of this property fixed by the DRV-codons. These values were tested by precise Fisher's criterion [18]. The results are in the Table 3.

The first line in Table 3 demonstrates the analysis of the stop-codons. The *E. coli* genetic code includes three stop-codons, namely TGA, TAG, TAA. All of them are the DRV-codons. That is why, $\mathbf{N}_{++}=3$, $\mathbf{N}_{+-}=0$, and also $\mathbf{N}_{-+}=18-3=15$ and $\mathbf{N}_{--}=64-18=46$ (Table 3). For these values $\mathbf{N}_{++}$, $\mathbf{N}_{-+}$, $\mathbf{N}_{+-}$ и $\mathbf{N}_{-}$, precise Fisher's criterion [18] gives, that the DRV-codons is significant for **encoding** the stop-codons ($\alpha<0.05$).

The same analysis gives, that the DRV-codons are also significant for **encoding** the following properties: the most frequent residues for coil (Property «Coil») and surface (Property «Surface») of protein globula charged residues (Property «Charge») and also residues which side chains can form H- and SS-bonds (Property «H-/SS-bonds»).

The one more correlation was revealed (Table 3). For example, residues {L, I, M, F, V} are the most frequent for nucleus (Property «Nucleus») of protein globula. In the *E. coli* genetic code, these residues are encoded by 16 codons none of them are the DRV-codon. That is why, $\mathbf{N}_{++}=0$, $\mathbf{N}_{-+}=18$, $\mathbf{N}_{+-}=16$ and $\mathbf{N}_{--}=64-18-16=30$ (Table 3). According to precise Fisher's criterion [18] the DRV-codons are significant for **non-encoding** this sort of residues ($\alpha<0.0025$).

It was also shown by this way, that the DRV-codons are significant for non-coding the most frequent residues for the protein β-structures (Property «β-structure»), aliphatic and izosructural residues, and also residues with their code degenerated (Property «Degeneration»).

Taking all together the results of the DRV-codons analysis (Table 3), we can conclude, that the DRV codons encode stop-codons and residues for protein surface formation and do not encode the residues for protein globular nucleus. The sensibility of RecA-filament to genetic code corresponds the well known fact, that the 90% of the *E. coli* genome encodes proteins.

According to the functional sites are usually on the protein surface, the RecA-filamets sensibility to the DRV codons can be explained as its sensibility to the gene regions encoding protein functional sites. Thus RecA-filament can recognize this sort of gene regions to protect them from damaging in the course of recombination. It fits the well-known property of functional site sequences to be more conservative than the protein secondary structure.

As soon as the DNA sequence specificity of the RecA-filament has been discovered [6, 7], the genetic recombination becomes a new perspective field for applying any computer methods for the DNA sequence analysis. The present investigation exemplifies the efficiency of the such computer analyses of experimental data on genetic recombination.

The genetic recombination is widely used in biotechnology and gene-engineering. That is why the computer analysis of experimental data on genetic recombination can help in developing of new approaches in these two modern fields of molecular biology and genetics.

ACKNOWLEDGMENTS

We are thankful to N. Omelyanchuk for translating this manuscript into English.

The research described in this publication was made possible in part by the travel grant PDP-#24 of the U.S. Civilian R&D Foundation for the Independent States of the Former Soviet Union, and also by a grant from the Russian Foundation for Basic Research.

REFERENCES

[1] Cox, M.M. (1993) *BioEssay*, **15**, 617.

[2] Kowalczykowski, S.C., et al. (1994) *Microbiol. Rev.*, **58**, 401.

[3] Stasiak, A., and Egelman, E.H. (1994) *Experientia (Basel)*, **50**, 192.

[4] West, S.C. (1994) *Cell*, **76**, 9.

[5] Ogawa, T., et al. (1993) *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 567.

[6] Mazin, A., and Kowalczykowski, S.C. (1996) *Proc. Natl. Acad. Sci. U.S.A.*, **93** 10673.

[7] Tracy, R.B., and Kowalczykowski, S.C. (1996) *Genes. Dev.*, **10**, 1890.

[8] Kel, A.E., et al. (1993) *Comput. Applic. Biosci.*, **9**, 617.

[9] Lindslay, J.E., and Cox, M.M. (1989) *J. Mol. Biol.*, **205**, 695.

[10] Cox, M.M., and Lehman, I.R. (1981) *Proc. Natl. Acad. Sci. U.S.A.*, **78**, 6018.

[11] Konforti, B.B., and Davis, R.W. (1992) *J. Mol. Biol.*, **227**, 38.

[12] Jwang, B., and Radding, C.M., (1992) *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 7596.

[13] Ayala, F., and Kiger, Jr, J. (1984) *Modern genetics*, The Benjamin/Com.Publ.Comp.

[14] Zadeh, L.A., (1965) *Information and Control*, **8**, 338.

[15] Forster, E., and Ronr, B. (1979) *Methoden der korrelations- und regressions analyse.* Verlag Die Wirtschaft, Berlin.

[16] Fishburn, P.C. (1970) *Utility theory for decision making*, Wiley, New York

[17] Hajek, J., and Sidak, Z. (1967) *Theory of rank tests.* Academia, Prague.

[18] Lehman, E.L. (1959) *Testing statistical hypotheses.* Willey. New York.

[19] Likes, J., and Laga, J. (1978) *Zakladni statisticke tabulky.* SNTL. Prague.

[20] Cohen, B.I., et al. (1991) *Methods in enzimology* (Langone, J.J., ed), **202**, 252.

[21] Karlin, S., et al. (1989) *Mathematical methods for DNA sequences*, Boca Raton, CRC Press, 133.