Doklady Biochemistry, Vol. 363, 1998, pp. 179–182. Translated from Doklady Akademii Nauk, Vol. 363, No. 1, 1998, pp. 122–125.
Original Russian Text Copyright © 1998 by M. Ponomarenko, Yu. Ponomarenko, Titov, Kolchanov, Mazin, Kowalczykowski.

BIOCHEMISTRY, BIOPHYSICS, AND MOLECULAR BIOLOGY

The Affinity of the RecA Filament to DNA Sequences Correlates with the Genetic Code

M. P. Ponomarenko, Yu. V. Ponomarenko, I. I. Titov, N. A. Kolchanov, A. V. Mazin, and S. C. Kowalczykowski

Presented by Academician V.K. Shumnyi December 15, 1997

Received December 30, 1997

RecA is the key protein of homologous recombination in *Escherichia coli* [1]. RecA binding to single-stranded DNA results in the formation of an RecA filament, in which one helix coil is composed of six RecA monomers and 18 nucleotides. This filament performs strand exchange during DNA homologous recombination [2]. To determine the preferences of the RecA filament for the DNA context [3], DNA properties that determine this affinity should be revealed.

In this work, we analyzed data on the RecA filament affinity for DNA [3] using the ACTIVITY computer system [4]. It was found that this affinity decreased with an increase in the concentration of trinucleotides DRV = {AAA, AAC, AGA, AGC, GAA, GAC, GGA, GGC, TAA, TGA, AAG, AGG, TAG, TGG, GAG, GGG, TAC, TGC}. These DRV nucleotides code for amino acids that usually form protein-globule surface functional sites but not a protein globular core, which correlates well with the known data on the conservatism of protein functional sites and block rearrangements of protein cores.

Data on the RecA filament affinity for DNA [3] are shown in Table 1. The data for 16 DNA fragments (each of 32 nucleotides) $S = s_1 \ldots s_{32}$ are presented. The RecA filament affinity for these fragments $F(S)$ varies from –0.51 to 1.20 logarithmic units. Since in the RecA filament, one RecA monomer interacts with three nucleotides [3], we studied the weighed concentrations of trinucleotides ($Z = z_1z_2z_3$):

$$X_{Zw}(S) = \sum_{i=1}^{L-2} w(i) \times \Delta(s_i \in z_1) \times \Delta(s_{i+1} \in z_2) \times \Delta(s_{i+2} \in z_3), \qquad (1)$$

where $z \in$ {A, T, G, C, W = A/T, R = A/G, M = A/C, K = T/G, Y = T/G, S = G/C, B = T/G/C, V = A/G/C, H = A/T/C, D = A/T/G, N = A/T/G/C}; Δ(truth) = 1; Δ(lie) = 0; $w(i)$ is the weight function that models the contribution of the trinucleotide Z at position i of the sequence S to the RecA filament affinity: "higher $w(i)$" ⇔ "greater contribution."

Figure 1 shows the $w(i)$ functions that model the maximum contribution to the RecA filament affinity to DNA for trinucleotides at the 5'-end of the DNA (solid line) and at position 19 (dashed line) located at a distance of 18 nucleotides from the DNA 5'-end, which comprises one coil of the RecA filament helix [3]. A total of 180 $w(i)$ functions were analyzed. Combining these functions with all trinucleotides $Z = z_1z_2z_3$ yielded $15^3 \times 180 \approx 10^6$ X_{Zw} concentrations. The $X_{Zw}(S_n)$ concentration determines the affinity F_n of the RecA filament to DNA (S_n), if all $\{X_{Zw}(S_n), F_n\}$ pairs fit the regression:

$$F_{Zw}(S_n) = a + b \times X_{Zw}(S_n), \qquad (2)$$

where a and b are regression coefficients [5] for $\{X_{Zw}(S_n), F_n\}$.

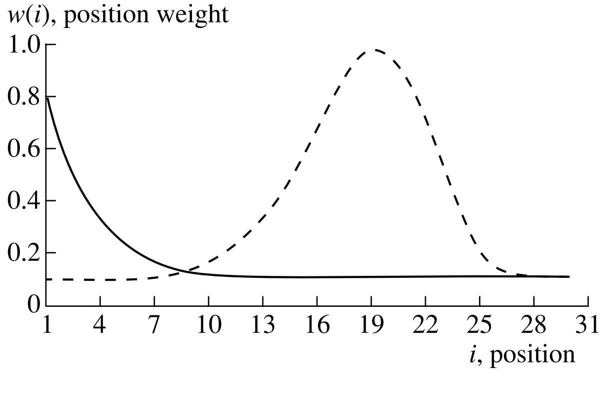

Fig. 1. Examples of $w(i)$ weight functions modeling the maximum contribution to the RecA filament affinity for 32-nucleotide DNAs for trinucleotides at the 5'-end (solid line) and at position 19 (dashed line) located at a distance of 18 nucleotides from the 5'-end, which comprises one coil of the RecA filament helix [3].

Institute of Cytology and Genetics, Siberian Division, Russian Academy of Sciences, pr. Akademika Lavrent'eva 10, Novosibirsk, 630090 Russia
California University at Davis, Hutchison Hall, Davis, CA 95616-8665, United States

0012-4958/98/1112-0179$20.00 © 1998 МАИК Наука / Interperiodica Publishing

Table 1. RecA filament affinity for single-stranded DNAs [3]

No.	Variant	DNA sequence, S_n	Affinity, F_n
1	A > T	CCTTCCGCTTTTTTGTCCTCTTTTCTTTTGGT	1.20
2	dC	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	0.54
3	#40	ACCACCACACACGCGCACACCACCACACACGC	0.48
4	htr#3	TTCACAAACGAATGGATCCTCATTAAAGCCAG	0.34
5	#39	GCGTGTGTGGTGGTGTGCGCGTGTGTGGTGGT	0.33
6	dT	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	0.09
7	G > C	CCATCCCCAAAAATCACCTCTTATCAAAACCA	0.03
8	IDENT	CCATCCGCAAAAATGACCTCTTATCAAAAGGA	0.00
9	htr#4	CATGGAGCAGGTCGCGGATTTCGACACAATTT	–0.02
10	G > T	CCATCCTCAAAAATTACCTCTTATCAAAATTA	–0.40
11	C > G	GGATGGGGAAAAATGAGGTGTTATGAAAAGGA	–1.00
12	C > T	TTATTTGTAAAAATGATTTTTTATTAAAAGGA	–1.20
13	#7	GGCGGGCGGCGCGGCCGGGCGGCGGGCGCGCG	–1.99
14	htr#2	AATTCTTCGAAGCTAGCCCTCAGGCCTAGGCA	–2.42
15	C > A	AAATAAGAAAAAATGAAATATTATAAAAAGGA	–3.40
16	dA	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	–5.01

The formula (2) predicts the affinity F_{Zw} of the RecA filament for DNA from the concentration X_{Zw}. To use this formula, the predicted affinity F_{Zw} should correlate with the known F value. A total of 22 correlations were verified, including the linear correlation between F_{Zw} and F, and the normality of deviations $\{F_{Zw} - F\}$. For the mth correlation [6], the level of its significance α_m ($1 \le m \le 22$) was estimated. Significant correlations ($\alpha_m < 0.05$) between F_{Zw} and F give concentration X_{Zw} (in terms of Zadeh's logic [7]—positive estimation $U_m(X_{Zw}, F) > 0$); insignificant correlations give a negative estimation. A total of 22 estimations $\{U_m(X_{Zw}, F)\}$ were obtained, whose averaging in terms of decision making theory [8] gave the value of the concentration X_{Zw} for predicting the RecA filament affinity for DNA (F):

$$U(X_{Zw}, F) = \sum_{m=1}^{22} U_m(X_{Zw}, F)/22. \qquad (3)$$

According to formula (3), the value $U(X_{Zw}, F)$ of the concentration X_{Zw} increases with the increase in the number of significant correlations between the predicted F_{Zw} and the known F values. The negative value $U(X_{Zw}, F) < 0$ means that the RecA filament affinity for DNA cannot be predicted based on X_{Zw}, since most correlations between the F_{Zw} and F are not significant. If the value is positive ($U(X_{Zw}, F) > 0$), the RecA filament affinity for DNA can be predicted based on X_{Zw}, since most correlations between the F_{Zw} and F are significant. According to binominal criterion [6], the probability of an occasional appearance of one X concentration with $U(X, F) > 0$, among 10^6 X_{Zw} concentrations can be determined by the following formula:

$$p[U(X, F) > 0]$$
$$= 10^6 \times \sum_{m=12}^{22} C_{22}^m \times 0.05^m \times 0.95^{22-m} < 10^{-4}.$$

The maximum $U(X_{Zw}, F) > 0$ indicates the concentration X_{Zw} that determines the RecA filament affinity for DNA.

Ten DNA sequences from Table 1 were analyzed (nos. 2, 3, 4, 5, 6, 8, 9, 13, 14, and 16); the other six sequences were used as a control. For these ten DNAs (S_n), all 10^6 possible concentrations $X_{Zw}(S_n)$ were calculated by formula (1). Based on the known RecA filament affinity for DNA (F_n), regressions (2) were plotted for predicting $F_{Zw}(S_n)$ based on $X_{Zw}(S_n)$. A comparison of $F_{Zw}(S_n)$ to F_n by formula (3) yielded the value $U(X_{Zw}, F)$ for the concentration X_{Zw}. A total of 10^6 value $U(X_{Zw}, F)$ values were obtained; only five of them were positive (Table 2).

The maximum value ($X_{\mathrm{DRV}w1}$, F) = 0.270 was obtained for the trinucleotides DRV = {AAA, AGA, TAA, TGA, GAA, GGA, AAG, AGG, TAG, TGG, GAG, GGG, AAC, AGC, TAC, TGC, GAC, GGC} weighed by the function $w1(i)$ with the maximum at the DNA 5'-end (Fig.1, solid line). This indicates that the RecA filament affinity for DNA is determined by the concentration of DRV trinucleotides near its 5'-end. Table 2 shows that two other trinucleotides with positive values are cyclic rearrangements (RVD and VDR) of the best DRV trinucleotide weighed by the same

Table 2. Revealed trinucleotide concentrations X_{Zw} with the positive value $U(X_{Zw}, F)$ for predicting the RecA filament affinity for single-stranded DNA

No.	Revealed trinucleotide concentrations X_{Zw}			Relationship with the best trinucleotide
	trinucleotide Z	weight function $w(i)$ (see Fig. 1)	value $U(X_{Zw}, F)$ ± s.d.	
1	DRV	Solid line	**0.270 ± 0.011**	**The best**
2	RVD	″	0.229 ± 0.021	Cyclic rearrangement
3	VDR	″	0.170 ± 0.033	
4	RRV	Dashed line	0.191 ± 0.033	Particular case
5	RRM	″	0.180 ± 0.012	

Table 3. Projection of the DRV trinucleotides on the genetic code

Amino acid reside		Genetic code (**DRV**)	Number	
Amino acid	Designation		of codons	of DRV trinucleotides
Alanine	A	GCG GCA GCT GCC	4	
Arginine	R	**AGG AGA** CGG CGA CGT CGC	6	**2**
Asparagine	N	AAT **AAC**	2	**1**
Aspartic acid	D	GAT **GAC**	2	**1**
Cysteine	C	TGT **TGC**	2	**1**
Glutamine	Q	CAG CAA	2	
Glutamic acid	E	**GAG GAA**	2	**2**
Glycine	G	**GGG GGA** GGT **GGC**	4	**3**
Histidine	H	CAT CAC	2	
Isoleucine	I	ATA ATT ATC	3	
Leucine	L	TTG TTA CTG CTA CTT CTC	6	
Lysine	K	**AAG AAA**	2	**2**
Methionine	M	ATG	1	
Phenylalanine	F	TTT TTC	2	
Proline	P	CCG CCA CCT CCC	4	
Serine	S	AGT **AGC** TCG TCA TCT TCC	6	**1**
Threonine	T	ACG ACA ACT ACC	4	
Valine	V	GTG GTA GTT GTG	4	
Tryptophan	W	**TGG**	1	**1**
Tyrosine	Y	TAT **TAC**	2	**1**

$w1(i)$ function. The two remaining trinucleotides (RRV and RRM) are particular cases of DRV weighed by the $w2(i)$ function with the maximum at position 19 (Fig. 1, dashed line) that is located at a distance of 18 nucleotides from the 5'-end, which comprises one coil of the RecA filament [3]. This indicates the importance of DRV nucleotides for the RecA filament affinity for DNA. This affinity was predicted by the simple regression:

$$F_{\mathrm{DRV}w1}(S) = 0.54 - 1.03 \times X_{\mathrm{DRV}w1}(S). \qquad (4)$$

The RecA filament affinity for DNA predicted by the formula (4) is shown in Fig. 2 (open circles, ten analyzed DNAs; closed circles, six control DNAs). For the control DNAs, the coefficient of linear regression (r) between the predicted $F_{\mathrm{DRV}w1}$ and known F values was 0.812 ($\alpha < 0.05$). Therefore, formula (4) significantly predicts the RecA filament affinity for DNA, as demonstrated for independent control data, which unambiguously indicates the importance of DRV trinucleotides for this affinity.

Since the major part of the *E. coli* genome codes for proteins, it was interesting to compare the DRV trinucleotides that determine the RecA filament affinity for DNA to the genetic code (Table 3). DRV trinucleotides were found to coincide with arginine, asparagine, lysine, glycine, cysteine, serine, tyrosine, aspartic acis,

Table 4. Correlation between the amino acid properties [9, 10] and DRV trinucleotides

Amino acid properties	Amino acids		Number of codons				Significance*
	DRV trinucleotide		DRV		other		
	yes	no	yes	no	yes	no	
DRV trinucleotides code for							
random coil [9]	YDNEKG	AV	10	5	12	34	0.010
surface amino acids [10]	DEKNR	HQ	8	7	10	36	0.025
charged amino acids [10]	DEKR	H	7	8	7	39	0.025
DRV trinucleotides do not code for							
protein core amino acids [9]	–	LIMFV	0	15	16	30	0.005

* Significance α was estimated by Fischer’s criterion [6].

glutamic acid, and tryptophan codons. Four significant correlations between amino acid properties [9, 10] and DRV trinucleotides were found (Table 4): DRV trinucleotides code for amino acids that determine the protein charge ($\alpha < 0.025$) and form a random coil ($\alpha < 0.01$) or protein surface ($\alpha < 0.025$) but not the protein core ($\alpha < 0.005$).

Therefore, the RecA filament preference for the DNA context correlates with the genetic code: the RecA filament has an increased affinity for DRV-poor regions (i.e., regions coding for protein core amino acids) and decreased affinity for DRV-enriched regions (i.e., those encoding amino acids of the protein surface and functional sites). This result correlates well with the commonly known fact that, in the course of evolution, protein cores undergo block rearrangements, whereas functional sites remain conserved.

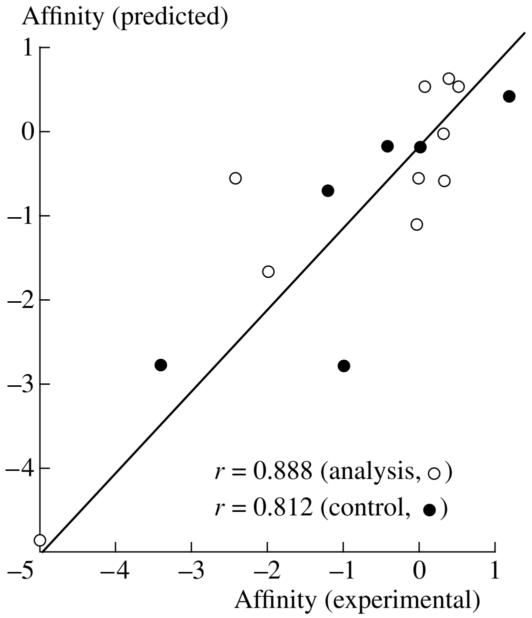

Fig. 2. Comparison of the predicted values (formula (4)) of the RecA filament affinity for DNA with the experimental values [3]; open circles, ten analyzed DNAs; closed circles, six control DNAs.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research and the Human Genome Project.

REFERENCES

1. Cox, M.M., *BioEssays*, 1993, vol. 15, pp. 617–623.
2. West, S.C., *Cell* (Cambridge, Mass.), 1994, vol. 76, pp. 9–15.
3. Mazin, A.V. and Kowalczykowski, S.C., *Proc. Natl. Acad. Sci. USA*, 1996, vol. 93, pp. 10673–10678.
4. Ponomarenko, M.P. *et al.*, *J. Comput. Biol.*, 1997, vol. 4, pp. 83–90.
5. Forster, E. and Ronr, B., *Methoden Der Korrelations- und Regressionsanalyse*, Berlin: Verlag Die Wirtschaft, 1979.
6. Lehman, E.L., *Testing Statistical Hypotheses*, New York: Wiley, 1959.
7. Zadeh, L.A., *Information Control*, 1965, vol. 8, pp. 338–353.
8. Fishburn, P.C., *Utility Theory for Decision Making*, New York: Wiley, 1970.
9. Karlin, S. *et al.*, *Mathematical Methods for DNA Sequences*, Boca Raton: CRC, 1989, p. 133.
10. Cohen, B.I. *et al.*, *Methods Enzymol.*, 1991, vol. 202, pp. 252–278.