Nucleic Acids Research, 1999, Vol. 27, No. 22 **4491–4500**

Long W tracts are over-represented in the *Escherichia coli* and *Haemophilus influenzae* genomes

Benny Shomer[1,2] and Gad Yagil[1,2,*]

[1]The European Bioinformatics Institute, Hinxton, Cambridge, UK and [2]Department of Molecular Cell Biology, The Weizmann Institute of Science, Rehovot 76100, Israel

Received June 22, 1999; Revised and Accepted September 27, 1999

ABSTRACT

The occurrence of DNA tracts of the three binary base combinations: R.Y, K.M and W;S has been mapped in the complete genomes of *Haemophilus influenzae* and *Escherichia coli*. A highly significant over-representation of W tracts is observed in both bacteria. The excess of W tracts is particularly striking in the 10% intercoding regions. Subdivision of intercoding regions into divergent (promoting), convergent (terminating) and sequential subregions shows that the excess of W tracts is most concentrated in the promoter regions. A particularly high excess of W tracts is observed in the first 200 bases 5′ upstream of coding start sites. The data suggest that W tracts have a role in promoter function. A function as unwinding centers, analogous to the role of R.Y tracts in eukaryotes, is proposed. R.Y and K.M tracts are only modestly over-represented in the two bacteria.

INTRODUCTION

Nucleotide base tracts consisting of only two bases ('binary tracts') can be found in just three combinations: (i) tracts made of purines on one strand and pyrimidines on the complementary one ('R.Y tracts'); (ii) tracts made of G,T on one strand and A,C on the other ('K.M tracts'); and (iii) the W;S pair which consists of either A,T or G,C tracts, each complementing itself. It has been known for some time that R.Y tracts are highly over-represented in higher eukaryotic DNA (1–6). More recently, we documented that R.Y tracts are also over-represented in a lower eukaryote, *Saccharomyces cerevisiae* (7).

The over-representation of R.Y tracts in eukaryotes was found to be particularly high in regulatory regions. Thus, in chromosome III of yeast, intercoding regions contain R.Y tracts, which are longer than 15 nt, 32 times more than expected in uniform (random) DNA of the same composition. When only intercoding regions up to 200 bases upstream from a gene are considered this excess increases to 46-fold (7)! This observation suggests that the excess of R.Y tracts may be connected to promoter and terminator functions.

In earlier work, Kowalski and co-workers (8,9) demonstrated that DNA in A,T-rich regions of yeast can easily unwind and can thus serve as DNA unwinding elements (DUEs). Experimental work from our lab (10) showed that A,T-rich elements are not the only potential DUEs. We found that two *S.cerevisiae* promoters containing long R.Y tracts (*CYC1* and *DED1*) are attacked by single strand-specific nucleases in the supercoiled but not in the linear state. These observations, supported by 2D topoisomer analysis, indicate that in yeast promoters R.Y tracts have a similar tendency to assume an unwound (paranemic) state. It thus seems that, in yeast, both W tracts and R.Y tracts can serve as DUEs and support the notion that these binary tracts can readily form unwinding elements.

Escherichia coli is long known to be free of the excessive R.Y tracts present in the higher eukaryotes (1), but that its promoters are rich in A,T tracts (11–13). Studies by Blattner, Kornberg, Kowalski and their colleagues (14–16) indicated that unwinding elements may play a regulatory role in bacteria, and that these elements are A,T-rather than R.Y-rich. Classical DNA melting theory actually suggests that A,T-rich tracts are the first ones to unwind. Kowalski *et al*. proposed an algorithm to predict unwinding centers based on their A,T content (17,18). This approach has been refined to include the effect of superhelicity (19). The availability of the complete sequences of *E.coli* and *Haemophilus influenzae* makes it now possible to map the occurrence of the binary tracts in the entire genome of these prokaryotic organisms.

In this study we apply several unique DNA analysis programs (GENTRACTS, ANEX and DIVCON), in addition to TRACTS (5), to analyze the occurrence and distribution of long binary tracts in the *E.coli* and *H.influenzae* genomes. It is found that W tracts are in as large an excess in these two prokaryotes as are the R.Y tracts in eukaryotes. R.Y and K.M tracts are in only moderate excess. W tracts are thus the dominating excessive binary theme in both bacteria. It is further shown that the over-representation of W tracts, and to some extent also of R.Y and K.M tracts, is particularly high in promoter regions. This observation strengthens the proposition that in prokaryotes W tracts serve as the principal unwinding elements and may thus play a crucial role in prokaryotic gene regulation.

MATERIALS AND METHODS

The complete sequences of *E.coli* (20), GenBank entry U00096, and of *H.influenzae* (21), entry HIL42023, were analyzed.

The following programs were used:

*To whom correspondence should be addressed at: Department of Molecular Cell Biology, The Weizmann Institute of Science, Rehovot 76100, Israel. Tel: +972 89 342 275; Fax: +972 89 344 125; Email: lcyagil@wiccmail.weizmann.ac.il

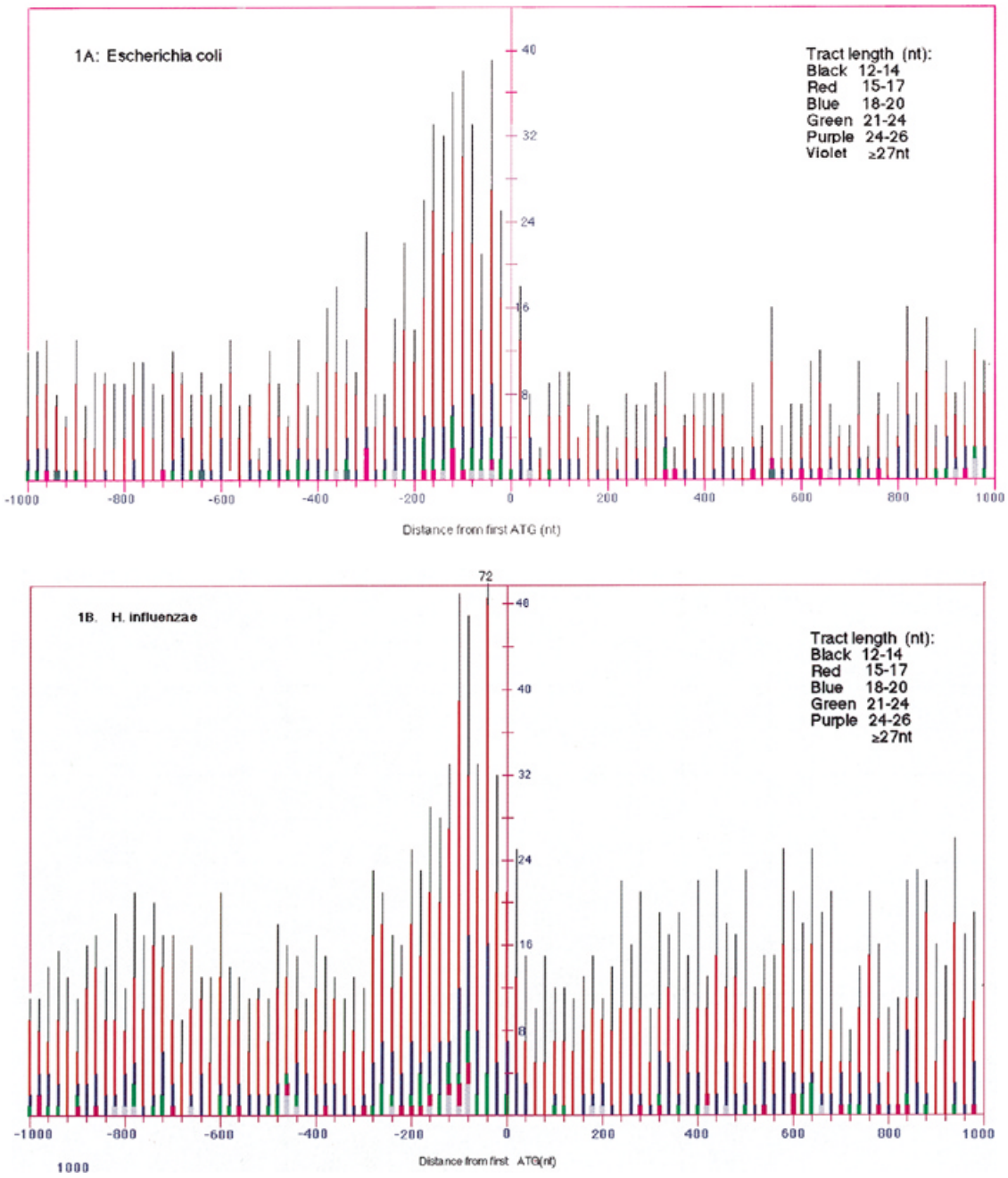

Figure 1. The number of long W tracts in bins of 20 nt (from x to x + 20) plotted against the distance x from the first translated nucleotide (A when ATG). The length of the tracts is color coded as listed. (**A**) *E.coli*; (**B**) *H.influenzae.*

(i) Program GENTRACTS, written in python with an extension module in C (B.S.), reads multiple files from an external source (EMBL) and combines them into a logical annotated genome. It then computes the positions of the features selected, identifies binary or other tracts and determines the distance to the selected features (start or stop sites of the closest genes).

(ii) Program PLOTRACTS, written in python (B.S.) creates a graphical representation of the frequencies of various tracts generated by GENTRACTS versus their distances from the closest ORFs in a 'pyramid' form, as shown in Figure 1A and B. PLOTRACTS provides a clickable www interface which enables each section on the graph to be linked to vital information on the associated genes and their products.

(iii) Program ANEX, written in FORTRAN (G.Y.), parses the annotation data from a GenBank flat file and generates a file of gene start and stop sites. The file also lists the designation, length and a 50 letter description of each annotated gene.

(iv) Program TRACTS (formerly PUR) (5) calculates and lists the frequencies of tracts of each length, and lists all tracts above a certain length. Version 6.1 of TRACTS has been

extended to calculate separate tract frequencies in coding and non-coding regions. This is performed by reading the output of ANEX and determining which bases are within ORFs and which are intercoding (mostly intergenic, but, as transcription start sites are presently mostly unavailable, 5′ and 3′ UTS are scored as intercoding). rRNA and tRNA regions are treated as genes.

(v) Program DIVCON (G.Y.) reads the lists of all tracts longer than a given length, l, generated by TRACTS, as well as the start and stop sites provided by ANEX, and assigns each intercoding region into one of four classes: divergent, convergent, or sequential of two kinds: 'www' when between two genes both coding on the GenBank listed strand, or 'ccc', when between two genes both coding on the complementary strand. DIVCON then calculates the number of tracts in each class and lists the cumulative number of bases in these tracts.

*Binary base frequencies expected in random DNA.*The number of tracts of length l and longer, $n(\geq l)$, expected in randomized DNA (with fraction of e.g. purines p, so that p + q = 1) are calculated as previously described (5), by:

$$n(\geq l) = L[(pq^l) + (qp^l)] \quad \mathbf{1}$$

where *L* is the number of bases in the input sequence (4 639 221 for *E.coli*). The number of *bases* in tracts ≥l, $N(\geq l)$, is:

$$N(\geq l) = L\,\{[(p + lq)p^l] + [(q + lp)q^l]\} \quad \mathbf{2}$$

Controls. As control, two random DNA sequences of the length and composition of *H.influenzae* were generated, using IMSL routine GGUD. The average ratios of found over expected tracts were: for W;S: 0.98,1.05; for K.M: 0.96;0.97; for R.Y: 1.03,0.99 (the ratio expected for randomized DNA is of course unity).

These averages are for tracts from 10 nt to the longest consecutive tract found in the randomized genome (of 19;20 nt for R.Y and K.M, and of 25 nt for W;S); the randomized sequence had the same base composition as the studied sequence, e.g. 0.62 for *H.influenzae*.

RESULTS

W tract distribution along *E.coli* and *H.influenzae* genomes

The distribution of long W tracts between intercoding and coding regions of *E.coli* and *H.influenzae* is shown in Figure 1A and B, respectively. The number of tracts of each length, from 12 nt upwards, in bins of 20 nt along the sequence, is plotted against the distance of that bin from the first coding position (ATG) of the closest gene. The different colors represent tracts of increasing lengths, in steps of three, e.g. red represents tracts of 15–17 nt. A distinct peak is observed between positions –200 and –1, which makes it evident that a significant concentration of long W tracts (≥12 nt) is present in the first 200–250 bases upstream to the first ATG, in both *E.coli* and *H.influenzae*. This peak is relative to a background frequency averaging, for tracts ≥15 nt (red color), 4 nt for *E.coli* or 8 nt for *H.influenzae*. The excess of long W tracts over the background is increasingly evident as tract size increases. However, statistics become less significant as tract size increases as discussed in greater detail

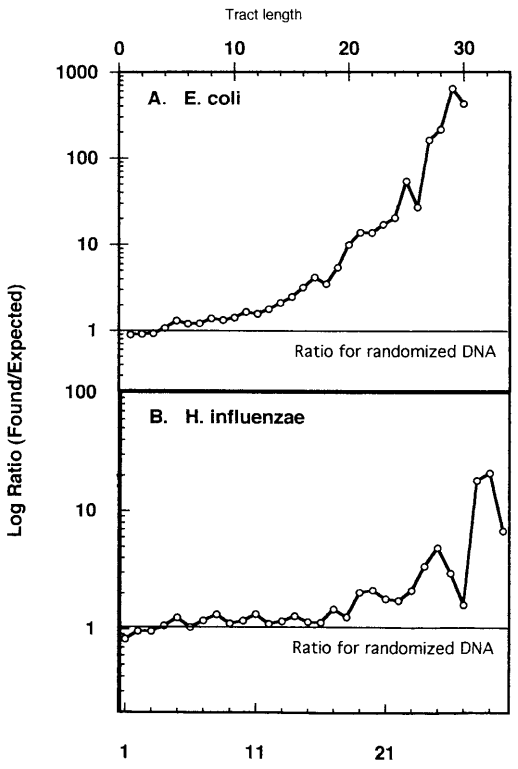

Figure 2. The log ratio of bases found in W tracts of a particular length over the expected base number, plotted against tract length. (**A**) *E.coli*; (**B**) *H.influenzae*.

below. A similar but much less significant excess was observed for R and Y tracts (not shown).

The high excess of W tracts in the first 200 bases upstream, is a strong indication of a possible role in promoter function. Since no systematic data on transcription start sites are available for either bacteria, part of the intercoding regions can be transcribed (UTS) and are not really intergenic. Also, many of the intercoding regions, especially the shorter ones, reside within operons, and therefore are probably not transcription promoters. Salgado *et al*. (22) list 292 operons in *E.coli*. If we assume that each operon contains on average two intercoding regions then 20–25% of the intercoding regions are within operons. UTS may contribute to the somewhat lesser concentration between positions –100 and –1 relative to –200 and –100 in *E.coli*. Altogether, one can expect most of the long W tracts to occur in promoter regions. This is a strong indication that the long tracts may have a role in promoting transcription.

The extent of binary tract overrepresentation

To obtain more quantitative information about the excessive W tracts, as well as of other binary tracts, program TRACTS was applied to the genomes of both *E.coli* (22) and *H.influenzae* (20). The results are shown in Table 1 and are plotted in Figure 2A and B. The number of bases in binary tracts of every length found in the genomes of *E.coli* and *H.influenzae* are listed in the tables. The length expected in randomized DNA of the same composition is also shown (Table 1, columns 3, 6 and 9). Also listed are the ratios between these two values (Table 1, columns 4, 7 and 10); these ratios give a direct measure of the

Table 1. Binary tract distribution

Tract length	Bases found W;S	Bases expected	Ratio	Bases found R.Y	Bases expected	Ratio	Bases found K.M	Bases expected	Ratio
A. *Escherichia coli*									
1	1 030 925	1 159 516	0.89	203 233	115 9805	1.04	1 085 342	1 159 803	0.94
2	1 056 204	1 159 228	0.91	403 474	115 9805	1.21	1 157 000	1 159 804	1.00
3	801 114	869 421	0.92	773 253	869 854	0.89	914 310	869 852	1.05
4	619 632	579 758	1.07	483 124	579 902	0.83	549 544	579 902	0.95
5	476 695	362 529	1.31	318 665	362 439	0.88	377 155	362 439	1.04
6	261 180	217 680	1.20	185 688	217 463	0.85	236 760	217 464	1.09
7	154 434	127 106	1.21	112 231	126 854	0.88	137 571	126 854	1.08
8	101 088	72 722	1.39	70 184	72 488	0.97	79 608	72 488	1.10
9	54 567	40 967	1.33	36 936	40 774	0.91	45 099	40 774	1.11
10	32 310	22 799	1.42	22 730	22 652	1.00	24 400	22 652	1.08
11	20 878	12 564	1.66	13 255	12 459	1.06	14 289	12 459	1.15
12	10 860	6868	1.58	7116	6795	1.05	7392	6795	1.09
13	6669	3729	1.79	3679	3681	1.00	4446	3681	1.21
14	4242	2013	2.11	2408	1982	1.21	2856	1982	1.44
15	2685	1082	2.48	1275	1062	1.20	1515	1061	1.43
16	1856	579	3.21	832	566	1.47	800	566	1.41
17	1292	308	4.19	493	301	1.64	442	300	1.47
18	576	164	3.51	126	159	0.79	144	159	0.90
19	475	87	5.47	152	84	1.81	209	84	2.49
20	460	46	10.0	140	44	3.16	180	44	4.07
21	336	24	13.9	126	23	5.42	42	23	1.81
22	176	13	13.8	44	12	3.62	22	12	1.81
23	115	6.70	17.2	–	–	–	23	6	3.62
24	72	3.51	20.5	–	–	–	72	3	21.69
25	100	1.84	54.4	–	–	–	–	–	–
26	26	0.96	27.1	–	–	–	–	–	–
27	81	0.50	161	–	–	–	–	–	–
28	56	0.26	214	28	0.24	115	–	–	–
29	87	0.14	638	29	0.13	231	–	–	–
30	30	0.07	423	–	–	–	–	–	–
	%A,T = 0.492			%A,G = 0.500			%A,C = 0.500		

over-representations at each length, and are plotted in Figure 2 against the respective tract lengths. W tracts of every length up to 30 nt are found in both *E.coli* (Table 1A) and *H.influenzae* (Table 1B). It is seen that in both bacteria, stand alone W and S bases (l = 1) are under-represented (r = 0.89; 0.82), while W tracts of every length above four nt are over-represented to an increasing extent, up to enormous excesses for the longest tracts. Thus in *E.coli*, W tracts of l = 25 (4 tracts, 100 bases) are found at 54-fold excess over the average number expected in random DNA.

The most over-represented binary pair in *E.coli* is clearly W;S. A consideration of the full output of TRACTS shows, however, that only W tracts are involved; the longest S tract is a single 22 nt tract, while seven W tracts of that length are found. The longest W tract is of 30 nt, expected only 0.07/30 times in the entire *E.coli* genome of 4 693 221 nt, a 423-fold excess! The longest W tract expected in a random genome of that length is of 21 bases (24/21 = 1.14 tracts, see Table 1A). The detailed outputs of TRACTS can be seen on the web site http://www.weizmann.ac.il/~lcyagil

R.Y tracts of every length up to 22 nt are found in *E.coli*. The two tracts of 22 nt are 3.62 times the number expected in random DNA. Two isolated tracts of 28 and 29 nt are also present. R.Y tracts up to 10 nt are actually under-represented (ratio below unity). K.M tracts are also moderately over-represented

Table 1. *Continued*

Tract length	Bases found W;S	Bases expected	Ratio	Bases found R.Y	Bases expected	Ratio	Bases found K.M	Bases expected	Ratio
B. *Haemophilus influenzae*									
1	353 389	431 808	0.82	431 204	457 505	0.93	382 882	457 499	0.84
2	392 616	407 555	0.96	466 718	457 505	1.03	396 012	457 499	0.87
3	293 970	305 666	0.96	332 244	343 127	0.93	356 808	343 120	1.04
4	230 596	215 223	1.07	220 652	228 752	0.92	231 860	228 749	1.01
5	185 400	148 821	1.25	144 900	142 970	1.01	172 825	142 972	1.21
6	106 800	102 410	1.04	91 092	85 782	1.01	113 826	85 787	1.33
7	83 125	70 317	1.18	56 735	50 040	1.1	69 006	50 045	1.38
8	63 840	48 143	1.33	36 872	28 594	1.2	42 192	28 599	1.48
9	36 702	32 828	1.12	20 790	16 084	1.2	26 523	16 088	1.65
10	26 340	22 276	1.18	11 620	8936	1.3	15 290	8939	1.71
11	20 119	15 036	1.34	7304	4915	1.4	9141	4917	1.86
12	11 220	10 096	1.11	3972	2681	1.4	5580	2682	2.08
13	7865	6744	1.17	2535	1452	1.7	3016	1453	2.08
14	5796	4484	1.29	1526	782	1.9	1820	783	2.33
15	3405	2968	1.15	825	419	1.9	1005	420	2.40
16	2224	1956	1.14	400	223	1.7	768	224	3.43
17	1887	1285	1.47	340	119	2.8	476	119	4.01
18	1062	841	1.26	54	63	0.8	234	63	3.72
19	1121	549	2.04	114	33	3.4	76	33	2.29
20	760	357	2.13	60	17	3.4	100	17	5.72
21	420	232	1.81	63	9	6.8	42	9	4.57
22	264	150	1.75	–	–	–	–	–	–
23	207	97	2.13	–	–	–	69	2.5	27.4
24	216	63	3.44	–	–	–	–	–	–
25	200	40	4.95	–	–	–	–	–	–
26	78	26	3.00	–	–	–	–	–	–
27	27	17	1.62	–	–	–	–	–	–
28	196	11	18.3	–	–	–	–	–	–
29	145	6.9	21.1	–	–	–	–	–	–
30	30	4.4	6.83	–	–	–	–	–	–
68	–	–	–	68	<E–3	>E6	–	–	–
78	–	–	–	78	<E–3	>E6	–	–	–
85	–	–	–	85	<E–3	>E6	–	–	–
87	–	–	–	87	<E–3	>E6	–	–	–
151	–	–	–	151	<E–3	>E6	–	–	–
	%A,T = 0.62			%A,G = 0.50			%A,C = 0.50		

(22-fold for the longest tract), but continuously from 5 nt up. In brief, a moderate excess of R.Y and K.M tracts is observed, much less pronounced than for the W tracts. W tracts are thus the dominant excessive binary motif in *E.coli*.

A similar situation is evident in *H.influenzae* (Fig. 2B). W tracts of every length up to 30 nt are found. The 30 nt W tract is only 6.8 times over-represented, due to the high (62%) A,T content of *H.influenzae*. In spite of this high A,T content, W tracts are continuously over-represented from 4 nt up. Up to 21 nt the over-representation of R.Y tracts is marginal. K.M tracts are in a continuous high excess, also up to 21 nt (Table 1B). Five extremely long K.M tracts, of 68–151 nt, are

Table 2. *Escherichia coli* sequences longer than 24 nt

Region	nt	From	To	Sequence
ING	29	84 064	84 092	ATTAAATATATAAATTAATTATTAAATAA
CDS	28	536 772	536 799	TATTAAAAATAATATTTTTATTTTATTT
CDS	25	563 057	563 081	AATTATAATTAATATTATATTAATT
ING	26	953 835	953 860	AAATAAAAATAAATTTTAAAAATTAA
CDS	29	1 528 094	1 528 122	AAAAAAATATTATTTATAAAATAATTAAT
ING	30	1 639 020	1 639 049	TATTTTTTATATTTTAATAATATATTTAAA
ING	25	2 627 646	2 627 670	AAATAAATATAAAATTAATATATAT
ING	27	2 986 428	2 986 454	ATAAATATAAAATTAATATATATTTAT
ING	25	2 993 218	2 993 242	AAATATAATTAAAAATAAAATTTTT
ING	29	3 281 758	3 281 786	AATAATATATTTAAAAAAATTATATTATT
ING	25	3 411 217	3 411 241	TTAAATAAATAATATATATTATTTA
ING	28	3 767 415	3 767 442	TTAATTTTATTTAAAAATATATTAATAA
ING	27	4 041 666	4 041 692	TTTTTTATTATTTAATAAATATAAATA
ING	27	4 371 837	4 371 863	AATTAAATAATTAATTTTAATTTATAA

found, as often encountered in mammalian genomes. The composition of all these long tracts is $(AACC)_n$ on one strand and $(GGTT)_n$ on the other. They are thus true microsatellites and require a special explanation. We should emphasize that the great majority of the tracts mapped by TRACTS have no particular repetitive or other symmetric feature, most of them are composed of just random mixtures of the two bases, as can be seen when all *E.coli* W tracts ≥25 nt are inspected (Table 2). A more detailed analysis is planned.

Coding versus non-coding regions

Is over-representation evenly distributed over the genome, or is there a difference between coding and non-coding regions? Coding regions compose 89% of the *E.coli* genome and 87% of the *H.influenzae* genome (Table 3). In Table 3 we see that W tracts ≥15 are somewhat less over-represented in the coding regions than in the total genome (see ratios). However, in the intercoding regions (11 and 13% of the genomes) W tracts are represented at a much higher degree than in the whole genomes: tracts 15 nt and longer ($l \geq 15$) reach a 17.63-fold excess in *E.coli* over the value expected in uniform DNA (Table 3A). The over-representation in *H.influenzae* (6.38) is of a somewhat lesser magnitude, but still highly significant (Table 3B). The vast excess of long W tracts in intercoding regions is a further indication that W binary tracts may have a regulatory function in the bacterial genome. The excess of W tracts is evident whether one examines tracts ≥12 or ≥15 nt (there are 497 such tracts in *E.coli*; Table 3 lists the number of bases in these tracts). K.M and R.Y tracts also show a significant excess in the intercoding regions of both bacteria. The excess of K.M tracts in *H.influenzae* (7.84 for K.M ≥15; 794 tracts) is particularly notable.

Over-representation is highest in promoting regions

To determine whether the excess of long tracts is connected to promoting, the 4398 intercoding regions of *E.coli*, as well as the 1818 ones of *H.influenzae*, were dissected into four subclasses: (i) divergent intercoding subregions, which are promoting in both directions (on opposite strands); (ii) convergent subregions which are terminating in both directions, and consecutive subregions, comprising of (iii) 'www' regions, which are between two ORFs coded on the analyzed (GenBank listed) strand and (iv) 'ccc' regions which are between ORFs coded on the opposite strand. Consecutive regions have both terminator and promoter elements. The division was done with the program DIVCON, which parses the data in the gene list produced by ANEX. A similar dissection into subregions has recently been carried out in conjunction with termination signals (23). DIVCON assigns each tract to the proper class, and counts tracts, as well as bases within tracts, in each class in two ways as follows.

The first way is to consider the entire intercoding region as a potential promoter or terminator region. The data in Table 4A show that the *E.coli* genome has 645 divergent and 645 convergent intercoding regions. While 27% of the divergent regions contain at least one W tract ≥12 nt, only 9.5% of the convergent regions contain at least one of these longer tracts. Similarly low percentages (10.2; 9.9%) are observed for the www and ccc regions. Only 27% of all promoter regions include a long tract, but it should be borne in mind that many intercoding regions are quite short, often only a few bases, often within operons where no promoting features should be expected. If only tracts ≥15 are considered (next three rows), the excess in divergent regions becomes even more pronounced. However, the percentage of intercoding regions having long W tracts is now smaller, indicating that a tract of length 12 nt, or possibly shorter (or incomplete), may already fill the functional role, whatever it may be. It should be added that the number of tracts expected in random DNA is, in all subclasses, ~20% of those found, so that we are dealing with excessive, subclass-specific, tracts.

The second way to assess subclass distribution is to assume that promoting or terminating regions can extend into the preceding or following genes. To examine this possibility, tract frequencies 200 bases upstream from each ATG, whether extending into an upstream gene or not, as well as downstream

Table 3. Number of bases in long binary tracts, coding and intercoding regions

	Total	Coding	Intercoding	Total	Coding	Intercoding
A. *Escherichia coli* (bases)						
Genome:	4 639 221	4 133 682	505 539	4 639 221	4 133 682	505 539
	l ≥ 12 nt			**l ≥ 15 nt**		
	W (%G,C = 0.508)					
Found	30 194	18 822	11 372	8423	3971	4452
Expected	14 930	13 302	1627	2317	2065	252
Ratio	2.02	1.41	6.99	3.64	1.92	17.63
	K.M (%A,C = 0.50)					
Found	18 143	14 709	3434	3449	2585	864
Expected	14 724	13 120	1605	2265	2018	247
Ratio	1.23	1.12	2.14	1.52	1.28	3.50
	R.Y (%A,G = 0.50)					
Found	16 448	11 510	4938	3245	1929	1316
Expected	14 724	13 120	1604	2265	2018	247
Ratio	1.12	0.88	3.08	1.43	0.96	5.33
B. *Haemophilus influenzae* (bases)						
Genome:	1 830 135	1 622 885	207 250	1 830 135	1 622 885	207 250
	l ≥ 12 nt			**l ≥ 15 nt**		
	W (%G,C = 0.38)					
Found	37 116	23 113	14 003	12 236	6011	6225
Expected	29 947	26 556	3391	8617	7641	976
Ratio	1.24	0.87	4.13	1.72	0.79	6.38
	K.M (%A,C = 0.50)					
Found	13 627	11 629	1998	3223	2429	794
Expected	5813	5151	658	895	793	101
Ratio	2.34	2.26	3.04	3.60	3.06	7.84
	R.Y (%A,G = 0.50)					
Found	9877	7147	2730	1856	1266	590
Expected	5809	5151	658	894	792	101
Ratio	1.70	1.39	4.15	2.08	1.60	5.83

from each terminating codon, were counted. The results are also shown in Table 4 (rows 9–15 in each half table). In that case the percentage of convergent regions having tracts was somewhat increased (16%; 3.6% for ≥12 or ≥15 nt in *E.coli*), but was still significantly less than in the divergent regions.

As to *H.influenzae* (Table 4B), an even higher percentage of the diverging regions (56%) have at least one W tract ≥12 nt but the convergents also have a large amount of these tracts, so that the case in favor of promoters as a major unwinding sink is less strong than for *E.coli*, but still significant; in particular when the ±200 nt range is considered. The data indicate, nevertheless, that long W tracts are present in terminator regions as well, often at the 3′ end of the RNA, or just beyond at the polyadenylation site. These regions are well known to contain W-rich elements which have been proposed to control polyadenylation and mRNA stability (19,24).

Concerning R.Y tracts, it was previously noted (5) that both the *lac* operon and pBR322, an *E.coli* derived plasmid, tend to have their few R.Y tracts concentrated in regulatory regions. Divergent intercoding regions contain three times as many R.Y tracts as convergent ones (Table 5). With K.M tracts divergent regions have nearly twice as many long tracts as convergent ones when examining all intercoding ±200 bases, but not beyond. It may be summarized that a certain amount of excessive R.Y and K.M tracts are present in *E.coli* promoters and also in terminators, but the significance is less obvious than with W tracts. *Haemophilus influenzae* also shows a certain excess of R.Y and K.M tracts, mainly in the divergent regions (not shown). Many promoters contain more than one binary tract, e.g. the *ilv* promoter (25) which has a W18 tract at –110 and an R11 tract at –155 from the first codon.

DISCUSSION

The three main findings described are: (i) a very high over-representation of long W (A,T) tracts occurs in the *E.coli* and

Table 4. W tracts in different subclasses of intercoding regions

	Divergent ← →	Convergent → ←	wwww → →	cccc ← ←	Total
A. *Escherichia coli*					
Number of IC regions	645	645	1515	1593	4398
IC only (up to previous ORF)					
Bases within ICs	153407	67574	141714	146795	509460[b]
Bases in W tracts ≥12 within ICs	3631	1055	3346	3055	11087
ICs having W tracts ≥12	173	60	156	156	545
% ICs having tracts ≥12	27	9.5	10.2	9.9	12.4
Bases in W tracts ≥15 within ICs	1395	480	1371	1077	4323
ICs having W tracts ≥15	74	23	65	52	214
% ICs having tracts ≥15	11.4	3.5	4.3	3.2	12.4
Bases 200 nt upstream (even if entering previous ORF)					
Bases within ± 200 nt[a]	258000	258000	606000	637200	1759200
Bases in W tracts ≥12 ± 200 nt from ORF	3331	1654	4894	5433	15312
±200 nt regions having W tracts ≥12	167	104	262	306	839
% ±200 regions having tracts ≥12	26	16	17	19.2	19
Bases in W tracts ≥15 ± 200 nt from ORF	1209	473	1867	1724	5273
±200 nt regions having W tracts ≥15	65	23	96	91	275
% ±200 regions having tracts ≥15	10.0	3.6	6.3	6.0	19
B. *Haemophilus influenzae*					
Number of IC regions	245	245	664	664	1818
IC only (up to previous ORF)					
Bases within ICs	51392	30203	60526	62645	204766
Bases in W tracts ≥12 within ICs	3300	2692	4410	3719	14121
ICs having W tracts ≥12	138	100	183	152	573
% ICs having tracts ≥12	56	41	28	23	32
Bases in W tracts ≥15 within ICs	1611	1113	2040	1507	6271
ICs having W tracts ≥15	68	47	88	71	274
% ICs having tracts ≥15	27.8	19.2	13.3	10.7	15.1
Bases ±200 nt upstream (even if entering previous ORF)					
Bases within ± 200 nt[a]	98000	98000	265000	265000	726000
Bases in W tracts ≥12 ± 200 nt from ORF	3555	3636	6670	7190	21051
±200 nt regions having W tracts ≥12	158	155	318	335	966
% ±200 regions having tracts ≥12	64.5	63.3	47.9	50.5	53.1
Bases in W tracts ≥15 ± 200 nt from ORF	1608	1364	2565	2498	8035
±200 nt regions having W tracts ≥15	68	63	125	129	385
% ±200 regions having tracts ≥15	27.8	25.7	18.8	19.4	21.2

[a]Often partly overlapping.
[b]This number is larger than the parallel number in Table 3A because negative ICs (when a gene starts within the preceding gene) are not subtracted here.

Table 5. R.Y and K.M tracts in *E.coli* intercoding subregions

	Divergent ← →	Convergent → ←	wwww → →	cccc ← ←	Total
R.Y ≥ 12					
Number of ICs	645	645	1515	1593	4398
Bases within ICs	153 407	67 574	141 714	146 795	509 490
Bases in tracts within	1384	397	1471	1474	4726
Having tracts ≥12	88	29	93	102	312
% having long IC tracts	13.64	4.50	6.14	6.40	7.09
Bases ±200	258 000	258 000	606 000	637 200	1 759 200
Bases in tracts ±200	1491	789	2870	2527	7677
Having W tracts ≥12	99	58	196	176	529
% having long IC tracts	15.35	8.87	12.94	11.05	12.03
K.M ≥ 12					
Number of ICs	645	645	1515	1593	4398
Bases within ICs	153 407	64 574	141 714	146 795	509 490
Bases in tracts within	1073	464	831	855	3223
Having tracts ≥12	75	32	58	58	223
% having long IC tracts	11.6	5.0	3.8	3.6	5.1
Bases ±200	258 000	258 000	606 000	637 200	1 759 200
Bases in tracts ±200	1170	1101	2234	2422	6927
Having W tracts ≥12	82	81	164	181	508
% having long IC tracts	12.7	12.6	10.8	11.4	11.6

H.influenzae genomes, as compared with random DNA; (ii) W tract over-representation is particularly high in promoter regions and, to a certain extent, in terminator and other intercoding regions; (iii) a high fraction of all promoter regions contain one or several binary tracts.

The two bacteria studied here significantly differ from the eukaryotic genomes previously studied by us and others (3,4,9,10,26) in that W tracts are the most excessive binary theme. In the genomes of the higher eukaryotes, R.Y tracts were found to be dominating, while W and S tracts were at a marginal excess at most. Kowalski and co-workers (9,17,18) have demonstrated in yeast that A,T-rich regions tend to form DUEs in autonomous replication sequences (ARS) and in several yeast gene promoter regions. R.Y tracts were also shown to serve as unwinding centers, e.g. in the *CYC1* and *DED1* promoters (19). Yeast thus occupied an intermediate position, with all three binary motifs (except S) being in a large excess (7). As to archaea, the data for *Methanococcus janaschii* (unpublished) behave like eukaryotes rather than like the prokaryotes described here. Further organisms will have to be analyzed to verify the conclusion that an excess of W tracts characterizes prokaryotes in general.

What could the function of these W tracts be? If the excessive W tracts had no function, one would expect a more modest excess in compact genomes like the bacterial ones. Previous work on eukaryotic genomes, computational and experimental, has suggested that the binary tracts may serve as DNA unwinding centers in both transcription and replication control (27). The seminal study in this direction was by Larsen and Weintraub (28), who detected single-strand-specific DNA cleavage in active chick globin promoters. Many other susceptible promoter regions have been detected since, the theme common to most of them being the binary homopurine·homopyrimidine theme, although other binary themes do occur (summarized in ref. 27).

May the W tracts serve as unwinding centers as well? A,T-rich regions are well known to be the most readily melting form of DNA. Evidence in favor of melting of W tracts as a factor in gene activation in both *E.coli* and yeast exists. Susceptibility to cleavage by single strand-specific nucleases showed that A,T-rich regions in *E.coli* associated elements (phage lambda and pBR322) can serve as DUEs (14,29). Studies concerning the *ori c* replication origin of *E.coli* (15,16) led to the same conclusion. *ori c* unwinding occurs in preparation for replication, another major cellular process requiring a certain degree of DNA unwinding.

The propensity of W tracts to unwind in the two bacteria could thus be the parallel of the propensity of R.Y tracts to unwind in higher eukaryotes. Do R.Y tracts play any role in bacteria? As seen in Table 5, in *E.coli* the R.Y tracts are in a certain, yet small, excess, a situation paralleling the situation with W tracts in the higher eukaryotes. As for K.M tracts

(A,C·G,T), these show almost random ratios in *E.coli* (found over expected ratio = ~1), but are systematically over-represented in *H.influenzae*. This supports the possibility that all DNA sequences made of only two bases have a propensity to unwind into a paranemic state (6).

A general structural basis for a propensity of binary tracts to unwind is not available at present, but in the case of W tracts is in line with classical melting theory (30,31), which leads us to expect the W tracts to separate readily. Recent procedures to include the effect of supercoiling (19,32) strengthen that view and the presence of W-rich unwinding centers in certain bacterial promoters, such as the *ilv* promoter, has been experimentally documented (25). A structural basis for ready melting of R.Y tracts is less obvious and their melting under supercoiling tension deserves further investigation. It should be added that in eukaryotes other functions have been proposed for A,T-rich elements, including signals that control mRNA degradation or polyadenylation (at the 3′ end of the gene) (24), to serve as nuclear matrix attachment sites (MAR/SARs) (32) or even as preferred nucleosome attachment sites (4). These possibilities may explain some of the observed excessive tracts. The preponderance of the W tracts in divergent (promoter) regions (Table 4) speaks nevertheless in favor of a DUE role as the major function of W tracts.

Margalit *et al*. (33) found no region of particular helix instability in *E.coli* promoters beyond the –35 to –10 sites. The explanation could be that the W tracts need not reside at a particular distance from the origin. Application of TRACTS to the first 75 bases upstream (34; results not shown) shows an abundance of longer W tracts in the –45 to –75 region. The 'A,T richness' of this region was indeed noted by several previous researchers (11–13,34,35). Excess of W tracts is thus a feature of well studied *E.coli* promoters and not a special feature of a functionally yet unidentified set of ORFs.

An unwinding region need not be located at an exact site or orientation. A linking deficiency may be formed in a remote location, far from the initiation site proper, and be transiently stabilized by single strand-specific proteins. Upon a proper signal the linking deficiency can first be transformed into negative superhelicity distributed along an entire constrained chromosome loop, and finally, upon arrival of a second set of factors, reconcentrate at the transcription/replication initiation site and permit unwinding when and where needed for entry of the copying machinery. Thus, in the *lac* operon, one W tract of 17 nt is found at the very end of the operon, i.e. at the termination of the *lacA* gene. This raises the possibility that a torsional sink may exist also at the end of a transcribing unit, remigrating to the initiation site by the supercoiling/decoiling mechanism just mentioned. All these inferences can be readily put to experimental examination. A regulatory role associated with unwinding events may open new alternatives for DNA expression control mechanisms.

ACKNOWLEDGEMENTS

This work was initiated at the European Bioinformatics Institute, Hinxton, UK, when G.Y. was the recipient of a Visiting Fellowship. The authors are indebted to Drs M. Ashburner and C. Sander for hospitality, and to many members of the EBI for helpful discussions and assistance. We thank Dr E. Yagil for his comments on the manuscript.

REFERENCES

1. Chargaff,E. (1963) *Essays in Nucleic Acids*. Elsevier, Amsterdam, volume 1, pp. 1–126.
2. Birnboim,H.C., Sederoff,R.R. and Paterson,M.C. (1979) *Eur. J. Biochem.*, **98**, 301–307.
3. Behe,M.J. (1987) *Biochemistry*, **26**, 7870–7875.
4. Behe,M.J. (1995) *Nucleic Acids Res.*, **23**, 689–695.
5. Bucher,P. and Yagil,G. (1991) *DNA sequence*, **1**, 27–43.
6. Yagil,G. (1993) *J. Mol. Evol.*, **37**, 123–130.
7. Yagil,G. (1994) *Yeast*, **10**, 603–611.
8. Umek,R.M., Eddy,M.J. and Kowalski,D. (1988) *Cancer Cells*, **6**, 473–478.
9. Umek,R.M. and Kowalski,D. (1990) *Nucleic Acids Res.*, **18**, 6601–6605.
10. Yagil,G., Shimron,F. and Tal,M. (1998) *Gene*, **225**, 152–163.
11. Nussinov,R. (1980) *J. Theor. Biol.*, **85**, 285–291.
12. Hawley,D.K. and McClure,W.R. (1983) *Nucleic Acids Res.*, **11**, 2237–2255.
13. Burge,C., Campbell,A.M. and Karlin,S. (1993) *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.
14. Schnos,M., Zahn,K., Inman,R.B. and Blattner,F.R. (1988) *Cell*, **52**, 385–395.
15. Bramhill,D. and Kornberg,A. (1988) *Cell*, **5**, 915–917.
16. Kowalski,D. and Eddy,M.J. (1989) *EMBO J.*, **8**, 4335–4339.
17. Kowalski,D., Natale,D.R. and Eddy,M.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 9464–9468.
18. Natale,D.R., Umek,R.M. and Kowalski,D. (1993) *Nucleic Acids Res.*, **21**, 555–560.
19. Benham,C.J. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 2999–3003.
20. Blattner,F.R. *et al.* (1997) *Science*, **277**, 1453–1462.
21. Fleischmann,R.D. *et al.* (1995) *Science*, **269**, 504–512.
22. Salgado,H., Santos,A., Garza-Ramos,U., van Helden,J., Diaz E. and Collado-Vides,J. (1999) *Nucleic Acids Res.*, **27**, 59–60.
23. Washio,T., Sasayama,J. and Tomita,M. (1998) *Nucleic Acids Res.*, **26**, 5456–5463.
24. Zubiaga,A.M., Belasco,J.G. and Greenberger,M.E. (1995) *Mol. Cell. Biol.*, **15**, 2219–2230.
25. Sheridan,S.D., Benham,C.J. and Hatfield,G.W. (1998) *J. Biol. Chem.*, **273**, 21298–21308.
26. Shapiro,H.S., Rudner,R., Miura,K.-I. and Chargaff,E. (1965) *Nature*, **205**, 1068–1070.
27. Yagil,G. (1991) *Crit. Rev. Biochem. Mol. Biol.*, **26**, 475–559.
28. Larsen,A. and Weintraub,H. (1982) *Cell*, **29**, 609–616.
29. Sheflin,L.G. and Kowalski,D. (1985) *Nucleic Acids Res.*, **13**, 6137–6153.
30. Orenstein,R.L. and Fresco,J.R. (1983) *Biopolymers*, **22**, 1979–2000.
31. Breslauer,K.J., Frank,R., Blocker,H. and Marky,L.A. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 3748–3750.
32. Benham,C., Kohwi-Shigematsu,T. and Bode,J. (1997) *J. Mol. Biol.*, **274**, 181–196.
33. Margalit,H., Shapiro,B.A., Nussinov,R., Owens,J. and Jernigan,R.L. (1988) *Biochemistry*, **27**, 5179–5188.
34. Lisser,S. and Margalit,H. (1993) *Nucleic Acids Res.*, **21**, 1507–1516.
35. Galas,D.J., Eggert,M. and Waterman,M.S. (1985) *J. Mol. Biol.*, **186**, 117–128.

NOTE ADDED IN PROOF

The identification in *E.coli* promoters of a UP element, containing a W_{11} tract consensus site, has been brought to our attention, thanks to Dr D. Charlier of Brussels. See Estrem,S.T. *et al.* (1999) *Genes Dev.*, **13**, 2134–2447.