3682–3685 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg625

TRACTS: a program to map oligopurine. oligopyrimidine and other binary DNA tracts

Moshe Gal, Tzvi Katz, Amir Ovadia and Gad Yagil[1],*

BioMining Ltd, POB 2526, Kadima, Israel and [1]Department of Molecular Cell Biology,
Weizmann Institute of Science, Rehovot, Israel 76100

Received February 18, 2003; Revised and Accepted April 2, 2003

ABSTRACT

A program to map the locations and frequencies of DNA tracts composed of only two bases ('Binary DNA') is described. The program, TRACTS (URL http://bioportal.weizmann.ac.il/tracts/tracts.html and/or http://bip.weizmann.ac.il/miwbin/servers/tracts) is of interest because long tracts composed of only two bases are highly over-represented in most genomes. In eukaryotes, oligopurine.oligopyrimidine tracts ('R.Y tracts') are found in the highest excess. In prokaryotes, W tracts predominate (A,T 'rich'). A pre-program, ANEX, parses database annotation files of GenBank and EMBL, to produce a convenient one-line list of every gene (exon, intron) in a genome. The main unit lists and analyzes tracts of the three possible binary pairs (R.Y, K.M and S;W). As an example, the results of R.Y tract mapping of mammalian gene p53 is described.

INTRODUCTION

Much attention has been given to genomic base composition as expressed by %A,T (W) or %G,C (S), and that for good reason (1). Much less attention has been given to the other two possible binary DNA compositions (2), namely to the percentage purines (pyrimidines) or to the percentage of G,T (complemented by A,C). Both compositions are nevertheless very interesting, not because of their variation (they are always very close to 50%) but because in most genomes, long tracts made up of those binary pairs are present in huge excess over the amount expected in random DNA (Yagil, manuscript in preparation).

The over-representation of oligopurine.oligopyrimidine tracts ('R.Y tracts') was first discovered by Chargaff and coworkers (3,4), even before the double helix was known. R.Y tract over-representation was confirmed in detail when sequence data began to accumulate (5–7). Of the two other binary DNA pairs, long K.M tracts (G,T on one strand and A,C on the complementary one) were also found to be vastly over-represented (7) in eukaryotes, while long W tracts are in high excess mainly in bacteria (8). W and S tracts are autocomplementary, S tracts playing a role in GC islands.

The function of the excessive binary tracts has yet to be established. A series of experiments from this laboratory (9) and from others (10,11) indicate that a DNA unwinding role may be involved (reviewed in 12). DNA unwinding, accompanied by complete or partial strand separation, is necessary for replication, transcription etc. and can be expected for the low melting bacterial W tracts (13,14). Early melting for R.Y or K.M tracts is less expected, but these binary motives are nevertheless found in particular high excess in 5′ promoter regions of yeast and many mammalian genomes (5,15). Consequently, a program able to map and quantify the occurrence of the various binary tracts ought to be available to the bioinformatic community. This web version of TRACTS was written with that purpose in mind.

THE PROGRAM

The program resides presently at URL: http://bioportal.weizmann.ac.il/tracts/tracts.html and/or at: http://bip.weizmann.ac.il/miwbin/servers/tracts. TRACTS consists of three main modules: (i) an html/cgi interface module; (ii) ANEX—a parser for the annotation data, a convenient gene list with one line for each gene (exon and intron) is produced by ANEX; (iii) the main unit, which identifies binary tracts, generates lists of these tracts and analyzes the data including their distribution in genomic subregions (exons, introns, etc.). The package was originally written in Fortran (5,8) and is rewritten in Perl 5.6.1 using HTML–CGI procedures. The package resides on a Unix server machine and can run up to 10 Mb of sequence at the present stage. A 'How to use' feature is accessible from the package.

Input

The program requires flat EMBL or GenBank files (.gbk) to be inserted. Versions accepting the gff format and certain XML formats are in preparation. User supplied sequences can also be analyzed but annotation features can be obtained only when annotation is supplied in GenBank or EMBL formats.

*To whom correspondence should be addressed. Tel: +972 89 460 918; Fax: +972 89 344 125; Email: gad.yagil@weizmann.ac.il

Output

Five output files are generated and can be chosen from:

1. Tract list: a list of all the binary DNA tracts longer than or equal to a certain length that is specified by the user. The list shows for each tract its length, the start and end positions, the match level (see below) and the base sequence of the tract listed.
2. Tract frequencies: a table that shows the number of tracts (and number of bases in these tracts) found for every length from one to the longest tract observed, as well as the number of tracts expected in random DNA of the same length and base composition (the formulas are given below). The table also shows the ratios between the numbers of found and expected tracts ('ratios').
3. Subregion distribution: a table giving the number of found and expected tracts in the different genomic subregions (exon, intron or intergenic) as well as the ratio between the found and expected numbers. The subregional distribution table is considered the more informative output. When run under 'mRNA', the 5′ UTR and 3′ UTR subregions are included in the exons. When run under 'CDS', these subregions are counted and listed as intergenic (strictly: 'intercoding').
4. Gene summary table: a one line entry for each gene (exon, intron) giving the name of the gene and feature, its direction (+ or −), start and end of the feature and a short functional description of the gene.
5. Annotated sequence: the full sequence analyzed is presented in a convenient 100 base 'landscape' format. The minimum length of a tract to be colored is user supplied (see options). Exons and introns are identified by their background colors and access to additional data is obtainable by mouse-activated links.

Options and parameters

In addition to the flat input files, the following parameters are entered by means of buttons, a window or a drop down menu:

1. The binary motive: the user can decide whether R.Y, K.M, or S;W motives are to be run, as pairs or individually. R and Y tracts as well as K and M tracts are generally combined, because when one of them is on the plus strand (the strand listed in GenBank), its pair mate will complement it on the minus strand. The autocomplementary S and W motifs can be individually run, which can be useful to users interested in GC islands. Poly A, poly G, poly C and poly T tracts can also be mapped. When the user chooses these or the 'none' button, the program will produce just the annotated sequence and the gene list.
2. Match level: tracts consisting of less then 100% of the designated binary pair can also be identified and listed by TRACTS. Thus, for instance, a 90% match level means that one 'outsider' base in 10 nt, or 3 in 30 nt, are tolerated and counted.
3. Genomic features: the user has to choose whether mRNA or CDS data will be extracted from the annotation table and processed. UTR regions will be identified only when the 'mRNA' parameter is chosen. Choice of mRNA is important when 5′ promoters are of interest. However, many GenBank entries (e.g. yeast chromosomes) do not have yet mRNA entries. A combined analysis of both features, enabling separate UTR identification, is planned.
4. The minimum tract length to be displayed in the tract list and the annotated sequence, as well as a tract length for which to calculate subregional distribution, can be specified in the drop down menus.

Expected binary tract frequencies

The expressions by which frequencies of binary pairs expected in random DNA are calculated are as following [$N(l)$ gives the number of *tracts* of length l expected in randomized DNA of the same length L and base composition p as the analyzed DNA sequence]:

$$N(l) = L(p^l * q^2 + q^l * p^2)p + q = 1 \qquad \mathbf{1}$$

where p, q are the fractions of the participating base pairs (e.g. p is the fraction of A + G).

The number of *bases* expected in tracts of length $n(l)$ is simply:

$$n(l) = l * N(l) \qquad \mathbf{2}$$

To calculate expected values for only one motive in a pair, only one member of each sum is to be used. The expected number of *tracts* equal or greater than a given length l, $N(\geq l)$, can be shown to be (5,8):

$$N(\geq l) = L(p * q^l + q * p^l) \qquad \mathbf{3}$$

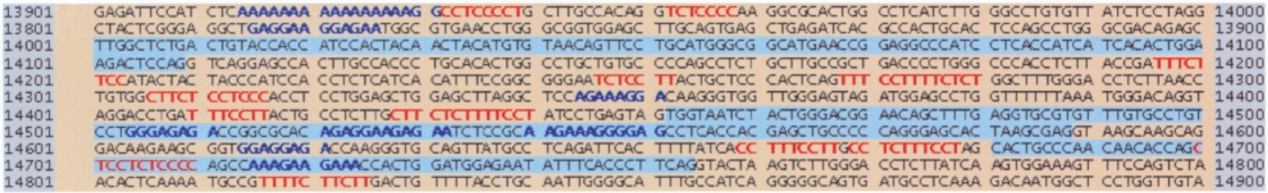
The expected number of *bases* in tracts $\geq l$, $n(\geq l)$, is:

$$n(\geq l) = L\{(p + ql)p^l + (q + pl)q^l\} \qquad \mathbf{4}$$

These four expressions are valid only for tracts with no outsider bases (i.e. at the 100% match level). The validity of these expressions was tested by generating and running random DNA sequences of given base compositions and length L.

EXAMPLE—p53

As an example, the human oncostatic gene p53 (entry HSP53G, accession no. X54156, 20 303 bases) will be brought. p53 has 11 exons; about 50% of the sequence is in the first intron (10 738 nt). In Figure 1, part of the annotated sequence, including exons 6–9, is shown, giving an impression of R.Y occurrences—mainly in the introns. The 'tract frequencies' table, the main output, is shown in Table 1, for the R.Y tracts. In column 5 it can be seen that the longest tract fully-expected in randomized p53-like DNA would be 13 nt long, while 35 R.Y tracts longer than 13 nt are actually observed (column 4). Every tract length up to 26 nt (except



```
13901   GAGATTCCAT CTCAAAAAAA AAAAAAAAAG GCCTCCCCTG CTTGCCACAG GTCTCCCCAA GGCGCACTGG CCTCATCTTG GGCCTGTGTT ATCTCCTAGG 14000
13801   CTACTCGGGA GGCTGAGGAA GGAGAATGGC GTGAACCTGG GCGGTGGAGC TTGCAGTGAG CTGAGATCAC GCCACTGCAC TCCAGCCTGG GCGACAGAGC 13900
14001   TTGGCTCTGA CTGTACCACC ATCCACTACA ACTACATGTG TAACAGTTCC TGCATGGGCG GCATGAACCG GAGGCCCATC CTCACCATCA TCACACTGGA 14100
14101   AGACTCCAGG TCAGGAGCCA CTTGCCACCC TGCACACTGG CCTGCTGTGC CCCAGCCTCT GCTTGCCGCT GACCCCTGGG CCCACCTCTT ACCGATTTCT 14200
14201   TCCATACTAC TACCCATCCA CCTCTCATCA CATTTCCGGC GGGAATCTCC TTACTGCTCC CACTCAGTTT CCTTTTCTCT GGCTTTGGGA CCTCTTAACC 14300
14301   TGTGGCTTCT CCTCCCACCT CCTGGAGCTG GAGCTTAGGC TCCAGAAAGG ACAAGGGTGG TTGGGAGTAG ATGGAGCCTG GTTTTTTAAA TGGGACAGGT 14400
14401   AGGACCTGAT TTCCTTACTG CCTCTTGCTT CTCTTTTCCT ATCCTGAGTA GTGGTAATCT ACTGGGACGG AACAGCTTTG AGGTGCGTGT TTGTGCCTGT 14500
14501   CCTGGGAGAG ACCGGCGCAC AGAGGAAGAG AATCTCCGCA AGAAAGGGGA GCCTCACCAC GAGCTGCCCC CAGGGAGCAC TAAGCGAGGT AAGCAAGCAG 14600
14601   GACAAGAAGC GGTGGAGGAG ACCAAGGGTG CAGTTATGCC TCAGATTCAC TTTTATCACC TTTCCTTGCC TCTTTCCTAG CACTGCCCAA CAACACCAGC 14700
14701   TCCTCTCCCC AGCCAAAGAA GAAACCACTG GATGGAGAAT ATTTCACCCT TCAGGTACTA AGTCTTGGGA CCTCTTATCA AGTGGAAAGT TTCCAGTCTA 14800
14801   ACACTCAAAA TGCCGTTTTC TTCTTGACTG TTTTACCTGC AATTGGGGCA TTTGCCATCA GGGGGCAGTG ATGCCTCAAA GACAATGGCT CCTGGTTGTA 14900
```

Figure 1. Exons 5–7 of human oncostatic protein p53. R tracts are in bold red letters, Y tracts in bold blue. Exons are on a light blue background, introns on a light brown background.

Table 1. Frequencies of R.Y tracts in human p53

Tract length	Tracts				Bases in the tracts			Bases in tracts eq. AND longer		
	R	Y	R + Y (f)	Expected (e)	Found	Expected	Ratio (f/e)	Found	Expected	Ratio (f/e)
1	1905	1893	3798	5073.8	3798	5073.8	0.75	20 303	20 303	1.00
2	1006	1028	2034	2535.9	4068	5071.8	0.80	16505	15229.2	1.08
3	518	584	1102	1267.9	3306	3803.8	0.87	12437	10157.5	1.22
4	399	408	807	634.2	3228	2536.9	1.27	9131	6353.7	1.44
5	200	169	369	317.4	1845	1586.8	1.16	5903	3816.8	1.55
6	130	96	226	158.9	1356	953.2	1.42	4058	2230.0	1.82
7	60	47	107	79.6	749	556.9	1.35	2702	1276.8	2.12
8	21	31	52	39.9	416	318.9	1.31	1953	719.9	2.71
9	20	16	36	20.0	324	179.8	1.80	1537	401.0	3.83
10	6	7	13	10.0	130	100.2	1.30	1213	221.3	5.48
11	6	7	13	5.0	143	55.3	2.59	1083	121.1	8.94
12	9	4	13	2.5	156	30.2	5.16	940	65.9	14.3
13	4	3	7	1.3	91	16.4	5.53	784	35.6	22.0
14	2	2	4	0.6	56	8.9	6.30	693	19.2	36.2
15	5	3	8	0.3	120	4.8	25.1	637	10.3	62.0
16	2	3	5	0.2	80	2.6	31.2	517	5.5	94.2
17	0	1	1	0.1	17	1.4	12.4	437	2.9	150
18	1	1	2	0	36	0.7	49.3	420	1.6	271
19	2	0	2	0	38	0.4	98.1	384	0.8	467
21	1	0	1	0	21	0.1	194	346	0.2	1510
22	1	1	2	0	44	0.1	769	325	0.1	2693
23	1	0	1	0	23	0	763	281	0.1	4428
24	0	1	1	0	24	0	1514	258	0	7743
25	2	0	2	0	50	0	6009	234	0	13 391
26	2	0	2	0	52	0	11 916	184	0	20 103
29	1	0	1	0	29	0	46 386	132	0	1 00 955
32	2	0	2	0	64	0	7 20 300	103	0	5.5 * E5
39	0	1	1	0	39	0	4.2 * E6	39	0	2.1 * E7

20 nt) is found, and three longer tracts, up to 39 nt, are present. The over-representation of the 13 nt tracts is already 5.53-fold (column 8, the ratio column). The smooth increase of the ratios from $l = 3$ on, means that over-representation does not depend on a phenomenon related to a single length category.

In Table 2, a 'tract list' output, for all R.Y tracts in p53 longer than 17 nt, is shown. A number of 'simple' motifs can be identified in these tracts; other tracts are nevertheless as scrambled, or cryptic, as possible. Table 3 shows the subregion distribution for the p53 tracts (output 'subregions'); generally, coding regions are relatively poor in long tracts, as can be expected because of the load it imposes on the protein coding capacity. The last three long Y tracts in Table 3 are actually in the 3′ UTR part of the gene, which is often loaded with binary tracts. The 5′ intergenic region that includes the promoter region of p53, not a strong promoter, is atypically poor in binary tracts. The total excess of bases longer than 15 nt is 62-fold (last column of Tables 1 and 3).

CONCLUSION

Results from human chromosomes as well as from *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana* genomes (Yagil, manuscript in preparation) show a similar general result—all binary tracts except for the S motif are heavily over-represented in all eukaryotes so far tested. Results for yeast (15), for collections of vertebrate genes (7), for globin (7) and *Drosophila* (16) have been previously published and a more detailed discussion of the issues involved can be found in those publications. We hope that the web version of TRACTS will stimulate further studies on the perplexing phenomenon of the high binary DNA over-representation.

ACKNOWLEDGEMENT

The authors wish to express their thanks to Dr Eitan Rubin, Dr Marilyn Safran, Dr Shifra Ben-Dor and many other members

Table 2. R.Y tracts of gene p53 longer than 17 nt

From	To	Length	Sequence
1706	1730	25	GAGAGGGGAGGAGAGAGAGAGAAAA
2006	2024	17	TCTTTTTTTTTTTTTTT
4604	4629	26	AGAAAAAAAAAGAAAGAAAGAAAAAA
5377	5398	22	AAAAAAAGAAAAAGAAAAAGGA
6101	6132	32	AAAAAAAAAAAAAAAAAAAAAAGAAAAGAAAA
6518	6540	23	AAAAAAAAAAAAAGAAAAAGAAA
7140	7171	32	AAAAAAAAAAAAAAAAAAGGAAAGAAAAAAAA
9478	9506	29	GAAAAAAAAAAAAAGAAAAAGAAAGAGAG
9846	9870	25	AAAAAAAAAAAGAAAAAGAAAAAGA
10 019	10 044	26	AAAAGAAAAAAGAAAGAAAGAAAGAA
12 883	12 903	21	AAAAAAAAAAAAAAAGAAAAG
13 914	13 931	18	AAAAAAAAAAAAAAAAGG
15 149	15 170	22	CTTTTTTTTTTTTTTTTTTTTT
16 975	16 993	19	AGAAAAAAAAGAAAAGAAA
19 200	19 217	18	TTCCCTCTCCCTCTCCCT
19 432	19 470	39	TTTCTTTTTTCTTTTTTTTTTTTTTTTTTCTTTTTCTTT
19 802	19 825	24	CCCTTCCCCTCCTTCTCCCTTTTT

Table 3. Bases found in R.Y tracts GE.15 in p53 subregions (mRNA)

	Coding	Intergenic	Introns	Total seq.
Sequence (nt)	2510	1269	16 524	20 303
Bases found (nt)	81	15	541	637
Bases expected (nt)	1.27	0.64	8.4	10.3
Ratio	63.8	23.4	64.7	62.0

of the Biological Computing Unit at the Weizmann Institute for their assistance in the presented endeavor.

REFERENCES

1. Bernardi,G., Mouchiroud,D., Gautier,C. and Bernardi,G. (1988) Compositional patterns in vertebrate genomes, conservation and change in evolution. *J. Mol. Evol.*, **28**, 7–18.
2. Burge,C., Campbell,A.M. and Karlin,S. (1993) Over and underrepresentation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.
3. Tamm,C., Shapiro,H.S., Lipshitz,R. and Chargaff,E. (1952) Distribution density of nucleotides within a deoxyribonucleic acid chain. *J. Biol. Chem.*, **203**, 673–698.
4. Chargaff,E. (1963) *Essays in Nucleic Acids.* Elsevier, Amsterdam, The Netherlands, p. 226.
5. Bucher,P. and Yagil,G. (1991) The occurrence of oligopurine–oligopyrimidine tracts in eukaryotic and prokaryotic genes. *DNA Sequence*, **1**, 27–43.
6. Behe,M.J. (1995) An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. *Nucleic Acids Res.*, **23**, 689–695.
7. Yagil,G. (1993) The frequency of two-base tracts in eukaryotic genomes. *J. Mol. Evol.*, **37**, 123–130.
8. Shomer,B. and Yagil,G. (1999) Long W tracts are over-represented in the *E.coli* and *H.influenzae* genomes. *Nucleic Acid Res.*, **27**, 4491–4480.
9. Yagil,G., Shimron,F. and Tal,M. (1998) DNA unwinding in the CYC1 and DED1 yeast promoters. *Gene*, **225**, 152–163.
10. Larsen,A. and Weintraub,H. (1982) An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin. *Cell*, **29**, 609–616.
11. Hentschel,C.C. (1982) Homocopolymer sequences in the spacer of a sea urchin histone gene repeat are sensitive to S1 nuclease. *Nature*, **295**, 714–716.
12. Yagil,G. (1991) Paranemic structures of DNA and their role in DNA unwinding. *Crit. Revs. Biochem. Mol. Biol.*, **26**, 475–559.
13. Bramhill,D. and Kornberg,A. (1988) A model for initiation at origins of DNA initiation. *Cell*, **5**, 915–917.
14. Kowalski,D. and Eddy,M.J. (1989) The DNA unwinding element, a novel, cis acting component that facilitates the opening of the *E.coli* replication origin. *EMBO J.*, **8**, 4335–4339.
15. Yagil,G. (1994) The frequency of oligopurine–oligopyrimidine and of other two-base tracts in yeast chromosome III. *Yeast*, **10**, 603–611.
16. Yagil,G. (2001) Binary DNA tracts can serve as DNA unwinding centers. *J. Biomol. Struct. Dyn.*, **18**, 911.