Nucleic Acids Research, Vol. 20, No. 3 551–556

Enrichment of oligonucleotide sets with transcription control signals II: mammalian DNA

S.Volinia+, C.Scapoli, R.Gambari, R.Barale and I.Barrai*
Dipartimento di Biologia Evolutiva e Istituto di Chimica Biologica – Universita' di Ferrara, Via Borsari 46, 44100 Ferrara, Italy

Received September 23, 1991; Revised and Accepted January 9, 1992

ABSTRACT

We studied the frequency distribution of oligonucleotides 10 bp long in a sample of 1.6 Mb of mammalian genes, containing 579 sequences from GenBank(R) 55.0, with the aim of detecting transcription control signals. 2216 decamers had a frequency higher than 10 times the mean and were subjected to further statistical analysis. For each of the 2216 decamers (parents), we counted the individual frequencies of the 30 decamers differing from the parent by one base mutation (progeny) and then calculated two variance/mean chi squares for the progeny, with and without the parent. We then studied the distribution of the ratio between the two chi squares. Out of 2216 decamers, 346 had a chi square ratio of 1.9 or larger. In this final set, which corresponds to less than 0.033 per cent of all possible decamers, 18 were found to contain 23 eukaryotic transcription control elements 5 – 10 bp of length, such as Sp1 and others. Furthermore, when compared to 210 random sets containing 346 decamers, this set contains a highly significant excess of the longer signals.

INTRODUCTION

Computer analysis has been used to identify regulatory DNA sequences mostly by means of homology searches (1–3) and also by considering the absolute frequency of oligonucleotides in large sets of genes or ultimately in complete genomes (4–12). If the oligonucleotide is for example a decamer, it would be expected to occur about once in a million base pairs. But if that same decamer plays a functional role in a large number of different genes, its frequency should then be higher than expected according to compositional models.

After observing that the NF-kB binding site behaves according to this simple hypothesis in animal viruses, we developed computer methods to identify and characterize frequent decamers in a sample of Genbank(R) entries (13).

We studied, as a preliminary approach, a sample of sequences from mammalian and avian viruses, since (a) the majority of viral genomes displays only 'functional' sequences, either coding or regulatory, and (b) viral genomes contain DNA elements that interact with host proteins involved in cellular tropism and gene activation (14, 15). With our method we were able to considerably enrich with signals a set of 479 viral decamers (13).

The aim of the present work was to apply our computer methodology to mammalian DNA, with the purpose of 1) testing if the method has any validity with DNA from complex genomes and 2) identifying a set of decamers enriched of transcription signals.

The experimental design we devised to detect frequent decamers putatively playing functional roles is described in detail in (13), but we will briefly repeat it here.

At first, we calculated the frequency of each decanucleotide in a mammalian sample from GenBank(R) and collected the most frequent decamers having frequency higher than ten times the mean, into a preliminary subset, which we call the 'query' set.

Secondly, for each of the selected decamers, namely for each parent, we determined the frequencies of the 30 progeny decamers which differ from the parent only by one base substitution.

Thirdly, we calculated the chi square for equality of the frequencies including the parent, with 30 degrees of freedom, and excluding the parent, with 29 degrees of freedom, so that decamers were classified in two groups: (a) those whose progeny mutants have frequency similar to the parent and (b) those where the parent is much more frequent than any individual member of the progeny.

This last class of decanucleotides could be of interest, since it might contain DNA elements that play biological roles depending upon the uniqueness of their sequence. This second set of frequent decamers was then examined for presence of known transcription signals. It was selected on the basis of the chi square ratio, and being the last set considered, we call it the 'residual' set.

MATERIALS AND METHODS

A sample of 579 sequences of mammalian DNA, for a total length of 1.601 Mb, was selected from Genbank(R) release 55.0. The selected entries are reported in Table I. The criteria for selecting

* To whom correspondence should be addressed
+ Present address: Ludwig Institute for Cancer Research, 91 Riding House Street, London W1P 8BT, UK

Table I. The 579 mammalian sequences used in this study

HUMA1ATP	HUMG6PD	HUMMHCP52	HUMUK	HUMENKB1
HUMACCYBA	HUMGCB	HUMMHCW3	HUMVPNP	HUMENKPH1
HUMACHRA7	HUMGCRA	HUMMHDC3B	HUMVWFM	HUMERRNA
HUMAGG	HUMGCRB	HUMMHDCB	HUMA1ACM	HUMETRD
HUMALBAF4	HUMGG	HUMMHDRB1	HUMA1ATM	HUMFBRB
HUMALBGC	HUMGHN	HUMMHDRB2	HUMA1ATZ	HUMFIXA
HUMANFA	HUMGLUTRN	HUMMHDRB3	HUMA2TPI	HUMFMSN
HUMANG	HUMHBA1	HUMMHDRHA	HUMACBPA1	HUMFOL1
HUMAPOA2I	HUMHBA3	HUMMHSBB1	HUMACBPB1	HUMFXM
HUMAPOAI1	HUMHBA4	HUMMIS	HUMACTASK	HUMG3PD
HUMAPOAII	HUMHBBRT	HUMMYCC	HUMACTCA2	HUMG3PDP
HUMAPOB	HUMHMGCOA	HUMNGFB	HUMADAM2	HUMGAST2
HUMAPOBA	HUMHP2FS	HUMOAT	HUMADHIAB	HUMGC
HUMAPOCII	HUMHPARS1	HUMOPS	HUMADHIG	HUMGCBL
HUMAPOE4	HUMHPARS2	HUMP53R	HUMALAD	HUMGFI2
HUMBHA	HUMIFNA04	HUMP971	HUMALBF1	HUMGFIB
HUMC3	HUMIFNAGS	HUMPGKPS	HUMALDBX	HUMHBA2
HUMC9	HUMIFNB3	HUMPGKPX	HUMALDH1	HUMHBAPS
HUMCERP	HUMIGCC4	HUMPHH	HUMALDH2	HUMHCII
HUMCG1PA1	HUMIGCC8	HUMPLA	HUMAMYAP	HUMHMG17
HUMCG1PAT	HUMIGCD1	HUMPOMC	HUMAMYAS	HUMHPA1S
HUMCG2A1	HUMIGCD2	HUMPRCA	HUMARS2A	HUMHPA2B
HUMCINHP	HUMIGCD6	HUMRASH	HUMAS1PS	HUMHPA2R
HUMCRF	HUMIGCD7	HUMRENA3	HUMAS3PS	HUMHPRT
HUMCRPG	HUMIGHAD	HUMRENA4	HUMASA	HUMHXM
HUMCS1	HUMIGKVI1	HUMRGEA	HUMASGPR1	HUMIFNA01
HUMCTHD	HUMIL1P	HUMRGM	HUMASGPR2	HUMIFNA03
HUMCYP145	HUMIL2A	HUMRGNTSA	HUMATCT4	HUMIFNA20
HUMCYPNO	HUMIL2RA	HUMRPS14	HUMBLYM1	HUMIFNAA
HUMEGFRN	HUMINS1	HUMRSKP08	HUMC1A22	HUMIFNAB
HUMEGFRS	HUMINSPR	HUMRSKPNA	HUMC1AIN1	HUMIFNAD
HUMENKPH2	HUMINSRA	HUMSISM	HUMC1INHA	HUMIFNAH2
HUMER41	HUMKEREP	HUMSOMI	HUMC4BP	HUMIFNAI
HUMERMCF	HUMKIN10	HUMSPI	HUMC5	HUMIFNAPS
HUMERP	HUMLAMA	HUMTBB1P	HUMCATF	HUMIFNATA
HUMESTR	HUMLDLR18	HUMTBB5	HUMCG3A1M	HUMIFNATB
HUMFBRAA	HUMMETIA	HUMTBB7P	HUMCG5B	HUMIFNATC
HUMFBRG	HUMMETIF1	HUMTBBM40	HUMCG6B	HUMIFNB1F
HUMFBRGAB	HUMMG1	HUMTCBJB	HUMCMOS	HUMIL2R8
HUMFIXG6	HUMMH	HUMTFRR	HUMCN2	HUMKER56K
HUMFNMC	HUMMHA2	HUMTGFB	HUMCRAS2P	HUMKIN01
HUMFOL5	HUMMHA3	HUMTHBNB	HUMCSF1M3	HUMLAMC
HUMFOS	HUMMHBC	HUMTNFA	HUMDBP	HUMLCAT
HUMFTRRA	HUMMHBGEN	HUMTPAR	HUMEIF2A	HUMLDHA
HUMFVII	HUMMHC4A	HUMTROPFB	HUMENK1	HUMLDHAP
HUMFVIII	HUMMHCP42	HUMTUBAG	HUMENK2	HUMLHB
HUMFXI	HUMMHCP51	HUMTUBBM	HUMENKA	HUMLHRH
HUMLT	MUSCRYA5A	RATACCYB	RATSVSIVG	MUSMETII
HUMMET2	MUSCYP14X	RATACSKA	RATTATR	MUSMHBDA
HUMMETIE	MUSCYP345	RATAGPA1G	RATTGB	MUSMHC4T
HUMMG2	MUSEGF	RATALAC	RATTGDCL	MUSMHCQ1
HUMMG3	MUSERFV41	RATANF	RATTHY1G	MUSMHDB
HUMMHDCAM	MUSERP	RATAPOA01	RATTNTG	MUSMHSEC
HUMMHDR5B	MUSFOS	RATAPOA02	RATTUBAL2	MUSMLC13P
HUMMHDRB	MUSGFAPD	RATAPOA03	RATTUBAPS	MUSP53M
HUMMHDRB2	MUSGKAL1	RATAPOEA	RATUDPGTR	MUSPRIMP
HUMMHGM	MUSGPD	RATCATL	RATVPNPA	MUSRENIN
HUMMHSBA	MUSH	RATCPSI01	MUS68NPF	MUSRGE
HUMMHSBAP	MUSHBBH0	RATCRYG	MUSABL2	MUSRPLPSA
HUMMTLMC2	MUSHBBH3P	RATCTRPB	MUSACACM	MUSTAT
HUMOTC	MUSIL3C	RATCYC	MUSACSM	MUSTGDEG
HUMOTNPI	MUSINT1	RATCYCPD	MUSACCYB	MUSTNF
HUMP33	MUSKTEPII	RATCYPD45	MUSACHRD	MUSTUBA1M
HUMPNU	MUSLBPA	RATCYPRM	MUSADAM	RATALBM
HUMPOLB	MUSMAL	RATELAIII	MUSAMY1M	RATALDA
HUMPRL1	MUSMHAB3	RATENKB	MUSAMY2G	RATAMLS
HUMPRL3	MUSMHCQ3	RATFBA5E	MUSAMY2M	RATAPOAIV
HUMPRPC	MUSMHDD	RATFBB5E	MUSARAF	RATATCOX8
HUMPTH2	MUSMHIEAD	RATFBG5E	MUSASPATC	RATCASA
HUMRAF19	MUSMHEBD	RATFN3M1	MUSCAIIM	RATCASB1
HUMRAF2PS	MUSMHKK	RATGH1	MUSCCPA	RATCASB5
HUMRBP1	MUSMHTLAC	RATGLCB	MUSCPR	RATCBXPA
HUMRGLTIA	MUSMHTLPS	RATGLTPG	MUSCRYG2D	RATCKM
HUMRSAP3	MUSMOS	RATINSI	MUSCRYG42	RATCMOS
HUMRSH3	MUSMX	RATINSII	MUSCYC1MC	RATCRF
HUMRSSA1A	MUSMYBM	RATLCA1	MUSCYCP22	RATDBP
HUMSISA1	MUSMYCNA	RATMT12C	MUSCYCP4	RATEIF2A
HUMTBB11P	MUSMYSA	RATMT1PA	MUSCYP34A	RATELAI1
HUMTGKQL	MUSNGFAG2	RATMT1PB	MUSEIF4AL	RATFBRGA
HUMTGLU	MUSNGFAG4	RATMYHAB2	MUSERE1M	RATGAPDHB
HUMTHYS	MUSODC	RATMYL2G	MUSFO5E	RATGBA2US
HUMTK	MUSP53PG	RATPDI	MUSFOLTER	RATGNPAS
HUMTPI	MUSPER	RATPECG1	MUSGI	RATHDP
HUMTPIPSA	MUSPIM1	RATPKCI	MUSGS	RATIGFII1
HUMTPIPSB	MUSPKCD	RATPKCII	MUSHBA	RATLDH
HUMTROPA	MUSPRP	RATPKL	MUSHBBH1	RATLHB
MUS45SRNA	MUSPRPMPB	RATPLPX	MUSHBBH2P	RATLL
MUSACAP	MUSREN2IA	RATPRLHR4	MUSHBBY2E	RATMABPA1
MUSAFP	MUSRGEB3	RATPRLSD1	MUSHPRT	RATMABPA2
MUSAPOIVA	MUSRNRM1	RATPTRY24	MUSIFNA1M	RATMABPC
MUSASP	MUSRPL30	RATPTRYI	MUSIFNA7	RATMABPPS
MUSASPATM	MUSRPL3A	RATRGEB4	MUSIFNG	RATMAPA1
MUSBAND3	MUSRPOII2	RATRGMX	MUSIFNPS	RATMBP
MUSBFIGE	MUSRPS16	RATRHL1	MUSIL2REC	RATMT1PC
MUSC31	MUSRRM	RATRII51	MUSKTCEB	RATMYHCA
MUSC3B	MUSRSRP2	RATSCD	MUSL10	RATNMOR
MUSCMYC1	MUSUPA	RATSOM141	MUSMETI	RATNNE
RATOTC	BOVFGFB	BOVTHBNM	RABHBB3	
RATOXTNP	BOVFX	BOVTRNA1	RABHBB4	
RATP1A75	BOVGG	BOVTRNB	RABIFRCP	
RATPHH	BOVGH	BOVVP	RABLDLRM	
RATPECG6	BOVGLYAA3	DOGCK	RABMHI191	
RATPOLB	BOVGSAR	DOGINS	RABMYSAC	
RATPRLHR1	BOVHBB	GOTHBAI	RABMYSBC	
RATPTR1	BOVHBG	GOTHBAII	RABPRCAM	
RATRSIDC	BOVHBP1	GOTHBBEI	RABTNF	
RATS100	BOVHBP2	GOTHBBEII	RABUG	
RATSBP	BOVHBPP3	GOTHBBEVP	SEAMG3	
RATSPOT14	BOVIFNAC	GOTHBBPS1	SHPATPAA	
RATTGE3	BOVKERVIC	GOTHBBZPS	SHOCRF	
RATTGPK4	BOVKIN1HW	HRSHBA1	SHPKERB2A	
RATUG11	BOVLDLRX	PIGENKB	ATRB1GLOP	
RATWAP1	BOVLHB	PIGPOMC	BABA1AT	
BOVACHRA	BOVMIS	PIGUPA	BABHBDPS	
BOVACHRB	BOVOT	RABALDA	CHPHBAPS	
BOVCHYMOA	BOVPBC	RABATPAC	CHPHBBC	
BOVCNP	BOVPOMC7	RABCPK	CHPHBBPCH	
BOVCSAA1B	BOVPRLP1	RABCY450A	CHPRGMC	
BOVCYPC21	BOVPTHG	RABHBAPT	GORHBBPG	
BOVENKEPH	BOVS	RABHBB1		
BOVFBRB	BOVTG	RABHBB2PS		

GenBank(R) entries were: (a) sequences had to be longer than 1 Kb and shorter than 30 Kb; (b) all individual sequences available in the database under condition a) were included, taking care to avoid duplications and (c) the presence of a vast excess of mammalian, highly conserved similar genes was avoided. These latter, due to high degree of homology, could lead to a biased selection of regulatory regions.

We wrote and used three programs for our study. The program HYPERSCAN generates the frequency distributions of all the k-tuples up to k = 10 in any set of sequences from GenBank(R). The program ESTRAZ extracts the progeny with their observed frequencies, from a list of query decamers. BOOTMUT performs the bootstraps, namely it repeats for an arbitrary number of cycles the operations on the distribution of decamers to obtain a standard for significance tests on results, using random queries. The program LOCALIZ was also written and used to identify and count k-tuples in short DNA sequences.

Our programs were written in Turbo Pascal version 5.0 and accept entries from the Genbank(R) database release 55.0. We used an IBM PC AT with IBM 3363 optical disk, and two PS/2 PCs with Hitachi CDR 1503S CD ROMs.

Table II. Distribution of 1,048,576 decanucleotides in 1.601 Megabases of mammalian DNA.

Occurrences	Frequency	Occurrences	Frequency
0	440526	50	4
1	245538	51	3
2	142693	52	4
3	84179	53	4
4	50462	54	2
5	30591	55	1
6	18785	56	3
7	11829	57	1
8	7486	58	4
9	4914	59	4
10	3206	60	2
11	2158	61	4
12	1529	62	1
13	1056	63	3
14	780	71	2
15	628	75	1
16	435	76	2
17	319	77	1
18	260	78	1
19	216	81	1
20	140	82	3
21	112	87	2
22	90	88	1
23	73	90	1
24	55	97	1
25	63	98	1
26	46	99	1
27	36	101	1
28	38	105	1
29	22	107	1
30	29	109	1
31	28	114	2
32	23	116	2
33	18	127	1
34	17	128	1
35	9	139	1
36	11	150	1
37	11	152	1
38	4	156	1
39	12	158	2
40	10	160	1
41	6	168	1
42	2	250	1
43	9	258	1
44	5	356	1
45	6	362	1
46	6	365	2
47	3	493	1
48	9	501	1
49	7	743	1
		744	1

RESULTS AND DISCUSSION

Frequency distribution of decamers within the sample of mammalian sequences

The frequency distribution of all possible decanucleotides in the mammalian sample is given in Table II. The mean occurrence is 1.53 and its variance is 7.60; the ratio variance/mean is 4.98, much larger than unity; thus the Poisson distribution is excluded. The distribution of the frequency of occurrence is represented in Fig. 1 in logarithmic scale (base 2). Unity is added to all occurrences, to include the zero class in the diagram. The distribution shows a striking regularity of shape, up to 16–20 occurrences, then it has a long and irregular tail. There are 2216 decamers with frequency higher than 16, and these constitute our first set. The set of 2216 corresponds to about 1 in 473 of all possible decamers and to 1 in 274 of the observed ones.

Table III. The 56 decanucleotides which occur 60 or more times in the 1601 kilobases of 579 sequences of mammalian DNA selected from GenBank(R) 55.0

tctatctatc	60	gacaggggtg	101
gctgctgctg	60	gtgtggggac	105
ttttttttc	61	tgtggggaca	107
gaggaggagg	61	aggggtgtgg	109
cttccttcct	61	gtggggacag	114
cggggcgggg	61	ggtgtgggga	114
taaaaaaaaa	62	gggtgtgggg	116
tctgtctgtc	63	ggggtgtggg	116
caaaaaaaaa	63	ggcacaggca	127
ctgtctgtct	63	gcacaggcac	128
agaaagaaag	71	aggcacaggc	139
gaaagaaaga	71	tggggacagg	150
atttatttat	75	caggcacagg	152
aaaaaaaaag	76	ggacaggggt	158
tctttctttc	76	cacaggcaca	158
ttatttattt	77	ggggacaggg	160
ctttctttct	78	acaggcacag	168
tttatttatt	81	agagagagag	250
aaagaaagaa	82	gagagagaga	258
aagaaagaaa	82	acacacacac	356
ctttttttt	82	tctctctctc	362
atatatatat	87	cacacacaca	365
tttctttctt	87	ctctctctct	365
ttctttcttt	88	tgtgtgtgtg	493
tatttattta	90	gtgtgtgtgt	501
tatatatata	97	aaaaaaaaaa	743
acaggggtgt	98	ttttttttt	744
caggggtgtg	99		

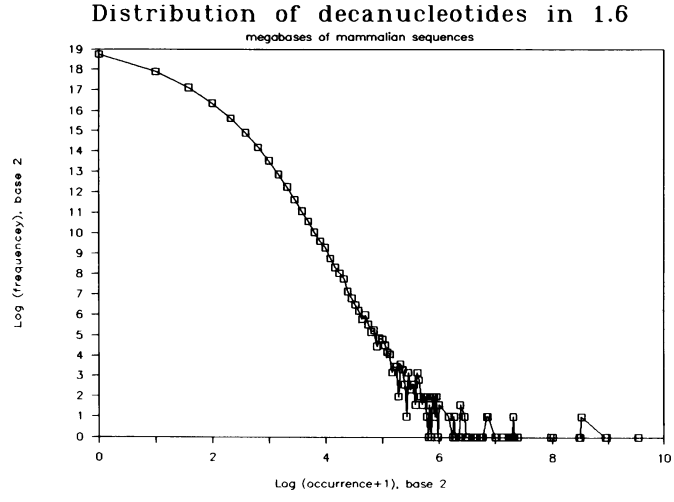

Figure 1. The distribution of the frequency of the occurrences of decanucleotides is represented in bi-logarithmic scale in the base 2. One unity was added to each occurrence to permit representation of the zero class of occurrence in the logarithmic scale. Note the regularity of the log-log distribution up to occurrences of 16–20.

The 56 most frequent decamers, which occur 60 or more times, are listed in Table III. Note the vast excess of deca-A and deca-T, and of the overlapping contigs poly-AG poly-GA, poly-AC poly-CA, poly-CT poly-TC, and poly-GT poly-TG. The last two were also the most frequent in animal viruses (13), indicating some correlation between infectors and their hosts also for higher order oligonucleotides (16).

Identification of a set of decamers enriched in signal sequences

We calculated two variance/mean chi squares for each parent decamer, one including both parent and progeny, with 30 degrees of freedom, and the other including the progeny only, with 29 degrees of freedom.

Plotting in a scatter diagram the two chi squares for each of the 2216 parent decamers practically gave a hardly readable L-

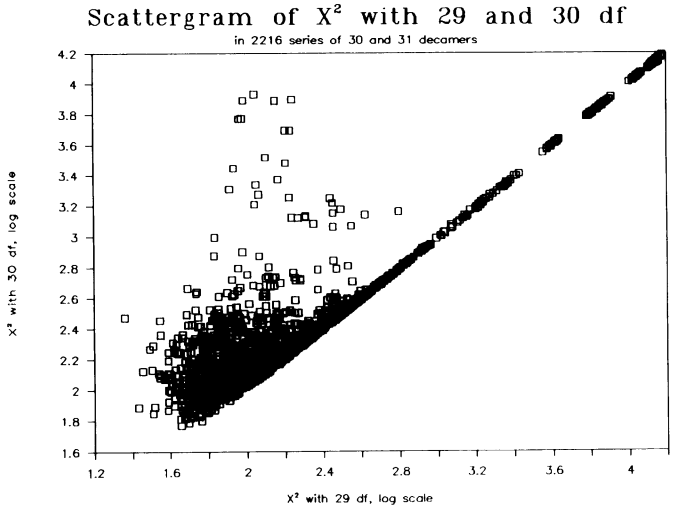

Figure 2. Scatter diagram of the chi squares with 29 and 30 degrees of freedom calculated for each of the 2216 decanucleotides with occurrence of 16 and above. Abscissa, chi square for progeny only; ordinate, chi square for progeny plus parent. Log scale.

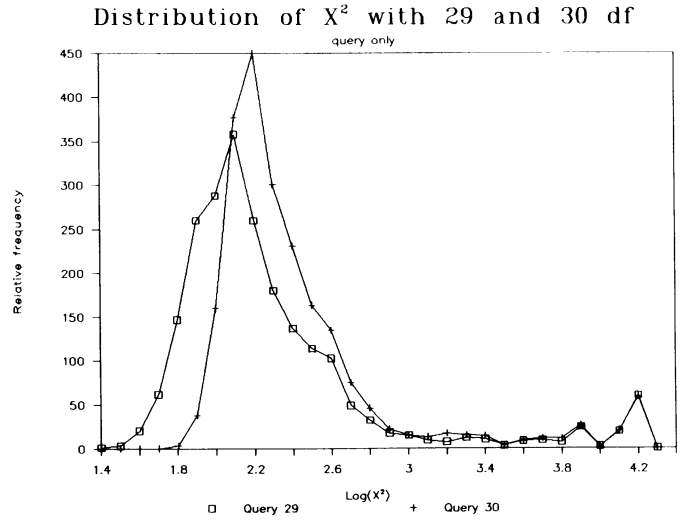

Figure 3. Frequency distributions of the chi squares with 29 and 30 degrees of freedom. The distributions for the query of 2216 decanucleotides are separated, but not as widely as in Viruses (ref. 13). The abscissa is in logarithmic scale to compact the drawing.

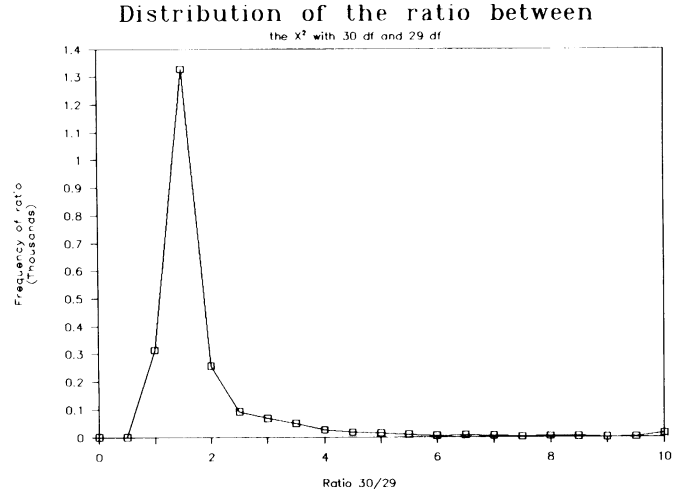

Figure 4. Frequency distribution of the ratio between the chi square with 30 and the chi square with 29 degrees of freedom in the query of 2216 decanucleotides. Out of 2216 ratios in the query, 346 or the 15.6% are greater than 1.9

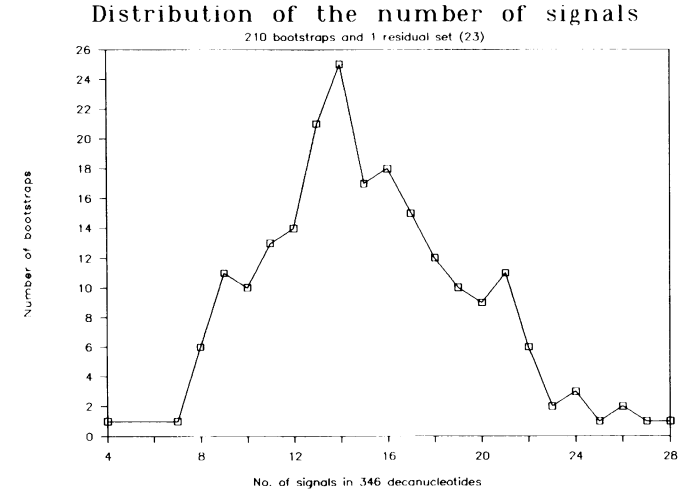

Figure 5. Frequency distribution of the number of signals in 210 bootstraps of 346 decanucleotides and in the residual set. The number of signals, 23, in the residual set is 1.82 standard deviations above the mean of the bootstraps, with empiric probability P=0.0477.

shaped graph, due to the large size of a few chi squares. Then, we logarithmically transformed the chi square values and, with some surprise, obtained a graph (Fig. 2) which is almost superimposable to the one obtained for viral decamers in the linear scale (13). Apparently a large fraction of the decanucleotides had almost equal chi squares with 29 and 30 df, but there was also a sizable fraction showing altered behaviour. This is indicated by the points above the diagonal in Fig. 2.

The distributions of the chi squares with 30 and 29 degrees of freedom are plotted in Figure 3, where the abscissa is in logarithmic scale. Since the two distributions were considerably separated by the single degree of freedom (consider that the scale is logarithmic), we decided not to compare them to bootstrap distributions, but to proceed directly to the study of the ratio between the chi square including the parent (30 df) and the chi square without it (29 df). In Fig. 4, the distributions of the ratios for the sample of 2216 decamers are plotted. It is apparent that, although most decanucleotides in the query set do not show an anomalous behaviour, there is a group with large 30/29 ratios which demands further study.

In analogy with our procedure with viruses we obtained the residual set using the ratio of 1.9 as a threshold. We accepted then the 346 decanucleotides with a 30/29 ratio of 1.9 or greater. A cursory reading of this residual set allowed identification of binding sites such as TGGGGA, GGGGCGGG and others.

Finally, to test the efficiency of the procedure in producing an enrichment for transcription signals, we compared the set of 346 decanucleotides with a sample of eukaryotic signals (Table

Table IV. Eukaryotic transcription regulating signals of 5 to 10 bp in length according to Wingender (1988). Alphabetic order.

aacccccc	cccccgcccc	gtcaaa
aagataagg	cccgcc	gtgacga
aagatggc	ccgccc	gtgacgt
aagggcgc	ccgcccg	gtggaaa
acagctg	ccgcccgcgg	gttttaa
accacctg	ccgtgcgg	tagaaca
accccgccca	cctgc	tataaaa
accctgctt	ccttaagagt	tcccca
actgaccgc	cgtgac	tgacgtca
actgttct	ctgagtgca	tgactaa
agaaatg	ctgggga	tgcacaccg
agaacagatg	ctgtaca	tgcgcccgg
agaacat	ctttgcat	tgcgctcgg
agccaat	gaacag	tgcgctcgt
aggggacg	gaagggaaaa	tgctttgcat
aggtaag	gaggagggg	tgggcgggac
agtccttt	gaggcgtggg	tgggcgtggt
agtgttct	gatgtcc	tgggga
agttgc	gattgg	tggtttg
atgaaaat	gatttc	tgtcct
atgagtcaga	gccaac	tgtggaaa
atgcaaat	gccacctg	tgtggtat
atgcaaatag	gccccgcccc	tgttct
attcctctgt	gcgccacc	tgttctcc
attgg	gcgcgaaa	tgtttgct
atttgcat	gctccgcctc	ttagtcag
atttgtat	ggacatc	tttcca
attttgtaat	gggactttcc	tttcgcgc
caaccac	gggcgg	
cagctggc	ggggagtggc	
caggata	ggggcggagc	
caggtggc	ggggcggcgc	
catgtggc	ggggcggg	
ccaat	ggggcggggc	
ccacacccg	gggggcgggc	

Table V. Data from the search in 346 random decamers of the signals listed in Table IV. One residual set and 210 bootstraps.

Signal length	5	6	7	8	9	10	Total
Bootstraps	1346	1393	376	76	6	1	3198
Residual set	3	12	1	6	0	1	23

Signals of length 9 and 10 pooled for the calculation of the X^2

1. Chi square for equality of frequencies of signals between the bootstrap and the residual set:
$X^2_{[4]}$ = 72.162, P vanishing
2. Chi square for equality of the fraction of signals in the 23/346 decamers of the residual set and in the 3198/72660 decamers of the bootstraps:
$X^2_{[1]}$ = 4.119 , P = 0.042
3. Chi square with 1 df testing for linear (17) increase of the frequency of the longer signals in the residual set:
$X^2_{[1]}$ = 26.128, P vanishing

IV) from the compilation by Wingender (1). In this table, there are 98 signals listed alphabetically. In the same table published in (13) there are four repetitions, and the present one, alphabetically ordered, corrects that error.

Out of 346 decamers, 18 matched 23 transcription signals. To evaluate the significance of this result, we bootstrapped 210 series of 346 decanucleotides and counted the compilation (1) signals present in these bootstraps.

On the average there were 15.2 signals per bootstrap, with a standard deviation of 4.29 signals. This locates the residual set, with its 23 signals, 1.82 standard deviations above randomness (Fig. 5). Out of 210 bootstraps, 10 had 23 or more signals, namely 23 or more signals were seen with a probability of 0.0477, which makes the excess of signals in the residual set significant. A slightly higher level of significance is obtained when one compares the 23/346 signals in the residual set to the 3198/72660 in the bootstrap. The X from the 2×2 table is 4.12 with 1 df, with P=0.042.

The numbers of signals of increasing length in the bootstrap and in the residual set are compared in Table V. Longer signals

Table VI. The set of 346 mammalian decamers highlighting the matches to the sample of 23 transcriptional elements, according to Wingender(1). The individual frequency is given at the right of the decamer.

aaaaaaaaaa	743	aggcacaggc	139	ttacagatga	20	tgggattaca	39
aaaaattagc	28	aggcccacca	19	ttacaggcat	19	tgggatgaga	16
aaaattagcc	20	agcagaaatc	17	tttatttatt	81	TGGGGAcagg	150
aaatacttcc	17	agcagatgtg	19	tttagtagag	22	tgggggtggg	37
aaattagcca	16	agcactttgg	22	ttttagtaga	24	tggggcacag	34
aaatcatgag	17	agccaaagtc	20	tttttagtag	25	tgcaaatgca	16
aaagtgctgg	31	agccatctct	24	ttttttttttt	744	tgcagtgagc	30
aaagtcacag	22	agcctgggca	34	tttttttgag	29	tgcactgtga	18
aaactgaggc	21	agcccttgtg	18	ttttttgaga	33	tgcactccag	27
aatacttcca	16	acaaacaaac	29	ttttgagaca	18	tgctgggatt	40
aattagccag	19	acaaggacca	18	tttgagacag	19	tgcctgggag	22
aatgaggact	18	acaacaacaa	27	tttgggaggc	34	tcaagccatc	16
aatcccagca	42	acagtgagga	23	tttcaccatg	17	tcattcattc	30
aagaaagaaa	82	acaggggtgt	98	tttctttctt	87	tcagagcaga	23
aagtgctggg	39	acaggggtcc	37	ttgttgttgt	34	tcaggctcag	30
aagtcaagga	16	acagggcacag	168	tttgtggggca	33	tcaccatgtt	17
aagtcacagg	25	acagccaaag	25	ttgtcagagc	17	tctatctatc	60
aaggaaggaa	50	acacacacac	356	ttgcagtgag	21	tctactaaaa	26
aaggttaggt	16	actaaaaata	26	ttcaacaaac	16	tctttctttc	76
aagccacaag	16	actttgggag	31	ttcattcatt	33	tcttcagcac	21
aacaacaaca	30	acttccaaag	16	ttcaccatgt	18	tctgtctgtc	63
aactcctgac	20	actgcactcc	27	ttctttcttt	88	tctcaaaaaa	30
atatatatat	87	actcatctgc	16	ttctCCTGCc	43	tctctactaa	23
attattatta	46	actccagcct	35	ttccttcctt	48	tctctctctc	362
atttatttat	75	acggcttgtg	25	tgaatgacct	17	tctccagggg	22
attcattcat	29	acgccacact	26	tgaaggacag	22	tctCCTGCct	45
attctCCTGC	31	accactgagc	18	tgagaacttc	18	tcggctcact	18
atgtatgtat	23	accttccaga	31	tgaggcagga	30	tccttccttc	46
atgctaaagg	16	taatcccagc	46	tgtaatccca	32	tcctgagaac	18
atctatctat	56	tatatatata	97	tgtattttta	28	tcccaaagtg	32
atcccagcac	31	tattattatt	56	tgttgttgtt	36	tcccagcact	33
agaaagaaag	71	tatttattta	90	tgtgagccac	20	tcccagctac	31
agagagagag	250	tatgtatgta	23	tgtgtgtgtg	493	tcccctccctc	41
agaccttcaa	29	tatgctaaag	16	tgTGGGGAca	107	gaaagaaaga	71
agttccagaa	26	tatctatcta	53	tgtggggcac	30	gaaggaagga	41
agtgcagtgg	28	tacaggcatg	24	tgtgcctggg	25	gaaccttcca	18
agtgctggga	40	tactaaaaat	27	tgtctgtctg	58	gattacaggc	39
aggaaggaag	44	ttattattat	52	TGTCCTccat	17	gatggatgga	31
aggactcatc	17	ttatttattt	77	tggatggatg	30	gagagagaga	258
AGGGGACGcc	27	ttagtagaga	20	tggagtgcag	27	gagaggtgcc	17
aggggtgtgg	109	ttaggttgta	16	tggtgaaacc	21	gagtgcagtg	27
gaggaggagg	61	ggtgTGGGGA	114	gcGGGGCGGG	49	ctactaaaaa	26
gaggttgtca	17	ggtggctcac	23	gcggcggcgg	40	ctttgggagg	33
gaggcaggag	39	gggattacag	40	gccaaagtca	18	ctttctttct	78
gaggctgagg	52	gggagggagg	46	gccactgcac	20	cttgtggggc	31
gacaggggtg	101	gggacagggg	156	gcctgtaatc	26	cttcagcaca	20
gacggcttgt	25	gggacggctt	19	gcctgggagg	23	cttctctcca	34
gacgccacac	26	gggacgccac	25	gcctcccaaa	34	cttccttcct	61
gtaatcccag	34	gggtttcacc	19	gccgccgccg	40	ctgagaactt	22
gtatttttag	24	gggtgtgggg	116	caaagaatca	16	ctgaggcagg	37
gtagctggga	28	ggggacaggg	160	caaagtgctg	31	ctgtaatccc	33
gtttgagacc	20	ggggacgcca	25	caagaaggtg	23	ctgtggtggt	21
gtttcaccat	18	ggggtttcac	17	caaggaggca	21	ctgtctgtct	63
gttgttgttg	25	ggggtgtggg	116	caagcgattc	16	ctggagtgca	31
gttgtcagag	19	ggggcacagg	37	caacaacaac	22	ctgggattac	38
gttgcagtga	16	GGGGCGGGGC	56	cattcattca	28	ctgggaggtt	19
gtgaaacccc	16	GGGCGGGggcg	41	catgaagatc	16	ctgcactcca	32
gtgatctgcc	20	ggcacaggca	127	catggcaaga	16	ctcaaaaaaa	26
gtgagccacc	25	ggctaatttt	23	catctctgtc	22	ctcaagccat	16
gtgtgtgtgt	501	ggcttgtggg	31	catcctggct	32	ctcaggacct	16
gtgTGGGGAc	105	ggcttcatga	16	cagaaatcat	21	ctcaggctca	30
gTGGGGAcag	114	ggctgaggca	32	cagttccaga	24	ctcagcctcc	49
gtggggcaca	33	ggctgtgagg	29	cagtgaggag	30	ctcactgcaa	25
gtggctcaca	21	ggctcactgc	25	caggaggccc	24	ctctactaaa	28
gtgcagtggc	29	ggcgGGGCGG	38	cAGGGGACGc	28	ctcttcagca	21
gtgctgggat	40	ggcggcggcg	45	caggggtgtg	99	ctctctctct	365
gtctgtctgt	54	ggccaaggtg	19	caggcacagg	152	ctcggctcac	18
gtcCTGGGGA	31	ggcctcccaa	30	cagcagatgt	20	ctcctgacct	29
gtcctccatg	17	gcaagaaggt	19	cagcagcagc	44	ctcccaaagt	33
ggaaagccac	16	gcacaggcac	128	cagcactttg	26	ctccctccct	54
ggaaggaagg	45	gcactttggg	24	cagccaaagt	23	cgtggagtgg	20
ggattacagg	43	gcactccagc	27	cacaggcaca	158	cGGGGCGGGg	61
ggagtgcagt	23	gctaattttt	25	cacagccaaa	25	cggcttgtgg	30
ggaggaggag	59	gcttgtgggg	29	cacacacaca	365	cggcggcggc	40
ggaggttgtc	18	gctgaggcag	39	cacaccttct	17	cgccacactc	24
ggagggaggg	43	gctggagtgc	28	cactttggga	28	cgccgccgcc	48
ggacaggggt	158	gctgggatta	49	cactgtgaca	16	ccaaagtgct	27
ggactcatct	18	gctgctgctg	60	cactgcactc	25	ccaaagtcac	18
ggacggcttg	25	gctcaggctc	30	caccctgagg	20	ccaagaccta	16
ggacgccaca	28	gctcacacct	19	ctaaaaatac	23	ccatgaagga	23
ggtgaaaccc	17	gctcactgca	29	ctatctatct	55	ccatctctgt	25
ccatcctggc	34	cctgacctca	18	ccgccgccgc	47		
ccAGGGGACG	25	cctgtaatcc	36	cccaaagtgc	28		
ccagcacttt	28	cctgggcaac	34	cccaggctgg	37		
ccagctactc	23	cctccaggaa	26	cccagcactt	28		
ccagcctggg	45	cctcctcctc	58	cccagctact	35		
ccacacagct	16	cctcccaaag	30	ccctccctcc	47		
ccactgcact	25	cctccctccc	57				
ccttccttcc	58	ccgggacggc	24				

are more common in the residual set than in the bootstrap. The chi square for increase in frequency of longer signals is $X^2_{[1]}$=26.128, which is very highly significant (17).

The residual set of mammalian decamers is enriched in signals.

The enrichment is significant at the bootstrap. The enrichment, however, is in no way comparable to the one obtained for mammalian viruses (13). The number of signals filtered through our deterministic procedure in the residual set is lower than, or equal to, that found in 10 over 210 random sets. However, one point makes the result of some interest: in the bootstraps there is an excess of the shorter signals, and a deficiency of the longer signals, in comparison to the residual set. This difference excludes the residual set from the bootstrap distribution ($X^2_{[1]}=26.128$, P vanishing). Other signals may be present in the remaining 328 decamers, and it may be worth exploring them biochemically.

In Table VI we list all decanucleotides in the residual set, and highlight the signals they contain according to Wingender's compilation. Overlapping signals within a decamer are not highlighted. The most frequent signal containing decamer, occurring 150 times, was TGGGGAcagg. It should be emphasized that some eukaryotic signal sequences, TGGGGA as well as CCTGC, are present in several different decamers. In the 23 matches we found several signals which occured several times. Furthermore Sp1 signals overlap a few time within the same decamer.

Thus the selection procedure we presented in (13) is also valid for mammalian DNA. However, it is apparent that the complexities of mammalian genomes permit only limited, albeit significant, enrichment of the residual set with transcription signals.

CONCLUSIONS

The results obtained show that our algorithm selects, from a sample of mammalian genes, a set of 346 decanucleotides which are significantly enriched in transcription signals. The yield of **known** transcription factors is 23/346, close to 7% of the sample. This is not as high as the enrichment of 12% which we found in animal viruses (13). Among the frequent decamers containing binding motifs for nuclear proteins, some are of interest, particularly those containing the Sp1 binding sites.

Our conclusion is that through the statistical analysis of the distribution of oligomers in mammalian sequences, specific oligonucleotides may be identified and tested biochemically for protein interaction and assessment of transcriptional activity.

We emphasize again that the method described in the present report only selects for DNA elements whose molecular function, through high sequence selectivity in their interaction with proteins, does not allow nucleotide substitution.

ACKNOWLEDGMENTS

This work was supported by MURST grants 60% and 40%, and National Research Council grants CNR.88.03586 and CNR.89.00853 to I.B.; R.G. is supported by Istituto Superiore di Sanita' (AIDS-90), by CNR PF Ingegneria Genetica and by AIRC. We wish to dedicate this work to the VI Centennial Jubilaeum of the University of Ferrara, 1391−1991.

REFERENCES

1. Wingender E. (1988) Nucleic Acids Res. 16: 1879−1889
2. Boss, J.M. and Strominger, J. (1986) Proc. Natl. Acad. Sci. USA 83: 19139−19144
3. Sullivan, K.E., Kalman, A.F., Nakanishi, M., Tsang, S.Y., Wang, Y., and Peterlin, B.M. (1987) Immunol. Today, 8: 289−292
4. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A.(1980) Nucleic Acids Res. 8: r49−r62
5. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res. 9: r43−r74
6. Barrai, I. (1983) J. Theor. Biol. 104: 633−645
7. Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Nucleic Acids Res. 11: 2205−2220
8. Sadler, J. R., Waterman, M.S. and Smith, T.F. (1983) Nucleic Acids Res. 11: 2221−2231
9. Claverie, J.-M. and Bouguerelet,L. (1986) Nucleic Acids Res. 14: 179−186
10. Volinia, S., Bernardi F., Gambari, R. and Barrai, I. (1988) J. Mol. Biol. 203: 385−390
11. Volinia, S., Gambari, R., Bernardi, F. and Barrai, I. (1989) CABIOS, 5: 33−40
12. Seto, M.H., Brunck, T.K. and R.L. Bernstein et al. (1989) Nucleic Acids Res. 17: 2783−2800
13. Volinia, S., Scapoli, C., Gambari, R., Barale, R., and Barrai, I. (1991) Nucleic Acids Res. 19: 3733−3740
14. Williams, J.L., Garcia, J., Harrich, D., Pearson, L., Wu,F.,and Gaynor, R. (1990) EMBO Journal 9: 4435−4442
15. Kim, S., Ikeuchi, K., Groopman, J., and Baltimore, D. (1990) J. Virol. 64: 5600−5604
16. Barrai, I., Scapoli, C., Barale, R. and Volinia, S. (1990) Nucleic Acids Res. 18: 3021−3025
17. Cochran, W.J. (1954) Biometrics, 10: 417−451.