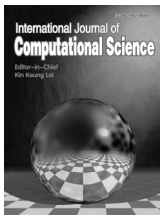
International Journal of Computational Science
1992-6669 (Print) 1992-6677 (Online) www.gip.hk/ijcs



Seeking Significant Oligomers via Set Partitions Expected Count

Stephen Sauchi Lee*

Department of Statistics, College of Science, University of Idaho, Moscow, ID 83844-1104, USA
stevel@uidaho.edu

Abstract. In order to determine significance of word counts of DNA sequences, it is of first importance to develop a baseline comparison so that the non-randomness of the observed word count can be measured. We developed a novel measure of oligomer expected count using the concept of set partitions. This expected count provides a baseline reference to reveal non-random DNA sequences. Non-randomness of oligomers is evaluated in terms of the amount of deviation from the derived expected count. As a consequence, the ratio of the observed count to the expected count will indicate the degree of under- or over-representation of the oligomers. The usefulness of the method is demonstrated when applied to two human chromosomes and an artificially generated random chromosome. Under- and over-represented oligomers are revealed in the human chromosomes but not in the random chromosome.

Keywords: Bell number, expected value, set partition.

1 Introduction

Analyses of biological sequences have been given much attention since the availability of massive accumulation of genomic data and the accessibility of super computing power.

Using a version of a pattern-discovery algorithm Rigoutsos *et al.* developed earlier [11], they sought variable-length motifs in the human intergenic and intronic regions that comprise a minimum of 16 nucleotides and appear a minimum of 40 times. The result is a set of 127,998 patterns, termed pyknons, which the authors found to be enriched in a statistically significant manner in

* Corresponding Author. Email: stevel@uidaho.edu.

genes involved in specific biological processes [12]. Hampikian *et al.* designed an algorithm [5] for identifying absent DNA sequences, called nullomers, and explore its usage in species identification and characterization. Arnau and Marin implemented an algorithm that allows the exhaustive determination of words of up to 12 nucleotides in DNA sequences [3].

Words that are, by some measure, over- or under-represented in the genome have been linked to biological relevance, meanings, and functions. These words are polynucleotide patterns, or oligomers (called k-mers in short below), which occur unexpectedly often or rare. Many works are produced based on either an underlying random model or a Markov model. A random model is a simple probabilistic model in which sequences are produced by a random source emitting nucleotides independently one by one according to a given stationary distribution. A Markov model assumes that the probability of the next nucleotide given all prior nucleotides history is the same as the probability given only the most recent prior m nucleotides. This is a one-step ahead (i.e., one nucleotide ahead) Markov model of order m. The random model is a special case of a one-step-ahead Markov model of order 0.

Apostolico *et al.* took the global approach of annotating a suffix tree with the expected values and variances based on the random model in [1], with the intent to filter patterns with unexpected occurrences for further scrutiny. They then later extended the result based on a Markov model of order one in [2] , and developed a software suite, named Verbumculus, for discovering unusual words. Phillips *et al.* [9] developed methods to test for accuracy in predicting observed frequencies of 2-mers through 6-mers in the E. coli DNA. They found that a Markov chain was more accurate than a random model based on independent mononucleotide. Cuticchia *et al.* [4] predicted the higher order oligonucleotide frequencies in Drosophila melanogaster DNA by a 3rd order Markov chain, a one nucleotide ahead Markov model based on the immediate past tri-nucleotides frequencies.

Statistical properties of words have been of considerable interest in the analysis of the genome. Reinert *et al.* [10] provided a quick overview on this subject. They claimed that to determine the statistical significance of a genomic word in a DNA sequence is more important than to know where they occur and how many times do they occur. Leung *et al.* [6] and Schbath [14] presented and compared different statistics to find over- and under-representation of words in DNA sequences, based essentially on the comparison of the observed count of a word in the sequence, with its expected count computed under a Markov chain model. The literatures published many works on one-step-ahead Markov chains with order m, and emphasis was placed on finding the order m given the data. A better reflection of reality would be to relax this assumption to a p-step ahead Markov model of order m, or even more flexibly to discrete combinations or continuous mixing of these Markov models with non-constant p and m values.

In order to determine the statistical significance of word count of DNA sequences, it is of first importance to develop a baseline comparison so that the non-randomness of the observed word count can be measured. This is the motivation for the present work of this paper. In searching for non-random words, our attempt is to compute a novel measure of expected count through the concept of set partitions. This will provide a baseline foundation to reveal non-random DNA se-

quences or subunits. Oligomer non-randomness is then measured, not only in terms of its absolute observed count of occurrences, but in terms of its relative amount of deviation from this expected value. When the observed count of a certain *k*-mer is compared to its expected count, non-random words will be revealed through the ratio *O/E*. This ratio can be evaluated for all *k*-mers. Ratios located at the extremes will correspond to non-random oligomers which are either under- or over-represented.

This paper is organized as follows. Section 2 formulates the methodology of computing expected value via set partitions. The method is evaluated in Section 3 when applied to two human chromosomes and an artificially generated chromosome. Section 4 concludes with remarks.

2 Methodology

We developed an expected count for a *k*-mer assuming it is formed randomly from its partition components. There are B_k*-1* ways to partition a *k*-mer into lower order components, where B_k is the *k*-th Bell number [13]. These partition components may or may not be contiguous. In combinatorics, the *k* th Bell number, B_k , named in honor of Eric Temple Bell, is the number of partitions of a set with *k* members.

The number of set partitions of a *k*-mer with exactly *q* $(1 \leq q \leq k)$ partition components is the Stirling number of the second kind and is denoted *S(k,q)*. It is the number of ways of partition a *k*-mer into exactly *q* partition components. It is known that $B_k = \sum_{q=1}^{k} S(k,q)$ and $S(k,q) = qS(k-1,q) + S(k-1, q-1)$. The recurrence relations form the basis of recursive algorithms for generating set partitions for any *k*-mers, starting from $k = 1$.

In a genome of length *n*, let us denote a certain *k*-mer by *x*. Consider a particular set partition of *x*, denoted by *(i)*. The expected count of *x* formed randomly through this partition is

$$E_{(i)}(x) = n \prod_{\substack{W:W \\ is \\ a \\ partition \\ component \\ in \\ set \\ partition \\ (i)}} p(W),$$

where *p(W)* is the probability of observing *W* in the genome, which is approximated by *#(W)/n* with # denotes the number of occurrence of *W*.

For example, there are 14 non-trivial partitions for a 4-mer *ABCD*. Consider a particular set partition *(i) = {A, B.D, C}*. Note that the partition component *B.D* is not contiguous; the **.** in between represents any nucleotide. The expected count of *x* formed randomly through this partition is

$$E_{(i)}(x) = E_{(i)}(ABCD) = n\ p_A\ p_{B.D}\ p_C$$

where *n* is the number of nucleotides in the genome, p_A equals *#(A)/#(.) = #(A)/n*, $p_{B.D}$ *equals #(B.D)/#(...) = #(B.D)/(n-2) ≈ #(B.D)/n, and* p_C *equals #(C)/#(.) = #(C)/n.*

We compute all expected values under each of these B_k *-1* partitions. The overall expected value is a single measure summarizing all these expected values. Weighted average will be prefer if we have some kind of prior knowledge regarding which partition are more important or more-likely. Since there is no prior knowledge to favor one partition over the other, we treat all of them equal likely. The most nature way to aggregate these expected values is to compute the arithmetic average of all of these B_k *-1* expected values, i.e.,

$$E(x) = \sum_{i=1}^{Bk-1} E_{(i)}(x) / (B_k - 1)$$

The ratio *O(x)/E(x)* where *O* and *E* are, respectively, the observed and expected counts of the *k-mer, x,* will reveal its non-randomness. The further away from 1.00 on either side will indicate the degree of over- or under-representation of the *k*-mer.

3 Application

To evaluate the usefulness of the methodology, we applied the above methodology to the human genome chromosomes 1 and 19, version hg18 at the UCSC genome databases. As a control for comparison, we generate an artificial genome which is totally random, but have the same uni-nucleotide marginal distributions as the human genome. We chose chromosome 1 in the human genome because it is the longest human chromosome. Chromosome 19 was chosen because of all the human chromosomes, its uni-nucleotide percentages are closest to an equal distribution of ¼.

We computed the expected values and the corresponding ratios on all di-mers, tri-mers, tetra-mers, penta-mers, and hexa-mers, i.e., from *k*=2 to 6. We plot them on the following figures, with ratios shown in logarithmic scale. Human chromosomes 1 and 19 were plotted with symbols + and o, respectively. The random chromosome which serves as a control was plotted with symbol x. The amount of deviation from the horizontal line with value equals 1.00 indicates the magnitude of non-randomness of the oligomers. The observed to expected ratios are plotted against the alphabetical order of the *k*-mers (e.g., AA, AC, AG, AT, … in 2-mers). This allows a visual presentation to compare the ratios for the same *k*-mer in all 3 chromosomes. Another way to present the results which will show how the ordered ratios profile deviate from 1.00 for the entire chromosome is to plot the ratios from the smallest to the largest. The forms a ratio profile for the 3 chromosomes. Both graphs are presented side by side as follows.

In all *k*-mers with *k*=2 to 6, the amount of deviation from 1 is increasing for the two human chromosomes, while the random chromosome remains close to the horizontal line of 1.00. It means that under- and over-represented oligomers are revealed in the two human chromosomes but not in the random chromosome.

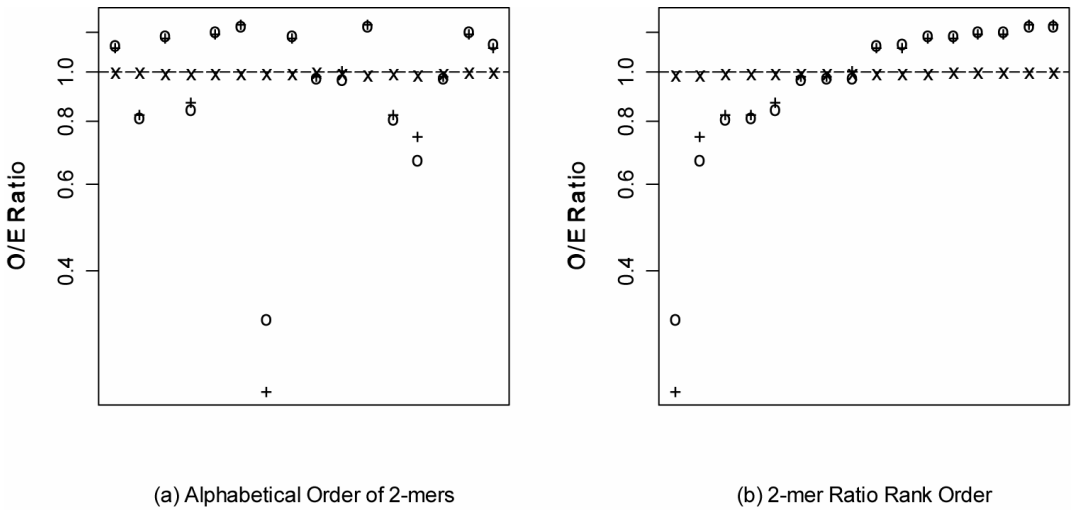


Fig. 1. All $4^2 = 16$ di-mers observed-to-expected ratios for human chromosome 1 (+), human chromosome 19 (o), and artificial random chromosome (x) are plotted (a) in alphabetic order (i.e., from AA, AC, AG, AT, and so on); (b) in ratio ranked order (i.e., from smallest ratio to largest). The y axis is in logarithmic scale.

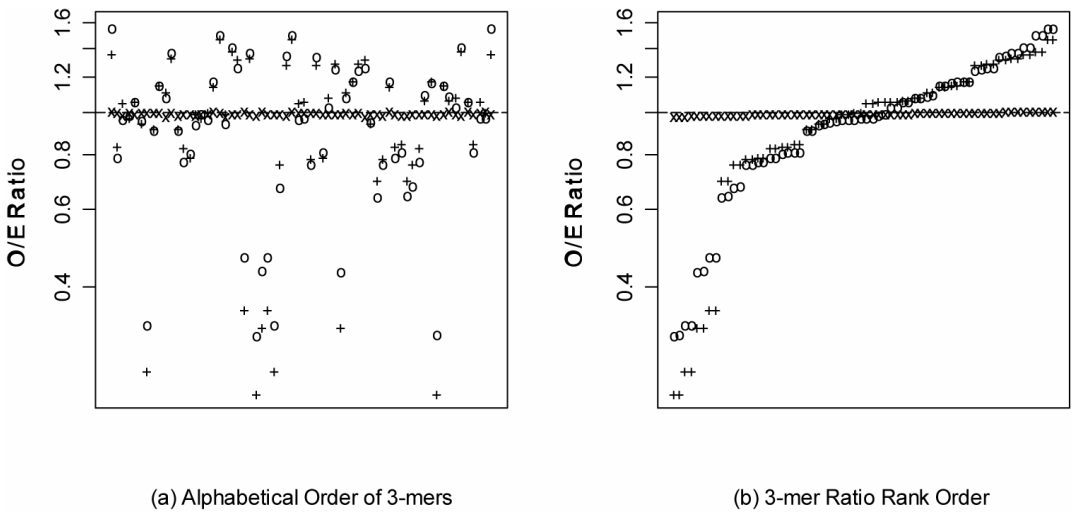


Fig. 2. All $4^3 = 64$ tri-mers observed-to-expected ratios for human chromosome 1 (+), human chromosome 19 (o), and artificial random chromosome (x) are plotted (a) in alphabetic order (i.e., from AAA, AAC, AAG, AAT, and so on); (b) in ratio ranked order (i.e., from smallest ratio to largest). The y axis is in logarithmic scale.

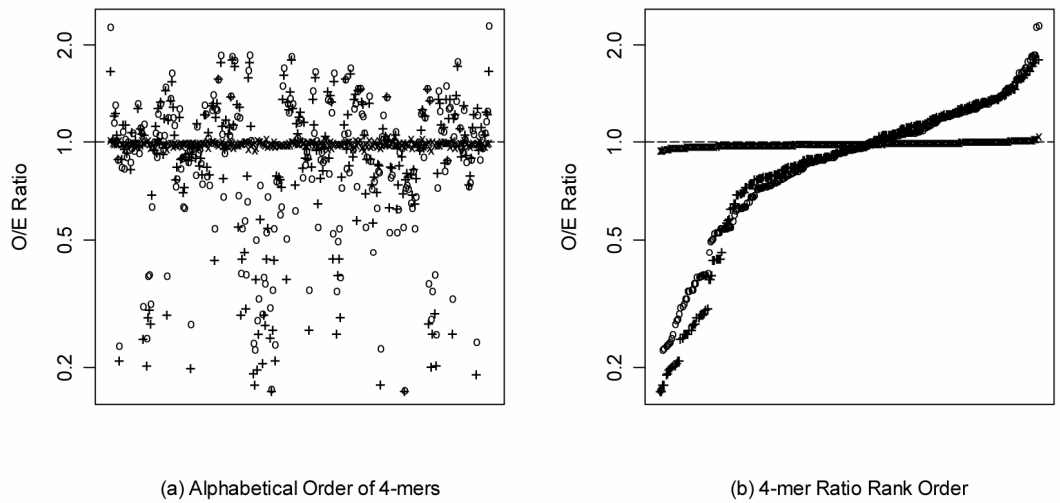


Fig. 3. All 4^4 = 256 tetra-mers observed-to-expected ratios for human chromosome 1 (+), human chromosome 19 (o), and artificial random chromosome (x) are plotted (a) in alphabetic order (i.e., from AAAA, AAAC, AAAG, AAAT, and so on); (b) in ratio ranked order (i.e., from smallest ratio to largest). The y axis is in logarithmic scale.

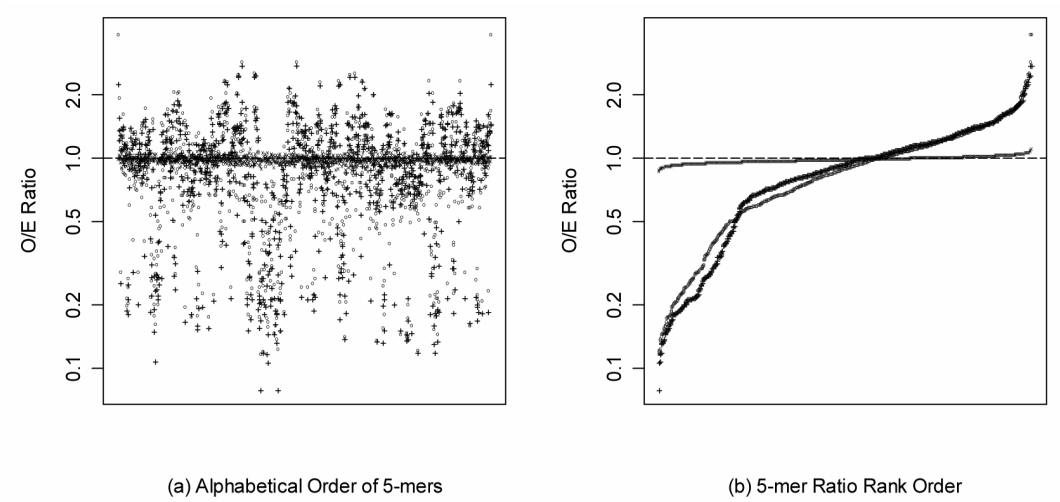


Fig. 4. All 4^5 = 1024 penta-mers observed-to-expected ratios for human chromosome 1 (+), human chromosome 19 (o), and artificial random chromosome (x) are plotted (a) in alphabetic order (i.e., from AAAAA, AAAAC, AAAAG, AAAAT, and so on); (b) in ratio ranked order (i.e., from smallest ratio to largest). The y axis is in logarithmic scale.

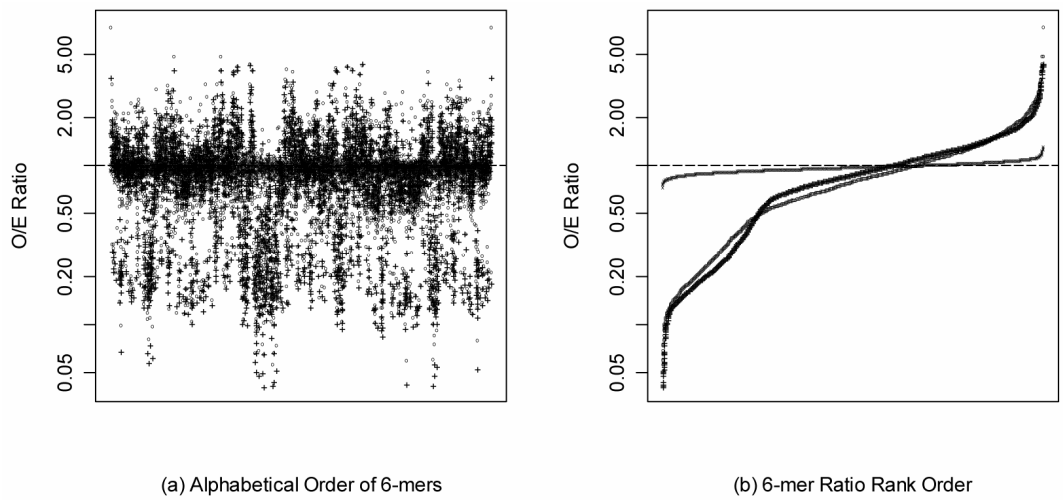


Fig. 5. All 4^6 = 4096 hexa-mers observed-to-expected ratios for human chromosome 1 (+), human chromosome 19 (o), and artificial random chromosome (x) are plotted (a) in alphabetic order (i.e., from AAAAAA, AAAAAC, AAAAAG, AAAAAT, and so on); (b) in ratio ranked order (i.e., from smallest ratio to largest). The y axis is in logarithmic scale.

Table 1. All 4^2 = 16 di-mers: observed counts, expected counts (round to whole number), and ratios in human chromosome 1 and 19 in ascending order of the ratios O/E.

O1	E1	2-mer	O1/E1	O19	E19	2-mer	O19/E19
2281713	9801335	**CG**	0.2328	1057112	3262218	**CG**	0.3240
14293281	19090643	**TA**	0.7487	2508619	3718437	**TA**	0.6746
11301414	13670294	**AC**	0.8267	2846196	3491858	**GT**	0.8151
11320734	13687638	**GT**	0.8271	2833985	3473898	**AC**	0.8158
16766098	19090645	**AT**	0.8782	3165323	3718437	**AT**	0.8513
13465343	13690047	**TC**	0.9836	3180474	3262219	**GC**	0.9749
13448102	13667887	**GA**	0.9839	3414357	3483369	**TC**	0.9802
9938922	9801347	**GC**	1.0140	3423863	3482364	**GA**	0.9832
21438891	19118230	**TT**	1.1214	4229414	3708327	**AA**	1.1405
21387422	19063097	**AA**	1.1219	4260148	3728575	**TT**	1.1426
16060830	13690040	**CT**	1.1732	4150572	3483368	**CT**	1.1915
16036982	13667879	**AG**	1.1733	4154304	3482364	**AG**	1.1930
16363107	13670285	**CA**	1.1970	4239115	3491858	**TG**	1.2140
16389034	13687629	**TG**	1.1974	4221129	3473897	**CA**	1.2151
12248730	9799616	**GG**	1.2499	4056078	3270169	**GG**	1.2403
12259077	9803067	**CC**	1.2505	4044958	3254288	**CC**	1.2430

Table 2. The 16 most under-represented and 16 most over-represented (from all $4^3 = 64$) tri-mers: observed counts, expected counts (each expected count is an overall average of B_3*-1 = 4* expectation counts from set partitions, then round to whole number), and ratios in human chromosome 1 and 19 in ascending order of the ratios O/E.

O1	E1	3-mer	O1/E1	O19	E19	3-mer	O19/E19
504568	2199578	**CGA**	0.2294	207608	664955	**CGA**	0.3122
505447	2202675	**TCG**	0.2295	208793	666636	**TCG**	0.3132
558646	2160289	**ACG**	0.2586	213446	646833	**ACG**	0.3300
560283	2163592	**CGT**	0.2590	214467	648514	**CGT**	0.3307
562063	1727259	**GCG**	0.3254	296265	679196	**GCG**	0.4362
562630	1727691	**CGC**	0.3257	296707	677966	**CGC**	0.4376
654232	1830570	**CGG**	0.3574	338330	719505	**CGG**	0.4702
655557	1831207	**CCG**	0.3580	338608	718275	**CCG**	0.4714
2506774	3566074	**GTA**	0.7030	503346	776985	**GTA**	0.6478
2508789	3566083	**TAC**	0.7035	501792	774385	**TAC**	0.6480
2889906	3774471	**TAG**	0.7656	568578	835529	**CTA**	0.6805
2890580	3774760	**CTA**	0.7658	571738	838421	**TAG**	0.6819
2163927	2737275	**GTC**	0.7905	613849	803171	**GAC**	0.7643
2161182	2733303	**GAC**	0.7907	616029	804578	**GTC**	0.7657
2972111	3733633	**GAT**	0.7960	698148	901275	**TAT**	0.7746
2972927	3733855	**ATC**	0.7962	697499	898615	**ATA**	0.7762
… intermediates omitted for clarity…							
3582893	3027832	**TCC**	1.1833	1044349	882327	**GGA**	1.1836
3579473	3022381	**GGA**	1.1843	1036848	875482	**CAC**	1.1843
3955885	3063200	**CTC**	1.2914	1048185	836427	**GGC**	1.2532
3951890	3057876	**GAG**	1.2924	1051639	834913	**GCC**	1.2596
2842554	2191545	**GGC**	1.2971	1164539	914750	**GGG**	1.2731
2845766	2192246	**GCC**	1.2981	1158760	909551	**CCC**	1.2740
3141630	2367598	**GGG**	1.3269	1229210	909924	**GAG**	1.3509
3148036	2369605	**CCC**	1.3285	1226789	907491	**CTC**	1.3518
4186728	3138284	**CCT**	1.3341	1266135	921047	**AGG**	1.3747
4179170	3132404	**AGG**	1.3342	1265801	919394	**CCT**	1.3768
8356466	6135638	**TTT**	1.3620	1287074	911342	**TGG**	1.4123
8333904	6111072	**AAA**	1.3637	1281789	904734	**CCA**	1.4168
4268753	3077969	**CCA**	1.3869	1423347	945637	**CTG**	1.5052
4273698	3081319	**TGG**	1.3870	1421113	943595	**CAG**	1.5061
4717054	3186229	**CAG**	1.4805	1681598	1076671	**AAA**	1.5618
4725132	3190990	**CTG**	1.4808	1699626	1086699	**TTT**	1.5640

Table 3. The 16 most under-represented and 16 most over-represented (from all $4^4 = 256$) tetra-mers: observed counts, expected counts (each expected count is an overall average of $B_4 - 1 = 14$ expected counts from set partitions, then round to whole number), and ratios in human chromosome 1 and 19 in ascending order of the ratios O/E.

O1	E1	4-mer	O1/E1	O19	E19	4-mer	O19/E19
94070	548131	**CGTA**	0.1716	24175	139820	**TACG**	0.1729
94289	548129	**TACG**	0.1720	24350	139911	**CGTA**	0.1740
78709	437436	**CGAC**	0.1799	35227	153038	**CGAC**	0.2302
79107	437918	**GTCG**	0.1806	35868	153693	**GTCG**	0.2334
120394	625570	**TTCG**	0.1925	37041	155753	**AACG**	0.2378
121593	623929	**CGAA**	0.1949	37483	156589	**CGTT**	0.2394
113194	563126	**CGAT**	0.2010	39578	164042	**CGAA**	0.2413
114163	563209	**ATCG**	0.2027	40085	164921	**TTCG**	0.2431
119312	584955	**TCGA**	0.2040	39840	158578	**ACGA**	0.2512
122606	594315	**ACGA**	0.2063	40902	159482	**TCGT**	0.2565
123320	596071	**TCGT**	0.2069	41183	148671	**ATCG**	0.2770
60756	288819	**CGCG**	0.2104	42168	148669	**CGAT**	0.2836
128242	598622	**CGTT**	0.2142	44885	156995	**TCGA**	0.2859
128027	596866	**AACG**	0.2145	40997	138415	**CGCG**	0.2962
115390	461911	**ACCG**	0.2498	49877	159782	**CGGT**	0.3122
116217	462502	**CGGT**	0.2513	50099	159308	**ACCG**	0.3145

… intermediates omitted for clarity …

1050679	672531	**AGGC**	1.5623	374106	235335	**GCTG**	1.5897
1053192	673763	**GCCT**	1.5631	372879	234551	**CAGC**	1.5898
1102816	691086	**GCTG**	1.5958	379343	229390	**AGGC**	1.6537
1102042	690314	**CAGC**	1.5964	381382	229677	**GCCT**	1.6605
1142776	709847	**GAGG**	1.6099	406124	243514	**GAGG**	1.6678
1146050	711356	**CCTC**	1.6111	407743	242635	**CCTC**	1.6805
1180110	718510	**CTCC**	1.6424	419797	244601	**CTCC**	1.7162
1179340	717172	**GGAG**	1.6444	422588	245871	**GGAG**	1.7187
3378917	2016710	**TTTT**	1.6755	438699	243803	**TGGG**	1.7994
3371285	2006585	**AAAA**	1.6801	437559	241722	**CCCA**	1.8102
1240905	711728	**TGGG**	1.7435	468203	250659	**CCAG**	1.8679
1241774	711273	**CCCA**	1.7458	470511	251750	**CTGG**	1.8690
1326541	742828	**CAGG**	1.7858	481498	256126	**CAGG**	1.8799
1329127	744109	**CCTG**	1.7862	482127	256245	**CCTG**	1.8815
1326768	727970	**CCAG**	1.8226	763886	332206	**AAAA**	2.2994
1330315	728958	**CTGG**	1.8250	777671	336564	**TTTT**	2.3106

Table 4. The 16 most under-represented and 16 most over-represented (from all $4^5 = 1024$) penta-mers: observed counts, expected counts (each expected count is an overall average of $B_5 - 1 = 51$ expected counts from set partitions, then round to whole number), and ratios in human chromosome 1 and 19 in ascending order of the ratios O/E.

O1	E1	5-mer	O1/E1	O19	E19	5-mer	O19/E19
6127	77266	**CGTCG**	0.0793	3803	31802	**CGACG**	0.1196
6121	77147	**CGACG**	0.0793	4487	36324	**CGATA**	0.1235
8472	79102	**CGCGT**	0.1071	4550	36424	**TATCG**	0.1249
8540	79001	**ACGCG**	0.1081	4037	31891	**CGTCG**	0.1266
9566	81242	**CGCGA**	0.1177	4622	33247	**GTACG**	0.1390
9687	81345	**TCGCG**	0.1191	4683	33187	**CGTAC**	0.1411
19025	158776	**TATCG**	0.1198	5034	34977	**GCGTA**	0.1439
19025	158564	**CGATA**	0.1200	5025	34871	**TACGC**	0.1441
15911	121024	**TCGAC**	0.1315	5282	36513	**TACGA**	0.1447
15108	113917	**CGTAC**	0.1326	5440	36637	**TCGTA**	0.1485
16098	120984	**GTCGA**	0.1331	5318	35761	**TAACG**	0.1487
15348	113902	**GTACG**	0.1347	5481	36354	**CGGTA**	0.1508
21634	159969	**TCGTA**	0.1352	5488	36304	**TACCG**	0.1512
21894	159743	**TACGA**	0.1371	5485	35879	**CGTTA**	0.1529
16522	118461	**TACGC**	0.1395	5808	37087	**TCGAC**	0.1566
16531	118423	**GCGTA**	0.1396	5884	37159	**GTCGA**	0.1583

… intermediates omitted for clarity …

336334	151881	**GAGGC**	2.2145	141942	60525	**GAGGC**	2.3452
337315	152236	**GCCTC**	2.2157	143827	60495	**GCCTC**	2.3775
362493	161380	**GCCTG**	2.2462	152004	63908	**GGCTG**	2.3785
362332	161146	**CAGGC**	2.2485	148629	62468	**TCCCA**	2.3793
1499045	665768	**TTTTT**	2.2516	150106	62999	**TGGGA**	2.3826
1495959	661766	**AAAAA**	2.2606	151904	63648	**CAGCC**	2.3866
371040	160479	**GGCTG**	2.3121	159088	62894	**CCAGC**	2.5295
371770	160428	**CAGCC**	2.3174	161032	63241	**GCTGG**	2.5463
386715	158990	**CCAGC**	2.4323	173184	67894	**CCTGG**	2.5508
412665	169434	**CCTCC**	2.4356	172972	67712	**CCAGG**	2.5545
387921	159139	**GCTGG**	2.4376	171413	66746	**GGAGG**	2.5681
412121	169030	**GGAGG**	2.4382	170911	66403	**CCTCC**	2.5738
420507	169824	**CCTGG**	2.4761	196927	67922	**CTGGG**	2.8993
419908	169551	**CCAGG**	2.4766	197067	67553	**CCCAG**	2.9172
470476	169800	**CCCAG**	2.7708	411373	105078	**AAAAA**	3.9149
471214	169952	**CTGGG**	2.7726	420829	106886	**TTTTT**	3.9372

Table 5. The 16 most under-represented and 16 most over-represented (from all $4^6 = 4096$) hexa-mers: observed counts, expected counts (each expected count is an overall average of $B_6 - 1 = 202$ expected counts from set partitions, then round to whole number), and ratios in human chromosome 1 and 19 in ascending order of the ratios O/E.

O1	E1	6-mer	O1/E1	O19	E19	6-mer	O19/E19
861	21224	**CGCGTA**	0.0406	403	6942	**CGTACG**	0.0581
839	20249	**CGTACG**	0.0414	456	7476	**TACGCG**	0.0610
903	21678	**TCGACG**	0.0417	484	7755	**TCGACG**	0.0624
890	21226	**TACGCG**	0.0419	495	7483	**CGCGTA**	0.0662
949	21679	**CGTCGA**	0.0438	524	7756	**CGAACG**	0.0676
983	20582	**CGATCG**	0.0478	526	7763	**CGTCGA**	0.0678
1126	23514	**TCGCGA**	0.0479	522	7377	**CGATCG**	0.0708
1182	23908	**CGCGAA**	0.0494	593	7797	**CGTTCG**	0.0761
1133	22271	**CGACGA**	0.0509	659	8600	**CGCGAA**	0.0766
1141	22099	**CGTTCG**	0.0516	673	8748	**TCGCGA**	0.0769
1267	23968	**TTCGCG**	0.0529	707	8640	**TTCGCG**	0.0818
1221	22333	**TCGTCG**	0.0547	699	8452	**ACGCGA**	0.0827
1318	23119	**TCGCGT**	0.0570	744	8495	**TCGCGT**	0.0876
1335	23059	**ACGCGA**	0.0579	703	8006	**CGACGA**	0.0878
1296	22044	**CGAACG**	0.0588	753	8049	**TCGTCG**	0.0936
1389	22207	**ACGTCG**	0.0625	781	8088	**CGACGT**	0.0966

… intermediates omitted for clarity …

780141	219110	**TTTTTT**	3.5605	67297	16785	**TCCCAG**	4.0093
779482	217570	**AAAAAA**	3.5827	67845	16873	**CTGGGA**	4.0209
141196	37321	**GGCTGG**	3.7833	72281	17683	**CCTGGG**	4.0875
141685	37314	**CCAGCC**	3.7971	71331	17399	**CAGGAG**	4.0996
147113	37445	**CCAGGC**	3.9288	72395	17616	**CCCAGG**	4.1096
147567	37501	**GCCTGG**	3.9350	71780	17382	**CTCCTG**	4.1295
153859	38483	**CCTGGG**	3.9981	66918	16177	**GGAGGC**	4.1366
191823	47906	**AGGCTG**	4.0042	67950	16149	**GCCTCC**	4.2077
142716	35611	**GCCTCC**	4.0076	74760	17258	**CCCAGC**	4.3318
154131	38441	**CCCAGG**	4.0095	75340	17381	**GCTGGG**	4.3345
192240	47940	**CAGCCT**	4.0100	79894	18131	**GGGAGG**	4.4064
142599	35522	**GGAGGC**	4.0144	80828	18033	**CCTCCC**	4.4822
158345	37684	**CCCAGC**	4.2019	86477	17439	**AGGCTG**	4.9589
159182	37710	**GCTGGG**	4.2212	86483	17407	**CAGCCT**	4.9684
174090	40083	**CCTCCC**	4.3432	251841	33283	**AAAAAA**	7.5665
173668	39967	**GGGAGG**	4.3453	258292	33992	**TTTTTT**	7.5987

Under- and over-represented oligomers in human chromosomes 1 and 19 are revealed. The longer the oligomers (larger *k*), the more extreme are the ratios deviate from 1.00. The above allows a visual comparison for all possible *k*mers for *k*=2 to 6, they do not show clearly which oligomers are under- or over-represented. It is informative to list some of them with additional information. We decide to include the actual observed counts, the computed expected count according to our method, and the resulting ratios. They are tabulated in ascending *O/E* values, i.e., from the most under-represented to the most over-represented ones. For $k = 2$, all 16 di-mers are shown. For $3 \leq k \leq 6$, only the 16 most under-represented and over-represented *k*-mers are tabulated in the following tables for clarity.

The observed counts in chromosome 1 are shown in the column labeled O1, the expected counts in E1, and so on. For chromosome 19, they are labeled correspondingly as O19, E19, etc.

4 Remarks

We use the concept of set partition to compute DNA oligomers expected counts. When the observed counts are compared to the expected counts, the ratio reveals non-random *k*-mers in both human chromosomes, but not in the random chromosome. This demonstrates the usefulness of the methodology. It needs more extensive comparisons to benchmark how this measure compares to other existing methods like Markov Chain approach.

A general feature which stands out clearly from the figures is its overall similarity at different scale levels - a fractal symmetry property. It is intriguing to see the biological genome displays some kind of fractal symmetry. You see the same motif when all figures (a) are overlaid together; and the same S profile graph for all figures (b). When the O/E ratios are plotted against the alphabetical orders of the *k*-mers, the patterns are preserved with increasing density and scales. We see the same when the ratios are plotted in increasing order. The range and standard deviation of the ratios increase for larger *k* indicates that these *k*-mers are getting more and more non-random for larger *k*. Their increasing order of non-randomness suggests the specificity of these over- and under-represented *k*-mers increases for larger *k*. Their possible connections with genomic words or word fragments await further scrutiny.

Not only are the overall ratios distribution and profile preserved as a whole, individually these non-random oligomers start to cluster into groupings, sometimes in a successive overlapping manner. For example,

In overlapping manner:

- AA, AAA, AAAA, AAAAA, and AAAAAA are all over-represented
- TT, TTT, TTTT, TTTTT, and TTTTTT are all over-represented
- CC, CCT, CCTC, CCTCC, CCTCCC, and their reverse compliments GG, AGG, GAGG, GGAGG, GGGAGG are all over-represented in both chromosomes
- CA, CAG, CCAG, CCCAG, CCCAGC, and their reverse compliments TG, CTG, CTGG, CTGGG, GCTGGG are all over-represented in both chromosomes

- CG, TA, CGT, CGTA, CGCGT, CGCGTA are all under-represented
- CG, TA, TCG, TACG, ACGCG, TACGCG are all under-represented

In non-overlapping manner:

- CC is over-represented, but not CCC, nor CCCC, CCCCC, CCCCCC
- GG is over-represented, but not GGG, nor GGGG, GGGGG, GGGGGG
- CG and CGC are under-represented, but not CGCG, nor CGCGC, CGCGCG

It is reassuring to see consistent observations. Many non-random under-represented *k*-mers contain CG, but do not contain CC or GG. Many non-random over-represented *k*-mers contain either CC or GG, but do not contain CG. These patterns are similar between the two different chromosomes, yet appear in different ranked orders. This is to say, if a *k*-mer is non-random in chromosome 1, than it is likely non-random in chromosome 19, though at different level. These may provide some insights into chromosome differences within species, and even be able to extend to genomic signatures between species.

A large portion of the human genome consists of repeated patterns like poly A's and T's. The reverse complimentary pairs AAAAAA and TTTTTT are the most numerous 6-mers in chromosome 1 and 19. It is also the most extremely over-represented in chromosome 19, and almost so in chromosome 1. This suggests that non-random oligomers may have some kind of relationship with genomic repeats. They may form some basic subunits upon which various repeated patterns are built from or modified into.

Reverse compliment of a *k*-mer is the partner sequence on the paired DNA strand. When the reverse compliment of a sequence is itself, it is called a palindrome. Palindromic sequences are thought to be important DNA motifs involved in the regulation of many cellular processes. Lisnic *et al*. [7] worked on the palindrome content of the yeast Saccharomyces cerevisiae genome. They discovered that while palindromes longer than 12 base pairs were over-represented, short palindromes (2- to 12-mers) were in fact underrepresented. It is interesting to observe similar phenomena here in human chromosomes 1 and 19 as shown in Tables 3, 5.

The complexity of the problem is exponentially increasing as we see the number of set partitions increase dramatically beyond 6. Instead of doing all set partitions to compute the expected value, we can estimate the overall expected value by a Monte Carlo simulation method, or find an approximate value by a first order approximation to reduce the Bell number of total set partitions possible.

Analyses of protein sequences using motif statistics are currently a popular topic in bioinformatics, for example [8]. The approach of computing amino acid sequence expected counts can be similarly applied to proteome analysis.

This paper captures the essence of the method and briefly demonstrates its initial success and potential usefulness. It needs more extensive comparisons to benchmark how this method compares to other existing methods. Our method is a data-based method, and many other methods are model-based, e.g., Markov Chain approach.

In summary, our method allows over- and under-represented *k*-mers to be found based on a statistical criterion, the *O/E* ratio, yet, there is a qualitative difference between statistical significance and biological significance. Specifically, the biological functions of these *k*-mers are not implied here. Statistical analysis of the genome is exploratory and illuminating. The correspondence between statistical and biological significance is an immense research frontier to be explored. The biological significance of these non-random *k*-mers is waiting to be investigated.

Acknowledgements

The author is grateful to John Sanford for introducing the problem, and to Barbara Friedman for proof-reading the manuscript.

References

1. Apostolico, A., Bock, M.E., Lonardi, S., Xu, X.: Efficient Detection of Unusual Words. *Journal of Computational Biology* (2002) 71-94.
2. Apostolico, A., Gong, F., Lonardi, S.: Verbumculus and the Discovery of Unusual Words. *Journal of Computer and Science Technology* 19 (2004) 22-41.
3. Arnau, V., Marin, I.: A Fast Algorithm for the Exhaustive Analysis of 12-Nucleotide-Long DNA Sequences. Applications to Human Genomics. *Proceedings of the International Parallel and Distributed Processing Symposium*. (2003)
4. Cuticchia, A.J., Ivarie, R., Arnold, J.: The Application of Markov Chain Analysis to Oligonucleotide Frequency Prediction and Physical Mapping of Drosophila Melanogaster. *Nucleic Acids Research* 20 (1992) 3651-3657.
5. Hampikian, G., Andersen, T.: Absent Sequences: Nullmomers and Primes. *Pacific Symposium on Biocomputing* 12 (2007) 355-366.
6. Leung, M. Y., Marsh, G. M., Speed, T. P.: Over and Underrepresentation of Short DNA Words in Herpesvirus Genomes. *Journal of Computational Biology* 3 (1996) 345-360.
7. Lisnic, B., Svetec, I.K., Saric, H., Nikolic, I., Zgaga, Z.: Palindrome Content of the Yeast Saccharomyces Cerevisiae Genome. *Current Genetics* 47 (2005) 289-297.
8. Nicodeme, P., Doerks, T., and Vingron, M.: Proteome Analysis Based on Motif Statistics. *Bioinformatics* 18 (2002) 161-171.
9. Phillips, G.J., Arnold, J., Ivarie, R.: Mono-through Hexanucleotide Composition of the Escherchia Coli Genome: A Markov Chain Analysis. *Nucleic Acids Research* 15 (1987) 2611-2626.
10. Reinert, G., Schbath, S., Michael, S., Waterman, M.S.: Probabilistic and Statistical Properties of Words: An Overview. *Journal of Computational Biology* 7 (2000) 1-46.
11. Rigoutsos, I., Floratos, A.: Combinational Pattern Discovery in Biological Sequences: the TEIRESIAS Algorithm. *Bioinformatics* 14 (1998) 55-67.

12. Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A., Platt, D.: Short Blocks from the Non-coding Parts of the Human Genome Have Instances Within Nearly All Known Genes and Relate to Biological Processes. *Proceedings of the National Academy of Sciences of the United States of America* 103 (2006) 6605-6610.
13. Rota, G.: The Number of Partitions of A Set. *American Mathematical Monthly* 71 (1964) 498-504.
14. Schbath, S.: An Efficient Statistics to Detect Over- and Under-represented Words in DNA Sequences. *Journal of Computational Biology* 4 (1997) 189-192.