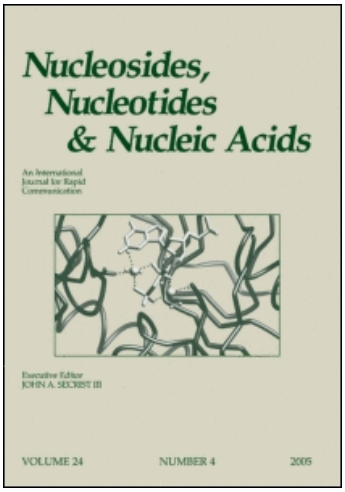


Nucleosides, Nucleotides and Nucleic Acids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

Finding Keywords for Intergenic and Gene Regions for Human Genome

Y. H. Qiao[a]; J. L. Liu[b]; C. G. Zhang[c]; Yanjun Zeng[a]
[a] Biomechanics and Medical Information Institute, Beijing University of Technology, Beijing, China [b] Institute of Computer, Beijing University of Technology, Beijing, China [c] Beijing Institute of Radiation Medicine, Beijing, China

To cite this Article Qiao, Y. H. , Liu, J. L. , Zhang, C. G. and Zeng, Yanjun(2005) 'Finding Keywords for Intergenic and Gene Regions for Human Genome', Nucleosides, Nucleotides and Nucleic Acids, 24: 3, 191 — 198
To link to this Article: DOI: 10.1081/NCN-200055714
URL: http://dx.doi.org/10.1081/NCN-200055714

Nucleosides, Nucleotides, and Nucleic Acids, 24 (3):191–198, (2005)

ISSN: 1525-7770 print/ 1532-2335 online
DOI: 10.1081/NCN-200055714

Taylor & Francis
Taylor & Francis Group

FINDING KEYWORDS FOR INTERGENIC AND GENE REGIONS FOR HUMAN GENOME

Y. H. Qiao ▫ *Biomechanics and Medical Information Institute, Beijing University of Technology, Beijing, China*

J. L. Liu ▫ *Institute of Computer, Beijing University of Technology, Beijing, China*

C. G. Zhang ▫ *Beijing Institute of Radiation Medicine, Beijing, China*

Yanjun Zeng ▫ *Biomechanics and Medical Information Institute, Beijing University of Technology, Beijing, China*

▫ *The analysis of functionally related sequences for conserved patterns is important for further research of different functional regions. This paper presents an analysis of genes and intergenic sequences from the point of view of linguistics analysis, where gene and intergenic regions are regarded as two different subjects written in the four-letter alphabet {A, C, G, T} and high-frequency simple sequences are taken as keywords. A measurement $\alpha[l(\tau)]$ was introduced to describe the relative repeat ratio of simple sequences. Cutoff values were found for keywords selection. After eliminating "noise," 87 short sequences were selected as keywords for intergenic regions and 76 for gene regions.*

INTRODUCTION

The Human Genome Project (HGP), which aimed to identify all of the genomic DNA sequences of human chromosomes, is regarded as a major enterprise for science in the 20th century. As the various genome sequencing projects moving on rapidly, millions of nucleotide of genomic DNA sequences are obtained daily, which makes interpreting the meaning of the genome more important. For this purpose, many annotation tools[1–4] have been developed based on several key simple sequences that were selected from the known samples, and nucleotide, dinucleotide, and trinucleotide cause much attention.[5–7] However, some key simple sequences are usually too small or too short to represent the

Received 30 January 2004, accepted 31 January 2005.
Due to the equal important contribution made to this paper, both Y. H. Qiao and J. L. Liu are first authors.
Address correspondence to Yanjun Zeng, Biomechanics and Medical Information Institute, Beijing University of Technology, Beijing 100022, China; E-mail: yjzeng@bjpu.edu.cn

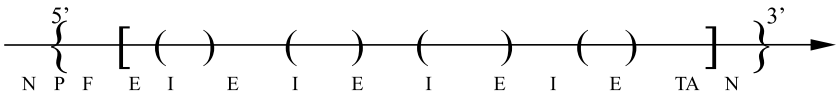
information of certain regions, and there is no universal measurement for the word "key."

This article focuses on finding keywords for a family of related sequences, which means short sequences that represents common features to most members of the family. It is thus of great importance to determine keywords for a family of related sequences, which arise in various applications within computational biology. For example, in the test for DNA sequencing, a keyword can be only determined after sequencing multiple times for the same region of a whole genome. Special keywords selected using computational methods can also provide a convenient index for the common elements in a family of sequences. Based on linguistics analysis and statistics, we put forward a novel method for finding keywords for gene and intergenic regions, which is completely different from multiple sequence alignment.

METHODS

From the point of view of linguistics, the DNA sequence could be read as a sentence composed of the four-letter alphabet {A, C, G, T}. Intergenic and gene regions are two different regions in DNA sequences (Figure 1) which can be regarded as two different subjects written in the four-letter alphabet whose lexical features should be different from each other. In principle, short sequences occurring in high frequency in intergenic sequences provide sufficient information about intergenic regions, and short sequences in high frequency in gene sequences will represent the feature of gene regions, because they have different short key sequences that we define as keywords characterizing the two different regions.

Here we try to find out these keywords based on their discrepancy in these two regions. Keywords are in fact nucleotide strings composed of {A, C, G, T} of variance lengths. Generally speaking, very long strings occur in low frequency in any DNA sequence, so we just consider simple short sequences in a length of less than 7. Sliding windows of length 2, 3, 4, 5 and 6 are used. We denote the length of the sliding window as l. The window is initially located at the start of a DNA sequence. l Letters are read one by one in order. After a nucleotide string of length l is obtained, by sliding the window forward following each nucleotide, a new string will be obtained, and so on. The process was continued until reaching the end of the sequence. With this method a serial of window sequences deduced from



Gene: from P to TA, intergenic region: N, P: promoter,
F: 5'UTR, E: exon , I: intron TA: transcriptional end site

FIGURE 1 Sketch map of gene structure.

different sliding windows are obtained. We denote $l(\tau)$ as short sequences of length l obtained by sliding windows, and τ are some nucleotide strings such as ACG, GTAGG, etc.

Definition 1 Let $R_L[l(\tau)]$ be the measurement of repeats of $l(\tau)$ for a DNA sequence in Length L, satisfying $R_L[l(\tau)] = \frac{n_L[l(\tau)]}{N_L[l(\tau)]}$, where $n_L[l(\tau)]$ is the number of $l(\tau)$ occurrences in sequence L, $N_L[l(\tau)]$ is the number of window sequences of length l in DNA sequence in length L, and $N_L(l) = L-l + 1$.

Here $n_L[l(\tau)]$ is used to measure the frequency that a certain subnucleotide string occurs in a DNA sequence.

For a training sample dataset $\{L_i | i = 1,2,\ldots n\}$ consisting of n sequences, if τ occurs at least once in L_i, we say τ occurred in sequence L_i. Here L_i denotes the i sequence denote $S[l(\tau)]$ is used to denote as the number of sequences that τ occurs in the training sample dataset. Therefore, $\frac{S[l(\tau)]}{n}$ is the ratio that τ occurs in the sample dataset. For an intergenic dataset we denote this ratio as $\frac{S^I[l(\tau)]}{n}$ and for a gene dataset as $\frac{S^G[l(\tau)]}{n}$. In the process of finding keywords, the ratio is important because it provides the measurement of occurrences of $l(\tau)$ in a dataset. If the ratio is close to 1, then $l(\tau)$ are more abundant in a dataset, and if it is close to zero, then $l(\tau)$ are rarely present in a dataset. In order to measure the frequency of a simple sequence occurrence in a dataset, we have the following definition.

Definition 2 Let $R[l(\tau)]$ be the measurement of repeated occurrences of $l(\tau)$ for a training dataset $\{L_i | \mathrm{i} = 1,2\ldots \mathrm{n}\}$, satisfying $R[l(\tau)] = \frac{n[l(\tau)]}{N(l)} \cdot \frac{S[l(\tau)]}{n}$, where $n[l(\tau)]$ is the total number of $l(\tau)$ occurrences in the training dataset, $N(l)$ is the total number of window sequences of length l in the dataset, and $n[l(\tau)] = \sum_{i=1}^{n} n_{L_i}[l(\tau)], N(l) = \sum_{i=1}^{n} N_{L_i}(l)$.

Both definition 1 and definition 2 resemble the ones in genescan.[8] To measure the discrepancy that indicates a simple sequence occurrence in two training datasets, we make the following definition:

Definition 3 Let $R^I[l(\tau)]$ and $R^G[l(\tau)]$ denote the measurement of $l(\tau)$ occurrence in intergenic and gene region datasets respectively. We have the following definition:

$$\alpha[l(\tau)] = \frac{R^I[l(\tau)] - R^G[l(\tau)]}{R^I[l(\tau)] + R^G[l(\tau)]}$$

In the process of finding keywords, the absolute value of $\alpha[l(\tau)]$ is important. If the absolute value of $\alpha[l(\tau)]$ is larger, and the ratio of τ occurrence in a training dataset is over 60%, we deduce τ as a keyword; otherwise, we regard it as "noise." Therefore, there exists a cutoff value, which we denote as ω_0 dividing τ between keywords and "noise." We define the keywords for intergenic and gene regions according to a positive or minus sign of $\alpha(\tau)$.

As mentioned above, because the selection of ω_0 is equivalent to the choice of keywords, we focus on the choice of the ideal value of ω_0. For this purpose, a high-quality dataset is important. We built an intergenic and gene region dataset for testing this algorithm. A set of 613 human 5′ UTR sequences was obtained at first

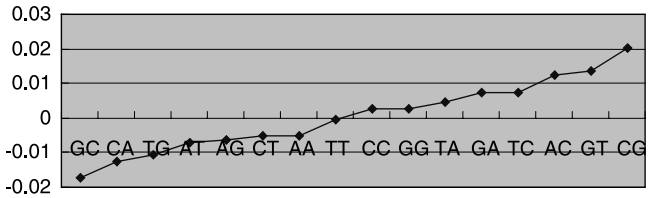


FIGURE 2 The change of $\alpha[l(\tau)]$ of simple sequences in length 2.

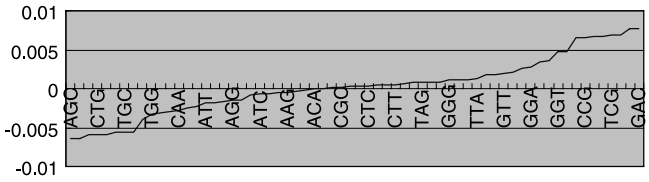


FIGURE 3 The change of $\alpha[l(\tau)]$ of simple sequences in length 3.

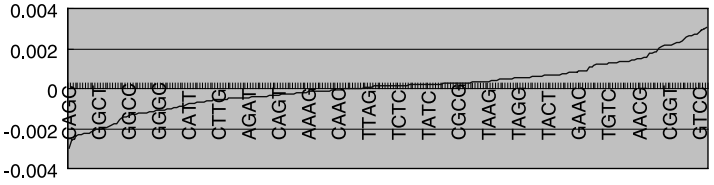


FIGURE 4 The change of $\alpha[l(\tau)]$ of simple sequences in length 4.

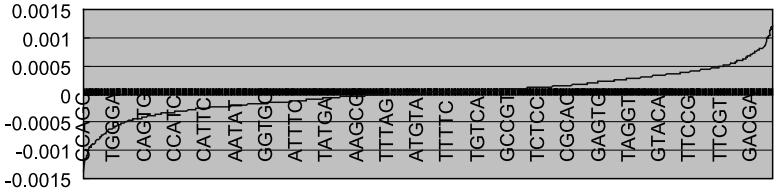


FIGURE 5 The change of $\alpha[l(\tau)]$ of simple sequences in length 5.

from the 5′-end-enriched cDNA library (Suzuki et al., 1997, 2000), the 613 5′ UTR sequences located at human chromosome 22 were mapped to their corresponding genomic sequences using the BLAT program[9] (http://genome.ucsc.edu/cgi-bin/hgBlat). The downstream of 5′-UTR of those genes from chromosome 22 was analyzed, and the annotated neighbor exons and introns from the downstream of 5′ UTR were lined out using the SIM4 program[10] (http://biom3.univ-lyon1.fr/sim4.php) up to the transcription end site (usually indicated by TA). The entire sequence from 5′-UTR to transcription end site was extracted as a gene region. Hence a total number of 613 gene sequences were obtained as the dataset for

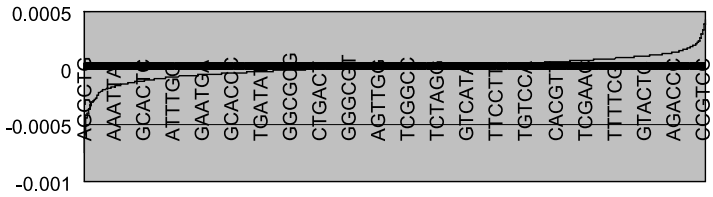


FIGURE 6 The change of $\alpha[l(\tau)]$ of simple sequences in length 6.

analysis (some of them are only segments of a complete structure of a gene, especially those with 5′ UTR which are not found from 5′-end-enriched cDNA library). The downstream sequence from the former transcription end site to the start of ′-UTR of next gene is spontaneously considered as an intergenic sequence.

Two programs were developed for this study: one is for $\alpha[l(\tau)]$ computing based on the definitions and the other is a sorting program to sort $l(\tau)s$ by their $\alpha[l(\tau)]$ values. Looking from the y-axis of Figures 2–6, the magnitude of the discrepancy of $\alpha[l(\tau)]$ for the sequences among $l(\tau)s$ in different lengths is from two to four; therefore, different w_0 are considered for $l(\tau)$ in different lengths.

RESULTS

The detailed results for the value of $\alpha[l(\tau)]$ for nucleotide string sequences of length 2 are listed in Table 1. As shown in Table 1, because the absolute values of $\alpha[l(\tau)]$ for GC, CA, TG, CG, GT, and AC are much larger than that of other sequences in the table, the value 0.01 can be denoted as the cutoff value for defining keywords in length 2. The data in Figure 2 also display the discrepancy among the 16 short sequences. For nucleotide string sequences of length 3, 4, 5, and 6, we sort them in ascending order according to the absolute value of $\alpha[l(\tau)]$. The dispersion of the absolute value of $\alpha[l(\tau)]$ of the latter nucleotide string and the former are obtained one by one; the largest among them was selected. The absolute value of $\alpha[l(\tau)]$ of the latter sequences corresponding to the largest dispersion was used as cutoff value w_0. Those short sequences with absolute value of $\alpha[l(\tau)]$ over w_0 are

TABLE 1 The Value of $\alpha[l(\tau)]$ of the Nucleotide Sequences in Length 2

Alkali base sequences	$\alpha[l(\tau)]$	Alkali base sequences	$\alpha[l(\tau)]$	$\lvert\alpha[l(\tau)]\rvert$
CC	0.00257165	TT	− 0.00029786	0.00029786
GG	0.00261753	AA	− 0.00497144	0.00497144
TA	0.0045099	CT	− 0.00510368	0.00510368
GA	0.00726288	AG	− 0.00627386	0.00627386
TC	0.0074633	AT	− 0.00719976	0.00719976
AC	0.01257825	TG	− 0.01076279	0.01076279
GT	0.01364047	CA	− 0.0124671	0.01246710
CG	0.02033975	GC	− 0.01743213	0.01743213

TABLE 2 The Threshold Values for Simple Sequences in Lengths 2, 3, 4, 5, and 6

Window length	2	3	4	5	6
ω_0	0.01	0.0048	0.002	0.00072	0.0003

TABLE 3 The Simple Key Sequences Selected for Intergenic Region.

The length of keywords	Simple key sequences
2	CG, GT, AC
3	GAC, GTC, CGA, TCG, ACG, CGT, CCG, CGG, ACC, GGT
4	CGAC, GGAC, GTCG, GTCC, GACC, GGTC, GACG, CGTC, CGGA, TCCG, CCGA, ACCG, TCGG, ACGA, ACGT, CGGT, ACGG, CCGT, TCGT, TCGA
5	GACCC, GGACC, GGGTC, GGTCC, CGACC, CCGAC, GTCGG, GGTCG, CGGAC, GTCCG, GGACG, GACGG, CGGAG, ACGTC, TCGGA, CCGTC, ACGAC, TCCGA, ACGGA, CTCCG, AGGAC, CGTCC, GACGT, TCGAC, TCCGT, GACCG, ACCGA, GTCCT, CGGTC, GTCGT
6	GGGTCC, GGACCC, GAGGAC, CGACCC, GTCCTC, GTCGGA, GGGTCG, TCCGAC, CCGACC, CGGACC, CGGAGG, AGGGTC, GACCCT, GGTCGG, GGTCCG, GACGGA, GGACGG, CCTCCG, GGACCG, GACCCG, GAGGGT, CCGTCC, CGGGTC, TCCGTC

TABLE 4 The Simple Key Sequences Selected for Gene Region

The length of keywords	Simple key sequences
2	GC, CA, TG
3	AGC, CAG, GCA, CTG, GCT, GCC, TGC, GGC
4	CAGC, GCTG, AGGC, GCCT, CCAG, CAGG, AGCC, GCAG, GCCA, CTGC, TGCA, CCTG, GGCT, AGCA, AGCT, CTGG
5	CCAGC, CAGCC, GCTGG, CAGGC, GCCTG, GGCTG, CCCAG, CTGGG, AGGCT, AGCCA, CAGCT, AGCCT, CAGCA, GAGGC, AGGCA, CTGCA, CCAGG, AGCTG, AAAAA, GCCAG, TGCAG, GCCTC, TGCCT, CCTGG, GGCAG, TGCTG, TGGCT
6	AGGCTG, CAGCCT, CCCAGC, AAAAAA, GCTGGG, GCCTGG, CCAGCC, CCAGGC, GGCTGG, AGCCTG, CAGGCT, GCCAGG, GGAGGC, CCAGCT, AGGCAG, GCTGAG, GCAGTG, CTCAGC, CACTGC, CCTGGC, CCTGGC, AGCTGG

collected as keywords. Therefore, we chose $w_0 = 0.01$ as the cutoff value for short sequences in length 2. The corresponding tables for short sequences in length 3, 4, 5, and 6 with $\alpha[l(\tau)]$ are too large to be listed in the article. The trendline of short sequences of length 3, 4, 5, and 6 with the change of $\alpha[l(\tau)]$ are illustrated in Figures 3, 4, 5, and 6. The cutoff values are given in Table 2. Using w_0s, the keywords for intergenic and gene regions, which are listed in Tables 3 and 4,

respectively are selected and the keywords are ordered by the absolute value of $\alpha[l(\tau)]$.

DISCUSSION

A new method for finding keywords for certain genome regions is introduced based on linguistics analysis and statistics. It is a complement to the usually used motif finding technique such as sequencing and multiple sequences alignment. A measurement $\alpha[l(\tau)]$ is introduced to describe the relative repeat ratio of simple sequences. Using this method for intergenic and gene regions, consensus sequences were analyzed in a dataset. Results demonstrated that a total of 87 for intergenic regions and 76 for gene regions in length 2, 3, 4, 5 and 6 were found.

Further analysis shows that when increasing or decreasing the cutoff value w_0, the number of selected keywords is increasing or deceasing accordingly, but the number of selected keywords from intergenic regions is always larger than that from gene regions, in accordance with another group's result.[5]

Among simple sequences repeats, triplet repeats are of special significance because some of them have been linked to various genetic disorders.[11,12] However, further research on the significance is needed because the sextuplet repeats contain more information than that of triplet repeats. In this article, about 32% CNNNNC box and 28% GNNNNNG box repeats are found in consensus sequences for gene regions, but 21 and 21%, respectively, for intergenic regions. About 19% ANNNNG and 14% CNNNNT occurred in gene regions, yet they rarely occurred in intergenic regions. Moreover, about 13% CNNNNG, GNNNNT, TNNNNC, respectively, occurred in intergenic regions, yet rarely occurred in gene regions. The information is obviously useful in future for predicting the gene and intergenic regions.

It should be pointed out that the simple sequences from the gene and intergenic regions are based on the training data which we extracted by mapping 5′-UTR sequences to their genome sequences in our analysis. It is possible that there are certain 5′-UTR sequences in the genome, which are not predicted or annotated precisely, and some 5′-UTR sequences on chromosome 22 are not found. Therefore, the keywords within gene and intergenic regions are tentative. It is thus expected that whenever the entire human gene regions are fully available and the gene boundary is more exactly determined, more exact information on the human genome and other eukaryotic genome could be found in the future.

REFERENCES

1. Davuluri, R.V.; Suzuki, Y.; Sugano, S.; Zhang, M.Q. CART classification of human 5′UTR sequences. Genome Res. **2000**, 1807–1816.
2. Burge, C.; Karlin, S. Prediction of complete gene structure in human genomic DNA. J. Mol. Biol. **1997**, 268, 78–94.
3. Rogic, S.; Francis, B.F.; Mackworth, A.K. Improving gene recognition accuracy by combining predictions from two gene-finding programs. **2002**, 18, 1034–1045.

4. Gao, F.; Zhang, C.T. Comparison of various algorithms for recognizing short coding sequences of human genes. Bioinformatics **2004**, 20(5), 673–681.
5. Goldman, N. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations. Nucleic Acids Res. **1993**, 21(10), 2478–2491.
6. Torres, A.; Nieto, J.J. The fuzzy polynucletide space: basic properties. Bioinformatics **2003**, 19(5), 587–592.
7. Bazzan, A.L.C.; Engel, P.M.; Schroeder, L.F.; da Silva, S.C. Automated annotation of keywords for proteins related to mycoplasmataceae using machine learing techniques. Bioinformatics **2002**, 18, 355–435.
8. Burge, C.; Karlin, S. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **1997**, 268, 78–94.
9. Kent, W.J. BLAT–the BLAST-like alignment tool. Genome Res. **2002**, 12, 656–664.
10. Florea, L.; Hartzell, G.; Zhang, Z.; Rubin, G.M.; Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. **1998**, 8, 967–974.
11. Subramanian, S.; Adgula, V.M.; et al. Triplet repeats in human genome: distribution and their association with genes and other regions. Bioinformatics **2003**, 19(5), 549–552.
12. Baldi, P.; Brunak, S.; Chauvin, Y.; Pedersen, A.G. Structural basis for triplet disorders: a computational analysis. Bioinformatics **1999**, 15, 918–929.