Computers in Biology and Medicine 42 (2012) 935–941

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A statistical fat-tail test of predicting regulatory regions in the *Drosophila* genome

Jian-Jun Shu*, Yajing Li

School of Mechanical & Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 2 September 2010
Accepted 18 July 2012

Keywords:
Statistical approach
Transcription factor binding sites (TFBSs)
Cis-regulatory modules (CRMs)

ABSTRACT

A statistical study of *cis*-regulatory modules (CRMs) is presented based on the estimation of similar-word set distribution. It is observed that CRMs tend to have a fat-tail distribution. A new statistical fat-tail test with two kurtosis-based fatness coefficients is proposed to distinguish CRMs from non-CRMs. As compared with the existing fluffy-tail test, the first fatness coefficient is designed to reduce computational time, making the novel fat-tail test very suitable for long sequences and large database analysis in the post-genome time and the second one to improve separation accuracy between CRMs and non-CRMs. These two fatness coefficients may serve as valuable filtering indexes to predict CRMs experimentally.

1. Introduction

The identification of transcription factor binding sites (TFBS's) and *cis*-regulatory modules (CRMs) is a crucial step in studying gene regulation. Computational methods of predicting CRMs can be classified into three types: (1) TFBS-based methods; (2) homology-based methods and (3) content-based methods. TFBS-based methods, such as ClusterBuster [12] and MCAST [2], use information about known TFBS's to identify potential CRMs. Methods of this type are generally unable to be applied to genes for which TFBS's have not yet been studied experimentally. Homology-based methods use information contained in the pattern of conservation among related sequences. The related sequences can come from single species [23], two species [14] and multiple species [8]. Methods of this type using the pattern of conservation alone are limited in their performance because TFBS conservation necessary to maintain regulatory function in binding sequences may not be significantly higher than in non-binding sequences [11]. In addition, it still remains an open question that how many genomes are sufficient to the reliable extraction of regulatory regions. Content-based methods assume that different genome regions (CRMs, exons and NCNRs) have different rates of evolutionary micro changes; therefore, they exhibit different statistical properties in nucleotide composition. TFBS's often occur together in clusters as CRMs [7,15]. The binding site cluster causes a biased word distribution within CRMs, and this bias leaves a distinct "signature" in nucleotide composition. Content-based methods detect this signature by statistical techniques [16,1] or machine learning techniques [9], in order to distinguish CRMs from non-CRMs. Methods of this type may be used to predict the CRMs which have not yet been observed experimentally. A large number of CRM search tools have been reported in the literature, but the computational method attempting to identify CRMs still remains a challenging problem due to the limited knowledge of specific interactions involved [22].

The fluffy-tail test [1] is one of content-based methods. It is a bootstrapping procedure to identify CRMs by checking the statistical difference between the size distribution of the largest group of similar-words obtained for the randomized shuffled sequences and the corresponding size distribution for the original input nucleotide sequence. If there are no statistical differences, it is concluded that the original input nucleotide sequence probably is a coding (exon) region or a non-coding non-regulatory (NCNR) region.

In the work that follows, the fluffy-tail test is re-examined by considering the following two issues: (1) Due to its bootstrapping procedure, the computational time of calculating the fluffiness coefficient is determined by the number of randomization. In order to get reliable results statistically, the number of randomization is usually set very large in the fluffy-tail test, so the computational time is expensive, especially for long sequences. This limits the use of the fluffy-tail test under the situation when more and more DNA sequences need to be analyzed in the post-genome time. (2) The fluffy-tail test looks only at the subsequence with the highest incidence in the CRMs. Therefore, the fluffy-tail test may not capture the statistical features caused by heterotypic TFBS clusters in the regulatory regions. It is an interest to address these two issues of the fluffy-tail test and to develop a more efficient and effective CRM prediction method.

This paper is to explore some statistical properties of DNA composition due to the multiple occurrences of TFBS's of the same or different types in CRMs. For an enumeration purpose, a consensus

* Corresponding author. Tel.: +65 6790 4459; fax: +65 6791 1859.
E-mail address: mjjshu@ntu.edu.sg (J.-J. Shu).

0010-4825/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2012.07.007

sequence is used as a motif representation, *i.e.*, using a similar-word set to represent a motif. The main concern is to explore specific properties in similar-word set distribution for CRMs, and to identify suitable parameters in order to distinguish CRMs from non-CRMs.

2. Materials and methods

2.1. Training datasets

To explore statistical parameters to distinguish CRMs from non-CRMs, three training datasets are used in this paper. The positive training set is a collection of 60 experimentally-verified functional *Drosophila melanogaster* regulatory regions [17,16]. This set consists of CRMs located far from gene coding regions and transcription start sites. It contains many binding sites and site clusters, including *abdominal-b*, *bicoid*, *caudal*, *deformed*, *distal-less*, *engrailed*, *even-skipped*, *fushi tarazu*, *giant*, *hairy*, *huckebein*, *hunchback*, *knirps*, *krüppel*, *odd-paired*, *pleiohomeotic*, *runt*, *tailless*, *tramtrack*, *twist*, *wingless* and *zeste*. The total size of positive training sets comprises about 99 kilobase (kb) sequences. The two negative training sets are (1) 60 randomly-picked *D. melanogaster* exons; and (2) 60 randomly-picked *D. melanogaster* NCNRs: the exons and NCNRs of length 1 kb upstream and downstream of genes are excluded by using the Ensembl genome browser. The exon training set contains 85 kb sequences, and the NCNR training set contains 90 kb sequences. All sequences with tandem repeats in the three training datasets are masked by using a tandem repeats finder program [6] before processing.

2.2. Formulation of the fat-tail test

The fat-tail test is based on the assumption that each word (binding site) recognized by a given transcription factor belongs to its own family of similar-word sets (binding site motifs) found in the same enhancer sequence and the redundancy of binding sites within CRMs leaves distinct "signatures" in similar-word set distribution. For a given m-letter segment W_m as a seed-word, all m-letter words that differ from W_m by no more than j substitution comprise a corresponding similar-word set $N_j(W_m)$. Because the core of TFBS's is relatively short [24], a five-letter seed-word is selected, allowing for one mismatch, that is, $m=5$ and $j=1$. The fat-tail test is adopted to study the similar-word set distribution and to predict the probable function of the original input sequence. A flow chart of the fat-tail test is shown in Fig. 1.

Step 1: Number of similar-words with the same seed-word (n)
As an example, consider a stretch of DNA: ACGACGCCGACT. For $m=5$ and $j=1$, all five-letter segment W_5 is selected as a seed-word, that is, ACGAC, CGACG,…,CGACT. For each seed-word W_m, all m-letter words with no more than j substitution comprise a corresponding similar-word set $N_j(W_m)$. In this example, the first seed-word W_5, ACGAC, has three similar-words with no more than one mismatch: ACGAC, ACGCC, CCGAC; n is the cardinality, $n=|N_j(W_m)|=|N_1(\text{ACGAC})|=3$, and forms the X-axis in Figs. 2–7.
Step 2: Number of seed-words with the same number of similar-words (f)
$f(n)$ is the number of seed-words containing n similar-words and forms the Y-axis in Figs. 2–9.
Step 3: Kurtosis (k)
The kurtosis k of similar-word set distribution $f(n)$ is evaluated as

$$k=\frac{\sum_{n=1}^{N}[f(n)-\mu]^4}{(N-1)\sigma^4}-3 \tag{1}$$

where μ and σ are the mean and standard deviation respectively.

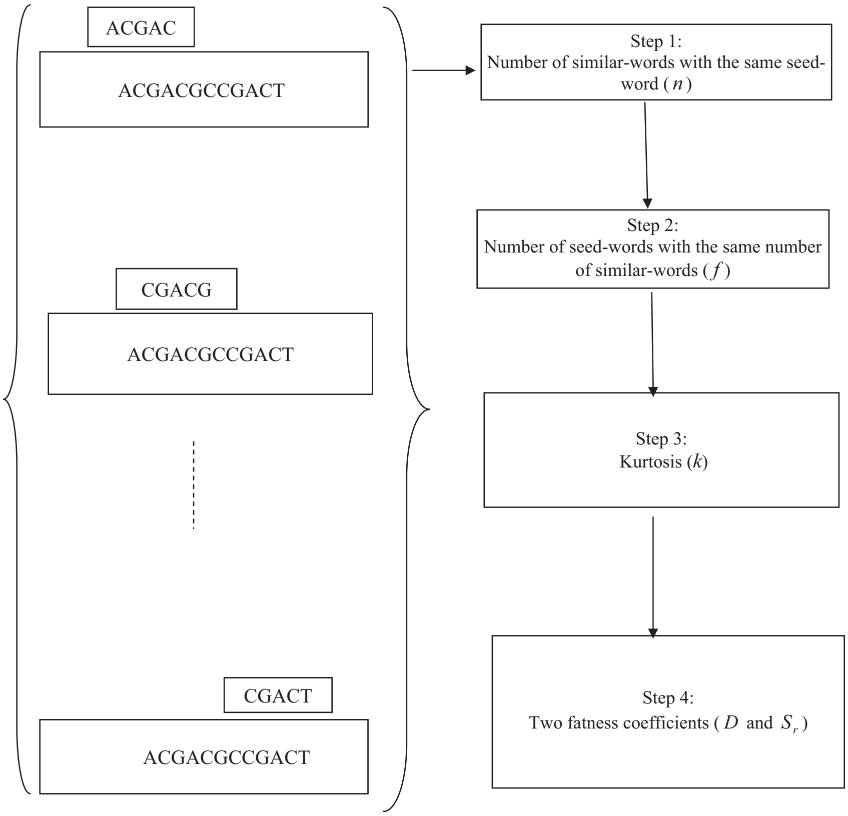
ACGAC
ACGACGCCGACT
CGACG
ACGACGCCGACT
CGACT
ACGACGCCGACT
Step 1:
Number of similar-words with the same seed-word (n)
Step 2:
Number of seed-words with the same number of similar-words (f)
Step 3:
Kurtosis (k)
Step 4:
Two fatness coefficients (D and S_r)

Fig. 1. A flow chart of the fat-tail test.

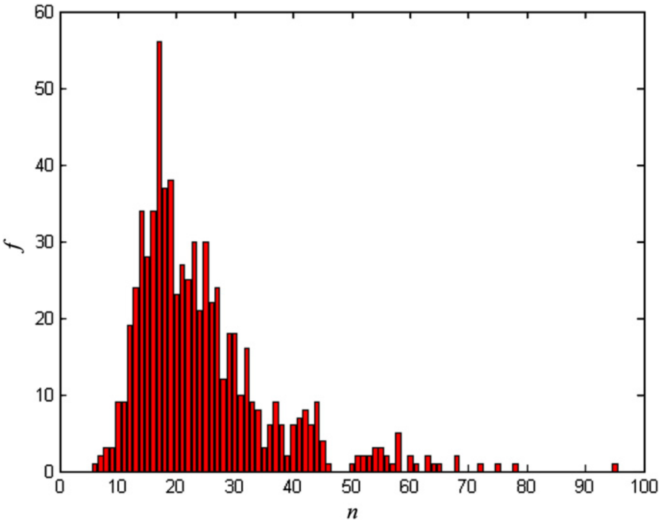

Fig. 2. Histogram of *Drosophila* CRMs ($m=5$, $j=1$, $k=4.19$, $\mu=24.4$, $\sigma=11.7$).

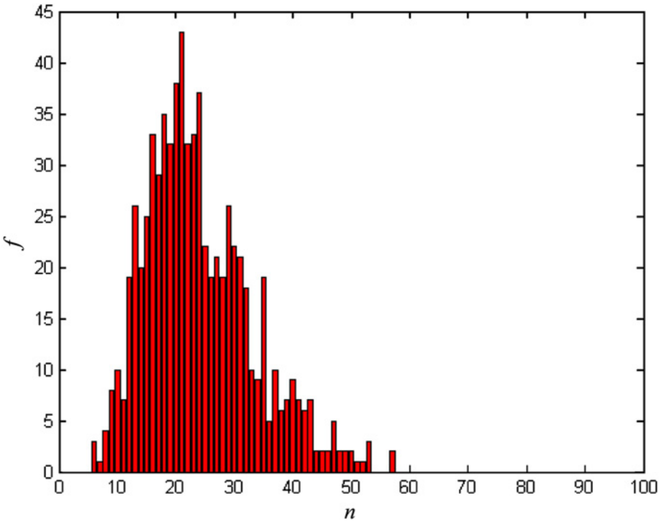

Fig. 3. Histogram of *Drosophila* CRMs ($m=5$, $j=1$, $k=0.19$, $\mu=23.9$, $\sigma=7.7$) after randomly-shuffling.

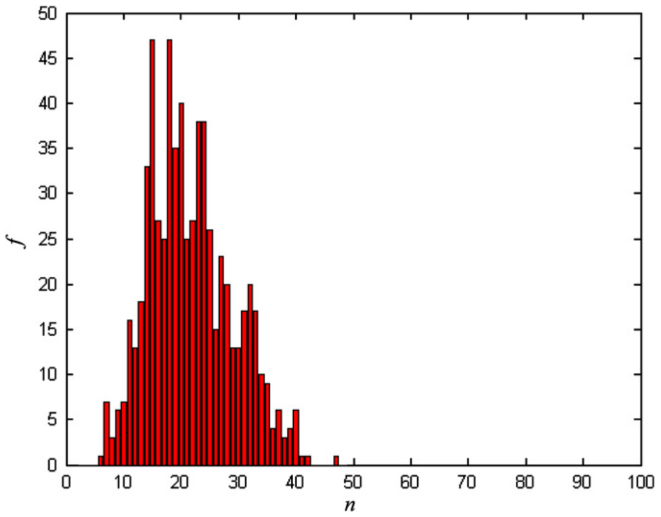

Fig. 4. Histogram of *Drosophila* exons ($m=5$, $j=1$, $k=-0.28$, $\mu=21.73$, $\sigma=7.33$).

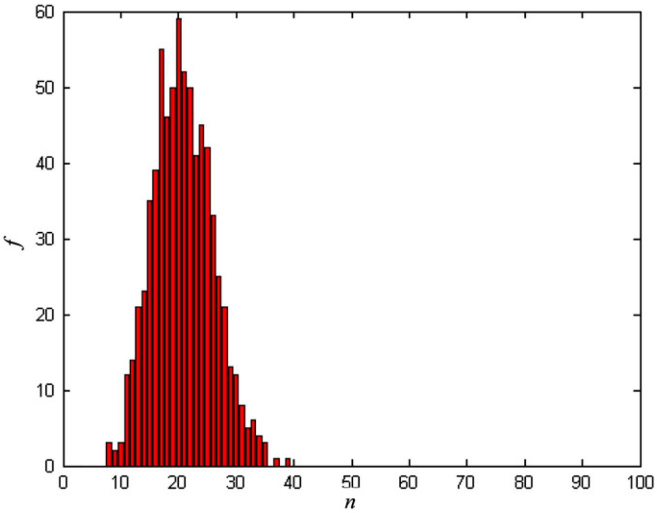

Fig. 5. Histogram of *Drosophila* exons ($m=5$, $j=1$, $k=0.35$, $\mu=21.4$, $\sigma=7.19$) after randomly-shuffling.

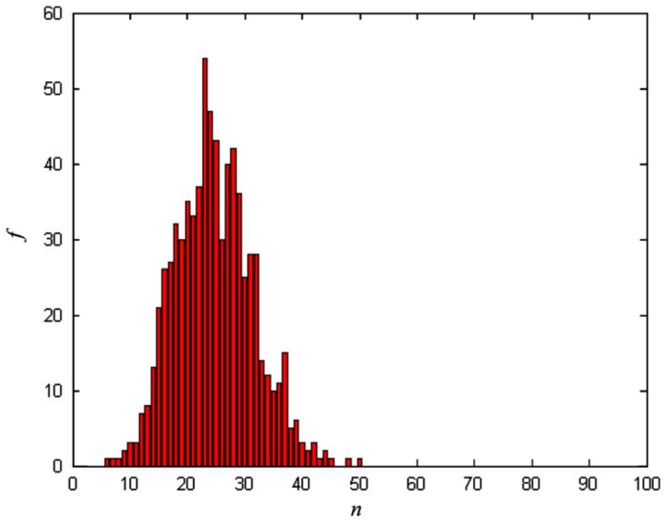

Fig. 6. Histogram of *Drosophila* NCNRs ($m=5$, $j=1$, $k=0.09$, $\mu=24.66$, $\sigma=6.82$).

Step 4: Two fatness coefficients (D *and* S_r)

The first fatness coefficient D is defined as

$$D=\frac{k_0+2\varepsilon}{4\varepsilon} \tag{2}$$

Here k_0 denotes the kurtosis k of the original input sequence without randomly-shuffling and ε is the standard error calculated by

$$\varepsilon=2\sqrt{\frac{6}{N}} \tag{3}$$

D is used to measure how strong the similar-word set distribution of CRMs deviates from normal distribution. The 95% confidence interval is set between -2ε and 2ε.

To measure how strong the similar-word set distribution of CRMs deviate from randomness, the second fatness coefficient S_r is computed by comparing with all randomized r-time shuffled sequence versions of the original input sequence:

$$S_r=\frac{k_0-k_r}{\sigma_r} \tag{4}$$

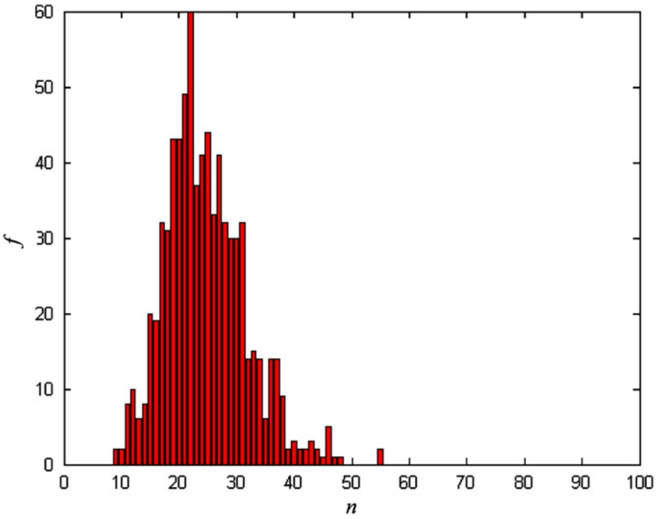

Fig. 7. Histogram of *Drosophila* NCNRs (m=5, j=1, k=0.25, μ=24.32, σ=6.59) after randomly-shuffling.

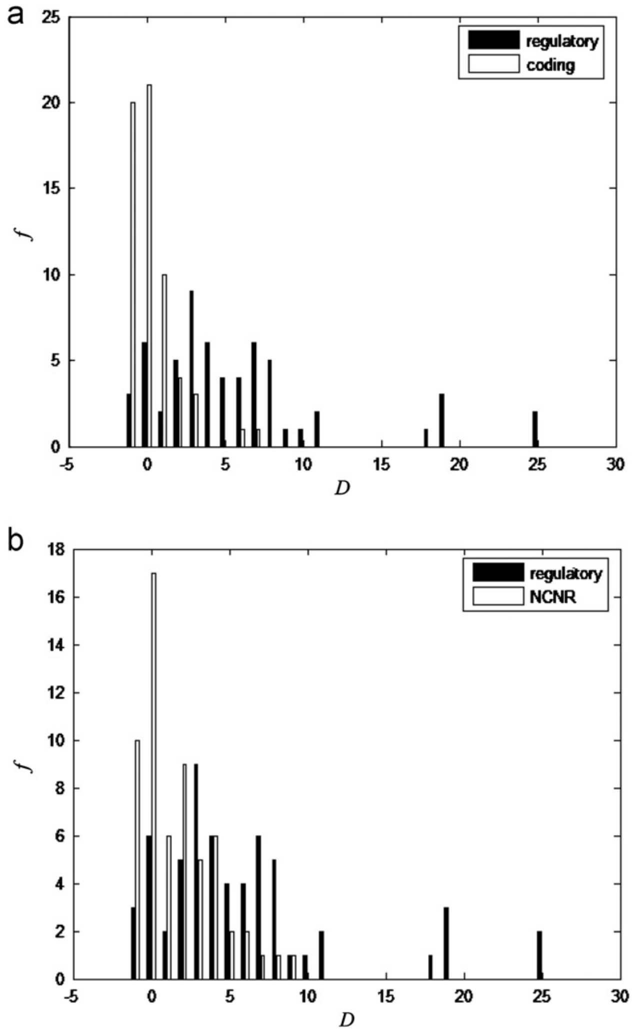

Fig. 8. Histogram for CRMs, exons and NCNRs classified by D (m=5, j=1). (a) CRMs vs. exons and (b) CRMs vs. NCNRs.

Here a sequence is called "random" if it is obtained from the original input sequence by shuffling it, preserving its single nucleotide composition. S_r can be regarded as measuring the degree of difference between signal and noise, where the signal is regarded as the original input sequence, and the noise is regarded as randomized sequences.

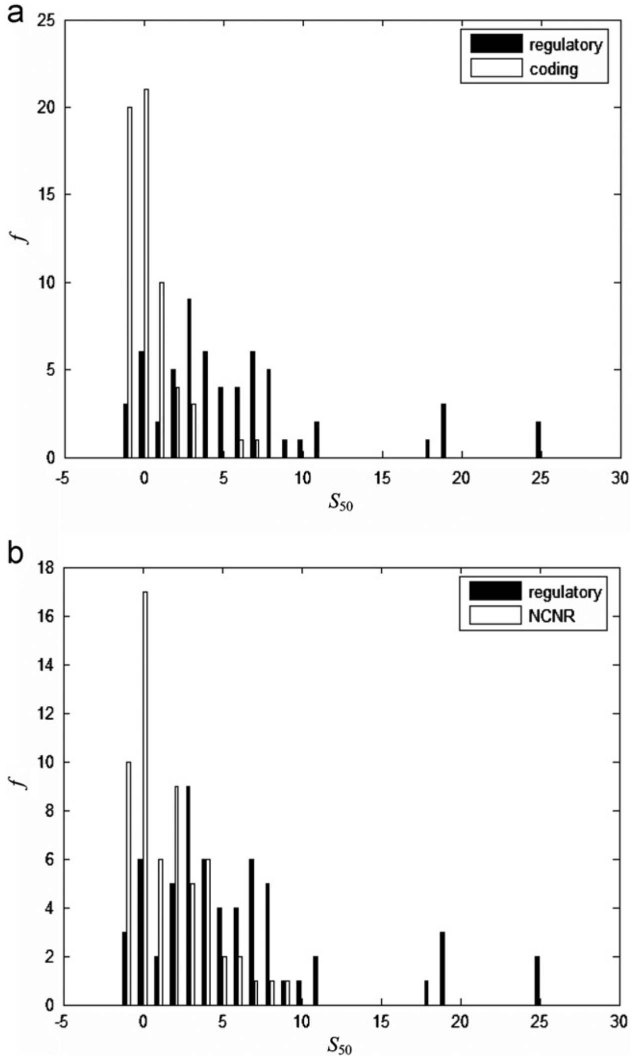

Fig. 9. Histogram for CRMs, exons and NCNRs classified by S_{50} (m=5, j=1). (a) CRMs vs. exons and (b) CRMs vs. NCNRs.

In the fluffy-tail test [1], the fluffiness coefficient F_r is defined as

$$F_r = \frac{L_0 - L_r}{\sigma_r} \tag{5}$$

where L_r is the number of seed-words with the maximal similar-words for r-time shuffled sequences. Here it is worth to mention to this end that CRMs tend to have a fat-tail distribution in Fig. 2, as compared with that of the randomized sequence in Fig. 3. Since kurtosis measures the tail heaviness of a distribution relative to that of normal distribution, the second fatness coefficient S_r based on the kurtosis k_r should be a more reasonable index than the fluffiness coefficient F_r based on the maximal number L_r in order to predict CRMs.

3. Results

3.1. Distribution for CRMs

For the training datasets of CRMs, Fig. 2 shows a similar-word set distribution for a region of *D. melanogaster hunchback* CRMs.

It can be seen that the most frequent similar-word set occurs 10–40 times and some similar-word sets occur about 95 times. If the original input sequence is characterized by the presence of an unusually-high number of over-represented similar-words, the similar-word set distribution is expected to have a long right tail in comparison with that of a random sequence, in view of that (k_0=4.19) is far greater than (k=0) for the normal distribution.

To obtain a random distribution, the original input sequence is shuffled 50 times by using the Fisher–Yates shuffle algorithm. Fig. 3 shows a typical example of similar-word set distribution after randomly-shuffling. As compared with the original input sequence in Fig. 2, the randomized sequence in Fig. 3 lacks a long right tail, and is nearly the normal distribution, in view of (k_r=0.19) around 0.

3.2. Distribution for exons

For the training datasets of randomly-picked *D. melanogaster* exons, Fig. 4 shows a similar-word set distribution for a region of *D. melanogaster CG8229* exons. The absence of long right tail is noted in Fig. 4 in view of that (k_0=−0.28) is around 0. Fig. 5 shows a typical example of similar-word set distribution after randomly-shuffling with (k_r=0.35) around 0. The kurtosis k_0 of similar-word set distribution for the original input sequence does not differ significantly from k_r of the randomized version (k_0=−0.28) *vs.* (k_r=0.35).

3.3. Distribution for NCNRs

For the training datasets of randomly-picked *D. melanogaster* NCNRs, Fig. 6 shows a similar-word set distribution for a region of *D. melanogaster* NCNRs. The presence of short right tail is noted in Fig. 6 in view of that (k_0=0.09) is around 0. Fig. 7 shows a typical example of similar-word set distribution after randomly-shuffling with (k_r=0.25) around 0. The kurtosis k_0 of similar-word set distribution for the original input sequence does not differ significantly from k_r of the randomized version (k_0=0.09) *vs.* (k_r=0.25).

3.4. The fat-tail test

In order to distinguish CRMs from non-CRMs, D and S_r are calculated for 180 sequences in three training datasets. Fig. 8 shows that CRMs tend to have a greater D than exons and NCNRs. Table 1(a) lists functional classification based on D. Nearly 75% CRMs have $D>2$, while only 18.3% exons have $D>2$, and 53.3% NCNRs have $D>2$. Fig. 9 shows S_{50} for CRMs, exons and NCNRs. For each sequence, its (r=50)-time shuffled versions are generated to calculate $S_{50.}$ It can be seen that CRMs intend to have greater S_{50} than exons and NCNRs. Table 1(b) lists functional classification based on S_{50}. Nearly 76.7% CRMs have $S_{50}>2$, while only 11.7% exons have $S_{50}>2$, and 36.7% NCNRs have $S_{50}>2$.

3.5. Large CRM datasets

The fat-tail algorithm has been tested on the current version 3 of *REDfly* database [13], which contains 894 experimentally-verified CRMs from *Drosophila*. Results show that 63.1% CRMs have $D>2$ and 59.5% CRMs have $S_{50}>2$ passing the fat-tail test. The low pass rate may be due to the stringent threshold value. Another possible reason is that some CRMs do not contain binding site cluster. This directs future study: (1) to check if the binding site clustering is the common feature of all CRMs; (2) to optimize the threshold to get more reliable results. It is worth to mention to the point that the fluffy-tail algorithm has never been tested on the large CRM datasets.

4. Discussion

Some statistical properties of similar-word set distribution in three training datasets have been explored. Results show that CRMs have a fat-tail distribution, *i.e.*, tend to have high fatness coefficients ($D>2,S_r>2$), while exons lack a fat-tail distribution, *i.e.*, tend to have low fatness coefficients. However, NCNRs tend to have median fatness coefficients. Thus, D and S_r can be used to distinguish between CRMs and exons effectively. CRMs are predominant if ($D>2,S_r>2$), while exons are prevailing if ($D<2$, $S_r<2$). Thus, the regions with ($D>2,S_r>2$) are CRMs and those with ($D<2,S_r<2$) are exons.

4.1. Comparison with the fluffy-tail test

The fat-tail test is evaluated by comparison with the fluffy-tail test [1]. The performance of three parameters is assessed: (1) the first fatness coefficient D; (2) the second fatness coefficient S_r; and (3) the fluffiness coefficient F_r based on separation between CRMs and exons, and between CRMs and NCNRs.

The training datasets are employed to evaluate the above three parameters. For comparison, the original input sequence is shuffled 50 times to calculate S_{50} and F_{50}. The thresholds of D, S_{50} and F_{50} are all set as two. For the fat-tail test, the original input DNA sequence is considered with $D>2$ predicted as CRMs, $D<2$ as predicted exons, and $S_{50}>2$ as predicted CRMs, $S_{50}<2$ as predicted exons. For the fluffy-tail test, the original input DNA sequence is considered with $F_{50}>2$ as predicted CRMs, $F_{50}<2$ as predicted exons. The classification result of 180 sequences in the training datasets by F_{50} is listed in Table 1(c). The fluffy-tail test F_{50} identified 42 out of 60 CRMs in the positive training datasets, while the fat-tail test identified 45 and 46 CRMs with D and S_{50} respectively (see Table 1). For each parameter, sensitivity (SN) (number of true positive/number of positive), specificity (SP) (number of true negative/number of negative) and accuracy (number of true positive+number of true negative)/(number of positive+number of negative) are calculated to distinguish CRMs from exons and NCNRs (Table 2).

For distinguishing CRMs from exons, the fat-tail test with S_{50} has the best accuracy (82.5%), as compared with the other two parameters (D: 78.3%; F_{50}: 78.3%). Thus, the fat-tail test with S_{50} can effectively distinguish between CRMs and exons. Moreover, S_{50} (SN=76.7%) can more efficiently identify CRMs than D (SN=75%) and F_{50} (SN=70%), as well as S_{50} (SP=88.3%) can more efficiently identify exons than F_{50} (SP=86.7%) and D (SP=81.7%). The fat-tail test with D has the same accuracy as the fluffy-tail

Table 1
Classification of 180 sequences.

Functional type	$D>2$	$D<2$	Positive rate (%)	Negative rate (%)
(a) The fat-tail test with D				
CRMs	45	15	75	25
Exons	11	49	18.3	81.7
NCNRs	32	28	53.3	46.7
Functional type	$S_{50}>2$	$S_{50}<2$	Positive rate (%)	Negative rate (%)
(b) The fat-tail test with S_{50}				
CRMs	46	14	76.7	23.3
Exons	7	53	11.7	88.3
NCNRs	22	38	36.7	63.3
Functional type	$F_{50}>2$	$F_{50}<2$	Positive rate (%)	Negative rate (%)
(c) The fluffy-tail test				
CRMs	42	18	70	30
Exons	8	52	13.3	86.7
NCNRs	21	39	35	65

Table 2
Evaluation of D, S_{50} and F_{50}.

	The fat-tail test		The fluffy-tail test
	D (%)	S_{50} (%)	F_{50} (%)
(a) Distinguishing CRMs from exons			
SN	75	76.7	70
SP	81.7	88.3	86.7
Accuracy	78.3	82.5	78.3
	The fat-tail test		The fluffy-tail test
	D (%)	S_{50} (%)	F_{50} (%)
(b) Distinguishing CRMs from NCNRs			
SN	75	76.6	70
SP	46.7	63.3	65
Accuracy	60.8	70	67.5
	The fat-tail test		The fluffy-tail test
	D (s)	S_{50} (s)	F_{50} (s)
(c) CPU time for a sequence length of 1000			
CPU time	6.2	310	310

Table 3
Sensitivity of S_r to choice of r for CRMs (k=4.19).

r	S_r	k_r	σ_r
50	3.63	0.26	0.67
100	5.31	0.2	0.47
500	4.28	0.2	0.58

test. However, the computational time (CPU time) of calculating D for an original input DNA sequence length of 1000 is 50 times faster than those of calculating F_{50} and S_{50} for the same original input sequence, because of no 50-time randomly-shuffling is required for calculating D. Thus, the fat-tail test with D is very suitable for long sequences and large database. For distinguishing CRMs from NCNRs, the results show that the accuracy (67.5%) of the fluffy-tail test with F_{50} is worse than (70%) of the fat-tail test with S_{50}, but better than (60.8%) of the fat-tail test with D.

4.2. Time complexity

Table 3 shows that the value of the fat-tail kurtosis coefficient S_r is affected by the number of randomization r. In order to get more reliable estimation of S_r, a large r is needed, so that high computational time is expected. For reliable result within reasonable computational time, the original input sequence is shuffled by 50 times to calculate S_r.

The algorithm used for shuffling is the Fisher–Yates shuffle algorithm, which is linear on the sequence length N, so that the time complexity of calculating D is $O(N)$ and the time complexity of calculating S_r and F_r is $O(Nr)$. In Table 2(c), the computational time (CPU time) of calculating D is 50 times faster than those of calculating F_{50} and S_{50} due to no sequence shuffling. All computations are run on a 3.2 GHz Pentium IV processor with 1 G physical memory.

4.3. Tandem repeat region

The results show that the most frequent similar-word set usually corresponds to the word of “TTTTT” or “AAAAA” for CRMs and NCNRs. These phenomena are due to the poly N (such as TTT…) occurrence in CRMs and NCNRs and affect greatly the maximal number L_r. Thus, true CRMs cannot be distinguished from NCNRs effectively in the fluffy-tail test. The motifs corresponding to experimentally-verified TFBS’s usually occur more than the mean value of similar-word set distribution and locate around the right tail, so that the prediction accuracy using the kurtosis-based fatness coefficient S_r is improved. It is worth to mention to this end that the phenomenon of motif fat-tail distribution can be also observed in protein sequences [3–5,10,18,19,21].

5. Conclusion

The redundancy of binding sites within CRMs causes the bias base composition and leaves distinct “signatures” in similar-word set distribution. The fluffy-tail test captured this characteristic by searching the most frequent similar-word. However, the real binding site motif may be the moderate similar-word sets. In this paper, the fat-tail test is proposed to distinguish CRMs from non-CRMs. In the fat-tail test, characteristics are investigated by examining distribution pattern, using datasets of 180 DNA sequences (60 for CRMs, 60 for exons and 60 for NCNRs). Results show that the similar-word set distribution of CRMs tends to be a fat-tail distribution as compared with those of exons and NCNRs. Based on this observation, two kurtosis-based fatness coefficients D and S_r are introduced here. The fat-tail test with D has comparable accuracy to, but r times faster than the fluffy-tail test, because of no r-time randomly-shuffling required. The fat-tail test with S_r has better accuracy of distinguishing CRMs from exons and NCNRs than the fluffy-tail test. Thus, the novel fat-tail test greatly simplifies the functional prediction of an original input DNA sequence and can guide future experiments aimed at finding new CRMs in the post-genome time [20].

Conflict of interest statement

None declared.

References

[1] I. Abnizova, R. te Boekhorst, K. Walter, W.R. Gilks, Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the *Drosophila* genome: the fluffy-tail test, BMC Bioinform. 6 (2005) 109.

[2] T.L. Bailey, W.S. Noble, Searching for statistically significant regulatory modules, Bioinformatics 19 (2) (2003) II16–II25.

[3] O. Bastien, J.-C. Aude, S. Roy, E. Maréchal, Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of *Z*-value statistics, Bioinformatics 20 (4) (2004) 534–537.

[4] O. Bastien, A simple derivation of the distribution of pairwise local protein sequence alignment scores, Evol. Bioinform. 4 (2008) 41–45.

[5] O. Bastien, E. Maréchal, Evolution of biological sequences implies an extreme value distribution of type I for both global and local pairwise alignment scores, BMC Bioinform. 9 (2008) 332.

[6] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (2) (1999) 573–580.

[7] B.P. Berman, Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, M.B. Eisen, Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome, Proc. Natl. Acad. Sci. USA 99 (2) (2002) 757–762.

[8] D. Boffelli, J. McAuliffe, D. Ovcharenko, K.D. Lewis, I. Ovcharenko, L. Pachter, E.M. Rubin, Phylogenetic shadowing of primate sequences to find functional regions of the human genome, Science 299 (5611) (2003) 1391–1394.

[9] B.Y. Chan, D. Kibler, Using hexamers to predict *cis*-regulatory motifs in *Drosophila*, BMC Bioinform. 6 (2005) 262.

[10] J.P. Comet, J.C. Aude, E. Glémet, J.L. Risler, A. Hénaut, P.P. Slonimski, J.J. Codani, Significance of *Z*-value statistics of Smith–Waterman scores for protein alignments, Comput. Chem. 23 (3–4) (1999) 317–331.

[11] E. Emberly, N. Rajewsky, E.D. Siggia, Conservation of regulatory elements between two species of *Drosophila*, BMC Bioinform. 4 (2003) 57.

[12] M.C. Frith, M.C. Li, Z.P. Weng, Cluster-Buster: finding dense clusters of motifs in DNA sequences, Nucleic Acids Res. 31 (13) (2003) 3666–3668.

[13] S.M. Gallo, D.T. Gerrard, D. Miner, M. Simich, B. Des Soye, C.M. Bergman, M.S. Halfon, *REDfly* v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*, Nucleic Acids Res. 39 (1) (2011) D118–D123.

[14] Y.H. Grad, F.P. Roth, M.S. Halfon, G.M. Church, Prediction of similarly acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D. pseudoobscura*, Bioinformatics 20 (16) (2004) 2738–2750.

[15] A.P. Lifanov, V.J. Makeev, A.G. Nazina, D.A. Papatsenko, Homotypic regulatory clusters in *Drosophila*, Genome Res. 13 (4) (2003) 579–588.

[16] A.G. Nazina, D.A. Papatsenko, Statistical extraction of *Drosophila cis*-regulatory modules using exhaustive assessment of local word frequency, BMC Bioinform. 4 (2003) 65.

[17] D.A. Papatsenko, V.J. Makeev, A.P. Lifanov, M. Régnier, A.G. Nazina, C. Desplan, Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers, Genome Res. 12 (3) (2002) 470–481.

[18] J.-J. Shu, L.S. Ouw, Pairwise alignment of the DNA sequence using hypercomplex number representation, Bull. Math. Biol. 66 (5) (2004) 1423–1438.

[19] J.-J. Shu, Y. Li, Hypercomplex cross-correlation of DNA sequences, J. Biol. Syst. 18 (4) (2010) 711–725.

[20] J.-J. Shu, Q.-W. Wang, K.-Y. Yong, DNA-based computing of strategic assignment problems, Phys. Rev. Lett. 106 (18) (2011) 188702.

[21] J.-J. Shu, K.Y. Yong, W.K. Chan, An improved scoring matrix for multiple sequence alignment, Math. Probl. Eng. 2012 (490649) (2012) 1–9.

[22] J. Su, S.A. Teichmann, T.A. Down, Assessing computational methods of *cis*-regulatory module prediction, PLoS Comput. Biol. 6 (12) (2010) 1001020.

[23] J. van Helden, B. André, J. Collado-Vides, Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies, J. Mol. Biol. 281 (5) (1998) 827–842.

[24] L.J. Zhu, R.G. Christensen, M. Kazemian, C.J. Hull, M.S. Enuameh, M.D. Basciotta, J.A. Brasefield, C. Zhu, Y. Asriyan, D.S. Lapointe, S. Sinha, S.A. Wolfe, M.H. Brodsky, FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system, Nucleic Acids Res. 39 (1) (2011) D111–D117.