Enrichment of transcriptional regulatory sites in non-coding genomic region

Wen Xue¹, Jin Wang¹,* , Zhirong Shen¹ and Huaiqiu Zhu²

¹The State Key Laboratory of Pharmaceutical Biotechnology, School of Life Science, Nanjing University, Nanjing 210093, China and ²The Center for Theoretical Biology, Beijing University, Beijing 100781, China

Received on July 22, 2003; revised on September 3, 2003; accepted on September 26, 2003
Advance Access publication January 22, 2004

ABSTRACT

Motivation: Over-represented k-mers in non-coding genomic regions often lead to identification of potential transcriptional regulatory sites (TRS). This phenomenon has been employed by many algorithms to predict TRS in silico. Yet, the improvement of these algorithms should be based on deeper understanding of the enrichment feature. To obtain a general distributional profile of TRS in different regions of genomes as well as in different genomes, we here performed a systematic analysis on the over-representation of TRS in intergenic regions and gene upstream regions of yeasts and viral genomes, and the distributional pattern of TRS in intergenic and intron regions of the Drosophila genome. We also explored the way to evaluate the accuracy of TRS consensus sequences by measuring their enrichment.

Results: To measure enrichment, a statistical background model was introduced by comparing TRS frequency in certain regions of genome to either the frequency in the whole genome or the frequency in exon region. This model was applied to different classes of non-coding genomic regions in four genomes. Most of the TRS were observed to be over-represented in the intergenic regions of the Saccharomyces cerevisiae, Schizosaccharomyces pombe and Epstein-Barr virus (EBV) genomes. The enrichment of S.cerevisiae TRS in the 600bp upstream region of genes was also significant. In Drosophila genome, TRS did not show enrichment in intergenic and intron regions when TRS frequency in the whole genome was taken as background, as we did in other genomes. However, when we took TRS frequency in exon region as background, over 70% TRS are over-represented in those two classes of non-coding regions.

In addition, databases and literatures often report inconsistent consensus sequences for the same TRS. For example, the binding sequence for GCR1 was reported to be ‘rgcttcwc’ (r = a, g; w = a, t) by TRANSFAC (Matys et al., 2003) and ‘cwttc’ by SCPD (Zhu and Zhang, 1999). A simple method to evaluate the accuracy of this kind of sequences will be valuable to TRS predicting programs based on pattern searching.

Enrichment can be measured by comparing the frequency of k-mers to the frequency of their one base mismatch neighbors.

Availability: Free programs are available at http://dii.nju.edu.cn/~xuewen/enrichment/
Contact: jwang@nju.edu.cn

INTRODUCTION

Over-represented k-mers in non-coding genomic regions are often associated with potential transcriptional regulatory sites (TRS) (Hutchinson, 1996). This helps to design algorithms to extract TRS from oligonucleotides that are enriched in non-coding intergenic regions or upstream promoter regions (Sinha and Tompa, 2002). However, some essential questions arise during the improvement of these algorithms, e.g. whether all the TRS are enriched in non-coding regions or to what extent they are enriched.

Most TRS predicting programs based on enrichment can yield satisfactory results in simple genomes like that of the yeast (Xie et al., 1999). Whether the distributional preference of TRS in non-coding regions is valid in large and complicated genomes like the fruitfly is still elusive.

Distributional preference of TRS in different classes of non-coding genomic regions has recently been unveiled by both experimental and computational effort. Brown’s lab proved that most in vivo DNA-binding sites of Rap1, a Saccharomyces cerevisiae transcriptional factor, are located in intergenic region of yeast genome (Lieb et al., 2001). Levy et al. (2001) discovered the enrichment of regulatory signals in conserved non-coding region of human and murine genome. A systematic analysis of TRS over-representation is required to integrate the results from various independent studies.

In addition, databases and literatures often report inconsistent consensus sequences for the same TRS. For example, the binding sequence for S.cerevisiae transcriptional factor GCR1 was reported to be ‘rgcttcwc’ (r = a, g; w = a, t) by TRANSFAC (Matys et al., 2003) and ‘cwttc’ by SCPD (Zhu and Zhang, 1999). A simple method to evaluate the accuracy of this kind of sequences will be valuable to TRS predicting programs based on pattern searching.

Enrichment can be measured by comparing the frequency of k-mers to the frequency of their one base mismatch neighbors.

*To whom correspondence should be addressed.
(Hampson et al., 2002), or by comparing the frequency of k-mers in the promoter region of co-regulated genes to their frequency over all genes (Van Helden et al., 1998). Other programs search for k-mers that occur more frequently in upstream regions than in the randomly chosen reference regions in the genome (Brazma et al., 1998). Here, we introduced a measure of enrichment that compares TRS frequency in different classes of non-coding regions to the frequency either in whole genome or in exon region. This algorithm also takes into account the diversity of the genomic nucleotide frequency to minimize the effect of the nucleotide bias in coding and non-coding genomic regions.

We studied TRS enrichment in the genomes of two yeasts, one virus and one fruitfly. More than 68% consensus sequences of yeasts’ and EBV TRS are over-represented in intergenic regions. Most S. cerevisiae TRS are also enriched in 600 bp upstream region. In Drosophila genome, we did not observe TRS enrichment in intergenic or intron regions over the background of the whole genome. However, the enrichment of TRS over the background of the exon region is significant. Further study revealed that intron region has a certain content of transcriptional regulatory signals and plays a role in the distribution pattern of TRS in the genome. A simple method to evaluate the accuracy of contradictory TRS sequences from different sources was produced from this study through measuring their enrichment in non-coding genomic regions.

METHODS

Scoring the enrichment

The enrichment of k-mer in non-coding intergenic region is measured by \( S_n \) (Hampson et al., 2002), the ratio of \( C_n \), the k-mer’s actual occurrence, with \( C_n(E) \), its expected occurrence if the k-mer displays a uniform distribution in the genome. \( S_n > 1 \) correlates with over-representation of k-mer, and the larger \( S_n \) is, the more significantly this k-mer is enriched.

\[
S_n = C_n / C_n(E) \quad (1)
\]

On a long single strand DNA with nucleotide percent \( P_a, P_g, P_c, P_t \) for the four nucleotides respectively, the theoretical frequency of k-mer with its sequence composition \( \Lambda_nG_nC_mT_n \) can be estimated as:

\[
f(T) = P_a^1 * P_g^1 * P_c^m * P_t^n \quad (2)
\]

In a model if the k-mer distributes randomly in the genome, the theoretical occurrence in non-coding region and the whole genome can be estimated as \( L_n * f(T_n) \) and \( L_g * f(T_g) \) respectively. \( L_n \) and \( L_g \) correspond to the length of these two regions. \( f(T_n) \) and \( f(T_g) \) are the theoretical frequency in the two regions. With the nucleotide percent in non-coding region, \( f(T_n) \) can be obtained with Equation (2). Because Equation (2) calculates the frequency on a single strand, we sum \( f_n(T) \) of the k-mer and \( f_n(T) \) of its reverse complement to get the total \( f(T_n) \) on double strands. \( f_g(T) \) is calculated similarly.

Consequently, if k-mer distributes uniformly in the genome, the theoretical ratio of k-mer occurrence in non-coding region against that of the whole genome is the ratio of \( L_n * f_n(T_n) \) and \( L_g * f_g(T_g) \). In this way, \( C_n(E) \) is calculated by Equation (3).

\[
C_n(E) = C_n * \frac{L_n * f_n(T)}{L_g * f_g(T)} \quad (3)
\]

This algorithm can be applied to measure the over-representation of k-mer in non-coding region, 600 bp upstream region, exon region or intron region over the background of TRS frequency in the whole genome. The enrichment scores are represented as \( S_n, S_{600}, S_{exon}, S_{intron} \) respectively.

A simple calculation proves that \( S_n / S_{exon} \) is equivalent to the measure of enrichment using the frequency in exon region as background [Equation (4)], which includes the theoretical ratio of k-mer occurrence in the non-coding region against the exon region.

\[
S_n / S_{exon} = C_n \left( \frac{L_n * f_n(E)}{L_{exon} * f_{exon}(E)} \right) \quad (4)
\]

Genomic data

Complete genome sequences of S. cerevisiae (accession nos NC_001133–NC_001148; Goffeau et al., 1996), Schizosaccharomyces pombe (accession nos NC_003421, NC_003423, NC_003424; Wood et al., 2002) and Epstein-Barr Virus (EBV) (accession no. NC_001345; Baer et al., 1984) were obtained from GenBank. Drosophila genome was taken from Berkeley Drosophila genome project (release 2.0; Adams et al., 2000). The indeterminate bases encoded with ‘N’ in S. pombe and Drosophila genomes were ignored in calculating sequence length and nucleotide percentage.

Statistics of TRS occurrence around translation start sites

Occurrence of TRS in −1.5 to 1.5 kb region around the start codon (ATG) of all genes in the S. cerevisiae genome was counted and plotted with respect to the distance to the start codon (0 being the translation start). Window length is 50 bp.

Definition of non-coding genomic regions

We extracted the coding sequences of all 6200 genes annotated in S. cerevisiae genome and combined the overlapping region. The rest of the genome sequence was collected as intergenic region. Intergenic region in the other genomes were generated in the same way except that intergenic region starts from the start codon (ATG) in yeasts and EBV genomes and from transcription start site (TSS) in Drosophila genome.

The majority of S. cerevisiae TRS in SCPD are located within 10–700 bp upstream of start codon (Zhu and Zhang,
Distribution of transcriptional regulatory sites

Earlier works used 600 bp as the length of upstream region for TRS prediction (Brazma et al., 1998). We extracted 600 bp upstream sequences of all genes as upstream region of interest. Overlapping sequences were combined and the intersection of these sequences with previously defined intergenic region was pooled as 600 bp upstream region.

To obtain exon region in Drosophila genome, all the exon sequences were collected and combined to eliminate overlapping region. Intron region was generated similarly.

TRS representation and pattern finding

Calculations were performed on 80 S. cerevisiae TRS consensus sequences from databases and literatures. Among them 47 were taken from SCPD, 20 from TRANSFAC and the other 13 from literatures (Simon et al., 2001; Iyer et al., 2001; Lieb et al., 2001; Lee et al., 2002). Some TRS have iterant entries from different sources. For example, there are three entries altogether for ACE2. One comes from SCPD and the other two from literatures. As a result, the collection of 80 sequences belongs to 49 kinds of TRS. All the TRS from literatures were determined by chromatin immunoprecipitation (ChIP) assays. To create a control set, we randomly extract 80 sequences (6–8 bp in length) from the S. cerevisiae genome and compare their enrichment with that of the 80 TRS.

Forty-three Drosophila TRS, 14 S. pombe TRS and 19 EBV TRS are collected from TRANSFAC and literatures (Berman et al., 2002; Carey et al., 1992; Lehman et al., 1998; Markstein et al., 2002). Random sequences are generated as in S. cerevisiae.

TRS are represented as consensus sequences in this study. We searched the genome for exact matches on both strand. Non-overlapping sites were counted so that palindromes and their reverse complement sequences were not counted iteratively.

RESULTS

Distributional pattern of TRS around the start codon of S. cerevisiae genes

Current biochemical model agrees that TRS regulate gene transcription from sites upstream of their target genes (Ptashne and Gann, 1997; Wang et al., 1999). However, since most TRS are reported to be 5–25 bp in length, these short sequences are discovered in both coding and non-coding genomic regions by computational search. A statistics of the occurrence of yeast TA TA box near the start codon reveals that this sequence distributes in both the upstream non-coding region and downstream coding region (Fig. 1A). Nevertheless, there is relatively more occurrence in upstream non-coding region than in coding region (Fig. 1A). Another yeast TRS, SCB, also shows over-representation in upstream region (Fig. 1B). Considering earlier reports that TRS distributed unevenly in the genome with preference in different classes of non-coding regions (Hutchinson, 1996), we set out to investigate the enrichment of a collection of 80 TRS in the intergenic and 600 bp upstream regions in the S. cerevisiae genome.

Over-representation of TRS in intergenic and 600 bp upstream region of yeast and EBV genomes

We measured the enrichment score $S_{nc}$ in intergenic region by calculating the ratio of the actual occurrence, $C_{nc}$, over the expected occurrence, $C_{nc}(E)$. In this model, $S_{nc} > 1$ refers to TRS enrichment in intergenic region over the background of TRS frequency in the whole genome.

We calculated the $S_{nc}$ of 80 S. cerevisiae TRS and the result showed that over 70% TRS have $S_{nc} > 1$ (Fig. 2). Less than 50% random sequences are enriched. The curve of TRS is above that of the random sequences. This result confirms that most TRS are enriched in intergenic region when compared to their frequency in the whole genome.

We next analyzed the enrichment of TRS in 600 bp upstream region of S. cerevisiae genome. $S_{600}$ was calculated to measure the over-representation of TRS and random sequences in this
Fig. 2. Enrichment analysis of *S. cerevisiae* TRS in intergenic region. $S_{nc}$ for 80 TRS are sorted and plotted according to their numerical value. The horizontal line at $S_{nc} = 1$ is drawn. Dots above the line ($S_{nc} > 1$) mean the corresponding TRS are enriched in intergenic region over the background of TRS frequency in the whole genome. Eighty random sequences (6–8 bp in length) extracted from the genome are plotted in the same manner as control set. Fifty-seven TRS (71%) and 34 random sequences (42%) show $S_{nc} > 1$.

Fig. 3. Enrichment analysis of *S. cerevisiae* TRS in 600 bp upstream region of gene. $S_{600}$ for 80 TRS and the same number of random sequences are plotted as in Figure 2. Sixty-three TRS (79%) and 31 random sequences (39%) have $S_{600} > 1$.

Table 1. Statistics of TRS sequences enriched in intergenic regions of four genomes

<table>
<thead>
<tr>
<th>Genomes</th>
<th>No. of sequences</th>
<th>Enriched TRS (%)</th>
<th>Enriched random (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>80</td>
<td>71</td>
<td>42</td>
</tr>
<tr>
<td><em>S. pombe</em></td>
<td>14</td>
<td>92</td>
<td>43</td>
</tr>
<tr>
<td>EBV</td>
<td>19</td>
<td>68</td>
<td>47</td>
</tr>
<tr>
<td><em>Drosophila</em></td>
<td>43</td>
<td>49</td>
<td>35</td>
</tr>
</tbody>
</table>

*Random sequences extracted from the corresponding genomes were used as control.

50% random sequences are enriched in this region. These calculations confirm that enrichment is significant in intergenic region of simple genomes.

**Distribution pattern of TRS in intergenic and intron regions of *Drosophila* genome**

Calculation of $S_{nc}$ for a collection of 43 *Drosophila* TRS revealed that less than 50% of TRS have $S_{nc} > 1$, as shown in Figure 4A and Table 1. It seems that a large proportion of TRS are not over-represented in intergenic region over the background of the whole genome.

To clarify if the under-representation of those TRS in intergenic region is caused by the longer intergenic region of *Drosophila* genome which have more repeated elements than its yeasts and EBV counterparts, we calculated $S_{600}$ for *Drosophila* TRS. However, the result turned out to be the same under-representation in this region (data not shown).

To exclude the influence of intron region in the *Drosophila* genome, we calculated the enrichment score of TRS in non-coding sequences by taking TRS frequency in exon region as background. In our model, $S_{nc}/S_{exon}$ means the comparison of TRS frequency in intergenic region to the frequency in exon region. Calculation of $S_{nc}/S_{exon}$ for 43 TRS reveals that 74% TRS have $S_{nc}/S_{exon} > 1$ and this ratio is significantly above that of the random sequences (Fig. 4B). This result manifests the enrichment of most TRS in intergenic region over the background of the exon region, which can not be observed on the background of the whole genome.

We then investigated the role of intron region in the distributional pattern of *Drosophila* TRS. Values of $S_{intron}$ showed insignificant enrichment of TRS over the background of the whole genome (Fig. 5A). However, when we calculated the enrichment in introns over the background of TRS frequency in exon region by calculating the ratio of $S_{intron}/S_{exon}$, 79% TRS have $S_{intron}/S_{exon} > 1$, which means that most of the *Drosophila* TRS distribute more frequently in intron region than in exon region (Fig. 5B). This data indicates the existence of transcriptional regulatory signals in intron region, which agrees with early reports that some functional transcriptional regulatory signals may reside in introns.
Fig. 4. Enrichment analysis of 43 Drosophila TRS in intergenic region. (A) $S_{nc}$ of TRS. Twenty-one TRS (49%) and 15 random sequences (35%) have $S_{nc} > 1$. Random sequences are 8–9 bp in length from the genome. (B) $S_{nc}/S_{exon}$ of TRS. Thirty-two TRS (74%) and 22 random sequences (51%) have $S_{nc}/S_{exon} > 1$.

From this analysis, we established the enrichment of TRS in intergenic or intron region over the background of TRS frequency in exon region. This feature can be used to device computational approaches for TRS evaluation in Drosophila genome.

Evaluation of TRS based on the enrichment in non-coding genomic region

Using the model above, we explored the possibility to evaluate the accuracy of yeast TRS consensus sequences by measuring their enrichment in intergenic region. Calculation of $S_{nc}$ showed that 23 S.cerevisiae TRS are not enriched in intergenic region. Eight of these un-enriched TRS were found to have alternative consensus sequences that do enrich in intergenic region (Table 2). These enriched consensus sequences were found to come from different sources. For example, $S_{nc}$ for the ACE2 consensus sequence reported by SCPD (Zhu and Zhang, 1999) is 0.68, while the other two ACE2 sequences determined by ChIP experiment (Simon et al., 2001) have $S_{nc} > 1$.

What is noticeable is that all 13 TRS determined by ChIP assays have the feature of $S_{nc} > 1$. As ChIP is a novel technique that reflects the in vivo binding site of transcriptional factors, it can provide more precise sequence information than the TRS sequences determined by traditional ways such as mutagenesis assay or affinity experiment (Iyer et al., 2001). We suspect that some of the under-represented TRS sequences might incompletely describe the transcriptional factor binding sites due to our lack of experimental data. Their validity requires further evaluation.

DISCUSSION

The asymmetric distribution of TRS in the genome was presented through a systematic analysis of TRS in four
calculated in estimating whole genome or in exon region. Nucleotide percentage is the frequency of TRS in non-coding region to the frequency in the whole genome.

The result from random sequences because they display a balanced tendency toward enrichment and under-representation using our algorithm. The random sequences, which have large enrichment scores, may be associated with potential regulatory signals.

Our background model works for pattern finding algorithms using simple k-mers and consensus sequences with indeterminate bases. Similar model can be developed for algorithms using position weight matrix (PWM). In this study, some of the consensus sequences contain indeterminate nucleotides. The searching for TRS based on low stringent consensus sequence may include more false positive counts than high stringent PWM (Hertz and Stormo, 1999) and Gibbs sampling method (Hughes et al., 2000). However, the most sophisticated algorithm to this date is not precise enough to fit the in vivo binding sites determined by experiments because chromatin structure and combinatorial partner transcriptional factors (TFs) will alter the binding preference of a certain TF to its consensus sequence (Iyer et al., 2001).

Although most known TRS consensus sequences are enriched in non-coding region, a large number of potential TRS sites are located in coding and exon regions. Whether these sites could bind TF and to what extent they will compete with the sites in non-coding region for TF binding is not clearly characterized. However, enrichment of TRS in non-coding regulatory region is reasonable to compete for TF activity and enhance regulation efficiency. Most biochemical theories propose that TF regulate transcription from sites that are upstream of the gene coding sequence. How TRS in introns regulate gene expression is not well understood. The finding of TRS occurrence in intron region imposes strong fascination upon future study both in bioinformatics and in molecular biology.

Our result proposed that computational methods for TRS evaluation could be based on TRS over-representation in intergenic or upstream genomic regions. Since many algorithms used for TRS prediction are based on pattern finding (Berman et al., 2002), a simple method to evaluate the accuracy of TRS will be of practical value. However, further study is required to prove the feasibility of this method.

Many algorithms search for TRS in upstream region of co-regulated genes. It will be interesting to study the enrichment of these signals in upstream region surrounding classes of co-regulated genes (Pilpel et al., 2001). On the other hand, since over-represented TRS is a distinct feature of non-coding genomic region, it is possible that this feature can be used to discriminate non-coding region from the rest of the genome.
so as to improve the accuracy of gene discovery and exon prediction tools in an era of large-scale genomic sequencing.

ACKNOWLEDGEMENTS

We thank Peter Markstein at the Hewlett-Packard Lab for valuable discussions. This work is supported by the National Science Foundation of China (Nos 90208021 and 39970184).

REFERENCES


