ELSEVIER

Knowledge-Based Systems 15 (2002) 243–250

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

A data mining approach to discover unusual folding regions in genome sequences

Shu-Yun Le[a,*], Wei-min Liu[b], Jacob V. Maizel Jr.[a]

[a]*Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, NIH, Building 469, Room 151, Frederick, MD 21702, USA*

[b]*Department of Computer and Information Science, Indiana University, 723 W. Michigan St., Indianapolis, IN 46202-3216, USA*

Received 15 March 2001; revised 4 May 2001; accepted 31 May 2001

Abstract

Numerous experiments and analyses of RNA structures have revealed that the local distinct structure closely correlates with the biological function. In this study, we present a data mining approach to discover such unusual folding regions (UFRs) in genome sequences. Our approach is a three-step procedure. During the first step, the quality of a local structure different from a random folding in a genomic sequence is evaluated by two *z*-scores, significance score (SIGSCR) and stability score (STBSCR) of the local segment. The two scores are computed by sliding a fixed window stepped a base along the sequence from the start to end position. Next, based on the *non-central Student's t distribution theory* we derive a linearly transformed *non-central Student's t distribution* (LTNSTD) to describe the distribution of SIGSCR and STBSCR computed in the sequence. In the third step, we extract these significant UFRs from the sequence whose SIGSCR and/or STBSCR are greater or less than a given threshold calculated from the derived LTNSTD. Our data mining approach is successfully applied to the complete genome of *Mycoplasma genitalium* (*M. gen*) and discovers these statistical extremes in the genome. By comparisons with the two scores computed from randomly shuffled sequences of the entire *M. gen* genome, our results demonstrate that the UFRs in the *M. gen* sequence are not selected by chance. These UFRs may imply an important structure role involved in their sequence information.

Keywords: Data mining; Statistical model; RNA/DNA folding; UFR

1. Introduction

Complete genomic sequence data are being accumulated at an unprecedented pace. A wide variety of computational methods for analyzing genomic sequences have been developed [1,2]. Most of the problems in these methods are essentially statistical. Computational analyses of the distinct sequence pattern can help to understand the structure and function of genomic sequences. The discovery of biological knowledge from sequence data consisting of bases A, C, G, T/U in biological databases, such as Genbank, is especially important in a post-genomics age.

RNA is a single-stranded conformationally polymorphic macromolecule with its nucleotide sequence identical to that of one of the DNA strands except for a base replacement of T to U. The RNA sequence often folds back on itself between complementary segments to form various local structures guided by Watson–Crick rules. In addition to the Watson–Crick A–U and G–C base pairs, wobble G–U base pairs also contribute to the thermodynamic stability of an RNA structure. It has been demonstrated that some structures folded by local RNA segments are functional elements of the control for gene regulations in different levels [3,4]. These functional elements are often closely associated with unusual folding regions (UFRs) where the folding free energy of the UFR is significantly lower than that expected by chance [5–12]. The development of an efficient data mining approach to extract these potentially functional structured elements in the sequence database is highly desirable.

Knowledge discovery of functional structured elements in a genomic sequence is an important step to reach our goal from genome data to biological knowledge. The thermodynamic stability of an RNA/DNA fragment in the genome is often measured by the free energy of the formation of the folded RNA/DNA segment. Based on accumulated data [3,4,13], UFRs in an RNA sequence are assessed by the two *z*-scores, significant score (SIGSCR) and stability score (STBSCR) [13,14]. SIGSCR signifies the difference

* Corresponding author. Tel.: +1-301-846-5576; fax: +1-301-846-5598.
E-mail address: shuyun@orleans.ncifcrf.gov (S.-Y. Le).

0950-7051/02/$ - see front matter
PII: S0950-7051(01)00146-0

of thermodynamic stability between a local, natural RNA fragment and the average of its randomly shuffled sequences. Similarly, STBSCR indicates the difference of the stability between a specific fragment at a given place and the average from all other fragments of the same size in the sequence. As an example of our data mining, we analyze the complete genome sequence of *Mycoplasma genitalium* (*M. gen*).

Our data mining approach consists of three steps. In the first stage, we compute SIGSCR and STBSCR by sliding a fixed window with a step of one base along the sequence from the start to end position. Our statistical analysis shows that the distributions of the two z-scores in the sequence do not follow a simple normal distribution. In order to obtain useful information from an extraordinarily large number of sample observations in the analysis, we have to derive a reliable statistical model to describe the distributions of the two z-scores. In the second step we develop a linearly transformed *non-central Student's t* statistical model to delineate the distributions of SIGSCR and STBSCR in the entire genomic sequence by means of a *non-central Student's t distribution* theory [15]. Statistical tests show that the linearly transformed *non-central Student's t distribution* (LTNSTD) is a good statistical model to describe the distributions of the two scores computed in the genome. In the last step, the significant UFRs that are either much more stable or unstable than expected by chance are discovered based on the derived, well-fitted LTNSTD.

As a comparison, we also compute the distributions of SIGSCR and STBSCR in the randomly shuffled sequence of the complete *M. gen* genome. Our results further demonstrate that the statistical extremes of UFRs are not selected by chance in *M. gen*. The UFRs in the genome may imply the biological functions of the primary sequence data and provide useful information in further searching for functional structured elements involved in the control of regulatory genes [5–12].

2. Mathematical background

2.1. SIGSCR and STBSCR of a folding segment

The quality of a local structure in a DNA/RNA sequence is often evaluated by the thermodynamic stability of the structured segment. The greater the free energy of the formation of the structure in negative numbers, the more stable the folded structure of a fragment. In this study, the biological information of such structured fragments in an RNA sequence is evaluated by SIGSCR and STBSCR of a local segment. SIGSCR and STBSCR are a standard z-score and given by

$$\mathrm{SIGSCR} = (E - E_{\mathrm{r}})/std_{\mathrm{r}}$$

and

$$\mathrm{STBSCR} = (E - E_{\mathrm{w}})/std_{\mathrm{w}}$$

where E is the folded lowest free energy computed from a local segment in the sequence, E_{r} is the sample mean and std_{r} is the sample standard deviation of the lowest free energies computed from folding a large number of randomly shuffled segments of the same size and same base compositions as the local segment. Similarly, E_{w} and std_{w} are the sample mean and standard deviation of the lowest free energies obtained by folding all segments of the same size that are generated by taking successive, overlapping, fixed length segments stepped one base at a time from the start to end position of the sequence [13,14].

2.2. Linearly transformed non-central Student's t distribution

The *non-central Student's t distribution* (NSTD) is an asymmetric continuous distribution with range $(-\infty. +\infty)$. It has two parameters [15]: the degree of freedom, f (a positive integer); and the non-centrality parameter, δ (a real number). Its probability density function (PDF) [16] can be expressed as

$$P(x;f,\delta) = \frac{1}{2^{(f+1)/2}\Gamma(f/2)\sqrt{\pi f}} \int_0^{\infty} y^{(f-1)/2} \times \exp\left[-\frac{1}{2}y - \frac{1}{2}\left(x\sqrt{\frac{y}{f}} - \delta\right)^2\right] \mathrm{d}y \quad (1)$$

where x is a random variable, y is an integration variable and Γ is the well-known gamma function. Its r-th moment about the origin [15] is

$$\mu'_r = \left(\frac{f}{2}\right)^{r/2} \frac{\Gamma((f-r)/2)}{\Gamma(f/2)} \sum_{j=0}^{r/2} \binom{r}{2j} \frac{(2j)!}{2^j j!} \delta^{r-2j}, (f > r). \quad (2)$$

To simplify the notation, we introduce

$$g(f) = \frac{\Gamma((f-1)/2)}{\Gamma(f/2)}. \quad (3)$$

Let the observed data, SIGSCR and STBSCR be $\{y_i, 1 \le i \le n\}$. Consider a linear transformation:

$$y_i = ax_i + b, a > 0. \quad (4)$$

Let x_i, $1 \le i \le n$ be distributed as an NSTD t, whose degree of freedom is f and non-centrality parameter is δ. For a given degree of freedom f, we estimate the parameters a, b and δ by assuming the sample mean, sample variance and sample coefficient of skewness (k) of variables x_i ($1 \le i \le n$) are equal to the mean, variance and coefficient of skewness of the NSTD, that are shown in the following three equations [15,23]:

$$\frac{\bar{y} - b}{a} = \sqrt{\frac{f}{2}} g(f)\delta, \quad (5)$$

$$\frac{s_y^2}{a^2} = \frac{f}{f-2}\left[1 + \delta^2\left(1 - \frac{f-2}{2}g^2(f)\right)\right], \quad (6)$$