GBA Manager: An Online Tool for Querying Low-Complexity Regions in Proteins

NIRMALYA BANDYOPADHYAY and TAMER KAHVECI

ABSTRACT

We developed GBA Manager, an online software that facilitates the Graph-Based Algorithm (GBA) we proposed in our earlier work. GBA identifies the low-complexity regions (LCR) of protein sequences. GBA exploits a similarity matrix, such as BLOSUM62, to compute the complexity of the subsequences of the input protein sequence. It uses a graph-based algorithm to accurately compute the regions that have low complexities. GBA Manager is a user-friendly web-service that enables online querying of protein sequences using GBA. In addition to querying capabilities of the existing GBA algorithm, GBA Manager computes the p-values of the LCR identified. The p-value gives an estimate of the possibility that the region appears by chance. GBA Manager presents the output in three different understandable formats. GBA Manager is freely accessible at http://bioinformatics.cise.ufl.edu/GBA/GBA.htm.

Key words: biochemical networks, databases, gene expression, gene networks, protein structure, sequence analysis.

1. INTRODUCTION

L ow-complexity regions (LCR) of a protein sequence are subsequences of that protein sequence with biased amino acid composition. The sources of LCRs can be attributed to three kinds of repeats: cryptic, tandem, and interspersed repeats. Let $\Gamma$ represent the alphabet of amino acids. Two letters from $\Gamma$ are similar, if their similarity score is more than a certain cutoff value according to some scoring matrix (e.g., BLOSUM62). Let $\Gamma^*$ be an amino acid sequence with letters from $\Gamma$. Let $x = \Gamma^*s_1\Gamma^*s_2\Gamma^*\cdots\Gamma^*s_k\Gamma^*$ be a protein sequence. We denote $s_1, s_2, \cdots, s_k$ as repeats of one another if the following three conditions satisfy: (1) $s_1, s_2, \cdots, s_k$ are similar sequences, (2) each $s_i$ is longer than a cutoff, and (3) each $\Gamma^*$ is shorter than a cutoff. Identifying and studying the LCRs is an important problem, as a significant percentage of the amino acids are part of LCRs, and more than 50% of the proteins have at least one LCR according to a statistical study.

In our earlier work (Li and Kahveci, 2006), we developed an algorithm, named Graph-Based Algorithm (GBA), that identifies the LCRs in a protein sequence. In that work, we proposed new complexity measures that takes a scoring matrix such as BLOSUM62 and the permutation of the amino acids into consideration. Thus, unlike other methods that identify LCRs, the complexity measure of GBA takes care of the length,
composition, and ordering of amino acids in the protein sequence. We refer the readers to the original GBA article for a discussion on other literature (Li and Kahveci, 2006). Briefly, GBA develops a graph model for a user-supplied protein sequence. This model can explain all LCRs regardless of the length and location of the LCRs. In this model, each vertex represents a pair of similar amino acids. Each edge joins two pairs of amino acids that can be combined together to create a longer repeat. GBA finds short subsequences by traversing this graph. Then it expands them to find longer subsequences that possibly contains full repeats with low complexities.

Although GBA identifies LCRs more accurately than existing code, its use was limited in practice for several reasons. First, the source code was available; however, running it on a local machine and transferring the results to a useful format required installing several additional software. Second, it did not report statistical confidence measure for the LCRs it found.

Here, we have addressed the above-mentioned problems. We developed a web service, called GBA Manager. This service provides a user friendly graphical interface for querying LCRs in protein sequences. GBA Manager also calculates the statistical significance of the identified LCRs. To do that, it computes the p-value of each LCR. The p-value of an LCR shows the probability of having a subsequence with complexity at most as much as that of LCR if the subsequence was constructed randomly with the same background distribution as the input protein sequence. We have used PHP and Java to develop the GBA Manager code. The program runs on an Apache web server on an AMD Opteron X2 machine with Ubuntu operating system. GBA Manager is available at http://bioinformatics.cise.ufl.edu/GBA/GBA.htm. Figure 1 shows two snapshots of GBA Manager.

**FIG. 1.** (a) A snapshot of Graph-Based Algorithm (GBA) Manager query interface. It shows both the text area and the file upload as the two options for providing input sequence for a query. It also has four parameters: the forget rate for R and NR matrix, and three threshold values. (b) Part of the result generated by GBA manager. The upper rectangular block contains the starting and ending positions of LCRs along with the p-values. The lower rectangular block underlines the LCRs in the query sequence.
2. SYSTEM ARCHITECTURE AND FEATURES

In this section, we discuss the architecture details and the features of GBA Manager. In Section 2.1, we briefly discuss the GBA algorithm. In Section 2.2, we discuss the computation of the p-value of an LCR. We describe the architecture of web application in Section 2.3.

2.1. Graph based algorithm

The GBA applies to a single protein sequence. It consists of four main steps:

1. GBA starts by constructing a directed, acyclic and unweighted graph for each query protein. This includes both constructions of vertices and edges. It takes three thresholds parameters—\( t_1 \), \( t_2 \), and \( t_3 \)—as input. These thresholds define the properties of the repeat patterns that can be found from this graph in following steps.
2. The vertices joined by edges represents short repeats that can be grouped to make longer repeats. GBA finds the longest path in every connected subgraph to get the longest repeating pattern.
3. GBA discards the short intervals found in the last step and extends the remaining ones in left and right direction to find longer intervals containing full repeats with low complexities.
4. Extended intervals may contain non-LCRs. GBA applies a post-processing strategy to filter the non-LCRs with higher complexities.

![FIG. 1.](image)
2.2. p-Value of low-complexity region

The original GBA algorithm finds the LCRs and reports their complexities. GBA Manager takes this one step further and computes a statistical measure, p-value, for each reported LCR. We define the p-value of an LCR as follows. Let us denote the LCR and the sequence that contains this LCR using $x$ and $s$, respectively.

We define a random variable $Y$ as the complexity of a sequence that is generated randomly with the same background distribution of amino acids as $s$, and has the same number of amino acids as $x$.

Let us denote the complexity of $x$ with $c$. p-value of $x$ is the probability that value of $Y$ is at most $c$ (i.e., $\text{Prob}(Y\leq c)$).

To calculate the p-value of an LCR, we need to compute the cumulative distribution function of $Y$. The random variable $Y$ follows an inverse Gaussian distribution. We use Monte Carlo simulations to learn the parameters that define the distribution of $Y$ as follows. We generate 500 random sequences, each having the same number of amino acids as $x$. We do this by randomly generating amino acids with Bernoulli trials with the frequency of the amino acids the same as that in $s$. We compute the complexity of all these random sequences using 2-gram normalized complexity measure. Let us denote the mean and standard deviation of these complexities using $\mu$ and $\sigma$, respectively. We compute the p-value of $x$ as

$$\frac{1}{2}(1 + \frac{2}{\sqrt{\pi}} \int_0^{v_1} e^{-t^2} dt) + e^{\frac{\nu_2^2}{2}} \frac{1}{2}(1 + \frac{2}{\sqrt{\pi}} \int_0^{v_2} e^{-t^2} dt),$$

where

$$\nu_1 = (\frac{c}{\mu} - 1)\sqrt{\frac{\mu^3}{2c\sigma^2}} \quad \text{and} \quad \nu_2 = -(\frac{c}{\mu} + 1)\sqrt{\frac{\mu^3}{2c\sigma^2}}.$$

2.3. Web architecture

GBA Manager takes protein sequences only in Fasta format. The user can print one or more protein sequence on the text box or he can upload an input file. There are four parameters to fine tune the algorithm:

- **Forget rate for R and NR matrices ($z$):** The algorithm uses this parameter to calculate Repeat (R) and Non-Repeat (NR) matrices. These two matrices are used to filter the vertices constructed for similar letter pairs when they are in the same window by chance (i.e., false positives). R and NR are 20x20 matrices show statistical information on frequencies of pairs of letters in repeat and non-repeat regions, respectively. The values $R(\gamma_i, \gamma_j)$ and $R(\gamma_i, \gamma_j)$ represent average probabilities that $\gamma_i$ and $\gamma_j$ appear together in a repeat region and non-repeat region, respectively. Forget rate ($z$) is used as a parameter to calculate R and NR ($0 \leq z \leq 1$). The default value of $z$ is 0.95.

- **Window size ($t_1$):** This parameter specifies the length of the sliding window. Large values of $w$ can find repeats that are far away, at the expense of increased chance of false positives. The default value of $t_1$ is 15.

- **Indel thresholds ($t_2, t_3$):** These two thresholds determine a bound on the number of gaps for tandem and cryptic repeats. The default values of $t_2$ and $t_3$ are 3 and 5, respectively.

The GBA Manager returns results in a few seconds for small number of proteins. However, if the user wants to submit a very large number of proteins at once, and is unwilling to wait in front of the browser for results, s/he can optionally provide an email address. In this case, the GBA Manager sends a link to the query result which remains available for one month after the result is created.

GBA Manager reports the results in three different formats:

- **Masked LCRs:** The LCRs are masked with letter “X.” The user can simply copy these results and use for further analysis such as BlastP queries.

- **Range of LCRs:** The starting and ending position of each LCR is reported along with their p-values.

- **Underlined LCRs:** The LCRs are underlined. This allows the user to visually inspect the results for manual curation.

If the query contains more than one protein sequences, GBA Manager reports the results of them one by one.
DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCE


Address correspondence to:
Dr. Tamer Kahveci
Department of CISE
CSE Building, Room E436
University of Florida
Gainesville, FL 32611

E-mail: tamer@cise.ufl.edu