Video Article

Genomic MRI - a Public Resource for Studying Sequence Patterns within Genomic DNA

Ashwin Prakash[1], Jason Bechtel[1], Alexei Fedorov[1]

[1]Department of Medicine, University of Toledo Health Science Campus

Correspondence to: Alexei Fedorov at alexei.fedorov@utoledo.edu

URL: https://www.jove.com/video/2663
DOI: doi:10.3791/2663

Keywords: Genetics, Issue 51, bioinformatics, computational biology, genomics, non-randomness, signals, gene regulation, DNA conformation

Date Published: 5/9/2011

Citation: Prakash, A., Bechtel, J., Fedorov, A. Genomic MRI - a Public Resource for Studying Sequence Patterns within Genomic DNA. *J. Vis. Exp.* (51), e2663, doi:10.3791/2663 (2011).

Abstract

Non-coding genomic regions in complex eukaryotes, including intergenic areas, introns, and untranslated segments of exons, are profoundly non-random in their nucleotide composition and consist of a complex mosaic of sequence patterns. These patterns include so-called Mid-Range Inhomogeneity (MRI) regions -- sequences 30-10000 nucleotides in length that are enriched by a particular base or combination of bases (e.g. (G+T)-rich, purine-rich, etc.). MRI regions are associated with unusual (non-B-form) DNA structures that are often involved in regulation of gene expression, recombination, and other genetic processes (Fedorova & Fedorov 2010). The existence of a strong fixation bias within MRI regions against mutations that tend to reduce their sequence inhomogeneity additionally supports the functionality and importance of these genomic sequences (Prakash *et al.* 2009).

Here we demonstrate a freely available Internet resource -- the *Genomic MRI* program package -- designed for computational analysis of genomic sequences in order to find and characterize various MRI patterns within them (Bechtel *et al.* 2008). This package also allows generation of randomized sequences with various properties and level of correspondence to the natural input DNA sequences. The main goal of this resource is to facilitate examination of vast regions of non-coding DNA that are still scarcely investigated and await thorough exploration and recognition.

Video Link

The video component of this article can be found at https://www.jove.com/video/2663/

Protocol

All the used programs in the paper have been written using perl, and all the web pages have been created using PHP.

1. Starting Point:

Open the home page of the online Genomic MRI package at *http://mco321125.meduohio.edu/~jbechtel/gmri/* . The web resource also provides instructions/explanations on the programs in the "*Help (How-to/README)*" link, while all published materials on *Genomic MRI* and similar algorithms are listed in the "*Links to relevant resources*" link.

2. Preparation and Uploading of Input Sequence(s).

Create a file with FASTA-formatted sequence(s) to start a GMRI analysis session. Each nucleotide sequence in this format should be preceded with a single line starting with the ">" character that represents an identifier, followed on the same line by a short description of this sequence. Nucleotide sequences for GMRI analysis also permits characters like R, Y, N, X, etc. Hwever, non-A, T, C, G characters will not be processed by the program and will be skipped. Sequences in which repetitive elements have been "masked" (replaced by "N"s) can be used as input. Note that sequence characters are case insensitive.

1. Begin a GMRI session by clicking on the "***Start or Resume***" button on the *Genomic MRI* home page. This takes the user to a page where nucleotide sequences can be uploaded.
2. Copy-and-paste your FASTA-formatted sequences or upload a file containing the sequences from your local computer using the "***choose file***" button.
3. Click on the "***start new session with this file***" button. A confirmation message should appear above the input window stating that "*Your sequence has been successfully uploaded*" and you should also get an alphanumeric "GMRI identifier" [the site calls it a "session label"] for your session (e.g. b16yMj), which can be used to retrieve and continue a session for up to two weeks after first use.

NOTE: Henceforth the input sequences are referred to as "userfile".

3. Get an Oligonucleotide Frequency Distribution of the Input Sequences (optional).

Click on the "***SRI Analyzer***" tab (top row) in order to get a distribution of oligonucleotide frequencies for the entire set of input sequences. The acronym SRI stands for short-range inhomogeneity. At this juncture, the user may specify the highest length of oligonucleotides (from 2 up to 9 nucleotides, default 6 nts) for which frequencies will be calculated. This selection is made by clicking on the desired option within the "***Maximum oligomer size***" list box. Then press the "***Analyze File***" button to initiate computation. A rough representation of the input sequence composition will immediately appear as a short table in the middle of this web page and downloadable as "*userfile.comp.tbl*". This table represents only the most and the least abundant oligonucleotides within the input sequences.

The entire frequency table for all possible oligonucleotides is generated as a file named "*userfile.comp*", which can be obtained via the "***Download composition file***" link.

NOTE: SRI analyzer counts the entire set of all overlapping oligonucleotides.

4. Generate Random Sequences with the Same Oligonucleotide Composition As in the Input Sequences (optional).

(Completion of step 3 of the protocol is required for this task).

1. Click on the "***SRI Generator***" tab (top row) to open up a new web page that creates random sequences. Choose the number of samples of random sequences to be generated using the list box on this web page. Each of these sample files will contain random sequences of the same number and length as the input sequences in "*userfile*". Moreover, if an input sequence contains non-A, T, C, or G characters, the random sequence will have "N"s at exactly the same positions as in the input sequence.
2. Choose the longest length of oligonucleotides for which frequencies will be approximated in the random sequences. This can be chosen by checking the radio button for the desired oligomer level (e.g. "***4-mers"*** for four-base oligonucleotides) in the table at the center of the screen. It is to be noted here that random sequences will consist of not only the approximate frequencies at the chosen oligomer level, but also the corresponding frequencies of shorter oligomer levels, as in the input sequences. Small fluctuations in the oligonucleotide frequencies of input and random sequences are possible due to the Markov Model procedure applied for the generation of random sequences.
3. Start the program by clicking the "***Generate File***" button. If the input sequences are large it could take a couple of minutes to generate random sequences. Thus, a user should wait until blue "Download" links appear at the bottom of this page. The random sets are placed in files with names such as "*userfile.randX_Y*" where *X* is the number of the random set and *Y* is the chosen oligomer level (e.g. "*userfile_rand2_4*").

5. Analysis of Mid-Range Inhomogeneity (MRI) of Input and Random Sequences.

1. Click on the "***MRI Analyzer***" tab (top row), which opens up a new web page that analyzes the mid-range inhomogeneity of the nucleotide composition of sequences.
2. Select a sequence to be analyzed from the "***File to analyze***" list box (a choice between the input sequence and generated sets of random sequences can be made here).
3. Choose the ***content type*** of MRI to be analyzed via the provided list box. (Seven content options are available: G+C; G+A; G+T; A; G; C; or T.)
4. Choose the length of the window for which content-rich and content-poor sequences will be examined via the ***"Window size"*** list box (default is 50 nucleotides; the valid range is from 30 to 1000).
5. Choose the ***upper threshold*** and ***lower threshold*** for content-rich and content-poor regions, respectively. These thresholds can be defined by the exact number of particular nucleotides in the current window (using the ***by number*** option in the list box) or by percentage of these nucleotides in the window (using the ***by percentage*** option)
6. After all five choices have been made (for example: Sequence = "*userfile*"; Content = *GC*; Window size = *50*; Upper threshold = *35*; Lower threshold = *15*), invoke the program by pressing the ***Analyze File*** button. The program scans through all sequences from the selected input consecutively. At each step it obtains a segment of the current sequence with length equal to the specified window size and computes whether the number or percentage of nucleotides of the chosen content is above the upper threshold or below the lower threshold. If the window does not match either criteria, the next overlapping window (shifted by one nucleotide) is selected for the same analysis. When a window is found where the sequence meets one of the threshold requirements for content-rich or -poor composition, the program saves the sequence of this window in the output file and generates a spike on the graphical output. After this, the program jumps to the next non-overlapping adjacent window and resumes the scanning process until the end of the sequence is reached.
7. After completion of the program, a link to the output file (with name "*userfile_GC_50_35..15*" for the example above) appears and a graphical representation of the results is displayed in the middle of the web page (see Figure 1). On this graphical display all input sequences from the *userfile* are concatenated into a single string and presented as a horizontal black line on the X axis, with length in kilobases (kb) shown below. All content-rich regions along input sequences are marked as blue "upward" spikes, and content-poor regions as red "downward" spikes. The total numbers of content-rich and content-poor windows are shown in parenthses in the legend at the bottom of this figure (32 and 19, respectively). The figure serves to illustrate the relative abundance and the arrangement of MRI regions. Meanwhile specific details are presented in the output file (see Figure 3). In this file, all nucleotide sequence segments that match content-rich or -poor criteria and their coordinates are available to a user as a list according to their consecutive positions along the input file.
8. After completion of MRI analysis for the chosen sequence a user can start a new process in the same web page by making changes to parameters and/or input files. For example, in order to examine the previously generated random sample #1 with the same MRI parameters, the user only needs to change the ***File to analyze***option and select the "*userfile_rand1_4"* file, and then press the ***Analyze File*** button again. A new file and graphical display will replace the old one. The results and figures of ALL examinations under each "session lable" (GMRI

identifier) will be saved and be available for two weeks from the last activity. In order to save the results/figures permanently, the user should select the "***Download Files***" tab (top row) and download the entire session or individual files, as needed.

9. With this ***MRI Analyzer*** web page a user can study
   - (G+C)-rich and (A+T)-rich regions
   - Purine (A+G)-rich and Pyrimidine (C+T)-rich regions
   - Keto (G+T)-rich and aMino (A+C)-rich regions
   - A-rich and A-poor regions
   - G-rich and G-poor regions
   - T-rich and T-poor regions
   - C-rich and C-poor regions

10. The latest release of *Genomic MRI* has a new option for studying regions rich with Purine(R)/Pyrimidine(Y) alternation patterns that might form Z-DNA conformations. Currently, this option is available from the link "***Z-DNA***" and it works on the same basis as other aforementioned MRI regions. A user should select upper and lower thresholds for the number of (RY+YR) overlapping dinucleotides in the scanning window. The program produces a similar graphical output and a file of DNA segments enriched and depleted by alternating purines and pyrimidines. The putative Z-DNA regions must be highly enriched by alternating R/Y bases (see review F&F 2011).

6. Additional Programs Within the *Genomic MRI* Package (optional).

The *Genomic MRI* resource also has two advanced options for generation of very specific random sequences. They are available through the "***MRI Generator***" and "***CDS Generator***" tabs in the top row.

1. ***MRI generator*** creates randomized sequences with the same oligonucleotide composition as the input file (similar to ***SRI generator***). However, in addition, randomized sequences mimic a particular MRI pattern specified by the user. Within this web page a user should specify from a list box a particular MRI pattern to be mimicked. The list box contains all patterns that have been examined in this session by ***MRI analyzer*** (e.g. "*userfile_GC_50_35..15").* A random sequence generated with this option will have the same oligonucleotide composition as the selected input file and also the same GC-rich and -poor patterns as seen in "*userfile_GC_50_35..15".*
2. ***CDS generator*** is used for randomization of protein coding sequences. It preserves the same amino acid sequence as the one coded by the user-specified input. In addition the program retains the same codon and di-codon biases as specified in the user-chosen input table. The online version of the ***CDS generator*** also accepts a protein sequence as an input. All other options for the program are offered only via stand-alone Perl scripts available for download from the main Genomic MRI web page.

7. Representative Results

This protocol allows a user to study compositional inhomogeneity of nucleotide sequences. Importantly, it also supports the generation of a variety of randomized sequences with an oligonucleotide composition approximating that of the input sequences. Usually, genomic sequences of complex eukaryotes are not homogeneous in composition, but rather represent a complex mosaic of sequence segments enriched by particular nucleotides (for example, purine-rich, (G+T)-rich, (A+T)-rich, etc.). These patterns at mid-range scale (30-1000 bp) are visualized by the graphical output of ***MRI analyzer*** that shows selected content-rich segments as upper blue spikes and content-poor segments as lower red spikes (see Figures 1 and 2). Typically, the number of any content-rich and content-poor regions in a natural sequence (Figure 1) is on the order of times higher than the number of the same types of regions in corresponding randomized sequences (Figure 2) having the same oligonucleotide composition. These sequence segments with mid-range inhomogeneity in nucleotide composition may be of interest to the user. They are available from the *Genomic MRI* output files for further investigation.

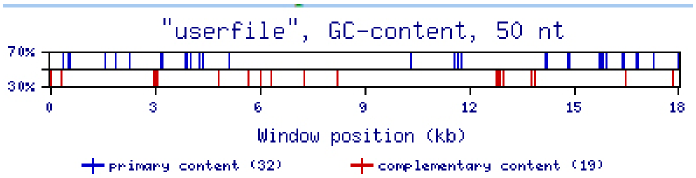


Figure 1. An example of the *MRI analyzer* graphical output from step 5.7. The results have been obtained on a sample of 44 human introns. Blue bars represent positions of GC-rich regions along these introns. Red bars represent GC-poor (or AT-rich) MRI regions. The y-axis contains upper and lower thresholds for the given content type.

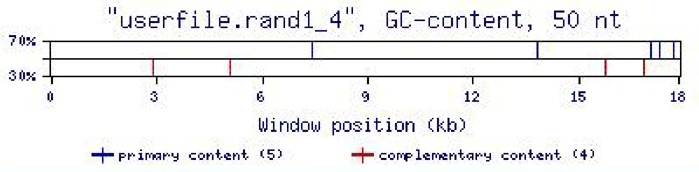
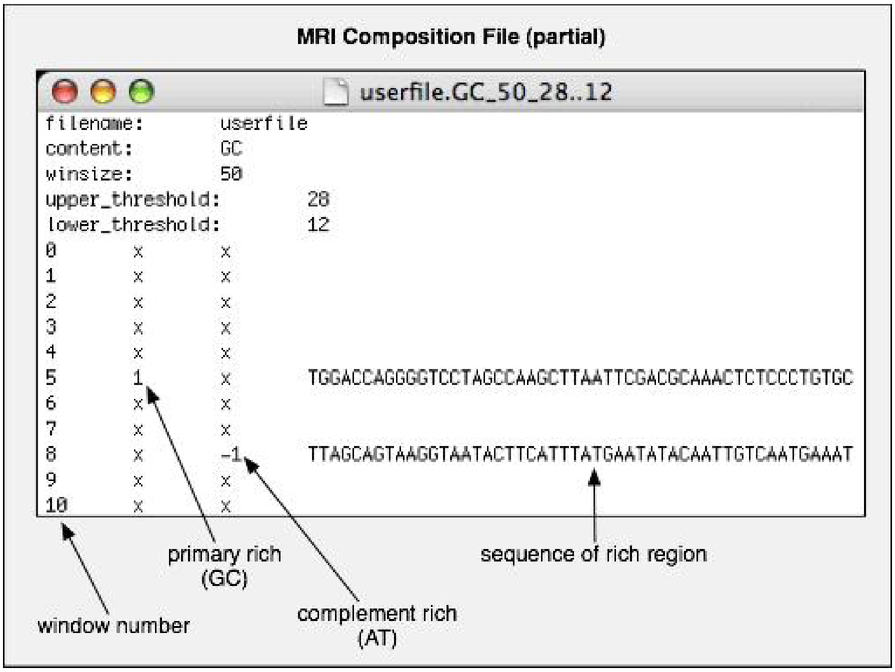


Figure 2.*MRI analyzer* output for the random sequence "userfile.rand1_4".
The graphical representation of MRI within a randomly generated sequence using the SRI generator program.



MRI Composition File (partial)

userfile.GC_50_28..12

```
filename:       userfile
content:        GC
winsize:        50
upper_threshold:        28
lower_threshold:        12
0       x       x
1       x       x
2       x       x
3       x       x
4       x       x
5       1       x       TGGACCAGGGGTCCTAGCCAAGCTTAATTCGACGCAAACTCTCCCTGTGC
6       x       x
7       x       x
8       x       -1      TTAGCAGTAAGGTAATACTTCATTTATGAATATACAATTGTCAATGAAAT
9       x       x
10      x       x
```

Figure 3. An example of the beginning of a textual output file from *MRI analyzer*.
All content-rich and content-poor sequences detected by the program are presented in the last (fourth) column. Their relative positions, measured in the number of windows, are shown in the first column. The second and third columns are indicators for content-rich and content-poor regions, respectively.

Discussion

Regions with inhomogeneous nucleotide composition at mid-range scales (30-1000 nucleotides) are overabundant in the genomes of complex eukaryotes and can be found anywhere (intergenic regions, introns, untranslated regions of exons, repetitive elements). These regions are frequently associated with unusual DNA conformations. For instance, purine-/pyrimidine-rich sequences tend to form DNA triplexes (H-DNA); sequences with alternating purine/pyrimidine bases are associated with Z-DNA conformations; (G+C)-rich regions exhibit structural abnormalities in B-DNA and could be prone to backbone cleavage; (A+T)-rich regions might form an unusual structure - a DNA unwinding element; etc. (reviewed by Fedorov & Fedorova 2010). Some of these mid-range patterns (e.g. (G+T)-rich regions) are scarcely investigated and still await thorough exploration and recognition. The main aim of our *Genomic MRI* web resource is to help users in the identification of these MRI regions for their further experimental analysis and for exploration of their possible functions. Knowledge of MRI regions could be incorporated into and improve the new generation of gene predictor programs (Shepard 2010) and advance our understanding of genome functions and properties.

Disclosures

No conflicts of interest declared.

Acknowledgements

We are thankful to Samuel Shepard, Peter Bazeley, and John David Bell for the administration of the *Genomic MRI* web pages. This work was supported by National Science Foundation Career award "Investigation of intron cellular roles" [grant number MCB-0643542].

References

1. Bechtel, J.M., Wittenschlaeger, T., Dwyer, T., Song, J., Arunachalam, S., Ramakrishnan, S.K., Shepard, S., Fedorov, A. Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures. BMC *Genomics* 9 : 284 (2008).
2. Prakash, A., Shepard, S., Mileyeva-Biebesheimer, O., He, J., Hart, B., Chen, M., Amarachiniha, S., Bechtel, J., Fedorov, A. "Molecular forces shaping human genomic sequence at mid-range scales", *BMC Genomics* 10 : 513 (2009).
3. Fedorov, A., Fedorova, L. "An Intricate Mosaic of Genomic Patterns at Mid-range Scale" Chapter 3, pp. 65-91, In "Advances in Genomic Sequence Analysis and Pattern Discovery" in print (2010).
4. Shepard, S. S. "Binary-abstracted Markov models and their application to sequence classification", chapter 4, pp 75-157, In PhD thesis "The characterization and utilization of middle-range sequence patterns within human genome", The University of Toledo (2010).