Characterization of Genomic Mid-Range Inhomogeneity

Submitted by: Jason Bechtel

In partial fulfillment of the requirements for the degree of Master of Science in Biomedical Sciences

Examination Committee

Major Advisor: Alexei Fedorov, Ph.D.  _____________________________

Academic Advisory Committee:
  Robert Trumbly, Ph.D.  _____________________________
  Robert Blumenthal, Ph.D.  _____________________________
  Neocles Leontis, Ph.D.  _____________________________

Senior Associate Dean
College of Graduate Studies
  Michael S. Bisesi, Ph.D.  _____________________________

Date of Defense: May 28, 2008
Characterization of Genomic Mid-Range Inhomogeneity

Jason M. Bechtel

University of Toledo -- Health Science Campus

2008
Acknowledgments

I would like to thank my research advisor, Dr. Alexei Fedorov, for giving me the opportunity to take a leading role in this research and for his instruction, patience and guidance. He has trained me to think as a biologist. The other members of my committee, Dr. Robert Blumenthal, Dr. Robert Trumbly, and Dr. Neocles Leontis, also deserve to be acknowledged for their patience and guidance. Dr. Sadik Khuder, while not a member of my committee, has always been a source of support and reassurance and I was pleased and honored to have him represent the faculty at my defense.

I am indebted to my co-authors for their work both on the manuscript and with the original research, and especially Tom Wittenschlaeger for his elegant CDS Generator code.

I must also acknowledge the support, encouragement and hard work of everyone in the BPG lab, but especially: Peter Bazeley, for maintaining the Linux cluster and the lab, but more so for his steadfast guidance and support and for helping me to think things through; Samuel Shepard, for his work on the manuscript, the GMRI website, the algorithm development, coding and code review, and for his support and friendship.

Finally, I am most grateful to my family and friends whose love and support have carried me through.
# Table of Contents

Acknowledgments............................................................................................................................................................iii
Table of Contents..............................................................................................................................................................iv
Introduction........................................................................................................................................................................1
Literature.............................................................................................................................................................................3
Manuscripts.........................................................................................................................................................................13
  Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures............................................14
Materials..............................................................................................................................................................................52
Methods..............................................................................................................................................................................52
Results................................................................................................................................................................................59
  The Genomic MRI Internet Resource.................................................................................................................................59
  MRI and Strong Local Secondary Structures.....................................................................................................................73
Discussion..............................................................................................................................................................................86
Conclusions..........................................................................................................................................................................91
Summary...............................................................................................................................................................................92
References.............................................................................................................................................................................93
Abstract...............................................................................................................................................................................99
Introduction

Genomes consist of long strings of nucleotides. The particular arrangement of these nucleotides serves many biological functions at multiple levels. The sequence can specify RNA structures, DNA binding sites, transcription initiation sites, splicing signals, polyadenylation sites, amino acid sequences, intracellular localization motifs, and much more. Above the sequence level, the bulk properties of the nucleotides can direct transcriptional availability (Antequera 2003; Bernardi 2000), nucleosome positioning (Kogan et al. 2006, Segal et al. 2006), static potential (Polozov et al. 1999) and recombination (Forsdyke 1995). To the extent that the DNA substrate codes for these signals and properties, this informational content is detectable as non-randomness or inhomogeneity in the distribution of the four nucleotide bases in the linear genomic sequence.

Prior studies of the informational content of genomes have focused mostly on the extremes: the relative frequency of occurrence of small oligonucleotides in any genome--what we call short-range inhomogeneity (SRI)--and the so-called "isochore" structure of most amniote genomes--what we call long-range inhomogeneity (LRI) (Bernardi 2000; Karlin et al. 1998; Press and Robins, 2006). Short-range inhomogeneity corresponds to various "preferences" that a genome has acquired during its evolution. These short-range
correlations tend to most strongly reflect the influence of overrepresented (and underrepresented) dinucleotides, such as CpG islands (Antequera 2003; Hackenberg et al. 2006; Karlin et al. 1998; Kozhukhin and Pevzner 1991). In coding sequences, SRI is dominated by coding signals (codon bias, dicodon bias, and amino acid bias) (Eskesen et al. 2004; Fedorov et al. 2002). However, short-range correlations can also reflect signals for DNA flexibility (to facilitate nucleosome wrapping) (Kogan et al. 2006; Segal et al. 2006) and the presence of a particular oligonucleotide can be an indicator of a retroviral pathogen or transposon (Laprevotte et al. 2001). Long-range inhomogeneity has been described as the "mosaic" structure of eukaryotic genomes, referring to the patchwork arrangement of the genome in large (on the order of $10^5$-$10^6$ base pairs) sections of relatively uniform consistency in multiple diverse properties (isochores) (Bernardi 2007). Isochores have been related to gene density, intron and UTR size, rate of transcription (chromatin structure), rate of recombination, replication timing, G+C content (i.e. the overall proportion of a segment of nucleotides consisting of guanine (G) and cytosine (C)) and G+C heterogeneity (Bernardi 2007; Melodelima et al. 2006). Because G+C content is a convenient property for direct measurement and bioinformatic analysis, isochores are commonly defined in terms of this measure. As discussed in Oliver et al. (2002), sliding window analyses have drawbacks for identification of isochores. The researchers in that paper identified isochores using a "hierarchical segmentation" method capable of locating
isochores at single base pair resolution (Oliver et al. 2002).

One recent study (Bechtel et al., pending publication, manuscript included below) found a two-fold to 10-fold enrichment of G+C-rich and A+T-rich "islands" in all parts of the human genome, including exons, introns, 5'-UTRs, 3'-UTRs and intergenic regions, by searching for putative strong local RNA secondary structures. The size of these islands, being based on the approximate size of an RNA hairpin with sufficient strength to overcome single-stranded RNA (ssRNA)-binding proteins, is, by necessity, larger than 20 nucleotides and yet much smaller than an isochore. Prompted by these and other findings, the authors hypothesized the existence of a mid-range inhomogeneity (MRI) throughout the human genome, henceforth referred to as genomic MRI (GMRI). This thesis seeks to expand upon that initial discovery by characterizing these hypothetical genomic features.

**Literature**

The literature is replete with studies of genomic content, organization and structure, but almost all of these investigations can be categorized as dealing exclusively with either SRI or LRI. Studies of SRI are usually conducted with methods involving straightforward statistical measures of oligonucleotide frequency (Karlin et al. 1998; Kozhukhin and Pevzner 1991). These methods cannot produce meaningful results when the number of occurrences of each oligonucleotide drops into the single digits. Thus, these studies are
practically limited to a range from one to approximately 15 nucleotides, the point beyond which each oligonucleotide can be expected to be statistically unique in a sufficiently large genome.

Support for the biological significance of this limit can be inferred from the highly conserved lower boundary on the size of small non-coding RNAs involved in RNA interference (RNAi). For example, microRNAs (miRNAs) and small interfering RNAs (siRNAs) encode their targets with no fewer than 16 nucleotides, and yet they are highly specific (Carrington and Ambros 2003). MicroRNAs have been found in all forms of life on this planet from viruses, slime molds and protists to plants and animals (Griffiths-Jones 2004; Griffiths-Jones et al. 2006; Griffiths-Jones et al. 2008) and do not violate this 16 nt lower boundary on size. It is logical that miRNAs and siRNAs must be highly specific because they have the potential to completely silence gene expression for any mRNA matching their sequence. The scope of evolutionary conservation of this information-coding minimum confirms the biological significance of a general 16 nt minimum for oligonucleotide uniqueness. Because the minimum length for high specificity is roughly proportional to total genome size, and over 99 percent of miRNAs are at least 20 nt in length, a safe approximation for the beginning of mid-range scale is 20 nt.
Most studies of LRI are concerned with very large scale non-randomness, on the order of hundreds of thousands to millions of bases. The precise definition of what constitutes an isochore is still a matter of some dispute, but there is general agreement that they are not smaller than 100,000 nucleotides (Melodelima et al. 2006; Nekrutenko and Li 2000; Oliver et al. 2002). One study of LRI (Häring and Kypr 2001), however, mentions phenomena that extend down to the range of hundreds of nucleotides. As this overlaps with the mid-range scale, this warrants careful attention.

Häring and Kypr (2001) examined nucleotide distributions using a 1 kb window and nucleotide correlations using a 10 kb window. They then used k-means clustering to identify two groups of segments in human chromosomes 21 and 22. Their segments were mostly (80%) explained by the correlations of A with T and C with G. Importantly, these segment types did not differ significantly in G+C content or (G−C/G+C) skew, or in the occurrence of genes. This indicates that their features are not only different from isochores, but that they are also different from the MRI phenomenon reported herein. Nevertheless, their suppositions regarding the source of their features are potentially applicable to MRI. They suggest that "base composition causes a base substitution bias." Indeed, this phenomenon has been observed experimentally and is a plausible mechanism by which MRI may be maintained.
Some studies encompass all scales at once. These tend to use methods such as "detrended fluctuation analysis" and a "Brownian walk" to uncover relationships such as "power law correlations" and "exponential decays," which assess the "scaling behavior" of a system. This scaling behavior is related to fractal geometry and deals with "self-similarity," defined as the property of resembling a subset of oneself. Earlier investigations of this kind generally confined themselves to clusters of purines and pyrimidines, but later studies shifted to examining G+C and A+T clusters for the thermodynamic implications of their pair-binding (Cheng and Zhang 2005; Cheng et al. 2007; Havlin et al. 1995; Nicolay et al. 2004; Peng et al. 1992; Peng et al. 1995).

I was only able to locate two papers dealing directly with DNA sequence non-randomness at the intermediate scale of tens to thousands of bases: Mrážek and Kypr (1995) discussed a "middle-range clustering" of single nucleotides on the order of hundreds of base pairs. Nikolaou and Almirantis (2002) investigated "middle-scale nucleotide clustering" using a block size of 20 nt with a "halo" of 500 nt. Each of these papers will now be meticulously examined to make note of the findings and shortcomings of each.

Mrážek and Kypr (1995) analyze the "middle-range clustering" of nucleotides in genomes. They present a simple algorithm for detecting clustering. Their algorithm involves counting the number of occurrences of some nucleotide or oligonucleotide within
some distance of a reference nucleotide (e.g.: 'A'), throughout a set of sequences. Each
sequence has a randomized partner generated using the reference sequence's dinucleotide
frequencies. This is almost exactly the randomization method used in SRI Generator. One
exception is that SRI Generator does not choose the starting oligonucleotide at random, but
rather based on the oligonucleotide frequencies. The other major difference is that I used
tetramers in this investigation instead of dinucleotides. Tetramers will capture (and thus
control for) any dinucleotide bias, trinucleotide bias and tetranucleotide bias.

The scanning method employed by Mrázek and Kypr is quite different from that used
in this work, however. It uses a fixed window size to look upstream and downstream of
every occurrence of a particular nucleotide of interest. My MRI Analyzer performs a
sliding window scan looking for overall content of the window and allows the window to
"stretch" to encompass more material until the content criteria are no longer met. This
allows for the identification of MRI features with sizes or positions that do not happen to be
convenient for the arbitrarily chosen window size.

Based on their algorithm, what Mrázek and Kypr are scanning for is, in fact,
nucleotide-oligonucleotide correlations. A subset of these are the correlations of a
nucleotide (say, 'A') with itself, which are revealed to be prominent features of primate
genomes. This is an interesting finding related to the topic under consideration in this thesis.

What a simple scanning algorithm may detect as G+C-richness may in fact only be C-richness or G-richness. Mrázek and Kypr found "overlapping clusters, several hundred base pairs in length" which are rich in a single nucleotide. They also noted that A tended to be rare in clusters of C and vice-versa, and that G tended to be rare in clusters of T and vice-versa. This phenomenon could be at work beneath the MRI features that described in this thesis, and thus should be carefully considered. Nonetheless, I have visually inspected the sequences being identified by MRI Analyzer and there are many that are not only C- or G-rich.

Importantly, Mrázek and Kypr can observe a range of sizes of these clusters using their method. They conclude that there are no measurable correlations between individual nucleotides separated by more than 1000 bp, but that the CG doublet does cluster at distances above this arbitrary cutoff. While there is not yet a consensus for the parameters defining CpG islands, by some measures they can correspond to distances over 1000 bp, especially in promoter regions (Hackenberg et al. 2006; Wang and Leung 2004). The authors also note the prevalence of clustering of the TA doublet. They also reason that
their results indicate a lack of correlations at distances ranging from one to tens of kilobases. The work presented in this thesis confirms this observation.

An interesting part of the discussion in this paper refers to the highly invariant cluster size of a few hundred nucleotides and its prevalence in eukaryotes. The authors posit that this size may have its origins in a relationship to nucleosomes, a prescient statement for 1995!

Of great import in this article is that it characterized dinucleotide clusters, such as CG and TA, as well as trinucleotide clustering (e.g. CNG, TNA), and that it established an upper bound on the sizes of these clusters. While the algorithm employed did not scan for bulk G+C content, this paper is a pioneering work and provides data that is useful for choosing parameters and confirming related results.

Nikolaou and Almirantis (2002) investigated "middle-scale nucleotide clustering" using a modified standard deviation method. This method considers a sequence of length $L$ divided into $N$ contiguous segments of length $m$ and a "halo" of length $h$ around each one. For each segment, they calculate the nucleotide frequencies for both the segment and its halo. The $Q^2$ quantity is then defined for each nucleotide as the square of the difference between the frequency in the segment and in its local halo. Then the MSD($m, h$) is calculated as the square root of the average of these $Q^2$ values over all $N$ segments,
averaged over the four nucleotides. An alternative measure, denoted \( \text{MSD}_w(m,h) \), uses a weighted average over the four nucleotides based on their relative frequency in the whole sequence. Yet another variation on the measure uses overlapping ("superposed") blocks. The authors note that for most sequences, the weighted averaging yields better results and that superposed blocks invariably yield better results, where "better" is defined as the ability to distinguish coding from non-coding sequences. They also examined the effect of varying the block size \( m \) and found that a block size of 20 generally performed best, but that sizes up to 50 or 100 nt saw no significant loss of power in their measure. For an \( m \) of 20, values of \( h \) in the range 400 to 800 nt yielded good results. The authors therefore chose to standardize on the parameters \( m=20, h=500 \) for the presented results. These values can also be considered as important cues for the work in this thesis, which also seeks to avoid the effects of both short-range signals and large-scale inhomogeneity.

Nikolaou and Almirantis (2002) observed a clear separation between coding and non-coding sequences using their MSD measure (see their Figure 2).
As the authors note in their text, 87% of the exons responsible for the section of the "long eukaryotic CDSs" curve that coincides with non-coding sequences are terminal exons, which are only translated for one fifth of their length, on average. Importantly, the overall conclusion of this work is two-fold: that this method readily distinguishes sequences based on their function (i.e. exons, introns, promoters, and coding sequences); and that intronic and intergenic regions of eukaryotic genomes possess high levels of nucleotide clustering.
Key to note in these conclusions is that this work does not distinguish between clustering of the four individual nucleotides and does not even consider co-clustering of two nucleotides, such as guanine and cytosine, which have well known biological roles in genomes. One might then question the decision to lump together the four nucleotides in their measure when each is presumably free to cluster separately. Thus, this work leaves the door open for further studies examining the nature of particular forms of nucleotide clustering at the scale of 20 to 1000 bases. More importantly, while this work is able to find an association between overall levels of nucleotide clustering and genomic role (exons, introns, etc.), it does not claim to relate this clustering to any particular biological process and does not present any evidence of evolutionary mechanisms or selection pressures to create or maintain this clustering.
Manuscripts

Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures

Authors: Jason M. Bechtel\textsuperscript{1}, Thomas Wittenschlaeger\textsuperscript{1,6}, Trisha Dwyer\textsuperscript{1,2}, Jun Song\textsuperscript{1,2}, Sasi Arunachalam\textsuperscript{1,3}, Sadeesh K. Ramakrishnan\textsuperscript{1,4}, Samuel Shepard\textsuperscript{1,5}, Alexei Fedorov\textsuperscript{1,5}

\textsuperscript{1}Program in Bioinformatics and Proteomics/Genomics, \textsuperscript{2}Dept. of Physiology and Pharmacology, \textsuperscript{3}Dept. of Neuroscience, \textsuperscript{4}Department of Cardiovascular and Metabolic Diseases, \textsuperscript{5}Dept. of Medicine, University of Toledo Health Science Campus, Toledo, OH 43614, USA.

\textsuperscript{6}Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403, USA.

Contributions: I created two of the five main programs (MRI Analyzer, and MRI Generator) as well as MRI visualizer, which produces the "spike graph" visualizations of MRI distribution in a sequence. I developed most of the GMRI website functionality. I prepared figures 3 through 6. I played a central role in data generation and collection. I developed the idea of "optimal contrast" and wrote scripts to find the best parameters for the analysis.
Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures

Jason M. Bechtel1†, Thomas Wittenslaeger1,6+, Trisha Dwyer1,2, Jun Song1,2*, Sasi Arunachalam1,3, Sadeesh K. Ramakrishnan1,4, Samuel Shepard1,5, Alexei Fedorov1,5§

1Program in Bioinformatics and Proteomics/Genomics, 2Dept. of Physiology and Pharmacology, 3Dept. of Neuroscience, 4Department of Cardiovascular and Metabolic Diseases, 5Dept. of Medicine, University of Toledo Health Science Campus, Toledo, OH 43614, USA.

6Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403, USA.

†Present address: Luye Pharmaceutical LTD, Rm1107, Zhubang 2000 Business Center, Chaoyang District, Beijing 100025, China.

†These two authors have equal contribution; §Corresponding author

Email addresses:

JMB: jason.bechtel@utoledo.edu
TW: tmwitten@gmail.com
TD: trisha.dwyer@utoledo.edu
JS: songjun@luye-pharm.com
SA: sasi.arunachalam@utoledo.edu
SKR: sadeesh.ramakrishnan@utoledo.edu
SS: samuel.shepard@utoledo.edu
AF: alexei.fedorov@utoledo.edu


(© 2008 Bechtel et al.; reproduced here according to the terms of the Creative Commons Attribution License 2.0 - http://creativecommons.org/licenses/by/2.0)
Abstract

**Background.** Genomes possess different levels of non-randomness, in particular, an inhomogeneity in their nucleotide composition. Inhomogeneity is manifest from the short-range where neighboring nucleotides influence the choice of base at a site, to the long-range, commonly known as isochores, where a particular base composition can span millions of nucleotides. A separate genomic issue that has yet to be thoroughly elucidated is the role that RNA secondary structure (SS) plays in gene expression.

**Results.** We present novel data and approaches that show that a mid-range inhomogeneity (~30 to 1000 nt) not only exists in mammalian genomes but is also significantly associated with strong RNA SS. A whole-genome bioinformatics investigation of local SS in a set of 11,315 non-redundant human pre-mRNA sequences has been carried out. Four distinct components of these molecules (5´-UTRs, exons, introns and 3´-UTRs) were considered separately, since they differ in overall nucleotide composition, sequence motifs and periodicities. For each pre-mRNA component, the abundance of strong local SS (< -25 kcal/mol) was a factor of two to ten greater than a random expectation model. The randomization process preserves the short-range inhomogeneity of the corresponding natural sequences, thus, eliminating short-range signals as possible contributors to any observed phenomena.
Conclusions. We demonstrate that the excess of strong local SS in pre-mRNAs is linked to the little explored phenomenon of genomic mid-range inhomogeneity (MRI). MRI is an interdependence between nucleotide choice and base composition over a distance of 20-1000 nt. Additionally, we have created a public computational resource to support further study of genomic MRI.

Background

RNA secondary structures

Secondary structures (SS) are crucial elements for the biosynthesis and/or correct action of non-coding RNAs in mammals and other eukaryotes. Moreover, they are key regulators in the function and turnover of mRNA molecules. SS in pre-mRNAs regulate the splicing process [1-3]. In mature mRNAs, SS located in 5΄-untranslated regions (5΄-UTRs) signal for translational control [4, 5] and those located in 3΄-untranslated regions (3΄-UTRs) regulate sub-cellular localization and stability [6-8]. SS located inside protein-coding sequences could play a role in translational speed and stability [9, 10].

Prior studies of the strength of computer-predicted SS in mRNA have had conflicting conclusions [11, 12]. More importantly, these studies did not investigate the abundance of SS and considered only coding sequences. This spurred us to perform a bioinformatics investigation into the abundance of SS throughout mammalian genomes. Our results show
that the existence of many energetically-strong SS is associated with the phenomenon of global mid-range inhomogeneity (MRI), manifest as nucleotide compositional relationships at a scale of 20 to 1000 bases throughout the genome. MRI appears as a strong tendency for the clustering of particular bases (e.g. C and G nucleotides, or G and A nucleotides) inside short regions of genomic sequences. This paper provides new approaches and tools to gain insights into this form of genomic inhomogeneity.

**Short-range inhomogeneity**

It is well established that the particular base (A, G, C, or T) that appears in a given position of a genomic sequence significantly depends upon the nearest bases surrounding its position [13, 14]. Consequently, the frequency (F) of a dinucleotide XY is often not equal to the product of the individual frequencies of nucleotides X and Y (F_{XY} \neq F_X \cdot F_Y). The highest interdependence of base frequencies is always observed for adjacent nucleotides. The ratio (F_{XY} / (F_X \cdot F_Y)) for adjacent bases X and Y is known as a "genomic signature" [14]. Genomic signatures as low as 0.22 (for the CG dinucleotide in mouse) and as high as 1.75 (for the GC dinucleotide in *Campylobacter jejuni*) have been recorded [15].

The interdependence of base frequencies sharply drops with increasing distance. When the distance between nucleotides X and Y is more than six bases, their occurrence interdependency becomes negligible. Here, we refer to this type of interdependency...
between nucleotides separated from each other by a few positions as short-range inhomogeneity (SRI).

**Long-range inhomogeneity**

Also well recognized are long-range interdependencies in nucleotide frequencies on a scale of up to millions of bases, known as genomic isochores [16]. It has been shown that isochores can be generally categorized according to their level of G+C content. Isochores defined by G+C content correspond to many other genomic phenomena. GC-rich isochores replicate later in S-phase, contain higher concentrations of genes, and have genes with shorter introns and untranslated regions. Moreover, GC-rich isochores tend to have an "open" chromatin structure and thus have higher rates of transcription [17]. Higher G+C content isochores also experience higher recombination rates—perhaps lending support to the notion that higher recombination rates led to the creation of isochores through biased gene conversion [18]. While the evolution and maintenance of isochores is subject to debate, their presence is indeed evidence of existing interdependencies in nucleotide composition on the scale of tens of thousands to millions of nucleotides. We will refer to this form of non-randomness in genomic nucleotide composition as long-range inhomogeneity.
**Mid-range inhomogeneity**

The compositional non-randomness between the two extremes described above we call *mid-range inhomogeneity* or MRI. MRI has yet to be thoroughly investigated. The only well-known manifestation of mid-range inhomogeneity is CpG islands. Most attempts to define CpG islands set hard requirements for region size (at least 200 or 500 bases long), G+C content (> 50% or 55%), and CpG observed/expected ratio (> 0.6 or 0.65) [19, 20, respectively]. CpG-islands are found near 60% of human genes, including all housekeeping genes and about half of the tissue-specific genes [21]. Here we demonstrate that MRI can be observed for regions from 30-1000 bp and is significant not only for G+C content but for other nucleotide pairings (A+G and G+T) as well as for the individual bases.

**Results**

**Analysis of strong local SS within pre-mRNAs**

Distinct parts of mRNAs and introns have large variations in nucleotide composition (from 35% to 60% of GC-content, see Table 1). Due to this difference the analyses of SS distribution were performed on four separate regions: GC-rich 5´-UTR regions, GC-poor introns and 3´-UTR regions, and intermediate GC-content protein-coding regions of mRNAs. In addition, vertebrate and invertebrate species have considerable variations in
their mRNA nucleotide composition. Within the mammalian class, however, the variation in GC-content is negligible (Table 1). For this reason we demonstrate only results for human sequences, although the observed trends are applicable to all mammals.

We begin by looking at SS created by the interactions of nucleotides less than 50 bases apart (local SS). Prediction of local SS is more reliable than prediction of global RNA structures, which can span hundreds of nucleotides [22]. We examined these structures in 11,315 non-redundant human gene sequences (see Methods section) and calculated their strengths using the RNALfold program of the Vienna RNA package [23]. Figure 1 illustrates the distribution of local SS according to folding strength in distinct parts of mRNAs and introns. We concentrate our study on the stronger secondary structures. Such strong local SS could withstand competition with the many RNA-binding proteins that cover mRNAs and, thus, are more likely to persist in vivo. Indeed, hairpin structures with an mfe of -30 kcal/mol situated close to the mRNA cap significantly impede ribosome scanning, while hairpin structures with an mfe upwards of -50 kcal/mol inhibit elongation [5]. Strong secondary structures are also required for the regulation of splicing. Kishore and Stamm showed that the 18 nt-long HBII-52 snoRNA antisense element interacts with a pre-mRNA segment with an mfe of -27.6 kcal/mol and thereby determines the fate of alternative splicing in the serotonin receptor gene [24]. Finally, a majority of human miRNA genes from the microRNAdb [25] have strong interaction energies (< -25 kcal/mol)
with dozens of targets within mRNAs. Thus, "strong" local SS is defined as having a minimum free energy (mfe) value of less than or equal to -25 kcal/mol, so as to encompass all known functional local SS.

**Analysis of strong local SS in randomized sequences**

To evaluate the abundance of local SS, one must compare their prevalence in naturally occurring mRNAs with their levels in reference sequences having no selection for SS. In most research, reference sequences are randomly generated to have nucleotide compositions approximating those of the naturally occurring mRNAs. In order to properly compare local SS in natural and randomized sequences one needs to preserve short-range inhomogeneity (SRI), as discussed previously by Workman and Krogh [12].

Thus, to most accurately preserve SRI in randomized sequences we created a public resource for generating randomized sequences while taking into account the SRI of a given set of natural sequences. Our algorithm can take into account not only relative dinucleotide frequencies, but also frequencies of longer oligonucleotides (up to 9-mers). We first applied our SRI-analyzer program (see Methods section) [26] to a set of natural mRNA sequences to obtain their oligonucleotide composition, shown in Table 2. Our second program, SRI-generator [26], then uses these oligonucleotide frequency tables to generate random sequences with approximately the same oligonucleotide distribution as the
natural sequences but without any similarity in their sequence alignments. Table 2 demonstrates the oligonucleotide frequencies for human 5´-UTRs and two independent SRI-generated sets of sequences. Notably, the oligonucleotide compositions of the SRI-generated sequences are very close to those of the natural sequences, with only small fluctuations due to the inherently random nature of the sequence generation process.

Pseudocode and further explanation of SRI-generator (Algorithm 1) is provided for the reader as a supplementary figure.

Figure 2 A-C demonstrate the distribution of strong local SS in human 5´-UTR, 3´-UTR and intronic sequences and in their corresponding randomized counterpart sequences. Prediction of local SS was computed on a non-redundant sample of 11,315 human genes (see M&M) by the RNALfold program [23]. All SRI-randomization was performed based on the tetramer frequency tables of the corresponding natural sequences. Tetrarmers reflect almost all short-range non-randomness since the major influence on SRI comes from adjacent bases [27]. Remarkably, the number of local SS in the natural sequences exceeds the number of structures of the same strength in the random sequences by a factor of at least 2 to 10. Because we processed eleven thousand genes, the significance of this difference is unquestionable: the chi-square goodness-of-fit test gives a p-value less than 10^{-200}. Having observed the difference between introns and their SRI-generated counterparts, we also examined sequences from intergenic regions (located between
protein-coding genes and having an overall nucleotide composition similar to that of introns) and detected the same trend (Figure 2D). Moreover, the abundance of SS in natural sequences has no relation to genomic repetitive elements. Masking all human repeats with the RepeatMasker program (Smit et al. 2004, http://www.repeatmasker.org) [28] even mildly enhanced the difference in strong local SS between natural and SRI-generated sequences (see Fig. S1). This observation simply reflects the fact that DNA repeats do not have an excess of strong SS, although some repeats have a distinct oligonucleotide composition and are enriched by C/G bases (e.g. Alu family).

Protein-coding sequences (CDS) contain a profound 3 nt periodicity and other non-randomness associated with translational properties [29]. All of this information would be lost in SRI-generated sequences. To overcome this problem we created CDS-generator, a public resource for the randomization of protein-coding sequences [26]. CDS-generator changes only the variable nucleotides in the third codon position, which do not change the coded amino acids. Additionally, CDS-generator maintains the codon and di-codon biases of a given set of natural coding sequences. Thus, randomization by CDS-generator is much weaker than randomization by SRI-generator since it retains > 70% sequence identity between the natural and random sequences. On the other hand, maintaining a considerable level of sequence identity is useful because it preserves the major periodicity characteristics of the source coding sequences. Figure 2E demonstrates
that natural coding sequences have twice the number of strong local SS as randomized sequences obtained by CDS-generator. The chi-square test confirms that the difference is statistically significant (p < 10⁻²⁰⁰).

**Mid-range inhomogeneity in natural genomic sequences**

To understand the observed abundance of strong local SS in mRNAs, we examined dozens of these structures in natural sequences, a typical example of which is shown in Figure 3. This structure, from a human 3′-UTR of the *KIAA1751* gene, has an mfe of -27.2 kcal/mol and represents a hairpin stem-loop configuration. The sequence of this SS is GC-rich (67%) and is neighbored by several other alternating short AT-rich and GC-rich regions, as highlighted in Figure 3A. In contrast, such frequent alternation in GC-composition is practically absent in SRI-generated sequences. Statistical examination revealed that strong local SS with mfe values in the range -25 to -30 kcal/mol in human mRNAs have a mean GC-composition of 70%, which is much higher than the average GC-composition of the mRNA, introns, or intergenic regions presented (Table 1). The observed GC-enrichment within strong local SS can be explained by thermodynamics (G-C base pairs are about twice as strong as A-T pairs) and by combinatorics (random base-pairing is more frequent in GC-rich strands than when GC-composition is around 50%). These notions have led to the hypothesis that natural sequences have a profound mid-range
inhomogeneity, that is, they are enriched by short GC-rich regions (30-100 nt) alternating with adjacent AT-rich regions. In other words, we theorize the non-random clustering of G/C and A/T bases on the scale of ~50 nucleotides and the over-abundance of such clusters in natural sequences.

To test this hypothesis we created a program named MRI-analyzer [26]. This program scans input sequences with a mid-range window size (the default, utilized for the results presented, is a 50 nt window). When the GC-content of the sequence in this window reaches the upper threshold, MRI-analyzer generates a blue top spike on the output graph (Figure 4). Similarly, when the GC-content of the window reaches the lower threshold, MRI-analyzer generates a red bottom spike. The upper and lower thresholds are flexible parameters defined by the user. MRI-analyzer output for natural and SRI-generated 3'-UTR and intronic sequences is shown in Figure 4. Here, the graph clearly shows a 35-fold enrichment of GC-rich (≥ 70%) 50 nt-long regions in natural 3'-UTR sequences over the randomized SRI-generated sequences. For the 319 kb single extra-large intron in Figure 4 C-D, GC-rich regions are enriched by a factor of 14. Similar to the results shown in Figure 4, we observed comparable contrasts in GC-rich and GC-poor regions for 5'-UTR, intronic, and intergenic sequences for a wide range of scanning window sizes (30-1000 nt) (Figure 6). Additionally, MRI was generally observed with other base combinations and single bases (Figure 6). An example of the MRI pattern for
GA- and CT-rich regions is shown in Figures 5A and 5B; GT- and AC-rich regions are presented in Figures 5C and 5D. The enrichment factor can be considered as the contrast between real and randomized sequences. We thus use the term "contrast" to refer to the ratio of the number of content-rich regions in a real sequence to the number in its SRI-generated counterpart. "Optimal contrast" is defined as the highest contrast observed over all thresholds for a given content type (for example, GC-content) and window size. Since we do not yet know the properties of mid-range inhomogeneity in the human genome, we probed for the optimal contrast by repeating the above analysis for all thresholds starting from one standard deviation above (and below) the mean GC-content and proceeding until the number of GC-rich (or GC-poor) regions decreases to about ten. Figure 7 shows the distribution of optimal contrasts for all content types and a range of window sizes for the longest first intron of the \textit{DMD} gene.

\textbf{Association of MRI with the over-abundance of strong local SS}

Finally, we created a program named MRI-generator [26] for obtaining random sequences having the same oligonucleotide composition and also the same MRI pattern in GC-composition as a specified set of natural sequences. This program works by producing an excessively long SRI-generated sequence and then discarding segments with intermediate GC-content to obtain the desired pattern of GC-rich and CG-poor regions.
Thus, the output sequence from MRI-generator has a Genomic-MRI pattern of GC-rich and GC-poor regions very similar to that of the natural sequence.

Comparison of natural sequences with their MRI-generated counterparts for each genomic sequence category (5´-UTRs, 3´-UTRs, introns, intergenic regions, and CDS) shows that they each have approximately the same number (5-10% difference) of strong local SS, as illustrated in Figure 2F for 3´-UTRs. This finding supports the conclusion that the abundance of strong SS in all parts of the mammalian genome (mRNA, introns, intergenic regions) is associated with the MRI of these sequences.

**DNA repetitive elements and genomic MRI**

Even though human interspersed repeats do not show an excess of strong SS as discussed above, they do influence the patterns of MRI as demonstrated in Figure S2. This figure illustrates the MRI patterns of the extra-large first intron of the *DMD* gene (shown in Figure 4) after masking its repetitive elements by RepeatMasker. Unsurprisingly, the number of MRI regions in the masked sequence is a fraction of those in its non-masked counterpart. The masked sequence contains 41% N’s instead of A, G, C, or T bases. The current version of MRI-analyzer skips a window containing any non-A, G, C, or T character. For a proper comparison of MRI patterns in a masked sequence, one should compare it to the SRI-generated random sequence based on the masked sequence. Such
random sequences contain the same number of N’s at exactly at the same positions as the
natural masked sequence. Figure S2 demonstrates that the masked sequence of the first
*DMD* intron has 3 to 12 times the number of MRI peaks compared to its random
counterpart. This particular example with the *DMD* intron presents an AT-rich sequence
(67% of A+T), which is typical for extra-large introns [17]. Accordingly, we set the upper
threshold for GC-composition to 60% in studying this sequence. Under such conditions,
we observe GC-rich MRI peaks overlapping various portions of Alu-repeats. This
overlapping of MRI regions with repetitive elements seems to depend on the threshold used
and the G+C-composition of the region under analysis.

**Discussion**

We have demonstrated an association between MRI in GC composition and the
abundance of strong SS in genomic sequences. There are at least two possible
interpretations of these results. First, one can argue that MRI causes the abundance of
strong SS. The second possibility is that selection for strong SS was the reason for the
appearance of MRI. Both views have merit and we thus include a discussion of the
supporting evidence.

Central to this discussion is the observation that MRI exists not only in mRNA
sequences but also in introns and intergenic regions. If selection were limited to transcripts
or to mature mRNAs, there would be no way for evolution to directly drive the creation of SS in non-transcribed regions. This would leave MRI in GC composition as a potential mediator of strong SS enrichment. However, some experimental evidence suggests that much more of the genome is transcribed than previously thought [30]. It has also been suggested for some time that SS play a role in the initiation of recombination. This theory predicts positive selection for SS throughout genomes and especially within introns and intergenic regions [31-33]. Moreover, studies of coding sequences in mammals have found that synonymous substitutions tend to increase the strength of SS and regulate mRNA stability [34-36]. Thus, SS could have emerged first due to selection for DNA hairpins to facilitate homologous recombination and for stable mRNA SS signals, yielding MRI in GC content as a by-product. On the other hand, MRI is also observed for AG- and GT-content as well as for the individual nucleotides (see Figure 5 and 7), so it is also possible that selection for MRI is a fundamental force driving genome organization and composition.

It is of special interest to investigate possible biological roles for MRI in the structural and functional organization of mammalian genomes. To address this important issue, we have studied 3.3 million point mutations occurring over the last 10 million years in humans and over 3.9 million SNPs in the MRI-regions and outside them. These results will be detailed in our next publication (under preparation). Based on the preliminary results of these investigations, we can state that MRI patterns are formed by a combination of...
processes. Some patterns (e.g. A+T-rich regions) are like cellular automata, based on non-selection biases in nucleotide changes at genomic regions with specific base compositions, while other patterns are formed by a strong fixation bias (presumably positive selection of functional regions) that preserve particular base enrichments in corresponding regions (e.g. G+C-, purine-, and pyrimidine-rich). These forces drive mid-range non-randomness, shaping the human genome and potentially imparting additional layers of organizational complexity.

Indeed, an important feature of the human genome is that its vast array of genes is differentially expressed in hundreds of different cell types and subtypes. Moreover, at different stages of development and in response to diverse extracellular stimuli, gene expression must be finely tuned. To perform the enormous task of creating a human body composed of trillions of cells, the genome must contain a vast number of signals for gene regulation, the majority of which have yet to be discovered. We hypothesize that MRI could represent a novel class of genomic signals, based on overall composition and clustering of nucleotides rather than particular sequence motifs. To facilitate the testing of this hypothesis, we created a free, public Internet resource called "Genomic MRI" that allows one to run all programs described here without any programming knowledge. Additionally, all of these programs are freely available for downloading and off-line usage, primarily for computational biologists.
Methods

The programs SRI-analyzer, SRI-generator, MRI-analyzer, MRI-generator, and CDS-generator are available at the following URL:

http://mco321125.meduohio.edu/~jbechtel/gmri/. A link to the current location of the website will be maintained at our departmental project site at

http://hsc.utoledo.edu/depts/bioinfo/gmri/.

Sequence randomization algorithm (SRI-generator)

There are several possible approaches for randomizing nucleotide sequences while maintaining their N-mer oligonucleotide frequency composition. The simplest approach would be to randomly choose N-mer oligonucleotides based on their frequency composition and tile them one after the other. However, this approach does not necessarily preserve the frequencies of shorter length oligonucleotides that one may observe in natural sequences. For example, the random concatenation of N-mers as tiles artificially introduces dinucleotide composition bias created from the border of two adjacent oligonucleotide tiles—producing an overrepresentation of CpG dinucleotides and the like that do not match the SRI natural sequences. Therefore we chose a different approach which generates a randomized sequence one nucleotide at a time moving in a 5´ to 3´ direction.
We generate randomized sequences in the following manner. First we choose the largest oligonucleotide size (N) that is sufficiently sampled. In practice, this means avoiding sizes for which some of the oligonucleotides are never encountered in the input sequence (i.e. occur with zero frequency). Throughout our study we used 4-mer oligonucleotides (N=4) because they were consistently well sampled across all of our input sequences, including a single large intron in the Figure 2C. The starting oligonucleotide is chosen at random, abiding by the frequency table for oligonucleotides of the chosen size (N). Next, we take the last (N-1) bases of our sequence, and append a base to the 3’ end, following the N-mer oligonucleotide frequencies. For example, if N=4 and GTC were the last three bases in the growing random sequence, the frequencies of GTCA, GTCT, GTCC, and GTCG would be used in randomly adding the next base. For instance, suppose these four oligomers have relative frequencies of 0.5, 0.1, 0.2, and 0.2, respectively. Then the random number generator will append A with a probability of 0.5, T with a probability of 0.1, C with a probability of 0.2, and G with probability of 0.2. This final step is then repeated until the randomized sequence reaches the length of the input sequence. In contrast to the tiling method, our approach preserves the frequencies of short oligonucleotides in addition to preserving the N-mer frequency composition.

Finally, we made our SRI-generator work properly with sequences that have masked repetitive elements (where all sequences of DNA repeats are replaced by N’s by the
RepeatMasker program). Any non-A, T, C, or G bases are copied from the source sequence over the output sequence. The random sequences thus contain the same number of N’s (or other non-A, T, C, or G bases) in the same positions as in the natural sequences provided as input.

The pseudocode for SRI-generator is presented in the supplementary file (Algorithm 1), while the source code (written in Perl) is freely available for downloading from our project’s website.

**CDS-generator**

Several sophisticated algorithms are already available for the randomization of coding sequences [37-38]. However, here we used our own randomization approach developed by AF in 2001 while working on a context-dependent codon bias project in the Walter Gilbert lab [29]. We stayed with our program because we are familiar with the peculiarities of this type of randomization. In addition our approach gets the dicodon distribution of randomized sequences very close to that of the natural CDS.

**Program notes**

1) We observe a gradual diminution of the difference between real and randomized sequences when using progressively larger oligonucleotides with the randomized sequence generation programs (**SRI-generator** and **MRI-generator**). The difference is not
considerable, but it is noticeable. Therefore, we recommend the use of longer oligonucleotides in the construction of randomized sequences—to maximize the retention of short-range inhomogeneity—as long as the rarest oligonucleotide in the corresponding frequency table occurs at least ten times. We use tetramer frequency tables throughout the manuscript for the sake of consistency and since they can safely be used for analyses of individual loci having as little as 100 kb.

WARNING: In MRI-generator it is easy to shift the nucleotide content level of generated sequences by using thresholds that do not balance the number of content-rich and content-poor regions. One must experiment with the thresholds and use SRI-analyzer to confirm that the content of the MRI-generated sequence approximates that of the source sequence.

2) The graphical output provided with the online version of MRI-analyzer serves only as a quick visual aid. The true output is represented by large tab-delimited files containing a record for each window in the analysis. Each record contains flags indicating a content-rich or content-poor window and, for those records where one of the thresholds has been crossed, the corresponding sequence.

3) All programs are written in Perl and may be freely downloaded from the website. They are licensed under version 3 of the GNU General Public License (GPL).
4) The RNAlfold program from version 1.6.1 of the Vienna RNA package was utilized locally on our computers with default parameters.

**Source for gene sample set**

Our sample of 11,315 non-redundant human genes (with < 50% sequence identities between each other) was obtained from the human Exon-Intron Database, release 35p1 [39]. Samples of intergenic regions were obtained from Genbank human genome files build 36 based on the records from the Feature Tables. We used only plus strands for calculations because there are only fluctuation differences between plus and minus strands in the non-coding regions of mammalian genomes. Also, plus and minus strands have the same G+C and A+T compositions. All these samples are available from our departmental project site at http://hsc.utoledo.edu/depts/bioinfo/gmri/.

**List of Abbreviations**

SS = secondary structure(s)

MRI = mid-range inhomogeneity

SRI = short-range inhomogeneity

UTR = untranslated region

CDS = coding sequence(s)

nt = nucleotide(s)
Authors' Contributions

JMB, TW, TD, JS, SA, SKR, SS were responsible for computational processing of all datasets and creation of all described programs. JMB and SS also created the Genomic MRI web resource. AF supervised the project, provided guidance and wrote the draft. All authors have read and approved the final manuscript.

Acknowledgments

This project is supported by NSF Career award MCB-0643542. We thank Peter Bazeley, University of Toledo, for his computational support and discussion of our algorithms.

References


22. Mathews DH: **Predicting a set of minimal free energy RNA secondary structures common to two sequences.** *Bioinformatics* 2005, **21**:2246-2253.


26. All described programs are freely available via our web site:

[http://hsc.utoledo.edu/depts/bioinfo/gmri/]


[http://www.repeatmasker.org]


34. Chamary JV, Hurst LD: **Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals.** *Genome Biol* 2005, **6:**R75.


**Figures**

*Figure 1 - Distribution of local SS with respect to folding energy in mRNA components and introns*

Number of structures was measured within 1 kcal/mol intervals and normalized by 1,000 nucleotides of analyzed sequences.
Figure 2 - Distribution of strong local SS with respect to folding energy in mRNAs and genomic sequences

Number of structures was measured within 1 kcal/mol intervals and normalized by 1,000 nucleotides of analyzed sequences. (A) 5’-UTRs (blue) and two independent SRI-
generated sequences (gray); (B) 3’-UTRs (yellow) and two independent SRI-generated sequences (gray); (C) introns (green) and two independent SRI-generated sequences (gray); (D) intergenic regions from chromosome 17 (red) and two independent SRI-generated sequences (gray); (E) CDS (burgundy) and two independent CDS-generated sequences (gray); (F) 3’-UTRs (yellow), random MRI-generated counterpart sequences (black), and random SRI-generated counterpart sequences (gray).
Figure 3 - Example of a strong local SS in the 3'-UTR of the human KIAA1751 gene [GenBank:NM_001080484]

(A) Nucleotide sequence of the entire 3'-UTR region in which a segment exemplifying strong local SS (mfe = -27.2 kcal/mol) is shown in red and its schematic base-pairing is shown in dot-bracket notation (Hofacker 2003) below the sequence. Other GC-rich regions
are highlighted in blue and GC-poor regions are underlined. (B) 2-D representation of this strong SS.

Figure 4 - Visualization of MRI-analyzer output for GC-composition of two 300 kb samples using a 50-nt window

Upper and lower thresholds are specified on the y-axis as a percentage of the window size. (A) A sequential sample of human 3´-UTRs from chromosomes 1 and 2 (EID ids 1745_NT_004487 through 2327_NT_022184); (B) a random SRI-generated set based on the tetramer oligonucleotide frequency table of 11,315 human 3´-UTR sequences (15). (C)
The first 300 kb of the first intron of the \textit{DMD} gene; (D) a random SRI-generated set based on the tetramer oligonucleotide frequency table of the first intron of the \textit{DMD} gene.

\textbf{Figure 5 - Comparison of MRI-analyses of GC-content for various window sizes and genomic contexts}

(A-F) The first 300 kb of the first intron from the \textit{DMD} gene, and its SRI-generated counterpart, analyzed for optimal visual contrast over a range of window sizes (30, 50, 100, 200, 500, 1000) (cf. Figures 4 and 6); (G) The first 300 kb of a sample of human 5´-UTRs and its SRI-generated counterpart; (H) The second 300 kb from a sample of intergenic sequences from human chromosome 17 and a corresponding SRI-generated sequence.
This figure is the "XY conditioning plot" (from Rcmdr 1.2) of the optimal contrasts (see text) for regions of high and low content for all seven possible content types over a range of window sizes (30, 50, 100, 200, 300, ... 1000). The sample sequence is the first
300 kb of the first intron from the *DMD* gene. The SRI-generated counterpart is constructed from the tetramer frequency table derived from the entire intron.

**Tables**

*Table 1. Percentage of GC-composition in different regions of pre-mRNA for diverse animal species.*

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>5´-UTRs</th>
<th>CDS</th>
<th>3´-UTRs</th>
<th>Introns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>60%</td>
<td>52%</td>
<td>44%</td>
<td>41%</td>
</tr>
<tr>
<td>Mouse</td>
<td>59</td>
<td>52</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>Cow</td>
<td>60</td>
<td>54</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>Chicken</td>
<td>57</td>
<td>51</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>45</td>
<td>50</td>
<td>37</td>
<td>35</td>
</tr>
<tr>
<td>Drosophila</td>
<td>45</td>
<td>54</td>
<td>36</td>
<td>40</td>
</tr>
</tbody>
</table>
Table 2 - Excerpt from oligonucleotide frequency table for 5’-UTRs of 11,315 human genes and two SRI-generated counterparts. The entire dataset is presented in Table S1.

<table>
<thead>
<tr>
<th>Oligonucleotide</th>
<th>Human 5’-UTRs</th>
<th>Random 1 SRI-generated</th>
<th>Random 2 SRI-generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.203</td>
<td>0.203</td>
<td>0.203</td>
</tr>
<tr>
<td>T</td>
<td>0.201</td>
<td>0.201</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.292</td>
<td>0.294</td>
<td>0.293</td>
</tr>
<tr>
<td>G</td>
<td>0.303</td>
<td>0.303</td>
<td>0.303</td>
</tr>
<tr>
<td>AA</td>
<td>0.0506</td>
<td>0.0503</td>
<td>0.0506</td>
</tr>
<tr>
<td>AT</td>
<td>0.0323</td>
<td>0.0323</td>
<td>0.0318</td>
</tr>
<tr>
<td>AC</td>
<td>0.0438</td>
<td>0.0437</td>
<td>0.0438</td>
</tr>
<tr>
<td>AG</td>
<td>0.0771</td>
<td>0.0767</td>
<td>0.077</td>
</tr>
<tr>
<td>TA</td>
<td>0.0258</td>
<td>0.0257</td>
<td>0.0256</td>
</tr>
<tr>
<td>TT</td>
<td>0.0511</td>
<td>0.051</td>
<td>0.0508</td>
</tr>
<tr>
<td>TC</td>
<td>0.0593</td>
<td>0.0591</td>
<td>0.0588</td>
</tr>
<tr>
<td>TG</td>
<td>0.0653</td>
<td>0.0651</td>
<td>0.0651</td>
</tr>
<tr>
<td>CA</td>
<td>0.0608</td>
<td>0.0609</td>
<td>0.0608</td>
</tr>
<tr>
<td>CT</td>
<td>0.0726</td>
<td>0.0726</td>
<td>0.0725</td>
</tr>
<tr>
<td>CC</td>
<td>0.0965</td>
<td>0.0973</td>
<td>0.0968</td>
</tr>
<tr>
<td>CG</td>
<td>0.0627</td>
<td>0.0627</td>
<td>0.0629</td>
</tr>
<tr>
<td>GA</td>
<td>0.0658</td>
<td>0.0661</td>
<td>0.0662</td>
</tr>
<tr>
<td>GT</td>
<td>0.0453</td>
<td>0.045</td>
<td>0.0452</td>
</tr>
<tr>
<td>GC</td>
<td>0.0932</td>
<td>0.0934</td>
<td>0.0936</td>
</tr>
<tr>
<td>GG</td>
<td>0.0977</td>
<td>0.098</td>
<td>0.0984</td>
</tr>
<tr>
<td>ACAA</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ACAT</td>
<td>0.002558</td>
<td>0.002507</td>
<td>0.002544</td>
</tr>
<tr>
<td>ACAC</td>
<td>0.002046</td>
<td>0.00205</td>
<td>0.002057</td>
</tr>
<tr>
<td>ACAG</td>
<td>0.002775</td>
<td>0.002731</td>
<td>0.0028</td>
</tr>
<tr>
<td>ACTA</td>
<td>0.000442</td>
<td>0.0004362</td>
<td>0.000442</td>
</tr>
<tr>
<td>ACTT</td>
<td>0.001396</td>
<td>0.001362</td>
<td>0.001388</td>
</tr>
<tr>
<td>ACTC</td>
<td>0.002893</td>
<td>0.002899</td>
<td>0.002881</td>
</tr>
<tr>
<td>ACTG</td>
<td>0.003028</td>
<td>0.002958</td>
<td>0.003063</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Materials

The sample of 11,315 non-redundant human genes was obtained from the human Exon-Intron Database, release 35p1 (Shepelev and Fedorov 2006; see manuscript above).

For initial experiments with intronic sequences, a sample of three well-known genes with very long introns and reliable nucleotide sequences were chosen. These are contained in a FASTA-formatted file referred to as "introns_3genes_BIG".

"a non-redundant collection of 67,686 human introns" ...

Methods

The methods used in the analysis of the data are encapsulated in the following Perl programs. All graphs are generated using the GD::Graph library (Verbruggen 2002), except where otherwise noted.

SRI_analyzer.pl – SRI Analyzer scans a set of nucleotide sequences and generates a table of oligonucleotide frequencies for all n-mers from one to N, where N is the user-specified "maximum oligonucleotide size." It does this by pre-generating tables of all oligonucleotides from one to N and then counting the number of occurrences of each oligonucleotide in the sequence. It then calculates the frequencies for each n-mer oligonucleotide by dividing by the total number of n-mer oligonucleotides in the sequence. The precision for this operation is set at n+2 decimal places (e.g. six decimal places for tetramers). Finally, the program writes the oligonucleotide frequency tables to a user-
specified file. In addition, it writes out a table of most and least common oligonucleotides for each n-mer level to a separate file. That table is described in the text below. The program is currently limited to nonamers to help avoid problems with scaling and extremely large output files. The current version also handles non-canonical positions (e.g. 'X', 'N', 'R', 'Y'). It identifies the canonical and non-canonical positions, splits each sequence at any non-canonical positions, and analyses the resulting sub-sequences (as opposed to stripping out the non-canonical positions and treating the remaining canonical positions as if they had been adjacent). A relative frequency and total count of the canonical and non-canonical positions is appended to the monomer table in the output file and a warning is issued on the command-line (standard error).

**SRI_generator.pl** – SRI Generator uses the oligonucleotide composition file produced by SRI Analyzer to generate a randomized set of sequences based on a user-supplied template. Under most circumstances that template should be the original sequences from which the oligonucleotide composition file was derived. SRI Generator uses only the oligonucleotide frequencies from the N-mer table, where N is specified by the user. Of course, n-mer frequencies contain within them the (n-1)-mer frequencies and the (n-2)-mer frequencies, etc. For each sequence in the template file, the program generates a randomized replacement by starting with an n-mer chosen at random based on the n-mer oligonucleotide frequencies. It then proceeds by considering all n-mers starting with the last n-1 nucleotides of the sequence generated so far. It chooses one of these four possibilities at random, based on the oligonucleotide frequencies and appends the corresponding final nucleotide to the growing sequence. If at any point the program enters
an indeterminate state (i.e. a condition in which all potential oligonucleotides have a frequency of zero) it simply chooses a nucleotide at random and continues. Once the program has reached the end of a sequence, it checks the template for non-canonical bases (e.g. 'N', 'X', '.', etc.). Any non-canonical bases are then copied from the template to the randomized sequence at the same positions to preserve this lack of knowledge.

**MRI_analyzer.pl** – MRI Analyzer scans a set of sequences for mid-range inhomogeneity using a set of user-specified parameters: window size, content type, and upper and lower thresholds for that content type. Window size must be greater than ten. Content type may be any non-empty subset the known alphabet ("ACGT") other than the entire alphabet. The upper and lower thresholds must allow for a non-zero "neutral" area between them. Each sequence is scanned base-by-base using the user-supplied window size until a threshold is satisfied. A window position is then calculated using the current starting position of the window and the window size. That window position is then flagged for that content type and the exact sequence in the current window is stored. Scanning then resumes after the end of the current window. A set of sequences is treated as one long sequence for simplicity. When the program reaches the end of the input, the output is written to a user-specified file. The output consists of header lines with the parameters used for the scan followed by one line for each window position. Each window position is printed with two flags: the first is '1' if the window satisfies the upper threshold and the second is '-1' if the window satisfies the lower threshold. The corresponding flag for each unsatisfied threshold is written as 'x'. Finally, for each window with a satisfied threshold,
the triggering sequence is written out as a fourth and final field. This output can be plotted as a "spike graph" using MRI_visualizer.pl.

**MRI_generator.pl** – MRI Generator uses the oligonucleotide composition file output from SRI Analyzer and the MRI composition output from MRI Analyzer and a user-specified oligomer level (N) to generate a single randomized sequence. The generated sequence mimics the MRI pattern in the MRI composition file while preserving the oligonucleotide frequencies in the oligonucleotide composition file. The program is merely an extension of SRI Generator that continues to generate random windows based on the N-mer frequency table until it generates one that satisfies the requirements of the next window in the MRI pattern.

**MRI_analyzer_detail_v3.pl** – The crude approximation of the position and size of MRI features available with MRI_analyzer.pl was insufficient for our more detailed studies. Therefore, I wrote this modified version, which implements what we call the "stretchy windows" algorithm. The program scans a set of user-supplied sequences using a user-specified "base window size," but when a threshold is satisfied it marks the starting position and continues to scan window-by-window until the threshold is no longer satisfied. Then the program scans backwards until the threshold is satisfied again and marks the ending position. The MRI feature is then recorded as the sequence/record number and the starting and ending position along with the threshold flags as described for MRI_analyzer.pl. This output is suitable for graphing using MRI_mega_visualizer.pl.
MRI_mega_visualizer.pl – MRI_mega_visualizer.pl takes as input the output file from MRI_analyzer_detail_v3.pl and the sequence file from which it was derived. It then maps each of the "stretchy" MRI features to its position on the corresponding sequence. The program then scans for "interesting" sequences to graph, where "interesting" is defined as having at least seven MRI features of either the primary or the complementary content type. The threshold for "interesting" was empirically determined for a particular dataset and can be adjusted in the source code. For each interesting sequence, the program generates an output file similar to that generated by MRI_analyzer_detail_v3.pl, only the granularity is much finer. By writing out a line for every position in the sequence specifying whether it is in a content-rich or content-poor region, a graph similar to a "spike graph" can be generated in which the location and size of the features is more representative of reality. Obviously, for very long sequences the granularity is more coarse as the number of nucleotides per pixel increases. MRI_mega_visualizer_grapher.pl is then called upon to produce the fine-grained plot of the sequence's MRI features using this output file. The program then identifies the MRI features within each interesting sequence using larger "windows" of 1000 nt. It then "zooms in" on them, generating a high resolution graph using MRI_mega_visualizer_grapher_zoom.pl. Finally, the graph of the interesting sequence and each of its zoomed in sub-sequences are assembled into one image using the 'montage' utility from the ImageMagick package.

MRI_mega_visualizer_grapher.pl – This is a specialized graphing program to display the output from MRI_mega_visualizer.pl. It dynamically scales the data so that the
sequence can be displayed in a reasonable width and includes the final scaling factor in the
title of the resulting graph.

**MRI_mega_visualizer_grapher_zoom.pl** – This is a specialized version of
MRI_mega_visualizer_grapher.pl for displaying a sub-sequence from the output of
MRI_mega_visualizer.pl.

**MRI_randomizer.pl** – This program reads in the output file from
MRI_analyzer_detail_v3.pl and the sequence file from which it was derived. It then
randomly distributes the "stretchy" MRI features across the sequences and writes out a new
output file with the randomized positions. It distributes the content-rich MRI features first,
followed by the content-poor MRI features. No MRI feature may overlap with another
MRI feature. The output from this program is used by sdist_machine_real_vs_random.pl.

**sdist_machine.pl** and **sdist_machine_real_vs_rand.pl** – The former program was
written by Samuel Shepard; the latter program is an extension of the former. Both
programs take as input the "stretchy windows" output files from
MRI_analyzer_detail_v3.pl and MRI_randomizer.pl. Both programs calculated the
distances between all adjacent MRI features on each sequence. Note that a sequence with
zero or one MRI feature contributes nothing to the results of these programs.
sdist_machine.pl reads in one input file and produces a file suitable for graphing with
graph_sdist_hist.pl. sdist_machine_real_vs_rand.pl reads in three input files (one derived
from real sequences and two from MRI_randomizer.pl) and produces a file suitable for
graphing with graph_sdist_hist_real_vs_rand.pl. Both programs have optional arguments
for bin size (default = 1) and truncation of the dataset at a certain maximum distance (default = no truncation).

**graph_sdist_hist.pl** and **graph_sdist_hist_real_vs_rand.pl** – Both programs produce histograms based on their corresponding input files (see **sdist_machine.pl** and **sdist_machine_real_vs_rand.pl**).

**find_MRI_clusters.pl** and **graph_cluster_size_hist.pl** – The former program takes as input a "stretchy windows" output file from **MRI_analyzer_detail_v3.pl** and produces an output file suitable for graphing with the latter program. **find_MRI_clusters.pl** scans the input for clusters of MRI features of the same content type. There is one user-specified parameter: the "content threshold" (default = 60%). This is the threshold of content required in order for a locus containing a collection of MRI features to constitute a cluster. The program starts building a cluster with the first MRI feature it encounters. It attempts to add each subsequent MRI feature until the resulting cluster would violate the content threshold. The cluster parameters (start, end, length, type, # features) are then written out to two output files, one similar to the input format and suitable for graphing with **graph_cluster_size_hist.pl**, and the other a tab-delimited file with a ".dat" extension suitable for reading into a spreadsheet or statistical calculation package. Output files of the latter type were used with the R statistical computing environment (Rizzo 2007) to generate Figures 18 and 19.
Results

The Genomic MRI Internet Resource

The main visible product of this work is the Genomic MRI Internet Resource, a public resource for the investigation of GMRI. The site is currently located at http://mco321125.meduhio.edu/~jbechtel/gmri/ and works with any modern web browser. The introductory page appears as in Figure 2.
**Genomic MRI v1.0**

Bechtel J M, Wittenschnaeger T, Dwyer T, Song J, Arunachalam S, Ramakrishnan S K, Shepard S and Fedorov A

Genomic Mid-Range Inhomogeneity (MRI) is the notion that nucleotides are related to one another over distances greater than the oligonucleotide level but less than at the isochore level. Specifically, we have observed these relationships at a range of 30-1000 nucleotides.

In order to isolate the effects of these relationships from the many forms of Short-Range Inhomogeneity (SRI), one can compare the properties of a natural sequence with the properties of a random sequence with the same oligonucleotide composition. This site contains a set of tools to enable any researcher to do exactly that.

SRI analyzer profiles the oligonucleotide composition of a FASTA-formatted file containing RNA/DNA sequences. SRI generator uses that composition profile to generate randomized sequences matching the dimensions of a given file. MRI analyzer searches real or random sequence files for regions of mid-range inhomogeneity. Finally, MRI visualizer allows one to visualize the distribution of these regions on the concatenated sequence. See the How-to/README page for a more detailed outline of these tools.

Click the button below to start a new session or to resume a prior session.

**Start or Resume a GMRI Session**

- Download GMRI-related code
- About this site (scope, mission)
- GMRI departmental site
- Help (How-to/README)
- Links to relevant resources
- Glossary (explanations of key terms)

The source code for all of the GMRI tools is available from a link on this page. A glossary and helpful explanations are also provided. Upon clicking the large central button the user is taken to the 'Start/Resume Session' page shown in Figure 3.
**Figure 3:** The Start/Resume Session page

To start a new session, you must provide a file containing RNA/DNA sequences in FASTA format. The file can be up to about 21 mebibytes in size. To process larger files, download our programs and run them locally. Realistically, the file should contain sequences totalling at least 100,000 bases.

Use the forms below to start a new session or to resume a prior session.

**EITHER (a)**

Input FASTA sequence:

**OR (b)** Choose a file:

Enter a session label:

A session label is a randomly generated string of 6 letters and numbers shown in red at the top of the page while working in this site.

From this page the user can upload any suitable sequence to the website for investigation. Upon successful upload the user is presented with confirmation of the total file size as well as the sequence length and the number of FASTA records. Each new session is also given a random identifier called the "session label," which can be used to
resume a session later. Now that a session has been created with some sequence data, the user can proceed to use the GMRI analyzer tools. Because the first step will likely be to analyze the short-range inhomogeneity of the sequence, a button is provided on the Start/Resume session confirmation screen for convenience (Figure 4).

**Figure 4:** Start/Resume Session confirmation

![Image of Start/Resume Session confirmation]

Your file, "demo_file.txt", was successfully uploaded! For confirmation, the filesize is 109,138 bytes (106 KiB). The FASTA-formatted sequences file consists of 106,376 nucleotides in 1 records. Your session label for this data is K0TPES. Please make a note of it. (Case matter!)

You can now proceed to using the GMRI tools. You can read the How-to/README page for a detailed outline of these tools, or you can proceed directly to the SRI Analyzer:

The session label is shown in a fixed-width font and in red.

The SRI Analyzer interface appears as in Figure 5. Note that the first section merely provides some session information and serves as a header across all of the GMRI tools (except for CDS Generator, for reasons explained below). The middle section is for the specification of parameters and the results will appear in the lower section after clicking the 'Analyze File' button.
The selected "file" (at first only the original sequence) will be scanned for all n-mers from individual nucleotides up to and including the "maximum oligomer size." The maximum "maximum oligomer size" is currently limited to nine, which is sufficient for all practical purposes and helps to avoid problems with scaling. After clicking the 'Analyze File' button, the lower section is populated with links to download the composition file and a summary table, which is also presented for the user (Figure 6).

For each oligomer size scanned, the summary table lists the most and least common oligonucleotide(s) with the corresponding number of occurrences. Where more than one oligonucleotide had that number of occurrences, all are included in the table for download, but only the first two are displayed in the web page, followed by an ellipsis (Figure 6).

Once an oligonucleotide "composition file" has been generated by SRI Analyzer, the SRI Generator tool can then be used to generate randomized versions of sequence files.
using a particular oligonucleotide composition. The SRI Generator interface is shown in Figure 7.

**Figure 6:** SRI Analyzer results

<table>
<thead>
<tr>
<th>oligomer size</th>
<th>most-common</th>
<th>least-common</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>count</td>
<td>oligo(s)</td>
</tr>
<tr>
<td>1-mers</td>
<td>32196</td>
<td>T</td>
</tr>
<tr>
<td>2-mers</td>
<td>10883</td>
<td>TT</td>
</tr>
<tr>
<td>3-mers</td>
<td>4414</td>
<td>TTT</td>
</tr>
<tr>
<td>4-mers</td>
<td>1867</td>
<td>TTTT</td>
</tr>
<tr>
<td>5-mers</td>
<td>825</td>
<td>TTTTT</td>
</tr>
<tr>
<td>6-mers</td>
<td>420</td>
<td>TTTTTT</td>
</tr>
</tbody>
</table>
The selection of the oligomer level to use in SRI Generator is of critical importance. As indicated on the left in the SRI Generator interface, some choices are "weak" or "poor" due to underrepresentation of some oligonucleotides in the sequence. A "weak" oligomer level is one in which the number of occurrences of the least common oligonucleotide(s) is less than 10; a "poor" oligomer level is one with some oligonucleotides completely absent. If the user were to proceed with the "poor" choice, the randomization algorithm could enter a situation in which it had no oligonucleotide frequencies with which to continue. In this case, it would be forced to choose a nucleotide at random, which decreases the fidelity of
the algorithm to the supplied oligonucleotide composition. A weak oligomer level, on the other hand, merely implies that the oligonucleotide frequencies may be statistically weak.

While avoiding any suboptimal oligomer levels at the high end, the user will generally want to choose the highest reasonable oligomer level in order to preserve (control for) as much SRI as possible. This value is pre-selected as a convenience, but all choices are available to the user.

Once a "randomized sequence file" has been generated by SRI Generator, MRI Analyzer can be used to view the difference in MRI between the original sequence and its randomized counterpart. The MRI Analyzer interface is shown in Figure 8. The user now has two sequence files to choose from as well as seven different content types. The window size is left to the user, but a default value of 50 nt is pre-selected. The program will scan for content-rich and content-poor MRI using the user-specified window size and thresholds. The user can choose to enter the thresholds in whole nucleotides or as percentages, which would then be internally converted to whole nucleotides.
For each combination of parameters used, an "MRI composition file" is generated and stored along with a visualization of the MRI distribution. A link to the MRI composition file and the corresponding visualization are displayed in the lower section after clicking the 'Analyze File' button or selecting a previously analyzed combination of parameters (Figure 9). This allows for relatively fast switching between results.

The visualization used in the current release of the GMRI website we term a "spike graph." Each fixed-width non-overlapping window in which an MRI feature begins is marked by a spike in the graph, blue and above the baseline for content-rich, red and below the baseline for content-poor. Due to the arbitrary size and contiguity of the user-supplied sequences, the entire sequence is represented proportionally in a single fixed-width graph. This usually results in each window being represented by only a fraction of a pixel-width in the visualization output. Thus, due to both the fixed-width nature of the resulting MRI

**Figure 8: MRI Analyzer interface**

<table>
<thead>
<tr>
<th>Start/Resume Session</th>
<th>SRI Analyzer</th>
<th>SRI Generator</th>
<th>MRI Analyzer</th>
<th>MRI Generator</th>
<th>CDS Generator</th>
<th>Download Files</th>
</tr>
</thead>
</table>

session KOTYPES: "demo_file.txt" | 109,138 bytes (106 KiB) | 106,376 nt in 1 records/sequences

File to analyze: demo_file.txt

Content type: [GC Select]

Upper threshold: [50 Select by nucleotide]

Lower threshold: [Select by nucleotide]

[Analyze File] (Help)
composition file and the coarseness of the visualization, the spike graphs presented in MRI Analyzer are only intended to be a rough approximation of the MRI content. Because the representation of each window in pixels is not even, we provide a count of the number of spikes in each category in the legend at the bottom of the graph (Figure 9). This also allows the user to calculate a ratio (fold-change) in the number of MRI spikes between the original sequence and a randomized counterpart. For example, Figure 10 shows the results for a randomized version of the original sequence whose results are shown in Figure 9. The real:random ratio for GC-rich MRI features using a window size of 50 nt and a threshold of 60% (or 30 nt) is 315 / 80 = 3.9 and the ratio for GC-poor MRI features using a window size of 50 nt and a threshold of 20% (or 10 nt) is 254 / 66 = 3.8. This ratio is discussed below as both "fold-change" and "contrast ratio."
Now that an MRI composition file has been generated, MRI Generator can be used to create a randomized "mimic" of the original sequence using a combination of the
oligonucleotide composition and the MRI composition (Figure 11). The oligomer level must be chosen again, as in SRI Generator. Note also that MRI Generator may take a longer time to complete (possibly several minutes or more) depending on the size of the sequence and the thresholds chosen for the MRI composition. More stringent thresholds can result in requirements that are stochastically unreachable, such as a 200 nt window with 199 'G's.
When MRI Generator has finished, it provides a link to download the "mimic" sequence file. At this point the user may want to download the three types of sequence files (original, SRI-generated and MRI-generated) and analyze them using another method, such as the local RNA secondary structure analysis conducted in the manuscript.

Downloading individual files or the entire session can be accomplished using the "Download Files" tool. For the purposes of demonstration, however, if the user were to return to MRI Analyzer and now select the "mimic" sequence, the results would be

---

**Figure 11:** MRI Generator interface

---

Note: MRI-generator is more computationally expensive than the other programs.

---

**Session Notes:** "demo_file.txt" | 109,138 bytes (106 Kib) | 186,176 total records/sequences
somewhat different from the first MRI Analyzer results (Figure 12). The explanation for this difference is that MRI Generator uses the fixed-width non-overlapping window representation of the MRI composition from MRI Analyzer, yet this is only a rough approximation of the positions of the MRI features. Then this discrepancy is compounded when MRI Analyzer scans the mimic sequence using a fixed-width non-overlapping window again. New MRI spikes can be identified in the mimic due to the chance arrangement of content in adjacent randomly generated windows. Therefore the mimic sequences can only be regarded as rough approximations of MRI composition.

**Figure 12:** MRI Analyzer results for MRI-generated "mimic" sequence

Note that the numbers of spikes reported in the legend differ from those reported for the original sequence in Figure 9. See the text for discussion.

The fifth tool featured in the GMRI Internet Resource is CDS Generator. It was written by Tom Wittenschlaeger (see manuscript) for the special purpose of generating randomized coding sequences while preserving (controlling for) dicodon bias. For the
moment, this is a stand-alone tool (i.e. not linked with the GMRI Internet Resource and its session management) because it is species-specific and requires amino acid sequences as input. This tool may be integrated more fully into the GMRI Internet Resource in a future release.

MRI and Strong Local Secondary Structures

As presented in the manuscript, MRI Analyzer was written to scan a sequence and report the locations of MRI features in that sequence. This function was manually confirmed by visual inspection of the regions targeted as being content-rich or content-poor. By running MRI Analyzer on naturally occurring ("real") sequences and their randomized counterparts created with SRI Generator, one is able to observe the fold-change in the number of MRI features due to encoded genomic signals above the SRI level. By using MRI Generator and then scanning for putative RNA secondary structures, one can observe that it is the mid-range inhomogeneity in G+C-content that is overwhelmingly responsible for the abundance of strong putative RNA secondary structures.

Analysis of a sample of 11 large introns from three genes ("introns_3genes_BIG" includes introns from genes GRM8, DMD and HTR2C) allowed for quick and convenient testing of the algorithms. The next dataset was a non-redundant collection of 67,686 human introns. These were sorted into five groups according to G+C-content. The
boundaries of these five groups were taken from the accepted boundaries for the five human isochore classes (see Table 1). This is not meant to imply that these introns originate from genes in these isochore classes. In fact, the more G+C-rich isochores tend to have a mixture of G+C-rich and A+T-rich genes (Nikolaou and Almirantis 2002). Future studies may attempt to group introns by their true isochore associations, but it was more important for this study not to mix sequences of widely varying G+C-content. Including a wide range of G+C-content sequences would make it impossible to select generally applicable thresholds.

By experimenting with thresholds to obtain a number of MRI features that was enough to generate statistically significant results and yet not so many that they stopped being representative of the phenomenon, I was able to generate reliable histograms of the distances between MRI features within introns. The thresholds for each intron group are listed in Table 1.
Table 1: Intron groups with thresholds used to scan for MRI features

<table>
<thead>
<tr>
<th>isochore label*</th>
<th>G+C range (%)</th>
<th>number of introns</th>
<th>G+C-rich threshold (%)</th>
<th>G+C-poor threshold (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>32 to 37</td>
<td>13,089</td>
<td>74</td>
<td>8</td>
</tr>
<tr>
<td>L2</td>
<td>37 to 41</td>
<td>17,147</td>
<td>84</td>
<td>8</td>
</tr>
<tr>
<td>H1</td>
<td>41 to 46</td>
<td>16,233</td>
<td>86</td>
<td>12</td>
</tr>
<tr>
<td>H2</td>
<td>46 to 53</td>
<td>14,176</td>
<td>86</td>
<td>14</td>
</tr>
<tr>
<td>H3</td>
<td>53 to 60</td>
<td>7,041</td>
<td>86</td>
<td>24</td>
</tr>
</tbody>
</table>

[*Introns do not physically belong in the listed isochore classes; isochore labels were only used for convenient, consensus G+C-content thresholds.]

I then applied a simple randomization algorithm to distribute the MRI features randomly across the pool of introns and compared the naturally occurring distribution to the random distribution. These histograms seemed to show that G+C-rich MRI features tend to cluster together, while A+T-rich MRI features show a slight preference to spread out. (See Figures 13 and 14 below.)
Figure 13: Histograms of Distances Between G+C-rich MRI Features

Note that the real distribution of MRI features (red) shows a pronounced tendency toward clustering when compared to the randomized distributions of the same features on the same collection of intron sequences (green and blue).
Figure 14: Histograms of Distances Between A+T-rich MRI Features

Intrigued by the apparent clustering tendency of G+C-rich MRI features, I wrote

Note that the real distribution of MRI features (red) matches closely the randomized distributions of the same features on the same collection of intron sequences (green and blue), except for perhaps a slight tendency toward disaggregation.
Figure 15: Visualization of MRI features in some 53-60% G+C-content introns

G+C-rich MRI features are indicated in blue; A+T-rich MRI features are indicated in red. The scale for each sequence is independent and is given in its sub-heading in nucleotides per pixel.
The clusters of G+C-rich MRI features seemed to be separated by very small G+C-neutral gaps. Therefore, I next investigated whether increasing the base window size (in the stretchy window algorithm) would cause these gaps to be subsumed as the MRI features were allowed to expand a bit under these relaxed conditions. However, when enlarging the base window size, one must be careful to relax the content thresholds. If one were to maintain the thresholds at their relatively high levels for a window size of 50 while the window size grows to 200, the requirement for that very high level of content would eliminate most MRI features. Indeed, this is what was observed, and thus as the window size was increased to 200 nt and then to 400 nt, the content thresholds were concomitantly reduced to yield approximately the same number of distances between features. [Remember, these sequences are individual introns and one can only measure the distances between two features when they occupy the same intron.] While the larger window sizes confirmed the pattern of clustering in G+C-rich MRI features as well as the tendency to avoid clustering in A+T-rich MRI features, it did not produce noticeable agglomeration of the G+C-rich MRI features. The short G+C-neutral gaps remained (Figure 16 and Figure 17).
Figure 16: Visualization of MRI features in some 53-60% G+C-content introns using a 200 nt base window size.

G+C-rich MRI features are indicated in blue; A+T-rich MRI features are indicated in red. The scale for each sequence is independent and is given in its sub-heading in nucleotides per pixel.
**Figure 17:** Visualization of MRI features in some 53-60% G+C-content introns using a 400 nt base window size

*G+C-rich MRI features are indicated in blue; A+T-rich MRI features are indicated in red. The scale for each sequence is independent and is given in its sub-heading in nucleotides per pixel.*
Wanting to quantify the parameters of the observed clusters of G+C-rich MRI features, I wrote a program to scan for these clusters (find_MRI_clusters.pl) and another to produce histograms of their size and G+C-content (graph_cluster_size_hist.pl). The scanning procedure involves identifying groups of at least one MRI feature occupying a locus with at least a certain threshold of G+C-content. The algorithm attempts to add each successive MRI feature to the growing cluster as long as the resulting G+C-content of the newly expanded locus does not fall below the threshold. These scans were performed using thresholds of 40, 50, 60, and 70 percent G+C-content. The resulting histograms for cluster size indicate that the vast majority of MRI G+C-rich clusters are less than 500 nt with a simple majority being less than 250 nt (see Figure 18). The resulting histograms for cluster G+C-content show nearly all clusters having at least 70% G+C-content (see Figure 19).
Figure 18: Histogram of sizes of clusters of G+C-rich MRI features

The MRI features themselves are pooled from all five groups presented in Table 1. Note that 2 kb clusters are very rare, with the vast majority residing below the 500 bp mark.
**Figure 19:** Histogram of the G+C content of clusters of G+C-rich MRI features

The MRI features themselves are pooled from all five groups presented in Table 1. Note that the clusters tend to have a G+C content over 80%, which is very close to the high G+C content of the MRI features themselves.
Discussion

MRI was initially probed using a fixed window with sizes of 30 and 50 nt. These parameters were chosen somewhat arbitrarily but also because they were large enough to avoid sensitivity to most short-range signals and small enough so as not to miss any shorter MRI features by dilution. Based on these initial results, a window size of 50 nt was adopted for the rest of the studies, except where indicated.

Due to the fact that "local" (small) secondary structures are more reliable than "global" (large) secondary structures, but also for consistency, putative local RNA secondary structures were predicted using a window size of 50 nt. The Vienna RNA package was selected for this task for its superior applications programming interface (API), which allows one to write programs in C that directly use the Vienna RNA functions and resulting data structures.

The method for randomizing sequences while preserving oligonucleotide frequencies (SRI) is similar to that used in prior studies (Down et al. 2006; Fedorov et al. 2003; Forsdyke 1995; Katz and Burge 2003), differing in only a few uncontroversial details. The method for generating randomized sequences while preserving both SRI and MRI is novel, but it is simply an iterative application of SRI Generator that selects only segments with the appropriate MRI properties. The main innovation in the above work is in MRI Analyzer.
While it began as a fixed-size sliding window algorithm, it eventually acquired more sophistication. It was apparent that the arbitrary size of a fixed window would inevitably have an adverse effect on the detection of MRI. Thus, I changed the scanning algorithm to allow for a fixed "base" window size that then grows to encompass any region that continues to embody the target properties. We dubbed this algorithm "stretchy windows."

The fixed window algorithm was used for investigations in the manuscript and the stretchy window algorithm was used for all investigations thereafter.

The first study into the nature of GMRI concerns the distribution of MRI features in the human genome. We did not know at the outset whether any pattern would be apparent. As with the prior studies with putative local SS, we chose to compare real and random data. I wrote a program to randomize the positions of the MRI features in the same set of introns, with the qualifications that G+C-rich MRI features were distributed first and that no MRI features were allowed to overlap. The distances between each pair of adjacent MRI features of the same type (G+C- or A+T-rich) were plotted in a series of histograms at various scales. The choices of bin size and upper cut-off (truncation distance) result in an effective scale for viewing the distribution of the MRI features in a histogram plot. It was immediately apparent that the G+C-rich and A+T-rich MRI features were distributed differently. G+C-rich MRI features displayed a nice, monotonically decreasing histogram profile only at the smallest scale (bins of 50 nt distances), whereas A+T-rich MRI features
displayed a monotonically decreasing histogram profile only at large scales (bins of at least 1000 nt distances). By comparing each feature set to two randomized versions of itself at all scales, we could resolve that the G+C-rich features exhibit a tendency to cluster together and that the A+T-rich features were closer to a random distribution with perhaps a slight aversion to clustering.

To confirm our interpretation of the histograms as implying clustering of G+C-rich MRI features, I wrote MRI_mega_visualizer.pl, a program similar to MRI visualizer only specialized for "stretchy" MRI features. Instead of treating each MRI feature as a single spike, it scales the number of spikes to match the width of feature. This also allows one to zoom in on particular regions of interest, such as clusters of features. By carefully setting a threshold to select only those sequences (introns) with an interesting number of MRI features, a quick glimpse of the most profound clustering can be obtained, as seen in Figure 4. The clusters were then visually apparent.

Wanting to quantify the parameters of the observed clusters of G+C-rich MRI features, I wrote a program to scan for these clusters (find_MRI_clusters.pl) and another to produce histograms of their size and G+C-content (graph_cluster_size_hist.pl). The scanning procedure involves identifying groups of at least one MRI feature occupying a locus with at least a certain threshold of G+C-content. The algorithm attempts to add each successive...
MRI feature to the growing cluster as long as the resulting G+C-content of the newly expanded locus does not fall below the threshold. These scans were performed using thresholds of 40, 50, 60, and 70 percent G+C-content. The resulting histograms did not vary much between the different thresholds, so 60 percent was chosen as a reasonable cut-off. Perhaps of interest is that the histograms of G+C-content of the clusters revealed a possible bimodal peak, though the significance of this observation has not yet been investigated.

Most current research into genome compositional structure and organization focuses on the very small and very large scales of oligonucleotide frequencies and isochores, respectively. While some prior studies have covered "mid-range inhomogeneity" (MRI) in some way, it has never before been approached with the method I have developed. By creating a modified sliding window analysis ("stretchy window") that allows for features larger than the base window size, a more flexible characterization of MRI was possible. This work details the initial investigations into the nature of MRI: its prevalence and properties. Follow-up studies (already underway in our lab) are looking at the effects of various mutations on MRI-defined regions using a human-chimp-macaque triple alignment and the SNP database (a.k.a. dbSNP) (Sherry et al. 2001).
The forces for creating MRI have yet to be established, but the potential implications for genomic organization are great. The human genome, for example, is responsible for coordinating the behavior of trillions of cells organized into hundreds of tissue types and subtypes. It must specify the fine machinations of development, the topology of the brain, and the complex regulatory networks that govern homeostasis in multiple systems over the lifetime of an individual. The current annotation of the complete human genome does not appear to account for the quantity of information necessary to direct this sophisticated program. Thus, the remaining information must be present in currently unannotated forms. MRI may represent one of those cryptic forms of information, based not in particular sequence motifs but rather in a non-random organization of mid-range nucleotide distribution. Thus, MRI could signal a new kind of genomic information requiring a new layer of genome annotation.
Conclusions

1. Mid-range inhomogeneity (MRI) exists throughout the human genome in all four individual nucleotides as well as in all pairwise combinations of nucleotides (G+C, A+G, and G+T) and their complements.

2. MRI in G+C-content and A+T-content is associated with an abundance of strong local secondary structures (SS).

3. G+C-rich MRI features occur in clusters while A+T-rich MRI features do not.

4. G+C-rich MRI clusters have an average size of approximately 100 nt and a maximum size of approximately 2000 nt.

5. A public resource has been created to support further investigations of genomic MRI.
Summary

Mid-range inhomogeneity (MRI) is a mostly overlooked genomic property, except for CpG islands. Relevant prior studies have not gone beyond quantification of the level of inhomogeneity in some known sequences. The work laid out in this thesis quantified the nature of MRI. It established an approximate range of sizes for these features and quantified their distribution patterns in the non-coding regions of the human genome. The cumulative evidence seems to support the notion that MRI is a real feature of the human genome and that G+C- and A+G-content reveals its presence in both coding and non-coding genomic regions.
References


Oligonucleotides, and a Probable Important Role of Scrambled Stepwise
Duplications/Deletions in Molecular Evolution. Molecular Biology and Evolution
18, 1231-1245.

prediction of isochores based on hidden Markov models. Gene, Section
Evolutionary Genomics 385, 41-49.

CABIOS 11, 195-199.


clustering in DNA sequences of various origin and functionality, by means of a

27. Nicolay, S., Argoul. F., Touchon, M., d'Aubenton-Carafa, Y., Thermes, C., and
to the Organization of Gene Location and Orientation? Physical Review Letters 93,
108101(4).


Abstract

Genomes encode many signals at multiple levels. They must therefore manifest inhomogeneity in their nucleotide composition. Inhomogeneity is identifiable from the short-range where neighboring nucleotides influence the choice of base at a site, to the long-range, commonly known as isochores, where a particular base composition can span millions of nucleotides. This work explores the mostly overlooked phenomenon of genomic mid-range inhomogeneity (MRI). MRI is defined as an interdependence between nucleotide choice and base composition over a distance of 30-1000 base pairs. Through the investigation of thousands of human genes, the results presented herein establish the existence of MRI throughout mammalian genomes, establish a link with strong local secondary structures and characterize the properties of MRI. A public computational resource has been created to support further study of genomic MRI.