4-2003

# Identifying Patterns in DNA Change

Jason R. Gilder
*Wright State University - Main Campus*

Dan E. Krane
*Wright State University - Main Campus*, dan.krane@wright.edu

Travis E. Doom
*Wright State University - Main Campus*, travis.doom@wright.edu

Michael L. Raymer
*Wright State University - Main Campus*, michael.raymer@wright.edu

# Identifying Patterns in DNA Change

Jason R. Gilder, Dan E. Krane, Travis E. Doom, and Michael L. Raymer

*Abstract*— **Now that a draft sequence of the human genome is nearly complete, questions regarding both the information contained within our genetic blueprints as well as the manner in which that information content changes over time can be addressed in ways that had not previously been possible. By their very nature, some of the nucleotide sequences present within our genome allow detailed examination of the mode and pattern of evolution that has shaped our genetic instructions over time spans of tens of millions of years. *Alu* repeats are one example. Using these relatively short, ubiquitous DNA sequences we explore the problem of attempting to predict the relative abundance of a variety of different possible substitution events that have accumulated over the past 20 million years. To perform well when applied to biological sequence data, computational methods must have the ability to tolerate both natural variation in the data and noise introduced in data measurement. As a result and due to their ability to search complex, noisy search spaces, Evolutionary computation techniques are particularly promising for the analysis of nucleotide sequence data and other biological data sets. We have used these techniques to address a key question in understanding the process of evolution: the effect of genomic context on substitutions (the degree to which the genomic information surrounding a particular region of a chromosome affects the changes to that region over time). We utilized genetic programming to predict changes in these DNA sequences over time. These approaches reveal that a significant proportion of DNA nucleotide substitutions within a given region are governed by a model that takes into consideration only the GC-content of the DNA sequences surrounding the region being considered.**

***Index Terms***—**Alu Repeats, Bioinformatics, Classification, Genetic Programming, Substitution Rates**

Manuscript received March 1, 2003. This work was supported in part by the National Science Foundation (grant #EIA-0122582), and by the Dayton Area Graduate Studies Institute.

J. Gilder is with the Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435-0001 USA (e-mail: jgilder@cs.wright.edu).

D. Krane is with the Department of Biological Sciences, Wright State University, Dayton, OH 45435-0001 USA (e-mail: dan.krane@wright.edu).

T. Doom is with the Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435-0001 USA (937-775-5105; fax 937-775-5133; e-mail: travis.doom@wright.edu).

M. Raymer is with the Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435-0001 USA (e-mail: michael.raymer@wright.edu).

## I. Introduction

### *A. Exploring the Human Genome*

A genome is the sum total of an organism's heritable information that can be passed from one generation to the next. The bulk of that information is stored in the specific order in which four different chemical units (commonly abbreviated as G, A, T and C) called nucleotides are linked together in long chains to make DNA molecules. With the determination of the sequence of roughly three billion nucleotides that comprise the human genome nearly complete, we are presented with new opportunities to examine our fundamental makeup. One such problem is exploring and identifying factors that govern how our genome has changed and is continuing to change over time. We can determine the changes that have taken place by comparing a sequence of our own genome with a homologous region ( a region that is derived from a single sequence in a common ancestor) in another organism. DNA sequences that are functionally constrained change very little due to the fact that a mutation often limits (in extreme cases, by death) the affected organism's ability to pass that mutation onto subsequent generations. Sequences that are not functionally constrained are free to change. Analyses of homologous sequences, such as those of *Alu* repeats, which are free of selective constraint, have the potential to give insights into underlying boundaries associated with mutational processes.

### *B. Alu Repeats*

The genome of every mammalian order (such as primates, rodents, carnivores, and artiodactyls) studied to date have been found to possess their own characteristic family of short interspersed repetitive elements (SINEs) (reviewed in Deininger and Batzer, 1993). *Alu* repeats are the predominant SINE in primate genomes (Deininger and Batzer, 1993). Like typical SINEs, *Alu* repeats have an average length of about 280 bp and account for roughly 10% of the primate genomes where they are found (Houck, Rinehart and Schmid, 1979; Sun *et al.,* 1984; Hwu *et al.*, 1986).

*Alu* repeats, like other SINEs, have been propagated throughout primate evolution by a process known as retrotransposition (Schmid and Shen, 1985; Weiner, Deininger and Efstradiatis, 1986) in which a "master" copy of the repeat is transcribed (made into an RNA copy by an enzyme called RNA polymersase), reverse transcribed (by an enzyme called reverse transcriptase such as those typically associated with retroviruses like HIV) and then reinserted into the genome at a distant site. The result has been an expansion to an estimated 500,000 to 1,000,000 copies of *Alu* repeats (Rinehart *et al.,* 1981; Jurka *et al.,* 1993) within the human genome. Copies of these *Alu* repeats are generally free of selective constraint (Labuda and Striker, 1989; Batzer *et al.,* 1990) and remain stably inserted for at least tens of millions of years (Koop *et al.,* 1986; Sawada and Schmid, 1986).

*C. Predicting Substitution Rates*

Over time, errors in DNA replication and repair introduce changes into a genome. At the level of the four possible nucleotides at any particular position within a genome, only a relatively small number of changes in state (substitutions) can be observed. Our goal is to produce a function capable of accusatively predicting the number of such changes in a given region (over time). Such a function could provide significant insight into the biological factors that drive substitutions across the entire genome and have implications for the study of disease-causing mutations that continue to accumulate.

For each of the twelve possible changes of state (G to A, T, or C; C to A, G, or T, etc.) the number of changes is predicted. Correct predictions are classified by simple difference, using the test below:

$$|\text{PREDICTED SUBSTITUTIONS} - \text{ACTUAL SUBSTITUTIONS}| < 0.5$$

Rounding will account for any absolute error less than 0.5. A fitness function can utilize either the absolute error or the classification rate (percentage of substitutions correctly predicted).

*D. Feature Selection and Extraction*

Since all members of a particular *Alu* subfamily of repeats are considered to have begun with exactly the same nucleotide sequence at the time of its propagation, no information about the progenitor itself is used in predicting the type and quantity of substitutions for a given repeat. Rather, each *Alu* repeat is characterized according to 16 features of the repeat itself and its genomic context: the length of the repeat, the number of A's, G's, C's, and T's within the repeat copy, the GC-content of the repeat itself, and the GC-content (the fraction of nucleotides that were either G or C as opposed to A or T) of ten flanking regions of various sizes.

## II. Methods

CENSOR (Jurka *et al.,* 1996; GIRI, 2003) was utilized to identify *Alu* repeats in the human genome (GenBank release 133.0). 6,749 repeats belonging to the *Alu*-Y subfamily were obtained from the 274,400,000 bp of sequence available from human chromosome 1. All changes in the *Alu*-Y family were recorded as well as the GC-content for five flanking regions on each side of the repeat (500, 1000, 5000, 10000, and 20000 nucleotides).

*2.1 Genetic Programming*

Genetic programming (GP) attempts to create an equation that solves a given problem by creating several randomly generated expression trees.

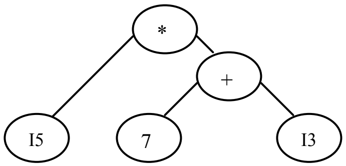

***Figure 2.1.1:*** A GP tree for the expression (I5 * (7 + I3))

GP expression trees are made up of input, operator, and constant nodes (see Figure 2.1.1). Inputs are features that are evaluated directly from the data set. Operators come from a predefined list of possible operations that can be performed inside an equation. Like genetic algorithms, GP employs mutation and crossover during reproduction. Crossover randomly chooses two GP expressions, randomly cuts a link in each tree, and combines the results, forming a new expression. Mutation replaces a randomly chosen node with a new randomly generated node of the same type.

GP provides a very flexible environment for optimizing discriminant functions for pattern classification, in that it does not require a fixed form for the equations that it creates (Bäck and Schwefel, 1996; Koza, 1992). A wide variety of nodes can be utilized to create an equation tree of any complexity. This flexibility has allowed GP-based classifiers to be developed for a variety of biological data types (Raymer *et al.*, 1996). Other methods of evolutionary computation are being used to solve biological problems as well (Raymer *et al.*, 1997). Genetic Programming also offers dimensionality reduction by selecting the features that provide the greatest fitness (Raymer *et al.,* 1997, Raymer *et al.,* 2000). For this application, the fitness function utilized the absolute value of the error between the number of substitutions predicted by the GP and the number of actual substitutions, which were previously calculated in the data set.

*2.2 Functions and Terminals*

The types of nodes available dictate what type of equation can be created. The traditional operations of add, subtract, multiply, and divide form the base of the set. The following functions were also included:

***Min*** *– returns the minimum of two nodes or subtrees*

***Max*** *– returns the maximum of two nodes or subtrees*

***Cos*** *– if the connected nodes are x and y, it returns x * Cos(y)*

***Sin*** *– if the connected nodes are x and y, it returns x * Sin(y)*

***Ave*** *– if the connected nodes are x and y, it returns (x + y)/2*

***Log****– if the connected nodes are x and y, it returns x * Log(y)*

*2.3 Mask Operator Terminals*

In addition, a set of mask operator terminals was added.

Each terminal contains a mutable binary mask that selects the features to be used. Mask operator nodes only occur in the leaves of a tree and do not contain any leaves of their own. The mask operator terminals are as follows:

**Summation**: $\sum_i f_i m_i$

**Multiplication**: $\prod_i f_i m_i$

**SumSquareRoot**: $\sqrt{\sum_i f_i m_i}$

Where ***m*** is a binary mask vector of length 16. Each bit in ***m*** is associated with a particular feature. For example, if ***m*** = [0000000000001111] for a summation node, the value of the node would be the sum of the feature values for features 13 through 16.

### *2.4 Initial GP Experiments*

The data set of 6,749 examples was broken into 5,000 examples for training and 1,749 examples for holdout testing. The fitness function was the absolute average error across all training examples. The first classification problem attempted was predicting the number of C's that were previously G's. Classification rates were observed around 46% in the first few generations. Unfortunately, little improvement of the initial classification rates was realized during subsequent generations. The classification rate remained at around 46% and the average of the absolute error (the fitness value) converged at around 0.75.

### *2.5 Theta Correction: The Offset Factor*

The average absolute error value of 0.75 was nonetheless encouraging, because an absolute error of ≤ 0.5 can be corrected by rounding each predicted value to the nearest integer. Thus, if the average absolute error was lessened by only 0.26, many more examples would possibly be classified correctly.

If an equation is consistently 0.75 off of the desired result, then we can achieve the correct result by simply applying a constant bias to all predicted results. The fitness function was changed to the following for a correct classification:

$$\left|\,|\text{RESULT} - \text{DESIRED}| - \text{OFFSET}\,\right| < 0.5$$

An offset of 0.30 was used as an initial test to see how well the examples would be classified. The change was dramatic, as the classification rate jumped to around 66%, indicating that the theta correction factor was performing as intended.

Determining the offset factor could be done experimentally, through a mutation operator, or simply exhaustively. The latter was chosen because the ideal offset is likely between zero and one as the solution converges. With each tree evaluation, the fitness is calculated with every possible theta between 0 and 1 in 0.01 intervals. The theta that results in the most correctly classified training examples is used.

### *2.6 Experiments Utilizing the Theta Offset Factor*

The previous experiment of predicting the number of G to C substitutions was performed using the theta offset factor. Training and testing data sets were chosen at random at the start of each experiment. A population and offspring size of 700 individuals each was utilized. Initial experiments yielded trees with fewer than 10 nodes, so a mutation rate of 0.20 was used. Initial trees were chosen to be between 3 and 50 nodes. The GP consistently chose an offset between 0.5 and 0.6 and resulted in a classification rate converging at 79%.

| | | ***Progenitor Sequence*** | | | |
|---|---|---|---|---|---|
| | | **A** | **C** | **G** | **T** |
| ***Alu Repeat*** | **A** | - | 76, 76 | 23, 21 | 94, 95 |
| | **C** | 96, 96 | - | 76, 76 | 80, 79 |
| | **G** | 81, 82 | 79, 79 | - | 96, 96 |
| | **T** | 91, 91 | 28, 28 | 66, 66 | - |

***Figure 2.6.1:*** Classification rates for all substitutions. Classification rates listed as [training rate, test rate].

Further experiments yielded impressive results for classifying other possible substitutions (see Figure 2.6.1). Nine out of twelve holdout classification rates were above 75%, with four above 90%. The two lowest classification rates (C to T and G to A) can be accounted for due to these changes of state occurring primarily as a result of a distinct and separate substitution process related to methylation of human DNA sequences. Substitutions involving 5'-CG-3' dinucleotides are known to be heavily influenced by methylation-related mutagenesis (Coulondre *et al*., 1978; Razin and Riggs, 1980). Methylation within the human genome occurs only at the C's of 5'-CG-3' dinucleotides. Oxidative deamination (a common form of DNA damage) of methylated C's typically causes T's to be put in place of those C's or A's in place of their associated G's during the process of DNA replication.

### *2.7 Narrowing the Feature Set*

Examining the expression trees from the previous experiments showed that there was a heavy reliance on flanking GC-content. This suggested the intriguing idea that perhaps all substitutions could be predicted through context information alone. All twelve experiments were redone, this time using only the GC-content of 10 flanking regions.

| | | ***Progenitor Sequence*** | | | |
|---|---|---|---|---|---|
| | | **A** | **C** | **G** | **T** |
| ***Alu Repeat*** | **A** | - | 76, 76<br>(91, 91) | 20, 19<br>(49,51) | 94, 95 |
| | **C** | 96, 96 | - | 73, 73<br>(80,80) | 80, 79 |

| | | | | | |
|---|---|---|---|---|---|
| | **G** | 81, 82 | 78, 76<br>(91, 91) | - | 96, 96 |
| | **T** | 91, 91 | 22, 25<br>(60, 64) | 66, 66<br>(86, 85) | - |

***Figure 2.7.1:*** Classification rates based on GC context. Classification rates listed as [training rate, test rate]. Classification rates in parentheses are classifying CpG masked rates.

The classification rates remained largely the same as the previous experiments, demonstrating that content information of the repeat itself is largely an unimportant feature of the predictive model (see Figure 2.7.1). The only changes in classification were in classifying C to G substitutions (from 79, 79 to 78, 76) and G to C substitutions (from 76, 76 to 73, 73). There was no apparent bias as to which scale of flanking GC-content was preferred, and these values are in fact highly correlated with one another.

As in previous experiments, the CpG-dinucleotides again proved to be a problem, this time resulting in even lower prediction rates. Masking CpG-dinucleotides from the analysis resulted in a substantial improvement in the classification rates. Classification accuracy for C to T substitutions jumped from (22, 25) to (60, 64) and accuracy for G to A substitutions went from (20, 19) to (49, 50). The increases were fairly substantial, but their rates still constituted the lowest of all twelve predictions. Interestingly, G to C substitutions were classified with 80% accuracy, and both C to G substitution and C to A substitution classification rates were above 90%. G's to T's, which was previously the third lowest classification rate, increased to (85, 86), placing it in the same range as the other prediction rates. In fact, the overall classification rates were 80% or above for ten out of twelve predictions, with six reaching more than 90%.

Surprisingly, the size of the GP trees did not balloon as the number of generations increased (a problem common to GP). The smallest tree consisted of only three nodes, while the largest contained 25. The average tree size for all twelve substitution models was ten nodes.

## III. Conclusions

The GP-optimized discriminant functions illustrated that the most important features available for classification of substitution rates were the region's flanking GC-contents. No information on the *Alu* repeat itself was needed to predict the repeat's substitution pattern, a fact both surprising and illuminating. These results suggest that *Alu*'s must either undergo specific changes that are dependent upon their surroundings or *Alu*'s are inserted in specific locations based on their current configuration (an alternative that is substantially at odds with the current model of how the repeats are retrotransposed from a single master progenitor).

Substitutions from A or T were fairly consistent, while those from C or G were much more varied. Much of this has to do with independent, competing trends, such as methylation-related mutagenesis. Masking CG-dinucleotides improved classification rates dramatically.

We are currently exploring *Alu* repeats in other families throughout the entire genome to see if the same classification trends can be established. We are also working on comparing the models generated by the GP to see if any generalizations can be made about mutation processes. Genetic algorithms are also being explored as a means to solve this problem. The work will be presented in a later manuscript.

*Alu* repeats are often thought of as "junk" DNA in that they do not seem to add anything to the functionality of organisms in which they are found, but they contain a wealth of information about the evolutionary history of primates. Interpreting that information in these regions free of selective constraint can also provide insights into the changes that have taken place in functionally constrained regions, such as those associated with genes. Ultimately, a better appreciation of the constraints on models of the substitution process should yield improved understanding of the mutation and evolution process that has operated and continues to operate upon the nucleotide sequences of the human genome.

## VI. References

Batzer, M. A., G. E. Kilroy, P. E. Richard, T. H. Shaikh, T. D. Desselle, C. L. Hoppens and P. L. Deininger, "Structure and variability of recently inserted *Alu* family members," Nucleic Acids Res. **18:**6793-6798, 1990.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival and F. Rodier, "The mosaic genome of warm-blooded vertebrates," Science **228:**953-958, 1985.

Coulondre, C., J. H. Miller, P. J. Farabaugh and W. Gilbert, "Molecular basis of base substitution hotspots in *Escherichia coli*," Nature **274:**775-780, 1978.

The Genetic Information Research Institute (GIRI), CENSOR Server, http://www.girinst.org/Censor_Server.html, 2003.

Houck, C. M., F. P. Rinehart and C. W. Schmid, 1979, "A ubiquitous family of repeated DNA sequences in the human genome," J. Mol. Biol. **132:**289-306, 1979.

Hwu, H. R., J. W. Roberts, E. H. Davidson and R. J. Britten, "Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution," Proc. Natl. Acad. Sci USA **83:**3875-3879, 1986.

Jurka, J., P. Klonowski, V. Dagman and P. Pelton, "CENSOR - a program for identification and elimination of repetitive elements from DNA sequences," Computers Chem. **20:**119-122, 1996.

Koop, B. F., M. M. Miyamoto, J. E. Embury, M. Goodman, J. Czelusniak and J. L. Slightom, "Nucleotide sequence and evolution of the orangutan ε globin gene region and surrounding *Alu* repeats," J. Mol. Evol. **24:**94-102, 1986.

Koza, J. R. *Genetic Programming*, Cambridge, MA: MIT Press, 1992.

Krane, D. and M. Raymer, *Fundamental Concepts of Bioinformatics*, San Francisco, CA: Benjamin Cummings, 2003.

Labuda, D. and G. Striker, "Sequence conservation in Alu evolution," Nucleic Acids Res. **17:**23477-2491, 1989.

Pei, M., E.D. Goodman, W.F. Punch, and Y. Ding, "Genetic algorithms for classification and feature extraction," presented at *Classification Society Conference*, June 1995.

Raymer, M.L. W.F. Punch, E.D. Goodman, and L.A. Kuhn, "Genetic programming for improved data mining – application to the biochemistry of protein interactions," in Genetic Programming 1996: Proceedings of the First Annual Conference, (Stanford University, CA), pp. 375--380, MIT Press, 28--31 July 1996.

Raymer, M.L., P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman, and L. A. Kuhn, "Predicting conserved water-mediated and polar ligand interactions in proteins using a k-nearest-neighbors genetic algorithm," *J. Mol. Biol.*, vol. 265, pp. 445–464, 1997.

Raymer, M.L., W. F. Punch, E. D. Goodman, P. C. Sanschagrin, and L. A. Kuhn, "Simultaneous feature scaling and selection using a genetic algorithm," in *Proc. Seventh Int. Conf. Genetic Algorithms (ICGA)* (T. Bäck, ed.), (San Francisco), pp. 561–567, Morgan Kaufmann Publishers, 1997.

Raymer, M.L., W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computing*, vol. 4, no. 2, pp. 164–171, 2000.

Razin, A. and A. D. Riggs, "DNA methylation and gene function," Science **210:**604-610, 1980.

Sawada, I. and C. W. Schmid, "Primate evolution of the α-globin gene cluster and its *Alu*-like repeats," J. Mol. Biol. **192:**693-709, 1986.

Schmid, C. W. and C. K. J. Shen, "The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates," In: MacIntyre R. J. (ed) Molecular evolutionary genetics. Plenum Press, New York, pp. 323-358, 1985.