DNA Visual And Analytic Data Mining

Patrick Hoffman[1], Georges Grinstein[1]
Kenneth Marx[2] , Ivo Grosse[3], Eugene Stanley[3]

[1]Institute for Visualization and Perception Research
Department of Computer Science
University of Massachusetts Lowell
Lowell, MA 01854

[2]Center for Intelligent Biomaterials
Department of Chemistry
University of Massachusetts Lowell
Lowell, MA 01854

[3]Center for Polymer Studies and Department of Physics
Boston University
Boston, MA 02215

Abstract

In this paper we describe data exploration techniques designed to classify DNA sequences. Several visualization and data mining techniques were used to validate and attempt to discover new methods for distinguishing coding DNA sequences, or exons, from non-coding DNA sequences, or introns. The goal of the data mining was to see whether some other possibly non-linear combination of the fundamental position dependent DNA nucleotide frequency values could be a better predictor than the AMI[6]. We tried many different classification techniques including rule-based classifiers and neural networks. We also used visualization of both the original data and the results of the data mining to help verify patterns and to understand the distinction between the different types of data and classifications. In particular, the visualization helped us develop refinements to neural network classifiers, which have accuracy's as high as any known method. In the conclusion, we discuss the interactions between visualization and data mining and suggest an integrated approach.

1 Introduction

The international effort called the Human Genome Project is rapidly sequencing the complete DNA sequences of all 24 human chromosomes. As well, the chromosomes from a number of other organisms are being entirely sequenced. The DNA component of chromosomes are long linear molecules comprised of strings of the four nucleotides (A, C, T, G), the information bearing chemical units. Coding sequences (exons) are interspersed by non-coding sequences(introns) along the chromosomes whose information encodes protein structures. Transcription of the coding DNA sequence into mRNA, which is then translated into proteins in the cell comprise the general flow of information. This process is responsible for all normal cellular functions as diverse as development into multicellular organisms, organ development, the immune system, to name a few, as well as abnormal function such as cancer, birth defects, etc.

0-8186-8262-0/97 $10.00 Copyright 1997 IEEE.

The current approach for finding genes (protein coding sequences) is both experimental and computational. Any small increase in the accuracy of computer classification can therefore result in substantial time and cost savings. In this paper we describe our experiences to harness data exploration techniques to classify DNA sequences.

In order to use visualization and data mining techniques to develop new methods for distinguishing coding DNA sequences (exons) from non-coding DNA sequences (introns), it is necessary to represent symbolic DNA sequences by numbers or vectors. It has been demonstrated by Fickett et al.[4] that the proper choice of this representation is as important as the later processing of the numbers by neural nets or other classification schemes. The representation of DNA sequences we chose was guided by the recent discovery that a non-linear correlation statistic for DNA sequences, called the average mutual information (AMI), [6], is capable of distinguishing coding from non-coding DNA sequences in all taxonomic classes ranging from the most simple to the most complex organisms. mathematically, the AMI is a non-linear function based on the vector of 12 frequencies p_i^k by which the nucleotide i = a, c, g, t appears in position k = 1, 2, 3 relative to a given reading frame in a small segment of a DNA.

The goal of our study was to see whether some possibly non-linear combination of these position dependent DNA frequency values could be a better predictor than the AMI.

After a brief introduction to the various data mining techniques, we discuss our analytic approach, the visualizations we used and developed, and the results of these analyses. We also briefly suggest how visualization and data mining could be integrated in the future.

2 Data Mining

2.1 Tools

A number of data mining tools are available. Some of the ones we used include Tooldiag[12], and Stuttgart Neural Network Simulator SNNS [15,17] . We concentrated our efforts on applying Clementine[2]. Clementine is a data mining suite based on the data flow visual programming paradigm similar to AVS or IBM's Data Explorer. It provides four machine learning modules, two rule-based algorithms, a standard neural net (multi-layer Perceptron), and a Kohonen neural net for clustering, each with default settings. Elaborate tuning is possible but not necessary to get some early results. One rule-based classifier in Clementine is the C4.5 algorithm by Quinlan [11].

2.2 Data Mining DNA Sequences

Fickett [4] developed databases consisting of sequences of known exons and introns and described the accuracy of several classification methods. Some methods depend on knowing the particular starting and ending sequences and most require elaborate training on exon and intron data sets from the organism under consideration. In [6] the Mutual Information function was developed and studied. It provides a non-linear measure of the correlation between a particular nucleotide and another ***n*** nucleotides away.

For biochemical reasons coding sequences possess a triplet codon information structure. Thus nucleotides at positions 3, 6, 9, and generally ***3n*** , positions away from each other have higher correlations. Thus we only need to look at the frequencies of A, C, T, G extracted from a small segment of DNA, at positions 1, 2, and 3. [7] defines the AMI as a particular combination of these 12 values and uses it as a predictor for distinguishing exons from introns. It classifies DNA sequences with a high degree of accuracy (76 for 108 bp and 81% for 162 bp). Is it possible that there are other functions of these values that can even better distinguish exons from introns?

We decided to use data mining to help find these functions. In our initial study we examined several thousands of Fickett's sequences of various length exons and introns. Our first task was to divide the data into training and test sets. The training sets were used to build a classifier, similar to the rule-based ID3 [10], and a neural net. The test sets were used to evaluate the accuracy of the classifier.

To date the various classification programs have reached accuracy's between 73 and 81 percent[4]. This provided a baseline that we were comparing against. Since there are only two classes of data any classifier should reach at least a minimum accuracy of 50%.

Initially, only 200 points (100 exons and 100 introns) were used to train Clementine's 2 rule-based classifiers and neural net. We used DNA sequences of 162 base pairs for all of our classifiers. The default NN used 12 input nodes , 4 hidden layers and 1 output node. After training the rule-based classifiers were correct 93 and 94 percent of the time while the NN was only about 80% accurate on the small training data. However, with 800 non-training samples (400 exons and 400 introns) we obtained the following accuracy:

	Correct	Wrong
Neural Net	638 (79.55%)	164 (20.45%)
C4.5 RULE	551 (68.70%)	251 (31.30%)
Clementine Rule	573 (71.45%)	229 (28.55%)

Here, rule based classifiers were inferior to the NN classifiers. These initial results were very promising for the neural network, since we knew we could potentially tune the network for better performance, and we still were using a small training set.

At this point we were ready to explore other packages, but more pressing was trying to understand in greater detail the structure of the data we had generated. Unfortunately, neural networks and classification rules do not easily reveal their insights into the data. We wanted to "see" these differences! Visualization was needed.

3 Visualization

We used several visualization approaches to look both at the original Fickett data as well as processed data.

3.1 Radial Visualizations

Spring constants can be used to represent relational values between points [1,9]. We developed a radial visualization(Radviz), similar in spirit to parallel coordinates (lossless visualization), in which ***n***-dimensional data points are laid out as points equally spaced around the perimeter of a circle. One end of ***n*** springs are attached to these ***n*** perimeter points. The other ends of the springs are attached to a data point. The spring constant K_i equals the values of the i-th coordinate of the fixed point. Each data point is then displayed where the sum of the spring forces equals 0. All the data point values are usually normalized to have values between 0 and 1.

For example if all ***n*** coordinates have the same value the data point will lie exactly in the center of the circle. If the point is a unit vector then that point will lie exactly at the fixed point on the edge of the circle (where the spring for that dimension is fixed). Many points can map to the same position as in the Exvis displays [3]. This represents a non-linear transformation of the data which preserves certain symmetries and which produces an intuitive display. Some features of this visualization include:

- points with approximately equal coordinate values will lie close to the center
- points with similar values whose dimensions are opposite each other on the circle will lie near the center
- points which have one or two coordinate values greater than the others lie closer to those dimensions
- An n-dimensional line will map to a line
- A sphere will map to an ellipse
- An n-dimensional plane maps to a bounded polygon

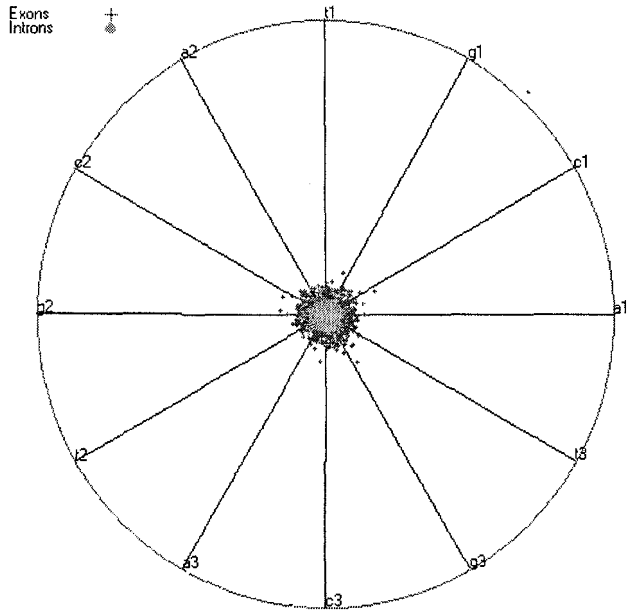

Figure 1 Radviz 1000 exons(red) 1000 introns(green)

Figure 1 displays 2000 points using Radviz: red points are exons and green points are introns. Most points lie close to the center implying equal forces. However, introns lie closer to the center than exons.

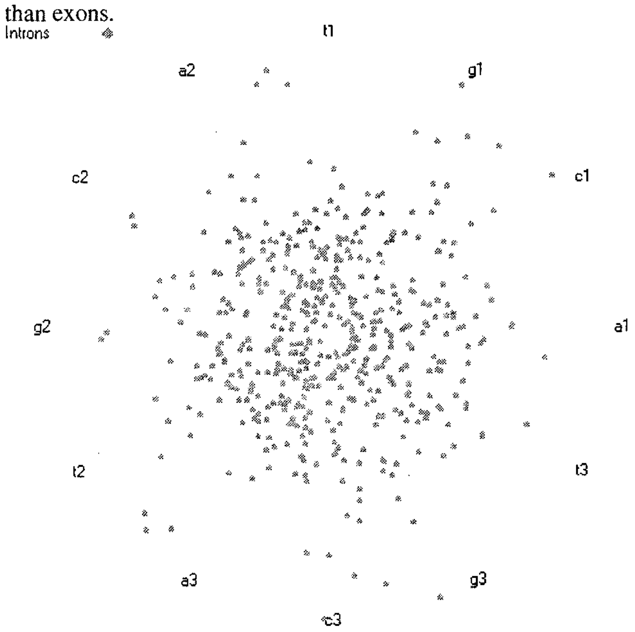

Figure 2 500 Introns - expanded X 10

In Figure 2, 500 Introns are displayed, zoomed by a factor of 10 with the point size increased. In this picture we discovered a "symmetry" of the data around a line drawn between dimensions c2 & g2 and a1 & t3. This mirror image is a consequence of the complementary pair nature of Fickett DNA sequences: consecutive lines of the sequences were the reverse of the previous lines. This symmetry needed to be corrected for some aspects of the data mining.

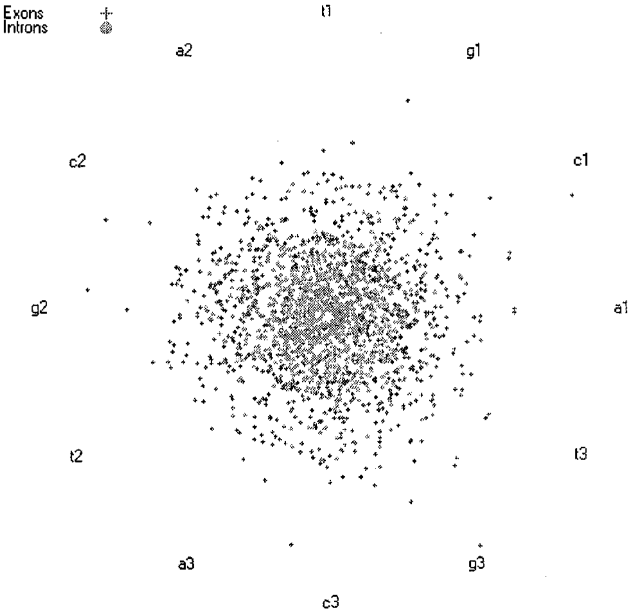

Figure 3 2000 pts expanded 5 times

Figure 3 displays 2000 points zoomed up by a factor of 5 . In this picture we can see that the exons (red +) are more spread out, and the introns (green) are closer to the center of the circle.

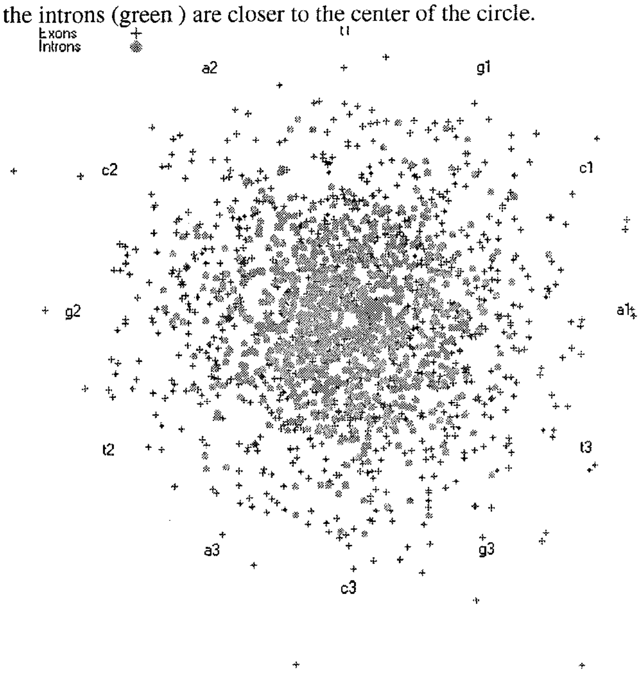

Figure 4 Expanded X 8

Figure 4 represents the same data with a zoom factor of 8 and larger points. Notice that zooming produces points well outside the circle which is not possible with real springs. The explanation for the spreading of exons is that they are not as random as introns. The random frequency distribution tends to make the forces balance, hence points closer to the center.

Radviz can effectively display many dimensions. Figures 5 and 6 show exons/introns using 16 dinucleotide dimensions and 48 dimensions of dinucleotides in 3 frames.

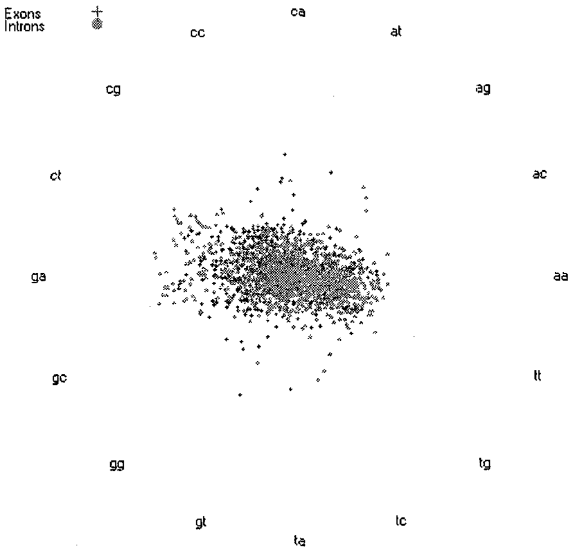

Figure 5 16 dimensions (dinucleotides)

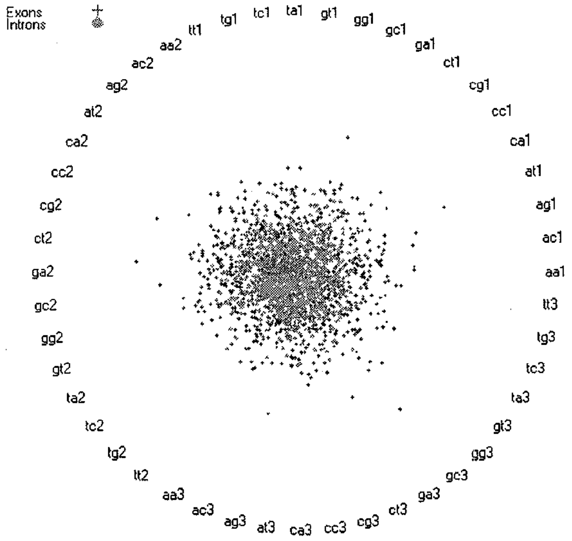

Figure 6 48 dimensions (3 X 16 dinucleotides)

3.2 Parallel Coordinates

As a comparison with Radviz, we used the Parallel Coordinate implementation in Xmdv [16]. In Figure 7 Exons are highlighted in red and Introns are in green. Notice some spreading of Exons is still apparent. Also notice the symmetry line can still be seen between c2 and g2 (ignoring the last dimension, used for brushing between exons and introns).

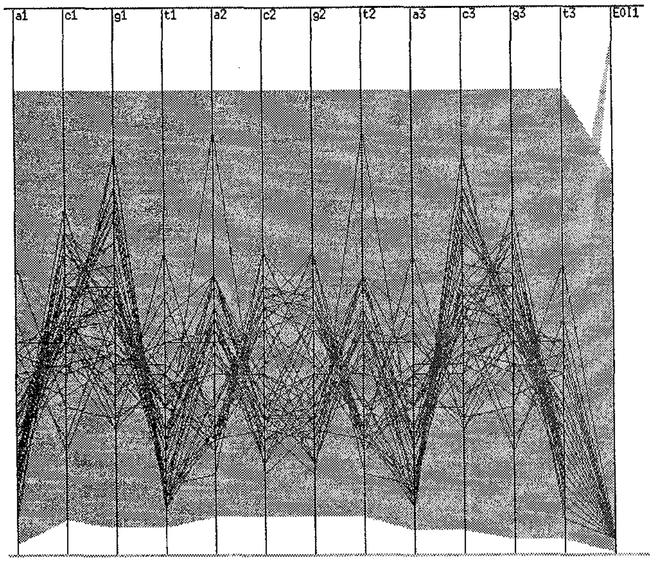

Figure 7 Parallel Coordinates (100 pts)

3.3 Sammon Plots

Another interesting visualization is a Sammon plot [14]. We used Tooldiag to produce Figure 8.

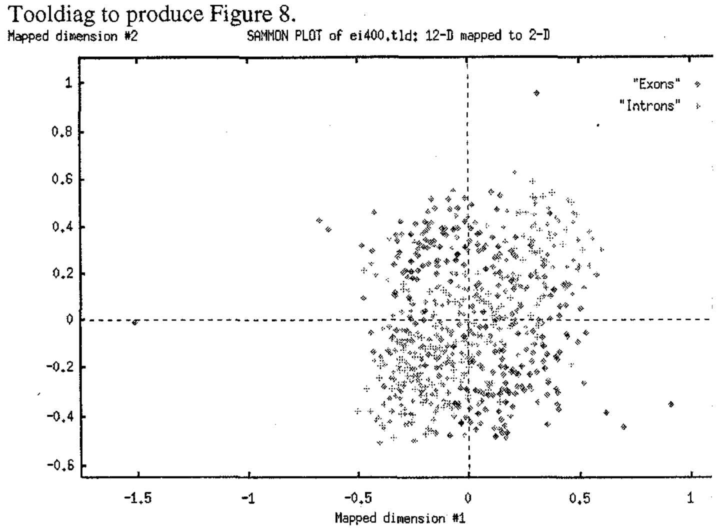

Figure 8 Sammon Plot (600 pts)

The Sammon plot reduces dimensions by trying to preserve the distances between data points. Notice there is some spreading of exons but that the exact symmetry is lost.

Both Radviz and Parallel Coordinates let one see how individual dimensions affect the data. When using Sammon plots or other Multidimensional Scaling techniques like Kohonen nets, it is difficult to preserve any of the original dimensional information.

4 Further Analysis

The previous visualizations helped us cleanse the Fickett datasets by eliminating the redundancies. Additionally the outside/inside pattern of exons and introns from the visualizations also suggested a highly non-linear discriminating function was needed. This would imply that more hidden nodes were needed in the Neural Net. The table below shows NN accuracy results from using 3000 exons and 3000 introns for training. The test data shown below (a and b are similar test sets) included over 7000 exons and 34000 introns from two of the Fickett test sets.

	5 hidden nodes	20 hidden nodes	
Exons - a	80.27% (3076)	80.27% (3076)	
Exons - b	76.95% (4178)	76.59% (4178)	
Introns - a	86.80% (17000)	87.64% (17000)	
Introns - b	86.58% (17000)	87.78% (17000)	
avg.	82.65 %	83.07%	

From the data, it is clear that increasing the hidden node numbers from 5 to 20, as the visualization suggested, resulted in more accurate prediction. These are promising results, showing that input coding, (the 12 extracted values), number of NN nodes, and Visualization are very important for successful Data Mining.

5 Conclusion & Future directions

This case study demonstrates the need for integrating multi-dimensional visualization tools with data mining tools. Many packages provide standard scatter plots, and some have 3-d plots. However, reducing data to 2 or 3 dimensions from many is a difficult task (which dimensions to select) and one which always produces the feeling that something is missing (is the other dimension more important ?). Analytic tools do help. Neural nets, classifying and clustering algorithms, are clearly powerful but they still need to be guided by human insight. However, when the only output of analytic results presented is a few numbers such as "82% accuracy", or some 500 line rule of nested "if statements", the user is left stranded. Does the user understand the rules? Can the user believe the accuracy?

Thus, there is a need to integrate the analytic with the visual [5]. Such integration with intelligent visualizations which automatically map data dimensions to the "best" display parameters such as color, texture, or the coordinate systems will prove attractive [e.g., 13].

Acknowledgments

This work is funded in part by a grant from Pfizer to the Center for Intelligent Biomaterials and the Institute for Visualization and Perception Research at the University of Massachusetts Lowell and Boston University's Center for Polymer Studies and the Department of Physics; and in part by a grant from the National Institute for Standards and
Technology.

References

[1] M. Ankerst, D. A. Keim, H. P. Kriegel Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets, *IEEE Visualization'96 Proceedings*, Hot Topic, San Francisco, CA, 1996.

[2] *Clementine.* http://www.isl.co.uk/clem.html

[3] R Erbacher and G. Grinstein, Issues in the development of 3D Icons, *Proceedings of the Fifth Eurographics Workshop on Visualization in Scientific Computing*, Springer-Verlag Publishers, pp109-131, 1994.

[4] J.W. Fickett and Chang-Shung Tung, Nucl. Acids Res. 20 (1992) 6441

[5] G. Grinstein (1996) Harnessing the Human in Knowledge Discovery , Proceedings of the second International Conference on Knowledge Discovery and Data Mining, August 1996, Portland, Simoudis, Han, and Fayyad (eds), pp384-385.

[6] I. Grosse, K. Marx, S. Buldyrev, G. Grinstein, H. Herzel, P. Hoffman, A. Li, C. Meneses, and H.E. Stanley. Data Mining of Large Gene Datasets Using the Mutual Information Function. to appear in *Journal of Biomolecular Structure and Dynamics.*

[7] I. Grosse, H. Herzel, S. Buldyrev, and H.E Stanley, Mutual information of coding and noncoding DNA. To appear in *Nature.*

[8] P. Hoffman. *Radviz.* http://www.cs.uml.edu/~phoffman/viz

[9] K.A. Olsen, R.R. Korfhage, K.M. Sochats, M.B. Spring and J.G. Williams,Visualisation of a Document Collection: The VIBE System, Information Processing and Management, Vol. 29, No. 1, pp. 69-81, Pergamon Press Ltd, 1993.

[10] J R. Quinlan. Induction of deciscion trees. In Machine learning, volume 1, pages 81-106. Kluwer Academic Publishers, 1986.

[11] J.R. Quinlan, C4.5: *Programs for Machine Learning*, Morgan Kaufmann, 1993.

[12] T.W. Rauber. *Tooldiag.* Universidade de Lisboa, Dept. of Electrical Engineering. http://www.uninova.pt/~tr/home/tooldiag.html

[13] S.F. Roth. The SAGE Project.
http://www.cs.cmu.edu/Web/Groups/sage/sage.html

[14] J. W. Sammon, Jr. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, C-18(5):401-409, May 1969.

[15] (SNNS)http://www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html

[16] M. Ward, A. Martin, High Dimensional Brushing for Interactive Exploration of Multivariate Data. *Visualization'95*, Atlanta, GA , 1995

[17] A. Zell, G. Mamier, M. Vogt, N. Mache A. Der Stuttgarter Neuronale Netze Simulator, in G. Dorffner, K. Möller, G. Paaß, S. Vogel (Hrsg.): *Konnektionismus und Neuronale Netze*, GMD-Studien Nr. 272, Okt. 1995, pp. 335-350.