ABSTRACT

COX, DAVID N. Finding Patterns in DNA Sequences through Visualization with Symbolic Scatter Plots. (Under the direction of Dr. Alan L. Tharp).

Visualization is frequently mentioned as a technique for analyzing large amounts of data. It has been widely anticipated for many years that visualization would become a major tool for the analysis of rapidly growing genomic databases. However, beyond the dot plot which was introduced in 1981 there have been few successful attempts at visualizing this data.

In this thesis a new technique for visualizing DNA sequences, the *symbolic scatter plot*, is introduced. It is shown how the symbolic scatter plot addresses the problems of 1) finding complex patterns in DNA sequences and 2) the comparison of sequences. Second, the symbolic scatter plot is analyzed in terms of human visual perception – particularly in terms of Gestalt theory and pre-attentive visual processing. Third, examples of how specific pre-attentive visual cues can be manipulated or added to find motifs and visualize information content (i.e. entropy) are presented. Fourth, the practicality of symbolic scatter plots is demonstrated by using them to visualize and compare the human and chimpanzee genes responsible for Huntington's disease.

Finding Patterns in DNA Sequences through Visualization with Symbolic Scatter Plots

by
David N Cox

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2010

APPROVED BY:

| | |
|---|---|
| _____ | _____ |
| Dr. Alan L. Tharp<br>Committee Chair | Dr. Donald L. Bitzer |
| _____ | _____ |
| Dr. Christopher G. Healey | Dr. Steffen Heber |

# DEDICATION

To my parents, Charles and Barbara Cox

To my wife and the love of my life, Helen

# BIOGRAPHY

David Cox is a native of Pennsylvania. He attended the Pennsylvania State University where he received a B.S. in Biochemistry and the Rochester Institute of Technology where he received a M.S. in Computer Science.

David worked first as a biochemistry laboratory technician and then as a programmer/analyst at the University of Rochester. He also worked for several years as a software engineer developing software for a line of blood analyzers at the Eastman Kodak Company and as a software engineer developing middleware at Xerox Corp. David currently works as a Principle Scientist in the U.S. Corporate Research Center for ABB Corporation.

David is married to his wife, Helen and resides in Raleigh, NC.

# ACKNOWLEDGEMENTS

To Professor Alan Tharp, words cannot express my gratitude nearly enough. Time flies and I have lost count of all the times we have met and exchanged email. Not only have you offered your guidance, you have been patient and kind and are a true friend. As I go forward from here, you will always be in my thoughts. In whatever I do in the future, I hope I can do it with the same grace and dignity that I have found in you. I came to learn how to conduct research and I have found so much more.

Thank you to Dr. Steffen Heber and Dr. Christopher Healey. Both taught me – one in computational methods in molecular biology and one in computer graphics. I am grateful for our conversations and discussions. Your ideas, suggestions, and encouragement were invaluable to me. To Professor Donald Bitzer, we ran into each other often. You always had warm words of advice and encouragement. Thank you.

To my Dad, Charles Cox, you instilled in me a tremendous respect for education. From the days of my first chemistry set when we had to throw open the windows and doors in the dead of winter after we concocted too much sulfur dioxide until today, you always made learning and life fun. Thank you, Dad, for helping me make it this far.

To my Mom, Barbara Cox, who is not with me today, I love you, Mom and wish so much that you could be here. You always encouraged me to do my best.

Finally, to my wife Helen, I don't know what I would do without you. You stood by me when I worked on my master's degree and you have unwaveringly supported me

again. I remember the words you gave me more than twenty years ago, “You don’t know what you can’t do until you try.” With you, Helen, I know that I will always try.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# 1 Introduction

Deciphering DNA is an important and open research goal. Now that we have the sequences for whole genomes, the goal is to understand what the sequences represent. This goal exists partly to satisfy human curiosity – most people want to understand the nature of life. But this goal also exists for practical reasons. Many if not all diseases are genetically based. These include cancer, heart disease, and diseases that we know have specific genetic causes such as Downs syndrome and Huntington's disease. Developing treatments and cures will require understanding a disease at the genetic level. Understanding how different people respond to medications will require understanding these differences at the genetic level. Developing accurate diagnostic tests also requires understanding diseases at the genetic level.

Deciphering DNA sequences requires understanding how DNA is transcribed and then translated into proteins. It requires understanding how some proteins bind to DNA to enhance or inhibit transcription as part of a molecular feedback loop. It requires understanding how DNA folds into various structures which in turn enhance or inhibit transcription. Lastly, it requires understanding how DNA is chemically modified by processes such as methylation to turn transcription and translation on or off.

The predominant tool for analyzing DNA sequences is statistics and the predominant approach is comparative genomics to find statistically significant similarities and differences between DNA sequences. Other statistical approaches include looking for patterns (or motifs) in DNA using Bayesian or hidden Markov models and looking for

repetitive patterns. Aside from comparing sequences and searching for patterns, statistics is employed to analyze how genes are expressed in different cell types and how organisms are related to each other genetically.

A major motivation for using statistical approaches and algorithms that automatically find similarities between sequences is the vast amount of data that must be analyzed. It is well known that the amount of genomic data has been growing exponentially and that humans can process only very small amounts of it manually in a short period of time (Tao, Liu, Friedman, & Lussier, 2004). Computers can analyze vast amounts of data quickly and perform statistical analyses and string comparisons efficiently and accurately.

It has been recognized for some time that data visualization is an effective strategy for analyzing large amounts of data. Healey (Healey, Effective visualization of large multidimensional datasets, 1996) wrote, "...the desire for computer-based data visualization arose from the need to analyze larger and more complex datasets. Scientific visualization has grown rapidly in recent years as a direct result of the overwhelming amount of data being generated. New visualization techniques need to be developed that address this 'fire hose of information' if users hope to analyze even a small portion of their data repositories."

(Tao, Liu, Friedman, & Lussier, 2004) propose that the high bandwidth of human vision could also be exploited for processing genomic data. However, it has not been clear what visualization techniques are needed or how useful they can be. As early as 1981 dot

plots were introduced to compare two sequences (Maizel & Lenk, 1981). Nevertheless, for many years there have been few other developments in visualizing DNA and genomic data.

At the IEEE Visualization conference in 2001 several researchers participated in a panel discussion entitled, "Visualization for Bio- and Chem-Informatics: are you being served?" The panel discussion was introduced with a simple but challenging question: "To meet the computing challenges (of analyzing biological data), visualization plays a key role. Or does it?" Panelist Georges Grinstein stated, "In the past, I have argued that visualization must be used at every stage of the knowledge discovery process. I now argue that visualization today is not just one of the main keys to knowledge discovery but that it is still the most underutilized component of that discovery process and that so much more can be done to support that process." The panelists expressed frustrations that visualization hasn't taken a larger role particularly in the area of bioinformatics.

Indeed, these frustrations have been largely born out. In the decade since this conference very little progress has been made in using visualization for DNA or protein sequence analysis. In the mid 1980's the National Science Foundation "convened a panel to report on the potential of visualization as a new technology" (Johnson, Moorhead, Munzner, Pfister, Rheingans, & Yoo, 2006). Twenty years later in 2004 the NSF partnered with the National Institutes of Health to convene the Visualization Research Challenges Executive Committee to write an updated report on visualization. That report (Johnson, Moorhead, Munzner, Pfister, Rheingans, & Yoo, 2006) was issued in 2006 with just two paragraphs devoted to bioinformatics visualization. The authors discussed what they hoped

will be accomplished with visualization without mention of a single previous accomplishment.

At the Workshop on Ultrascale Visualization in November 2008, the primary visualization technique noted for comparing DNA sequences was still the dot plot (Samatova, Breimyer, Hendrix, Schmidt, & Rhyne, 2008). Compared to statistical methods, visualization remains a distant second when analyzing DNA whether for comparing sequences or finding patterns.

The research that is the object of this dissertation investigates how to fill this gap. This dissertation introduces a novel graphical representation of DNA sequences referred to as *symbolic scatter plots*. Symbolic scatter plots are evaluated in terms of several bioinformatics tasks: comparison of sequences, identification of repeats, identification of motifs, and the display of information content (i.e. entropy) for DNA sequences. The potential utility of symbolic scatter plots is demonstrated by a visual comparison and the discovery of potentially significant differences between the huntingtin genes for humans and chimpanzees. Lastly, this research evaluates symbolic scatter plots in terms of what is known and generally accepted about human visual perception. No other technique for the visual analysis of biological sequences – dot plots, DNA walks, spectrograms, sequence logos, etc. – has ever been described or evaluated in terms of how the human visual system responds to visual stimuli. It is perhaps this latter result that is the most significant of this research. Only by considering how the human visual system works will we be able to

exploit its power and develop visualization techniques that 1) rival their statistical cousins and 2) open the door of data analysis to non-specialists.

## 1.1 Thesis Organization

Chapter two provides background pattern discovery, pattern perception, and how humans process visual information. It reviews the "central dogma of molecular biology" and traditional methods of sequence analysis. Chapter two concludes with a review of techniques developed by other researchers for visualizing DNA.

Chapter three introduces a technique for visualizing the sieve of Eratosthenes. This chapter is included because this technique was the inspiration for symbolic scatter plots. Furthermore, the technique was a significant result in its own right by providing a novel visualization of a two thousand year old mathematical approach for finding prime numbers.

Chapter four describes symbolic scatter plots and how to produce them. Various aspects of the plots are described. How patterns are perceived in the scatter plots is discussed in terms of Gestalt principles and visual processing.

Chapter five explores visualizing copies of DNA in larger sequences. The copies can be either exact or inexact. Results are compared to those generated by the predominant statistical technique, Tandem Repeats Finder.

Chapter six presents an application of using symbolic scatter plots – visualization of the huntingtin gene which is responsible for Huntington's disease in humans.

Chapter seven explores in more detail why we see patterns in symbolic scatter plots and how various visual cues can be manipulated to reveal specific patterns.

Chapter eight presents conclusions and chapter nine discusses thoughts for future research.

# 2 Background

This research crosses several boundaries including computer science, biology, perception, pattern discovery, and information visualization. This chapter provides background on these areas. Sections 2.1, 2.2, and 2.3 discuss pattern discovery, Gestalt laws of pattern perception, and pre-attentive visual processing. The section on pattern discovery defines what is meant by a pattern. The section on Gestalt laws explains how we segment images into discernable objects or groups of objects. The section on pre-attentive processing explains why certain features grab our attention. The material in these sections helps to explain why we see patterns in graphical representations of DNA. It also suggests how we can change our graphical representations to improve our ability to see patterns in DNA.

Following these sections is a discussion about the "central dogma of molecular biology," a term introduced by Francis Crick in 1958 and formally stated in an article in the journal Nature (Crick, 1970). The role of DNA sequence analysis is described and traditional algorithms for analyzing DNA are explained.

Lastly, there has also been some research into representing DNA graphically and the various techniques are reviewed and critiqued.

## 2.1 Pattern Discovery

Hardy once wrote, "A mathematician, like a painter or poet, is a maker of patterns" (Hardy, 1940). In his book, *Mathematics as a Science of Patterns*, Resnick states, "…in mathematics the primary subject-matter is not the individual mathematical objects but rather the structures in which they are arranged" (Resnick, 1997). Patterns are important but what exactly are they?

The first set of results returned by Google when searching for the word *pattern* is a set of images. The first few of these are shown here:

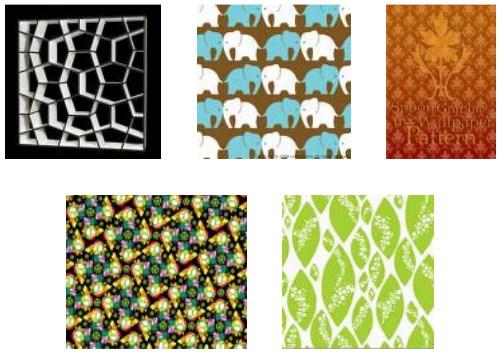


Figure 2.1: Examples illustrating how repetition influences our notion what is or isn't a pattern. (Google image results of patterns, 2009).

In his book, "Pattern Discovery in Bioinformatics," Parida states that a pattern is a non-unique phenomenon observed in a set of input data (Parida, 2008). In this definition repetition is central and repetition is certainly seen in the examples returned by Google. Even common usage of the word, pattern, suggests that repetition is central to its definition. For example, "I think I see a pattern here," is commonly exclaimed after examining a series of events implying that something has been seen repeatedly.

Patterns, however, involve more than repetition. In his book, "Information Visualization: Perception for Design," Colin Ware refers to a pattern as something that can be visualized as a coherent whole (Ware, 2004). There are two images in Figure 2.2. Both contain about two thousand randomly placed points. However, certain points on the right were painted white to render them invisible. The repetition in these images kept to a minimum. Even so, the image on the right contains what most people would perceive to be a singular white circle.

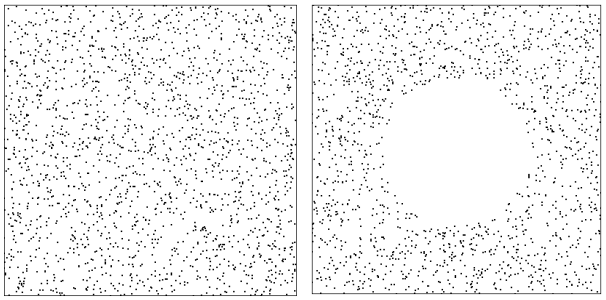


Figure 2.2: Illustration of a pattern with minimal repetition.

Ware asks, "What does it take for us to see a group? How can 2D space be divided into perceptually distinct regions? Under what conditions are two patterns recognized as similar? What constitutes a visual connection between objects?" To answer these questions, Ware refers to the Gestalt laws of pattern perception (indeed, the German word gestalt means "pattern").

## 2.2 Gestalt Laws

The Gestalt school of thought was a group of German psychologists (Max Westheimer, Kurt Koffka, and Wolfgang Kohler) founded in 1912. Although founded

nearly a hundred years ago, the Gestalt laws of pattern perception remain valid today. The laws describe several visual cues that we use to organize what we see. These visual cues are proximity, similarity, connectedness, continuity, symmetry, closure, relative size, and figure and ground.

### 2.2.1 Proximity

Our visual system forms groups for objects that are near each other. This happens automatically during an early stage of visual processing. In the following two images, the distances between the dots differ slightly. In the image on the left, the rows are closer together than the columns causing most people to see the dots organized into columns. In the right image, the columns are closer together than the rows causing most people to see the dots organized into rows.

Figure 2.3: Effect of proximity on object grouping. On the left the dots are closer vertically than horizontally and on the right the reverse is true. (Ware, 2004).

### 2.2.2 Similarity

Similar visual elements tend to be grouped together. The objects in the left of Figure 2.4 are rendered as circles and squares. The distances between the objects are the same thus ruling out proximity as a grouping factor. Most people group the objects by their shapes and thus see rows of similar objects. Objects of similar color and texture also tend

to be grouped together. The objects in the right of Figure 2.4 have the same shape but the color is varied. Again, the shapes are organized into rows based on their similar colors.

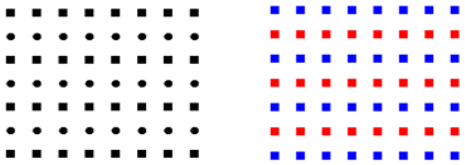


Figure 2.4: Effect of similarity on grouping. On the left objects are grouped by shape while on the right they are grouped by color. (Ware, 2004).

### 2.2.3 Connectedness

According to Ware, connectedness was introduced by Palmer and Rock as a fundamental Gestalt organizing principle. Here the objects connected by a line tend to form groups.

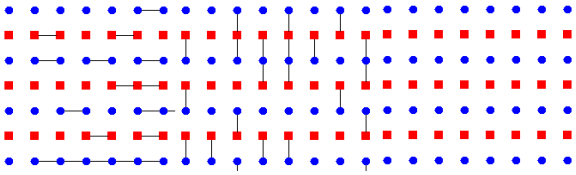


Figure 2.5: We tend to group objects that are somehow connected. Here objects are connected with horizontal and vertical lines. (Ware, 2004)

### 2.2.4 Continuity, Symmetry, Closure, and Relative Size

In Figure 2.6 people tend to see the left most figure as two crossed lines rather than four individual lines that meet at one point. This illustrates the principle of **continuity** where we construct objects out of visual cues that are smooth and continuous.

We tend to perceive two **symmetrical** objects as a single object. The two wavy lines immediately to the right of the cross tend to appear as two distinct lines. However, if one of the lines is flipped, then the resulting **symmetry** suggests that the lines belong together as a single object.

**Closed** contours tend to be seen as a single object. We also tend to see contours as closed even though they have gaps. The blue circle in Figure 2.6 is perceived to be whole and partially hidden by the green rectangle rather than as three quarters of a pie. **Relatively** smaller objects tend to be perceived as objects in preference over larger objects. At the far right of Figure 2.6 most people see blue objects over a red background rather than red objects on a blue background.

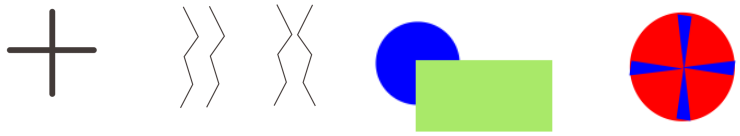


Figure 2.6: Examples of continuity, symmetry, closure, and relative size. (Ware, 2004)

### 2.2.5 Figure and Ground

The following is a classic example where the image is sometimes perceived to be a single vase and sometimes perceived to be two faces. What is perceived depends on how we respond to the image. Sometimes we see parts of an image as being in the foreground (i.e. figure) and sometimes as being in the background (i.e. ground).

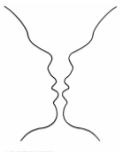


Figure 2.7: A classic example of figure and ground. Is this a vase or two faces (Ware, 2004) ?

## 2.3 Pre-attentive Processing

The Gestalt Laws help to explain how we perceive patterns. Pre-attentive processing helps to explain why some patterns "pop out" and are distinguishable from other patterns (Healey, Perception in Visualization, 2009). Pre-attentive processing occurs early in visual perception and determines what visual features grab our attention.

An example is color. The red circle below is immediately noticed by our visual system. Another example is a counting exercise. On the right of Figure 2.8 are several rows of numbers. Some of the 3's are colored red. When given the task of counting the 3's, subjects were able to count the red threes in constant time regardless of the number of other distracting numbers. When the 3's were the same color as the distracters, the time to count them increased linearly as the number of distracters increased.

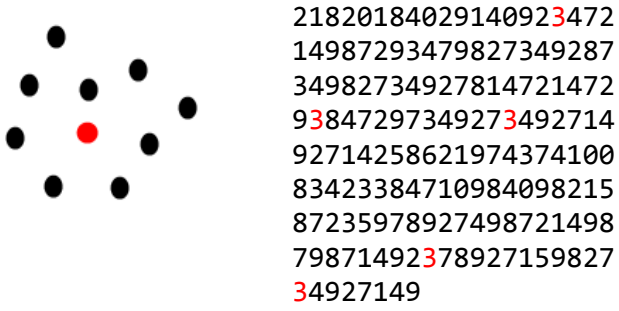



Figure 2.8: Color is pre-attentively processed and allows us to locate objects quickly. (Healey, Perception in Visualization, 2009) (Ware, 2004)

Using such experiments, it has been determined that features that are pre-attentively processed fall into several categories: form, color, motion, and spatial position. According to Healey we pre-attentively process features such as lines (orientation, length, width, and co-linearity), size, curvature, color (hue and intensity), motion, and spatial position as well as several others.

Examples cited by Healey are presented in

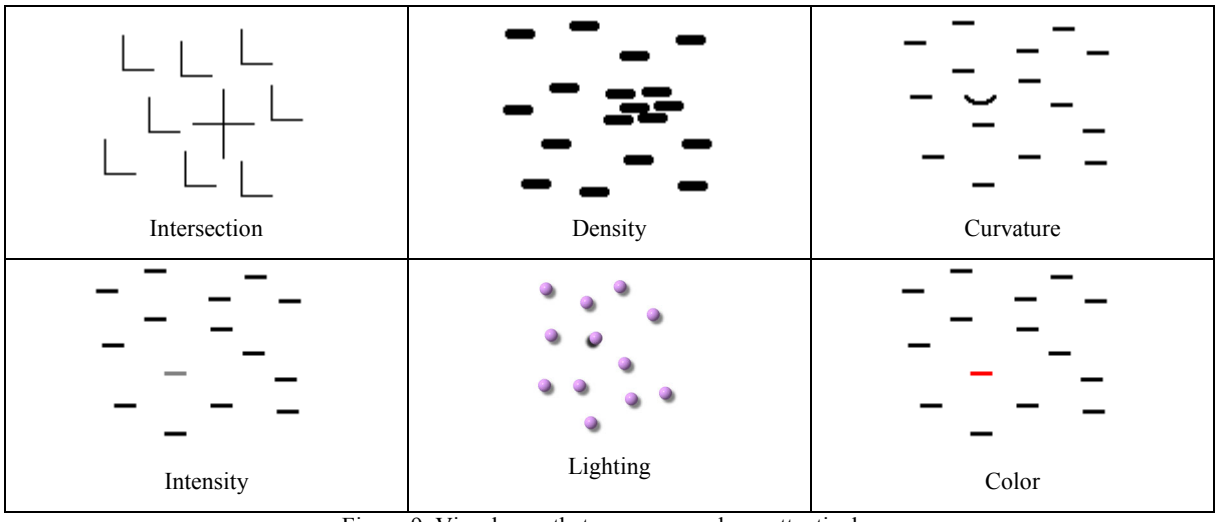



Figure 9: Visual cues that are processed pre-attentively.
(Healey, Perception in Visualization, 2009)

Additional examples cited by Ware (Ware, 2004) are presented in Figure 2.10 and Table 2.1

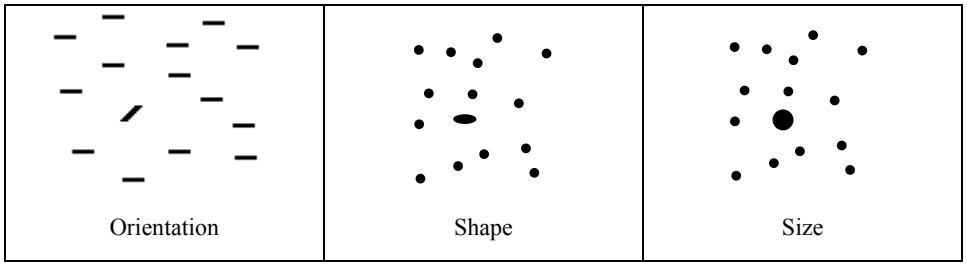



Figure 2.10: Examples of visual cues that are pre-attentively processed. (Ware, 2004)

Table 2.1: Visual cues that are pre-processed by the human visual system (Ware, 2004).

| | |
|---|---|
| • *Form*<br>  ○ Line orientation<br>  ○ Line length<br>  ○ Line width<br>  ○ Line collinearity<br>  ○ Size<br>  ○ Curvature<br>  ○ Spatial grouping<br>  ○ Blur<br>  ○ Added marks<br>  ○ Numerosity | • *Color*<br>  ○ Hue<br>  ○ Intensity<br>• *Motion*<br>  ○ Flicker<br>  ○ Direction of motion<br>• *Spatial Position*<br>  ○ 2D position<br>  ○ Stereoscopic depth<br>  ○ Convex/concave shape from shading |

Pre-attentive processing is also fast and accurate. Pre-attentive tasks such as searching for particular features happen in 200 ms or less and this time is constant regardless of the number of elements displayed (Healey, Perceptual techniques for scientific visualization, 1999). A typical computer display contains about one million pixels. When properly displayed for pre-attentive processing, a single computer monitor could easily represent 100,000 nucleotides of a DNA sequence while using only ten percent of the available display area.

Considering that a complex protein assembly might consist of a half dozen peptides each ranging in size from several hundred to several thousand nucleotides, then it should be possible to examine a visual representation of such a complex in less than a minute. A single strand of the human Y-chromosome contains about 58 million nucleotides. Allowing about one second per screen, a person could scan the entire chromosome in a little more than nine minutes. Combined with multiple observers and multiple/larger displays, visual

examination of DNA sequences could prove extremely fast and accurate when compared to other methods.

## 2.4 The Central Dogma of Molecular Biology

What is life? What distinguishes a living thing from a non-living thing? After many years of observation and experimentation we have reached several conclusions. Living things consist of microscopic cells. Certain species such as bacteria consist of single cells while others such as humans consist of billions of cells. Cells differ from each other and this is particularly evident in multi-cellular organisms. Blood cells differ from nerve cells which differ from muscle cells which differ from skin cells.

Cells divide to produce more cells. Humans arise from many cell divisions beginning with a single cell – the fertilized egg. When the egg divides it produces identical copies of itself. Yet, with time those copies change allowing the cells to differentiate into many different cell types. We now know that each new cell receives an exact copy of its parent's DNA. We also know that DNA controls the production of proteins and it is these proteins that give each cell its particular characteristics.

We know in considerable detail how proteins arise from DNA. DNA is a polymer made from four basic molecules: adenine, cytosine, guanine, and thymine. For short we refer to these "nucleotides" as A, C, G, and T. A single DNA molecule consists of these nucleotides strung together like a string of beads. A small example is AAGTAGGCCTACT. A typical DNA molecule will consist of thousands to millions of nucleotides. Each human cell contains 24 pairs of DNA molecules called chromosomes that together account for some three billion nucleotides.

DNA serves as a template for creating proteins. First, small sections of DNA are copied to produce another molecule called RNA. RNA consists of the same nucleotides as DNA except that uracil (U) is used in place of thymine (T). Typically, a few hundred to several thousand DNA nucleotides are *transcribed* to produce a single RNA molecule. An A in DNA will result in a U in RNA. C will be transcribed to G, G to C, and T to A.

Once created, the RNA molecule serves as a template to produce a protein – another polymer consisting of a chain or sequence of amino acids. In this case triplets of nucleotides in RNA are *translated* into amino acids. With four nucleotides, there are 64 possible triplets. However, there are only 20 amino acids. Consequently, there is some redundancy where more than one triplet will be translated to the same amino acid.

The interplay between DNA, RNA and protein is complex. Most commonly DNA serves as a template for creating RNA and RNA serves as a template for creating protein. In some cases RNA serves as a template for creating DNA. When cells divide, DNA serves as a template for exact copies of DNA to be given to the newly created daughter cells. We have also found that cells can copy RNA to produce other RNA. In contrast, we never see protein serving as a template for other proteins. Nor do we see protein serving as a template for RNA or DNA. Francis Crick summarized these relationships in a famous paper in 1970 with the following figure:

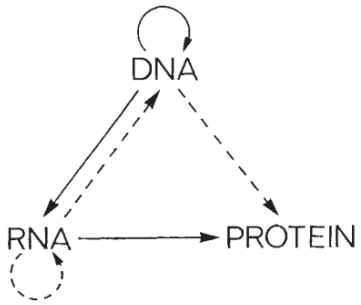



Figure 2.11: The central dogma of molecular biology. (Crick, 1970)

The solid arrows refer to the most common transfers while the dotted arrows refer to specialized transfers. This figure is referred to as the *central dogma of molecular biology* and was published in Crick's paper of the same name, "Central Dogma of Molecular Biology" (Crick, 1970).

The interplay between DNA, RNA, and protein can be viewed as an exchange of information. Thus, the central dogma can be viewed as a flow of information with messages being encoded as DNA, RNA, or protein. The way that RNA is created from DNA or the way that protein is created from RNA can be viewed as a complex communication system. Understanding the details of this system is one of the main goals of molecular biologists.

For example, not all human DNA is transcribed into RNA and not all RNA is translated into protein. Less than three percent of human DNA ultimately is transcribed into RNA and then translated into sequences of amino acids. When human DNA is transcribed into RNA, the resulting RNA consists of regions known as exons interspersed by regions known as introns. Through a process called splicing, introns are removed and the

exons are spliced together to form a final RNA molecule that is then translated into protein. Different exons can be spliced together to form different proteins.

Another complexity is that some proteins bind to DNA affecting how DNA is subsequently transcribed into RNA. These proteins can either inhibit or enhance DNA transcription. Through this complex feedback loop certain proteins can affect the creation of other proteins.

We know that DNA twists and folds and that the final shape of DNA can also affect transcription. Folding allows certain regions of DNA to be exposed to the machinery that transcribes DNA to RNA. Folding also hides certain regions of DNA from this machinery preventing transcription. Folding seems to depend on the sequence of nucleotides with different sequences resulting in different shapes. However, the binding of proteins such as histones also plays a role because they can hide or expose portions of DNA.

We are making progress at understanding the details. Through some very clever chemistry we now know all of the nucleotides in our DNA. However, we do not know all of the more intricate details. We cannot accurately predict which DNA is transcribed into RNA. We can't predict which regions are exons and which are introns. We can't predict where certain proteins will bind to DNA or how that binding will affect transcription. We also cannot accurately predict how DNA will fold or under what conditions it will fold or predict how this folding affects transcription. Yet, being able to make these predictions is key to understanding how cells differentiate into different cell types and aggregate to form

tissues and organs. It is also key to understanding why diseases such as cancer arise and key to understanding how to control and cure them.

## 2.5 DNA Sequence Analysis

Now that we know the sequences of nucleotides in the human chromosomes as well as the sequences of nucleotides for the chromosomes of other species, we can begin to analyze the data. The goal of DNA sequence analysis is to help unravel the details of the central dogma of molecular biology – to determine exactly how information flows between DNA, RNA, and proteins. There are several ways to do this analysis and what follows is a review of several common approaches. This brief review is based on material in the texts by (Mount, 2004), (Pevsner, 2003), (Durbin, Eddy, Krogh, & Mitchison, 1998) and (Jones & Pevzner, 2004).

Perhaps the most obvious way to analyze DNA is to compare sequences to find similarities and differences. These similarities and differences help us to infer evolutionary relationships between species. They also help us to develop rules to locate genes, exons, and introns. We can compare sequences two at a time or we can compare several at a time. Comparing sequences involves creating a pairwise sequence alignment or a multiple sequence alignment depending on the number of sequences involved.

According to (Jones & Pevzner, 2004), an *alignment* of two sequences v and w is a two-row matrix such that the first row contains the characters of v in order and the second row contains the characters of w in order. Each character occupies a column in each row. Spaces may be interspersed throughout the matrix. However, spaces cannot occupy both rows of the same column. Columns that contain the same characters are called matches while columns that contain different characters are mismatches. A column containing a

space in the first row is called an insertion while a column containing a space in the second row is called a deletion. Generally speaking, an optimal alignment maximizes the number of matches while minimizing the number of mismatches, insertions, and deletions. Multiple sequence alignments extend this concept to a matrix with multiple rows for multiple sequences.

Similar sequences are believed to share an evolutionary relationship. Over many generations DNA sequences will have diverged due to nucleotide insertions, deletions, and substitutions. These changes affect the flow of information between DNA, RNA, and protein resulting in both subtle and not so subtle changes to cells, tissues, and organs.

Finding both closely related (well aligned) and distantly related (poorly aligned) sequences is important because it is thought that related (and, therefore, well aligned) sequences perform common functions. Knowing how a sequence behaves in one species can provide clues about how a related sequence will behave in another species. The ability to make these comparisons is particularly helpful for studies related to humans. When it is unethical or impractical to perform experiments on human subjects, we can gain insights from animal studies and can extrapolate those insights when human DNA is found to align well with that of another species.

Comparing and contrasting sequences also helps us to develop statistical models for sequences. If we know that several sequences represent genes, then we can look for features that appear frequently in them and infrequently elsewhere. We can convert the frequencies of these features into probabilities and use them with varying degrees of

success to distinguish between protein coding and non-protein coding regions. In other cases we can develop more complex statistical models such as hidden Markov models to capture dependencies between regions of DNA. For example, such models attempt to distinguish between exons and introns or attempt to identify promoter regions where proteins bind to enhance or inhibit transcription.

Another method for analyzing DNA looks for either underrepresented or overrepresented strings of nucleotides in a larger DNA sequence or across several sequences. Searching for these strings is complicated because they might not be exactly the same due to mutations. Nevertheless, the idea is that such strings represent biologically important sequences – possibly regulatory motifs that serve as protein binding sites to either enhance or inhibit transcription.

Many genetic diseases are associated with abnormally repetitive sequences. For example, Huntington's disease is associated with an excessive number of CAG's in the huntingtin gene. Methods that look for repetitive sequences in DNA represent yet another way to analyze DNA.

A common theme with these approaches is that they rely heavily on statistical analysis and string processing algorithms. However, these need not be the only approaches. The research of this thesis investigates the use of computer graphics and data visualization as tools to analyze DNA. While others have attempted to use graphics and visualization, those attempts are not nearly as extensive as the statistical approaches. The next section is a review of these attempts.

## 2.6 Visualization Techniques: A Review

There have been few attempts to visualize patterns in DNA over the past 30 years. These approaches include dot plots, chaos game representations, DNA walks, color coding, color merging, repeat graphs, pygrams, spectral analysis, and sequence logos. Examples are presented in Figure 2.12.

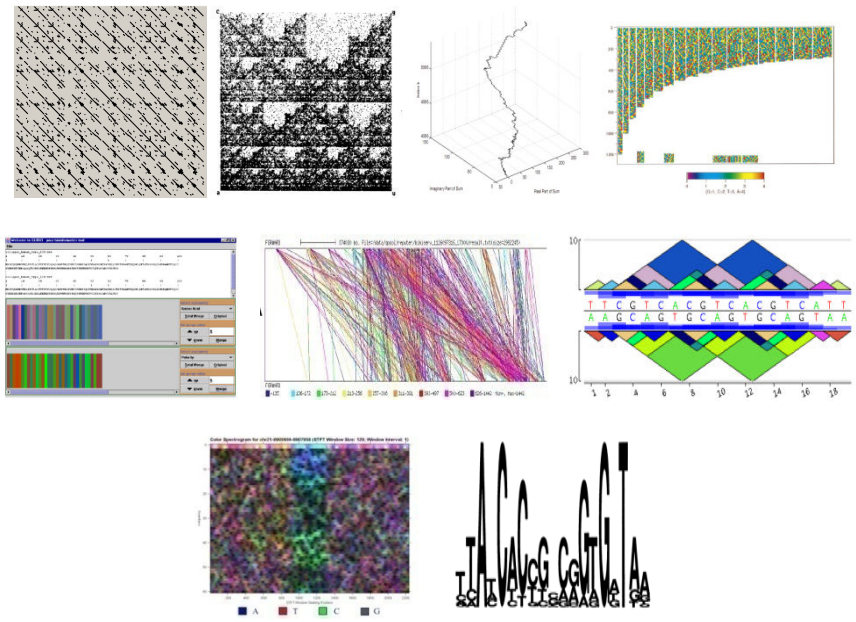


Figure 2.12: Examples of DNA visualization. Top row: dot plots, chaos game representation, DNA walk, color coding; middle row: color merging, repeat graph, pygram; bottom row: spectrogram, sequence logos. (Maizel & Lenk, 1981) (Jeffrey, 1990) (Yoshida, Obata, & Oosawa, 2000) (Berger, Mitra, Carli, & Neri, 2004) (Alston, Johnson, & Robinson, 2003) (Durand, Mahe, Valin, & Nicolas, 2006) (Dimitrova, Cheung, & Zhang, 2006) (Schneider & Stephens, 1990)

Dot plots (Maizel & Lenk, 1981) are rendered using the same visual device as symbolic scatter plots – namely a scatter plot. The x- and y-axes correspond to two sequences in which a point is plotted where the two sequences match. The result is a scatter plot where diagonal lines correspond to regions conserved in the two sequences. Using a

dot plot a sequence may be compared to another sequence or to itself. Because dot plots frequently fill with a great number of points, filtering techniques and heuristics are employed to reduce the number of points to reveal diagonals that indicate regions in common in the two sequences. A limitation of dot plots is that they require the two sequences to have regions in common. If the sequences lack commonality, then the plot consists of points where matches occur at random thus providing no result.

Chaos game representations (Jeffrey, 1990) are also scatter plots. Fractals emerge from an iterative process of placing points in the plot. Construction begins by randomly placing an initial point at the center of a small set of vertices. Subsequent points are plotted by randomly choosing one of the vertices and plotting a point half way between the vertex and the previously plotted point. Using this process and an initial set of three vertices produces a Siepinski triangle. The technique is applied to a DNA sequence by using the nucleotides in the sequence to choose the next vertex. The resulting scatter plot serves as a unique signature for the sequence. As with a dot plot, chaos game representations quickly fill with points that must be filtered to reduce the noise in the image. This is particularly true with large sequences. Chaos game representations have the added problem that it isn't clear how features in the plots correspond to features in the sequences that are of interest to biologists.

Color coding (Yoshida, Obata, & Oosawa, 2000) produces several scatter plots. Each scatter plot has a different number of rows and columns. For example, a sequence of 599 nucleotides is used to produce several scatter plots that are 120 rows x 5 cols, 100 rows

x 6 cols, etc. through 28 rows x 21 cols. Each point corresponds to a nucleotide in the sequence. For example the point at row 5, column 5 corresponds to the 25th nucleotide. Each point is colored in one of four colors. Examining the plots for diagonals reveals the presence of repeating nucleotides. However, the plots are filled with points and these patterns are hard to discern from the background.

Color merging (Alston, Johnson, & Robinson, 2003) represents nucleotides as colored vertical bars. Color merging allows one to zoom in or out of a sequence. When zoomed in, each nucleotide is rendered in its respective color. When zoomed out, colors are merged according to a formula. The technique has been tested to visually differentiate regions of hydrophobicity (i.e. attraction or repulsion to water). It isn't clear how color merging would be used for general pattern finding.

Repeat graphs and pygrams (Durand, Mahe, Valin, & Nicolas, 2006) visualize the structure of repeats. Repeat graphs rely on numerous colored lines connecting the repeats between two sequences. Pygrams rely on colored triangles superimposed on a sequence where the base of a triangle spans a repeat. Multiple triangles of the same area and color reflect the same repeats in the sequence. Each is used after another technique performs repeats analysis. Neither is used alone and consequently each has the same limitations as the algorithm that was used to find the repeats.

Spectral analysis (Dimitrova, Cheung, & Zhang, 2006) converts a sequence of characters to a numeric sequence and then performs a Fourier analysis to create and render a spectrogram. A spectrogram reveals periodicities in a sequence that might be of

biological interest. The spectrogram alone, however, cannot reveal which characters are repeating nor is it clear how the patterns in a spectrogram map to specific patterns in a sequence.

Sequence logos (Schneider & Stephens, 1990) are constructed from consensus sequences. Different sized characters represent the nucleotides with large characters reflecting strong consensus and small characters reflecting weak consensus. Interesting biological patterns are revealed only if the multiple sequences being compared have regions in common.

Of these techniques dot plots and sequence logos have had the greatest acceptance by far. Dot plots are specifically useful for comparing two sequences whereas sequence logos are used extensively with multiple sequence alignments and the frequency profiles that are calculated from them. Both are fairly easy to understand and interpret which likely contributes to their popularity.

None of the techniques, however, have been analyzed in terms of human perception. There are no discussions in the literature about what patterns are perceptible using these techniques nor why. In retrospect we can say that dot plots produce linear patterns (i.e. diagonal lines) when regions of two sequences match. Sequence logos take advantage of several visual cues including size, shape, and color. Unfortunately, without some consideration of the rules of visual perception, there is little or no insight about how to improve these visualizations to reveal additional patterns or to interpret their meaning.

# 3 Inspiration from the Sieve of Eratosthenes

*The material in this chapter appeared as a featured communication of Notices of the American Mathematical Society in May, 2008.* (Cox, Visualizing the sieve of Eratosthenes, 2008)

The integers are the ultimate sequential data. Prior to working with DNA sequences, a technique to visualize the relationship of the integers to their divisors was created. The result was a scatter plot that is a graphical variation of the sieve of Eratosthenes – a technique developed 2,000 years ago for finding prime numbers. This novel visualization of the sieve of Eratosthenes directly led to the visualization of DNA sequences with symbolic scatter plots. This chapter presents result.

Every so often new technology is applied to an age-old problem to produce unexpected results. This article re-examines the sieve of Eratosthenes. The sieve is a one-dimensional device for finding prime numbers. The numbers from 2 to n are written as a single, long sequence. Then the multiples of 2 are crossed out but leaving 2. Then the multiples of 3 are crossed out but leaving 3. And so on until the only numbers remaining are the primes. This paper explores what happens if the procedure is converted from one dimension to two. Rather than a single sequence, a matrix is constructed such that in the first row every column is marked with a dot. In the second row, every other column is marked with a dot. In the third row, every third column is marked with a dot. In general, in the $n^{th}$ row, every $n^{th}$ column is marked with a dot. In this fashion a two-dimensional image is built for all $n^2$ cells. Results of this procedure as generated by computer software are

presented. Despite the simplicity of this method, when enough dots are generated, the resulting image turns out to be stunning. This article demonstrates well that computerized visualization can shed new light on old subjects—even those more than 2,000 years old.

According to (Gullberg, 1997) Eratosthenes lived from 276 to 194 BC. Only fragments of Eratosthenes's original documents have survived. However, a description of his sieve method for finding prime numbers was described in "Introduction to Arithmetic" by Nicomedes written sometime prior to 210 BC(Cojocaru & Murty, 2005).

To use the method, imagine a written sequence of numbers from 2 to *n*. Starting at 2 cross off every other number in the sequence except for 2 itself. When done, repeat for 3 (which will be the next remaining number in the sequence) by crossing off every third number. When done, the next number remaining in the sequence will be 5. Repeat the process for every fifth number in the sequence. Continue with this process until you reach the end of the sequence. At the end of the process the numbers remaining in the sequence will be the primes. Here is an example.

Start with a sequence of integers:

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Keep 2 but eliminate every second number beyond 2:

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Keep 2 and 3 but eliminate every third number beyond 3:

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Keep 2, 3 and 5 but eliminate every fifth number beyond 5:

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

At completing the sieve of Eratosthenes the following numbers remain:

2 3 5 7 11 13 17 19 23 29

These prime numbers have a few interesting properties. They are not divisible by any other numbers except 1 and themselves. All other numbers ultimately are divisible by some subset of the primes. Because of these fundamental properties both the prime numbers and the sieve of Eratosthenes have been studied intensely for 2,000 years. One would think that everything there is to know about the method has long since been discovered. Indeed, there are advanced sieve methods and optimization methods. However, these are significant variations of the original method and do not provide any additional characterization of the original method itself. After 2,000 years what more could be said?

This chapter explores the use of computerized visualization to further characterize the sieve of Eratosthenes. After all, Eratosthenes didn't have a computer and computer graphics and visualization have only been widely available for the past 20 or 30 of those 2,000 years. With a simple extension of the sieve we arrive at novel result. In the next chapter this result is extended to DNA sequences.

## 3.1 The Method

The method extends the sieve of Eratosthenes from a one dimensional sequence to a two dimensional matrix. The method constructs a matrix of dots that can be easily viewed on a computer screen. The method is as follows. In the first row of the matrix every

column contains a dot. This would correspond to crossing off every number in Eratosthenes original sequence. In the original method this is not done. However, in two dimensions it proves useful.

In the second row of the matrix, every other column contains a dot starting with the second column. This corresponds to crossing off every other number in the original sequence. In the third row, every third column contains a dot starting with the third column. Again, this corresponds to crossing off every third number in the original sequence. In general, in the nth row, every $n^{th}$ column contains a dot starting with the nth column.

Aside from extending the one dimensional sequence of length n to a two dimensional matrix of size $n$ x $n$, the process of crossing off numbers (using dots) remains faithful to the original method with two differences. In the first row every number is marked with a dot. In the remaining rows every nth number is marked with a dot including n itself. Consequently, even though 2 is not crossed off in Erathosthenes's original sequence, it is marked with a dot in row 2. The same holds for rows 3, 5, 7, etc. The implication of this is that a prime number corresponds to a column containing two dots – one in the first row (division by 1) and one in the $n^{th}$ row (division by the number itself).

Using this method a computer program was created to generate a matrix containing 1 million columns and the first 1,000 rows. The matrix is easily converted into a binary bitmap for viewing. Because it is not possible to view the entire matrix on a computer screen, a scrolling facility was provided to traverse through the columns and rows.

The computer algorithm builds the image of the matrix with the first row at the top of the screen and the first column at the left. The only reason for doing this is because most computers address pixels beginning with (0,0) at the top left. Other orientations would also work. As will be shown, the resulting image has several interesting features.

## 3.2 Results

Figure 3.1 shows several hundred columns and rows beginning with column 1 on the left. The most striking feature is the set of diagonals. Close inspection of these diagonals reveals a pattern. The main diagonal has a slope of 1 and consists of contiguous dots. The adjacent diagonal has a slope of 2 and has a dot in every other column. The third diagonal has a slope of 3 and has a dot in every third column. And so on. Remarkably, these diagonals are *constructed in exactly the same manner* as the rows from which the image was constructed.

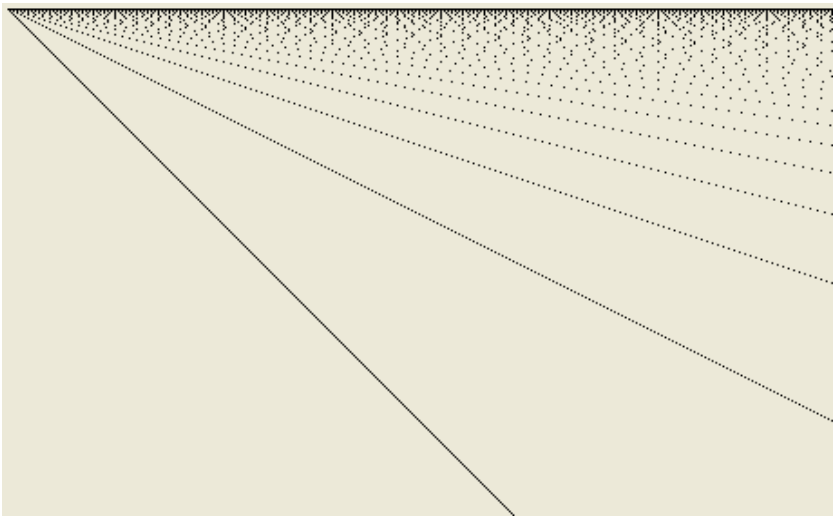


Figure 3.1: A visual representation of the sieve of Eratosthenes beginning with the number 1. Remarkably, the diagonals are constructed in exactly the same manner as the rows. Dots appear in every column for the diagonal with slope 1. They appear in every other column for the diagonal with slope 2. They appear in every third column for the diagonal of slope 3. And, so on.

Figure 3.2 shows a portion of the matrix beginning at number 17918. Diagonals are apparent at the top of the image. Near the bottom the dots merge into other patterns. From row 1 there are smaller diagonals radiating out from the top row. Several rows below

appear parabolic-like structures (these structures have, in fact, been proven to be parabolas and it has been proven that they are all oriented in the same direction).

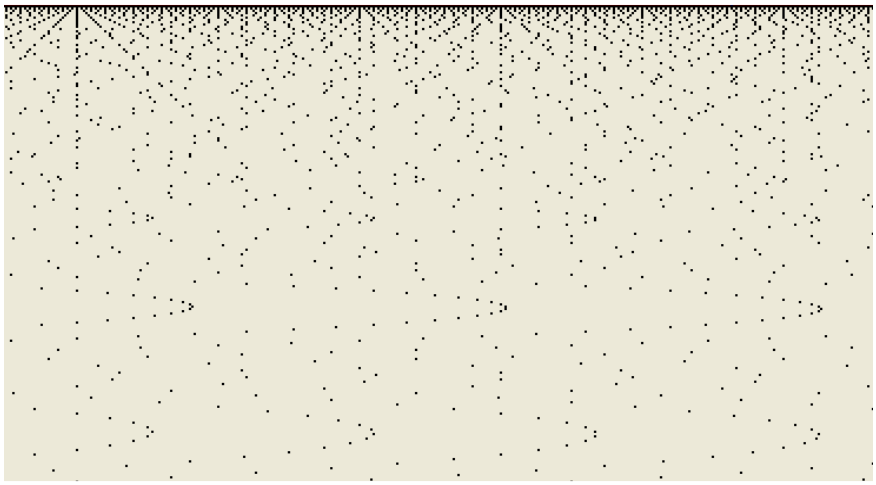


Figure 3.2: Further along in the matrix diagonals merge to reveal left facing parabolas.

Some of the diagonals radiating out from the first row will be very prominent. An example is 327600 as shown in Figure 3.3. This prominence is related to the number of dots in the central column. The more dots plotted, the more prominent the diagonals. This is easy to understand by considering that a dot represents that column $c$ is evenly divisible by row $r$. Consider two rows, $r_1$ and $r_2$, that both evenly divide column $c$. If $c$ is divisible by $r_1$, then so is $c + r_1$. Similarly, $c + r_2$ is divisible by $r_2$. Consequently a dot will be plotted at ($r_1$, $c$+$r_1$) and another dot will be plotted at ($r_2$, $c$+$r_2$). These points lie on a line with a slope of 1. The same reasoning can be extended to a dot at ($r_1$, $c$+2$r_1$) and another at ($r_2$, $c$+2$r_2$) which both lie on a line with a slope of 2.

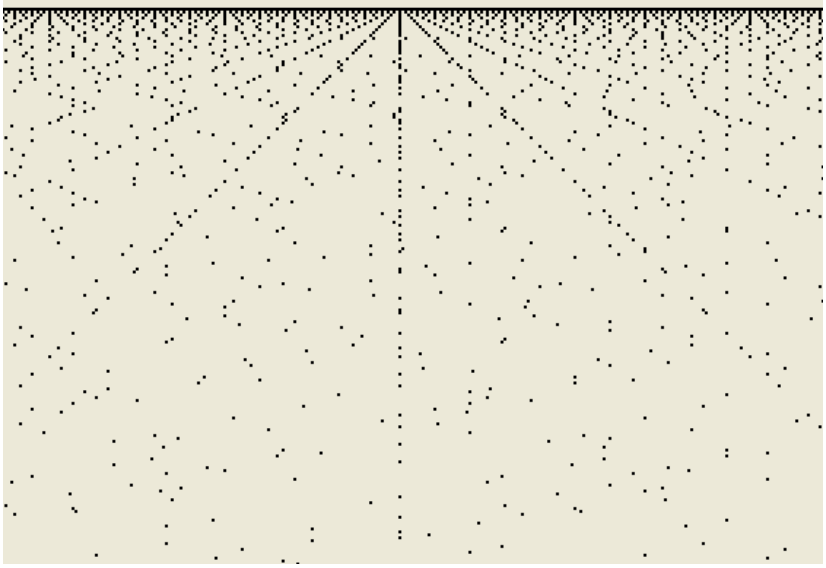


Figure 3.3: Diagonals radiate from the first row of every column. Some are more visible than others. Those most visible correspond to numbers with the most divisors. The pattern of dots in each diagonal is identical. Only the spacing increases as the diagonals approach the horizontal.

Thus, while it is not immediately obvious, every column of points will have corresponding diagonals. What about the diagonals in Figure 3.1? What central column of points do they correspond to? If the procedure used to construct the matrix is extended to the left, we arrive at a result that is illustrated in Figure 3.4. There is a central column where every row contains a dot. The image to the left of this column is the mirror image of that to the right. Furthermore, the central column corresponds to the number 0.

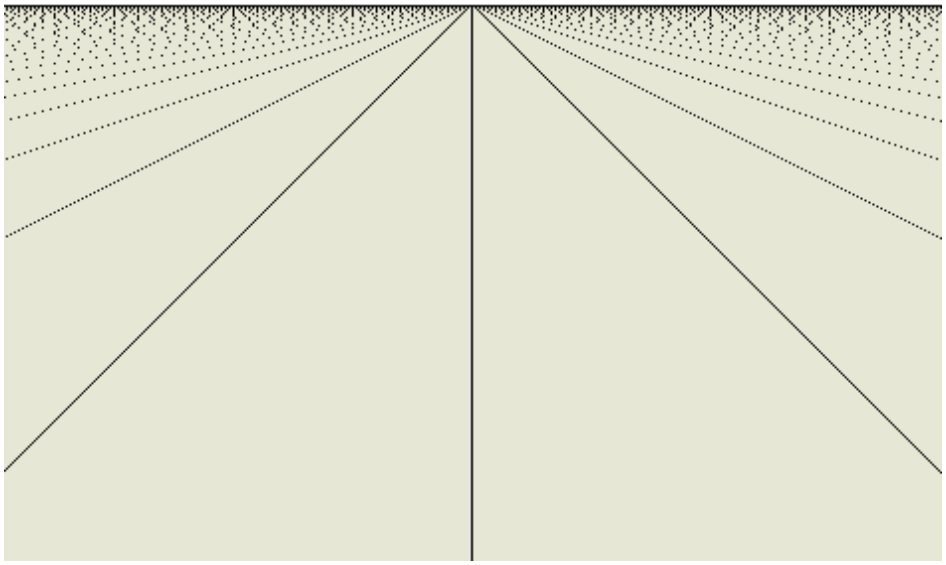


Figure 3.4: The matrix as it appears from column 0.

Another remarkable feature of Figure 3.4 is that the diagonals all seem to converge to a point. In fact, the convergence is real and the point is (0, 0).

Any binary image is easily represented numerically using 1's and 0's. Consequently, Figure 1 can be represented as in Figure 3.5.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Figure 3.5: Representing a portion of the matrix as 1's and 0's.

Carrying this idea further, these numbers can be replaced by remainders. In other words, each cell will contain the value $c$ mod $r$ where $c$ and $r$ are the column and row respectively. Doing so gives the result in Figure 3.6.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 3 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 0 |
| 4 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 0 | 1 | 2 |
| 5 | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 | 0 |
| 6 | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | 0 |
| 7 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 0 | 1 | 2 |
| 8 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 0 | 1 | 2 | 3 |
| 10 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 |
| 11 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 12 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 13 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 0 | 1 | 2 | 3 | 4 |
| 14 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 0 | 1 | 2 |
| 15 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 0 |
| 16 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 17 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 18 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 19 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 20 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 21 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 22 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 23 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 24 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 25 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 0 | 1 | 2 | 3 | 4 | 5 |
| 26 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 0 | 1 | 2 | 3 | 4 |
| 27 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 0 | 1 | 2 | 3 |
| 28 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 0 | 1 | 2 |
| 29 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 0 | 1 |
| 30 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 0 |

Figure 3.6: Illustration of how the remainders are duplicated in each column. Again, the spacing increases as each diagonal approaches the horizontal.

The central column of red numbers are the remainders of 15 divided by 1, 2, 3, ..., 15. Notice that these same remainders appear in the diagonals which are highlighted in red, yellow, and blue for easier reading. What is interesting is that every number will have a column of divisors which will be repeated in corresponding diagonals radiating out from the first row.

## 3.3 Discussion

This simple method for visualizing the sieve of Eratosthenes has resulted in surprisingly complex patterns. The set of dots in each column represents a set of divisors for that column. Extending out from each column is a set of diagonals with slopes of 1, 2, 3, etc. containing sets of dots that map 1-to-1 to the divisors in the corresponding column. This is true for every column.

Every column except for column 0 has a finite set of dots and every integer has a finite set of divisors while 0 is divisible by everything. Hence, column 0 contains an infinite set of dots.

The original sieve is used to find prime numbers. In this method, the prime numbers are represented in the image as columns containing exactly two dots. Column c corresponds to a prime number if it contains a dot in row 1 and row c and nowhere else.

Alternatively, these images can be represented numerically using a matrix whose cells are filled with 1's and 0's. However, it is not necessary to limit the numerical representation to 1's and 0's. The cells of the matrix can also be filled with remainders found by dividing each column by each row. Doing so reveals copies of each number's divisors along diagonals extending out from the first row. These diagonals have slopes of 1, 2, 3, etc.

These images are instructive in that they reveal that the divisors are not distributed randomly. There is repetition of each number's divisors along the diagonals that radiate from each number. Although a number's divisors radiate diagonally and are intertwined

with those of other numbers, they do not interfere with another number's divisors. Interestingly, the spacing of the divisors along these diagonals mirrors the spacing of dots used to create the initial images.

Additionally, there are other parabolic patterns that emerge in these images. These patterns have been shown to be true parabolas and that they are all oriented in one direction.

Perhaps the most interesting result is that the process of building the image had nothing to do with diagonals, parabolas, or other features. These features emerged from the plotting of points in each row following the method of Eratosthenes' sieve. We can explain why we see these particular features using a number of Gestalt laws. Perhaps the most relevant are the laws of symmetry and closure. Closure, in particular, comes into play as we perceive complete lines and curves from the collections of points.

These images illustrate the wonderful nature of the integers. Moreover, these images illustrate that even with a method more than 2,000 years old, a surprising new way of viewing the results can be found through the use of computerized visualization.

Can this technique be used to visualize patterns in other sequences such as DNA? The next chapter explains how this technique has been adapted to DNA sequences to show that, yes, a variety of patterns can be visualized in such sequences. Subsequent chapters explore the possible utility and meaning of these patterns.

# 4 Symbolic Scatter Plots

Chapter 3 introduced a technique for visualizing patterns in a sequence of numbers. This chapter introduces a modification of this technique to visualize patterns in DNA. DNA is a large polymer constructed as a sequence of small molecules called nucleotides. There are four possible nucleotides: adenine, cytosine, guanine, and thymine. Typically, these are represented with the letters A, C, G, and T. A typical DNA sequence consists of thousands of nucleotides. A sequence beginning with AGGATC... and continuing with many additional nucleotides would be one example.

The strategy for visualizing a DNA sequence begins by converting the DNA from a sequence of characters to a sequence of numbers. Once converted, the strategy is as in chapter 3. Divide each number by a set of integers to calculate a set of remainders. If a remainder is 0, then plot a point. Otherwise, don't.

A simple way to convert a DNA sequence to a sequence of integers is to represent each nucleotide as a number from 1 to 4. As an example, AGGATC would be converted into 133142. Each number in the sequence would be tested for divisibility by the numbers 1 to 4. If a number in the sequence is divisible by one of these divisors, then a point is plotted at the corresponding row and column.

Unfortunately, this technique produces images with only four rows of points resulting in plots that are hard to see and interpret. To address this limitation the technique is modified to allow points to be distributed over a larger number of rows. In other DNA processing techniques it is common to divide a DNA sequence into short overlapping

strings called *k*-mers. Each *k*-mer contains exactly *k* characters. Because there are only four possible characters (A, C, G, and T) to choose from, the number of possible *k*-mers is $4^k$. Thus, there are 16 2-mers, 64 3-mers, 256 4-mers, and so on. By plotting points that correspond to *k*-mers rather than nucleotides, plots can be created with a larger number of rows.

Using 3-mers as an example, each overlapping 3-mer of a DNA sequence is converted to an integer ranging from 1 to 64. Zero is not used because it is divisible by everything. A plot is produced where each column corresponds to a position in the sequence. Each row corresponds to a divisor ranging from 1 to 64. If the 3-mer is divisible by a divisor, then a point is plotted in the row corresponding to the divisor and the column corresponding to the position of the 3-mer in the sequence. Two examples are shown in Figure 4.1 illustrating patterns formed from points distributed over 64 rows.

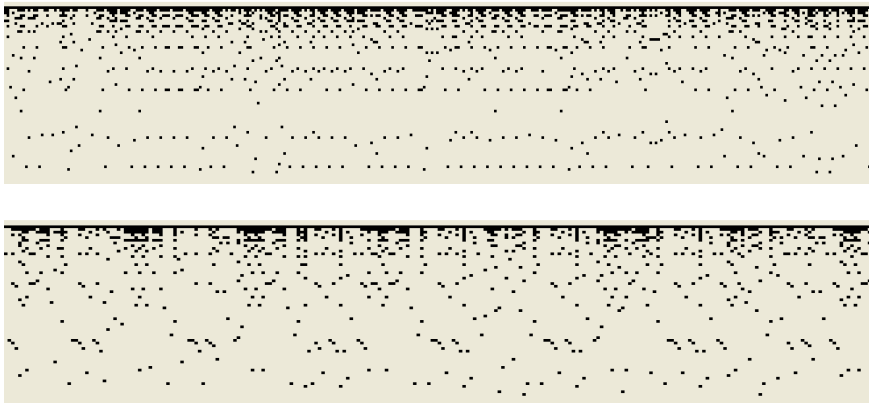


Figure 4.1: Examples of symbolic scatter plots.

These images along with many others not shown here demonstrate that non-random patterns can be produced from DNA sequences using this technique. The images are intriguing. However, the technique has a shortcoming that could conceivably produce patterns that do not really exist. Different 3-mers will be divisible by different numbers of divisors. For example, assume that the 3-mer TCA maps to 12. In turn, 12 is divisible by 1, 2, 3, 4, and 6 leading to five points in every column that corresponds to TCA. In contrast if TTT maps to 1, then it is only divisible by 1 leading to a single point in its column. This difference would bias the observer towards certain 3-mers over others potentially leading the observer to find patterns correlated to the number of divisors rather than to the distribution of nucleotides in the DNA. To eliminate this bias, the technique was modified to plot one and only one point per column. Specifically, a point is plotted in the row corresponding to the integral value of a 3-mer. In the above example, TCA maps to 12. Thus a single point is plotted in row 12. Similarly, a single point is plotted in row 1 for TTT.

3-mers can be mapped to the numbers ranging from 1 to 64 using a lookup table. However, many bioinformatics techniques will, in fact, map 3-mers to values ranging from 0 to 63. This mapping is done by assigning two bits to each nucleotide and concatenating the bits to form six bit binary number. For example by letting T map to 00 the 3-mer TTT will map to the binary number 000000 which, of course, is 0 decimal. If A maps to 11, then AAA maps to the binary number 111111 which is 63 decimal. If a TTT is encountered in the sequence, then a single point is plotted in row 0 of the corresponding column. Similarly

for AAA a single point is plotted in row 63. An example of such an image is shown in Figure 4.2. It was observed from this and other images that modifying the technique to eliminate any bias in the plots did not eliminate the patterns.

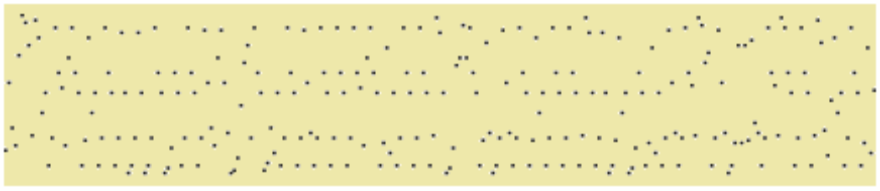


Figure 4.2: Effect of plotting one point per 3-mer.

Is there any relevance to the y-value? The y-values allow the points to be distributed vertically to produce patterns that draw the observer's attention. Otherwise, they have no particular meaning. For 3-mers there are 64 y-values. The 3-mers could be mapped to these y-values in 64! ways. Different mappings will change the proximity of points to each other. As we have seen, proximity does affect perception. It will be shown later how different mappings can be exploited to help reveal different patterns. It will also be shown how different visual cues such as color can be added to the plots to enhance their usability. For now, several examples of the patterns revealed by these symbolic scatter plots are presented. All of these were created using DNA sequences from the human Y-chromosome.

The symbolic scatter plot in Figure 4.3 shows several repeated 3-mers in the center of the image flanked on either side by two additional features (indicated by arrows) that do not resemble their surroundings.

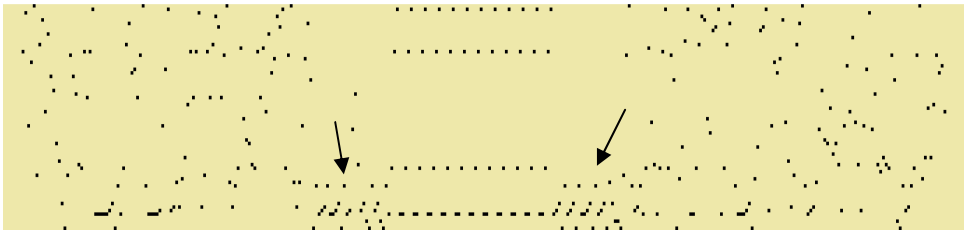


Figure 4.3: An example of repeated 3-mers in a symbolic scatter plot.

The repeats in the middle correspond to the sequence

AAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAA.

While those on the left correspond to

GGAAAGAAAAGGAAAGCAAGGGAAGGGAAA

and those on the right correspond to

GGAAAGGAAAAGGAAAGGAGGGGATGAAGAAATCAAATAAT.

Putting them all together gives:

GGAAAGAAAAGGAAAGCAAGGGAAGGGAAAAAATAAAATAAAATAAAATAAAATAAAATAAA
ATAAAATAAAATAAAATAAAATAAAATAAAGGAAAGGAAAAGGAAAGGAGGGGATGAAGAAA
TCAAATAAT

While it is possible to see repetitions in the sequence from the text alone, it is much more of a challenge to see how the ends of the sequence differ from the middle and how much they resemble each other. In contrast it is much easier to distinguish these three regions in the symbolic scatter plot.

This next image also shows a feature in the center (highlighted in orange) that does not resemble its surroundings. However, this feature doesn't have the high number of repeats as the in the previous image. Nevertheless, it is still distinguishable from its surroundings.

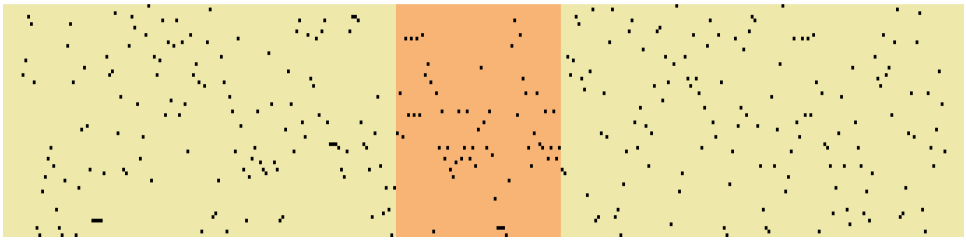


Figure 4.4: An illustration of how one region of a symbolic scatter plot is visually different from its surroundings.

The next image shows a structure highlighted in orange that appears frequently in DNA sequences. It corresponds to the sequence ATATATATATATAT. The arrows point to repeats of A's. These repeats of A's would normally go unnoticed even by programs that search for repeats because they are so short. However, the fact that they appear to be symmetrically arranged about the ATAT repeat makes them more interesting and easily noticeable.

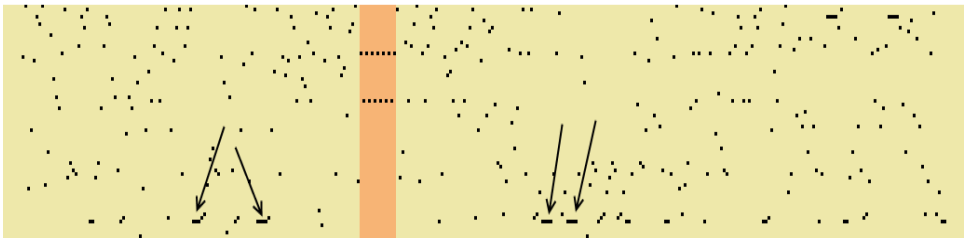


Figure 4.5: What makes a portion of a DNA sequence interesting? Here repeats of AAA's are arranged symmetrically around a central set of repeats. Is this a coincidence or is there some biological relevance?

The next image shows several short repeats of A's and other small features. Again, due to their size they would normally go unnoticed. However, in the symbolic scatter plot their relatively large numbers become apparent.

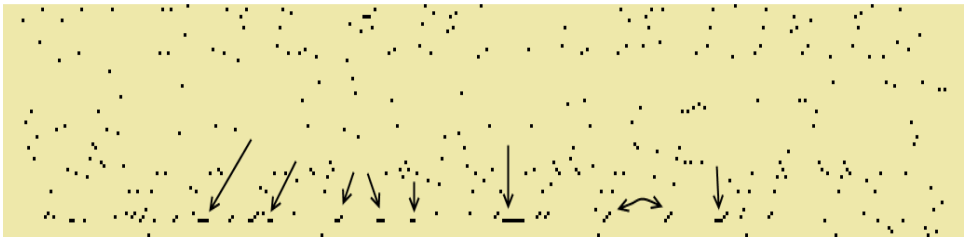


Figure 4.6: Small patterns such as these repeats would normally go unnoticed. However, their relatively large numbers make them visually conspicuous.

This next image shows a region that contains a moderate number of repeats interspersed with 3-mers that do not repeat or repeat in an irregular fashion. The highlighted region appears noticeably different from the surrounding regions.

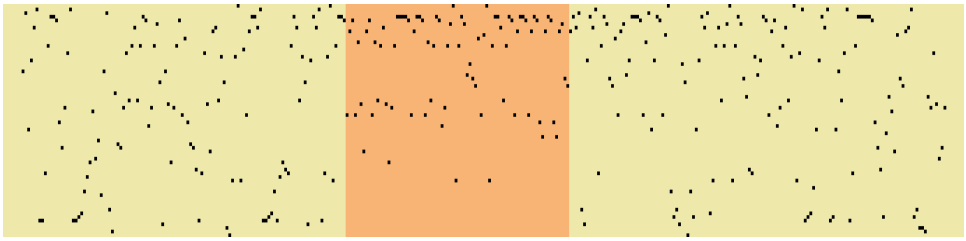


Figure 4.7: An example of a slightly repetitive region that is distinct from its surroundings.

The next image shows a single region that consists of several regions of repeats. Algorithms that search for repeats will correctly identify the smaller regions as repeats but will fail to recognize that taken together they form a larger region that is distinct from its surroundings. Here we take advantage of our ability to effortlessly group objects to form a coherent whole.

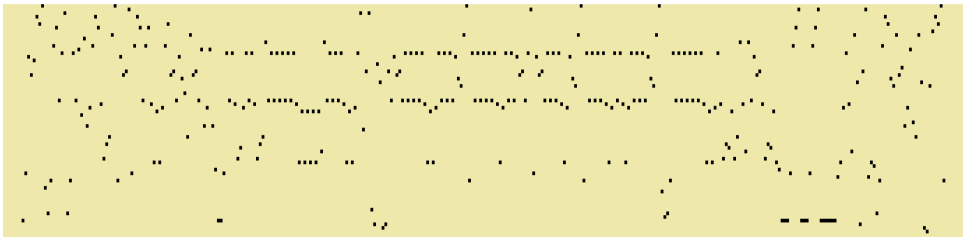


Figure 4.8: The human visual system groups objects at different levels. Here a series of several small groups of repeats are grouped to form a single whole. Such larger groupings are not easily identified algorithmically.

Our ability to group similar objects is also apparent in the following image. We can easily distinguish three different groupings while at the same time perceiving them as forming a single larger object based on their proximity to each other.

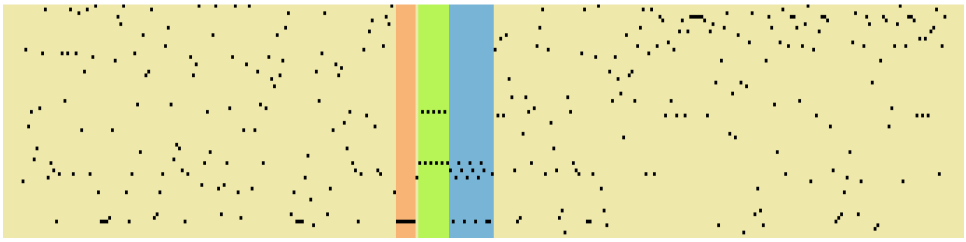


Figure 4.9: Here three very different groups of patterns are perceived by their close proximity to form a single objet.

Figure 4.10 contains two small clusters of dots arranged symmetrically about a center region. To the right are several small repeats that would normally go unnoticed. Further to the left is a regular little feature that also would normally go unnoticed. As for the region in the middle, can you tell that it consists of two regions arranged symmetrically about a center by glancing at the following text?

AAACAAACAAACAGGAAGCAACAGCAACAACAACAAAAAAGACCCCACAAAAACCCCCATTCA
AAGGCCAGCAACCTCAAAGATTGAAGATAAACCCACAAA

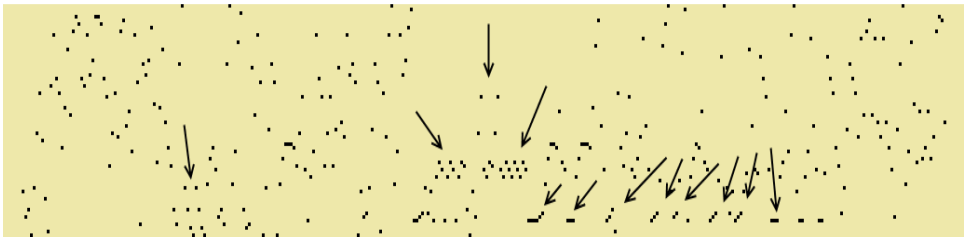


Figure 4.10: An example of many small features that are visually conspicuous. Visual cues such as proximity and symmetry make them stand out.

Patterns such as these occur throughout the DNA sequences that have been examined. The human Y-chromosome alone contains thousands of them. Some correspond to well know patterns of repetitive nucleotides. An example is the repetition of the nucleotide pair, AT, which is known as a "TATA" box. Most, however, correspond to patterns of nucleotides whose functions are unknown.

The remainder of this thesis examines the patterns of symbolic scatter plots in more detail. The patterns visible in symbolic scatter plots are compared to the repetitive patterns found by algorithms such as Tandem Repeats Finder. The utility of symbolic scatter plots is explored by using them to visualize the gene responsible for Huntington's disease. Lastly, the patterns are explained in terms of human visual perception. In particular, certain visual cues are manipulated to reveal specific types of patterns and to augment the scatter plots with additional information.

# 5 Sequence Analysis with Symbolic Scatter Plots

*The material in this chapter is based on the paper, "An Analysis of DNA Sequences Using Symbolic Scatter Plots"* (Cox & Dagnino, An analysis of DNA sequences using symbolic scatter plots, 2009) *presented at the 2009 International Conference on Bioinformatics & Computational Biology*.

## 5.1 Introduction

Chapter 4 introduced the technique for creating symbolic scatter plots and presented several examples of patterns visible in the plots. This chapter considers the effectiveness of using symbolic scatter plots for sequence analysis. The types of tasks considered are 1) identifying tandem repeats, 2) identifying complex patterns, and 3) comparing sequences.

The vast majority of methods used by researchers to look for patterns in DNA sequences are non-visual. Some look for patterns by comparing sequences and finding regions that are statistically similar. Usually the similarity is measured in terms of the edit distance between two strings. The results of such methods are tables of coordinates for locating the similar regions and a listing of the regions' nucleotides. Examples include BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990), PatternHunter (Ma, Tromp, & Li, 2002), MUMmer (Delcher, Kasif, Fleischmann, Peterson, White, & Salzberg, 1999), and ClustalW (Thompson, Higgins, & Gibson, 1994). The matching regions are usually characterized only by the probability that the regions could have matched by chance. Those regions that have a low probability of matching by chance are likely to play some important biological role. An obvious limitation of sequence comparison is that if two sequences

have nothing in common, then comparing them will find no significant patterns (Benson, 1999)(Schneider & Stephens, 1990).

Other methods look for regions called motifs that are statistically over-represented either within a single sequence or across several sequences. Examples of software of this type include MEME (Bailey, Williams, Misleh, & Li, 2006) and CONSENSUS (Stormo & Hartzell, 1989). A limitation of this approach is that if a pattern is not overrepresented in a sequence, then no motifs will be found (Frith, Fu, Chen, Hansen, & Weng, 2004)(Keich & Pevzner, 2002).

Some methods look for repetitive patterns. A popular (and possibly the most popular) tool for finding repetitive patterns in DNA sequences is Tandem Repeats Finder created by Gary Benson (Benson, 1999). A tandem repeat is a pattern that exists as many copies occupying consecutive locations within a sequence. These copies are sometimes exact but most often are inexact – that is the "copies" differ from each other through insertions, deletions, or substitutions of nucleotides. Tools such Tandem Repeats Finder identify the repetitive pattern, determine all the places where it is located, determine its periodicity, and characterize it statistically (e.g. to determine if the pattern could have arisen by chance).

A tool such as Tandem Repeats Finder is limited by the heuristics it applies to handle inexact copies. Hauth and Joseph note two issues that complicate repeat identification: imperfect conservation of patterns and complex pattern structures (Hauth & Joseph, 2002). Lack of conservation refers to differences in the repeats due to insertions,

deletions, and substitutions. Complex pattern structures are those formed by multiple simple patterns that are each repeated a variable number of times. For example, let $P = (p_1)_{c1}\ (p2)_{c2} \ldots (p_n)_{cn}$ where $p_i$ is some simple repeated pattern that is repeated $c_i$ times. For a simple sequence repeat (SSR) the number of times that each simple pattern is repeated is constant. For a variable length tandem repeat (VLTR) the number of one or more $p_i$ is variable. Multiple period tandem repeats (MPTR) allow for variable numbers of substitutions.

For example, CAGTA CAGCA CAATA CAGCA CAGTA CAGCA CAATA CAGCA could be considered as eight repeats of $CA^{G}/_{A}{}^{T}/_{C}A$ allowing substitutions in the third and fourth positions. However, the pattern could also be considered to be four copies of ($CA^{G}/_{A}TA\ CA^{G}/_{A}CA$) or two copies of (CAGTA CAGCA CAATA CAGCA). As noted by (Hauth & Joseph, 2002), Tandem Repeats Finder fails to identify and/or accurately characterize the complex patterns associated with VLTRs and MPTRs.

To find repeats of complex patterns, Hauth and Joseph developed a heuristic algorithm to address several small problems not considered by Tandem Repeats Finder. It is worth reviewing some of the details of this algorithm because they illustrate the general approach taken by researchers to find patterns in sequences.

Hauth's and Joseph's algorithm begins with a preprocessing step to build a distance array. Each entry in the array maps to a *k*-mer in the DNA sequence and contains the distance to the next identical *k*-mer. Histograms are constructed where identical distances

are placed in the same bin. For each peak in the histogram potential tandem repeats are identified.

The DNA sequence is then segregated into regions using "period analysis." This analysis assigns a base period to each region. Region patterns are then constructed for each region. These region patterns are constructed in one of several ways depending upon a class designation for the region. These region patterns are designated as base patterns and are used to evaluate the previously identified potential tandem repeats.

The algorithm continues by iterating over the potential tandem repeats and comparing each to the base pattern associated with the potential tandem repeat's region. Each potential tandem repeat is characterized by aligning it with its base pattern using wraparound dynamic programming. If the potential tandem repeat is characterized as a true tandem repeat, then the algorithm continues by identifying a tandem repeat region. The periodicities within each tandem repeat region are identified. For each periodicity another pattern is constructed and this new pattern is used to characterize the tandem repeat region.

Hauth and Joseph illustrate their algorithm with three examples. In one of their examples they identified two copies of a repeat about 10,000 nucleotides apart and reported the results using the following graphic:

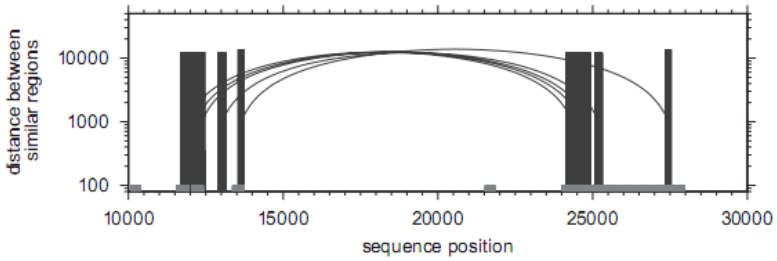



Figure 5.1: Matching regions from the yeast chromosome. (Hauth & Joseph, 2002)

The arcs connect matching regions within the chromosome. Unfortunately, it is impossible to conclude if these matches are good or not from this graphic. Furthermore, Hauth and Joseph provide no results such as numerical data to illustrate why these particular matches are good. Nor did they present any results of a comprehensive comparison of their algorithm with Tandem Repeats Finder or other algorithms for finding repeats.

Other researchers have also addressed the shortcomings of Tandem Repeats Finder with a variety of heuristic approaches. These include a spectral-statistical approach (Chaley & Kutyrkin, 2008), an application of hash functions (Reneker & Shyu, 2005), an application of quaternions and signal processing (Brodzik, 2007), and a combinatorial approach (Kolpakov, Bana, & Kucherov, 2003). These techniques universally begin by finding small repeating strings that act as guideposts for further analysis. The techniques differ by how they proceed with that analysis. For Hauth and Johnson, analysis proceeds by calculating the distance array and histogram. Chaley's and Kutyrkin's analysis proceeds by performing a statistical analysis and extracting spectral information. Brodzik assigns

nucleotides to complex numbers and then combining the complex numbers to convert the sequence to quaternions. The quaternions are then analyzed using signal analysis. Kolpakov, et al begin with *k*-mers which are analyzed for certain combinatorial properties.

These various techniques illustrate that there is no precise definition of a "tandem repeat" or agreed upon approach for finding them. All start with short strings and apply various tactics until an acceptable result is reached. Ideally, one would like an algorithm that reveals interesting patterns including repeats while imposing as few constraints as possible on the search. Graphical approaches have long been employed to find patterns in data. The scatter plot has been used for decades to find correlations between data sets. Indeed, scatter plots are frequently used to compare DNA sequences with Dot Plots being a prime example. Such approaches impose minimal search constraints allowing interesting patterns to be revealed in the final image.

A goal of this research is to apply a similar minimalist approach to finding patterns in DNA sequences. As described in chapter 4, symbolic scatter plots also begin with short *k*-mers. These *k*-mers are converted directly into an image and not analyzed further by machine. In contrast to other techniques, subsequent analysis relies entirely on the human visual system. Rather than trying to define tandem repeats and complex patterns rigorously, this approach defines them not at all and relies on the visual system to locate them.

To demonstrate the applicability and efficiency of symbolic scatter plots, the repeats found with Tandem Repeats Finder are compared to what is visible in symbolic scatter

plots. Symbolic scatter plots are examined for their ability to identify complex patterns. Lastly, comparing sequences by comparing their symbolic scatter plots is considered.

## 5.2 Results

Three categories of results are presented. The first compares symbolic scatter plots with Tandem Repeats Finder. The second examines complex patterns found in symbolic scatter plots. The third explores using symbolic scatter plots for sequence comparisons. In total more than 100 sequences were analyzed and representative examples are presented here. About three-quarters of these sequences and their scatter plots are available at http://www.guffy.net/ssp.

### 5.2.1 Repeats Analysis

Earlier it was explained how visualizing the sieve of Eratosthenes was the inspiration for symbolic scatter plots. Dividing the sequence into *k*-mers was used as a way to convert the DNA into a sequence of integers that could then be plotted. It has also been explained how traditional tools for analyzing DNA also begin by dividing the DNA into *k*-mers. There are more than cursory similarities between these tools and symbolic scatter plots. These similarities are particularly evident with Tandem Repeats Finder.

Tandem Repeats Finder (Benson, 1999) begins its analysis of DNA by building a hash table entry for each *k*-mer in the sequence. Benson provides an example with *k* equal to 5. The possible 5-mers are AAAAA … TTTTT. A window of length 5 slides along the sequence and the position of the window is noted for each 5-mer residing in the window. These positions are added to the corresponding 5-mer's hash table entry (referred to by Benson as the 5-mer's "history list").

When a new position is added to a history list, the position is compared to other positions already in the list. The hypothesis is that each position corresponds to a possible tandem repeat. To validate the hypothesis the distance *d* between two positions *i* and *j* is calculated. To determine if repeats are present at *i* and *j*, other 5-mers between *i* and *j* are examined. If these 5-mers are also found to have a copy a distance *d* away, then a repeat is flagged as detected. To allow for possible substitutions, insertions, and deletions, a sum of heads distribution, a random walk distribution, and an apparent size distribution are applied to evaluate the two potential repeats. If the required heuristics are met, the repeats are reported to the user. The process is depicted in Figure 5.2.

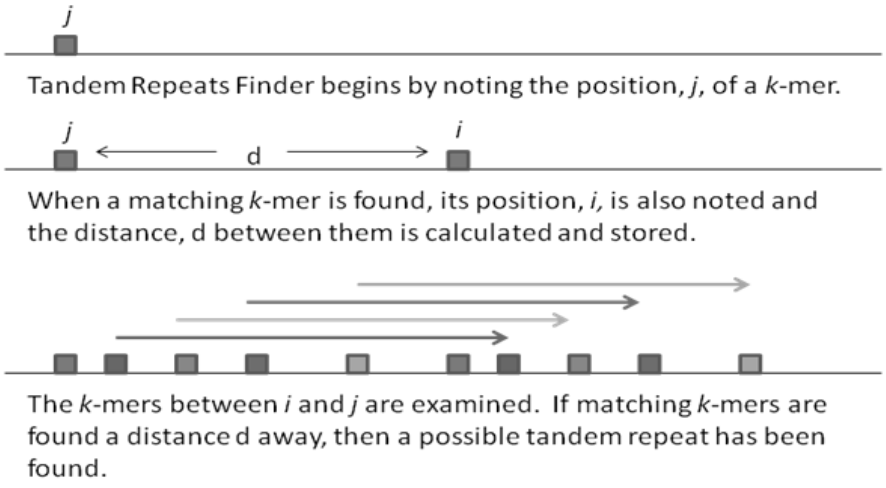
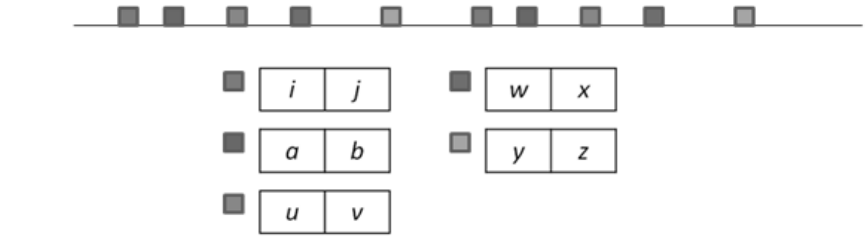



Figure 5.2: How Tandem Repeats Finder uses hashing to find tandem repeats.

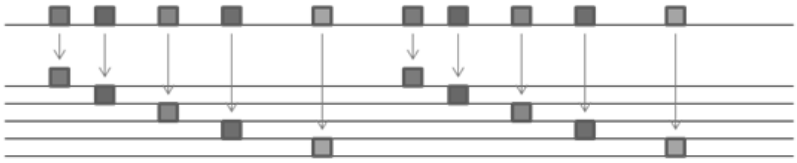
In comparison symbolic scatter plots are created by first determining the positions of *k*-mers in a sequence. Rather than *k*-mers of size 5, the typical symbolic scatter plot uses

size 3. The height of a symbolic scatter plot depends on the size of *k*. When *k* is 3, the scatter plot contains 64 rows. If *k* were 5, then the height of the scatter plot is 1024. As will be explained later, the patterns visible in symbolic scatter plots depend on the proximity of the points to each other. With 1024 rows, the points tend to be widely dispersed making it difficult to perceive groupings of data points. Aside from this proximity problem, Nature itself likes to group nucleotides three at a time. Nearly all living things map groups of three nucleotides to amino acids. Thus, groups of three nucleotides produce reasonably sized scatter plots and groups of three nucleotides represent a natural unit of information for life as we know it.

History lists for each 3-mer are constructed visually. The y-axis corresponds to the entire set of possible 3-mers where each 3-mer occupies a single row. A history list corresponds to each row. If AAA is the 100th 3-mer in the sequence, then a point is plotted in the 100th column of row AAA. It is the collection of points that are plotted in each row that represents the history list for the 3-mer. The end result is a scatter plot of points that map one-to-one to the positions recorded in the history lists of Tandem Repeats Finder. These differences are illustrated in Figure 5.3.




Tandem Repeats Finder notes the positions each *k*-mer in each *k*-mer's "history list" or hash table. Distances between *k*-mers are calculated by machine.



In contrast, Symbolic Scatter Plots hash the positions as points in a graph. Each row represents the "history list" for a given *k*-mer. The positions correspond to the columns in which the *k*-mers are found. The distances between *k*-mers are determined visually.

Figure 5.3: How hashing is used to create a symbolic scatter plot.

The chromosomal contig, NT_008183.18, for the ARFGEF1 gene was submitted to Tandem Repeats Finder which identified 126 repeats. One of these repeats begins at nucleotide 71680. The results are reported in Figure 5.4.

```
71670 GCCTTCTCTC

71680 TCTT TCTT TCTT TCTT TCTT TC-T T-TT TCTT TCTT TCTT TCTT TCTT
    1 TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT

71726 TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT
    1 TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT

              *    *    *    *    *    *    *
71774 TCCTT TCCT TCCT TCCT TCCT TCCT TCCT TCCT TCTT TCTTT TCTTT TCTTT
    1 T-CTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TCTT TC-TT TC-TT TC-TT

71826 TCTTT TCTTT TCTTT TCTTT TCTTT TCTTT TCTTT TCTTT TCTTT TCTTT TCTTT
    1 TC-TT TC-TT TC-TT TC-TT TC-TT TC-TT TC-TT TC-TT TC-TT TC-TT TC-TT

                                    *
71876 TCTT T-TCT T-TT TCTT T-TT CTTTT TCTT TCTT T
    1 TCTT TCT-T TCTT TCTT TCTT -TCTT TCTT TCTT T

71908 TTGTTTTAAG
```

Figure 5.4: An analysis of chromosomal contig, NT_008183.18 using Tandem Repeats Finder.

The first row beginning at 71670 is a short sequence that precedes the repeats. The last row beginning at 71908 is a short sequence that follows the repeats. Neither is actually part of the repeats that begin at 71680. For each pair of rows, the top row is from the actual sequence. The bottom row is a consensus sequence. The asterisks correspond to substitutions and the dashes correspond to insertions or deletions. The repetitive pattern is given as TCTT and the overall length of the region where the pattern occurs is calculated as 71908 – 71680 which equals 228. The symbolic scatter plot for the region containing this repeat is in Figure 5.5.

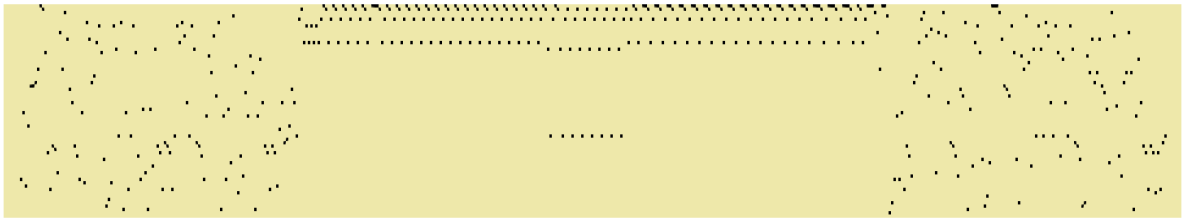


Figure 5.5: A symbolic scatter plot visualizing the same region as Figure 5.4.

The repetitive nature of the region is quite apparent. The middle of this region shows the substitutions evident from the Tandem Repeats Finder report. The effect of substitutions is a vertical shift in the points. A shift of points is also apparent of insertions and deletions because the result is a transformation of a 3-mer into another 3-mer.

Something that the symbolic scatter plot illustrates that is not obvious from Tandem Repeats Finder is the nearly symmetrical nature of the region. Not only is there a middle region resulting from substitutions, there is also a clustering of repeats at the left and right ends of the region.

Another example from the same sequence begins at 80175. The results from Tandem Repeats Finder are in Figure 5.6.

```
80165  ACGGGA

80175  GAGGGGGAGGGG
    1  GAGGGGGAGAGG

80187  GAGGGGGAGAGG
    1  GAGGGGGAGAGG

80199  GAGAGGGAGA
    1  GAGGGGGAGA

80209  CCTAACGGTT
```

Figure 5.6: Results of Tandem Repeats Finder for another portion of chromosomal contig, NT_008183.18. Contrast this result with the symbolic scatter plot in the next figure.

Here the repeats are of size 12. There are two whole repeats at 80175 and 10187. At 80199 there is a partial repeat consisting of the first ten nucleotides. Consequently, the total number of repeats is reported as 2.8. Figure 5.7 presents the symbolic scatter plot for this region.

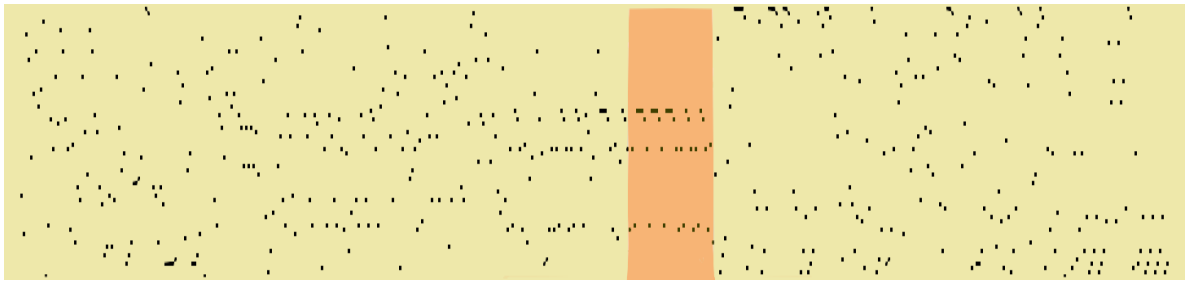


Figure 5.7: This symbolic scatter plot illustrates that Tandem Repeats Finder failed to find all of the repetitive 3-mers in this portion of the sequence. TRF reported only the highlighted region and missed the repeats to the left.

The repeats reported by Tandem Repeats Finder are contained in the highlighted area. Notice the additional repeats to the left of the highlighted area. These repeats were not reported by Tandem Repeats Finder but are evident in the symbolic scatter plot. Notice that these additional repeats lack the same regular spacing as those in the highlighted area indicating that the pattern is not as well conserved as on the right. Despite the lack of regular spacing, the human visual system is able to adapt and recognize them as additional repeats.

In many cases Tandem Repeats Finder reports many overlapping repeats. In Figure 5.8, the repeats listed in the first two rows fall within two overlapping regions ranging from position 358857 to 358991.

| Indices | Period Size | Copy Number | Consensus Size | Percent Matches | Percent Indels | Score | A | C | G | T | Entropy (0-2) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 358869--358971 | 14 | 7.5 | 14 | 84 | 12 | 133 | 50 | 22 | 0 | 27 | 1.49 |
| 358857--358991 | 16 | 8.9 | 16 | 79 | 9 | 133 | 50 | 22 | 0 | 27 | 1.49 |
| 359358--359404 | 2 | 23.5 | 2 | 95 | 0 | 85 | 51 | 46 | 0 | 2 | 1.13 |
| 360490--360576 | 22 | 4.0 | 22 | 84 | 12 | 119 | 81 | 1 | 17 | 0 | 0.75 |
| 360483--360571 | 43 | 2.1 | 41 | 89 | 4 | 133 | 83 | 1 | 15 | 0 | 0.71 |
| 360508--360569 | 10 | 6.3 | 10 | 85 | 9 | 83 | 82 | 1 | 16 | 0 | 0.75 |

Figure 5.8: Tandem Repeats Finder identifies several often overlapping regions of repeats.

Now, consider the symbolic scatter plot for the same region in Figure 5.9. Scrolling to the right or the left of the image reveals that it is distinct from the surrounding sequence. The different rows of points illustrate that the 3-mers are not evenly distributed.

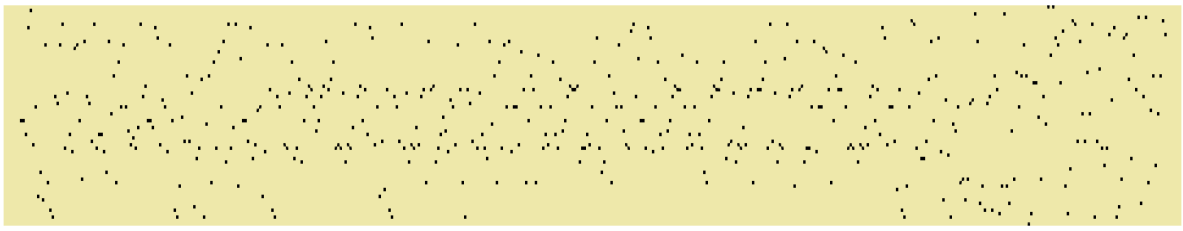


Figure 5.9: As illustrated in Figure 5.8, the repeats in this symbolic scatter plot are reported as several overlapping regions. Visually, the region appears as a single object.

In many cases repeats are evident in symbolic scatter plots but are not reported as repeats by Tandem Repeats Finder. Figure 5.10 presents an example beginning at 97179 extending to position 98169. The middle of this plot shows several repeats of the 3-mer TGA (represented by the dashes at the middle bottom of the plot indicated by the arrow) that correspond to STOP codons. These repeats were not reported by Tandem Repeats Finder.

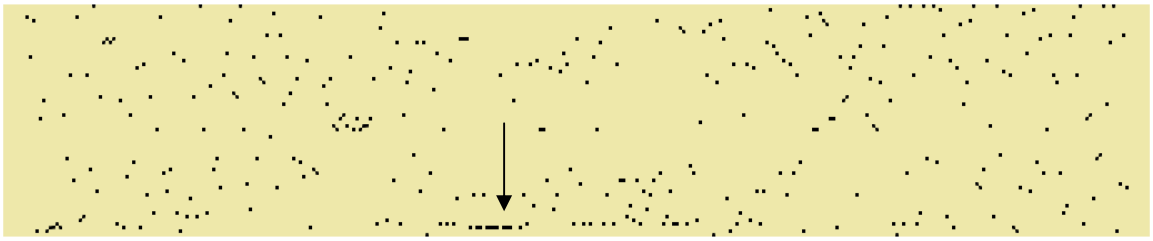


Figure 5.10: At the bottom of this plot are several repeats corresponding to STOP codons. These repeats were not reported by Tandem Repeats Finder but are evident visually.

Figure 5.11 presents another example that is part of the symbolic scatter plot for NT_029419.11. The triplet at the top corresponds to the 3-mer TAT while the doublet at the bottom corresponds to the 3-mer ATA. These repetitive patterns together correspond to a TATA box which typically exists in the promoter region of DNA and is often the binding

site of RNA polymerase II. TATA boxes are an example of a biologically important, repetitive sequence. In this case the pattern for the TATA box is readily visible in the symbolic scatter plot but is not listed as a repetitive element by Tandem Repeats Finder.

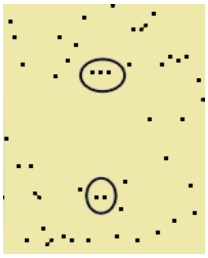


Figure 5.11: Here is a short repeat with an easily recognizable pattern. These short repeats are often not reported by programs such as Tandem Repeats Finder.

A symbolic scatter plot presents many repeats for NM_006191.2 (for the PA2G4 gene). Although Tandem Repeats Finder reported, "No Repeats Found!" for the sequence, there are many doublets and triplets of 3-mers and groups of 3-mers as in the following:

Figure 5.12: In this figure are additional examples of short repeats not reported by Tandem Repeats Finder. Are they biologically significant?

These and other groupings are visible in the symbolic scatter plot of Figure 5.13. The region at the left in the dotted circle is an example of a region of DNA that is very different from the surrounding sequence. It is not a tandem repeat. Yet it is not random. Symbolic scatter plots are very useful in revealing such patterns that escape detection by other tools.

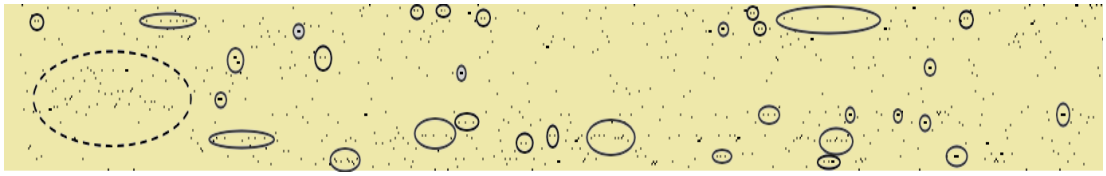


Figure 5.13: A zoo of visual patterns. Statistically we would expect to find many short repeats. However, both the number of repeats and the regularity of the spacing between them are visual cues that allow some patterns to stand out from others.

### 5.2.2 Identifying Complex Patterns

Symbolic scatter plots can be used to find complex patterns. Figure 5.14 presents the symbolic scatter plot for AF249277 (top panel) and a corresponding entropy profile (bottom panel).

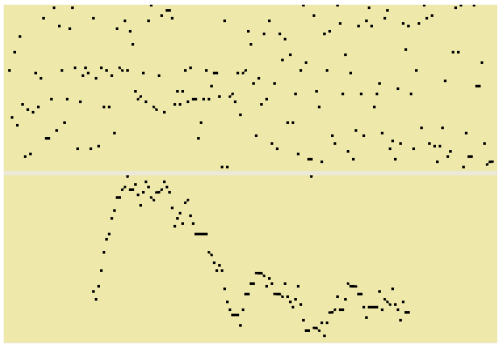


Figure 5.14: Different visual patterns correspond to different information content. Here a symbolic scatter plot is compared to an entropy profile for the same sequence.

The region to the left of center differs from its surroundings. The entropy profile at the bottom demonstrates that the region corresponds to an area of low entropy (low entropy is towards the top). Although there are few discernable repeats, the region of minimum entropy is distinguishable from the surrounding regions.

Figure 5.15 presents a symbolic scatter plot for a portion of the DAP gene. The plot shows patterns of repeating 3-mers (several are indicated by the arrows). However, the pattern is complex because the repeating 3-mers are irregularly spaced. In fact, Tandem Repeats Finder reported no repeats found for for this portion of the sequence. The only repeats identified by Tandem Repeats Finder for the entire gene were a series of about 20 A's (known as a poly-A tail) at the end of the sequence (not shown).

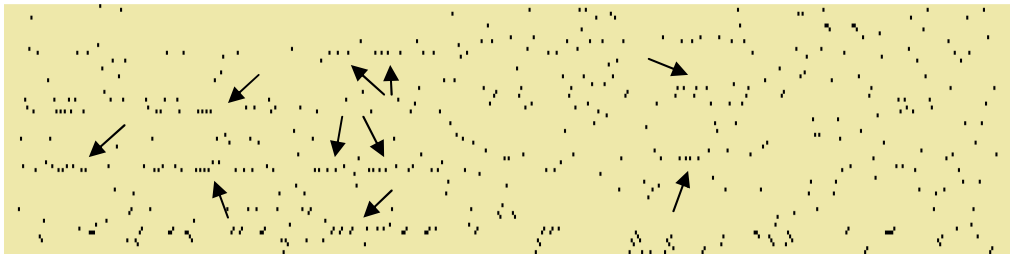


Figure 5.15: A complex pattern of irregularly spaced repeating 3-mers. Tandem Repeats Finder failed to report these repeats even though they are fairly evident to the eye.

The next examples presents a sybolic scatter plot for the RPS6KA5 gene. Again the symbolic scatter plot (top panel) presents a complex pattern of repeating 3-mers that were not reported by Tandem Repeats Finder as tandem repeats. The remaining panels of Figure 5.16 present complex patterns for the JAK2, CDC42EP3, and CHN2 genes. In each case Tandem Repeats Finder reported finding no repeats.

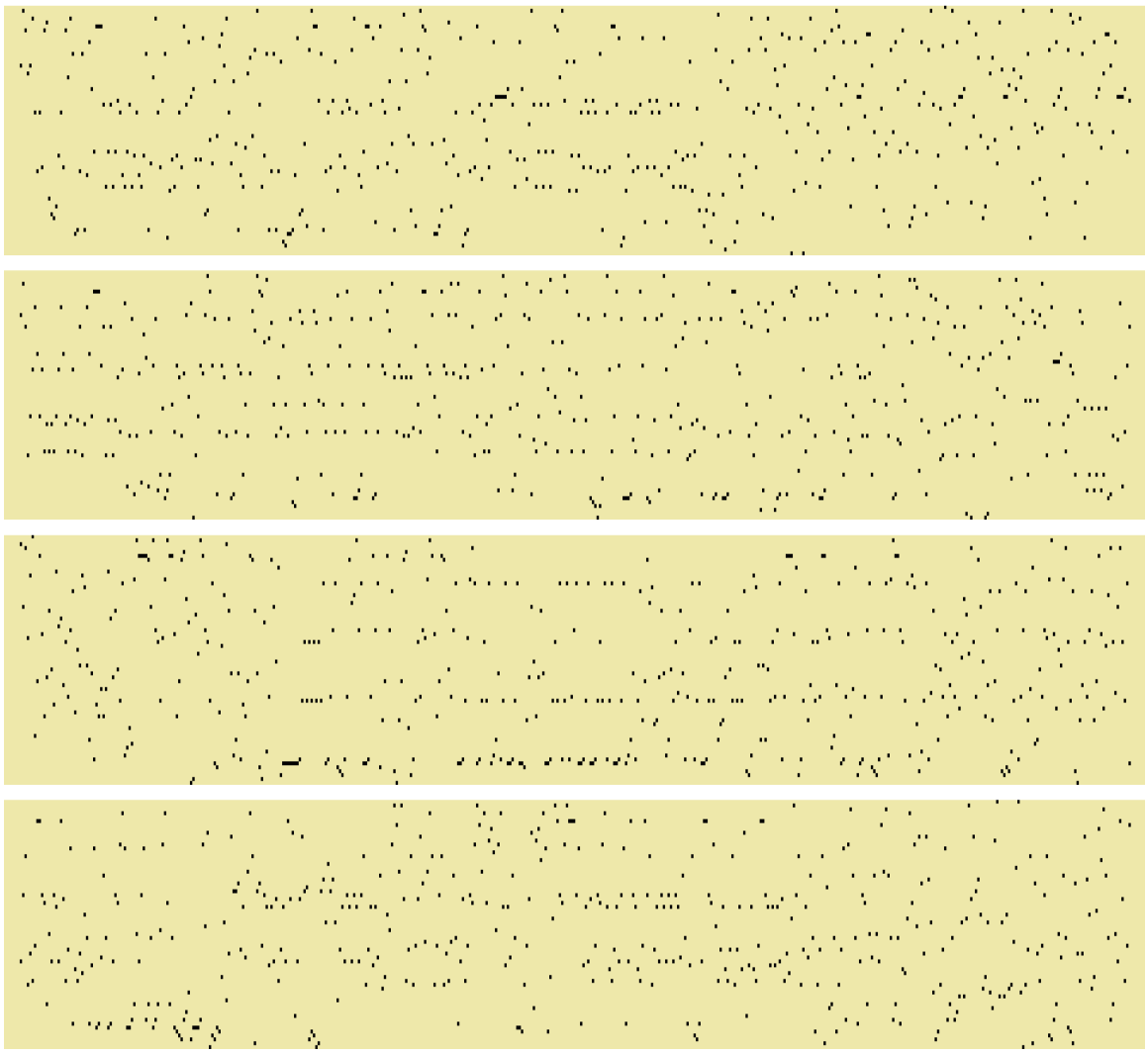


Figure 5.16: Examples of visibly complex patterns consisting of repeating 3-mers.
No repeats were found by Tandem Repeats Finder.

### 5.2.3 Comparing Sequences with Symbolic Scatter Plots

Algorithms that compare sequences insert spaces as needed to find an alignment of the sequences such that the final strings (with spaces) has a minimal edit distance. An observer can compare symbolic scatter plots visually by taking note of similar features. To

demonstrate, the human insulin gene is aligned with the cow insulin gene using four traditional sequence alignment algorithms: FASTA, ClustalW, TCoffee, and Needleman-Wunsch. Figure 5.17 illustrates the insertions to the human gene after alignment with the various methods. Figure 5.18 illustrates the same results using symbolic scatter plots.

```
No Alignment:        AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGGTCTGTTCCAAGGGCCTTTGCGTCAGGT
FASTA:                                                CCATCAAGCAGGTCTGTTCCAAGGGCCTTTGCGTCAGGT
ClustalW:            ---------------------------------------------AGCCCTCCAGGAC-AGGCTGCATCAG
TCoffee Alignment:   AGCCCTCCAGG--------A--CAGG--------------CTGCATCAGAA-------GAGGCCATCAAG
Needleman-Wunsch:    AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGGTCTGTTCCAAGGGCCTTTGCGTCAGGT
```

Figure 5.17: Different alignments produced by different alignment algorithms.

The different alignment results illustrated in Figure 43 is a common and frustrating problem. Lunter, et al report their finding that more than 15% of aligned nucleotides are incorrect in whole genome alignments (Lunter, Rocco, Mimouni, Heger, Caldeira, & Hein, 2008) . They and other researchers (Dewey & Pachter, 2006) state that existing algorithms show considerable disagreement. They argue that this disagreement is the result of "limited information available in extant sequences, from which different algorithms infer distinct but equally plausible homologies."

These disagreements are apparent in Figure 5.17. The FASTA alignment deletes several nucleotides and begins the alignment as shown with CCAT... ClustalW inserts several spaces at the beginning of the sequence. TCoffee inserts spaces throughout the beginning of the sequence. Needleman-Wunsch doesn't insert any spaces or delete any characters at the beginning of the sequence. Differences in the number and location of spaces are seen throughout the aligned sequences. In Figure 5.18 we see the symbolic scatter plots for these alignments.

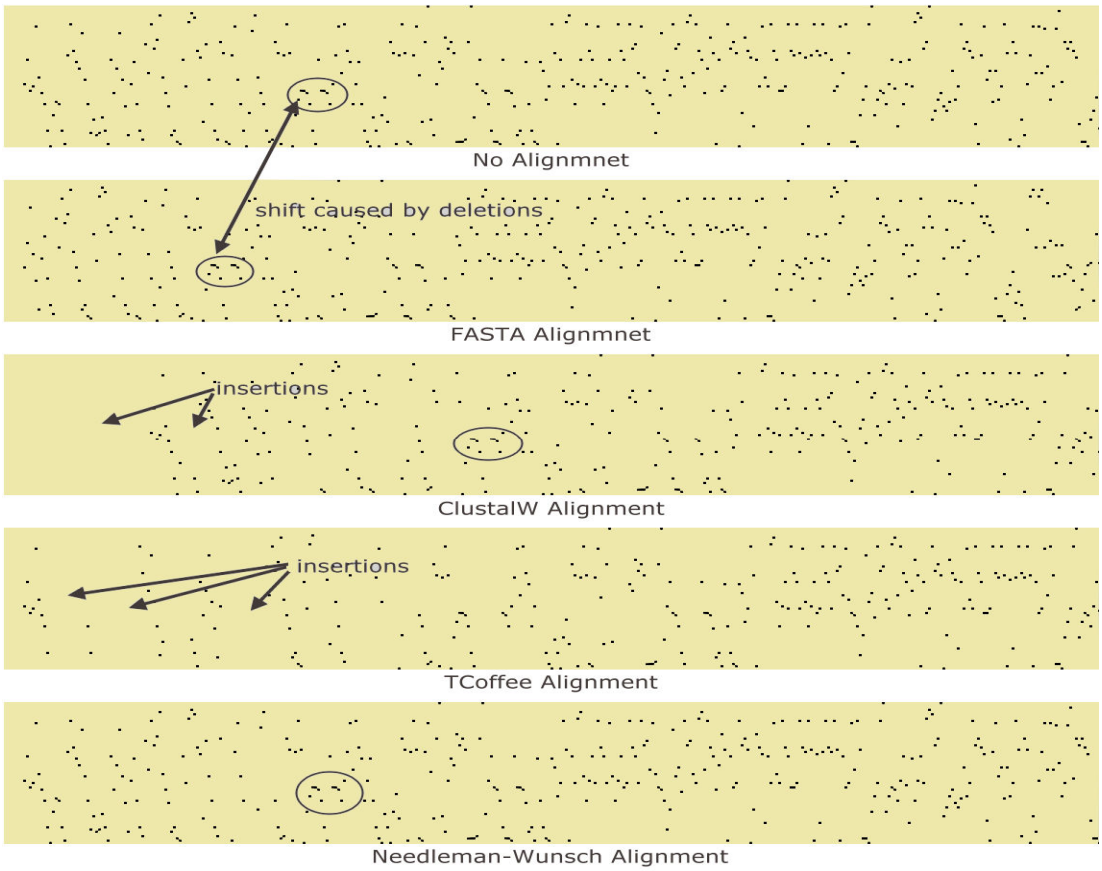



Figure 5.18: Comparison of alignments using symbolic scatter plots.

These scatter plots illustrate how the different alignment algorithms inserted spaces into the sequence and how they shifted and affected the various features. Notice the small feature circled in four of the plots. Examination of this feature and the surrounding points strongly suggests that the plots should be aligned on this feature. Feature-by-feature alignment is possible when aligning the scatter plots visually. However, such feature-by-feature alignment is not possible with current alignment algorithms. In fact, the insertions of the TCoffee alignment obliterate this particular feature in Figure 5.18.

Another example is presented in Figure 5.19 for the cDNA and chromosomal DNA of the CCNA1 gene. The top panel corresponds to the sequence NM_003914 and the

bottom to the sequence NT_024524.13. NM_003914 corresponds to the messenger RNA after NT_024524.13 (bottom panel) has been edited. The circled regions in the top panel are, by definition, contained in exons. The region circled at the left in the bottom panel corresponds to an intron that has been removed during transcription. Notice how it does not match the region in the top panel.

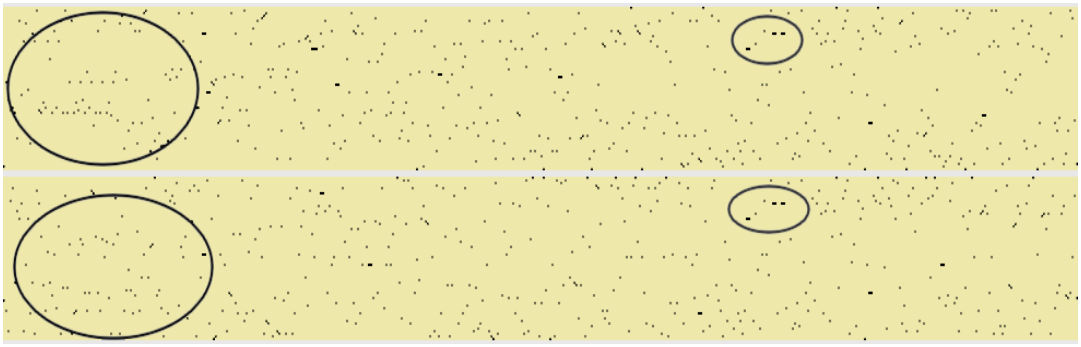


Figure 5.19: Visual patterns can be used as markers to assist in the alignment of sequences. Here the small patterns on the right suggest a possible alignment. A closer examination does show that the right half of the sequences do align well. However, the sequences do not align at the left as indicated by the circled patterns.

Scrolling through the scatter plot for NT_024524.13 one comes across the region that corresponds to the one circled in for NM_003914. This is presented in Figure 5.20.

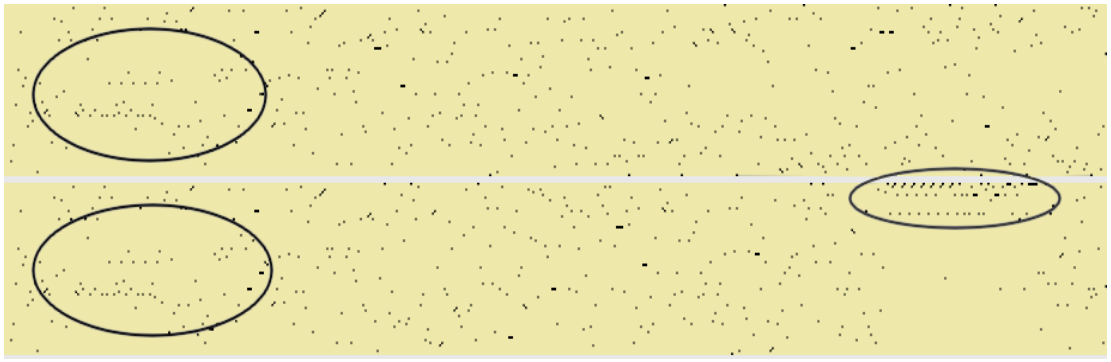


Figure 5.20: Here the bottom sequence was scrolled until a pattern matching the top left pattern was found. Notice the differences, however, on the right.

Again, the symbolic scatter plot for NM_003914 is presented in the top panel and the plot for NT_024524.13 is presented in the bottom panel. Now the regions circled at the left agree. However, there is now a region circled at the right that is not present in the plot for NM_003914. This region must, by definition, belong to an intron. Does this particular pattern have some special influence over transcription? Surely it is a region worthy of further investigation.

## 5.3 Discussion

More than 100 sequences were examined with symbolic scatter plots and fully 74 of these sequences were also checked for tandem repeats using Tandem Repeats Finder. Many of the examples presented in this chapter illustrate that Tandem Repeats Finder fails to find many kinds of complex repetitive patterns. In some cases Tandem Repeats Finder finds part of a pattern. In still other cases Tandem Repeats Finder finds many tandem repeats. However, when examined visually it becomes apparent that these many separate repeats constitute a single complex pattern.

It is quite difficult to rigorously define such patterns. Many researchers have offered widely different definitions using various heuristics. To date there is no universally agreed upon approach. In contrast, examination of a symbolic scatter plot relies on the human visual system to find complex patterns. The technique purposely offers no definition of these patterns and, yet, the eye can find them.

There are many alignment algorithms available for comparing sequences. As illustrated in this chapter, different algorithms produce different alignments. This begs the question, "which alignment is correct?" Visualizing the sequences using symbolic scatter plots reveals that the plots have conspicuous features suggesting that these features chould serve as guides for aligning the sequences. Visualization allows for such feature-by-feature comparisons that are not possible with currently available alignment algorithms.

Someday we will likely fully understand how the visual system works. Someday we will likely be able to build systems that see as well as we do. However, until that day the human visual system will remain unmatched in its ability to find patterns.

# 6 Visualizing the Huntingtin Gene

## 6.1 Introduction

Huntington's disease is a neurological disorder characterized by progressive uncontrolled movements, behavioral changes, and dementia. Despite efforts to find a cure, no effective therapy exists. Since 1993 it has been recognized that Huntington's disease is associated with an excessive number of CAG repeats in the gene that codes for the huntingtin protein (Macdonald, 1993). The normal human gene contains from 10 to 35 copies of the CAG triplet while Huntington's disease is associated with a gene containing 40 or more copies.

The CAG repeats reside in the coding region of the gene and are translated to become a series of polyglutamines referred to as a polyglutamine tract. The additional CAGs in the defective gene translate into a much longer polyglutamine tract. It has been demonstrated that these longer polyglutamine tracts lead to an aggregation of N-terminal fragments of the protein and their accumulation has been implicated as a pathogenic mechanism in Huntington's disease in transgenic mice and monkeys (Yang, et al., 2008). These experiments were performed by inserting the defective *human* huntingtin gene and observing the result. The conclusions were not drawn from defective huntingtin genes native to these species.

What do the huntingtin genes native to these and other species tell us about Huntington's disease? A first step to answering this question is to compare the DNA sequences of these genes. The object of this research was to compare the DNA sequences from human and *Pan troglodytes* (chimpanzee) using symbolic scatter plots. These plots immediately revealed that the huntingtin gene in the chimpanzee genome assembly has two regions of CAG repeats rather than the one region found in humans and other species.

The presence of two regions of CAG repeats is completely unexpected and has never been previously reported even though the chimpanzee genome has been available since 2006 (Chimp (Pan troglodytes) Genome Browser Gateway). What does this duplication mean for Huntington's disease? As always when experimenting with living things, it is important to first verify the result with additional tests and experimentation. It is possible that the duplication is the result of an experimental or computational error. Nevertheless, it is noteworthy that the duplication is easily revealed with a symbolic scatter plot that was otherwise missed by inspecting the text of the sequence and by current automatic analysis algorithms.

## 6.2 Results

When applied to the human huntingtin gene (NC_000004:3046206-3215485 Homo sapiens chromosome 4, reference assembly, complete sequence) the symbolic scatter plot of Figure 6.1 is produced showing the first 500 3-mers. The CAG repeats in the center of this plot are visible as three horizontal lines of dots. Visual inspection of the entire gene reveals only this one region of CAG repeats.

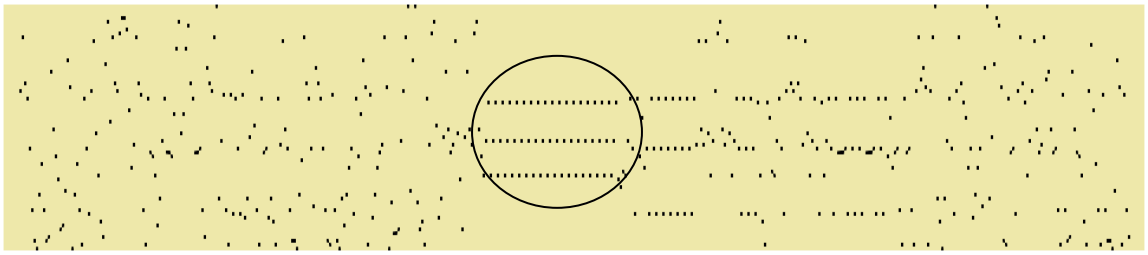


Figure 6.1: The CAG repeats are visible as three horizontal lines of dots in the center of this scatter plot for the human huntingtin gene.

Figure 6.2 shows the symbolic scatter plot for the chimpanzee huntingtin gene (NC_006471:3191346-3367586 Pan troglodytes chromosome 4, reference assembly). Again, the CAG repeats are visible as another set three parallel lines of dots. The similarity of this region to the human gene is quite apparent.

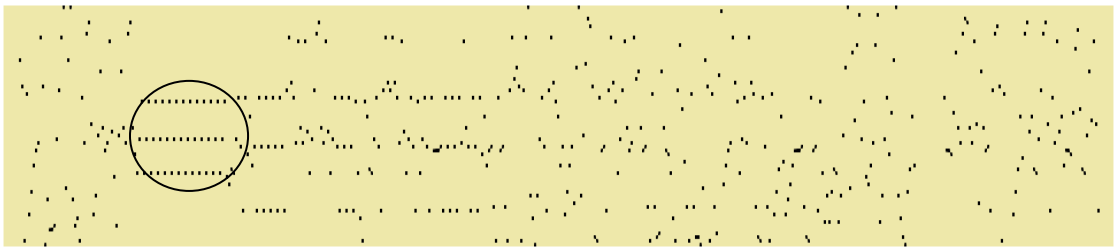


Figure 6.2: A symbolic scatter plot for the first 500 3-mers of the chimpanzee huntingtin gene. Although shorter, the CAG repeats are also evident at the left of this plot.

Again, Figure 6.2 shows only the first 500 3-mers. Scrolling to the right reveals a second region of CAG repeats as illustrated in Figure 6.3. These repeats begin at position 887 of the sequence whereas the first region of repeats begins at position 51.

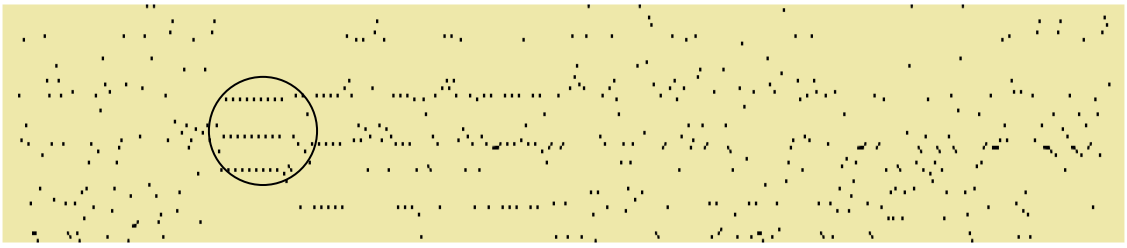


Figure 6.3: Remarkably, the chimpanzee huntingtin gene shows a second group of CAG repeats that visually is very similar to the first. Comparison of the lengths of the repeats shows that the regions do differ.

Visual inspection shows that this region is not quite the same as the first region. The most striking difference is that the number of CAG repeats is fewer than in the first region. Such duplication does not exist in the human huntingtin gene.

It is possible that the duplication is an artifact introduced by mistake into the chimpanzee genome assembly. A genome assembly is created by aligning many small overlapping pieces of DNA each about 500 nucleotides long. The sequences for these smaller pieces are stored as raw data in a database referred to as the NCBI trace archive. Note that each cell contains pairs of chromosomes with one copy of each chromosome from each parent. There is no guarantee that each copy is identical. It is quite likely that the two copies of any given gene differ slightly because the parents are not identical individuals. Thus, we might expect to find two slightly different huntingtin genes for any given individual. The trace archive will contain sequences of the pieces of the huntingtin gene

and there will possibly be two slightly different variations of some of these pieces in the archive.

For humans we would expect to find at most two pieces in the archive corresponding to the CAG repeat region of the huntingtin gene. One would come from the father and the other from the mother. If the huntingtin gene of chimpanzees also contains one region of CAG repeats, then we would also expect to find at most two sequences in the trace archive for chimpanzees. Remarkably, a search of the trace archive for the chimpanzee returns four matches suggesting that the huntingtin gene of the chimpanzee could contain two regions of CAG repeats. The symbolic scatter plots for these matches are presented in Figure 6.4.

The blue lines in Figure 6.4 highlight the CAG repeats. The top trace differs from all the others because it has a CAG repeat at the far left that is not in the other traces. The arrow in the second trace points out that the trace has more CAG repeats than the others. The arrows in the bottom two traces point to regions that differ in the two traces. The differences are subtle but easily noticeable. This visual inspection reveals that the four traces are unique and distinct from each other.

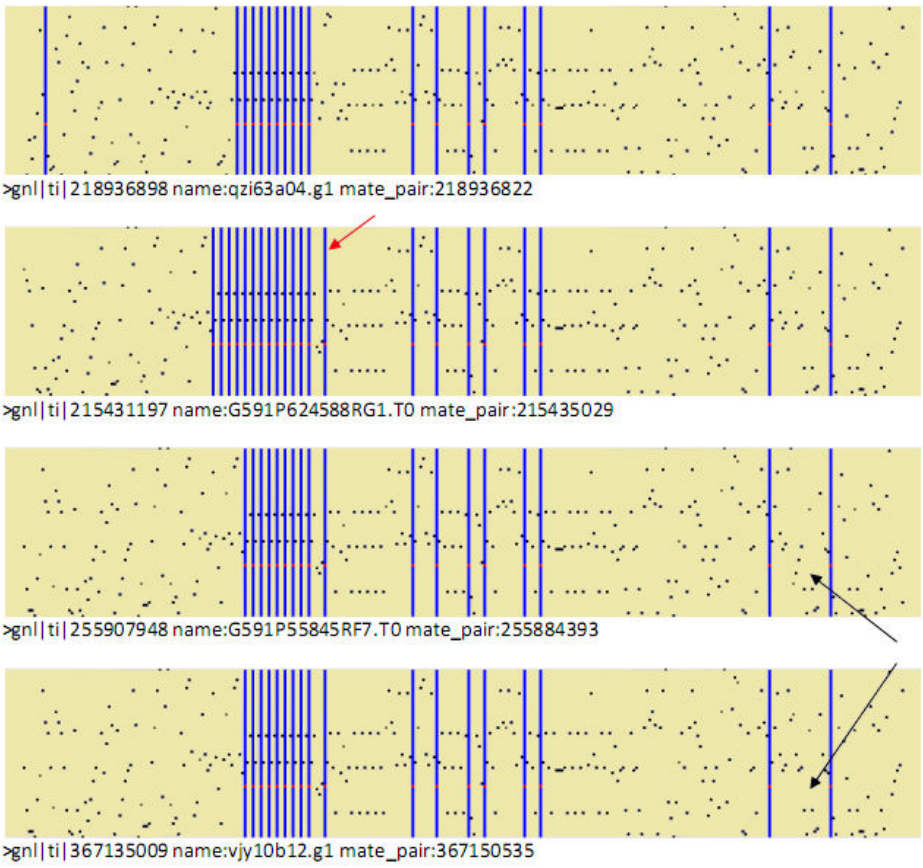



Figure 6.4: A comparison of the four sequences found in the chimpanzee trace archive. Each is different from the rest suggesting the possibility that the chimpanzee huntingtin gene might really contain two sets of CAG repeats in contrast to the human gene which contains only one.

## 6.3 Discussion

Visualizing the huntingtin gene in the genome assembly of the chimpanzee reveals that it might contain two regions of CAG repeats. There is the possibility that the genome assembly contains an artifact and that the duplication might not be real. However, a search of the chimpanzee trace archive from which the genome assembly was derived reveals four distinct sequences that closely match the huntingtin gene. If the chimpanzee huntingtin gene contains only one region of CAG repeats, then one would expect at most two trace sequences – one for each allele.

This study demonstrates the power of visualization. The chimpanzee genome assembly has been available since 2006. Yet, the presence of two CAG repeat regions in the huntingtin gene has not been reported. Nobody has *seen* the duplication. With symbolic scatter plots it was simply a matter of looking to see the duplication.

# 7 Exploiting Visual Cues

*The material in this chapter is based on a paper that has been accepted for publication in the book, "Advances in Computational Biology" to be published by Springer in 2010* (Cox, Towards a Visualization of DNA Sequences, 2010).

## 7.1 Introduction

Symbolic scatter plots use the proximity of the points in the plots to reveal patterns. Here we explore altering proximity and adding other visual cues to reveal more patterns.

## 7.2 Proximity

The human visual system uses several cues to group objects. The most fundamental of these visual cues is proximity (Kubovy, Holcombe, & Wagemans, 1998). Kubovy, et al experimented with lattices of points. A dot lattice is a collection of points in the plane that is invariant under two translations. The translation vectors, **a** and **b,** describe the parallelogram that is the basic unit of the lattice. The experiments that they performed quantified the attraction between points in the lattice. The greater the attraction, the greater the likelihood the points would be grouped into lines of particular orientations. From their experiments they were able to rule out orientation as a determining factor. The perceived line orientations in the lattices were determined entirely by the proximity of the points.

Although symbolic scatter plots are not exactly dot lattices, there are strong similarities. Thus, we most likely respond to patterns in symbolic scatter plots largely because of the proximity of points to each other.

Can we manipulate proximity to reveal more patterns or to enhance those already present in the plots? There are two ways to manipulate proximity in these plots. One is to vary the x-coordinate of the points and the other is to manipulate the y-coordinates. It will be shown that bringing points into closer proximity allows the visual system to respond to patterns that were not otherwise apparent. It will be shown that this technique is useful for finding motifs within DNA sequences – an important task commonly performed in bioinformatics.

### 7.2.1 Manipulating Horizontal Proximity

The Gestalt law of proximity tells us that we tend to perceive patterns from objects that are near each other rather than far apart. By relaxing the rule that a single point of a symbolic scatter plot occupies a column we can move the points in an image closer to each other. The effect is to "squish" the points horizontally.

Figure 7.1 illustrates the result. The top image is an original symbolic scatter plot that appears to display a random array of points. In the subsequent scatter plots, more than one 3-mer is plotted in a column. The second panel of Figure 7.1 plots two points per column. The bottom two panels plot four and eight points per column respectively. If there are copies of a 3-mer in a group, a single point in the column represents all copies. While this does represent a loss of information, the closer proximity of the entire set of points does allow one to see patterns that would not otherwise be visible. Repetitive patterns not obvious in the first panel become much more noticeable in the subsequent panels. The

bottom panel even shows how one pattern ends abruptly with a random looking region of points followed by another more orderly region.

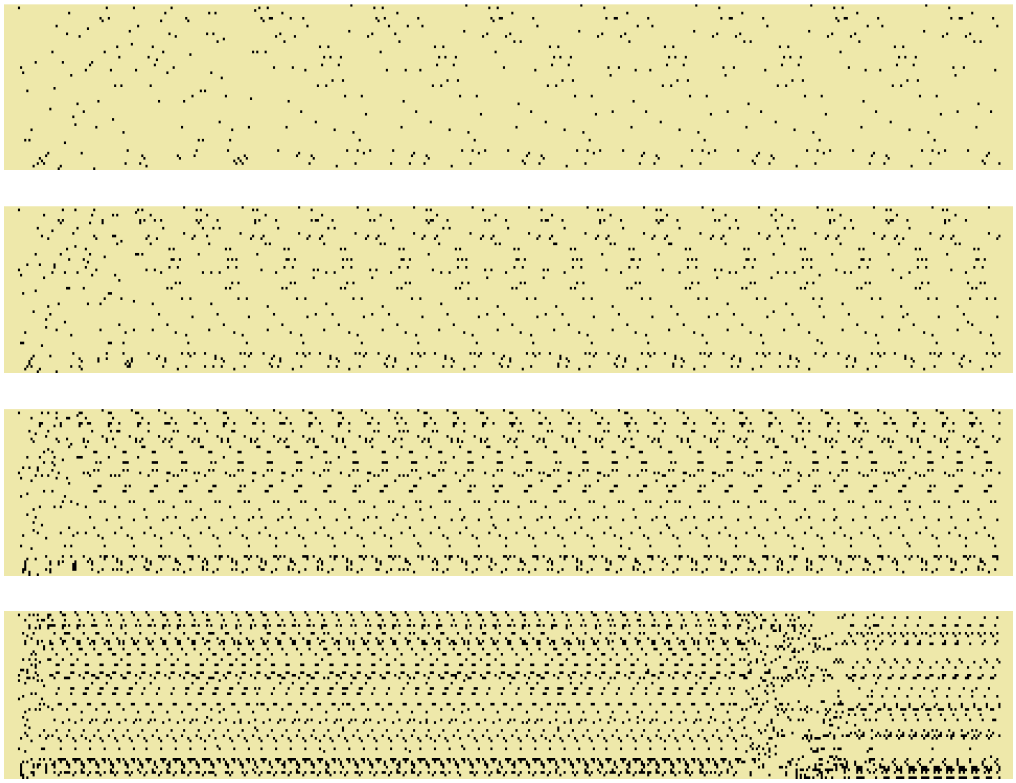


Figure 7.1: Manipulating visual cues can render patterns more visible. Here proximity is manipulated by bringing the points closer together horizontally. This "squishing" of the points transforms an apparent random display of points into a highly ordered array ordered array of 3-mers.

This example and others not shown suggest that bringing points closer together horizontally allows one to see patterns that would not otherwise be apparent. Bringing points closer together also allows one to examine much longer stretches of DNA. A typical computer monitor will display about 1,000 pixels across. By compressing the image horizontally symbolic scatter plots can represent considerably longer sequences. In Figure 7.1 the symbolic scatter plots were magnified to where each point occupies 3x3 pixels. Uncompressed, the plot represents 484 3-mers or 1452 nucleotides. Compressed by a

factor of 8, the bottom panel represents 11,616 nucleotides. By wrapping the scatter plot a single computer monitor one could easily visualize a sequence of 50,000 nucleotides.

Compressing a scatter plot in this fashion allows us to exploit our tendency to group objects that are far apart by bringing them into closer proximity. Before compression two identical 3-mers would occupy the same row of the symbolic scatter plot but in different columns. Thus, the two points would be distinguishable. After compression, the points continue to occupy the same row but now occupy the same column. When rendered, the 3-mers appear as only a single point is. This loss of information is a trade-off for the ability to see longer range relationships in the data.

### 7.2.2 Manipulating Vertical Proximity

The x-value for a point in a symbolic scatter plot corresponds to the position on an underlying 3-mer in the sequence. The y-value is not assigned according to any particular criteria. The only constraint is that each of the 64 possible 3-mers is assigned a distinct value. Thus, for a given plot all TTT's are assigned some value such as 0. However, for another plot, the TTT's could be assigned a different value.

There are 64! different ways to map 3-mers to the y-axis and there is no rule that states that one mapping is better than another. We are free to choose any mapping that suits a particular task. In these experiments the 3-mers of a user specified string are remapped to bring them into close proximity. As in section 7.1.1 the goal is to exploit proximity to accomplish a particular task.

A common task in bioinformatics is to find motifs. A motif is a short sequence often only about ten nucleotides long. Motifs are thought to be biologically important and are found by comparing several sequences. Several sequences might be very different but will have short spans of DNA that are either identical or very similar to each other. Usually it is the latter case because mutations will have occurred to cause sequences to diverge over the years. Those spans of DNA that remain most similar are thought to have been preserved by evolution because they serve an important biological function. Even though a particular span of DNA might have been preserved by evolution, there might have been some mutations that would render the span slightly different in different sequences.

Motifs are usually specified using a frequency profile. For example, after comparing several sequences a span is found where the nucleotides occur with the following frequencies:

```
A  [50 49 53 18  1 93 98  0  0 53  6  3 25 38 56 31 ]
C  [26 24  6 30 46  5  1  0  2  0 39 39 21 23  8 14 ]
G  [ 9 12 39 44  0  2  0  1  5 46 40  6 30 16 15  8 ]
T  [14 16  3  8 53  0  0 98 93  1 15 53 25 23 21 46 ]
     A  A  A  G  T  A  A  T  T  A  G  T  G  A  A  T
```

Figure 7.2: Example of a frequency profile often used for finding motifs.

The sequence, AAAGTAATTAGTGAAT, contains the nucleotides that occur most frequently and represents the consensus sequence for the motif. If this motif is biologically interesting, we can divide it into overlapping 3-mers and assign the 3-mers consecutive y-values in a symbolic scatter plot. The expectation is that if the motif is present in a DNA sequence, it will be visible as a diagonal line of points in the plot. Using the Gestalt

principle that we tend to group items that are in close proximity, we should be able to easily perceive such clusters in the plot.

To illustrate, the top panel of Figure 7.3 shows an apparently random set of points. A closer look does show some repetition in the plot. Using the motif AGATAGAGAGCAG we can re-map the y-values for the 3-mers AGA, GAT, ATA, TAG, GAG, AGC, and CAG to force them to occupy the center of the plot. The remaining 3-mers are mapped to other y-values in no particular order. The bottom panel of Figure 7.3 shows the result. The middle region that these 3-mers occupy is highlighted to enhance their visibility. The diagonal clusters in this highlighted area indicate approximate matches of this motif at several locations in the sequence. By clicking on a diagonal one can immediately see the exact sequence and can determine to what degree it matches the provided motif.

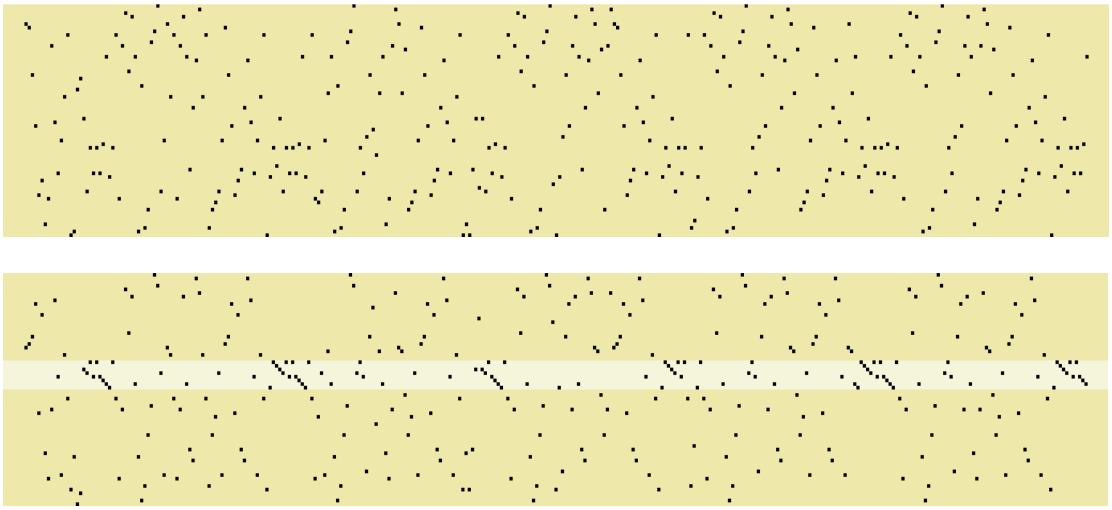


Figure 7.3: Manipulation of vertical proximity can reveal specific sequences including those that do not match exactly.

This technique is also useful for remapping sequences within a DNA sequence and not just motifs. Consider the next example in Figure 7.4. By clicking on the plot we can examine the sequence at that point. Clicking about a third of the way from the left reveals that the sequence at that point begins with ATGGGGTCCCTTTCCATA. Dividing this string into overlapping 3-mers and remapping the 3-mers to force them to occupy the middle of the plot gives the result in the middle panel of Figure 7.4. The long diagonal corresponds to this sequence. Its jagged appearance is because there are two GGG's in the string and two TCC's. The first TCC in the string determines the 3-mer's y-value. The second TCC in the string maps to the same y-value as the first causing the jagged appearance.

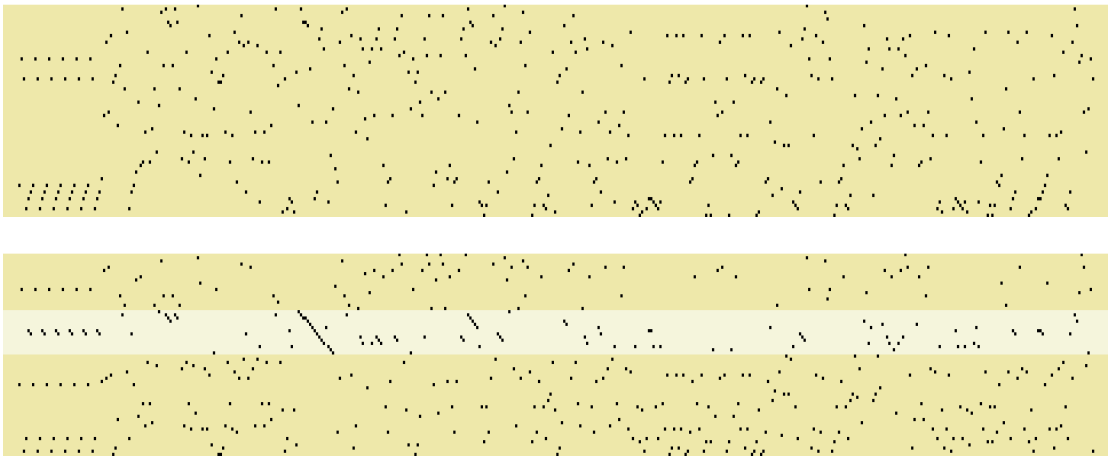


Figure 7.4: Manipulation of vertical proximity reveals specific sequences.

To the right is a shorter diagonal consisting of five points. Examining this string reveals that it matches a portion of string used for the remapping. Here is the original string

along with shorter string showing how they match. Notice that the matching string has a deletion and is missing a G:

Original string: ATGGGGTCCCTTTCCATA

Matching string: TTGGGTCCACACTGCCTT

Scrolling through the plot one can easily find other matches of varying length. The longer the length, the more likely the match is statistically significant. For example, another match is presented in Figure 7.5. The two matching regions are the same length. However, they differ by one nucleotide.

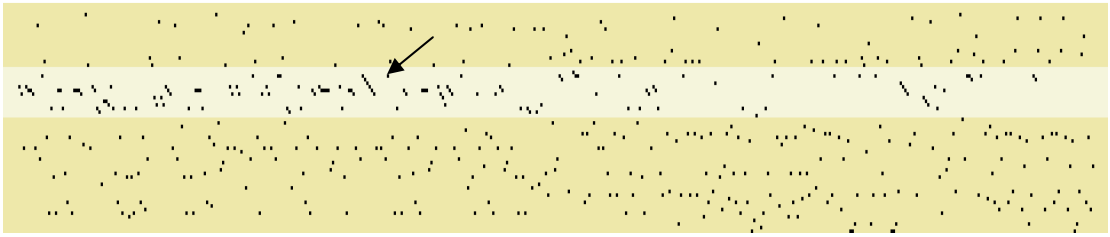


Figure 7.5: An example of an approximate match revealed by changing the vertical proximity of 3-mers.

Original string: ATGGGGTCCCTTTCCATA

Matching string: CGGGTCCCTCTCGGGA

Are these matches significant? In this example the lengths of the matching strings are 11 nucleotides long. The chance that ten of these nucleotides match is less than one in one million.

One last example in Figure 7.6 is quite interesting. Scrolling further in the symbolic scatter plot, one comes across another matching string. Again the matching strings are 11

nucleotides long and they match exactly. This time the chance of them matching is less than one in four million.

Original string: ATGGGGTCCCTTTCCATA

Matching string: ATGGGGTCCCTGCGAGA

What makes this example interesting is the region of the scatter plot to the left of the matching string. Like the original string, the matching string is preceded by a sequence of repeats. Perhaps this is a coincidence. However, it raises suspicions that the pattern has some important biological function.

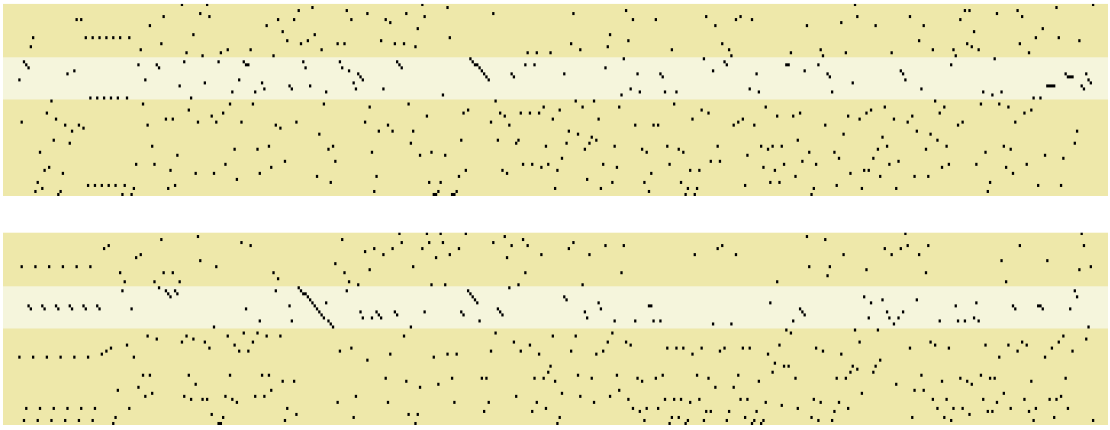


Figure 7.6: Co-location of a feature near another feature can raise suspicion that the feature has some biological function.

### 7.2.3 Discussion

When symbolic scatter plots consist of widely dispersed points the eye has difficulty distinguishing patterns in them from a random background. The visual cue, proximity, works against us in these cases. However, by bringing points closer together in a prescribed way allows proximity to work for us rather than against us.

In this chapter we explored changing the horizontal and vertical proximity of points in a scatter plot. Horizontally, points were translated allowing more than one to occupy a single column. This compression allowed a much larger DNA sequence to be viewed at one time and allowed the eye to see patterns that were missed because the points were too far apart.

On the vertical, the y-values of points were re-mapped to force 3-mers of interest to occupy consecutive positions. An application for re-mapping y-values is the motif finding problem. By re-mapping the 3-mers in a motif it was demonstrated how the motif becomes conspicuously visible as a diagonal arrangement of points in the center of a symbolic scatter. Because the eye can distinguish both perfect and imperfect diagonals, it is possible to locate both exact and inexact matches to a motif. By clicking on the scatter plot it is easy to see the underlying sequence and to assess how closely the sequence matches the motif.

## 7.3 Animation

Animation is not limited to entertainment. Examples abound demonstrating the usefulness of animation in data analysis. Animations are created by variations in visual parameters such as position, shape, size, color, or orientation. A classic example is the variation in all of these parameters to represent blood flow in medical imaging. Other examples include animations of weather and seismic data.

Within molecular biology and genomics, animation has improved students' achievements in the study of molecular genetics (Marbach-Ad, Rotbain, & Stavy, 2008). However, animation is not limited to teaching. It also conveys scientific information about molecular structures as surveyed by McGill (McGill, 2008). It describes the time course of microarray data (Ishiwata, Morioka, Ogishima, & Tanaka, 2009). It illustrates the movement of a DNA strand through a nanopore in the presence of an electric field (Sigalov, Comer, Timp, & Aksimentiev, 2008).

For sequence analysis Seevolution (Esteban-Marcos, Darling, & Ragan, 2009) animates sequence inversions, transpositions, insertions, deletions, and substitutions. Seevolution relies on other software and algorithms to identify these events and then integrates them into a single animation.

Smoot et al (Smoot, Guerlain, & Pearson, 2004) simultaneously display multiple near-optimal alignment paths through a scoring matrix and use color transitions to cycle through the different paths.

Santo and Dimitrova (Santo & Dimitrova, 2007) animate spectrograms to analyze the harmonic properties of genomic sequences. These "SpectroVideos" require enormous computational resources and are difficult to interpret and correlate to specific sequence locations and functions.

Other tools such as Jalview (Clamp, Cuff, Searle, & Barton, 2004), Seaview (Galtier, Gouy, & Gautier, 1996), ClustalX (Larkin, et al., 2007), Fingerprint (Lou & Golding, 2007), and Chroma (Goodstadt & Ponting, 2001) depict sequence alignments. All except Fingerprint present aligned sequences using columns of text with coloring to highlight similarities and differences. Fingerprint presents alignments as colored barcodes with the lines representing different characteristics ranging from identification of consensus residues to hydrophobicity (i.e. attraction or repulsion to water molecules). None employs animation.

This section presents a method to animate sequence alignments using symbolic scatter plots. Mutational events (insertions, deletions, and substitutions) are easily recognized in these animations. They not only show the locations of mutations but also show patterns (such as repeats) associated with each sequence's nucleotides and how those patterns change from sequence to sequence. These patterns provide additional information with which to judge sequence alignments and to make improvements to them.

### 7.3.1 Method

Sequence alignments reveal the substitutions, insertions, and deletions (i.e. mutations) needed to convert one DNA sequence into another. The output of sequence

alignment software is universally textual and uses a five character alphabet {A, C, G, T, -}. A typical alignment is shown in Figure 7.7.

```
AAAAAAGAAATGAAGTTCTCTTGGTCACATCCTAAAAGTGACCAGCTCCC
.|||||     |||||||||||||||||.|||||||||||||||||||||
TAAAAA-----GAAGTTCTCTTGGTCACGTCCTAAAAGTGACCAGCTCCC
```

Figure 7.7: Typical textual representation of a sequence alignment

The vertical lines indicate nucleotides that match and the dots indicate substitutions. The dashes indicate nucleotides that would have to be deleted from the top sequence allowing it to match the bottom sequence.

For sequence alignments, symbolic scatter plots are easily extended. If a dash (-) is encountered in a 3-mer, then an empty column is inserted in the scatter plot. Consider the following sequence: AAT - ATC. A point is plotted for AAT. However, a point is not plotted for AT -, T - A, or - AT. The columns for these 3-mers are skipped to leave them blank. Notice that there is no loss of information about the underlying sequence because the points resume with ATC. When applied to an alignment the result is similar to that in Figure 7.8.

```
AAAGAAATGAA       AAA AAG AGA GAA AAA AAT ATG TGA GAA
|||     ||| ==>   AAA                           GAA
AAA-----GAA
```

Figure 7.8:An alignment of overlapping 3-mers.

Each of the codons for the top sequence is mapped to a point in a symbolic scatter plot. The codons in the bottom sequence are mapped to a second symbolic scatter plot. Where there are no codons, gaps will appear in the symbolic scatter plot. By cycling between the two scatter plots an animation is created.

### 7.3.2 Results

Animated symbolic scatter plots contain some points that appear stationary and others that appear to move. Some movements can be predicted. For example, a single substitution will cause three points to move. The substitution transforms the overlapping 3-mer to the left, the 3-mer in the middle and the 3-mer to the right. Two adjacent mutations will cause four points to move, etc. An insertion or deletion produces gaps causing a region of points to disappear and then reappear.

Figure 7.9 illustrates a single substitution for the two sequences: AAAAA and AAGAA. The table lists the overlapping 3-mers for the two sequences. The narrow strips depict the points corresponding to the 3-mers. A single substitution transforms all three overlapping 3-mers. Both AAA and AAG map to lysine and intentionally map to points that are neighbors. AGA, however, maps to arginine and GAA maps to glutamine. Consequently, those points are intentionally positioned further away.

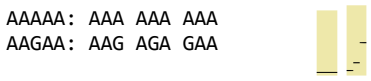



Figure 7.9: Correspondence between an alignment and a symbolic scatter plot

Figure 7.10 illustrates portions of the insulin genes for chimpanzees (top panel) and humans (bottom panel).

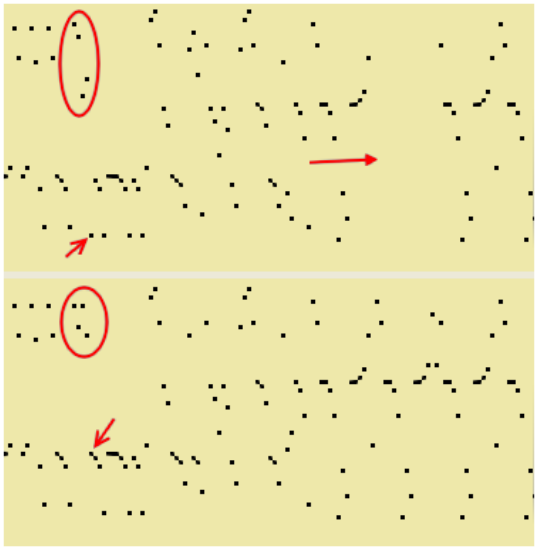


Figure 7.10: Subtle and not so subtle differences between symbolic scatter plots become noticeable with animation.

An area where points are deleted is indicated by the right-most arrow. Cycling through these two panels shows the points disappearing and then reappearing. At the same time, the points highlighted by the circles and the left-most arrows also differ and become animated when cycling between the two images.

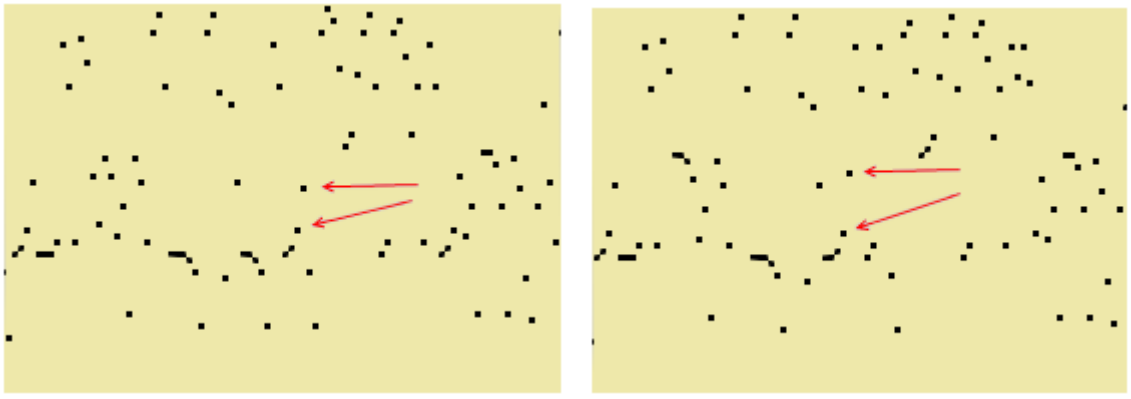


Figure 7.11: Subtle differences between the chimpanzee (left) and human (right) huntingtin genes are much more noticeable with animation.

Figure 7.11 (above) illustrates another example using the huntingtin genes for chimpanzees (left panel) and humans (right panel). The arrows refer to a pair of 3-mers that are seen in the animation to oscillate from right to left.

Sequences that contain large numbers of repeats can be difficult to align statistically. The following is a sequence alignment produced by FASTA. The alignment appears valid given the large number of matches. However, it is not valid.

```
GCCTCCGGGGACTGCCGTGCCGGGCGGGAGACCGCCATGGCGACCCTGGA
||||||||||||||||||||||||||||||||||||||||||||||||||
GCCTCCGGGGACTGCCGTGCCGGGCGGGAGACCGCCATGGCGACCCTGGA

AAAGCTGATGAAGGCCTTCGAGTCCCTCAAGTCCTTCCAGCAGCAGCAGC
|||||||||||||||||||||||||||||||||||
AAAGCTGATGAAGGCCTTCGAGTCCCTCAAGTCCTTC-------------

AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAG
                  |||||||||||||||||||||||||||||||||
----------------CAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAG

CCGCCACCGCCGCCGCCGCCGCCGCCTCCTCAGCTTCCTCAGCCGCC
||||||   |||||||||||||||.||||||||||||||||||||||
CCGCCA---CCGCCGCCGCCGCCGCTGCCTCCTCAGCTTCCTCAGCCGCC

GCCGCAGGCACAGCCGCTGCTGCCTCAGCCGCAGCCGCCCCCGCCGCCGC
||||||||||||||||||||||||||||||||||||||||||||||||||
GCCGCAGGCACAGCCGCTGCTGCCTCAGCCGCAGCCGCCCCCGCCGCCGC
```

When visualized using symbolic scatter plots and animated there is a horizontal shift of the right halves of the two sequences. This shift is apparent in Figure 7.12. The blue lines highlight the positions of all GCC's. These lines illustrate that the right halves of the sequences are actually identical. Additional insertions to the right of center bring these two regions into alignment.

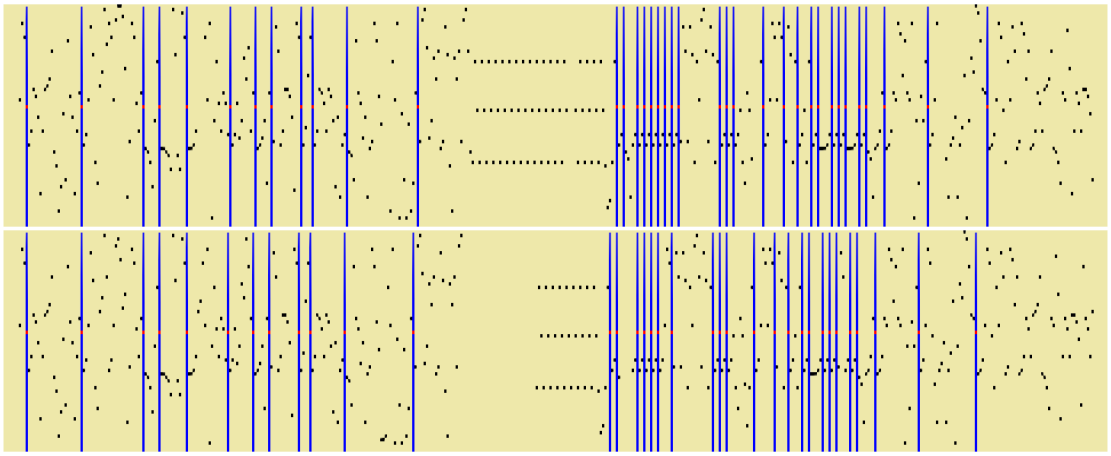


Figure 7.12: Shifting noticed in an animation can reveal misaligned sequences.

This example illustrates that the technique is particularly good at helping to align sequences with repeats. Statistically two sequences with repeats can appear to be well

aligned but in reality are shifted and out of phase with each other. Visualizing them allows one to immediately recognize such phase shifts.

### 7.3.3 Discussion

Symbolic scatter plots can be modified to account for the insertions, deletions, and substitutions necessary to produce aligned sequences. By cycling through a set of these scatter plots animations are produced. These animations exploit our visual system's preprocessing of motion. Even very small movements of a single point are immediately noticeable. The motions of specific points give an indication the type and number of mutations and whether or not those mutations might result in amino acid changes. Large changes in position hint of amino acid changes while small changes hint that the mutations did not result in amino acid changes. Seeing these motions can help the biologist decide which parts of sequences to focus further research on. As was demonstrated with the huntingtin gene repeats in a sequence can result in several alignments with very similar statistics. Animation highlights these differences in a symbolic scatter plot and can help the biologist choose among these variations.

## 7.4 Color

Color is pre-attentively processed. An important application of color is to indicate categories of information (Ware, 2004). Other applications include emphasizing or highlighting certain features. Color helps people to perceive the spatial layout of patterns in their data and can be used to represent quantitative values. Here color is used to selectively highlight 3-mers in a symbolic scatter plot and to represent entropy values.

### 7.4.1 Feature Highlighting

Color is useful for highlighting specific features and two uses of color have been explored with symbolic scatter plots. The first use is simple but very effective. Color is used to highlight specific 3-mers within a symbolic scatter plot. Figure 7.13 provides an example.

The blue lines tell us where in the sequence specific 3-mers are located. They are obviously useful for pinpointing specific 3-mers in the sequence. However, these lines also form patterns in their own right. The top panel of Figure 7.13 presents an unremarkable symbolic scatter plot for a portion of the human Y-chromosome. The bottom panel highlights TCT's present in the sequence. The proximity of the blue lines illustrates a pattern of alternating pairs of TCT's

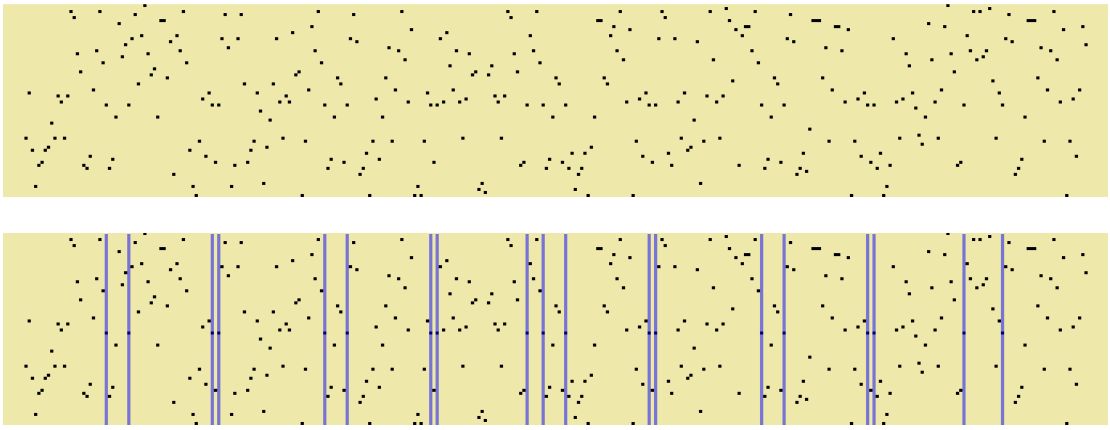


Figure 7.13: Color is useful for highlighting specific features. Here, blue lines are used to highlight TCT's to reveal an interesting pattern of pairs of 3-mers.

A report by Tandem Repeats Finder reveals a repeat of length 68 and shows how the consensus sequence aligns with consecutive segments of the sequence. Figure 7.14 presents two of these alignments. The top row contains the nucleotides from the sequence and the bottom row contains the nucleotides from the consensus sequence. The asterisks indicate mismatches between the two sequences. The report states that there are 14.2 copies of this repeating unit in the sequence.

```
        *                 *             *       *        *   *
  1 TCTTGCCTAGGCTCTGCCTACATGGGCATTGTGACACATCTCTGAACTGATCAACCAAGTGATGT
  1 TCTTGTCTAGGCTCTGCCTACAGGGGCATTGTGACATATCTCTGCACTGATCACCCAGGTGATGT

    *                       *  *                                    *
 69 CCTTGTCTAGGCTCTGCCTACAGGAGCTTTGTGACATATCTCTGCACTGATCACCCAGGTGATGG
  1 TCTTGTCTAGGCTCTGCCTACAGGGGCATTGTGACATATCTCTGCACTGATCACCCAGGTGATGT

     *                      *                    *  * *    *
137 TTTTGTCTAGGCTCTGCCTACAGTGGCATTGTGACATATCTCTACAGTAATCAACCAGGTGATGT
  1 TCTTGTCTAGGCTCTGCCTACAGGGGCATTGTGACATATCTCTGCACTGATCACCCAGGTGATGT
```

Figure 7.14: Algorithms that report tandem repeats often present alternatives with very similar scores. Is this repeat of length 68 and a score of 994 better or worse than one of length 205 and score 957? Visualization provides additional information to help answer such questions.

Tandem Repeats Finder also reports a much longer alternative consensus sequence of length 205. This latter result is given a score of 957 while the former result was given a score of 994. It is left to the user to judge between the scores and the textual output of the report which might be better.

In contrast, exploring the same region of sequence with a symbolic scatter plot and a few mouse clicks reveals repetitive patterns that easily indicate that the repeat length is shorter rather than longer. The spacing of the blue lines also indicates at a glance how these repeats could be aligned by using the pairs of TCT's as markers.

### 7.4.2 Visualizing Additional Dimensions of Information

Color can also be used to add information to a visualization. From a computer science perspective a DNA sequence is a string of characters. Like any string of characters it can be encoded using some number of bits. Shannon entropy tells us how many bits we can use. In these experiments a window containing 64 3-mers is used to calculate Shannon entropy. A length of 64 is used because there are 64 possible 3-mers. If each possible 3-mer were in the window, then a full 64 bits would be needed to encode the sequence contained in the window. Fewer than 64 distinct 3-mers would require some number less than 64 bits. The entropy is assigned to the 3-mer located in the center of the window (specifically, the 32$^{nd}$ 3-mer). This entropy is then used to color the 3-mer at that location in the sequence.

Figure 7.15 illustrates the result. The top panel shows a symbolic scatter plot without color to highlight entropy. The bottom panel shows the same plot with

highlighting. A linear color gradient extending from pure green for the highest entropy to pure red for the lowest entropy is applied. The points themselves are not colored. Instead the color "bleeds" vertically away from each point. Thus, the color not only highlights each point but also follows the positioning of the points in the scatter plot. This is both aesthetically appealing and informative.

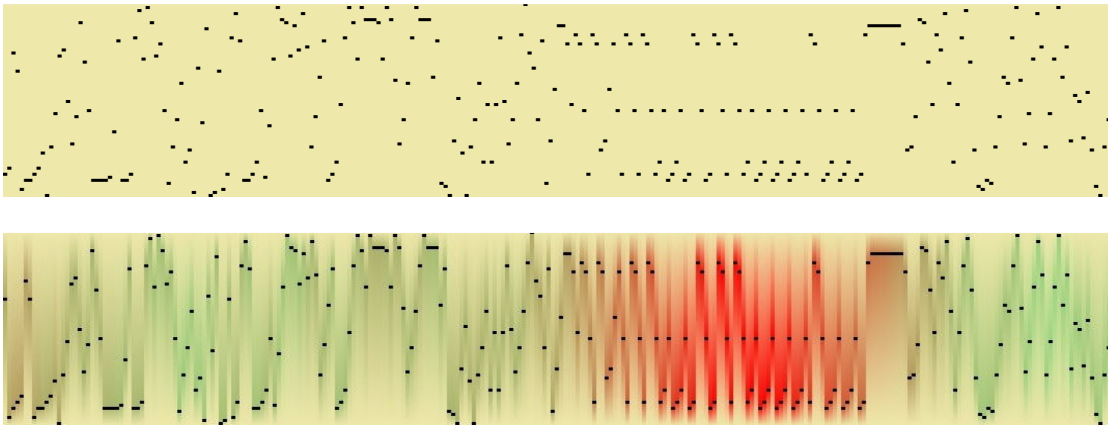


Figure 7.15: Example of color to emphasize the entropy of a DNA sequence while retaining the distribution of 3-mers. Red indicates low entropy while green represents high entropy.

The bright red area in Figure 7.15 highlights an area of low entropy which corresponds to a visually ordered region in the symbolic scatter plot. However, it is not always obvious which areas in a symbolic scatter plot have higher or lower entropy. Figure 7.16 presents another example where the difference is not so clear. The highlighted version of the plot does a better job of indicating the different entropies than the version that is not highlighted.

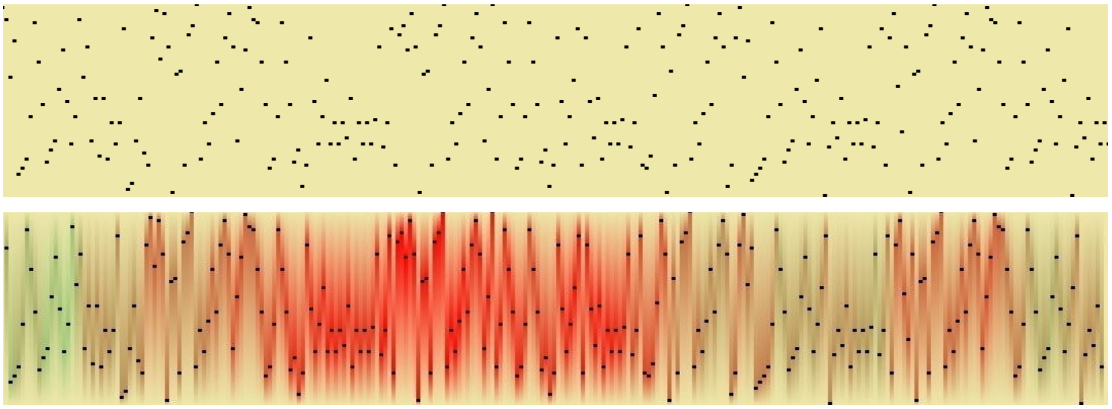


Figure 7.16: Regions of high and low entropy are not always obvious. Here is another example where color helps make these differences more obvious.

### 7.4.3 Discussion

Color is a useful visual cue for highlighting specific features. Here we explored using color to highlight the entropy of a DNA sequence when the sequence is rendered as a symbolic scatter plot. Allowing the color to "bleed" away from the individual points of a scatter plot preserved the scatter plot and allowed the color to both highlight differences in entropy and to follow the positioning of the points in the plot. There are many other physical measures associated with DNA sequences beyond entropy. These include quantities that indicate the shape of DNA and it chemical properties. Although not attempted here, it is expected that color would be a useful tool to convey any of these quantities.

## 7.5 Summary

This research is the first to evaluate a technique for visualizing DNA sequences in terms of the science of human visual perception. Thousands of researchers have conducted experiments and developed a body of theory about how we respond to and organize visual stimuli. Although the research presented in this chapter does not expand on those theories, it does use them to explain the patterns in symbolic scatter plots. Furthermore, this research demonstrates how to exploit those theories to achieve specific goals in analyzing DNA sequences. We believe that this research is the first to do so.

To summarize, we can extend a visualization technique by understanding the basic principles involved. For symbolic scatter plots a key principle was the visual system's preference to group objects that are in close proximity. By exploiting the horizontal

proximity of points in a scatter plot, it is possible to see some large scale patterns in DNA sequences. By exploiting the vertical proximity of points, it is possible to visualize specific motifs and patterns similar to those motifs.

Another pre-attentive visual cue that can be exploited is animation. A simple modification of symbolic scatter plots permits the visualization of aligned sequences. By cycling through these scatter plots, regions that differ become animated allowing the user to easily see both differences and similarities in the sequences. These animations allow the user to quickly scan large amount of aligned sequences to explore those regions that look the most interesting.

Color is useful for highlighting specific features. With it one can see complex patterns associated with specific 3-mers. One can also use color to highlight specific properties of DNA as a way of visualizing multiple dimensions of information at the same time.

The key to successfully applying visualization to any data analysis requires understanding how our brains interpret visual information, understanding the analytical goals, and then tailoring our visualizations to achieve those goals.

# 8 Conclusion

This research began with an observation that a novel type of scatter plot could present interesting visual patterns in sequential data. This ability was demonstrated with the sieve of Eratosthenes and with DNA sequences. In the case of the sieve of Eratosthenes interesting linear and parabolic relationships between the divisors of natural numbers were revealed. While the patterns displayed for DNA sequences were interesting, it remained to determine why the patterns arise, what they represent, and if the patterns are useful.

The literature on human visual perception was reviewed. The patterns recognized in these "symbolic" scatter plots were explained in the context of what is currently known and accepted about visual perception. No new theory regarding visual perception was presented. However, existing theory was used to help understand how symbolic scatter plots could be used to perform several common bioinformatics tasks.

An important task in bioinformatics is the identification of tandem repeats. Points arranged linearly into horizontal and diagonal lines are pre-attentively identified by our visual system. Horizontal and diagonal arrangements of points are prominent in symbolic scatter plots and are easily recognized. Horizontal arrangements of points correspond to repeats of 3-mers. Experiments with Tandem Repeats Finder demonstrated that many of these features correspond to simple tandem repeats. However, these experiments also demonstrated that many other features correspond to considerably more complex patterns including hierarchically arranged repeats. In many cases patterns of repeats were clearly evident in symbolic scatter plots but were either partially reported by Tandem Repeats

Finder or not reported at all. In a large number of cases visually identifying tandem repeats and complex patterns was superior to the heuristic approach of Tandem Repeats Finder.

Another important task in bioinformatics is searching for motifs in DNA sequences. Traditionally this is done with statistical or machine learning techniques. Here we wanted to determine if something similar could be done visually by exploiting how we respond to proximity. Things that are near to each other appear related to each other. Those that are far away appear distinct from each other. By reducing the vertical proximity of points corresponding to 3-mers in a motif, we were able to represent the motif as an easily recognizable diagonal line in a symbolic scatter plot. Inexact matches were identified by broken but still recognizable diagonal lines. Thus, by exploiting how we respond to proximity, we were able to create a simple visual representation that could be used to find both exact and inexact matches of a motif.

Long range patterns in DNA sequences can be difficult to identify and one region of DNA can seem very similar to another. By reducing the horizontal proximity of points in a symbolic scatter plot, one can view much larger sequences in a single computer display. Moreover, patterns that were not apparent in the original image popped out as the proximity between the points was decreased. Regions that originally were not distinguishable became visually distinct often with very distinct boundaries.

Comparing DNA sequences is another vital task of DNA sequence analysis. This research investigated the utility of comparing sequences with the help of color and animation in symbolic scatter plots. Highlighting a specific 3-mer by coloring the

corresponding columns of a scatter plot permits aligning sequences by how the 3-mer is distributed in two or more sequences. Color combined with conspicuous groups of points allows sequences to be aligned by comparing visual features. When several traditional alignment algorithms produce different results, visualization by symbolic scatter plots helps to choose a correct result.

Sequences were also compared by combining their symbolic scatter plots into an animation. In this case the animation consists of motion – i.e., apparent changes in the positions of points as the animation cycles between different plots. Motion is a visual cue that is processed pre-attentively. As expected it immediately draws the attention to those regions where the points are moving. These movements are useful to understand the degree of difference between two sequences and also to understand their degree of similarity. Points that do not move are made conspicuous by those that do. If these stationary regions contain visually interesting features that perhaps are linear, repetitive, or highly symmetrical, then the viewer gains confidence in the alignment of the sequences. Indeed, some alignments of highly repetitive regions might be statistically correct but are misaligned even slightly when examined visually. Visualization helps the observer to immediately see such misalignments which would be difficult to detect from tabular data.

This research also investigated using color to visualize multiple dimensions of information. Color blended with symbolic scatter plots was used to highlight regions of high and low entropy within DNA sequences. The color helps to associate different patterns of points in the plots with this additional information. In many cases the color

verifies that what we perceive as ordered patterns of points correspond to low entropy (and vice versa). In other cases when it is hard to distinguish regions of high and low entropy from the points alone, color helps to differentiate those regions.

Throughout the course of this work many real DNA sequences were examined. Perhaps the most interesting of these was the huntingtin gene that is responsible for the incurable Huntington's disease that gradually destroys a victim's brain. The huntingtin gene is interesting because it is characterized by repeats of the CAG 3-mer and the disease is highly correlated with a large number of repeats. The human huntingtin gene contains one region of these CAG repeats. Examination of the huntingtin gene for the chimpanzee immediately revealed two regions of CAG repeats – a fact that had gone completely unnoticed despite a healthy community of medical researchers working to find a cure. In 2007 – 2008 the Huntington's Disease Society of American alone spent nearly $2.5M on research. Nevertheless, although the DNA sequence for the chimpanzee huntingtin gene has been available since 2006, as explained in a private communication with Dr. Albert La Spada, a specialist in Huntington's disease in the Department of Pediatrics at UCSD (LaSpada, 2009), not a single researcher has reported the presence of two regions of CAG repeats in the chimpanzee gene. Dr. La Spada was completely unaware of the duplication and found the result fascinating. Such a significant difference in the huntingtin gene of our closest relative, the chimpanzee, could help to explain the disease and provide insights for a possible cure. At this time it is unclear if the duplication observed in the sequence really exists in the chimpanzee genome or if it represents an error in the chimpanzee genome

assembly. Recently, the chimpanzee genome has been sequenced with greater accuracy. When those results are available it will be interesting to see if the duplication is still present. Despite the result, it is clear that a symbolic scatter plot of the chimpanzee huntingtin gene easily made the duplication visible suggesting that there is utility for people visualize and look at DNA sequences and not rely too heavily on automatic analysis.

In summary, symbolic scatter plots are a novel visualization of DNA sequences. They can be used to visualize a wide variety of tandem repeats and complex patterns in DNA sequences. When repeats are too long to be recognized, we can take advantage of how we pre-attentively process visual information to alter symbolic scatter plots to make those patterns "pop out." Similarly, we can adjust symbolic scatter plots to reveal motifs that would otherwise be hidden from view.

Symbolic scatter plots can be modified to accommodate insertions, deletions, and substitutions of nucleotides that are introduced when aligning sequences. With this modification, symbolic scatter plots can be used to evaluate sequence alignments. Rather than relying solely on a traditional cost-minimization of string changes, biologists can compare higher level features that are easily recognizable in symbolic scatter plots. It is suggested that comparison of these high level features is a possible way of solving the problem that different alignment algorithms produce considerably different results.

Perhaps the greatest contribution of this research was to place the visualization of DNA sequences in the context of the science behind human visualization. All previous attempts to visualize DNA sequences never considered why we see what we see. By

manipulating or adding visual cues it was demonstrated that additional information could be obtained from symbolic scatter plots. Because there are many more visual cues to consider, there is certainly much more work to be done in this area.

In summary, the primary contributions of this research are:

- A novel method for visualizing the sieve of Eratosthenes and its adaptation for visualizing DNA sequences.
- A proposed explanation of the visualization method in terms of Gestalt theory and pre-attentive processing.
- How to manipulate visual cues to produce visualizations that achieve specific purposes such as finding motifs and finding long range patterns in DNA.
- How to add pre-attentively processed visual cues such as animation and color to reveal additional information including patterns of entropy and distribution of specific 3-mers.
- How to use visualization to evaluate DNA sequence alignments. In particular, it was demonstrated how visualization can be used to align sequences based on conspicuously visible features which has not been achieved with current methods. It was also demonstrated how animation can be used to evaluate alignments of DNA sequences.
- How to use symbolic scatter plots to find complex repetitive patterns in DNA sequences that can be used in conjunction with tools such as Tandem Repeats Finder. Symbolic scatter plots can be used to evaluate the results of Tandem Repeats Finder and can be used to find patterns that are missed by Tandem Repeats Finder.

- How to use visualization to quickly examine large amounts of DNA sequence data. By taking advantage of pre-attentive processing, it is possible to find patterns in as little as 250 ms in sequences with several thousands of nucleotides using a standard computer monitor. Although a "manual" process, sequences containing hundreds of thousands or millions of nucleotides could be quickly scanned by a small number of people within minutes using several monitors such as large, high resolution displays found in control rooms.

# 9 Future Research

There are many avenues for future research. One avenue is to explore in more detail the psychophysical aspects of visualizing DNA sequences. This avenue could lead to visualizations other than symbolic scatter plots or enhancements to symbolic scatter plots. A second avenue would explore the patterns visible in symbolic scatter plots to determine what they represent. A third avenue could look at sequences other than DNA.

For example, the points of symbolic scatter plots occupy only a fraction of the available pixels in a display. Moreover, the points are monochromatic. Theoretically, we should be able to pack significantly more information into a single display. With the combined use of color we could theoretically represent as many as a million nucleotides in one computer display and potentially a few million data points.

There is also a tendency to forget that we could visually analyze DNA sequences using more than one display. For example, it is common in the power industry to cover entire walls with several large displays in order to visualize the power grid for a large geographic region. A wall of computer displays could conceivably present tens of millions of data elements covering as much as an entire chromosome. Doing so addresses a frequent criticism that visualization is a "manual" process. Implied in this criticism is that examining DNA sequences visually is slow. However, given the speed at which pre-attentive processing occurs, several pairs of eyes could easily scan such displays within seconds to locate interesting features potentially with greater accuracy than any machine.

Symbolic scatter plots exploit some of what is known about visual perception. However, there is much that they do not exploit. For example, consider animating symbolic scatter plots with flicker. In the research presented here all of the points of a single scatter plot are displayed at the same time. When the animation switches to the next scatter plot all of the points displayed are overwritten by the new set of points at the same time. We could draw the observer's attention to different features by altering this behavior. One possibility is using variable rates of flicker for different sets of points. Other types of animation such as changes in motion or changes in color as well as visual cues other than animation could produce still other results.

In some cases, as in this research, we need to experiment with visualizations to discover what we can and cannot see. In other cases we can begin by asking, "what do we want to discover?" Do we want to find introns and exons? Do we want to find regulatory sites? Do we want to map some physical property? Do we want to correlate certain properties? Answering these questions will allow us to experiment with visualizations that quickly draw our attention to those features.

Other research can ask, "what do the patterns in symbolic scatter plots represent?" There are several possible answers. Some patterns might represent regions of DNA that fold into particular shapes. Some patterns might be associated with regions that bind proteins that enhance or inhibit transcription. Some patterns might be associated with introns while other are associated with exons. Still others might be associated with regions that serve no biological function at all.

Knowing all the letters of a DNA sequence tells us nothing about what the DNA does. Similarly, knowing the patterns present in a corresponding symbolic scatter plot does not tell us what the DNA does. However, they do serve as guideposts. If a particular pattern stands out or we see a particular pattern often enough, then we begin to wonder what it represents. We can work backwards from a pattern to its underlying DNA sequence and then work with it in the laboratory. We can experiment with it. For example, using microarrays we could determine if the DNA associated with a frequently occurring pattern in a symbolic scatter plot is expressed by cancer cells. Perhaps some patterns are only seen in growing children but not in adults. Still others might be present in stem cells but not in the cells of fully differentiated tissues (or vice versa). Only experimentation will be able to answer these questions. But once we know the answers, then seeing the patterns will help us to immediately identify the function of a DNA sequence.

Soon new techniques will be used to map the human methylome. It has been known for some time that certain C's (those immediately adjacent to G's) are chemically modified by a process called methylation. It is thought that methylation of cytosine is a mechanism for turning genes on and off. Different cytosines will be methylated in different cell types. The cytosines methylated in skin cells might be different from those that are methylated in liver cells. Methylation might also play a role in transforming cells into cancer. Mapping which C's are methylated and which are not in each cell type is a major undertaking promising to generate terabytes of new data. Symbolic scatter plots might help to make

sense of this data. It would be very exciting to find correlations between the patterns in symbolic scatter plots and the patterns of methylation in our DNA.

Lastly, DNA is but one type of sequence. Biologically there are RNA and protein sequences. However, there are also an endless number of non-biological sequences. The work that led to symbolic scatter plots began with a visualization of the integers. Could symbolic scatter plots be applied to other physical phenomena such as electrocardiograms showing the sequential electrical activity of a beating heart? Could they be used to find patterns in minute-by-minute trading of stocks and bonds? What other visualization techniques can be applied to the analysis of inherently sequential information?

The most surprising aspect of symbolic scatter plots is perhaps their simplicity. They are easy to construct and seeing patterns in them is almost effortless. Despite this simplicity, there appears to be no question about the novelty of the technique and the surprise that they haven't been tried before. Perhaps the most interesting research question of all is, "what else are we failing to see?"

# 10 Works Cited

Alston, M., Johnson, C. G., & Robinson, G. (2003). Colour merging for the visualization of biomolecular sequence data. *Proceedings of the Seventh International Conference on Information Visualization* .

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology , 215*, 403-410.

Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleid Acids Research , 34*, W369-W373.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research , 27* (2), 573-580.

Berger, J. A., Mitra, S. K., Carli, M., & Neri, A. (2004). Visualization and analysis of DNA sequences using DNA walks. *Journal of the Franklin Institute , 341*, 37-53.

Brodzik, A. K. (2007). Quaternionic periodicity transform: an algebraic solution to the tandem repeat detection problem. *Bioinformatics , 23* (6), 694-700.

Chaley, M., & Kutyrkin, V. (2008). odel of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences. *Mathematical Biosciences , 211*, 186-204.

*Chimp (Pan troglodytes) Genome Browser Gateway*. (n.d.). Retrieved from http://genome.ucsc.edu/cgi-bin/hgGateway?db=panTro2

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics , 20* (3), 426-427.

Cojocaru, A., & Murty, M. R. (2005). *An Introduction to Sieve Methods and Their Applications*. Cambridge University Press.

Cox, D. N. (2010). Towards a Visualization of DNA Sequences. In *Advances in Biological Computing.* CSREA Press.

Cox, D. N. (2008, May). Visualizing the sieve of Eratosthenes. *Notices of the American Mathematical Society* , 579-582.

Cox, D. N., & Dagnino, L. (2009). An analysis of DNA sequences using symbolic scatter plots. *BIOCOMP '09* .

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature , 227*, 561-563.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., & Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research , 27* (11), 2369-2376.

Dewey, C. N., & Pachter, L. (2006). Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Human Molecular Genetics , 15* (1), R51-R56.

Dimitrova, N., Cheung, Y. H., & Zhang, M. (2006). Analysis and visualization of DNA spectrograms: open possibilities for the genome research. *ACM Multimedia 2006 Conference* , 1017-1024.

Durand, P., Mahe, F., Valin, A., & Nicolas, J. (2006). Browsing repeats in genomes: Pygram and an application to non-coding region analysis. *BMC Bioinformatics , 7* (477).

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis.* Cambridge, UK: Cambridge University Press.

Esteban-Marcos, A., Darling, A. E., & Ragan, M. A. (2009). Seevolution: visualizing chromosome evolution. *Bioinformatics , 25* (7), 960-961.

Frith, M. C., Fu, L., Chen, J., Hansen, U., & Weng, Z. (2004). Detection of functional DNA motifs via statistical overrepresentation. *Nucleic Acids Research , 32*.

Galtier, N., Gouy, M., & Gautier, C. (1996). SEA VIEW and PHYLO_ WIN: two graphic tools for sequence alignment and molecular phylogeny. *CABIOS , 12* (6), 543-548.

Goodstadt, L., & Ponting, C. P. (2001). CHROMA: Consensus-based colouring of multiple alignments for publication. *Bioinformatics , 17* (9), 845-846.

*Google image results of patterns*. (2009). Retrieved 2009, from Google: www.google.com

Gullberg, J. (1997). *Mathematics from the Birth of Numbers.* New York: W.W. Norton & Company.

Hardy, G. H. (1940). *A Mathematician's Apology.* Cambridge: University Press.

Hauth, A. M., & Joseph, D. A. (2002). Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics , 18*, S31-S37.

Healey, C. G. (1996). Effective visualization of large multidimensional datasets. *Ph.D. Dissertation* .

Healey, C. G. (2009, May). *Perception in Visualization*. Retrieved from Dr. Christopher Healey: http://www.csc.ncsu.edu/faculty/healey/PP/index.html

Healey, C. G. (1999). Perceptual techniques for scientific visualization. *SIGGRAPH '99 Course 6* , 1-2.

Ishiwata, R. R., Morioka, M. S., Ogishima, S., & Tanaka, H. (2009). BioCichlid: central dogma-based 3D visualization system of time-course microarray data on a hierarchical biological network. *Bioinformatics , 25* (4), 543-544.

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research , 18* (8), 2163-2170.

Johnson, C., Moorhead, R., Munzner, T., Pfister, H., Rheingans, P., & Yoo, T. S. (2006). *NIH/NSF Visualization Research Challenges.* IEEE.

Jones, N. C., & Pevzner, P. A. (2004). *An Introduction to Bioinformatics Algorithms.* Cambridge, MA: The MIT Press.

Keich, U., & Pevzner, P. A. (2002). Finding motifs in the twilight zone. *Sixth Annual International Conference on Computational Biology* , 196-204.

Kolpakov, R., Bana, G., & Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research , 31* (13), 3672-3678.

Kubovy, M., Holcombe, A. O., & Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology , 35*, 71-98.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics , 23*, 2947-2948.

LaSpada, A. (2009, February). Professor, Department of Pediatrics, UCSD.

Lou, M., & Golding, B. (2007). FINGERPRINT: visual depiction of variation in multiple sequence alignments. *Molecular Ecology Endnotes , 7* (6), 908-914.

Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., & Hein, J. (2008). Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Research , 18*, 298-309.

Ma, B., Tromp, J., & Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics , 18* (3), 440-445.

Macdonald, M. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell , 72* (6), 971-983.

Maizel, J. V., & Lenk, R. P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proceedings of the National Academy of Sciences, USA , 78* (12), 7665-7669.

Marbach-Ad, G., Rotbain, Y., & Stavy, R. (2008). Using computer animation and illustration to improve high school students' achievement in molecular genetics. *Journal of Research in Science Teaching , 45* (3), 273-292.

McGill, G. (2008). Molecular movies: coming to a lecture near you. *Cell , 133*, 1127-1132.

Mount, D. W. (2004). *Bioinformatics Sequence and Genome Analysis* (Second ed.). Cold Spring Harber: Cold Spring Harbor Laboratory Press.

Parida, L. (2008). *Pattern Discovery in Bioinformatics.* Boca Raton: Chapman & Hall/CRC.

Pevsner, J. (2003). *Bioinformatics and Functional Genomics.* Hoboken, NJ: John Wiley & Sons, Inc.

Reneker, J., & Shyu, C. (2005). Refined repetitive sequence searches utilizing a fast hash function and cross species information retrievals. *BMC Bioinformatics , 6*.

Resnick, M. D. (1997). *Mathematics as a Science of Patterns.* Oxford: Oxford University Press.

Samatova, N. F., Breimyer, P., Hendrix, W., Schmidt, M. C., & Rhyne, T. M. (2008, November). An outlook into ultra-scale visualization of large-scale biological data. *Workshop on Ultrascale Visualization, 2008* , 29-39.

Santo, E., & Dimitrova, N. (2007). Improvement of spectral analysis as a genomic analysis tool. *IEEE International Workshop on Genomic Signal Processing and Statistics, 2007. GENSIPS 2007* .

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research , 18* (20), 6097-6100.

Sigalov, G., Comer, J., Timp, G., & Aksimentiev, A. (2008). Detection of DNA sequences using an alternating electric field in a nanopore capacitor. *Nano Lett. , 8* (1), 56-63.

Smoot, M. E., Guerlain, S. A., & Pearson, W. R. (2004). Visualization of near-optimal sequence alignments. *Bioinformatics , 20* (6), 953-958.

Stormo, G. D., & Hartzell, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences, USA , 86*, 1183-1187.

Tao, Y., Liu, Y., Friedman, C., & Lussier, Y. A. (2004). Information visualization techniques in bioinformatics during the postgenomic era. *Drug Discovery Today , 2* (6), 237-245.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic Acids Research , 22* (22), 4673-4680.

Ware, C. (2004). *Information Visualization: Perception for Design.* San Francisco: Morgan Kaufmann Publishers.

Yang, S. H., Cheng, P. H., Banta, H., Piotrowska-Nitsche, K., Yang, J. J., Cheng, E., et al. (2008). Towards a transgenic model of Huntington's disease in a non-human primate. *Nature , 453* (7192), 921-924.

Yoshida, T., Obata, N., & Oosawa, K. (2000). Color-coding reveals tandem repeats in the Escherichia coli genome. *Journal of Molecular Biology , 298*, 343-349.

Notes:
The sources for the examples of patterns in Figure 2.1 that were returned from Google are (from left to right):
http://www.alfredo-haeberli.com/products/pattern/1.html, http://printpattern.blogspot.com/2008_02_01_archive.html,
http://www.blog.spoongraphics.co.uk/freebies/free-ornate-wallpaper-pattern
http://evildesign.com/2007/04/wzrd_pattern_wzrd_pattern.html,
http://printpattern.blogspot.com/2007_05_01_archive.html