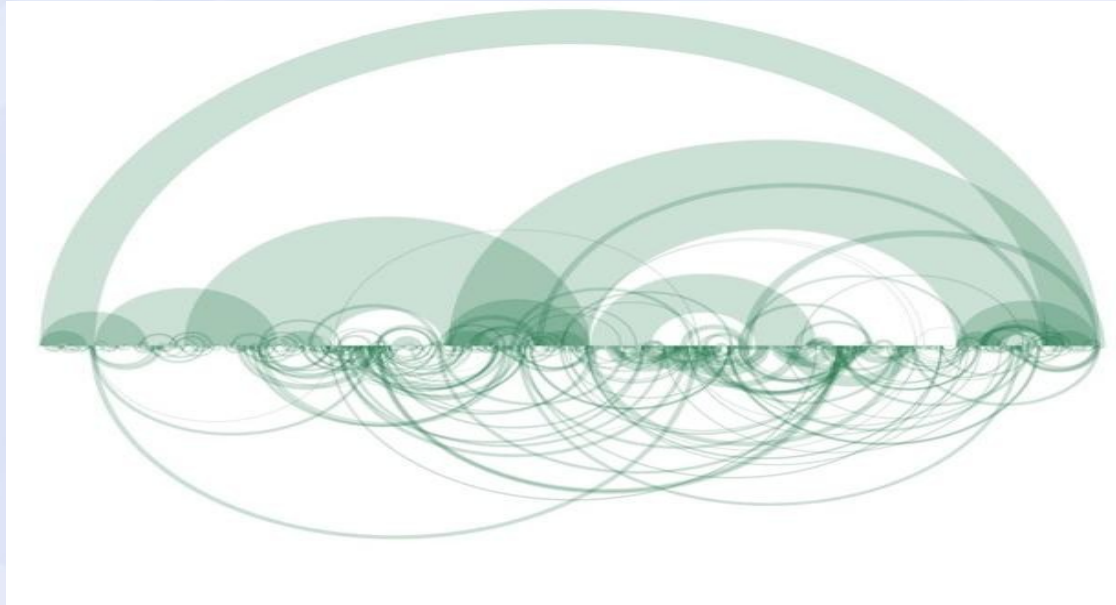


Arc Diagrams

Martin Wattenberg



Rajesh Bhasin
CS 598 Information Visualization
University of Illinois at Urbana Champaign

Overview

- Introduction
- Related Work
- Arc Diagrams
- Applications
- Future Work
- Critiques

About the Author

- Computer Scientist and Artist. Founded IBM's Visual Communication Lab.[IBM TJ Watson]
- Works on novel ways of Visualization, Visual Tools .[Project]
- Best known for his Visualization based artwork-on display in museums in London .New York and other cities .



1.Introduction

- Novel method of visualizing complex patterns of repetition in data .
- Data sets come in forms of strings. Eg :-Melodies(Repeated Strings).
- Existing methods have drawbacks for complex strings.
- Paper describes the design and implementation of this novel method.

2.Existing Methods

- H-curve and W-curve-Sequence into 3D space[Hamori and Ruskin].
- Letter chain representations of DNA chains.
- What happens for long chains?
- Map DNA sequences into 3-D space.

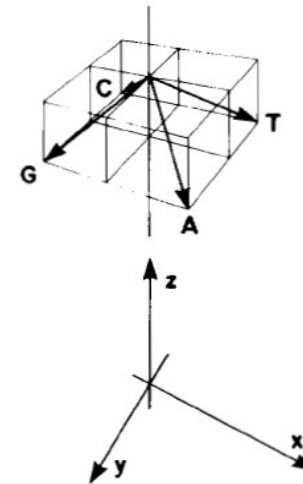
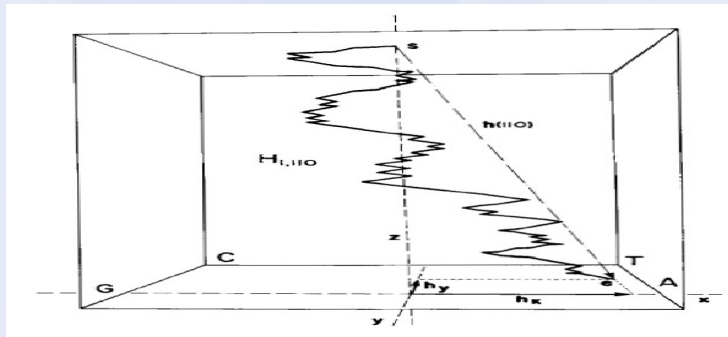


FIG. 1. The selected base vector assignments of the 4 DNA nucleotides A, G, C, and T.

- [+] Can show fine detail.
- [-] Hard to interpret.

- Chaos Game Representation(CGR) for gene structure. Investigation of patterns , visually revealing unknown structures.
- Based on techniques from chaotic dynamics.

Sierpinski Triangle

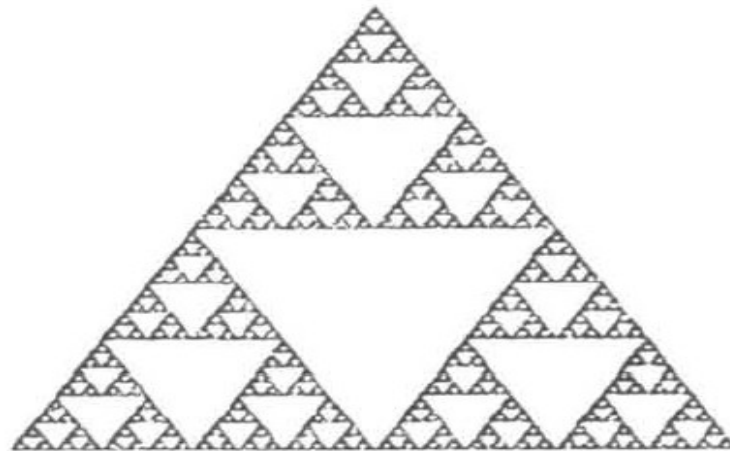
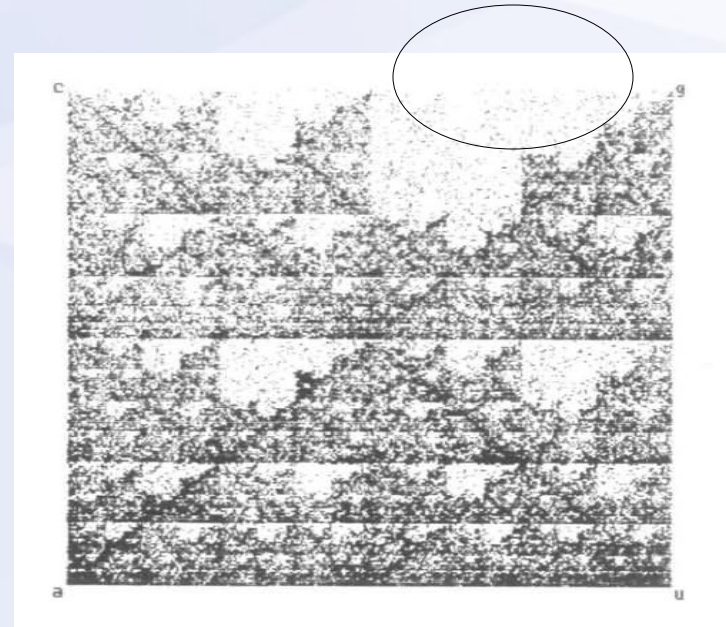
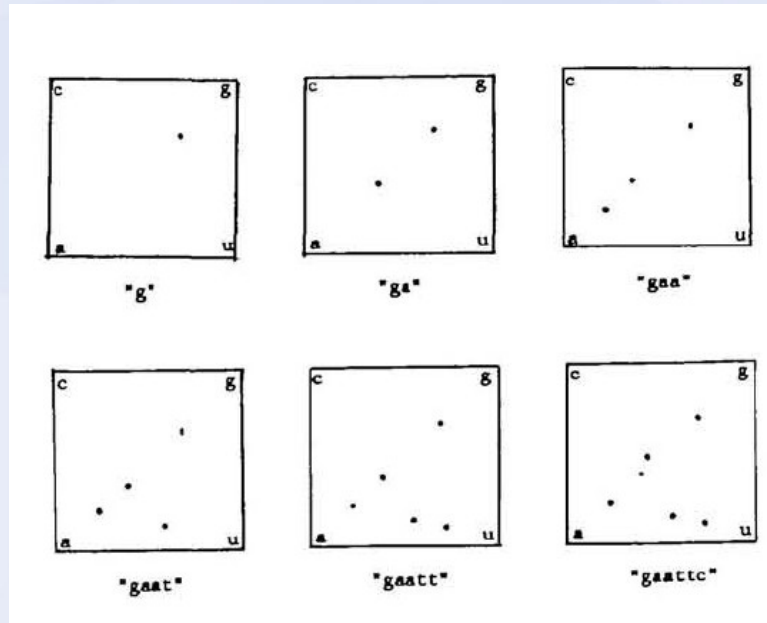


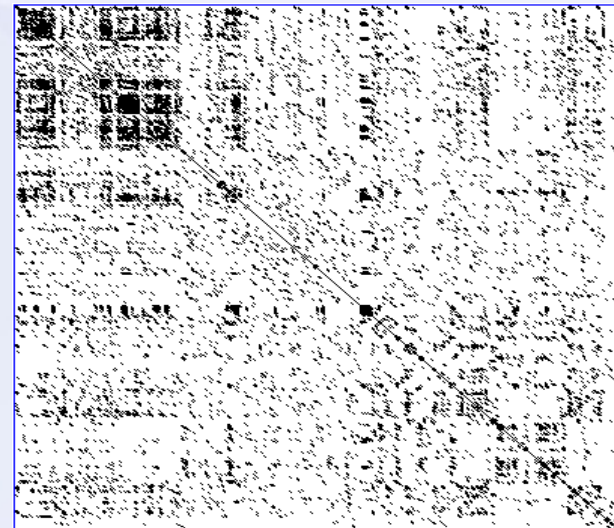
Figure 1. The Result of the Chaos Game on Three Points.

- [+] Works well for small sub-strings.
- [-] Remove Ordering.



- Dotplot for DNA sequences-One of the most popular methods.
- [+]Can handle large data sets,resistant to noise and can show large structures.
- [-] Any guesses on drawback ?
- $O(n^2)$ visual features.Can be confusing if applied to frequently used substrings.

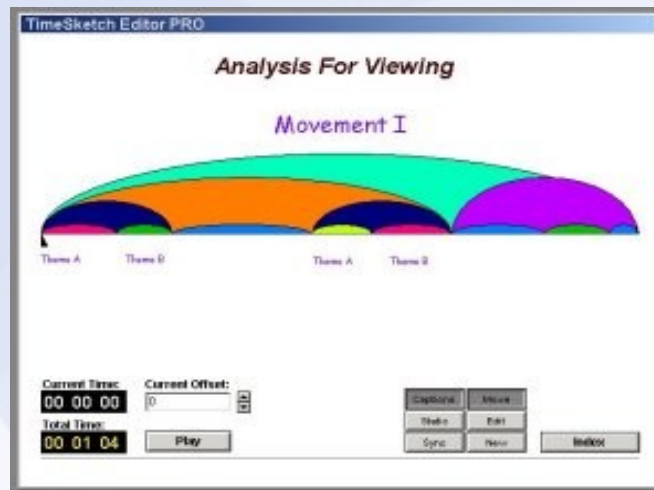
A DNA dot plot of a human zinc finger transcription factor(Source: Wikipedia)



- Non Visual method -Describing which sub-sequences are repeated. Eg: AABB.
- [+]Easy to understand.Gives broad view.
- [-]Looses smaller details.
- Music Theorists - Schenkerian Diagrams. Subjective and Manual. Unsuitable for automation.



- Time Sketches-Uses color for differentiation .
- [-]Requires human definition of 'related passages'. Colors do not scale well for large quantities.



3.Arc Diagram

- The idea is to visualize only a subset of all possible pairs of matching substrings.
- Highlights only the sub-sequences essential to understand the string's structure.
- Avoids Quadratic problem faced by dotplot.

How is it different?

Dotplot	<ul style="list-style-type: none">•Can efficiently represent sequences where individual sub-sequences are repeated many times.
Timesketch	<ul style="list-style-type: none">•an arc diagram can be constructed automatically•can represent the structure of a sequence with many different repeated subsequences and multiple scales of repetition

Definitions

1. Maximal Matching Pair:

- *Identical*- Same sequence of numbers.
- *Non-Overlapping*-Do not intersect.
- *Consecutive*-No identical string should start between them.
- *Maximal*-A longer sub-sequence cannot be found .

10101010??

- Hence , we may not want to base the visualization on Maximal pairs alone.

2. Repetition Region-

- Region where a sub-sequence is repeated in immediate succession.
- Each repetition is called fundamental sub-string.

Now we can define the precise set of sub-strings we want to use.

3. Essential Matching pair-

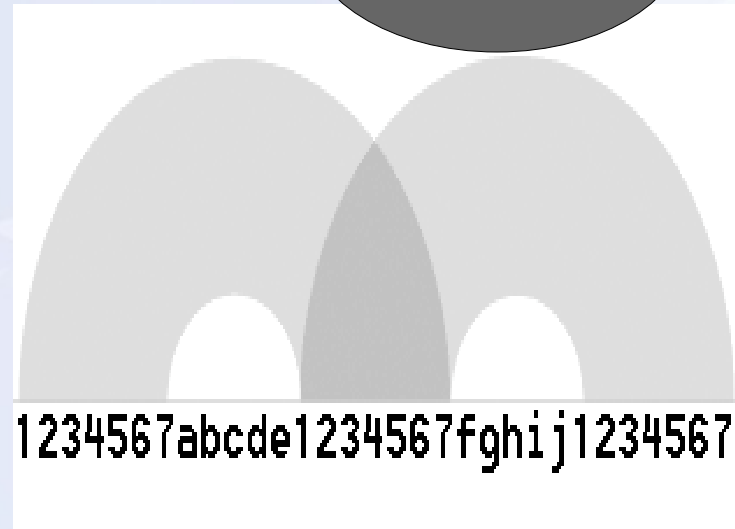
- maximal matching pair not in any repetition region.
- a maximal matching pair contained in the same fundamental substring of any repetition region that contains it
- two consecutive fundamental substrings for a repetition region

Procedure

- Define the mapping from string to x-axis.
- m th symbol is at m/N position on x-axis.
- String corresponds to an interval.
- Connect essential matching pairs with thick arcs.
- Height is proportional to the distance.

Examples

What do you notice here?





1010101010101010

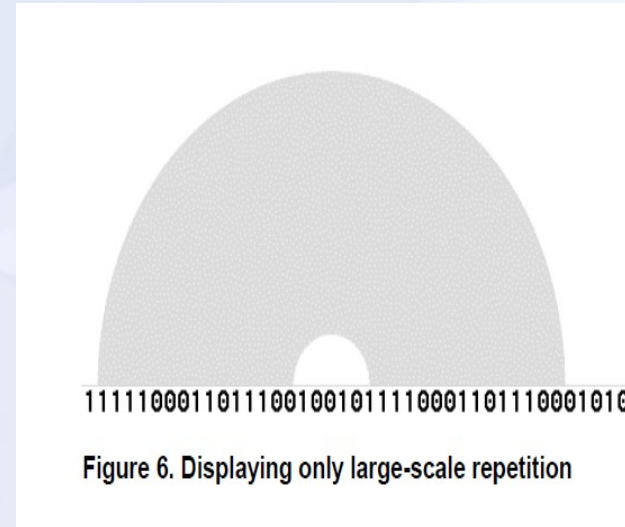
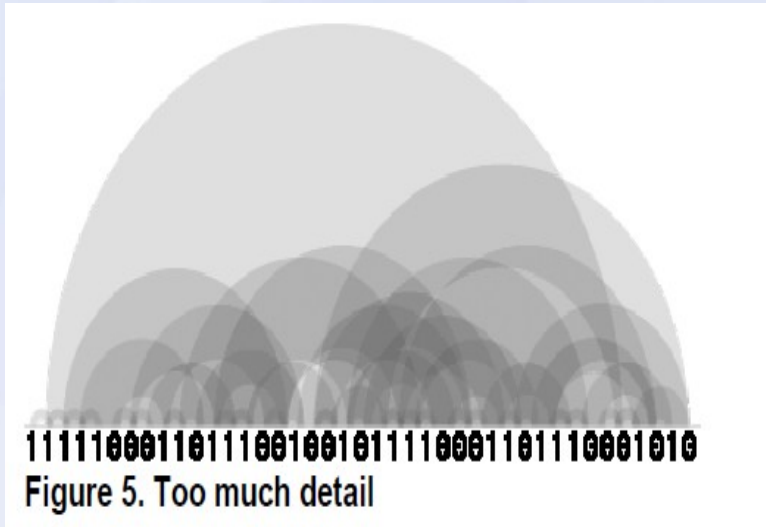
The diagram shows a binary string '1010101010101010' with a horizontal line above it. Six overlapping, semi-transparent grey arches are positioned above the line, each spanning two characters of the string. This illustrates the concept of immediate repetition, where a substring is repeated consecutively.

Figure 3. Immediate repetition

According to Definition 3(c)

Lets Try

11111000110111001001011110001101110001010

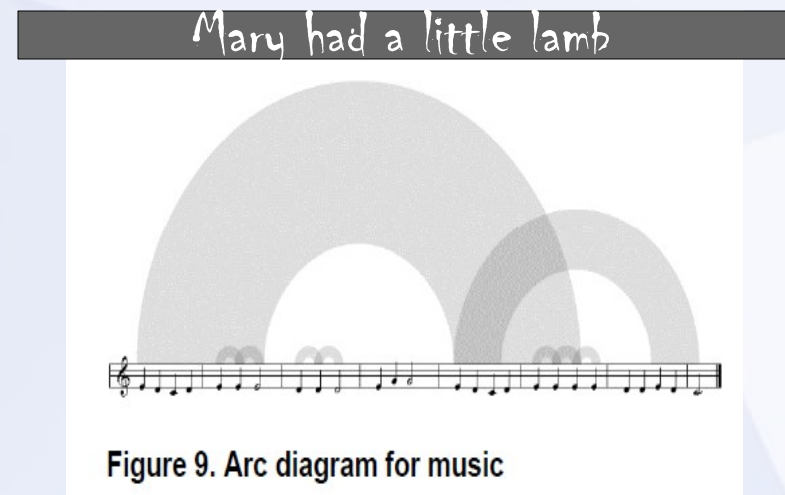


Implementation

- Using Java, low-end 266 MHz Pentium II.
- To enumerate repeated patterns construct a suffix tree and traverse it twice.
- In the first pass you find the repetition regions
- In the second pass you find the essential matching sub-strings.

4.Applications

- *Music*-Most promising application. Reveals the structure in musical compositions.
- Same sequence of pitches.
- Resulting matching diagrams reveal an intricate and beautiful structure.



- “AABB” structure .Picks out structures corresponding to regular music analysis.[Minuet]
- Relationship between A and B?

[Overlap]

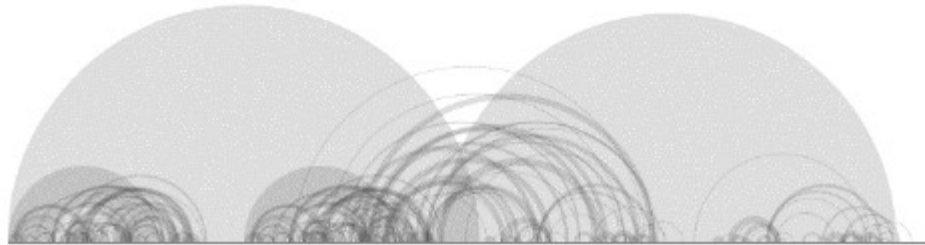


Figure 12. Minuet in G Major, Bach

Text and Compiled Code-

- 2 main divisions
- Proportions



Figure 13. Java class file (bytecode)

- DNA- Very important application.
- Used to find patterns in UTR(Upstream Transcription Region)
- [+]Provides both information and clarity.
- [-]Noisy data (repetition on a large scale is uncommon due to mutations)
- Motifs supposed to play a role in regulating the gene.

- Recent finding that regulatory motifs tend to appear in a restricted area.[Atleast 1 instance]



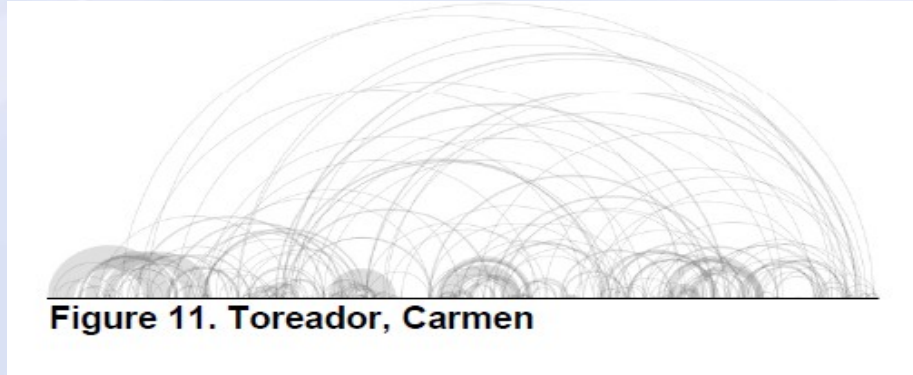
5.Future Work

- Interactivity to the diagrams.
- Sliders to control detail.
- Drawing particular sub-sequence/Playing corresponding music on clicking.(drill down into detail)
- Additional Variables into Visualization.
- Different Hues/intensity to make the change evident.

6. Critiques

- Large amount of information using simple techniques.
- Show the symmetry and are aesthetically pleasing.
- Generic Nature -can be applied to many more areas.
- Can be used to detect the nature of code, especially with XML (Unmatched code).
- Detecting particular patterns (say in Youtube videos). Protecting copyrighted material.

- Misses a deeper level of structure at places[Eg-Toreador ,Carmen].



- Could incorporate similarity rather than just repetition.
- Come up with a heuristic/formula for deciding thresholds.
- Adding colors , extending to 3-D , using the lower portion of the x-axis.
- Opacity and Lighting.

- DNA sequences seems to be one of the most useful applications of the visualizations.
- Fuzzy pattern logic helps. Especially with noisy data.
- Could be applied to other forms - arts , pictures
- Detect design ,structure or even plagiarism.
- Could be extended to real time.

Demo

<http://www.turbulence.org/Works/song/mono.html>

Thank You

Questions??