TCGR: A Novel DNA/RNA Visualization Technique

Donya Quick and Margaret H. Dunham
Department of Computer Science and Engineering
Southern Methodist University
Dallas, Texas 75275
dquick@mail.smu.edu, *mhd@engr.smu.edu*

Table of Contents

FCGR

•A	•C
•G	•T

•a) Nucleotides

•AA	•AC	•CA	•CC
•AG	•AT	•CG	•CT
•GA	•GC	•TA	•TC
•GG	•GT	•TG	•TT

•b) Dinucleotides

•AAA	•AAC	•ACA	•ACC	•CAA	•CAC	•CCA	•CCC
•AAG	•AAT	•ACG	•ACT	•CAG	•CAT	•CCG	•CCT
•AGA	•AGC	•ATA	•ATC	•CGA	•CGC	•CTA	•CTC
•AGG	•AGT	•ATG	•ATT	•CGG	•CGT	•CTG	•CTT
•GAA	•GAC	•GCA	•GCC	•TAA	•TAC	•TCA	•TCC
•GAG	•GAT	•GCG	•GCT	•TAG	•TAT	•TCG	•TCT
•GGA	•GGC	•GTA	•GTC	•TGA	•TGC	•TTA	•TTC
•GGG	•GGT	•GTG	•GTT	•TGG	•TGT	•TTG	•TTT

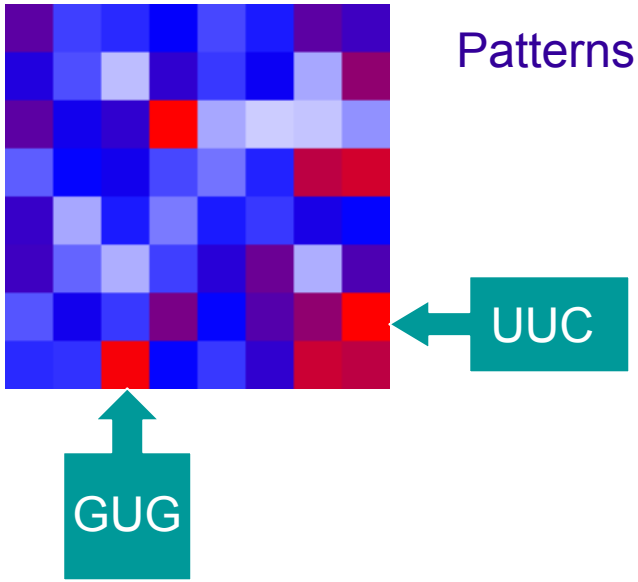
•c) Trinucletides

Courtesy of Eamonn Keogh, UCR

FCGR Example

Homo sapiens – all mature miRNA

Patterns of length 3



What is TCGR?

- Temporal Chaos Game Representation (TCGR)
- A visual and numerical representation of data
- Can be applied to DNA sequence data as well as other data types
- Shows general structure of sequences
- Structure is represented as distribution of subsequence over sequence length.

Temporal CGR (TCGR)

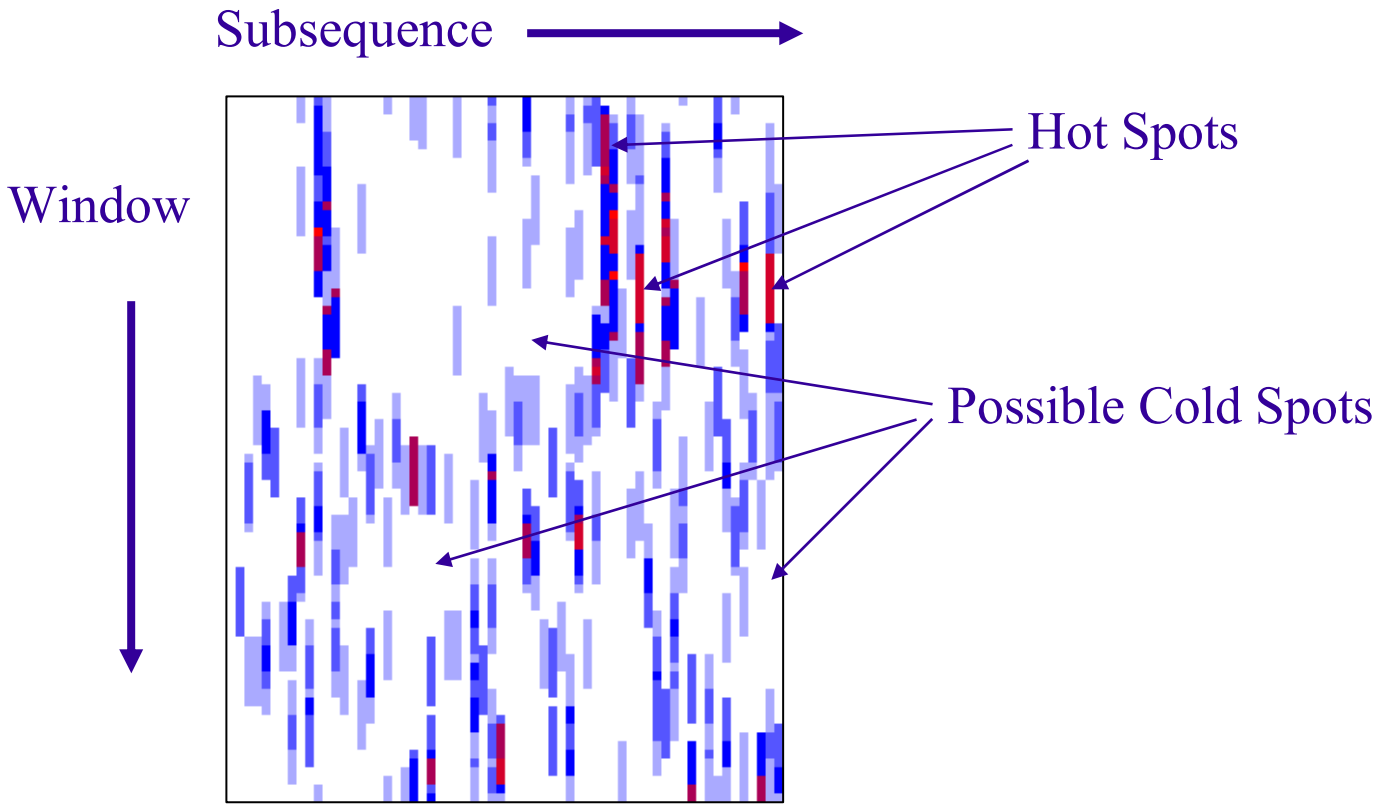
- Temporal version of Frequency CGR
  - *In our context temporal means the starting location of a window*
- 2D Array
  - *Each Row represents counts for a particular window in sequence*
    - *First row – first window*
    - *Last row – last window*
    - *We start successive windows at the next character location*
  - *Each Column represents the counts for the associated pattern in that window*
    - *Initially we have assumed order of patterns is alphabetic*
  - *Size of TCGR depends primarily on sequence length and subsequence size*
- As sequence sizes vary, we only examine complete windows
- We only count patterns completely contained in each window.

Frequency

- Instead of actual frequencies, a normalization (based on largest frequency in sequence) is used.
- 0.0 means a subsequence did not occur in a window
- 1.0 for a subsequence means it is the most frequently occurring in the data set.
- Color schemes for visualization:

Frequency	Color	Grayscale
0.0	White (“cold spot”)	White
0.5	Blue	Gray
1.0	Red (“hot spot”)	Black

TCGR Representation



TCGR Algorithm Overview

1. Counting
   While windows are left:
   - *Count all subsequences present for all strings in current window*
   - *Move window down by specified overlap and repeat*
2. Frequency conversion
   - *Divide all subsequence counts by maximum to scale to [0,1].*

TCGR Algorithm Overview

```
CAAAAGAGTCAGGG
AAAATTCAGGGAT_
CAAAAGAGTCAGGG
AAAATTCAGGGAT_
   . . .
CAAAAGAGTCAGGG
AAAATTCAGGGAT_
```

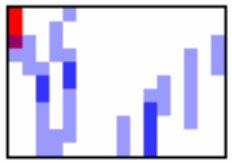
Counting Process

	AA	AC	AG	...	TT
Win0:	5	0	0		0
Win1:	5	1	0		0
...					
Win10:	0	0	1	...	0

Counts Array

	AA	AC	AG	...	TT
Win0:	1.0	0	0		0
Win1:	1.0	0.2	0		0
...					
Win10:	0	0.2	0.2	...	0

Frequency Array



TCGR output

TCGR Parameters

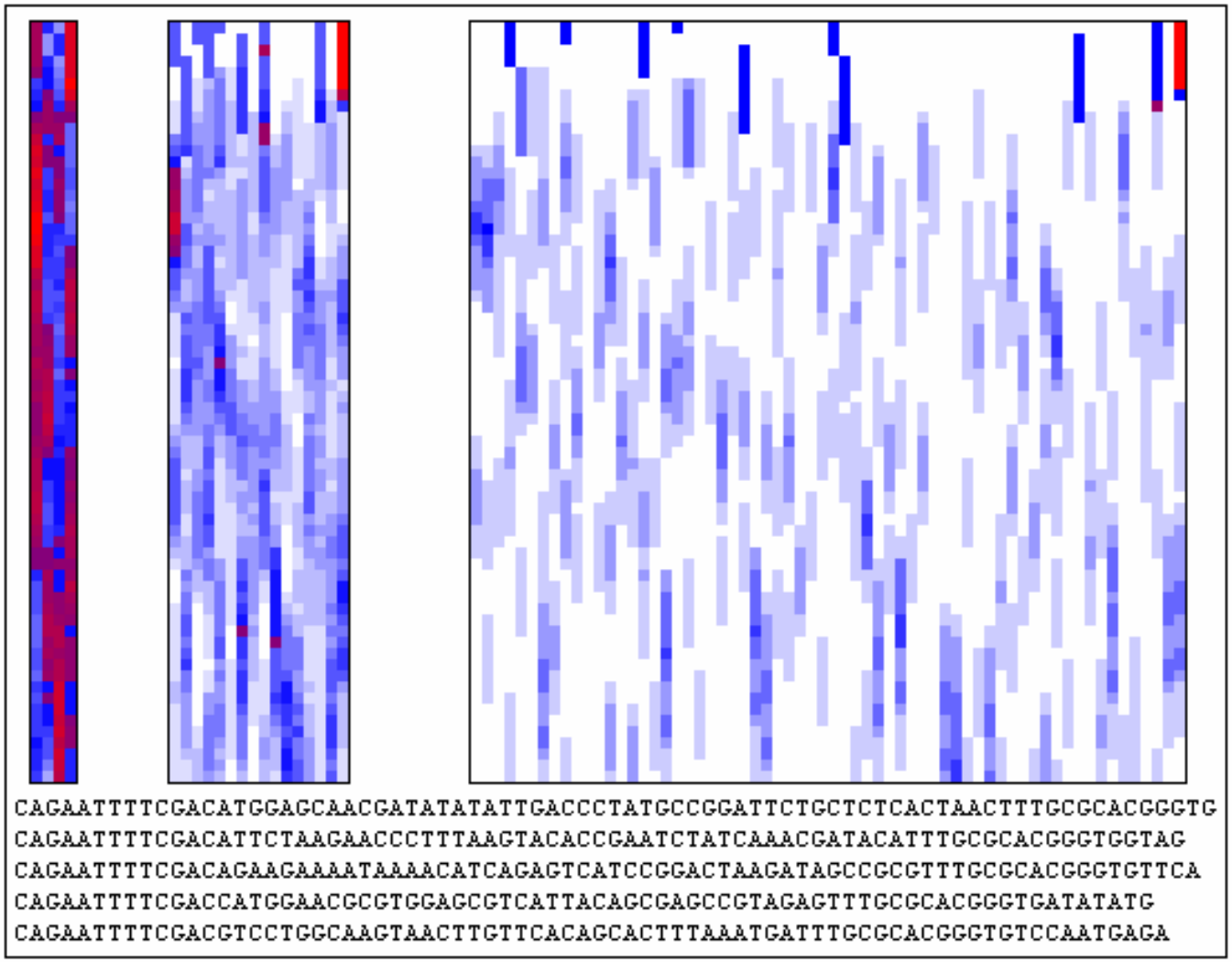
- Subsequence size (SS)
- Maximum Count Value
- Window Length (WL)
- Window Overlap (WO)

Effects of Subsequence Size

- Number of columns is 4^n
- For a constant window length and overlap and increasing subsequence size:
  - *The number of columns will increase exponentially*
  - *The TCGR will become less dense (more white space)*
  - *As density decreases, white space holds less potential meaning.*

Effects of Subsequence Size

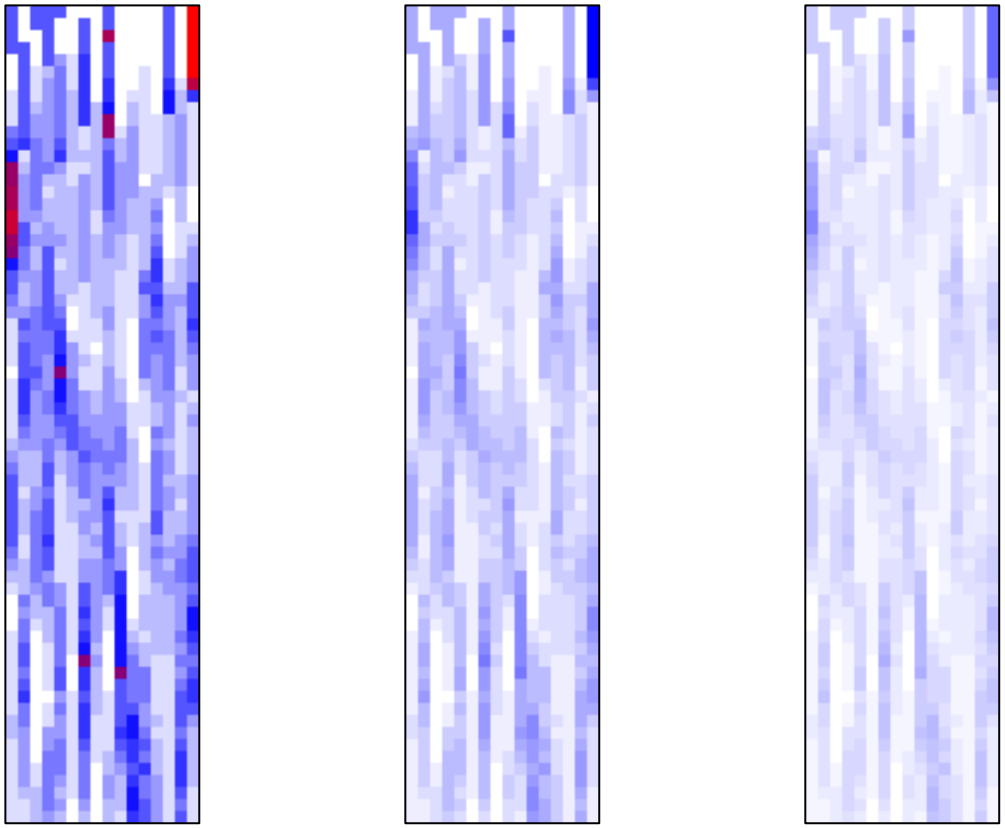
SS=1 SS=2 SS=3



Effects of Maximum Count Value

- Affects the scaling of the data at the frequency level.
- When the maximum count value is low, small differences in frequency are more visible.
- If comparing TCGRs for two different sequences, should scale both to the same maximum count value to avoid false hot spots.
- If comparing TCGRs where each represents a set of many sequences, using the default scaling may be better to compare relative structure.

Effects of Maximum Count Value



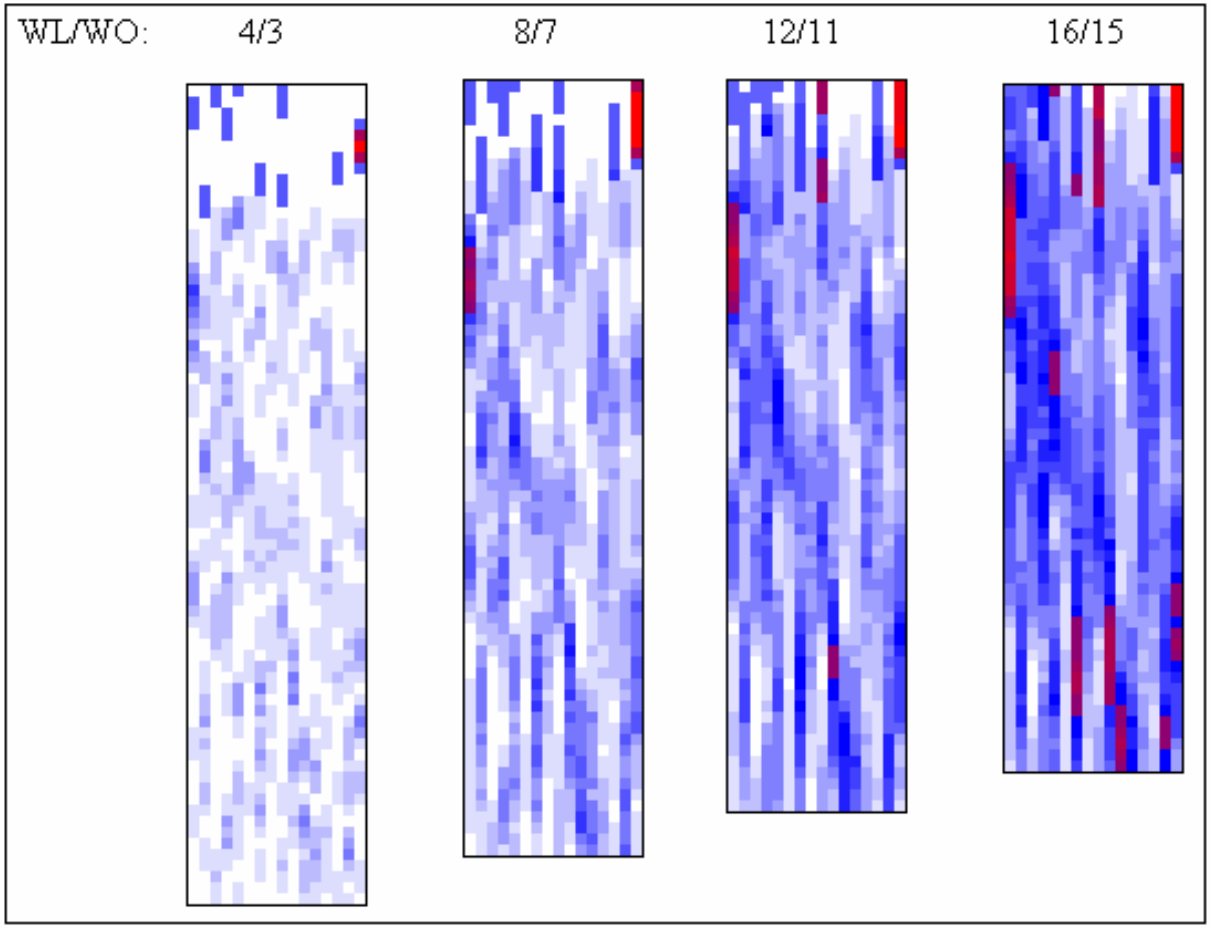
Max=23 (default) Max=30 Max=50

(data from slide 15, multiple sequences)

Effects of Window Length

- For a constant SS and maximal WO:
  - *The output becomes denser*
  - *Cold spots may become more meaningful*
  - *Total number of rows will decrease*

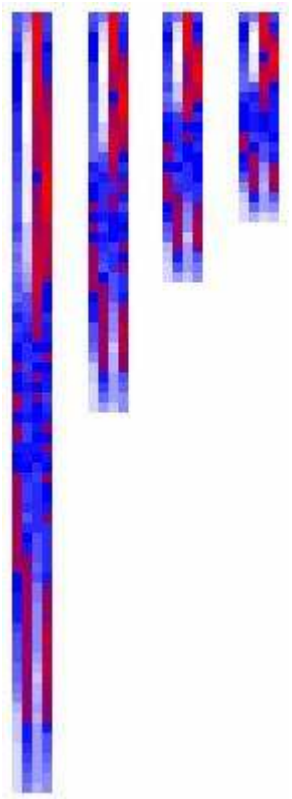
Effects of Window Length



Effects of Window Overlap

- Gives best results when maximized
- Risks associated with decreasing WO:
  - *Boundary anomaly can occur if last window is only partially filled*
  - *Maximum count values may be missed*
  - *Scaling may be off due to missed maximum counts*

Effects of Window Overlap



SS=1, WL=10

WO = 9, 8, 7, and 6 respectively

GGGTGAGGTAGTAGGTTGTATAGTTTGGGGCTCTGCCCTGCTATGGGATAACTATACAATCTACTGTCTTTCCT_____

TCAGAGTGAGGTAGTAGATTGTATAGTTGTGGGGTAGTGATTTTACCCTGTTCAGGAGATAACTATACAATCTATTGCCTTCCCTGA

CTGGCTGAGGTAGTAGTTTGTGCTGTTGGTCGGGTTGTGACATTGCCCGCTGTGGAGATAACTGCGCAAGCTACTGCCTTGCTA___

AGGTTGAGGTAGTAGGTTGTATAGTTTAGAATTACATCAAGGGAGATAACTGTACAGCCTCCTAGCTTTCCT_____

CCCGGGCTGAGGTAGGAGGTTGTATAGTTGAGGAGGACACCCAAGGAGATCACTATACGGCCTCCTAGCTTTCCCCAGG_____

(Xu et al.)

Effects of Sequence Alignment

- If used before performing TCGR:
  - *Can result in more accurate data representation*
  - *Hot spots will not be missed due to being misaligned*
  - *Rows may increase, particularly if gaps are allowed*

Effects of Sequence Alignment

A synthetic data set:

CAGAATTTTCGACATGGAGCAACGATATATATTGACCCTATGCCGGATTCTGCTCTCACTAACTTTGCGCACGGGTG
CAGAATTTTCGACATTCTAAGAACCCTTTAAGTACACCGAATCTATCAAACGATACATTTGCGCACGGGTGGTAG
CAGAATTTTCGACAGAAGAAAATAAAACATCAGAGTCATCCGGACTAAGATAGCCGCGTTTGCGCACGGGTGTTCA
CAGAATTTTCGACCATGGAACGCGTGGAGCGTCATTACAGCGAGCCGTAGAGTTTGCGCACGGGTGATATATG
CAGAATTTTCGACGTCCTGGCAAGTAACTTGTTCACAGCACTTTAAATGATTTGCGCACGGGTGTCCAATGAGA

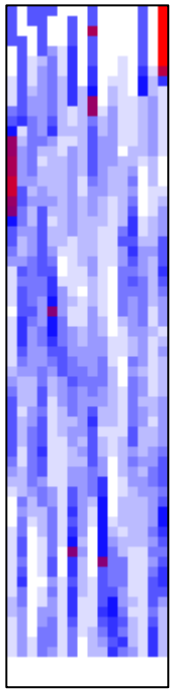
Conserved regions are marked in red.

Sample alignment of the data:

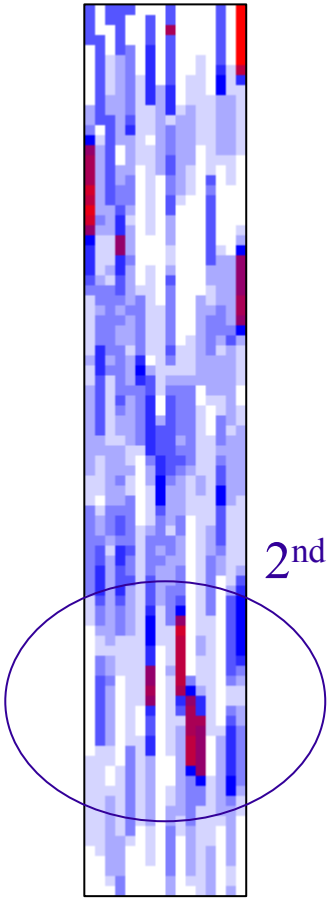
```
CAGAATTTTCGACATTCTAAGAAC_C____C_TTTAAGTAC_ACCGAA_TCTATCA__AACGATACATTTGC_GCACGGGTGG__TAG__________
CAGAATTTTCGACGTCCTGGCAAG_TAA__C_TTG__TT_C_ACAGCA_CTT_T_A__AATGAT_T_TGCGC_ACGGGTGTCCAATGAGA________
CAGAATTTTCGACAG___AAGAAAATAAAACATCAGAGTC__ATCCGGACT_AAGAT_AGCCGCGTTTGCGC_ACGGGTGTTCA_____________
CAGAATTTTCGACATGGAGCAACGATATAT_ATTGACCCTATGCCGGATTCTGCTCTCACTAACTTTGCGC__ACGGGTG___________________
_____________________________CAGAATTTTCGACCATGGAACGCGTGGAGCGTCATTACAGCGAGCCGTAGAGTTTGCGCACGGGTGATATATG
```

Effects of Sequence Alignment

Data unaligned



Data aligned



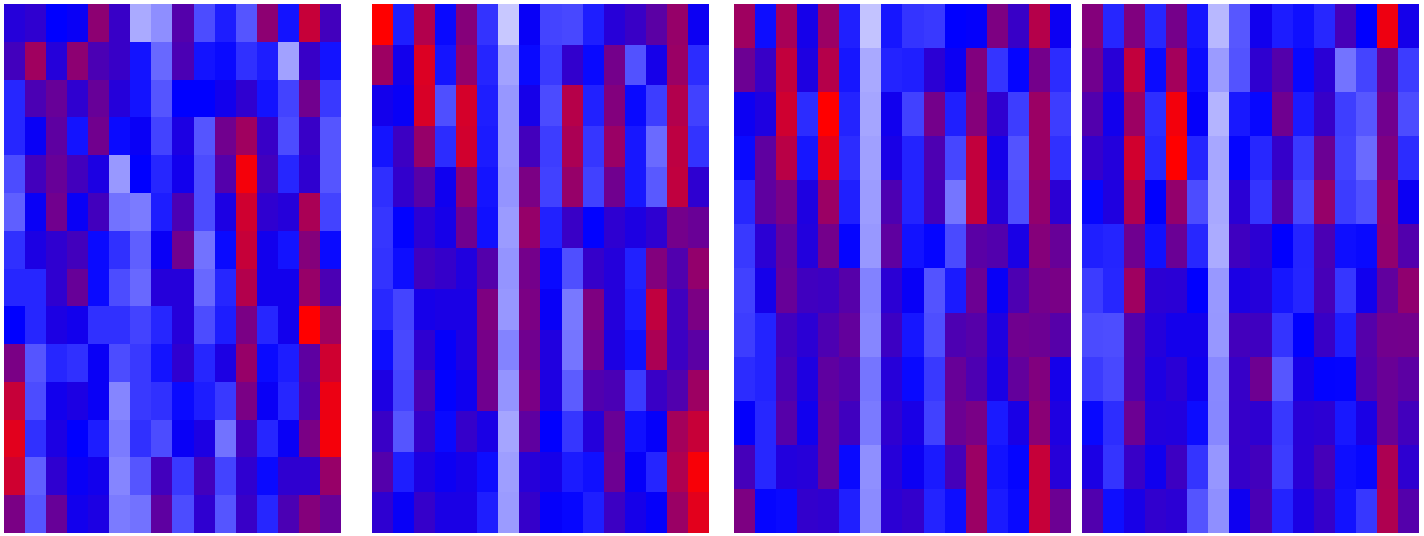
2nd set of hot spots

TCGR Applications

- Visualize Structure
  - *Identify motifs or conserved regions*
- Predict locations of DNA/RNA features
  - *miRNA*
  - *miRNA binding site*
- May be generalized to non DNA/RNA strings (temporal spatial data)
- Has been linked to a modeling prediction technique - EMM

Visualize Structure

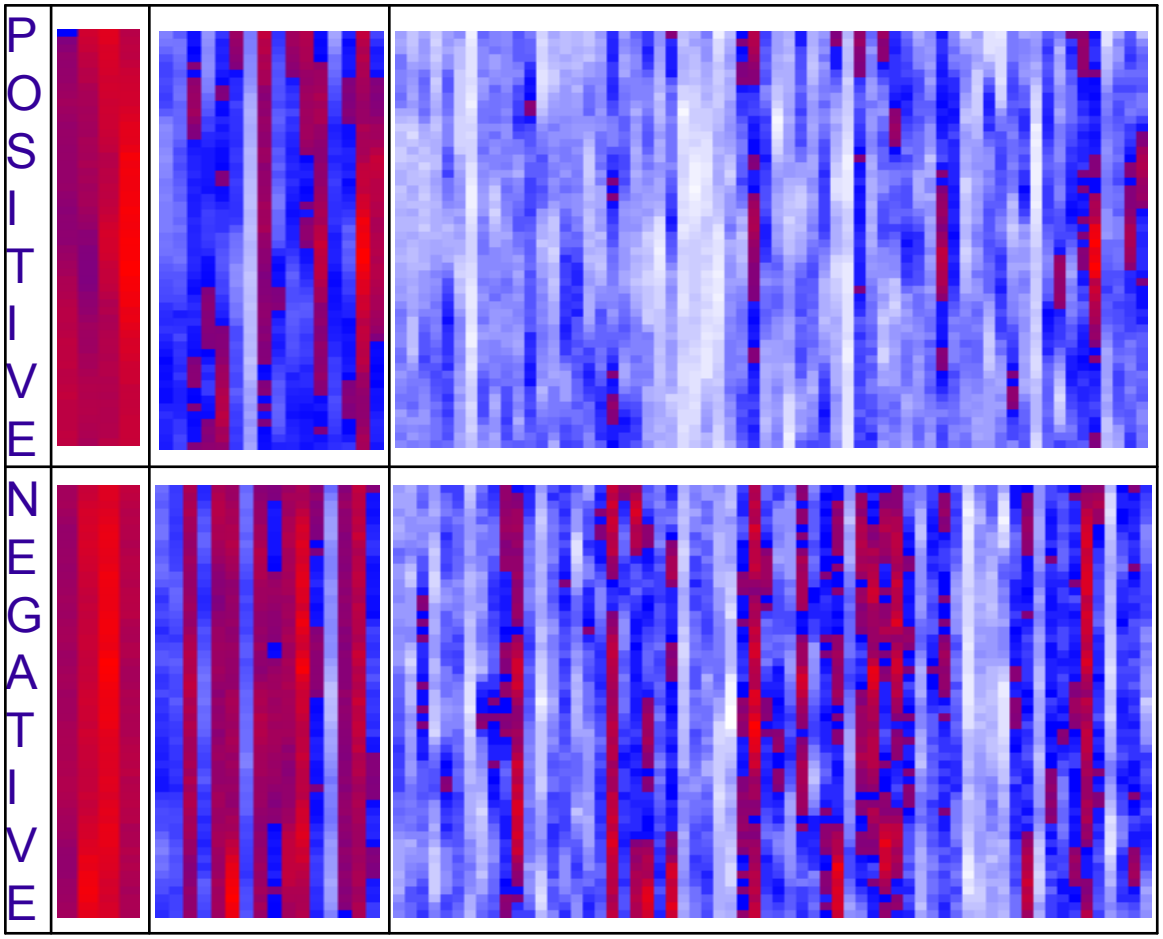
TCGR – Mature miRNA
(Window=5; Pattern=2)



C Elegans | Homo Sapiens | Mus Musculus | All Mature

All higher level animals' miRNA have a noticeable CG cold streak

Predict miRNA site



Data from: C. Xue, F. Li, T. He, G. Liu, Y. Li, nad X. Zhang, "Classification of Real and Pseudo MicroRNA Precursors using Local Structure-Sequence Features and Support Vector Machine," *BMC Bioinformatics*, vol 6, no 310.

TCGR - Not Just a Pretty Picture

1. Represent potential miRNA sequence with TCGR sequence of count vectors
2. Create dynamic Markov chain, EMM, using count vectors for known miRNA (miRNA stem loops, miRNA targets)
3. Predict unknown sequence to be miRNA (miRNA stem loop, miRNA target) based on normalized product of transition probabilities along clustering path in EMM

EMM Creation

1: <18,10,3,3,1,0,0>

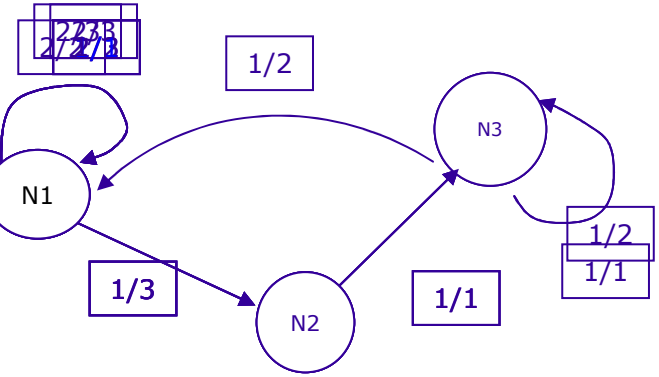
2: <17,10,2,3,1,0,0>

3: <16,9,2,3,1,0,0>

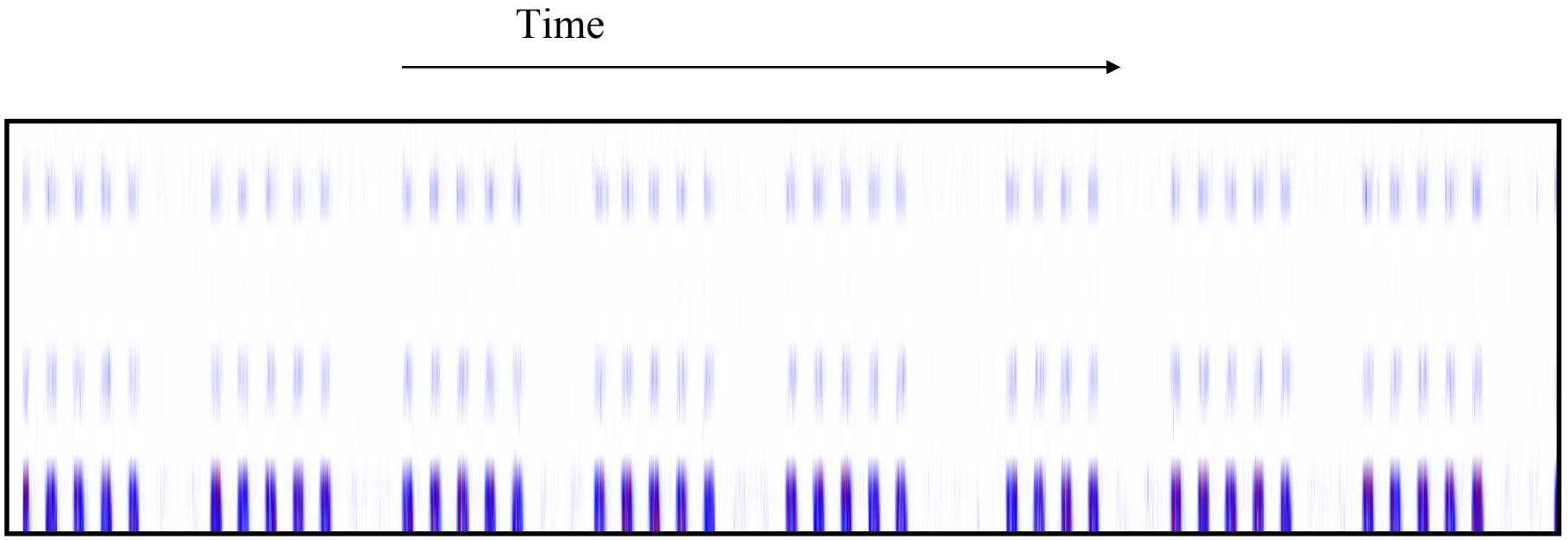
4: <14,8,2,3,1,0,0>

5: <14,8,2,3,0,0,0>

6: <18,10,3,3,1,1,0.>

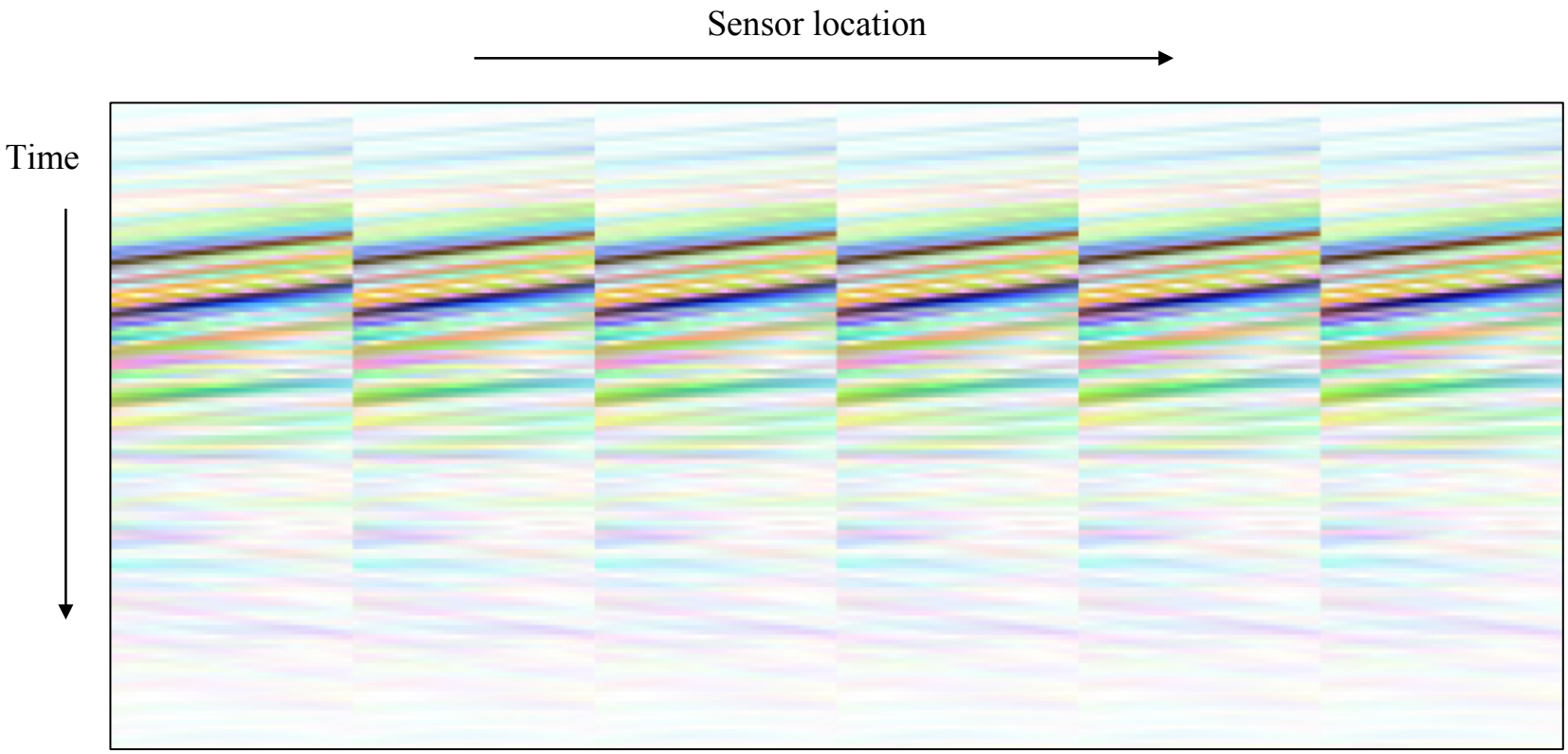


Cisco – Internal VoIP Traffic Data



VoIP traffic data was provided by Cisco Systems and represents logged VoIP traffic in their Richardson, Texas facility from Mon Sep 22 12:17:32 2003 to Mon Nov 17 11:29:11 2003.

Seismic Data Example



Conclusions

- TCGR is a useful new tool for data where composition varies with respect to distance or time.
- TCGR can be applied to data mining for event detection.
- Potential applications of TCGR to biological data include motif detection.
- Careful use of parameters makes TCGR more useful.

References

[1] C. S. a. A. Consortium, “Initial sequence of the chimpanzee genome and comparison with the human genome,” *Nature*, vol. 437, pp. 69-87, 2005.

[2] N. Rajewsky, "microRNA target predictions in animals," *Nat Genet*, vol. 38 Suppl 1, pp. S8-S13, 2006.

[3] H. J. Jeffrey, “Chaos Game Representation of Gene Structure*,” Nucleic Acids Research*, 1990, vol 18, pp 2163-2170.

[4] P.J., Deschavanne, A. Giron, J. Vilain, G. Fagot and B. Fertil, Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences,” *Molecular Biol. Evol, 1999, vol 16, pp 1391-1399.*

[5] Dunham et al, “Visualization of DNA/RNA Structure using Temporal CGRs,” *IEEE BIBE Conference Proceedings,* pp171-178, 2006.

[6] Margaret Dunham, Yu Meng, and Jie Huang, “Extensible Markov Model”, *Proc. IEEE Int’l Conf. Data Mining* (ICDM 04), 2004.

[7] E. Berezikov, E. Cuppen, and R. H. Plasterk, "Approaches to microRNA discovery," *Nat Genet*, vol. 38 Suppl 1, pp. S2-7, 2006.

[8] I. Bentwich, A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, E. Sharon, Y. Spector, and Z. Bentwich, "Identification of hundreds of conserved and nonconserved human microRNAs," *Nat Genet*, vol. 37, pp. 766-70, 2005.

[9] J. W. Nam, K. R. Shin, J. Han, Y. Lee, V. N. Kim, and B. T. Zhang, "Human microRNA prediction through a probabilistic co-learning model of sequence and structure," *Nucleic Acids Res*, vol. 33, pp. 3570-81, 2005.