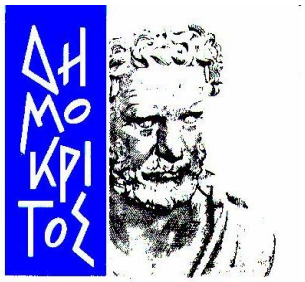


University of Crete

Analysis and Classification of Constrained DNA Elements with N-gram Graphs and Genomic Signatures

Dimitris Polychronopoulos[1,4], Anastasia Krithara[2], Christoforos Nikolaou[3], Giorgos Paliouras[2], Yannis Almirantis[1], and Giorgos Giannakopoulos[2] *

[1] Institute of Biosciences and Applications, NCSR Demokritos, 15310 Athens, Greece.
[2] Institute of Informatics and Telecommunications, NCSR Demokritos, 15310 Athens, Greece.
[3] Department of Biology, University of Crete, 71409 Heraklion, Greece.
[4] Department of Biochemistry and Molecular Biology, Faculty of Biology, National and Kapodistrian University of Athens, 15701 Athens, Greece.

* Corresponding author: ggianna@iit.demokritos.gr

1st International Conference on Algorithms for Computational Biology, AlCoB 2014
Tarragona, Spain, July 1-3, 2014

Constrained Elements and their Importance

- Only a small percent of the human genome has an assigned function
- Constrained Elements: Exons & Conserved Non-coding Elements (CNEs)
- Sequence conservation is a major indication of functionality

Curious DNA compositional preferences of CNEs

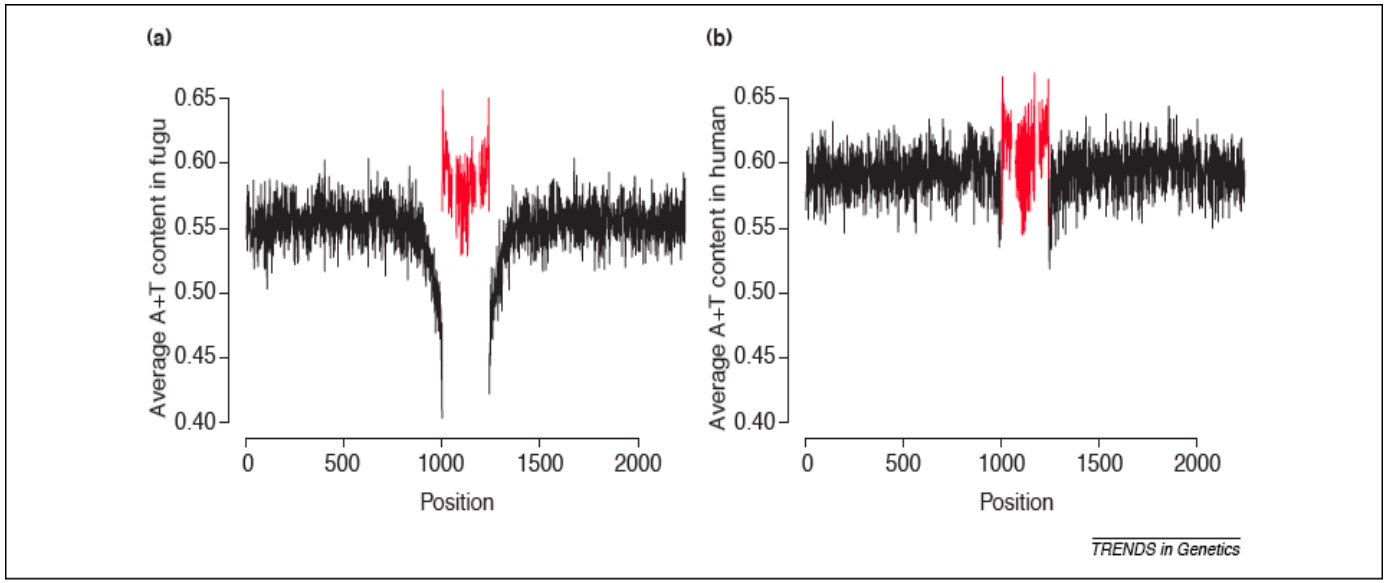


Figure 1. A+T bias in and around Fugu and human CNEs. **(a)** The average A+T content in 2200 columns of Fugu sequences that contain 1000 bp in the 5′ flanking region, 50 bp from the 5′ end of the CNE, 100 bp from the centre of the CNE, 50 bp from the 3′ end of the CNE and 1000 bp in the 3′ flanking region. A small gap indicates the border between boundary and central sequences. Panel **(b)** shows average A+T content of similarly compiled human CNE sequences. CNEs are shown in red and flanking regions are in black.

>Non-homogeneous distribution between chromosomes of both CNEs & exons

>Non-homogeneous DNA composition compared to their flanking regions (CNEs)

>**Overall**: AT rich but found in regions with a surrounding lower AT content

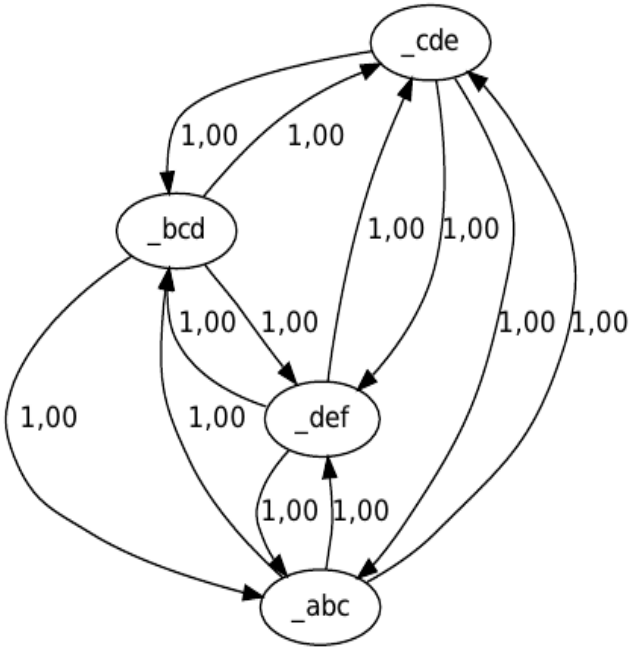
Problem Definition

1. To which extent are we able to distinguish bulk genomic sequences from different genomes?
2. Could we distinguish Constrained Elements from the bulk of the same genome?
3. To which extent are CNEs or exons from different organisms distinguishable?

Genomic Signatures - GS (Karlin *et al. 1995*)

- Standard Methodology for classifying and distinguishing genomes based on computing the odds ratios for dinucleotides.
- **Karlin *et al.* first proposed that these quantities differentiate between different genomes according to their evolutionary distance.**
- They have assigned to the vector of ten “first neighbor preferences” the name of Genomic Signatures.
- This is the first application of GS to the classification of short genomic sequences (< 50 kb)

n-gram graphs



- A graph which is constructed by:
  - Assigning each n-gram to a node
  - Assigning an edge between two nodes that are in proximity (within a certain distance in the symbol series)
  - Assigning a weight to each edge based on the frequency/count of co-occurrence (proximity)
- Describes symbol proximity (e.g. nucleotides)
- Edges are important
- Weights signify frequency of co-occurrence

Constructing an n-gram graph (NGG)

- Calculate n-gram content of various orders (L)
- Note proximity between each n-gram and all others within a distance D.
- Assignment of weights (may be distance-dependent or independent)
- Example:
  - A symbol sequence: atcag
  - 3-grams (L_{min}=L_{max}=3): atc, tca, cag
  - Distance (D_{win}=1): atc-tca, tca-cag
  - Weights (frequency of co-occurrence): atc-tca (1.0), tca-cag (1.0)

Graph functions

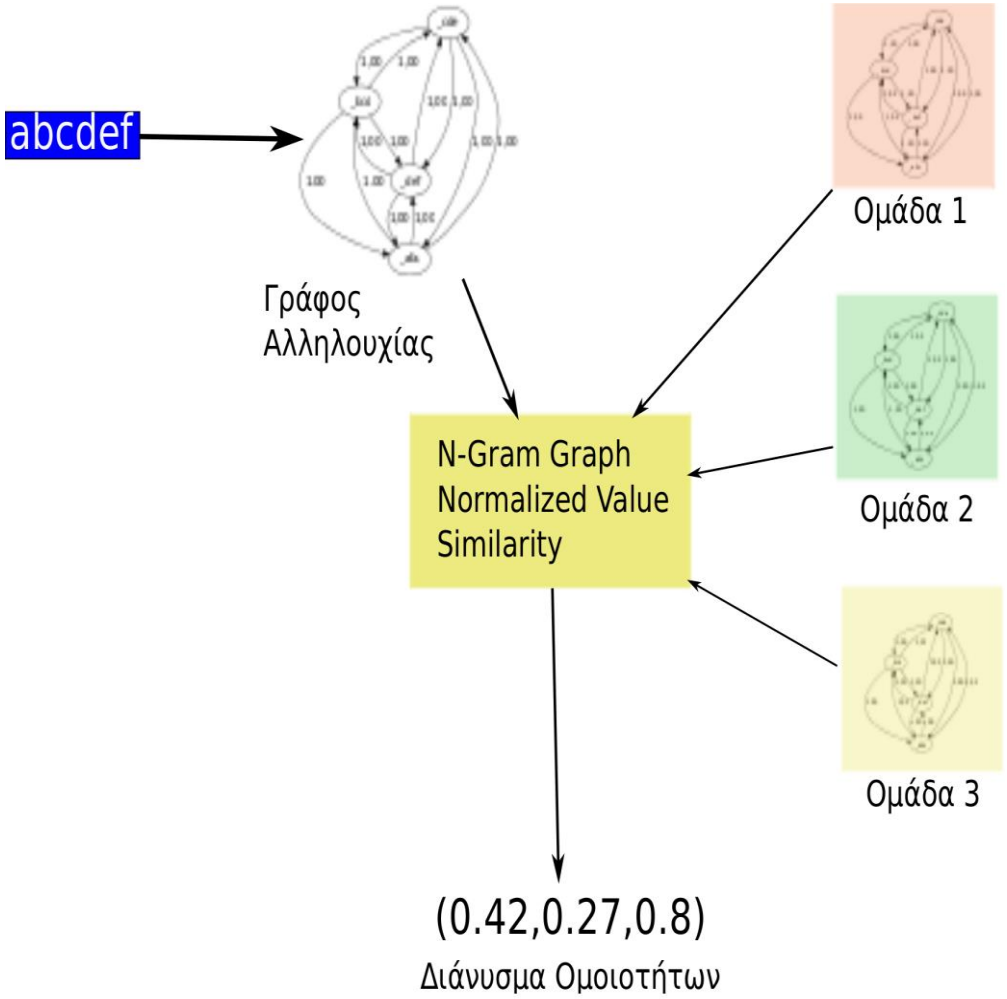
- Intersection (keeping only common edges between two graphs)
- **Merging (creating an average graph assigning mean weights over common edges)**
- Update (update the average graph for more than two cases)
- **Similarity (quantified comparison of two graphs)**

Operators allow:

- The representation of categories of symbol series in a single graph (Merging)
- Comparison of categorical graphs (Similarity)

Representation of sequence categories with n-gram graphs

- Each sequence yields a NGG
- Each functional category is merged into one average NGG
- Each sequence NGG is then described on the basis of similarity with the average NGG of different categories



Highlighting features in common and differences between NGGs and GS

	N-gram Graphs	Genomic Signatures
Description of co-existence / neighborhood	✔	✖
Arbitrary Fuziness	✔	✖
Lossy compression	✔	✖
Similarity measurements	✔	✖
Histogram of co-occurences	✔	✔
Influenced by GC content	✔	✖

NGGs vs…

- **Bag of words** More information
- **Probabilistic sequential models** Not probabilistic (by default), not strictly sequential
- **Automata** No explicit transitions
- **Neural Networks** No input / output; just representation

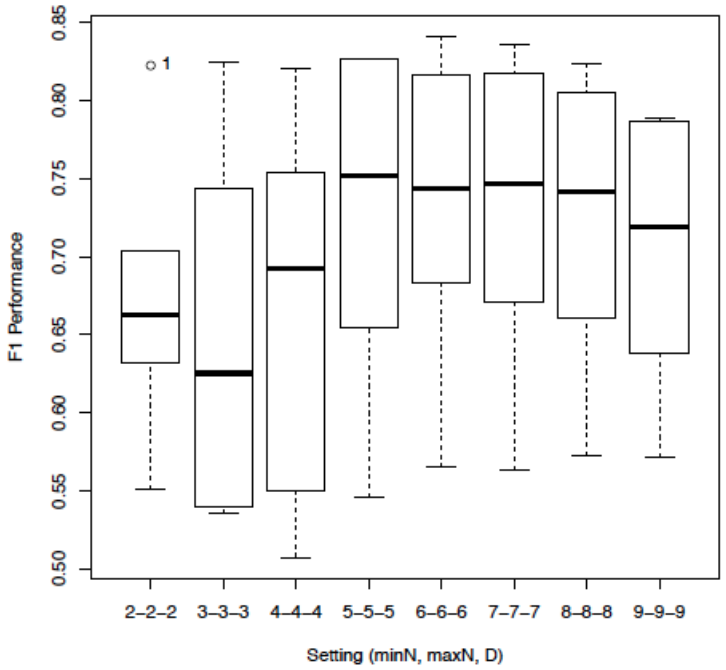
Experimental Setup: Datasets

- UltraConserved Noncoding Elements (UCNEs)
- EU100 nonexonic CNEs (EU100nx CNEs)
- Amniotic and Mammalian CNEs
- Worm and Insect UCNEs

Experimental Setup

- 26 pairwise experiments involving comparisons between different classes of sequences with NGG and GS approaches.
- Output of the analysis (similarity vectors) used as input vectors to train a well-known, rule-based classifier: JRIP (RIPPER)
- Similar results with other classifiers, such as Random Forest, SVM, etc
- 10-fold cross validation
- Performance measured by means of the F-measure
- Based on all experiments performed, we propose 5-5-5 as the best parameter settings for analyzing short genomic sequences.

Parameters tuning



- m: minimum length of the n-grams into which a sequence is split, M: maximum length of the n-grams into which a sequence is split, D: maximum distance within which we consider the n-grams (n-bases) to be neighbors.
- Keeping m=M=D simplifies the analysis step reducing the required time, while not significantly altering the results.

Results: Inter-Species Comparisons of Background Sequences

- Q: Are N-gram Graphs efficient in distinguishing bulk genomic sequences from different genomes?
- A: Yes, but not as efficient as Genomic Signatures.

Results: Inter-Species Comparisons of Background Sequences

Exp	Description	Average length	Average GC	NGG	GS
#1	surrogates for human exons	167.837	0.5155	83.86	85.98
	surrogates for worm exons	169.318	0.4049		
#14	surrogates for human UCNEs	86.094	0.3651	79.38	84.05
	surrogates for insect UCNEs	86.582	0.3949		
#20	surrogates for insect exons	169.318	0.5202	80.48	87.49
	surrogates for human exons	169.816	0.5087		
#22	surrogates for worm exons	213.365	0.4194	73.50	70.35
	surrogates for insect exons	212.858	0.5194		
#23	surrogates for human UCNEs	82.932	0.3648	80.35	83.75
	surrogates for worm UCNEs	82.875	0.4297		
#13	surrogates for worm UCNEs	83.407	0.4265	58.79	64
	surrogates for insect UCNEs	86.582	0.3949		
	Average			**76.06**	**79.27**

- Surrogate sequences: Same GC% and length with the examined sequences.
- Comparisons involving *H.sapiens* yield always the best classification rates.
- This might be understood on the grounds of the high difference of neighbor preferences (CpG and TpA) between human and invertebrates.

Results: Classification of Constrained DNA Sequences vs their background surrogates

- Q: Could we distinguish Constrained Elements from the bulk of the same genome?
- A: Yes and NGGs perform considerably better.

Results: Classification of Constrained DNA Sequences vs their background surrogates

Exp	Description	Average length	Average GC	NGG	GS
#2	worm exons	213.365	0.4243	57.7	61.05
	surrogates	213.365	0.4239		
#3	human exons	169.816	0.5190	74.3	63.41
	surrogates	169.816	0.5183		
#17	insect exons	388.822	0.5412	52.36	59.45
	surrogates	381.557	0.5389		
#4	worm UCNEs	82.875	0.4309	56.76	55.62
	surrogates	82.875	0.4308		
#5a	human UCNEs	326.923	0.3676	82.63	72.00
	surrogates	326.923	0.3676		
#5b	human EU100nx CNEs	155.499	0.3783	76.43	63.75
	surrogates	155.499	0.3783		
#5c	amniotic CNEs	289.061	0.3756	78.62	63.00
	surrogates	289.061	0.3756		
#5d	mammalian CNEs	246.488	0.4015	75.85	55.65
	surrogates	246.488	0.4018		
#12	insect UCNEs	86.582	0.3949	64.15	62.65
	surrogates	86.582	0.3949		
	Average			**68.76**	**61.84**

- Surrogate sequences: Same GC% and length with the examined sequences
- six last rows: CNEs more conserved than exons (they also bear overlapping TFBS)

Results: Classification of Functional Sequences Between Genomes

- Q: To which extent are CNEs or exons from different organisms distinguishable?
- A: They may be distinguished quite clearly, with the GS performing better (as usually) in inter-genomic classification tasks.

Results: Classification of Functional Sequences Between Genomes

Exp	Description	Average length	Average GC	NGG	GS
#9	worm exons	169.318	0.3886	74.68	82.21
	human exons	169.816	0.5190		
#10	worm UCNEs	83.113	0.4277	77.77	82.29
	human UCNEs	82.640	0.3728		
#15	worm UCNEs	83.086	0.4285	70.89	74.95
	insect UCNEs	86.582	0.3949		
#16	human UCNEs	86.094	0.3704	82.08	86.70
	insect UCNEs	86.582	0.3949		
#19	insect exons	169.318	0.5148	70.03	81.29
	human exons	169.816	0.5089		
#21	worm exons	213.365	0.4196	71.39	72.35
	insect exons	212.858	0.5093		
	Average			**74.47**	**79.97**

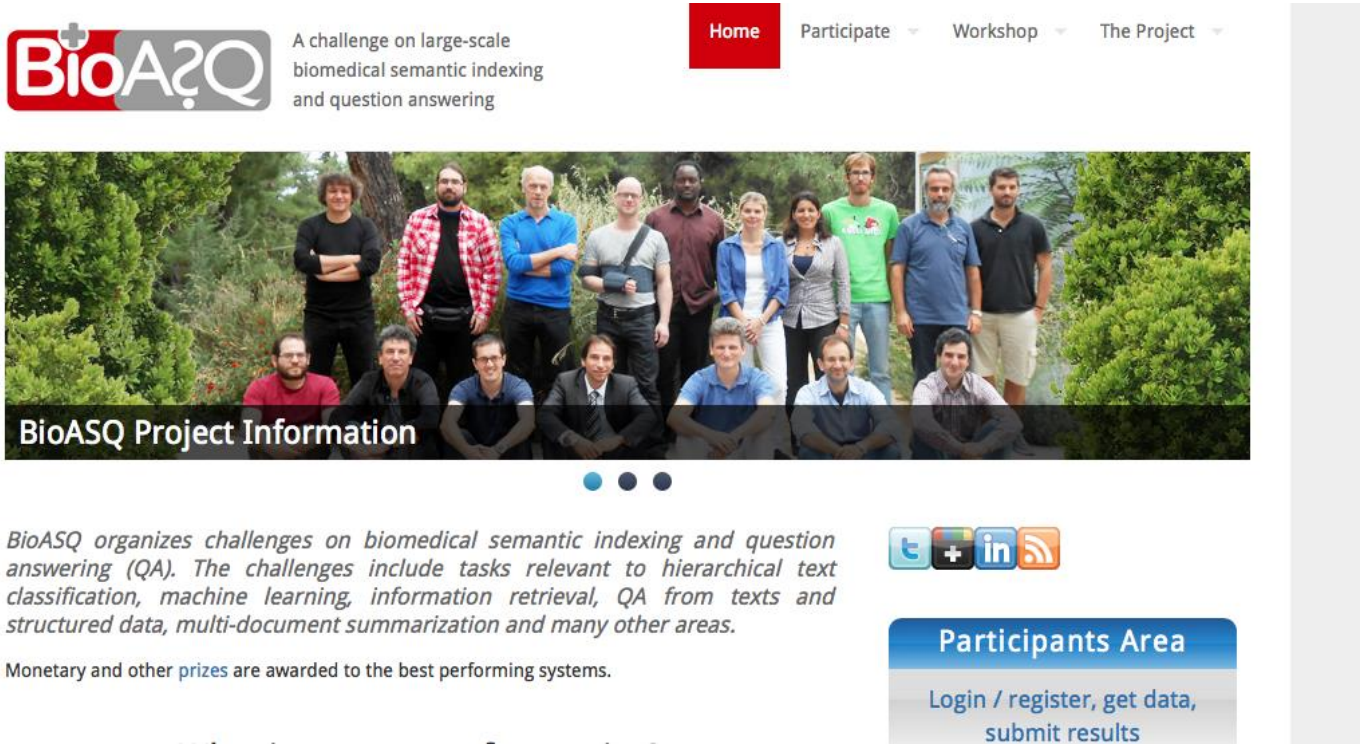
Conclusions

- First time application of the method of NGGs (initially designed for summaries extraction) to the field of genomics.
- First time application of Karlin's method of GS to the classification of short genomic elements (< 50 kb) and of functional importance rather than whole genomes.
- GS perform particularly well when it comes to inter-species comparisons.
- NGGs exceeds GS performance in intra – species comparisons, discriminating with high efficiency genomic sequences with expected function against the bulk of the genome.

Future Work

- Combination of the two methods (NGGs and GS) is expected to yield better results in terms of classification rates.
- Comparison with other methods such as HMMs widely used in topics such as gene finding.
- Application of the NGG framework to the field of analysis of metagenomes.

Thank you



EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

EMBO

1st International Conference on Algorithms for Computational Biology, AlCoB 2014
Tarragona, Spain, July 1-3, 2014