Graph-Based Genomic Signatures

Lenwood S. Heath

CS Research Seminar

Department of Computer Science

Virginia Tech
January 22, 2010
Outline

1. Introduction
2. Previous Signatures
3. Assumption
4. Evaluating a signature
5. Graph-based signatures
   - de Bruijn graphs
   - The $\theta^{oed}$ signature
   - The $\theta^{dbc}$ signature
6. Results
   - Empirical Results
   - Theoretical Results
7. Conclusions and Future Work
A *genomic signature* is a mathematical structure $\theta(H)$ derived from the genomic sequence $H$ of a genome $G$ such that

- $\theta(H)$ is efficiently computable
- $\theta(H)$ is sufficiently well-conserved in $G$
- $\theta(H)$ is sufficiently different from $\theta$ signatures computed from genomic sequences of other genomes
- $\theta(H)$ can be stored much more efficiently than $H$
Modeling genomic sequences computationally

- DNA alphabet $\Sigma_{\text{DNA}} = \{A, C, G, T\}$
- Genomic sequence $H = \text{String over } \Sigma_{\text{DNA}}^*$
- A genome $\mathcal{G} = \text{A set of one or more genomic sequences}$
DNA words

1. DNA word: A string in $\Sigma_{\text{DNA}}^*$
2. Reading DNA words in a DNA sequence using a sliding window
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window
DNA word: A string in \( \Sigma_{\text{DNA}}^* \)

Reading DNA words in a DNA sequence using a sliding window
**DNA words**

- DNA word: A string in $\Sigma_{DNA}^*$
- Reading DNA words in a DNA sequence using a sliding window

**ACCGTTAAGG**
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window

ACCGTTAAGG
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window

ACCGTTAAGG
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window

ACCGTTAAGG
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window

ACCGTTAAGG
DNA words

- DNA word: A string in $\Sigma_{DNA}^*$
- Reading DNA words in a DNA sequence using a sliding window

$S = \text{ACCGTTAAGGACAGTTCTTAAACCCGGGCT}$

With $\Sigma = \{A, C, G, T\}$ and $w = 2$, the lexicographic order is

\[\langle AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT \rangle.\]

and the word count vector is

\[\langle 3, 3, 2, 0, 1, 3, 1, 2, 1, 1, 2, 2, 2, 1, 0, 3 \rangle.\]
DNA words

- DNA word: A string in $\Sigma_{\text{DNA}}^*$
- Reading DNA words in a DNA sequence using a sliding window

$$S = \text{ACCGTTAAGGACAGTTCTTAAACCCGGGCT}$$

With $\Sigma = \{A, C, G, T\}$ and $w = 2$, the lexicographic order is

$$\langle AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT \rangle.$$ 

and the word count vector is

$$\langle 3, 3, 2, 0, 1, 3, 1, 2, 1, 1, 2, 2, 2, 1, 0, 3 \rangle.$$
Existing genomic signatures ...

- **Word count vector** $\theta^{wcv}$
  - Used by Deschavanne et al. 1999, Fertil et al. 2005
  - Previously analyzed by Heath and Pati, 2007
  - Applications: TETRA

- **The dinucleotide odds ratio** $\theta^{dor}$
  \[
  \rho_{XY}(H) = \frac{f_{XY}(H)}{f_X(H)f_Y(H)},
  \]
  - Uses $\delta$-distance to compute variation
  - Performs better than word count vectors
  - Conserved over a wide range of sequence lengths

- Chaos game representation (CGR) image signatures
Assumption

- Each genome $G$ is generated by a characteristic discrete-time, finite-state Markov chain $\mathcal{M}_G$.
- A discrete-time, finite-state Markov chain is a stochastic process with states $\{X_i\}$ at discrete, consecutive time-points of observation $i = 0, 1, \ldots$, respectively. The states take on values from a finite state space $S = \{x_1, x_2, \ldots, x_n\}$, while following the Markov property as described below:

$$
\Pr[X_{i+1} = x_{i+1} \mid X_i = x_i, X_{i-1} = x_{i-1}, \ldots, X_0 = x_0] = \Pr[X_{i+1} = x_{i+1} \mid X_i = x_i].
$$
Evaluating a signature

- Grading criteria
  - Performance for short sequences (few kilobases)
  - Comparison with other signatures
  - Differentiating between far-away species as well as species that are closely-related

- So, why do we need other signatures?
  - No sound mathematical framework for above results
  - Existing signatures fail for short sequences
  - Satisfactory power to discriminate is lacking

Graph-based signatures
- de Bruijn graphs
- The $\theta^{\text{occ}}$ signature
- The $\theta^{\text{disc}}$ signature

Results
- Empirical Results
- Theoretical Results

Conclusions and Future Work
Separation between signatures of sequences generated by the same Markov chain
Separation between signatures of sequences generated by the same Markov chain

\[ |\theta(H_1) - \theta(H_2)| < d \]
\[ |\theta(H_2) - \theta(H_3)| < d \]
\[ |\theta(H_1) - \theta(H_3)| < d \]
Separation between signatures of sequences generated by the same Markov chain

\[ \Pr[|\theta(H_1) - \theta(H_2)| > d] \leq f(M) \]
\[ \Pr[|\theta(H_2) - \theta(H_3)| > d] \leq f(M) \]
\[ \Pr[|\theta(H_1) - \theta(H_3)| > d] \leq f(M) \]
Separation between signatures of sequences generated by different Markov chains

\[
|\theta(H_1) - \theta(H_2)| > d_1 \\
|\theta(H_2) - \theta(H_3)| > d_2 \\
|\theta(H_1) - \theta(H_3)| > d_3
\]
Separation between signatures of sequences generated by different Markov chains
Separation between signatures of sequences generated by different Markov chains

\[ P(|\theta(H_1) - \theta(H_2)| > d) \geq g(M_1, M_2) \]
\[ P(|\theta(H_2) - \theta(H_3)| > d) \geq g(M_2, M_3) \]
\[ P(|\theta(H_1) - \theta(H_3)| > d) \geq g(M_1, M_2) \]
Graph-based signatures
de Bruijn graphs

- Alphabet $\Sigma$, Order $w$
- Vertex set of a de Bruijn graph of order $w$ is $\Sigma^w$
- Edge set of a de Bruijn graph of order $w$ is $\Sigma^{w+1}$

Figure: The de Bruijn graph $DB^3_{\Sigma_B}$ with the sequence $S = 0001110111000101$ defining a walk on it.
A de Bruijn chain (DBC) ... is a de Bruijn graph that includes a finite Markov chain.
Edge deletion cycle

Minimum edge count = 1067
Edges deleted = 0
Vertices isolated = 0
Vertex isolation order = {}
Number of connected components = 1
Edge deletion cycle

Minimum edge count = 1067
Edges deleted = 0
Vertices isolated = 0
Vertex isolation order = {}
Number of connected components = 1
Genomic Signatures

Lenwood S. Heath

Introduction

Previous Signatures

Assumption

Evaluating a signature

Graph-based signatures

de Bruijn graphs

The $\theta_{ooe}$ signature

The $\theta_{abc}$ signature

Results

Empirical Results

Theoretical Results

Conclusions and Future Work

Virginia Tech
Introduction
Previous Signatures
Assumption
Evaluating a signature
Graph-based signatures
de Bruijn graphs
The $\theta_{oct}$ signature
The $\theta_{dcb}$ signature
Results
Empirical Results
Theoretical Results
Conclusions and Future Work

Minimum edge count = 35000
Edges deleted = 2
Vertices isolated = 0
Vertex isolation order = {}
Number of connected components = 1
Genomic Signatures

Lenwood S. Heath

Introduction

Previous Signatures

Assumption

Evaluating a signature

Graph-based signatures

de Bruijn graphs
The $\theta_{oed}$ signature
The $\theta_{abc}$ signature

Results

Empirical Results
Theoretical Results

Conclusions and Future Work

Virginia Tech

**Edge deletion cycle**

- Minimum edge count = 45897
- Edges deleted = 3
- Vertices isolated = 0
- Vertex isolation order = {}
- Number of connected components = 1
**Edge deletion cycle**

Minimum edge count = 51772
Edges deleted = 4
Vertices isolated = 0
Vertex isolation order = {}
Number of connected components = 1

![Graph with labeled vertices and edges showing edge deletion cycle](image-url)
**Edge deletion cycle**

Minimum edge count = 54000
Edges deleted = 5
Vertices isolated = 0
Vertex isolation order = {}
Number of connected components = 1

- Edge deletion cycle diagram with vertices and edges labeled.
**Graph-based signatures**

- de Bruijn graphs
- The $\theta^{odd}$ signature
- The $\theta^{abc}$ signature

**Results**

- Empirical Results
- Theoretical Results

**Conclusions and Future Work**
Edge deletion cycle

Minimum edge count = 56743
Edges deleted = 7
Vertices isolated = 1
Vertex isolation order = {111}
Number of connected components = 2
Edge deletion cycle

Minimum edge count = 64532
Edges deleted = 8
Vertices isolated = 2
Vertex isolation order = {111, 010}
Number of connected components = 3
Minimum edge count = 73423
Edges deleted = 9
Vertices isolated = 2
Vertex isolation order = {111, 010}
Number of connected components = 3
Edge deletion cycle

Minimum edge count = 75614
Edges deleted = 10
Vertices isolated = 2
Vertex isolation order = \{111, 010\}
Number of connected components = 4
Edge deletion cycle

Minimum edge count = 89776
Edges deleted = 11
Vertices isolated = 2
Vertex isolation order = \{111, 010\}
Number of connected components = 4
Edge deletion cycle

Minimum edge count = 99876
Edges deleted = 12
Vertices isolated = 2
Vertex isolation order = \{111, 010\}
Number of connected components = 4
Minimum edge count = 100123
Edges deleted = 13
Vertices isolated = 3
Vertex isolation order = \{111, 010, 001\}
Number of connected components = 5
Edge deletion cycle

Minimum edge count = 156432
Edges deleted = 14
Vertices isolated = 4
Vertex isolation order = \{111, 010, 001, 011\}
Number of connected components = 6
Edge deletion cycle

Minimum edge count = 159988
Edges deleted = 15
Vertices isolated = 6
Vertex isolation order = \{111, 010, 001, 011, 000, 100\}
Number of connected components = 7
Minimum edge count = NA
Edges deleted = 16
Vertices isolated = 8
Vertex isolation order = \{111, 010, 001, 011, 000, 100, 101, 110\}
Number of connected components = 8
Ordered vertex-based edge deletion vector signature $\theta^\text{oed}_3$

The ordered vertex-based edge deletion vector $\theta^\text{oed}$ is the $4^w$-vector whose $i^{th}$ component is the total number of edge deletions required to isolate the vertex $x_i$, where $x_i$ is the $i^{th}$ element of $\sum^w$ in lexicographic order.
Comparison of $\theta^w_{\text{wcv}}$, $\theta^o_{\text{oed}}$, and $\theta^d_{\text{dor}}$: Far-away species

**Figure:** Performance of $\theta^w_{\text{wcv}}$, $\theta^o_{\text{oed}}$, and $\theta^d_{\text{dor}}$ in sequence identification of random 50 Kb segments taken from the species on the x-axis.
**Comparison of $\theta^w_{2c}, \theta^o_{2ed},$ and $\theta^d_{cor}$:**

Closely-related species

**Figure:** Performance of $\theta^w_{2c}, \theta^o_{2ed},$ and $\theta^d_{cor}$ in sequence identification of random 50 Kb segments taken from the species on the x-axis.
Observations

- Summary: $\theta^{wcv}$, $\theta^{dor}$: word frequency based, $\theta^{oed}$: graph based
- 20 $\alpha$-proteobacterial genomes
- Comparable performances

\[
\begin{array}{ccc}
\theta^{oed}_2 > \theta^{wcv}_2 : 11 & \theta^{oed}_2 < \theta^{wcv}_2 : 8 & \theta^{oed}_2 = \theta^{wcv}_2 : 1 \\
\theta^{oed}_2 > \theta^{dor}_2 : 10 & \theta^{oed}_2 < \theta^{dor}_2 : 9 & \theta^{oed}_2 = \theta^{dor}_2 : 0
\end{array}
\]
Ordered vertex isolation frequencies $\theta^{ovif}$: Improving $\theta^{oed}$

Graph-based signatures
- de Bruijn graphs
- The $\theta^{oed}$ signature
- The $\theta^{abc}$ signature

Results
- Empirical Results
- Theoretical Results

Conclusions and Future Work

Virginia Tech
Genomic Signatures

Lenwood S. Heath

Introduction

Previous Signatures

Graph-based signatures

Assumption

Evaluating a signature

Graph-based signatures

De Bruijn graphs

Results

Empirical Results

Theoretical Results

Conclusions and Future Work

Virginia Tech
Using the stationary distribution

Slight perturbations in the probability transition matrices cause the stationary distribution to be perturbed by negligible amounts, 

\[ |\pi - \pi'| < \epsilon, \epsilon \to 0. \]

Origin prediction using the $\theta^{\text{dbc}}$ signature

DNA of unknown origin

Sequencing

DBC signature

Graph-based signatures
de Bruijn graphs
The $\theta^{\text{co}}$ signature
The $\theta^{\text{dbc}}$ signature

Database of barcodes

Predicted target

Conclusions and Future Work

Results
Empirical Results
Theoretical Results

Graph-based signatures
de Bruijn graphs
The $\theta^{\text{co}}$ signature
The $\theta^{\text{dbc}}$ signature

Introductory Signatures
Results
Databases used for experiments

- $\mathcal{D}^{DIV}_{12}$: Database of 12 diverse eukaryotic and bacterial species
- $\mathcal{D}^{APB}_{67}$: Database of 67 $\alpha$-proteobacterial species
- $\mathcal{D}^{DIV}_{50}$: Database of 50 diverse species consisting of 10 archaeal, 20 bacterial, and 20 eukaryotic genomic sequences
- $\mathcal{D}^{GPB}_{77}$: Database of 77 $\gamma$-proteobacterial species
Accuracy of the $\theta_{2}^{dbc}$ signature: Self hits vs. Other hits

Identifies originating organism accurately even for sequence sizes as low as 10 Kb.

Figure: Database used: $D_{12}^{DIV}$. The 12 species are on the x-axis. The small box and whisker plots near the top (with associated circles) represent the distribution of correlations of $\theta_{2}^{dbc}$s of the 100 samples with the $\theta_{2}^{dbc}$ of their origin. The larger box and whisker plots represent the distribution of correlations of $\theta_{2}^{dbc}$s with $\theta_{2}^{dbc}$s of other genomes.
Accuracy of first hits of the $\theta_{2}^{\text{dbc}}$ signature
Comparison of accuracy of different signatures using $D_{50}^{D_{IV}} + D_{77}^{GPB}$
Comparison of accuracy of different signatures using $D_{50}^{DIV} + D_{77}^{GPB}$

Results
Empirical Results

Graph-based signatures
de Bruijn graphs
The $\theta_{ocw}$ signature
The $\theta_{dcw}$ signature

Conclusions and Future Work
Separation between $\theta^{dbc}_2$ signatures of sequences generated by the same DBC

\[ \Pr \left[ d \left( \theta^{dbc}_1, \theta^{dbc}_2 \right) > 64\tau \right] < \]

\[ 2 \cdot \sum_{\beta \in S^2} (L^\pi (\beta) + U^\pi (\beta)) + 2\varsigma^2 \sum_{\beta \in S^2} \left( L^{ovif} (\beta) + U^{ovif} (\beta) \right). \]
\[ L^\pi(x) = \exp \left( \frac{T^2}{2n\pi x} \right) \]

and

\[ U^\pi(x) = \left( \frac{T}{e^{n\pi x}} \right)^{n\pi x} \left( 1 + \frac{T}{n\pi x} \right)^{1 + \frac{T}{n\pi x}} . \]

\[ L^{ovif}(\beta) = e^{-n\pi \beta} \left( \exp \left( \exp \left( -8\tau^2 \frac{\pi \beta}{\pi \alpha} \right) (n\pi \beta) \right) - 1 \right) \]

and

\[ U^{ovif}(\beta) = e^{-n\pi \beta} \left( \exp \left( \left( \frac{4\tau^4 \pi \beta}{\pi \alpha} \right)^{\frac{\pi \alpha}{\pi \beta}} \left( 1 + \frac{4\tau \pi \beta}{\pi \alpha} \right)^{1 + \frac{4\tau \pi \beta}{\pi \alpha}} \right) (n\pi \beta) \right) - 1 \]
Separation between $\theta_2^{dbc}$ signatures of sequences generated by different DBCs

\[ \Pr \left[ d \left( \theta_1^{dbc}, \theta_2^{dbc} \right) \geq 2 \cdot 16\tau \right] \geq 1 - \Pr \left[ d \left( \theta_1^{dbc}, \mathbb{E} \left[ \theta_1^{dbc} \right] \right) \geq 2 \cdot 16\tau \right] - \Pr \left[ d \left( \theta_2^{dbc}, \mathbb{E} \left[ \theta_2^{dbc} \right] \right) \geq 2 \cdot 16\tau \right]. \]
Conclusions and Future Work

Conclusions
- Compared existing signatures
- Explored graph-based signatures
- Integrated features into new improved signature demonstrably better than existing signatures
- Characterized signatures within a mathematical framework

Future work
- Further explore sequence size requirements
- Explore higher order signatures using subsets of state space
- Compare phylogeny calculated from signatures with that calculated from alignment of 16s rRNA


Acknowledgements

- Amrita Pati
- NSF ITR grant