GRAPH-BASED GENOMIC SIGNATURES

Amrita Pati, Virginia Tech, apati@vt.edu

Problem and Motivation

A ***genomic signature*** is a mathematical structure, typically a vector of numbers, which is used to denote unique characteristics of a DNA sequence. A ***DNA sequence*** can be defined as a long string over the alphabet Σ ={A,C,G,T} of bases Adenine (A), Guanine (G), Cytosine (C), and Thymine (T), that constitute DNA. The ***genome G*** of a species is defined as the set of all its chromosomes. Each chromosome is a genomic sequence, as is any subsequence of it. Let S be a sufficiently long (greater than a few kilobases (kb)) genomic sequence taken from genome *G*. Then, the genomic signature $\theta(S)$ of S is a vector of numerical quantities computed from the manner in which characters from Σ constitute S such that:

- $\theta(S)$ is efficiently computed,
- $\theta(S)$ is significantly similar to the signature $\theta(S')$ of another genomic sequence S' taken from the same genome *G*, for sufficiently long S',
- $\theta(S)$ is significantly different from the signature $\theta(S'')$ of a genomic sequence S'', taken from a different genome *G′*, and
- $\theta(S)$ requires much less space to store than S itself.

The species from which a DNA sequence is derived is called the ***origin*** of that sequence. In genomic research, scientists often encounter short DNA sequences of unknown origin. A genomic signature that is highly conserved within the genome of a species and differs between genomes of different species is extremely useful in the identification of such short DNA segments. For a given short sequence S , $\theta(S)$ is computed and then matched with an existing database of θ signatures of all known species with sequences genomes using a suitable similarity measure. The species with the closest signature to $\theta(S)$ is predicted as the origin of S. Other species having signatures with high similarities to $\theta(S)$ are predicted as close relatives. In the absence of any close signatures in the database, the discovery of a new species is hypothesized. The application of a good genomic signature to the short sequence origin identification problem can be effectively used for metagenomics, infectious microbial sequence identification, global cataloging of species, and building of a star-trek-like genome tricorder with scientific, industrial, and domestic applications. While several genomic signatures have been defined in the scientific literature, none of them have demonstrated high accuracy in predicting the origins of short DNA sequences.

In this work, we propose the ***de Bruijn chain (DBC) signature*** θ^{dbc}, a powerful genomic signature computed by treating a genomic sequence as a walk over a de Bruijn chain. Information from the structure of the de Bruijn graph and the properties of the underlying Markov chain is then extracted into numerical quantities that constitute the θ^{dbc} signature. Fig. 1 illustrates the use of DBCs in origin prediction of short DNA sequences.

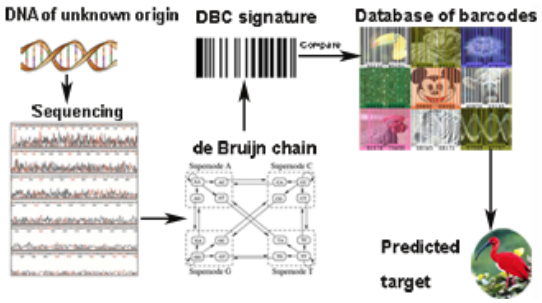

Fig. 1. Using de Bruijn chains to predict the origin of short DNA sequences.

We demonstrate using both theoretical and empirical results that the θ^{dbc} signature is information-rich, efficient, sufficiently representative of the sequence from which it is derived, and superior to existing genomic signatures such as the dinucleotide odds ratio and word frequency based signatures. We develop a mathematical framework to elucidate the power of the θ^{dbc} signature to distinguish between sequences hypothesized to be generated by DBCs of distinct parameters.

Background and Related Work

A ***DNA word*** of length *w* is a string over Σ^w. Two other sequence-based genomic signatures have been proposed in the scientific literature. These are the ***word count vector*** θ^{wcv} (WCV) [1] and the ***dinucleotide odds ratio vector*** θ^{dor} (DOR) [2], respectively. Let $\mathcal{S}^w$ denote the set of all strings in Σ^w. For strings *x* and *y* with |*x*| <= |*y*|, *occ(x, y)* is the count of occurrences of *x* as a substring of *y*. The frequency of *x* in *y* is *freq(x,y)=occ(x,y)/(|y|-|x|+1)*. Given a word length *w* and a genomic sequence *H*, the corresponding WCV signature is a 4^w-long vector with i^{th} component equal to $occ(x_i,H)$, where x_i is the i^{th} string in lexicographic order in Σ^w. A DNA word of length *2* is called a ***dinucleotide***. Given *H*, the DOR signature is a *16*-long vector with the entries for dinucleotides in lexicographic order. The entry corresponding to the dinucleotide *XY* is computed by the expression *freq(XY,H)/(freq(X,H)*freq(Y,H))*. Although the θ^{wcv} and θ^{dor}

signatures have been studied with respect to conservation within a species and variation between species, their effectiveness in identifying short DNA sequences accurately has not been studied systematically. Also, a formal mathematical framework within which separation between the above signatures for different species is characterized, is lacking in the literature.

Uniqueness of the Approach

The de Bruijn chain (DBC) signature was proposed by us in [3]. The ***order-w de Bruijn graph*** $\mathcal{DB}^w = (\mathcal{S}^w, E)$ over alphabet Σ is a directed graph, where $(x_i, x_j) \in E$ when $x_i\sigma = \iota x_j$, for some $\sigma, \iota \in \Sigma$; such an edge is labeled σ. Fig. 2 depicts a binary de Bruijn graph of order *3*. Let *H* be a genomic sequence of length *n*. We think of *H* as tracing a walk in $\mathcal{DB}^w = (\mathcal{S}^w, E)$. The ***edge count*** of edge $(x_i, x_j) \in E$ in *H*, where $x_i\sigma = \iota x_j$, is $\mathrm{occ}(x_i\sigma, H)$. Now consider the Markov chain underlying the above de Bruijn graph $\mathcal{DB}^w$. The said Markov chain has state space $\mathcal{S}^w$ and a sparse transition probability matrix with nonzero transition probabilities only for edges in $\mathcal{DB}^w$; such a Markov chain is called an ***order-w de Bruijn chain (DBC)***. Assuming that all DBCs computed from genomic sequences are ergodic and hence that there is a unique ***stationary distribution*** $\pi = (\pi_i)$ on $\mathcal{S}^w$ satisfying $\pi P = \pi$, for word length *w>= 1*, we obtain $\mathcal{DB}^w(H)$ with associated edge counts. Let $\psi \geq 0$ be a positive integer ***threshold***. Let $E^{\leq\psi} = \{(i,j) \in E \mid \mathrm{ec}((i,j), H) \leq \psi\}$ be the set of edges with counts at most ψ. Then ***edge deletion*** is the process of deleting edges $E^{\leq\psi}$ from $\mathcal{DB}^w(H)$ while varying ψ from *0* to $\Xi = \max\{\mathrm{ec}((i,j), H) \mid (i,j) \in E\}$ and deleting edges with tied counts in arbitrary order. As ψ increases from *0* to Ξ, the number of isolated vertices increases from *0* to 4^w. The ***ordered vertex isolation frequency vector (OVIF)*** θ^{ovif} is the 4^w-vector whose i^{th} component is the frequency of the last edge whose deletion isolates the vertex labeled with x_i, the i^{th} string in lexicographic order. The ***de Bruijn chain signature*** θ^{dbc} is the $2 \cdot 4^w$-vector $\pi_w \cdot \theta_w^{ovif}/4^{w-1}$, where π_w is the estimated stationary distribution for the order-*w* DBC and '·' represents vector concatenation. Fig. 3 depicts the construction of the order-*2* DBC signature.

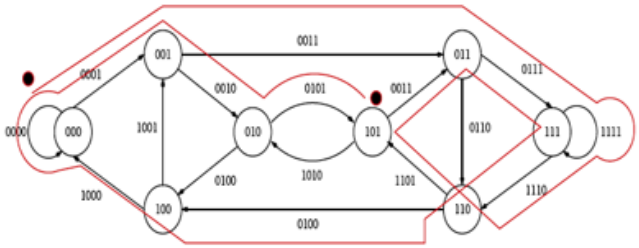

Fig. 2. de Bruijn graph of order *3* over the binary alphabet *{0,1}* The red line depicts the walk traced by the sequence *0001110111000101* on the graph.

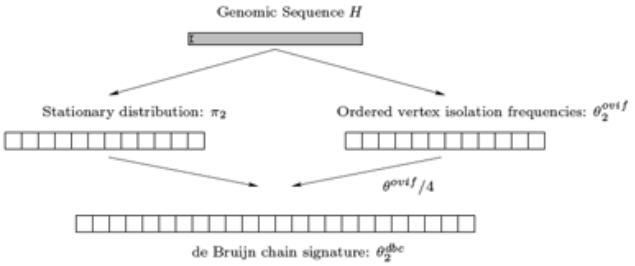

Fig. 3. Construction of the DBC signature of order *2*.

Results and Contributions

Including both theoretical and empirical results that we obtain, we make the following contributions:

1. ***Within a probabilistic mathematical framework, we prove that the separation between the*** θ_2^{dbc} ***signatures of sequences generated by the same DBC is small with high probability, while the separation between the*** θ_2^{dbc} ***signatures of sequences generated by different DBCs is large with high probability [5].***

   Theorem 1. Let H_1 and H_2 be genomic sequences of length *n* generated independently by the same order-*2* DBC with underlying stationary distribution $\bar{\pi}$. Let $\hat{\pi}^1$ and $\hat{\pi}^2$ be the order-*2* stationary distributions derived from the respective transition matrices of H_1 and H_2. Assume that the number of occurrences of a dinucleotide *x* has a Poisson distribution with mean $n\hat{\pi}_x$. Then for $\tau > 0$ and $T = n\tau$,

$$\Pr\left[d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) > 64\tau\right] < 2 \cdot \sum_{\beta \in \mathcal{S}^2} \left(\mathcal{L}^{\pi}(\beta) + \mathcal{U}^{\pi}(\beta)\right) + 2\varsigma^2 \sum_{\beta \in \mathcal{S}^2} \left(\mathcal{L}^{ovif}(\beta) + \mathcal{U}^{ovif}(\beta)\right),$$

   where,

and

$$\mathcal{L}^{\pi}(x) = \exp\left(\frac{-T^2}{2n\pi_x}\right) \text{ and } \mathcal{U}^{\pi}(x) = \left(\frac{e^{\frac{T}{n\pi_x}}}{\left(1+\frac{T}{n\pi_x}\right)^{1+\frac{T}{n\pi_x}}}.\right)^{n\pi_x}$$

$$\mathcal{L}^{ovif}(\beta) = e^{-n\pi_\beta}\left(\exp\left(\exp\left(-8\tau^2\frac{\pi_\beta}{\pi_\alpha}\right)(n\pi_\beta)\right)-1\right) \text{ and}$$

$$\mathcal{U}^{ovif}(\beta) = e^{-n\pi_\beta}\left(\exp\left(\left(\frac{e^{\frac{4\tau\pi_\beta}{\pi_\alpha}}}{\left(1+\frac{4\tau\pi_\beta}{\pi_\alpha}\right)^{1+\frac{4\tau\pi_\beta}{\pi_\alpha}}}\right)^{\frac{\pi_\alpha}{\pi_\beta}}(n\pi_\beta)\right)-1\right).$$

Through empirical simulations, we show that the individual expressions $\mathcal{L}^{\pi}(x)$, $\mathcal{U}^{\pi}(x)$, $\mathcal{L}^{ovif}(\beta)$, and $\mathcal{U}^{ovif}(\beta)$ are extremely small numbers close to zero, and their sum, which is the R.H.S. of the bound presented in Theorem 1, is a very small probability as well.

Theorem 2. Let H_1 and H_2 be genomic sequences of length n generated independently by different order-*2* DBCs. Let θ_1^{dbc} and θ_2^{dbc} be their respective DBC signatures. Let $\hat{\pi}^1$ and $\hat{\pi}^2$ be the respective stationary distributions and let θ_1^{ovif} and θ_2^{ovif} be the respective OVIF signatures computed from H_1 and H_2. Then assuming that $d\left(\mathbf{E}\left[\hat{\pi}^1\right], \mathbf{E}\left[\hat{\pi}^2\right]\right) > 3\cdot 16\tau$ and $d\left(\mathbf{E}\left[\theta_1^{ovif}\right], \mathbf{E}\left[\theta_2^{ovif}\right]\right) > 3\cdot 16\tau$, the distance $d\left(\theta_1^{dbc}, \theta_2^{dbc}\right)$ distinguishes between the two DBCs generating H_1 and H_2, and can be bounded as

$$\Pr\left[d\left(\theta_1^{dbc}, \theta_2^{dbc}\right) \geq 2\cdot 16\tau\right] \geq 1 - \Pr\left[d\left(\theta_1^{dbc}, \mathbf{E}\left[\theta_1^{dbc}\right]\right) \geq 2\cdot 16\tau\right] - \Pr\left[d\left(\theta_2^{dbc}, \mathbf{E}\left[\theta_2^{dbc}\right]\right) \geq 2\cdot 16\tau\right].$$

Through empirical simulations, we show that the negative bounds on the R.H.S. are very small numerical quantities and hence, the R.H.S. is evaluated to a large probability.

2. ***We demonstrate empirically that an algorithm using the order-2 DBC signatures can predict the origin of short DNA sequences with high accuracy while distinguishing between both far-away and closely-related species signatures in a pre-computed database effectively [3,4].***

We propose two datasets, L_1: a list of *50* diverse species including archaea, bacteria, and eukaryotes, and L_2: a list of *74* closely related γ-proteobacterial species, to serve as standard datasets in comparing accuracies of all available sequence-based genomic signatures in the literature. The intention of the first list is to test the ability of a signature to distinguish between diverse species while the intention of the second list is to test the ability of a signature to distinguish between closely-related species. First, the order-*2* DBC signatures were computed for all species in both lists L_1 and L_2, and stored in a database D. We test the accuracy of the DBC signatures in identifying the origin of short DNA sequences using DNA sequences of lengths 5 kb, 10 kb, 25 kb, 50 kb, and 100 kb. For each of the five lengths listed above, from each species in both lists L_1 and L_2, *100* subsequences of that length are randomly sampled. The order-*2* DBC signature of each subsequence is computed and compared with all signatures in the database D using the Pearson correlation coefficient. The species whose signature matches with the highest correlation is the predicted target. The number of predicted targets among *100* sequences, which correspond to the true origin is the ***accuracy***. Fig. 5 illustrates the accuracy of the order-*2* DBC signature using short DNA sequences sampled from the species on the *x*-axis, for different sample lengths. Observe that the DBC signature is extremely well-conserved (high accuracy) among archaea (the first 10 species on the *x*-axis), bacteria (the next 20 species on the *x*-axis), and most eukaryotes (the last 20 species on the *x*-axis). As the sample length goes down, so does the accuracy, as expected. Even at sequences as small as 10 kb, a median accuracy greater than 80% is observed.

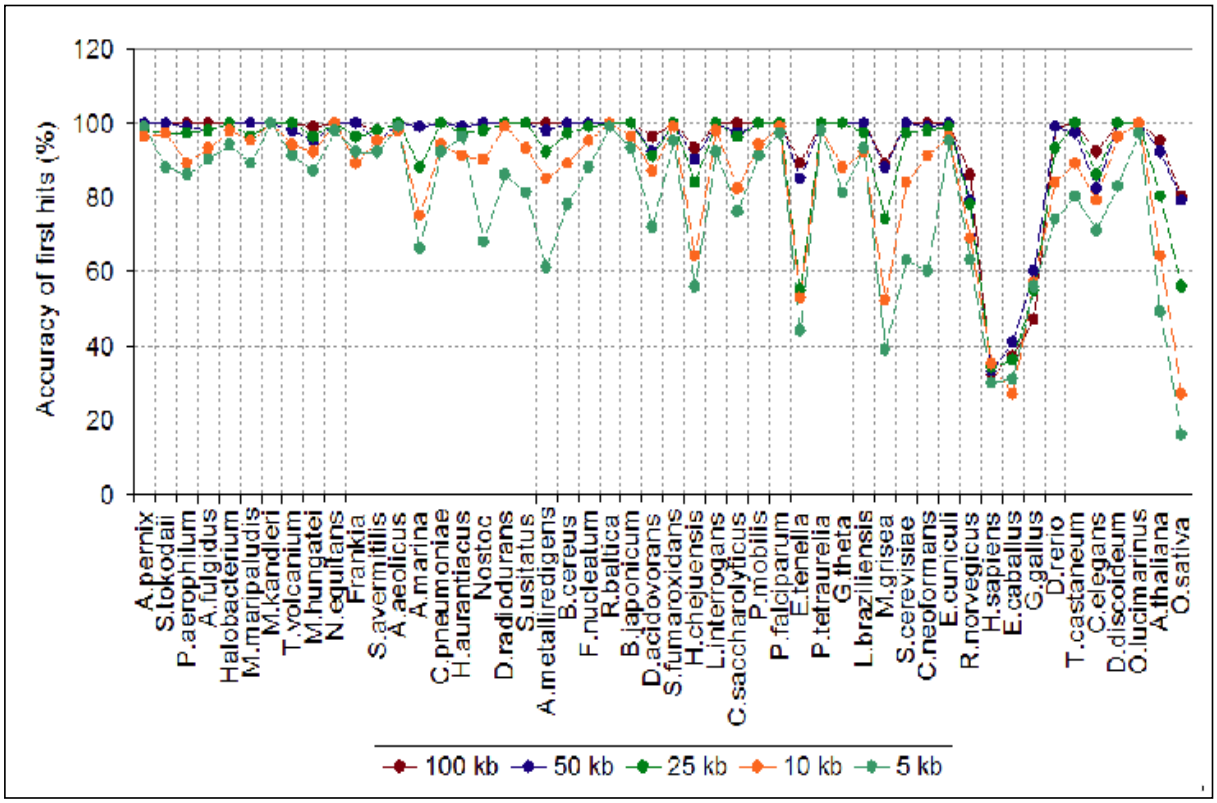

Fig. 5. Accuracy of origin prediction of order-2 DBC signatures for DNA sequences of various lengths.

3. ***We demonstrate empirically that the order-2 DBC signature can predict the origin of short DNA sequences with much higher accuracy than the order-2 DOR and WCV signatures, and the margin by which the accuracy of the DBC signature is higher increases with decreasing length of the DNA sequence [5,6].***

In Fig. 5, we plot the median accuracies of origin identification among all species in list L_1 for each signature, while in Fig. 6, we do the same for all species in list L_2. The ***combo*** signature is the combination of the DBC and DOR signatures. Observe that the median accuracy of the DBC signature is higher than the median accuracies of the DOR and WCV signatures for all sequence lengths, and especially so for shorter sequences. As the sequence length decreases, the margin by which the DBC signature outperforms the DOR and WCV signatures increases. Overall, the combination of the DBC and DOR signatures performs the best, and demonstrates the highest median accuracy.

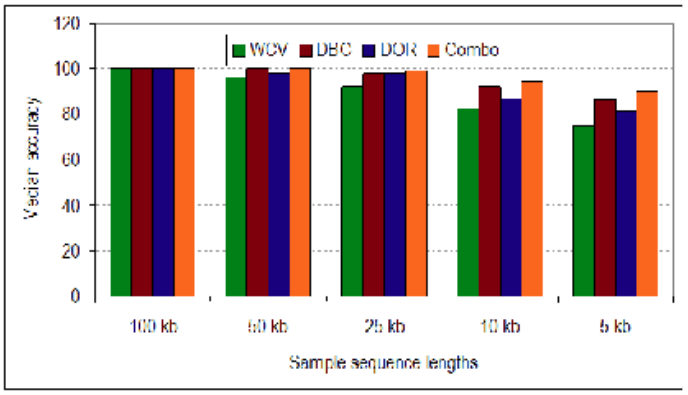

Fig. 5. Median accuracies of DBC, DOR, WCV, and combo signatures using diverse species.

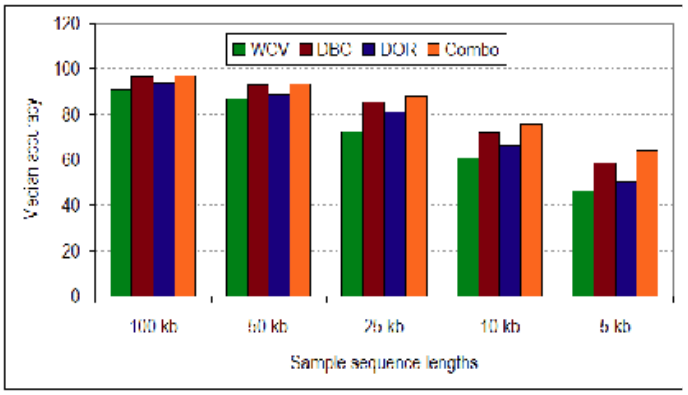

Fig. 6. Median accuracies of DBC, DOR, WCV, and combo signatures using closely-related species.

Conclusions and Future Work

We have examined genomic signatures from the point of view of accurate identification of the origin of short unknown DNA sequences. The genomic signatures introduced in this paper are derived from the structure and properties of de Bruijn chains. When a sample

sequence is sufficiently long, the target organism for the sample can be retrieved by querying a database of signatures. Given an unknown DNA sequence, its possible high-level location in the phylogenetic tree can be predicted using the combination of the DBC and DOR signatures, after which its origin and closest relatives can be predicted using the DBC signature alone. We have demonstrated both theoretically and empirically that the DBC signature is a powerful signature, able to efficiently identify the origin of an unknown genomic sequence as short as a few kilobases. This implies that the origin and the closest relatives of an unknown sequence can be identified with very little actual sequencing. We also observed the effect of order on efficiency of the DBC signature. In continuing work, we are exploring the effect of size of the signature database on short sequence target prediction efficiency. We are also studying the phylogeny implied by distances between DBC signatures and the extent to which this phylogenetic structure is conserved on random sampling of short sequences for phylogenetic reconstruction. A software system that predicts the origin using signatures computed from available genomic sequences of species in NCBI's taxonomy database is under construction.

References

1. 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 Biology and Evolution, 16(10):1391-1399, 1999.
2. S. Karlin and C. Burge. Dinucleotide relative abundance extremes: A genomic signature. Trends in Genetics, 11(7):283-290, 1995.
3. Lenwood S. Heath and Amrita Pati. Genomic signatures from DNA word graphs. In Lecture Notes in Bioinformatics, volume 4463, pages 317-328. Springer-Verlag, 2007.
4. Lenwood S. Heath and Amrita Pati. Genomic signatures in de Bruijn chains. In Lecture Notes in Bioinformatics: Algorithms in Bioinformatics (WABI 2007), volume 4645, pages 216-227. Springer-Verlag, 2007.
5. Lenwood S. Heath and Amrita Pati. Computing genomic signatures using de Bruijn chains. Submitted to IEEE/ACM TCBB.
6. Lenwood S. Heath and Amrita Pati. On the accuracy of origin prediction: A comparison of sequence-based genomic signatures. Submitted to Bioinformatics.