MORE POWERFUL DISCRIMINANTS FOR CLASSIFYING PHYLOGENETIC SIGNALS IN DINUCLEOTIDE FREQUENCIES

Robert H. Baran[1] *Changwon Jeon*[1] *David K. Han*[2] *Hanseok Ko*[1]

[1]Department of Electronics and Computer Engineering, Korea University, Seoul, Korea
hsko@korea.ac.kr

[2]United States Naval Academy, MD, USA

ABSTRACT

Microbial DNA fragments are classified according to species using compositional features and "genomic signatures" the oldest of which is the dinucleotide relative abundance profile defined by Karlin et al. More informative features, including higher order signatures, have demonstrated greater species-specificity in comparison to the baseline established by the dinucleotide signature using "delta-distance" to assess dissimilarity; but lack of standard methods has precluded rigorous comparison. We describe a new method for classifier evaluation that reduces any number of pair-wise inter-genomic comparisons to a single performance measure. To illustrate the method, we compare delta-distance to quadratic and linear discriminants prescribed by elementary pattern recognition theory, and find that the quadratic form is significantly more powerful.

Index Terms: Biomedical signal processing, DNA, Error analysis, Pattern classification, Software performance.

1. INTRODUCTION

Pre-genomic investigations found that dinucleotide relative abundance values are fairly constant in the DNA of a given microbial species and more highly variable between species. As complete prokaryotic genome sequences became available in the 1990s, this phenomenon was carefully studied by Karlin and co-workers, who developed it as a basis for phylogeny construction. The dinucleotide relative abundance profile was called a "genomic signature" by Karlin and Burge [1] because an organism can generally be identified by computing it from any 50 kilobase (kb) or longer segment of the genome sequence. A suitable measure of dissimilarity between the signatures of two whole genomes provides a useful measure of their evolutionary distance as confirmed by a recent survey of 334 prokaryotes in which it was found to be essentially concordant with more standard phylogenetic measures like 16S ribosomal DNA identity [2].

Dinucleotide relative abundance is computed as follows. When a DNA strand of length n is scanned in one direction, there are n_{xy} transitions (base steps) from x to $y \in$ {A, G, C, T}, and $f_{xy} = n_{xy}/(n-1)$ is the normalized frequency of dinucleotide xy. Scanning the complementary strand in the reverse direction produces $f_{xy}^{(-c)}$. The 4x4 matrix of elements $f_{xy}^{*} = f_{xy} + f_{xy}^{(-c)}$ exhibits counter-diagonal symmetry when the bases are indexed alphabetically, as {A, C, G, T} ↔ {1, 2, 3, 4}, by Watson-Crick base pairing. Dividing by the product of the marginal frequencies gives $\rho_{xy}^{*} = f_{xy}^{*}/(f_x^{*} f_y^{*})$ in the usual notation [1]. These 16 quotients comprise the dinucleotide relative abundance profile, $\boldsymbol{\rho}^{*}$, but six of them are redundant. The dissimilarity measure introduced by Karlin et al. is the dinucleotide relative abundance distance ("delta-distance") between sequences G and H,

$$\delta^{*}(G,H) = \delta^{*}(H,G) = (1/16)\sum_{i,j=1}^{4} |\,\rho_{ij}^{*}(H) - \rho_{ij}^{*}(G)\,|.$$

When G and H are the complete genomes of species A and B, respectively, the delta-distance can be taken as a monotonic (increasing) function

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2008**	2. REPORT TYPE	3. DATES COVERED **00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE **More Powerful Discriminants for Classifying Phylogenetic Signals in Dinucleotide Frequencies**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Electronics and Computer Engineering,Korea University,Seoul, Korea, ,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

See also ADM002091. Presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Held in Las Vegas, Nevada on March 30-April 4, 2008. Government or Federal Purpose Rights License

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**4**	

of the time since their divergence from a common ancestor.

The ability to resolve genomic signatures in DNA sequences shorter than 50 kb would serve some current objectives including the detection of bacterial genes acquired from foreign species by horizontal gene transfer [3] and the classification of fragments that have been directly sequenced from the environment in metagenomic exploration [4]. While the dinucleotide signature pervades its genome on scales down to 1 kb and less [5], the "phylogenetic signal [6]" that it represents is typically too weak for reliable identification in single genes and gene-sized fragments. The average size of protein-coding genes in most prokaryotes is around 1 kb [7]. Because genomic signatures are indistinct on this small scale, more powerful discriminants are needed to reliably detect foreign genes in a known genome or to associate unknown genes with genome sequences that are under construction.

The search for better discriminants has led investigators to species-specific features in codon usage [3], in the base compositions at the three codon positions, and in higher order genomic signatures [6,8] obtained from frequencies of overlapping (frame-independent) short oligonucleotides of length >2. This search proceeds without a "ground truth" list of all the foreign genes in any single species and without a standard dataset and metric for assessing the accuracy of a fragment-to-genome classifier. Progress in this field produces fragment classifiers and foreign gene detectors that perform above the baseline level that is attained with the dinucleotide signature; but this improvement is merely relative because the baseline is not an absolute benchmark.

While classifiers and detectors based on higher order compositional features have been the focus of considerable research, the optimality of δ^* for recognizing dinucleotide signatures in short genomic segments has never really been asserted. It is impossible say what functional form is best in the absence of a generally accepted stochastic model. Many practical problems in signal detection and pattern classification have likewise been approached without a model of the data source and this situation often motivates the use quadratic discriminant analysis. The quadratic discriminant would be optimal if the components of $\boldsymbol{\rho}^*$ are normally distributed—perhaps after a suitable nonlinear transformation—and the quadratic form would reduce to a linear form only if the covariance matrix were approximately constant for all species. Finding no previous comparison of this kind, we formulate the problem in the next section. After that, we demonstrate a method for assessing the accuracies of alternative discriminants based on a manageable number of pair-wise intergenomic comparisons.

2. METHODS

For any query sequence (gene or fragment) q and any genome G, we define:

> $\boldsymbol{v}(q)$ = a row vector of ten components, the non-redundant elements of $\ln[\boldsymbol{\rho}^*(q)]$;
>
> $\boldsymbol{\mu}(G)$ = E[$\boldsymbol{v}(g)$], where the expectation operator E takes the average over all contiguous fragments g of genome G;
>
> $[\boldsymbol{v}(g) - \boldsymbol{\mu}(G)]$ = a matrix with 10 columns and one row for each contig of G;
>
> $\boldsymbol{S}(G) = \mathrm{E}\{[\boldsymbol{v}(g) - \boldsymbol{v}(G)]^{\mathrm{T}}[\boldsymbol{v}(g) - \boldsymbol{v}(G)]\}$, in which the superscript ($^{\mathrm{T}}$) takes the transpose, is a 10x10 covariance matrix; and
>
> $\boldsymbol{S}^{-1}(G)$ = the matrix inverse.

With these definitions, the quadratic discriminant function is

$$d_2(q,G) = [\boldsymbol{v}(q) - \boldsymbol{v}(G)]\boldsymbol{S}^{-1}(G)[\boldsymbol{v}(g) - \boldsymbol{v}(G)]^{\mathrm{T}}$$

in which $\boldsymbol{v}(G)$ consists of the ten non-redundant elements of $\ln[\boldsymbol{\rho}^*(G)]$. When the quadratic form d_2 reduces to a linear form, it can be expressed as a weighted correlation coefficient (corr). For simplicity, we assume equal weights and use the function

$$d_1(q,G) = 1 - \mathrm{corr}[\boldsymbol{v}(q), \boldsymbol{v}(G)]$$

for a sub-optimal approximation. Note $0 \leq d_1 \leq 2$ with smaller values indicating greater similarity.

Two species A and B are selected from the public database and their respective genomes G and H are broken into fragments. For a reproducible experiment, a 1 kb window is displaced by increments of 1 kb across each published sequence, and each displacement produces a fragment g of G (or h of H). The 2-way classifier uses discriminant function d to measure dissimilarity between the fragment and each genome. For each fragment q, the classifier calculates two values of d, assigning q to G if $d(q,G) < d(q,H)$ and vice versa. One error is counted every time $d(g,G) > d(g,H)$ or $d(h,H) > d(h,G)$. The pair-wise classification error rate ε(A,B) is computed as the number of errors divided by the total number of fragments without regard for the unequal sizes of G and H. One-half error is counted when the discriminants are *exactly* equal so that $\varepsilon = ½$ (instead of 1) in the event that $G = H$. But pair-wise classification error rates (CERs) will only be computed for each unordered pair of different genomes.

When two species A and B are randomly selected from the database of k complete genomes, ε(A,B) is a random variable that depends on which of $k(k - 1)/2$ unordered pairs is chosen. We have discriminants d_1 and d_2, possibly involving different compositional features, and corresponding error rates ε_1 and ε_2 are computed for each pair of species. To claim that d_2 is preferable to d_1, it will be sufficient to show that $\varepsilon_1 \geq \varepsilon_2$ for a clear preponderance of pairs, and statistical significance can be assessed with reference to a binomial model. But if ε(A,B) is essentially determined by the evolutionary distance $\mathit{\Delta}$(A,B) then we expect the relation between ε and $\mathit{\Delta}$ to show a clear decreasing trend. Taking $\mathit{\Delta}(A,B) \equiv 1000\delta^*(G,H)$, we obtain a scatter plot in which the decreasing trend is evident. A monotonic transformation of the error rate captures this trend as the slope of a regression line that intersects the origin. Thus the resolving power of the discriminant is expressed as a single number—the (negative) slope of the regression line—which is expected to approach a fixed limit as the number of pair-wise comparisons increases.

3. RESULTS

The baseline classifier takes $\boldsymbol{\rho}^*(q)$ as its feature vectors and uses delta-distances $d_0(q,G) \equiv \delta^*(q,G)$ and $d_0(q,H)$ for discrimination. For pair-wise comparisons among seven selected bacterial species, a subset of those considered in [7], the CERs are plotted against the corresponding values of $\mathit{\Delta}$. This scatter plot of ε versus $\mathit{\Delta}$ (not shown) has a clear decreasing trend but is apparently nonlinear. Since ε estimates a probability, the logistic transformation $y(\varepsilon) = \log(1/\varepsilon - 1)$ is the canonical link for linearizing the data. Note that y is the logarithm of the odds ratio $(1-\varepsilon)/\varepsilon$ and that $y(1/2) = 0$. The transformed scatter plot of Figure 1 shows y versus $\mathit{\Delta}$ using the base 10 logarithm, and the trend is reasonably described by the simple linear regression line or least squares fit. The line is forced through the origin because we must have $\varepsilon = ½$ if $G = H$. The estimated slope of the regression line (x1000) is –10.4 ± 0.4 and the simple correlation coefficient is –0.986.

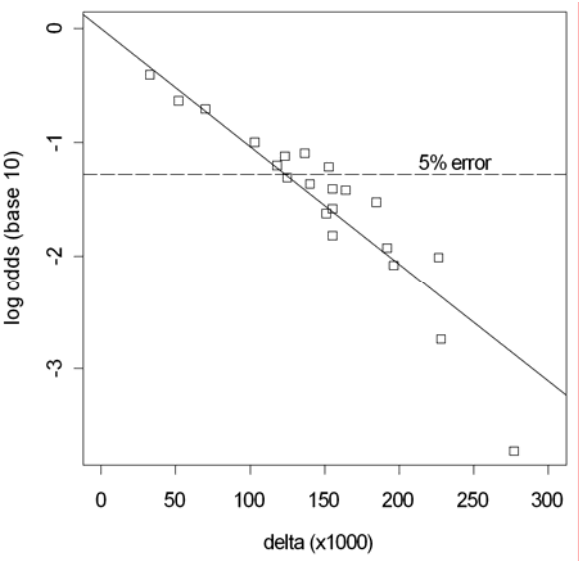

Figure 1. Scatter plot of transformed CER versus $\mathit{\Delta}$ for 21 pair-wise inter-genomic comparisons using (d_0) the delta-distance discriminant.

The CERs ε_2 and ε_1 achieved by the new discriminants are now compared to the baseline results. For the quadratic discriminant, they mostly fall below the baseline; and we find that $\varepsilon_2 < \varepsilon_0$ in 19/21 cases ($p < 10^{-5}$), where the p-value of success rate x/21 is the probability of x or more successes in 21 binomial trials. For the linear discriminant, the results compare unfavorably with the baseline CERs, as $\varepsilon_1 < \varepsilon_0$ in 4/21 cases ($p > 0.99$). Finally, since $\varepsilon_2 < \varepsilon_1$ in 21/21 cases ($p = 0$), these results imply that d_2 is substantially better than d_0 which is better than d_1.

In order to compare discriminants objectively, and reduce performance to a single number, the CERs are transformed by the logit link and plotted against evolutionary distances in Figure 2. Handling the baseline CERs this way produced a least squares fit in Figure 1 which is now copied as a dashed line onto Figure 2. The CERs produced by the new discriminants are fitted by regression lines that straddle the baseline. Their respective slopes (x1000) are (for d_1) −9.3 ± 0.3 and (for d_2) −12.5 ± 0.5, which imply the same rank ordering as the binomial p-values.

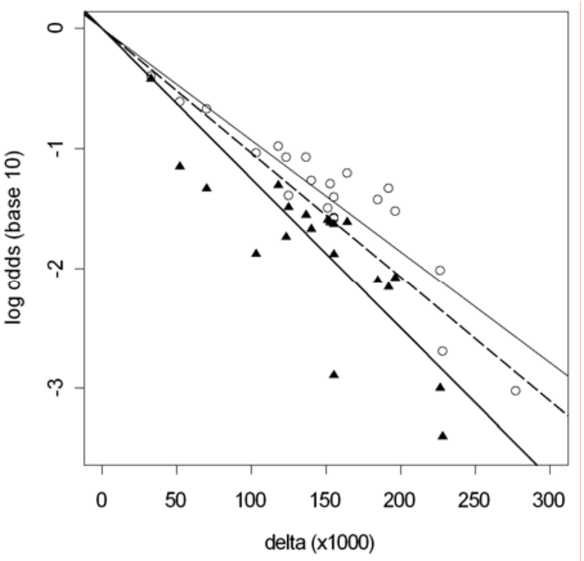

Figure 2. Scatter plots of transformed CERs versus Δ using linear (open circles) and quadratic (solid triangles) discriminants. Baseline data fit the dashed regression line.

When the experiment was repeated for a different set of seven species, the estimated slope parameters were all within 1.5 standard deviations of the stated results. After pooling all 42 data points, the final estimate in each case fell within one standard deviation.

4. DISCUSSION

It has been theorized that higher order genomic signatures are inherently more species-specific than lower order signatures and hence that tetranucleotide frequencies, computed without reference to a reading frame, convey more information than codon usage. The species-specificity of the tetranucleotide signature has been claimed "even in DNA fragments as short as 1 kb [8]." Yet other investigators found that it "works quite well for sequences in the range of 40 kb" but "is certainly not suited for the analysis of single-read end-sequences, which are usually shorter than 1 kb [6]." Such apparently divergent claims may be reconciled by understanding that discrimination accuracy increases with both fragment length and evolutionary distance.

Although this point is generally understood, and previous analyses have stratified the problem accordingly, our new method formalizes and quantifies the dependence more explicitly. In this way it yields consistent performance estimates based on a small fraction of the pair-wise comparisons that can be selected from the growing public database and avoids the tendency to summarize results in terms of averages that fail to generalize from one experiment to another.

REFERENCES

1. S. Karlin and C. Burge (1995), "Dinucleotide relative abundance extremes: a genomic signature," *Trends in Genetics* 11: 283-290.

2. M.W.J. van Passel, E.E. Kuramae, A.C.M. Luyf, A. Bart and T. Boekhout (2006), "The reach of the genome signature in prokaryotes," *BMC Evolutionary Biol.* 6: 84.

3. R.K. Azad and J.G. Lawrence (2007), "Detecting laterally transferred genes: use of entropic clustering methods and genome position," *Nucleic Acids Res.* 35: 4629 - 4639.

4. K. Chen and L. Pachter (2005), "Bioinformatics for whole-genome shotgun sequencing of microbial communities," *PLoS Computational Biol.* 1(2): e24.

5. R.W. Jernigan and R.H. Baran (2002), "Pervasive properties of the genomic signature," *BMC Genomics* 3: 23.

6. H. Teeling, J. Waldmann, T. Lombardot, M. Bauer and F.O. Glockner (2004), "TETRA: a web-based service for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics* 5: 163.

7. R.H. Baran, H. Ko and R.W. Jernigan (2003), "Methods for comparing sources of strand compositional asymmetry in microbial chromosomes," *DNA Res.* 10: 85-95.

8. C. Dufraigne B. Fertil, S. Lespinats, A. Giron and P. Deschavanne (2005), "Detection and classification of horizontal transfers in prokaryotes using genomic signature," *Nucleic Acids Res.* 33(1): e6.