Signal processing and statistical methods in analysis of text and DNA

Matthew J. Berryman[a], Andrew Allison[a], Pedro Carpena[b], and Derek Abbott[a]

[a]Center for Biomedical Engineering and
School of Electrical and Electronic Engineering,
The University of Adelaide, SA 5005, Australia
[b]Departmenta de Fisica Aplicada II,
Universidad de Málaga, E-29071, Spain

ABSTRACT

A number of signal processing and statistical methods can be used in analyzing either pieces of text or DNA sequences. These techniques can be used in a number of ways, such as determining authorship of documents, finding genes in DNA, and determining phylogenetic and linguistic trees. Signal processing methods such as spectrograms provide useful new tools in the area of genomic information science. In particular, fractal analysis of DNA "signals" has provided a new way of classifying organisms.

Keywords: phylogenetic trees, stylography, fractal, DNA sequences

1. INTRODUCTION

The Human Genome Project[1] together with a number of other projects has produced the DNA sequences for a large number of organisms, from humans and mice, to zebrafish, yeast, and over eighty bacteria. There has been a great deal of work done in applying signal processing and statistical methods to DNA recently, and our work has been looking at the application of these methods to not only DNA but in the field of text analysis as well.

A number of interesting statistical techniques have been explored in recent times in the area of text analysis. These attempt to answer questions regarding what the text is about (by extracting relevant keywords)[2] or relationships between texts or languages.[3] In this paper we show significant results of analysis performed on ancient Greek texts, and give a detailed explanation of an F-statistic method for keyword extraction we have been developing.

In the field of DNA analysis, techniques such as the discrete Fourier transform[4] and multifractal analysis[5] have been explored. In this paper we present exciting new applications of those methods to the areas of sequence analysis[6] and phylogenetic trees[7] (those showing the relationships between organisms) respectively.

2. TEXT ANALYSIS

2.1. Introduction to text analysis

In this section we explore new methods from two main areas of text analysis, namely stylography and keyword extraction. Stylography or more general style analysis using statistical methods can highlight differences or similarities between authors, and between languages. Although the idea of using statistical properties of texts in extracting keywords and categorizing texts is not new,[8,9] here we use a new method of analyzing the spacing between words for both determining style similarities and finding keywords. We begin by reviewing the seminal work of Ortuño *et al.*,[2] and then introduce a new keyword extraction technique based on an F-statistic method.

Further author information: (Send correspondence to Derek Abbott)
Derek Abbott: E-mail: dabbbott@eleceng.adelaide.edu.au, Telephone: +61 8 8303 5748

Biomedical Applications of Micro- and Nanoengineering, Dan V. Nicolau, Abraham P. Lee, Editors,
Proceedings of SPIE Vol. 4937 (2002) © 2002 SPIE · 0277-786X/02/$15.00

2.2. Stylography

Ortuño *et al.*[2] suggest using standard deviation of the inter-word spacing to characterize word distributions and extract keywords, as opposed to using a frequency count of each word. By inter-word spacing, we mean the number of words in between successive occurrences of a keyword (non-inclusive), for example if the keyword is "the", then the spacing in "The cat sat on the mat" is three, and the spacing in "The cat is the best cat." is two as there are two and three words respectively between the two occurrences of the word "the". Initial results of plotting the standard deviation of word spacing for (almost) all the words in a given text against the plot of those from other texts by the same author revealed an unexpected result; namely that works by the same author have a similar distribution of inter-word spacings. A striking result is obtained when we plot the gospels of *Matthew* and *Luke*, and the book of *Acts* from the *Koine* Greek New Testament. This suggests a statistical approach to stylography could be taken, and we demonstrate the use of this for both texts by known authors, and for texts where the historical authorship is unclear. The following method can be used to give an indication of the similarity in style between two texts. We conjecture that this can give a valuable insight into the authorship of texts.

Given a set of word spacings $\{x_1, \ldots, x_n\}$ for a given word, we compute the scaled standard deviation of word spacings

$$\hat{\sigma} = \frac{1}{\bar{x}} \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}. \tag{1}$$

We repeat this calculation for all the words in a text, giving us a set of standard deviations $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_m\}$. In order to generate the graphs we then rank these $\hat{\sigma}_j$ in order and plot standard deviation vs. $\log_{10}(\text{rank})$. We omit those words occurring five or fewer times as being statistically insignificant. The technique of using inter-word spacing in this way and plotting the ranked standard deviations on a logarithmic scale was suggested by Carpena *et al.*,[10] however here we show some striking similarities between graphs of standard deviations of words in texts by the same author. Figure 1 shows the result of applying this to the gospels of *Matthew* and *Luke*, and the book of *Acts*. Note that we have used the *Koine* Greek sources for the New Testament[11] to eliminate any changes in style due to translation. Figure 2 shows the similarity between works by Charles Dickens (*Great Expectations* and *Barnaby Rudge*) and works by Thomas Hardy (*Tess of the d'Urbervilles* and *Far From the Madding Crowd*). This illustrates that the technique is not restricted to only Greek texts.

2.3. Keyword extraction

Here we detail the method used by Ortuño *et al.*[2] and introduce a new method for extracting keywords.

As per the standard deviation graphs, we determine the set of standard deviations of word spacings for all the words in a text, $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_m\}$. Again, we rank the words from highest standard deviation from highest to lowest, but keeping all the words. We thus obtain a list of words ranked from high relevance to low relevance.

Another method we have examined for keyword extraction is using the F-statistic on the word spacings, assuming a geometric distribution. The F-statistic detects word-spacing with excess variance (relative to a maximal-entropy or "geometric" distribution). The F-statistic behaves asymptotically like a Gaussian random variable (when the number of inter-word spacing samples is large) with mean of 0 and variance of 1 so the statistical tests for relevant keywords are very easy. Given the set of word spacings $\{x_1, \ldots, x_n\}$ we use the F-statistic

$$\frac{1}{2} \ln(n) \left(\frac{s^2}{\bar{x}(1+\bar{x})} - 1 \right), \tag{2}$$

where s is the normal sample standard deviation. Note the similarity between the F-statistic and the square of the scaled standard deviation. Assuming the null hypothesis then we get a maximum likelihood estimate of the parameter a in the geometric pdf $p(x) = (1-a)(a^x)$, and can hence estimate the variance of the distribution and we compare this with the standard unbiased estimator for variance. The $\log(n)$ term is for scaling (to deal with the accuracy of the F-statistic for different sample sizes). Other terms are corrections as detailed in Abramowitz and Stegun.[12]

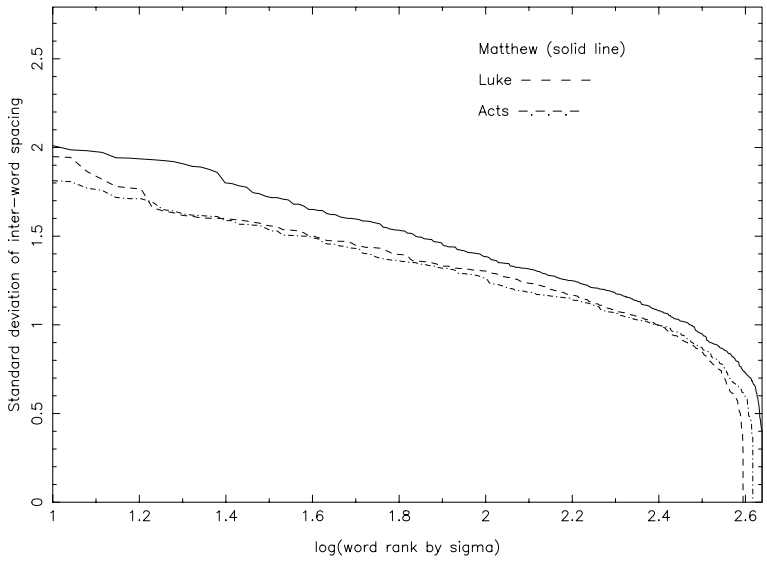


Figure 1. The scaled standard deviation of the inter-word spacing (y-axis) for each word is ranked in descending order on a logarithmic scale (x-axis). Using the original *Koine* Greek text, a remarkably close match is obtained between the gospel of *Luke* and the book of *Acts* in the New Testament, which were written by the same author. For reference, a curve of a different author is shown (the book of *Matthew*) illustrating a distinct difference (this is the upper curve). Although the match between *Luke* and *Acts* deviates for a log rank < 1.2, this represents less than four per cent of the total curve (due to the base-ten logarithmic scale). Note that uncommon words occurring less than 5 times in each text are not included in the ranking, as their scaled standard deviation values are not significant.

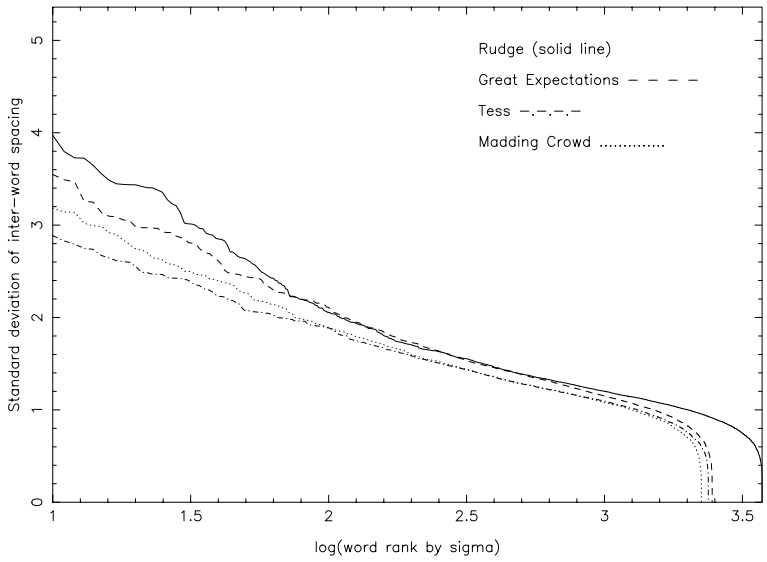


Figure 2. Standard deviation vs. log(rank) for books by Charles Dickens and Thomas Hardy. For much of the length of the plots the graphs for the Dickens texts are nearly coincident, and likewise for the pair of Hardy texts. Although the plots are all quite different for the region 1 to 1.7, note that this apparently large region is quite small due the logarithmic scale on the x-axis. The extra length on the plot for *Barnarby Rudge* is simply due to the larger number of different words in this text when compared with the other three texts.

Table 1. Rankings according to frequency, standard deviation and F-statistic for words in the *The Hound of the Baskervilles.*

Rank	Freq. word	Freq. val.	$\hat{\sigma}$ word	$\hat{\sigma}$ val.	F-stat. word	F-stat. val.
1	the	3327	her	3.23	her	24.02
2	and	1628	hotel	3.11	she	17.83
3	of	1592	she	2.88	we	15.08
4	i	1465	mortimer	2.75	mortimer	14.54
5	to	1448	we	2.44	hotel	13.96
6	a	1306	hound	2.41	you	10.82
7	that	1133	body	2.38	hound	10.18
8	it	979	boot	2.31	charles	6.92
9	he	914	hugo	2.28	body	6.77
10	in	913	alley	2.19	barrymore	6.61
11	you	826	heir	2.15	boot	6.41
12	was	803	cab	2.13	i	6.08
13	his	689	high	2.13	hugo	5.84
14	is	622	miss	2.11	your	5.62
15	have	541	you	2.07	cab	5.38

Table 1 shows words from Sir Arthur Conan Doyle's *The Hound of the Baskervilles.* As can be seen in Table 1, the standard deviation and F-statistic methods are much better than a simple frequency count, which tends only to pick out the conjunctive words.

Qualitative tests need to be carried out to establish whether the standard deviation or F-statistic method performs best at extracting relevant keywords. We are currently working on a web search engine which ranks pages by the score of the keywords the user is searching for (using any of the methods described above or a combination of them).

3. DNA ANALYSIS

3.1. Introduction to DNA analysis

DNA is like text, but instead of coding for human thoughts it codes for the building blocks of all living things. Here we present several new techniques for analyzing regions of DNA and classifying organisms based on their whole genome. For the purposes of our analysis, we simply treat DNA as a long string with letters from an alphabet $A = a, t, c, g$. We then map that alphabet into numerical sequences, and then use standard signal processing and statistical methods to analyze those sequences.

The first method shows how color spectrograms give a visual feel for various properties of the DNA sequences. The second method then explores how multifractal methods and spacing methods like we have used for text analysis above can be used in classifying bacteria.

3.2. Analysis of coding regions in DNA using color spectrograms

Color spectrograms are a useful tool in visualizing aspects of signals occurring in time, for example one can see the noise present in the audio recording of the moon landing and then design an efficient filter to remove it, using only the visual information provided in the spectrogram to determine the noise.

Distinguishing coding from non-coding regions is an important problem in genetics. Others such as Bernaola-Galván *et al.* have explored entropy-based methods for separating the coding from the non-coding regions.[13] Can color spectrograms give a visual guide to these regions? Here we show some results which suggest this is possible.

Anastassiou[4] has explored the use of color spectrograms in analyzing DNA sequences. As detailed further in Anastassiou, for a sequence of bases numbered $1, \ldots, n, \ldots, N$, we can define the following sequences:

$$\begin{aligned} x_r[n] &= \tfrac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n]), \\ x_g[n] &= \tfrac{\sqrt{6}}{3}(u_C[n] - u_G[n]), \\ x_b[n] &= \tfrac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n]), \end{aligned} \tag{3}$$

where $u_X[n] = 1$ if the base at position n is X, zero otherwise. The sequences x_r, x_g, x_b are used in generating red, green, and blue color components of pixels (the squares) in the spectrograms. The mapping of a base at position n, from the set $\{a, t, c, g\}$ onto the sequences x_r, x_g, x_b is done so as to maximize the differences between the sequences at that position n, which results in more vivid colorings of the spectrogram. To color the spectrogram we compute the discrete Fourier transform (DFT)

$$X_c[k] = \sum_{n=0}^{59} x_c[l] e^{-2\pi jnk/60}, \tag{4}$$

for $k = 1, \ldots, 30$ and where c represents one of the color sequences. Note we do not scale the DFT expression in Eq. 4 with a $1/N$ term as usual, as we have to scale the resulting $|X_c[k]|$ into the color component range $\{0, \ldots, 255\}$. We use a DFT block size of 60 as opposed to a power of two (which would enable us to use the FFT algorithm to improve the computational complexity) since 60 has a large number of integer dividers, some of which correspond to common repeat lengths in DNA, for example the frequency $k = 20$ corresponds to the codon length $3 = 60/20$. Repeats of length two and six are also common in sections of DNA, these have frequencies of $k = 60/2 = 30$ and $k = 60/2 = 30$. These repeat lengths thus give rise to center-cell frequencies, so there is no sidelobe leakage for these repeat lengths.

To illustrate the usefulness of this technique in identifying regions of DNA, we show in Figure 3 the color spectrograms of DNA sequences in *Staphylococcus aureus Mu50*[14] and *Homo sapiens.*[1] Figure 3 suggests there are differences in the spectrograms between coding regions of DNA and regions with both coding and non-coding regions. The lack of fine-grained resolution of the spectrograms is problematic, and prevents easy visualization of the borders between coding and non-coding regions.

3.3. Multifractal analysis

A method useful in comparing different organisms is to use a multifractal method. We use the fractal method as detailed by Yu *et al.*[15] namely to each possible substring $s = s_1 \ldots s_k$, $s_i \in A$ of DNA of length K, we assign a unique set $[x_l, x_r)$ given by

$$x_l(s) = \sum_{i=1}^{K} \frac{x_i}{4^i}, \tag{5}$$

where

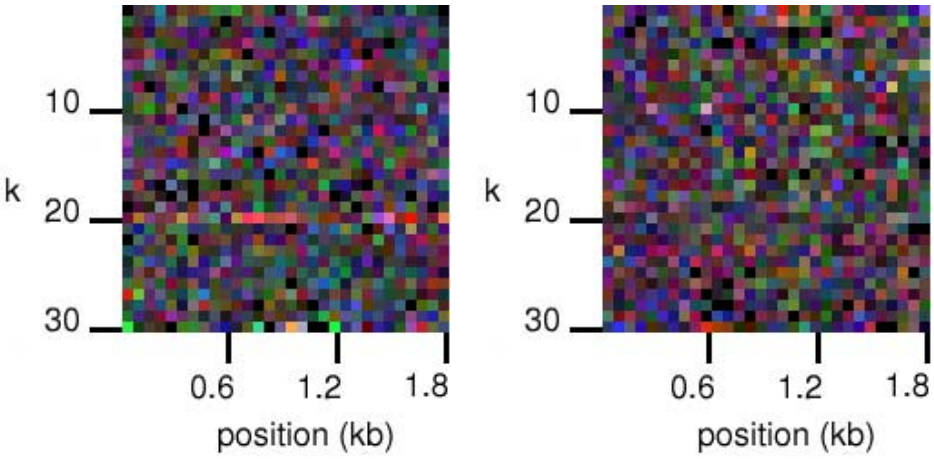
$$x_i = \begin{cases} 0, & s_i = a, \\ 1, & s_i = c, \\ 2, & s_i = g, \\ 3, & s_i = t, \end{cases} \tag{6}$$

and

$$x_r(s) = x_l(s) + \frac{1}{4^K}. \tag{7}$$

Then

$$F(s) = \frac{N(s)}{L - K + 1}, \tag{8}$$



(a) Color spectrogram of a gene in *S. aureus*

(b) Color spectrogram of a gene in *H. sapiens*

Figure 3. Figure 3(a) shows the color spectrogram for a 1.8 kb region in the *gyrA* gene of *S. aureus Mu50*. Note the band at a frequency $k = 20$, indicating a section with codons occurring repeatedly in a long sequence. This differs to Figure 3(b), which shows a 1.8 kb region in the *Q9BZA9* gene of *H. sapiens*, indicating less repetition of codons, perhaps because of the presence of introns.

where $N(s)$ is the number of occurrences of the substring s in the string of length L of the whole genome. The fractal measure is then

$$\mu_K(dx) = Y_K(x)dx, \tag{9}$$

where

$$Y_K(x) = 4^K F_K(s), x \in [x_l(s), x_r(s)). \tag{10}$$

The partition sum is

$$Z_\epsilon(q) = \begin{cases} \sum_{\mu(B)\neq 0} [\mu(B)]^q, & q \neq 1, \\ \sum_{\mu(B)\neq 0} \mu(B) \ln \mu(B), & \text{otherwise.} \end{cases} \tag{11}$$

Here we run over all non-empty boxes $B = [n\epsilon, (n+1)\epsilon)$ where ϵ is $\epsilon = 4^{-K}$ and $n = 1, \ldots, 4^K - 1$. Since $\mu(B) \in \mathbf{R}$ and addition is commutative in the reals, the ordering of the $\mu(B)$ given by Eq. 6 is unimportant in calculating Eq. 11. It is therefore unimportant in calculating the Rényi dimension D_q for $q \in \mathbf{R}$, given by

$$D_q = \begin{cases} \lim_{\epsilon\to 0} \frac{\ln Z_\epsilon(q)}{(q-1)\ln\epsilon}, & q \neq 1, \\ \lim_{\epsilon\to 0} \frac{Z_\epsilon(q)}{\ln\epsilon}, & q = 1 \end{cases} \tag{12}$$

Note that although the method used by Yu *et al.* doesn't show long-range correlations in the DNA sequence, we are considering the information content in the sequence and not the correlations. If you consider the case where $q = 1$, then the Rényi dimension D_1 is the same as the Shannon entropy.[16] As differences in and similarities in G+C content can indicate relationships between organisms,[17] here we are using the Rényi dimensions to determine if this is reflected in a useful way in an uneven distribution of the segments – the ordering is

unimportant here, since we are only comparing the unevenness of the distribution, and not properties relating to the ordering.

The multifractal $D(q)$ plot for *Campylobacter jejuni*[18] is shown in Figure 4. As with the Yu *et al.* we found that a segment size of $K = 8$ works best in classifying bacteria. The near linearity of the $D(q)$ plot around $q = 0$ suggests that we can assign to each bacteria a point in $\mathbf{R}^2$ or $\mathbf{R}^3$ given by (D_{-1}, D_1) or (D_{-1}, D_1, D_{-2}). Yu *et al.* found that phylogenetically close bacteria are close in the two spaces. We use the space (D_{-1}, D_1, D_{-2}) in conjunction with the minimal-span tree algorithm[19] to generate phylogenetic trees in the following subsection.

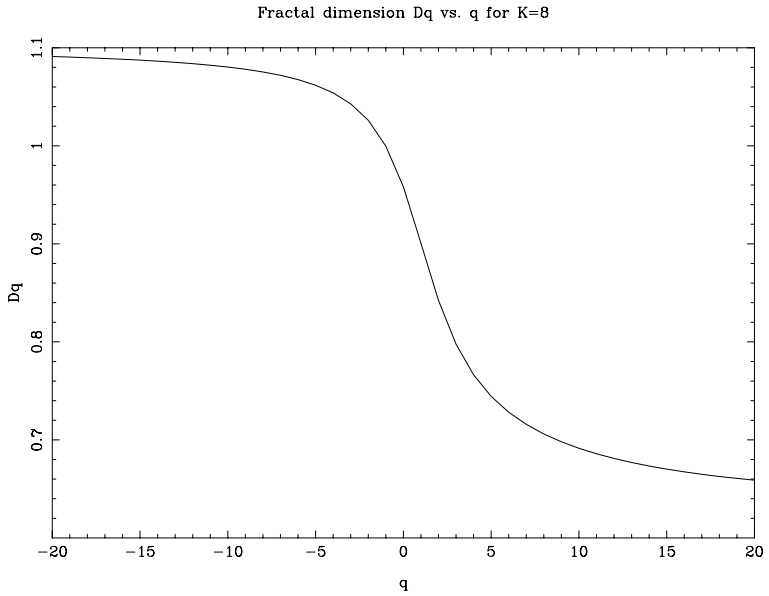


Figure 4. Multifractal Rényi Dimension plot (K=8) for the bacteria *C. jejuni*. Note that the value D_1 is the Shannon entropy of the genome for a symbol size of 8. The graph is relatively linear in the region $[-2, 1]$ which suggest these values of D_q can be used in vectors in a Euclidean space.

3.4. Phylogenetic trees

For each pair $(\mathbf{x}, \mathbf{y})$ of genomes, we compute the vectors in Euclidean $\mathbf{R}^3$ space

$$\mathbf{r_x} = (D_{-1}(\mathbf{x}), D_1(\mathbf{x}), D_{-2}(\mathbf{x}))\,, \tag{13}$$

and

$$\mathbf{r_y} = (D_{-1}(\mathbf{y}), D_1(\mathbf{y}), D_{-2}(\mathbf{y}))\,. \tag{14}$$

Then we compute the metric

$$d^2_{\mathbf{xy}} = \|\mathbf{r_y} - \mathbf{r_x}\|^2 \tag{15}$$

and use this in the minimal-span tree algorithm[19] to generate binary phylogenetic trees. We have used this approach to generate the phylogenetic tree for members of the proteo-bacteria and hyperthermophile families of bacteria as shown in Figure 5(a).

Another method we have been exploring in relation to both text and DNA is a quantitative chi-squared method which computes a metric with lower scores indicating a closer match. Similar to the inter-word spacing technique for text, for DNA we compute a scaled standard deviation of spacing, in this case for codons. For example, the spacing for the codon *gat* in the sequence *gat agg gcg gat* is two. Note we simply break the sequence into groups of three bases, starting at the beginning, to form codons; while not correct in the sense

of the true biology of gene reading we ignore this problem as we are only interested in large scale properties of the sequence. We compute the scaled standard deviations as per Eq. 1.

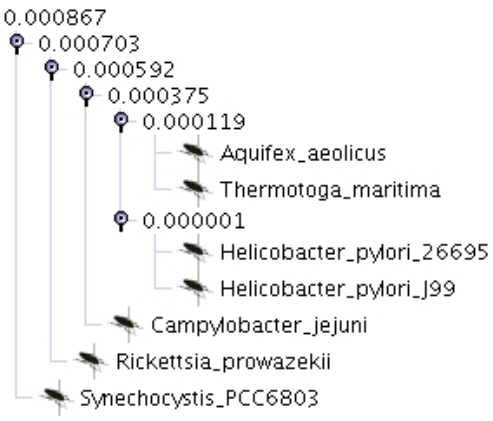
This gives sets of variances of codon spacings for all the genomes, $\{\hat{\sigma}_{11}^2, \ldots, \hat{\sigma}_{I1}^2\}, \ldots, \{\hat{\sigma}_{1J}^2, \ldots, \hat{\sigma}_{IJ}^2\}$, for all possible codons, labelled $i = 1, \ldots, M$ and genomes $j = 1, \ldots, J$. Then we use a formula for χ^2 as given in Kullback[20] for a pair of genomes $(k, l) \in \{1, \ldots, J\} \times \{1, \ldots, J\}$.

$$\chi_{kl}^2 = \frac{1}{N_k N_l} \sum_{i=1}^{I} \frac{\left(N_l \hat{\sigma}_{ik}^2 - N_k \hat{\sigma}_{il}\right)^2}{\hat{\sigma}_{ik}^2 + \hat{\sigma}_{il}^2}, \tag{16}$$

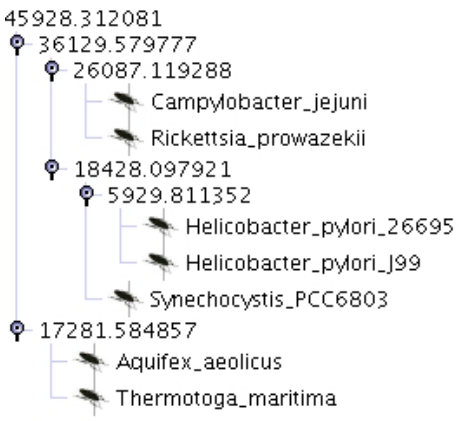
$$N_k = \sum_{i=1}^{I} \hat{\sigma}_{ik}^2, \tag{17}$$

$$N_l = \sum_{i=1}^{I} \hat{\sigma}_{il}^2. \tag{18}$$

We thus generate a set of χ^2 values for each pair of genomes. As with the multifractal metric, the chi-squared values can be combined with the minimal-span tree algorithm to produce a phylogenetic tree. For comparison between the trees generated between the metric, see Figure 5



(a) Phylogenetic tree obtained using the multifractal metric



(b) Phylogenetic tree obtained using the chi-squared metric

Figure 5. The result of applying the minimal-span tree algorithm to the mulitfractal distance metric in Eq. 15 is shown for several members of the proteo-bacteria family in Fig. 5(a). Using the chi-squared metric in Eq. 16 instead results in the tree shown in Fig. 5(b). The miniature bug icons represent the organisms we see today, the circles represent the branches of the tree (where our software thinks the species diverged), and the numbers represent the metric scores used to separate the families of bacteria at that point. Clearly the two *H. pylori*[21] strains group together correctly for both metrics. A comparison with trees obtained by a detailed analysis of proteins,[7] indicates the *Thermatoga maritima*[22] and *Aquifex aeolicus*[23] as also closely related, and indeed these group together in our trees. Of the two trees, the one using the chi-squared metric appears more correct when compared with ones generated from the more usual metrics and tree algorithms used in the study of phylogenetic relationships.[7,24]

4. CONCLUSIONS

We have presented a number of interesting methods for analyzing both text and DNA. The graphs of inter-word spacing standard deviation highlight some interesting results, we are pursuing further work in extracting

relevant features of the graphs. As we have mentioned, a web search engine is currently under development to give a qualitative analysis of different keyword extraction techniques. The color spectrogram technique shows promise, we are working on improving the color contrast between important features and the general background colors.

ACKNOWLEDGMENTS

We gratefully acknowledge funding from The University of Adelaide.

REFERENCES

1. International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature* **409**, pp. 860–921, 2001.
2. M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, and A. Somoza, "Keyword detection in natural languages and DNA," *Europhysics Letters* **57**(5), pp. 759–764, 2002.
3. D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters* **88**, Jan. 2002.
4. D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine* **18**, pp. 8–20, July 2001.
5. Z. Yu, V. Anh, and K. Lau, "Measure representation and multifractal analysis of complete genomes," *Physical Review E* **64**, pp. 031903/1–9, Sept. 2001.
6. J. Fitch and B. Sokhansanj, "Genomic engineering: moving beyond dna sequence to function," *Proceedings of the IEEE* **88**, pp. 1949–1971, Dec. 2000.
7. J. Brown, C. Douady, M. Italia, W. Marshall, and M. Stanhope, "Universal trees based on large combined protein sequence data sets," *Nature Genetics* **28**, pp. 281–285, 2001.
8. A. Bookstein and D. Swanson, "Probabilistic models for automatic indexing," *Journal of the American Society for Information Science* **25**, pp. 312–318, 1974.
9. A. Bookstein and D. Swanson, "A decision theoretic foundation for indexing," *Journal of the American Society for Information Science* **26**, pp. 45–50, 1975.
10. P. Carpena and J. O'Vari. Private communication, 2001.
11. E. Nestle, K. Aland, M. Black, *et al.*, *Novum Testamentum Graece*, Deutsche Bibelgesellschaft, 26th ed., 1979.
12. M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, U.S. Govt. Printing Office, 1964.
13. P. Bernaola-Galván, I. Grosse, P. Carpena, J. Oliver, R. Román-Roldán, and H. Stanley, "Finding borders between coding and noncoding dna regions by an entropic segmentation method," *Physical Review Letters* **85**, pp. 1342–1345, Aug. 2000.
14. M. Kuroda, T. Ohta, I. Uchiyama, T. Baba, H. Yuzawa, *et al.*, "Whole genome sequencing of meticillin-resistant staphylococcus aureus," *Lancet* **357**, pp. 1225–1240, 2001.
15. V. Anh, K. Lau, and Z. Yu, "Multifractal characterisation of complete genomes," *Journal of Physics A* **34**, pp. 7127–7139, Sept. 2001.
16. C. Shannon, IEEE Press, 1993.
17. A. da Silva, J. Ferro, F. Reinach, C. Farah, and L. Furlan, "Comparison of the genomes of two xanthomonas pathogens with differing host specificities," *Nature* **417**, pp. 459–463, 2002.
18. J. Parkhill, B. Wren, K. Mungall, J. Ketley, and C. Churcher, "The genome sequence of the food-borne pathogen campylobacter jejuni reveals hypervariable sequences," *Nature* **403**, pp. 665–668, 2000.
19. P. Winter, "Steiner problem in networks: a survey," *Networks* **17**(2), pp. 129–167, 1987.
20. S. Kullback, *Information Theory and Statistics*, Dover Publications, 1968.
21. R. Alm, L. Ling, D. Moir, B. King, E. Brown, *et al.*, "Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen helicobacter pylori," *Nature* **397**, pp. 176–180, 1999.
22. K. Nelson, R. Clayton, S. Gill, M. Gwinn, R. Dodson, *et al.*, "Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima," *Nature* **399**, pp. 323–329, 1999.

23. G. Deckert, P. Warren, T. Gaasterland, W. Young, A. Lenox, *et al.*, "The complete genome of the hyperthermophilic bacterium aquifex aeolicus," *Nature* **392**, pp. 353–385, 1998.
24. B. Snel, P. Bork, and M. Huynen, "Genome phylogeny based on gene content," *Nature Genetics* **21**, pp. 108–110, 1999.