Classifying ancient texts by inter-word spacing

In the 1940s, Claude Shannon applied statistical concepts to both written text and genetics, pioneering the field of information theory. The seminal paper that exploited Shannon’s concept of information entropy for the study of DNA sequences was by Mantegna et al. Due to the mapping of the human genome, there is renewed interest in statistical techniques for classification and data mining of DNA sequences. Given that DNA sequences can be viewed as possessing a 4 letter alphabet {A, C, G, T}, it is not surprising that data mining and classification of text has inspired techniques for DNA and vice versa. Recent advances in these areas have been demonstrated by Ortuño et al and Benedetto et al., for example.

Ortuño et al. have explored both written texts and DNA using a powerful new idea of inter-word spacing for extracting keywords. The inter-word spacing is defined as the word count between a word and the next occurrence of the same word in a text. All the inter-word spaces, for each case, are then counted up and the standard deviation is computed. This is then repeated for different words – the words are then ranked according to the standard deviation values, the highest first. The standard deviation is then plotted versus the logarithm of the rank. Ortuño et al. have found that words with the highest σ ranking tend to make better search engine keywords, as opposed to words with high hit counts.

We demonstrate a striking new result; in Figure 1, where standard deviation versus log of rank is plotted for Koine Greek source texts of the New Testament. For simplicity only the books of Luke, Matthew and Acts are plotted here. The close match between the curves for Acts and Luke, compared to other books, appears to add weight to what has always been accepted by scholars; namely that the author of Luke was identical to that of Acts.

To check the significance of this match we introduce the idea of a quantitative measure obtained by comparing variances in spacing between words common to all the texts under examination. We use a two-distribution χ² measure on the variances σ² of top-ranked keywords; the χ² values for pairs of texts are put into the table below, with a lower χ² indicating a closer match. This gives the following result for the gospels and Acts.

|        | Matt | Mark | Luke | John | Acts |
|--------|------|------|------|------|------|
| Matt   | 0.00 | 3.91 | 2.20 | 6.05 | 3.95 |
| Mark   | 3.91 | 0.00 | 3.21 | 5.53 | 4.90 |
| Luke   | 2.20 | 3.21 | 0.00 | 2.42 | 2.02 |
| John   | 6.05 | 5.53 | 2.42 | 0.00 | 3.17 |
| Acts   | 3.95 | 4.90 | 2.02 | 3.17 | 0.00 |

As a check against a known benchmark, the following table compares works by Charles Dickens (Great Expectations and Barnaby Rudge) and Thomas Hardy (Jude the Obscure and Tess of the D’Urbervilles):

|        | Jude | Tess | Barn | GE  |
|--------|------|------|------|-----|
| Jude   | 0.00 | 1.05 | 4.92 | 8.24|
| Tess   | 1.05 | 0.00 | 2.34 | 4.59|
| Barn   | 4.92 | 2.34 | 0.00 | 1.86|
| GE     | 8.24 | 4.59 | 1.86 | 0.00|

As expected, lowest χ² scores are obtained for the correct author match.

In conclusion, our results add weight to the generally accepted hypothesis of a common author between the books of Luke and Acts. Future developments in this area may shed some light on a number of historical debates surrounding the question of authorship. Applying these types of tests to DNA may be of interest in the study of phylogenetic relationships.

Matthew J. Berryman,
Andrew Allison, Derek Abbott
Centre for Biomedical Engineering,
Department of Electronic and Electrical Engineering, University of Adelaide,
Adelaide, SA 5005, Australia
e-mail: dabbott@eleceng.adelaide.edu.au

1. Calderbank, R., Sloane, N.J.A., Nature, 410, 768 (2001).
2. Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M., Stanley, H.E., Physics Review E, 52, 2939-2950 (1995).
3. Birney, E., Bateman, A., Clamp, M.E., Hubbard, T.J., Nature, 409, 827-828 (2001).
4. Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, Somoza, A.M., Europhysics Letters, 57, 759-764 (2002).
5. Benedetto, D., Caglioti, E., Loreto, V., Physical Review Letters, 88, 048702 (2002).
6. Nettle, E., Aland, K., Black, M., Martini, C., Metzger, B., Wikipren, A., Novum Testamentum Graece, 26th edition (Deutsche Bibelgesellschaft, Stuttgart, 1979).