Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 329 (2003) 473–483
PHYSICA A
www.elsevier.com/locate/physa

Information categorization approach to literary authorship disputes

Albert C.-C. Yang[a,b,c,*], C.-K. Peng[a], H.-W. Yien[b,c], Ary L. Goldberger[a]

[a] *Cardiovascular Division and Margret and H.A Rey Institute for Nonlinear Dynamics in Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA 02215, USA*
[b] *School of Medicine, National Yang-Ming University, Taipei, Taiwan*
[c] *Taipei Veterans General Hospital, Taipei, Taiwan*

Received 28 April 2003

Abstract

Scientific analysis of the linguistic styles of different authors has generated considerable interest. We present a generic approach to measuring the similarity of two symbolic sequences that requires minimal background knowledge about a given human language. Our analysis is based on word rank order–frequency statistics and phylogenetic tree construction. We demonstrate the applicability of this method to historic authorship questions related to the classic Chinese novel "The Dream of the Red Chamber," to the plays of William Shakespeare, and to the Federalist papers. This method may also provide a simple approach to other large databases based on their information content.

PACS: 05.10.−a; 89.70.+c

Keywords: Linguistic analysis; Authorship; Shannon entropy

1. Introduction

Symbolic sequences as information carriers are commonly seen in nature. Some are unique human creations, such as language and music; and others are generated by

* Corresponding author. Cardiovascular Division and Margret and H.A Rey Institute for Nonlinear Dynamics in Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA 02215, USA.
E-mail address: ccyang@physionet.org (A.C.-C. Yang).

0378-4371/$ - see front matter
doi:10.1016/S0378-4371(03)00622-8

natural processes, such as genetic codes and neural transmission signals. One central problem in the analysis of these sequences is how to effectively categorize their information content based on their origins. For example, musical compositions by different composers usually display distinct and differentiable styles that can be recognized by experienced listeners. However, this type of categorization is often non-quantifiable. Here we introduce a generic approach to study the problems of information categorization. We illustrate the important features of our approach by applying this method to authorship disputes, since "forensic" text analysis presents generic challenges common to other information categorization problems.

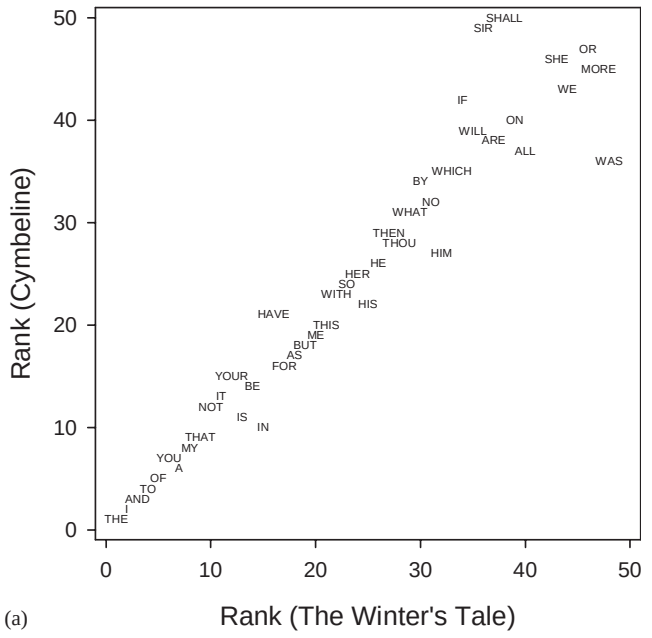
Our approach is based on the concept that the information content in any symbolic sequence is primarily determined by the repetitive usage of its basic elements. Recently, we developed an algorithm to quantify the similarity between symbolic sequences mapped from physiologic time series [1] based on statistical comparisons of the rank and frequency of repetitive elements. Here we show the generalization and application of this approach to the challenge of determining the authorship of written texts.

For written texts this challenge would correspond to comparing the writing styles of different authors. Prior attempts to quantify writing styles have relied primarily on statistical properties of certain linguistic features [2], such as function word frequencies [3,4], word lengths [5,6], sentence lengths [7,8], and vocabulary richness [9–11]. An alternative approach to the authorship problem is to consider written texts as a special case of information-carrying sequences of symbols. In the case of human languages, the basic repetitive elements used in these types of symbolic sequences are called words.

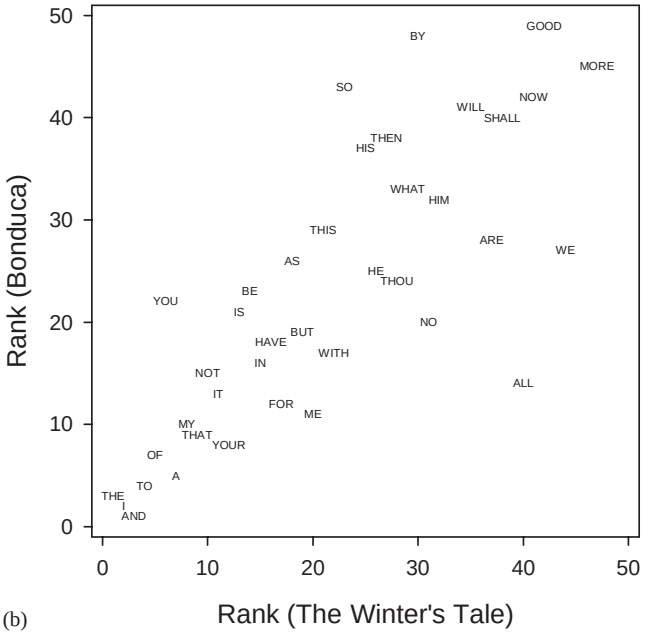
To adapt our new method to written texts, we count the occurrences of each word, and then sort them by descending frequency. The resulting rank–frequency distribution represents the statistical hierarchy of word usage of the original text. For example, the first ranked words correspond to the most frequently used words such as "the", "and", "to", or pronouns in literary English. In contrast, the last ranked words define the rarest used words in the text. Each author has his/her own vocabulary "database". Therefore, when comparing the writings of two different authors, there are *shared words* used by both, as well as *unique words* used specifically by one author but not the other. A simple implementation of our method is to consider only the shared words. Although, unique words contain important information about an author's style, these words could also be related to the specific topic of the text (e.g., names of characters or places in the text).

The rank order difference between two texts can be visualized by plotting the rank number of each shared word in the first text against that of the second text. Fig. 1a shows the comparison between two of Shakespeare's plays, *The Winter's Tale* and *Cymbeline*. The words are tightly centered along the diagonal line, indicating the rank of each word is very similar in these two plays. In contrast, the words are more widely scattered when the comparison is made between one play by Shakespeare, *The Winter's Tale*, and one play by John Fletcher, *Bonduca* (Fig. 1b).

As demonstrated by the above examples, the "distance" (or dissimilarity) between any two texts can be quantified by measuring the scatter of these points from the diagonal line in the rank order comparison plot. Greater distance indicates less similarity and vice versa. Therefore, we can define the distance (D) between two texts, T_1 and



(a)



(b)

Fig. 1. Rank order comparison of top ranking words for: (a) two of Shakespeare's plays: *The Winter's Tale* versus *Cymbeline*; and (b) *The Winter's Tale* versus John Fletcher's *Bonduca*. Words from the two Shakespeare plays fall close to the diagonal, indicating nearly identical ranking. In contrast, the Shakespeare–Fletcher comparison yields greater scatter of words around the diagonal. For example, the word "the" is the most frequently used word in the two Shakespeare's plays, whereas the same word is ranked third in John Fletcher's *Bonduca*. For graphical clarity, we plot only the top ranking words shared by both texts. In the actual analysis, all shared words have been used.

T_2, as

$$D(T_1, T_2) = \frac{1}{N_{12}} \sum_{k=1}^{N_{12}} |R_1(w_k) - R_2(w_k)| F(w_k) \, . \tag{1}$$

Here $R_1(w_k)$ and $R_2(w_k)$ represent the rank of a specific word, w_k, in texts T_1 and T_2, respectively. N_{12} is the number of total shared words used in texts T_1 and T_2. The absolute difference of ranks, $|R_1(w_k) - R_2(w_k)|$, is proportional to the Euclidean distance from a scattered point to the diagonal line. This term is then multiplied by a weighting function, $F(w_k)$, to take into account that not all points on the rank order comparison plot are equal. The more frequently used words should contribute more to defining the style of an author.

We select the weighting function $F(w_k)$ to be the sum of Shannon's entropy [12] for w_k in texts T_1 and T_2:

$$F(w_k) = [- p_1(w_k) \log(p_1(w_k)) - p_2(w_k) \log(p_2(w_k))]/Z \, ,$$

where Z is the normalization factor such that $\sum_{k=1}^{N_{12}} F(w_k)=1$. Here $p_1(w_k)$, and $p_2(w_k)$ represent the probability of a specific word, w_k, in texts T_1 and T_2, respectively. The selection of Shannon's entropy to be the weighting function is to ensure that words occurring with higher probability will be more heavily weighted. We note that this similarity measurement is an empirical index which does not necessarily obey the triangular inequality criterion of a distance measure. Therefore, the triangular inequality test is required before generating a phylogenetic tree. When applied to the literary texts here, no violation of the triangular inequality was observed.

For initial validation of the rank–frequency method, we first apply it to a database containing 16 texts by eight well-known authors used in a previous study [11]. We find the method can unambiguously classify all of authors without error (Fig. 2). In comparison, previous approaches for authorship identification [11,13,14] misidentified at least one author.

2. Chinese literature: the dream of the red chamber

We next apply the distance measure, defined in Eq. (1), to study a controversy surrounding the authorship of the classic 18th century novel *The Dream of the Red Chamber* (http://cls.admin.yzu.edu.tw/hlm). This novel is considered one of the most influential texts in Chinese literature. However, the original manuscript was lost and only incomplete hand-written copies survived. A century's old debate centers on the consistency of first 80 chapters and last 40 chapters [15]. To analyze the entire text of the book, we group sequences of 10 chapters into segments, so the first segment contains chapters 1–10, the second segment chapters 11–20, and so on. We then calculate the distance between each pair of segments. The resulting distance matrix is shown in Fig. 3a. We find that the distance between the first 80 and the last 40 chapters is generally larger than the distance within both parts. Furthermore, using this distance metric, we arrange each segment in a phylogenetic tree [16,17] to visualize the result qualitatively. The structure of the tree (Fig. 3b) reveals that last 40 chapters are

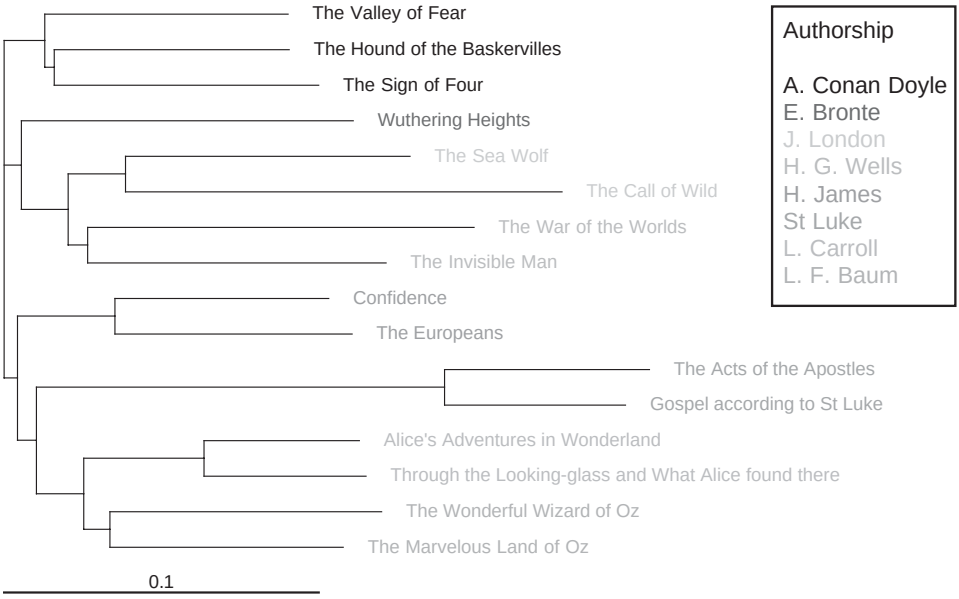


Fig. 2. Initial validation of the rank–frequency method. We apply our method to a database containing 16 texts by eight well-known authors used in a previous study (1). The word rank–frequency list of each work is generated from the first 20 000 words in each text. The neighbor-joining phylogenetic tree is then constructed based on a distance matrix (see text). Each text has been color-coded according to its authorship. We find that the rank–frequency method can unambiguously classify all the texts according to their authors.

arranged along a separate branch clearly distinguishable from other chapters. Similarly, the first 20 chapters are also arranged along another individual branch. This result is consistent with the recent consensus of literary critics that the original manuscript underwent multiple editings by different authors [15].

3. English literature: Shakespeare's plays

We next apply this method to questions concerning Shakespeare's plays. One question is whether certain plays, in particular *Edward III* and *The Two Noble Kinsmen*, were actually written by Shakespeare. The former was first printed anonymously in 1596, and recently has been controversially restored to the Shakespearean canon [18]. The latter has been considered as written collaboratively by Shakespeare and John Fletcher. However, Shakespeare's role in this play's composition is still debated [18].

The second more general question regards the identity of Shakespeare. More than 80 Elizabethans have been proposed since the middle of the 18th century as the "true Shakespeare." Four continue to receive serious consideration: Sir Francis Bacon (Lord Verulam), Christopher Marlowe, William Stanley (Sixth Earl of Derby), and Edward de Vere (17th Earl of Oxford). To demonstrate how our method may provide some useful insights to both questions simultaneously, we construct a database of 50 plays including

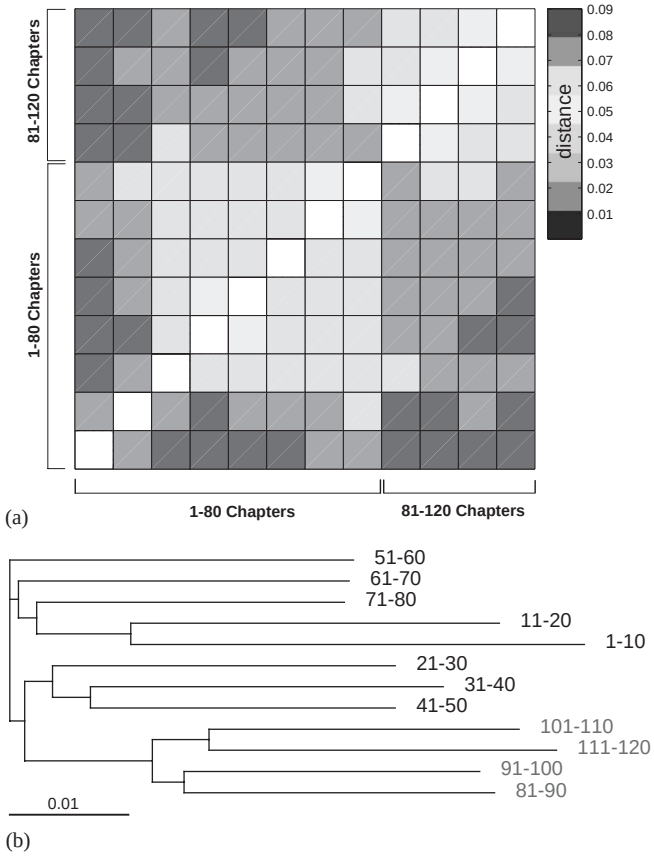


Fig. 3. Rank frequency analysis of *The Dream of the Red Chamber*. (a) Pairwise distance matrix of the entire book. Each box represents a segment containing 10 consecutive chapters. The distance between each pair of segments is represented on a graded color scale. (b) The neighbor-joining phylogenetic tree constructed by the pairwise distance matrix. The distance between each segment on the tree is the summation of the total horizontal lengths along the connection path. Both representations support the view that the first 80 chapters and last 40 chapters are by different authors.

Shakespeare's canon and selected works by other authors (http://www.chadwyck.com/products/pt-product-Lion.shtml; http://the-tech.mit.edu/Shakespeare). No existing dramatic works are attributed to Bacon, Stanley, or de Vere. Therefore, we include seven of Marlowe's plays, as well as two of Ben Jonson's plays, a lesser candidate for the true Shakespeare. We also include two of John Fletcher's sole-authorship plays to compare with the style of Shakespeare. The phylogenetic tree of these texts based on rank–frequency statistics is shown in Fig. 4. All of the Marlowe, Jonson, and Fletcher texts are distinct from Shakespeare's plays, suggesting that the plays by the former three authors were not composed by the author of Shakespeare's plays. There are two interesting findings regarding the controversial plays: (i) *Edward III* is classified under

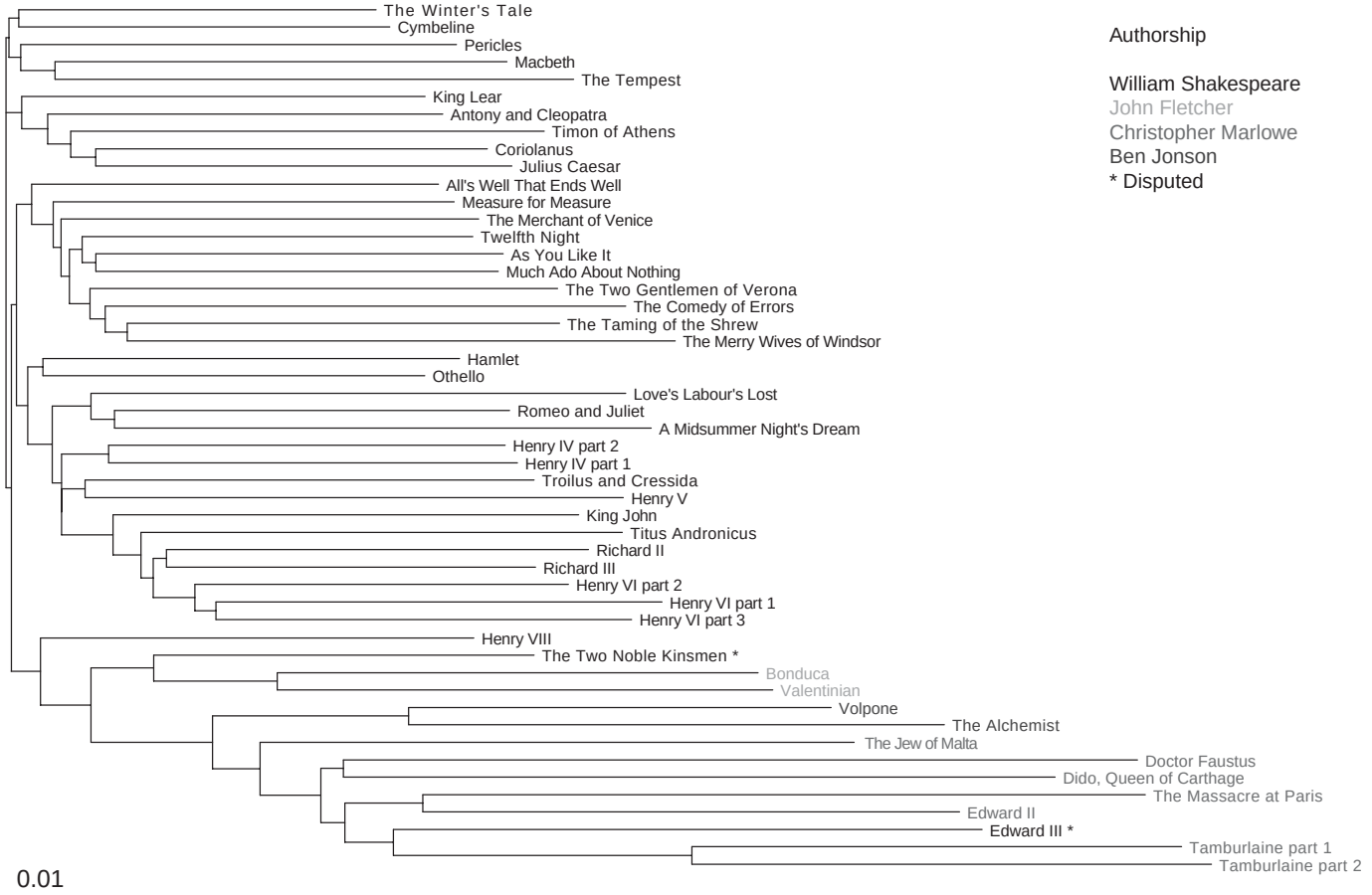


Fig. 4. Neighbor-joining phylogenetic tree based on 50 plays, including those attributed to William Shakespeare ($n = 37$), John Fletcher ($n = 2$) Christopher Marlowe ($n = 7$), Ben Jonson ($n = 2$) and other disputed works ($n = 2$). Each play has been coded according to widely accepted authorship assessments.

Marlowe's branch, and (ii) *The Two Noble Kinsmen* appears between John Fletcher's and Shakespeare's branches. The first finding is compatible with another study showing that the play has a Marlovian framework [19]. The second finding agrees with the view [18] that this later work may have been co-authored by Fletcher. We also find that the above results are qualitatively similar using different versions of Shakespeare's canon, including the First Folio and Globe editions (http://etext.lib.virginia.edu). We also find, when using the version of plays from the First Folio, the early historic plays by Shakespeare, such as the Henry VI series, are arranged closer to the Marlowe branch, consistent with the purported influence of this playwright on those early Shakespeare's works [18].

For further analysis, we can adapt this method to study successive pairs of words. Successive word pairs (coupled words) may be informative because they relate to the syntactic structure of the literature. To reduce the effect of coupled words appearing simply by chance, we apply the distance measurement Eq. (1) to coupled words that appear at least twice in a text. The result of this coupled word analysis on the Shakespeare identity question is consistent with that of the single word analysis presented in Fig. 4. Furthermore, when considering Shakespeare's canon by itself, we note that the phylogenetic tree constructed using coupled words robustly classifies plays by different genre (Fig. 5).

4. English literature: the federalist papers

Finally, we apply the rank–frequency method to authorship problems concerning the *Federalist Papers*, a series of 85 essays written on the proposed new US Constitution and the nature of republican government (http://www.constitution.org). The papers are believed to have been written between 1787 and 1788 by Alexander Hamilton, James Madison, and John Jay. Since 1788, the consensus has been that Alexander Hamilton was the sole author of 51 papers, John Jay of five, James Madison of 14, and that Hamilton and Madison collaborated on another three [20]. The authorship of the remaining 12 papers (numbers 49–58, 62, 63) has been in dispute. Analysis using the word rank–frequency method supports Mosteller and Wallace's conclusion [3,21] that Madison was the author of all 12 disputed papers.

5. Discussion

The rank–frequency method applied here complements traditional approaches to text analysis because it incorporates certain salient features not accounted for by other techniques. Our method is based on a simple assumption, namely that different authors have a preference for certain words that they use with higher frequency. This behavior is unlikely to be consciously manipulated by the author and may serve as a robust stylistic signature. The method was originally developed for studying symbolic sequences, and, therefore, assumes minimal knowledge about any specific language. We demonstrate

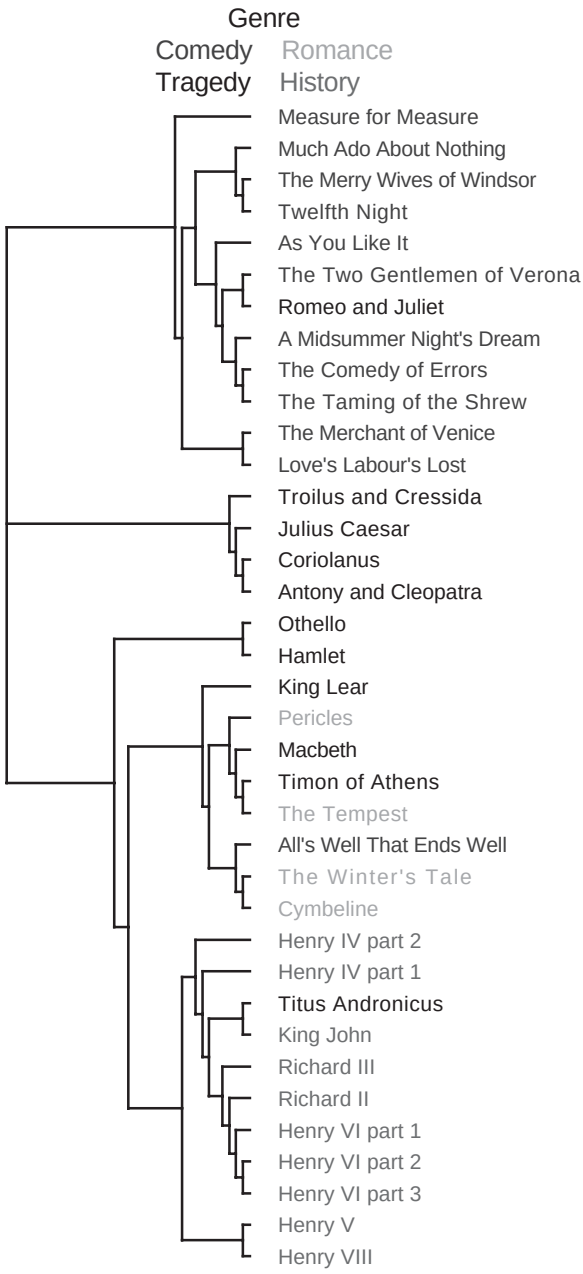


Fig. 5. Phylogenetic tree of Shakespeare's canon based on coupled word rank–frequency lists. Each play is coded by genre: comedy, tragedy, history, and romance.

that the method can be readily applied both to Chinese and English literary texts and also to authorship disputes in very different eras.

The new distance measure, defined in Eq. (1), yields enhanced discrimination when compared to previously proposed methods [11,13,14] because it incorporates both a probabilistic weighting factor given by Shannon's entropy and a term related to the number of shared words. Additional discriminatory information may be obtained by incorporating certain classes of unique words. However, even without reference to such specific identifiers, our method yields robust separation among authorial styles, thereby providing a generic, readily implemented approach to automated textual analysis.

Additional issues regarding the application of this method to authorship problems are (i) the minimum length of text needed to perform the analysis, and (ii) the effect of differences in text length on the distance measurement. In the case of *The Dream of the Red Chamber*, the method can detect textual differences using single chapters as the minimum segments (average length $\sim$6100 words). For English literary texts, the method also obtains consistent results with the relatively short *Federalist papers* (average length $\sim$2200 words). However, for dramatic works, the method requires longer datasets to reach a qualitatively stable result, possibly because of the diverse topics of the plays we analyzed and the multiplicity of candidate authors. The 50 plays related to the Shakespearean authorship question (Fig. 4) have different lengths (coefficient of variance: 21%). However, the tree structures based on the entire texts are similar to the one obtained based on equal length texts (e.g., using the first 10,000 words).

In summary, the rank–frequency method provides a generic analytic tool to measure the similarity between datasets based on their information content. The method is applicable not only to assess the authorship of literary texts, but also for categorizing physiologic time series [1], and, potentially, for analyzing a wide range of symbolic sequences such as genetic codes, musical compositions, and large internet databases.

Acknowledgements

We thank L. Glass, I.C. Henry, J. Healey, and C.-K. Hu for valuable discussions. We gratefully acknowledge the support from the NIH/NCRR (P41-RR13622), the NIH/NIA OAIC (P60-AG08812), the G. Harold and Leila Y. Mathers Charitable Foundation, the National Science Council of Taiwan (NSC90-2314-B-075-127), and the Academia Sinica (Taipei).

References

[1] C.C. Yang, et al., Phys. Rev. Lett. 90 (2003) 108 103.
[2] D.I. Holmes, Lit. Linguist. Comput. 13 (1998) 111.
[3] F. Mosteller, D.L. Wallace, J. Am. Stat. Assoc. 58 (1963) 275.
[4] J. Burrows, Comput. Humanit. 37 (2003) 5.
[5] C.S. Brinegar, JASA 58 (1963) 85.

[6] C.B. William, Biometrika 62 (1975) 207.
[7] A.Q. Morton, J. Roy. Statist. Soc. A 128 (1965) 169.
[8] C.B. William, Biometrika 31 (1940) 356.
[9] D.I. Holmes, J. Roy. Statist. Soc. A 155 (1992) 91.
[10] R. Thisted, B. Efron, Biometrika 74 (1987) 445.
[11] F.J. Tweedie, R.H. Baayen, Comput. Humanit. 32 (1998) 323.
[12] C.E. Shannon, Bell Labs Tech. J. 27 (1948) 379.
[13] S. Havlin, Physica A 216 (1995) 148.
[14] B. Vilensky, Physica A 231 (1996) 705.
[15] S. Hu, et al., Collected Chinese Papers: Textual Research of The Dream of the Red Chamber, The Far East Book Company, Taipei, Taiwan, 1985.
[16] J. Felsenstein, Computer Program, PHYLIP (Phylogeny Inference Package) version 3.5c, Department of Genetics, University of Washington, Seattle, 1993.
[17] N. Saitou, M. Nei, Mol. Biol. Evol. 4 (1987) 406.
[18] H. Bloom, Shakespeare: The Invention of Human, Riverhead Books, New York, 1998.
[19] T. Merriam, Lit. Linguist. Comput. 15 (2000) 157.
[20] J.E.E. Cooke, The Federalist, Word Publishing Company, Meridian Books, Cleveland, OH, 1961.
[21] F. Mosteller, D.L. Wallace, Applied Bayesian and Classical Inference: The Case of the Federalist Papers, 2nd Edition, Springer, New York, 1984.