



Cluster Analysis and Its Applications

Part 2

Albert C. Yang, M.D., Ph.D.

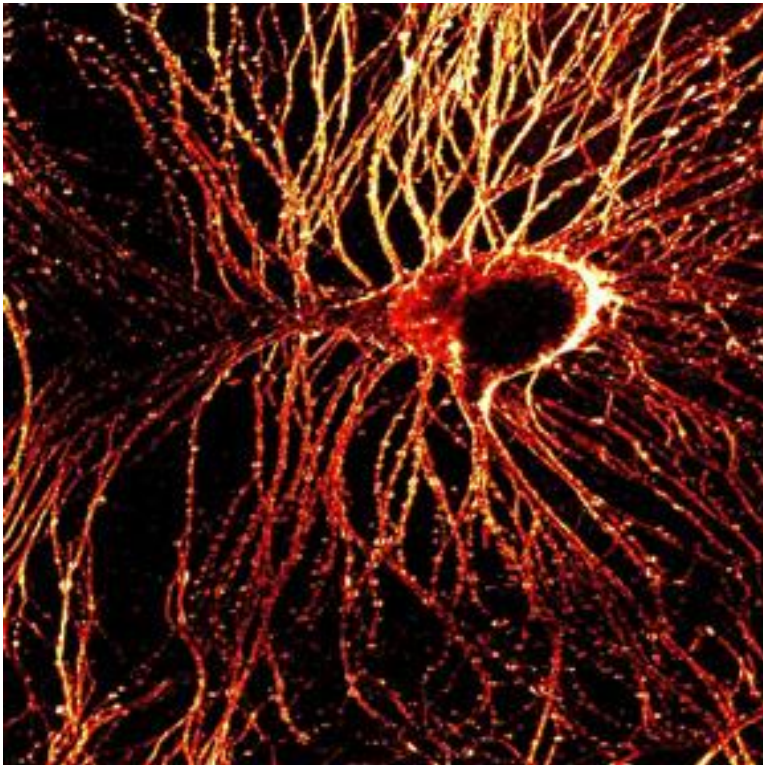
Institutes of Brain Science/Digital Medicine Center
National Yang-Ming University

May 7, 2020

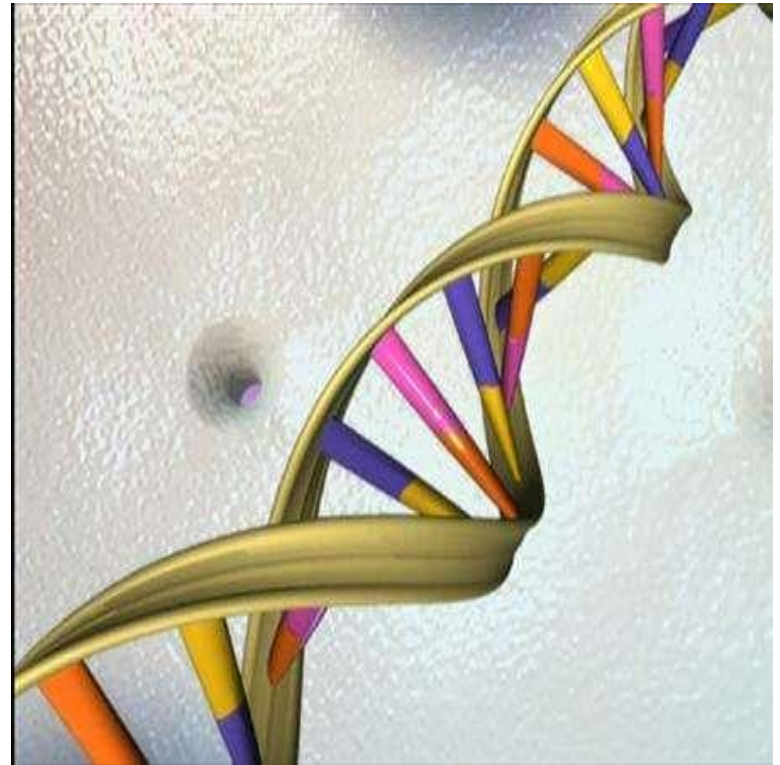
accyang@gmail.com

Information Created by Biological Systems

Neuronal Impulse



Genetic Codes



Information Created by Biological Systems

Human Heartbeats



Human Creations

Earliest record of paintings by human



Lascaux Cave France 20000 BC

Human Creations

Writing Systems



Cuneiform script Sumerians Iraq 2600 BC

Challenge

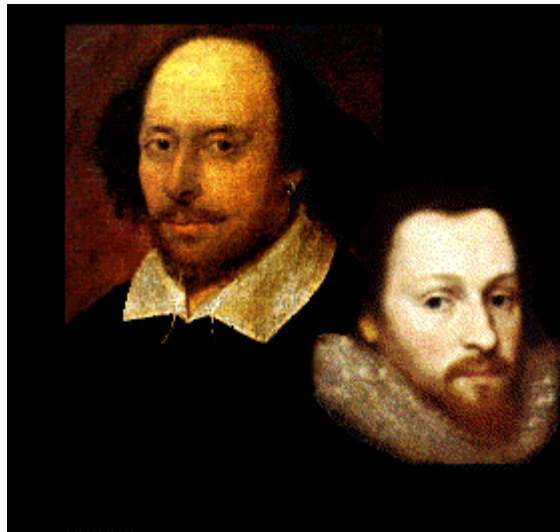
How to effectively categorize information of different origins?



Repetitive Patterns
Human Genome vs. Chimpanzee Genome

Information Categorization Method

Comparison of human literary texts



Repetitive patterns: words

Frequency and Rank Order Statistics

a. The Winter's Tale (William Shakespeare)

Word	Rank	Frequency
The	1	857
I	2	701
And	3	657
To	4	635
Of	5	475
You	6	472
A	7	419
My	8	405
That	9	338
Not	10	305
...

Total different words: 3703

b. Cymbeline (William Shakespeare)

Word	Rank	Frequency
The	1	966
I	2	771
And	3	713
To	4	671
Of	5	525
A	6	459
You	7	424
My	8	383
That	9	381
In	10	320
...

Total different words: 4042

Frequency and Rank Order Statistics

a. The Winter's Tale (William Shakespeare)

Word	Rank	Frequency
The	1	857
I	2	701
And	3	657
To	4	635
Of	5	475
You	6	472
A	7	419
My	8	405
That	9	338
Not	10	305
...

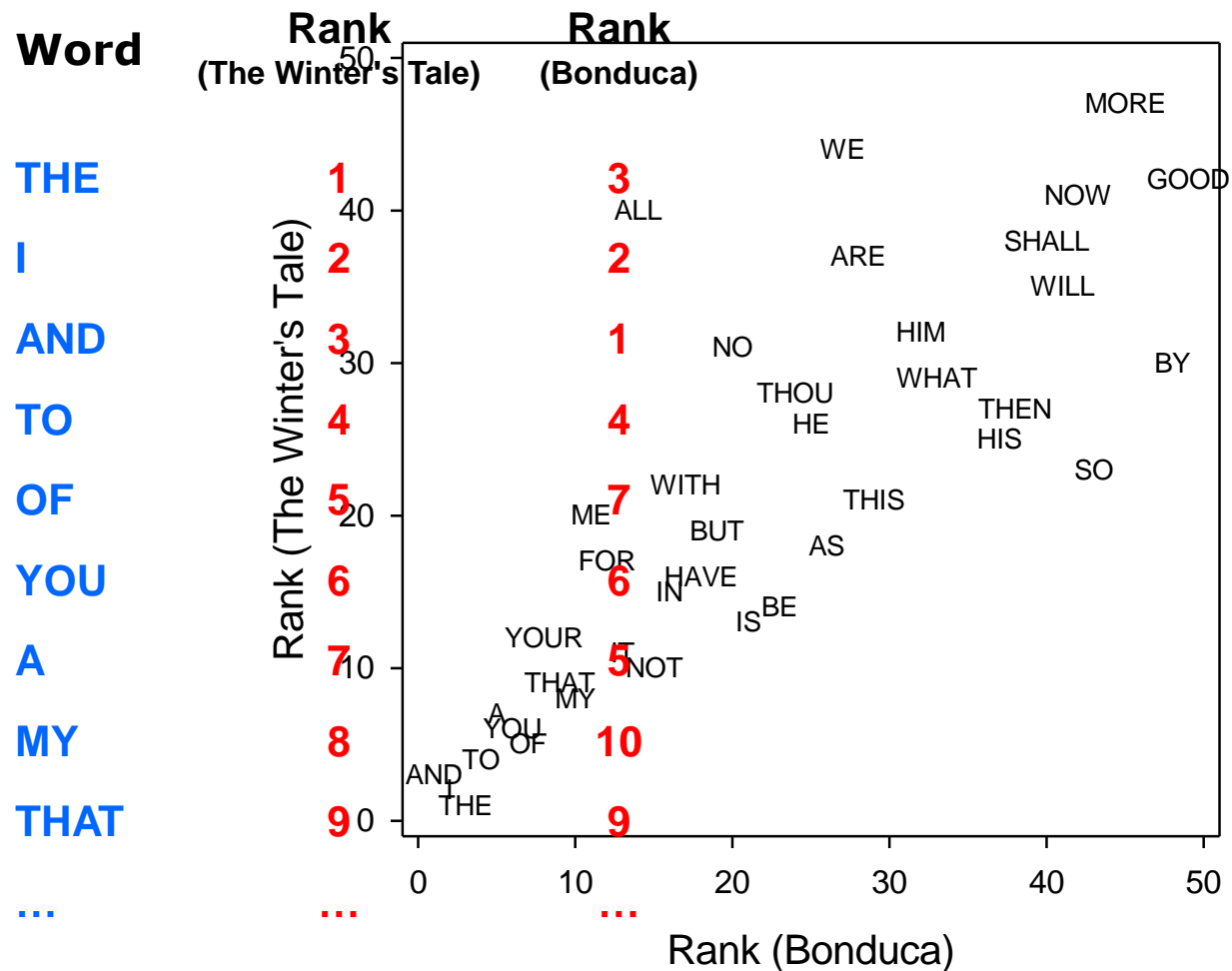
Total different words: 3703

c. Bonduca (John Fletcher)

Word	Rank	Frequency
And	1	667
The	2	584
I	3	552
To	4	432
A	5	393
You	6	286
Of	7	273
Petillius	8	224
Your	9	214
That	10	203
...

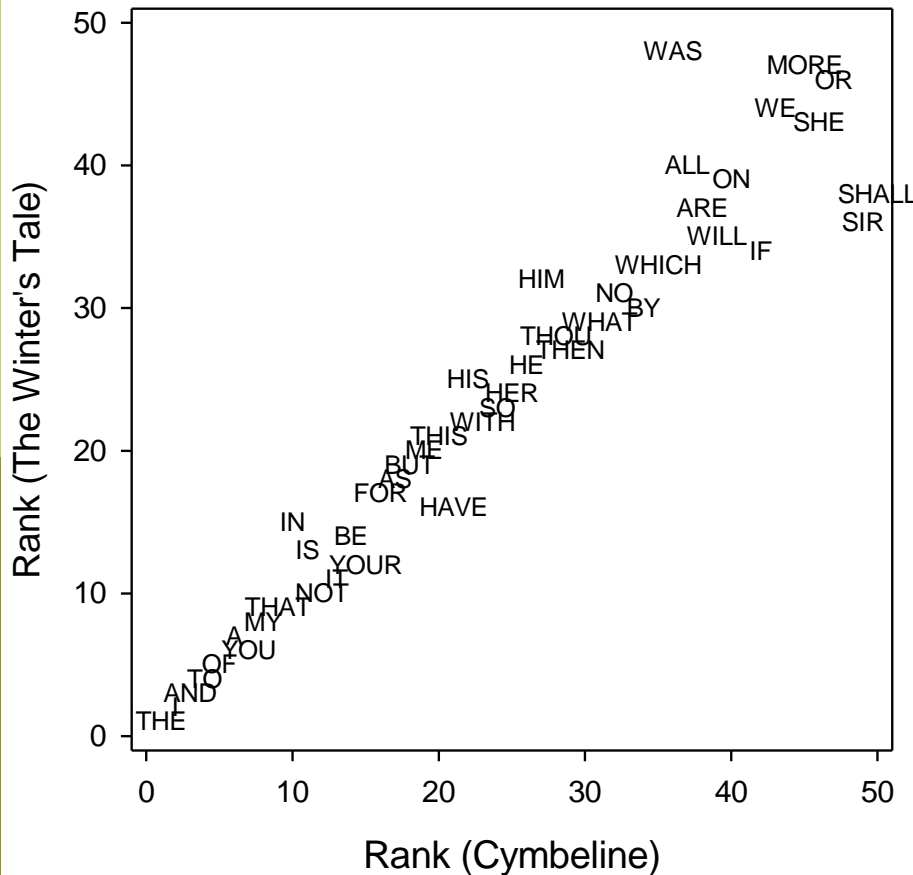
Total different words: 3124

Rank Comparison Map

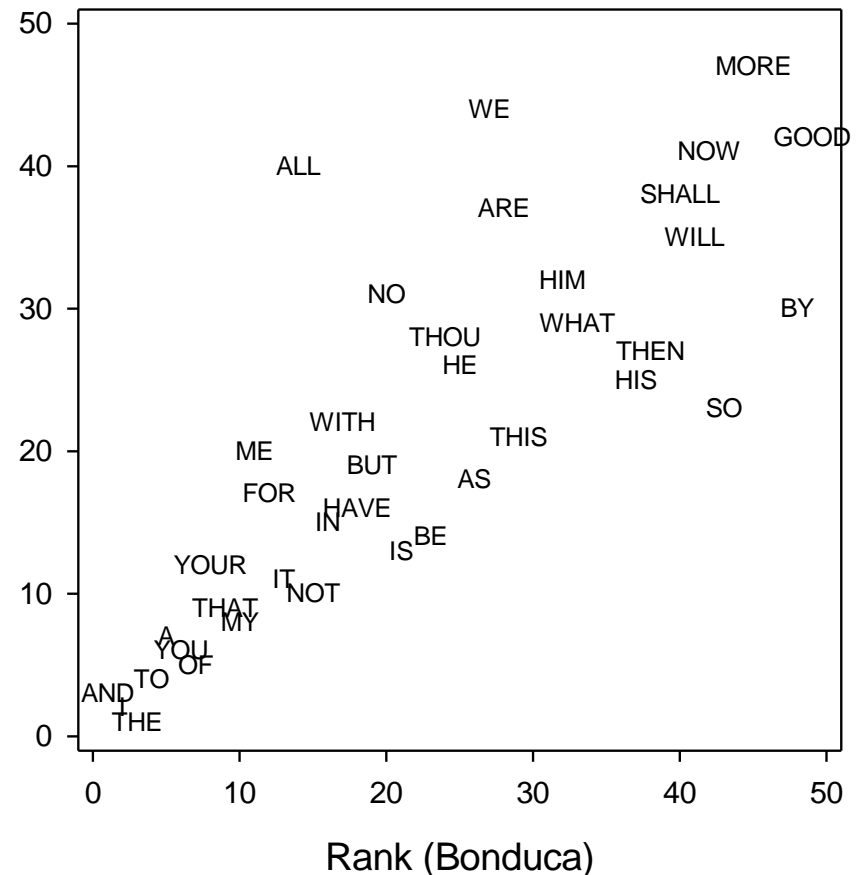


Rank Comparison Maps

Shakespeare vs. Shakespeare



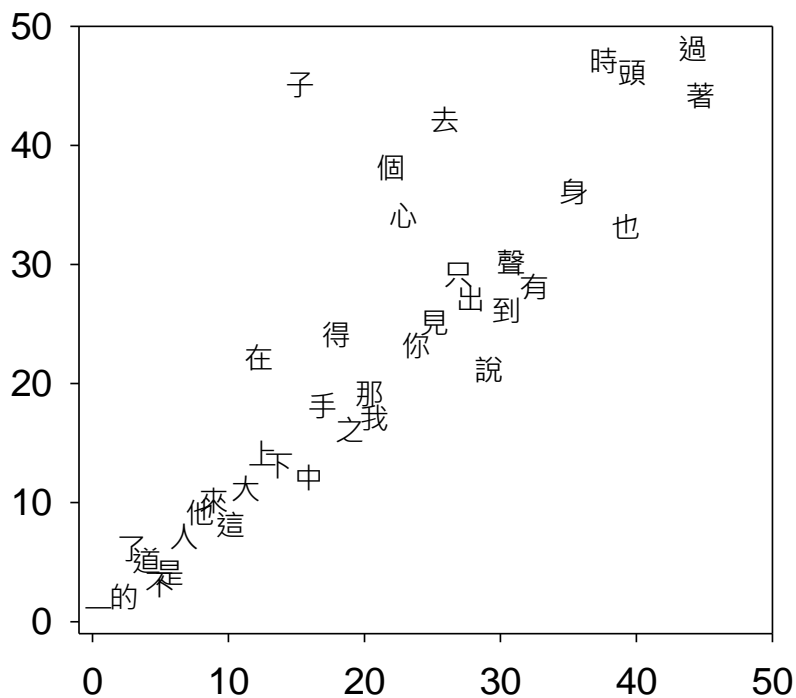
Shakespeare vs. Fletcher



Rank Comparison Maps

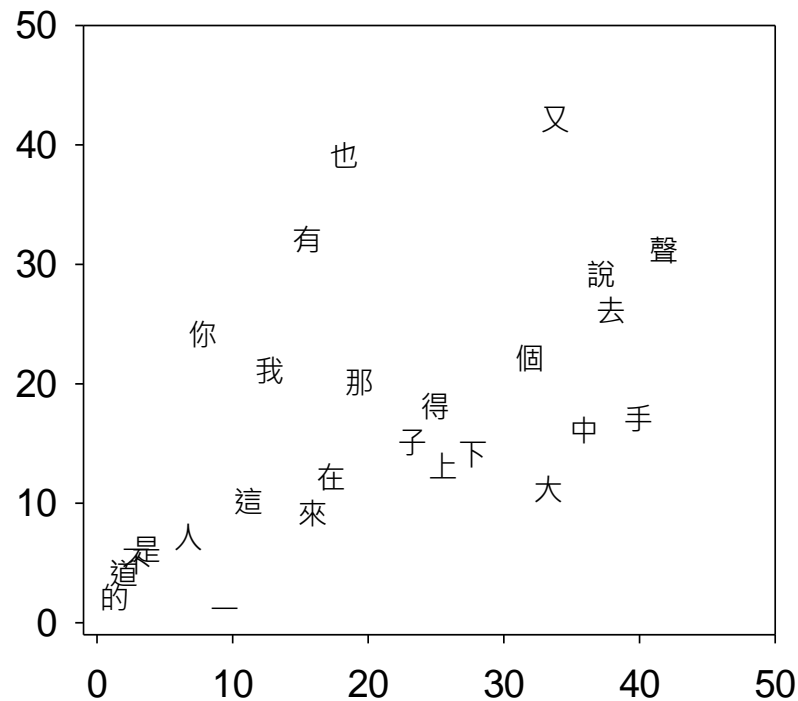
金庸 vs. 金庸

射雕英雄傳



倚天屠龍記

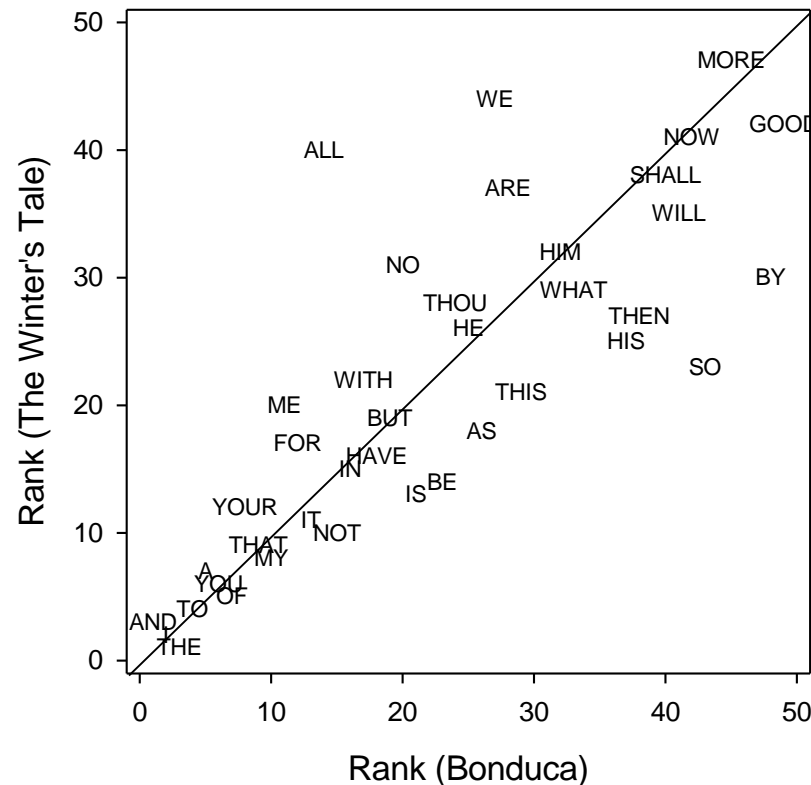
金庸 vs. 古龍



楚留香傳奇

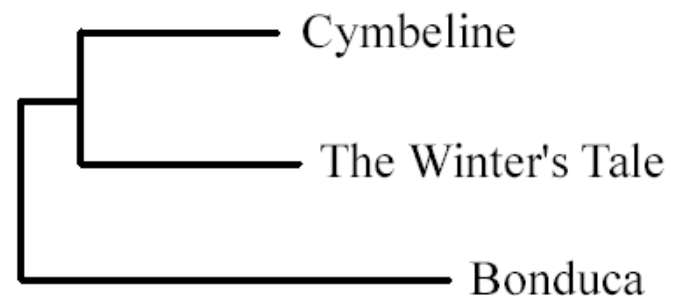
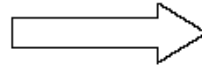
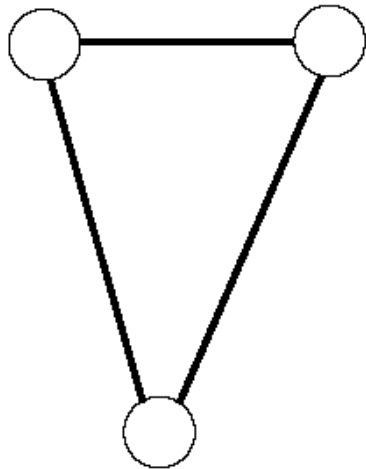
Information-Based Similarity Index

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



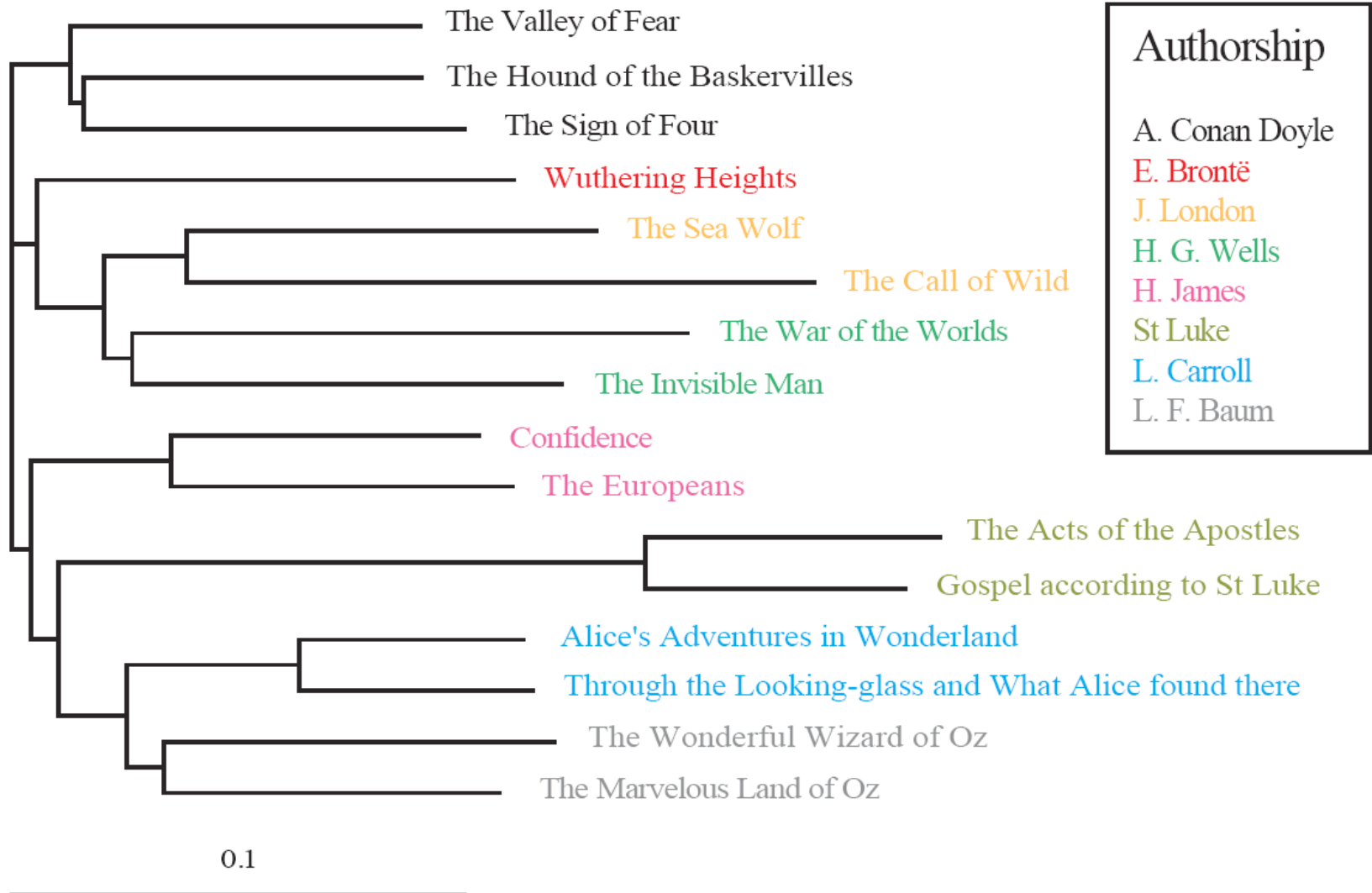
Cluster Analysis

Cymbeline The Winter's Tale



Bonduca

Known Authorship Classification



Chinese Authorship Debate

Dream of the Red Chamber



紅樓夢

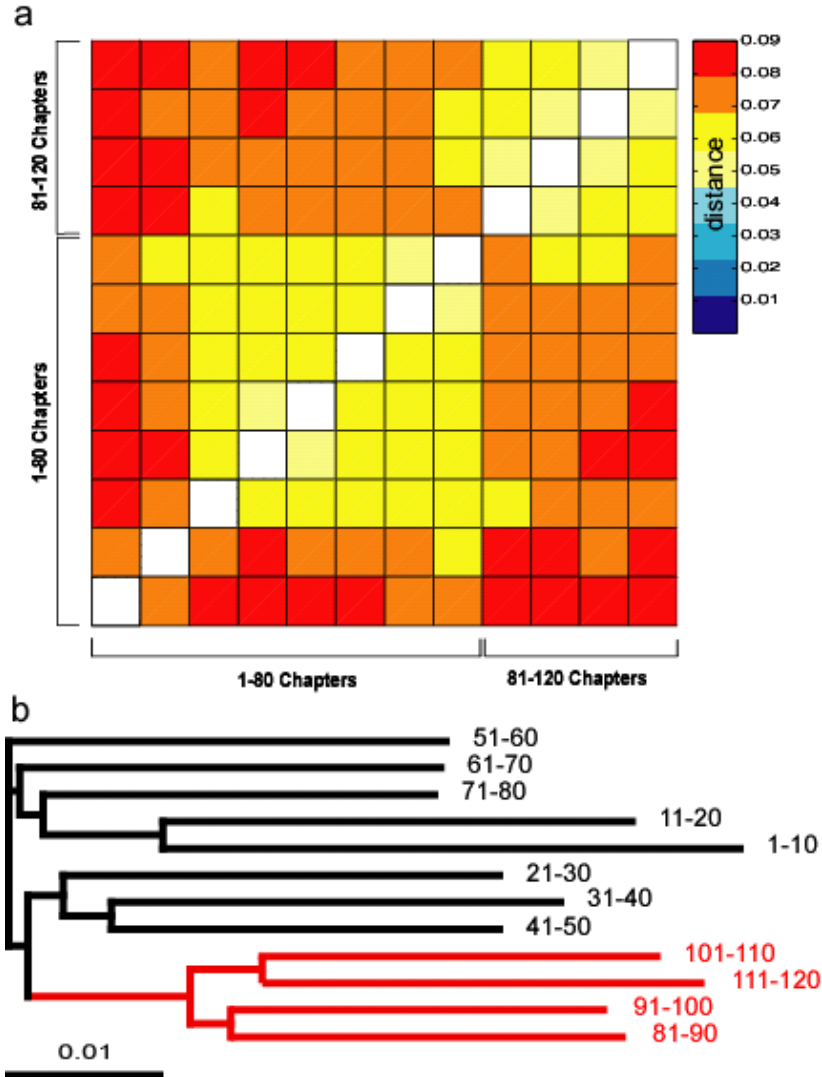
Dream of Red Chamber 紅樓夢

- One of China's four great classical novels.
- Written by Cao Xueqin in the middle of the 18th century during the early Qing Dynasty.
- 80 Chapters in original manuscript copies.
- Gao E and Cheng Weiyuan added 40 additional chapters to complete the novel.

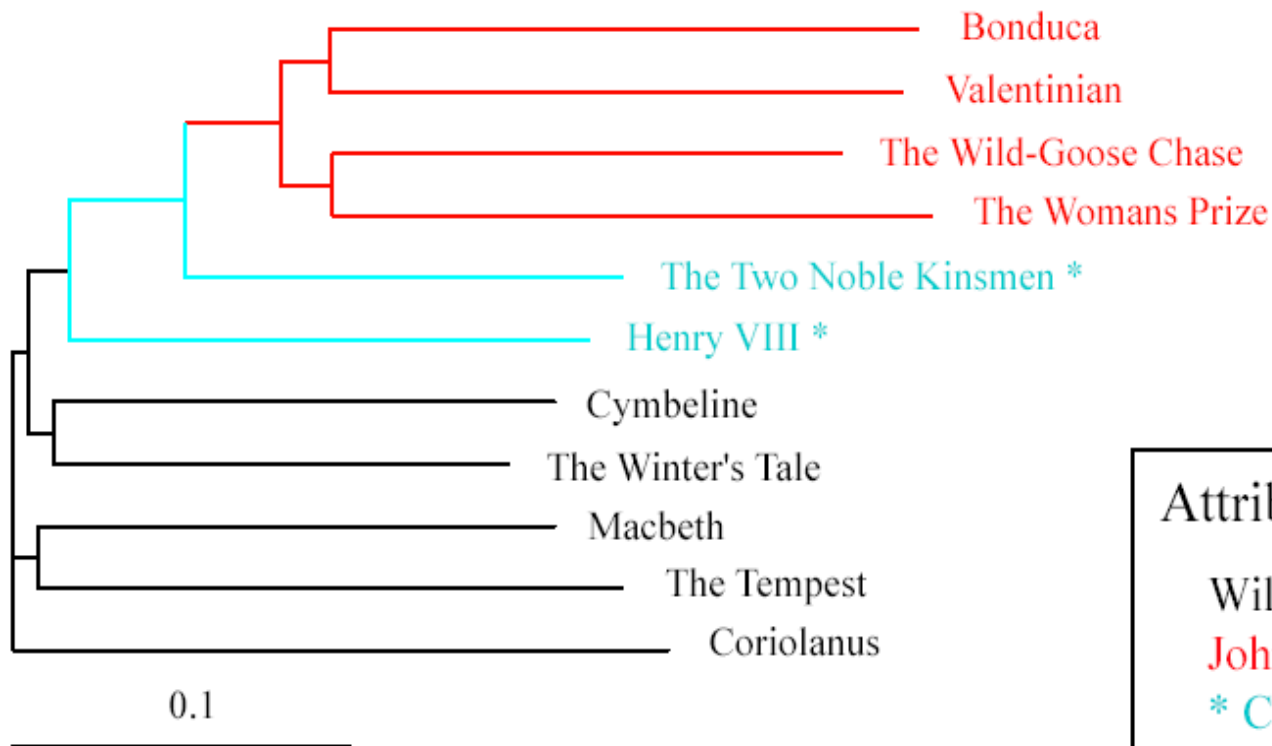
Authorship Debate (紅樓夢)

Rank	1-40		41-80		81-120	
	Word	Frequency	Word	Frequency	Word	Frequency
1	了	6250	了	8301	了	6946
2	不	4505	不	5676	的	5499
3	的	4010	的	5539	不	5009
4	一	3891	一	4942	來	3944
5	道	3683	來	4097	道	3756
6	來	3563	人	3892	是	3741
7	人	3139	我	3769	人	3644
8	我	2843	是	3720	一	3461
9	是	2833	道	3683	說	3391
10	說	2805	說	3637	我	2743

Authorship Debate (紅樓夢)



Shakespeare versus Fletcher



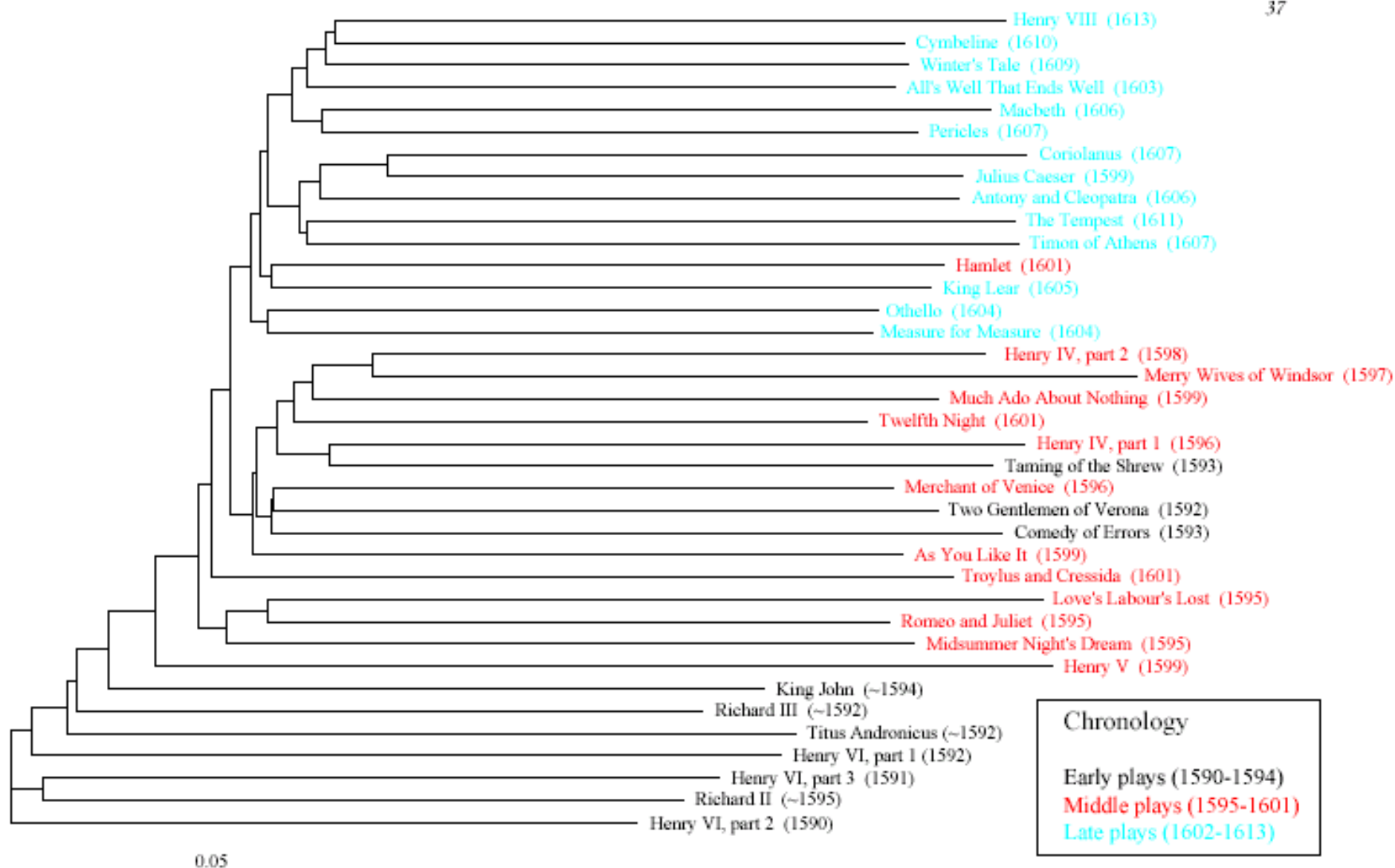
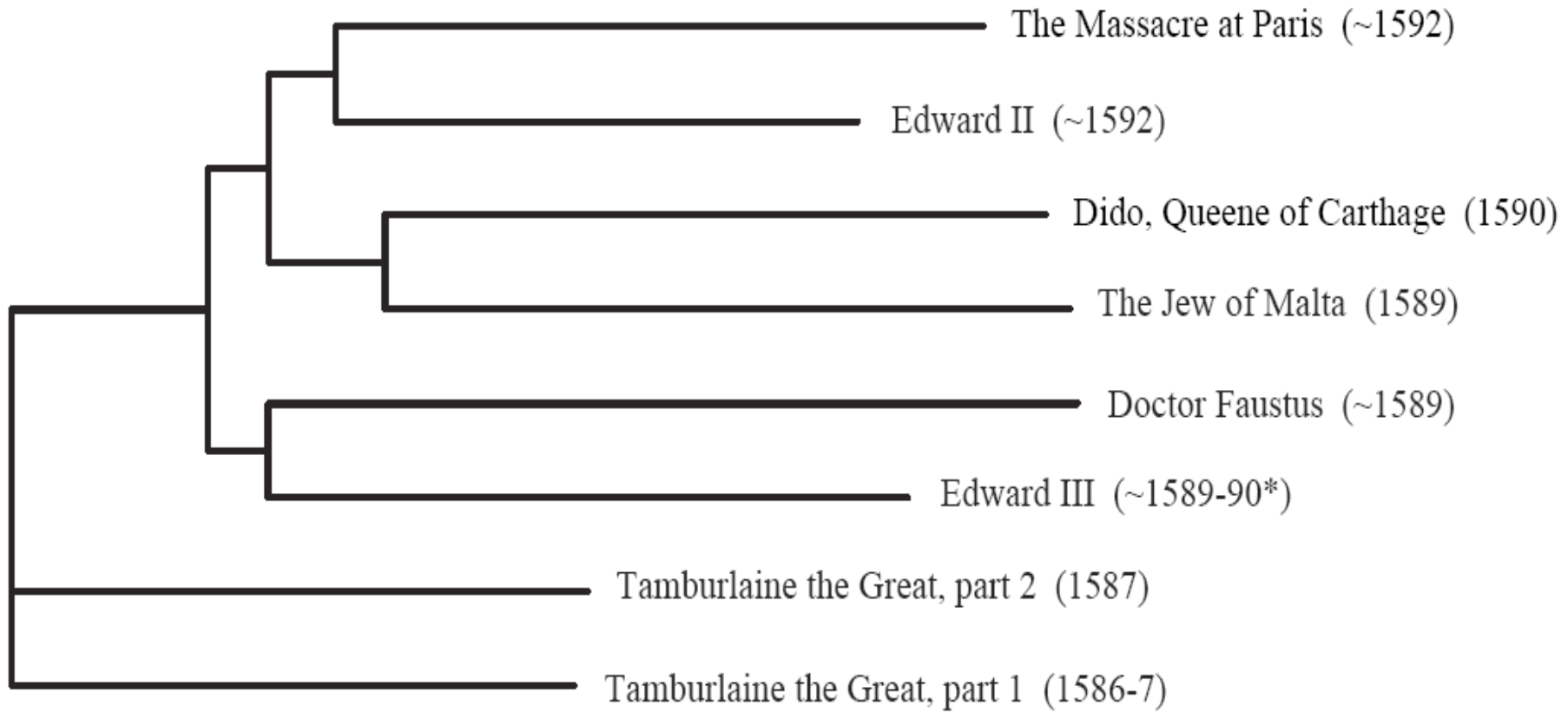


Figure 8. Evolutionary tree of Shakespeare plays. We assign the root (first play) as the second part of the *Henry VI* series as proposed by Wells and Taylor³³. The resulting Shakespeare dramatic phylogeny reasonably places the early plays (1590-94), middle plays (1595-1601), and late plays (1602-1613) from the bottom to the top of the tree.



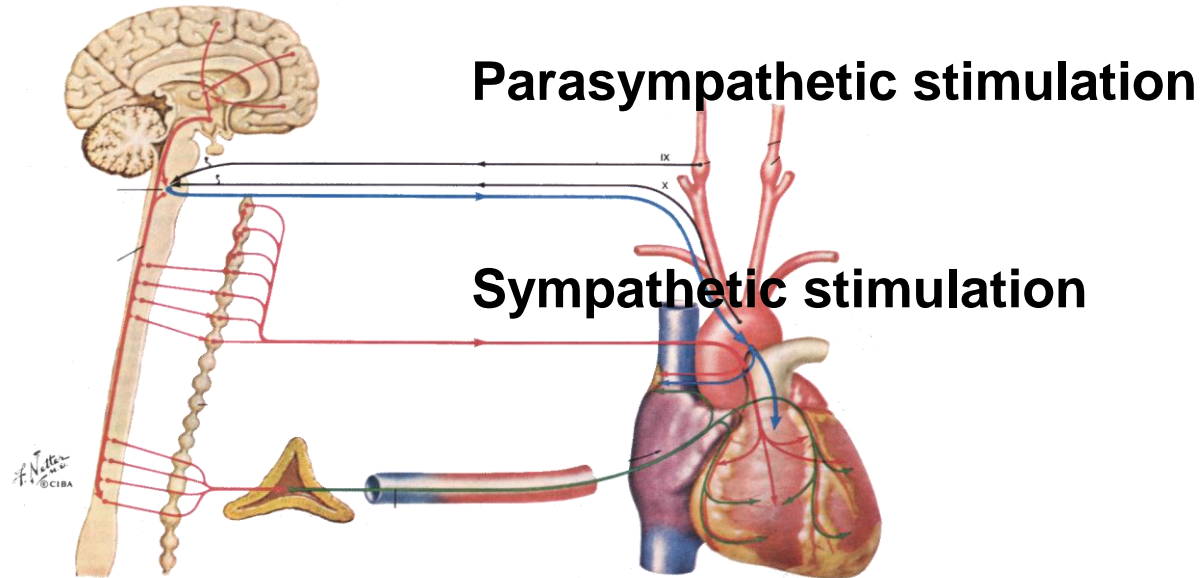
0.1

* Proposed date of play (Wentersdorf 1960)

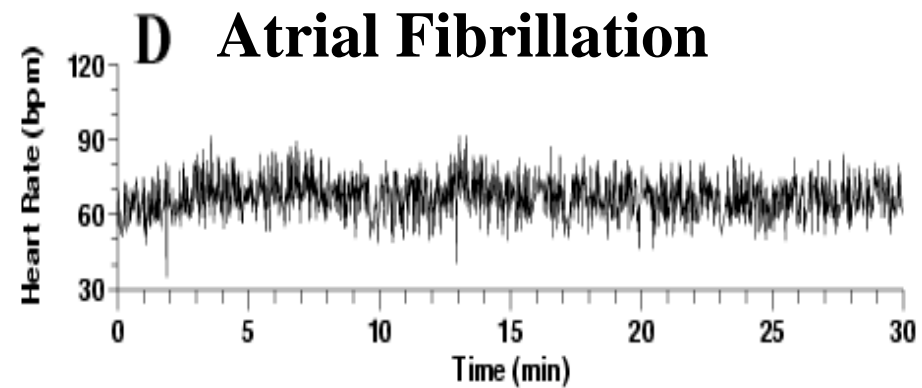
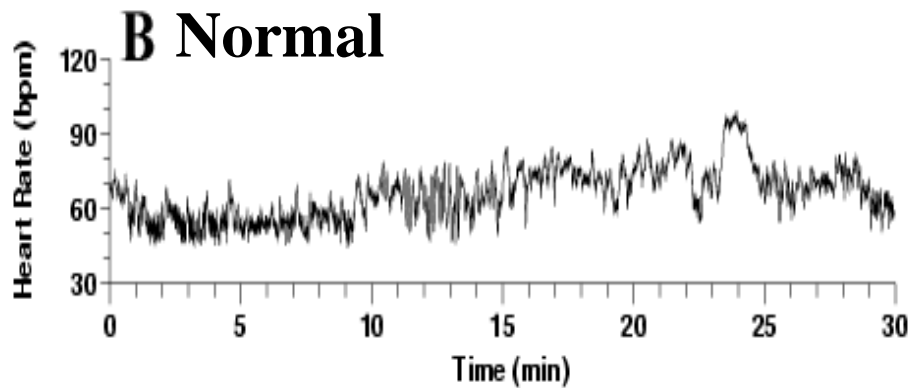
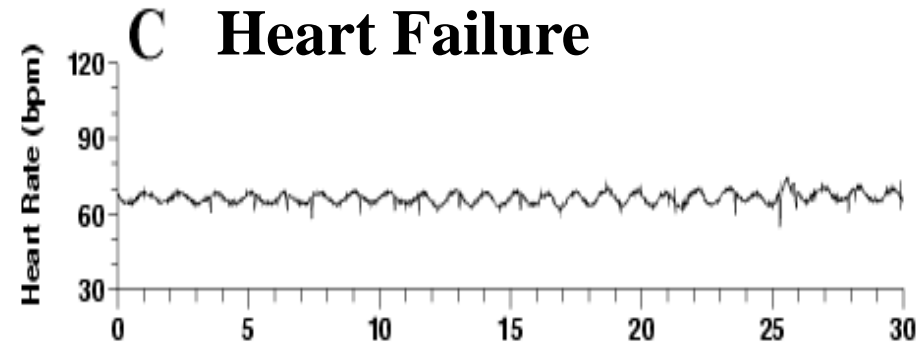
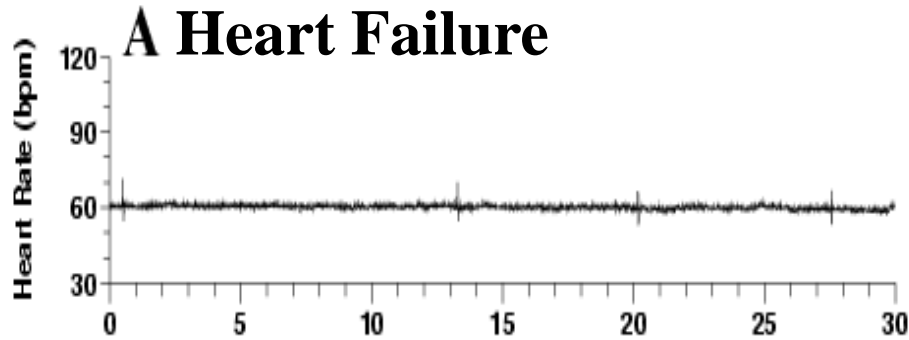
Application to Human Heartbeat



Heart rate dynamics

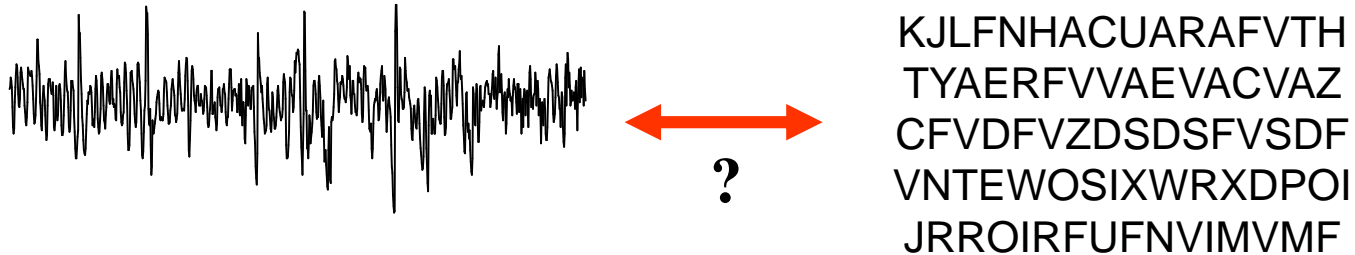


Which Heart Rate Pattern is Healthy?



Technical Challenges

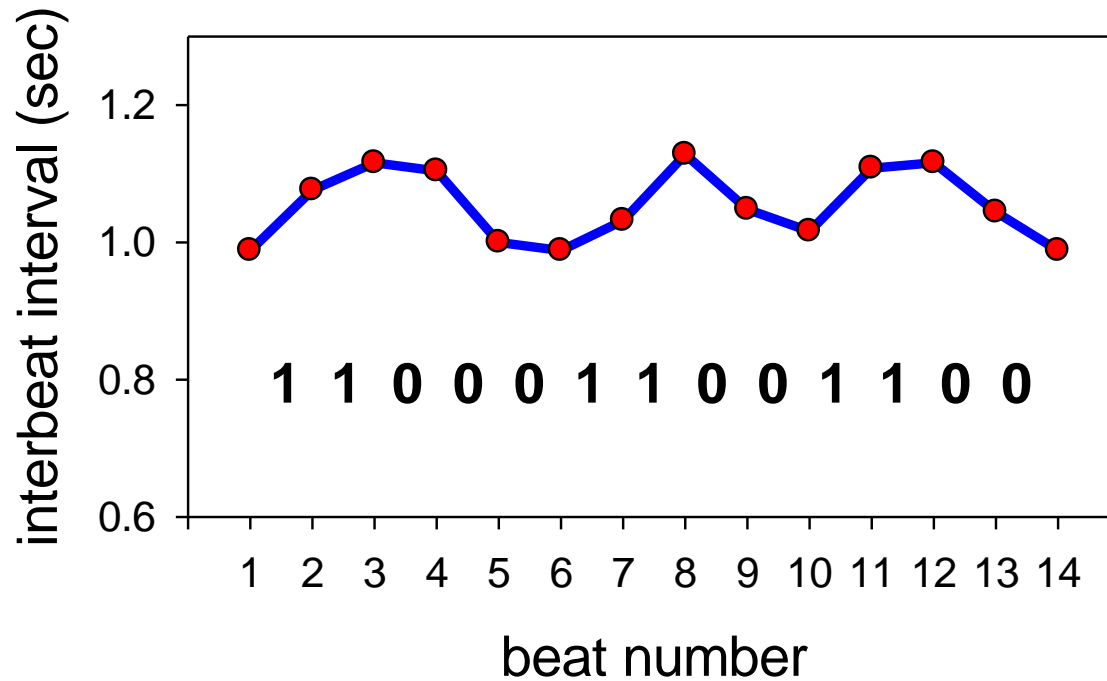
- How to map a heart rate time series to a symbolic sequence?



- How to define **words** in heart rate symbolic sequences?

KJ LFN HACUA RAFVT HTY AER FV
VA EVAC VAZ CFVDF VZ D SDSFV
SDFV NTEWOSI XW RXDP OIJR RO
IRFU FNV IMV MF

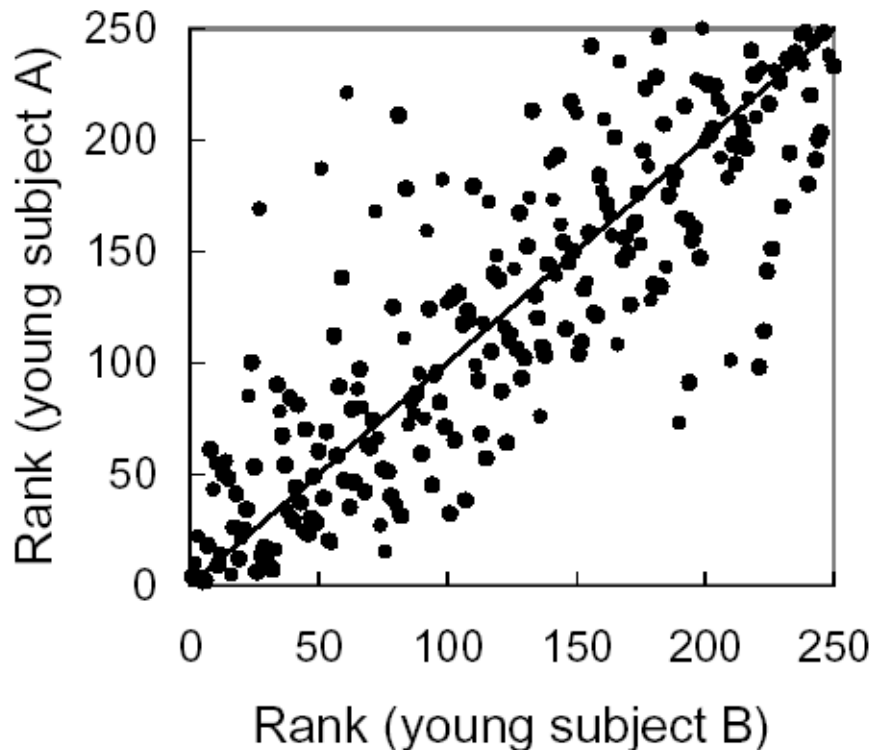
Symbolic Mapping



8-bit word: **11000110**, **10001100**, **00011001**

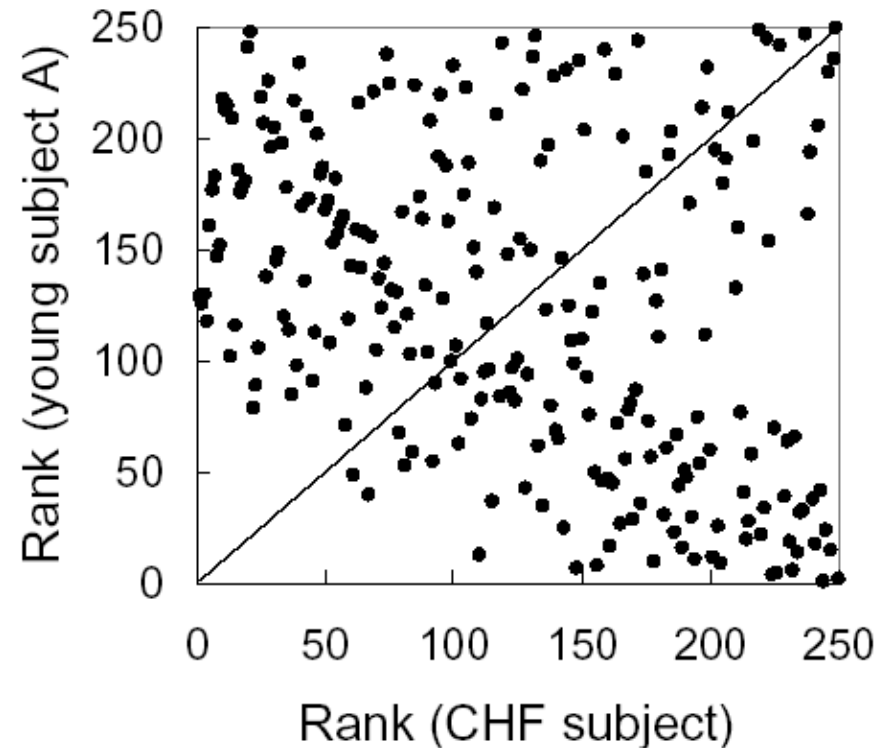
Comparison of Human Heartbeat

Health vs. Health



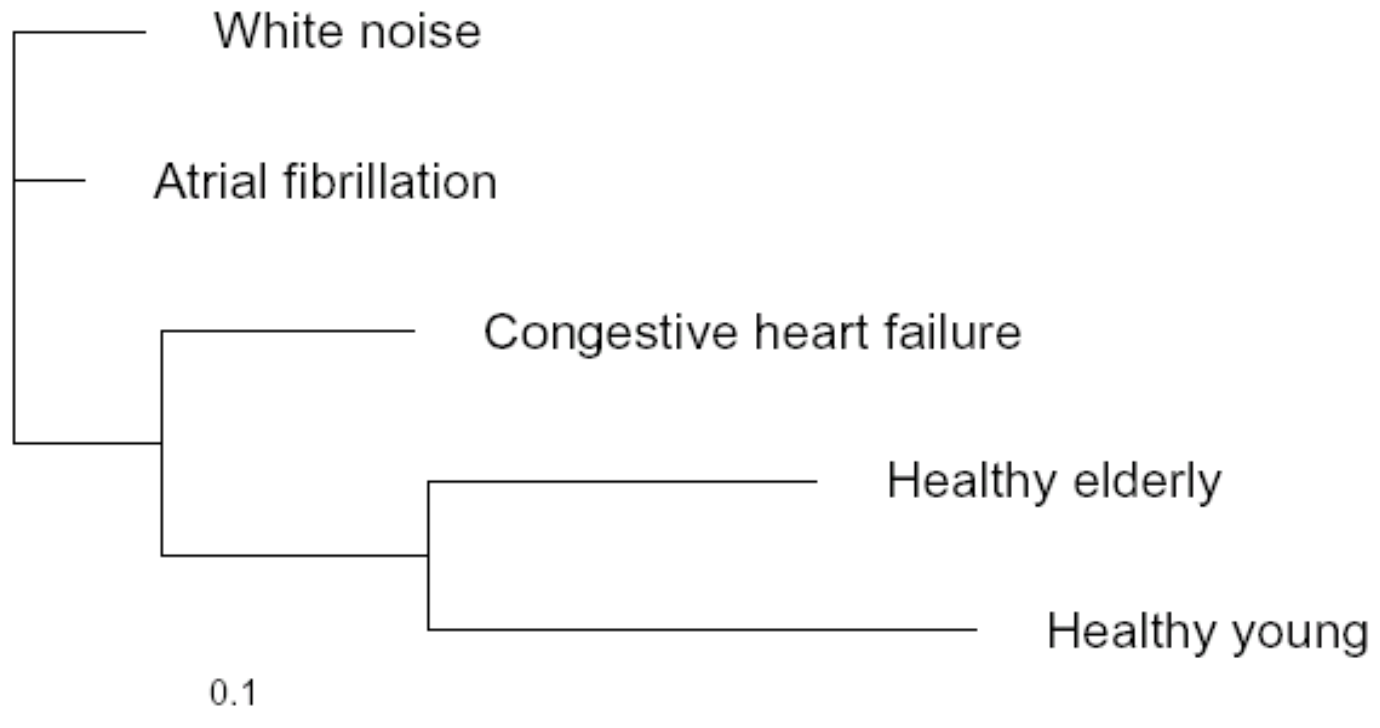
$D = 0.10$

Health vs. Disease



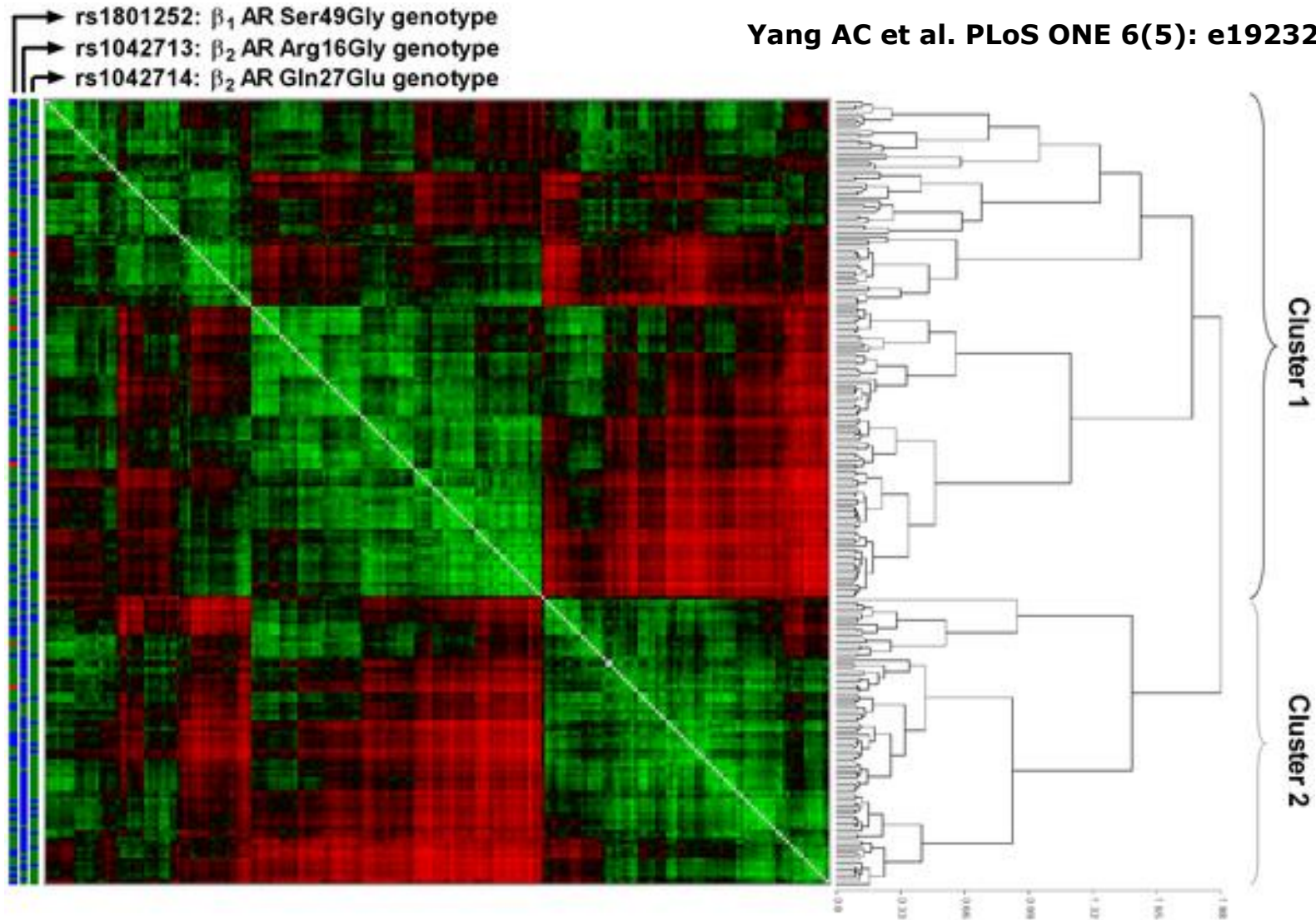
$D = 0.25$

Phylogenetic Tree of Human Heartbeat

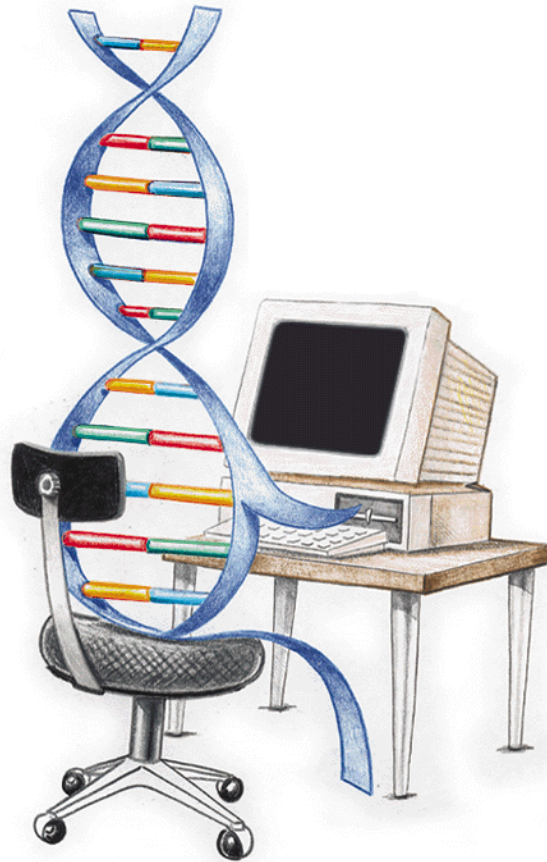


Clustering of Human Heartbeat Is Associated with β_2 -AR Gene Polymorphisms

Yang AC et al. PLoS ONE 6(5): e19232 (2011)



Application to Genetic Sequences



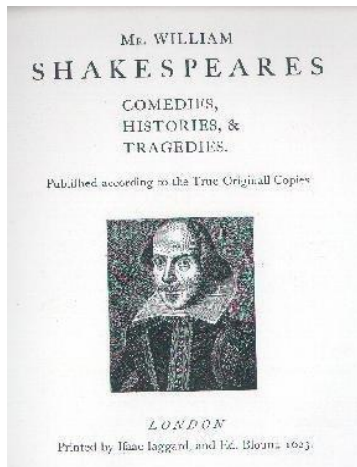
Analogy to Natural Languages

ATATTAGGTTTTTACCTACCCAGGAAAAGCCAACC
AACCTCGATCTCTGTAGATCTGTTCTCTAAACGA
ACTTTAAAATCTGTGTAGCTGTCGCTCGGCTGCATG
CCTAGTGCACCTACGCAGTATAAACAATAATAAA
TTTTACTGTCGTTGACAAGAAACGAGTAACCTCGTCC
CTCTTCTGCAGACTGCTTACGGTTTTCTGCCGTGT
TGCAGTCGATCATCAGCATACTAGGTTTCCGTCCGG
GTGTGACCGAAAAGGTAAGATGGAGAGCCTTGTTT
TTGGTGTCAACGAGAAAAACACACGTCCAACCTCAGT
TTGCCTGTCCTTCAGGTTAGAGACGTGCTAGTGCG
TGGCTTCGGGGACTCTGTGGAAGAGGCCCTATCGG
AGGCACGTGAACACCTCAAAAATGGCACTTGTGGT

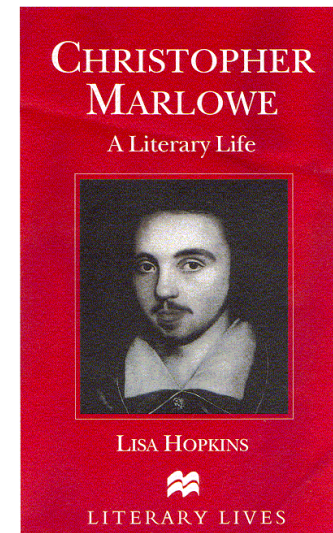
Similarity



ACTTAAGTACCTTATCTATCTACAGATAGAAAAGT
TGCTTTTAGACTTTGTGTCTACTTTTCTCAACTA
AACGAAATTTTGTATGCCCAGGATCTTTGATGCT
GGAGTCGTAGTGAATTGAAATTCATTGGGTT
GCAACAGTTTGGAAAGCAAGTGCTGTGTGTCCTAGT
CTAAGGGTTTCGTGTTCCGTCACGAGATTCCATT
TACAAACGCCTTACTCGAGGTTCCGTCCTGTGTTT
TGTGGAAGCAAAGTTCTGTCTTTGTGGAACCAG
TAACTGTTCCCTAATGGCCTGCAACCGTGTGACACT
TGCCGTAGCAAGTGATTCTGAAATTTCTGCAAATG
GCTGTTCTACTATTGCGCAAGCCGTCGCCGTTATA
CGGAGGCCGCTAGCAATGGTTTTAGGGCATGCCG



Similarity



DNA “Words”

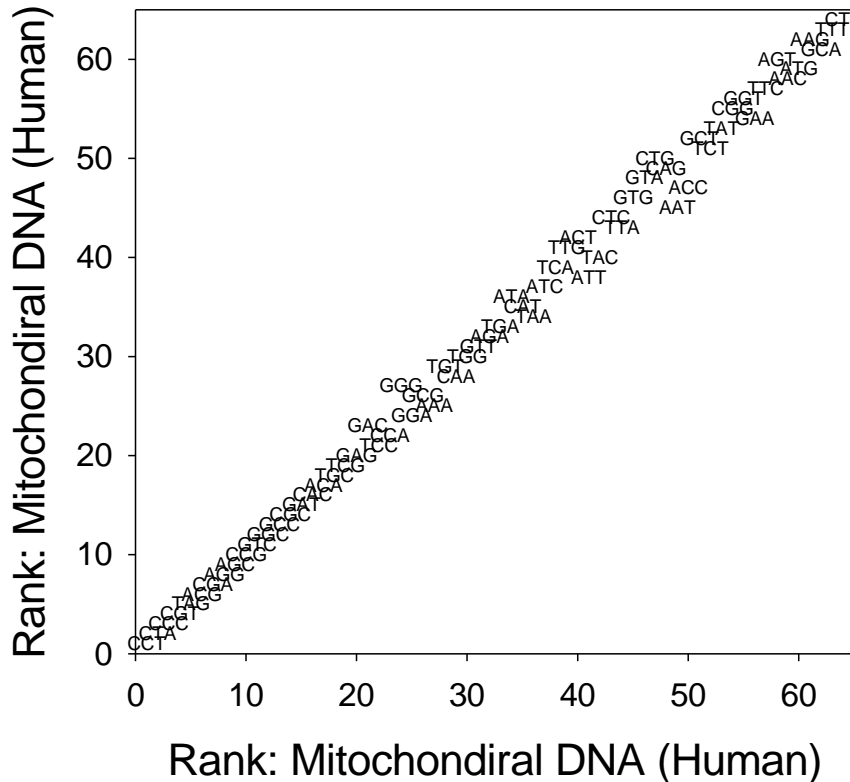
TACCCCCACTGTCAACCCAACACAGGCATG.....



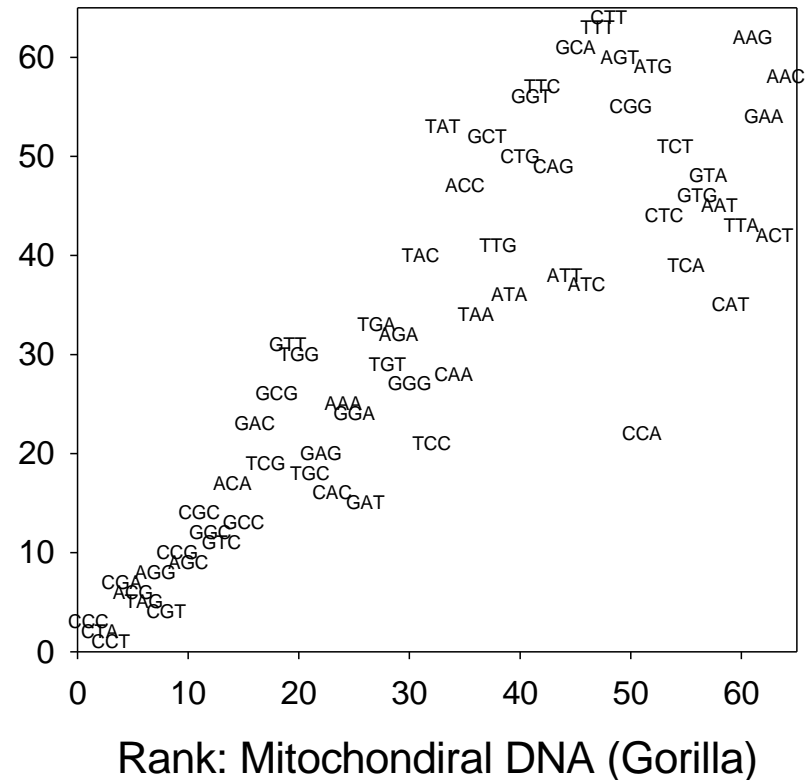
Word	Frequency	Rank
CCC	633	1
CCT	543	2
CTA	526	3
AAA	524	4
ACC	515	5
...

Rank Comparison Maps

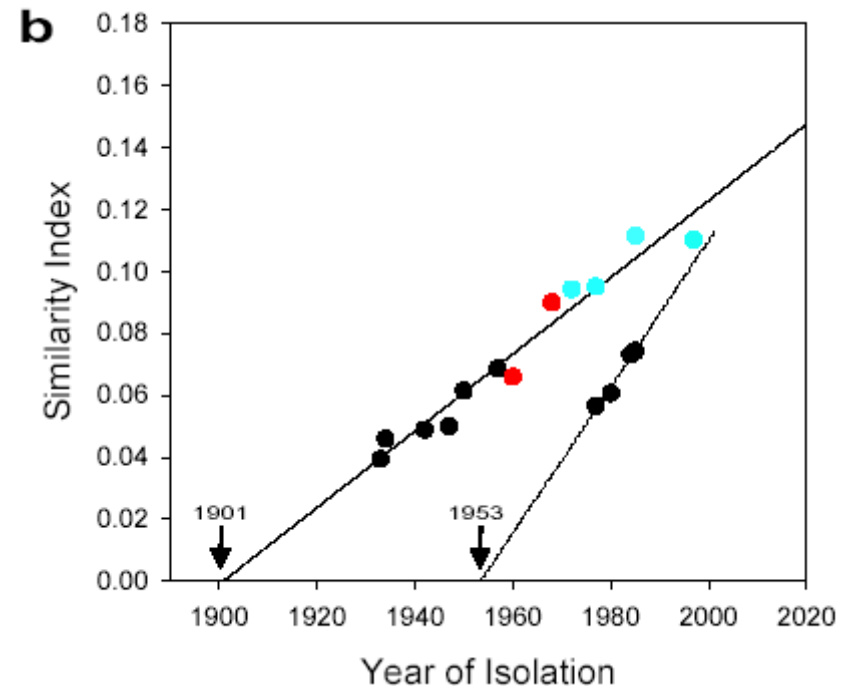
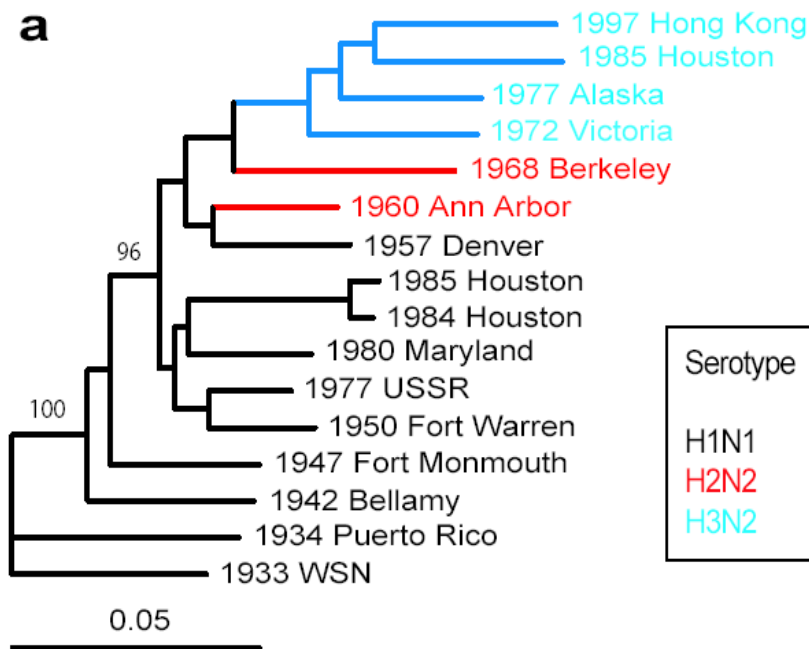
Same Species



Different Species

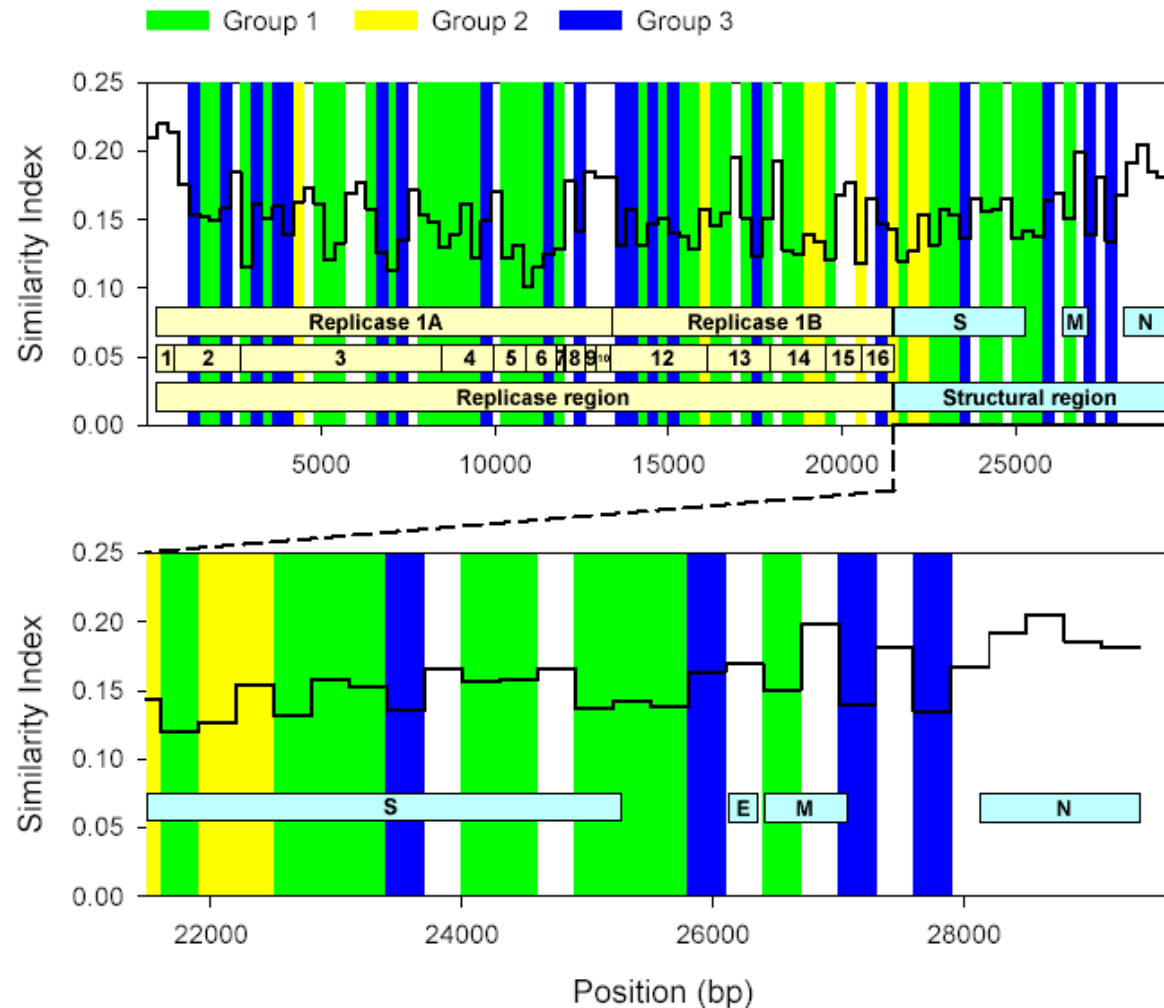


Human Influenza Virus

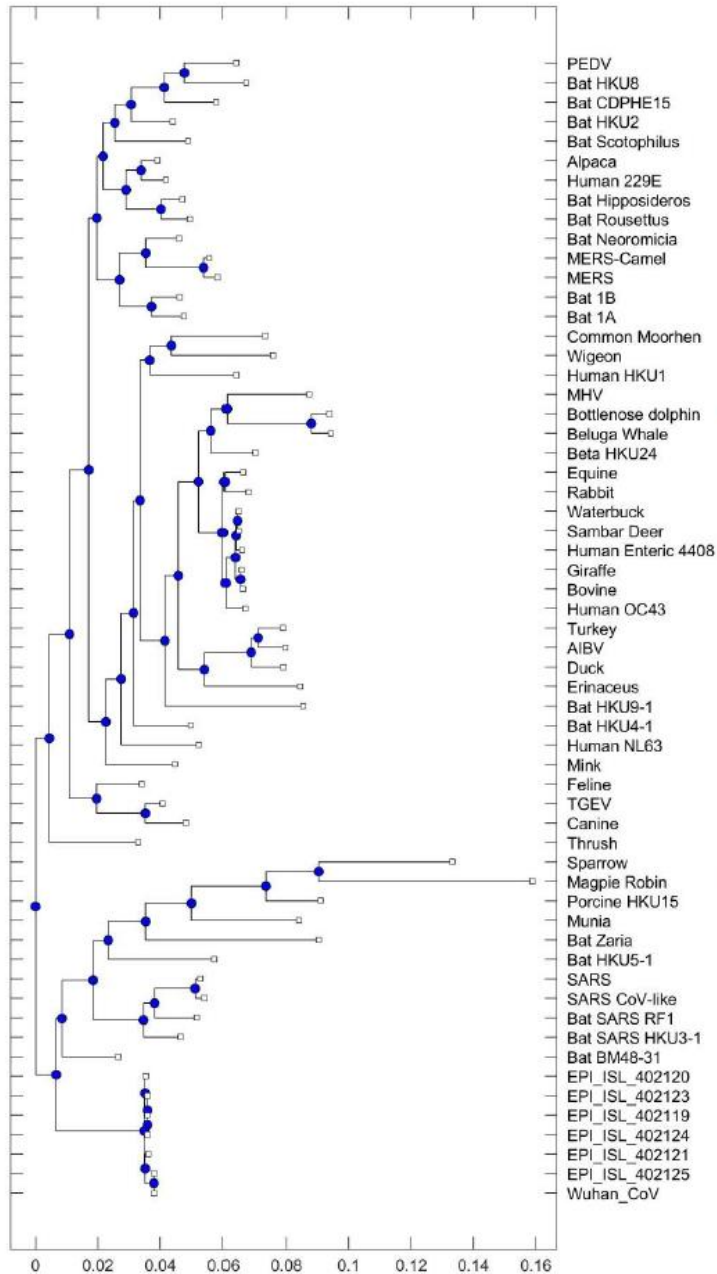


Our result is consistent with previous finding based on sequence alignment technique (*Science* 1986; **232**: 980)

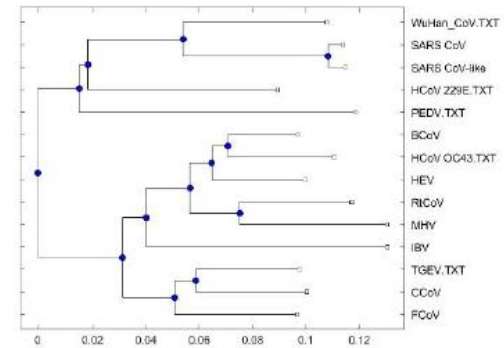
Genome-wide Sequence Comparison (SARS Coronavirus)



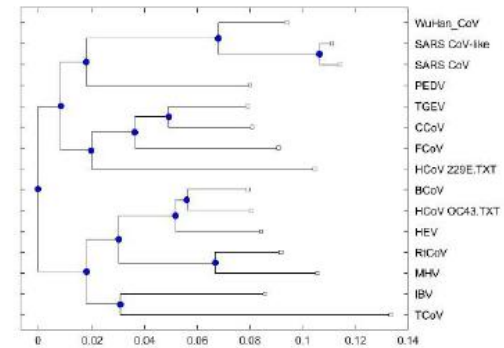
A Complete Genome



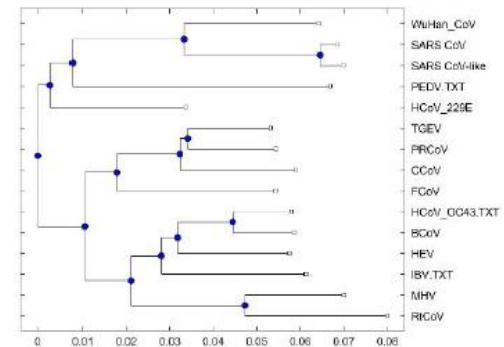
B Membrane Protein



C Nucleocapsid Protein

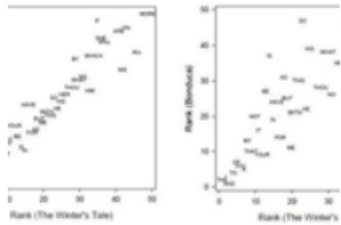


D Surface Protein



<https://www.mathworks.com/matlabcentral/fileexchange/46691-information-based-similarity-toolbox>

MATLAB Central ▾ | [Files](#) | [Authors](#) | [My File Exchange](#) | [Contribute](#) | [About](#)



Information-based Similarity Toolbox

version 1.2.0.0 (4.47 MB) by [Albert Yang](#)

Information-based similarity index is an analysis of measuring distance between symbolic sequences

Overview

[Functions](#)

The Information-based Similarity (IBS) method was developed to effectively categorize symbolic sequences according to their information content. The method has been fully described and validated (4), with applications to heart rate time series (1), literary authorship disputes (2), and genetic sequences (3).

This toolbox provides an array of MATLAB functions for quantifying the distance (or dis-similarity) between a set of symbolic sequences, and for displaying the results in graphical form such as dendrogram. The type of symbolic sequences can be binary sequences mapping from a time series, written texts of any given language, or genetic sequences.

References:

1. Yang AC, Hseu SS, Yien HW, Goldberger AL, Peng CK. Linguistic analysis of human heartbeats using frequency and rank order statistics. *Phys. Rev. Lett.* 90, 108103 (2003).

<http://reylab.bidmc.harvard.edu/pubs/2003/prl-2003-90-108103.pdf>