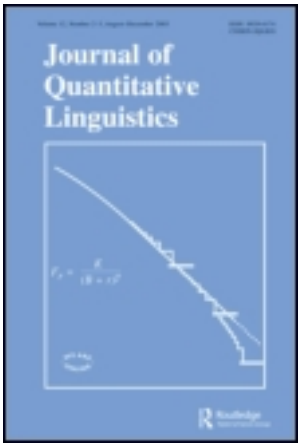

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/njql20

Historical Linguistics and Evolutionary Genetics. Based on Symbol Frequencies in Tamil Texts and DNA Sequences

S. Naranan [a]
[a] Chennai, India

Available online: 17 Nov 2011

To cite this article: S. Naranan (2011): Historical Linguistics and Evolutionary Genetics. Based on Symbol Frequencies in Tamil Texts and DNA Sequences, Journal of Quantitative Linguistics, 18:4, 359-380

To link to this article: http://dx.doi.org/10.1080/09296174.2011.608607

Journal of Quantitative Linguistics
2011, Volume 18, Number 4, pp. 337–358
http://dx.doi.org/10.1080/09296174.2011.608607

Routledge
Taylor & Francis Group

Historical Linguistics and Evolutionary Genetics. Based on Symbol Frequencies in Tamil Texts and DNA Sequences*

S. Naranan
Chennai, India

ABSTRACT

We have studied the rank frequency distribution (RFD) of letters of the alphabet in Tamil language texts. In a novel application of rank frequencies, we have defined a simple intuitive distance parameter between a pair of strings (text or DNA sequence of codons). This distance correlates well with age difference in historical linguistics and evolutionary genetics. Using a distance matrix of a set of strings, we derive evolutionary trees that are broadly in agreement with historical evidence. The method has potential for refinement and application in evolutionary studies to complement other approaches to evolution. The RFD in a single string conforms to a law called the CMPL (Cumulative Modified Power Law), which we had formulated and applied to RFD's of diverse symbol sets.

1. INTRODUCTION

We consider two different kinds of strings of symbols: language texts and DNA sequences. A language text consists of N symbols called ''tokens'' chosen from a set of V symbols called ''types''. For written text, the symbols are the letters of the alphabet. For speech, the symbols are phonemes that are the smallest units of speech. String size or the number of word tokens is variable but the symbol set V is

*Address correspondence to: S. Naranan, 20A/3 Second Cross Street, Jayaramnagar, Thiruvanmiyur, Chennai 600-041, India. Email: snaranan@gmail.com

constant for all the strings. DNA molecules, the genetic material of life are – at the most fundamental level – sequences of ''bases'', which belong to a set of four base molecules represented as A, G, C, T. Actual genetic information is, however, coded as triplet bases (AGC, GTC, etc.) known as codons. The 64 possible codons form the ''alphabet'' of DNA sequences.

For a given string of size N with V different symbols, the relative frequency of occurrence of the V symbols is a very valuable source of information. The statistical function that describes them has been explored extensively. But in this paper, we focus our attention on *a pair of strings* and define a measure of their ''separation'' or ''distance apart'' in terms of the differences in probability of a symbol in the two strings of the pair. This distance measure is expected to correlate positively with the chronological age difference in the evolutionary history of the strings. This idea was first tested for DNA molecules with an evolutionary time span of a few billion years. Relative codon frequencies were examined in 20 different species spanning this period and a positive correlation was seen between the ''distance'' and the difference in the evolutionary age of the two species.

A similar examination of evolutionary history in the literary texts of a language is possible using the relative letter frequencies of the alphabet in a pair of texts. The distance measure in this case can be expected to correlate with the chronological age difference of the texts. For applying this new tool for ''historical linguistics'', one needs literary texts spanning a long period. The most suitable language for this effort is Tamil, which has an almost uninterrupted record of literary works spread over the last 2500 years. Tamil is perhaps the only language which has its basic structure relatively stable over such a long period. We use data from seven Tamil literary texts (Sections 3, 4). There is significant agreement between the historical ages (which are somewhat uncertain) and the distance measures. It is interesting that the two types of strings we have considered – DNA sequences and language texts – have time scales differing by a factor of over 1000 (billions of years vs. thousands of years). Yet the proposed tool of ''distance'' measure promises to be of significant value in both cases.

The relative probability of letter frequencies in a *single text* is also examined (Section 5). In Section 6, the entropies of Tamil texts as defined by Shannon's Information Theory are discussed. We conclude with a discussion and summary of the results.

2. TAMIL LANGUAGE, ALPHABET, TEXTS AND LETTER FREQUENCIES

Tamil is one of the oldest living languages and is spoken by over 80 million people spread over 40 countries. It is mainly spoken in South India and is one of the 22 scheduled languages of the Indian constitution. It belongs to the Dravidian family of 22 languages, which are distinct from the Indo-European Languages. The four languages Tamil, Malayalam, Kannada and Telugu have literary histories. The earliest epigraphic records of Tamil date back to 300 BCE, making it one of the oldest languages in the world. The earliest surviving literature is also in Tamil, dating to pre-Christian Era; the highly acclaimed *Tholkappiyam* and the *Sangam* literature are attributed to the period 300 BCE–300 CE. There is an uninterrupted record of Tamil literature spread over the last 2500 years. ''Tamil is perhaps the only example of an ancient classical tongue which has survived as a spoken language for more than 2500 years with its basic structure almost unchanged'' (*Encyclopaedia Britannica*, 1973 edition). In recognition of its antiquity the Government of India declared Tamil a ''classical language'' in 2004. The Tamil Diaspora is worldwide and Tamil websites on the Internet are growing in popularity. The Government of the State of Tamil Nadu organized a World Classical Tamil Conference, in June 2010 lasting five days and on a mammoth scale, to highlight classical as well as modern aspects of Tamil. For details, see reference at the end. Tamil is perhaps the only classical language which is still spoken extensively unlike the other classical languages such as Sanskrit and Latin.

The Tamil alphabet consists of 30 letters, 12 vowels and 18 consonants. The vowels are listed first, followed by consonants in Table 1, column 2. The number of vowel-consonants is 216; each is regarded as a syllable of two letters, a consonant and a vowel. The consonants ''*l*'' ''*n*'' ''*ṭ*'' (pronounced ''*zh*'' ''*ṉ*'' ''*ṟ*''), unique in Tamil are not represented fully in other Dravidian and Indian (Indic, Sanskritic) languages. Modern Tamil has six more letters, the *Grantha* characters borrowed from Sanskrit (*j*, *ṣ*, *s*, *h*, *kṣ*, *śrī*). Tamil alphabet is the most compact among the Dravidian languages, with only 30 letters since it excludes separate characters for closely related aspirated and unaspirated sounds: e.g. the character ''*k*'' represents also ''*kh*'', ''*g*'' and ''*gh*'' which occur in Sanskrit loanwords.

The author's father had acquired a Tamil typewriter (Bijou) made by a German company in 1937. The company had claimed it could

Table 1. Relative frequencies of letters in seven Tamil texts.

SER no.	Translation in English	1 MOD	2 BHR	3 KUK	4 KYK	5 SLP	6 TKL	7 TLK	2–7 POE	1–7 TAM	Huffman code TAM
1	**a**	150	129	156	145	142	126	139	837	**987**	010
2	**a:**	47	43	47	43	35	60	29	257	**304**	1110
3	**I**	78	75	71	69	68	64	83	430	**508**	0011
4	**I:**	4	7	6	5	7	7	3	35	**39**	1100001
5	**u**	77	70	58	72	73	74	87	434	**511**	0010
6	**u:**	4	3	6	6	6	5	5	31	**35**	11000000
7	**e**	17	22	20	23	22	24	19	130	**147**	11111
8	**e:**	13	20	12	11	11	11	14	79	**92**	110001
9	**ai**	27	32	28	23	29	30	29	171	**198**	10100
10	**o**	10	13	9	12	13	13	14	74	**84**	111100
11	**o:**	7	12	11	11	12	9	9	64	**71**	111101
12	**au:**	0	0	0	0	0	0	0	0	**0**	110000011
13	**k**	79	49	63	58	67	71	56	364	**443**	1001
14	**ng**	6	5	9	9	11	7	9	50	**56**	1100000
15	**c**	22	24	19	17	18	21	16	115	**137**	000110
16	**nj**	1	3	4	2	3	4	5	21	**22**	110000010
17	**t:**	36	25	25	27	32	25	25	159	**195**	10101
18	**n:**	11	16	16	16	15	20	15	98	**109**	100010
19	**th**	71	73	65	68	63	57	61	387	**458**	0111
20	**nh**	21	27	22	22	24	26	20	141	**162**	11001
21	**p**	44	40	33	33	39	42	49	236	**280**	00001
22	**m**	42	57	48	44	44	45	47	285	**327**	1101
23	**y**	29	36	35	36	38	34	38	217	**246**	01100
24	**r**	45	38	40	47	46	39	39	249	**294**	00000
25	**l**	31	36	34	32	33	35	33	203	**234**	10000
26	**v**	35	42	41	39	38	35	44	239	**274**	00010

(*continued*)

Table 1. (*Continued*)

SER no.	Translation in English	1 MOD	2 BHR	3 KUK	4 KYK	5 SLP	6 TKL	7 TLK	2–7 POE	1–7 TAM	Huffman code TAM
27	**l-**	7	10	9	9	12	8	16	64	**71**	0001110
28	**l:**	21	16	12	16	12	12	12	80	**101**	100011
29	**t**	27	24	32	36	36	43	41	212	**239**	01101
30	**n**	38	53	68	68	51	53	44	337	**375**	1011
	Total	1000	1000	999	999	1000	1000	1001	5999	**6999**	
	NTOT	22855	5056	16283	11855	4165	6355	12430	56144	**78999**	

1. MOD: Modern Prose (1946–57).
2. BHR: Bharathi.
3. KUK: Kamba Ramayanam (Uthara Kandam).
4. KYK Kamba Ramayanam (Yuddha Kandam)
5. SLP: Silappathigaram.
6. TKL: Thirukkural.
7. TLK: Tholkappiyam (Porul athikaram).

2–7 POE: Texts nos 2–7. 1–7. TAM: All Tamil texts (bold).
The frequencies are per thousand. Their " Total" does not add up to 1000 due to round-off errors. NTOT is the number of tokens in each text.
Column 2: Transliterated letters in English.

manufacture typewriters in any language with an alphabet that could fit into a standard English keyboard. Tamil was the only Indian language for which this was possible because of its compact alphabet. Lessons were devised to type in Tamil in blind touch at age 13. The typewriter is still a valued possession. The keyboard design has some novel, ingenious features.

This article is based on data on Tamil letter frequencies for seven Tamil texts given by Gift Siromoney (1963). The data are reproduced in Table 1. In column 2, the English transliterations of the Tamil letters, adopted by Siromoney are given; they mostly conform to the International Phonetic Alphabet (IPA) code. Samples were taken from Tamil poetry from six different works belonging to periods ranging from the beginning of the Christian Era to the modern period.

As the dates of these works cannot be fixed with any certainty the (accepted) order in which they were written is followed. *Tholkappiyam* (*Porul athigaram*) is the oldest and *Bharathi* is the most recent. Modern prose is a sample of over 20,000 letters from the prose works published in Madras State in 1946–1947. The seven texts are given in Table 1. All except the first are poetical works. The last five are ancient texts ranging from approximately third century to 12th century CE. BHR is early 20th century and MOD is from mid-20th century. The relative frequencies of letters in a text are given in Table 1 (cols 3 to 9) as number of occurrences per 1000. The total number of letters (N) in the sample text (letter tokens) is given in the last row (Siromoney, 1963). Whereas the ancient texts contain only 30 letters, the modern texts have in addition six Grantha characters, which are ignored in the counting.

From the seven texts, we have cumulated the data from all poetical works (2–7) in a new text POE (column 10). This is done by adding the frequencies of a given letter in all the six texts. The same is repeated for all the seven texts cumulated as a new text TAM (column 11).

3. DISTANCE MATRIX OF TAMIL TEXTS – A POSSIBLE NEW TOOL FOR HISTORICAL LINGUISTICS

We consider how different texts vary in the occurrence frequency of a letter, in particular how we can quantify the difference between two texts X and Y in terms of the letter frequencies in the two texts. For this purpose we focus on the different ranks of the same letter in two texts.

Here we adopt the "distance measure" between two species which we formulated in Balasubrahmanyan and Naranan (2000) based on codon frequencies. (Hereafter we abbreviate references to our previous works as BN or NB.)

Let x_j and y_j represent the ranks of the jth letter in the texts X and Y. The index j ranges over the n letters of the alphabet ($j = 1, 2, 3, \ldots, n$). Each text is thus represented by n numbers and so is an n-dimensional vector. We denote the vectors corresponding to X and Y as $\boldsymbol{X}$ and $\boldsymbol{Y}$ with components $\boldsymbol{X}$: $\{x_j\}$ and $\boldsymbol{Y}$: $\{y_j\}$, $j = 1, 2, 3, \ldots, n$.

The vector magnitudes $|X|$ and $|Y|$ are given by:

$$|X|^2 = \sum x_j^2 \qquad |Y|^2 = \sum y_j^2 \tag{1}$$

The summation Σ is over $j = 1, 2, \ldots, n$ in this section. The n components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are simply different permutations of ranks 1, 2, 3, $\ldots$, n. Hence

$$R_n^2 = |X|^2 = |Y|^2 = \sum j^2 = n(n+1)(2n+1)/6 \tag{2}$$

Therefore $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent points on an n-dimensional sphere of radius R_n. The distance D_n between these two vector points is given by

$$D_n^2 = \sum (x_j - y_j)^2 \tag{3}$$

The angle θ_n between the vectors is given by

$$\sin(\theta_n/2) = D_n/2R_n \tag{4}$$

What is the maximum possible distance $D_{n,\max}$? It corresponds to the circumstance when $\{x_j\} = (1, 2, 3, \ldots, n)$ and $\{y_j\} = (n, n-1, \ldots, 2, 1)$ – the ranks in one text appear reversed in their order in the other text. It can be shown from elementary algebra

$$D_{n,\max}^2 = n(n^2 - 1)/3 \tag{5}$$

We define a normalized distance

$$D_n^* = D_n/D_{n,\max} \tag{6}$$

so that D_n^* has a value 0 to 1. From Equations (2), (4), (5) and (6)

$$\sin(\theta_n/2) = D_n^* \sqrt{(n-1)/[2(2n+1)]} \quad (7)$$

D_n^* can be calculated from the vector components x_j, y_j ($j = 1, 2, 3, \ldots, n$). θ_n the angle between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (in radians) is obtained from Equation (7). For $n = 30$

$$\sin(\theta_n/2) = 0.488\ D_n^*$$

3.1 Distance Matrix for Tamil Texts

All the seven vectors corresponding to the texts 1 to 7 have the same magnitude R and the points lie on a 30-dimensional sphere of radius R. The angle θ_{ij} between the text i and text j (in degrees) for $i, j = 1, 2, \ldots, 7$ is the angular distance matrix or simply the "distance matrix". The matrix is given in Table 2. For instance the distance between text 1 (MOD) and text 5 (SLP) $\theta_{15} = 8.17$ degrees. The matrix is symmetric since $\theta_{ij} = \theta_{ji}$. The diagonal elements $\theta_{ii} = 0$. We observe the following salient features:

1. The largest distance is $\theta_{17} = 9.36$ degrees and the smallest is $\theta_{34} = 4.72$. Since the texts are indexed 1 to 7 in increasing order of their historical age, it is interesting that the maximum distance is between the texts farthest apart in age (1 MOD and 7 TLK). Similarly the minimum distance is θ_{34} is between neighbouring texts KUK (3) and KYK (4).
2. To examine the correlation between "distance" and "age" difference of two texts i and j we plot θ_{ij} *vs* $|i-j|$, the absolute value of $(i-j)$ in Figure 1. Each point in the figure is labeled with i, j. Circles are for the texts 3–7. There is some evidence of positive correlation.
3. The correlation improves if we remove "stragglers". In Equation (3) $(x_j - y_j)^2$ represents the contribution of the jth symbol to D_n^2. If a *single* symbol contributes more than 0.4 D_n^2 that symbol is defined as a straggler. Stragglers represent extreme cases which can mask an otherwise systematic trend in correlation. There are three such straggler symbols in θ_{26}, θ_{56}, θ_{67}, all of which have text 6 (TKL) common. The symbol is "*a:*" in θ_{56}, θ_{67} and "*t*" in θ_{26}. These two symbols have a very low rank (high frequency of occurrence) in TKL

compared to other texts. They contribute 45%, 42% and 46% respectively of D_n^2 in θ_{56}, θ_{67}, θ_{26}. The revised distances are indicated in Table 2. They are $\theta_{26} = 6.04$ (7.82), $\theta_{56} = 5.41$ (7.03), $\theta_{67} = 6.53$ (8.17). The numbers in parenthesis are the unrevised distances. Linear regression gives

$$\theta_{ij} = a + b|i-j| \qquad (i \neq j) \tag{8}$$

Table 2. "Distance matrix" (Theta) of seven Tamil texts.

	Theta	1 MOD	2 BHR	3 KUK	4 KYK	5 SLP	6 TKL	7 TLK
1	MOD	0.00	7.27	7.91	7.64	8.17	7.46	9.36
2	BHR		0.00	4.93	6.88	7.73	*6.04*	8.26
3	KUK			*0.00*	**4.72**	**6.82**	**6.88**	**8.59**
4	KYK			4.72	*0.00*	**5.14**	**6.40**	**7.91**
5	SLP			6.46	5.50	*0.00*	***5.41***	**5.78**
6	TKL			7.45	6.49	5.15	*0.00*	***6.53***
7	TLK			8.44	7.48	6.14	6.57	*0.00*

Theta (2, 6), (5, 6), (6, 7) in italics, exclude "stragglers" (see text).
Distance matrix of five ancient Tamil texts is shown in the box. ($i, j = 3, 4, 5, 6, 7$). In this box, distances above the diagonal, in bold are the "observed" distances, distances below the diagonal are the corresponding distances from the "evolutionary" tree. Both are presented in the same matrix for easy comparison.

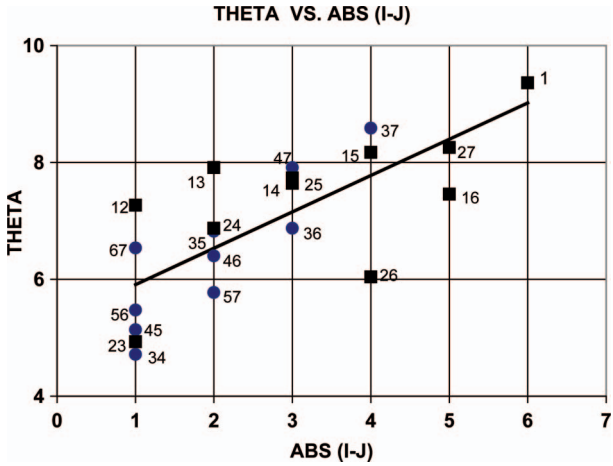

Fig. 1. Theta vs. abs($i - j$) for Tamil texts.

$a = 5.29 \pm 0.39$ and $b = 0.62 \pm 0.13$. Since the texts 1 to 7 cover a period of about 1800 years, *on the average* each step in $|i - j|$ corresponds to about 300 years and a distance interval of 0.62 degrees. Or every one-degree change in θ corresponds *on the average* to a change of age about 500 years. This is a very rough approximation since the age difference is not strictly proportional to $|i - j|$.

4. With larger sample sizes N and more texts, especially of known historical age, one can expect stronger correlation between distance and age. If so, one can infer the age of a text of unknown date using the distance-age relation.

3.2 Historical Evolutionary Tree of Texts Using the Distance Matrix

The positive correlation between the distance θ_{ij} and $|i - j|$ of two texts (i, j), has motivated us to use a more sophisticated analysis of the distance matrix that leads to a binary evolutionary tree of texts. For this analysis we focused only on the five ancient texts (3–7). The reasons for excluding MOD (1) and BHR (2) are as follows. For tree analysis the criterion of homogeneity among texts is crucial. MOD is the only prose text and so is excluded. Both MOD and BHR being modern, their alphabet includes six *Grantha* characters unlike the ancient texts. Although they may be only a small fraction of the texts and are not counted, it is not clear how their use affects the frequencies of the other letters closely related to the *Grantha* characters. So it is safe to exclude both MOD and BHR. The distance matrix of the ancient texts (3–7) is shown enclosed in a box in Table 2 (upper diagonal terms). The straggler symbols have been excluded (see previous section).

3.3 The Tree Algorithm

The algorithm for generating an evolutionary tree from an $m \times m$ distance matrix of m species is called the UPGMA algorithm commonly used in evolutionary genetics to obtain phylogenic trees of evolution of biological species (Ewens & Grant, 2001). The acronym UPGMA stands for ''Unweighted Pair-Group Method using arithmetic Averages''. Here is a brief description of the algorithm.

1. First, the smallest matrix element in the distance matrix is picked (say θ_{ab}). It is assumed that the species a and b diverged from a common ancestor species r_1. This means $\theta_{ab} = \theta_{ar1} + \theta_{br1}$.

2. The matrix is replaced by a new matrix in which the rows and columns corresponding to species *a, b* are deleted and replaced by a new row and column for the "species" r_1. This requires calculating θ_{zr1} where z represents all species other than *a, b*.

$$\theta_{zr1} = (\theta_{za} + \theta_{zb} - \theta_{ab})/2 \qquad z \neq a, b \tag{9}$$

3. The steps 1 and 2 are now applied to the new matrix which has a new "species" r_1 instead of *a, b*. A new common ancestor is found (r_2) and a new matrix is created with r_2.
4. The above procedure is iterated $m - 1$ times giving rise to $m - 1$ common ancestors $r_1, r_2, \ldots, r_{m-1}$ which are presumed extinct. The basic assumption here is that all the extant (current) species evolved as bifurcations of common ancestors along a time-line of biological evolution.

3.4 Evolutionary Tree of Five Ancient Tamil Texts

The evolutionary tree derived from the distance matrix of five ancient texts is shown in Figure 2. The smallest distance in the matrix (Table 2) is θ_{34} (4.72) between KUK and KYK. So r_1 is the "common ancestor" of KYK and KUK. Successive iterations create r_2 as the common ancestor

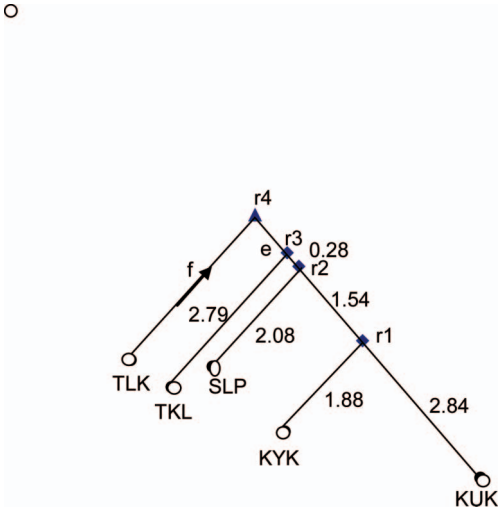

Fig. 2. Evolutionary tree of five ancient Tamil texts.

of SLP and r_1; r_3 as the common ancestor of TKL and r_2; r_4 as the common ancestor of TLK and r_3. The best distance estimate between the nodes (in this case r_1, r_2, r_3, r_4) and the texts can be calculated and are shown (in degrees) in the figure. Note that $e\ (=\theta_{r4r3})$ and $f\ (=\theta_{r4r7})$ are not known individually but only their sum $e+f\ (=3.78)$. The distance matrix of the "best-fit" tree is also given in Table 2 (lower diagonal elements) for easy comparison of the two types of distances. The agreement between the tree matrix elements and the observed ones is very good. The maximum difference between the observed and tree values is for θ_{47}: observed 7.91 vs. tree value 7.48, a difference of 0.43 degrees. Here are some interesting comparisons between the optimum historical evolutionary tree derived from the observed distance matrix and known historical facts about the texts.

1. The two texts KYK, KUK are closest in time and KUK is later than KYK (Table 2). Both are attributed to 800 – 1100 CE, KYK written by Kamban and KUK by a different author at a later time.
2. TLK is the oldest text, farther in distance from any other text (see the last column in Table 2). TLK is the earliest known Tamil literary text ("*tol*" means "original" and "*kappiyam*" means literary epic) and is attributed to "*Sangam*" period of secular (non-religious) poetry, early CE (c 0 – 300 CE).
3. TKL and SLP are of an intermediate historical age between TLK and KYK. Both are believed to belong to 500 – 700 CE. *Thirukkural*, TKL a didactic collection of cryptic aphorisms (couplets of seven words each) is highly esteemed worldwide. Though it is thought that TKL preceded SLP the tree, as well as observation, suggest TKL may be later than SLP. A clearer picture may emerge by increasing the sample size of texts for distance analysis. TKL is a short text of 9310 words – 1330 couplets of seven words each. The current sample of letter-tokens is perhaps only a 10th of the total size.
4. The tree analysis of distance matrix yields ages for texts broadly in agreement with historical facts. The approach certainly has potential as a tool for research in historical linguistics. While the older texts are difficult to date, there are many literary texts of intermediate age with known dates – e.g. Pallava (6th to 9th century CE), Chola (9th to 12th century CE) and Nayak (13th to 17th century CE) periods. They can be used to calibrate the historical time scale. It is necessary

to emphasize here that the distances indicated in Figure 2 may not be proportional to time – in other words the evolutionary clock marking variation in letter usage in texts may not be a regular uniformly ticking clock at different epochs of history. Distance matrix analysis can complement historical evidence, and each can motivate research in the other to help resolve some uncertainties.

4. DNA SEQUENCES: DISTANCE MATRIX AS A POSSIBLE TOOL FOR EVOLUTIONARY GENETICS

4.1 DNA sequences

DNA molecules in living cells carry genetic material (genes) with information that determines structures and regulation of proteins vital for life's functions. DNA contains up to millions of bases in a linear sequence. There are four types of bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Hence DNA sequence is a string of letters A, G, C, T from a four-letter alphabet. However, genetic information is carried by codons each made up of three bases. The 64 possible triplet bases (AGC, CAT etc.) form an alphabet of 64 symbols. This is the level that is used for our analysis: DNA as a string of 64 codons. The genetic code maps 64 codons onto 20 amino acids and a STOP symbol for translation of a gene into its corresponding protein which is a string of 20 amino acids. The STOP codon is needed to signal the termination of translation process. The genetic code is almost universal across all species.

Organisms are classified into two types: 1. Prokaryotes (bacteria). 2. Eukaryotes (non-bacterial). Eukaryotes are complex organisms with a nucleus in the cell surrounded by cytoplasm, both with their own brand of DNA. Nuclear DNA sequences are long (up to millions of bases). Cytoplasm has short stretches of DNA, the mitochondrial DNA (mtDNA); they are specialized parts of the cell (organelles) for generation, release and storage of energy. Each cell has thousands of copies of mtDNA. Bacteria do not have mtDNA.

We have analysed DNA sequences of 20 different species spread over the last four billion years of evolution classified into 10 phyla (NB, 2000; BN, 2000). The codon frequencies are given by Wada et al. (1992). The 20 species are divided into two sets A and B. Set A consists of 10 species one from each phylum with the largest sample of codons in Wada's data.

Set B is similar to set A but has species with the second largest number of codons. The species are:

Set A: HUM, MUS, BOV, CHK, DRO, YSC, ECO, FLA, PT4, RIC CP
Set B: CHP, RAT, RAB, XEL, TRB, MZE, BAC, VAC, LAM, TOB CP

In addition mtDNA for seven species have been analysed (HUM, MUS, RAT, BOV, DRO, YSC, MZE). For details see Balasubrahmanyan and Naranan (2000).

4.2 Phylogenic Tree of Six Species in Biology from the Distance Matrix

The first application of the concept of distance between two species based on rank frequency distribution of 64 codons of DNA sequences was to study the evolutionary age of species relative to HUM (*Homo sapiens*) over a period of four billion years. We obtained θ_{HX} the distance between HUM and species X, X standing for any one of the 20 species mentioned in Section 4 (Balasubrahmanyan & Naranan, 2000). The species were indexed: $i = 1, 2, 3, \ldots, 10$, $i = 1$ representing HUM and $i = 10$, the most ancient organelles. A plot of θ_{Hi} ($i = 2,3, \ldots, 10$) showed good linear correlation

$$\theta_{Hi} = 4.37\ (i - 1). \tag{10}$$

This is to be compared with Equation (8) for Tamil texts. For species the evolutionary time scale is four billion years; Equation (10) indicates that on the average θ increases by one degree every 100 million years. In comparison for Tamil texts one degree increase corresponds to an increase in average age of about 500 years.

With a distance matrix of m species θ_{ij} ($i, j = 1, 2, \ldots, m$), we have the means to derive an evolutionary phylogenic tree (Section 3). Here we present the distance matrix for 6 mammalian species including three great apes, *chimp* (CHP), *gorilla* (GOR) and *orangutan* (ORA). The other three are *Homo sapiens* (HUM), *Mouse* (MUS) and *Rat* (RAT). In Section 4, we mentioned two types of DNA, nuclear and mitochondrial (mtDNA) in cells. For evolutionary studies mtDNA is preferred since it is inherited only from the mother. So we have chosen to analyse the mtDNA of the six species mentioned. Further, instead of using the 64 codons of the DNA sequence we have used the corresponding 20 amino acids and one STOP symbol (as dictated by the genetic code), as the symbols for rank frequency analysis. The data is from Kazusa (2004).

The distance matrix based on 21 symbols RFD is given in Table 3. The phylogenic tree derived from the matrix using the UPGMA algorithm (Section 3) is given in Figure 3. The distances derived from the tree are given in Table 3 (lower diagonal terms). The agreement between the observed and derived distances is very good. The root mean square of the difference between observed and derived distances is 0.52. We mention only some of the interesting features of the tree.

1. The smallest distance is between CHP and GOR (3.27) and the next to smallest distance is between MUS and RAT (6.36).

Table 3. "Distance matrix" (Theta) of six mammalian species.

	Theta	1 HUM	2 CHP	3 GOR	4 ORA	5 MUS	6 RAT
1	HUM	0.00	**9.00**	**8.02**	**6.70**	**9.57**	**10.03**
2	CHP	8.93	0.00	**3.27**	**14.44**	**14.73**	**14.95**
3	GOR	8.14	3.32	0.00	**13.46**	**14.32**	**14.15**
4	ORA	7.34	13.74	12.96	0.00	**7.28**	**8.25**
5	MUS	8.95	15.36	14.57	7.58	0.00	***6.36***
6	RAT	9.45	15.86	15.07	8.08	6.36	0.00

Distances above the diagonal, in bold are the "observed" distances. Distances below the diagonal are the corresponding distances from the "phylogenic" tree. Both are presented in the same matrix for easy comparison.

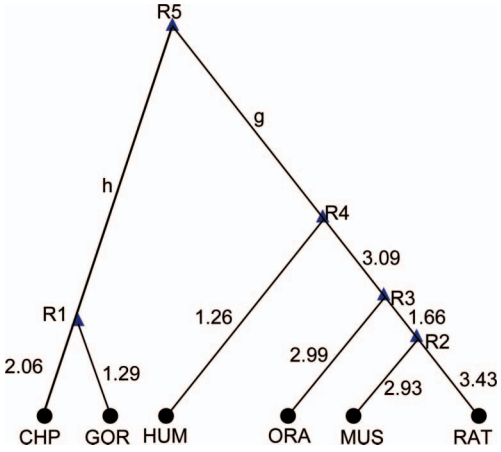

Fig. 3. Phylogenic tree of six mammalian species.

2. HUM is closest to ORA (6.70). HUM to CHP is 9.00 but ORA and CHP are far apart (14.44). CHP and GOR belong to a branch different from ORA.
3. The branches of the tree from the nodes or the common ancestors (r_1 to r_5) all terminate on the same base line since all the species are extant. Their varying lengths reflect different rates of evolution after bifurcations occurred at common ancestors. For example, the common ancestor of (MUS, RAT) and HUM is r_4. θ_{r4H} is 1.26 whereas θ_{r4MUS} is 7.68 and θ_{r4RAT} is 8.18. This suggests that MUS and RAT evolved about 6 times faster than HUM after the split at r_4 if we assume the rate of change of distance with time is uniform.

Phylogenic tree analysis in Evolutionary Genetics is very complex and different approaches are known to sometimes yield contradictory results. The method used here based on a random sample of the whole genome has its merits and demerits. The analysis is very simple and databases of codon frequencies in DNA sequences are readily available. Variations across individual genes are averaged out in deriving an over-all gross evolutionary picture; but the time resolution is likely to be poor for estimates of evolutionary age. Obviously the method can only complement other already currently used methods. A common method used in evolutionary genetics is to count the number of base changes in a conserved gene across a long period of evolution, as a measure of evolutionary age since the base changes occur due to mutations which are accumulated over time. It is worth noting that the RFD of codons in single genes can also be used instead of a random sample of the whole genome. The evolution of the changes in ranks of codons with time is also closely related to the base changes caused by mutations.

5. RANK FREQUENCY DISTRIBUTION (RFD) OF SYMBOLS IN LANGUAGE TEXTS AND DNA SEQUENCES

So far we have seen how the relative frequencies of symbols occurring in a pair of strings, can be used as a measure of distance or "separation" between them. In Tamil texts, the symbols are the letters of the alphabet, and in DNA sequences the symbols are the 64 codons. For protein sequences the symbols are the amino acids and the STOP symbol.

What can we learn about the RFD of symbols in a single string? For each string the frequencies of symbols are ordered in descending order such that the symbol with the highest occurrence has rank 1. The frequency of the symbol of rank r is $p(r)$, $r = 1, 2, \ldots, V$, where V the total number of symbols is a constant. Earlier, we had deduced a mathematical function for $p(r)$, drawing upon concepts from Shannon's Information Theory and Statistical Physics (NB, 1993, 2005; BN, 1996, 2000). It was successful in describing the RFD of (1) phonemes in six spoken Indian languages, (2) S-words (articles, prepositions and conjunctions) in English texts and (3) codons in DNA sequences of 20 species. The function called the cumulative modified power law (CMPL) is given by

$$p(r) = \sum_{i=r}^{i=V} d(i) \tag{11a}$$

Here $d(i) = p(i) - p(i+1)$ for $i = 1, 2, \ldots, V-1$ and $d(V) = p(V)$. Further

$$d(i) = De^{-v/i} i^{-\delta} \tag{11b}$$

where D, v, δ are constants. D depends only on the string size (N) and is a scaling or normalizing constant. v, δ determine the shape of the function. $p(i)$ is a sum of "modified power laws", which are the terms $e^{-v/i} i^{-\delta}$.

For each of the nine texts in Table 1, we obtained the RFD of letters, $p(r)$ *vs.* r. They fit the Equations (11a, b) very well according to the Kolmogorov-Smirnov test (Keeping, 1962). The (v, δ) parameters for the texts 1, 2, ... , 7, TAM, POE are respectively

$$(-2.01, 0.52), (-2.96, 0.13), (-3.55, 0.10), (-3.08, 0.19),$$
$$(-3.7, -0.01)$$
$$(-2.82, 0.15), (-2.38, 0.35), (-3.25, 0.11) \text{ and } (-3.36, 0.08).$$

For details see NB (2006). A remarkable feature of the parameters v and δ is that they are statistically correlated. $v/(\delta - 1) \approx b$. b is a constant ≈ 3.2.

We had studied the RFD of 64 codons in DNA and mtDNA sequences of 20 species (Section 4.1) and found a similar feature: $v/(\delta - 1) \approx 3.1$ (BN2000). The spread of (v, δ) in Tamil texts is however smaller ($\delta = 0$ to

0.52) than the corresponding spread in DNA sequences ($\delta = -1$ to 2.0). There are some chance coincidences: (ν, δ) are nearly the same for the pair TAM (-3.25, 0.11) and a common bacterium E-Coli or ECO (-3.2, 0.12). Another close pair is TLK (-2.38, 0.35) and mtDNA of cow, BOV (-2.36, 0.37). The wide range of (ν, δ) suggests the possibility of distinguishing different texts from the values of ν and δ. In other words, the values of ν and δ are signatures characterizing an author, a text, a language or a species. Finally we note that CMPL is not system-specific and may have a wider application in all sciences. It is derived from a model based on information theory and statistical physics and is not an exercise in curve fitting.

Closely related to CMPL is the function MPL, modified power law, which fits very well word frequency distributions in language texts.

$$W(k) = C\, e^{-\mu/k}\, k^{-\gamma} \tag{12}$$

Here $W(k)$ is the number of word-types that occur k times in the text. C is a size-dependent normalizing constant and (μ, γ) are shape parameters (NB, 1992a, 1992b, 1998, 2005; BN, 2002,). Baayen's book *Word Frequency Distributions* (Baayen, 2001) compares various mathematical functions describing WFD, beginning with the Zipf's Law: $W(k) = C/k^2$ (Zipf, 1935, 1949). Among other WFD's, MPL is also tested for several English literary classics (Lewis Carroll, H.G. Wells, Conan Doyle), a large corpus of about 6×10^6 words in English and a Dutch text. The sample size varies from 25,000 to 100,000 in most cases. For most texts $\gamma \approx 2.0$ and μ is in the range [0, 1]. Baayen refers to MPL as Naranan-Balasubrahmanyan-Zipf (NBZ) distribution. A CD ROM accompanying the book has several computer programs including one to fit the NBZ distribution to the observed WFD (*spectfit*).

Notice the close resemblance between MPL (Equation (12)) and CMPL (Equation (11b)). Our model was first used to derive the MPL and then extended to CMPL. A natural language text can be considered as a string of symbols of different kinds. When the symbols are words (vocabulary) we have the MPL for the WFD of word-types. The number of word-types grows with the size N of the text. When the symbol set is the alphabet, phonemes or S-words, the number of symbols V is constant, independent of N. In such cases we deal with the RFD of symbols which is well fit by the CMPL. For alphabet and phonemes $V \approx 20$–40 and for S-words in English $V = 71$ (BN, 1996).

Zipf's Law is valid for many European languages; however, it has not been tested for Indian languages, presumably because the word types are difficult to identify in a computer-based automatic analysis of word frequencies. Therefore, the letters of the alphabet as symbols are suitable for analysis of rank frequencies. Their number is fixed and small (20–40). CMPL for rank frequency distribution of letters, is the equivalent of MPL (NBZ) for word frequency distribution.

6. ENTROPIES OF TAMIL TEXTS

In Shannon's Information theory (Shannon, 1948), entropy or information is defined for a string of symbols as follows

$$H_s = -\sum_{i=1}^{n} P_i \lg P_i \tag{13}$$

Here $P_i > 0$ ($i = 1, 2, 3, \ldots, n$) ($\Sigma\ P_i = 1$), are the probabilities of occurrence of the n symbols and lg is logarithm to base 2. Shannon entropy H_s has a very practical application for communication: it is the minimum average binary code length per symbol, required for efficient communication. The code length of a symbol is determined by its probability; most frequently occurring symbols have short codes and the least frequent ones have long codes. The actual code assignments can be determined from P_i using an elegant algorithm by Huffman (1952). For a discussion of the relationship between entropy, information and complexity, see Balasubrahmanyan and Naranan (2005).

Siromoney (1963) calculated the entropy of Tamil prose (MOD) as $H_s = 4.34$ bits and gave the code assignments for different symbols. We have determined H_s for TAM as 4.428 bits and the code assignments are given in Table 1 (last col.). The average Huffman code length is 4.460 close to H_s. Information theory guarantees that this is the best possible code.

7. DISCUSSION AND SUMMARY

For strings with a small symbol set, the rank frequency distribution (RFD) is the basic data from which all the quantitative estimates of

parameters described, have been determined. For a given string, RFD of symbols can be described by CMPL with two parameters v and δ. There are indications that only one of them is an independent parameter. If the chosen text is a randomized sample of the whole text, it is to be expected that v, δ will be independent of sample size N. Since v, δ vary widely from text to text, they may serve as signatures to distinguish different texts (Section 5). The range of variation is greater in DNA sequences compared to language texts.

The difference in the ranks of a symbol in two strings is used to define a "distance" between the two strings. Note that this does not depend on any model such as CMPL for RFD of symbols; it is based on raw frequency data and a very intuitive concept of Euclidean distance generalized to n dimensions (n is the total number of symbols). A noteworthy feature is that only the *difference* in the ranks are used and not the ranks themselves: e.g. a symbol with ranks 3 and 4 in two texts contributes the same distance as a symbol with ranks 20 and 21, thereby giving equal weights to all symbols irrespective of their probability of occurrence. The angular separation between the two strings (DNA sequences or Tamil texts) shows good correlation with their age difference indicating that the method is a promising tool in evolutionary genetics and historical linguistics. This is illustrated by deriving an evolutionary binary tree from the distance matrix using a tree algorithm. Trees for five ancient Tamil texts and six mammalian species are broadly in agreement with known facts from history of literary texts and evolution of species.

The RFD of symbols in strings is a tool with interesting potential in evolutionary studies with possibilities of refining the technique for better results. By including strings of known age (texts or DNA species) one can calibrate the time scale involved. One might ask why such a correlation between distance and age is to be expected. In the case of DNA sequences, mutations of the four bases of DNA cause the relative codon frequencies to change with time and accumulate at a rate related to the average rate of mutation (estimated as about 1 in 100,000 years). Similarly in Tamil texts, as word usage changes over historical times, connected changes occur in the relative frequencies of letters of the words. Such changes also accumulate in time showing a net positive correlation between distance and age. The heartening aspect of this type of analysis is that string sizes (genes, literary texts) can be sufficiently large to yield parameter estimates of high statistical accuracy. In this

respect mtDNA suffers in comparison to nuclear DNA since mtDNA's are short in length. But mtDNA is preferred for evolutionary studies since it is inherited only from the mother. It will be interesting to compare evolutionary histories of individual genes conserved over long periods of evolution; from gene to gene and from mtDNA to nuclear DNA. For the choice of symbol one has 64 codons or 20 amino acids + STOP symbol (21 symbols). Evolutionary tree from a distance matrix based on RFD of symbols can complement other evidence, e.g. fossil, linguistic and anthropological evidence.

ACKNOWLEDGMENT

I thank my daughter Gomathy Naranan for a careful reading of the paper and valuable suggestions.

REFERENCES

Baayen, R. H. (2001). *Word frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Balasubrahmanyan, V. K., & Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, *3*, 177–228.

Balasubrahmanyan, V. K., & Naranan, S. (2000). Information theory and algorithmic complexity: Applications to language discourses and DNA sequences as complex systems: Part II: Complexity of DNA sequences, analogy with linguistic discourses. *Journal of Quantitative Linguistics*, *7*, 153–183.

Balasubrahmanyan, V. K., & Naranan, S. (2002). Algorithmic information, complexity and Zipf″s law. *Glottometrics*, *4*, 1–26.

Balasubrahmanyan, V. K., & Naranan, S. (2005). Entropy, information and complexity. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds), *An International Handbook of Quantitative Linguistics* (Chap. 61, pp. 878–891). Berlin, New York: Walter de Gruyter.

Ewens, W. J., & Grant, G. R. (2001). *Statistical Methods in Bioinformatics – an Introduction*. Berlin, New York: Springer Verlag.

Huffman, D. A. (1952). A method for the construction of minimum redundancy codes. *Proceedings of the Institute of Radio Engineers*, *40*, 1098–1101.

Kazusa (2004). Codon Usage Tables. Retrieved January 24, 2004, from http:www.kasuza.or.jp/codon.

Keeping, E. S. (1962). *Introduction to Statistical Inference*. New York: van Nostrand.

Naranan, S., & Balasubrahmanyan, V. K. (1992a). Information theoretic models in statistical linguistics – Part I: A model for word frequencies. *Current Science*, *63*, 261–269.

Naranan, S., & Balasubrahmanyan, V. K. (1992b). Information theoretic models in statistical linguistics – Part II: Word frequencies and hierarchical structure in language – statistical tests. *Current Science*, *63*, 297–306.

Naranan, S., & Balasubrahmanyan, V. K. (1993). Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *Journal of Scientific and Industrial Research*, *52*, 728–738.

Naranan, S., & Balasubrahmanyan, V. K. (1998). Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics*, *5*, 35–61.

Naranan, S., & Balasubrahmanyan, V. K. (2000). Information theory and algorithmic complexity: Applications to language discourses and DNA sequences as complex systems: Part I: Efficiency of the genetic code of DNA. *Journal of Quantitative Linguistics*, *7*, 129–152.

Naranan, S., & Balasubrahmanyan, V. K. (2005). Power laws in statistical linguistics and related systems. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds), *An International Handbook of Quantitative Linguistics* (Chap. 50, pp. 716–738). Berlin, New York: Walter de Gruyter.

Naranan, S., & Balasubrahmanyan, V. K. (2006). Statistical analogs in DNA sequences and Tamil language texts. Rank frequency distribution of symbols and their application to Evolutionary Genetics and Historical Lnguistics. *Festschrift in honor of Professor G. Altmann on his 75th birthday* (pp. 483–496). Berlin, New York: Walter de Gruyter.

Shannon, C. E. (1948). A mathematical theory of communication. I, II. *Bell System Technical Journal*, *27*, 379–423, 623–656. Reprinted in Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois.

Siromoney, G. (1963). Entropy of Tamil Prose. *Information and Control*, *6*, 297–300.

Wada, K., Wada, Y., Ishibashi, F., Gojobori, T., & Ikemura, T. (1992). Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Research*, *20*, 2111–2118.

World Classical Tamil Conference 2010. Retrieved July 29, 2010, from http:en.wikipedia.org/wiki/World_Classical_Tamil_Conference_2010.

Zipf, G. K. (1935). *The psychobiology of language*. New York: Houghton Mifflin Co. Reprinted (1968). Cambridge: MIT Press.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Reading: Addison-Wesley.