JOURNAL OF APPLIED
COMPUTER SCIENCE
Vol. 24 No. 1 (2016), pp. 37-49

Evolutionary Ordering of the Mitochondrion-Encoded Proteins

Bohdan Kozarzewski

University of Information Technology and Management
Faculty of Applied Computer Science
H. Sucharskiego 2, 35-225 Rzeszow, Poland
bkozarzewski@wsiz.rzeszow.pl

Abstract. *The parsing of a symbolic sequence into a set of short substrings called words invented by the author is used for a new definition of the distance between sequences. No sequence alignment is necessary. The most frequent among spectra of multiprotein sequences are selected and considered as a reference spectrum of the sequences. The distance between the reference spectrum and protein sequences is considered as the estimation of the evolutionary distance of the protein. As an application, amino acid sequences of the several mitochondria-encoded proteins of mammal species are ordered according to their evolutionary distance. Statistical distribution of the distances between exhibits some structures related to the evolutionary rate in the past.*

Keywords: *symbolic sequence, parsing, distance.*

1. Introduction

The main objective of this article is to explore a new measure of the distance between symbolic sequences.The majority of DNA or protein sequence analyses rely on previous aligning the corresponding sequences. Then the evolutionary distance is defined as the number of substitutions per site which have occurred since a pair of sequences diverged from a common ancestral sequence. However, the alignment algorithm suffers from inherent drawbacks, in particular for long sequences, see, for example, [1] and references therein.

Recent accumulation of the nucleotide and amino acid sequences data combined with the computational power have opened to realise of Zuckerkandl and Pauling [2], of reconstructing amino acid sequences of ancestral proteins by tracing changes in the sequences of related proteins found in contemporary organism. A prerequisite to the reconstruction is chronological ordering of amino acid sequences with the use of some evolutionary distance. All method used so far for chronological ordering of genetic sequences need as an input both the multiple sequence alignment of the available sequences and the corresponding phylogenetic tree. They output a statistical inference of the paternal sequence at any internal node of the phylogenetic tree. Any uncertainties associated with phylogenetic hypotheses and methodological issues associated with the inference of extinct protein sequences can lead to a false reconstruction, [3], [4]. Williams *et al.,* [5] performed computational population evolution simulations and compared the thermodynamic properties of the true ancient sequences with the properties of the sequences inferred by the mentioned methods. They concluded that the methods may sometimes lead to an incorrect reconstruction of of the functional properties of an ancestral sequence.

In the present paper a new method for inferring an evolutionary order among amino acid sequences based solely on a set of sequences of related proteins present in an extant species, is proposed. No phylogenetic tree or evolutionary model is necessary. The tree can be inferred from the distance (or similarity) matrix of the set, as a by-product, as shown in [6]. The method starts with the mapping sequence onto a vector of short strings of different length called words. The notion of the word in the present paper differs significantly from the word defined so far, e.g. in [7]. The words follow from the parsing of the symbolic sequence proposed by Ke and Tong [8] to define a measure of complexity of the binary sequence. In [6] the algorithm, after modifications was used as a tool for discovering a set of words which are considered as patterns representing a symbolic sequence over an arbitrary alphabet. The set, which will be called a word spectrum, consists of ordered, distinct, non-overlapping words of size greater than two characters. The parsing algorithm and pattern matching approach together make a new tool for the symbolic sequence analysis.

DNA and protein sequences contain a large amount of information about their history. The information includes chronological order of related sequences. In the present paper the method is used for estimating the evolutionary chronology amino acid sequences of mitochondria-encoded protein families of primates as well as several other mammal species. How correctly chronological ordering of sequences is revealed depends on the set of extant taken into account. The more extant sequences are included in the set the more accurate order of sequences would be. The protein of a species may be analysed separately or as combined data set. The separate approach means that the estimate evolutionary time is computed for each protein and the average of the estimates over all proteins is used as a final estimate. However, it was shown by Nei *et al.,*[9] that the average estimate generally has an upward bias. A more reliable estimate is obtain by prior merging all thirteen protein sequences for each species into one longer multiprotein sequence of amino acids. This approach will be used in what follows. The set (as long as the average word spectrum) of the most frequent words among spectra of multiprotein sequences are selected and defined as a reference set of words. The distance between the reference set and protein spectra is considered as the estimation of the evolutionary distance of the protein. A statistical analysis of the distance between protein sequences found in contemporary species and the reference set of words considered as random variable x is performed. In particular, the cumulative distribution function of distance x reveals some structures related to the variability of evolutionary rate resulting from environmental changes in the past. It tells, among others, how many proteins in the contemporary species from the considered set is closer to to the reference set than x. There is some relation (so far unknown) between evolutionary distance and time of protein divergence. If it becomes known, the dating with the use of the evolutionary distance presented would be compared with the dating based on the fossil records.

2. Methods

2.1. Algorithm for extracting words

Suppose there is primary sequence C of symbols $c_1, c_2, ..., c_n$. Suppose S_t is a set of words parsed so far and the first symbol of the new word w is c_i . The word formed as a result of a procedure of appending symbol c_i by the following symbols in three steps.

Step 1. String $Q = c_i$ is neither periodic nor chaotic because there is only one symbol in it. So it has to be appended by the next symbol. Appending is continued

until some symbol c_{i+j+l} repeats one of the symbols, say k -th, in the string $Q = c_i, ..., c_{i+j}$.
Step 2. Let $P = c_k$ and $R = c_{i+j+1}$, so far they are equal. Both strings are appended $P = c_k c_{k+1}$, $R = c_{i+j+1}c_{i+j+2}$ and so on, until they become different. Then string Q found in Step 1 is appended by string R, and the new string is $Q = QR$.
Step 3. Set S_t of words is searched for the presence of string Q. If string Q is found, it is appended by the following (next to the last symbol of Q) symbol of C becoming $Q = Qc_{i+j+k+1}$. The appending is continued until some string $Qc_{i+j+k+l}$ does not replicate any word from S_t. The string $w = Qc_{i+j+k+l}$ becomes the new word of the spectrum representing sequence C. When several last symbols of C cannot be processed by the above replication, they are discarded (code available on request).

The result of the sequence decomposition is an ordered set of consecutive, distinct and non-overlapping, longer than two characters words which will be called the word spectrum of the symbolic sequence C. The code of the parsing algorithm is available on request. The spectrum is a very rich resource of information on the symbolic sequence over any alphabet.

2.2. Distance between sequences

Measuring the distance between symbolic sequences is essential in many data analyses. Let S_1 and S_2 be spectra of two sequences C_1 and C_2. The most natural distance measure is defined as the number of words in the set which includes words from S_1 and S_2 that are not in the intersection of S_1 and S_2 . It is convenient to work with normalized distance, which can be written as

$$dist(C_1, C_2) = 1 - \frac{2l(intersect(S_1, S_2))}{l(S_1) + l(S_2)}. \tag{1}$$

Here $intersect(S_1, S_2)$ is a set of words that the two spectra share (set theory intersection of S_1 and S_2), and $l(A)$ denotes the length (number of words) of set A . The value of distance measure $dist(C_1, C_2)$ varies between 0 when sequences C_1 and C_2 are mutual copies and 1 when the spectra are disjoint sets.

2.3. Union set of spectra

The sequence analysis often has to deal with a set C of sequences $C_1, C_2, ..., C_n$. The union set is defined as a set theory union of spectra $S_1, S, ..., S_n$ of all sequences $C_1, C_2, ..., C_n$

$$U = union(S_1, S_2, ..., S_n). \tag{2}$$

The union set includes words from all spectra but with no repetitions. The union set U of words representing set C plays a crucial role in the construction of the reference set of words.

2.4. Algorithm for the reference set

The present day proteins result from a long evolution of ancient life forms whose proteins have not survived long lasting destructive processes. Nevertheless, ancient proteins can be studied by means of appropriate mathematical techniques applied to protein sequences of contemporary species.

The reference set can help understand the evolutionary processes and mechanisms by which proteins have acquired their present functions. The fundamental assumption of the present approach says that the more frequent is the word in the spectra of extant species the older it is. The more old words are present in the spectrum of the particular sequence the less distant from the reference set it is and earlier the sequence diverged. The reference set plays the role of an outgroup (used in phylogenetic analyses) to determine the reference point on the distance axis for the remaining sequences.

The distance of a sequence can be considered as an instance of a random variable. If so, cumulative and partial distribution functions of the variable can be defined. Both functions carry information about some past events in the evolution of the protein sequences of species. In this way the distribution of distances of the protein sequences of contemporary species to reference set can provide some insight into the evolutionary history of their predecessors.

Let $C = C_1, C_2, ..., C_n$ be a set of extant sequences. The algorithm for building an reference set of set C consists of the following steps:

Step 1. Find word spectra $S_1, S_2, ..., S_n$ of all sequences.

Step 2. Generate the union set of the spectra, suppose it is the set of words $U = w_1, w_2, ..., w_N$. It is convenient to represent the set as a column N vector.

Step 3. Determine intersection of spectrum S_1 with the union set. It is a set of words the spectrum shares with U. It is convenient to represent the set as a numeric column N vector, it is the sparse vector. Its i-th nonzero component is equal to the index of word w_1 in spectrum S_1. The vectors corresponding to all the spectra form n columns of N rows each. The vector of union words from Step 2 appended by them forms the table of N rows and n+1 columns, that comprises all available data

on the words present in the set of n sequences.
Step 4. Find the mode, frequency and conservation number (it indicates how many spectra possess this particular word - cn in short) of every word from the union set. They all form a table consisting of N rows and 4 columns.
Step 5. Select a word of the highest frequency from the group of words of mode one. If there are several such words, then select one of the highest conservation numbers. Make the word and its attributes (mode, frequency and cn) the first row of a table A of 4 columns. Continue for words of mode 2, 3, .., until the mode approximately equals n. All the words in the first column of table A make a references spectrum of set C of extant sequences.

3. Results

3.1. Protein of 65 primate species

Thirteen families of protein sequences of N = 385 mammal species, including 65 of primates sequences were downloaded from NCBI, http://www.ncbi.nih.gov/. Their accession numbers and species names are listed in the Supplementary material, file Table1. For each family, the reference set of 385 species was found. Then the distance between 65 primate sequences and the reference spectrum were calculated and the primate species were ordered against the decreasing distance. Selected part of the results is given in Table 1

Table 1: Index of selected primates sequences in several protein sets

primate species	COX1	COX2	CYTB	NADH1	merged
Hylobates lar	23	58	42	35	29
Pan paniscus	38	65	39	40	35
Homo heidelbergensis	31	59	34	48	37
Homo sapiens	34	62	35	49	38
Homo s. neanderthal.	35	64	44	42	39
Homo s. Denisova	33	63	37	36	43
Pongo abelii	40	35	64	61	44
Gorilla gorilla	41	54	43	46	46
Papio hamadrya	56	38	58	63	58
Macaca mulatta	61	25	59	37	64

The last column shows corresponding indices when all thirteen protein sequences for each species were merged into one sequence. It follows from Table 1 that no protein evolved at the same rate over a long evolutionary time. Otherwise, the indices in all the columns of one row would not differ. A reasonable hypothesis is that if all the thirteen protein sequences are considered as a single sequence representing a species, the resulting evolutionary ordering of species would be more reliable. In what follows all mitochondrial-encoded protein of the species are merged into one multiprotein sequence. The details of ordering multiproteins are presented in file Table2 in the Supplementary material. Selected part of the results is given in Table 2.

Table 2: Index of selected primate multiproteins

species	index	dist	species	index	dist
Prolemur simus	1	0.668	*Homo s. Denisova*	43	0.785
Propithecus coquereli	2	0.671	*Pongo abelii*	44	0.786
Hylobates lar	29	0.768	*Gorilla gorilla*	46	0.788
Pan paniscus	35	0.780	*Pongo pygmaeus*	57	0.803
Homo heidelbergensis	37	0.781	*Papio hamadryas*	58	0.804
Homo sapiens	38	0.782	*Macaca mulatta*	64	0.822
Homo s. neanderthal	39	0.782	*Lophocebus aterrimus*	65	0.836

From the table it follows that *Pan paniscus* (bonobo chimpanzee) is less distant to reference set by 0.002 than the human which, on the other hand, is less distant than *Pongo abelii* (Sumatran orang-utan) by 0.003. It can be considered as one more opinion in the so far unsolved debate whether chimpanzees or orang-utans are our closest living relatives [9].

When the distances are considered as instances of a random variable, the experimental cumulative distribution function ($ecdf(s)$) of the variable can be introduced and extracted from the distance table. The $ecdf(s)$ function indicates how many species bear proteins that are equal or less distant than s from the reference spectrum. Fig. 1 presents *ecdf* function for the multiprotein of primate species.

All the distances between the reference spectrum and protein fall into (0.65, 0.85) range. It means that in the considered set of primates there are no contemporary species less distant than approximately 0.65 and more distant than approximately 0.85. Besides, let us note that there are ranges of distance where there are no species at all. The longest one is in the range (0.73,0.76) approximately 0.03

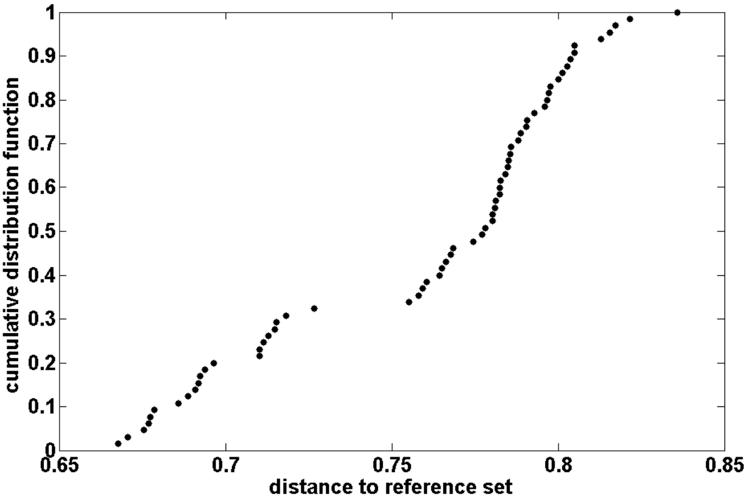

Figure 1: Experimental cumulative distribution function

wide. It is obvious that there is some relation between the defined distance and the real physical time elapsed since some unknown event in the past. There may be several reasons of the absence of extant species in some ranges of the distance (time) axis. For example:
No one living being can have mitochondrial-encoded proteins of these distances or the species have diverged in the related period of time but have not survived due to some disaster that had happened in the past.

If speciation and extinction rates were constant through time, the cumulative distribution function would rise linearly with distance, with a slope that estimates the rate of the DNA substitution process. Then, with the use of dates from the fossil record a calibration rate giving the amount of genetic change expected per unit of evolutionary time would be available [10].

3.2. Multiproteins of 385 mammal species

Now the distance between all the 385 sequences and the reference spectrum were calculated and species were ordered against the decreasing distance. The full list is presented in file Table1 in the Supplementary material. The upper plot in Fig.

2 presents experimental *cdf* function for all species (dots) and theoretical fit (line). All the distances between the reference spectrum and protein fall to approximately (0.3,0.85) range. It also follows from the plot that in two ranges of distance, the distribution function can be roughly approximated by linear function.

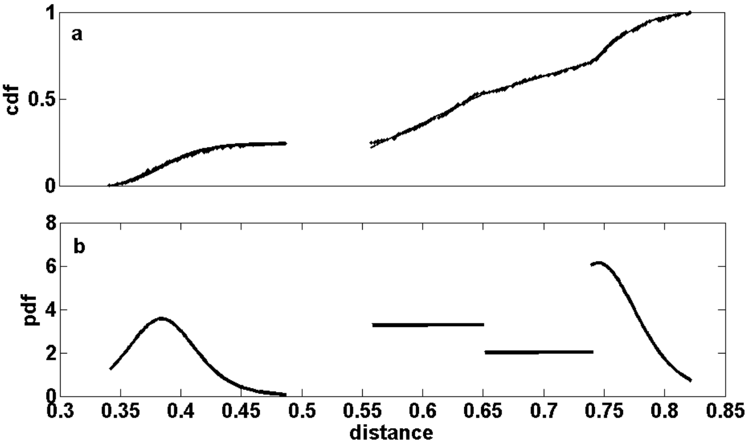

Figure 2: *ecdf* and fit (up), *pdf* of multiprotein sequences of 385 sequences (down)

In general, equation *pdf*(s) = d*cdf(s)*/ds relates the cumulative distribution function to the partial distribution function which indicates how many sequences $\Delta cdf(s)$ in in the vicinity of distance s d falls into the range Δs. Therefore, in the considered ranges the experimental partial distribution function is uniform (constant) and equals $\Delta cdf(s)/\Delta s$. In the examples above *pdf* = 3.66 within the range (0.55,0.66) and *pdf* = 2.05 within the range (0.66,0.74). It means that if the total number of extant species considered is 385 then the number of species per unit distance that are distant from the reference spectrum in the ranges mentioned is approximately 400/unit distance and 700/unit distance, respectively. In two other ranges a good fit to the experimental distribution is obtained with the logistic distributions

$$cdf(s) = b + \frac{a}{1 + e^{[(s-m)/\sigma]}}, \tag{3}$$

where the constant m is the mean value of the random variable, $1/\sigma$ defines the growth rate and a is the carrying capacity (limiting the population size).

Table 3: Parameters and ranges of applicability for logistic distributions

a	b	m	σ	range
0.271	-0.029	0.384	0.019	(0.34,0.49)
0.540	0.472	0.745	0.025	(0.74,1.00)

As follows from the plot, the overall fit is quite satisfactory. However, in the small scale the fitting accuracy becomes worse. The experimental distribution function consists of many small groups of close lying points separated by narrow ranges of distance not allowed for proteins. Some of the groups can be locally successfully fitted by logistic distributions. The experimental distribution function of distances between species and representative set of words seems to be built of logistic distributions in large and small scales. The experimental *cdf* is not linear in the whole distance range, which means that the molecular clock hypothesis [11] does not hold for multiprotein sequences of mammal species.

3.3. Evolutionary rearrangement of words in protein sequences

In the course of evolution words that are components of a protein sequence evolve as well. They change their position, are discarded and replaced by newborn words. The index of a word in the spectrum of every protein sequence of any species introduced in Section 3.2 can be analysed against the distance between the species and the representative set of words. It is sufficient to rearrange the columns of the matrix mentioned in Step 3 of the algorithm for an ancestral sequence. The data in a column are related to some species. In further analysis a convenient order of columns is achieved when the first column corresponds to the least distant species and the last one to the most distant species. If ordering has been done, it is easy to find the index of every word from the union of words in a sequence of protein of every species as the function of species distance. The statistical data presented in the second subplots need grouping sequences in bins of ten consecutive species.

The ways the words index evolves are diverse. The index of the most frequent words at low distances usually varies slowly then a kind of random walking among the neighbouring ten indices is observed. Besides, a word often disappears from more distant sequences so that the number of that particular word in bins decreases. In the figures that follow the upper plots show the index of a word in every species against the distance to the reference set. The lower one shows the number of species bearing the word in the bin against the distance of the first species in the bin. As an example the evolution of the index of two words from COX1 protein is presented.

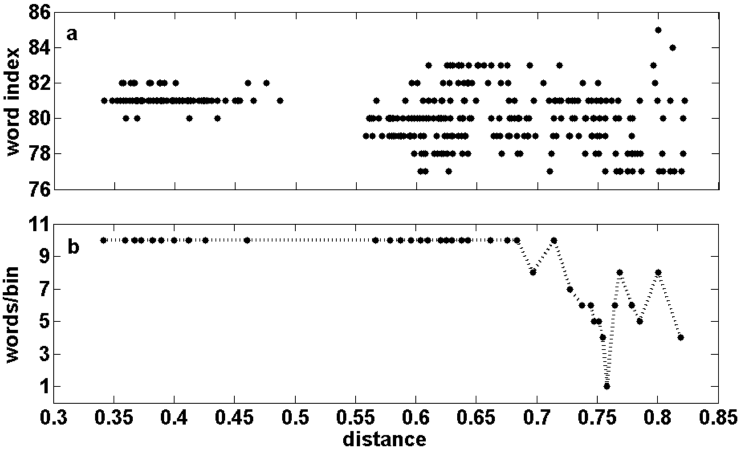

Figure 3: Index of 'HTFEEP' word (up) and its number per bin (down)

The upper subplot in Fig. 3 shows the evolution of the word 'HTFEEP' index in the spectrum of COX1 protein and the lower one the number of words in consecutive bins of ten sequences long, both against the distance to the reference sequence. It follows from the upper subplot that at the beginning the word is present in every sequence until the distance 0.69. Its index varies at the beginning between three then between seven neighbouring indices. There is no word in the distance range (0.49,0.56) as there is no sequences in that range. Starting from distance 0.69, the word becomes less frequent and if the trend were continued, it would be discarded from COX1 sequences.

At some distance the new word may also happen to come into view and then in a more distant species is discarded from the protein sequence.

The word is not present in the species of distance less than approximately 0.6. Then, it becomes more frequent and starting from sequences more distant than 0.65 its frequency declines and above distance 0.77 the word ASSM' is no more seen in COX1 sequences.

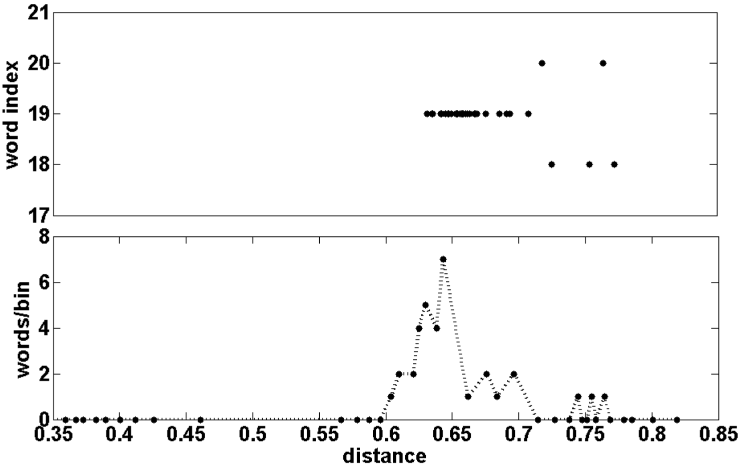

Figure 4: Index of ’ASSM’ word (up) and its number per bin (down)

4. Conclusion

Given a distance measure it is relatively easy to find all pairwise distances (or similarities) between all sequences of a DNA or a protein. However, it is much more difficult to arrange all sequences as a list ordered according to distance, because it needs some sequence as a reference point on the distance axis. In fact, adopting the distance measure defined in the present paper, it is sufficient to have a set of old words called a reference set. It was assumed that the most common words among word spectra of all the sequences are the oldest ones. The reference set of the oldest words, counting approximately as many words as the average length of the word spectrum, has been found. The list of all sequences arranged according to the decreasing distance to the reference set has been built. The list can be exploited in phylogenetic tree construction as additional data which help to root a tree. A reliable tree and good estimate of the branch length are required in order to correct the estimate of the divergence time, so is the use of long sequences. Both lists for primates and 385 multiprotein sequences look reasonable but will be discussed in details only when similar lists based on mitochondrial genes are available. The genome is a much longer sequence, but the method presented is free of any restrictions on the sequence length. The list is currently under preparation.

References

[1] Kumar, S. and Filipski, A., *Multiple sequence alignment: In pursuit of homologous DNA positions*, Genome Research, Vol. 17, 2007, pp. 127–135.

[2] Zuckerkandl, E. and Pauling, L., *Molecular restoration studies of extinct forms of life*, Acta Chemica Scandinavica, Vol. 17, 1963, pp. 9–16.

[3] Schluter, D., *Uncertainty in ancient phylogenies*, Nature, Vol. 377, 1995, pp. 108–110.

[4] Rosenberg, M., S., *Multiple sequence alignment accuracy and evolutionary distance estimation*, Bioinformatics, Vol. 6, 2005, pp. 278–288.

[5] Williams, P., D., Pollock, D., D., Blackburne, B., P., and Goldstein, R., A., *Assessing the accuracy of ancestral protein reconstruction methods*, PLoS Comput. Biol., Vol. 2, 2006, pp. 598–605.

[6] Kozarzewski, B., *A method for nucleotide sequence analysis*, Computational Methods in Science and Technology, Vol. 18, No. 2, 2012, pp. 5–10.

[7] Vinga, P. and Almeida, J., *Alignment-free sequence comparison - a review*, Bioinformatics, Vol. 19, 2003, pp. 513–523.

[8] Ke, D., G. and Tong, Q., Y., *Easily adaptable complexity measure for finite time series*, Phys. Rev., Vol. E77, 2008, pp. 513066215–23.

[9] Nei, M., Xu, P., and Glazko, G., *Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms*, Proc. Nat. Acad. Sciences, Vol. 98, 2001, pp. 2497–2502.

[10] Caswell, J., Mallick, S., Richter, D., Neubauer, J., Schirmer, C. end Gnerre, S., and Reich, D., *Analysis of Chimpanzee History Based on Genome Sequence Alignments*, PLoS Genetics, Vol. 4, No. 4, 2008.

[11] Bromham, L. and Penny, D., *The modern molecular clock*, Nature Reviews, Genetics, Vol. 4, pp. 216–224.