LINKFINDER: AN EXPERT SYSTEM THAT CONSTRUCTS PHYLOGENIC TREES

James Inglehart and Peter C. Nelson

Department of Electrical Engineering and Computer Science (M/C 154)
The University of Illinois at Chicago
Chicago, Illinois 60680 e-mail: inglehar@uicbert.eecs.uic.edu

Abstract. An expert system has been developed using CLIPS that automates the process of constructing DNA sequence-based phylogenies—trees or lineages that indicate evolutionary relationships. LinkFinder takes as input homologous DNA sequences from distinct individual organisms. It measures variations between the sequences, selects appropriate proportionality constants, and estimates (if possible) the time that has passed since each pair of organisms diverged from a common ancestor. It then designs and outputs a phylogenic map summarizing these results.

LinkFinder can find genetic relationships between different species, and between individuals of the same species, including humans. It was designed to take advantage of the vast amount of sequence data being produced by the celebrated Genome Project, and should be of great value to evolution theorists who wish to utilize this data, but who have no formal training in molecular genetics.

The mathematical basis of LinkFinder's DNA sequence analysis is remarkably simple. Evolutionary theory holds that distinct organisms carrying a common gene inherited that gene from a common ancestor. Homologous genes vary from individual to individual and species to species, and the amount of variation is now believed to be directly proportional to the time that has passed since divergence from a common ancestor. The proportionality constant must be determined experimentally; it varies considerably with the types of organisms and DNA molecules under study. Given an appropriate constant, and the variation between two DNA sequences, a simple linear equation gives the divergence time.

INTRODUCTION

Plants and animals have long been classified according to the similarities and differences in their form and structure. When the concept of evolution was first proposed, biologists naturally used these morphological features to establish phylogenies—evolutionary trees. Proposed lineages were modeled as paths (from root to leaf) through the tree, and closely related species were shown as parallel branches emanating from a common ancestor node.

Constructing a tree based on morphology has always been highly subjective, more of an art than a science. Most classifications are based on anatomy, but microscopic species, such as bacteria, are more commonly distinguished by chemical analyses. Morphology is good for classifying evolutionary relationships at certain scales, but it indicates neither the large-scale structure of evolution nor the fine details. Mice are clearly much closer cousins to humans than bacteria, but the details of how we diverged from mice are obscure, despite numerous anatomical similarities between mice and men. And bacteria seem so alien to us (despite some common features) that the details of how we diverged from them seem impossible to determine from morphology alone.

Fortunately, sophisticated new methods of genetic analysis have arisen to challenge morphology as the prime determinant of family trees. We now understand that all life processes are ultimately controlled by DNA. This self-replicating molecule is found in every living thing, and it is the key to the structure and complexity of all life on earth. Because of recent technological advances, biologists and geneticists are now able to ascertain the atomic structure of an individual organism's DNA molecules. Since the common ancestor of all life on earth is believed to have been a single, primal molecule of DNA, evolution can be viewed as simply the development of new forms of DNA through accidental mistakes in duplication.

Because all DNA is constructed exactly the same way (only the specific base sequence differs from individual to individual), DNA analysis provides an objective basis for discerning evolutionary relationships. Closely related species should have closely similar DNA sequences. Distantly related species will have far more dissimilarities. But if all life is truly related (through some primal common ancestor) then even the most disparate life forms should share some similarities. All life shares the same chemical basis, so total DNA sequence divergence (to the point of zero resemblance between distantly related organisms) should not occur (Doolittle et al. 1986).

Some phylogenic trees based on genetic analyses are radically different from traditional morphological trees, challenging our traditional views. For example, comparative studies of the Ursidæ, or bear family, and Procyonidæ, or raccoon family, indicate that the giant panda belongs in the bear family, whereas the red panda belongs in the raccoon family (O'Brien et al. 1985). Phylogenically, then, there is really no such thing as a "panda," whereas bears and raccoons really exist.

Similar results have been obtained in primate studies. Comparative analyses of the beta-globulin genes of humans and the great apes (the chimpanzee, gorilla, and orangutan) indicate that humans are more closely related to chimpanzees than chimpanzees are to the other great apes (Miyamoto et al. 1988). This contradicts the common notion (based on the morphological similarities of apes) that humans diverged from the common ancestor of the great apes. In fact, the semantic label "ape" has now lost its phylogenic connotation, since it makes more sense to lump humans and chimpanzees together than to group chimpanzees with the other great apes (Ueda et al. 1986).

The most controversial result (cf. Latorre et al. 1986) has been the "Mother Eve hypothesis" of Rebecca Cann and her colleagues. Their studies of worldwide human mitochondrial DNA indicate that all humans alive today have a common ancestor, a woman, who lived in Africa roughly 200,000 years ago (Cann et al. 1987). Prior to this discovery, anthropologists generally assumed that the the most recent common human ancestor must have lived closer to one million years ago (Stoneking et al. 1986).

The Genetic Code

The genetic information in DNA is encoded within strings of nitrogenous bases. There are four of these: adenine (A), cytosine (C), guanine (G), and thymine (T). A DNA *sequence* can be thought of as simply a long string of these four letters in some combination.

A string of three bases codes for an amino acid; there are exactly twenty of these. Proteins are simply long strings of amino acids. Thus, for any protein, there is some corresponding base sequence that acts as a template for that protein; this template is called a *gene*. Genes code for protein production, and chromosomes are simply long strings of genes. A *genome* is the entire gene set of a single organism.

During reproduction, the genome is copied, to be passed on to future generations. Usually the copy is perfect, but sometimes an incorrect base is substituted. Or a gene may be damaged by something in its environment. If the mutation is a minor one (non-fatal) it will be copied and passed on to future generations. Over time, these inheritable mutations will lead to the development of new genotypes within a species, and ultimately to new species.

Algorithmic Tree Construction

These considerations have led to the development of phylogenic tree construction algorithms, which take as input DNA sequence data from living organisms. The input organisms are then "clustered" into larger related units on the basis of their genetic similarities. The more distantly related clusters are then iteratively clustered in turn, until a complete tree is formed. Each internal node on this tree represents a hypothetical ancestor form, a "missing link," joining separate lineages. Leaf nodes represent the input organisms living today. The root node represents the common ancestor of all the leaves on the tree.

Detailed family trees of certain organisms, such as pedigreed animals, are already known. These known trees can be used to test tree construction algorithms. Given genetic data from present-day forms, a good algorithm should reconstruct the known trees. Fruit flies are good test cases, since they have been bred in the laboratory for decades. Some human genealogies extend more than 1,000 years. But no known trees go back far enough in time to link distinct species, such as humans and chimpanzees.

Because known trees are so limited, tree construction algorithms are usually tested on made-up data. An single ancestor DNA sequence is chosen at random, and descendent generations are iteratively created by random base substitutions (Tateno 1985). A good algorithm will correctly reconstruct the entire made-up tree from the DNA sequences of its final generation.

In an accurate morphologically-based tree, the vertical height of lineages is made proportional to the passage of time. Divergence times are deduced from accepted geological time scales by examining the fossil record. Ideally we would like trees constructed by algorithm to also show how long ago each pair of lineages diverged. But finding the correct time scale is difficult, and highly organism-dependent. Bacteria, for example, can mutate much faster than humans. Because of these difficulties, the vertical height of divergent lineages in trees constructed by algorithm is usually made proportional to the "genetic distance" between divergent pairs of organisms. This "distance" is the calculated (or estimated) percentage by which the genomes of the two organisms diverge. When a tree is constructed from genetic data, some attempt is usually made to convert this "distance scale" to a time scale. This conversion usually requires expert knowledge concerning the fossil record, the types of organisms under study, and the types of genetic data being used.

LinkFinder

LinkFinder automates the process of tree construction in two ways: 1) it automatically constructs a tree from genetic data, and 2) it converts (if possible) the distance scale of the initial tree to a time scale. The topology of the final tree is entirely determined by the input data and the tree construction algorithm. But to convert the tree from a distance scale to a time scale, LinkFinder requires an expert system. Our CLIPS-based system takes into account the specific nature of the input data in choosing a conversion, considering both the fossil record and known mutation rates before making a decision.

The fossil record remains the primary source of authoritative evidence on the divergence times of major lineages, and LinkFinder's knowledge base is mostly derived from the published literature on the fossil record. Its knowledge is thus limited to areas where evidence of divergence times has already been found. Since the fossil record is incomplete, there are considerable gaps in the knowledge base, which hopefully will be filled in the future. Other techniques of estimating genetic divergence rates also exist (e.g., Nei and Tajima 1983, Ferris et al. 1983, and Stoneking et al. 1986), and some of these estimates are now being added to LinkFinder's knowledge base. Our primary reason for developing LinkFinder was to take advantage of the explosion of new genetic data being produced by the Genome Project—the ongoing worldwide effort to map and sequence the entire human genome, as well as the entire genomes of several other organisms. We hope to continue developing LinkFinder as the Genome Project progresses, adding new information to its knowledge base as it becomes available. In its present form, LinkFinder is a powerful tool for tree estimation, but it will not be a true, general purpose tree constructor until more complete data is available on rates of divergence.

HOW LINKFINDER WORKS

LinkFinder constructs a phylogenic tree from input genetic data in two distinct stages:

1. The topology of the tree is determined by the unweighted pair-group (UPG) method. At this stage, branch lengths in the tree are proportional to the calculated percent difference between the clustered genotypes.

2. A CLIPS-based expert system attempts to determine an explicit time scale for the tree. It considers known mutation rates and the fossil record (if any) of the organisms in the tree before making a decision.

Estimating Tree Topology

There is as yet no known algorithmic method of tree construction which can reproduce known phylogenic trees with unfailing accuracy. Reconstructed trees are therefore phylogenic *estimates* at best.

The UPG method utilized by LinkFinder is the simplest well-known tree construction algorithm (first proposed by Sokal and Michener in 1958), but it has stood well the test of time. Numerous, far more complicated tree construction algorithms have since been proposed (e.g., Fitch and Margoliash 1967, Farris 1972, Moore et al. 1973, and Tateno et al. 1982) but the overall performance of the UPG method still compares favorably with these (Li 1981).

Genetic data is input to LinkFinder as a two-dimensional array, with each row containing sequence data from a single operational taxonomic unit (OTU), which can be a gene, an individual, a population, a species, or a taxa of higher rank (Moore et al. 1973). Each row also contains a unique label identifying the OTU.

Another input file classifies each OTU according to kingdom, phylum, class, etc. This taxonomic information will be needed by LinkFinder's expert system.

Sequence data is coded either as a string of amino acids, or (more typically) as a string of bases. The four possible bases can be coded with the four letters A, C, G, and T, and the twenty possible amino acids can be coded using any convenient choice of twenty distinct characters. OTU labels are distinct name strings chosen to identify the OTUs under study, but in the examples below single letters will be used as labels for hypothetical OTUs.

The UPG method utilized by LinkFinder makes the following assumptions about its OTU sequence data:

1. Genetic sequences have been chosen to be *homologous* between OTUs. I.e., if the sequence is a gene coding for some protein which varies from OTU to OTU, each OTU must have inherited that gene from some common ancestor, so that each individual contemporary form represents divergence from the same ancestral form (Tateno 1985). For example, the gene for hemoglobin, found in some form in all mammals, is homologous in mammals.

2. Contemporary homologous OTU sequences are assumed to have diverged from the ancestral form because of *random* base (amino acid) substitutions in succeeding generations. (Note that changing a single base in a group of three can also change the amino acid that the group codes for.) These substitutions are assumed to be random both in the choice of base (amino acid) and with respect to position in the sequence. This assumption is well supported by our current understanding of genetic mutations (Tateno 1985).

3. The number of base substitutions in all lineages is assumed to be linear over time, i.e., all lineages are assumed to evolve at the same constant rate. This simplifying assumption is realistic in many cases, but it is not strictly true. The actual number of base substitutions in an evolutionary lineage over time is believed by many geneticists to follow a Poisson distribution whose mean is the expected number of base substitutions in that OTU over time. This implies that the actual numbers of base

substitutions in two lineages can differ considerably due to stochastic error (Tateno 1985). When the actual rates for different lineages are very different, tree estimation by the UPG method is sometimes in error. This has motivated the development of more complex tree construction algorithms. However, the UPG method is much simpler to implement, and it works well in the majority of cases (Li 1981).

Assumptions (1) and (2) combined imply that all input genetic sequences must be exactly the same length, with each containing the same number of bases (amino acids). All three assumptions must hold for the constructed tree to be considered valid.

Given an input array of homologous sequence data for n distinct OTUs, LinkFinder starts by computing the genetic distance between every pair of distinct OTUs, and loading these values into an $n \times n$ distance matrix. Genetic distance is simply the percent difference between two distinct sequences, which is calculated by direct comparison. If the pair of bases (amino acids) in the ith position of two sequences don't match, a counter is bumped, and the final count for that pair is divided by the total number of bases (amino acids) in a sequence.

The tree topology is generated from the distance matrix by the following iterative algorithm (from Li 1981):

1. Choose the smallest non-zero distance in the distance matrix. These two closest OTUs will now be clustered together into a single OTU. E.g., if d_{AB} is the shortest distance (as in Table 1), then A and B are the closest OTUs, and the new OTU will be labeled (AB).

2. Draw vertical lines from the chosen nodes A and B to their presumed common ancestor node (as in Figure 1). Make the lines proportional in length to $d_{\mathrm{AB}}/2$.

3. Construct a new, smaller distance matrix from the old one by taking the distance between AB and any other OTU, say C, to be the arithmetic average of d_{AC} and d_{BC}—i.e., $d_{(\mathrm{AB})\mathrm{C}} = (d_{\mathrm{AC}} + d_{\mathrm{BC}})/2$ (as in Table 2).

4. Continue the process (1, 2, and 3) until all the initial OTUs are clustered into a single binary tree. The root node of this tree will represent the common ancestor of all the initial OTUs, and the height of the tree will be proportional to the genetic distance (percent divergence) between the hypothetical ancestor and all of its descendent leaf nodes.

In the tables and figures, seven contemporary OTUs are labeled A, B, C, D, E, F, and G, and their initial distance matrix (Table 1) indicates that (AB) should be the first cluster, since the sequences of A and B differ by the smallest amount (3%). The common ancestor of A and B should thus differ 1.5% from each of its descendants, so the parent node linking A and B is placed at a height of 1.5 percentage units (Figure 1).

The second distance matrix (Table 2) gives us (EF) as a cluster with a height of 2.5 units (Figure 2). The third (Table 3) joins (AB) with D to produce the cluster ((AB)D) with an overall height of 3.75 units (Figure 3). Next we get ((EF)G), also 3.75 units high (Figure 4). Then (((AB)D)C), 4.38 units high (Figure 5). We now have only two clusters left, so our hypothetical common ancestor must lie between them. If we call this root node

Table 1. Initial distance matrix.

	A	B	C	D	E	F	G
A	0	**3**	8	7	10	9	13
B		0	9	8	11	10	14
C			0	9	12	11	15
D				0	9	8	12
E					0	5	9
F						0	6
G							0

Table 2. After one iteration.

	(AB)	C	D	E	F	G
(AB)	0	8.5	7.5	10.5	9.5	13.5
C		0	9	12	11	15
D			0	9	8	12
E				0	**5**	9
F					0	6
G						0

Table 3. After two iterations.

	(AB)	C	D	(EF)	G
(AB)	0	8.5	**7.5**	10	13.5
C		0	9	11.5	15
D			0	8.5	12
(EF)				0	7.5
G					0

Table 4. After three iterations.

	((AB)D)	C	(EF)	G
((AB)D)	0	8.75	9.25	12.75
C		0	11.5	15
(EF)			0	**7.5**
G				0

Table 5. After four iterations.

	((AB)D)	C	((EF)G)
((AB)D)	0	**8.75**	11
C		0	13.25
((EF)G)			0

Table 6. After final iteration.

	(((AB)D)C)	((EF)G)
(((AB)D)C)	0	**12.125**
((EF)G)		0

Figure 1. Initial cluster: (AB).

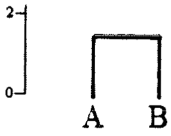


Figure 2. Resultant cluster: (EF).

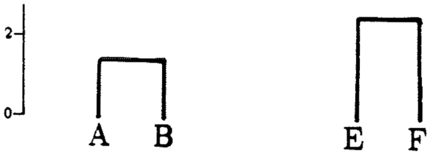


Figure 3. Resultant cluster: ((AB)D).

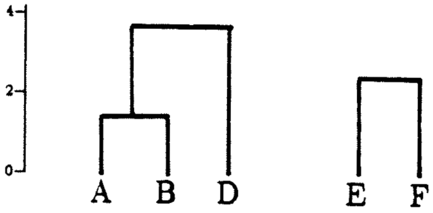


Figure 4. Resultant cluster: ((EF)G).

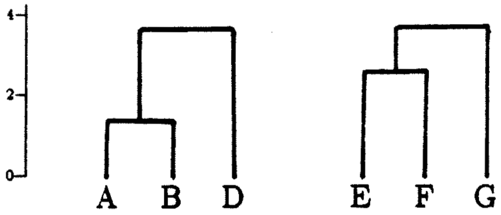


Figure 5. Resultant cluster: (((AB)D)C).

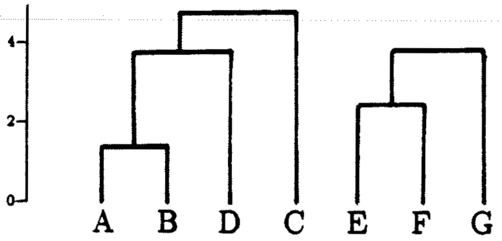
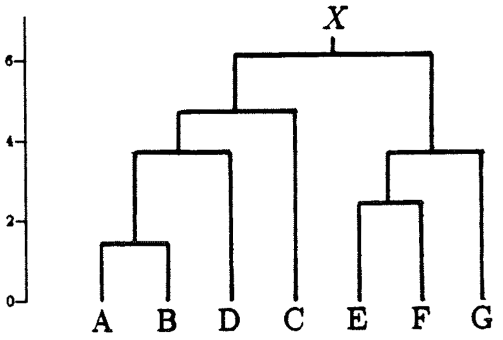


Figure 6. Resultant tree:



X, then our final tree (Figure 6) can be written in line-form as (((AB)D)C)X((EF)G). Its height of 6.06 units gives the average divergence of the seven present-day OTUs from their presumed common ancestor: about 6%.

Estimating Divergence Times

LinkFinder's method of phylogenic tree estimation is a straightforward implementation of the UPG method. It can produce a tree for any reasonable input set of sequence data. However, the vertical axis of a true phylogenic tree represents time, not percent divergence. To produce a true phylogenic tree, LinkFinder must convert the vertical axis to units of time.

Unfortunately, there is no straightforward way to do this. Divergence rates vary greatly for different organisms and different types of DNA. To make a reasonable conversion, LinkFinder must proceed on a case-by-case basis. It must consider the available experimental evidence on divergence rates for the specific data under study. It needs expert advice.

It was to solve this problem that LinkFinder was equipped with a CLIPS-based expert system. LinkFinder's knowledge base contains current data on divergence rates and times for many basic types of organisms. These data have been culled from selected books and papers on evolution, the fossil record, and genetics. Given the taxa under consideration, LinkFinder can usually make educated guesses about divergence times.

For example, it has been estimated that prokaryotic organisms (like bacteria) diverged from eukaryotic organisms (like plants and animals) roughly 1.8 billion years ago. Plants probably diverged from animals about 1 billion years ago, and animals diverged into vertebrates and invertebrates about 500 million years ago. These facts (from Doolittle et al. 1986) have been entered into LinkFinder's knowledge base, and LinkFinder can use them to obtain conversion factors. Given homologous gene sequences from two OTUs, one human, one bacterial, LinkFinder will assume that their linking ancestor existed 1.8 billion years ago. Based on this knowledge, LinkFinder will postulate a conversion factor for a divergence-scaled tree containing these two OTUs. E.g., if the pair of homologous genes (one human, one bacterial) coded for the metabolic enzyme ***triosephosphate isomerase*** (found in both bacteria and humans), LinkFinder would find a 54% divergence between the pair (Doolittle et al. 1986). This suggests a conversion factor of 33 million years per percent divergence for a tree containing these two OTUs, which is exactly what LinkFinder would propose. If the same tree contained other OTU pairs which were also present in LinkFinder's knowledge base, then other possible conversion factors would also be reported, along with the specific OTUs upon which each conversion was based. If there is good agreement between the various calculated factors, this is strong evidence in favor of an overall time-scale conversion. If the various factors do not agree, then various individual clusters within the tree should probably be assigned their own separate time scales based on the recommended conversions. In its present form, LinkFinder only recommends possible conversions. It leaves the final decision on how to scale the overall tree to the user.

LinkFinder's knowledge base is organized according to the usual taxonomic distinctions. By examining the classification file for each OTU, it can quickly position each OTU within the general biological categories in its knowledge base: prokaryote-eukaryote, plant-animal, vertebrate-invertebrate, fish-amphibian-reptile-mammal, etc. Once it has categorized each OTU, it searches for known divergence times for all OTU pairs. Each divergence time found is reported as a possible time-scale conversion factor for the percent divergence-scaled tree.

LinkFinder's knowledge base is intended to be augmented as new information becomes available. There are tens of millions of different species upon the earth, and divergence rates in general are not known for a given pair of OTUs. Even if they were, it would take considerable time to add so much information to LinkFinder's knowledge base. For now, we have concentrated on entering divergence times of the most basic taxonomic units—the kingdoms, phyla, and classes. In certain taxonomic areas (notably primates/humans) more detailed information has been entered on individual species. Our main purpose in creating LinkFinder has been to develop a prototype of the automated expert phylogenic tree constructor of the future. The great volumes of sequence data being generated by the Genome Project will be valueless without the proper analytic software tools, and detailed phylogenic analyses of these data will be needed before we can determine with any certainty the actual divergence paths taken by the myriad forms of life on earth.

REFERENCES

Cann, R.L., Stoneking, M., and Wilson, A.C. (1987). Mitochondrial DNA and human evolution. *Nature* **325**(1 Jan), 31–36.

Doolittle, R.F., Feng, D.F., Johnson, M.S., and McClure, M.A. (1986). Relationships of human protein sequences to those of other organisms. *Molecular Biology of Homo Sapiens: Cold Spring Harbor Symposium on Quantitative Biology* **51**(part 1), 447–455.

Farris, J.S. (1972). Estimating phylogenetic trees from distance matrices. *Am. Nat.* **106**, 645–668

Ferris, S.D., Sage, R.D., Prager, E.M., Ritte, U., and Wilson, A.C. (1983). Mitochondrial DNA evolution in mice. *Genetics* **105**, 681–721.

Fitch, W.M. and Margoliash, E. (1967). Construction of phylogenic trees. *Science* **155**, 279–284.

Latorre, A., Moya, A., and Ayala, F.J. (1986). Evolution of mitochondrial DNA in *Drosophilia subobscura. Proc. Natl. Acad. Sci. U.S.* **83**, 8649–8653.

Li, W.-H. (1981). Simple method for constructing phylogenic trees from distance matrices. *Proc. Natl. Acad. Sci. U.S.* **78**, 1085–1089.

Miyamoto, M.M., Koop, B.F., Slightom, J.L., Goodman, M., and Tennant, M.R. (1988). Molecular systematics of higher primates: Genealogical relations and classification. *Proc. Natl. Acad. Sci. U.S.* **85**, 7627–7631.

Moore, G.W., Goodman, M., and Barnabas, J. (1973). An iterative approach from the standpoint of the additive hypothesis to the dendogram problem posed by molecular data sets. *J. Theor. Biol.* **38**, 423–457.

Nei, M. and Tajima, F. (1983). Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. *Genetics* **105**, 207–217.

O'Brien, S.J., Nash, W.G., Wildt, D.E., Bush, M.E., and Benveniste, R.E. (1985). A molecular solution to the riddle of the giant panda's phylogeny. *Nature* **317**(12 Sep), 140–144.

Sokal, R.R. and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **38**, 1409–1438.

Stoneking, M., Bhatia, K., and Wilson, A.C. (1986). Rate of sequence divergence estimated from restriction maps of mitochondrial DNAs from Papua New Guinea. *Molecular Biology of Homo Sapiens: Cold Spring Harbor Symposium on Quantitative Biology* **51**(part 1), 433–439.

Tateno, Y., Nei, M., and Tajima, F. (1982). Accuracy of estimated phylogenetic trees from molecular data. *J. Mol. Evol.* **18**, 387–404.

Tateno, Y. (1985). Theoretical aspects of molecular tree estimation. *Population Genetics and Molecular Evolution* (Ohta, T. and Aoki, K., ed.), 293–312, Japan Sci. Soc. Press, Tokyo/Springer-Verlag, Berlin.

Ueda, S., Watanabe, Y., Hayashida, H., Miyata, T., Matsuda, F. and Honjo, T. (1986). Hominoid evolution based on the structures of immunoglobulin epsilon and alpha genes. *Molecular Biology of Homo Sapiens: Cold Spring Harbor Symposium on Quantitative Biology* **51**(part 1), 429–432.