# An Evolutionary Trend Discovery Algorithm Based on Cubic Spline Interpolation

ZHIQIANG LI and PETER Z. REVESZ
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
USA
zli@cse.unl.edu, revesz@cse.unl.edu, http://cse.unl.edu/~revesz/

*Abstract:* - The speed of evolution, measured by the number of mutations over a fixed number of years, varies greatly in various branches of the evolutionary tree. This paper proposes an evolutionary trend discovery algorithm that reveals the distinguishing characteristics of any branch of the evolutionary tree. The evolutionary trend discovery algorithm is designed to work with either fossil-based data or automatically generated data about the age of the internal nodes in the evolutionary tree. The evolutionary trend discovery algorithm estimates the missing age data using cubic spline interpolation. The evolutionary trend discovery algorithm identifies, for example, that human evolution seems to be currently speeding up while the evolution of chickens is slowing down.

*Key-Words:* - Common mutations similarity matrix (CMSM), cubic splice interpolation, evolutionary tree, neighbor joining, phylogenetics, UPGMA

## 1 Introduction

Whereas evolutionary biologists in the past could be satisfied with piecing together an evolutionary tree of related species, now they can ask deeper questions, such as when was the evolutionary change most rapid or slow in any branch of the evolutionary tree [11]. In the past, evolutionary biologists could make only relatively subjective statements about the speed of evolution. However, the DNA data available today in many genome databases [5], [14] for an increasing number of living species and even from ancient DNA from fossils enables modern evolutionary biologists to make more precise and measurable statements about the speed of evolution. This is because the speed of biological evolution from an ancestor species to a descendant species can be measured in the number of genetic mutations.

Data analytics has the potential to make many fascinating discoveries about the evolutionary trends and their causes. Such a data analytics would bring together evolutionary biologists and data scientists. Towards this goal, we describe in this paper a novel evolutionary trend discovery (ETD) algorithm. The ETD algorithm estimates the different trends of evolution for various branches of the evolutionary tree. Our work already brings together cubic spline interpolation from numerical analysis, phylogenetic tree algorithms, and evolutionary biology.

This paper is organized as follows. Section 2 presents some related work. Section 3 describes a method to find evolutionary trends using the fossil record-based age estimates of ancestral species. Section 4 describes the experimental results from Section 3. Section 5 proposes a way to automatically estimate the age of internal nodes in the evolutionary tree. With this alternative age estimation, the evolutionary trend discovery algorithm is shown to be yielding a result similar to the result in Section 3. Finally Section 6 gives some conclusions and directions for future work

## 2 Related Work

Given the genes of a set of related species, a hypothetical evolutionary tree, also called a phylogenetic tree, can be constructed using several different algorithms.

The UPGMA [12] and the Neighbor Joining (NJ) [10] algorithms are the most commonly used phylogenetic tree algorithms. The maximum likelihood method is also well known, although it less frequently used that UPGMA and Neighbor Joining because it requires more computational time. The Common Mutations Similarity Matrix (CMSM) algorithm of Revesz [6], the Incremental Phylogenetics by Repeated Insertions (IPRI) algorithm of Revesz and Li [4], and Wang's method [15] are some recently proposed

phylogenetic tree algorithms. Many phylogenetic tree algorithms are reviewed in the textbooks [1]-[3].

The phylogenetic tree algorithms generate for a given set of genes of a set of related species a common ancestor/root node as well as internal nodes that correspond to the ancestral forms of various branches in the evolutionary tree. Moreover, the algorithms also associate with the root and each internal node an estimated gene (DNA sequence) based on all the descendant species.

There is a strong relationship between the biological classification of species and their evolutionary tree. In fact, the biological classification is often updated to match closer the constantly improving understanding of biological evolution.

For example, Table 1 lists the biological classification of fourteen vertebrate species. A phylogenetic tree generated for these fourteen species using the Common Mutations Similarity Matrix algorithm of Revesz [6] is shown in Fig. 1.

Table 1. The biological classification of fourteen vertebrate species

| Species | Phylum | Class | Order | Family |
|---|---|---|---|---|
| Human (*Homo sapiens*) | Chordata | Mammalia | Primates | Hominidae |
| Cattle (*Bos taurus*) | Chordata | Mammalia | Cetartiodactyla | Bovidae |
| Dog (*Canis familiaris*) | Chordata | Mammalia | Carnivora | Canidae |
| Brown rat (*Rattus norvegicus*) | Chordata | Mammalia | Rodentia | Muridae |
| Mouse (*Mus musculus*) | Chordata | Mammalia | Rodentia | Muridae |
| Hamster (*Mesocricetus auratus*) | Chordata | Mammalia | Rodentia | Cricetidae |
| Chicken (*Gallus gallus*) | Chordata | Aves | Galliformes | Phasianidae |
| Japanese quail (*Coturfnix japonica*) | Chordata | Aves | Galliformes | Phasianidae |
| African clawed frog (*Xenopus laevis*) | Chordata | Amphibia | Anura | Pipidae |
| Japanese puffer fish (*Takifugu rubripes*) | Chordata | Actinopterygii | Tetraodontiformes | Tetraodontidae |
| Estuary cod fish (*Epinephelus coioides*) | Chordata | Actinopterygii | Perciformes | Serranidae |
| Ricefish (*Oryzias melastigma*) | Chordata | Actinopterygii | Beloniformes | Adrianichthyidae |
| Japanese ricefish (*Oryzias latipes*) | Chordata | Actinopterygii | Beloniformes | Adrianichthyidae |
| Zebrafish (*Danio rerio*) | Chordata | Actinopterygii | Cypriniformes | Cyprinidae |

The biological classification in Table 1 and the phylogenetic tree in Fig. 1 correspond well with each other. In particular, the root, which is node 27, corresponds to the ancestors of vertebrates, the Chordata phylum. Node 24 represents the ancestor of fish, while node 26 represents the ancestor of every other vertebrate. Node 25 represents the ancestor of mammals, and node 21 represents the ancestor of rodents, etc. Biologists have used the extensive fossil record of vertebrates to estimate when each ancestor form existed. For example, the ancestor of all vertebrates is estimated to have lived about 525 million years ago. Some of the other

Table 2. Fossil-based age estimates of ancestral species.

| Node Number | Classification | Million Years Ago |
|---|---|---|
| 15 | Beloniformes | N/A |
| 16 | Galliformes | 85 |
| 18 | Muridae | N/A |
| 21 | Rodentia | 66 |
| 22 | Amphibia | 370 |
| 23 | Primates | 56 |
| 24 | Actinopterygii | 420 |
| 25 | Mammalia | 225 |
| 26 | Non-Actinopterygii | 420 |
| 27 | Chordata | 525 |

known estimates of evolutionary biologists are listed in Table 2.

Our data analytics method also uses the cubic spline interpolation method from numerical analysis. A review and recent extension of the cubic spline method can be found in [7].

## 3 The Evolutionary Trend Discovery Algorithm

In this section we describe our *Evolutionary Trend Discovery* (ETD) algorithm. The pseudocode of our ETD algorithm is shown below. The ETD algorithm takes as input the following:

1. An evolutionary tree *E*.
2. A function *T* from internal nodes of *E* to millions of years ago, where for any internal node *N*, the value of *T(N)* is the estimated evolutionary time from the root of *E* to *N*. The root *R* is always assumed to be at time 0.
3. A function *A* from nodes of *E* to amino acid sequences or DNA sequences.
4. A specific leaf node *L*.

The output of the ETD algorithm is the discovered evolutionary trend function *D*. In our description, the function *D* is a cubic spline interpolation function based on the combination of the genetic and temporal data that is associated with the path from the root to a leaf *L*. However, in theory, the trend function could be generated by several other numerical interpolation methods. Therefore the cubic spline interpolation is used here as an example of this general idea. Cubic spline interpolation gives an interpolating polynomial that is smoother than some other interpolating polynomials such as Lagrange polynomial and Newton polynomial.

**ALGORITHM ETD**(E, T, A, L, D)
1 Find the tree E2 that is the same as E except that pointers from the parents to the children are reversed.
2 Create arrays D1 and D2, and initialize i = path_length(L);
3 current_node = L;
4 **while** (current_node ≠ R)
5 D1[i] = T[current_node];
6 mutation_number = Hamming(A(current_node), A(R));
7 D2[i] = mutation_number;
8 current_node = current_node.next(in E2);
9 i = i-1;
10 **end**
11 D1[i] = T(R);
12 D2[i] = 0;
13 D = Cubic_Spline(D1,D2);
14 Return D;

In the ETD algorithm we assume that we have available as a subroutine Hamming, which computes the Hamming Distance between two strings, and Cubic_Spline, which finds the cubic spine interpolation function with time D1 and corresponding values D2. The ETD algorithm allows us to investigate the evolutionary trend of a given species of interest using the changes in the number of mutations from the root to the leaf node corresponding to that species.

**Example.** Suppose that the ETD algorithm is called with the parameters where the tree is in Fig.1, the function *T* is in Table 2, the function *A* is the amino acids that are returned for each internal node by the CMSM algorithm and for each leaf node. In Fig. 1 the *TERT amino acid* (which is discussed in detail in Section 4), and the leaf node is 1. As can be seen in Fig. 1, here L = node 1 and R = node 27.

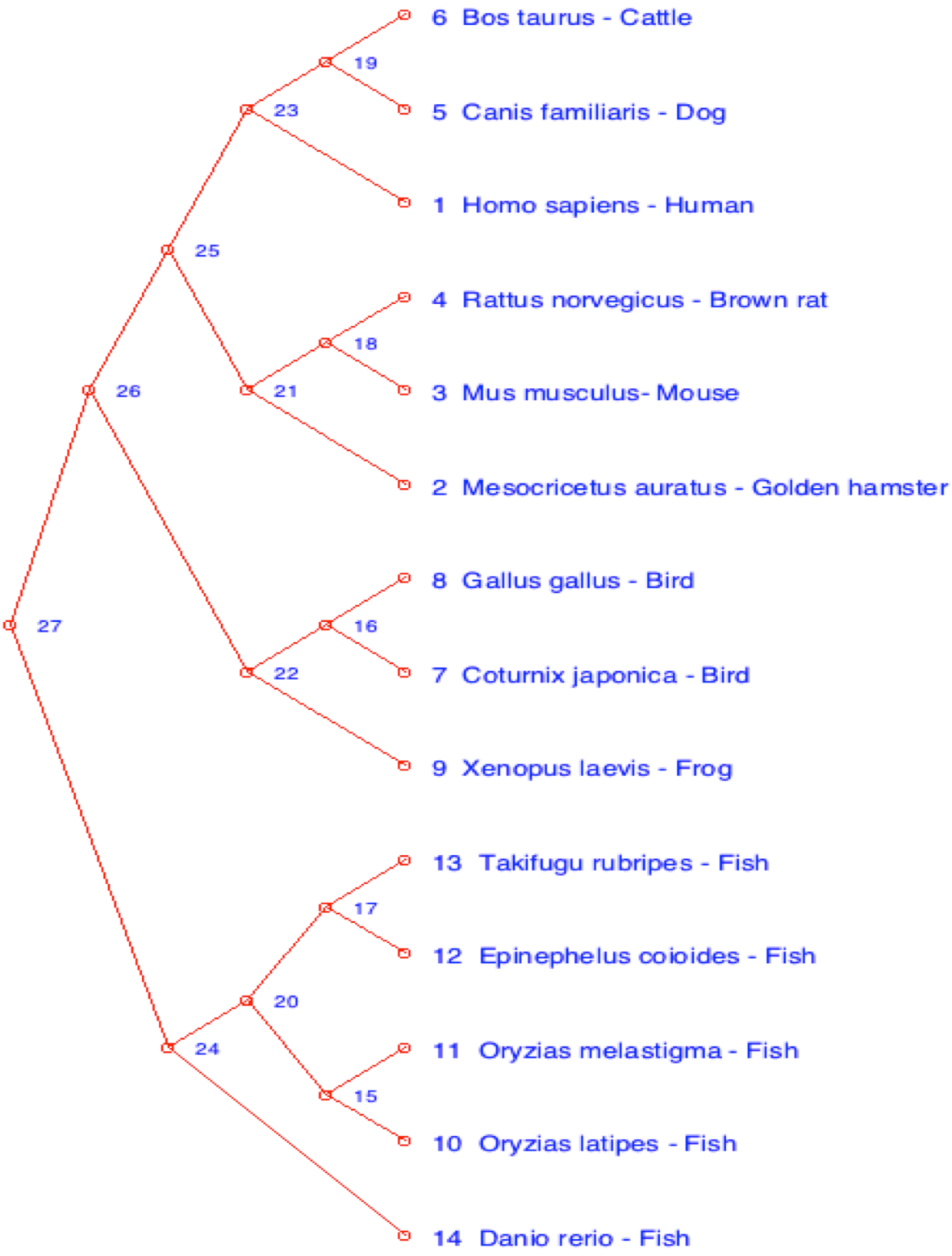


Fig. 1. The CMSM phylogenetic tree based on vertebrate telomerase protein data.

The path from L, which corresponds to humans, to R, which is the ancestral vertebrate, is the following: 1 → 23 → 25 → 26 → 27. Hence the ETD algorithm will work as follows:

D1[4] = T(node 1)
D2[4] = Hamming(A(node 1), A(R)) = 83

D1[3] = T(node 23)
D2[3] = Hamming(A(node 23), A(R)) = 28

D1[2] = T(node 25)
D2[2] = Hamming(A(node 25), A(R)) = 11

D1[1] = T(node 26)
D2[1] = Hamming(A(node 26), A(R)) = 4

D1[0] = T(node 27)
D2[0] = 0

Build a cubic spline that satisfies D2 = D (D1).

## 4 Experimental Results

As an example, we build an evolutionary tree based on the telomerase (TERT) protein family using the CMSM algorithm. Telomerase help protect eukaryote chromosomes during duplication and is generally present protein in eukaryotes. From the website http://telomerase.asu.edu we obtained 14 vertebrate telomerase proteins as our input data. After alignment, the length of each amino acid sequence was 1353. Fig. 1 shows the evolution tree from CMSM.

We evaluate our evolutionary trend discovery algorithm using as test TERT data related to human and chicken evolution. Fig. 2 shows the cubic spline interpolation results for both humans and chickens based on fossil records.

Each unit on the x axis in Fig. 2 is 1 million years. Both the human and the chicken evolutionary trend functions indicate that the overall number of mutations is increasing with time but at different rates. There are some small periods that can be considered errors in the interpolation because the number of mutations should always increase. These blips of errors non-withstanding, the overall trends seem quite reasonable.

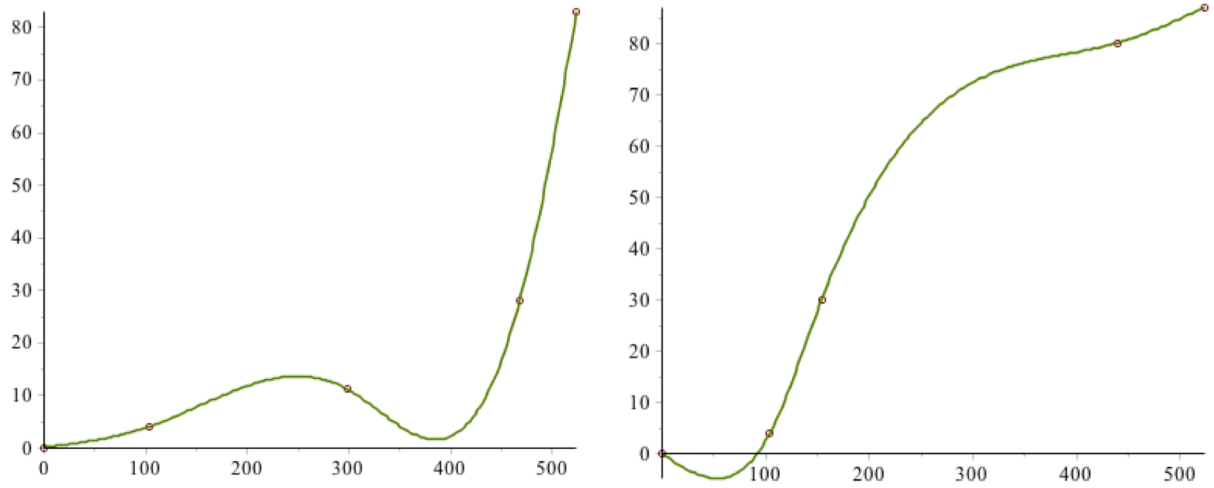


Fig. 2. Cubic spline interpolation of the number of evolutionary mutations for human (left) and chicken (right). The x-axis is already known millions of years since the common ancestor of all vertebrates, and the y-axis is the number of mutations in the TERT proteins.

In order to check better the evolutionary trends, we also draw the curves of the first derivatives for the evolutionary trend functions as shown in Fig. 3. The red curve stands for human evolution, and the purple curve represents chicken evolution from an ancestral vertebrate that lived around 500 million years ago. Fig. 3 suggests that the evolution of humans involved a speeding up of the rate of evolutionary mutations.

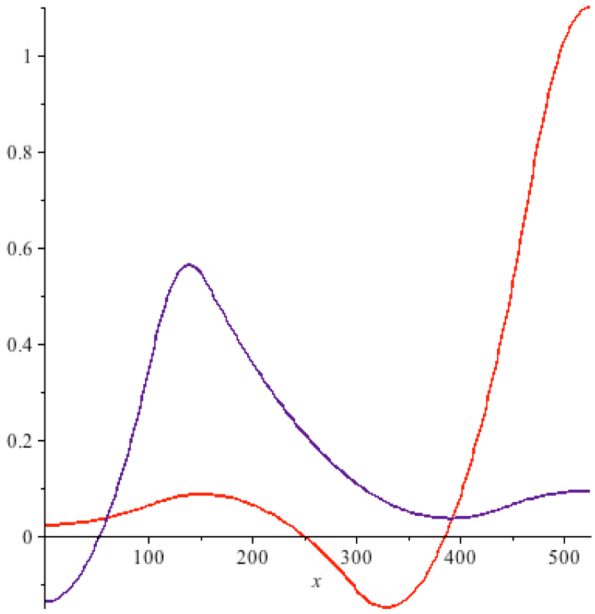


Fig. 3. The above figure shows the first derivatives of the functions in Fig. 2 based on fossil records. The red curve shows the evolutionary trend function associated with humans and the purple curves shows the evolutionary trend function associated with chickens.

In contrast, the rate of evolutionary mutations slowed down for chickens around 200 million years ago. This result agrees with our intuition with biological evolution as birds have evolved from dinosaurs millions of years ago, and mammals have evolved more recently. There seems to be a more rapid evolution at the beginning of the appearance of radically new forms of species and adaptations to new modes of living, such as flying for birds, and then a decline in the rate of mutations and adaptations after a period of establishment of the new form. It would be interesting to check whether this pattern also appears when considering other genome or protein families and other examples of vertebrates from the mammalian and bird phyla.

We implemented our evolutionary tree algorithms and generated evolutionary trees using Java and MATLAB. In addition, we implemented in Maple the cubic spline interplation function and its visualization that is shown in various figures in our paper. These programs are freely available for any researchers who request them.

## 5 Estimating the Branching Times

Unfortunately, the evolutionary trend discovery algorithm described in Section 3 cannot be applied when there is no available time estimate for each branching that occurs in the evolutionary tree. The time period of some internal evolutionary tree nodes are not always possible to estimate based on fossil data. In the special case of the vertibrates, some fossil record can be found for each internal node. However, if you consider bateria, for example, then we do not have available fossil records. Therefore, the internal nodes of bateria evolution are usually inferred from genetic similarity of extant species. A well-known example of inferring age estimates from genetic data is in a case of primates[13], [16]. Even in this case, there are still several types of errors which result in an estimation errors.

In this section, we propose a new method for estimating the time of each internal node. Before describing our method, we introduce some notation.

Let L(N) represent the length of the longest path from node N to any leaf that is a descendant of N.

$$\mathrm{L(N)} = \mathit{length\ of}\ \text{longest path from N to any leaf}$$

Let Age(L) be the age of any leaf node L. We assume that any leaf node represents a current species. Hence we have:

$$\mathrm{Age(L)} = 0$$

Let Age(R) be the age of the root node. This is the only constant that we expect to know from the beginning.

$$\text{Age(R)} = \text{age of R in millions of years ago}$$

Let Age(I) be the age of any internal node I. For any internal node I we define Age(I) as follows:

$$\text{Age(I)} = \text{Age(R)} \times \frac{\text{L(I)}}{\text{L(R)}} \qquad (1)$$

T(I) is a variable that is the elapsed time from the root to the beginning of internal node I. That is,

$$\text{T(I)} = \text{Age(R)} - \text{Age(I)} \qquad (2)$$

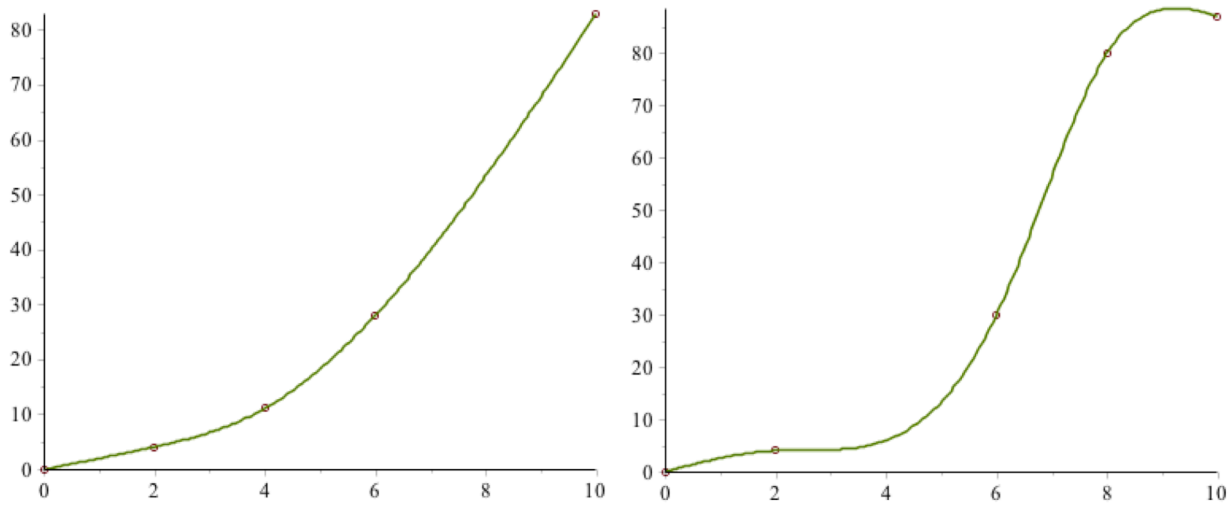


Fig. 4 Cubic spline interpolation of the number of evolutionary mutations for human (left) and chicken (right). The x-axis is automatically estimated millions of years since the common ancestor of all vertebrates, and the y-axis is the number of mutations in the TERT proteins.

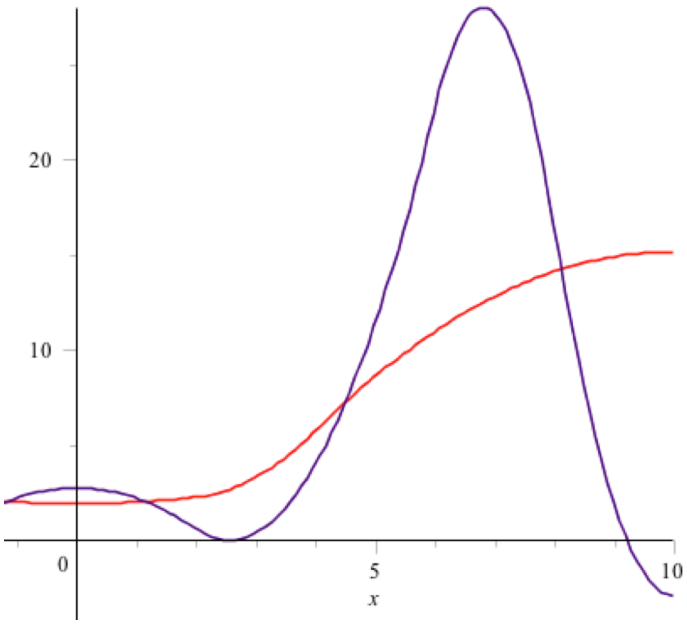


Fig. 5 The above figure shows the first derivatives of the functions in Fig. 4. based on our time estimation. The red curve shows the evolutionary trend function associated with humans and the purple curves shows the evolutionary trend function associated with chickens.

Fig. 4 shows the cubic spline interpolation results for both humans and chickens based on our time estimation method. Fig. 5 shows the first derivative of the evolutionary function from Fig. 4. The unit of x-axis is one million years.

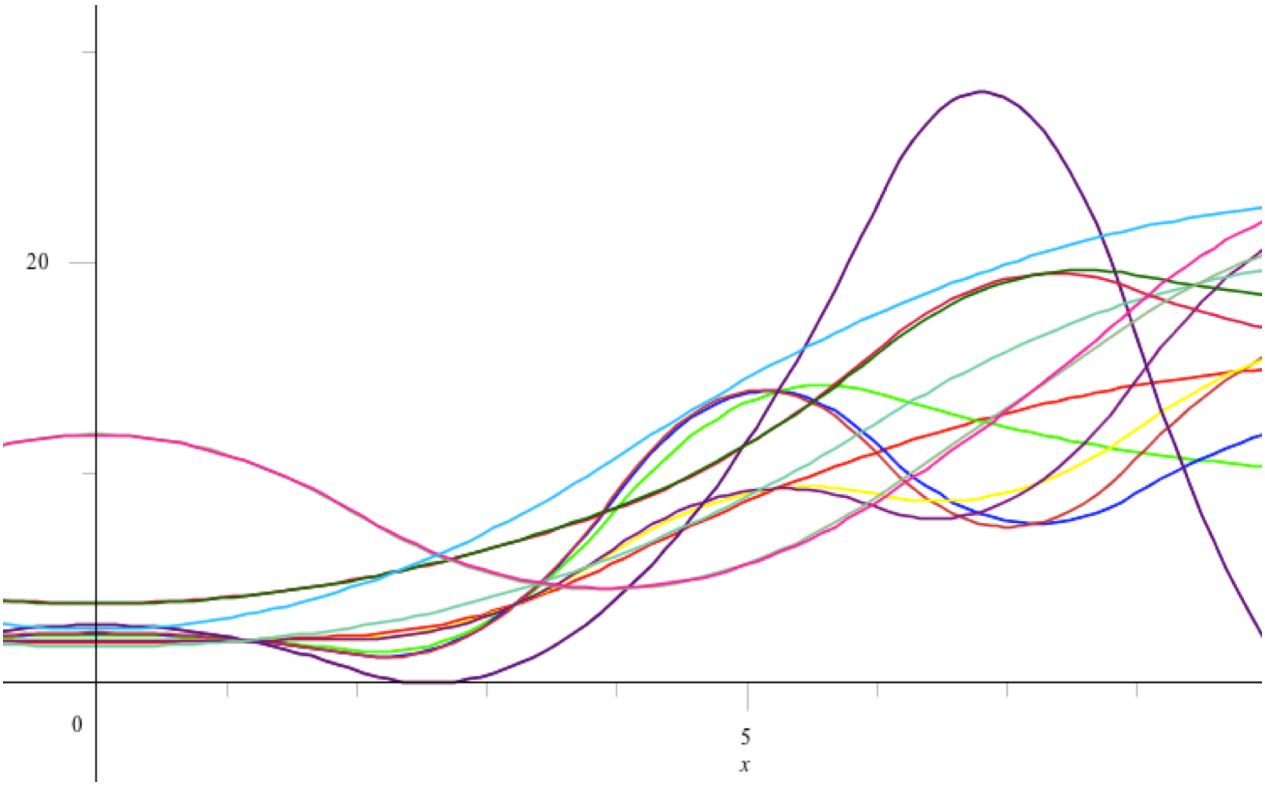


Fig. 6. The first derivative of the evolutionary function for each branch in Fig. 1. The colors refer to the following: human (red), golden hamster (green), mouse (blue), brown rat (orange), dog (yellow), cattle (purple), Japanese quail (tomato), chicken (indigo), frog (medium aquamarine), Japanese ricefish (crimson), ricefish (dark green), estuary cod fish (dark sea green), Japanese puffer fish (deep pink) and zebrafish (deep sky blue).

Now we can use T(I) as the estimated time. Based on T(I), we can again use a cubic spline interpolation to build a function f:T→M to represent the relationship between time and the number of accumulated mutations.

Comparing Fig. 2 with Fig. 4, and Fig. 3 with Fig. 5, we find that the evolutionary trend discovery algorithm gives very similar results with the fossil-based age estimates and the automatically generated age estimates. In particular, in both Fig. 3 and Fig. 5, the derivative of the mutational changes is larger for humans than for chickens toward the right end, i.e., the most recent evolutionary times, and it is larger for chickens than for the predecessors of humans in the middle of the graph, i.e., when birds were evolutionarily introduced and undergone wide adoptive radiation while spreading to the air and reaching various continents and habitat areas in the world. Meanwhile the vertebrate ancestors of humans and other mammals were more restricted in diversity according to the fossil record.

This similarity between the two results gives some confidence in the automatic age estimation method. In terms of the fossil-based method, looking up the years of each node is required, if there are no existing records to check (such as bacteria), then this method fails. However, the automatic time estimation method can overcome this issue. In fact, Fig. 6 shows the first derivatives of the evolutionary mutation functions for each branch of Fig. 1. The various curves show a great variety. The rate of mutation varies from near zero to almost thirty, and the peak mutation rate occurs at various times, although there is a general tendency of accelerated mutation in the recent times.

# 6 Conclusions and Future Work

We plan to apply the ETD algorithm to other protein and genome families for both eukaryotes and bacteria. In the ETD algorithm, we also plan to use other estimated time function *T*. Some possibilities include the estimates obtained by the UPGMA algorithm that returns not only an evolutionary tree but also a time estimate for each internal node of the tree. Many other phylogenetic tree algorithms also a time function that may be useful. It remains to be seen which of these estimates is the best and what is the degree of consistency in the results when using all of these different estimates of *T*. The estimating of the time function *T* by some method is especially important in the case of species that do not have available as extensive fossil records as for the vertebrates.

In addition, in the future more complex data analytics would need to correlate the overall evolutionary trends with significant known events in the history of the earth, such as gradual changes in the atmospheric concentrations of carbon dioxide, oxygen and water vapor, temperature changes, water elevation changes etc. These may enable a deeper data analytics and temporal classifications [8]-[9] that identify the significant factors that drive the speed of evolution and other characteristics of evolution over time.

***References:***


[1] D. Baum and S. Smith, *Tree Thinking: An Introduction to Phylogenetic Biology*, (Roberts and Company Publishers, 2012)

[2] B.G. Hall, *Phylogenetic Trees Made Easy: A How to Manual*, 4th edition, (Sinauer Associates, 2011)

[3] P. Lerney, M. Salemi, and A.-M Vandamme, editors. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd edition, (Cambridge University Press, 2009)

[4] P.Z. Revesz and Z. Li, "Incremental phylogenetics by repeated insertions: An evolutionary tree algorithm," *International Journal of Biology and Biomedical Engineering*, **10**, 1, 148-158, (2015)

[5] P.Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, (Springer, New York, 2010)

[6] P.Z. Revesz, "An algorithm for constructing hypothetical evolutionary trees using common mutations similarity matrices," *Proc. 4th ACM International Conference on Bioinformatics and Computational Biology*, 731-734, (ACM Press, New York, 2013)

[7] P.Z. Revesz, "A recurrence equation-based solution for the cubic spline interpolation problem," *International Journal of Mathematical Models and Methods in Applied Sciences*, **9**, 1, 446-452, (2015)

[8] P.Z. Revesz and T. Triplet, "Classification integration and reclassification using constraint databases," *Artificial Intelligence in Medicine*, **49,** 2, 79-91, (2010)

[9] P.Z. Revesz, and T. Triplet, "Temporal data classification using linear classifiers," *Information Systems*, **36**, 1, 30-41, (2011)

[10] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Molecular Biological Evolution*, **4**, 406-425, (1987)

[11] M. Shortridge, T. Triplet, P.Z. Revesz, M. Griep, and R. Powers, "Bacterial protein structures reveal phylum dependent divergence," *Computational Biology and Chemistry*, **35**, 1, 24-33 (2011)

[12] R.R. Sokal, and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, **38**, 1409-1438, (1958)

[13] P. Soares, L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Röhl, A. Salas, S. Oppenheimer, V. Macaulay and M.B. Richards, "Correcting for purifying selection: an improved human mitochondrial molecular clock", *American Journal of Human Genetics,* (2009)

[14] T. Triplet, M. Shortridge, M. Griep, J. Stark, R. Powers, and P.Z. Revesz, "PROFESS: A protein function, evolution, structure and sequence database," *Database -- The Journal of Biological Databases and Curation*, (2010)

[15] S. Zhang and T. Wang, "A new distance-based approach for phylogenetic analysis of protein sequences," *International Journal of Biology and Biomedical Engineering*, **3,** 3, 35-42, (2009)

[16] L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes and A.C. Wilson, "African populations and the evolution of human mitochondrial DNA", Science, 253 (5027): 1503–7 (1991)