ON THE DETERMINATION OF INTERNAL NODES OF AN EVOLUTIONARY DENDROGRAM

J. TOHÁ and M. A. SOTO
Biofísica, Departamento de Física, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 487–3, Santiago, Chile

(Received 17 October 1996)

Abstract. In this communication the determination of intermediary nodes in evolutionary dendrograms of proteins is analyzed, based on the molecular mechanisms involved in these transitions and the probability of acceptance of the amino acid substitutions.

1. Introduction

Evolutionary dendrograms generally define the well-known extant species as external vertices and the distances mediating among species can be determined with relative accuracy. Nevertheless, the inference of the identity and distances of internal (intermediate) nodes in an evolutionary tree is not always reliable.

In usual binary trees, when species derived from a common node differ in amino acids that have codons dissimilar in only one base, it is acceptable to assign to the ancestor (defined by the most parsimonious pathway) an structure similar to one of the descendant species. However, if the amino acid substitution corresponds to codons with more than one base mutation, for the structure of the intermediate more than one alternative are possibles, which can be analyzed in terms of probability of codon substitution and amino acid expression.

The aim of this communication is to analyze the doubtful ancestor structure, considering not only the feasibility of amino acid substitution, backed in general by its phylogenetic representation, but the probability of codon transitions induced by erroneous copy of the original template.

There are in the literature references on the probability of amino acid and codon evolutionary substitutions, but most of them are based on mathematical models with symmetric fixed parameters of transition which not always represent faithfully the particular physico chemical conditions permitting these exchanges (Holmquist, 1976; Fitch, 1980; Kimura, 1981; Felsenstein, 1984; Tajima, 1984; Zharkikh, 1994).

2. Methods

In a well-known communication, Topal & Fresco (1976) discussed the complementary base pairing and the origin of substitution mutations on the basis of

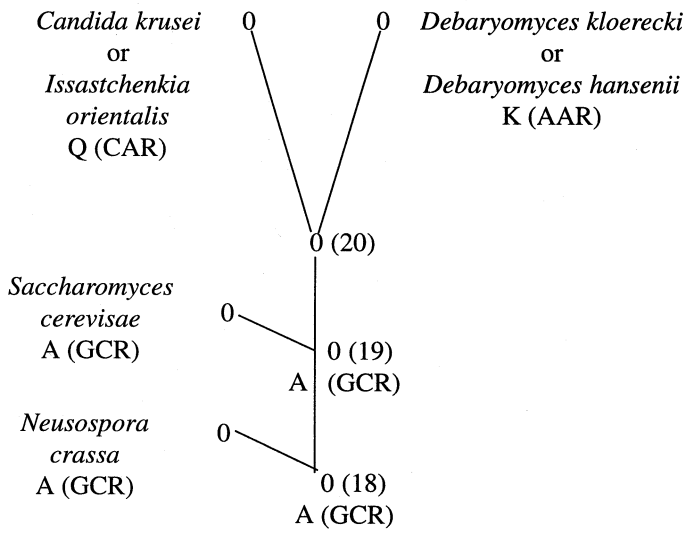
Origins of Life and Evolution of the Biosphere **28:** 97–103, 1998.

chemical considerations, model building and experimental observations. Transitions are defined by purine-pyrimidine or pyrimidine-purine mispairs, induced by spontaneous tautomeric enol or imine forms of bases and transversions are induced by purine-purine or pyrimidine-pyrimidine mispairs, which demand a tautomeric form of the template base and a syn conformation in the copy. Nevertheless, the pyrimidine-pyrimidine mispairs could not be satisfactorily built into a regular DNA helix as base pairs and experimentally they do not occur with any significant frequency. In fact, in a matrix of amino acid transition in a group of protein families (Hasegawa and Yano, 1975; Argyle, 1980) substitutions involving pyrimidine-pyrimidine mispairs are absent. On the other hand, is easy to see at the genetic code classic representation that codon transitions from the upper right quarter to the left bottom quarter and vice versa, involve at least two base transversions at the two first positions of their codons. These exchanges obligate to a non canonical pyrimidine-pyrimidine mispairs, as shown:

	First possibility				Second possibility			
	< – – – – – – – >				< – – – – – – – – – – >			
First position	(Pyr	– – –	Pyr)	Pur		Pyr	(Pur	Pur)
Second position	(Pur		Pur)	Pyr	or	Pur	(Pyr	– – – Pyr)
Third position	R/Y		Y/R	R/Y		R/Y	Y/R	R/Y
	Original		Copy	Final		Original	Copy	Final

Among all the amino acid possible transitions, there are 32 out of 380 exchanges corresponding to these substitutions. These involve in general sustitutions of chemical and structural dissimilar amino acids. Nevertheless, there are cases between near species where, at the same sequence position, this situation of dissimilar amino acids is present. The difficult transition observed in these species can be resolved if at the same position in their common ancestor (intermediate node) is present an amino acid with a codon belonging to the upper left quarter or to the bottom right quarter of the genetic code, with two purine or pyrimidine bases at the first and second position of the codon or vice-versa. This structure allows a canonical substitution.

Some demostrative examples can be analyzed in a branch of the evolutionary tree of cytochrome c. (Dayhoff *et al.* 1972), where two neighbour species: *Candida krusei* or *Issatchenkia orientalis* (Swiss Protein Bank) and *Debaryomyces kloerecki* or *Debaryomyces hansenii* (Swiss Protein Bank) share a common ancestor (node number 20 in the dendrogram). In the branch, *Saccharomyces cerevisae* derives from node number 19 and *Neurospora crassa* from node number 18.



At position 8th of the cytochrome c sequence *C.krusei* or *I. orientalis* display a Q residue and *D. kloerecki* or *hansenii* a K residue. Node number 20 appears open in the branch and *Saccharomyces cerevisae, Neurospora crassa* as well as nodes 19 and 18 have an A residue at that position. The transition of A (node 19) to Q (*I.orientalis*) is a difficult one because its involves a purine-purine and a pyrimidine-pyrimidine mispairings, as shown:

First possibility				Second possibility			
(G		G)	C	G	(C	− − −	C)
(C	− − −	T)	A	C	(G		A)
R		Y	R	R	Y		R
A codon		Copy	Q codon	A codon	Copy		Q codon

In both possibilities there is an improbable pyrimidine-pyrimidine mispairing and also a purine-purine abnormal complementarity. This situation could be circumvented, if at node number 20 a K residue is incorporated. The transition from K to Q residues could occur as follows:

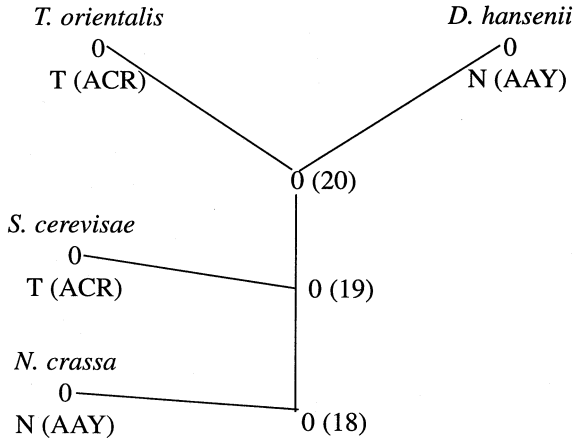
First possibility				Second possibility			
(A	G)	C		A	(T	− − −	C)
(A	T)	A		A	T		A
R	Y	R		R	Y		R
K codon	Copy	Q codon		K codon	Copy		Q codon

The first possibility has a probability of transition of the order of 10^{-8} to 10^{-9}, following Topal & Fresco (1976), because there is only one purine-purine mispair-

ing. Moreover the transition from A residue of node number 19 to K at node 20 is not forbidden:

First possibility				Second possibility			
G		T	A	G	(C	A)	
(C	– – –	T)	A	C	(G	A)	
R		Y	R	R	Y	R	
A codon		Copy	K codon	A codon	Copy	K codon	

At the same branch of the cytochrome c evolutionary tree, at position 16, *I. orientalis* and *S. cerevisae* have a T residue and *D. hansenii* and *N. crassa* an N residue. Nodes 18, 19 and 20 remain open in the original tree.



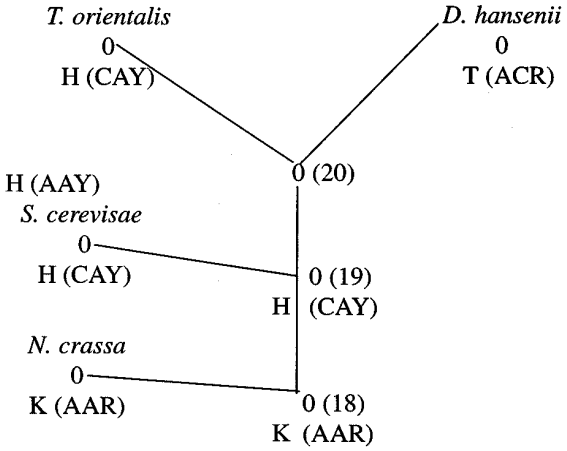
A solution considering an N residue at position 16 of nodes 18, 19 and 20 could be described as follows:

First possibility				Second possibility			
A	T	A		A	T		A
(A	G)	C		A	(T	– – –	C)
Y	R	Y		Y	R		Y
N codon	Copy	T codon		N codon	Copy		T codon

The probability of mispairing of the first possibility, involving only one transition, would have a probability of 10^{-8}- 10^{-9} (Topal and Fresco, 1976).

A similar solution is obtained with a T residue in this position at the internal nodes 18, 19 and 20.

Also in the branch under study of the cytochrome c evolutionary tree, at position 47, there are transitions where the probability to be accomplished could be improved. *I.orientalis* has an H residue at that position and *D.hansenii* a T. Nodes 19, 20 and *S. cerevisae* have an H residue and node 18 as well as *N.crassa* a K.



The transition from H (node 20) to T in *D.hansenii* has a very low probability of occurrence due to the nature of the involved mispairings, a purine-purine and an improbable pyrimidine-pyrimidine mispairing.

The inclusion of N residue (or K) at the level of node 20 facilitates these transitions because the N substitution by T is feasible, as shown:

First possibility			Second possibility		
A	T	A	A	T	A
(A	G)	C	A	(T −−−	C)
Y	R	Y	Y	R	Y
N codon	Copy	T codon	N codon	Copy	T codon

The first possibility has a probability of substitution about 10^{-8} to 10^{-9} due to the presence of a purine-purine mispairing. It is interesting to quote that N and T residues belong to the same group of polar but not charged amino acids and that at different positions of the cytochrome c sequence there are evolutionary substitutions of this kind. On the other hand, the transition of H at node 19 to an N at node 20 is acceptable following the scheme here shown:

First possibility			Second possibility		
(C −−−	T)	A	C	(G	A)
A	T	A	A	T	A
Y	R	Y	Y	R	Y
N codon	Copy	N codon	H codon	Copy	N codon

The probability of transition following the second possibility corresponds to a purine-purine mispairing and its probability of ocurrence is 10^{-8} to 10^{-9}.

3. Discussion

The examples analyzed above correspond to amino acid substitutions accomplished between intermediate nodes and neighbour species, which are characterized by the different chemical properties of the amino acids involved and also by the differences in their codons (more than one base mutation) avoiding an easy exchange.

In general, these transitions correspond to substitutions between amino acids with codons located respectively at the upper right quarter and at the bottom left quarter of the genetic code representation:

TheGeneticCode

	U		C		A		G		
	UUU	Phe	UCU		UAU	Tyr	UGU	Cys	U
U	UUC		UCC	Ser	UAC		UGC		C
	UUA	Leu	UCA		UAA	Ter	UGA	Ter	A
	UUG		UCG		UAG		UGG	Trp	G
	CUU		CCU		CAU	His	CGU		U
C	CUC	Leu	CCC	Pro	CAC		CGC	Arg	C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
	AUU		ACU		AAU	Asn	AGU	Ser	U
A	AUC	Ile	ACC	Thr	AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
	GUU		GCU		GAU	Asp	GGU		U
G	GUC	Val	GCC	Ala	GAC		GGC	Gly	C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

These substitutions involve at least two base changes, as shown, at the first two positions of the codons:

$$\begin{array}{ccc} \text{Purine} & & \text{Pyrimidine} \\ & \longleftrightarrow & \\ \text{Pyrimidine} & & \text{Purine} \end{array}$$

This can be seen, for example, at position 110 of cytochrome c dendrogram where node 19 has a T residue (ACR) and the derived *S. cerevisae* has a C residue (UGY). These transitions induce pyrimidine-pyrimidine mispairings, which are not satisfactory built as normal base pairs into a regular DNA helix. On the contrary, amino acids located at the upper left quarter and bottom right quarter of the genetic code differing also in at least two bases at the first positions of codons, can originate

non restricted transitions. Furthermore, they serve, as shown in the examples, as good intermediates for forbidden substitutions.

The analysis performed in this communication is based on the study of the frequency of spontaneous base mispairing (Topal and Fresco, 1976) considering in the case of transitions that the probability of the necessary tautomeric protonation of the bases involved in the mispairing is around 10^{-8} to 10^{-9}. In the case of transversions, the probability of the necessary tautomeric protonation of the template base plus a concomitant presence of a syn structure at the copied base is, for purine-purine mispairings, about 10^{-9} to 5×10^{-12}.

A more accurate solution for the study of the probability of base mispairing can be done experimentally by construction of small complementary oligonucleotides, including the triplet involved in the substitution and measuring the degree of hybridization with the original DNA section. This test can be performed following spectrophotometric, chromatographic or electrophoretic changes against temperature variation (Tm measurements of oligonucleotide complementarity) (Borer *et al*., 1974; Vilenchik *et al*., 1994).

References

Argyle, E.: 1980, *Origins of Life* **10**, 357.
Borer, P. N., Dengler, B., Tinoco Jr, I. and Uhlenbeck, O. C.: 1974, *J. Mol. Biol.* **86**, 843.
Dayhoff, M. O.: 1972, Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, p. D-367.
Felsenstein, J.: 1984, *Evolution* **38**, 16.
Fitch, W. M.: 1980, *J. Mol. Evol.* **16**, 153.
Holmquist, R.: 1976, *J. Mol. Evol.* **8**, 337.
Kimura, M.: 1981, *Proc. Natl. Acad. Sci. USA.* **78**, 454.
Tajima, F. and Nei, M.: 1984, *Mol. Biol. Evol.* **1**, 269.
Tinoco Jr, Y., Uhlenbeck, O. C. and Levine, M. D.: 1971, *Nature* **230**, 362.
Topal, M. D. and Fresco, J. R.: 1976, *Nature* **263**, 285.
Vilenchick, M., Belenky, A. and Cohen, A. S.: 1994, *J. Chromatogr. A* **663**, 105.
Zhaikikh, A.: 1994, *J. Mol. Evol.* **39**, 315.