Statistical Alignment of Retropseudogenes and Their Functional Paralogs

Miklós Csűrös* István Miklós†

March 16, 2005

Abstract

We describe a model for the sequence evolution of a processed pseudogene and its paralog from a common protein-coding ancestor. The model accounts for substitutions, insertions and deletions, and combines nucleotide- and codon-level mutation models. We give a dynamic programming method for calculating the likelihood of homology between two sequences in the model, and describe the accompanying alignment algorithm. We also describe how ancestral codons can be computed when the same gene produced multiple pseudogene homologs. We apply our methods to the evolution of human cytochrome *c*.

Key words molecular evolution, pairwise sequence alignment, processed pseudogenes, Thorne-Kishino-Felsenstein model, cytochrome c.

Running title Statistical alignment of retropseudogenes

Abbreviations

- **HCS** human somatic cytochrome *c*
- **JC** Jukes-Cantor [distance]
- **K2P** Kimura's two parameter [distance]
- **Mya** million years ago
- **TKF** Thorne-Kishino-Felsenstein [model]
- **OWM, NWM** Old World monkeys, New World monkeys

*Department of Computer Science and Operations Research, Université de Montréal, C. P. 6128, succ. Centre-Ville, Montréal, Québec H3C 3J7, Canada. E-mail: csuros@iro.umontreal.ca. Phone: +1 (514) 343-6111 ext. 1655, Fax: +1 (514) 343-5834. Corresponding author.

†Department of Plant Taxonomy and Ecology, Eötvös Lóránd University, 1117 Budapest, Pázmány Péter Sétány 1/c, Hungary. E-mail: miklosi@ramet.elte.hu

1 Introduction

Processed pseudogenes (Vanin 1985; Mighell et al. 2000), or *retropseudogenes*, are created by reverse transcription from mRNA. They are almost always non-functional even at the time point when they are first integrated into a chromosome (Graur and Li 2000), due to the likely lack of regulatory elements at the point of insertion. As a result, pseudogenes are generally not subject to natural selection, and thus are ideal for studying neutral evolution (Li et al. 1981). In some rare, but notable cases, the retrotransposition leads to functional elements (Balakirev and Ayala 2003): the present work is concerned only with non-functional processed pseudogenes. Retropseudogenes are free to accumulate mutations, including nucleotide substitutions, deletions and insertions. They are a valuable resource for evolutionary and comparative studies, in that they reveal evolutionary histories within and across species, help us understand mechanisms of DNA repair, and provide information about ancient transcripts.

Processed pseudogenes are typically found by aligning them to their functional paralogs: retropseudogenes are then classified as such based on evidence of retrotranscription consisting of frameshifts, mid-sequence stop codons, characteristic truncations at the 5' end, and polyadenilate tails (Harrison et al. 2001; Harrison et al. 2002). While rapid alignment heuristics are sufficient for constructing genome-wide catalogues of pseudogenes (Zhang et al. 2003; Torrents et al. 2003; Zhang et al. 2004), high-quality alignment techniques are essential for studying mutation patterns. Insights into the dynamics of molecular evolution can be drawn only from an accurate assessment of substitution rates (Yang and Nielsen 1998), and of indel events (Zhang and Gerstein 2003b). Furthermore, functionality tests employed in the identification of pseudogenes (Torrents et al. 2003; Coin and Durbin 2004), and estimation of non-functionalization time (Fleishman et al. 2004) rely on alignments of homologous characters.

When aligning the coding homolog with the non-coding pseudogene sequence, one needs to design various policies to deal with codon- vs. nucleotide-level mismatches, and with gaps that result in frameshifts. A few heuristics have been proposed for similar purposes (e.g., Gotoh 2000; Zhang et al. 1997). This paper aims to introduce a statistical model for the specific purpose of aligning the DNA sequence of a retropseudogene with the coding DNA sequence (CDS in Genbank) of a paralog protein-coding gene. Our model is based on the framework introduced by Thorne et al. (1991), in which substitutions, insertions, and deletions are generated by stochastic processes. The Thorne-Kishino-Felsenstein framework has been extended and applied to multiple sequences (Hein et al. 2000; Hein 2001; Steel and Hein 2001; Miklós 2002; Lunter et al. 2003; Hein et al. 2003). All these extensions apply to the case of aligning sequences from the same "alphabet:" nucleotides, amino acids, or codons. The main purpose of this work is to demonstrate how statistical alignment techniques can be employed when sequences come from different alphabets: codons and nucleotides. A rigorous statistical alignment framework has several advantages, which include the possibility of parameter inference, hypothesis testing, and quantitative assessment of the alignment (Hein et al. 2000; Lunter et al. 2005).

2 Materials and methods

2.1 A statistical model for retropseudogene evolution

Let ΨG be a retropseudogene and G its functional paralog (see Figure 1). Assume that they share a common ancestor G_0 (as it will be clear, this assumption is not restrictive, as evolution on the $G_0 - G$ branch is reversible). Sequence evolution on the $G_0 - G$ branch is subject to natural selection, and occurs at the codon-level. Evolution on the $G_0 - \Psi G$ branch is neutral, and operates at the nucleotide level.

[Figure 1 about here.]

Our sequence evolution model is based on the Thorne-Kishino-Felsenstein (TKF) model (1991). First, we describe the TKF model of sequence evolution, and then discuss the substitution models in detail. In the TKF model, a sequence consists of a finite string of *characters* (such as nucleotides or codons), separated by *links*. The sequence starts with a so-called *immortal link*, followed by the first character, followed by the first *mortal link*, followed by the second character, and so on, alternating characters and mortal links so that there is one link to the right of each character. Insertions and deletions are produced by time-continuous Markov processes. Deletions of single characters are generated by the mortal links, at a rate of μ per unit time and link. Insertions of single characters are generated by all links, at a rate of λ per unit time and link. In an insertion event at a link, a character is inserted to its right, along with the new character's link. Conversely, a deletion event removes a link, with the character to its left. The immortal link is never deleted. In parallel to the birth-death process of character-link pairs, individual characters mutate according to a continuous-time mutation process.

An alignment of an ancestral sequence with its descendant shows the fate of every ancestral character. The following categories are considered:

$C\underbrace{\#\cdots\#}_{n \text{ times}}$	ancestral character survives, along with $n \geq 0$ characters generated by its link;
$\square\underbrace{\#\cdots\#}_{n \text{ times}}$	ancestral character does not survive, leaves $n > 0$ characters generated by its link;
$\square$	ancestral character dies out and leaves no descendants;
$\star\underbrace{\#\cdots\#}_{n \text{ times}}$	immortal link generates n descendants.

Here, C stands for homologous, # stands for non-homologous characters, $\square$ denotes a [deletion] gap, and $\star$ represents the immortal link.

Let t denote the time (branch length) and let λ, μ be the parameters of the TKF birth-death process on the $G_0 - G$ branch. On the $G_0 - \Psi G$ branch, the corresponding parameters are t_Ψ, λ_Ψ, and μ_Ψ, respectively. Define

$$\beta = \frac{1 - e^{(\lambda-\mu)t}}{\mu - \lambda e^{(\lambda-\mu)t}}; \tag{1a}$$

$$B = \lambda\beta \qquad E = \mu\beta \qquad J = 1 - \lambda\beta \tag{1b}$$
$$H = e^{-\mu t}(1 - \lambda\beta) \qquad N = (1 - e^{-\mu t} - \mu\beta)(1 - \lambda\beta); \tag{1c}$$

$$\beta_\Psi = \frac{1 - e^{(\lambda_\Psi - \mu_\Psi)t_\Psi}}{\mu_\Psi - \lambda_\Psi e^{(\lambda_\Psi - \mu_\Psi)t_\Psi}}; \tag{2a}$$

$$B_\Psi = \lambda_\Psi\beta_\Psi \qquad E_\Psi = \mu_\Psi\beta_\Psi \qquad J_\Psi = 1 - \lambda_\Psi\beta_\Psi \tag{2b}$$
$$H_\Psi = e^{-\mu_\Psi t_\Psi}(1 - \lambda_\Psi\beta_\Psi) \qquad N_\Psi = (1 - e^{-\mu_\Psi t_\Psi} - \mu_\Psi\beta_\Psi)(1 - \lambda_\Psi\beta_\Psi). \tag{2c}$$

Probabilities for different fates can be calculated as follows (Lunter et al. 2005).

Fate	$G_0 \to G$ probability	$G_0 \to \Psi G$ probability
$C \to C\#^{n-1}$	HB^{n-1}	$H_\Psi B_\Psi^{n-1}$
$C \to \square\#^n$	NB^{n-1}	$N_\Psi B_\Psi^{n-1}$
$C \to \square$	E	E_Ψ
$* \to \#^n$	JB^n	$J_\Psi B_\Psi^n$

Notice that these probabilities do not include substitutions. Characters on the $G_0 - G$ branch are codons, while on the $G_0 - \Psi G$ branch, they are nucleotides.

We assume that the processes guiding the evolution on the $G_0 - G$ branch are reversible, and thus both G and G_0 are at equilibrium. As a consequence, their length distribution (measured in codons) is geometric with parameter $\gamma = \lambda/\mu$. Accordingly, the expected length of G and G_0 is $\frac{1-\gamma}{\gamma}$ codons.

The indel processes for the two branches are intertwined: there is a link in every third nucleotide position on the gene branch, and there is a link in every nucleotide position on the pseudogene branch. Not surprisingly, the evolutionary history of the sequences can be rather convoluted in this model, see Fig. 2 for an example. The alignment procedure needs to find homologous positions between the sequences of G and ΨG.

[Figure 2 about here.]

2.2 Recursions for the alignment

The following notation is used in describing the alignment procedures. Let $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ be the DNA alphabet, and $\Sigma' = \Sigma \cup \{\square\}$ where $\square$ indicates a gap (i.e., a deceased ancestral character). The coding DNA sequence of the gene G is $g^1 \cdots g^L$ (where L is a multiple of three), and the DNA sequence of the pseudogene ΨG is $h^1 \cdots h^\ell$ (with no restriction on ℓ). For the codon ending with the i-th nucleotide we use the shorthand notation $g^{i-2..i} = g^{i-2}g^{i-1}g^i$. Let $f^1 \cdots f^{L_0}$ denote the (unknown) ancestral gene sequence, where $L_0/3$ is geometrically distributed with parameter γ: $\mathbb{P}\{L_0 = 3k\} = (1-\gamma)\gamma^{k-1}$ for all $k = 1, 2, \ldots$, and f^i are independent identically distributed nucleotides.

At first sight, calculating the likelihood of homology may seem unfeasible since an unbounded number of ancestral codons may die out without descendants, which means that

an infinite number of histories need to be considered. For that reason, we design a dynamic programming approach that works with blocks formed by modern descendants of the same ancestral codon, and include the infinite summation of histories between such blocks using a closed formula. Another complication stems from the fact that the probability of observing the three nucleotides of a codon can only be calculated when the homologous positions in the two sequences are determined, while insertions on the pseudogene branch may occur between homologous positions within a codon.

First, we give recursions for computing the likelihood of homology between prefixes of the two input sequences while ignoring substitutions. Different variables are introduced to track the phases within a coding frame. The principal variable is $\mathsf{A}(i,j)$, which is the likelihood of observing $g^1 \cdots g^i$ and $h^1 \cdots h^j$ as being all the living descendants of some ancestral prefix $f^1 \cdots f^k$ with complete codons, where k and i are integer multiples of three. Figure 3 shows the recursions[1] between the variables: we track homologous positions of the three sequences together. Notice that we impose that insertions on the $G_0 \to \Psi G$ branch precede those on the $G_0 \to G$ branch within each aligned codon position. (The alternative would be to place insertions generated by a codon link in the third codon position after codon insertions on the gene branch.)

[Figure 3 about here.]

The probability of multiple visits to state EEE while passing from state Z to state A can be summed (Miklós 2002; Steel and Hein 2001) to obtain $\mathsf{A}(i,j) = \mathsf{Z}(i,j)\delta$, where

$$\delta = \sum_{k=0}^{\infty} \left(\gamma E E_{\Psi}^{3}\right)^{k} = \frac{1}{1-\gamma\mu\beta(\mu_{\Psi}\beta_{\Psi})^{3}}.$$

In order to reduce the number of variables, we make the following observations.

- Passages through states for which an ancestral character dies out in ΨG ($\mathsf{SE}i, \mathsf{EE}i, \mathsf{EA}i$) can be incorporated into the recursions, and thus they become obsolete.
- State pairs $\mathsf{SH}k$ and $\mathsf{SN}k$ can be merged into one single state $\mathsf{S}k$ for each k. The original codon phase for $\mathsf{S}k(i,j)$ is is the remainder of i when divided by three, and thus $i \bmod 3 = 1$ for $\mathsf{S1}(i,j)$, $i \bmod 3 = 2$ for $\mathsf{S2}(i,j)$ and $i \bmod 3 = 0$ for $\mathsf{S3}(i,j)$.
- State pairs $\mathsf{EH}k$ and $\mathsf{EN}k$ can be merged into one state $\mathsf{E}k$ for each i.
- State Z can be eliminated.

We need to track the variables $\mathsf{S}k(i,j), \mathsf{E}k(i,j), \mathsf{NNN}(i,j), \mathsf{A}(i,j)$ where $k = 1,2,3$, $i \bmod 3 = 0$ for $\mathsf{E}k, \mathsf{NNN}, \mathsf{A}$, and $i \bmod 3 = k \bmod 3$ for $\mathsf{S}k$. The full probability of observing the modern sequences is $\mathsf{A}(L,\ell)$.

[1]Equations for these initial recurrences are shown in Supplemental Material.

We use the state NNN to handle the characters generated from the immortal link (see Fig. 4). Accordingly, we pose

$$\mathsf{NNN}(0,0) = (1-\gamma)JJ_\Psi; \tag{3a}$$

$$\mathsf{NNN}(0,j) = \mathsf{NNN}(0,j-1)B_\Psi\pi_\Psi(h^j) \qquad \text{for } j>0; \tag{3b}$$

and $\mathsf{NNN}(i,j) = 0$ if $i<0$ and/or $j<0$.

[Figure 4 about here.]

We formalize the assumptions about the substitution processes.

- The conditional probability of having $g_1g_2g_3$ in G, given that $f_1f_2f_3$ is its ancestral codon in G_0, equals $\tau(f_1f_2f_3 \to g_1g_2g_3)$ where $\tau\colon \Sigma^3\times\Sigma^3 \mapsto [0,1]$ is the transition probability function. New codons are inserted, and the ancestral codons are selected according to the distribution π. The assumption that π is the stationary distribution is not used in the upcoming recursions, but is logical if G and G_0 are to have the same codon distribution, and includes the case when the functional gene that gave rise to ΨG is not an ancestor of G.

- The conditional probability of having nucleotide h in ΨG, given that its ancestral homologue is f in G_0, equals $\tau_\Psi(f\to h)$. New nucleotides are inserted according to the distribution π_Ψ.

As a consequence of the codon-level evolutionary process, substitution probabilities can only be calculated when a whole codon is aligned. In order to track the nucleotides of ΨG aligned in the different codon positions, we use the variables $\mathsf{S}1(i,j,h), \mathsf{S}2(i,j,h'h), \mathsf{S}3(i,j)$, where $h,h'\in\Sigma'$, and for each $\mathsf{S}k$, $i=k,k+3,\ldots,L+3-k$ and $j=1,2,\ell$. We also use the variables $\mathsf{E}1(i,j,h), \mathsf{E}2(i,j,h'h), \mathsf{E}3(i,j)$, where $h,h'\in\Sigma'$, and for each $\mathsf{E}k$, $i=0,3,\ldots,L$ and $j=1,2,\ldots,\ell$. Finally we need the variables $\mathsf{A}(i,j)$, and $\mathsf{NNN}(i,j)$, where $i=0,3,\ldots,L$ and $j=0,1,\ldots,\ell$. The upcoming recursions reflect the principle that $H, H_\Psi, B, B_\Psi, \gamma$ factors are included as soon as possible, and E, E_Ψ factors are included when it is known that the

ancestral character has no descendant. The recursions are the following.

$$\begin{aligned}
\mathsf{S1}(i,j,h) = \quad & \mathsf{A}(i-1,j-1)\Big(\{h=h^j\}H_\Psi + \{h=\square\}N_\Psi\pi_\Psi(h^j)\Big) \\
& +\mathsf{S1}(i,j-1,h)B_\Psi\pi_\Psi(h^j)
\end{aligned} \tag{4a}$$

$$\begin{aligned}
\mathsf{S2}(i,j,h'h) = \quad & \mathsf{S1}(i-1,j-1,h')\Big(\{h=h^j\}H_\Psi + \{h=\square\}N_\Psi\pi_\Psi(h^j)\Big) \\
& +E_\Psi\mathsf{A}(i-2,j-1)\Big(\{h=h^j,h'=\square\}H_\Psi + \{h=h'=\square\}N_\Psi\pi_\Psi(h^j)\Big) \\
& +\mathsf{S2}(i,j-1,h'h)B_\Psi\pi_\Psi(h^j)
\end{aligned} \tag{4b}$$

$$\begin{aligned}
\mathsf{S3}(i,j) = \quad & \sum_{h,h'\in\Sigma'}\mathsf{S2}(i-1,j-1,h'h)\Big(H_\Psi p(g^{i-2..i},h'hh^j) + N_\Psi\pi_\Psi(h^j)p(g^{i-2..i},h'h\square)\Big) \\
& +E_\Psi\sum_{h\in\Sigma'}\mathsf{S1}(i-2,j-1,h)\Big(H_\Psi p(g^{i-2..i},h\square h^j) + N_\Psi\pi_\Psi(h^j)p(g^{i-2..i},h\square\square)\Big) \\
& +E_\Psi^2\mathsf{A}(i-3,j-1)\Big(H_\Psi p(g^{i-2..i},\square\square h^j) + N_\Psi\pi_\Psi(h^j)p(g^{i-2..i},\square\square\square)\Big) \\
& +\mathsf{S3}(i,j-1)B_\Psi\pi_\Psi(h^j).
\end{aligned} \tag{4c}$$

$$\begin{aligned}
\mathsf{E1}(i,j,h) = \quad & \mathsf{A}(i,j-1)\Big(\{h=h^j\}H_\Psi + \{h=\square\}N_\Psi\pi_\Psi(h^j)\Big) \\
& +\mathsf{E1}(i,j-1,h)B_\Psi\pi_\Psi(h^j)
\end{aligned} \tag{5a}$$

$$\begin{aligned}
\mathsf{E2}(i,j,h'h) = \quad & \mathsf{E1}(i,j-1,h')\Big(\{h=h^j\}H_\Psi + \{h=\square\}N_\Psi\pi_\Psi(h^j)\Big) \\
& +E_\Psi\mathsf{A}(i,j-1)\Big(\{h=h^j,h'=\square\}H_\Psi + \{h=\square,h'=\square\}N_\Psi\pi_\Psi(h^j)\Big) \\
& +\mathsf{E2}(i,j-1,h'h)B_\Psi\pi_\Psi(h^j)
\end{aligned} \tag{5b}$$

$$\begin{aligned}
\mathsf{E3}(i,j) = \quad & \sum_{h,h'\in\Sigma'}\mathsf{E2}(i,j-1,h'h)\Big(H_\Psi p(\square\square\square,hh'h^j) + N_\Psi\pi_\Psi(h^j)p(\square\square\square,h'h\square)\Big) \\
& +E_\Psi\sum_{h\in\Sigma'}\mathsf{E1}(i,j-1,h)\Big(H_\Psi p(\square\square\square,h\square h^j) + N_\Psi\pi_\Psi(h^j)p(\square\square\square,h\square\square)\Big) \\
& +E_\Psi^2\mathsf{A}(i,j-1)\Big(H_\Psi p(\square\square\square,\square\square h^j) + N_\Psi\pi_\Psi(h^j)p(\square\square\square,\square\square\square)\Big) \\
& +\mathsf{E3}(i,j-1)B_\Psi\pi_\Psi(h^j).
\end{aligned} \tag{5c}$$

$$\mathsf{NNN}(i{+}3,j) = \pi(g^{i+1..i+3})\Big(B\Big(\mathsf{S3}(i,j){+}\mathsf{NNN}(i,j){+}E_\Psi \sum_{h,h'\in\Sigma'} \mathsf{S2}(i{-}1,j,h'h)p(g^{i-2..i},h'h\square) + E_\Psi^2 \sum_{h\in\Sigma'} \mathsf{S1}(i-2,j,h)p(g^{i-2..i},h\square\square) + E_\Psi^3\mathsf{A}(i-3,j)p(g^{i-2..i},\square\square\square)\Big) \quad (6)$$
$$+ N\Big(\mathsf{E3}(i,j) + E_\Psi \sum_{h,h'\in\Sigma'} \mathsf{E2}(i,j,h'h)p(\square\square\square,h'h\square) + E_\Psi^2 \sum_{h\in\Sigma'} \mathsf{E1}(i,j,h)p(\square\square\square,h\square\square) + E_\Psi^3\mathsf{A}(i,j)p(\square\square\square,\square\square\square)\Big)\Big);$$

$$\mathsf{A}(i,j) = \delta\Big(\mathsf{S3}(i,j) + \mathsf{NNN}(i,j) + E_\Psi \sum_{h,h'\in\Sigma'} \mathsf{S2}(i-1,j,h'h)p(g^{i-2..i},h'h\square) + E_\Psi^2 \sum_{h\in\Sigma'} \mathsf{S1}(i-2,j,h)p(g^{i-2..i},h\square\square) + E_\Psi^3\mathsf{A}(i-3,j)p(g^{i-2..i},\square\square\square)$$
$$+E\Big(\mathsf{E3}(i,j) + E_\Psi \sum_{h,h'\in\Sigma'} +\mathsf{E2}(i,j,h'h)p(\square\square\square,h'h\square) + E_\Psi^2 \sum_{h\in\Sigma'} \mathsf{E1}(i,j,h)p(\square\square\square,h\square\square)\Big)\Big). \quad (7)$$

In Eqs. (4–7), $\{\cdot\}$ denote indicator functions (or guards): $\{\langle\text{condition}\rangle\} = 1$ if $\langle\text{condition}\rangle$ is true, otherwise the value is zero.

The expression $p(g_1g_2g_3, h_1h_2h_3)$ is the probability of seeing the corresponding nucleotides (or gaps) in homologous positions of a codon. Either $g_1 = g_2 = g_3 = \square$ or all three are nucleotides. These values can be calculated prior to computing the recursions in the following manner. Recall that $\pi(f_1f_2f_3)$ gives the probability of a codon $f_1f_2f_3$ in a specific position of the ancestral sequence: for all $k = 1, 2, \ldots, L_0/3$ and $f_1f_2, f_3 \in \Sigma$, $\mathbb{P}\{f^{3k-2}f^{3k-1}f^{3k} = f_1f_2f_3\} = \pi(f_1f_2f_3)$. For notational convenience, we extend the distribution to stop codons: $\pi(f_1f_2f_3) = 0$ and $\tau_\Psi(f_1f_2f_3 \to g_1g_2g_3) = \prod_k\{f_k = g_k\}$ if $f_1f_2f_3$ is a stop codon. Let

$$\tau_\Psi^{(3)}(f_1f_2f_3 \to h_1h_2h_3) = \prod_{i=1}^{3}\Big(\{h_i = \square\} + \{h_i \in \Sigma\}\tau_\Psi(f_i \to h_i)\Big)$$

for all $f_1, f_2, f_3 \in \Sigma$ and $h_1, h_2, h_3 \in \Sigma'$. Then

$$p(\square\square\square, h_1h_2h_3) = \gamma \sum_{f_1f_2f_3\in\Sigma^3} \pi(f_1f_2f_3)\tau_\Psi^{(3)}(f_1f_2f_3 \to h_1h_2h_3); \quad (8a)$$

$$p(g_1g_2g_3, h_1h_2h_3) = \gamma \sum_{f_1f_2f_3\in\Sigma^3} \pi(f_1f_2f_3)H\tau(f_1f_2f_3 \to g_1g_2g_3)\tau_\Psi^{(3)}(f_1f_2f_3 \to h_1h_2h_3) \quad (8b)$$

where $h_1, h_2, h_3, \in \Sigma'$. In practice, these values can be computed very quickly by multiplying the matrices formed by the τ and $\tau_\Psi^{(3)}$ mutation probabilities.

Observe that the subsitution models appear only in Eqs. (8a–b). In what follows we describe specific codon- and nucleotide-substitution models that we employed in experiments, but the recursions of Eqs. (4–8) are valid for any Markov model. In fact, they can be readily adapted even to the case of aligning the protein sequence of a gene with the DNA sequence of its retropseudogene.

2.3 Modeling codon substitutions on the gene branch

Substitutions on the $G_0 - G$ branch occur according to a continuous-time Markov process at the codon-level. Each codon is assumed to evolve independently, as determined by the same instantaneous rate matrix $\mathbf{Q}$. Let $\mathcal{C}$ denote the alphabet of sense codons. The matrix $\mathbf{Q}$ is of size $|\mathcal{C}| \times |\mathcal{C}|$, and the matrix exponential $\mathbf{M} = e^{\mathbf{Q}\tau}$ gives the transition probabilities for the Markov process for an interval of length τ. If X' and X are the random codons in homologous positions of G_0 and G, respectively, then the probability of seeing c in the descendant, conditioned on the ancestral character state, is

$$\tau(c' \to c) = \mathbb{P}\Big\{ X = c \;\Big|\; X' = c' \Big\} = \mathbf{M}[c', c].$$

Assume that G_0 is a sequence of independent and identically distributed (i. i. d.) codons, and each codon is drawn according to the distribution π, and thus $\mathbb{P}\{X' = c'\} = \pi(c')$. By the assumption of time-reversibility on the gene branch, $\sum_{c' \in \mathcal{C}} \pi(c')\tau(c' \to c) = \pi(c)$ for all codons c.

Imposing time-reversibility has the computational advantage that the matrix exponentiation can be carried out without much difficulty. The rate matrix is then related to a symmetric matrix $\mathbf{S}$ defined as $\mathbf{S}[c', c] = \mathbf{Q}[c', c]\sqrt{\frac{\pi(c')}{\pi(c)}}$, and thus

$$\tau(c' \to c) = \sqrt{\frac{\pi(c)}{\pi(c')}} \sum_k u^{(k)}[c] e^{\nu^{(k)}\tau} u^{(k)}[c'],$$

where $\{u^{(k)}\}$ are the orthonormal eigenvectors of $\mathbf{S}$ and $\{\nu^{(k)}\}$ are the corresponding eigenvalues (Holmes and Rubin 2002; Schadt and Lange 2002). The importance of this observation is that there exist efficient numerical methods that find the eigenvectors and eigenvalues of a real-valued symmetric matrix (Press et al. 1997).

A codon substitution model for protein-coding sequences has to account for possible changes in evolutionary fitness accompanying amino acid alterations: discourage them in case of purifying selection, or favor them in case of Darwinian selection. A pioneering model was proposed by Goldman and Yang (1994), in which the rate matrix is determined by a parameter related to the transition-transversion ratio, a selection coefficient for replacement amino acid substitutions, and an amino acid distance function. Another model was introduced at the same time by Muse and Gaut (1994), which imposes different rates for synonymous and nonsynonymous codon changes, but works with much fewer parameters (the stationary nucleotide distribution, in addition to two rate factors). Given the large number of amino acid distance parameters (one for each 190 amino acid pair; Grantham 1974), subsequent models (Yang and Nielsen 1998; Yang and Nielsen 2000; Bustamante et al. 2002) simplify the Goldman-Yang rate matrix. Schadt and Lange (2002) describe a general method of combining nucleotide substitution models with selection coefficients to derive codon substitution models. We use a rate matrix introduced by Yang and Nielsen (1998), which is determined by the codon frequencies, and two parameters, κ and ω. Parameter κ is related

to the ratio of transitions ($\mathtt{A} \leftrightarrow \mathtt{G}$ or $\mathtt{C} \leftrightarrow \mathtt{T}$) and transversions; parameter ω is a factor affecting non-synonymous mutations.

For two sense codons $c' \neq c$:

$$\mathbf{Q}[c', c] = \begin{cases} 0 & \text{if } c' \text{ and } c \text{ differ in more than one nucleotides} \\ \alpha\pi(c) & \text{if they differ by one synonymous transversion} \\ \alpha\kappa\pi(c) & \text{if they differ by one synonymous transition} \\ \alpha\omega\pi(c) & \text{if they differ by one non-synonymous transversion} \\ \alpha\omega\kappa\pi(c) & \text{if they differ by one non-synonymous transition.} \end{cases} \tag{9}$$

Entries on the diagonal are set to $\mathbf{Q}[c', c'] = -\sum_{c \in \mathcal{C}} \mathbf{Q}[c', c]$, and thus all row sums are zero. The value α is a scaling parameter, usually chosen so that the expected rate of substitutions per codon equals one: $\sum_{c'} \pi(c') \sum_{c \neq c'} \mathbf{Q}(c', c) = 1$.

2.4 Modeling nucleotide substitutions on the pseudogene branch

We use the substitution model described by Felsenstein and Churchill (1996), generally called the F84 model. The rate matrix of the F84 model can be determined by the parameters κ_Ψ and α_Ψ, in addition to the stationary probabilities π_Ψ, where κ_Ψ is the transition-transversion ratio, and α_Ψ is the rate of substitution per nucleotide site. The rates in the F84 model are

$$\begin{array}{c|cccc} & \mathtt{A} & \mathtt{G} & \mathtt{C} & \mathtt{T} \\ \mathtt{A} & \cdot & \alpha_\Psi\kappa_\Psi\pi_\Psi(\mathtt{G}) & \alpha_\Psi\pi_\Psi(\mathtt{C}) & \alpha_\Psi\pi_\Psi(\mathtt{T}) \\ \mathtt{G} & \alpha_\Psi\kappa_\Psi\pi_\Psi(\mathtt{A}) & \cdot & \alpha_\Psi\pi_\Psi(\mathtt{C}) & \alpha_\Psi\pi_\Psi(\mathtt{T}) \\ \mathtt{C} & \alpha_\Psi\pi_\Psi(\mathtt{A}) & \alpha_\Psi\pi_\Psi(\mathtt{G}) & \cdot & \alpha_\Psi\kappa_\Psi\pi_\Psi(\mathtt{T}) \\ \mathtt{T} & \alpha_\Psi\pi_\Psi(\mathtt{A}) & \alpha_\Psi\pi_\Psi(\mathtt{G}) & \alpha_\Psi\kappa_\Psi\pi_\Psi(\mathtt{C}) & \cdot \end{array}, \tag{10}$$

where entries along the diagonal are set so that the row sums are 0, and α_Ψ is a scaling parameter chosen so that the expected rate of substitutions equals one per site. (Hasegawa et al. (1985) proposed an equivalent rate matrix. The parametrization of Felsenstein and Churchill (1996) is particularly advantageous in computations.)

2.5 Implementation issues

Our recursions show how to calculate the likelihood of homology between a hypothetical pair of pseudogene and its functional paralog. For every codon-nucleotide position pair $(3k-2..3k, j)$: $k = 1, \ldots, L/3, j = 1, \ldots, \ell$ of the two sequences, there are 64 variables to compute. There is no need, however, to keep all the values in memory at the same time. For every k, the algorithm calculates $\mathsf{S1}(3k-2, j, h)$, $\mathsf{S2}(3k-1, j, h'h)$, $\mathsf{S3}(3k, j)$, $\mathsf{E1}(3k, j, h)$, $\mathsf{E2}(3k, j, h'h)$, $\mathsf{E3}(3k, j)$, $\mathsf{A}(3k, j)$, $\mathsf{NNN}(3k+3, j)$, in this order, in an iteration over j. As a consequence, it suffices to keep these 64 variables for each j, along with a copy of the $\mathsf{A}(i-3, j)$. Thus, the likelihood can be calculated by using $65(\ell+1)$ floating-point variables. While it may seem to be a subtle issue, it is important to calculate $\mathsf{NNN}(i+3, j)$

right after $\mathsf{A}(i,j)$: if we wanted to calculate $\mathsf{NNN}(i,j)$ right before calculating $\mathsf{A}(i,j)$, then we would need to store many more previous values, as $\mathsf{A}(i-6,j), \mathsf{S1}(i-5,j,h)$, and similar terms would appear in the recursion. Additionally, it is necessary to perform some scaling to avoid underflow during the calculations: we found that it is enough in practice to keep one scaling factor for every (i,j)-pair, which is an integer variable $s(i,j)$: the actual probabilities equal $10^{s(i,j)}$ times the stored floating-point value. Given the need to access variables for $(i-3,\cdot)..(i,\cdot)$ in the recursions, $4(\ell+1)$ integer values are needed for storing the scaling factors.

It is straightforward to modify the recursions if one wants to compute the most likely alignment between the two sequences: summation of terms is replaced by $\max\{\cdot\}$, and traceback pointers need to be stored. Since the number of terms on the right-hand side of every recursion is below 256, one byte is enough per traceback pointer, and thus in addition to the floating-point variables $21\frac{1}{3}L\ell + 2\ell + O(1)$ bytes are needed for storing the traceback values. As a consequence, sequences with up to a few thousand nucleotides in length can be aligned without memory problems on a desktop computer.

We tested the computational performance of the method using a test implementation and found that it is practical, as two sequences of length 1700 and 2000 can be aligned in about a minute and a half, and sequences with a few hundred nucleotides are aligned within a few seconds. (The implementation was tested on an Apple PowerBook G4 1.25 GHz; the memory footprint was below 256Mbytes for the mentioned cases when looking for the most likely alignment.)

2.6 Computing ancestral codons

Suppose that the same gene produces many processed pseudogenes in the course of its evolution. These pseudogene sequences provide "snapshots" of the gene's evolution at different time points. An alignment of the pseudogene sequences with the modern gene sequence can be used to date the non-functionalization time of the pseudogenes. Moreover, the nucleotides that are aligned in homologous positions with the functional gene can be used to deduce ancestral codons. Pupko et al. (2000) give a linear-time algorithm for computing ancestral sequences by maximum likelihood on a phylogeny with known topology. We describe here a similar approach that exploits the comb-like topology of the phylogeny of the retropseudogenes and their functional paralog, and combines the codon- and nucleotide-specific substitution processes. Formally, the problem is defined as follows. Given a gene sequence $\mathbf{g}_0 = g_0^{1..L}$, an *aligned pseudogene sequence* $\mathbf{h} = h^{1..L}$ is a sequence of L characters over the extended alphabet $\Sigma' = \Sigma \cup \{\square\}$ where h^i is either the nucleotide in a homologous position with g_0^i or it is the gap character $\square$. Let $\mathbf{h}_1, \ldots, \mathbf{h}_n$ be a set of aligned pseudogene sequences, with divergence times $t_1, \ldots, t_n$. We assume a molecular clock in the sense that the evolution of every pseudogene sequence is determined by the same [nucleotide] rate matrix $\mathbf{Q}_\Psi$, and the gene sequence's evolution is determined by a [codon] rate matrix $\mathbf{Q}$. Accordingly, the rate matrices are scaled in such a way that the probability of a $g \to h$ nucleotide substitution is $e^{\mathbf{Q}_\Psi t}[g,h]$ for divergence time t, and the probability of a $ff'f'' \to gg'g''$ codon substitution is $e^{\mathbf{Q}t}[ff'f'', gg'g'']$. We restrict our attention to the case

when the gene sequence evolution has not included codon insertions or deletions. Our aim is to characterize the ancestral gene sequences $\mathbf{g}_1, \ldots, \mathbf{g}_n$ at times $t_1, \ldots, t_n$, by computing posterior probabilities for the ancestral codons. We assume that $t_1 < t_2 < \cdots < t_n$, and for the sake of notational uniformity, we pose $t_0 = 0$. Since codons evolve independently, it suffices to consider every codon position k separately. Define the *recent evolution probabilities* $A_i(ff'f'')$ and the *ancient evolution probabilities* $B_i(ff'f'')$ for every sense codon $ff'f''$, and $i = 0, \ldots, n$ as follows. The recent evolution probability $A_i(ff'f'')$ equals the probability of observing the pseudogene sequences $\mathbf{h}_1, \ldots, \mathbf{h}_i$ and the modern gene sequence $\mathbf{g}_0$ given that at time t_i the gene had the codon $ff'f''$ in positions $k..k+2$. In a similar vein, $B_i(ff'f'')$ is the probability of observing pseudogene sequences $\mathbf{h}_{i+1}, \ldots, \mathbf{h}_n$, given that at time t_i the gene had the codon $ff'f''$ in positions $k..k+2$. These values can be computed using dynamic programming via the following recursions.

$$A_0(ff'f'') = \begin{cases} 0 & \text{if } ff'f'' \neq g_0^{k..k+2}; \\ 1 & \text{if } ff'f'' = g_0^{k..k+2}; \end{cases}$$

$$A_i(ff'f'') = \prod_{j=1}^{3} H^{t_i}(f^j, h_i^{k+j-1}) \sum_{gg'g'' \in \mathcal{C}} A_{i-1}(gg'g'') G^{t_i - t_{i-1}}(ff'f'', gg'g'') \qquad (i > 0);$$

$$B_n(ff'f'') = 1;$$

$$B_i(ff'f'') = \prod_{j=1}^{3} H^{t_{i+1}}(f^j, h_{i+1}^{k+j-1}) \sum_{gg'g'' \in \mathcal{C}} B_{i+1}(gg'g'') G^{t_{i+1} - t_i}(gg'g'', ff'f'') \qquad (i < n).$$

In the equations, $G^t(ff'f'', gg'g'')$ is the codon substitution probability, given by the corresponding entry of the matrix $e^{\mathbf{Q}t}$. The nucleotide substitution probability $H^t(f, h)$ equals the corresponding entry of the matrix $e^{\mathbf{Q}_\Psi t}$ when $h \neq \square$, otherwise, it is 1. The probability that the modern codons are observed, given that the ancestral codon was $ff'f''$ in the homologous position at time t_i, equals

$$C_i(ff'f'') = \frac{A_i(ff'f'')B_i(ff'f'')}{\sum_{gg'g'' \in \mathcal{C}} A_i(gg'g'')B_i(gg'g'')}. \tag{11}$$

3 Results

3.1 Recent history of cytochrome *c*

The cytochrome *c* gene is one of the most widely studied genes (Banci et al. 1999; Grossman et al. 2001; Evans and Scarpulla 1988; Zhang and Gerstein 2003a). It plays a central role in the electron transport chain, and is ubiquitous among living organisms. In addition, its primary sequence is highly conserved among eukaryotes, exhibiting a 45% amino acid similarity between human and yeast, and more than 93% similarity between chicken and mouse (Banci et al. 1999; Zhang and Gerstein 2003a). It is particularly important in the context of the evolution of primates: the enlarged brain and prolonged fetal life required important changes in proteins related to energy transport, including cytochrome *c*. A period

of accelerated rate of evolution was postulated by Grossman et al. (2001), which period took place in our anthropoid ancestors preceding the platyrrhine-catarrhine divergence about 40 million years ago. During this time the somatic gene underwent nine amino acid changes. The cytochrome *c* gene produced many processed pseudogenes. Some of them were identified by screening genomic libraries using DNA hybridization (Evans and Scarpulla 1988). Most of them, however, were discovered by Zhang and Gerstein (2003a) using purely computational techniques consisting of a search for cytochrome *c* protein sequence homologs in the human genome sequence. Zhang and Gerstein (2003a) named the identified pseudogenes as hcp1, hcp2, ..., hcp49. They found a single pseudogene hcp9, which is a disabled relic of a testis-specific cytochrome *c* gene, and identified 48 other pseudogenes that are products of the somatic cytochrome *c* gene. These latter can be classified into two classes, demarcating the period of accelerated evolution. The difference between estimated ages of pseudogenes in the two classes restricts the length of that period to about 15 million years (Evans and Scarpulla 1988).

We applied our methods to date the origin of these pseudogenes, and to analyze the evolution of the cytochrome *c*. For each pseudogene sequence, we carried out the likelihood calculations when aligned with the human gene sequence, and optimized the alignment parameters for maximizing the likelihoods. In a second stage, we evaluated how well gene branch lengths t and deviations from neutral selection (ω) can be detected by simultaneous optimization with the pseudogene branch lengths. We used the optimized parameters to realign the sequences and date their origin. Nucleotides in homologous positions were then used to compute ancestral codons.

Prior to the analysis, we verified the presence of the pseudogenes in the latest human genome assembly (IHGSC 2004). Six pseudogene sequences are missing from the genome annotation at NCBI. Three of those can be localized on the genome. The three remaining ones (hcp13, hcp31, and hcp47) are apparently due to errors in a previous genome assembly, as BLAST (Altschul et al. 1997) does not find them in the current assembly. Zhang and Gerstein (2003a) characterize all three of them as duplicated copies of other pseudogenes based on near-identity of sequences: some assembly errors appear as segmental duplications. (A handful[2] of likely cytochrome *c* pseudogenes different from hcp1–hcp49 can be found by searching for similarities with an HCS mRNA sequence, but we do not use them in this study.)

In the analysis of the cytochrome *c* pseudogenes, we came across a number of instances that illustrate the advantage of explicitly modeling codon evolution in the alignment. First, in a similar study, one needs to decide whether to align the protein or the nucleotide sequence of the functioning gene with the candidate pseudogene sequences. Consider the alignment $\begin{bmatrix} \texttt{.ACA.CTG.ATG.} \\ \texttt{.AC-.CAG.ATG.} \end{bmatrix}$. (Gene sequence is on the top, dots denote codon boundaries.) If the gene protein sequence is used, then the `ACA` codon gets most probably aligned with `ACC` because they code for the same amino acid, while `CTG` and `CAG` do not. Looking at the nucleotide sequences, however, the current alignment may have a better score than the alternative.

[2]The RetroGene track annotation in the UCSC genome browser includes at least five of them.

As another example, a different alignment computed by our program contains the block $\begin{bmatrix} \texttt{.AAG.AAG.AAG.GAA.GAA.AGG.} \\ \texttt{.AAG.--G.CA-.GAA.-A-.AGG.} \end{bmatrix}$. An alternative alignment for this part is $\begin{bmatrix} \texttt{.AAG.AAG.AAG.GAA.GAA.AGG.} \\ \texttt{.AAG.--G.CAG.-AA.--A.AGG.} \end{bmatrix}$. The latter has the advantage over the first that the two gaps in the fifth codon are put next to each other. Aside from the gap penalties, the first has an advantage over the second one: when summing over the possible codons at the time point when the pseudogene is created, the first alignment considers codons that are "closer" to `GAA` (`GAG` is a synonymous codon). The same argument holds for placing the single gap between the third and the fourth codon positions: the first alignment gets a better score by penalizing the synonymous codon mutation `AAG` $\rightarrow$ `AAA` less severely.

3.2 Parameter optimization and divergence times

In a first phase of the experiments, we calculated plausible parameters for modeling the evolution of cytochrome c pseudogenes. Since the gene has the same length in all known mammalian sequences (104 aa), we have imposed a very low level of indel rates on the gene sequence: $\lambda = 10^{-8}$. (From these and other experiments that we conducted, it seems that the in general, indel rates cannot be optimized simultaneously on both branches.)

We used 35 of the 46 known pseudogenes, those that are not truncated and have no large deletions. For parameter optimizations, we used Powell's (or, in case of optimizing for a single parameter, Brent's) method (Press et al. 1997). After an initial set of experiments[3], we selected a common pseudogene insertion rate $\lambda_\Psi = 0.036$ and transition-transversion ratio $\kappa_\Psi = 1.32$ to be used in all alignment calculations. Based on results by Bustamante et al. (2002), we set the transition/transversion ratio to be the same on both branches. We used the codon substitution model of Equation (9), in conjunction with the codon frequencies observed in the human protein-coding genes[4]. We optimized the pseudogene deletion rate μ_Ψ in each pairwise alignment independently, in order to avoid too much distortion on the pseudogene branch lengths due to large deletions.

From our experiments, and previous results (Evans and Scarpulla 1988; Zhang and Gerstein 2003a), it is clear that the pseudogenes generated by the somatic cytochrome c gene can be clustered into two classes. The younger three pseudogenes of Class I likely had a progenitor functional gene identical in amino acid composition to the modern gene, while the older pseudogenes of Class II give a remarkably consistent picture about their progenitor differing from the modern gene in nine amino acid positions. Relying on the strong consensus in the pseudogene sequences, we constructed a plausible progenitor DNA sequence (see Supplementary Material), and used it to compute divergence times that are less affected by the functional gene sequence evolution. We used thus two gene sequences: HCS (human

[3]In these initial experiments, we fixed the gene branch length at $t = 0.01$. We found that λ_Ψ and μ_Ψ are difficult to optimize for, due to the fact that the sequences are relatively short, and that a number of them have no obvious indels with respect to the gene sequence. The optimization for κ_Ψ seemed more stable, with an average of 1.44 and a standard deviation of 0.68. For about two thirds of the pseudogene sequences, the optimum was reached with κ_Ψ between 1 and 2, and only four of them had an optimum κ_Ψ above 2. Based on the observed values, we used the medians in subsequent experiments.

[4]See `http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=Homo+sapiens+[gbpri]`.

somatic cytochrome c), and preHCS (hypothetical progenitor of Class II pseudogenes) that differs in nine non-synonymous and three synonymous codon substitutions from HCS. By maximizing the likelihood of alignment between HCS and preHCS, we found that $\omega = 0.8$ and $t_\Psi = 0.14$ are the "correct" values for the alignment of Class II pseudogenes. We carried out pseudogene branch length optimization in the following setups.

[**OPT1**] alignment with progenitor; gene branch length fixed at $t = 0.01$

[**OPT2**] alignment with modern gene; t and ω are fixed at their "correct" values

[**OPT3**] alignment with modern gene; t and ω are optimized for

In addition, we computed the sequence divergence by Kimura's two parameter (K2P) model (Kimura 1980) between the gene and pseudogene sequences from each pairwise global alignment. Alignments were obtained by using the `align` program in the Fasta v2 suite (Pearson and Lipman 1988; Myers and Miller 1988). Figure 5 plots the pseudogene branch length estimates. Variance estimates of the K2P distance range from 0.01 (distances ≤ 0.06) to 0.04 (distances around 0.23). Based on simulations, maximum likelihood estimates have similar uncertainty if indels are excluded. For example, if the true pseudogene branch length $t_\Psi = 0.05$, then in 95% of the cases, the branch estimate is in the range 0.03–0.08 for OPT1.

There are some consistent patterns that can be observed. Not surprisingly, K2P estimates decrease for Class II pseudogenes when preHCS is used in the alignment. There are slight changes in most cases when likelihood optimization OPT1 is used with our sequence evolution model (partially due to the 0.01 divergence attributed to the gene branch). In some cases, however, when the alignment has a large number of indels, the ML estimate is significantly larger than the K2P distance, since the latter does not include indels in the distance calculation. The most prominent such cases are hcp12, hcp44, and hcp26. When the changes on the gene branch are uncertain (OPT2), the divergence time estimates become more variable, depending on the amount of nonsynonymous-synonymous codon differences between the sequences, and clear indications of non-functionality. When even the gene branch length is optimized for (OPT3), the differences are distributed between the two branches, depending again on the amount of indications for non-functionality. For instance, the hcp1 and hcp23 sequences have no frameshifts and stop codons, and the optimization assigns most of the changes to the gene branch, resulting in the underestimation of divergence times. Generally, the pseudogene branch length is overestimated for Class II pseudogenes, as the mutation in the gene sequence are accounted for on the pseudogene branch in the optimization. It is notable that for most Class II pseudogenes, the complete optimization (OPT3) found a large ω, indicating positive selection. In a rough calculation, $\omega > 0.5$ results in a non-synonymous substitution rate that exceeds the synonymous substitution rate in Eq. (9), since changes in the first two codon positions typically alter the amino acid, while the third one does not. After plugging in κ and the static codon probabilities, we calculated this threshold more carefully to get $\omega > 0.38$. The ML-OPT3 optimization settled on an ω above 0.76 for more than two thirds of the Class II pseudogenes.

[Figure 5 about here.]

Calculated divergence times can sometimes be validated by analyzing conserved syntenies: if a pseudogene sequence is in a conserved syntenic region with another organism, then the divergence time can be directly compared to the date of the split. For example, we verified that pseudogene hcp9, which was postulated to be a disabled ortholog of rodent testis-specific cytochrome *c* genes (Zhang and Gerstein 2003a), does fall into a syntenic region (see Supplementary Material) with that gene in the mouse genome (Karolchik et al. 2003; Blanchette et al. 2004). The pseudogene hcp14 also falls into a syntenic region with the mouse, rat, and dog genomes (see Supplementary Material). This pseudogene had the largest divergence time estimates in our experiments: 0.44 ± 0.07 by K2P and 0.51 by the ML-OPT1 method. The next oldest pseudogenes hcp5 and hcp38 do not appear in conserved syntenies with the mouse, rat, and dog genomes.

3.3 Ancestral codons and a molecular clock

We calculated ancestral codons plugging the divergence times and our alignments into Equation (11). We wanted to map the changes to a real time scale, for which we needed a molecular clock.

A molecular clock can be calibrated using the following time points.

(a) All pseudogene sequences that we looked it, including the most recent hcp21, have homologs in syntenic regions with the chimpanzee genome (see Supplementary Material).

(b) The Class I pseudogenes hcp15, hcp21 and hcp45 do not appear in conserved syntenies in the preliminary assembly of the rhesus macaque genome (see Supplementary Material).

(c) The Class I pseudogene hcp15 is disrupted into two fragments by an AluY insertion that has a 5.8% divergence from the consensus.

(d) Class II pseudogenes seem to predate the Old World monkey-New World monkey (OWM-NWM) split, since the spider monkey somatic cytochrome *c* agrees with the modern human amino acid sequence in six positions out of the nine amino acid differences between the progenitors of Class I and Class II pseudogenes. (See Figure 7.)

For a real time scale, we used the dates of 6 Mya, 25 Mya, and 40 Mya for the last common ancestor (LCA) of human-chimpanzee, human macaque, and OWM-NWM, respectively (Goodman 1999; Goodman et al. 1998); AluY sequences are about 19 million years old (Kapitonov and Jurka 1996).

The constraints do not imply an obvious molecular clock. (See Supplementary Material for illustration.) Graur and Li (2000) cite a rate of $3.9 \cdot 10^{-9}$ that was calculated using human-murid sequence comparisons, which is compatible with the calibration points (a–c) but puts Class II pseudogenes too early. It seems plausible that the rate for our pseudogenes is lower than that. Li (1997) gives [Table 8.5] a rate of around $1.5 \cdot 10^{-9}$ for intronic regions

that applies for human-OWM comparisons. We used this latter value to project pseudogene branch lengths to a real time scale, mostly for illustrative pusposes only. It is interesting to notice that the youngest Class II pseudogenes have branch lengths comparable to or even lower than those of Class I pseudogenes, although the two classes are separated by about 15 million years that passed between the splits with New World monkeys and Old World monkeys. The numerical values are not that curious if one takes into account the large variance due to the shortness of the sequences, uncertainties of calibration points, and fluctuations in substitution rates in different parts of the genome. The closeness of the branch estimates, in any case, pinpoints the rapid evolution of the cytochrome *c* gene. For that matter, the speed of gene evolution can be assessed based on the very fact that no pseudogene recorded an intermediate stage (with the possible exception of codon 58 in hcp46, see Figure 7), by estimating the time separating the two classes. Using the distribution for the ages of around 8000 retropseudogenes identified in the human genome (Zhang et al. 2003), we generated random samples of 46 pseudogenes in order to measure the maximum distance between generation times of Class I and Class II pseudogenes[5]. We found that the median distance was 0.0024, that in more than 95% of the cases, the distance between the oldest Class I and youngest Class II pseudogene was less than 0.01, and that in 99% of the cases, that distance was below 0.015. With a $1.5 \cdot 10^{-9}$ per year substitution rate, these distances translate to 1.6, 6.3 and 10 million years, respectively. In other words, the gene underwent eight or nine amino acid changes in a period of about 2–6 million years.

In order to compute the posterior probabilities for ancestral codons, we imposed a molecular clock on the gene branch. We used a 5% factor for codon evolution (other factors such as 1% and 10% give similar results), i.e., codon substitution rate on the gene branch was 5% of the nucleotide substitution rate on the pseudogene branch. Figure 6 shows the computed posterior probabilities[6] for codons that are present with a probability above 0.05 at some time.

[Figure 6 about here.]

[Figure 7 about here.]

The posterior probabilities corroborate previous hypotheses on cytochrome *c* evolution (Banci et al. 1999; Grossman et al. 2001). Namely, the gene underwent non-synonymous

[5]More precisely, we first computed the cumulative distribution function for the K2P distances between pseudogenes and their closest functional paralog provided by Zhang et al. (2003), in order to be able to generate random values following the age distribution of pseudogenes. In each round, we selected 46 independent random distances, so that the three smallest ones were below 0.05 (simulating Class I pseudogenes), and the largest one was below 0.25 (simulating Class II pseudogenes that are more recent than the primate-rodent split). We then collected statistics on the distance between the third and fourth smallest value in one million rounds.

[6]The underlying logic may appear circular, since the pseudogene branch lengths are based on alignments with a hypothetical progenitor, which is in turn evidenced by the changes in the posterior probabilities. Posterior probabilities, however, give a similar picture even if the pseudogene branch lengths are calculated in the ML-OPT3 scheme. After an ML-OPT3 round, the hypothetical ancestral gene sequence was constructed, and then used in the ML-OPT1 scheme.

codon changes in nine codon positions, fairly recently, right around the split between Old World monkeys and hominids (apes and human). Figure 7 illustrates that, in most cases, homologous sequences from different species support the codon substitution histories told by pseudogenes. The hypothetical ancestor sequence preHCS has GCCGTT in codon positions 43–44, and, thus, the relatively frequent ATT in codon position 44 may be due to an elevated probability of G → A mutations in the CG dinucleotide at the codon boundaries. Interestingly, codon CTT has a high posterior probability in the progenitor of hcp21 (the youngest pseudogene), which accounts for an intermediate stage between the modern CCT and the ancestral GTT. Two of the presumed synonymous changes (codon 36 and codon 94) are very recent, a third (codon 30) is contemporaneous with the non-synonymous changes. For one, or possibly even two codons, (codon 36 and codon 30), back-substitutions are also observed. For instance, the chimpanzee ortholog differs in four codons from the human gene (accession numbers NM_018947.4 and AY268594.1). Among the four, codon 36 is TTC in chimpanzee, and it is TTT in human. The pseudogene sequences suggest that this is a recent change in the human sequence, which occurred after the split with the chimpanzee. The mouse ortholog also has TTC in that position. This codon probably underwent two or three substitutions, as 17 older pseudogenes also have TTT in that position, and 14 younger ones have TTC. The ambiguity about the history of the 58-th amino acid (possibly Thr → Ile → Thr → Ile) becomes apparent only when OWM and NWM sequences are included, the pseudogenes on their own tell a similar history as that of other codons with one non-simultaneous change.

4 Discussion and future work

We described a sequence evolution model specific to pseudogenes and their protein-coding paralogs. The model can be used in conjunction with statistical alignment techniques, including parameter optimization, hypothesis testing, and alignment computation.

We showed how to calculate the homology likelihood, and the most likely alignment given the model parameters. The optimizable parameters include branch lengths, and synonymous/non-synonymous codon substitution rates, which are particularly interesting in the context of molecular evolution.

It seems impossible to estimate all the model parameters at the same time by comparing two sequences. This is a common feature of several substitution models, starting with the relatively simple Jukes-Cantor+Γ model (Nick Goldman, personal communication). The explanation of this feature is that the model has more parameters than the degree of freedom of the system. For example, the JC+Γ model has two parameters, while a gap-free alignment of two sequences has only one degree of freedom in this model, which is the percentage of mismatched columns. As a consequence, there is a continuous set of parameters (a ridge on the likelihood surface) that have maximum likelihood value, and the true evolutionary parameters cannot be recovered with likelihood maximization. If, however, one of the parameters is set to the true value, then the other parameter can be properly estimated. Even if the degree of freedom is slightly greater than the number of parameters, the ML estimation of all parameters might fail (Nick Goldman, personal communication), since different sce-

narios yielding the same output data might have almost the same likelihood under different parameter values. In our case, a substitution might be a result of a mutation either on the gene or on the pseudogene branch. When we fixed the value of ω estimated in a preliminary analysis, other substitution parameters could be estimated in a ML framework.

The degree of freedom (that is, the number of possible patterns in the sequence alignment that are equivalent with respect to the model) grows exponentially with the number of sequences, while the number of evolutionary parameters grows only linearly with the number of sequences. Consequently, the non-identifiability problem of parameters disappears as soon as more sequences are involved in the analysis. Our model can be naturally extended to many sequences. The running time and memory usage grow exponentially with the number of sequences in the analysis, hence dynamic programming becomes unfeasible for more than about four sequences. By restricting the search space heuristically (for example, bounding the number of indels), the algorithms can handle larger number of sequences (Hein et al. 2000). Another obvious solution is to employ Markov Chain Monte Carlo techniques, which have proven useful in a number statistical alignment problems (Holmes and Bruno 2001; Lunter et al. 2003; Lunter et al. 2005).

The experiments involving cytochrome c pseudogenes demonstrate that, in most cases, our maximum likelihood framework yields a reliable tool for alignment. The real value of the framework, however, is that it represents an important first step toward a statistically sound and biologically realistic approach to analyzing mixed data of coding and non-coding homologous sequences.

It should also be pointed out that retropseudogenes have features that cannot be captured by the alignment to the coding sequence only. The retrotranscribed mRNA sequences often include untranslated regions and poly(A) tails. It would be interesting to see how the current model can be extended to consider these and other features of retrotranscription.

We conclude with an open problem, relating to the results of Section 2.6. In our study, we separated the question of estimating divergence times from that of computing ancestral characters. Using similarities between the pseudogene sequences, however, one should be able to estimate pseudogene divergence times better than what can be achieved by comparing each pseudogene sequences to the gene sequence separately. In general, the problem of computing a phylogeny and ancestral sequences that together maximize the likelihood, is computationally intractable (Addario-Berry et al. 2004). Excluding segmental duplications, the comb-like evolutionary tree topology for a set of homologous retropseudogenes is determined solely by the order of their age, and the corresponding ancestral maximal likelihood problem may be tractable. The problem is thus whether and how one can efficiently compute divergence times and ancestral sequences from a modern gene sequence and a set of its independently evolved pseudogene homologs.

5 Acknowledgements

This work has been supported by grants from from the Natural Sciences and Engineering Research Council of Canada and the Fonds québécois de la recherche sur la nature et les

technologies to M.Cs, and a György Békésy postdoctoral fellowship to I. M. We thank Nick Goldman for helpful discussions concerning parameter optimization. We are also grateful to Alan Harris, Andrew Jackson, and Aleksandar Milosavjevic for help with the macaque genome data.

Supplemental Material Supplemental Material is available in the document `http://www.iro.umontreal.ca/~csuros/pseudogenes/pseudo-ali-supplement.pdf`.

References

Addario-Berry, L., B. Chor, M. Hallett, J. Lagergren, A. Panconesi, and T. Wareham (2004). Ancestral maximum likelihood of evolutionary trees is hard. *Journal of Bioinformatics and Computational Biology 2*(2), 257–271.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research 25*(17), 3389–3402.

Balakirev, E. S. and F. J. Ayala (2003). Pseudogenes: Are they "junk" or functional DNA? *Annual Review of Genetics 37*(1), 123–151.

Banci, L., I. Bertini, A. Rosato, and G. Varani (1999). Mitochondrial cytochrome *c*: A comparative analysis. *J. Biol. Inorg. Chem. 4*, 824–837.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller (2004). Aligning multiple genomic sequences with the Threaded Blockset Aligner. *Genome Research 14*(4), 708–715.

Bustamante, C. D., R. Nielsen, and D. L. Hartl (2002). A maximum likelihood method for analyzing pseudogene evolution: Implications for silent site evolution in humans and rodents. *Molecular Biology and Evolution 19*(1), 110–117.

Coin, L. and R. Durbin (2004). Improved techniques for the identicfication of pseudogenes. *Bioinformatics 20*, i94–i100.

Evans, M. J. and R. C. Scarpulla (1988). The human somatic cytochrome c gene: Two classes of processed pseudogenes demarcate a period of rapid molecular evolution. *Proceedings of the National Academy of Sciences of the USA 85*(24), 9625–9629.

Felsenstein, J. and G. A. Churchill (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution 13*(1), 93–104.

Fleishman, S. J., T. Dagan, and D. Graur (2004). pAnt: A method for pairwise assessment of nonfunctionalization times of processed pseudogenes. *Molecular Biology and Evolution 20*(11), 1876–1880.

Goldman, N. and Z. Yang (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution 11*(5), 725–736.

Goodman, M. (1999). The genomic record of humankind's evolutionary roots. *American Journal of Human Genetics 64*, 31–39.

Goodman, M., C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C. P. Groves (1998). Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Molecular Phylogenetics and Evolution 9*, 585–598.

Gotoh, O. (2000). Homology-based gene structure prediction: Simplified matching algorithm using translated codon (tron) and improved accuracy by allowing long gaps. *Bioinformatics 16*(3), 190–202.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science 185*(4154), 862–864.

Graur, D. and W.-H. Li (2000). *Fundamentals of Molecular Evolution* (Second ed.). Sunderland, Mass.: Sinauer Associates, Inc.

Grossman, L. I., T. R. Schmidt, D. E. Wildman, and M. Goodman (2001). Molecular evolution of aerobic energy metabolism in primates. *Molecular Phylogenetics and Evolution 18*(1), 26–36.

Harrison, P. M., N. Echols, and M. B. Gerstein (2001). Digging for dead genes: An analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Research 29*(3), 818–830.

Harrison, P. M., H. Hegyi, S. Balasubramanian, N. M. Luscombe, P. Bertone, N. Echols, T. Johnson, and M. Gerstein (2002). Molecular fossils in the human genome: Identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Research 12*(2), 272–280.

Hasegawa, M., H. Kishino, and T. Yano (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution 22*, 160–174.

Hein, J. (2001). An algorithm for statistical alignment of sequences related by a binary tree. In R. B. Altman, A. K. Dunker, L. Hunker, K. Lauderdale, and T. E. Klein (Eds.), *Biocomputing: Proc. of the Pacific Symposium*, pp. 179–190. World Scientific Publishing.

Hein, J., J. L. Jensen, and C. N. S. Pedersen (2003). Recursions for statistical multiple alignment. *Proceedings of the National Academy of Sciences of the USA 100*(25), 14960–14965.

Hein, J., C. Wiuf, B. Knudsen, M. B. Møller, and G. Wibling (2000). Statistical alignment: Computational properties, homology testing and goodness-of-fit. *Journal of Molecular Biology 302*, 265–279.

Holmes, I. and W. J. Bruno (2001). Evolutionary HMMs: A Bayesian approach to multiple alignment. *Bioinformatics 17*(9), 803–820.

Holmes, I. and G. M. Rubin (2002). An expectation maximization algorithm for training hidden substitution models. *Journal of Molecular Biology 317*, 753–764.

IHGSC (2004). Finishing the euchromatic sequence of the human genome. *Nature 431*, 931–945.

Kapitonov, V. and J. Jurka (1996). The age of Alu subfamilies. *Journal of Molecular Evolution 42*, 59–65.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent (2003). The UCSC genome browser database. *Nucleic Acids Research 31*(1), 51–54.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution 16*, 116–120.

Li, W.-H. (1997). *Molecular Evolution*. Sunderland, Mass.: Sinauer Associates.

Li, W.-H., T. Gojobori, and M. Nei (1981). Pseudogenes as paradigm of neutral evolution. *Nature 292*(5820), 237–239.

Lunter, G., A. J. Drummond, I. Miklós, and J. Hein (2005). Statistical alignment: Recent progress, new applications, and challenges. In R. Nielsen (Ed.), *Statistical Methods in Molecular Evolution*, Chapter 14. Heidelberg: Springer-Verlag. (Forthcoming.).

Lunter, G., I. Miklós, A. Drummond, J. Jensen, and J. Hein (2003). Bayesian phylogenetic inference under a statistical indel model. In *Proc. Workshop on Algorithms in Bioinformatics (WABI)*, Volume 2812 of *LNCS*, pp. 228–244. Springer-Verlag.

Lunter, G. A., I. Miklós, Y. S. Song, and J. Hein (2003). An efficient algorithm for statistical multiple alignment of arbitrary phylogenetic trees. *Journal of Computational Biology 10*(6), 869–889.

Mighell, A. J., N. R. Smith, P. A. Robinson, and A. F. Markham (2000). Vertebrate pseudogenes. *FEBS Letters 468*, 109–114.

Miklós, I. (2002). An improved algorithm for statistical alignment of sequences related by a star tree. *Bulletin of Mathematical Biology 64*, 771–779.

Muse, S. V. and B. S. Gaut (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution 11*(5), 715–724.

Myers, E. and W. Miller (1988). Optimal alignments in linear space. *Computer Applications in the Biosciences 4*, 11–17.

Pearson, W. R. and D. J. Lipman (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA 85*(8), 2444–2448.

Press, W. H., S. A. Teukolsky, W. V. Vetterling, and B. P. Flannery (1997). *Numerical Recipes in C: The Art of Scientific Computing* (Second ed.). Cambridge UniversIty Press.

Pupko, T., I. Pe'er, R. Shamir, and D. Graur (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution 17*(6), 890–896.

Schadt, E. and K. Lange (2002). Codon and rate variation models in molecular phylogeny. *Molecular Biology and Evolution 19*(9), 1534–1549.

Steel, M. and J. Hein (2001). Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. *Applied Mathematics Letters 14*, 679–684.

Thorne, J. L., H. Kishino, and J. Felsenstein (1991). An evolutionary model for maximum likelihod alignment of DNA sequences. *Journal of Molecular Evolution 33*, 114–124.

Torrents, D., M. Suyama, E. Zdobnov, and P. Bork (2003). A genome-wide survey of human pseudogenes. *Genome Research 13*(12), 2559–2567.

Vanin, E. F. (1985). Processed pseudogenes: Characteristics and evolution. *Annual Review of Genetics 19*, 253–272.

Yang, Z. and R. Nielsen (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution 46*(4), 409–418.

Yang, Z. and R. Nielsen (2000). Estimating synonymous and nonsynonymous rates under realistic evolutionary models. *Molecular Biology and Evolution 17*(1), 32–43.

Zhang, Z., N. Carriero, and M. Gerstein (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends in Genetics 20*, 62–67.

Zhang, Z. and M. Gerstein (2003a). The human genome has 49 cytochrome *c* pseudogenes, including a relic of a primordial gene that still functions in mouse. *Gene 312*, 61–72.

Zhang, Z. and M. Gerstein (2003b). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research 31*(18), 5338–5348.

Zhang, Z., P. M. Harrison, Y. Liu, and M. Gerstein (2003). Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Research 13*(12), 2541–2558.

Zhang, Z., W. R. Pearson, and W. Miller (1997). Aligning a DNA sequence with a protein sequence. In *Proc. First Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 337–343. ACM Press.

List of Figures

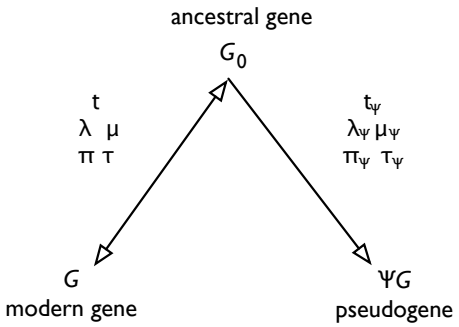

Figure 1: Evolution of a processed pseudogene ΨG and its paralog G from a common ancestor G_0. On the $G_0 - G$ branch of length t, the TKF birth-death process is characterized by λ and μ, and a codon substitution model is employed with stationary distribution π and transition probabilities τ. On the pseudogene branch of length t_Ψ, the birth-death process has parameters λ_Ψ and μ_Ψ, and a nucleotide substitution model is employed.

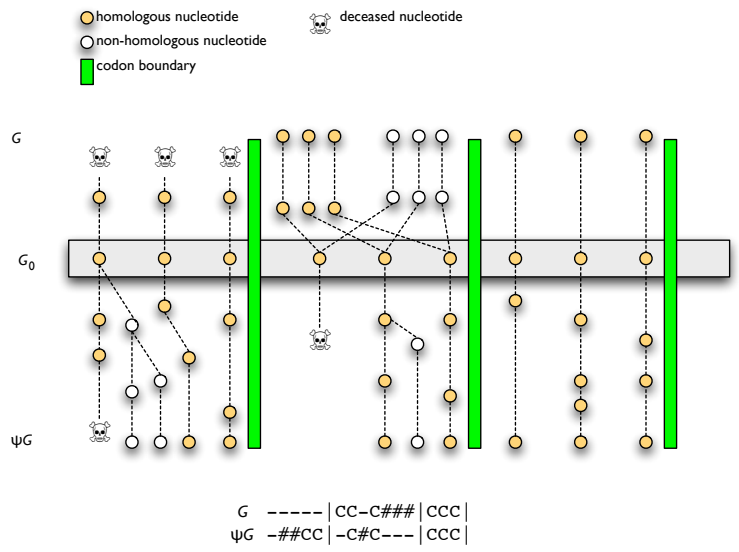

Figure 2: Example of an evolutionary history and the corresponding alignment.

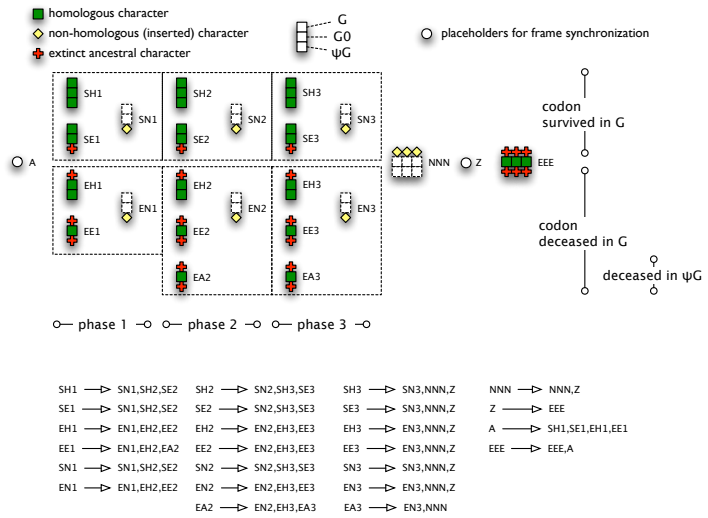

Figure 3: Recursions after a mortal codon link that has at least one modern descendant. Variables EA are introduced to force the inclusion of at least one modern character on the path $\mathsf{A} \rightarrow \mathsf{Z}$. Variables on the left-hand side of the arrows appear as terms in the recursions for the variables on the right-hand side. Multipliers on the arrows are not shown.

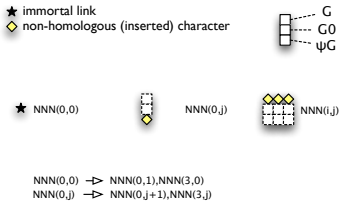

Figure 4: States for the immortal link

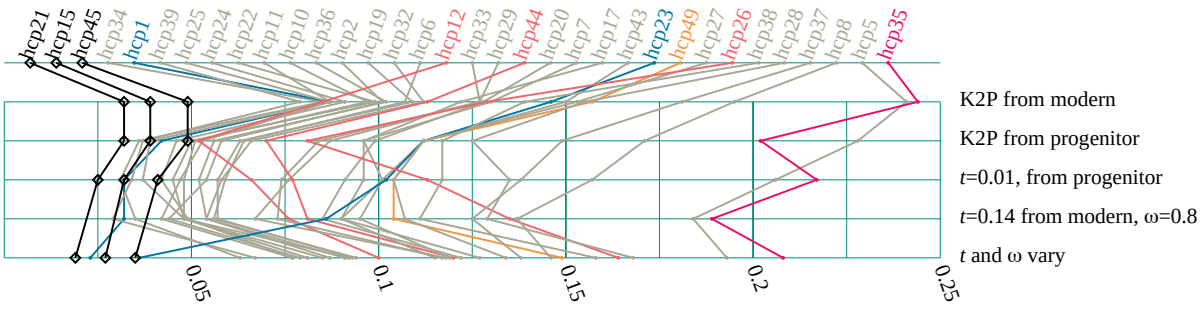

Figure 5: Pseudogene branch lengths computed by different methods. The first two values are Kimura's two parameter distance from HCS and preHCS, respectively. The lower three values are obtained through likelihood maximization in our model (OPT1, OPT2, and OPT3, respectively).

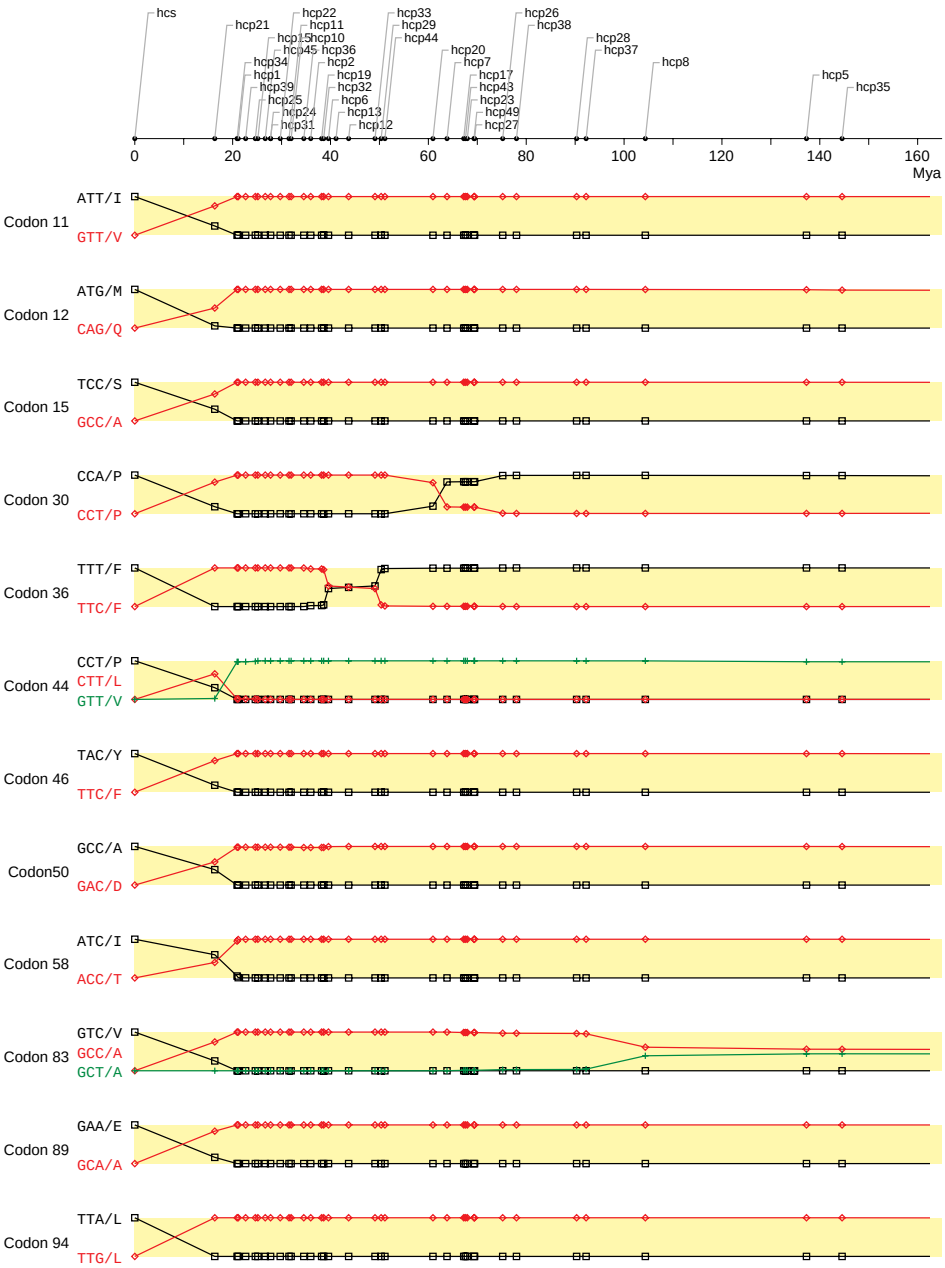

Figure 6: Posterior probabilities for codons in cytochrome *c* sequence evolution. Each box shows the probabilities for codons that have a non-negligible probability at some point.

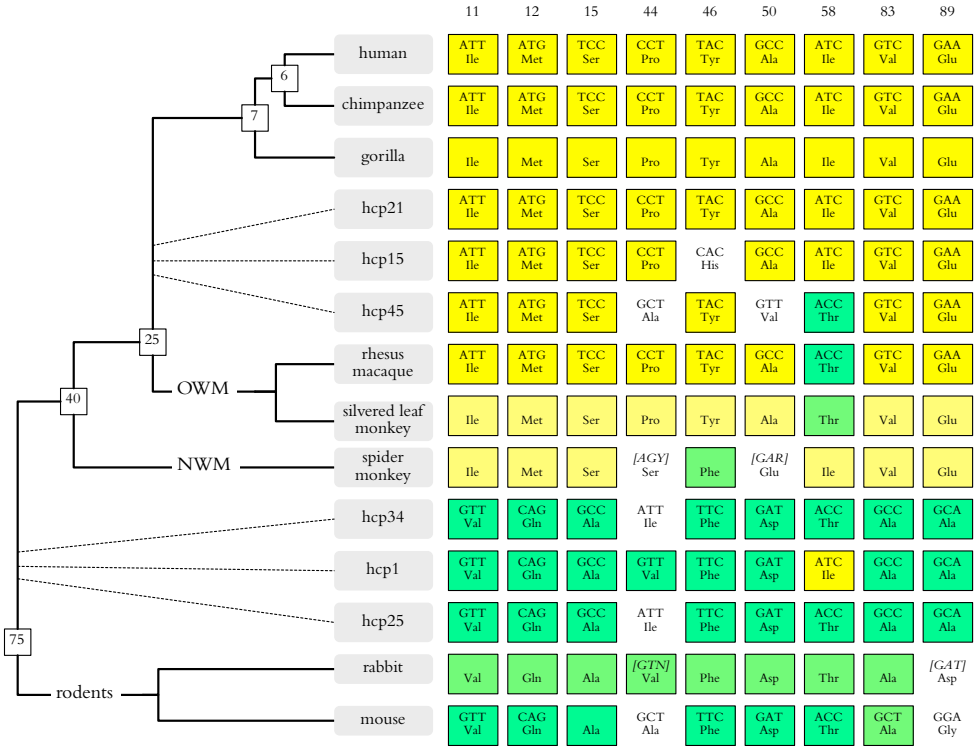

Figure 7: Homologous codons in different cytochrome *c*-related sequences. Numbers at internal nodes give approximate dates for the splits in million years.