Fitting Models of Intron Evolution to
Aldehyde Dehydrogenase Data

Andrey Rzhetsky, Francisco José Ayala, Lily C. Hsu, Cheng Chang,
and Akira Yoshida

ABSTRACT. Whether or not nuclear introns predate the divergence of bacteria and eukaryotes is the central argument between the proponents of the "introns-early" and "introns-late" theories. In this study we compared the "goodness-of-fit" of each theory using a probabilistic model of exon/intron evolution and new genomic sequences of non-allelic genes encoding human aldehyde dehydrogenases (ALDH). Using a reconstructed phylogenetic tree of ALDH genes we computed the likelihoods of obtaining the present-day ALDH sequences under the assumptions of each competing theory. Although *on the grounds of their own assumptions* each theory fit the ALDH data significantly better than its rival, the model corresponding to the "introns-early" theory required extensive "intron slippage," and the estimated slippage rates were too high to be consistent with previously reported correlations between the boundaries of ancient protein modules and the ends of ancient exons. Arguing that the molecular mechanisms proposed for explaining intron slippage are incapable of providing such high slippage rates and are incompatible with the observed intron distribution in higher eukaryotes, we concluded that the ALDH data support the "introns-late" theory.

1. Introduction

The "introns-early" theory suggests that the "genes in pieces" structure of eukaryotic genes emerged long before the Eubacteria, Archaebacteria, and Eukaryota diverged as separate groups (1, 2, 3). According to this theory, (*i*) the present-day exon/intron structures originated through the aggregation of short primordial mini-genes (15-20 amino acids) which were critically important for generating protein diversity through "exon shuffling," (*ii*) the apparent absence of spliceosomal introns in bacterial and organelle genomes resulted from their secondary loss, and (*iii*) the nuclear splicing machinery is as ancient as are the nuclear introns themselves. Furthermore, the theory presumes that introns can be easily lost and an "intron slippage" mechanism exists which

can displace introns for short distances (1-12 nucleotides; see 4) while leaving the coding sequence intact.

The alternative "introns-late" theory (5, 6, 7) states that (*i*) split genes appeared by random intron insertion into primordial continuous protein-coding regions, (*ii*) the genes of cellular organelles and those of bacteria never had spliceosomal introns, and (*iii*) the spliceosomal machinery emerged through coevolution of group II self-splicing introns with eukaryotic proteins (8, 6, 9). Although the "introns-late" theory denounces the shuffling of primordial exons, it does not deny neither the possibility of recent exon shuffling within eukaryotic lineages nor early protein evolution by fusion, duplication, and permutation of primordial protein modules. However, the "introns-late" theory does not permit intron slippage and thus regards all introns occupying different sites within related proteins as non-homologous.

The following lines of argument have been used to either support or reject the two theories. (*i*) A few introns were found in homologous positions in genes duplicated *before* the separation of eukaryotes and bacteria (supporting "introns-early") (10), although the distribution of the vast majority of introns in such genes seems to be better explained by intron insertion (9). (*ii*) "Introns-early" supporters correctly predicted the position of a new intron in a gene of mosquito *Culex tarsalis* (11), although this was later argued to be a lucky coincidence (12, 13, 14). (*iii*) "Introns-early" supporters have claimed that exon/intron boundaries statistically correlate with the ends of units of protein three-dimensional structure (ancient "modules," *e.g.*, see 15), although this conclusion was also vigorously challenged (14, 16). (*iv*) Multigene analyses of the distribution of intron phase indicated a significant excess of exons and exon groups with the same intron phase at both ends (which was presented as evidence for the "introns-early" theory; 17, 18), but this could have resulted from recent exon shuffling events, and is thus compatible with both theories (13). (*v*) Parsimonious reconstructions of the evolution of the exon/intron structure in eukaryotes supported the "introns-late" view (9, 14) but the possibility of intron slippage was completely discarded in these analyses.

We present here a new method aimed at (*i*) qualitatively analyzing new data under the respective assumptions of the two competing theories, (*ii*) scrutinizing the internal consistency of the results of each analysis, and (*iii*) evaluating factual support for the assumptions underlying each theory. We illustrate the application of this new method with an analysis of aldehyde dehydrogenase (ALDH) genes.

2. Human ALDH genes

2.1. Human ALDH genes are ancient. Aldehyde dehydrogenases are enzymes catalyzing the conversion of biogenic and foodstuff aldehydes into acid metabolites (19, 20, 21). Humans have at least ten homologous ALDH genes that apparently emerged from a series of duplications of a single ancestral gene (22, 23, 24, 25, 26, 27, 28, 29), and which have surprisingly diverse exon/intron structures (fig. 1, 2). Although all known human ALDH's are nuclearly encoded, at least three of them (ALDH2, ALDH5, and methylmalonate-semialdehyde dehydrogenase, MMSDH) have leader peptides and are transported to the mitochondria after synthesis.

A neighbor-joining tree of *ALDH*-like sequences from several eukaryotic and prokaryotic species yielded four well-defined clusters of eukaryotic genes (fig. 3).

```
                                    0  0                    0  0
            1     11    21    31   41    51    61    71   81    91
1  ----------------MSSSGTPDLPVLLTDLKIQYTKIFINNEWHDSVSGKKFPVFNPATEEELCQVEEGDKDVDKAVKAARQAFQIGSPW-RTMDAS
6  -----MATANGAVENGQPDGKPPALPRPIRNLEVKFTKIFINNEWHESKSGKKFATCNPSTRBQICEVEEGDKPDVDKAVEAAQVAFQRGSPW-RRLDAL
2  MLRAAARFGPRLGRRLLSAAATQAVPAPNQQPEVFCNQIFINNEWHDAVSRKTFPTVNPSTGEVICQVAEGDKEDVDKAVKAARAAFQLGSPW-RRMDAS
5  MLRFLAPRLLSLQGRTALYSSAAALPSPILNPDIPYNQLFINNEWQDAVSKKTFPTVNPTTGEVIGHVAEGDRADVDRAVKAAREAFRLGSPW-RRMDAS
3  ----------------------------------------------------------------------MSKISEAVKRAPAAFSSGRTRPLQFRIQ
10 ----------------------------------------------------------------------MELEVRRVRQAFL---SGRSRPLRFRLQ
7  ----------------------------------------------------------------------MDPLGDTLRRLREAFHAGRTRPAEFRAA
8  ----------------------------------------------------------------------MDPFEDTLRRLREAFNAGRTRPAEFRAA
9  -----------------------MSTGTFVVSQPLNYRGGAAGAGGRSGTEKAFEPATGRVIATFTCSGEKEVNLAVQNAKAAFKI---W-SQKSGM
```

```
               0  0   0        2       1   1    2       A   0  0
      101    111   121   131   141   151   161   171   181   191
1  ERGRLLYKLADLIERDRLLATMESMNGGKLYSNAYLNDIAGCIKTTRYCAGWADKIQGRTIPIKQNFFTYTRHEPIGVCGQIIFPWNFPLVMLIWKIGPA
6  SRGRLLHQLADLVERDRATLAALETMDTGKPFLHAFFIDLEGCIRTLRYFAGWADKIQGKTIPTDDNVVCFTRHEPIGVCATTPWNFPLIMLVWKLAPA
2  HRGRLLNRLADLIERDRTYLAALETLDNGKPYVISYLVDLDMVLKCLRYYAGWADKYHGKTIPIDGDFFSYTRHEPVGVCGQIIPWNFPLLMQAWKLGPA
5  ERGRLLNLLADLVERDRVYLASLETILDNGKPFQESYALDLDEVIKVYRYFAGWADKWHGKTIPMHGQHFCFTRHEPVGVCGQIIPWNFPLVMQGWKLAPA
3  QLEALQRLIQBQBQELVGALAADLHKNEWNAYYEEVVYVLEEIEYMIQKLPEWAADEPVEKTPQTQQDELYIHSEPLGVLVIGTWNYPFNLTIQPMVGA
10 QLEALRRMVQEREKDILTAIAADLCKSEEFNVYSQEVITVLGEIDFMLENLPEWVTAKPVKKNVLIMLDEAYIQPQPLGVLIIGAWNYPFVLTIQPLIGA
7  QLQGLGRFLQENKQLIHDALAQDLHKSAFESEVSEVAISQGEVTLALRNLRAWMKDERVPKNLATQLDSAFIRKEPFGVLIIIAPWNYPLNLTLVPLVGA
8  QLQGLGHFLQENKLLRDVLAQDLHKTAFEADISELILCQNEVDYALKNLQAWMKDEPRSTNLFMKLLDSVFIWKEPFGLVLIIAPWNYPLNLTLVLLVGT
9  ERCRILLEAARIIRERENDEIATMECINNGKSIFEARL-DIDISWQCLEYYTGLAASMAGEHIQLPGGSFGYTRREPLGVCVGIGAWNYPFQIASWKSAPA
```

```
       1                        0            0  0  0            1        0    0
      201    211   221   231   241   251   261   271   281   291
1  LSCGNTVVVKPAEQTPLTALHVASLIKEAGFPPGVVNIVPGYGPTAGAAISSHMDIDKVAFTGSTIEVGHLIKEAAGKSNLKKVTLELGCKSPCIVLADAD
6  LCCGNTMVLKPAEQTPLTALYLGSLIKEAGFPPGVVNIVPGFGPTVGAAISSHPQINKIAFTGSTIEVGKLVKEAASRSNLKRVTLELGCKNPQIVCADAD
2  LATGNVVVMKVAEQTPLTALYVANLIKEAGFPPGVVNIITGYGPTAGAAIAQHMDVDKVAFTGSTIEVGHLIQKAAGDSNLKRVTLELGCKSPSIVLADAD
5  LATGNTVVMKVAEQTPLSALYLASLIKEAGFPPGVVNIITGYGPTAGAAISHEDVDKVAFTGSTIEIGRVIQVAAGSSNLKRVTLELGCKSPNIIMSDAD
3  IAAGNSVVLKPSELSENMASLLATIIPQYLDKDLYPVINGGVPETTELLKER---FDHILYTGGTGKI IMTAAAK-HLTPVTLELGCKSPCYVDKNCD
10 IAAGNAVIIKPSELSENTAKILAKLLPQYLDQDLYIVINGGVPETTELLKQR---FDHIFYTGTIASCIKVMEAAAK-HLTPVTLELGCKSPCYIDKDCN
7  LAAGNCVVLKPSELSKNVEKILAEVLPQYVDQSCFAVVLGGPQETGQLLEHR---FDYIFFTGSPHRVGKIVMTAAAK-HLTPVTLELGCKNPCYVDDNCD
8  LPAGNCVVLKPSELSQGTEKVLAEVLPQYLDQSCFAVVLGGPQETGQLLEHK---LDYIFFTGSPHRVGKIVMTAATK-HLTPVTLELGCKNPCYVDDNCD
9  LACGNAMVFKPSPFTPVSALLLAEIYSEAGVPPGLFNVVQG-GAATGQFLCQHPDVAKVSFTGSVPEVGKMIME-MSAKGIKPVTLELGCKSPLIIFSDCD
```

```
   1  11  12            D           0  0            0        2  0  1            0
      301    311   321   331   341   351   361   371   381   391
1  LLNAVEFAHHGVFYHQGQCCIAASRIFVEESIYDEFVRRSVERAKKYIILGNPLTPGVTQGPQIDKEQYDKILDLLIESGKKEGAKLECGGGPWGN------
6  LLDLAVECAHQGVFFNQGQCCTAASRVFVEEQVYSEFVRRSVEYAKKRPVGDPFDVKTBQGPQIDQKQFDKILELIESGKKEGAKLECGGSAMED------
2  MDWAVEQAHFALFFNQGQCCAGSRTFVQEDIYDEFVERSVARAKSRVVGNPFDSKTBQGPQVDETQFKKILGYINTGKQEGAKLLCGGGIAAD------
5  MEHAVEQCHEALFFNMGQCCCAGSRTFVEESIYNEFLERTVEKAKQRKVGNPFELDTQQGPQVDKEQFERVLGYIQLGQKEGAKLLCGGERFGE------
3  LDVACRRIAWGKFMNSGQTICVAPDYIILCDPSIQNQIVEK-LKKSLKEFYGEDAKKSRDYGRIISARHFQRVMGLIE-QGK-----VAYGGTGDA------
10 LDIVCRRTITWGKYMNCGQTICIAPDYILCEASLQNQIVWK-IKETVKEFYGENIKESPDYERIINLRHFKRILSLLE-QGK-----IAFGGETDE------
7  PQTVANRVAWFRYFNAGQTICVAPDYVLCSPEMQERLLPA-LQSTITRFYGDDPQSSPNLGRIINQKPFQRLRALLG-CGR-----VAIGGQSDE------
8  PQTVANRVAWFCYFNAGQTICVAPDYVLCSPEMQERLLPA-LQSTITRFYGDDPQSSPNLGRIINQKPFQRLRALLG-CGR-----VAIGGQSNE------
9  MNNAVKGALMANFLTQGQVCCNGTRVFVQKEILDKFTEEVVKQTQRIKIGDPLLEDTRMGPLINRPHLERVLGFVKVAKEQGAKVLCGGDIYVPEDPKLK
```

```
      2  1              0  0  0   0                    0            2  2
      401    411   421   431   441   451   461   471   481   491
1  KGYFVQPTVFSNVIDEMRIAKEEIFGPVQQIMKFKSLDDVIKRANNTFYGLSAGVFTKDILKAITISSALQAGTVWVNCYGVVSAQC--PFGGFKMSGNG
6  KGLFIKPTVFSEVIDNMRIAKEEIFGPVQPILKFKSIEEVIKRANSTDYGLTAAVFTKNLDKALKLASALESGTVWINCYNALYAQA--PFGGFKMSGNG
2  RGYFIQPTVFGDVQDGMTIAKEEIFGPVMQILKFKTIEEVVGRANNSTDYGLAAAVFTKDLDKANYLSQALQAGTVWVNCYDVFGAQS--PFGGYKMSGSG
5  RGFFIKPTVFGGVQDDMRIAKEEIFGPVQPLFKFKKIEEVVERANNTRYGLAAAVFTKDLDKAMYFTQALQAGTVWVNTYNIVTCHT--PFGGFKESGNG
3  ATRYIAPTLITDVDPQSPFVMQEEIFGPVLPIVCVRSLEEAIQFINQREKPLALYMFSSNDKVIKKMIAETSSGGVAANDVVHITLHSLPFGGVGNSGMG
10 ATRYIAPTVLTDVDPKTKVMQEEIFGPILPIVPVKNVDEAINFINEREKPLALYVFSHNHKLIKRMIDETSSGGVTGNDVIMHFTLNSFPFGGVGSSGMG
7  SDRYIAPTVLVDVQEMEPVMQEEIFGPILPIVNVQSLDEAIEFINRREKPLALYAFSNSSQVVKRVLTQTSSSGFCGNDGFMHMTLASLPFGGVGASGMG
8  SDRYIAPTVLVDVQETEPVMQEEIFGPILPIVNVQSVDEVIKFINRQEKPLALYAFSNSRQVVNQMLERTSSSGFGGNBGFTYISLLSVPFGGVGHSGMG
9  DGYYMRPCVLTNCRDDMTCVKEEEIFGPVMSILSFDTEAEVLERANDTTFGLAAGVFTRDIQRAHRVVAELQAGTCFINNYNVSPVE--LPFGGYKKSGFG
```

```
   12  1  1                            2  0  0  0
      501    511   521   531   541   551   561   571
1  RELGEYGFHEYTEVRTVTVKISQKNS--------------------------------------------------
6  RELGEYALAEYTEVKTVTIKLGDKNP--------------------------------------------------
2  RELGEYGLQAYTEVKTVTIVKVPQKNS-------------------------------------------------
5  RELGEDGLKAYTEVKTVTIKVPQKNS--------------------------------------------------
3  SYHGKKSFETFSHRRSCLVRPLMNDEGLKVRYPPSPAKMTQH----------------------------------
10 AYHGKHSFDTFSHQRPCLLKSLKREGANKLRYPPNSQSKVDWGKFFLLKRFNKEKLGLLLLTFLGIVAAVLVKAEYY
7  RYHGKFSFDTFSHHRACLLRSPGMEKINALRYPPQSPRRLRMLLVAMEAQGCSCTLL-------------------
8  RYHGKFTFDTFSHHRTCLLAPSGLEKLKEIRYPPYTDWNQQLLRWGMGSQSCTLL--------------------
9  RENGRVTIEYYSQLKTVCVEMGDVESAF------------------------------------------------
                    0
```

Fig. 1. Positions of introns within eight human genes superimposed with alignment of corresponding protein sequences. Intron positions (arrowheads) for ALDH1/2/5/6/10 and ALDH3/7/8/9 groups are shown above and below the alignment, respectively. The number shown next to each arrowhead (0, 1, or 2) indicates the "phase" of corresponding intron with respect to the reading frame (0 - exon/intron boundaries are between codons, 1 - after the first codon position, and 2 - after the second nucleotide in the codon). Open boxes emphasize conservative sites; closed boxes show amino acid sites with established function. Although we were able to find alternative plausible alignments, all of them predicted the same relative arrangement of introns.
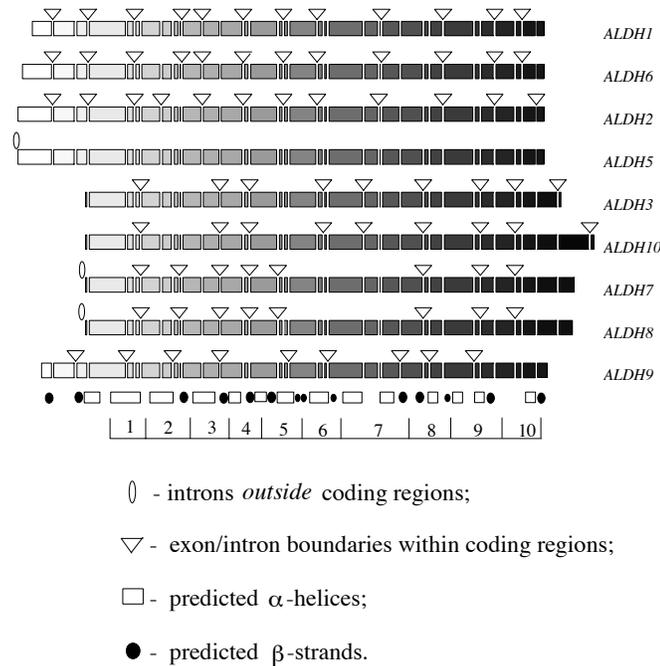
Fig. 2. Exon/intron structures of human ALDH genes mapped to the alignment of their amino acid sequences. The sequences themselves are shown as shaded rectangles where the shading intensity increases in direction from N- to C-terminus of the protein; deletions and insertions are not shown. Each discontinuity in rectangles corresponds to an exon/intron boundary observed in at least one of the genes; triangles and ellipses indicate only those introns that are actually found in the corresponding gene. The figure also shows the predicted secondary structure that is assumed to be similar for all compared proteins: the hatched boxes indicate α-helices and the filled circles correspond to β-strands. The secondary structure was predicted with a neural network algorithm implemented in program PHD (30, 31). The bottom of the figure shows an artificial segmentation of the protein into ten domains which was used in computation of the likelihood values.

Although the average substitution rate in the group IV cluster (*ALDH3/7/8/10*) was twice as large as the rate in the group I cluster (*ALDH1/2/5/6*), each of the two groups *separately* conformed to a "molecular clock" (see figs. 3 and 5 A) allowing an estimation of the divergence times between genes (see 32, 33). These estimates (fig. 5 A) indicated that duplications in group I were likely to have occurred much earlier than the duplications in group IV. In our reconstruction, diversification within group I happened during the Neoproterozoic period (34), while duplications in group IV seemed to arise much later, in the Phanerozoic period, when diverse vertebrate and invertebrate animals were already abundant. The latest two duplications in group IV probably took place near 212 and 87 million years ago, respectively (fig. 5 A), dates which roughly correspond to the appearance and then subsequent radiation of mammals. Finally, the existence of at
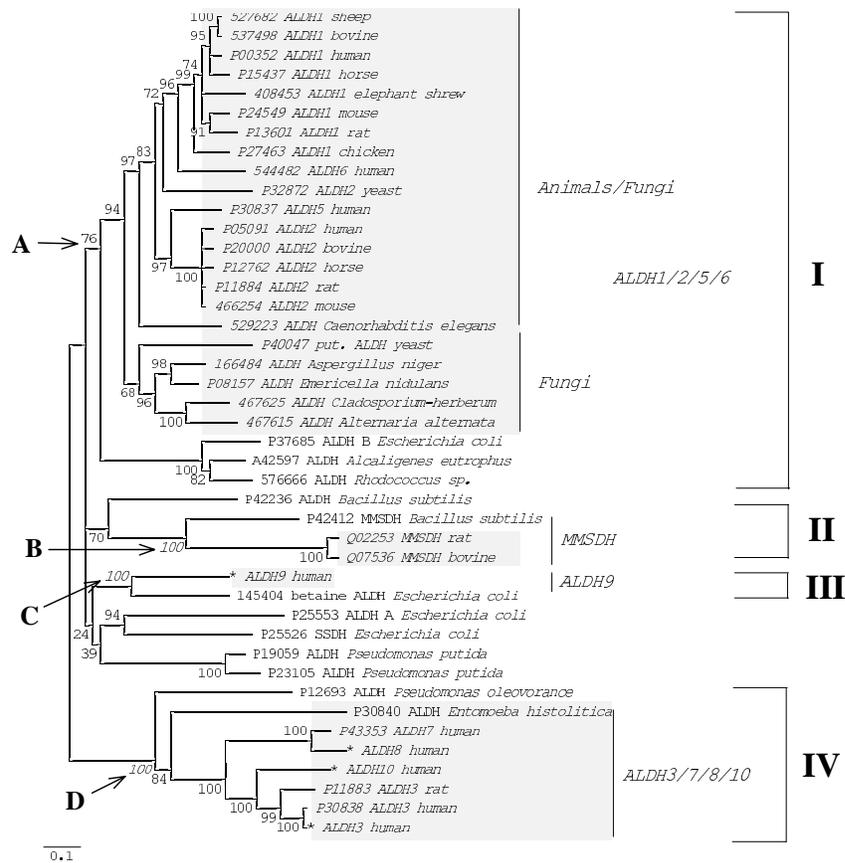
Fig. 3. A neighbor-joining tree (35) computed and visualized with MEGA (36) from 43 ALDH-like protein sequences using the Poisson correction (37, for ALDH data virtually all currently available corrections for multiple hits give essentially the same tree) for multiple hits; the branch lengths are given in terms of the number of amino acid substitutions per site. All sites with deletions or insertions were excluded from the analysis. The shaded areas indicate sequences from eukaryotic organisms. We deliberately excluded plant ALDHs from the analysis to facilitate interpretation of the resulting phylogeny. Bootstrap *p*-values are shown next to the corresponding interior branches; the interior branches which were supported with 20% or less out of 500 bootstrap replications (39) were set to zero. Description of each protein sequence includes either SwissProt or GenBank accession number (asterisks indicate new sequences) and protein and species names. **A**, **B**, **C**, and **D** indicate the interior branches defining four stable clusters of proteins from both eukaryotes and eubacteria. The tree may indicate that at least four ALDH-like genes (*ALDH1/2/5/6*-like, *ALDH3/7/8/10*-like, *ALDH9*-like, and *MMSDH*-like genes) pre-existed the divergence of eukaryotes and eubacteria. SSDH stands for succinate-semilaldehyde dehydrogenase.

Q02252 MMSDH human
Q02253 MMSDH rat
Q07536 MMSDH bovine
MMSDH
P42412 MMSDH *B. subtilis*
1788042 ALDH *E. coli*
913941 BADH *Brassica napus*
P46562 ALDH (put.) *C. elegans*
1353248 ALDH4 human
P42236 ALDH (put.) *B. subtilis*
P39634 ALDH *B. subtiis*
ALDH4
P33008 ALDH (put.) *Pseudomonas sp.*
1742508 GABDH *E. coli*
P43503 ALDH *P. putida*
1790871 ALDH *Comamonas testosteroni*
P25553 ALDHA *E. coli*
556221 ALDH11 human
ALDH11
P25526 SSDH *E. coli*
P38067 ALDH (put.) yeast
P23883 ALDH (put.) *E. coli*
P19059 ALDH *Pseudomonas putida*
P23105 ALDH *Pseudmonas putida*
587110 ALDH *E. coli*
A42597 ALDH *Alcaligenes eutrophus*
P23240 ALDH *Vibrio cholerae*
576666 ALDH *Rhodococcus sp.*
1790014 ALDH B *E. coli*
P37685 ALDHB *E. coli*
927643 BADH *Hordeum vulgare*
520546 BADH *Sorghum bicolor*
520544 BADH *Sorghum bicolor*
118492 BADH *Spinacia oleracea*
1813538 BADH *Spinacia oleracea*
17936 BADH *Beta vulgaris*
118490 BADH *Beta vulgaris*
17934 BADH *Beta vulgaris*
Plant betaine ALDH
166484 ALDH *Aspergillus niger*
P08157 ALDH *Emericella nidulans*
467625 ALDH *Cladosporium herberum*
467615 ALDH *Alternaria alternata*
1749700 ALDH *S. pombe*
P40047 ALDH (put.) yeast
P22281 ALDH1 yeast
P32872 ALDH2 yeast
529223 ALDH *C. elegans*
P11884 ALDH2 rat
466254 ALDH2 mouse
P05091 ALDH2 human
P20000 ALDH2 bovine
P12762 ALDH2 horse
P30837 ALDH5 human
1743354 ALDH *Nicotiana tabacum*
544482 ALDH6 human
408453 ALDH1 shrew
P27463 ALDH1 chicken
P13601 ALDH1 rat
P15437 ALDH1 horse
P00352 ALDH1 human
P24549 ALDH1 mouse
527682 ALDH1 sheep
537498 ALDH1 bovine
Fungus/Plant /Animal ALDH1/2/5/6
P46329 ALDH (put.) *B. subtilis*
P12693 ALDH *Pseudomonas oleovorance*
P30840 ALDH *Entomoeba histolitica*
P43353 ALDH7 human
ALDH8 human
ALDH3 human
P11883 ALDH3 rat
ALDH10 human
P30839 ALDH10 rat
ALDH3/7/8/10
ALDH9 human
118491 BADH *E. coli*
145405 BADH *E. coli*
145404 BADH *E. coli*
ALDH9
145402 choline ALDH *E. coli*
114919 choline ALDH *E. coli*

0.25

Fig. 4. A neighbor-joining tree computed with MEGA (36) and visualized with TREEVIEW program by K. Tamura from expanded sample of ALDH-like protein sequences including plant sequences. As in the previous figure, the Poisson correction (37) for multiple hits was applied; the branch lengths are given in terms of the number of amino acid substitutions per site. All sites with deletions or insertions were excluded from the analysis. Although this tree is considerably less stable than the one on the previous figure when tested with bootstrap (data not shown), it clearly shows that there is either fewer different non-allelic ALDH genes in plants than in animals, or many of the plant ALDH genes are not discovered yet.
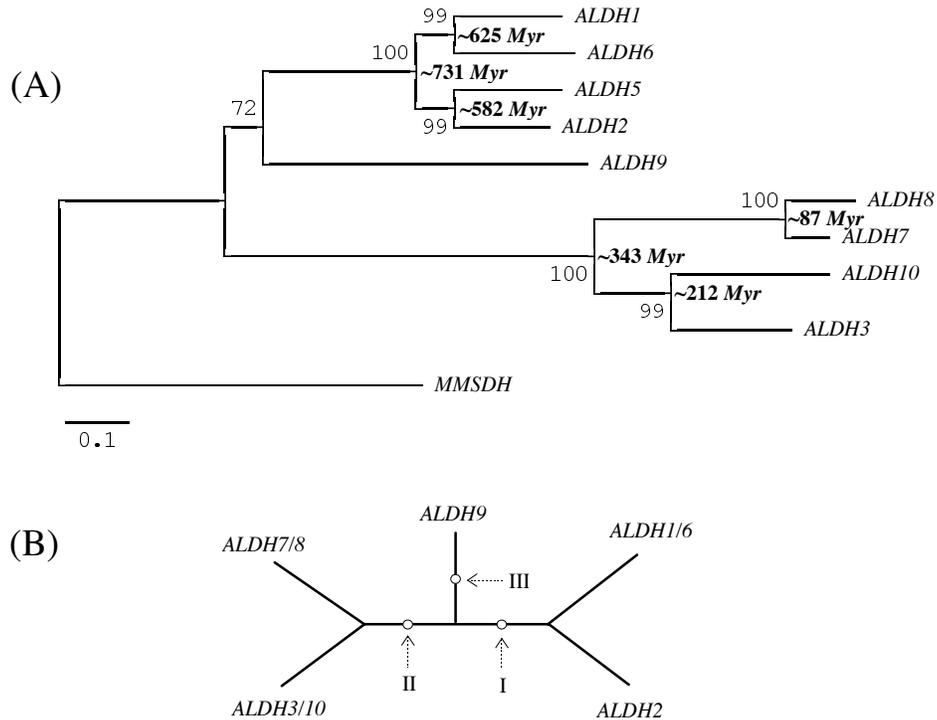
Fig. 5. (A) A neighbor-joining tree (35) computed and visualized with MEGA (36) from eleven human protein sequences using the Poisson correction for multiple hits (37). The per cent of bootstrap (39) resamplings (out of 500) supporting each sequence partition is shown next to the corresponding interior branch; the times of divergence between human genes were estimated with the "linearized tree" algorithm (32). In this estimation we used known ALDH protein sequences from Rodents, Primates, and Artyodactyls and the divergence time 104 *Myr* (33) for Rodentia/Artiodactyla bifurcation. (B) The unrooted tree topology that was used in the maximum likelihood analysis of exon/intron organization of ALDH genes. There are three pairs of ALDH genes which have identical exon/intron patterns within each pair: *ALDH1* and *ALDH6*, *ALDH3* and *ALDH10*, and *ALDH7* and *ALDH8*. The unrooted tree topology is the same as the neighbor-joining trees in figs. 3 A and 4. The arrows show three alternative positions of the tree root (in the maximum likelihood computation we refer to these rooted trees as tree I, tree II, and tree III, see Table 1).

least four clusters of ALDH genes where bacterial and eukaryotic genes are grouped together (see figs. 3, 4) suggested that the divergence times of the four clusters are greater than 2110 *Myr*, the estimated age of the oldest known extinct eukaryote, *Crypania spiralis* (38).

Our phylogenetic reconstruction thus indicated that the "progenote," the common ancestor of eukaryotes and bacteria, was likely to have at least four distinct ALDH genes, since animal and eubacterial genes were grouped together with high bootstrap (39) support. (At least some of the hypothetically ancient homologous ALDH genes are found in five kingdoms of living organisms, Bacteria, Protozoa, Plants, Fungi, and Animals, although plant ALDH genes seem to be studied less extensively than the animal genes, see fig. 4.) An alternative explanation of the same tree would require three or more "late" (after the eukaryotes/bacteria divergence) lateral gene transfers between animals and

bacteria. Either explanation should be compatible with the maximum likelihood analysis presented in the following section.

2.2. Intron evolution in ALDH genes: comparison of competing theories. We developed a model that was flexible enough to account for analyses under each rival theory. Our model incorporates the following assumptions. (*i*) There are three major types of elementary events causing changes in exon/intron patterns: intron insertion, intron deletion, and intron slippage, where an intron slippage is a hypothetical short-range "jump" of an intron within the same gene. (*ii*) The actual number of elementary events along a tree branch follows a Poisson distribution. (*iii*) The probability of each new evolutionary event *given a fixed exon/intron arrangement* does not depend on either the order or the number of past events in the evolutionary history of the gene. (*iv*) Rates of intron insertion, deletion and slippage are fixed along each branch of the tree, but can differ among branches. We also assumed that the correct *unrooted* tree topology for human ALDH genes is known and can be meaningfully rooted in three alternative ways (fig. 5 B).

Starting with the above assumptions, we applied a standard set of matrix manipulations (40) used in the theory of Markov chains for deriving transition probabilities between different exon/intron patterns. These probabilities were then used to compute the conditional probability ("the likelihood given data") of observing the present-day gene structures given a specified tree and a fixed set of the model parameter values (41). First, we defined instantaneous transition rate matrices corresponding to a first-order Markov chain description of intron evolution. The entries of each matrix were assigned rate parameters λ, μ, or ϕ whenever the corresponding pair of intron arrangements was separated by a *single* intron insertion, deletion, or slippage, respectively; the matrix entries were set to zero whenever the distance between corresponding intron arrangements exceeded one elementary event.   The diagonal elements of each rate matrix were chosen to ensure that the sum of elements in each row is equal to zero. For example, for a hypothetical gene with only *two* sites potentially hosting introns, there are four possible intron/exon configurations: 00, 01, 10, and 11, where zero and one stand for intron absence and presence, respectively.   Thus, the transition from configuration 00 to configuration 01 corresponds to an intron insertion; the transition from 01 to 00 indicates intron loss; a transition from 10 to 01 denotes an intron slippage. The resulting instantaneous transition rate matrix, **Q**, is then written as follows

$$\mathbf{Q} = \begin{bmatrix} -2\lambda & \lambda & \lambda & 0 \\ \mu & -\lambda-\mu-\varphi & \varphi & \lambda \\ \mu & \varphi & -\lambda-\mu-\varphi & \lambda \\ 0 & \mu & \mu & -2\mu \end{bmatrix} \begin{matrix} 00 \\ 01 \\ 10 \\ 11 \end{matrix},$$
$$ 00 01 10 11$$

where λ, μ, and ϕ stand for the instantaneous rates of intron insertion, deletion, and slippage, respectively. Second, the matrices of transition probabilities between exon/intron arrangements were computed numerically as matrix exponentials of the corresponding instantaneous transition rate matrices. This operation produces a matrix of transition probabilities between gene arrangement states during time t (expressed in terms of the expected number of events of each type), and is symbolically expressed as $e^{\mathbf{Q}t}$. Third, the likelihood value was calculated as described by J. Felsenstein (41), treating the number of ancestral introns at the "root" of the tree and the mean rates of intron

rearrangement along each tree branch as model parameters. All numerical computations were performed with the MATLAB® 4.0 package produced by Math Works Inc. (To make the required computations feasible we divided the ALDH genes into ten domains (see fig. 2) assuming that intron slippage was prohibited between domains. We defined the boundaries between these domains to minimize the number of the ancestral introns required to explain the present-day genes, as is commonly done in "introns-early" analyses. Without this segmentation the computation of likelihood functions would be effectively impossible because of the large number of intermediate sequence states at each node of the tree. Indeed, each sequence with n potentially intron-bearing sites can be observed in 2^n different binary states, where 0 stands for an intron absence, and 1 for an intron presence. This is a very large number even for a moderate n (*e.g.*, more than 10^9 for $n = 30$), and the likelihood values have to be computed by evaluating transition probabilities through each of 2^n states for each interior node of the tree. Fortunately, it was possible to compute an *approximate* likelihood value by assuming that intron slippages can move introns only *within* each of the ten domains shown in the figure. Only the present-day intron positions were used for the computation.) Finally, we used multidimensional simplex numerical optimization to find a set of parameter values maximizing the likelihood value. (We eliminated *ALDH5* from the analysis because this gene apparently resulted from a single processed mRNA reverse transcription event.)

With this model we were able to directly compare the fit of each alternative theory to the actual ALDH data. The fit of any two models to the data set can be objectively compared with the Akaike Information Criterion (AIC; ref. 42). The AIC value is computed for each rival model according to a simple formula, $AIC_i = 2 N_i - 2 \log L_i$, where N_i is the number of parameters used in the ith model, and $\log L_i$ is the logarithm of the maximum likelihood value obtained under the model. The criterion is designed such that the models that fit the data *better* have *smaller* AIC values.

The results of each analysis were completely different for each set of assumptions. Comparison of AIC values (Table 1, scenarios A, B, and C) showed that the probability of generating the actual ALDH data under the "no-slippage" assumption and the "insertions only" model (= "introns-late") was almost 10^6 times as large as the analogous probability under the "deletions only" (= "introns-early") model. To our surprise, reanalysis of the same data allowing for "intron slippages" ("introns-early" assumption, see Table 1, D, E, and F) resulted in a complete reversal of the conclusion. That is, the model "deletions + slippages" (= "introns-early") became the best with a large advantage in AIC values.

Thus, the "intron slippage" assumption is critical for discriminating between the two theories. Below we scrutinize the consistency of the available experimental data with the parameter estimates obtained in our maximum likelihood analysis. We demonstrate that although the model with the smallest ("best") AIC value corresponds to the "introns-early" theory (see Table 1 D), the parameter estimates obtained under this model appear incompatible with both available experimental data and previous arguments in favor of the "introns-early" theory.

Table 1. Comparison of alternative scenarios of exon/intron evolution in the maximum likelihood analysis.

Scenario	N[a]	$\ln L$[b]	AIC[c]	Ancestral intron number	max branch		
					slip[f]	ins[g]	del[h]
A. Only deletion							
Tree I/II/III	8	-176.75	369.51	31[d]	0.	0.	1.68
B. Only insertion							
Tree I	8 + 1	-165.75	**349.50**	0[e]	0.	0.02	0.
Tree II/III		-167.16	**352.32**		0.	0.02	0.
C. Insertion + Deletion							
Tree I	8 + 8 + 1	-165.75	365.50	0[e]	0.	0.02	0.
Tree II/III		-167.16	368.32		0.	0.02	0.
D. Deletion + Slippage							
Tree I		-92.68	**219.35**		11.6	0.	0.23
Tree II	8 + 8 + 1	-92.67	**219.34**	10[e]	7.9	0.	0.23
Tree III		-92.69	**219.38**		11.7	0.	0.23
E. Insertion + Slippage							
Tree I		-165.320	364.64		0.	0.02	0.
Tree II	8 + 8 + 1	-163.882	361.76	0[e]	0.	0.02	0.
Tree III		-165.992	365.98		0.	0.02	0.
F. Deletion + Insertion + Slippage							
Tree I	8 + 8 + 8 + 1	-92.66	235.33	10[e]	19.7	0.	0.22
Tree II/III		-92.67	235.34		22.4	0.	0.22

Note - [a] the number of model parameters, [b] the natural logarithm of the likelihood value, [c] Akaike Information Criterion, [d] the number of ancestral introns was pre-set rather than estimated, [e] the estimated number of ancestral introns, [f], [g], and [h] the maximum likelihood estimates of the rates of intron slippage, intron insertion, and intron deletion, respectively, expressed per site per branch of the tree.

The results of our analyses turned out to be completely different depending on the assumptions used. Comparison of AIC values (Table 1, scenarios A, B, and C) showed that the probability of generating the actual ALDH data under "no-slippage" assumption and the "insertions only" model (= "introns-late") was almost 10^6 times as large as the analogous probability under the "deletions only" (= "introns-early") model. To our surprise, the re-analysis of the same data allowing for "intron slippages" ("introns-early" assumption, see Table 1, D, E, and F) resulted in complete reversal of conclusion. That is, the model "deletions + slippages" (= "introns-early") became "the best" with a large advantage in AIC values. Apparently, the "intron slippage" assumption is critical for discriminating between the two theories and it is important to scrutinize the consistency of the available experimental data with the parameter estimates obtained in our maximum likelihood analysis.

Below we demonstrate that although the model with the smallest ("best") AIC value corresponds to the "introns-early" theory (see Table 1 D), parameter estimates obtained

under this model appear incompatible with both available experimental data and previous arguments in favor of the "introns-early" theory.
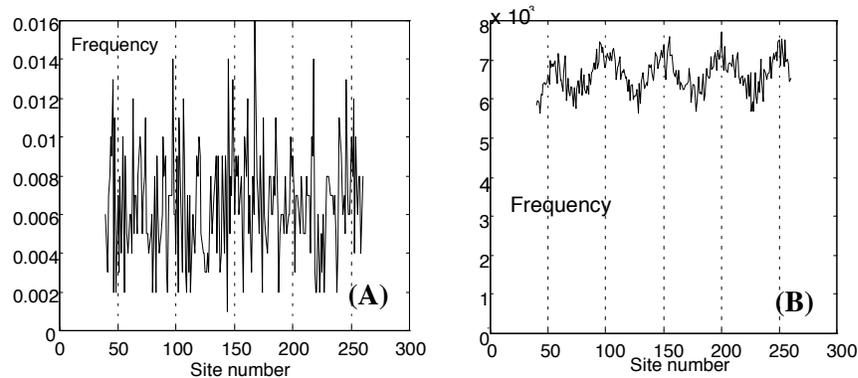
Fig. 6. (A) The non-randomness of intron distribution along gene is effectively undetectable when slippage rates are as high as was estimated in our maximum likelihood analysis and the number of independent present-day genes under analysis is not unrealistically large (<10,000). The figure shows a frequency distribution of introns which was computed through averaging *1000* "present-day" genes obtained in computer simulation. Each "present-day" gene independently "evolved" from a hypothetical ancestral gene with multiple "introns" separated by fifty-nucleotide "exons" (dashed lines indicate positions of the "ancestral" introns). Then, approximately two thirds of the "ancestral" introns were randomly deleted, the remaining introns were subjected to "slippages" at rate 9 slippages per site. Direction, 5' or 3', of each slippage event was chosen randomly; the length of each "leap" was sampled from a uniform distribution defined on interval [1, 12]. Note that in our simulation the "present-day" genes were assumed to evolve independently; the phylogenetic non-independence of actual present-day genes should additionally *increase* the variance of intron distribution. (B) To significantly prove non-randomness of intron distribution for the same model and parameter values one needs a very large sample of present day genes. This frequency distribution of intron positions was obtained by "averaging" over *100,000* (rather than 1000 in fig. 6 A) "present-day" genes generated as described above.

2.3. Mechanism of intron slippage and distribution of introns in human genes. At present there is no plausible known molecular mechanism to account for frequent intron slippage and that would be consistent with the actual patterns of intron distribution in eukaryotes. The simplest explanation for intron slippage is deletion of several nucleotides at the 3' end of one exon and insertion of the same number of nucleotides at the 5' end of the following exon. Since each of the two rearrangements by itself must be extremely deleterious (and putative intron slippages frequently leave the coding region undamaged) this mechanism appears to be inappropriate for explaining frequent slippage. Martinez *et al.*(43) suggested a more sophisticated mechanism based on the "single-intron-deletion" scenario of Fink (44). Fink's mechanism included the following steps: (*i*) a normal excision of an intron from pre-mRNA, (*ii*) reverse transcription of the modified pre-mRNA, and (*iii*) homologous recombination of the resulting cDNA with the original gene. Martinez *et al.* (43) hypothesized an additional event which may follow step (*i*): imprecise re-insertion of an excised intron back into the pre-mRNA (see ref. 45 for experimental evidence of reverse splicing). The advantage of the modified mechanism is that it accounts for a "clean" displacement of an intron within a coding region, although leaving the supposed twelve base-pair limit for intron slippage (4) unexplained. Since there is no direct evidence for reverse transcription of cellular RNAs

in eukaryotic cells, and since reverse transcription in retroviruses occurs only within the viral particle isolating cellular RNAs from the virus enzyme, the Fink-Martinez mechanism requires the presence of a *defective* retrovirus with a mutation in the packaging signal (44). (The simultaneous loss of several introns, as in the human *ALDH5* gene, can be explained by re-integration of the reverse-transcribed mRNA back into the genome (44, 46).)

Unless there exists strong selection preserving the number and/or spatial distribution of introns, evolution under the Fink-Martinez mechanism should result in very specific exon/intron structures (44): (a) Intron deletion should be more frequent than intron slippage, leading to a paucity of introns. (b) The retained introns should be concentrated near the 5' end of each gene because (*i*) reverse transcription begins at the 3' poly(A) tract of mRNA but rarely extends completely to the 5' end, and (*ii*) recombination between genes and cDNAs affects the ends of genes less frequently than the middle. (c) As a consequence of (a) and (b), intron slippages should be rarely observed at the ends of genes, especially at the 5' end. These predictions are in good accord with the exon/intron structures observed in yeast (44) but are clearly inconsistent with human ALDH genes: human genes have numerous introns which are *uniformly* distributed along the coding regions (see fig. 2), and to fit the "introns-early" theory hypothetical intron slippages have to be invoked at both ends of genes.

Unlike intron slippage, intron deletion can result from a one-step mutation event (rather than coincidence of two or more low-probability events), and, in the absence of counteracting selection, should be observed more frequently than intron slippage.

Finally, there are at least two hypothetical mechanisms explaining intron insertion. One is reverse splicing of an excised intron into a non-homologous pre-mRNA, followed by reverse transcription and homologous recombination (44). Another possible mechanism involves invasion of a group II intron (from organelles) into the nuclear genome, followed by a one-mutation transformation of the intron into a regular nucleosomal intron (6, 47, 48, 49, 50): only a single nucleotide substitution is required to convert "(U/C)A ... GU" dinucleotides flanking group II introns into canonical "GA ... GT" dinucleotides flanking nuclear introns, and it was recently discovered (51) that group II introns from yeast mitochondria can integrate *directly* into double-stranded genomic DNA . Therefore, the integration of group II introns into the genome is a one-step event where all molecular machinery is provided by the intron itself.

Thus, according to plausible evolutionary scenarios and the experimental evidence available today, intron slippage should be considerably less likely than intron deletion. In contrast, our maximum likelihood analysis under the "introns-early" assumptions (see Table 1 D, E, and F) suggested that to explain real data under this theory intron slippage has to be two orders of magnitude more frequent than intron deletion.

To demonstrate that the estimated rates if intron slippage contradict support for the "introns-early" theory based on a putative correlation between the ends of ancestral protein "modules" and the boundaries of proto-exons (*e.g.*, see 52), we performed a computer simulation built on the assumptions of the "introns-early" theory. This simulation (see fig. 6 A, B) demonstrated that the reported correlation cannot be detected from any *reasonable* sample of present-day genes (say, < 10,000) if intron slippage rates were as high as estimated in our analysis (see fig. 6 A, B). In our simulation "present-day" genes independently "evolved" from a hypothetical ancestral gene with multiple introns separated by fifty-nucleotide exons (dashed lines indicate positions of the "ancestral" introns). Approximately two thirds of the "ancestral" introns were then randomly deleted,

and the remaining introns were subjected to slippage at a rate of 9 slippages per site. The direction of each slippage event (either 5' or 3') was chosen randomly; the length of each "leap" was sampled from a uniform distribution defined by the interval [1, 12]. Figure 6 shows the resulting distribution of introns from a sample of 1000 genes (fig. 6 A) and 100,000 genes (fig. 6 B). (In our simulation the "present-day" genes were assumed to evolve independently; the phylogenetic non-independence of actual present-day genes should *increase* the variance of intron distribution.)

Assuming that introns were inserted into coding sequences relatively recently, how can one explain the non-randomness of intron distribution? Recent experimental data (*e.g.*, 53) indicate that nuclear DNA of eukaryotes is non-uniformly protected by proteins maintaining chromosome structure. For example, it was shown that during transcription of the *Dam* gene in yeast, each nucleosome associated with *Dam* selectively shielded approximately eighty base pairs of yeast DNA while allowing methylation enzymes to freely access DNA in internucleosome "linkers" (53). Therefore, we hypothesize that a non-uniform distribution of introns in eukaryotic genes could have been caused by preferential intron insertion into stretches of DNA that were temporary liberated from nucleosome protection.

3. Conclusion

The "intron slippage" assumption is the cornerstone of many lines of defense of the "introns-early" theory, yet, according to our analysis of ALDH genes it is precisely this assumption that leads to an internal contradiction between the arguments supporting the theory: first, contrary to expectation, the estimated intron slippage rates are much higher than the estimated intron deletion rates; second, high intron slippage rates question the reported correlation between the boundaries of the "ancient protein modules" and the ends of "proto-exons" (15). Indeed, if intron slippages are allowed, *each* putative ancestral intron had to move *at least once* to arrive at the present-day exon/intron arrangement in human ALDH genes. This is because *all* intron positions between groups *ALDH1/2/6* and *ALDH3/7/8/10*, and *ALDH3/7/8/10* and *ALDH9* are different (see fig. 2) and only *one* out of nine intron positions is conserved between *ALDH9* and *ALDH3/7/10*. Therefore, it is hardly surprising that the rates of intron slippage estimated in our maximum likelihood analysis are very high (Table 1).

In summary, the assumption of frequent intron slippage leads to inconsistencies with both the available body of experimental evidence and the data analyses provided by proponents of the "introns-early" theory; without this assumption the human ALDH data support the "introns-late" theory. The methods illustrated in this article can be readily applied to other data sets to test the generality of the conclusions drawn from the ALDH data.

4. References

1. Darnell, J.E., & Doolittle, W.F., *Speculations on the early course of evolution*, Proc. Natl. Acad. Sci. USA **83** (1986), 1271-1275.
2. Gilbert, W., Marchionni, M., & McKnight, G., *On the antiquity of introns*, Cell **46** (1986), 151-154.
3. Dorit, R.J., Schorndach, L, & Gilbert, W., *How big is the universe of exons?* Science **250** (1990), 1377-1382.
4. Cerff, R. *Tracing biological evolution in protein and gene structures,* eds. Gō, M., & Schimmel, P. Elsevier, New York, 1995, pp. 205-227.
5. Rogers, J.H. *The role of introns in evolution.* FEBS Lett. **268** (1990)**,** 339-343.
6. Cavallier-Smith, T. *Intron phylogeny: a new hypothesis*, Trends Genet. **7** (1991)**,** 145-148.
7. Patthy, L., *Exons -- original building blocks of proteins?* Bioessays **13** (1991)**,** 187-192.
8. Sharp, P.A., *"Five easy pieces,"* Science **254** (1991)**,** 663.

9. Palmer, J.D., & Logsdon, J.M., Jr., *The recent origins of introns*, Curr. Opin. Genet. Dev. **1** (1991)**,** 470-477.

10. Kersanach, R., Brinkmann, H., Liaud, M.-F., Zhang, D.-X., Martin, W., & Cerff, R., *Five identical intron positions in ancient duplicated genes of eubacterial origin*. Nature (London) **367** (1994)**,** 387-389.

11. Tittiger, C., Whyard, S., & Walker, V.K., *A novel intron site in the triosephosphate isomerase gene from the mosquito* Culex tarsalis, Nature (London) **361** (1993)**,** 470-472.

12. Kwiatowski, J., Krawczyk, M., Kornacki, M., Bailey, K., & Ayala, F., *Evidence against the exon theory of genes derived from the triose-phosphate isomerase gene*. Proc. Natl. Acad. Sci. USA **92** (1995)**,** 8503-8506.

13. Hurst, L.D., & McVean, G.T., *A difficult phase for introns-early*, Curr. Biol. **6** (1996)**,** 533-536.

14. Logsdon, J.M., Jr., Tyshenko, M.G., Dixon, C., D.-Jafari, J., Walker, V.K., & Palmer, J.D., *Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory*, Proc. Natl. Acad. USA **92** (1995)**,** 8507-8511.

15. Noguti, T., & Gō, M. *Tracing biological evolution in protein and gene structures*. eds., Gō, M., & Schimmel, P., Elsevier, New York, 1995, pp. 161-174.

16. Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M. Jr., & Doolittle, W.F., *Testing the exon theory of genes: the evidence from protein structure*, Science **265** (1994)**,** 202-207.

17. Fedorov, A., Suboch, G., Bujakov, M., & Fedorova, L., *Analysis of nonuniformity in intron phase distribution*, Nucl. Acid Res. **20** (1992)**,** 2553-2557.

18. Long, M.Y., Rosenberg, C., & Gilbert, W., *Intron phase correlations and the evolution of the intron/exon structure of genes*, Proc. Natl. Acad. Sci. USA **92** (1995)**,** 12495-12499.

19. Ambroziak, W., & Pietruszko, R. in *Enzymology and molecular biology of carbonyl metabolism 4*, eds. Weiner, H., Crabb, D.W., & Flynn, T.G., Plenum, New York, 1993, pp. 5-15.

20. Harrington, M.C., Henehan, G.T.M. & Tipton, K.F., *The roles of human aldehyde dehydrogenase isozymes in ethanol metabolism.* Prog. Clin. Biol. Res. **232** (1987)**,** 111-125.

21. Jakoby, W.B. & Ziegler, D.M., *The enzymes of detoxification*, J. Biol. Chem. **265** (1990)**,** 20715-20718.

22. Hsu, L.C., Chang, W.-C., & Yoshida, A., *Genomic structure of the human cytosolic aldehyde dehydrogenase gene*, Genomics **5** (1989)**,** 857-865.

23. Hsu, L.C., Bendel, R.E., & Yoshida, A., *Genomic structure of the human mitochondrial aldehyde dehydrogenase gene*, Genomics **2** (1988)**,** 57-65.

24. Hsu, L.C., Chang, W.-C., Shibuya, A., & Yoshida, A., *Human stomach aldehyde dehydrogenase cDNA and genomic cloning, primary structure, and expression in* Escherichia coli, J. Biol. Chem. **267** (1992)**,** 3030-3037.

25. Hu, C.A., Lin, W.W., & Valle, D., *Cloning, characterization and expression of cDNAs encoding human Δ 1'-pyrroline-5-carboxylate dehydrogenase*, J. Biol. Chem. **271** (1996)**,** 9795-9800.

26. Hsu, L.C., Chang, W.-C., Hiraoka, L.R., & Hsieh, C.L., *Molecular cloning, genomic organization, and chromosomal organization of additional human aldehyde dehydrogenase gene,* ALDH6, Genomics **24** (1994)**,** 333-341.

27. Hsu, L.C., Chang, W.-C., & Yoshida, A., *Cloning a cDNA encoding human ALDH7, a new member of aldehyde dehydrogenase family*, Gene **151** (1994)**,** 285-289.

28. Lin, S.W., Chen, J.C., Hsu, C.L., & Yoshida, A., *Human gamma-aminobutyraldehyde dehydrogenase (ALDH9): cDNA sequence, genomic organization, polymorphism, chromosomal localization, and tissue expression*, Genomics **34** (1996)**,** 376-380.

29. De Laurenzi, V., Rogers, G.R., Hamrock, D.J., Marekov, L.N., Steinert, P.M., Compton, J., Markova, N., & Rizzo, W.B., *Sjogren-Larsson syndrome is caused by mutations in the fatty aldehyde dehydrogenase gene*, Nature Gen. **12** (1996)**,** 52-57.

30. Rost, B., & Sander, C., *Prediction of protein secondary structure at better than 70% accuracy*, J. Mol. Biol. **232** (1993)**,** 584-599.

31. Rost, B., & Sander, C., *Conservation and prediction of solvent accessibility in protein families*, Proteins, **20** (1994)**,** 216-226.

32. Takezaki, N., Rzhetsky, A., & Nei, M., *Phylogenetic test of the molecular clock and linearized trees*, Mol. Biol. Evol. **12** (1995)**,** 823-833.

33. Hedges, S.B., Parker, P.H., Sibley, C.G., & Kumar, S., *Continental breakup and the ordinal diversification of birds and mammals*, Nature (London) **381** (1996)**,** 226-229.

34. Knoll, A.H., *End of the Proterozoic Eon*, Sci. Am. **265** (1991)**,** 64-73.

35. Saitou, N., & Nei, M., *The neighbor-joining method: a new method for reconstructing phylogenetic trees*, Mol. Biol. Evol. **4** (1987)**,** 406-425.

36. Kumar, S., Tamura, K., & Nei, M., *MEGA: Molecular Evolutionary Genetics Analysis*, 1993, The Pennsylvania State University, University Park, PA, Version 1.0.

37. Zuckerkandl, E., & Pauling, L. in *Evolving Genes and Proteins*, eds. Bryson, V., & Vogel, H. J., Academic Press, New York, 1965, pp. 97-166.

38. Han, T.-M., & Runnegar, B., *Megascopic eukaryotic algae from the 2.1-billion-year-old Negaunce iron-formation, Michigan* Science **257** (1992)**,** 232-235.

39. Felsenstein, J., *Confidence limits on phylogenies: an approach using the bootstrap*, Evolution **39** (1985)**,** 783-791.

40. Keilson, J., *Markov chain models - rarity and exponentiality*, Springer-Verlag, New York, 1979.

41. Felsenstein, J., *Evolutionary trees from DNA sequences: a maximum likelihood approach*, J. Mol. Evol. **17** (1981)**,** 368-376.

42. Akaike, H., *A new look at the statistical model identification*, IEEE Trans. Autom. Contr. **AC-19** (1974)**,** 761-723.

43. Martinez, P., Martin, W., & Cerff, R., *Structure, evolution and anaerobic regulation of a nuclear gene encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase from maize*, J. Mol. Biol. **208** (1989)**,** 551-565.

44. Fink, G.R., *Pseudogenes in yeast?* Cell **49** (1987)**,** 5-6.

45. Jarrell, K.A., *Inverse splicing of a group II intron*, Proc. Natl. Acad. Sci. USA **90** (1993)**,** 8624-8627.

46. Nugent, J.M., & Palmer, J.D., *RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution*, Cell **66** (1991)**,** 473-481.

47. Cech, T.R. *Five easy pieces*, Cell **72** (1986)**,** 161-164.

48. Weiner, A.M., *mRNA splicing and autocatalytic introns: distant cousins or the products of chemical determinism?* Cell **72** (1993)**,** 161-164.

49. Ferat, J.-L., & Michel, F., *Group II self-splicing introns in bacteria,* Nature (London) **354** (1993)**,** 358-361.

50. Zimmerly, S., Guo, H., Perlman, P.S., & Lambowitz, A.M., *A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility*, Cell **82** (1995)**,** 545-554.

51. Yang, J., Zimmerly, S., Perlman, P.S., & Lambowitz, A.M., *Efficient integration of an intron RNA into double-stranded DNA by reverse splicing*, Nature (London) **381** (1996)**,** 332-335.

52. Fukami-Kobayashi, K., Mizutani, M., & Gō, M. (1995) in *Tracing biological evolution in protein and gene structures*. eds., Gō, M., & Schimmel, P. , Elsevier, New York, pp. 271-282.

53. Kladde, M.P., & Simpson, R.T., *Positioned nucleosomes inhibit* Dam *methylation* in vivo, Proc. Natl. Acad. Sci. USA **91** (1994)**,** 1361-1365.

Andrey Rzhetsky
    Current address: COLUMBIA GENOME CENTER, COLUMBIA UNIVERSITY, 630 WEST 168TH STREET -BB 16-1611, NEW YORK, NY 10032
    E-mail address: andrey@genome2.cpmc.columbia.edu

Francisco José Ayala
    Current address: INSTITUTE OF MOLECULAR EVOLUTIONARY GENETICS AND DEPARTMENT OF BIOLOGY, PENNSYLVANIA STATE UNIVERSITY, UNIVERSITY PARK, PA 16802
    E-mail address: zeayala@psu.edu

Lily C. Hsu, Cheng Chang, Akira Yoshida
    Current address: DEPARTMENT OF BIOCHEMICAL GENETICS, BECKMAN RESEARCH INSTITUTE OF THE CITY OF HOPE, DUARTE, CA 91010
    E-mail address: ayoshida@smtplink.coh.org, lchsu@smtplink.coh.org.