The α-Mannosidases: Phylogeny and Adaptive Diversification

Daniel S. Gonzalez and I. King Jordan†*

*Department of Medical Microbiology, University of Georgia; and †Department of Biological Sciences,
University of Nevada at Las Vegas

α-Mannosidase enzymes comprise a class of gylcoside hydrolases involved in the maturation and degradation of glycoprotein-linked oligosaccharides. Various α-mannosidase enzymatic activities are encoded by an ancient and ubiquitous gene superfamily. A comparative sequence analysis was employed here to characterize the evolutionary relationships and dynamics of the α-mannosidase superfamily. A series of lineage-specific BLAST searches recovered the first ever recognized archaean and eubacterial α-mannosidase sequences, in addition to numerous eukaryotic sequences. Motif-based alignment and subsequent phylogenetic analysis of the entire superfamily revealed the presence of three well-supported monophyletic clades that represent discrete α-mannosidase families. The comparative method was used to evaluate the phylogenetic distribution of α-mannosidase functional variants within families. Results of this analysis demonstrate a pattern of functional diversification of α-mannosidase paralogs followed by conservation of function among orthologs. Nucleotide polymorphism among the most closely related pair of duplicated genes was analyzed to evaluate the role of natural selection in the functional diversification of α-mannosidase paralogs. Ratios of nonsynonymous and synonymous variation show an increase in the rate of nonsynonymous change after duplication and a relative excess of fixed nonsynonymous changes between the two groups of paralogs. These data point to a possible role for positive Darwinian selection in the evolution of α-mannosidase functional diversification following gene duplication.

Introduction

Comparisons of recently completed whole-genome sequences have confirmed the extent to which genic organization is characterized by a hierarchy of gene families, subfamilies, and superfamilies (Tatusov, Koonin, and Lipman 1997). Despite the rapid accumulation of gene sequences, the rate of new gene family discovery continues to drop (Henikoff et al. 1997). It is foreseeable that the entire spectrum of genic diversity will soon fall into a limited number of gene families. Gene families consist of homologous sequences that are designated either orthologs or paralogs. Orthologs are related to a common ancestor by speciation, while paralogs are related by duplication (Fitch 1970). The evolutionary significance of gene duplication has long been recognized (Haldane 1932; Muller 1935). While orthologs tend to encode the same function, paralogs often evolve novel activities. Ohno (1970) formulated the provocative hypothesis that gene duplication is a prerequisite for the evolution of any new gene function. He recognized that natural selection is inherently conservative and postulated that only the redundancy created by gene duplication could allow a gene copy to escape the pressure of negative selection and evolve a new function. Findings during the ensuing decades have revealed that while there are in fact other ways to evolve new genes and functions, gene duplication remains the most important mechanism for generating genic novelty (Li 1997).

The classification scheme of glycoside hydrolases provides an archetype example of the hierarchical relationships among a widespread evolutionarily and/or functionally related superfamily of enzymes (Henrissat 1991, 1998; Henrissat and Bairoch 1993, 1996; Henrissat and Romeu 1995; Henrissat and Davies 1997). Glycoside hydrolases are enzymes that hydrolyze the glycosidic bond between carbohydrates or between a carbohydrate and a noncarbohydrate moiety. The innovative sequence-based classification system originally proposed by Henrissat (1991) currently consists of 66 families of glycoside hydrolases (see http://expasy.hcuge.ch/cgi-bin/lists?glycosid.txt). Structural and functional characteristics that indicate relationships between members of different families have resulted in the designation of clans that are composed of two or more families.

α-Mannosidases are glycoside hydrolases involved in both the maturation and the degradation of Asn-linked oligosaccharides (Dewald and Touster 1973; Tulsiani et al. 1982; Lal et al. 1994; Liao, Lal, and Moremen 1996). The glycoprotein maturation and degradation pathways are very conserved, and α-mannosidase activities have been detected in all eukaryotes assayed. α-Mannosidase–encoding genes have been isolated and their products characterized from a diverse group of eukaryotes, including the protozoan *Trypanosoma cruzi*, the yeast *Saccharomyces cerevisiae*, and the metazoans *Drosophila melanogaster* and *Homo sapiens* (Camirand et al. 1991; Kerscher et al. 1995; Liao, Lal, and Moremen 1996; Vandersall-Nairn et al. 1998). Traditionally, α-mannosidases have been organized into two classes (I and II) based on both functional characteristics and sequence homology (Moremen, Trimble, and Herscovics 1994; Henrissat 1998). The cellular compartment where they catalyze mannose hydrolysis (e.g., endoplasmic reticulum, Golgi, or lysosome) further distinguishes different α-mannosidase enzymes.

[1] Present address: Department of Microbiology and Center for Computational Biology, Montana State University.

Abbreviations: ER, endoplasmic reticulum; MIR, motif intervening region; OSM, ordered series of motifs.

Key words: α-mannosidase, glycoside hydrolase, gene duplication, positive selection, adaptation.

Address for correspondence and reprints: I. King Jordan, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, Bethesda, MD 20894. E-mail: ikingjordan@hotmail.com.

Mol. Biol. Evol. 17(2):292–300. 2000

Class I α-mannosidase enzymes thus far characterized are all involved in the maturation of Asn-linked oligosaccharides. These enzymes all process the trimming of $Man_9GlcNAc_2$ to $Man_5GlcNAc_2$. While class I α-mannosidase enzymes only hydrolyze α-1,2 mannose bonds, they differ in their stereospecificities (Lal et al. 1994). Class I α-mannosidase enzymes are localized to either the endoplasmic reticulum or the Golgi complex. The majority of class II α-mannosidase enzymes that have been characterized catalyze the degradation of Asn-linked oligosaccharides. Class II α-mannosidase enzymes show less biochemical specificity, as they possess α-1,3, α-1,6, and α-1,2 hydrolytic activity. Enzymes of this class also have a wider range of cellular compartmentalization and can be localized to the cytosol and lysosomes in addition to the Golgi complex.

The rapid accumulation of genomic and cDNA nucleotide sequences in the various public databases has facilitated the in silico discovery of several putative α-mannosidase sequences with statistically significant similarities to classically cloned and functionally characterized α-mannosidase genes. Using a diverse representative set of α-mannosidase amino acid query sequences, an exhaustive search for and analysis of α-mannosidase homologous sequences was performed. Sequence retrieval, alignment, and phylogenetic analysis allowed a determination of the range and extent of α-mannosidase variation. The relationship among previously characterized and novel putative α-mannosidase sequences is defined and a classification consistent with the Henrissat scheme is proposed. The comparative method was used to assess the correlation between phylogenetic relationship and cellular localization of biochemically characterized α-mannosidase sequences. Finally, the existence of closely related orthologs and paralogs in the α-mannosidase IA-IB clade allowed a test for positive Darwinian selection for altered function following gene duplication.

Materials and Methods

Data Retrieval

All of the sequences used in this study were retrieved from the National Center for Biotechnology Information (NCBI) GenBank database. An initial data set of previously published and functionally characterized α-mannosidase amino acid sequences was retrieved manually from Entrez (http://www.ncbi.nlm.nig.gov/Entrez). These representative sequences were isolated from a wide phylogenetic range of eukaryotes and possessed a variety of biochemical activities and intracellular localizations. An in-house program (Cluster) was used to group these sequences by amino acid identity into clusters. Divergent α-mannosidase amino acid sequences with representatives from each cluster were used as queries in a series of lineage-specific BLAST (http://www.ncbi.nlm.nih.gov/BLAST) searches. Lineage-specific searches with a diverse set of query sequences allowed for a maximum amount of sequence space to be covered. NCBI's gapped and ungapped tblastn searches were run using the default BLOSUM62 (gapped) and the PAM250 (ungapped) distance matrices.

The sequences analyzed here correspond to the GenBank accession numbers listed below. The accessions include both nucleotide and amino acid sequences. The sequence abbreviations and their corresponding accession numbers follow the species names. Sequence abbreviations consist of two letters that represent the species binomial followed by a single-letter designation that indicates the cellular location of activity for functionally characterized sequences or a ''p'' to indicate putative α-mannosidase sequences that have not yet been biochemically characterized. Cellular compartmentalization abbreviations are as follows: E, endoplasmic reticulum; M, membrane-associated; G, Golgi apparatus; L, lysosome; X, extracellular; and V, vacuolar. Roman numerals that indicate the identity of the α-mannosidase family to which a sequence belongs make up the final component of the sequence abbreviations. The accession number list is as follows: *Arabidopsis thaliana*—AT-LII, Y11767; *Aspergillus saitoi*—AS-XI, D49827; *Bos taurus*—BT-LII, L31373; *Caenorhabditis elegans*—CE-pI.1, Z78012; CE-pI.2, Z73906; CE-pI.5, Z81497; CE-pI.7, Z68882; CE-pII.1, U40948; CE-pI.3, Z68270; CE-pI.6, Z47073; CE-pI.4, U41272; CE-pII.3, U97015; CE-pII.2, Z75954; *Dictyostelium discoideum*—DD-LII, M82822; *Drosophila melanogaster*—DM-GI, X82641; DM-pI, AL021086; DM-GII.1, X77652; DM-GII.2, AB018079; *Escherichia coli*—EC-pIII, AE000176; *Emericella nidulans*—EN-pIII, AF016850; *Felis catus*—FC-LII, AF010191; *Homo sapiens*—HS-EIII, AF044414; HS-LII, U68567; HS-GII.1, U31520; HS-GII.2, D55649; HS-pI, D86967; HS-GIA, X74837; HS-GIB, AF027156; *Mus musculus*—MM-MII, AB006458; MM-GIA, U04299; MM-LII, U87240; MM-GIB, AF078095; MM-GII, X61172; *Mycobacterium tuberculosis*—MT-pIII, Z92772; *Oryctolagus cuniculus*—OC-GIA, U04301; *Penicillium citrinum*—PC-XI, D45839; *Pyrococcus horikoshii*—PH-pIII, AP000003; *Rattus norvegicus*—RN-EIII, M57547; RN-GII, M24353; *Saccharomyces cerevisiae*—SC-EI, Z49631; SC-VIII, M29146; SC-pI.1, U00030; SC-pI.2, Z73229; *Schizosaccharomyces pombe*—SP-pI, AL021813; *Spodoptera frugiperda*—SF-GI, AF005035; SF-GII, AF005034; *Sus scrofa*—SS-MII, D28521; SS-GIA, Y12503; *Synechocystis*—SY-pIII, D63999; *Trypanosma cruzi*—TC-LII, AF077741.

Sequence Alignment

The CLUSTAL W (Thompson, Higgins, and Gibson 1994) and PROBE (Neuwald et al. 1997) programs were used to align amino acid sequences. Initial alignments of the total data set performed with CLUSTAL W revealed a highly divergent group of sequences, and therefore CLUSTAL W was not able to obtain an optimal global alignment. To effectively align the total amino acid data set, the PROBE program was used to identify a common ordered series of motifs (OSM) among all sequences. An alignment of diagnostic sites was extracted from the total OSM alignment. Diagnositic sites were chosen from sites that a PAUP* 4.0b parsimony reconstruction classified as apomorphies with a consistency index of 1 and that supported internal branches leading to the three main clades (families). Higher levels of amino acid sequence identity within families allowed the use of CLUSTAL W for within-family multiple alignments. CLUSTAL W was run with the PAM250 distance matrix and default gap penalty options. The relatively high sequence identity and the use of CLUSTAL W for within-family multiple alignment allowed for the inclusion of motif intervening regions (MIRs). MIRs contain additional information necessary to obtain accurate within-family phylogenetic reconstructions (McClure and Kowalski 1999).

Phylogenetic Analysis

Within-family and among-families amino acid alignments were used with the PAUP* 4.0b package (Swofford 1998) to reconstruct the phylogenies reported here. Both parsimony and the neighbor-joining (Saitou and Nei 1987) distance method were used in phylogenetic reconstruction. Parsimony heuristic searches were conducted with 10 replicates of random stepwise addition and tree bisection reconnection. Distance-based and parsimony methods gave virtually identical results. All topologies reported here are based on the neighbor-joining method. Trees were rooted with midpoint rooting along the longest branch. One hundred bootstrap replicates were performed using the full heuristic bootstrap option.

Amino Acid Sequence Diversity

Average percentages of amino acid identity and standard deviations within and between families were calculated using a subset of 12 sequences (four representatives from each family). The PAUP* 4.0b program was used to calculate the mean character difference distance matrix. Mean character differences were converted to percentages of identity and averaged within and between the three families.

Nucleotide Sequence Diversity

Closely related amino acid sequences for the Golgi α-mannosidase IA-IB clade were aligned using CLUSTAL W with the default options. The Golgi α-mannosidase IA-IB phylogeny was reconstructed as described above. Nucleotide sequences of the same taxa were aligned to correspond to the amino acid alignment using the DNA Stacks program (Eernisse 1992). Ancestral nucleotide sequences were inferred with parsimony using PAUP* 4.0b. The DnaSP program (Rozas and Rozas 1997) was used to calculate K_a and K_s values according to the method of Nei and Gojobori (1986) and to perform the McDonald-Kreitman test of neutrality (McDonald and Kreitman 1991). The time elapsed since IA-IB duplication (T_D) was calculated with K_s values using the method of Li (1997). This calculation was calibrated using the time elapsed since human-mouse speciation (T_S) (Kumar and Hedges 1998).

Results and Discussion

Motif Detection and Alignment

A total of 51 amino acid sequences were retrieved from the GenBank database. These sequences included 30 previously functionally characterized α-mannosidase enzymes, as well as 21 novel putative α-mannosidase sequences. Because of the sensitivity of the lineage-specific BLAST approach (see *Materials and Methods*), this data set included many distantly related sequences. Such weakly conserved, distantly related sequences are characteristic of many protein families and superfamilies. Sequence conservation among all members of a protein family is often limited to small islands of relatively high similarity, referred to as motifs (Dayhoff, Schwartz, and Orcutt 1978). When these regions of high similarity occur in conserved order among a related set of proteins, they are referred to as an OSM (McClure 1991). The OSM makes up a unique signature that characterizes a protein family (Hudak and McClure 1999). Identification of the OSM facilitates multiple alignment of distantly related amino acid data sets such as those analyzed here (McClure, Vasi, and Fitch 1994). A recent comparative analysis of a number of motif detection methods reported that the PROBE program performs the best (Hudak and McClure 1999). The PROBE program was used to identify and align the conserved OSMs common to all the α-mannosidase sequences. PROBE identified 10 motifs that range from 14 to 36 amino acid residues in length, with an average length of ~23 residues. The OSM alignment consists of 276 total sites, 231 of which are phylogenetically informative. An alignment of the total data set was generated using CLUSTAL W. This suboptimal (see *Materials and Methods*) alignment consisted of 1,571 total and 1,164 informative sites.

Phylogeny and Classification

The OSM alignment was used to reconstruct a global α-mannosidase phylogeny (fig. 1). The resulting neighbor-joining tree reveals three robust clades of α-mannosidase sequences. A parsimony phylogeny reconstructed using the OSM alignment showed an identical topology with the exception of a few weakly supported branches within the three main clades. Distance-based and parsimony phylogenetic reconstructions based on the suboptimal CLUSTAL W alignment also gave qualitatively similar results, with the same three major clades and topological differences in weakly supported terminal nodes. Figure 2 shows an alignment of diagnostic sites that clearly distinguish the three major α-mannosidase clades. Average percentages of amino acid identity within and between families (table 1) are also consistent with the existence of three distinct α-mannosidase clades.

Clade I consists of eukaryotic α-mannosidase sequences, including fungal and metazoan representatives. Within this clade, there are two well-supported groups. One group consists mainly of functionally characterized sequences, and the other group is made up of novel sequences characterized in various genome-sequencing

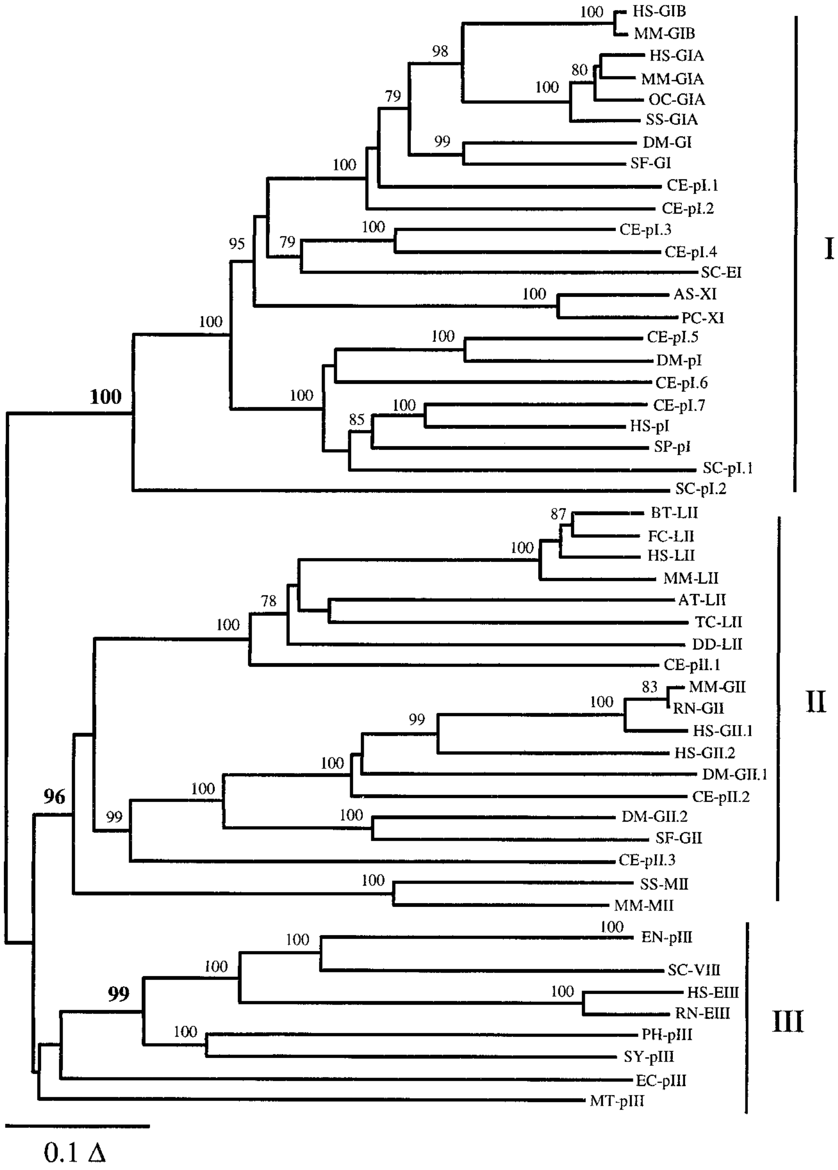


FIG. 1.—Neigbor-joining phylogeny of the α-mannosidase superfamily. The motif-based PROBE alignment was used to reconstruct this phylogeny as described in *Materials and Methods.* Vertical bars and roman numerals delineate the three α-mannosidase families. Taxon name abbreviations are described in *Materials and Methods.* Bootstrap values are shown above the branches. A scale bar indicating the relative amount of change (Δ) along branches is shown.

projects. Among the functionally characterized group, there are sequences with endoplasmic reticulum (ER), Golgi apparatus, and extracellular enzymatic activity. Sequences of this clade have previously been classified as belonging to glycoside hydrolase family 47 (Henrissat 1991; Henrissat and Bairoch 1993).

Clade II is made up of a more diverse set of eukaryotic sequences. In addition to fungal and metazoan isolated sequences, there are also plant, slime mold, and protazoan representatives. Clade II also has the lowest average level of sequence identity (table 1). The taxonomic and sequence diversity that characterizes this clade is consistent with the varied biochemical specificities of its taxa. Among the functionally characterized members of this clade, there are sequences with lysosomal, membrane-associated, and Golgi activity. Clade II Golgi α-mannosidase enzymes show α-1,3 and α-1,6 mannose bond cleavage distinct from the α-1,2 mannosidase activity of clade I Golgi sequences (Moremen, Trimble, and Herscovics 1994). Clade II sequences have previously been placed into glycoside hydrolase family 38 (Henrissat 1991; Henrissat and Bairoch 1993).

In the unrooted global α-mannosidase tree, clade III falls approximately at the midpoint between clades I and II (fig. 3). This group represents the most diverse taxonomic assemblage of α-mannosidase homologous sequences. Seven of the nine clade III members form a well-supported monophyletic group (fig. 1). Among these seven sequences, there are metazoan, fungal, and archaean representatives. The two most basal members of this clade are sequences from gram-negative (*E. coli*) and gram-positive (*M. tuberculosis*) eubacteria. How-

```
                             1111111222
                  122333336667890112477055
                 8828123460384297182534614
                 _________________________
HS-GIA           YLGAVDALTSERLAGGLSGNYETYF
CE-pI.2          YLGAIDAITSERLAGGLSGTYETYF
AS-XI            YLGAVDALTSERLSGGLSGNYESYF
DM-pI            YLSLIDALTSERLAGGISGDFESYF
HS-pI            YLSLVDALTQERLAGGVSGNYESYF

FC-LII           QQPRWHIDFSPNCLNFPEYGSGGTF
AT-LII           QRPRWQIDFGPDTFNFPVS.SGGKF
MM-GII           DIPRWAIDFHPYCQLFPDEHTGG..
SF-GII           TDPKWSIDFHPYCHLYPAESTGAQN
SS-MII           FRPQWQVDFPPFALQYSASATGLYF

EN-pIII          QSCTWLPDFNPFDELTQNEPPGLSR
SC-VIII          QKSKWLPDFNPFDLLTQMEPPG.SR
RN-EIII          QQCSWLPDFNPFDELTHNEPTGCSR
PH-pIII          QEPRYLPDFNPFSELFANEPPAPSR
SY-pIII          QASKWLPDFNPFDEFFANEPPGHTR
```

FIG. 2.—Alignment of representative sequences that shows diagnostic sites from the ordered series of motifs (OSM) that distinguish the three α-mannosidase families. The diagnostic sites were chosen as described in *Materials and Methods.* Bold type for residues within families indicates that at least four out of five residues at a site are identical or similar (i.e., conservative changes). Numbers above the alignment indicate the positions in the OSM alignment at which the diagnostic sites occur.

ever, there is no significant bootstrap support for the branches that group these sequences with the other clade III members (fig. 1). These sequences branch off the most internal node in the global phylogeny (fig. 3). The phylogenetic location and the eubacterial origin of these sequences suggest that they may be ancestral proto-α-mannosidase enzymes. The fact that these putative ancestral sequences group most closely with clade III is consistent with the diversity of taxa in this clade and suggests that clade III represents the most ancestral α-mannosidase family.

Considered together with the glycoside hydrolase superfamily organization (Henrissat 1991), the topology of the global α-mannosidase phylogeny provides a heuristic for a coherent α-mannosidase classification scheme. Previous work based largely on biochemical characterization and, to a lesser extent, on sequence homology has suggested the existence of two distinct classes of α-mannosidase enzymes (Moremen, Trimble, and Herscovics 1994). However, a more recent study reporting the AN-III sequence revealed a new distinct group of closely related class II α-mannosidase sequences (AN-III, RN-EIII, and SC-VIII) that appeared to be distantly related to previously reported class II sequences (Eades et al. 1998). The present global α-mannosidase phylogenetic analysis incorporated six more sequences of this new group and compared them with sequences previously designated class I and II. The results of this comprehensive analysis are consistent with the Eades et al. (1998) study and clearly indicate that there are three distinct well-supported groups of α-mannosidase sequences (figs. 1 and 2 and table 1). Thus, in accordance with the Henrissat scheme, clade III α-mannosidase sequences are proposed to belong to a new family of glycoside hydrolase sequences.

Table 1
Average Percentages of Amino Acid Identity (±SD) Within and Between α-Mannosidase Families

Within families			
I	44.91 ± 0.08		
II		26.49 ± 0.11	
III			34.54 ± 0.06
Between families			
I–II	8.40 ± 0.03		
I–III		9.12 ± 0.02	
II–III			16.78 ± 0.02

NOTE.—Average percentages of amino acid identity and standard deviations were calculated as described in *Materials and Methods.*

Above the level of family, the Henrissat scheme includes the designation of clan to cover separate but related families (see http://afmb.cnrs-mrs.fr/~pedro/CAZY/ghf_intro.html). Levels of average percentages of amino acid identity within families (table 1) indicate that each family represents a distinct homologous group of sequences. Average percentages of amino acid identity between families (table 1), on the other hand, are very low. Such low identity values suggest that sequences between families cannot be considered homologous with statistical confidence (Dayhoff, Schwartz, and Orcutt 1978). However, several other criteria suggest the possibility that the three α-mannosidase families studied here share a common ancestor. For example, lineage-specific BLAST searches that use α-mannosidase sequences from one family as a query can detect α-mannosidase sequences from different families. In addition, the presence of the OSM signature is suggestive of homology between families. Thus, while the BLAST results and the OSM signature are suggestive of common ancestry but not definitive, the three families of α-man-

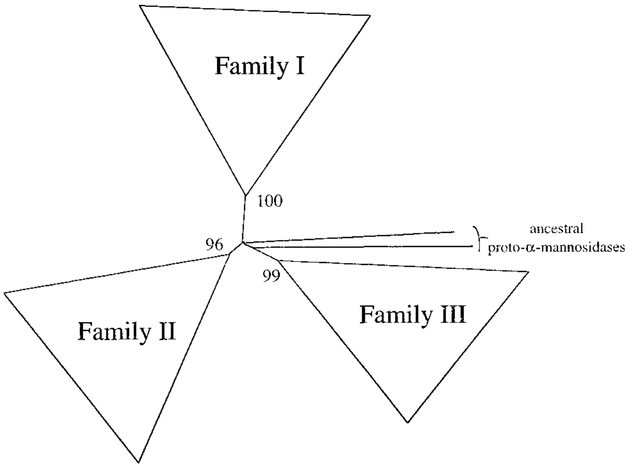


FIG. 3.—Schematic of the unrooted version of the global α-mannosidase phylogeny shown in figure 1. The relationship among the α-mannosidase families and the putative ancestral proto-α-mannosidase sequences is shown.

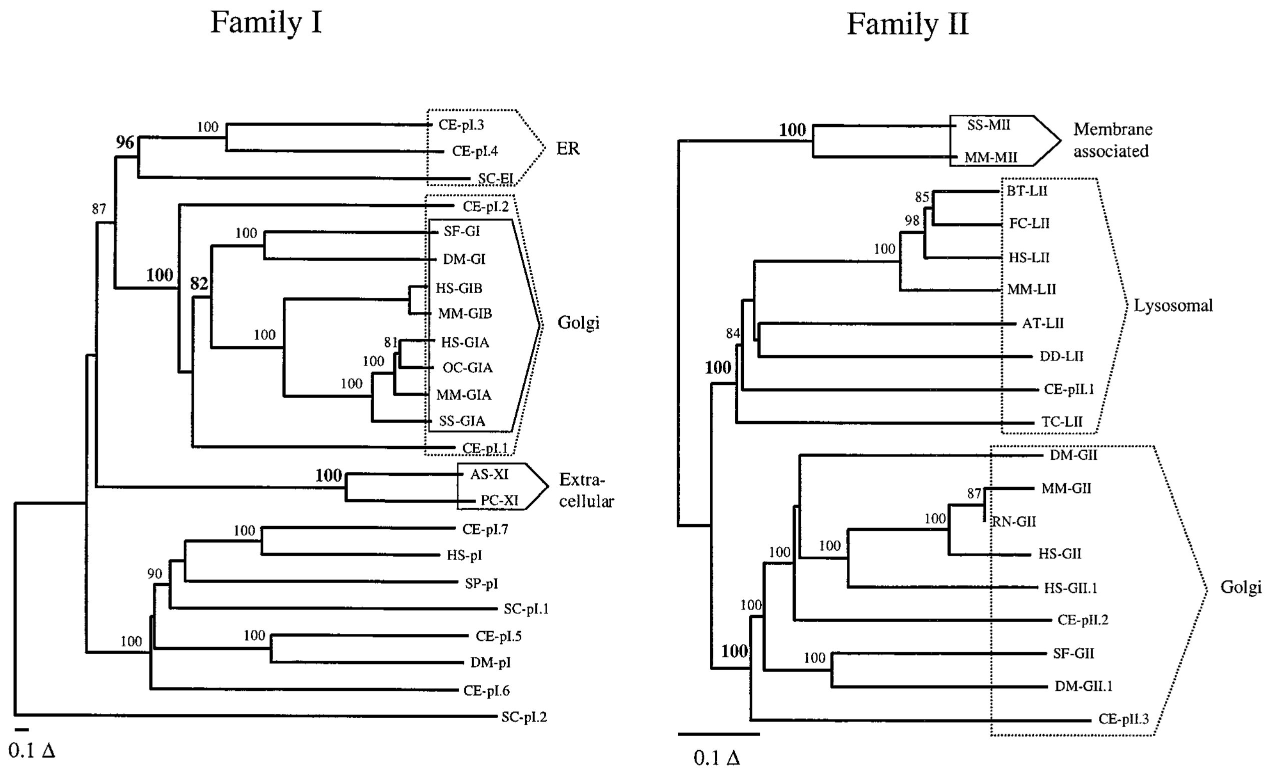


FIG. 4.—Within-family (I and II) phylogenetic reconstructions. Phylogenies were reconstructed using CLUSTAL W alignments as described in *Materials and Methods.* Taxa name abbreviations are the same as in figure 1. Monophyletic groups of sequences with similar cellular compartmentalizations are boxed, and the identities of the compartments are indicated. A solid line surrounds groups that include all characterized sequences, and a dashed line surrounds groups that include some putative sequences. Bootstrap values are shown above the branches. Scale bars indicating the relative amounts of change (Δ) along branches are shown.

nosidase sequences reported here are proposed to make up a α-mannosidase clan, or superfamily, based on the combination of this sequence evidence and the functional analogy of the α-mannosidase enzymes.

Gene Duplication and Functional Diversification

Gene duplication is a critical step in generating the functional diversification necessary for the evolution of complex organisms (Ohta 1991). According to the generally accepted view of gene duplication and evolution, the redundancy created by duplication allows paralogous gene copies to evolve new functions (Ohno 1970). However, once a paralog acquires a newly evolved function that enhances the fitness of its host, this function is likely to be constrained by negative selection (Goodman, Moore, and Matsuda 1975). One prediction of this hypothesis is that among the members of a gene family, orthologous copies are likely to encode the same functions, while paralogs will encode diverse functions. The α-mannosidase superfamily, with its abundance of functionally characterized sequences, provides an ideal system to test this prediction. Paralogous α-mannosidase sequences are known to encode slightly different biochemical activities. For example, among clade I, some sequences are ER-specific and some are Golgi-specific. If these discrete compartmentalized activities have evolved subsequent to duplication in the manner proposed above, then they should appear monophyletic when mapped onto a phylogeny of the sequences that encode them.

Within-family (clade I and II) α-mannosidase phylogenies were used to test this prediction. Relatively high levels of amino acid sequence homology within families allowed the alignment of MIRs in addition to the OSM. The inclusion of MIRs increased the phylogenetic resolution within families. Within-family phylogenies based on both OSM and MIR sequences showed topologies that were virtually identical (fig. 4) to the global α-mannosidase tree (fig. 1) based solely on the OSM alignment. The placement of only one sequence within each tree differed between the within-family and the among-families phylogenies. These results indicate that the OSM likely records an accurate phylogenetic history of the α-mannosidase superfamily despite the fact that it includes only a subset of the total sequence data. The increased resolution afforded by the inclusion of the MIR sequences manifested itself in a general increase in bootstrap support for the within-family trees. In both family I and family II, α-mannosidase sequences that encode enzymes with the same cellular compartmentalization group together in well-supported clades (fig. 4). These data support the hypothesis of gene duplication followed by functional diversification of paralogs and subsequent canalization of activity among orthologs. The presence of putative sequences in these clades suggests that these as yet uncharacterized se-

Table 2
α-Mannosidase Clade IA-IB Nonsynonymous (K_a) and Synonymous (K_s) Substitution Rates (×100)

	K_a	K_s	K_a/K_s
Extant sequences[a]	18.40	65.21	0.282
Node A–B[b]	17.57	43.12	0.407
Node A–HS-GIA	5.09	21.07	0.242
Node A–MM-GIA	4.65	22.74	0.204
Node B–HS-GIB	2.46	18.44	0.133
Node B–MM-GIB	2.99	19.79	0.151

NOTE.—K_a and K_s were calculated as described in *Materials and Methods*.

[a] Mean values were determined from comparisons among all extant sequences.

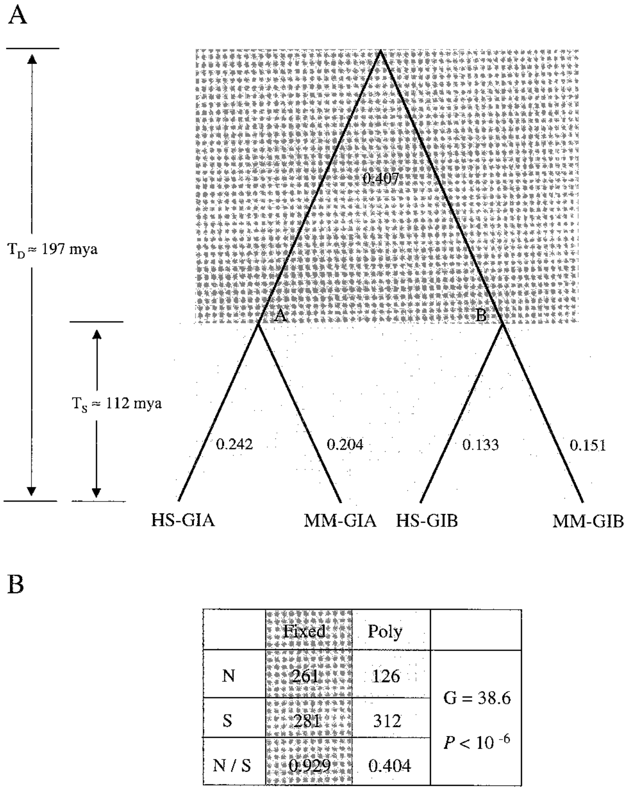
[b] Ancestral nodes as shown in figure 5.

quences will prove to have the same cellular compartmentalization patterns as their close relatives.

Positive Darwinian Selection After α-Mannosidase IA-IB Duplication

It is clear from the phylogenetic distribution of α-mannosidase cellular compartmentalization variants that gene duplication followed by functional diversification has been a pivotal mode of α-mannosidase superfamily evolution. In order to analyze in detail the molecular evolutionary dynamics of gene duplication and subsequent functional diversification, it is advantageous to investigate cases of relatively recent gene duplications. Recent duplications provide tractable levels of nucleotide polymorphism between paralogs and facilitate the accurate calculation of nonsynonymous and synonymous rates of substitution (i.e., avoid saturation of synonymous substitutions). The relatively shallow branch length between the α-mannosidase Golgi IA and IB clades (figs. 1 and 4) indicates that this is the most recent gene duplication yet detected in the α-mannosidase superfamily. α-Mannosidase IA- and IB-encoded enzymes have similar functions but differ in aspects of their substrate specificities and the structures of the hydrolysis products produced by their activity (Lal et al. 1998; Moremen, Trimble, and Herscovics 1994). The two genes also differ in their expression patterns. α-Mannosidase IA shows ubiquitous expression, while IB is primarily expressed in the placenta (unpublished data). The patterns of nucleotide polymorphism among the human and mouse IA and IB genes were analyzed in order to evaluate the role natural selection has played in IA-IB enzyme functional diversification subsequent to gene duplication.

Relative rates of K_a and K_s can yield improtant clues as to the nature of selection acting to shape nucleotide variation. A higher rate of K_a than K_s ($K_a/K_s > 1$) is generally considered unequivocal evidence of positive Darwinian selection (Kimura 1983; Hughes and Nei 1988; Sharp 1997). Levels of K_a and K_s for the IA-IB clade were analyzed to evaluate the role of selection following gene duplication (table 2). Comparison of extant IA-IB sequences shows an average $K_a/K_s < 1$. Such a pattern of variation demonstrates the prevalence of negative selection due to functional constraint on IA and IB amino acid sequences. This result is consistent with the expectations of the neutral theory (Kimura 1983) and is not surprising when the overall conservation and functional importance of α-mannosidase activity is considered.



	Fixed	Poly	
N	261	126	G = 38.6 $P < 10^{-6}$
S	281	312	
N / S	0.929	0.404	

FIG. 5.—*A*, Distribution of nucleotide variation on the Golgi α-mannosidase IA-IB phylogeny. The phylogeny was reconstructed as described in *Materials and Methods.* Taxon names are the same as in figure 1, and ancestral nodes (A and B) are indicated. Time since duplication (T_D) and time since speciation (T_S) were determined as described in *Materials and Methods.* Numbers next to the branches correspond to K_a/K_s for the branch (table 2). The dark gray shading represents the lineage since duplication and before speciation. Functional diversification of IA-IB paralogs occurred along this partition of the tree. The light gray shading represents the orthologous lineages subsequent to speciation. Evolution along this partition of the tree has been dominated by functional constraint. *B,* Results of the modified McDonald-Kreitman test comparing the relative amounts of nonsynonymous (*N*) and synonymous (*S*) change between the two phylogenetic partitions. Fixed substitutions (dark gray) occurred between paralogous lineages, and polymorphic substitutions (light gray) occurred among orthologous lineages.

Evaluation of extant sequences does not take into account the full historical context of the nucleotide substitution process. Nonsynonymous substitutions have been shown to accelerate during a period of functional differentiation following gene duplication (Li and Gojobori 1983; Ohta 1994). During this phase of evolution, positive selection may predominate. Once a new function has evolved, the changes involved in the emergence of the novel activity will be constrained by negative selection (Goodman, Moore, and Matsuda 1975). To evaluate the nature of nucleotide changes that occurred just subsequent to duplication separate from substitutions after the human-mouse speciation, ancestral IA and IB nucleotide sequences were inferred (fig. 5). Ancestral sequences provided for the partitioning of K_a/K_s on individual branches of the IA-IB tree. Previously, this approach revealed evidence of ancient episodes of adaptive evolution (Messier and Stewart 1997; Zhang, Rosenberg, and Nei 1998). In the present case, for every branch on the IA-IB tree, $K_a/K_s < 1$ (fig. 5). Thus, there is no unequivocal evidence for positive selection. However, the K_a/K_s ratio for the internal branch connecting the ancestral IA-IB nodes is higher than the ratios for the terminal branches within the IA and IB clades (fig. 5). This result indicates a possible postduplication increase in K_a due to positive selection. A number of studies have interpreted such a relative increase in K_a as evidence for positive selection following gene duplication (Long and Langley 1993; Ohta 1994; Schmidt et al. 1997). However, it is also possible that the increase in K_a is due to a period of decreased functional constraint (i.e., negative selection) after duplication.

To further evaluate the role of selection following gene duplication, a statistical analysis of the phylogenetic distribution of α-mannosidase IA-IB nucleotide changes based on the idea of the McDonald-Kreitman test (McDonald and Kreitman 1991) was used (as in Cirera and Aguade 1998). $K_a/K_s > 1$ is an extremely stringent criterion for evidence of positive selection (Wolfe and Sharp 1993). Under a scenario of accelerated adaptive evolution followed by negative selection, positive selection may predominate for only a fraction of the evolutionary history of a given lineage. Furthermore, the substitutions that are favored by positive selection likely represent a minority of the total sites. Therefore, discontinuous episodes of positive selection likely occur against a constant backdrop of negative selection that can easily obscure evidence of their existence. The McDonald-Kreitman test is sensitive in that it can provide evidence of adaptive evolution when $K_a/K_s < 1$ (Sharp 1997). In the analysis performed here based on the McDonald-Kreitman test, changes are partitioned on the IA-IB tree (fig. 5) to before (fixed) and after (polymorphic) speciation. In other words, sites that showed any variation within one or both of the paralogous groups (i.e., the HS-GIA, MM-GIA and/or the HS-GIB, MM-GIB group in fig. 5) are considered polymorphic, while sites with no variation within groups that differ between groups are considered fixed. According to the neutral theory, the ratio of nonsynonymous (*N*) to synonymous (*S*) changes should be the same for both classes of change (polymorphic and fixed). A *G*-test with Williams correction was used to compare polymorphic and fixed *N*/*S* (fig. 5). This allowed a test for evidence of positive selection following gene duplication. There is a significant departure from the expectations of neutrality due to an excess of fixed *N* changes. These data are consistent with the relative increase in K_a between ancestral IA and IB nodes and may be due to the fixation of *N* changes after duplication by positive Darwinian selection for functional diversification.

Conclusions

The tremendous accumulation of genomic sequence data provides the opportunity for increased resolution and power in molecular evolutionary studies. This potential has been exploited in the present analysis of the α-mannosidase enzyme superfamily. The combination of sensitive lineage-specific searches and a motif-based alignment approach has enabled a substantial expansion of the known α-mannosidase sequence space and revealed the presence of a new family of proteins. Consideration of functional specificity of characterized α-mannosidase sequences in a phylogenetic context indicated that functional diversification subsequent to gene duplication is a hallmark of α-mannosidase superfamily evolution. In at least one case, there is evidence that positive Darwinian selection may have acted following gene duplication. Currently, the distance between nearest paralogs prevents analysis of any other groups for a similar pattern of adaptive diversification. However, it is certainly possible that positive selection has played a role in the functional diversification of other α-mannosidase paralogs. A more robust sampling of sequences within closely related paralogous lineages would facilitate tests of this hypothesis. This is also true for the IA-IB clade characterized here. The time elapsed since the IA-IB duplication (fig. 5) suggests that virtually all mammals should have copies of each of these paralogs (barring loss). A denser IA-IB tree could facilitate a precise determination of the timing and nature of molecular adaptive events.

LITERATURE CITED

CAMIRAND, A., A. HEYSEN, B. GRONDIN, and A. HERSCOVICS. 1991. Glycoprotein biosynthesis in Saccharomyces cerevisiae. Isolation and characterization of the gene encoding a specific processing alpha-mannosidase. J. Biol. Chem. **266**: 15120–15127.

CIRERA, S., and M. AGUADE. 1998. Molecular evolution of a duplication: the sex-peptide (Acp70A) gene region of Drosophila subobscura and Drosophila madeirensis. Mol. Biol. Evol. **15**:988–996.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in protiens. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, D.C.

DEWALD, B. and O. TOUSTER. 1973. A new α-D-mannosidase occurring in Golgi membranes. J. Biol. Chem. **248**:7223–7233.

EADES, C. J., A. M. GILBERT, C. D. GOODMAN, and W. E. HINTZ. 1998. Identification and analysis of a class 2 alphamannosidase from Aspergillus nidulans. Glycobiology **8**: 17–33.

EERNISSE, D. J. 1992. DNA translator and aligner: HyperCard utilities to aid phylogenetic analysis of molecules. Comput. Appl. Biosci. **8**:177–184.

FITCH, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. **19**:99–113.

GOODMAN, M., G. W. MOORE, and G. MATSUDA. 1975. Darwinian evolution in the genealogy of haemoglobin. Nature **253**:603–608.

HALDANE, J. B. S. 1932. The causes of evolution. Longmans and Green, London.

HENIKOFF, S., E. A. GREENE, S. PIETROKOVSKI, P. BORK, T. K. ATTWOOD, and L. HOOD. 1997. Gene families: the taxonomy of protein paralogs and chimeras. Science **278**:609–614.

Henrissat, B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. **280**: 309–316.

———. 1998. Glycosidase families. Biochem. Soc. Trans. **26**: 153–156.

Henrissat, B., and A. Bairoch. 1993. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. **293**:781–788.

———. 1996. Updating the sequence-based classification of glycosyl hydrolases. Biochem. J. **316**:695–696.

Henrissat, B., and G. Davies. 1997. Structural and sequence-based classification of glycoside hydrolases. Curr. Opin. Struct. Biol. **7**:637–644.

Henrissat, B., and A. Romeu. 1995. Families, superfamilies and subfamilies of glycosyl hydrolases. Biochem. J. **311**: 350–351.

Hudak, J., and M. A. McClure. 1999. A comparative analysis of computational motif-detection methods. Pp. 138–149 *in* R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein, and K. Lauderdale, eds. Pacific Symposium on Biocomputing '99. World Scientific, Singapore.

Hughes, A. L., and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167–170.

Kerscher, S., S. Albert, D. Wucherpfennig, M. Heisenberg, and S. Schneuwly. 1995. Molecular and genetic analysis of the Drosophila mas-1 (mannosidase-1) gene which encodes a glycoprotein processing alpha 1,2-mannosidase. Dev. Biol. **168**:613–626.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, New York.

Kumar, S., and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917–920.

Lal, A., P. Pang, S. Kalelkar, P. A. Romero, A. Herscovics, and K. W. Moremen. 1998. Substrate specificities of recombinant murine Golgi alpha1, 2-mannosidases IA and IB and comparison with endoplasmic reticulum and Golgi processing alpha1,2-mannosidases. Glycobiology **8**:981–995.

Lal, A., J. S. Schutzbach, W. T. Forsee, P. J. Neame, and K. W. Moremen. 1994. Isolation and expression of murine and rabbit cDNAs encoding an alpha 1,2-mannosidase involved in the processing of asparagine-linked oligosaccharides. J. Biol. Chem. **269**:9872–9881.

Li, W. H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

Li, W. H., and T. Gojobori. 1983. Rapid evolution of goat and sheep globin genes following gene duplication. Mol. Biol. Evol. **1**:94–108.

Liao, Y. F., A. Lal, and K. W. Moremen. 1996. Cloning, expression, purification, and characterization of the human broad specificity lysosomal acid alpha-mannosidase. J. Biol. Chem. **271**:28348–28358.

Long, M., and C. H. Langley. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science **260**:91–95.

McClure, M. A. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol. Biol. Evol. **8**: 835–856.

McClure, M. A., and J. Kowalski. 1999. The effects of ordered-series-of-motifs anchoring and sub-class modeling on the generation of HMMs representing highly divergent protein sequences. Pp. 162–170 *in* R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein, and K. Lauderdale, eds. Pacific Symposium on Biocomputing '99. World Scientific, Singapore.

McClure, M. A., T. K. Vasi, and W. M. Fitch. 1994. Comparative analysis of multiple protein-sequence alignment methods. Mol. Biol. Evol. **11**:571–592.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature **351**:652–654.

Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151–154.

Moremen, K. W., R. B. Trimble, and A. Herscovics. 1994. Glycosidases of the asparagine-linked oligosaccharide processing pathway. Glycobiology **4**:113–125.

Muller, H. J. 1935. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. Genetics **17**:237–252.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

Neuwald, A. F., J. S. Liu, D. J. Lipman, and C. E. Lawrence. 1997. Extracting protein alignment models from the sequence database. Nucleic Acids Res. **25**:1665–1677.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, New York.

Ohta, T. 1991. Multigene families and the evolution of complexity. J. Mol. Evol. **33**:34–41.

———. 1994. Further examples of evolution by gene duplication revealed through DNA sequence comparisons. Genetics **138**:1331–1337.

Rozas, J., and R. Rozas. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13**:307–311.

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

Schmidt, T. R., S. A. Jaradat, M. Goodman, M. I. Lomax, and L. I. Grossman. 1997. Molecular evolution of cytochrome c oxidase: rate variation among subunit VIa isoforms. Mol. Biol. Evol. **14**:595–601.

Sharp, P. M. 1997. In search of molecular Darwinism. Nature **385**:111–112.

Swofford, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer, Sunderland, Mass.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. Science **278**:631–637.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Tulsiani, D. R. P., S. C. Hubbard, P. W. Robbins, and O. Touster. 1982. α-D-mannosidases of rat liver Golgi membranes. J. Biol. Chem. **257**:3660–3668.

Vandersall-Nairn, A. S., R. K. Merkle, K. O'Brien, T. N. Oeltmann, and K. W. Moremen. 1998. Cloning, expression, purification, and characterization of the acid alpha-mannosidase from Trypanosoma cruzi. Glycobiology **8**: 1183–1194.

Wolfe, K. H., and P. M. Sharp. 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. **37**:441–456.

Zhang, J., H. F. Rosenberg, and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**:3708–3713.

Pierre Capy, reviewing editor

Accepted November 1, 1999