J Mol Evol (2010) 70:313–324
DOI 10.1007/s00239-010-9320-8

Lineage-Specific Duplication and Loss of Pepsinogen Genes in Hominoid Evolution

Yuichi Narita · Sen-ichi Oda · Osamu Takenaka · Takashi Kageyama

Received: 7 August 2009 / Accepted: 8 January 2010 / Published online: 27 March 2010

Abstract Fourteen different pepsinogen-A cDNAs and one pepsinogen-C cDNA have been cloned from gastric mucosa of the orangutan, *Pongo pygmaeus*. Encoded pepsinogens A were classified into two groups, i.e., types A1 and A2, which are different in acidic character. The occurrence of 9 and 5 alleles of A1 and A2 genes (at least 5 and 3 loci), respectively was anticipated. Respective orthologous genes are present in the chimpanzee genome although their copy numbers are much smaller than those of the orangutan genes. Only A1 genes are present in the human probably due to the loss of the A2 gene. Molecular phylogenetic analyses showed that A1 and A2 genes diverged before the speciation of great hominoids. Further reduplications of respective genes occurred several times in the orangutan lineage, with much higher frequencies than those occurred in the chimpanzee and human lineages. The rates of non-synonymous substitutions were higher than those of synonymous ones in the lineage of A2 genes, implying the contribution of the positive selection on the encoded enzymes. Several sites of pepsin moieties were indeed found to be under positive selection, and most of them locate on the surface of the molecule, being involved in the conformational flexibility. Deduced from the known genomic structures of pepsinogen-A genes of primates and other mammals, the duplication/loss were frequent during their evolution. The extreme multiplication in the orangutan might be advantageous for digestion of herbaceous foods due to the increase in the level of enzymes in stomach and the diversification of enzyme specificity.

Keywords Pepsinogen · Hominoid · Gene duplication · Orangutan · Positive selection

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9320-8) contains supplementary material, which is available to authorized users.

Y. Narita · T. Kageyama
Center for Human Evolution Modeling Research, Primate Research Institute, Kyoto University, Inuyama 484-8506, Japan

S. Oda
Laboratory of Animal Management, Nagoya University School of Bio-Agricultural Sciences, Nagoya 464-0852, Japan

O. Takenaka
Department of Cellular and Molecular Biology, Primate Research Institute, Kyoto University, Inuyama 484-8506, Japan

Y. Narita (✉)
Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland
e-mail: yuichi.narita@fmi.ch

Introduction

Pepsinogens are zymogens of pepsins, the aspartic proteinases in vertebrates. They are synthesized and secreted from gastric mucosa. They can be activated into pepsins autocatalytically at acidic pH in the stomach, by releasing the activation segment from the N-terminal part. Pepsins work as the first digesting enzymes for proteins under the acidic condition in the stomach. To date, five major types are known in mammals, namely, pepsinogens A, B, C, F and Y (Foltmann 1981; Kageyama 2002). They are evolved from a common ancestral intracellular aspartic proteinase like cathepsin E, through several gene duplication events (Carginale et al. 2004; Borrelli et al. 2006). Pepsinogen A is the major zymogen in adults of most mammals, and pepsinogens B and C are minor ones although they have important roles in carnivores and rodents, respectively

(Narita et al. 2001; Suchodolski et al. 2002; Narita et al. 2002; Feng et al. 2008). Pepsinogens F and Y (prochymosins) function in the newborn and infant mainly for digesting milk proteins (Kageyama et al. 1990; Foltmann 1992). The occurrence of multiple forms is known in respective zymogens, especially in pepsinogen A in some mammalian species. The multiplicity is known to be extreme in primates. Six and five types of pepsinogen A have been purified from gastric mucosa of the human (Samloff 1971) and Japanese monkey (Kageyama and Takahashi 1976), respectively. The number of multiple isoforms is extreme in the orangutan, 14 forms being identified in orangutan gastric mucosal extract (Narita et al. 2000). This number is the largest hitherto known. Since the multiplicity of pepsinogen A has not been evidenced in New World monkeys (Kageyama 2000), the genes for pepsinogen A are thought to have duplicated several times in the catarrhine lineage after separation from platyrrhines. Gene duplications in hominoids are the most recent evolutionary events and thus are among the most important to the evolution of the human and great apes (Fortna et al. 2004; Dumas et al. 2007).

Pepsinogens work at the first stage of protein digestion of foods in stomach, and therefore food habits of animals might affect significantly the expression and evolution of pepsinogens. Primates are known to be largely plant-eating species although animal foods are occasionally taken (Harding 1981). Small-sized monkeys like the common marmoset are known to eat frequently animal foods like insects. Great apes including the gorilla, orangutan, and chimpanzee prefer fruits along with bark, leaves, flowers, and a variety of insects (Galdikas 1988). Although, since environments might also affect significantly on the variety of foods taken, it may be difficult to define their food habits simply, they are frugivorous in most cases, being typical in the orangutan. Chimpanzees frequently eat mammalian meat, thus occasionally being the most omnivorous like humans (Finch and Stanford 2004). Plant-eating species are thought to need high amount of pepsins to digest effectively food proteins since non-protein materials such as cellulose are rich in plant foods. Indeed, the level of pepsinogen in gastric mucosa has been shown to be much higher in plant-eating species than those in omnivorous and carnivorous species (Kageyama 2002; Narita and Kageyama 2003) (Supplementary Fig. 1). The increase of the number of pepsinogen genes might have occurred in the lineage of plant-eating species, and gene duplication might work as the major forces for such processes. Since the genes for digestive enzymes such as amylase have shown to increase their copy numbers to adapt starch digestion in foods (Perry et al. 2007), adaptive/positively selected evolution might occur in pepsinogen genes due to the diet shifts of great apes and the human.

Fig. 1 The amino acid alignments of orangutan pepsinogens A (**a**) and C (**b**) with other primate pepsinogens. The amino acid sequences were deduced from cDNA sequences determined in this study. The *dots* represent the amino acid identity of the reference human pepsinogens A (A) and C(B). Human pepsinogen A is the product of 12.0 kb gene (Evers et al. 1989). Porcine pepsin A numbering is used. The sites of insertions and deletions are shown with i and d, respectively, in the numbering line. Deleted residues are indicated with dashes. The regions of the signal peptide and the activation segment are shown by deep and faint shading, respectively. Two catalytic aspartic acids are shown in the filled boxes. The GenBank/EMBL/DDBJ accession numbers of orangutan pepsinogens A and C are as follows. Pepsinogens A-13; AB458307, A-19;AB458306, A-41; AB458314, A-71:AB458313, A-17:AB458310, A-28:AB458309, A-59:AB458308, A-36:AB458311, A-50:AB458312, A-14:AB458304, A-15:AB458303, A-35:AB458305, A-43:AB458302, and A-75:AB458301. Pepsinogen C:AB458315

In the present study, molecular cloning of cDNAs for orangutan pepsinogens A and C was carried out, and 14 pepsinogen-A cDNAs and a single pepsinogen-C cDNA were isolated. Pepsinogen-A cDNAs are classified into two types, one being close to human counterparts and the other being specific for the ape lineage. Molecular evolutional analyses showed these two genes separated before speciation of apes, and the A2 gene might be lost in the human. The high multiplicity of pepsinogen-A genes in the orangutan is thought to be correlated with the food habit. The arrangement of pepsinogen genes in the genome of the apes and human, along with other mammalian reference species, was discussed, with special reference the duplication and loss of the pepsinogen genes in the orangutan lineage.

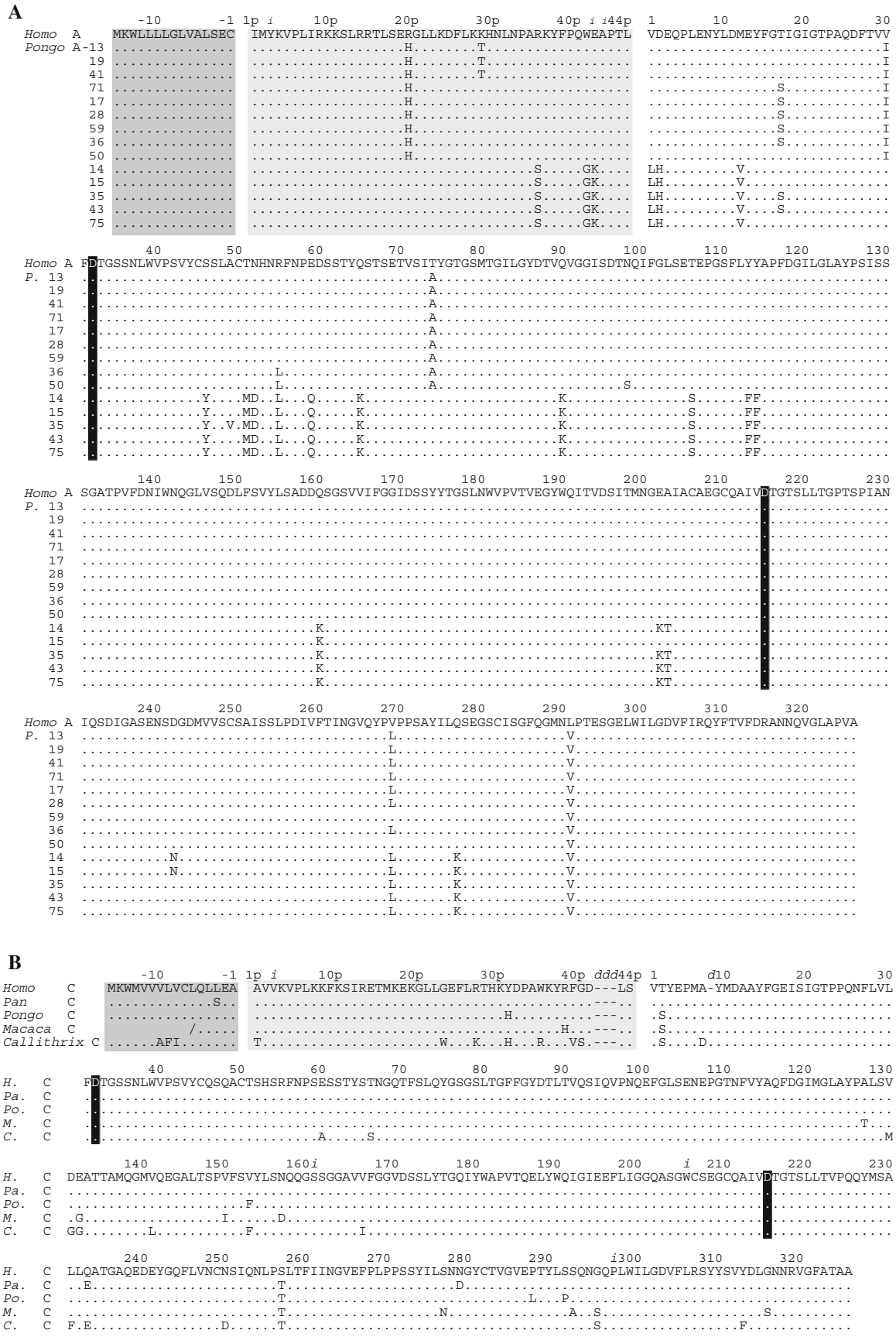
Materials and Methods

Materials

The mRNaid kit was purchased from BIO 101, Inc (Vista, CA), the SuperScript™ Choice System for cDNA-synthesis kit from Life Technologies, Inc. (Grand island, NY), λZapII DNA and in vitro packaging kit (Gigapack III Gold Packaging Extract) from Stratagene Cloning Systems (La Jolla, CA), and Thermo Sequenase™ cycle sequencing kit from Amersham (Cleveland, OH). All other chemicals were of reagent or analytical grade.

Specimens

A stomach of orangutan (*Pongo pygmaeus abelii*) was obtained from an individual (10 years-old male) died from pneumonia at the Primate Research Institute Kyoto University. The tissue was dissected immediately after death.


A
B

Molecular Cloning of Pepsinogen-A cDNAs and Nucleotide Sequence Determination

Total RNA was extracted and purified from gastric mucosa by the guanidium thiocyanate/cesium chloride centrifugation method (Chirgwin et al. 1979). Poly(A)-RNA was purified using the mRNaid kit. Double-stranded cDNA was prepared by the procedure of Gubler and Hoffman (1983) using the SuperScript™ Choice System. The cDNAs were ligated into the EcoR I site of λZapII. Pepsinogen cDNAs were screened by plaque hybridization with ^{32}P-labeled synthetic 45-base oligonucleotide as a probe. The sequence of the oligonucleotide probe was 5′-GCA GTA GAC TGA GGG CAC CCA CAG GTT GGA GCC GGT GTC AAA, which is complementary to the sequence that encodes the active-site region of one of the human pepsinogen-A isozymogens (Kageyama et al. 1990). The active-site sequence is highly conserved among pepsinogen subtypes and the probe has been successfully used to clone various types of pepsinogens (Kageyama et al. 1990; Narita et al. 1997; Kageyama 2000; Narita et el. 2002). Recombinant λZAP II clones that hybridized with both 45-base oligonucleotide were selected. The cDNA inserts of the hybridizing phages were subcloned and the whole nucleotide sequences were determined using the Thermo Sequenase™ cycle sequencing kit and DNA sequencer model LIC-4200S-1 from LI-COR Inc. (Lincoln, NE).

Sequence Data

The cDNA sequences of orangutan pepsinogens A and C were determined in this study. They were deposited in the GenBank/DDBJ/EMBL nucleotide sequence databases, and the accession numbers are given in the legend of Fig. 1. The cDNA sequences for pepsinogens A of the human (*Homo sapiens*) including those from a 12.0-kb gene (HPA-1) [J00279-J00287], and 15.0-kb [M26025] and 16.0-kb genes [M26032], Japanese monkey (*Macaca fuscata*) including pepsinogen A-1 (nucleotide JM196) [X59752], A-2/3 (JM72) [X59755], and A-4 (JM201) [X59753], rhesus monkey (*Macaca mulatta*) [M20788], and common marmoset (*Callithrix jacchus*) [AB038384], pig (*Sus scrofa*) [J04601], house musk shrew (*Suncus murinus*) [AB047243], bat (*Rhinolophus perdicus*) [AB047245], and dog (*Canis familiaris*) [AB047246], and for pepsinogens C of the human [J04443], Japanese monkey (JM642)[X59754], common marmoset [AB038385], cattle (*Bos taurus*) [XM_592361], rabbit (*Oryctolagus cuniculus*) [AB047250], and house musk shrew [AB047247] were obtained from the GenBank/DDBJ/EMBL nucleotide sequence databases and used in the molecular evolutional analyses. The genomic data of pepsinogens A for the human [NC_000011.8], chimpanzee (*Pan troglodytes*)[NC_006478], rhesus monkey [NC_007878.1], and dog [NC_006600.2] were obtained from the NCBI database. Although the sequences of the regions including pepsinogen-A genes have determined completely/nearly completely in the human and dog, they were determined partly in the chimpanzee and rhesus monkey.

Molecular Evolutional Analysis

The nucleotide sequences and deduced amino-acid sequences of pepsinogen-A cDNAs isolated in the present study were aligned together with those reported from other primates with the aid of DNASIS™-Mac ver.3.0. Phylogenetic trees were constructed by the neighbor-joining (NJ) method of ClustalW (Saitou and Nei 1987), the maximum parsimony (MP) method of PAUP 4.0 (Swofford 1998), and the maximum likelihood (ML) method of PUZZLE 5.2 (Strimmer and von Haeseler 1996). The sequence divergence was calculated among all pairwise comparisons following the Kimura 2-parameter model (Kimura 1980) for NJ, the HKY85 model (Hasegawa et al. 1985) for MP, and the SH (Schoniger and von Haeseler 1994) model for ML. Robustness of each node in the phylogenetic tree was assessed by bootstrap values (%), based on 1000 resampling (for NJ and MP), and quartet puzzling support values (%) based on 1000 quartet puzzling steps (for ML). Statistical test were carried out using the Kishino–Hasegawa test (Kishino and Hasegawa 1989) for MP and ML analyses, with the aid of PAUP 4.0 and PUZZLE 5.2. The rates of synonymous and non-synonymous nucleotide substitutions between primate lineages were compared using the relative-rate test program, RRTree (Robinson-Rechavi and Huchon 2000). The likelihood ratio tests of these substitutions for detecting positive selection were carried out according to Yang (1998) by the PAML program (Yang 2007).

Molecular Modeling

Tertiary structural model of orangutan pepsins were constructed with the program Modeller (Sali and Blundell 1993), using crystal structures of the complex between human pepsin A and a synthetic phosphonate inhibitor [PDB ID: 1QRP] and porcine pepsin A [PDB ID: 4CMS].

Results

Multiplicities for Orangutan Pepsinogens A and C

One hundred pepsinogen cDNA clones were randomly chosen from an orangutan gastric mucosal cDNA library. Since the oligonucleotide probe used in this study has a potential to hybridize various types of pepsinogens due to

conservativeness of the active-site nucleotide sequence, different types of pepsinogen cDNAs were anticipated to be chosen. Among 100 clones isolated, 95 clones were found to be pepsinogen-A clones and the rest 5 to be pepsinogen-C clones. The ratio of numbers of type-A and C clones is well consistent with that of the relative levels of pepsinogens in the orangutan stomach reported previously, in which pepsinogen A and pepsinogen C constituted 97, and 3%, respectively (Narita et al. 2000). Although no sequence heterogeneity was found in pepsinogen-C clones, 14 different nucleotide sequences were identified in pepsinogen-A clones (Supplementary Fig. 2). The pepsinogen-A clones were clearly separated into two groups. In respective groups, the sequences were very similar. The nucleotide substitutions of these two groups are thought to be the results of point mutations, showing little possibility of conversions between them. The deduced amino-acid sequences of these clones are given in Fig. 1. It is also clear that pepsinogens A were classified into two groups, one showing quite similar amino-acid sequences with that of human pepsinogen A and the other showing less similarities. The pepsin moieties of former types are highly acidic, containing 21 Asp and 14 Glu, but only 3 Arg and 1 His, and no Lys (isoelectric point = 3.16), for examples, in the case of A-13. While, the latter group members are less acidic, containing 20 Asp, 12 Glu, 2 Arg, 5 Lys, and 2 His (isoelectric point = 3.63), for example, in the case of A-14. These two types of pepsinogens are tentatively termed as pepsinogens A1 and A2, respectively. The difference in the number of Lys was thought to be essential to distinguish orangutan pepsinogens, being also called as low-lysine and high-lysine pepsinogens, respectively, in our

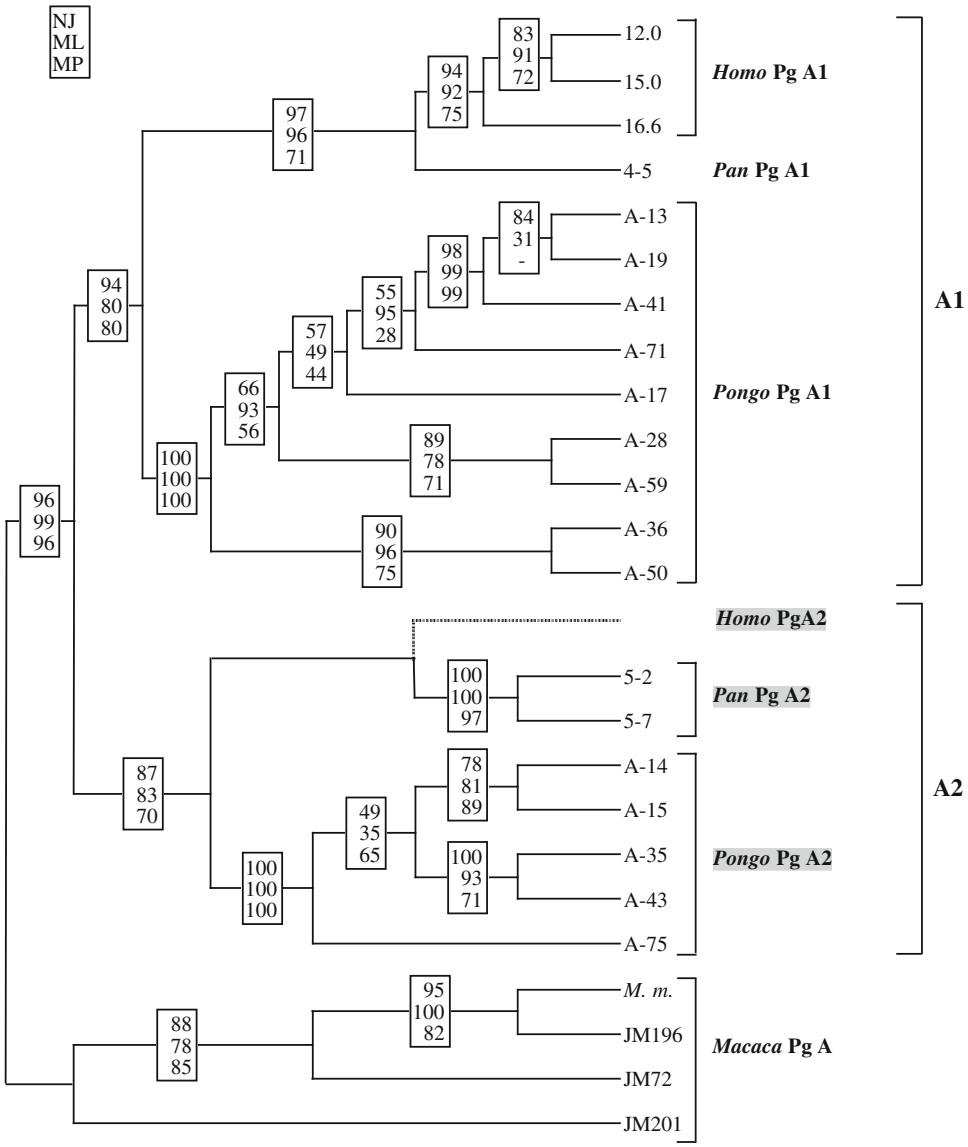


Fig. 2 Bootstrap values of a phylogenetic tree of catarrhine pepsinogens A inferred from their cDNA sequences. The tree topology is that of the NJ tree with bootstrap values (%) based on 1000 resamplings (*top*). Bootstrap values of the MP analysis (*bottom*), and quartet puzzling support values (%) based on 1000 quartet puzzling steps for ML (*middle*) are also shown. Human 12.0, 15.0, and 16.6 genes are referred from Evers et al. (1989). Chimpanzee gene terminology is based on the NCBI database. *M.m.* stands for rhesus monkey pepsinogen-A cDNA [M20788], and JM196 [X59752], JM72 [X59755], and JM201[XX59753] stand for cDNAs for Japanese monkey pepsinogen-A isozymogens. The gene for human pepsinogen A2 (*Homo* PgA2) is thought to be lost in the human lineage, and thus given with a dotted line

previous article (Narita et al. 2000). However, since the content of Lys is so variable between pepsins of other mammals, the terminology depending on the Lys content might not be appropriate. Among 14 pepsinogen-A clones, 9 and 5 clones were those of pepsinogens A1 and A2, respectively. Between A1 clones, A-13, A-19 and A-41, and A-17, A-28 and A-71 have the same amino acid sequences, respectively. The nucleotide sequences in the former 3 clones can be distinguished by 2 synonymous substitutions, while the latter 3 clones by 4 synonymous substitutions. As a result, 5 different types of A1 protein were identified. Between A2 clones, non-synonymous substitutions are frequent and, therefore, different A2 clones encoded different proteins. Totally, 10 different pepsinogen-A proteins were identified in the 100 cDNA clones chosen randomly. These results suggest that the occurrence of multiple genes for pepsinogens A1 and A2. Most of substitutions among them are point mutations and spread over the exons. No alternative splicing has been so far reported for pepsinogen genes (Evers et al. 1989). It might be also the case of orangutan pepsinogen-A genes. When we assume different alleles encode different pepsinogens, the occurrence of 9 and 5 alleles for pepsinogen A1 and A2 (thus, at least, 5 and 3 loci), respectively, is anticipated.

Phylogenetic Analyses of Primate Pepsinogens

We constructed the molecular evolutional tree of primate pepsinogens A based on their nucleotide sequences by NJ, ML, and MP methods. With any methods, the monophylies of orangutan pepsinogens A1 and A2 were clear with high bootstrap values. The trees constructed with three different methods showed that ape A1 and A2 genes diverged in the early period of ape evolution before the separation of each species (Fig. 2, Tree 1 of Table 1). Alternative tree topologies showing A1 and A2 genes diverged nearly simultaneously at the time of ape speciation (Tree 2) and orangutan pepsinogens A1 diverged first among ape pepsinogens, followed by the divergence of A2 genes (Tree 3) were, however, not significantly different from the Tree-1 topology with the ML and MP analyses (Table 1). Another tree showing that the orangutan, chimpanzee, and human diverged simultaneously, and A1 and A2 genes evolved independently in each species (Tree 4), the topology was denied with the MP method. Remaining alternative trees, showing that A1 and A2 genes separated independently in each species (Tree 5), and ape pepsinogen genes separated simultaneously (Tree 6) was denied strongly with both MP and ML methods. These results indicate that, although the divergence time of A1 and A2 genes is not dissolved completely, it is most probable that they diverged before the time of the orangutan divergence from an ancestral ape.

Contrary to pepsinogen A, only one species of pepsinogen-C cDNA was identified in an orangutan gastric cDNA library. This suggests the occurrence of a single gene for pepsinogen C. To date, the occurrence of the multiple genes for pepsinogen C has not evidenced (Kageyama 2002). Indeed, there are a single gene in the genomes of the human and chimpanzee. Phylogenetic trees of pepsinogens C of 5 primate species were constructed with the NJ, ML, and MP methods. They gave the same topologies, showing that, between apes, the orangutan diverged first, and the human and chimpanzee are closest relatives (Fig. 3).

Differences in the Nucleotide Substitution in Primate Lineages, and Test for Positive Selection

It was found that the branch lengths of pepsinogen-A trees constructed with the NJ and ML methods are largely different between pepsinogen-A taxa. Therefore, we compared nucleotide substitution rates at synonymous and non-synonymous sites between taxa (Table 2). The results

Table 1 Comparison of the log likelihood in ML analyses and number of steps in MP analyses for user-specified trees

Tree	ΔInL	ΔL
1. (((*Homo* A1, *Pan* A1), *Pongo* A1), (*Pan* A2, *Pongo* A2))	ML	MP
2. ((*Homo* A1, *Pan* A1), *Pongo* A1, (*Pan* A2, *Pongo* A2))	4.07 ± 4.05	3 ± 1.73
3. (((*Homo* A1, *Pan* A1), (*Pan* A2, *Pongo* A2)), *Pongo* A1)	2.17 ± 5.10	1 ± 2.23
4. ((*Homo* A1, *Pan* A1, *Pongo* A1), (*Pan* A2, *Pongo* A2))	7.96 ± 6.00	7 ± 2.65*
5. ((*Homo* A1, (*Pan* A1, *Pan* A2)), (*Pongo* A1, *Pongo* A2))	36.96 ± 17.82*	15 ± 5.73*
6. (*Homo* A1, *Pan* A1, *Pongo* A1, *Pan* A2, *Pongo* A2)	57.25 ± 16.42*	25 ± 5.52*

Comparison was carried out with the PUZZLE (Strimmer and von Haeseler 1996) and PAUP (Swofford 1998) programs. The maximum-likelihood and most-parsimonious trees are indicated as ML and MP, respectively. The differences in the log likelihood between ML and alternative trees and the differences in the number of steps between MP and alternative trees are shown with their SEs (after±). An asterisk indicates that the tree is significantly worse at $P = 0.05$ than the respective best trees with the Kishino–Hasegawa test (Kishino and Hasegawa 1989)

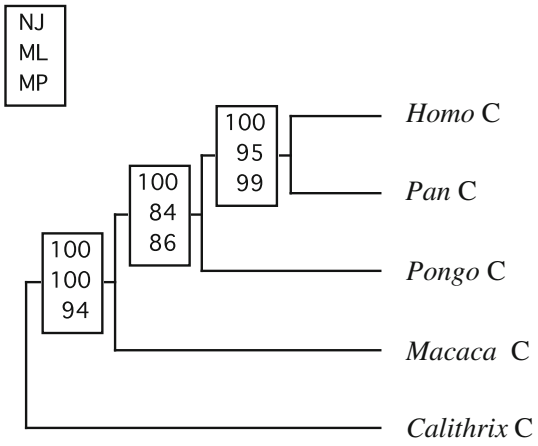


Fig. 3 Bootstrap values of the consensus phylogenetic tree of primate pepsinogens C

showed clearly that the rates at the non-synonymous site of the orangutan/chimpanzee A2 genes were significantly larger than those of A1 genes in any combinations. To know the excess of non-synonymous substitutions over synonymous ones in ape pepsinogens, the likelihood ratio test was carried out with the PAML program with 14 selected nucleotide sequences (Fig. 4). This small data set represents the four major clusters of Fig. 2, namely, (*Homo* & *Pan* Pg A1), (*Pongo* Pg A1), (*Pan* & *Pongo* Pg A2), and (*Macaca* Pg A). The log likelihood values of one-ratio and branch (free-ratio) models were −2612.77 and −2591.20, respectively. The difference was significant ($0.01 < P < 0.05$), indicating that d*N*/d*S* ratios are indeed different between lineages, where d*N* and d*S* are defined as the number of non-synonymous substitutions per non-synonymous site and that of synonymous substitutions per synonymous site, respectively. It was found that the d*N*/d*S* ratio only along the branch to the A2-gene lineage was 2.1, while values were less than one in other branches (Fig. 4). However, since the log likelihood value is –2608.13 when set the d*N*/d*S* ratio along the branch to 1.0, the difference was not significant ($0.05 < P < 0.1$), showing that positive selection along the branch was not supported. It is probable that non-synonymous substitutions are more frequent along the branch than along other branches.

High numbers of amino-acid substitutions were estimated to have occurred in the branch to A2 gene lineage with the PAML program, being consistent with the high d*N*/d*S* ratios of the branches. Apart from the lineage-specific positive selection, models of variable d*N*/d*S* ratios among sites were tested for the presence of site-specific positive selection with the same 14 selected nucleotide sequences. Site model using M3 (discrete)

Table 2 Relative rate tests between primate pepsinogens A

Taxa compared[a]	Δd*S*[b]	*P* value	Δd*N*[c]	*P* value
Homo A1–*Pan* A1	0.0029 ± 0.0090	0.7499	0.0024 ± 0.0024	0.3209
Homo A1–*Pan* A2	0.0000 ± 0.0124	0.9948	−0.0102 ± 0.0050	**0.0436***
Homo A1–*Pongo* A1	−0.0312 ± 0.0168	0.0639	0.0027 ± 0.0027	0.3275
Homo A1–*Pongo* A2	−0.0083 ± 0.0163	0.6099	−0.0132 ± 0.0054	**0.0143***
Homo A1–*Macaca* A	−0.0331 ± 0.0217	0.1270	0.0029 ± 0.0051	0.5672
Pan A1–*Pan* A2	−0.0028 ± 0.0111	0.7970	−0.0127 ± 0.0052	**0.0148***
Pan A1–*Pongo* A1	−0.0342 ± 0.0157	**0.0295***	0.0002 ± 0.0023	0.9123
Pan A1–*Pongo* A2	−0.0112 ± 0.0150	0.4552	−0.0157 ± 0.0057	**0.0059***
Pan A1–*Macaca* A	−0.0361 ± 0.0210	0.0855	0.0004 ± 0.0051	0.9259
Pongo A1–*Pan* A2	0.0313 ± 0.0170	0.0663	−0.0129 ± 0.0056	**0.0219***
Pongo A1–*Pongo* A2	0.0229 ± 0.0173	0.1857	−0.0159 ± 0.0059	**0.0073***
Pongo A1–*Macaca* A	−0.0019 ± 0.0235	0.9349	0.0002 ± 0.0049	0.9649
Pan A2–*Pongo* A2	−0.0083 ± 0.0133	0.5278	−0.0030 ± 0.0037	0.4235
Pan A2–*Macaca* A	−0.0332 ± 0.0212	0.1167	0.0131 ± 0.0057	**0.0213***
Pongo A2–*Macaca* A	−0.0248 ± 0.0201	0.2180	0.0161 ± 0.0061	**0.0082***

The relative-rate test was used to assess the heterogeneity of numbers of substitutions per site, which were estimated from the Kimura two-parameter method (Kimura 1980). The program RRTree (Robinson-Rechavi and Huchon 2000) was used. In hominoids, two types of pepsinogens A are distinguished with A1 and A2

[a] Mammalian pepsinogens including those of rabbit [M59237], pig [J04601], dog [AB047246], bat [AB047245], and shrew [AB047224] were used as outgroups

[b] Difference in the number of synonymous substitutions per synonymous sites

[c] Difference in the number of non-synonymous substitutions per non-synonymous sites

* Indicate that the rate of substitutions is significantly different between two taxa compared. Each value is highlighted with bold

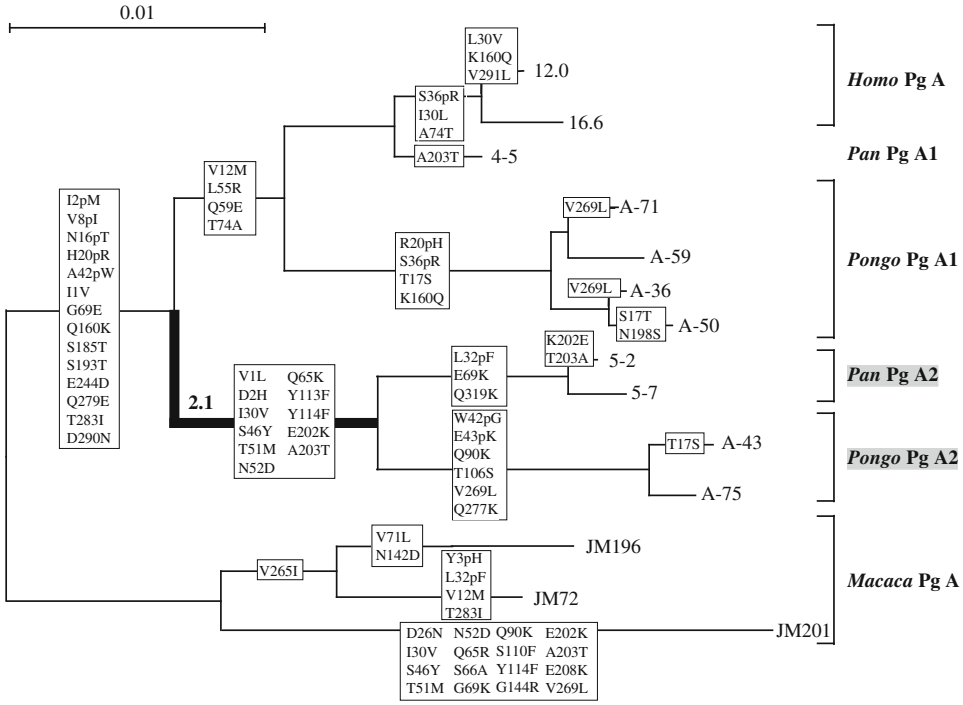


Fig. 4 Estimated amino-acid substitutions along the branches, and likelihood ratio tests applied to the pepsinogen-A evolution with the PAML program (Yang 2007). A subset of 14 pepsinogens A was chosen. An accepted phylogenetic tree was based on the NJ analysis. Evolutionary distance is given in the upper part of the tree. Only the branch to the lineage to pepsinogens A2 gave the d*N*/d*S* value over one, and highlighted with the thick line

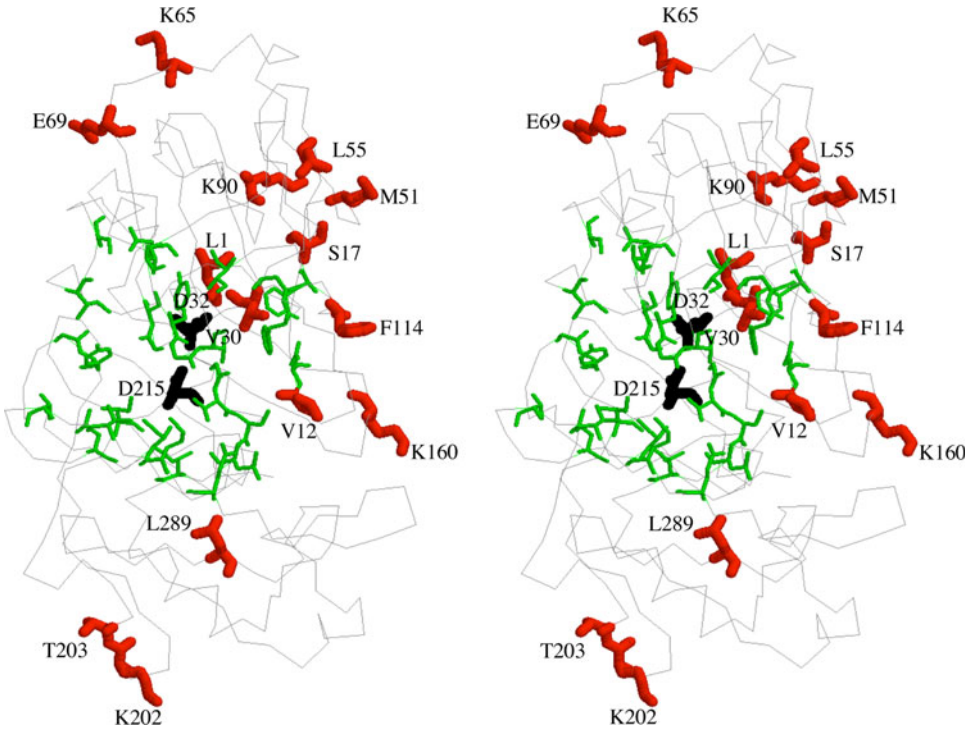


Fig. 5 Stereoview of the tertiary structures of a orangutan pepsinogen A2 (A-43). Catalytic aspartic acids (D32 and D215) are shown with thick black wireframes, active-site residues with thin green wireframes, and the residues under positive selection with thick red wireframes. Except these residues, only backbone atoms are given with gray. Residue numbers of positively selected residues and active-site Asps are given with porcine pepsin A numbering. (Color figure online)

(Yang 2007) was applied to the analysis. Resulting log likelihood and average d*N*/d*S* values were –2563.37 and 0.2417, respectively. Fourteen sites of the pepsin moiety were found to be under positive selection ($P < 0.05$) using the native empirical Bayes approach, including residue numbers 1, 12, 17, 30, 51, 55, 65, 69, 90, 114, 160, 202, 203, and 269. These sites are highlighted in the tertiary structure of typical orangutan pepsin A2 constructed by molecular modeling (Fig. 5). Four Lys residues unique to pepsins A2 were under positive selection, localizing on the surface of the molecule distal from the active site.

Genomic Structure of Ape Pepsinogens

In primates, the complete genome structure has been clarified in the human, and the analyses continue in other primates including the chimpanzee and rhesus monkey. In these species, the genomic structures around pepsinogen-A gene are rather well clarified. In the human, 3 pepsinogen-A genes localize in chromosome 11, repeated in tandem in the range of about 50 kb length (Fig. 6). The 5′- and 3′-neighbor genes are vacuolar protein sorting 37 homolog C (VPS37C), and von Willebrand factor C and EGF domains (VWCE). Such arrangement is common in the chimpanzee and rhesus monkey. In the rhesus monkey, the occurrence of an inactivated pepsinogen-F gene is obvious in the 5′-upstream of pepsinogen-A gene cluster, although such a gene was lost in the human and chimpanzee genomes. Respective active genes for pepsinogens A and F are typically found in the dog genome. The genome structures between VPS37C and VWCE of these 4 mammals were compared with Harr plot analyses (Supplementary Fig.3). High homologies of pepsinogen-A genes were obvious (data not shown). Alu repeats were found very frequently in primate genomes, especially abundantly in the 5′ upstream of the pepsinogen-A gene cluster. In mammals other than these 4 species, although it is common that pepsinogen-A and F genes localize between VPS37C and VWCE genes, the loss of A gene in the mouse genome [NC_000085.5] and several multiplications of F gene in the cattle genome [NC_007330.3] is obvious, suggesting the occurrence of dynamic duplications/losses of pepsinogen genes in this region between mammals.

Discussion

Lineage-Specific Gene Duplication

The occurrence of multiple genes for pepsinogen A in the orangutan was noteworthy, in contrast with a single gene for pepsinogen C. The level of pepsinogens A in gastric mucosa of the orangutan has been shown to be much higher than that of pepsinogen C (Narita et al. 2000), reflecting the difference in the number of genes. We here discuss the multiplicity of pepsinogen-A genes and their evolutionary background, since the repeated duplication in the orangutan lineage is quite unique. Orangutan pepsinogen-A genes were separated into two monophyletic genes, i.e., A1 and A2 genes. Respective orthologous genes are also found in the chimpanzee genome [NC_006478], while, in the human genome, only the A1 genes are present with the loss of the A2 gene [NC_000011.8] (Zelle et al. 1988; Evers et al. 1989). Phylogenetic analyses showed the divergence of the A1 and A2 genes occurred before speciation of great hominoids. In respective A1 and A2 clades, the divergence order of hominoids was (orangutan, (chimpanzee, human)), which is consistent with that reported with the analyses of various genes (Miyamoto et al. 1988; Goodman et al. 1994, 1998; O'hUigin et al. 2002; Steiper and Young 2006). Multiple copy numbers of A1 and A2 genes in the orangutan are thought to be generated by several-times gene-duplication events. The duplication frequency was much higher than those in human and chimpanzee, where only once or twice duplications have been anticipated (Zelle et al. 1988; Evers et al. 1989).

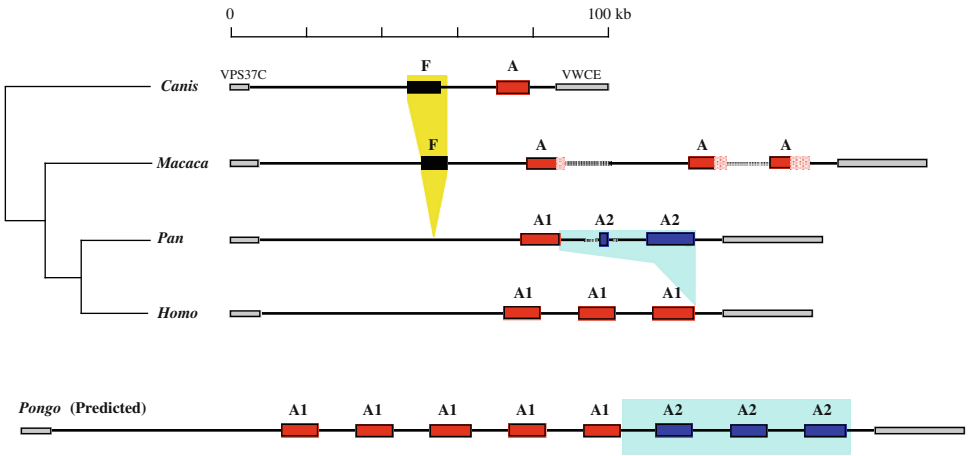


Fig. 6 Arrangement of pepsinogen-A genes in mammalian chromosomes. A and F stand for genes for pepsinogens A and F, respectively. In hominoids, A genes have diverged to A1 and A2 genes, the latter being shaded with blue. A2 genes are probably lost in the human. The numbers of these two genes in the orangutan are estimated based on the numbers of cDNAs isolated, being much larger than those of human and chimpanzee orthologues. Pepsinogen-F gene, the closest relative to pepsinogen-A gene (Kageyama 2002) is thought to be active in the dog, but inactive in the rhesus monkey, and lost in the chimpanzee and human. VPS37C and VWCE stand for vacuolar protein sorting 37 homolog C, and von Willebrand factor C and EGF domains, respectively, which locates in the nearest neighborhoods of pepsinogen-A genes in mammals. (Color figure online)

Site-Specific Positive Selection

Non-synonymous substitutions in the lineage of A2 genes were found to be frequent, giving the d*N*/d*S* ratio of 2.1 (Fig. 4). The value is markedly high given an average d*N*/d*S* ratio for genes of primates of 0.21 (Yang and Nielsen 1998). Although the value did not show the significant positive selection with likelihood ratio test, it strongly implies the contribution of positive selection. Indeed, positive selection acting on parts of the proteins was evident, highlighting several positively selected sites. The positively selected sites included 4 Lys residues unique to pepsins A2 that locate around the surface of the molecule distal from the active site. Since a distal Lys has been shown to affect enzymatic activity through the change in conformational flexibility in porcine pepsin A (Cottrell et al. 1995), similar modification of catalytic activities is anticipated in orangutan pepsins A2.

Physiological Significance of Pepsinogen Multiplicity

The multiplicity of pepsinogens A might be advantageous for food digestion in stomach. The occurrence of multiple forms of pepsinogens A has been reported in various mammals. Although the occurrence of multiple genes is the primary cause of multiplicity of pepsinogens as clarified in the human (Samloff 1971), Japanese monkey (Kageyama and Takahashi 1976), and rabbit (Kageyama and Takahashi 1984), post-translational modifications such as phosphorylation or glycosylation are also involved (Tang et al. 1973; Kageyama and Takahashi 1977). The expression of multiple genes might contribute to the increase in the pepsinogen level in stomach, which is favorable for gastric digestion. Indeed, the level of orangutan pepsinogens in stomach is highest between mammals (Narita and Kageyama 2003) (Supplementary Fig.1). Contrarily, pepsinogen genes have shown to be lost in platypus, a prototherian mammal, which has lost functional stomach during evolution (Ordonez et al. 2008). It is shown that different forms might have different hydrolytic specificities against substrate proteins/peptides (Foltmann 1981; Kageyama 2002). Human pepsin A isoforms have been shown to be generated from respective pepsinogens by different activation processes and give different pH-dependent activities (Athauda et al. 1989). Amino-acid substitutions of human pepsin A have changed its proteolytic specificity significantly (Kageyama 2004, 2006). From these instances, various proteins encoded by the multiple genes might have different specificities, being advantageous to the effective digestion. When we analyze the relationship between pepsinogen multiplicity and food habit between mammals, it is easy to say that the occurrence of multiple pepsinogens A are very frequent in herbivorous mammals including the orangutan (Narita et al. 2000), *Macaca* monkeys (Kageyama and Takahashi 1976), and rabbit (Kageyama and Takahashi 1984; Kageyama et al. 1990). The level decreases in omnivorous and carnivorous mammals (Narita et al. 2002; Kageyama 2002; Narita and Kageyama 2003) (Supplementary Fig. 1). High amounts of and a variety of pepsinogens might be needed in herbivorous mammals, since plants occasionally contain some substances that inhibit pepsin activity (Bankowska et al. 1998; Christeller et al. 1998). The orangutan is known to be herbivorous, eating leaf, fruit and bark frequently that are rich in non-digestive cellulose and plant fibers, the high level of gastric pepsins might be necessary for the digestion of food proteins efficiently (Galdikas 1988). Among great hominoids, the chimpanzee and human are better to be classified as the most omnivorous (Finch and Stanford 2004), thus possibly demanding less amount of gastric digestive enzymes.

Gene Duplication and Loss During Hominoid Evolution

The multiplicity of pepsinogen-A genes in the orangutan is thought to be the results of lineage-specific gene duplications. To date, lineage-specific increase/decrease in the copy number of various genes have been analyzed by cDNA array-based comparative genomic hybridization (Frazer et al. 2003; Fortna et al. 2004; Dumas et al. 2007). Human-lineage specific genes accounting for 84–134 genes increased their copy numbers significantly. These genes included those involved in brain function and endurance running, which have developed significantly in the human lineage (Dumas et al. 2007). In the orangutan, carbonic anhydrase genes amplified suggesting the affect on the physiology of the orangutan (Dumas et al. 2007). Regarding the food habit, the increase in the copy numbers of genes such as FLJ22004 and amylase genes have been reported in the gorilla (Fortna et al. 2004) and human (Perry et al. 2007), respectively, suggesting that dietary shifts affect the copy numbers of genes for digestive enzymes. The copy number of the salivary amylase gene (AMY1) has been shown to be correlated positively with the salivary amylase protein level (Perry et al. 2007). When compared genome structures of pepsinogens A genes for various mammals, the duplication and loss of genes were found to be frequent. Pepsinogen-A gene(s) and its closest relative pepsinogen-F gene(s) locate between VPS37C and VWCE genes commonly in mammalian chromosomes (Fig. 6). Although the primate ancestral genome structure might be close to the dog genome [NC_006600.2], it changed significantly during primate evolution with the repeated duplications of the A gene and the loss of the F gene. In primates, the separation of the A gene to the A1 and A2 genes in hominoids, and the extreme multiplicity of

A1 and A2 genes in the orangutan is noteworthy. These change in gene numbers of pepsinogens A might show that the genomic structures of genes regarding food digestion have changed very rapidly in primates, adapting to the diversity of food habit.

Conclusions

Before great-hominoid speciation, a major duplication of pepsinogen-A genes occurred to generate A1 and A2 genes. Each gene reduplicated further several times in the lineage of the orangutan. Site-specific positive selection was found to occur only in the A2 genes at the sites of surface molecules of encoded pepsins, potentially affecting enzyme flexibility. The multiple pepsinogen-A genes might be advantageous to the effective food digestion in stomach of the orangutan.

Acknowledgments This study was supported in part by Grants-in-Aid for Scientific Research (19370102 to T.K.) and the Global Center of Excellence Program "Formation of a Strategic Base for Biodiversity and Evolutionary Research: from Genome to Ecosystem" from the Ministry of Education, Science, Sports and Culture of Japan, and by Grants for the co-operative research program of the Primate Research Institute, Kyoto University.

References

Athauda SB, Tanji M, Kageyama T, Takahashi K (1989) A comparative study on the NH_2-terminal amino acid sequences and some other properties of six isozymic forms of human pepsinogens and pepsins. J Biochem 106:920–927

Bankowska A, Roszkowska-Jakimiec W, Worowki K (1998) Inhibitors of pepsin, trypsin and chymotrypsin in seeds of plants consumed by humans and animals. I. Evaluation of pepsin, trypsin, and chymotrypsin inhibitors activity in seeds of 26 plant species. Rocz Akad Med Bialymst 43:278–286

Borrelli L, De Stasio R, Filosa S, Parisi E, Riggio M, Scudiero R, Trinchella F (2006) Evolutionary fate of duplicate genes encoding aspartic proteinases. Nothepsin case study. Gene 368:101–109

Carginale V, Trinchella F, Capasso C, Scudiero R, Riggio M, Parisi E (2004) Adaptive evolution and functional divergence of pepsin gene family. Gene 333:81–90

Chirgwin JM, Przybyla AE, MacDonald RJ, Rutter WJ (1979) Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry 18:5294–5299

Christeller JT, Farley PC, Ramsay RJ, Sullivan PA, Laing WA (1998) Purification, characterization and cloning of an aspartic proteinase inhibitor from squash phloem exudate. Eur J Biochem 254:160–167

Cottrell TJ, Harris LJ, Tanaka T, Yada RY (1995) The sole lysine residue in porcine pepsin works as a key residue for catalysis and conformational flexibility. J Biol Chem 270:19974–19978

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM (2007) Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res 17:1266–1277

Evers MP, Zelle B, Bebelman JP, van Beusechem V, Kraakman L, Hoffer MJ, Pronk JC, Mager WH, Planta RJ, Eriksson AW, Frants RR (1989) Nucleotide sequence comparison of five human pepsinogen A (PGA) genes: evolution of the PGA multigene family. Genomics 4:232–239

Feng S, Li W, Lin H (2008) Characterization and expression of the pepsinogen C gene and determination of pepsin-like enzyme activity from orange-spotted grouper (Epinephelus coioides). Comp Biochem Physiol B Biochem Mol Biol 149:275–284

Finch CE, Stanford CB (2004) Meat-adaptive genes and the evolution of slower aging in humans. Q Rev Biol 79:3–50

Foltmann B (1981) Gastric proteinases—structure, function, evolution and mechanism of action. Essays Biochem 17:52–84

Foltmann B (1992) Chymosin: a short review on foetal and neonatal gastric proteases. Scand J Clin Lab Invest 52(Suppl. 210):65–79

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, Karimpour-Fard A, Glueck D, McGavran L, Berry R, Pollack J, Sikela JM (2004) Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol 2:E207

Frazer KA, Chen X, Hinds DA, Pant PV, Patil N, Cox DR (2003) Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. Genome Res 13: 341–346

Galdikas BMF (1988) Orangutan diet, range, and activity at Tanjung Puting, Central Borneo. Int J Primatol 9:1–35

Goodman M, Bailey WJ, Hayasaka K, Stanhope MJ, Slightom J, Czelusniak J (1994) Molecular evidence on primate phylogeny from DNA sequences. Am J Phys Anthropol 94:3–24

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9: 585–598

Gubler U, Hoffman BJ (1983) A simple and very efficient method for generating cDNA libraries. Gene 25:263–269

Harding RSO (1981) An order of omnivores: nonhuman primate diets in the wild. In: Harding RSO, Teleki G (eds) Omnivorous primates. Columbia University Press, New York, pp 191–214

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Kageyama T (2000) New world monkey pepsinogens A and C, and prochymosins. Purification, characterization of enzymatic properties, cDNA cloning, and molecular evolution. J Biochem 127:761–770

Kageyama T (2002) Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development. Cell Mol Life Sci 59:288–306

Kageyama T (2004) Role of S'1 loop residues in the substrate specificities of pepsin A and chymosin. Biochemistry 43:15122–15130

Kageyama T (2006) Roles of Tyr13 and Phe219 in the unique substrate specificity of pepsin B. Biochemistry 45:14415–14426

Kageyama T, Takahashi K (1976) Pepsinogens and pepsins from gastric mucosa of Japanese Monkey. Purification and characterization. J Biochem 79:455–468

Kageyama T, Takahashi K (1977) The carbohydrate moiety of Japanese monkey pepsinogens. Its composition and site of attachment to protein. Biochem Biophys Res Commun 74:789–795

Kageyama T, Takahashi K (1984) Rabbit pepsinogens. Purification, characterization, analysis of the conversion process to pepsin and determination of the NH_2-terminal amino-acid sequences. Eur J Biochem 141:261–269

Kageyama T, Tanabe K, Koiwai O (1990) Structure and development of rabbit pepsinogens. Stage-specific zymogens, nucleotide sequences of cDNAs, molecular evolution, and gene expression during development. J Biol Chem 265:17031–17038

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29:170–179

Miyamoto MM, Koop BF, Slightom JL, Goodman M, Tennant MR (1988) Molecular systematics of higher primates: genealogical relations and classification. Proc Natl Acad Sci USA 85:7627–7631

Narita Y, Kageyama T (2003) Diversity of ape pepsinogen genes (in Japanese with English summary). Primate Res 19:125–133

Narita Y, Oda S, Moriyama A, Takenaka O, Kageyama T (1997) Pepsinogens and pepsins from house musk shrew, Suncus murinus: purification, characterization, determination of the amino-acid sequences of the activation segments, and analysis of proteolytic specificities. J Biochem 121:1010–1017

Narita Y, Oda S, Takenaka O, Kageyama T (2000) Multiplicities and some enzymatic characteristics of ape pepsinogens and pepsins. J Med Primatol 29:402–410

Narita Y, Oda S, Takenaka O, Kageyama T (2001) Phylogenetic position of Eulipotyphla inferred from the cDNA sequences of pepsinogens A and C. Mol Phylogenet Evol 21:32–42

Narita Y, Oda S, Moriyama A, Kageyama T (2002) Primary structure, unique enzymatic properties, and molecular evolution of pepsinogen B and pepsin B. Arch Biochem Biophys 404:177–185

O'hUigin C, Satta Y, Takahata N, Klein J (2002) Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. Mol Biol Evol 19:1501–1513

Ordonez GR, Hillier LW, Warren WC, Grutzner F, Lopez-Otin C, Puente XS (2008) Loss of genes implicated in gastric function during platypus evolution. Genome Biol 9:R81

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256–1260

Robinson-Rechavi M, Huchon D (2000) RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. Bioinformatics 16:296–297

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

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815

Samloff IM (1971) Pepsinogens, pepsins, and pepsin inhibitors. Gastroenterology 60:586–604

Schoniger M, von Haeseler A (1994) A stochastic model for the evolution of autocorrelated DNA sequences. Mol Phylogenet Evol 3:240–247

Steiper ME, Young NM (2006) Primate molecular divergence dates. Mol Phylogenet Evol 41:384–394

Strimmer K, von Haeseler M (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Suchodolski JS, Steiner JM, Ruaux CG, Boari A, Williams DA (2002) Purification and partial characterization of canine pepsinogen A and B. Am J Vet Res 63:1585–1590

Swofford DL (1998) PAUP*. Phylogenetic analysis using parsimony and other methods. Sinauer Associates, Sunderland, MA

Tang J, Sepulveda P, Marciniszyn J Jr, Chen KC, Huang WY, Tao N, Liu D, Lanier JP (1973) Amino-acid sequence of porcine pepsin. Proc Natl Acad Sci U S A 70:3437–3439

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z (2007) PAML 4: a phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418

Zelle B, Evers MP, Groot PC, Bebelman JP, Mager WH, Planta RJ, Pronk JC, Meuwissen SG, Hofker MH, Eriksson AW, Frants RR (1988) Genomic structure and evolution of the human pepsinogen A multigene family. Hum Genet 78:79–82