Phylogenetic and Exon–Intron Structure Analysis of Fungal Subtilisins: Support for a Mixed Model of Intron Evolution

Chengshu Wang,[1] **Milton A. Typas,**[2] **Tariq M. Butt**[1]

[1] School of Biological Sciences, University of Wales Swansea, Swansea SA2 8PP, UK
[2] Department of Genetics and Biotechnology, Faculty of Biology, University of Athens, Athens 15701, Greece

Abstract. Phylogenetic and exon–intron structure analyses of intra- and interspecific fungal subtilisins in this study provided support for a mixed model of intron evolution: a synthetic theory of introns-early and introns-late speculations. Intraspecifically, there were three phase zero introns in *Pr1A* and its introns 1 and 2 located at the highly conserved positions were phylogentically congruent with coding region, which is in favor of the view of introns-early speculation, while intron 3 had two different sizes and was evolutionarily incongruent with coding region, the evidence for introns-late speculation. Noticeably, the subtilisin *Pr1J* gene from different strains of *M. ansiopliae* contained different number of introns, the strong evidence in support of introns-late theory. Interspecifically, phylogenetic analysis of 60 retrievable fungal subtilisins provided a clear relationship between amino acid sequence and gene exon–intron structure that the homogeneous sequences usually have a similar exon–infron structure. There were 10 intron positions inserted by highly biased phase zero introns across examined fungal subtilisin genes, half of these positions were highly conserved, while the others were species-specific, appearing to be of recent origins due to intron insertion, in favor of the introns-late theory. High conservations of positions 1 and 2 inserted by the high percentage of phase zero introns as well as the evidence of phylogenetic congruence between the evolutionary histories of intron sequences and coding region suggested that the introns at these two positions were primordial.

Key words: *Metarhizium anisopliae* — Subtilisin — Phylogenetic congruence — Exon–intron structure — Intron evolution

Introduction

The origin and evolution of spliceosomal introns are debatable since their discoveries, e.g., early or late, gain or loss. The dispute regarding the opposing introns-early (exon shuffling) versus introns-late (intron insertion) theory is very much alive (e.g., Stoltzfus et al. 1994; Logsdon 1998; O'Neill et al. 1998; de Souza et al. 1998; Roy et al. 2001; 2002; Wolf et al. 2001; Fedorova and Fedorov 2003; Roy 2003). The introns-early theory suggests that introns are ancient and present in the common ancestor of all life for exon shuffling to accelerate the formation of new genes (e.g., Hurst and McVean 1996; De Souza et al. 1998). The introns-late theory, however, proposes that ancient genes existed as uninterrupted exons, and introns were introduced during the course of gene evolution (e.g., Logsdon et al. 1995; Cho and Doolittle 1997; O'Neill et al. 1998). In recent years, the introns-early theory is becoming dominant (Fedor-

Correspondence to: Chengshu Wang; Department of Entomology, 4112 Plant Sciences Building, University of Maryland, College Park, MD 20742, USA; *email:* cwang4@umd.edu

Table 1. Background information on the examined strains with the sequenced *Pr1A* gene of *Metarhizium anisopliae*

No.	Isolate	Accession no.[a]	Original host	Geographical origin
1	ARSEF 2575	AJ416695	Coleoptera: Curculionidae	USA
2	ARSEF 820	AJ416688	Coleoptera: Curculionidae	France
3	5940	AB073327	Not known	Japan
4	990	AB073326	Not known	Japan
5	KM25	AB073345	Isolated from soil	Japan
6	511	AB073344	Lepidoptera: *Bombyx mori*	Japan
7	V275	AY389119*	Lepidoptera: *Cydia pomonella*	Austria
8	IMBST 9601	AY389120*	Coleoptera: *Melolontha melolontha*	Kramsach, Austria
9	V208	AY389121*	Orthoptera: Tettigonidae	Brazil
10	Ma1	AY389122*	Coleoptera: Ipidae	Guizhou, China
11	Ma5	AY389123*	Orthoptera: Tettigonioidae	Anhui, China
12	Ma6	AY389124*	Orthoptera: Tettigonioidae	Anhui, China
13	Ma9	AY389125*	Coleoptera: Scarabeidae	Anhui, China
14	PPRC-29	AY389126*	Coleoptera: *Pachnoda interrupta*	Ethiopia
15	V245	AY389127*	Hayfield soil	Finland
16	0-500B	AY389128*	Soil baited with *Tenebrio molitor*	Taiwan
17	687	AY389129*	Acari: *Ixodes ricinus*	Oxford, UK
18	IMI 330189	AY389130*	Orthoptera: locust	UK
19	CA-22	AY389131*	Avocado mulch baited with *Galleria*	California, USA
20	ESC-1	AY389132*	Blattodea: Blattidae	USA
21	Ma23	AY389133*	Coleoptera: *Conoderus* sp.	USA
22	ARSEF 3540	AY389134*	Baited with *Galleria* in soil	Vermont, USA
23	F-52	AY389135*	Homoptera: Aphididae	Taensa, USA

[a]Asterisks denote sequences from this study.

ova and Fedorov 2003; Roy 2003). However, the debate may last, with the evidence established from ongoing studies, though a synthetic theory is being considered compatible to make equal contributions to interpret intron origin/evolution (de Souza et al. 1998, 2003).

Serine protease subtilisins have been identified from bacteria, fungi, plant, invertebrates, and vertebrates and represent an ancient family of evolutionarily conserved proteins (Siezen and Leunissen 1997). Indeed, extracellular alkaline subtilisins of pathogenic fungi play an important role during the process of fungal pathogenesis. The subtilisin PR1A of *M. anisopliae* has been demonstrated to be one of the key virulence factors evident from which a transformant with multiple copies of the *Pr1A* gene exhibited an enhanced level of pathogenicity (St Leger et al. 1996), while a gene-deficient mutant lost its aggressiveness against an insect host (Wang et al. 2002). Similarly, the transformants of the nematode-trapping fungus *Arthrobotrys oligospora* developed a higher pathogenic ability when they contained the additional subtilisin gene *PII* (Åhman et al. 2002). In total, 11 subtilisin genes, *Pr1A* to *Pr1K*, have been identified from *M. anisopliae* (Bagga et al. 2004). In this study, phylogenetic and exon–intron structure analyses of retrievable fungal subtilisins were conducted at both intra- and interspecific levels and the findings support a mixed model of intron origins.

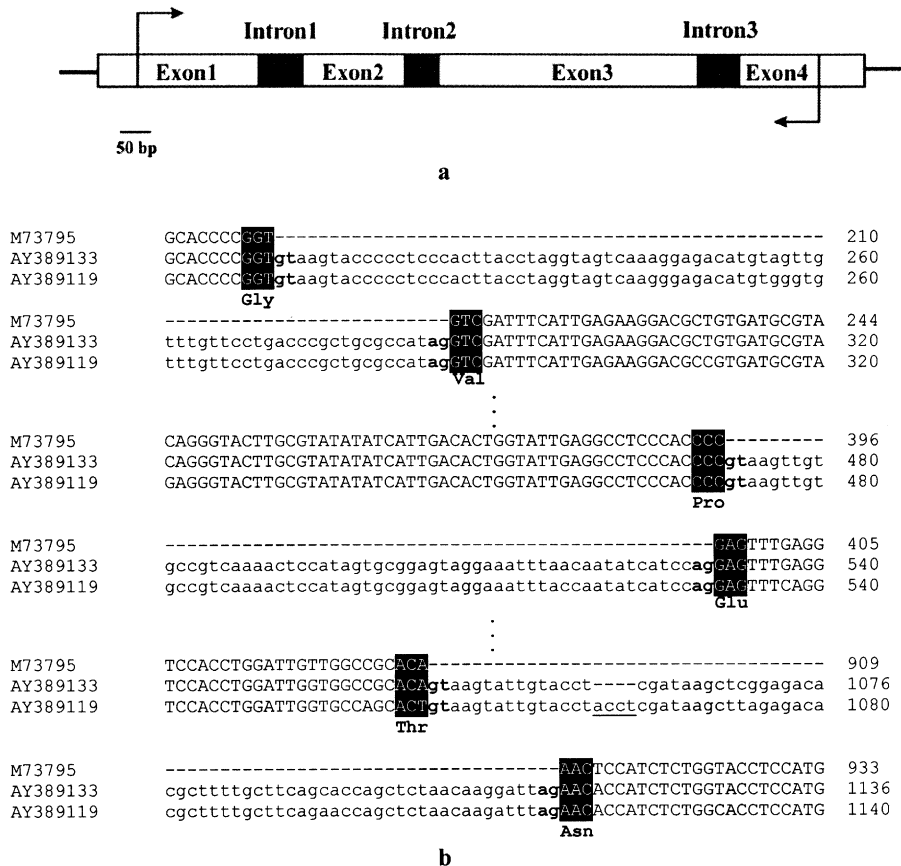## Materials and Methods

Gene Amplification

The primers for nested-PCR amplifications of the subtilisin gene *Pr1A* of *M. anisopliae* were described before (Wang et al. 2002). After analysis of the restriction map of a structure gene sequence (AJ416695) with the program Webcutter 2.0 (http://www.first-market.com/cutter/cut2.html), enzymes that had more than four cuttings were selected for restriction analysis. Five enzymes, *Dde*I, *Hinf*I, *Hae*III, *Msp*I, and *Rsa*I, were used for digestion analysis of the PCR products from 32 different original strains and the products from 17 strains with different restriction profiles from each other were cloned and sequenced subsequently (Table 1).

Cloning and Sequencing

The PCR products from the above selected strains were separated on 1.5 % agarose gels and the target bands were sliced off and purified with a Qiagen QIAEX-II purification kit according to the manufacturer's instructions. The purified products were quantified with a spectrophotometer (Beckman DU-600) and cloned with a Qiagen PCR cloning kit according to the provider's protocol. Plasmid DNA was extracted with a QIAprep Spin Miniprep kit (Qiagen) and sequenced with T7 (reverse) and Sp6 (forward) promoter primers. Automated sequencing reactions were performed on a Beckman Coulter CEQTM 2000XL DNA analysis system using a CEQ cycle sequencing kit (Beckman Coulter) according to the manufacturer's instructions.

Sequence analysis

The *Pr1A* sequences generated from 17 different original strains in this study and six other *Pr1A* sequences of *M. anisopliae*, i.e.,

a


```
M73795     GCACCCCGGT-------------------------------------------------  210
AY389133   GCACCCCGGTgtaagtacccctcccacttacctaggtagtcaaaggagacatgtagttg    260
AY389119   GCACCCCGGTgtaagtacccctcccacttacctaggtagtcaaggagacatgtgggtg     260
                  Gly

M73795     ------------------------GTCGATTTCATTGAGAAGGACGCTGTGATGCGTA     244
AY389133   tttgttcctgacccgctgcgccatagGTCGATTTCATTGAGAAGGACGCTGTGATGCGTA   320
AY389119   tttgttcctgacccgctgcgccatagGTCGATTTCATTGAGAAGGACGCCGTGATGCGTA   320
                                     Val
                                      .
                                      .
                                      .
M73795     CAGGGTACTTGCGTATATATCATTGACACTGGTATTGAGGCCTCCCACCCC---------  396
AY389133   CAGGGTACTTGCGTATATATCATTGACACTGGTATTGAGGCCTCCCACCCCgtaagttgt   480
AY389119   GAGGGTACTTGCGTATATATCATTGACACTGGTATTGAGGCCTCCCACCCCgtaagttgt   480
                                                          Pro

M73795     ----------------------------------------------GAGTTTGAGG      405
AY389133   gccgtcaaaactccatagtgcggagtaggaaatttaacaatatcatccagGAGTTTGAGG   540
AY389119   gccgtcaaaactccatagtgcggagtaggaaatttaccaatatcatccagGAGTTTCAGG   540
                                                            Glu
                                             .
                                             .
                                             .
M73795     TCCACCTGGATTGTTGGCCGCACA----------------------------------    909
AY389133   TCCACCTGGATTGGTGGCCGCACAGtaagtattgtacct----cgataagctcggagaca   1076
AY389119   TCCACCTGGATTGGTGGCCAGCACTgtaagtattgtacctacctcgataagcttagagaca 1080
                               Thr

M73795     ---------------------------------AAGTCCATCTCTGGTACCTCCATG     933
AY389133   cgcttttgcttcagcaccagctctaacaaggattagAACACCATCTCTGGTACCTCCATG   1136
AY389119   cgcttttgcttcagaaccagctctaacaagatttagAACACCATCTCTGGCACCTCCATG   1140
                                               Asn
```

b

Fig. 1. Schematic presentation of the *Metarhizium anisopliae Pr1A* gene (**a**) and the partial sequence alignment of representative sequences (**b**) showing the insertion and variation of introns. The region within the arrows (1175 of 1371 bp) was sequenced and analyzed from different original strains in this study. M73795, *Pr1A* cDNA sequence. The sizes of intron 1 (76 bp) and intron 2 (59 bp) are the same in all examined strains, while there are two types of intron 3 (72 bp, e.g., AY389119, and 68 bp, e.g., AY389133). All introns have typical consensus 5′ and 3′ boundaries of GT and AG of spliceosomal introns, respectively. All three introns lie between two codons, indicating that they belong to phase 0 introns.

AJ416688, AJ416695, AB073326, AB073327, AB073344, and AB073345, retrieved from GenBank by BLAST search were analyzed (Table 1). Nucleotide sequences were aligned with the program DIALIGN 2 (Morgenstern 1999) and edited by visual inspection. The individual intron sequence was isolated after alignment with a cDNA sequence (M73795).

Phylogenetic and Exon–Intron Structure Analyses

Neighbor-joining phylogenetic analysis was performed using the program MEGA 2.1 (Kumar et al. 2001). Parsimony bootstrap analysis was conducted with 1000 replicates and a 50% majority rule tree was produced. For nucleotide sequences, a Kimura two-parameter model was selected, with pairwise deletion of gap/missing data. For amino acid sequences, a Poisson correction model was selected.

Phylogenetic relationships between the topologies of coding region and intron-based trees were analyzed using a partition-homogeneity test with the program PAUP (v4.0 b10a; Swofford 2001). The aligned sequences were trimmed to the same size and combined with noted positions, e.g., gene 1, position 1 to position 968; and gene 2, position 968 to position 1044. The tests were carried out using a heuristic search with random addition of sequences and 1000 replicates. The estimated *P* value demonstrates the significance of the homogeneous or heterogeneous relationship between the data sets (Farris et al. 1994).

The correlation analysis of gene exon–intron (intron position and phase) and protein structure provides useful information on intron origin and evolution (e.g., Stoltzfus et al. 1994; Logsdon et al. 1995; Rzhetsky et al. 1997; Sánchez et al. 2003). Intron phase (the position of the intron within a codon: phase 0, an intron lying between two condons; phase 1, an intron lying after the first nucleotide of a codon; phase 2, an intron lying after the second nucleotide) and position have been considered important characters for evaluating intron evolution (Long et al. 1995, 1998). In this study, the subtilisin *Pr1A* gene of *M. anisopliae* together with retrievable fungal subtilisin genes with clearly annotated exon–intron structures from GenBank was examined to interpret intron origin/evolution. In total, 60 subtilisin genes/proteins were analyzed, including the subtilisin genes from *Pr1A* to *Pr1K* of *M. anisopliae*, among them the *Pr1J* gene from different strains containing different numbers of introns (Bagga et al. 2004) (Fig. 3).

Results

Gene Structure of Subtilisin Pr1A

The *Pr1A* gene (1175 of 1371 bp) from 23 different original strains of *M. anisopliae* was analyzed in this study (Fig. 1a). There were three introns conservatively inserted in the gene (Fig. 1b). Introns 1 (76 bp) and 2 (59 bp) had the same sizes in all examined strains, while intron 3 had two different sizes, 68 bp (AJ416695, AY389121, AY389133, and AY389134 from four American strains) and 72 bp (the others; Table 1), where the former were short four nucleotides, TACC, compared to the latter (Fig. 1b). All introns have typical 5′ and 3′ boundaries of GT and AG of spliceosomal introns. Decoding analysis indicated that these introns were phase 0 introns, i.e., the exons of *Pr1A* were symmetric (Fig. 1b).
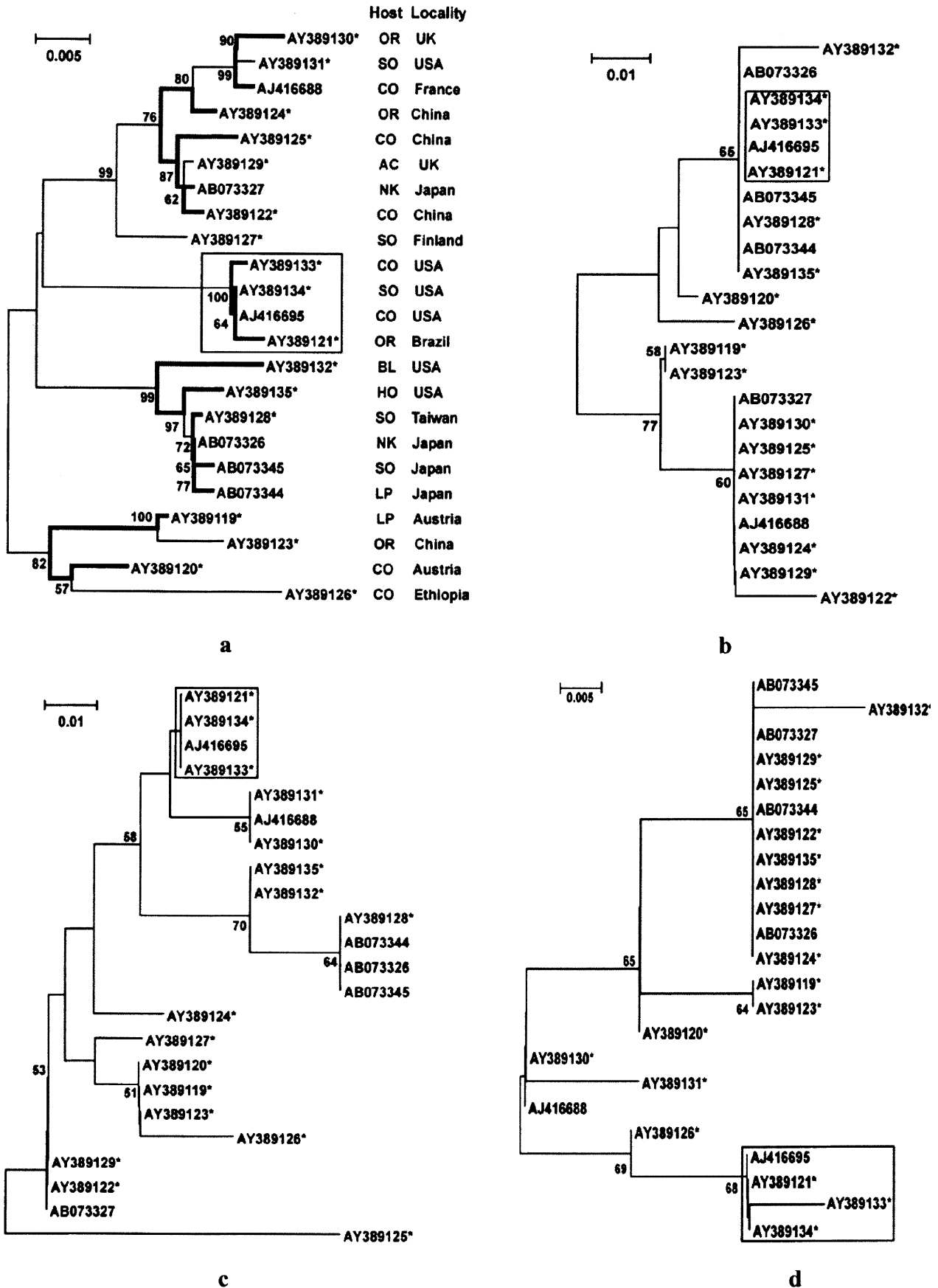
Fig. 2. Neighbor-joining phylogenetic trees of the *Pr1A* structure gene (**a**) and introns (**b**–**d**). Bootstrap values >50% are shown as numbers above branches. The clades with thick branches exhibited the sequences originally from the same localities of a country or a geographic region. Hosts: OR, Orthoptera; SO, isolated from soil; CO, Coleoptera; AC, Acari, tick; NK, not known; BL, Blattodea,

Table 2. Partition-homogeneity test between the cDNA and introns of *Metarhizium anisopliae* Pr1A[a]

	cDNA	Intron 1	Intron 2
Intron 1	0.106		
Intron 2	0.432	0.632	
Intron 3	0.015	0.020	0.725

[a]There are 1000 replicates in bootstrap analysis; $P < 0.05$ indicates that the topologies of the pairing tree are heterogeneous (Farris et al. 1994).

Phylogenetic Relationships

Phylogenetic analysis was performed in association with the *Pr1A* structure gene and its three introns (Fig. 2). There were five clades in the phylogenetic tree of the structure gene, which were clearly separated and supported by high bootstrap values (Fig. 2a). When lined with insect host orders and geographic localities of the host strains, the sequences from geographically associated localities (the same country or region) exhibited close evolutionary relationships (Fig. 2a). Particularly, the sequences from four American strains with a shorter intron 3 were monophyletic in all the trees, while the gene (AY389126) from an African strain was isolated from the others (Fig. 2).

The differences between the topologies of the phylogenetic trees were observed and verified by partition-homogeneity tests (Fig. 2, Table 2). The topologies of the pairing trees of cDNA and intron 1 and of cDNA and intron 2 were phylogenetically congruent ($P > 0.05$), while the topology of the intron 3 tree was significantly heterogeneous to the cDNA- ($P = 0.02$) and intron 1-based trees but, interestingly, was congruent with the intron 2 tree ($P = 0.752$) (Table 2).

Relationship Between Gene and Protein Structure

The 60 examined fungal subtilisin genes contained from zero to three introns (Fig. 3). The phase distribution of the total 133 introns was highly biased toward phase 0, with the ratio of phase 0:1:2 being 122:10:1 (Fig. 3). In addition, a high bias ($> 80\%$) of symmetric exons (flanked by introns in the same phase) was observed in fungal subtilisin genes (Fig. 3). Phylogenetic and exon–intron structure analyses indicated that the higher the similarity between subtilisin amino acid sequences, the more similar exon–intron structure they usually have (Fig. 3). The close relationship is independent of host fungal species, e.g., the subtilisins from three different species of *Aspergillus* sharing an identical gene structure (Fig. 3). One significant exception was that the *Pr1J* gene from three different strains of *M. ani-*
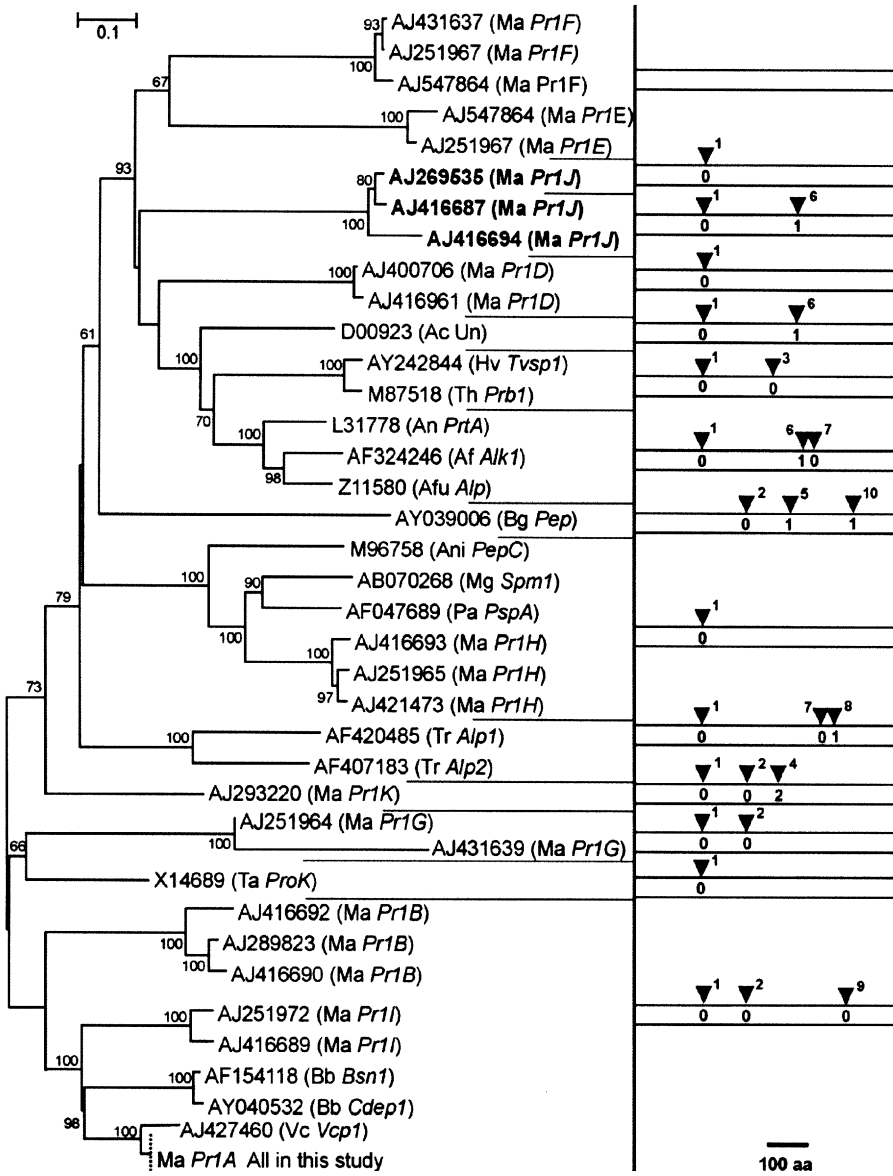
sopliae had different numbers of introns, with AJ269535 containing one intron (phase 0) and AJ416687 and AJ416694 having an additional phase 1 intron (Fig. 3). Nucleotide sequence alignment showed that the putative intron insertion site was exactly the same in AJ269535 (data not shown).

Amino acid sequence alignment of the selected proteins from each subclade indicated that there were 10 intron insertion positions among the examined fungal subtilisins. Positions 1, 2, 6, 8, and 10 were highly conserved, while the others fell into different subgroups (Figs. 3 and 4). Positions 1 and 2 had high percentages of intron occupancy, the frequencies being 90 and 45%, respectively. There was no position 1 in the intronless genes *Pr1E* and *Pr1F* of *M. anisopliae*, and the subtilisin gene *Pep* (AY039006) of *Blumeria graminis* contained no intron at position 1 (Figs. 3 and 4).
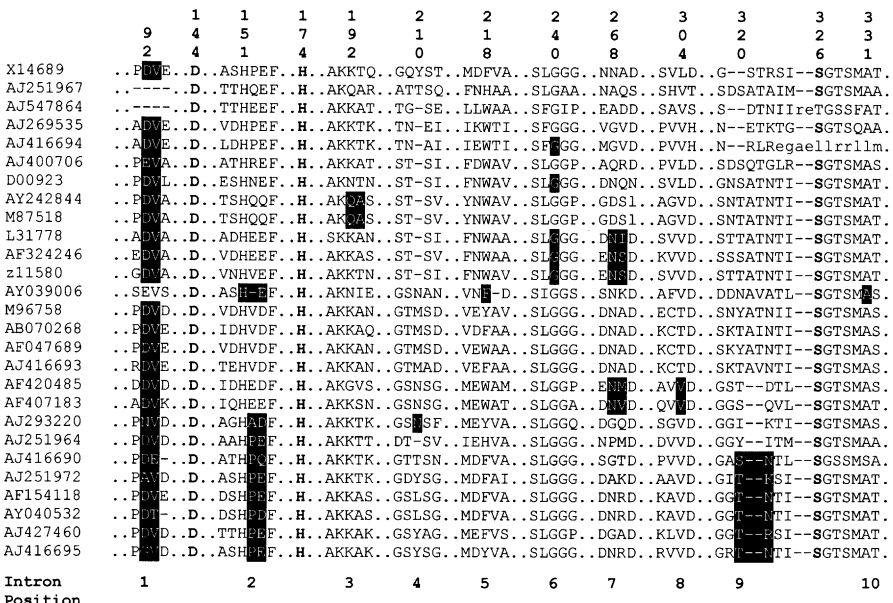
Partition-homogeneity tests were conducted using the intron sequences isolated from positions 1 and 2 and the corresponding coding region sequences and the P values were estimated as 0.075 and 0.557, respectively, indicating that the evolutionary histories between the introns at these two positions and the coding region were phylogenetically congruent.

Discussion

Phylogenetic and exon–intron structure analyses of intra- and interspecific fungal subtilisins in this study supported the theory of a mixed model of intron evolution: the evidence supported both introns-early and introns-late speculations. Intraspecifically, phylogenetic congruence or incongruence between different introns and coding region of *M. anisopliae* *Pr1A* indicates that no single scenario can the plain intron origin/evolution, which is similar to the evolution of group I introns (e.g., Wang et al. 2003). The congruence between the trees of introns 1 and 2, intron 1 and cDNA, and intron 2 and cDNA would suggest that these two introns were primordially present in the *Pr1A* gene and share a common evolutionary history with the coding region, supporting the introns-early speculation. The incongruence between intron 3 and cDNA as well as the two different sizes of intron 3 between strains suggests that the presence of intron 3 in *Pr1A* was not primordial, supporting the introns-late speculation. In addition, *Pr1A* introns 1 and 2 were located at highly conserved sites across fungal subtilisins, while the position (9) for intron 3 was subgroup specific (Fig. 4). Phylogenetic incongruence between intron 1 and intron 3 was expected, while the high topological homogeneity ($P = 0.725$) between the introns 2 and 3 trees was probably the result of coevolution of adjacent introns as speculated before (Schwaiger et al. 1993).

93 — AJ431637 (Ma Pr1F)
100 — AJ251967 (Ma Pr1F)
— AJ547864 (Ma Pr1F)
67
100 — AJ547864 (Ma Pr1E)
— AJ251967 (Ma Pr1E)
93
80 — AJ269535 (Ma Pr1J)
— AJ416687 (Ma Pr1J)
100
AJ416694 (Ma Pr1J)
100 AJ400706 (Ma Pr1D)
— AJ416961 (Ma Pr1D)
61
D00923 (Ac Un)
100
100 AY242844 (Hv Tvsp1)
— M87518 (Th Prb1)
70
100 L31778 (An PrtA)
AF324246 (Af Alk1)
98 Z11580 (Afu Alp)
AY039006 (Bg Pep)
M96758 (Ani PepC)
100
90 AB070268 (Mg Spm1)
AF047689 (Pa PspA)
100
100 AJ416693 (Ma Pr1H)
AJ251965 (Ma Pr1H)
79
97 AJ421473 (Ma Pr1H)
100 AF420485 (Tr Alp1)
73
AF407183 (Tr Alp2)
AJ293220 (Ma Pr1K)
AJ251964 (Ma Pr1G)
100
66 AJ431639 (Ma Pr1G)
X14689 (Ta ProK)
100 AJ416692 (Ma Pr1B)
100 AJ289823 (Ma Pr1B)
100 AJ416690 (Ma Pr1B)
100 AJ251972 (Ma Pr1I)
100 AJ416689 (Ma Pr1I)
100 AF154118 (Bb Bsn1)
98 AY040532 (Bb Cdep1)
100 AJ427460 (Vc Vcp1)
Ma Pr1A All in this study

0.1

100 aa

Fig. 3. Phylogenetic relationship between amino acid sequence and exon–intron structure of fungal subtilisins. Neighbor-joining tree with pairwise deletion of gap/missing data and 1000 bootstrap replicates. The right frame diagrams show the intron distribution (position and phase) in proteins that were trimmed into the same size after alignment. Arrowheads indicate the intron insertion position and the numbers in frame representing the intron phase 0, 1 or 2, respectively (see also Fig. 4). Fungal species: Ma, *Metarhizium anisopliae*; Ac, *Acremonium chrysogenum*; Th, *Trichoderma harzianum*; Hv, *Hypocrea virens*; An, *Aspergillus nidulans*; Af, *Aspergillus favus*; Afu, *Aspergillus fumigatus*; Bg, *Blumeria graminis*; Ani, *Aspergillus niger;* Mg, *Magnaporthe grisea*; Pa, *Pandora anserine;* Tr, *Trichophyton rubrum*; Ta, *Tritirachium album*; Bb, *Beauveria bassiana*; Vc, *Verticillium chlamydosporium*; Un, unnamed gene product.


```
                    1   1   1   1   2   2   2   2   3   3      3   3
                    9   4   5   7   9   1   1   4   6   0   2   2   3
                    2   4   4   7   9   1   1   4   6   4   0   2   6   1
X14689     ..PDVE..D..ASHPEF..H..AKKTQ..GQYST..MDFVA..SLGGG..NNAD..SVLD..G--STRSI--SGTSMAT..
AJ251967   ..-----..D..TTHQEF..H..AKQAR..ATTSQ..FNHAA..SLGAA..NAQS..SHVT..SDSATAIM--SGTSMAA..
AJ547864   ..-..--..D..TTHEEF..H..AKKAT..TG-SE..LLWAA..SFGIP..EADD..SAVS..S--DTNIIreTGSSFAT..
AJ269535   ..ADVE..D..VDHPEF..H..AKKTK..TN-EI..IKWTI..SLGGG..VGVD..PVVH..N--ETKTG--SGTSQAA..
AJ416694   ..ADVE..D..LDHPEF..H..AKKTK..TN-AI..IEWTI..SFGGG..MGVD..PVVH..N--RLRegaellrrllm..
AJ400706   ..PEVA..D..ATHREF..H..AKKAT..ST-SI..FDWAV..SLGGG..AQRD..PVLD..SDSQTGLR--SGTSMAS..
D00923     ..PDVL..D..ESHNEF..H..AKKTN..ST-SI..FNWAV..SLGGG..DNQN..SVLD..GNSATNTI--SGTSMAT..
AY242844   ..PDVA..D..TSHQQF..H..AKQAS..ST-SV..YNWAV..SLGGP..GDS1..AGVD..SNTATNTI--SGTSMAT..
M87518     ..PDVA..D..TSHQQF..H..AKQAS..ST-SV..YNWAV..SLGGP..GDS1..AGVD..SNTATNTI--SGTSMAT..
L31778     ..ADVA..D..ADHEEF..H..SKKAN..ST-SI..FNWAA..SLGGG..DATD..SVVD..STTATNTI--SGTSMAT..
AF324246   ..EDVA..D..VDHEEF..H..AKKAS..ST-SV..FNWAA..SLGGG..ENSD..KVVD..SSSATNTI--SGTSMAT..
z11580     ..GDVA..D..VNHVEF..H..AKKTN..ST-SI..FNWAV..SLGGG..ENSD..SVVD..STTATNTI--SGTSMAT..
AY039006   ..SEVS..D..ASH-EF..H..AKNIE..GSNAN..VN-D..SIGGS..SNKD..AFVD..DDNAVATL--SGTSMGS..
M96758     ..PDVE..D..IDHVDF..H..AKNAN..GTMSD..VEYAV..SLGGG..DNAD..ECTD..SNYATNTI--SGTSMAS..
AB070268   ..PDVE..D..IDHVDF..H..AKKAQ..GTMSD..VDFAA..SLGGG..DNAD..KCTD..SKTAINTI--SGTSMAS..
AF047689   ..PDVE..D..IDHVDF..H..AKKAN..GTMSD..VEWAA..SLGGG..DNAD..KCTD..SKYATNTI--SGTSMAS..
AJ416693   ..RDVE..D..TEHVDF..H..AKKAN..GTMAD..VEFAA..SLGGG..DNAD..KCTD..SKTAVNTI--SGTSMAS..
AF420485   ..DDVD..D..IDHEDF..H..AKGVS..GSNSG..MEWAM..SLGGP..ENRD..AVYD..GST--DTL--SGTSMAS..
AF407183   ..ADVK..D..IQHEEF..H..AKKSN..GSNSG..MEWAT..SLGGA..DKVD..QVVD..GGS--QVL--SGTSMAT..
AJ293220   ..PDVD..D..AGHADF..H..AKKTK..GSRSF..MEYVA..SLGGQ..DGQD..SGVD..GGI--KTI--SGTSMAS..
AJ251964   ..PDVD..D..AAHPEF..H..AKKTT..DT-SV..IEHVA..SLGGG..NPMD..DVVD..GGY--ITM--SGTSMAA..
AJ416690   ..PDF-..D..ATHPDF..H..AKKTK..GTTSN..MDFVA..SLGGG..SGTD..PVVD..GAS--NTL--SGSSMSA..
AJ251972   ..PAVD..D..ASHPEF..H..AKKTK..GDYSG..MDFAI..SLGGG..DAKD..AAVD..GIT--PSI--SGTSMAT..
AF154118   ..PDVE..D..DSHPEF..H..AKKAS..GSLSG..MDFVA..SLGGG..DNRD..KAVD..GGT--NTI--SGTSMAT..
AY040532   ..PDI-..D..DSHPDF..H..AKKAS..GSLSG..MEWAT..SLGGG..DNRD..KAVD..GGT--NTI--SGTSMAT..
AJ427460   ..PDVD..D..TTHPDF..H..AKKAK..GSYAG..MEFVS..SLGGP..DGAD..KLVD..GGT--PSI--SGTSMAT..
AJ416695   ..PDVD..D..ASH-DF..H..AKKAK..GSYSG..MDYVA..SLGGG..DNRD..RVVD..GRP--NTI--SGTSMAT..

Intron
Position            1       2       3       4       5       6       7       8       9       10
```

Fig. 4. Truncated amino acid sequence alignment of selected subtilisin proteins showing intron insertion positions. The amino acid positions correspond to *Tritirachium album* proteinase K. Dots indicate omission of amino acids and hyphens indicate gaps. Two shaded amino acids indicate the insertion of a phase 0 intron, while a single shaded amino acid indicates the insertion of a phase 1 (positions 5, 6, 8, and 10) or phase 2 (position 4) intron (see also Fig. 3). The boldface D, H, and S are the conserved catalytic triad of subtilisins.

In this study, direct support for the introns-late theory came from the observation that the *M. anisopliae Pr1J* gene could intraspecifically contain one intron (AJ269535) or two introns (AJ416687 and AJ416694). The high percentage of inton absence at position 6 (90%; the location of the *Pr1J* intron2) (Figs. 3 and 4) would suggest that it is more likely that intron insertions occurred in AJ416687 and AJ416694 rather than that AJ269535 lost an intron. This is consistent with the conclusion of a prevalence of intron gain over loss of splicesomal introns (Qiu et al. 2004; Babenko et al. 2004). However, it is debatable regarding the addition or deletion of introns from putative insertion sites during gene evolution, different studies reporting either that only random insertion and no loss had occurred (Cho and Doolittle 1997) or that just loss and no gain occurred (Roy et al. 2003) or both (Rogozin et al. 2003; Zhaxybayeva and Gogarten 2003). Intron presence–absence polymorphism in *Drosophila jingwei* genes was reported to result from intron loss, driven by positive Darwinian selection acting on the flanking sequences (Llopart et al. 2002). Whatever the scenario, it can be concluded that a recent event of intron insertion/deletion occurred in the *Pr1J* gene. A similar case in support of the introns-late view was the discovery of intron presence in the triose-phosphate isomerase (TPI) gene of *Culex* and *Aedes* mosquitoes but intron absence from those of *Anopheles* mosquitoes and the fly *Calliphora vicina* as well as the moth *Spodoptera littoralis* (Kwiatowski et al. 1995).

Interspecifically, the analysis of fungal subtilisin genes also supported a mixed model of intron origin/evolution. There were at least 10 intron insertion sites across fungal subtilisins, albeit at most only three introns were detected from a single gene. Half of these positions were highly conserved, and most significantly, 90% of fungal subtilisins contained a phase 0 intron at the highly conserved position 1, followed by a phase 0 intron at the highly conserved position 2 (45%) (Figs. 3 and 4), i.e., a significant excess of phase 0 introns and an excess of symmetric exons, strong evidence for exon shuffling (Long et al. 1995; De Souza et al. 1998; Roy et al. 1999; Long and Deutsch 1999; Fedorov et al. 2001, 2003). In addition, similarly to the discussion above, the partition-homogeneity test resulting in a phylogenetic congruence relationship between the evolutionary histories of introns and coding regions again supported the primordial presence of introns at these two positions. In contrast, positions 4, 5, and 10 (Figs. 3 and 4) were detected from only a single fungal species, appearing to be of recent origins, which is similar to the observations in TPI supporting the introns-late theory (Logsdon et al. 1995). For the introns-late theory, intron origin or the arising of new intron positions is

still perplexing; extensive genomic examinations demonstrated that intron transposition within genomes could not have occurred recently (Fedorov et al. 2003), however, intron sliding over short distances (1–15 bases) has been demonstrated (Rogozin et al. 2000), but its contribution to the diversity of intron positions could be ignored (Stoltzfus et al. 1997). Sequence analysis indicated that no significant similarity existed between subtilisin introns at different positions (data not shown), suggesting that intron drifting did not occur to contribute to the divergence of fungal subtilisin exon–intron structure. Regardless of functional magnitude, multiple subtilisins identified from a single fungal species demonstrated events of gene duplications (Bagga et al. 2004). A recent study suggested that gene duplication could boost intron insertion (Babenko et al. 2004), an alternative scenario that could be employed in this study to explain the diversity of intron positions.

The parameters of protein tertiary structure (e.g., centrality, diameter, or radius [Logsdon et al. 1995]) were usually used for correlation analysis of protein and gene structure to examine intron evolution. In this study, phylogenetic analysis based on amino acid sequences provided a fairly clear relationship between protein primary structure and gene exon–intron structure (Fig. 3). This could be an alternative method for future similar studies since the three-dimensional structures of a large number of proteins remain unknown. An exception that could call the gene–protein structure relationship into question was that the *Pr1J* gene from different strains of *M. anisopliae* had different exon–intron structures. There were some other exceptions of genes with the same exon–intron structure that were not clustered together, for example, the proteins with a single phase 0 intron at position 1 (Fig. 3). However, with the combination of protein size data, the difference between those proteins became clear, for example, proteinase K (X14689) of *Tritirachium album* has 384 amino acids, while the proteins in the subgroup including *Pr1H* of *M. anisopliae* have more than 530 amino acids (Fig. 3). The close relationship of gene–protein structure independent of fungal species suggests that subtilisin evolution and diversification have not been closely related to fungal specification. Supportively, the subtilisins *Pr1A–Pr1K* could be identified from a single strain of *M. anisopliae* as a paralogous gene family (Bagga et al. 2004). More interestingly, the host strains for *Pr1J* AJ269535 (one intron) and AJ416687 (two introns) belong to the subspecies *M. anisopliae* sf. *anisopliae*, while AJ416694 (two introns) belongs to *M. anisopliae* sf. *acridum* (Bagga et al. 2004), indicating that intron insertions at AJ416687 and AJ416694 occurred before the divergence of *M. anisopliae* into two subspecies but could be after the separation of *M.*

anisopliae sf. *anisopliae* strains. For the subtilisins identified from the genus *Aspergillus*, *PepC* (*A. niger*) has no introns at positions 6 and 7 compared with *PrtA* (*A. nidulans*), *Alk1* (*A. favus*), and *Alp* (*A. fumigatus*) (Fig. 3), again suggesting that intron insertion could occur before fungal specification.

In general, phylogenetic analyses of intra- and interspecific fungal subtilisins in this study provided evidence in support of the speculations of both the introns-early and the introns-late theories. However, no conclusion can be drawn from this study to resolve the issue of exon–intron origin/evolution. Literature review indicated that the arguments from some studies were more likely to be case associated, for example, studies of TPI genes (Kwiatowski et al. 1995; Logsdon et al. 1995), carbonic anhydrases (Hewett-Emmett and Tashian 1996), receptor tyrosine kinase genes (Gamulin et al. 1997), and aldehyde dehydrogenase genes (Rzhetsky et al. 1997), and even a large-scale examination of 10 protein families (Cho and Doolittle 1997), supported the introns-late theory, while a study of xylanase genes provided evidence for the view of ancient introns (Sato et al. 1999). Interestingly, a reexamination of TPI exon–intron structure recommended a reconciliation of the introns-early and introns-late views (Tyshenko and Walker 1997). More comprehensively, a statistical test of a data set of 44 ancient proteins with 988 introns indicated that about one-third of all introns were ancient and belonged to phase zero introns, while the others corresponded to intron addition/deletion, appearing equally in all three intron phases (De Souza et al. 1998). The examination of intron positions in 684 orthologous gene sets from 8 complete genomes suggests that many introns are ancient while lineage-specific intron loss/gain events have been apparently taking place (Rogozin et al. 2003). In this respect, a synthetic theory should be acceptable to provide insights that introns have been not only the remnants for exon shuffling to accelerate the formation of new genes, but also the mobile elements in gene evolution (De Souza 2003; Fedorova and Fedorov 2003).

References

Åhman J, Johansson T, Olsson M, Punt PJ, Van de Hondel CA, Tunlid A (2002) Improving the pathogenicity of a nematode-trapping fungus by genetic engineering of a subtilisin with nematotoxic activity. Appl Environ Microbiol 68:3408–3415

Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV (2004) Prevalence of intron gain over intron loss in the evolution of paralogous gene families. Nucleic Acids Res 32:3724–3733

Bagga S, Hu G, Screen SE, St Leger RJ (2004) Reconstructing the diversification of subtilisins in the pathogenic fungus *Metarhizium anisopliae*. Gene 324:159–169

Cho G, Doolittle RF (1997) Intron distribution in ancient paralogs supports random insertion and not random loss. J Mol Evol 44:573–584

De Souza SJ (2003) The emergence of a synthetic theory of intron evolution. Genetica 118:117–121

De Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W (1998) Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA 95:5094–5099

Farris JS, Kallersjo M, Kluge AG, Bult C (1994) Testing significance of congruence. Cladistics 10:315–320

Fedorov A, Cao X, Saxonov S, de Souza SJ, Roy SW, Gilbert W (2001) Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. Proc Natl Acad Sci USA 98:3117–3182

Fedorov A, Merican AF, Gilbert W (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. Proc Natl Acad Sci USA 99:16128–16133

Fedorov A, Roy S, Fedorova L, Gilbert W (2003) Mystery of intron gain. Genome Res 13:2236–2241

Fedorova L, Fedorov A (2003) Introns in gene evolution. Genetica 118:123–131

Gamulin V, Skorokhod A, Kavsan V, Muller IM, Muller WE (1997) Experimental indication in favor of the introns-late theory: The receptor tyrosine kinase gene from the sponge *Geodia cydonium*. J Mol Evol 44:242–252

Hewett-Emmett D, Tashian RE (1996) Functional diversity, conservation, and convergence in the evolution of the alpha-, beta-, and gamma-carbonic anhydrase gene families. Mol Phylogenet Evol 5:50–77

Hurst LD, McVean GT (1996) A difficult phase for introns-early. Molecular evolution. Curr Biol 6:533–536

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 17:244–1245

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

Logsdon JM Jr (1998) The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev 8:637–648

Logsdon JM Jr, Tyshenko MG, Dixon C, Jafari JD, Walker VK, Palmer JD (1995) Seven newly discovered intron positions in the triose-phosphate isomerase gene: Evidence for the introns-late theory. Proc Natl Acad Sci USA 92:8507–8511

Long M, Deutsch M (1999) Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol Biol Evol 16:528–1534

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

Long M, de Souza SJ, Rosenberg C, Gilbert W (1998) Relationship between "proto-splice sites" and intron phases: Evidence from dicodon analysis. Proc Natl Acad Sci USA 95:219–223

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M (2002) Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. Proc Natl Acad Sci 99:8121–8126

Morgenstern B (1999) DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211–218

O'Neill RJ, Brennan FE, Delbridge ML, Crozier RH, Graves JA (1998) De novo insertion of an intron into the mammalian sex determining gene, SRY. Proc Natl Acad Sci USA 95:1653–1657

Qiu WG, Schisler N, Stoltzfus A (2004) The evolutionary gain of spliceosomal introns: Sequence and phase preferences. Mol Biol Evol 21:1252–1263

Rogozin IB, Lyons-Weiler J, Koonin EV (2000) Intron sliding in conserved gene families. Trends Genet 16:430–432

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 13:1512–1517

Roy SW, Nosaka M, de Souza SJ, Gilbert W (1999) Centripetal modules and ancient introns. Gene 238:85–91

Roy SW (2003) Recent evidence for the exon theory of genes. Genetica 118:251–266

Roy SW, Lewis BP, Fedorov A, Gilbert W (2001) Footprints of primordial introns on the eukaryotic genome. Trends Genet 17:496–501

Roy SW, Fedorov A, Gilbert W (2002) The signal of ancient introns is obscured by intron density and homolog number. Proc Natl Acad Sci USA 99:5513–15517

Roy SW, Fedorov A, Gilbert W (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci USA 100:7158–7162

Rzhetsky A, Ayala FJ, Hsu LC, Chang C, Yoshida A (1997) Exon/intron structure of aldehyde dehydrogenase genes supports the "introns-late" theory. Proc Natl Acad Sci USA 94:6820–6825

Sánchez D, Ganfornina MD, Gutiérrez G, Marín A (2003) Exon-intron structure and evolution of the lipocalin gene family. Mol Biol Evol 20:775–783

Sato Y, Niimura Y, Yura K, Go M (1999) Module-intron correlation and intron sliding in family F/10 xylanase genes. Gene 238:93–101

Schwaiger FW, Weyers E, Epplen C, Brun J, Ruff G, Crawford A, Epplen JT (1993) The paradox of MHC-DRB exon/intron evolution: alpha-helix and beta-sheet encoding regions diverge while hypervariable intronic simple repeats coevolve with beta-sheet codons. J Mol Evol 37:260–272

Siezen RJ, Leunissen JAM (1997) Subtilases: The superfamily of subtilisin-like serine proteases. Protein Sci 6:501–523

St Leger RJ, Joshi L, Bidochka MJ, Roberts DR (1996) Construction of an improved mycoinsecticide overexpressing a toxic protease. Proc Natl Acad Sci USA 93:6349–6354

Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr (1994) Testing the exon theory of genes: The evidence from protein structure. Science 265:202–207

Stoltzfus A, Logsdon JM Jr, Palmer JD, Doolittle WF (1997) Intron "sliding" and the diversity of intron positions. Proc Natl Acad Sci USA 94:10739–10744

Swofford D (2001) PAUP (phylogenetic analysis using parsimony) v. 4.0 b8a. Sinauer, Sunderland, MA

Tyshenko MG, Walker VK (1997) Towards a reconciliation of the introns early or late views: Triosephosphate isomerase genes from insects. Biochim Biophys Acta 1353:131–136

Wang C-S, Typas MA, Butt TM (2002) Detection and characterisation of *pr1* gene deficiencies in the insect pathogenic fungus *Metarhizium anisopliae*. FEMS Microbiol Lett 213:251–255

Wang C-S, Li Z-Z, Typas MA, Butt TM (2003) Nuclear large subunit rDNA group I intron distribution in a population of *Beauveria bassiana* strains: Phylogenetic implications. Mycol Res 107:1189–1200

Wolf YI, Kondrashov FA, Koonin EV (2001) Footprints of primordial introns on the eukaryotic genome: Still no clear traces. Trends Genet 17:499–501

Zhaxybayeva O, Gogarten JP (2003) Spliceosomal introns: New insights into their evolution. Curr Biol 13:R764–R766