Available online at www.sciencedirect.com
SCIENCE DIRECT
Fungal Genetics and Biology 43 (2006) 20–33
FUNGAL GENETICS AND BIOLOGY
www.elsevier.com/locate/yfgbi

Computational and comparative analyses of 150 full-length cDNA sequences from the oomycete plant pathogen *Phytophthora infestans*

Joe Win, Thirumala-Devi Kanneganti, Trudy Torto-Alalibo, Sophien Kamoun *

Department of Plant Pathology, The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, OH 44691, USA

Received 28 March 2005; accepted 5 October 2005

Abstract

Phytophthora infestans is a devastating phytopathogenic oomycete that causes late blight on tomato and potato. Recent genome sequencing efforts of *P. infestans* and other *Phytophthora* species are generating vast amounts of sequence data providing opportunities to unlock the complex nature of pathogenesis. However, accurate annotation of *Phytophthora* genomes will be a significant challenge. Most of the information about gene structure in these species was gathered from a handful of genes resulting in significant limitations for development of ab initio gene-calling programs. In this study, we collected a total of 150 bioinformatically determined near full-length cDNA (FLcDNA) sequences of *P. infestans* that were predicted to contain full open reading frame sequences. We performed detailed computational analyses of these FLcDNA sequences to obtain a snapshot of *P. infestans* gene structure, gauge the degree of sequence conservation between *P. infestans* genes and those of *Phytophthora sojae* and *Phytophthora ramorum*, and identify patterns of gene conservation between *P. infestans* and various eukaryotes, particularly fungi, for which genome-wide translated protein sequences are available. These analyses helped us to define the structural characteristics of *P. infestans* genes using a validated data set. We also determined the degree of sequence conservation within the genus *Phytophthora* and identified a set of fast evolving genes. Finally, we identified a set of genes that are shared between *Phytophthora* and fungal phytopathogens but absent in animal fungal pathogens. These results confirm that plant pathogenic oomycetes and fungi share virulence components, and suggest that eukaryotic microbial pathogens that share similar lifestyles also share a similar set of genes independently of their phylogenetic relatedness.

Keywords: Oomycetes; *Phytopthora infestans*; Gene structure; Comparative genomics

1. Introduction

One of the challenges of the post-genomics era is the identification and annotation of genes from genome sequences. Accurate annotation is particularly challenging for eukaryote genomes, which are interspersed with non-coding DNA sequences and contain structurally complex genes (Stein, 2001). Characterization of transcripts in the form of full-length or near full-length cDNA (FLcDNA) sequences is an invaluable resource for gene identification and annotation (Haas et al., 2002). Computer programs can be developed to use FLcDNA sequence information to improve ab initio gene predictions (Ashurst and Collins, 2003; Brendel et al., 2004; Davuluri and Zhang, 2003). The importance of FLcDNAs in genomics is illustrated by the significant efforts undertaken by several research communities to construct and sequence collections of FLcDNAs with the aim of identifying the complete transcriptome of a given organism. For example, following the completion of the *Arabidopsis* and mouse genome sequences, large-scale FLcDNA sequencing projects were initiated resulting in the *Arabidopsis* FLcDNA database (Seki et al., 2002) and the RIKEN mouse genome encyclopedia (Hayashizaki, 2003b). The information collected in such databases, together with the corresponding genomic DNA sequences, has enabled the validation of existing genome annotations as well as the

* Corresponding author. Fax: +1 330 263 3841.
E-mail address: kamoun.1@osu.edu (S. Kamoun).

1087-1845/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.fgb.2005.10.003

development of new 'gene models' for these organisms (Castelli et al., 2004; Haas et al., 2002, 2003; Hayashizaki, 2003a,b; Okazaki et al., 2002; Wortman et al., 2003). In addition, FLcDNAs are critical reagents for functional expression assays and other types of functional analyses (Borevitz and Ecker, 2004; Frech and Joho, 1989; Kato et al., 1994; Ogihara et al., 2004).

Phytophthora infestans, the Irish potato famine pathogen, causes the devastating late blight disease of potato and tomato. Re-emergence of late blight during the late twentieth century ensured that this pathogen continues to be an important threat to agriculture (Fry and Goodwin, 1997). Late blight is estimated to cost growers billions of dollars annually through losses in crop production and increased fungicide expenses (Duncan, 1999). The study of *P. infestans* and other *Phytophthora* spp., such as *Phytophthora sojae* and *Phytophthora ramorum*, has entered the genomics age. The genome size of *P. infestans* was calculated to be approximately 237 Mb (Tooley and Therrien, 1987) with at least 51% repeated DNA sequences (Judelson and Randall, 1998). Due to this considerable genome size and complexity, early *P. infestans* genomics efforts focused on cDNA sequencing and generated collections of expressed sequence tags (ESTs) from a variety of developmental, stress, and infection conditions (Kamoun et al., 1999; Randall et al., 2005). So far, more than 75,000 ESTs and about 1X whole genome shotgun (WGS) sequence have been generated and assembled into about 18,000 genes (Randall et al., 2005). Genome sequencing, as well as the development of detailed genetic and physical maps, is also under way through work by an international consortium of researchers led by the Broad Institute, Boston, MA. In the near future, these efforts should lead to the complete sequencing of the *P. infestans* genome, thus adding to the genome sequence drafts of *P. sojae* and *P. ramorum*. However, accurate annotation of *Phytophthora* genomes will be a significant challenge. Most of the information about gene structure in these species was gathered from only a handful of genes resulting in significant limitations for development of ab initio gene-calling programs (Huitema et al., 2004; Kamoun, 2003). High quality genome annotation of the *P. infestans* genome sequence will also be essential for increasing the momentum of ongoing functional genomics studies (Birch and Whisson, 2001; Torto et al., 2003).

Phytophthora infestans belongs to the oomycetes, a deep branching group of eukaryotic microbes, that are related to diatoms and brown algae within the stramenopiles or heterokonts (Baldauf, 2003; Baldauf et al., 2000). The oomycetes share a number of morphological and physiological similarities with the fungi, such as filamentous growth habit, heterotrophic lifestyle, and specialized infection structures (Kamoun, 2003; Latijnhouwers et al., 2003). However, oomycetes and fungi represent some of the most highly divergent pathogenic eukaryotes and their pathogenic lifestyles have clearly evolved independently (Baldauf et al., 2000; Kamoun, 2003; Sogin and Silberman, 1998). The increased availability of genome sequences offers unique opportunities to perform comparative analyses among eukaryotes and improve our understanding of the evolution of pathogenic lifestyles in eukaryotic microbes (Huitema et al., 2004; Kamoun, 2003). Several pertinent questions can now be addressed. Are there common mechanisms of infection among pathogenic eukaryotes? How did the arsenal of pathogenicity genes emerge and evolve? To what extent do these genes vary between plant and animal pathogenic eukaryotes? In this study, we collected a total of 150 bioinformatically determined near full-length cDNA (FLcDNA) sequences for the oomycete plant pathogen *P. infestans*. The majority (131) of these FLcDNA sequences originated from our laboratory either as original sequences generated for this study or as part of earlier studies (Torto et al., 2003). We performed detailed computational analyses of these FLcDNA sequences to (1) obtain a snapshot of *P. infestans* gene structure, (2) gauge the degree of sequence conservation between *P. infestans* genes and those of *P. sojae* and *P. ramorum*, and (3) identify patterns of gene conservation between *P. infestans* and various eukaryotes, particularly fungi, for which genome-wide translated protein sequences are available. These analyses helped us to define the structural characteristics of *P. infestans* genes using a validated data set. We also determined the degree of sequence conservation within the genus *Phytophthora* and identified a set of fast evolving genes. Finally, we also identified a set of genes that are shared between *Phytophthora* and fungal phytopathogens but absent in animal fungal pathogens.

2. Materials and methods

2.1. Data set

We performed all our analyses on a data set that includes 150 *P. infestans* cDNA bioinformatically predicted to be near full-length and to contain complete open reading frames (ORF) (Table 1). The cDNA sequences either originated from the GenBank sequence database (Benson et al., 1999) or were sequenced in our laboratory by primer walking, and they covered a wide range of cellular functions. We enforced two requirements for including sequences in this data set: (1) the cDNAs must correspond to nuclear encoded genes, and (2) they must have high quality sequences and complete ORFs. The cDNAs were identified by scanning ESTs for 5′ proximal ATGs. A cDNA was deemed likely to be full-length or near full-length when it was the most 5′ proximal EST among assemblies and gave hits to the N-terminal portion of known proteins following BLAST searches against GenBank. Throughout the paper these near full-length cDNAs will be referred to as FLcDNAs.

2.2. Analysis of nucleotide composition and ATG start codon context

The sequences in the untranslated regions (UTRs) and coding regions were manually extracted. G + C contents and

Table 1
Identities of 150 *P. infestans* FLcDNA sequences and signal peptide predictions of the translated proteins

Categories and Accession	Identity of predicted protein[a]	Extracellular protein[b]	SignalP-HMM probability[c]
Small cysteine-rich			
AY961446[d]	Small cysteine-rich protein SCR74	Yes	0.997
AY961447[d]	Small cysteine-rich protein SCR74	Yes	0.997
AY961418[d]	Small cysteine-rich protein SCR91	Yes	0.986
AY961448[d]	Small cysteine-rich protein SCR91	Yes	0.986
AF424679[d]	Small cysteine-rich protein SCR50	Yes	1.000
AF424680[d]	Small cysteine-rich protein SCR58	Yes	0.998
AF424670[d]	Small cysteine-rich protein SCR76	No	0.638
AF424669[d]	Small cysteine-rich protein SCR108	Yes	0.999
AF424683	Small cysteine-rich protein SCR122	Yes	0.983
AY961427[d]	Cysteine-rich protein	Yes	1.000
AY961428[d]	Cysteine-rich protein	Yes	0.999
Elicitors and virulence			
U50844[d]	Elicitin INF1	Yes	0.998
AF004951	Elicitin INF2A	Yes	0.999
AF004952	Elicitin INF2B	Yes	1.000
AF419841[d]	Elicitin-like INF4	Yes	1.000
AF419842	Elicitin-like INF5	Yes	0.997
AF419843	Elicitin-like INF6	Yes	1.000
AF419844	Elicitin-like INF7	Yes	1.000
AY961426[d]	Elicitin-like protein	Yes	1.000
AY961417	NPP1-like protein	Yes	0.999
AY961431	NPP1-like protein	Yes	0.999
AY961432	NPP1-like protein	Yes	0.997
AY961430[d]	In planta-induced IPIO1	Yes	1.000
AY961429[d]	In planta-induced IPIB-like protein	Yes	0.999
AY961420[d]	CBEL-like protein	Yes	0.997
Extracellular protease inhibitors (EPI)			
AY586273	Kazal-like serine protease inhibitor EPI1	Yes	1.000
AY586274	Kazal-like serine protease inhibitor EPI2	Yes	0.999
AY586276	Kazal-like serine protease inhibitor EPI4	Yes	0.961
AY586281	Kazal-like serine protease inhibitor EPI9	Yes	1.000
AY935250[d]	Cysteine protease inhibitor EPIC1	Yes	0.999
AY935254[d]	Cysteine protease inhibitor EPIC4	Yes	0.999
Crinkling and necrosis inducing (CRN)			
AF424675[d]	Crinkling and necrosis-inducing protein CRN1	Yes	0.902
AF424677[d]	Crinkling and necrosis-inducing protein CRN2	Yes	0.977
AY961451[d]	CRN-like CRN3	Yes	0.963
AY961452[d]	CRN-like CRN4	No	0.541
AY961453[d]	CRN-like CRN5	No	0.554
AY961454[d]	CRN-like CRN6	Yes	0.955
AY961455[d]	CRN-like CRN7	Yes	0.970
AY961456	CRN-like CRN8	Yes	0.935
AY961457[d]	CRN-like CRN9	Yes	0.984
AY961458[d]	CRN-like CRN10	No	0.718
AY961459[d]	CRN-like CRN11	No	0.718
AY961460[d]	CRN-like CRN12	Yes	0.934
AY961461[d]	CRN-like CRN13	Yes	0.962
AY961462[d]	CRN-like CRN14	Yes	0.902
AY961463[d]	CRN-like CRN15	Yes	0.970
AY961464[d]	CRN-like CRN16	No	0.718
HAM34-like/low complexity			
AF424659	HAM34-like protein	Yes	1.000
AF424687	HAM34-like putative membrane protein	Yes	1.000
AY961416	HAM34-like putative membrane protein	Yes	1.000
AY961437	HAM34-like putative membrane protein	Yes	1.000
AY961438	HAM34-like putative membrane protein	Yes	1.000
AY961439	HAM34-like putative membrane protein	Yes	1.000
AY961440	HAM34-like putative membrane protein	Yes	1.000
AY961441	HAM34-like putative membrane protein	Yes	1.000
AY961442	HAM34-like putative membrane protein	Yes	1.000
AY961443	Low complexity protein	Yes	1.000

Table 1 (*continued*)

Categories and Accession	Identity of predicted protein[a]	Extracellular protein[b]	SignalP-HMM probability[c]
AY961444	Low complexity protein	Yes	1.000
AY961445	Low complexity protein	Yes	1.000
Metabolic enzymes			
AF394510	6-Phosphogluconate dehydrogenase	No	0.224
AF424638	Glutamine synthetase	No	0.000
AF424644	Dihydrodipicolinate synthase	No	0.214
AF424646	Fructose-1,6-biphosphatase	No	0.000
AF424647	Argininosuccinate lyase	No	0.000
AF424653	NADH dehydrogenase	No	0.303
AF424654	Enolase	No	0.000
AF424658	Peptidylprolyl isomerase	No	0.009
AF424660	Diaminopimelate decarboxylase	No	0.006
AF424661	Cystathionine β-synthase	No	0.103
AF424662	Thioredoxin peroxidase	No	0.000
AF424663	DAHP synthase	No	0.000
AF424664	*S*-Adenosyl methionine synthetase	No	0.001
AF424665	Transaldolase	No	0.000
AF533882	Glutamate dehydrogenase	No	0.000
AY098641	Glucose 6-phosphate isomerase	No	0.000
Hydrolytic enzymes			
AF424684	Acidic chitinase	Yes	0.998
AY052571	Endopolygalacturonase	Yes	1.000
AF352032	β-Glucosidase/xylosidase	Yes	0.992
AY961421	Cutinase	Yes	1.000
AY961419	Aspartic protease	Yes	0.999
AY961424	Cathepsin-like cysteine protease	Yes	1.000
AY961425	Cathepsin-like cysteine protease	Yes	1.000
AY961422	Cysteine protease	Yes	1.000
AY961423	Cysteine protease	Yes	1.000
AY961449	Trypsin protease GIP-like	Yes	0.989
AY961450	Trypsin protease GIP-like	Yes	0.995
Other enzymes			
AF424685	Peroxidase-like protein	Yes	0.994
AF424690	Peroxidase-like protein	Yes	0.999
AF424668	Protein disulfide-isomerase	Yes	1.000
AF424667	Peptidylprolyl isomerase	Yes	1.000
AY237405	Transglutaminase elicitor M81E	Yes	0.994
AY237403[d]	Transglutaminase elicitor M81C	Yes	0.990
AY237404[d]	Transglutaminase elicitor M81D	Yes	0.993
AF424650	Serine/threonine protein phosphatase	No	0.000
AF424649	Phosphoglycerate kinase	No	0.000
AF424648	GMP reductase	No	0.000
AF424645	Pyrophosphatase	No	0.406
AF424642	ADP/ATP translocase	No	0.455
AF424641	Ubiquitin-conjugating enzyme	No	0.000
AF424651	Ubiquitin-conjugating enzyme	No	0.009
Ribosomal proteins			
AF424686	Ribosomal protein	Yes	0.974
AF424689	Ribosomal protein	Yes	0.999
AY961480	Ribosomal protein L18	No	0.000
AY961465	Ribosomal protein L23	No	0.312
AY961477	Ribosomal protein L27	No	0.000
AY961473	Ribosomal protein L28	No	0.012
AY961479	Ribosomal protein L31	No	0.000
AY961475	Ribosomal protein L34	No	0.000
AY961476	Ribosomal protein L35	No	0.001
AY961471	Ribosomal protein L37	No	0.000
AY961469	Ribosomal protein L38	No	0.000
AY961478	Ribosomal protein L39	No	0.000
AY961467	Ribosomal protein S10	No	0.002
AY961470	Ribosomal protein S12	No	0.000
AY961468	Ribosomal protein S19	No	0.000
AY961474	Ribosomal protein S19	No	0.000

(*continued on next page*)

Table 1 (*continued*)

Categories and Accession	Identity of predicted protein[a]	Extracellular protein[b]	SignalP-HMM probability[c]
Unknown proteins			
AF424666	Unknown protein	No	0.331
AF424671	Unknown protein	No	0.767
AF424672	Unknown protein	Yes	1.000
AF424673[d]	Unknown protein	No	0.866
AF424674[d]	Unknown protein	Yes	1.000
AF424676	Unknown protein	Yes	0.997
AF424678[d]	Unknown protein	No	0.639
AF424681	Unknown protein	Yes	0.999
AF424682[d]	Unknown protein	Yes	0.998
AY961433[d]	Unknown protein	Yes	0.975
AY961434[d]	Unknown protein	Yes	1.000
AY961435[d]	Unknown protein	Yes	0.980
AY961436[d]	Unknown protein	Yes	0.918
AY961481	Unknown protein	No	0.007
Other proteins			
AF424688	Ammonium transporter	Yes	0.982
AF507054	Elicitin-like mating protein M25	Yes	1.000
AF507055	Elicitor-like mating protein M81	Yes	0.998
AF507059	Mating-induced protein M96	Yes	0.998
AY050538	G-protein β subunit 1	No	0.000
AY204515	G-protein β subunit 1	No	0.000
AY204881	Cell cycle protein cdc14	No	0.004
AF424657	GTP binding protein	No	0.002
AF424656	Argonaute-like protein	No	0.000
AF424655	Microtubial binding protein	No	0.000
AF424652	Electron transfer flavoprotein β subunit	No	0.000
AF424640	14-3-3-like protein	No	0.017
AF424639	K^+ channel protein	No	0.209
AF404749	Syntaxin 6	No	0.000
AJ133023	RIC1 protein	No	0.035
AY961466	Ubiquitin-like protein/ribosomal protein	No	0.002
AF424643	Proteasome subunit	No	0.000
AY961472	Proteasome subunit	No	0.004
AF507057	Croquemort-like mating protein M82	No	0.000
AF507056	Pumilio-like mating protein M90	No	0.000

[a] Determined by sequence similarity to the non-redundant GenBank protein database and literature searches.
[b] Predicted using the PexFinder algorithm (Torto et al., 2003) based on the SignalP signal peptide prediction program. A protein sequence is assigned "Yes" for extracellular protein if the protein is predicted to be signal peptide by SignalP Hidden Markov Model (SignalP HMM) with a probability greater than 0.900, and the SignalP Neural Network-predicted cleavage site between 10 and 40 amino acid residues. Otherwise, it is assigned "No."
[c] Calculated by SignalP Hidden Markov Model algorithm.
[d] *Phytophthora*-specific proteins.

lengths of the sequences were calculated using customized Perl scripts. Sequences surrounding the translational start codon ATG (−10 to +3) of the FLcDNA set were manually extracted, aligned, and submitted to WebLogo server (http://weblogo.berkeley.edu) (Crooks et al., 2004) to generate a sequence logo that graphically displays the consensus.

2.3. Analysis of polyadenylation signals

To determine the polyadenylation signals in the FLcDNA set, we searched the 3′-end sequences for known polyadenylation signals based on the conserved poly(A) signal (AAUAAA) found in animal (Manley and Proudfoot, 1994), yeast (Wahle, 1995), and angiosperm mRNAs (Rothnie et al., 1994). Putative polyadenylation signals were extracted and manually aligned to identify sequence conservation. Number of cDNA sequences in each variation in the motif "AAUAAA" was then counted and their representations were plotted graphically.

2.4. Codon usage

To calculate the codon usage of *P. infestans* based on the FLcDNA data set, we used "cusp" program from the European Molecular Biology Open Source Software Suite (EMBOSS). We generated a codon usage table specific for *P. infestans* from this analysis.

2.5. Intron sequences

To determine the gene structure and to obtain intron sequences in the genes corresponding to the FLcDNA

set, we performed BLASTN searches against WGS sequences of *P. infestans*, and retrieved top 10 genomic sequence hits (when available). The top hits in high scoring pairs were visually inspected for quality of match, length of coverage, and putative gaps. Genomic sequences that showed 100% nucleotide identity were then aligned with FLcDNA sequences to determine the positions of the introns.

2.6. Intron consensus sequences

To assess the conservation of nucleotides at the intron splice sites, we manually aligned nucleotide sequences (3 from exon region and 12 from intron region) covering each exon–intron junction for both 5′ and 3′ ends of introns. The sequence line-ups were submitted to the WebLogo server to calculate consensus sequences and relative conservation of nucleotides at their respective positions relative to the splice sites. The sequence conservations were shown as sequence logo plots in which the size of the letter representing the sequence is directly proportional to the degree of conservation of that particular sequence within the alignment (Crooks et al., 2004).

2.7. Comparative sequence analyses

To determine the conservation of the sequences in the FLcDNA data set within the *Phytophthora* species, we translated the ORFs into amino acid sequences and used TBLASTN against the genome sequences of *P. sojae* and *P. ramorum* (Joint Genome Institute, Department of Energy, http://genome.jgi-psf.org) with an Expect (*E*) value cut-off at 0.001 and without the low complexity filter. Top hits were collected in a spreadsheet and percent amino acid identities were plotted in a scatter graph. Additionally, we divided the *P. infestans* protein data set into cellular proteins or extracellular proteins based on presence or absence of a signal peptide predicted using the criteria of Torto et al. (2003). We plotted sequence conservation of cellular proteins and extracellular proteins among the *Phytophthora* spp. in scatter graphs.

To determine the conservation of the sequences in the FLcDNA data set within the phylogenetically distinct organisms, we used BLASTP to search against the amino acid sequences of predicted proteomes of several phylogenetically distinct organisms (Table 2) with an *E* value cut-off at 0.001 and with the low-complexity filter on query sequences. Top hit scores were collected in a spreadsheet and percent amino acid identities were plotted in a scatter graph. To reveal any distantly related conserved functional or structural domains in protein sequences that were present only in *P. infestans*, we used hmmpfam program (HMMER software, http://hmmer.wustl.edu) to search Pfam Hidden Markov Model profile database of protein domains (Bateman et al., 2004) with an *E* value cut-off at 1.0. Sequences were then annotated with this data.

3. Results

3.1. Full-length cDNA data set

At the time of this analysis, there were less than 100 *P. infestans* cDNA/mRNA sequences in GenBank. Among these, 77 contained complete sequence information suggesting near full-length cDNAs. These 77 sequences were incorporated into this study. Fifty-eight of these sequences originated from our laboratory, mainly as part of a control data set for an earlier study that included cDNAs encoding proteins with diverse functions and cellular locations (Torto et al., 2003). To increase the sample size, we completely sequenced additional 73 FLcDNA inserts by primer walking and submitted the sequences to GenBank. The full data set of 150 FLcDNAs is described in Table 1. The predicted proteins were annotated by sequence similarity and literature searches, and they possessed a variety of functions ranging from basic cellular activities to virulence (Table 1). A total of 73 sequences could be grouped into 17 protein families with family sizes ranging from 2 to 16. Most of the FLcDNAs were between 0.5 and 2 kb in length (Fig. 1). The majority (145) contained both 5′ and 3′ untranslated regions (Table 3) with 118 entries containing complete 3′ UTR sequences as indicated by the presence of 10 or more consecutive adenine bases at the 3′ ends of the sequences. The 3′ UTR sequences were longer than the 5′ UTR sequences possibly reflecting in part the lack of complete 5′ UTR sequences in many cases.

3.2. Nucleotide composition

Analysis of nucleotide composition of the FLcDNA sequences revealed that the coding regions of *P. infestans* genes have higher G + C content (57%) compared to noncoding regions (45% for 3′ UTR and 56% for 5′ UTR). The G + C content of *P. sojae* was similarly estimated at 58% based on EST sequences (Qutob et al., 2000).

3.3. ATG context

Alignment of nucleotide sequences (−10 to +6) around the ATG start codons of the FLcDNA showed weak but significant conservation (Fig. 2). The consensus sequence ACCATGA was similar to the Kozak sequence that is conserved in the majority of eukaryotic translation start sites (Kamoun, 2003).

3.4. Polyadenylation signal

Among the data set, 118 sequences contained 10 or more consecutive adenine residues (poly(A) track) at their 3′-ends. This enabled us to screen the 3′-end sequences preceding the poly(A) track for polyadenylation signals. We found that the eukaryotic polyadenylation signal (AAUAAA) was conserved in this data set. A synopsis of the molecular architec-

Table 2
Sequence databases used to analyze the *P. infestans* FLcDNA data set

Data set[a]	Subset[a]	Taxon/species	Data source[b]	No. peptide sequences
Fungi	Non-pathogenic fungi and saprophytes	*Saccharomyces cerevisiae*[c]	SGD	5855
		Schizosaccharomyces pombe[c]	Sanger	4992
		Neurospora crassa[c]	Broad	10620
		Phanerochaete chrysosporium[c]	JGI	10048
		Aspergillus nidulans[c]	Broad	9541
	Plant pathogenic fungi	*Magnaporthe grisea*[c]	Broad	11109
		Fusarium graminearum[c]	Broad	11640
		Ustilago maydis[c]	Broad	6522
	Animal pathogenic fungi (I)[d]	*Candida albicans*[c]	Stanford	9256
		Cryptococcus neoformans[c]	TIGR	5882
	Animal pathogenic fungi (II)[d]	*Aspergillus fumigatus*[c]	TIGR	9926
		Chaetomium globosum[c]	Broad	11124
Others	Plants	*Arabidopsis thaliana*[c]	TAIR	25545
		Oryza sativa[c]	GenBank	55299
	Animals	*Homo sapiens*[c]	GenBank	25319
		Mus musculus[c]	GenBank	25371
		Drosophila melanogaster[c]	FlyBase	18107
		Caenorhabditis elegans[c]	WormBase	17083
	Discicristates	*Euglenozoa*	GenBank	9176
	Alveolates	*Plasmodium falciparum*[c]	Sanger	5334
		Ciliates	GenBank	2249

[a] "Fungi" data set contains sequences from fungi with different growth habits and "Others" contains sequences from representative organisms of other major phyla.

[b] Protein sequences were acquired from the following resources: SGD, *Saccharomyces* genome database (http://www.yeastgenome.org); Sanger, The Sanger Institute (http://www.sanger.ac.uk); Broad, Broad Institute (http://www.broad.mit.edu); JGI, Joint Genome Institute, US Department of Energy (http://www.jgi.doe.gov); Stanford, The Stanford Genome Technology Center (http://www.sequence.stanford.edu/group/candida); TIGR, The Institute for Genome Research (http://www.tigr.org); TAIR, The Arabidopsis Information Resource (http://www.arabidopsis.org); GenBank, National Center for Biotechnology Information (http://www.ncbi.nih.gov); FlyBase, The FlyBase Consortium, Genetics Society of America (http://flybase.bio.indiana.edu); WormBase, WormBase website (http://www.wormbase.org).

[c] Organisms for which complete predicted proteomes are available.

[d] Animal pathogenic fungi are divided into two subgroups: (I) specialized animal pathogens and (II) opportunistic pathogens with substantial saprophytic stages on decaying plant materials.

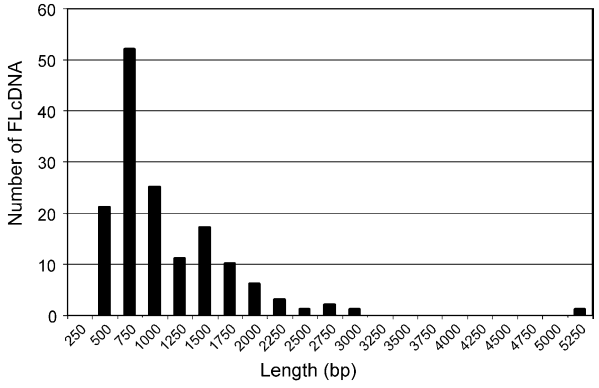

Fig. 1. Length distribution of the 150 *P. infestans* FLcDNA data set.

ture of putative polyadenylation signals in *P. infestans* is presented in Fig. 3. The 3′ UTR of 86 out of the 118 cDNAs contained the consensus hexanucleotide or a single nucleotide substitution in either AAUAAA or AAUGAA. Seven of the 86 cDNAs contained two or more putative polyadenylation sites. AAUAAA was the most prevalent of the identified polyadenylation motifs (Fig. 3).

Table 3
Features of 150 *P. infestans* FLcDNA sequences

Feature	Number	Average length (bp)	Shortest (bp)	Longest (bp)	G + C content (%)
FLcDNA	150	980	305	5121	55
ORF	150	811	153	4824	57
3′ UTR	145	110	6	444	45
5′ UTR	145	41	1	293	56

3.5. Codon usage

Nucleotide sequences of the coding regions in the FLcDNA set were used to calculate codon usage frequencies for *P. infestans* (Supplementary Table 1). This table was similar to a previously published codon usage table of *P. infestans* (Randall et al., 2005). Codon bias towards G and C nucleotides, particularly in the third 'wobble' position, was noted for almost all amino acids encoded by degenerate codons. This was also reflected in the high G + C content of the coding sequences. However, as reported for other GC-rich organisms, the codon 'GGG' encoding for

Fig. 2. Sequence conservation around translational start codons of *P. infestans*. Consensus sequences surrounding the ATG start codon (−10 to +3) were calculated based on 150 FLcDNA set using WebLogo server at http://weblogo.berkeley.edu (Crooks et al., 2004).

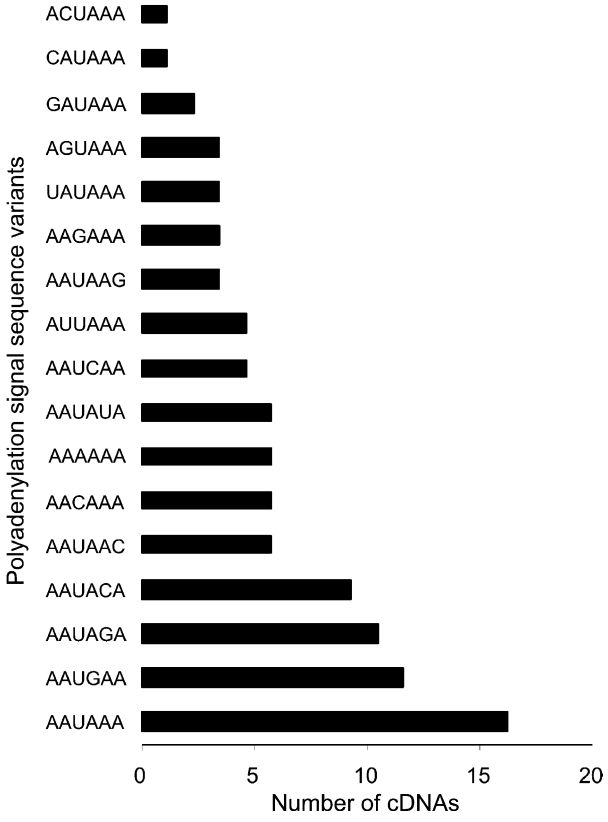

Fig. 3. Distribution of polyadenylation signal variants present in *P. infestans* FLcDNA data set. Polyadenylation signals were determined based on the eukaryotic consensus sequence of AAUAAA and its variants in the 3′ UTR of 118 *P. infestans* sequences.

glycine residue was under-represented suggesting potential selection against G-homotetramer (De Amicis and Marchetti, 2000).

3.6. Intron sequences

To identify intron junctions, we performed BLASTN searches of the FLcDNA sequences against *P. infestans* WGS reads (GenBank Trace Archives and Syngenta *Phytophthora* Consortium Database) and 10 BAC sequences (GenBank). Identical genome sequence matches were found for 101 FLcDNAs and the full gene sequence was recovered for 33 of these cDNAs. Among the 33 genes, eight genes (27%) had introns (six genes had one intron each and two genes had two introns each). Nine introns were found in seven of the 68 partially covered genes. Thus, a total of 19 introns were identified in 15 different genes. The average length of the introns was 84 bp with the majority of the introns between 61 and 80 bp (Fig. 4). G + C content of the introns was 44%, which was significantly less than the G + C content of coding regions (57%). We examined conservation at intron junctions based on 19 introns. The dinucleotides at each end of the introns (GT at the 5′ end and AG at the 3′ end) were totally conserved in all introns and intron sequences at the exon–intron junctions followed previously reported conserved sequence of 5′-GT RNGT…YAG-3′ (Fig. 5) (Kamoun, 2003). However, only six introns contained a conserved sequence CTRAC (Fig. 5) similar to the common branch point sequence (CUAAC) that is involved in lariat formation and intron splicing (Green, 1991). The introns of *P. infestans* did not show similarity to the genomic sequences of *P. sojae* and *P. ramorum* suggesting that intron sequences are diverse within the *Phytophthora* spp.

3.7. Phytophthora infestans protein sequence identities to P. sojae and P. ramorum sequences

To determine sequence conservation of the 150 *P. infestans* proteins within the *Phytophthora* genus, we used TBLASTN to search the WGS reads of *P. sojae* and *P. ramorum*. To identify trends and differences in the BLAST results, we extracted the amino acid identity of the top hits and visualized them using scatter plots (Fig. 6). Amino acid sequence identities between the *P. infestans* sequences and their top matches in these two species ranged from 27 to 100% over regions of significant similarity (Fig. 6). The average identity was 76% to *P. sojae* sequences and 75% to *P. ramorum*. The majority of the proteins were highly conserved within the *Phytophthora* species, with two-thirds showing more than 70% identity. Twenty *P. infestans* proteins were the most divergent with less than 50% amino acid identities to their best matches and can be considered as relatively fast evolving in *Phytophthora* (Supplementary Table 2). These included several secreted proteins that have been implicated in virulence, such as members of the CRN family of necrosis-inducing proteins, small cysteine-rich proteins, in planta-induced protein IPIO1, as well as a peroxidase-like protein, three proteins with

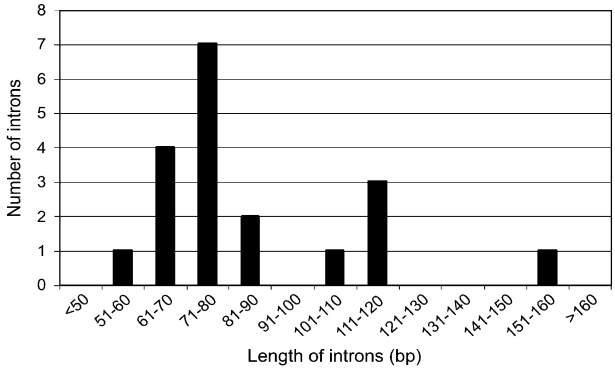

Fig. 4. Length distribution of the introns of 15 *P. infestans* genes.

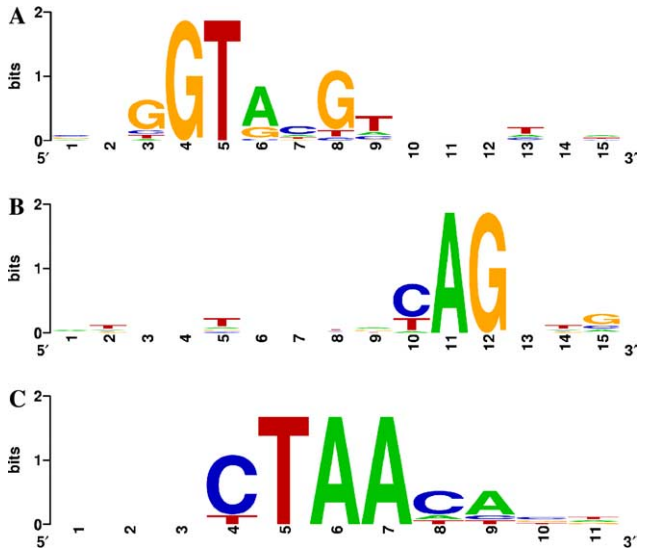

Fig. 5. Sequence conservation in *P. infestans* introns. Consensus sequences for (A) 5′ exon–intron junctions, (B) 3′ intron–exon junctions, and (C) putative branch point sequences were calculated based on 19 introns for the junctions and 6 introns for the branch point using WebLogo server at http://weblogo.berkeley.edu (Crooks et al., 2004). The intron sequences start at position 4 in (A) and end at position 12 in (B).

similarity to highly abundant mRNA protein 34 (HAM34-like), mating-induced protein M96, and two proteins of unknown function. There was only one protein, namely the small cysteine-rich protein SCR50 (GenBank Accession No. AF424679), that did not have a matching sequence above the cutoff in the *P. ramorum* genome, but showed 35% amino acid identity to a *P. sojae* sequence.

The plot of amino acid identities showed that the *P. infestans* proteins have more or less equal identities to their top matches in *P. sojae* and *P. ramorum* (Fig. 6). This is consistent with ribosomal gene phylogenetic analyses of *Phytophthora* that suggest that *P. infestans*, *P. sojae*, and *P. ramorum* fall in three distinct clusters (Cooke et al., 2000; Kroon et al., 2004).

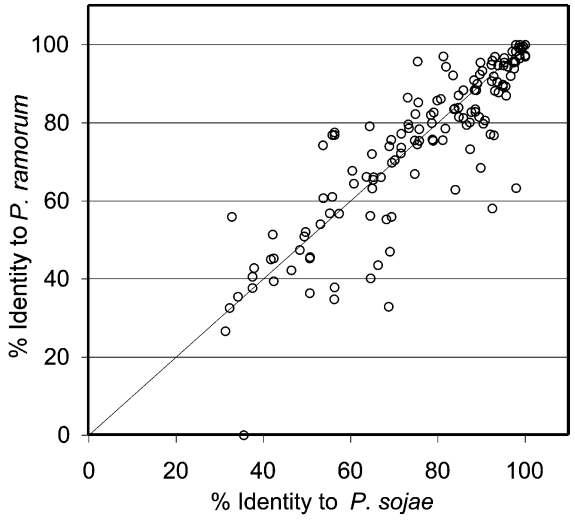

Fig. 6. Scatter plot of amino acid identities (%) of *P. infestans* proteins to *P. sojae* proteins vs. *P. ramorum* proteins. Amino acid identities were extracted from top high scoring pairs from TBLASTN searches of *P. infestans* protein sequences against WGS sequences of *P. sojae* and *P. ramorum*.

3.8. Extracellular proteins have evolved faster than cellular proteins in Phytophthora

To determine differences in divergence rates between cellular and extracellular proteins, we classified the *P. infestans* proteins based on the presence or absence of a secretory signal peptide. Using the criteria of Torto et al. (2003), 86 of the 150 proteins were predicted to contain signal peptides (Table 1). We observed significant differences in degree of conservation among *Phytophthora* species (Fig. 7). Extracellular proteins had average identity of 67% towards both *P. sojae* and *P. ramorum* proteins. This was considerably lower than cellular proteins, which had 90 and 88% identity to *P. sojae* and *P. ramorum* proteins, respectively. Fifty *P. infestans* proteins were the most divergent with less than 70% amino acid identities to their best matches against the *P. sojae* proteins (Supplementary Table 2). Forty-five of these 50 proteins were predicted to contain signal peptides indicating that the great majority of cellular proteins were highly conserved. These data suggest that, in average, genes encoding extracellular proteins are evolving faster than those encoding cellular products in *Phytophthora*.

3.9. Proteins shared between P. infestans and fungi but absent in other eukaryotes

Molecular phylogenies indicate that oomycetes are distantly related to fungi, despite apparent similarities. We took advantage of the availability of several eukaryotic genomes to investigate whether there are proteins that are uniquely shared by oomycetes and fungi. To this end, we put together a data set covering five major phyla of eukaryotes: fungi, animals, plants, alveolates, and discicristates (Table 2). The data included the complete proteomes of at least one species for all phyla except for discicristates. To identify *P. infestans* proteins that are conserved in fungi but absent in other eukaryotes, we plotted amino acid identities of BLASTP top matches to fungi vs. other eukaryotes (listed as "others" in Table 2) (Fig. 8). The majority of the data points scattered more or less along the diagonal line with no major bias towards either one of the two data sets. Average amino acid identity of *P. infestans* proteins was 47% to fungal sequences and 49% to other eukaryotes. A total of 43 *P. infestans* proteins did not show similarities to sequences in both data sets and will be referred to as putative "*Phytophthora*-specific proteins" (indicated with superscript "d" in Table 1). We also discovered *P. infestans* proteins that showed biased phylogenetic distribution (Table 4). Seven *P. infestans* proteins had matches in fungi but none in the other eukaryotes. Inversely, eight *P. infestans* proteins had matches in others but none in fungi. Proteins uniquely shared by *P. infestans* and fungi were annotated as the necrosis-inducing protein NPP1 and other NPP1-like proteins, cutinase, two peroxidase-like proteins, and a DAHP synthase (Table 4).

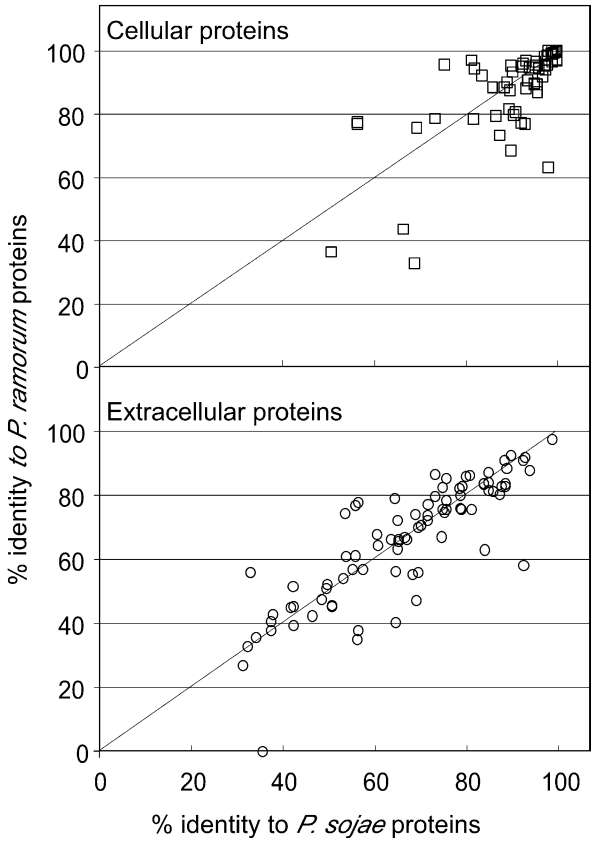

Fig. 7. Scatter plots of amino acid identities (%) of *P. infestans* cellular and extracellular proteins to *P. sojae* proteins vs. *P. ramorum* proteins. Amino acid identities were extracted from top high scoring pairs from TBLASTN searches of *P. infestans* protein sequences against WGS sequences of *P. sojae* and *P. ramorum*. (Top panel) Distribution of amino acid identities for cellular proteins. (Bottom panel) Distribution of amino acid identities for extracellular proteins. Extracellular proteins were determined by presence of a signal peptide predicted using the PexFinder algorithm (Torto et al., 2003).

3.10. *Phytophthora infestans* and phytopathogenic fungi share a set of proteins that are absent in specialized animal pathogenic fungi

We took advantage of the recent completion of the genome sequence of several fungal pathogens to address whether there are proteins that are shared by *P. infestans* and plant pathogenic fungi, but absent in animal pathogenic fungi. We divided the animal pathogenic fungi into two groups (Table 2) based on their adopted life styles. Animal fungi (I) contain fungi that are known to be specialized animal pathogens such as *Candida albicans* (Odds, 1988) and *Cryptococcus neoformans* (Casadevall and Perfect, 1998). Animal fungi (II) contain *Aspergillus fumigatus* and *Chaetomium globosum* that are known to exist largely as saprophytes on decaying plant materials and considered to be opportunistic pathogens (Sigler, 2003; Tekaia and Latgé, 2005). First, we compared the BLASTP hits of *P. infestans* protein data set against sequence databases of all plant pathogenic fungi vs. all animal pathogenic fungi (I + II) in a scattered graph (Fig. 9A) to survey the general conservation between the two groups. The majority of the data points scattered more or less along the diagonal line suggesting that majority of the *P. infestans* proteins in the data set hold similar degree of conservation towards both plant and animal pathogenic fungi. Only three proteins, two GIP-like proteins (AY961449 and AY961450) and a protein of unknown function (AF424671), were common to *P. infestans* and plant fungi but were missing in animal fungi. Second, considering that *A. fumigatus* and *C. globosum* are opportunistic fungi largely found in decaying plant materials, we removed these two species from the comparison and compared the BLASTP hits of *P. infestans* proteins to plant fungi vs. the hits to specialized animal fungi (Fig. 9B). Interestingly, we found 10 proteins that were common to *P. infestans* and the plant pathogenic species, but absent in specialized animal pathogenic fungi (Table 4). Remarkably, no sequences were uniquely shared by *P. infestans* and the specialized animal pathogens. The plant pathogen-specific proteins that were missing from specialized animal fungi included the necrosis-inducing protein NPP1, two NPP1-like proteins, a cutinase, two peroxidase-like proteins, a β-glucosidase/xylosidase, two trypsin protease glucanase inhibitor protein (GIP)-like proteins, and a protein of unknown function. However, all these proteins were found in saprophytic fungi such as *Neurospora crassa* and *Aspergillus nidulans*, which are usually associated with dead plants and are capable of feeding on decaying plant materials.

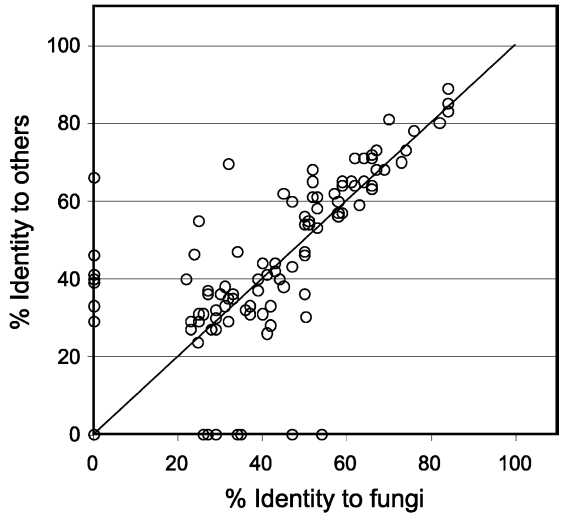

Fig. 8. Scatter plot of amino acid identities (%) of *P. infestans* proteins to fungal proteins vs. proteins from other eukaryotes. Amino acid identities were extracted from top high scoring pairs from BLASTP searches of *P. infestans* protein sequences against protein sequence databases of fungi and other eukaryotes as shown in Table 2.

4. Discussion

In this study, we put together a data set of 150 near full-length sequences for the oomycete plant pathogen *P. infestans* and performed a variety of computational analyses on

Table 4
Phytophthora infestans proteins that show biased phylogenetic distribution

Accession	Protein identity	Distribution[a]			
		Phytophthora infestans and animal fungi (I)[b]	*Phytophthora nfestans* and animal fungi (II)[b]	*Phytophthora infestans* and plant fungi[b]	*Phytophthora infestans* and others[b]
AF424663	DAHP synthase	λ	λ	λ	
AF424685	Peroxidase-like protein		λ	λ	
AF424690	Peroxidase-like protein		λ	λ	
AY961421	Cutinase		λ	λ	
AY961431	NPP1-like protein		λ	λ	
AY961432	NPP1-like protein		λ	λ	
AY961417	NPP1-like protein		λ	λ	
AF352032	β-Glucosidase/xylosidase		λ	λ	λ
AY961450	Trypsin protease GIP like			λ	λ
AY961449	Trypsin protease GIP like			λ	λ
AF424671	Unknown protein			λ	λ
AY961422	Cysteine protease		λ		λ
AY961424	Cathepsin-like cysteine protease		λ		λ
AY961425	Cathepsin-like cysteine protease		λ		λ
AF424684	Acidic chitinase				λ
AF507057	Croquemort-like mating protein M82				λ
AY586273	Kazal-like serine protease inhibitor EPI1				λ
AY586274	Kazal-like serine protease inhibitor EPI2				λ
AY586276	Kazal-like serine protease inhibitor EPI4				λ
AY586281	Kazal-like serine protease inhibitor EPI9				λ
AF424645	Pyrophosphatase				λ
AY961481	Unknown protein				λ

[a] Determined using BLASTP searches of *P. infestans* proteins against the databases of organisms described in Table 2 with an *E* value cut-off at 0.001. 'λ' denotes a protein that is found in both *P. infestans* and a group of organisms being compared. For example, peroxidase-like proteins (AF424685 and AF424690) and NPP1-like proteins (AY961432 and AY961417) from *P. infestans* have similar proteins in plant pathogenic fungi and opportunistic animal pathogenic fungi (II), but are missing in specialized animal pathogenic fungi (I) and other eukaryotes that are not fungi or oomycetes (see data set "Others" in Table 2).

[b] Detailed descriptions of the organisms in these groups are given in Table 2.

these sequences. These analyses helped to further define the structure of *P. infestans* genes with regard to G + C content, ATG context consensus, polyadenylation signals, codon usage, as well as intron size, composition, and junction consensus. Some of these features, such as G + C content of UTRs, polyadenylation signals, and intron sequence composition, have not been reported earlier for *Phytophthora* genes. Other features, such as consensus sequences at intron junctions, have been reported but were based on intron predictions collected from GenBank, most of which were not experimentally validated (Kamoun, 2003). In contrast, all the intron predictions described here were based on comparisons of cDNA and genomic sequences. Overall, these analyses help to improve our understanding of *Phytophthora* gene structure and complement existing partial cDNA sequences for the development of gene prediction and annotation programs for *Phytophthora*.

We confirmed previous reports that the average G + C content of transcribed sequences of *Phytophthora* is high at about 57–58% (Qutob et al., 2000). Considering that transcribed sequences of plants have only 42% average G + C content, high G + C content was exploited for identifying *Phytophthora* sequences in pools of expressed sequence tags (ESTs) derived from *Phytophthora*–plant interaction libraries (Hraber and Weller, 2001; Qutob et al., 2000). Interestingly, untranslated regions and introns have significantly lower G + C contents at 45 and 44%, respectively. This suggests that differences in G + C content across potential coding regions could be used as a criterion for defining gene models from genome sequences of *Phytophthora*. The codon usage determined in this study was similar to a previously published codon usage table of *P. infestans* (Randall et al., 2005). Additionally, codon usage for secreted proteins was found to be similar to the usage for non-secreted proteins in our follow-up analysis. The average 3′ UTR lengths were only slightly different between non-secreted protein cDNAs (132 bp) and secreted protein cDNAs (100 bp). This suggests that the gene structures corresponding to secreted proteins and non-secreted proteins are not significantly different in *P. infestans*. The majority of genes corresponding to the FLcDNA set are free of introns in line with previous estimates that at least two-thirds of *P. infestans* genes are intronless (Kamoun, 2003). In this study, we found introns in only 15 genes among 101 cDNA sequences that were covered at least in part by genomic sequences. Dinucleotide pairs (5′-GT…AG-3′) at the intron splice sites were perfectly conserved in *P. infestans* as previously reported (Kamoun, 2003). However, sequences flanking the dinucleotides were highly variable and their utility for reliable predictions is questionable. Introns were small with an average length of 84 bp. It was particularly interesting to note that all introns identified in this study

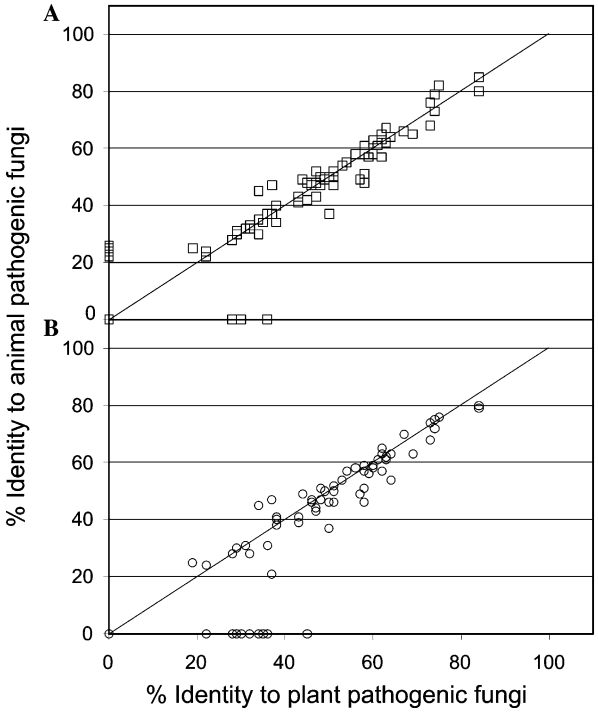

Fig. 9. (A) Scatter plot of amino acid identities (%) of *P. infestans* proteins to proteins of plant pathogenic fungi vs. proteins from animal pathogenic fungi (I + II). (B) Scatter plot of amino acid identities (%) of *P. infestans* proteins to proteins of plant pathogenic fungi vs. proteins from specialized animal pathogenic fungi (I). Amino acid identities were extracted from top high scoring pairs from BLASTP searches of *P. infestans* protein sequences against protein sequence databases of plant pathogenic fungi and animal pathogenic fungi (I) and (II) as shown in Table 2.

were less than 152 bp in length. Standard computer-generated gene finding programs often call introns of several hundred bp or longer for *Phytophthora* sequences. We recommend that users keep in mind that such predictions may not be realistic for *Phytophthora* genes. We did not discover any microexons or alternative splice sites within the examined data set.

We used the *P. infestans* FLcDNA sequences to perform comparative analyses within *Phytophthora*. Comparison of the FLcDNA data set to WGS sequences of *P. sojae* and *P. ramorum* revealed conserved proteins as well as divergent ones with an average amino acid identity of 75–76%. As expected, many of the conserved proteins, such as ribosomal proteins and metabolic enzymes, were predicted to carry housekeeping functions. Most of the fast evolving proteins carried signal peptides and were predicted to be extracellular. Indeed, 45 of the 50 *P. infestans* proteins with less than 70% identity to their top matches in *P. sojae* and *P. ramorum* were predicted to be extracellular. Of these, 20 proteins were identified as particularly fast evolving with less than 50% matches to their top hits. Several of these, such as CRN-like proteins (Torto et al., 2003), small cysteine-rich proteins (SCRs) (Bos et al., 2003; Liu et al., 2005), in planta-induced (IPI) proteins (Pieterse et al., 1994a,b; Van West et al., 1998), and Kazal-like serine protease inhibitors (EPIs) (Tian et al., 2004, 2005), have been previously implicated in infection as putative virulence factors. Our findings are consistent with the view that these genes may function in virulence. However, we would like to point out that the secreted proteins in this data were not selected entirely randomly, and are enriched in proteins involved in plant–pathogen interactions. Thus, it is reasonable to expect them to be more divergent than non-secreted proteins. Future whole genome-scale analyses will show whether our observations will hold for the entire proteome.

We exploited the FLcDNA sequences to perform comparative analyses between *Phytophthora* and other eukaryotes. We identified 43 proteins that are unique to *Phytophthora* and absent in the other examined eukaryotes (indicated with superscript "d" in Table 1). Several of these proteins were previously functionally characterized. These include CBEL (Gaulin et al., 2002; Mateos et al., 1997), CRN family of necrosis-inducing proteins (Torto et al., 2003), INF elicitins (Kamoun et al., 1997), as well as IPIB and IPIO proteins that are encoded by in planta-induced genes (Pieterse et al., 1994a,b; Van West et al., 1998). Other *Phytophthora*-specific proteins have unknown functions but include small cysteine-rich (SCR) proteins that are reminiscent of pathogen effectors. These include the products of the *scr74* and *scr91* genes, which are polymorphic and under diversifying selection in *P. infestans* (Bos et al., 2003; Liu et al., 2005).

Comparative analyses of the *P. infestans* sequences and the complete proteomes of fungal pathogens resulted in a notable finding. Ten proteins were uniquely conserved between the plant pathogens but were missing in specialized animal fungi (*C. albicans* and *C. neoformans*) despite the fact that *P. infestans* and the plant pathogenic fungi are highly divergent eukaryotes, whereas the plant and animal pathogenic fungi are relatively closely related. Remarkably, no sequences were uniquely shared by *P. infestans* and the specialized animal pathogens. These results suggest that eukaryotic microbial pathogens that share similar lifestyles also share a similar set of genes independently of their phylogenetic relatedness. A more comprehensive analysis with a larger number of specialized animal fungi proteomes, which will undoubtedly become available in near future, is needed to confirm this conclusion. It should be noted, however, that all 10 proteins identified here occur in saprophytic fungi and animal pathogenic fungi associated with decaying plant materials (non-specialized). This is consistent with prior observations that suggest that opportunistic animal pathogens, such as *A. fumigatus*, have to rely on plant materials for survival outside their animal hosts and therefore carry a number of genes associated with utilization of plant tissue (Tekaia and Latgé, 2005).

It is tempting to speculate that the proteins shared by the plant pathogens function in infection of plants and/or growth on plant tissues and are not relevant to interactions of specialized animal fungi with their animal hosts. Indeed, all 10 proteins in this group were predicted to con-

tain a secretory signal, suggesting that they could directly interact with host plants. In addition, some of the identified proteins, such as β-glucosidase/xylosidase and cutinase, are plant cell wall degrading enzymes that directly facilitate colonization of plant tissue. Other proteins in this group include trypsin-like proteases with similarity to *P. sojae* glucanase inhibitor GIP1 (Rose et al., 2002) and the NPP-like necrosis-inducing proteins that were also described in bacterial plant pathogens (Pemberton and Salmond, 2004). In summary, these findings confirm that there is significant overlap in the arsenal of virulence factors of plant pathogenic filamentous microbes (Latijnhouwers et al., 2003; Randall et al., 2005). The similarities could reflect convergent evolution between proteins from unrelated plant pathogens. Phylogenetically distant virulence factors may have evolved to share significant similarity, perhaps by targeting similar substrates. However, considering that these proteins occur in a wide range of saprophytic fungi and non-specialized animal pathogenic fungi, differential gene loss is the most parsimonious explanation for the observed biased phylogenetic distribution. Genes encoding the 10 proteins may have been preserved in both the oomycete and plant-associated fungal lineages, perhaps because they are essential for colonization of plant tissue. Loss of these genes in animal pathogen lineages (e.g., *C. albicans* and *C. neoformans*) may have accompanied host specialization and loss of saprophytic stages in these fungi. Alternatively, horizontal gene transfer events may have taken place in which plant pathogenic oomycetes acquired a set of genes encoding virulence factors from fungi. Future analyses will help test these various hypotheses.

Acknowledgments

We thank Shujing Dong, Diane Kinney, and several members of our laboratory for technical assistance, the Syngenta *Phytophthora* Consortium for access to sequences of *P. infestans*, and the editor for useful suggestions. This work was supported by the NSF Plant Genome Research Program Grant DBI-0211659. Salaries and research support were also provided by State and Federal Funds appropriated to the Ohio Agricultural Research and Development Center, the Ohio State University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fgb.2005.10.003.

References

Ashurst, J.L., Collins, J.E., 2003. Gene annotation: prediction and testing. Annu. Rev. Genomics Hum. Genet. 4, 69–88.

Baldauf, S.L., 2003. The deep roots of eukaryotes. Science 300, 1703–1706.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., Doolittle, W.F., 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290, 972–977.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al., 2004. The Pfam protein families database. Nucleic Acids Res. 32, D138–D141.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A., Wheeler, D.L., 1999. GenBank. Nucleic Acids Res. 27, 12–17.

Birch, P.R.J., Whisson, S.C., 2001. *Phytophthora infestans* enters the genomics era. Mol. Plant. Pathol. 2, 257–263.

Borevitz, J.O., Ecker, J.R., 2004. Plant genomics: the third wave. Annu. Rev. Genomics Hum. Genet. 5, 443–477.

Bos, J.I.B., Armstrong, M., Whisson, S.C., Torto, T.A., Ochwo, M., Birch, P.R.J., Kamoun, S., 2003. Intraspecific comparative genomics to identify avirulence genes from *Phytophthora*. New Phytol. 159, 63–72.

Brendel, V., Xing, L., Zhu, W., 2004. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics 20, 1157–1169.

Casadevall, A., Perfect, J.R., 1998. *Cryptococcus neoformans*. ASM Press, Washington, DC.

Castelli, V., Aury, J.M., Jaillon, O., Wincker, P., Clepet, C., Menard, M., Cruaud, C., Quetier, F., Scarpelli, C., Schachter, V., et al., 2004. Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. Genome Res. 14, 406–413.

Cooke, D.E., Drenth, A., Duncan, J.M., Wagels, G., Brasier, C.M., 2000. A molecular phylogeny of *Phytophthora* and related oomycetes. Fungal Genet. Biol. 30, 17–32.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: A sequence logo generator. Genome Res. 14, 1188–1190.

Davuluri, R.V., Zhang, M.Q., 2003. Computer software to find genes in plant genomic DNA. Methods Mol. Biol. 236, 87–108.

De Amicis, F., Marchetti, S., 2000. Intercodon dinucleotides affect codon choice in plant genes. Nucleic Acids Res. 27, 339–3345.

Duncan, J.M., 1999. *Phytophthora*—an abiding threat to our crops. Microbiol. Today 26, 114–116.

Frech, G.C., Joho, R.H., 1989. Construction of directional cDNA libraries enriched for full-length inserts in a transcription-competent vector. Gene Anal. Tech. 6, 33–38.

Fry, W.E., Goodwin, S.B., 1997. Re-emergence of potato and tomato late blight in the United States. Plant Dis. 81, 1349–1357.

Gaulin, E., Jauneau, A., Villalba, F., Rickauer, M., Esquerre-Tugaye, M.T., Bottin, A., 2002. The CBEL glycoprotein of *Phytophthora parasitica* var *nicotianae* is involved in cell wall deposition and adhesion to cellulosic substrates. J. Cell Sci. 115, 4565–4575.

Green, M.R., 1991. Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. Annu. Rev. Cell Biol. 7, 559–599.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr., R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al., 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31, 5654–5666.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., Salzberg, S.L., 2002. Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol. 3, research 0029.0021–0029.0012.

Hayashizaki, Y., 2003a. RIKEN mouse genome encyclopedia. Mech. Aging Dev. 124, 93–102.

Hayashizaki, Y., 2003b. The RIKEN mouse genome encyclopedia project. C. R. Biol. 326, 923–929.

Hraber, P., Weller, J.W., 2001. On the species of origin: diagnosing the source of symbiotic transcripts. Genome Biol. 2, research 0037.

Huitema, E., Bos, J.I., Tian, M., Win, J., Waugh, M.E., Kamoun, S., 2004. Linking sequence to phenotype in *Phytophthora*–plant interactions. Trends Microbiol. 12, 193–200.

Judelson, H.S., Randall, T.A., 1998. Families of repeated DNA in the oomycete *Phytophthora infestans* and their distribution within the genus. Genome 41, 605–615.

Kamoun, S., 2003. Molecular genetics of pathogenic oomycetes. Eukaryot. Cell 2, 191–199.

Kamoun, S., Hraber, P., Sobral, B., Nuss, D., Govers, F., 1999. Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences. Fungal Genet. Biol. 28, 94–106.

Kamoun, S., Lindqvist, H., Govers, F., 1997. A novel class of elicitin-like genes from *Phytophthora infestans*. Mol. Plant Microbe Interact. 10, 1028–1030.

Kato, S., Sekine, S., Oh, S.W., Kim, N.S., Umezawa, Y., Abe, N., Yokoyama-Kobayashi, M., Aoki, T., 1994. Construction of a human full-length cDNA bank. Gene 150, 243–250.

Kroon, L.P., Bakker, F.T., van den Bosch, G.B., Bonants, P.J., Flier, W.G., 2004. Phylogenetic analysis of *Phytophthora* species based on mitochondrial and nuclear DNA sequences. Fungal Genet. Biol. 41, 766–782.

Latijnhouwers, M., de Wit, P.J., Govers, F., 2003. Oomycetes and fungi: similar weaponry to attack plants. Trends Microbiol. 11, 462–469.

Liu, Z., Bos, J.I., Armstrong, M., Whisson, S.C., da Cunha, L., Torto-Alalibo, T., Win, J., Avrova, A.O., Wright, F., Birch, P.R., Kamoun, S., 2005. Patterns of diversifying selection in the phytotoxin-like *scr74* gene family of *Phytophthora infestans*. Mol. Biol. Evol. 22, 659–672.

Manley, J.L., Proudfoot, N.J., 1994. RNA 3′ ends: formation and function. Genes Dev. 8, 259–264.

Mateos, F.V., Rickauer, M.E., Esquerré-Tugayé, M.T., 1997. Cloning and characterization of a cDNA encoding an elicitor of *Phytophthora parasitica* var. *nicotianae* that shows cellulose-binding and lectin-like activities. Mol. Plant Microbe Interact. 10, 1045–1053.

Odds, F.C., 1988. Candida and Candidosis, second ed. Saunders, Philadelphia.

Ogihara, Y., Mochida, K., Kawaura, K., Murai, K., Seki, M., Kamiya, A., Shinozaki, K., Carninci, P., Hayashizaki, Y., Shin, I.T., et al., 2004. Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags. Genes Genet. Syst. 79, 227–232.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al., 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563–573.

Pemberton, C.L., Salmond, G.P.C., 2004. The Nep1-like proteins—a growing family of microbial elicitors of plant necrosis. Mol. Plant. Pathol. 5, 353–359.

Pieterse, C.M., Derksen, A.M., Folders, J., Govers, F., 1994a. Expression of the *Phytophthora infestans ipiB* and *ipiO* genes in planta and in vitro. Mol. Gen. Genet. 244, 269–277.

Pieterse, C.M., van West, P., Verbakel, H.M., Brasse, P.W., van den Berg-Velthuis, G.C., Govers, F., 1994b. Structure and genomic organization of the *ipiB* and *ipiO* gene clusters of *Phytophthora infestans*. Gene 138, 67–77.

Qutob, D., Hraber, P.T., Sobral, B.W., Gijzen, M., 2000. Comparative analysis of expressed sequences in *Phytophthora sojae*. Plant Physiol. 123, 243–254.

Randall, T.A., Dwyer, R.A., Huitema, E., Beyer, K., Cvitanich, C., Kelkar, H., Ah Fong, A.M.V., Gates, K., Roberts, S., Yatzkan, E., et al., 2005. Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol. Plant Microbe Interact. 18, 229–243.

Rose, J.K., Ham, K.S., Darvill, A.G., Albersheim, P., 2002. Molecular cloning and characterization of glucanase inhibitor proteins: coevolution of a counterdefense mechanism by plant pathogens. Plant Cell 14, 1329–1345.

Rothnie, H.M., Reid, J., Hohn, T., 1994. The contribution of AAUAAA and the upstream element UUUGUA to the efficiency of mRNA 3′-end formation in plants. EMBO J. 13, 2200–2210.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., et al., 2002. Functional annotation of a full-length *Arabidopsis* cDNA collection. Science 296, 141–145.

Sigler, L., 2003. Miscellaneous opportunistic fungi: Microascaceae and other Ascomycetes, Hyphomycetes, Coelomycetes, and Basidiomycetes. In: Howard, D.H. (Ed.), Pathogenic Fungi in Humans and Animals, second ed. Marcel Dekker, New York, pp. 637–676.

Sogin, M.L., Silberman, J.D., 1998. Evolution of the protists and protistan parasites from the perspective of molecular systematics. Int. J. Parasitol. 28, 11–20.

Stein, L., 2001. Genome annotation: from sequence to biology. Nat. Rev. Genet. 2, 493–503.

Tekaia, F., Latgé, J.P., 2005. *Aspergillus fumigatus*: saprophyte or pathogen? Curr. Opin. Microbiol. 8, 385–392.

Tian, M., Benedetti, B., Kamoun, S., 2005. A second Kazal-like protease inhibitor from *Phytophthora infestans* inhibits and interacts with the apoplastic pathogenesis-related protease P69B of tomato. Plant Physiol. 138, 1785–1793.

Tian, M., Huitema, E., da Cunha, L., Torto-Alalibo, T., Kamoun, S., 2004. A Kazal-like extracellular serine protease inhibitor from *Phytophthora infestans* targets the tomato pathogenesis-related protease P69B. J. Biol. Chem. 279, 26370–26377.

Tooley, P.W., Therrien, C.D., 1987. Cytophotometric determination of the nuclear DNA content of 23 Mexican and 18 non-Mexican isolates of *Phytophthora infestans*. Exp. Mycol. 11, 19–26.

Torto, T.A., Li, S., Styer, A., Huitema, E., Testa, A., Gow, N.A., van West, P., Kamoun, S., 2003. EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. Genome Res. 13, 1675–1685.

Van West, P., de Jong, A.J., Judelson, H.S., Emons, A.M., Govers, F., 1998. The *ipiO* gene of *Phytophthora infestans* is highly expressed in invading hyphae during infection. Fungal Genet. Biol. 23, 126–138.

Wahle, E., 1995. 3′-end cleavage and polyadenylation of mRNA precursors. Biochim. Biophys. Acta 1261, 183–194.

Wortman, J.R., Haas, B.J., Hannick, L.I., Smith Jr., R.K., Maiti, R., Ronning, C.M., Chan, A.P., Yu, C., Ayele, M., Whitelaw, C.A., et al., 2003. Annotation of the *Arabidopsis* genome. Plant Physiol. 132, 461–468.

Supplementary Table 1

Codon usage table for *P. infestans* based on the coding regions of 150 FLcDNA sequences

Codon	Aa[a]	Fract[b]	/1000[c]	Number
GCA	A	0.116	11.682	474
GCC	A	0.368	37.092	1505
GCG	A	0.202	20.407	828
GCT	A	0.314	31.694	1286
TGC	C	0.706	13.604	552
TGT	C	0.294	5.669	230
GAC	D	0.697	39.261	1593
GAT	D	0.303	17.104	694
GAA	E	0.289	16.315	662
GAG	E	0.711	40.074	1626
TTC	F	0.735	25.903	1051
TTT	F	0.265	9.316	378
GGA	G	0.165	11.953	485
GGC	G	0.480	34.726	1409
GGG	G	0.072	5.250	213
GGT	G	0.283	20.481	831
CAC	H	0.762	14.147	574
CAT	H	0.238	4.412	179
ATA	I	0.042	1.799	73
ATC	I	0.643	27.776	1127
ATT	I	0.316	13.654	554
AAA	K	0.185	10.721	435
AAG	K	0.815	47.172	1914

CTA	L	0.064	5.077	206
CTC	L	0.239	18.854	765
CTG	L	0.363	28.688	1164
CTT	L	0.144	11.386	462
TTA	L	0.028	2.243	91
TTG	L	0.161	12.693	515
ATG	M	1.000	22.945	931
AAC	N	0.781	27.579	1119
AAT	N	0.219	7.714	313
CCA	P	0.157	8.084	328
CCC	P	0.258	13.309	540
CCG	P	0.334	17.227	699
CCT	P	0.252	13.013	528
CAA	Q	0.254	7.887	320
CAG	Q	0.746	23.192	941
AGA	R	0.056	2.218	90
AGG	R	0.062	2.465	100
CGA	R	0.114	4.559	185
CGC	R	0.373	14.886	604
CGG	R	0.065	2.588	105
CGT	R	0.330	13.161	534
AGC	S	0.209	17.400	706
AGT	S	0.099	8.207	333
TCA	S	0.076	6.359	258
TCC	S	0.166	13.851	562
TCG	S	0.342	28.515	1157
TCT	S	0.108	8.971	364
ACA	T	0.122	9.094	369

ACC	T	0.319	23.783	965
ACG	T	0.380	28.293	1148
ACT	T	0.180	13.383	543
GTA	V	0.061	4.510	183
GTC	V	0.301	22.206	901
GTG	V	0.475	35.096	1424
GTT	V	0.163	12.052	489
TGG	W	1.000	11.091	450
TAC	Y	0.806	27.012	1096
TAT	Y	0.194	6.506	264
TAA	*	0.487	1.799	73
TAG	*	0.320	1.183	48
TGA	*	0.193	0.715	29

[a]Amino acids in one letter code. '*' denotes stop codons.

[b]The 'Fract' column gives that proportion of usage of a given codon among its redundant set (i.e. the set of codons which code for this codon's amino acid). For example, the sum of the 6 codons representing serine will add up to 1.00.

[c]The /1000 column represents the number of codons, given the input sequence(s), there are per 1000 bases. This will be an extrapolation if the sequence is shorter than 1000 bases.

Supplementary Table 2

P. infestans proteins that show less than 70% sequence identity to *P. sojae* and *P. ramorum* proteins

Accession	Identity of *P. infestans* protein[a]	% Identity to *P. sojae*[b]	% Identity to *P. ramorum*[b]	Extracellular protein[c]
AF424682	unknown protein[d]	31.33	26.67	Yes
AY961426	elicitin-like protein[d]	32.24	32.62	Yes
AF424675	crinkling and necrosis-inducing protein CRN1[d]	32.80	55.98	Yes
AY961430	*in planta*-induced IPIO1[d]	34.21	35.50	Yes
AF424679	small cysteine rich protein SCR50[d]	35.48	00.00	Yes
AY961446	small cysteine-rich protein SCR74[d]	37.50	37.68	Yes
AY961447	small cysteine-rich protein SCR74[d]	37.50	40.58	Yes
AF424680	small cysteine rich protein SCR58[d]	37.84	42.86	Yes
AY961438	HAM34-like putative membrane protein[d]	41.73	45.07	Yes
AF424685	peroxidase-like protein[d]	42.17	51.43	Yes
AY961437	HAM34-like putative membrane protein[d]	42.37	39.46	Yes
AY961456	CRN-like CRN8[d]	42.41	45.30	Yes
AF507059	mating-induced protein M96[d]	46.43	42.29	Yes
AY961433	unknown protein[d]	48.36	47.41	Yes
AY961442	HAM34-like putative membrane protein[d]	49.40	50.89	Yes
AY961427	cysteine-rich protein[d]	49.68	52.07	Yes
AY961452	CRN-like CRN4[d]	50.63	36.42	No
AY961418	small cysteine-rich protein SCR91[d]	50.65	45.33	Yes
AY961448	small cysteine-rich protein SCR91[d]	50.67	45.33	Yes
AY961453	CRN-like CRN5[d]	50.72	45.63	No
AY935250	cysteine protease inhibitor EPIC1	53.08	54.03	Yes

AY961432	NPP1-like protein	53.64	74.29	Yes
AY961461	CRN-like CRN13	53.77	60.71	Yes
AF507054	elicitin-like mating protein M25	55.20	56.86	Yes
AF424677	crinkling and necrosis-inducing protein CRN2	55.83	76.92	Yes
AY961451	CRN-like CRN3	55.83	76.92	Yes
AY586281	Kazal-like serine protease inhibitor EPI9	55.84	61.04	Yes
AY961454	CRN-like CRN6	56.29	34.82	Yes
AY961459	CRN-like CRN11	56.31	76.92	No
AY961455	CRN-like CRN7	56.31	77.56	Yes
AY961458	CRN-like CRN10	56.31	77.56	No
AY586274	Kazal-like serine protease inhibitor EPI2	56.36	37.88	Yes
AY961457	CRN-like CRN9	57.34	56.76	Yes
AY586276	Kazal-like serine protease inhibitor EPI4	60.42	67.75	Yes
AY935254	cysteine protease inhibitor EPIC4	60.74	64.42	Yes
AY961429	*in planta*-induced IPIB-like protein	63.64	66.15	Yes
AY961431	NPP1-like protein	64.37	79.10	Yes
AY961440	HAM34-like putative membrane protein	64.52	56.16	Yes
AY586273	Kazal-like serine protease inhibitor EPI1	64.58	40.16	Yes
AY961434	unknown protein	64.91	72.00	Yes
AY961450	trypsin protease GIP-like	64.96	63.18	Yes
AY961428	cysteine-rich protein	65.19	65.45	Yes
AF419841	elicitin-like INF4	65.25	66.10	Yes
AF424669	small cysteine rich protein SCR108	66.97	66.06	Yes
AY961460	CRN-like CRN12	68.18	55.25	Yes
AY961464	CRN-like CRN16	68.75	32.95	No
AF424683	small cysteine rich protein SCR122	68.91	73.98	Yes
AY961463	CRN-like CRN15	69.05	47.00	Yes
AY961462	CRN-like CRN14	69.42	55.98	Yes

AY961445	low complexity protein	69.48	69.80	Yes

[a] *P. infestans* proteins were sorted according to their % identity (from low to high) to *P. sojae* proteins.

[b] Percent (%) identities were extracted from top high scoring pairs of BLASTP matches.

[c] Predicted using the PexFinder algorithm (Torto et al., 2003) based on the SignalP signal peptide prediction program. A protein sequence is assigned "Yes" for extracellular protein if the protein is predicted to be signal peptide by SignalP Hidden Markov Model (SignalP HMM) with a probability greater than 0.900, and the SignalP Neural Network-predicted cleavage site between 10 and 40 amino acid residues. Otherwise, it is assigned "No".

[d] Fast-evolving proteins with less than 50 % identity among *Phytophthora* spp.