Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Microbiological Methods 54 (2003) 339–351
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

Distinguishing plant and fungal sequences in ESTs from infected plant tissues

Tom Hsiang*, Paul H. Goodwin

Department of Environmental Biology, University of Guelph, Guelph, Ontario, Canada N1G 2W1

Received 16 January 2003; received in revised form 11 February 2003; accepted 12 February 2003

Abstract

Expressed sequence tags (ESTs) from fungal-infected plant tissues are composed of a mixture of plant and fungal sequences. Using freely available software and tools, a novel procedure is described for distinguishing plant and fungal DNA sequences. Although the GenBank non-redundant (NR) database is larger and therefore one would presume that BLASTX analysis of it would be more accurate, superior resolution of 700 randomly selected fungal ESTs was found with Standalone TBLASTX analyses with a local matching database composed of a plant and a fungal genome. Standalone TBLASTX analyses of 3983 ESTs from nine different fungal-infected plant EST libraries also proved to be superior in identifying the origin of sequences as either plant or fungal compared to GenBank BLASTX analysis. Standalone TBLASTX with a matching database comprised of a single plant and a single fungal genome appears to be a faster and more accurate method than BLASTX searches of the GenBank non-redundant database to distinguish fungal and plant sequences in mixed EST collections.

Keywords: ESTs; Fungi; Plant diseases; Standalone BLAST

1. Introduction

Genomics is increasingly being applied to the study of plant diseases caused by fungi (Soanes et al., 2002). Single-pass, partial sequencing of cDNA clones to generate expressed sequence tags (ESTs) has become a common means of identifying and analyzing large numbers of genes that are involved in fungal–plant interactions (Soanes et al., 2002). While ESTs can be obtained from fungi grown in culture or from healthy plants, it is the genes expressed during the interaction that are generally considered to be most relevant to understanding the molecular basis of an interaction. These can be plant genes related to resistance and the response to infection, or fungal genes related to virulence and the ability to grow and reproduce in the plant. Because mRNA extracted from fungal-infected plant tissues will be a mixture originating from both the plant and fungus, which are both eukaryotes, one of the necessary steps in analyzing an EST collection in fungal-plant genomic studies is to distinguish the origin of the sequences as plant or fungal.

Several EST collections from fungal-infected plants have been done thus far (Fristensky et al., 1999; Kruger et al., 2002; Quotob et al., 2000). The

* Corresponding author. Tel.: +1-519-824-4120x52753; fax: +1-519-837-0442.
E-mail addresses: thsiang@uoguelph.ca (T. Hsiang), pgoodwin@uoguelph.ca (P.H. Goodwin).

0167-7012/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0167-7012(03)00067-8

most common approach to identifying the origin of the sequences is by using search programs such as BLAST (Altschul et al., 1990) to determine the highest match with sequences in genetic databases such as GenBank at the National Center for Biotechnology Information (NCBI) of the National Library of Medicine, Washington, DC, USA. The DNA DataBank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL) also house genetic databases, and these three organizations exchange data on a daily basis. Although a comparison with large databases, such as the non-redundant (NR) protein sequence database at GenBank, is currently the best method to identify the function of a gene based on its similarity to other sequences, it is biased by the availability of sequences, which is ultimately determined by the interests of researchers and, as a result, the best matches may not always properly place the taxonomic origin of the query sequence. An illustration of this can be seen in the analysis of a crenarcheaota species and *Escherichia coli* by Koski and Golding (2001), who found that the closest BLAST hit sometimes was not even a close phylogenetic neighbor. Their explanation for this was that the matches depended upon the availability of sequences from close relatives in the databases. This is relevant to the identification of fungal ESTs from infected plant tissues because publicly accessible databases contain considerably fewer fungal sequences than plant sequences.

Another resource available for comparison with ESTs from infected plants is the complete genomic sequences of several plant and fungal species, including the plants, arabidopsis (*Arabidopsis thaliana*; Arabidopsis Genome Initiative, 2000) and rice (*Oryza sativa*; Yu et al., 2002), as well as the fungi, yeast (*Saccharomyces cerevisiae*; Cherry et al., 1998), fission yeast (*Schizosaccharomyces pombe*; Wood et al., 2002) and *Neurospora crassa* (www-genome.wi.mit.edu). Furthermore, the genomic sequences of two plant pathogens, *Magnaporthe grisea* (www-genome.wi.mit.edu) and *Phanerochaete chrysosporium* (www.jgi.doe.gov/programs/whiterot.htm), are almost complete. The advantage of looking for matches between ESTs and an entire genome is that a complete genomic database is representative of all of the nuclear genes in an organism, including those that have not been well studied in any organism or have an unknown function.

The goal of this work was to develop a method using available databases to distinguish the taxonomic origin of fungal and plant sequences in a mixture. A comparison of the effectiveness and efficiency in distinguishing fungal and plant sequences was made between using BLAST analysis of the GenBank NR protein database and a locally available method with complete genome sequences of plants and fungi. With a diverse selection of described fungal and plant genes as query sequences, an examination was made of the choice of the particular entire genomes to form the local BLAST matching database, and then comparisons were made between Standalone BLASTN, TBLASTX or MegaBLAST. Based on these results, a method using Standalone BLAST with selected genomes was applied to several previously released EST collections from fungal-infected plants to examine its utility in detecting fungal and plant sequences, and this was compared to the results obtained with a conventional BLAST search of the GenBank NR protein database.

2. Materials and methods

2.1. Source of fungal and plant query sequences and genomic databases

For each of 20 plant and 20 fungal species, five described genes were selected from the GenBank NR database (Tables 1 and 2) and downloaded as nucleotide sequences. For EST collections from individual plant or fungal species, 100 sequences each were selected from the fungi, *Agaricus bisporus*, *Botryotinia fuckeliana*, *Glomus intraradices*, *Leptosphaeria maculans*, *M. grisea*, *Pleurotus ostreatus*, and *Schizophyllum commune* (Table 1), and the plants, *Hordeum vulgare*, *Malus* × *domestica*, *Medicago truncatula*, *Pinus pinaster*, *Solanum tuberosum*, *Sorghum bicolor* and *Triticum aestivum* (Table 2). Around 500 sequences were also selected for analysis from each of 9 EST libraries originating from fungal-infected plants (Table 3). The query sequences were compared to one or more of the genomic databases from the following organisms: *A. thaliana*, *M. grisea*, *N. crassa*, *O. sativa*, *P. chrysosporium* and *S. cerevisiae*. The genomic databases were downloaded from NCBI and other web sources (Tables 1 and 2). For clarity,

Table 1
Source of fungal DNA sequences used for query data sets or BLAST databases in this study

Species	GenBank accession no.
A. bisporus	100 EST sequences from GenBank AW324525–AW444286
Alternaria alternata	AB025309, AB047682, AF282320, U82437, X78225
Blumeria graminis	AF052515, AF189366, AF247001, AJ243654, X81961
B. fuckeliana	AF215732, AF243854, AF346594, AJ428403, Z69264 100 EST sequences from www.cogeme.man.ac.uk
Cercospora nicotianae	AF035619, AF121137, AF294268, AF306523, U03903
Claviceps purpurea	AF022911, AJ011963, AJ011964, AJ318517, AJ344050
Cochliobolus carbonum	AF032368, AF306764, L48982, L48994, M98024
Colletotrichum gloeosporioides	AF156983, AJ271152, AJ291494, AJ291495, AJ311709
Fusarium solani	AF403143, AF417005, D00809, U23722, X94315
Ganoderma lucidum	AF185275, AF185275, M58032, U56129, U56134
G. intraradices	100 EST sequences from GenBank BM959274–BM959379
L. maculans	AF192405, AF240001, AF290180, AF370014, U18793 100 EST sequences from GenBank BG370024–BG370123
M. grisea	AF264035, AF293848, AF325683, BAA34046, X61500 100 EST sequences from www.cogeme.man.ac.uk/ Genomic contigs from www-genome.wi.mit.edu/
Mycosphaerella graminicola	AF038152, AF329852, AF347058, AF440398, AF440399
N. crassa	Complete genome (ver. 3) from www-genome.wi.mit.edu
Ophiostoma novo-ulmi	AF052061, AF055293, AF378546, AF378551, Z80085
Phaeosphaeria nodorum	AJ009827, AJ133695, AJ249197, AJ271154, AJ271155
P. chrysosporium	Almost entire genome, 767 scaffold sequences from ftp://ftp.jgi-psf.org/pub/JGI_data/WhiteRot/
P. ostreatus	AF332138, AF355103, AF435445, AJ238148, U91642 100 EST sequences from http://www.cogeme.man.ac.uk
Puccinia graminis	L08126, L08127, U26597, X73529, X73529
Rhizoctonia solani	AB9028493, AF339929, S75056, Z54276, Z54277
S. cerevisiae	Complete genome from GenBank with 16 chromosomes
S. commune	AB066503, AF005405, AF125094, L43072, U17012 100 EST sequences from GenBank BF942484–BG550564
Ustilago maydis	AF487462, AF494288, AF495526, AF495526, X99718
Venturia inaequalis	AF047029, AF227914, AF227920, AF363785, M97951

genomic databases are referred to by genus, while the query sequences are referred to by their species name.

2.2. BLAST analyses of plant and fungal sequences

In addition to standard Internet access to BLAST at NCBI (called WWW BLAST), analyses were also done with Standalone BLAST, where both the BLAST programs (BLASTALL ver. 2.2.3) and the sequence databases were downloaded and used locally, and with Network BLAST, where BLAST was run locally (BLASTCL3 ver. 2.2.3) but the NCBI sequence database was accessed over the net. The Standalone BLAST program for Win32 systems was downloaded from NCBI (ftp://ftp.ncbi.nih.gov/blast/executables/blastz.exe, version 2.2.3 released 22 April 2002). Full descriptions of these different BLAST services and the subroutines (e.g., TBLASTX, BLASTN and MegaBLAST) can be found at the NCBI website. All BLAST analyses were run with default parameters, except for expectation values (e-values), which were specified for each analysis.

The 200 cDNA query sequences of diverse plant and fungal origin from described genes in the NCBI database were analyzed using the nucleotide matching programs BLASTN and MegaBLAST, and the protein matching program TBLASTX. They were matched in Standalone BLAST to a database composed of two genomes, *Arabidopsis* and *Neurospora*. The 100 fungal cDNA query sequences were also matched using Standalone BLAST to each of 4 fungal genomic databases separately (*Magnaporthe*, *Neurospora*, *Phanerochaete* and *Saccharomyces*). The 100 plant cDNA query sequences were matched against 2 plant genomic

Table 2
Source of plant DNA sequences used for query data sets or BLAST databases in this study

Species	GenBank accession no.
Allium cepa	AF004946, AF212157, AF268382, AF401622, L13365
Amaranthus tricolor	AB050117, AB050123, AB050124, AF290974, U48523
Apium graveolens	AF129423, AF393808, AF480069, U24561, Y12599
A. thaliana	Complete genome from GenBank with five chromosomes
Beta vulgaris	AJ422053, X55297, X55297, X81974, X98767
Betula pendula	AJ279687, AJ279689, X77273, X87153, Y07779
Citrus sinensis	AF255014, AJ319762, AY029198, AY098894, X66377
Coffea arabica	AB048793, AF363630, AF494411, AJ293305, AJ420082
Cucumis sativus	AB046596, AF104392, AF442485, AJ413314, X67695
Glycine max	AB083032, AF488307, M16772, S45035, V00452
Gossypium hirsutum	AF403367, AF469099, AF512539, AF513859, AY072783
Hedera helix	AF130203, AF347635, AJ235489, AJ319075, X68334
H. vulgare	AF460219, AF492370, X57526, X78876, Z99940, and 100 EST sequences from ftp.genome.clemson.edu/pub/barley/est/HVSMEa.lib.gz
Lactuca sativa	AB031206, AF321538, AF489964, AJ310450, X60092
Malus × domestica	100 EST sequences from GenBank AU223481–AU223580
M. trunculata	100 EST sequences from GenBank BQ750336–BQ50435
Musa acuminata	AF377948, AF470320, AF479832, AY083168, X96947
Nicotiana tabacum	AF110226, AF440272, AF506374, AY090039, AY096801
Olea europaea	AF384050, AF429429, AF492010, AJ428575, Y12428
O. sativa	Genomic sequences from the Beijing Genomics Institute (mirrored at: ccgb.umn.edu/rice)
P. pinaster	100 EST sequences from GenBank AL751175–AL751285
Populus tremuloides	AF072131, AF185574, AF206812, AF209658, AF349443
Prunus persica	AF318173, AF319165, AF362989, AF362990, U49454
Raphanus sativus	AB000706, AY052582, U18556, X68651, X78452
S. tuberosum	100 EST sequences from GenBank BM408324–BM408423
S. bicolor	88 EST sequences from GenBank AA607079–AA738549
T. aestivum	100 EST sequences from GenBank BQ744049–BQ744148
Vitis vinifera	AF280768, AF369827, AF373604, AF378126, AF439321

databases separately (*Arabidopsis* and *Oryza*). The number of matches with e-values $<10^{-1}$, 10^{-5} or 10^{-20} were tabulated for each set of query sequences.

A collection of 100 ESTs each from seven fungal libraries (Table 1) and seven plant libraries (Table 2) were tested using Standalone BLAST with combined plant and fungal genomic databases. For fungi, the genomic databases were *Neurospora* (ascomycete) and *Arabidopsis* or *Phanerochaete* (basidiomycete) and *Arabidopsis*. For plants, the genomic databases

Table 3
Source of fungal-infected plant EST sequences used for query data sets or BLAST databases in this study

Species	GenBank accession no.
Brassica napus–L. maculans	280 sequences from EST library www.bspp.org.uk/mppol/1999/0301FRISTENSKY
H. vulgare–Blumeria graminis	500 sequences from EST library ftp.genome.clemson.edu/pub/barley/est/HV_CEa.lib.gz
Medicago truncatula–G. intraradices	500 EST sequences from GenBank AJ311220–AL389765
Medicago truncatula–Phoma medicaginis	500 EST sequences from GenBank BQ140771–BQ141319
O. sativa–M. grisea	615 EST sequences from GenBank AT003334–AT003878, AF889432–BF889453, CA752582–CA752734
Pinus sylvestris–Heterobasidion annosum	500 EST sequences from GenBank BI416471–BM340258
Pisum sativum–Glomus mosseae	88 sequences from EST library GenBank AJ308122–AJ41971, U90031–U90034
T. aestivum–Fusarium graminearum	500 EST sequences from GenBank BQ161927–BQ162107
T. aestivum–Puccinia triticina	500 EST sequences from GenBank BQ620261–BQ620760

were *Arabidopsis* (dicot) and *Neurospora*, or *Oryza* (monocot) and *Neurospora*.

Based on the results of the previously described analyses, combinations of a single plant and a single fungal genomic database were selected that most closely matched the plant–fungal combination in each EST library from infected plants. Up to 500 sequences from each EST library from infected plants were then examined with Standalone TBLASTX at e-value $\leq 10^{-1}$. To assess the accuracy of the Standalone BLAST of the combined genomic databases, the EST sequences were also matched against the GenBank NR protein database using Network BLASTX. A further examination was made of the sequences identified as plant, fungal, other organism or no match from the Standalone BLAST by comparing these matches with those obtained using the GenBank NR database.

Another assessment of the accuracy of the Standalone TBLASTX method involved 615 sequences from an EST library of *O. sativa* infected by *M. grisea*. These sequences were individually matched against three databases: the *Oryza–Magnaporthe* genomic database, the *Oryza–Neurospora* genomic database and the GenBank NR database. The results from the latter two databases were cross-tabulated against the results of the *Oryza–Magnaporthe* database.

2.3. Data extraction

Perl scripts were written to extract the following relevant data from BLAST output: query sequence name, query sequence length, top match, length of top match, e-value for top match, number of matches with e-value $\leq 10^{-1}$ and a list of the next five top matches. Separate scripts were written to parse output from Standalone BLAST or Network BLAST, and they are available from www.uoguelph.ca/~thsiang/est/ or upon request to the senior author. The Perl scripts generated a single line of output from each BLAST output file. For each data set, the single lines were combined into a single file and the file imported into a spreadsheet program. The use of comma's for delimiters as specified in the Perl script allowed for the spreadsheet file to be parsed within the spreadsheet program into columns of relevant data listed above. For Standalone BLAST, the columns were then sorted by top matching sequence, and the number of matches with the fungal or the plant database could be counted. For Network BLAST, the top match for each sequence was manually classified as plant, fungal, other or none, and then sorted and counted. For each query sequence, the taxonomic origin was assigned based on the top match with e-value $\leq 10^{-1}$ or the selected threshold e-value.

3. Results

3.1. TBLASTX, BLASTN and MegaBLAST analyses

A wide range of plant and fungal species were selected for analysis to represent the diversity of species that might be expected to be studied in fungal-infected plant tissue (Tables 1–3). Standalone TBLASTX with genomic databases was able to correctly identify the origin of fungal sequences as fungal with over 80% accuracy, while Standalone BLASTN and Standalone MegaBLAST with genomic databases averaged just over 50% (Table 4). The differences between the three methods were less pronounced with the plant sequences, since all three procedures could identify plant sequences with 84–94% accuracy. TBLASTX still gave the most matches, and MegaBLAST was slightly poorer than BLASTN for identifying plant sequences.

3.2. Choice of genomic databases

Databases composed of the genomes of the filamentous ascomycetes, *N. crassa* and *M. grisea*, performed very similarly in their matches with the 100 cDNA sequences from 20 fungal species at maximum

Table 4
Comparison of Standalone TBLASTX, BLASTN and MegaBLAST of described plant or fungal genes as query sequences using a matching database composed of the genomes of *A. thaliana* and *N. crassa* at e-value $\leq 10^{-1}$

Query sequences	Percentage of plant or fungal matches					
	TBLASTX		BLASTN		MegaBLAST	
	Plant	Fungus	Plant	Fungus	Plant	Fungus
100 fungal cDNAs	8	86	14	57	16	53
100 plant cDNAs	92	4	85	2	84	2

e-values ranging from 10^{-1} to 10^{-20} (Table 5). They also showed more matches than did the databases of the basidiomycete, *P. chrysosporium*, or the more primitive ascomycete, *S. cerevisiae*. At the highest e-value (10^{-1}), the *Phanerochaete* database gave a similar performance to those of the two filamentous ascomycetes, but at lower e-values, the number of matches was lower by more than 15%. Compared to other databases, the *Saccharomyces* database provided the fewest matches at 74% at the highest threshold e-value (10^{-1}). With 100 cDNA sequences from 20 plant species, the *Arabidopsis* and *Oryza* genomic databases performed similarly at all e-values with >96% matching at e-value = 10^{-1} and 78% at 10^{-20} (Table 6).

The ability of the fungal genomes to match EST sequences from various fungi depended on their taxonomic affinities (Table 7). To assess the competitive performance of the fungal genomes, the *Arabidopsis* genomic database was included with the fungal genomic databases. The ascomycetes, *B. fuckeliana*, *L. maculans* and *M. grisea* had more matches with the *Neurospora* database, while the basidiomycetes, *A. bisporus*, *P. ostreatus* and *S. commune*, showed more matches with the *Phanerochaete* database. *G. intraradices* is a zygomycete, and showed the lowest number of matches with any database. However, *G. intraradices* sequences had higher matches with *Neurospora* than they did with *Phanerochaete*. By comparison, analysis of these sequences using BLASTX of the Genbank NR database resulted in consistently fewer fungal matches for all of these fungi (Table 7), with an average of 47% compared to 67% for the Standalone BLAST.

Despite its much smaller size, the *Arabidopsis* database performed similarly to that of *Oryza* for matching the EST sequences of *M. domestica* and *M. trunculata*, and was better at matching sequences of the dicot, *S. tuberosum* (Table 8). The difference in the two plant databases was more obvious for matching sequences from monocots, with the *Oryza* database yielding much higher levels of matches (90–98%) for *H. vulgare*, *S. bicolor* and *T. aestivum* compared to the *Arabidopsis* database (77–86%). However, neither database showed more than 70% match with sequences of the conifer *P. pinaster*, and therefore the complete genome of a gymnosperm may be required for a proper test involving gymnosperms. To assess the competitive performance of the plant genomes, the *Neurospora* database was included with the plant genomic databases. Unlike the results of the fungal EST analyses, BLASTX of the plant ESTs with the Genbank NR database showed equal or higher plant matches for most of these plants (Table 8), with an average of 91% compared to 81% for the *Arabidopsis* database and 86% for the *Oryza* database. This likely reflects the relatively large number of plant sequences available in the Genbank NR database.

Table 5
Comparison of different fungal genomic databases for matching 100 cDNA query sequences from 20 fungal species using Standalone TBLASTX

BLAST database (file size)	Percentage of fungal matches at each e-value		
	10^{-1}	10^{-5}	10^{-20}
M. grisea (38.9 Mb)	93	88	76
N. crassa (39.0 Mb)	92	87	77
P. chrysosporium (30.6 Mb)	92	70	52
S. cerevisiae (12.3 Mb)	74	59	49

Table 6
Comparison of different plant genomic databases for matching 100 cDNA query sequences from 20 plant species using Standalone TBLASTX

BLAST database (file size)	Percentage of plant matches at each e-value		
	10^{-1}	10^{-5}	10^{-20}
A. thaliana (120.8 Mb)	97	93	78
O. sativa (371.2 Mb)	96	90	78

3.3. ESTs from fungal-infected plant tissues

To date, only ESTs from the *O. sativa*–*M. grisea* interaction are available where both genomes of the plant and fungus have been sequenced. One would assume that BLAST analysis with a database of the *O. sativa* plus *M. grisea* genomes should correctly identify all EST sequences from this interaction. A set of 615 EST sequences from *O. sativa* infected with *M. grisea* were matched against the *Oryza* plus *Magnaporthe* genomic database, the *Oryza* plus *Neurospora* genomic database, or the GenBank NR database. Among the 615 EST sequences, the *Oryza* plus *Magnaporthe* database identified 75.6% as plant, 21.0% as fungal and 3.4% as no match, whereas the

Table 7
Comparison of the Network BLASTX of the GenBank NR database with Standalone TBLASTX of genomic databases for matching fungal EST sequences at e-value $\leq 10^{-1}$

Query sequences	Percentage of plant and fungal matches						
	Neurospora and *Arabidopsis*		*Phanerochaete* and *Arabidopsis*		GenBank NR		
	Plant	Fungal	Plant	Fungal	Plant	Fungal	Other[a]
100 *Agaricus* ESTs	18	47	13	67	4	29	31
100 *Botryotinia* ESTs	5	93	21	79	4	72[b]	23
100 *Glomus* ESTs	28	38	36	25	5	23	31
100 *Leptosphaeria* ESTs	10	75	16	59	3	43	35
100 *Magnaporthe* ESTs	2	96	13	83	1	78	21
100 *Pleurotus* ESTs	10	64	2	78	2	43	35
100 *Schizophyllum* ESTs	10	57	3	81	4	40	36

[a] In the GenBank NR matches, "other" refers to bacteria, animal or other matching sequences.
[b] Many of these *Botrytis cinerea* EST sequences were deposited in the GenBank NR database and so the hits matching at 100% were ignored.

Oryza plus *Neurospora* database identified 79.3% as plant, 8.6% as fungal and 12.0% as no match using an e-value of $\leq 10^{-1}$. By comparison, the use of the GenBank NR database yielded much higher levels of non-plant and non-fungal matches with the identification of 63.9% as plant, 9.1% as fungal, 17.2% as no match, 7.3% as animal, 0.2% as bacterial and 1.6% as other, mostly vector sequences. A stricter e-value of $\leq 10^{-5}$ increased the number with no match to 6.0% with the *Oryza* plus *Magnaporthe* database, 19.3% with the *Oryza* plus *Neurospora* database and 29.6% with the GenBank NR database.

Assuming that the results with the *Oryza* plus *Magnaporthe* database correctly identified the origin of the query sequences, a cross-tabulation of the results with those obtained with the *Oryza* plus *Neurospora* database revealed that the latter misidentified 0.5% of the plant sequences and 12.8% of the fungal sequences at an e-value $\leq 10^{-1}$. Among the 12.8% misidentified fungal sequences, 4.2% were misidentified as plant and 8.6% were misidentified as non-matches at an e-value $\leq 10^{-1}$. By decreasing the threshold e-value to 10^{-5}, most of the fungal sequences that had been misidentified as plant were now shifted to the no match category with the *Oryza* plus *Neurospora* database. A similar comparison of the results using the *Oryza* plus *Magnaporthe* versus the GenBank NR databases showed a higher level of misidentification from the GenBank NR database with 13.9% of the plant sequences and 13.4% of the fungal sequences misidentified at an e-value $\leq 10^{-1}$. Overall, analyses of the *Oryza* plus *Neurospora* database and the *Oryza* plus *Magnaporthe* database showed 86.6% and 86.3% agreement (plant as plant plus

Table 8
Comparison of the Network BLASTX of the GenBank NR database with Standalone TBLASTX of genomic databases for matching plant EST sequences at e-value $\leq 10^{-1}$

Query sequences	Percentage of plant and fungal matches						
	Arabidopsis and *Neurospora*		*Oryza* and *Neurospora*		GenBank NR		
	Plant	Fungal	Plant	Fungal	Plant	Fungal	Other[a]
100 *Hordeum* ESTs	86	8	98	2	89	0	10
100 *Malus* ESTs	81	7	82	6	81	0	13
100 *Medicago* ESTs	88	7	87	6	96	0	0
100 *Pinus* ESTs	66	10	69	10	83	0	2
100 *Solanum* ESTs	91	3	84	7	96	0	1
100 *Sorghum* ESTs	77	8	90	2	95	0	0
100 *Triticum* ESTs	77	12	94	2	94	2	2

[a] In the GenBank NR matches, "other" refers to bacteria, animal or other matching sequences.

fungus as fungus plus none as none) with at e-values of 10^{-1} and 10^{-5}, respectively, whereas the GenBank NR database only showed 71.8% and 70.5% agreement (at e-values of 10^{-1} and 10^{-5}, respectively) with the *Oryza* plus *Magnaporthe* database.

Based on previous analyses, ESTs from different fungal-infected plants were matched against the GenBank NR database and selected genomic databases: *Arabidopsis* for dicots, *Oryza* for monocots, *Neurospora* for ascomycetes and *Phanerochaete* for basidiomycetes. A comparison was made between Standalone TBLASTX analysis using the selected combinations of complete fungal and plant genomes, and BLASTX analysis with the GenBank NR protein database (Table 9). In all cases (at e-value $\leq 10^{-1}$), Standalone TBLASTX analysis with the selected genomic databases was able to place a greater number of sequences as plant (average 78%) or fungal (average 5.6%) than Network BLASTX analysis with the GenBank NR database (average 70% for plant and 1.7% for fungi). For several EST libraries, there were zero matches with fungal sequences using Network BLASTX of the GenBank NR protein database, whereas the lowest level of fungal matches was 1.8% with Standalone TBLASTX analyses.

A further examination of the analyses of the fungal-infected plant EST libraries showed that Standalone TBLASTX was able to detect nearly all the fungal matches that were detected by GenBank NR analysis (Table 10). However, many of the fungal sequences identified as ''other'' or ''no hit'' by GenBank NR analysis could be identified as fungal by Standalone TBLASTX analysis of the genomic databases. As an illustration, Kruger et al. (2002) found that 2% of 3546 EST sequences of *T. aestivum–F. graminearum* were of fungal origin using BLASTX with the GenBank NR database, whereas our analysis of a randomly chosen collection of 500 ESTs from this same library identified 4.6% as fungal using Standalone TBLASTX with the *Oryza* plus *Neurospora* genomes as the databases. This difference was not due to a bias in our subsample from this EST collection, since Network BLASTX of the GenBank NR database with these same 500 ESTs identified 1.8% of the ESTs as fungal. Kruger et al. (2002) also found that 49% of the *T. aestivum–F. graminearum* ESTs could not be matched using BLASTX of the GenBank NR protein database at an e-value $\leq 10^{-5}$, whereas our analysis with genomic Standalone TBLASTX of 500 ESTs from the same database gave less than 30% unmatched at an e-value $\leq 10^{-5}$ and less than 22% unmatched at an e-value $\leq 10^{-1}$.

A final assessment was made between the length of time required for the analysis of the same EST sequences using Standalone TBLASTX vs. GenBank BLASTX. On average, Standalone TBLASTX required 1–2 min to complete an analysis of a sequence, while the Network BLASTX required an

Table 9
Distinguishing plant and fungal sequences in EST libraries from infected plants using Standalone TBLASTX of genomic databases or BLASTX of the GenBank NR database at e-value ≤ 0.1

Query sequences (number)	Percentage of plant or fungal matches			
	Standalone TBLASTX		GenBank NR BLASTX	
	Plant	Fungal	Plant	Fungal
Brassica napus–L. maculans (280)	90.0	1.8[a]	79.1	0
H. vulgare–Blumeria graminis (500)	95.2	3.6[b]	89.5	0.2
Medicago truncatula–G. intraradices (500)	67.6	7.2[a]	66.6	2.0
Medicago truncatula–Phoma medicaginis (500)	78.4	6.4[a]	77.2	1.8
O. sativa–M. grisea (615)	79.3	8.6[b]	63.9	9.1
Pinus sylvestris–Heterobasidion annosum (500)	51.2	6.2[c]	45.5	0.2
Pisum sativum–G. intraradices (88)	69.3	8.0[a]	64.8	0
T. aestivum–Fusarium graminearum (500)	83.4	4.6[b]	67.0	1.8
T. aestivum–Puccinia triticina (500)	87.8	3.8[c]	79.4	0

[a] For standalone genomic database, ''a'' refers to *Arabidopsis* and *Neurospora*.
[b] For standalone genomic database, ''b'' refers to *Oryza* and *Neurospora*.
[c] For standalone genomic database, ''c ''refers to *Oryza* and *Phanerochaete*.

Table 10
GenBank NR BLASTX of fungal hits from Standalone TBLASTX of genomic databases

Query sequences	Standalone BLAST fungal hits[a]	GenBank NR hits[b]			
		Fungal	Plant	Other	None
Brassica napus–L. maculans	5	0	4	0	1
H. vulgare–Blumeria graminis	18	0	14	3	1
Medicago truncatula–G. intraradices	36	10	13	6	7
Medicago truncatula–Phoma medicaginis	32	9	12	2	9
O. sativa–M. grisea	53	41	1	4	7
Pinus sylvestris–Heterobasidion annosum	33	10	7	11	5
Pisum sativum–G. intraradices	7	0	7	0	0
T. aestivum–Fusarium graminearum	23	9	3	3	8
T. aestivum–Puccinia triticina	19	0	16	0	3

[a] Total number of fungal hits for each EST library from Table 9.
[b] The Standalone TBLASTX fungal hits were matched against the GenBank NR database using BLASTCL3 (Network BLASTX) and the subdivision of matches is shown. "Other" refers to animal, bacterial, vector or other non-fungal or non-plant matches.

average of 15 min on a variety of computers with CPUs ranging from 300 to 1700 MHz.

4. Discussion

ESTs provide a means of obtaining large inventories of gene sequences and, consequently, there has been rapid growth in the number of sequences publically available. However, this raises the need to organize and categorize these data sets. For ESTs from plant–fungal interactions, one of the first steps in this process is to determine which organism is the source of the sequence.

Codon usage patterns could possibly be used to distinguish the taxonomic source of coding sequences in fungal–plant interaction EST libraries. Codon usage biases have been used to study amino acid usage patterns (Lin et al., 2002; McInerney, 1997; McInerney, 1998) and assess phylogenies (Nesti et al., 1995). However, bias in codon usage not only differs between species, but differs greatly between different genes within a species. For example, codon usage bias within a genome may be important as a regulator of gene expression. Codons with abundant tRNAs are more common in those genes requiring optimal translational efficiency for the production of abundant proteins (Kendrew, 1994). Codon usage can also vary within a gene. Yu et al. (2002) proposed that gradients in amino acid usage (genes richer in GC at the 5′ than the 3′ end) found in grasses may be one reason why so many rice genes (50%) do not a have good match with arabidopsis genes. Therefore, the identification of the origin of ESTs by codon usage could be affected by the need for optimal translational efficiency as well as the portion of the gene that was sequenced for the EST.

Another option to identify the origin of ESTs from fungal–plant interactions is to compare them to the Genbank NR database. For a collection of ESTs from the *T. aestivum–F. graminearum* interaction, Kruger et al. (2002) performed a BLASTX analysis of the Genbank NR database and examined if the closest match was with a plant or fungal gene.BLASTX analysis of the GenBank NR protein database showed that a large number (49%) of the non-redundant sequences (i.e., unigenes) in this fungal-infected plant EST collection (e-value $\leq 10^{-5}$) were classified as unknown because of lack of any significant match to the GenBank NR protein database, and therefore the origin of these sequences could not be determined. Further analysis by matching with other EST collections generated from the axenic cultures of the fungus grown under different conditions revealed a number of additional fungal genes were identified that had not been previously identified using the GenBank NR database (Kruger et al., 2002).

Although the total size of the Genbank NR database is very large, it is still very limited in the number of available gene sequences for plant pathogenic fungi. For example, Soanes et al. (2002) searched the NCBI protein database for several commonly studied plant pathogenic fungi and found only 56 entries for *Blumeria graminis*, 25 for *M. graminicola* and 196 for *M. grisea*. For 2676 ESTs from the powdery mildew fungus, *B. graminis*, grown separately from the host, Thomas et al. (2001) found that 50% did not have a significant match (BLASTX at e-

value $\leq 10^{-5}$) with the GenBank NR database and Keon et al. (2000) found 42.2% of 986 ESTs for *M. graminicola* grown in culture did not have a significant match (BLASTX at e-value $\leq 10^{-5}$) with the GenBank NR database. Nelson et al. (1997) also observed that 66.5% of 838 unique ESTs from *N. crassa* had no match in the GenBank NR database (BLASTX at e-value $\leq 10^{-5}$). In our analyses of 1400 EST sequences from plants and fungi, 91% of the plant sequences had matches in the GenBank NR protein database while only 47% of the fungal sequences had matches. Clearly, the current GenBank NR protein database has significant deficiencies in its collection of sequences from plant pathogenic fungi compared to that of plants, and many more fungal genes remain to be characterized that could be expected to be expressed in a fungal–plant interaction.

On the other hand, the GenBank database is increasing very rapidly with more microbial, animal and plant sequences continually being added. However, the increasingly larger size of the GenBank NR database can also prove to be detrimental for quick taxonomic identification of a query sequence in at least two ways. First, increasing the number of sequences in the matching database will allow matches to have a lower e-value (and thus appear to be a better match) since a larger matching database decreases the probability value of matching by chance. Second, if the exact match is not present, then the closest match, which may not even be in the same phylum, will have the lowest e-value. With highly conserved genes, the matches between homologues in different phyla will be high (i.e., have a low e-value), leading to an assumption that a true match has been found. A very low e-value is generally assumed to represent homologous gene function; however, phylogenetic identity would be a more daring assumption because genes can be highly conserved across many taxonomic levels.

An additional option to identify ESTs from fungal–plant interactions is to use complete genomic databases. To differentiate fungal from plant ESTs in the *B. napus–L. maculans* interaction, Fristensky et al. (1999) compared the ESTs to the genomic sequence of *S. cerevisiae*, assuming that fungal sequences should be most closely related to those of another ascomycete. Only 2 out of 280 ESTs were identified as fungal based on matches with the *S. cerevisiae* genome, which may reflect a low level of fungal mRNA as the authors proposed, but this could also be due to an inability to match some fungal genes. Our results suggested that 5 out of the same 280 ESTs were fungal based on Standalone TBLASTX using the *Arabidopsis* and *Neurospora* genomic databases. It appears that comparisons of the ESTs with the genomic sequence of *S. cerevisiae* was less effective in identifying genes from plant pathogenic fungi than comparisons to the genome of a filamentous fungus.

This paper describes the use of genomic databases from selected plant and filamentous fungi in Standalone BLAST analysis, and the testing of the method with a diverse collection of 700 fungal, 700 plant and 3983 infected plant ESTs. The ability of this approach to distinguish plant and fungal ESTs was as high as 98.8% for the ESTs of the *H. vulgare–B. graminis* interaction, but as low as 57.4% for the ESTs of the *P. sylvestris–H. annosum* interaction. The quality of the level of identification is heavily influenced by the relatedness of the organisms under study to the organisms used in the genomic database. Therefore, the absence of a sequenced genome of a gymnosperm may explain the relatively low success in identifying *P. sylvestris* ESTs. However, more plant genomes are being sequenced and it is expected that the genomic sequences of a number of plant pathogenic fungi will become available in the foreseeable future (Soanes et al., 2002). The similarity in the ability of the genomic databases of *Neurospora* and *Magnaporthe* to identify fungal ESTs implies that the fungal component of the genomic database does not need to be from a plant pathogen or in the same genus as the organism being tested.

When we tested the *O. sativa–M. grisea* interaction ESTs against the *Oryza* plus *Magnaporthe* genomic databases, nearly all of the sequences (97%) were placed as either plant or fungal at e-value $\leq 10^{-1}$. Intuitively, a matching database comprised of the same genomes as those in the query sequences should give very high matches with few sequences left unmatched. Among the remaining 21 sequences with no matches at e-value $\leq 10^{-1}$, most sequences were relatively poor (i.e., 10–25% "N"s or shorter than 100 bp). Other reasons that apparently valid sequences were not matched may be that the cultivar of *O. sativa* from

one of the interaction EST libraries was not the same as that in the matching database, or that the genomic databases were not fully complete. In contrast, BLASTX with the GenBank NR database of these same query sequences resulted in 23% of the sequences not matched with either a plant or a fungus.

In this study, Standalone BLAST with genomic databases was selected because it provides the ability to create a customized database for BLAST searches. Although GenBank BLAST also allows some selection of databases, such as NR for the non-redundant database or EST for the entire collection of EST sequences or through the use of Genomic BLAST, some complete or mostly complete genome sequences are not available on GenBank or other centralized databases. The combination of Standalone BLAST with customized local genomic databases avoids the issues of having to maintain local copies of the complete GenBank databases and the extensive computing power needed to process sequence data against very large databases. An additional advantage of Standalone BLAST accessing a locally available database is that the analyses are dependent on local processing power and not web access, which can be slow during periods of high use.

Compared to the two nucleotide matching methods, Standalone BLASTN and Standalone MegaBLAST, Standalone TBLASTX was able to identify fungal sequences with much greater accuracy. BLASTN and MegaBLAST both involve DNA sequence comparisons, but MegaBLAST uses an algorithm that is designed to swiftly compare two large sets of nucleotide sequences that differ only slightly from one another, perhaps as a result of sequencing errors, supposedly making it 10 times faster than other BLAST programs (www.ncbi.nlm.nih.gov/Web/Newsltr/FallWinter2000/standalone.html). TBLASTX translates a nucleotide query sequence into all six frames, and compares these translations to nucleotide sequences in the database that are also translated in all six frames. The greater success of TBLASTX in the fungal matches likely was due to the protein sequences being generally more highly conserved than the corresponding nucleotide sequences (Lewin, 1997). However, all three methods showed relatively high success in matching plant sequences, perhaps because flowering plants, which arose around 150 million years ago, are evolutionarily less divergent than fungi, which arose 500 million years ago (Patterson, 1999).

A procedure similar to the one described here using selected genomes as matching databases can be performed with Genomic BLAST (Cummings et al., 2002) or with advanced options at the NCBI website when using WWW BLAST. For example, after selecting BLASTX, there are options further down the form for advanced blasting. In the box titled "Limit by entrez query", a smaller database can be selected such as "(*A. thaliana* [organism] or *N. crassa* [organism]) and chromosome". This tests the query sequence against a matching database composed of the entire genomes of the organisms listed. However, the limitation here is that relatively few complete genomes are available on GenBank and several complete or almost complete genomes, such as rice or *M. grisea*, have not been placed on GenBank and hence are not available for searching via WWW BLAST.

In many studies involving database searches, a match at 10^{-20} (either e-value or the previously used *p*-value) was considered a strong match, while matches below 10^{-5} were often the criterion for homology (Keon et al., 2000; Kruger et al., 2002; Thomas et al., 2001, 2002). However, in this study, e-values less than 10^{-1} were chosen as the criterion for a valid match, because we were not searching for homology between sequences to make claims concerning the function of the gene, but rather were using the genomes to detect phylogenetic identity of the EST at a phylum level. Some researchers also consider e-values less than 10^{-1} to represent biological significance and have used the e-value as a measure of statistical significance (Pertsemlidis and Fondon, 2002). Pearson (1998) states that an e-value of 0.02 could be used for inferring homology with only a 2% chance of a false positive.

Although WWW TBLASTX or Network TBLASTX is very useful in identifying potential proteins encoded by single pass read ESTs according to NCBI, this type of search is computationally intensive and time consuming, and therefore is not recommended by NCBI for search with numerous sequences (www.ncbi.nlm.nih.gov/BLAST/Why.html). Standalone TBLASTX is sufficiently efficient so that relatively large numbers of ESTs can be analyzed and large numbers of ESTs will likely be necessary to find any particular disease-

related gene. Ohlrogge and Benning (2000) pointed out that if a relevant mRNA is abundantly expressed at 0.1% of the total mRNA, then 3000 sequences will be needed for a 95% chance of discovering the gene. Many genes involved in plant–microbe interactions, particularly fungal genes, may be expressed at comparatively low levels, and therefore, one would expect that considerable numbers of ESTs would be needed to find all the genes relevant to a plant–microbe interaction. In addition, ESTs are limited by the treatment conditions at the time at which the mRNA is extracted. For scientists interested in plant–microbe interactions, this may require examining ESTs at several different time points while the fungus is growing inside susceptible plant tissue or at different stages of a resistance reaction, further increasing the number of sequences needed to be analyzed.

The method described in this paper provides a novel approach to identifying plant and fungal sequences from mixtures. Standalone BLAST with a database comprised of a single plant and a single fungal genome can be used to analyze large numbers of ESTs with less computing power than conventional BLAST of NR databases, requires no internet access after initial setup, and is widely applicable to diverse groups of fungal and plant species. It also appears to distinguish fungal and plant sequences in mixed EST collections better than BLAST searches of the GenBank NR database, thus significantly improving our ability to identify the taxonomic origin of such sequences.

Acknowledgements

Funding for this study was provided by the Natural Science and Engineering Research Council of Canada.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*. Nature 408, 796–815.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., Botstein, D., 1998. SGD: *Saccharomyces* Genome Database. Nucleic Acids Res. 26, 73–80.

Cummings, L., Riley, L., Black, L., Souvorov, A., Resenchuk, S., Dondoshansky, I., Tatusova, T., 2002. Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. FEMS Microbiol. Lett. 216, 133–138.

Fristensky, B., Balcerzak, M., He, D., Zhang, P., 1999. Expressed sequence tags from the defense response of *Brassica napus* to *Leptosphaeria maculans*. Mol. Plant Pathol. (On-Line. Ref. No. 1412).

Kendrew, J., 1994. The Encyclopedia of Molecular Biology. Blackwell, Cambridge, MA.

Keon, J., Bailey, A., Hargreaves, J., 2000. A group of expressed cDNA sequences from the wheat fungal leaf blotch pathogen, *Mycosphaerella graminicola* (*Septoria tritici*). Fungal Genet. Biol. 29, 118–133.

Koski, L.B., Golding, G.B., 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540–542.

Kruger, W.M., Pritsch, C., Chao, S., Muehlbauer, G.J., 2002. Functional and comparative bioinformatic analysis of expressed genes from wheat spikes infected with *Fusarium graminearum*. Mol. Plant–Microb. Interact. 15, 445–455.

Lewin, B., 1997. Genes VI. Oxford Univ. Press, New York, NY.

Lin, K., Kuang, Y., Joseph, J.S., Kolatkar, P.R., 2002. Conserved codon composition of ribosomal protein coding genes in *Escherichia coli*, *Mycobacterium tuberculosis*, and *Saccharomyces cerevisiae*: lessons from supervised machine learning in functional genomics. Nucleic Acids Res. 30, 2599–2607.

McInerney, J.O., 1997. Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb. Comp. Genomics 2, 1–10.

McInerney, J.O., 1998. GCUA: general codon usage analysis. Bioinform. Appl. Note 14, 372–373.

Nelson, M.A., Kang, S., Braun, E.L., et al., 1997. Expressed sequences from conidial, mycelial and sexual stages of *Neurospora crassa*. Fungal Genet. Biol. 21, 348–363.

Nesti, C., Poli, G., Chicca, M., Ambrosino, P., Scapoli, C., Barrai, I., 1995. Phylogeny inferred from codon usage patterns in 31 organisms. CABIOS 11, 167–171.

Ohlrogge, J., Benning, C., 2000. Unraveling plant metabolism by EST analysis. Curr. Opin. Plant Biol. 3, 224–228.

Patterson, C., 1999. Evolution, 2nd ed. Cornell Univ. Press, Ithaca, NY.

Pearson, W.R., 1998. Empirical statistical estimates for sequence similarity searches. J. Mol. Evol. 276, 71–84.

Pertsemlidis, A., Fondon, J.W., 2002. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol. 2 (10), 1–10.

Quotob, D., Hraber, P.T., Sobral, B.W.S., Gijzen, M., 2000. Comparative analysis of expressed sequences in *Phytophthora sojae*. Plant Physiol. 123, 243–253.

Soanes, D.M., Skinner, W., Keon, J., Hargreaves, J., Talbot, N.J., 2002. Genomes of phytopathogenic fungi and the development of bioinformatic resources. Mol. Plant–Microb. Interact. 15, 421–427.

Thomas, S.W., Rasmussen, S.W., Glaring, M.A., Rouster, J.A., Christiansen, S.K., Oliver, R.P., 2001. Gene identification in

the obligate fungal pathogen *Blumeria graminis* by expressed sequence tag analysis. Fungal Genet. Biol. 33, 195–211.

Thomas, S.W., Glaring, M.A., Rasmussen, S.W., Kinane, J.T., Oliver, R.P., 2002. Transcript profiling in the barley mildew pathogen *Blumeria graminis* by serial analysis of gene expression (SAGE). Mol. Plant–Microb. Interact. 15, 847–856.

Wood, V., Gwilliam, R., Rajadream, M.-A., et al., 2002. The genome sequence of *Schizosaccharomyces pombe*. Nature 415, 871–880.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. spp. *indica*). Science 296, 79–92.