Computational techniques for elucidating plant–pathogen interactions from large-scale experiments on fungi and oomycetes

Carol Soderlund

Submitted: 4th August 2009; Received (in revised form): 29th September 2009

Abstract

Eukaryotic plant pathogens are responsible for the destruction of billions of dollars worth of crops each year. With large-scale genomics of both pathogens and hosts and the corresponding computational analysis, biologists are now able to gain knowledge about many pathogenic and defense genes concurrently. To study the interactions between these two organism groups, it is necessary to design experiments to elucidate the genes being expressed during the invasion of the pathogen into the host. For the most part, this does not require new software development, though it does require the use of existing software in novel ways. This article provides a broad overview of several key and illustrative experiments and the corresponding computational analyses, outlining the knowledge gained in each. It goes on to describe databases for plant–pathogen data and important initiatives such as Plant-Associated Microbe Gene Ontology. It discusses how various emerging approaches will increase the power of computers in host-pathogen interaction studies.

Keywords: fungi; oomycetes; plant–pathogen interactions

INTRODUCTION

There are over 1.5 million members of the fungal community [1] and approximately 100,000 known fungal species, where only a small subset of them are pathogenic [2]. Single plant species are only hosts to a limited number of fungi, and fungi have a limited host range. This degree of specialization results in an enormous diversity of fungal species with an evolutionary history of over 900 million years [3]. The relationship between host and microbe can be mutually beneficial, as in the mycorrhizal relationship between the fungi and plant roots, or the microbe can be pathogenic causing infection in the plant. The degree of virulence can vary from weakening the crop through to significant annual losses or major epidemic. Since a small subset of fungi infect all major crop plants, they represent a major threat to global food security [4].

Most eukaryotic plant pathogens are fungi or oomycetes [5], where the fungal kingdom is split into the two major divisions of ascomycota and basidiomycota. Oomycetes were originally classified as fungi based on physical resemblance, however, they are phylogenetically closest to diatoms and golden-brown algae. Though there are many differences between oomycetes and fungi, they employ similar strategies for attacking plants [5], and consequently, similar strategies are used to detect interactions with their hosts, as discussed in this article.

The study of plant–pathogen interactions is necessary to develop management strategies for the destructive pathogen-induced diseases which cost the United States alone $33 billion a year [6]. The impact of global climate change is predicted to adversely affect the plant defense chemistry, further increasing the necessity for understanding and
controlling plant diseases [7]. Host–pathogen interactions are often similar across the different kingdoms [8, 9], so knowledge of plant–pathogen interactions will also aid our understanding of human pathogens.

There are a number of good reviews on plant–pathogen interactions, for example: Soanes et al. [10] review utilizing genomic sequence to elucidate plant–fungal and plant–oomycete pathogenic interactions. Latijnhousers et al. [5] and Meng et al. [11] compare the infection strategies of fungi and oomycetes. Van der Does et al. [12] review virulence genes and evolution. Xu et al. [13] provide an in-depth review of fungal pathogens. These reviews are written for biologists with a strong background in phytopathology. However, since the computation of plant–pathogen interactions is interdisciplinary, this review is written for the scientist who is new to phytopathology but who has a good understanding of bioinformatics. As terminology is an important aspect to understanding the literature, Table 1 provides common terms from the plant–pathogen literature.

The bioinformatics approaches for determining gene function in host–pathogen interactions are similar to those for studying abiotic stress or other conditions. The major differences in large-scale studies are: (i) an emphasis on secreted proteins; (ii) analyzing the host and pathogen transcriptomes during infection, where resulting ESTs must be partitioned into the host and pathogen expressed

Table 1: Salient terms and concepts in the plant–pathogen interaction literature.a

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td></td>
</tr>
<tr>
<td>Mycology</td>
<td>The study of fungi</td>
</tr>
<tr>
<td>Phytopathology</td>
<td>The study of plant diseases caused by pathogens and environmental conditions, i.e. plant pathology</td>
</tr>
<tr>
<td>Symbiosis</td>
<td>Close and often long-term interactions between different biological species</td>
</tr>
<tr>
<td>Symbiotic categoriesb</td>
<td></td>
</tr>
<tr>
<td>Parasitic</td>
<td>Parasite gains at the host's expense</td>
</tr>
<tr>
<td>Mutualisms</td>
<td>Both organisms benefit</td>
</tr>
<tr>
<td>Commensalisms</td>
<td>One benefits and the other is not harmed</td>
</tr>
<tr>
<td>Classifications</td>
<td></td>
</tr>
<tr>
<td>Autotrophic</td>
<td>Synthesizes its own organic substances from inorganic compounds using sources such as light (e.g. plants)</td>
</tr>
<tr>
<td>Heterotrophic</td>
<td>Uses organic substrates to get chemical energy for its life cycle (e.g. fungi)</td>
</tr>
<tr>
<td>Biotrophic</td>
<td>Requires living cells for proliferation, invades living host cells</td>
</tr>
<tr>
<td>Necrotrophic</td>
<td>Requires dead cells for proliferation, kills host cells before invasion</td>
</tr>
<tr>
<td>Hemibiotrophic</td>
<td>Initially forms an association with the host as a biotroph, and in later stages of infection becomes necrotrophic, actively killing host cells</td>
</tr>
<tr>
<td>Osmotrophic</td>
<td>Secretes enzymes into the environment to degrade polymers and then transports resulting sugars, amino acids and fatty acids into the cell for use</td>
</tr>
<tr>
<td>Saprophytic</td>
<td>Heterotrophs that obtain nutrients by consuming decomposed organic matter</td>
</tr>
<tr>
<td>Pathogen characteristicsc</td>
<td>May respond to the chemical and physical characteristics of the plant host</td>
</tr>
<tr>
<td>Cell surface receptors</td>
<td>Can be secreted enzymes in necrotrophic fungi that degrade the host–plant cell to enter the plant tissue</td>
</tr>
<tr>
<td>Secreted proteins</td>
<td>Secreted proteins that suppress the plant defenses and alter the cellular metabolism for the needs of the invading pathogen</td>
</tr>
<tr>
<td>Effectors</td>
<td>Their pathways may elicit toxins that cause plant cell death</td>
</tr>
<tr>
<td>Secondary metabolites</td>
<td></td>
</tr>
<tr>
<td>Pathogen cell types</td>
<td></td>
</tr>
<tr>
<td>Hypha (pl. hyphae)</td>
<td>Long branching filamentous cell</td>
</tr>
<tr>
<td>Mycelium (pl. mycelia)</td>
<td>Mass of tube-like hyphae that is the vegetative body of a fungus or oomycete</td>
</tr>
<tr>
<td>Appressorium</td>
<td>Specialized pathogen cells for leaf surface penetration</td>
</tr>
<tr>
<td>Conidium (pl. conidia)</td>
<td>Asexual spores produced by ascomycete fungi</td>
</tr>
<tr>
<td>Zoosporangia</td>
<td>Swimming asexual spores produced by oomycetes</td>
</tr>
<tr>
<td>Zoospores</td>
<td>Zoospores form walled cysts that germinate and infect plants</td>
</tr>
<tr>
<td>Plant cultivar categories</td>
<td></td>
</tr>
<tr>
<td>Susceptible</td>
<td>Compatible interaction with pathogen—usually results in disease</td>
</tr>
<tr>
<td>Resistant</td>
<td>Incompatible interaction with pathogen—no disease</td>
</tr>
<tr>
<td>Resistance</td>
<td></td>
</tr>
<tr>
<td>Virulence</td>
<td>Quantitative measure of pathogenicity</td>
</tr>
<tr>
<td>Resistance (R) genes</td>
<td>Host genes that encode receptor proteins that recognize and combat pathogens</td>
</tr>
<tr>
<td>Avirulence (avr) genes</td>
<td>May encode pathogen gene products recognized by R gene products, resulting in resistance</td>
</tr>
<tr>
<td>Hypersensitive response</td>
<td>Resistance response involving cell death (necrosis) of plant cells at the site of invasion</td>
</tr>
</tbody>
</table>

aTerms and concepts can have different connotations in different fungal communities [65].
bThese terms are not discrete, but are a continuum.
cOf genes found in many phytopathogenic fungi and oomycetes [10].
genes; and (iii) higher mutation rates in host–pathogen interacting genes. Since the basic approaches (i.e. EST assembly, gene identification, etc.) have been covered elsewhere, this article will concentrate on large-scale experiments and the corresponding computational analysis specific to finding genes involved in plant–pathogen interactions.

**INTERACTION EXPERIMENTS, ANALYSIS AND RESOURCES**

The following sections introduce a selected set of studies that require computational analysis, and are representative of those found in plant–pathogen interaction experiments. Results are given only as a means of providing an indication of outcomes that may be expected from such computation. For information on general analysis, the series on Fungal Genomics contains reviews on gene identification [14], comparative genomics [15], secreted proteins [16] and EST assembly [17].

The studies are from the following pathogens: the hemibiotrophic fungal pathogen *Magnaporthe grisea* (*oryzae*) that causes rice blast disease in rice; the fungal pathogen *Fusarium graminearum* (*Gibberalla zeae*) that causes head blight in wheat and barley; the biotrophic fungal pathogen *Ustilago maydis* that causes corn smut in maize; the two hemibiotrophic oomycete pathogens *Phytophthora sojae* and *Phytophthora ramorum* that cause root rot in soybean and sudden oak death in oak trees, respectively.

**GENOME ANALYSIS**

As of early 2000, only two fungal genomes had been sequenced. The Fungal Genome Initiative (FGI) [3] was formed to sequence fungal genomes, which were selected to: (i) represent important human pathogens, plant pathogens, saprophytes and model organisms; (ii) represent a range of evolutionary distances; and (iii) differ in terms of specific physiological traits such as pathogenicity. As of March 2009, the GOLD database [18] of draft and complete genome sequences contained 302 fungi of which 32 were plant pathogens, and 112 protists of which eight were plant pathogens. It has been possible to sequence this number of genomes as they are relatively small, e.g. *M. grisea* has one of the larger ascomyceta genomes of ∼40 Mb, whereas human is ∼3000 Mb and maize is ∼2300 Mb. These genomes are generally quite broadly diverged, e.g. *M. grisea* and *Neurospora crassa* are pathogenic and non-pathogenic ascomycetes respectively, which are considered closely related in fungi terms, yet are thought to share a common ancestor around 200 MYA [3]; in comparison, maize and sorghum are considered closely related plants, which diverged around 11.9 MYA [19].

As with any newly sequenced genome, one or more programs for gene identification are executed to determine candidate genes. As Grigoriev et al. [14] discuss, most gene prediction programs have been trained on human and do not automatically work for fungi, e.g. the coding region of the human genome is just a small fraction of the DNA and introns can be quite long (>10 kb); in comparison, fungi have coding densities ranging from 37% to 61% of the genome and intron sizes from 80 to 150 bp [3]. Consequently, gene prediction programs must be trained using known genes from the given genome or a closely related genome, where the alignment of ESTs and full-length cDNAs greatly aids the correct determination of genes and their gene structure. For example, for the genome sequence of *P. sojae*, the majority of the gene models were predicted using the FGENESH gene-finding software [20] trained with related EST sequences [21]. There is no essential computational difference in gene finding for pathogen versus non-pathogen genomes. However, if the ESTs used for training the gene-finding software are from infection libraries (see Transcriptome section), more pathogenic genes with a correct structure are likely to be found.

Initial functional annotation is usually computed by comparing (e.g. using BLASTx or BLASTp) the sequences of gene models obtained through a gene-finding analysis against sequences in a protein database, i.e. GenBank [22] nr (non-redundant) database, UniProt [23] (fungal only, plants only or all proteins), or KOG (eukaryotic orthologous groups) [24]. The nr and UniProt databases contain all known fungal and oomycetes genes; consequently, the results must be mapped into categories that distinguish pathogenic genes (see Plant-Associated Microbe Gene Ontology section). PHI-base [25, 26] has expert-curated molecular and biological information on genes proven to affect the outcome of host–pathogen interactions; comparing the gene models with the PHI-base sequences would flag putative pathogenic genes. As discussed below, additional annotation may be computed by comparing...
gene models with those of other organisms and through characterizing genes as potentially secreted.

COMPARATIVE GENOMICS
Gene models of two fungal or oomycete genomes may be compared to confirm gene status, transfer annotation and determine what genes are unique to each organism. For host–pathogen studies, genes that are not found in abundance in the non-pathogens but are found in pathogens are natural candidates for pathogenicity, which is illustrated in the following study. Dean et al. [27] compared (with BLASTp) the translated gene models of pathogen M. grisea with those of the non-pathogenic Emericella (Aspergillus) nidulans and N. cassa genomes. A G-protein–coupled receptor was previously found in M. grisea that is required for appressorium formation and pathogenesis; M. grisea has a large number of these receptors, whereas only one was detected in N. cassa and two in A. nidulans, which provides additional evidence that this gene is necessary for pathogenesis.

Cuomo et al. [28] have shown that evidence can be found for pathogenic genes by evaluating genomic patterns of SNPs (single nucleotide polymorphisms). They compared two strains of F. graminearum and found that genes expressed during plant infection were overrepresented in regions of high SNP density, supporting the hypothesis that pathogenic genes evolve more rapidly.

SECRETED PROTEINS
Secreted proteins are critical for nutrient acquisition for most fungi, as they secrete extracellular enzymes to break down potential food sources and then transport the resulting products into their cells. They have many roles in interactions with the plants; for example, some secreted proteins induce plant cell defenses and others block plant defenses to aid infection. Due to the importance of secreted proteins to the plant–pathogen interactions, and the fact that candidate secreted proteins can be computationally identified, these key proteins generally form part of most analytical strategies involving pathogen sequenced genomes and ESTs, as illustrated in the following studies.

Tyler et al. [21] analyzed the oomycetes P. sojae and P. ramorum, which have different host ranges, so it is expected that their pathogenic genes will have evolved rapidly in order to adapt to their specific host. It was found that their predicted secretome had evolved significantly more rapidly than the overall proteome; 17 and 11% of the secreted P. sojae and P. ramorum proteins are unique whereas only 9 and 4% are unique for the overall proteome, respectively.

Kamper et al. [29] found that the U. maydis genome had 426 predicted secreted proteins, of which 193 were specific to U. maydis and did not have a predicted function. Of these, 18% were found in clusters, and DNA array analysis showed up-regulation of many cluster genes in infected tissue.

Most studies in plant–pathogen interactions, including the above two, use the computational tool SignalP [30] to detect secreted proteins. SignalP has two different predictors, one uses a neural network and the other uses a hidden Markov model algorithm, where both have been trained on known secreted proteins. The WoLF PSORT software [31] is also occasionally used, as it classifies proteins into 10 locations including extracellular. It predicts localizations by converting the amino acid sequences of the secreted proteins into numerical vectors, which are then processed by a weighted k-nearest neighbor classifier.

TRANSCRIPTOME
As with the genomic study of any organism, it is typical to analyze ESTs and microarrays from various stages, conditions and tissues. When determining which genes are expressed in both the host and the pathogen during different time points of infection, a sample is extracted from the infected host that contains the RNA from both, resulting in a mixed library. A major problem with this approach is that during the early stages of host–pathogen interaction most cells are not infected, and so the extracted sample is saturated with RNA from normal plant cells. For example, Jantasuriyarat et al. [32] generated ESTs from rice infected with M. grisea and compared them using BLASTn with existing M. grisea ESTs; only four of 52 024 ESTs aligned, which was probably due to the early harvesting time (6 and 24 h after inoculation). To overcome this problem, Mosquera et al. [33] devised a reproducible method for extracting RNA from infected rice tissue that contained 20% M. grisea RNA at an early time point when the fungus was still growing in the first-invaded
plant cells; they selectively isolated infected plant cells and their near neighbors by dissection of plant tissue to eliminate much of the uninfected tissue as well as the fungal cells not involved in infection.

Once mixed EST libraries have been obtained, the host–pathogen ESTs must be partitioned into their host and pathogen constituents. As illustrated in the Jantasuriyarat et al. [32] study, the ESTs may be partitioned by comparing them to previously identified genes or related ESTs. A problem with this approach is that not all ESTs may be represented in previously sequenced sets [34]. Therefore, various software packages have been developed to partition the ESTs based on their nucleotide composition. Huitema et al. [35] used the fact that the GC content can sometimes be quite different between the host and pathogen. Maor et al. [36] developed the PF-IND method that uses codon usage bias. The problem with using codon usage is the need to find the in-frame coding sequence, and the fact that ESTs are also composed of untranslated regions. Therefore, Emmersen et al. [37] used triplet frequencies which are computed over both the coding and non-coding regions. The triplet frequencies are calculated using a sliding window over a training set of plant and pathogen ESTs, and a SVM (support vector machine) is used for the classifications.

Another approach to obtaining the sequence of pathogenic genes is to create libraries from conditions that are thought to mimic expression during infection, which avoids the problem of extracting RNA from cells during actual plant–pathogen interactions. For example, Trail et al. [38] generated "F. graminearum" ESTs from a carbon- and nitrogen-starved mycelian library, which had been shown to mimic infection in other fungal pathogens. The ESTs were searched against known virulent and pathogenic genes resulting in 16 matches (six of which were human pathogens), whereas ESTs from a library that was not expected to mimic infection only had one match.

When creating libraries to study interactions, it is common to have libraries from different stages of infection and to compare them to find which ESTs are unique at each stage. For example, Jantasuriyarat et al. [32] compared infected leaf tissue ESTs to uninoculated control ESTs, showing a large increase in the functional categories of defense and signal transduction genes in the infected sample.

ESTs are generally assembled into contigs of overlapping ESTs in order to determine the unique sets of expressed genes. The EST contigs result in a set of consensus sequences representing the putative transcripts, which are annotated using the same basic approach as for gene models. In addition, quite a few EST studies have determined categories of putative pathogenicity proteins based on their functional annotation [39, 40].

**GENOME-WIDE FUNCTIONAL ANALYSIS**

The above approaches all find ‘evidence’ of pathogenic genes. If the function has already been confirmed by an orthologous gene, then it does not necessarily need to be confirmed experimentally. However, there is still the issue of determining the function of genes with unknown function. With the relatively small genomes of fungi, it is possible to make collections of mutants that cover the majority of the genes; these mutants may then be used to test the effect of the mutated genes on pathogenicity. Computational tools are important for tracking, storing and displaying the data. Two recent studies developed large datasets of assayed mutants for "M. grisea". Jeon et al. [41] created 21 070 mutants, assayed them for various phenotypes, and recorded the results in the publicly available database ATMT (Agrobacterium tumefaciens-mediated transformation). Betts et al. [42] created over 55 000 mutants, entered the barcoded plates into a tracking database along with the assay results [43], and the final results were made queryable from the publicly available database MGOS (Magnaporthe grisea Oryza Sativa) [44].

**PLANT-ASSOCIATED MICROBE GENE ONTOLOGY**

An emerging important resource for functional studies of all genomes is the gene ontology (GO) [45], which is a standardization of terms for the functional characterization of genes and gene products across species. The Plant-Associated Microbe Gene Ontology (PAMGO) consortium collaborated with the GO Consortium to define GO terms to describe the biological processes common to plant and animal microbes [8]. As implied by the consortium name, they initially planned to develop terms for plant–pathogen interactions; however, since microbes of every domain face the same challenges, whether plant or animal pathogen, from mutualistic to
parasitic, the commonalities could be captured. As stated by Torto–Alalibo et al. [8], ‘All must initially attach to the host and breach a barrier or enter through openings to gain access to a nutritional source; all must suppress, evade, or tolerate host defenses for successful invasion.’ For example, a host–targeting signal was identified in the human malaria parasite *Plasmodium falciparum*, and a functionally equivalent signal was found to be conserved in the Irish potato famine pathogen *Phytophthora infestans* [46]. The following summarizes five of the PAMGO publications to illustrate how the terms are defined and used.

Over 700 GO terms have been added by the PAMGO consortium [8]. Most of the terms are under the parent ‘GO:0051704 multi-organism process’ and its children ‘GO:0044419 interspecies interaction between organisms’ and ‘GO:0051707 response to other organisms’. When both the host and pathogen are involved in a process, there are parallel terms, for example ‘GO:0044405 recognition of the host’ and ‘GO:0051855 recognition of the symbiont’. The specifications of the GO taxon field were modified to include the taxa of both the plant and pathogen involved in the interaction. For example, if an effector protein secreted by a pathogen triggers the hypersensitive response in the host, the effector gene would be assigned the GO term ‘GO:0034055 positive regulation by symbiont of host defense-related programmed cell death’ with the taxon ids of the pathogen and host.

Meng et al. [11] describe common processes in pathogenesis by both fungal and oomycete pathogens, and the corresponding GO terms. They cover spore dispersal, adhesion to the host, signal transduction during recognition of the host, penetration of the host, invasive growth and lesion development within the host.

When a GO term is assigned to a gene, it is given an ‘evidence code’, where there is one set of codes based on experimental evidence and another set based on computational evidence (see [47] for the explanations of the codes). When predicted genes are aligned to a protein database, e.g. InterPro [48] or UniProt [23], the results often contain associated GO terms which are then inherited and given the evidence code ‘IEA: Inferred from Electronic Annotation’. As discussed earlier, using sequence similarity is not as successful for finding fungal and oomycete virulence genes compared to most other gene types, as the high-mutation rate makes orthologs difficult to recognize [12].

In their annotation of the rice blast fungus, Meng et al. [49] provided automated GO assignments as follows. The GO database was downloaded, which had 50 GO-annotated fungi of which three were pathogens. The *M. grisea* predicted genes were then compared to the GO proteins to find reciprocal best hits using BLASTp [50] with e-value <10^-3. The GO proteins involved in significant hits were manually validated using published wet laboratory experiments and the NCBI Conserved Domain Database [51]. The *M. grisea* genes satisfying these requirements were assigned the corresponding GO term with an evidence code of ‘ISS: Inferred from Sequence Similarity’.

**DATABASES FOR FUNGAL AND OOMYCETES PLANT PATHOGENS**

As information gained for one organism can provide knowledge for another, an important resource for any plant–pathogen study are the existing web-based databases for plant fungi and oomycetes. Unfortunately, they are scattered around the world, have vastly different user interfaces, become out-of-date (e.g. annotated with an old version of UniProt), and new ones frequently emerge while others become obsolete. For example, COGEME [52, 53] was a major EST resource for pathogens, but the project has recently been terminated and the database will no longer be maintained. Many articles and reviews list the databases, but the information does not stay relevant due to the temporary nature of many web-based databases. There are multiple initiatives to create central portals, e.g. the site www.pathguide.org contains a listing of many of the online biological pathway resources [54], but there is no central resource for fungal databases (the ‘Links’ page at www.mgosdb.org has almost all current fungi-oomycetes databases, but may become out-of-date due to funding). Emerging approaches for sharing data between web-based databases using cyberinfrastructure network protocols and computational services [55] are now being investigated, though it will require support from the community, funding and initiative.

Many fungi-oomycetes databases from large laboratories are likely to stay active, where each represents multiple plant pathogens and one or more of the following features: comparative
genomes, gene search, genome browsers, secreted proteins and ESTs. The key databases are as follows: FGI (Fungal Genome Initiative) [3] at the Broad Institute, IMG (Integrated Microbial Genomes) [56] at JGI (Joint Genomic Institute), PEDANT (Protein Extraction, Description and Analysis Tool) [57] at MIPS (Munich Information Center for Protein Sequences), e-Fungi [58] at the University of Exeter and VMD (VBI Microbial Database) [59] at VBI (Virginia Bioinformatics Institute). There is overlap in data and features between these databases.

A few community databases are available for specific fungi: MGOS (M. grisea Oryza Sativa) [44], MIPS FGDB (F. graminearum DB) [60], and MIPS MUMDB (U. maydis DB) [29]. Of these, MGOS is the most complete database as it has ESTs, gene search, genome browser, secreted proteins, community annotation and mutants.

**DISCUSSION**

This article discusses combined computational and biological approaches for elucidating pathogenic genes from genomic and EST sequencing. In summary, the approaches use standard large-scale analysis programs, i.e. for genome sequence, genes are predicted and then annotated; for EST sequences, the sequences are assembled into contigs and then annotated. To adapt these approaches to plant–pathogen studies, potential pathogenic genes and transcripts are identified by annotation that includes sequence comparison to expertly curated genes in PHI-base and to genes in other databases annotated with PAMGO terms. Additionally, genes and transcripts are annotated as to whether they are predicted secreted proteins. The one deviation from other EST studies is that when libraries are created during infection, the plant–pathogen sequences have to be partitioned into plant and pathogen sources. Mutant studies are also important in plant–pathogen interaction studies, and the results of assays can be associated with gene predictions for additional annotation. Comparative analysis with a pathogen and related non–pathogen fungi can elucidate the unique genes in the pathogen that may be pathogenic genes. However, multiple studies have shown through comparative analysis of related pathogens that pathogenic genes have a higher mutation rate than other categories of genes, which in turn makes it harder to use comparative analysis to find pathogenic genes compared to more typical genome analysis.

As useful as these resources may be, genomic and EST sequences can only give evidence of pathogenicity, e.g. if a gene model is in a pathogen genome and not in a related non-pathogen genome; if an EST is in an infected library and not in a non-infected library; if a gene model or EST is found in PHI-base; if a gene model or EST is predicted to be secreted; if a gene model or EST has a functional annotation that is associated with a PAMGO term. As more evidence is accumulated, the greater the probability that a gene is truly pathogenic.

In the studies referenced in this article, Sanger sequencing was used where a single genome took considerable time and money to sequence, and the EST studies were generally limited to between 5000 and 25 000 ESTs. If one could sequence many related genomes, and if one could sequence many ESTs from different time points of infections and from different strains with variations in pathogenicity, the more quickly such evidence will accumulate. This is becoming a reality with NGS (next generation sequencing) machines, such as Roche (454 Titanium), Illumina (Genome Analyzer II) and ABI (AB Solid). The new 454 GS FLX Titanium is capable of generating over a million reads of 400 bases with reduced error [61], and the other two platforms are making similar gains. This increase in the number of sequences will provide both sequence and abundance for transcript studies (see Simon et al. [62] for a review of NGS for transcriptional analysis). Software and statistical measures have been developed to compare the abundance of ESTs from different stages, tissues and conditions; for example, the PAVE system [63] has a web interface that allows users to view only the contigs that have ESTs from a select set of libraries that are not in another selected set, and the R-statistic [64] identifies differentially expressed transcripts from the assembly of multiple EST libraries. These approaches will be much more meaningful with the large number of transcripts from a single experiment.

Future development of the GO and PAMGO projects is imperative to keep pace with the anticipated increase in information. When the same concept is described using different terms or phrases across the different organism-specific communities, it is very difficult to automate the transfer of
information from one organism to another. The GO/PAMGO initiatives are making this automatic transfer possible. However, it is essential for every microbiologist and computational scientist in the field to support this effort [65]. All scientists in this field should read the PAMGO articles and the recent special issue on ‘Gene ontology for the microbiologists’ in the journal Trends in Microbiology to fully understand how to use and contribute to the GO/PAMGO initiative. The shared terminology will have an enormous effect, as all the 100’s of little communities will start merging into one synergistic community studying life.

The last salient component is improved web resources; though the Internet has made searching for information faster, there is now more information to search. The web databases can be greatly improved, though as with the GO, it is going to take the community and funding agency support. The biologists must participate by adding data to the web databases, the computational scientists must work to make data flow seamlessly between web databases, and the funding agency must start providing more funds for these efforts, as the likely gains in crop performance will vastly outweigh the costs of the additional research.

The NGS + GO + Web will greatly accelerate all genomic research, including plant–pathogen interaction studies.

Key Points
- Almost all software-based analysis necessary to study plant–pathogen interactions can use existing software developed to study more general problems.
- The most significant differences in computation for host–pathogen studies are: (i) partitioning mixed EST libraries; (ii) an emphasis on secreted proteins; and (iii) the fast mutation rates of host–pathogen interacting genes, which make orthologs harder to detect.
- The PAMGO ontology will make it easier to share knowledge about interactions across species.
- Though plant–pathogen genome sequences are being generated at an increasing rate, databases and resources are not being built at a comparable rate, diminishing the utility of this data. Moreover, the resources that exist are scattered, forcing the scientist to search for needed data.

Acknowledgements
Many useful suggestions and corrections were made by Barbara Valent, Kevin Greer, Anupreet Kour, Anne Descour, Matthew Bomhoff, William Nelson and the reviewers of the article.

FUNDING
National Science Foundation under (Grant No. 0627159, partial).

References


