Nucleic Acids Research, 2006, Vol. 34, Database issue ***D459–D464***
doi:10.1093/nar/gkj047

PHI-base: a new database for pathogen host interactions

Rainer Winnenburg, Thomas K. Baldwin[1], Martin Urban[1], Chris Rawlings, Jacob Köhler and Kim E. Hammond-Kosack[1,*]

Biomathematics and Bioinformatics Division, Rothamsted Research, Harpenden, AL5 2JQ, UK and [1]Wheat Pathogenesis Programme, Plant-Pathogen Interactions Division, Rothamsted Research, Harpenden, AL5 2JQ, UK

Received August 15, 2005; Revised and Accepted October 4, 2005

ABSTRACT

To utilize effectively the growing number of verified genes that mediate an organism's ability to cause disease and/or to trigger host responses, we have developed PHI-base. This is a web-accessible database that currently catalogs 405 experimentally verified pathogenicity, virulence and effector genes from 54 fungal and Oomycete pathogens, of which 176 are from animal pathogens, 227 from plant pathogens and 3 from pathogens with a fungal host. PHI-base is the first on-line resource devoted to the identification and presentation of information on fungal and Oomycete pathogenicity genes and their host interactions. As such, PHI-base is a valuable resource for the discovery of candidate targets in medically and agronomically important fungal and Oomycete pathogens for intervention with synthetic chemistries and natural products. Each entry in PHI-base is curated by domain experts and supported by strong experimental evidence (gene/transcript disruption experiments) as well as literature references in which the experiments are described. Each gene in PHI-base is presented with its nucleotide and deduced amino acid sequence as well as a detailed description of the predicted protein's function during the host infection process. To facilitate data interoperability, we have annotated genes using controlled vocabularies (Gene Ontology terms, Enzyme Commission Numbers and so on), and provide links to other external data sources (e.g. NCBI taxonomy and EMBL). We welcome new data for inclusion in PHI-base, which is freely accessed at www4.rothamsted.bbsrc.ac.uk/phibase/.

INTRODUCTION

Pathogenic microbes cause disease in various host organisms including humans, animals and plants. In agriculture, ~10 000 fungal species are considered plant pathogenic (1), many causing severe disease epidemics and lower economic yields. In human health, there is growing concern over fungal infections in immuno-compromised patients. The number of genes confirmed by gene and/or transcript disruption experiments to be required for the disease causing ability of a microbe has gradually increased over the past 15 years. These genes are termed pathogenicity genes if the effect on the phenotype is qualitative, or virulence/aggressiveness genes if the effect is quantitative (2). However it is still difficult to access and compare the resulting data because this is mainly available in the literature or in the laboratories of individual investigators. The establishment of a web-accessible database to collate, cross-link and categorize genotypic and phenotypic information of individual pathogens and gene deletion/gene-silenced mutants will greatly facilitate our understanding of general pathogenicity mechanisms. For example, the identification of orthologs genes in multiple species could potentially reveal new generic targets for drug design and may result in the identification of novel strategies for disease control. However, a reliable source of collated data is an absolute prerequisite to achieve this objective. In addition, by using an accurate source of collated data the annotation of newly sequenced genomes can be achieved by using sequence similarity searches and homologs of verified pathogenicity genes detected. Typical users of this database will be medical and agricultural scientists, bioinformaticians and evolutionary biologists who need easy access to peer-reviewed data on multiple pathogen species from a single internet resource.

PHI-base is designed for hosting any type of pathogen host interaction and will not be restricted to certain pathogen species or a range of hosts. Currently several databases are

*To whom correspondence should be addressed. Tel: +44 1582 763133; Fax: +44 1582 760089; Email: kim.hammond-kosack@bbsrc.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

available which concentrate only on a specific host group like the Fish Pathogen Database (http://dbsdb.nus.edu.sg/fpdb/) or a sub-set of pathogen taxa such as Fungal Plant Pathogen Database (3). For instance the Phytopathogenic Fungi and Oomycete EST Database of COGEME (4) provides sequences of expressed sequence tags (ESTs) and unisequences (cluster assembled ESTs) from 15 plant pathogenic species. Gene sequences can be retrieved through text queries restricted to certain species based on pre-defined functional classification groups or on sequence similarity. PathoPlant (5) describes interactions of fungi, bacteria and viruses with plants at the whole organism level and combines this with information on the molecular basis of plant defence mechanisms. Links to external databases and a model signal transduction pathway are provided, but in contrast to PHI-base PathoPlant covers only phytopathogens. The fuGIMS database is being developed to integrate functional and sequence information from several plant and animal pathogenic fungi with similar information from *Saccharomyces cerevisiae* available from the GIMS database (6). Based on the study of microarray expression data, DRASTIC Insight (7) collates signal transduction information between plants, pathogens and the environment, including both biotic and abiotic influences on plant disease resistance at the molecular level. Ecological Database of the World's Insect Pathogens (EDWIP) (8) and Viral Diseases of Insects in the Literature (VIDIL) (8) offer literature-based information on fungi, viruses, protozoa, mollicutes, nematodes and bacteria which are infectious in insects, mites and related arthropods. Although a wealth of useful information can be found in EDWIP and VIDIL, these sites do not contain information on pathogenicity genes. Data management and analysis for key pathosystems are currently being developed within the generic framework PathPort (9), which envisions providing methods to validate candidate target sequences and predict host response models by exploiting and integrating data from several pathogen data sources through GRID technology. What makes PHI-base unique, is its focus on genes with functions that have been experimentally verified. These genes are compiled and curated in a way that can be used to bridge the genotype–phenotype gap underlying the interactions between hosts and pathogens.

In the following, we will describe PHI-base and the plans for future development of this resource.

IDENTIFICATION OF PATHOGENICITY GENES EXPERIMENTAL EVIDENCE

The only genes included in PHI-base are those which have been published in peer-reviewed journals as affecting pathogenicity in one or more gene disruption or gene silencing experiments. This approach is essential for accurately verifying a role in pathogenicity, and thus determines the design of the database and the curation process which are described in the next two sections of this publication. Several approaches have been taken by the scientific community to identify pathogenicity genes in fungal and Oomycete pathogens (Figure 1). Most of the genes included in PHI-base have arisen through forward or reverse genetics.

A forward genetics approach is often used to reveal novel pathogenicity/virulence genes. A library of mutants is created

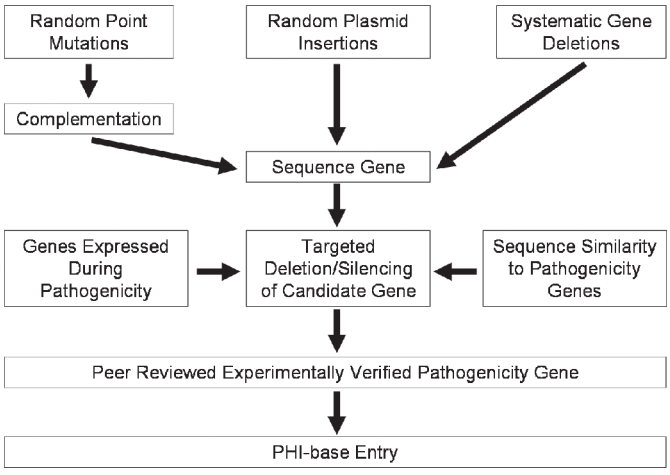

Figure 1. Representation of the experiments and criteria needed for a gene to be entered into PHI-base.

either by point mutation or insertional mutagenesis. Screening of large collections of mutants on host organisms can then reveal mutant strains that possess pathogenicity defects. The disrupted gene can then be identified, by complementation in the case of point mutations or by sequencing the genomic DNA flanking the mutagenic plasmid in the case of insertional mutations. This cloned and sequenced gene can only be classed as a putative pathogenicity gene. If a targeted deletion experiment in the wild-type strain results in the same pathogenicity defect, the gene is classified as a verified pathogenicity gene.

A reverse genetics approach has traditionally been used to identify pathogenicity/virulence gene homologs in multiple species. Orthologs of the yeast MAP kinases have been identified in many different fungal species. Sequence similarity also showed the yeast FUS3/KSS1 gene to be a pathogenicity factor in the rice blast fungal pathogen *Magnaporthe grisea* (PHI-base accession PHI:56). Since then, orthologs of this gene have been sequenced, disrupted and shown to affect pathogenicity in eight other plant and animal fungal pathogens.

Other approaches for identifying novel pathogenicity/virulence genes include the use of DNA microarrays, proteomics and metabolomics to identify the genes, proteins and metabolic pathways that are affected during pathogenicity. Fungal material can be collected for these studies either directly from an infected host, or from *in vitro* cultures designed to represent the conditions that exist within an infected host. Such approaches can reveal changes in gene transcription and translation and changes in metabolism indicative of a role in pathogenicity. This information can then be used to identify candidate genes for silencing, deletion and further characterization. The large collections of ESTs and fully sequenced genomes of many pathogenic fungi are also being used to identify novel pathogenicity/virulence genes by comparative genome analysis.

To date, >350 genes have been experimentally verified as pathogenicity or virulence genes. In addition, a smaller number of pathogen effector genes have been demonstrated to be required to trigger plant defence responses. The rate of gene discovery has increased from 9 in 1995 to 55 in 2004. This is

because the efficiency of pathogen transformation continues to increase, through the use of *Agrobacterium* and improved protoplast transformation, whilst new techniques such as gene silencing are employed to identify novel pathogenicity genes. In addition, the availability of genomic sequence simplifies the construction of disruption cassettes, and therefore the rate of gene function discovery can be expected to increase further.

IMPLEMENTATION AND CURATION

Curation

All data in PHI-base is carefully curated by a domain expert and supported by strong experimental evidence. PHI-base was initially populated with data on pathogen host interactions compiled by researchers of the plant–pathogen interaction group at Rothamsted Research. The pathway for this manual data retrieval is shown in Figure 2 (left-hand side). Keyword searches of the literature databases PubMed/MEDLINE and Web of Science use the following search string: (fung* or yeast) and (gene or factor) and (pathogenicity or virulen* or avirulence gene*).

As only ~10% of the returned articles covered pathogenicity genes, the relevant articles are manually curated by a domain expert and transferred into a spreadsheet. Further relevant publications are obtained from recommendations of internal and external colleagues. To validate and supplement this manual approach, automated text mining methods are being developed which operate on a local copy of the MEDLINE database containing more than 15 million references to biomedical publications. For this purpose, we have adapted and extended the ONDEX text mining and database integration framework (10) to identify and extract relevant pathogen host interactions from the scientific literature. Articles identified from the novel text mining methods are carefully curated and checked by a domain expert prior to inclusion in PHI-base to ensure the same high level of quality assurance as used in the manual approach.

New releases and updates of PHI-base are created by a parser which transfers the data from the spreadsheet where it is currently curated, to the relational database back-end of PHI-base. This parser also integrates further information from other external data sources into the spreadsheet. Nucleotide and protein sequences are extracted from the EMBL sequence databases and Gene Ontology (GO) annotations. Enzyme Commission (EC) numbers from the PHI-base curators are supplemented with GO and EC annotations which also come from EBI databases. The parser also generates hyperlinks to external resources such as the NCBI Taxonomy database, Pubmed links and GO terms. In addition, the parser checks and enforces syntactic correctness of the data in the spreadsheet before incorporating information into PHI-base.

Implementation

PHI-base was developed as a relational database using the database management system PostgreSQL. PHI-base can be accessed from any web browser through its web interface which is generated by the server-side scripting language PHP. Currently, PHI-base is installed on LINUX and uses the Apache web server, although it can be installed on any platform that supports PostgreSQL and PHP (i.e. LINUX, UNIX and Windows). The three-tier architecture of PHI-base is shown in Figure 2. Users submit database queries using the front-end via HTML forms. These are then processed by PHP (middle-tier) against the relational database (back-end). The result of each query is then presented to the user in the web browser.

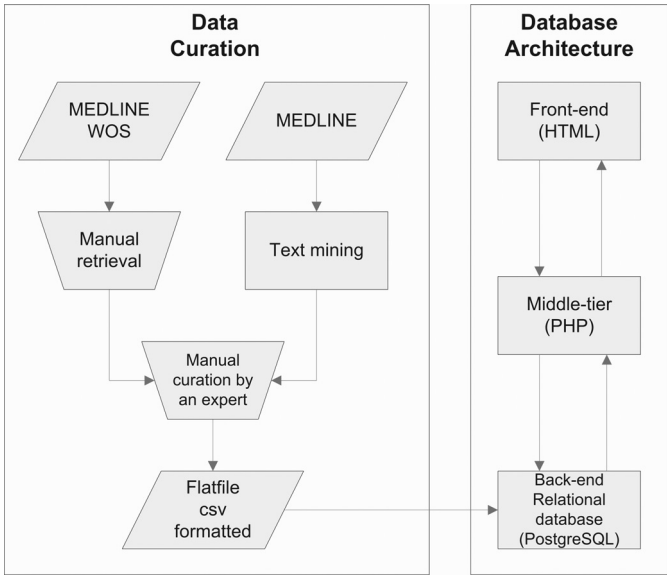

Figure 2. Data curation and architecture of PHI-base. Relevant papers are identified by searches and by text mining from MEDLINE and Web of Science (WOS). Domain experts curate the data which are then transferred to the relational back-end of PHI-base. Users can then query PHI-base via its front-end.

DATABASE DESIGN AND CONTENT

Gene disruption experiments are the experimental basis for PHI-base and this establishes which genes are essential or contribute to host infection and disease formation (see above). The database structure of PHI-base reflects this logic (see Entity Relationship diagram in Figure 3). PHI-base primarily consists of the main tables gene, interaction, species, disease and paper. Each gene in PHI-base is assigned a stable unique accession number which will never change between versions and thus serves as a central reference point in PHI-base. In addition, each gene is characterized by a number of attributes such as its name, nucleotide sequence and its function. Each gene in PHI-base was tested in one or several interactions with a host. For example, the aspartyl proteinase genes (SAP1-6) of *Candida albicans* were all disrupted and tested for pathogenicity defects in two different animal species, i.e. in two pathogen host interactions (PHI:68, PHI:72, PHI:73, PHI:125–127). Each interaction is supported by at least one paper. The disease table stores information on the diseases that are caused by a pathogen. Some pathogens can cause different diseases, depending on the host they infect. However, many different fungal pathogens can cause the same disease across many different hosts. Information on the host and pathogen species is stored in the species table. A full

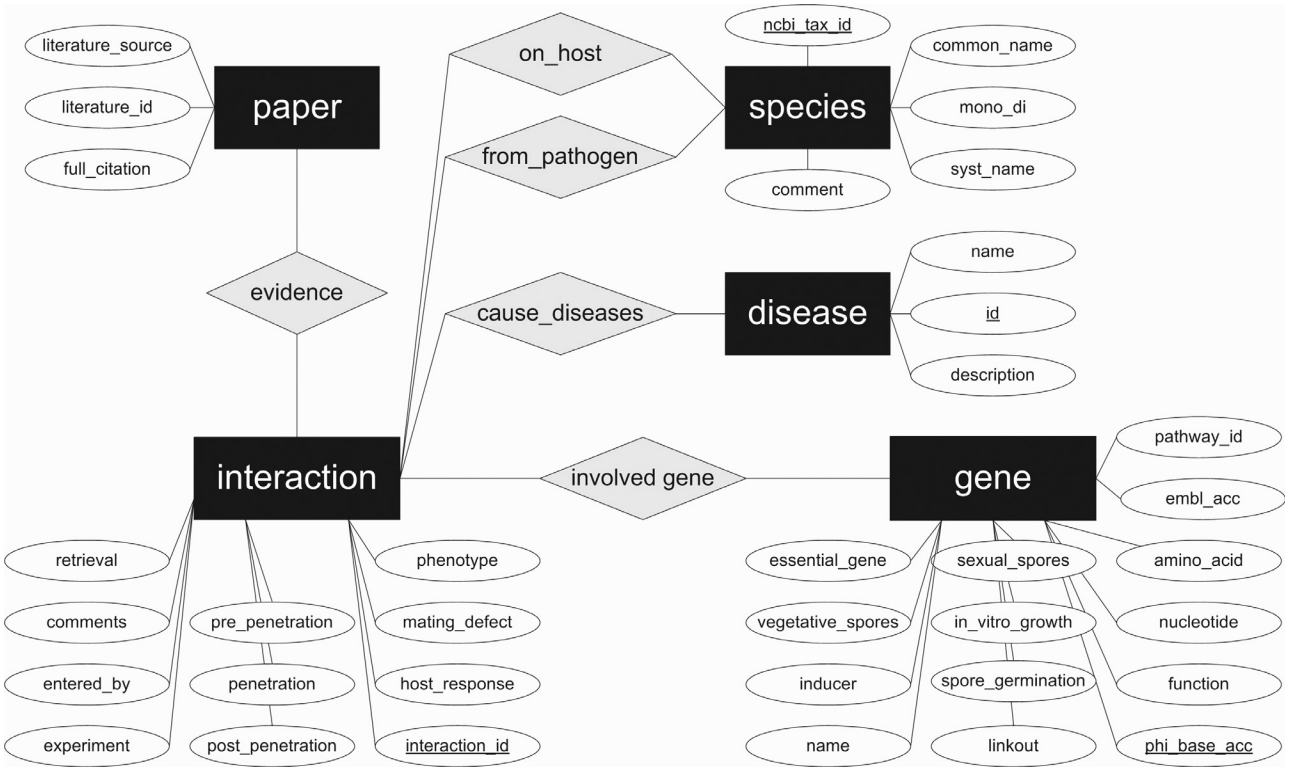

Figure 3. Entity relationship diagram of PHI-base. Each gene in PHI-base was tested in one or several interactions with a host. Each interaction is supported by at least one paper. The disease table stores information on the diseases which are caused by a pathogen host interaction. Information on the host and pathogen species is stored in the species table.

description of the database fields contained within each database table can be found in the online version of PHI-base.

From a technical point of view, we used generalized data structures for storing the different attributes of each table (11). This allows the addition of new data fields according to the changing requirements of the biological curators without having to change the database schema. The database schema is normalized, indexed and wherever it seemed to be sensible, the data is properly coded.

Arranging the information in such a logical and structured way facilitates searching the database, which is a prerequisite for a user-friendly interface. For instance the advanced search form will aid the user by providing drop-down menus giving options derived directly from the relational database structure and contents.

PHI-base contains information about each gene that has been disrupted and shown to affect pathogenicity in all animal, fungal and plant attacking fungal and Oomycete pathogens (Table 1). To date, the genes included in PHI-base are ones that have been verified as pathogenicity, virulence or effector genes. There have been many studies where a gene disruption has not affected pathogenicity, either owing to functional redundancy or simply because it is not involved in pathogenicity. Currently, very few of these studies have been included in PHI-base, but their inclusion is envisaged in future updates. It is important that comparative genomics and gene annotation projects include all the disruption phenotype information available, including studies which resulted in a wild-type phenotype (no effect on pathogenicity).

Table 1. Summary of the number of pathogen and host species and genes within Version 2.1 of PHI-base

Host	Host species	Pathogen species	Pathogenicity genes	Virulence genes	Effector genes
Animal	5*	10*	40	129	0
Plant	32	42	66	110	15
Fungal	3	2	0	3	0

*Five animal species infected by 10 pathogen species.

Each gene entry contains several categories of information. Where the molecular function of a gene product is known, this is included with the GO term or EC number where applicable and available. The phenotype of the pathogenicity defect is summarized in different categories depending on whether the gene disruption abolishes the ability to cause disease (pathogenicity gene), reduces the disease causing ability (virulence gene) or triggers host defence responses (effector). Further categories summarize where the pathogenicity defect occurs, i.e. before (pre-), during or after (post-) penetration of the host, or a combination of these defects (Figure 3).

DATA RETRIEVAL

PHI-base can be searched via its user-friendly interface from any web browser. Two different search options are provided. Full text searches can query the complete database or be restricted to specific database fields using the 'Quick Search' options. The 'Advanced Search' panel allows the user to

define specific database queries. In searches individual genes, diseases, hosts and pathogens can be selected from (initialized) drop-down menus and then be combined by AND or OR operators via radio buttons. Each search returns a result list of distinct pathogen host interactions matching the search criteria. The listed entries are referenced by a stable PHI-base identifier and consist of the interactions' key components as retrieved from the underlying publications, particularly the gene symbol of the experimentally disrupted gene, a functional annotation of this gene, the pathogen name, the phenotype that results from disrupting the gene, the disease that is caused and the host. As far as available, links to external data sources such as the EMBL Sequence Version Archive or NCBI Entrez Taxonomy are directly accessible from the search result list. For any record listed in the result table, the content of all data fields are shown in a separate view which can be accessed via a link in the first table column. In addition to the fields that further characterize the pathogenicity genes and their interactions, most notably the gene nucleotide sequence and the amino acid sequence of the gene product are given as well as links to the supporting publications in NCBI PubMed database. In addition, references to papers and links to supporting publications are included. These may contain information on gene expression, protein expression and cellular localization or altered interaction phenotypes conferred by variant gene sequences. Users can also download all PHI-base sequences as a FASTA file for further computational analysis and sequence similarity searches.

FUTURE DIRECTIONS

Future versions of PHI-base will also include host mutations that compromise or enhance host defence responses (12). We have already compiled from publications >100 *Arabidopsis thaliana* mutants and transgenic lines that modify different types of interactions (13). In addition, for many plant host species the molecular identity of various classes of disease resistance (*R*) genes are now known as well as the function of specific gene variants in the activation of plant responses (14).

The combined use of pathogen and host microarray data, as well as cross-comparisons to viral and bacterial microarray expression data could lead to the recognition of generic pathogen defence pathways and distinguish pathogen-specific mechanisms. Genes implicated by weaker types of correlative evidence stemming from microarray experiments or sequence similarity searches could be considered for inclusion in PHI-base in the future. To this end, the type of experimental evidence will always be annotated, so users can decide for themselves and specify in the query interface what kind of evidence they require.

Currently, each entry in PHI-base is annotated with up to 30 different attributes (Figure 3). In future we plan to increase the number of attributes according to user requests. It is also planned to improve interoperability with external data sources by providing linkouts or bidirectional links to related entries. This will cover the COGEME or fuGIMs databases, as well as pathway information in KEGG (15).

The main challenge for the future is to keep PHI-base up to date with the growing number of experimentally verified and published pathogenicity genes and to incorporate host mutants. Towards this goal, we will continue to improve the text mining support for database curation in terms of ease of use and precision and recall of the text mining methods. Development and application of improved methods for identifying genes in texts such as AbGene (16) or GAPSCORE (17) are of key importance. However, currently even the most successful text mining systems extract a certain amount of incorrect data, which suggests the need for an improved curation process by involving species specific domain experts. Assuring an effective involvement of external experts, we are currently developing guidelines as well as the computational infrastructure for external curation support. This will include advanced web interfaces for data submission and analysis, as well as methods for keeping track of changes by external curators.

DATABASE ACCESS

PHI-base can be freely accessed at http://www4.rothamsted.bbsrc.ac.uk/phibase/. User support can be obtained from this email phi-base@bbsrc.ac.uk. Please use the same email if you wish to provide new data for inclusion in PHI-base, are an interested expert willing to assist with curation or if you have suggestions for improvements.

ACKNOWLEDGEMENTS

R.W. and J.K. wish to acknowledge the advice and support from Ralf Hofestädt and Alexander Rüegg from the Bielefeld University. All authors are grateful to Gavin Harrison and the Rothamsted Computing Services group for efficient provision and support with the computational infrastructure for PHI-base. We also wish to thank Darren Soanes and Nick Talbot for beta-testing earlier versions of PHI-base, John Antoniw for technical support and the Rothamsted Research librarian Maggie Johnston. T.B. is supported by a CASE-studentship sponsored by Syngenta. Rothamsted Research receives grant aided support from the Biotechnology and Biological Sciences Research Council. Funding to pay the Open Access publication charges for this article was provided by the BBSRC.

Conflict of interest statement. None declared.

REFERENCES

1. Agrios,G.N. (1997) *Plant Pathology*. Academic Press, San Diego.
2. Shaner,G., Stromberg,E.L., Lacy,G.H., Barker,K.R. and Pirone,T.P. (1992) Nomenclature and concepts of pathogenicity and virulence. *Annu. Rev. Phytopathol*., **30**, 47–66.
3. Kang,S., Ayers,J.E., DeWolf,E.D., Geiser,D.M., Kuldau,G., Moorman,G.W., Mullins,E., Uddin,W., Correll,J.C., Deckert,G. *et al*. (2002) The internet-based fungal pathogen database: a proposed model. *Phytopathology*, **92**, 232–236.
4. Soanes,D.M., Skinner,W., Keon,J., Hargreaves,J. and Talbot,N.J. (2002) Genomics of phytopathogenic fungi and the development of bioinformatic resources. *Mol. Plant Microbe Interact*., **15**, 421–427.
5. Bulow,L., Schindler,M., Choi,C. and Hehl,R. (2004) PathoPlant: a database on plant–pathogen interactions. *In Silico Biol*., **4**, 529–536.
6. Cornell,M., Paton,N.W., Hedeler,C., Kirby,P., Delneri,D., Hayes,A. and Oliver,S.G. (2003) GIMS: an integrated data storage and analysis environment for genomic and functional data. *Yeast*, **20**, 1291–1306.

7. Newton,A.C., Lyon,G.D. and Marshall,B. (2002) DRASTIC: A database resource for analysis of signal transduction in cells. *BSPP Newsletter*, **42**, 36–37.
8. Braxton,S.M., Onstad,D.W., Dockter,D.E., Giordano,R., Larsson,R. and Humber,R.A. (2003) Description and analysis of two internet-based databases of insect pathogens: EDWIP and VIDIL. *J. Invertebr. Pathol.*, **83**, 185–195.
9. Eckart,J.D. and Sobral,B.W. (2003) A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework. *OMICS*, **7**, 79–88.
10. Köhler,J., Rawlings,C., Verrier,P., Mitchell,R., Skusa,A., Ruegg,A. and Philippi,S. (2005) Linking experimental results, biological networks and sequence analysis methods using ontologies and generalised data structures. *In Silico Biol.*, **5**, 33–44.
11. Philippi,S. (2003) Light-weight integration of molecular biological databases. *Bioinformatics*, **20**, 51–57.
12. Manger,I.D. and Relman,D.A. (2000) How the host 'sees' pathogens: global gene expression responses to infection. *Curr. Opin. Immunol.*, **12**, 215–218.
13. Hammond-Kosack,K.E. and Parker,J.E. (2003) Deciphering plant-pathogen communication: fresh perspectives for molecular resistance breeding. *Curr. Opin. Biotechnol.*, **14**, 177–193.
14. Meyers,B.C., Kaushik,S. and Nandety,R.S. (2005) Evolving disease resistance genes. *Curr. Opin. Plant Biol.*, **8**, 129–134.
15. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
16. Tanabe,L. and Wilbur,W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, **18**, 1124–1132.
17. Chang,J.T., Schutze,H. and Altman,R.B. (2004) GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, **20**, 216–225.