# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities




WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

## Interested in publishing with us?
## Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



# 9

# Mining Host-Pathogen Interactions

Dmitry Korkin, Thanh Thieu, Sneha Joshi and Samantha Warren
*University of Missouri, Columbia, USA*

## 1. Introduction

Infections are caused by a vast variety of pathogenic agents including viruses, bacteria, fungi, protozoa, multicellular parasites, and even proteins (Anderson and May 1979; Morse 1995; Bartlett 1997; Mandell and Townsend 1998) that target host organisms from virtually all kingdoms of life (Daszak, Cunningham et al. 2000; Williams, Yuill et al. 2002). Infectious diseases in humans account for 170 thousand deaths in the United States and 14,7 million deaths world-wide (2004; Rossi and Walker 2005). "Neglected diseases", a group of tropical diseases that are spread among the poorest segment of the world's population, account for a large portion of human infections (Ayoola 1987; Trouiller, Olliaro et al. 2002). With the reluctance of the pharmaceutical industry to invest in the development of drugs for neglected diseases, there is an increasing pressure on the scientific community in academia and non-profit organizations to obtain a fast and inexpensive cure (Trouiller, Torreele et al. 2001; Maurer, Rai et al. 2004; Fehr, Thurmann et al. 2006). In addition to human infections, infections in plant and animals have a multibillion dollar economic impact each year (Bowers, Bailey et al. 2001; Whitby 2001). Expanding the studies to the whole animal kingdom allows scientists to study the host-pathogen evolution of virulence mechanisms that are common among plant and animals, such as type III secretion system (T3SS), an elaborate protein-delivery system (Espinosa and Alfano 2004; Abramovitch, Anderson et al. 2006). Moreover, studying interactions between pathogens and simpler model organisms, such as drosophila, has led to important findings in mammalian systems and is critical for understanding human infections (Cherry and Silverman 2006). Recently another threat has come to scientists' attention: the potential use of some pathogens as bioweapons (Whitby 2001; Moran, Talan et al. 2008). The attacks can target population directly, or they can target strategic resources such as the world's most consumed crops. Studying HPIs may provide critical knowledge for the development of infection diagnosis and treatment for disaster planning in case of a bioterrorism event.

A pathogen causing an infectious disease generally exhibits extensive interactions with the host (Munter, Way et al. 2006). These complex crosstalks between a host and a pathogen may assist the pathogen in successfully invading the host organism, breaching its immune defence, as well as replicating and persisting within the organism. Systematic determination and analysis of HPIs is a challenging task from both experimental and computational approaches, and is critically dependent on the previously obtained knowledge about these interactions. The molecular mechanisms of host-pathogen interactions (HPIs) include

interactions between proteins, nucleotide sequences, and small ligands (Lengeling, Pfeffer et al. 2001; Kahn, Fu et al. 2002; Stebbins 2005; Forst 2006). The interactions between the pathogen and host proteins are one of the most important and therefore widely studied group of HPIs (Stebbins 2005). During the last decade, an increasing amount of experimental data on virulence factors, their structures, and their functions has become available (Sansonetti 2002; Stebbins 2005). The first steps towards large-scale systematic determination and analysis of molecular HPIs have recently emerged for important pathogens (Shapira, Gat-Viks et al. 2009; Dyer, Neff et al. 2010). Recent progress in data mining and bioinformatics allows scientists to accurately predict novel protein-protein interactions, structurally characterize individual proteins and protein complexes, and predict protein functions on a scale of an entire proteome (Thornton 2001; Russell, Alber et al. 2004; Shoemaker and Panchenko 2007). Unfortunately, there have been only a handful of methods designed to address the protein interactions between pathogenic agents and their hosts (Cherkasov and Jones 2004; Davis, Barkan et al. 2007; Dyer, Murali et al. 2007; Lee, Chan et al. 2008; Evans, Dampier et al. 2009; Tyagi, Krishnadev et al. 2009; Doolittle and Gomez 2011). As it is the case for many bioinformatics areas, collecting HPI data into a centralized repository is instrumental in developing accurate predictive methods. Recently, several such HPI repositories have been introduced, some are manually curated, while others are reliant on the existing databases (Winnenburg, Urban et al. 2008; Driscoll, Dyer et al. 2009; Kumar and Nanduri 2010). While this is a promising first step towards a large-scale HPI data collection, one of the largest and most comprehensive sources of experimentally verified HPI data remains largely underexplored: PubMed, a database of peer-reviewed biomedical literature, which includes abstracts of more than 20 million research papers and books (http://www.ncbi.nlm.nih.gov/pubmed/). Unfortunately, the comprehensive manual identification and data extraction of the abstracts containing HPI information from PubMed is not feasible due to the size of PubMed. Furthermore, no informatics approach currently available to do this automatically.

In this chapter, we discuss several possible solutions to the problem of automated HPI data collection from the publicly available literature. The chapter is organized as follows. First, we describe some of the popular HPI databases that are currently available publicly. Second, we discuss the state-of-the-art approaches to a related problem of mining general protein-protein interactions from the literature. Third, we propose three approaches to mine HPIs and discuss the advantages and disadvantages of these approaches. In conclusion, we discuss the future steps in the area of HPI text mining by highlighting factors that are critical for its successful development.

## 2. Host-pathogen interaction databases

During the last several years, a number of resources collecting HPI data have emerged (Snyder, Kampanya et al. 2007; Winnenburg, Urban et al. 2008; Driscoll, Dyer et al. 2009; Kumar and Nanduri 2010). Many resources rely on the automated post-processing of the large-scale databases for general protein-protein interactions, while some other obtain the HPI data by manually curating the biomedical literature. Often the resources focus on the human-pathogen interactions. Next, we will briefly describe some of the popular databases that include HPI data.

**HPIDB - Host-Pathogen Interaction DataBase.** One of the most recent HPI database, HPIDB (Kumar and Nanduri 2010) integrates the information from other HPI database, PIG

(Driscoll, Dyer et al. 2009), and more general protein-protein interaction databases, BIND (Gilbert 2005), GeneRIF(Mitchell, Aronson et al. 2003; Pruitt, Tatusova et al. 2003), IntAct (Aranda, Achuthan et al. 2010), MINT (Zanzoni, Montecchi-Palazzi et al. 2002), and Reactome (Matthews, Gopinath et al. 2009). Currently, the database has 22,841 protein-protein interactions between 49 host and 319 pathogen species (Kumar and Nanduri 2010). HPIDB is searchable via a keyword search, a BLAST search, or a homologous HPI search. For each query, the following output information is obtained: UniProt accession numbers of both host and pathogen proteins, host and pathogen names, detection method, author name, PubMed publication ID (PMID), interaction type, source database, and comments. The homologous HPI search option allows the user to do one or both of the following: search for a set of homologous host proteins, and search for a set of homologous pathogen proteins.

**PATRIC – PAThosystems Resource Integration Center.** PATRIC is a resource that integrates genomics, proteomics, and interactomics data on a comprehensive set of bacterial species as well as a set of data mining and comparative genomics tools (Snyder, Kampanya et al. 2007; Sullivan, Gabbard et al. 2010). The human-pathogen interaction data for 30 bacterial pathogens are also a part of the resource. Similar to HPIDB, the data are extracted and post-processed from a number of general protein-protein interaction databases including BIND (Gilbert 2005), DIP (Xenarios, Fernandez et al. 2001), IntAct (Aranda, Achuthan et al. 2010), and MINT (Zanzoni, Montecchi-Palazzi et al. 2002). With PATRIC a user selects a pathogen from the home page. The search can be refined by selecting specific interaction types (*e.g.*, "direct interaction", "colocalization"), detection methods (*e.g.*, "coimmunoprecipitation", "two hybrid"), or source databases. The results can be visualized as a network of interacting proteins with the colour nodes representing different species and weighted edges representing the number of independent experimental sources supporting the interaction. The Pathogen Interaction Gateway (PIG) is a part of PATRIC that is focused on collecting and analysing exclusively the protein-protein human-pathogen interactions and the corresponding interaction networks (Driscoll, Dyer et al. 2009). The PIG web interface allows mining the data using two query types: the BLAST search and text keyword search. PIG also has a utility that allows the user to visualize the network of protein-protein HPIs followed by the network comparison between the HPI networks extracted for two different pathogen genes.

**PHI-base – the Pathogen-Host Interaction dataBASE**. PHI-base collects information on experimentally verified pathogenicity, virulence and effector genes from bacterial, fungal, and Oomycete pathogens and includes a variety of infected hosts from plants, mammals, fungus, and insects (Winnenburg, Urban et al. 2008). All database entries are manually curated and are supported by experimental evidence and literature citations. The current version has a total of 1,065 gene entries participating in 1,335 interactions between 97 pathogens and 76 hosts, supported by 720 literature references. The interaction between a host and pathogen organism is considered in this database in a more general sense and often is not associated with any physical interaction between the host and pathogen proteins. Using the PHI-base web interface, a user can do either a simple quick search or an advanced search, where the user selects one or many of the following search terms: gene, disease (caused by pathogen), host, pathogen, anti-infective, phenotype, and experimental evidence. The search output is a list of interactions and their details including PHI-base accession number, gene name, EMBL accession number, phenotype of the mutant, pathogen species, disease name, and experimental host. The user can also obtain additional information on nucleotide and amino acid sequences of the pathogen gene, experimental evidence of the

interaction, gene ontology (pathogenesis, molecular function, and biological process), and a publication reference.

## 3. Current approaches for mining protein-protein interactions

Rapid growth of published biomedical research has resulted in the development of a number of methods for biomedical literature mining over the last decade (Krallinger and Valencia 2005; Rodriguez-Esteban 2009). The methods dealing with the biomolecular information can be generally divided into three categories based on the domain of biomedical knowledge they target: (i) automated protein or gene name identification in a text (Mika and Rost 2004; Seki and Mostafa 2005; Tanabe, Xie et al. 2005), (ii) literature-based functional annotation of genes and proteins (Chiang and Yu 2003; Jaeger, Gaudan et al. 2008), and (iii) extracting the information on the relationships between biological molecules, such as proteins and RNAs, or genes (Hu, Narayanaswamy et al. 2005; Shatkay, Hˆglund et al. 2007; Lee, Yi et al. 2008). The relationships detected by the third group of methods range from a co-occurrence of the genes and proteins in a text (Hoffmann and Valencia 2005) to detecting the protein-protein interactions (PPIs) (Blaschke and Valencia 2001; Marcotte, Xenarios et al. 2001; Donaldson, Martin et al. 2003) and identification of signal transduction networks and metabolic pathways (Friedman, Kra et al. 2001; Hoffmann, Krallinger et al. 2005; Santos and Eggle 2005). Being a special case of protein-protein interactions, HPIs could directly benefit from the advancements of the currently existing text mining methods.

Extraction of protein-protein interactions from the text has been one of the three main tasks for the recent BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenges, a community-wide effort for evaluating biological text mining and information retrieval systems (Hirschman, Yeh et al. 2005; Krallinger, Leitner et al. 2008). Three subtasks have been specified: (i) detection of protein-protein interactions relevant documents (interaction article subtask, IAS), (ii) identification of sentences with protein-protein interactions (interaction sentences subtask, ISS), and (iii) identification of interacting protein pairs (interaction pair subtask, IPS). A relevant problem, the protein interaction method subtask (IMS), is concerned with identification of the type of experimental data used to determine an interaction. Approaches that address these subtasks vary from supervised machine learning classifiers, to address the first subtask, to statistical language processing and grammar-based methods to address the second and third subtasks.

A simple approach to extract protein-protein interactions is to determine the co-existence of proteins in the same sentence (Stephens, Palakal et al. 2001; Hoffmann and Valencia 2005). However, this approach is insufficient to handle structured information of biomedical sentences. Therefore, pattern matching methods have been proposed that rely on either manually defined patterns (Leroy and Chen 2002; Corney, Buxton et al. 2004) or patterns that are automatically generated using dynamic programming (Huang, Zhu et al. 2004; Hao, Zhu et al. 2005). Another popular group of methods employs the natural language processing parsers. A basic approach, called shallow parsing, decomposes sentences into non-overlapping fragments and chunks, and defines the dependencies between the chunks without extracting their internal structure (Thomas, Milward et al. 2000; Leroy, Chen et al. 2003). Many shallow parsing approaches employ finite-state automata to recognize the interaction relationships between proteins or genes (Thomas, Milward et al. 2000; Leroy, Chen et al. 2003). One of the most prominent approaches relies on the deep parsing

techniques, where the entire structure of a sentence is extracted (Park, Kim et al. 2001; Ding, Berleant et al. 2003; Daraselia, Yuryev et al. 2004; Pyysalo, Ginter et al. 2004; Kim, Shin et al. 2008; Miyao, Sagae et al. 2009). Many deep parsing approaches have successfully employed link grammars (Sleator and Temperley 1995), context-free grammars that rely on a dictionary of rules (linking requirements) to connect, or "link", pairs of related words (Ahmed, Chidambaram et al. 2005; Seoud, Youssef et al. 2008; Yang, Lin et al. 2009).

Each of the above methods, while directly addressing the second and the third subtasks, can also solve the abstract classification problem from the first subtask, based on whether or not the method is able to extract any protein-protein interactions. The accuracy of such classification, however, depends on the accuracy of a more difficult subtask of protein-protein interaction extraction. Thus, several methods have been developed to directly address the problem of binary classification of protein-protein interaction relevant publications (Marcotte, Xenarios et al. 2001; Calli 2009; Kolchinsky, Abi-Haidar et al. 2010). The methods primarily rely on supervised and unsupervised feature-based classification techniques. Recently, the first method for classification of HPI-relevant documents has been introduced, which employs a Support Vector Machines (SVM) supervised classifier (Yin, Xu et al. 2010).

## 4. New approaches to detection and mining host-pathogen interactions from biomedical abstracts

HPI literature mining is related to a general problem of protein-protein interaction literature mining. However, the additional requirement that the interaction occurs exclusively between the host and pathogen proteins makes the task more challenging. The accuracy of an HPI mining method will depend on additional factors, such as its ability to correctly assign a host or pathogen organism to the interacting protein. Similar to the way the BioCreAtIvE initiative defines three types of protein-protein interaction mining problems (Hirschman, Yeh et al. 2005), the problem of HPI mining can be split into three specific tasks:

*HPI Mining Task 1*: Given a biomedical publication (a paper or an abstract), determine whether or not it contains information on HPIs.

*HPI Mining Task 2*: Given a biomedical publication containing HPI information, determine specific sentences that contain this information.

*HPI Mining Task 3*: Given a biomedical publication that contain HPI information, determine specific pairs of host and pathogen proteins participating in the interactions and the corresponding organisms.

The first task can be formulated as a standard classification problem, which is often addressed by machine learning methods and for which a number of the method assessment protocols have been developed. Here we rely on the following five basic measures. The first measure, accuracy, is calculated as $f_{AC} = \left(N_{TP} + N_{TN}\right)/N$, where $N_{TP}$ and $N_{TN}$ are the number of true positives and negatives, correspondingly, and $N$ is the number of classified interfaces. The other two related measures, precision and recall, are calculated as $f_{PR} = N_{TP}/\left(N_{TP} + N_{FP}\right)$ and $f_{RE} = N_{TP}/\left(N_{TP} + N_{FN}\right)$, correspondingly, where $N_{FP}$ and $N_{FN}$ are the number of false positives and negatives. F-score is calculated as $F = 2\frac{f_{PR}f_{RE}}{f_{PR} + f_{RE}}$. The last

measure, the Matthew correlation coefficient is calculated as $MCC = \frac{N_{TP}N_{TN} - N_{FP}N_{FN}}{\sqrt{(N_{TP}+N_{FP})(N_{TP}+N_{FN})(N_{TN}+N_{FP})(N_{TN}+N_{FN})}}$. Similarly, performance on the last task can be easily assessed based on the available information about the host and pathogen proteins and their respective organisms. Specifically, we use four different measures. The first two measures, $f_{ORG}$ and $f_{PRT}$, address the accuracy of detecting the pairs of interacting host and pathogen organisms as well as their proteins. Each measure is calculated as a percentage of the number of correctly detected pairs of organisms/proteins to the total number of pairs. The other two measures, $g_{ORG}$ and $g_{PRT}$, account for the partial detection of HPI information, when at least one of the two organisms or proteins is detected. Both measures are defined as the percentage of the total number of detected organisms/proteins to the total number of organisms/proteins in all HPIs.

Unfortunately, evaluating a method's performance for the second task is more challenging, since the HPI data are often (i) scattered across multiple sentences and (ii) redundant (for instance, the same interaction between two proteins can be mentioned in several sentences). The method assessment for the second task becomes even more challenging when multiple HPIs are present in the same abstract.

We next introduce several strategies that address the above tasks for the PubMed biomedical abstracts (here and below, we will always consider an abstract of the biomedical publication together with the publication's title; the latter often provides important information on HPIs). One of the main reasons behind extracting HPI information from the abstracts rather than entire papers is the fact that for many papers, the abstract is the only information that is freely available in PubMed. The first strategy is to rely on the existing methods for mining protein-protein interactions followed by additional post-processing to filter out the intra-species interactions. Another approach employs the language-based methods traditionally used in protein-protein interaction literature mining. The last approach introduces a supervised-learning feature-based methodology, which has recently emerged in the area of biomedical literature mining. While each of the approaches is applicable to each of the three tasks, here we will focus on assessing their performance for the first and third tasks.

### 4.1 Data collection

Collecting accurate, unbiased, non-redundant data on HPIs is a critical step for efficient training of a supervised method as well as for an accurate assessment of any literature mining approach. Both the positive set (abstracts containing HPI information) and the negative set (abstracts that do not contain HPI information) were manually selected and annotated. To obtain the set of potential candidates for the positive and negative sets we have combined of both searching the existing HPI databases and the PubMed database. Our positive set consisted of 175 HPI containing abstracts that include human and non-human hosts. The abstracts containing human-pathogen interactions were collected by searching and manually curating abstracts from PIG, a database of host-pathogen interactions manually extracted from the literature (Driscoll, Dyer et al. 2009). For each abstract, we required the presence of organism and protein names for both the host and the pathogen, resulting in 89 abstracts. Unfortunately, in its current form, PIG only has the abstracts with annotated human-pathogen interactions. Therefore to obtain the list of interactions between non-human hosts and their pathogens, we searched using an extensive PubMed query. We

required the presence in the same abstract of (i) at least one (non-human) host name, (ii) at least one pathogen name, (iii) and at least one interaction keyword. We then manually selected from the list another 86 abstracts that contained HPI information, adding them to the positive set.

To obtain candidates for the negative set, we performed an almost identical search strategy using the same PubMed query but including 'human' to the list of the host names. We again manually selected the abstracts to ensure that that they did not have any HPI information, even though they contained the important keywords. Note that it is significantly harder for a computational approach to distinguish between the abstracts from the obtained negative training set and those from the positive set, compared to a negative training set consisting of abstracts that were randomly chosen from PubMed. As a result, we selected 175 abstracts where no HPI information was found, although some of the abstracts included information on intra-species protein-protein interactions. The list of manually curated positive and negative sets of PubMed abstracts can be found at: http://korkinlab.org/datasets/philm/philm_data.html

### 4.2 A naïve approach based on literature mining of protein-protein interactions

In a simple naïve approach, we first establish whether an abstract contains any information on a protein-protein interaction using the existing state-of-the-art literature mining methods followed by extraction of the pair of interacting proteins (Fig. 1A). We rely on the PIE system, which integrates the natural language processing and machine learning methods to determine the sentences that contain protein-protein interactions in a PubMed abstract and extract the corresponding protein names and the interaction keywords (Kim, Shin et al. 2008). Next, for each interacting protein we identify its corresponding organism by applying NLProt protein/gene tagging software (Mika and Rost 2004). NLProt uses a number of techniques, such as the dictionary search, rule-based detection, and feature-based supervised learning, to extract the names of proteins and genes and tag them using SWISS-PROT or TrEMBL identifiers (Boeckmann, Bairoch et al. 2003). The method also predicts the most likely organisms associated with these proteins/genes. It was reported to have a precision of 75% and a recall of 76% on detecting protein/gene names (Mika and Rost 2004). Finally, for each sentence identified as containing a protein-protein interaction by the PIE system, we determine if this interaction is a HPI. Specifically, if each of the two proteins forming a protein-protein interaction belongs to a different organism, and these organisms can be assigned the host-pathogen roles, then the interaction is classified as an HPI. To assign the host-pathogen roles, we use our manually curated dictionaries of host and pathogen organism names (Table 1).

We assessed the naïve approach by applying it to our testing set of 88 abstracts, 44 positive and 44 negative examples. As a result in addressing Task 1, the obtained accuracy was 0.53, precision was 1.0, and recall was 0.07 for the classification of HPI-containing abstracts (Task 1); F-score and Matthews Correlation Coefficient were 0.13 and 0.19, correspondingly. We found that the method almost completely failed to detect the abstracts containing HPI information; the contribution to the accuracy came primarily from the true negative hits, containing 44 (out of 44) abstracts from the negative testing set. Interestingly, both high precision and low recall values could be attributed to the same property of the naïve approach: it failed to accurately detect the protein-protein interactions. Indeed, all 41 false negatives were not due to the approach's failure to assign the host and pathogen roles to the identified organisms, but due to its failure to identify a protein-protein interation in the abstract.

It is also not surprising that the naïve approach performed poorly when addressing Task 3: the method was able to detect only two proteins out of 44 protein pairs and none of the 44 pairs of organisms, resulting in the only non-zero score of $g_{PRT} = 0.02$; the other three scores, $f_{ORG}$, $f_{PRT}$, and $g_{ORG}$ were equal to zero.

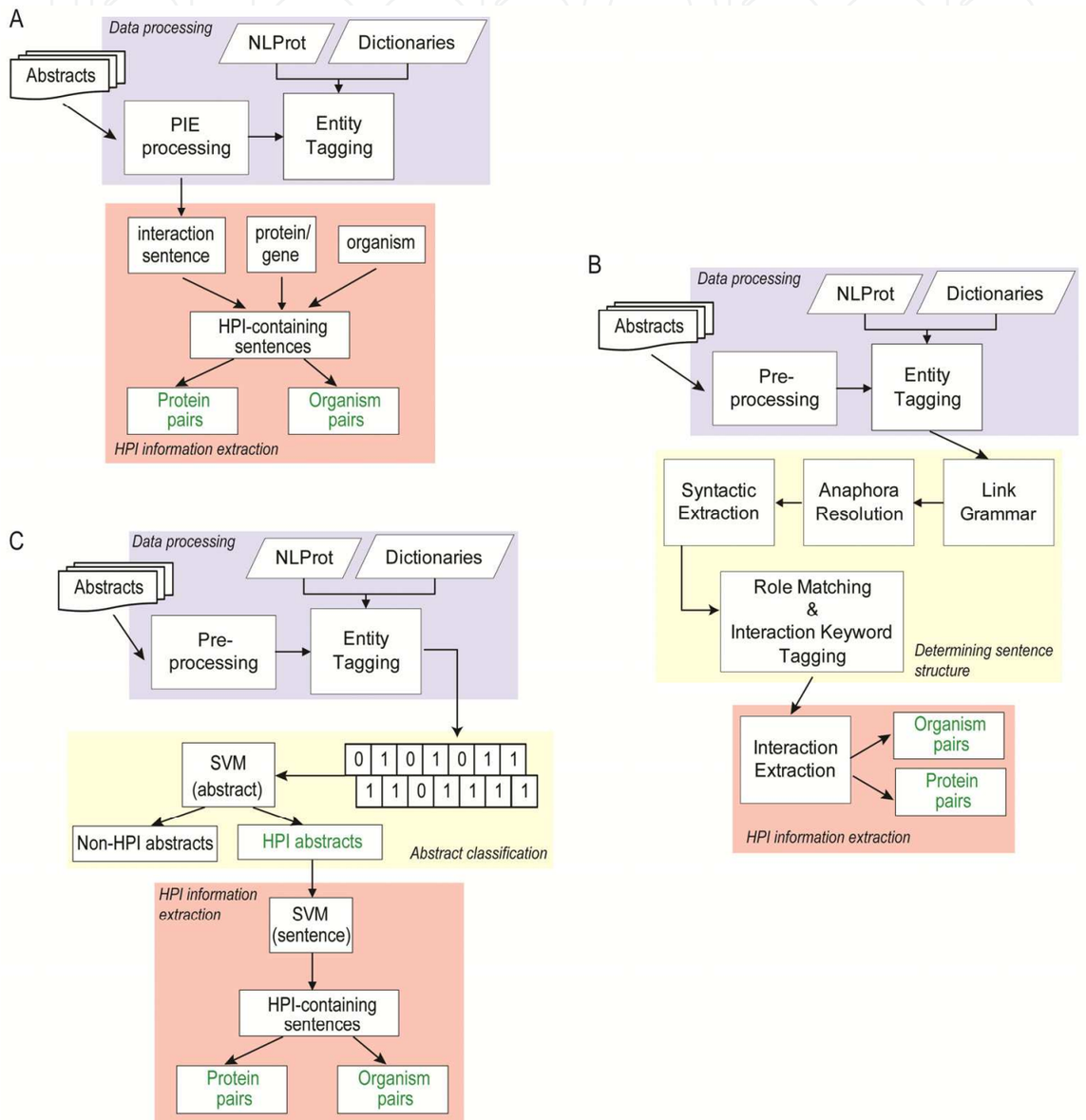



Fig. 1. Three HPI literature mining approaches. (A) Naïve approach. (B) Language-based approach (C) Feature-based supervised machine learning approach.

| Dictionary name | *N* | Examples |
|---|---|---|
| Interaction keywords | 54 | *Interact, associate, bind* |
| Experimental keywords | 28 | *Yeast two-hybrid, chemical crosslinking* |
| Negation keywords | 11 | *Not, neither, inability* |
| HPI specific keywords | 17 | *Virulence, effectors, infection* |
| Host names | 309 | *Host, plant, human* |
| Pathogen names | 349 | *Listeria monocytogenes, Hepatitis virus* |

Table 1. Dictionaries of keywords used by all three approaches. N is the number of unique entries for each dictionary.

### 4.3 A language-based approach

Our second approach is inspired by the language-based methods in biomedical text mining, which are also widely used in mining protein-protein interactions. In HPI text mining, we are faced with additional challenges such as correctly associating the organism name for each protein, ensuring that the extracted interaction is inter- and not intra-species interaction, and combining the information about an HPI from multiple sentences. As a result, these additional challenges necessitate adding new modules to the computational pipeline of our approach compared with a pipeline for extracting general protein-protein interactions. The HPI mining pipeline consists of the following 7 steps (Fig. 1B): (1) text preprocessing, (2) entity tagging, where we identify protein/gene and organism names, (3) grammar parsing, where we parse the input text into dependency structures (4) anaphora resolution, where we identity references to pronouns, (5) syntactic extraction, where we split a complex sentence into simple ones, (6) role matching, where we identify semantic roles in each simple sentence, (7) interaction keyword tagging, and (8) extraction of the actual HPI information. We note that this approach directly addresses Tasks 2 and 3 by finding the sentences containing HPI information and extracting the corresponding pairs of host and pathogen organisms and the interacting proteins/genes. Task 1 is addressed by classifying each abstract based on whether there was at least one HPI with the complete information extracted from the abstract's text.

**Entity tagging.** The entity tagging module identifies named entities in a abstract, such as protein/gene names and the corresponding organism names. For a language-based text mining approach, it is critical that all named entities are accurately identified. Thus, our language–based approach for HPI literature mining has the most elaborate entity tagging module of all three approaches introduced here. Specifically, the module includes three stages: (i) protein/gene name tagging using NLProt, (ii) host/pathogen organism dictionary match, and (iii) post-processing. First, we apply the NLProt tagger to identify the names of all proteins/genes occurring in the text and the corresponding organism names (Mika and Rost 2004). We note that in a case when a protein with the same name exists for multiple species, NLProt assigns the most likely organism for each entry of this protein. Second, we find a UniProt accession number (Bairoch, Apweiler et al. 2005) for each identified protein followed by grouping the proteins/genes with the same accession number into a protein/gene entity. Third we search for the organisms missed by NLProt using expanded versions of our host and pathogen organism dictionaries that include synonyms for each

organism name and group the organisms under NCBI Taxonomy IDs (Wheeler, Barrett et al. 2006). Since NLProt may not identify all organisms in the abstract, our module rescans the abstract text again to find the remaining host and pathogen organisms. Finally, the system revisits the entity tagging module again after the next module, Link grammar parsing, provides the internal structure of the sentences in terms of its basic units, phrases. The idea is that we can use the internal sentence structure to (i) find additional host/pathogen information that is not present in the dictionary, and (ii) reassign protein/gene name to its correct organism, if needed. This stage plays an important role in the entity tagging module, since our host and pathogen dictionaries are potentially incomplete (not all organisms provided by NLProt may be covered); in addition, the dictionaries overlap with each other (the same organism can be both, a host and a pathogen). If an organism name suggested by NLProt for a protein is not found in our dictionary, the entity tagging module nevertheless tries to assign the organism's role as a host or pathogen. It does so by searching for generic keywords (such as "host", "pathogen", "pathogenic", "pathogenesis", *etc.*), in each phrase containing the organism name. Similarly, the module checks the organism name suggested by NLProt for a protein/gene by identifying the organism's name in the phrase that contains a protein/gene name. To do so the module relies on two search patterns:

1. Organism name + protein name (*e.g.*, "*Arabidopsis* RIN4 protein");
2. Protein name + preposition + organism name (*e.g.*, "RXLX of human").

The newly obtained information about the organism assignment then replaces the current suggestions provided by NLProt. For instance, in the phrase "the *Arabidopsis* RIN4 protein", NLProt associates RIN4 with a pathogenic organism, while the dictionary search matches *Arabidopsis* as a host organism and identifies this phrase as pattern P1. Therefore, *Arabidopsis* is assigned as the organism for RIN4 protein, followed by the correct assignment of RIN4 as a host protein.

**Link grammar parsing.** In our next module, we use natural language processing methods to determine the intrinsic structure of each sentence in the abstract. In our approach, all grammatical constructions are based on the link grammar, a context-free grammar that relies on the dependency structure of natural language (Sleator and Temperley 1995). In link grammar, every word has a linking requirement, which specifies which types of other words or phrases can link to it. Two words can only be linked if their linking requirements match. A link is represented as an arc above the two words (Fig. 2). The linking requirements are organized into a dictionary that the grammar parser refers to when analyzing a sentence. The principal structure in link grammar is the linkage, a set of links that completely connect all words in a sequence. Such a sequence of words is called a link grammar sentence if it satisfies three conditions: (i) the links do not cross (planarity), (ii) each word is connected to at least another word by a link (connectivity), and (iii) the linking requirements for each word in the sentence are not violated (satisfaction). For example, the linkage for the sentence "Avirulence protein B targets the Arabidopsis RIN4 protein" is shown in Fig. 2. In total, the link grammar has 107 main links, each of which can derive many sub-links. We implemented the module using an open source link grammar parser from AbiWord project (http://www.abisource.com/projects/link-grammar/). This project implements the original link grammar (Sleator and Temperley 1995), combining it with additional features such as adaptation of the parser to the biomedical sublanguage, BioLG (Pyysalo, Salakoski et al. 2006) and an English-language semantic dependency relationship extractor, RelEx (Fundel, Kuffner et al. 2007).

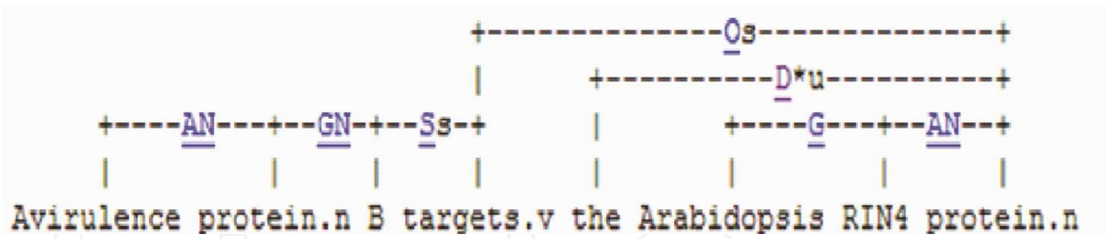



Fig. 2. Internal sentence structure annotated by a link grammar parser for an HPI relevant sentence. Words are labelled with the part-of-speech tags: .n (noun) and .v (verb). A link between two words can be formed to specify a dependency relation. Each dependency type has its own unique label: AN, GN, Ss, Os, D*u, G.

**Anaphora resolution.** In the anaphora resolution module, we determine semantic meaning for pronouns (it, they, he, she), and other language structures in the sentences. Unlike the case of intra-species protein-protein interactions, the information on HPIs often spans multiple sentences, with the pronouns often replacing the names of organisms or proteins/genes. Therefore, to extract the complete information on a HPI, it is critical to have an accurate anaphora resolution module. The module relies on the RelEx anaphora resolution method, which employs Hobbs' pronoun resolution algorithm (Hobbs 1978). For example, in the sentence "The *Pseudomonas syringae* type III effector protein avirulence protein B (AvrB) is delivered into plant cells, where it targets the *Arabidopsis* RIN4 protein", the anaphora resolution module resolves 'it' as 'The *Pseudomonas syringae* type III effector protein avirulence protein B (AvrB)'.

**Syntactic extraction.** Our syntactic extraction module splits each sentence into one or more simple sentences, where a simple sentence consists of four components organized into the following structure:

Subject (S) + Verb (V) + Object (O) + Modifying phrase of verb (M).

The module is built based on the automated extractor InTex (Ahmed, Chidambaram et al. 2005); it scans a sentence to find all links of the following four types. The first type, *S*-link, connects a subject to a verb, where the subject is located before the verb in the sentence. The second type, *RS*-link connects a verb to a subject, *i.e.*, the subject is located after the verb in the sentence. The third type, *O*-link, connects a verb to an object. Finally, the fourth type, *MV*-link, connects a verb to a modifying phrase. The module first determines the beginning of each simple sentence, which can be either an S-link or an RS-link. Following each verb from an *S*- or *RS*-link, the module determines the verb range by including all possible verb phrases, adverb phrases, or adjective phrase, before and after the verb. Finally, for each simple sentence the module determines the objects and modifying phrases for the verb in the corresponding verb range by identifying possible O-links and MV-links. For example, the modules split sentence "The Pseudomonas syringae type III effector protein avirulence protein B (AvrB) is delivered into plant cells, where it targets the Arabidopsis RIN4 protein" into two simple sentences: "The Pseudomonas syringae type III effector protein avirulence protein B (AvrB) is delivered into plant cells" and "The Pseudomonas syringae type III effector protein avirulence protein B (AvrB) targets the Arabidopsis RIN4 protein".

**Interaction keyword tagging.** In this module, the interaction keywords are tagged by searching (i) our manually curated dictionary of interaction keyword stems, to reduce the search time, and (ii) lexical database WordNet, which contains nouns, verbs, adjectives, and

adverbs grouped by semantic concepts, and which uses a morphological function to infer the stem of a word (Fellbaum 1998). In the previous example, the module identifies interaction keywords that are found in our dictionary: "delivered" (the stem is "deliver") and "targets" (the stem is "target").

**Role type matching.** In this module, we specify the role of each syntactic component depending on whether the component contains complete information about an HPI. Here, we consider three types of roles: elementary, partial, and complete. A component of the elementary type is defined to be a host entity, a pathogen entity, or an interaction keyword. A component of the partial type includes any two distinct components of the elementary type. Finally, a syntactic component of complete type includes components of all three elementary types.

**Interaction extraction**. Once the role of each syntactic component is identified, the components are searched against a set of interaction patterns. We first select components of the complete type, since they contain complete information about an HPI occurring between two proteins/genes. Next, we combine the elementary and partial components such that they provide the complete HPI information.

An interaction pattern is defined as *LS*=*RS*. The left side (*LS*) is used to match the complete type from syntactic component(s), and the right side (*RS*) is used to extract the interaction information from each component. For example, the pattern S<E>V<E>O<E> = P<S>I<V>H<O> indicates that if a simple sentence includes three components, each of elementary type: subject, verb, and object, then the sentence contains (i) a pathogen entity in the subject, (ii) an interaction keyword in the verb, and (iii) a host entity in the object. Note that both sides include a matching part S-V-O. In this work, for our patterns we considered the following seven matching parts: S-V-O, S-O, S-V-M, S-M, S, O, and M (for abbreviations, see *Syntactic extraction* subsection). In addition to the above patterns, we use a set of three template-based filters that allows us to remove those simple sentences that although satisfy an interaction pattern, do not have a semantic connection between the host entity, pathogen entity, and interaction keyword. The introduced templates are similar to those employed by RelEx:

*Pattern 1*: A + interaction verb + B

*Pattern 2*: Interaction noun + 'between' + A + 'and' + B.

*Pattern 3*: Interaction noun + 'of' + A + 'by' + B,

where an interaction keyword can be either the interaction verb or interaction noun.

**Interaction Normalization.** When mining HPI information from literature, there are several sources for ambiguous information. First, there may be multiple HPIs in the same abstract. Second, the information about a single HPI may be spread over multiple sentences. Finally, the sentences may contain duplicate information about the same HPI. Our last module ensures that all sentences containing duplicate HPIs are accounted for and each HPI is reported only once. To do so, we first extract all HPIs and then determine the duplicate pairs. We define two HPIs as duplicate if they have the same host entity and the same pathogen entity. We note that two duplicate HPIs may still have different interaction keywords. To detect the duplication in HPIs, the module refers to the normalized protein/gene names (in terms of UniProt accession numbers) and organism names (in terms of taxonomy ids) obtained at the entity tagging module.

**Performance of the language-based approach.** To compare with the feature-based approach, the language-based approach was evaluated using the same testing set of 44 positive and 44 negative examples. We first assessed the method's performance in

addressing Task 1. The method was able to classify the abstracts with 0.65 accuracy, 0.84 precision, and 0.36 recall. The F-score and Matthew correlation coefficient measures were 0.51 and 0.36, correspondingly. The performance of the approach on a more difficult Task 3 was significantly better than of the naïve approach, especially in partial predictions: $f_{ORG}$ = 0.18, $f_{PRT}$ = 0.14, $g_{ORG}$ = 0.25, and $g_{PRT}$ = 0.25. With the pre-calculated NLProt annotation, the average running time of the system on a single abstract was 36.3 sec. on a 2.4 Ghz Intel workstation. The computationally most expensive, link grammar parsing, module used 99.95% of the total running time.

### 4.4 A feature-based machine learning approach

The basic idea behind the feature-based approach introduced here is to extract a set of characteristic features that provide sufficient information for discriminating between an abstract containing HPI information and another abstract that does not. Using a training set of pre-annotated abstracts, the system can then learn how to efficiently discriminate between these two abstract types. Moreover, the same characteristic features can be calculated for the individual sentences in the abstract. Thus, we can use the same supervised-learning approach to solve Tasks 1 and 2. Finally, to solve Task 3 one can use a simple dictionary-based search for each sentence classified as containing HPI information. Our feature-based approach consists of four basic stages (Fig. 1C). First, each abstract is pre-processed to find each protein/gene in the abstract and identify its organism name. Second, for each abstract a feature vector is generated. Third, our supervised learning system is trained by providing the feature vectors generated from the positive and negative sets. Finally, the trained system is used on an independent testing set of HPI and non-HPI abstracts to assess the approach.

**Text preprocessing.** We first add the publication title to the abstract as its first sentence. The abstract is then further split into individual sentences by detecting the sentence termination patterns. A basic pattern of a period (.), followed by a space and capitalized letter can be directly used to distinguish sentences in a standard text. However, there are known challenges when preprocessing a biomedical (or any scientific) publication. For instance, the above simple approach is not always applicable, since the periods are often used in the names of proteins, abbreviations such as "*i.e.*", "*e.g.*", "*vs.*", and others. We first identify such cases using a predefined dictionary, replace periods in these words by spaces, and then apply the above basic pattern. The next steps of the preprocessing stage concerns with detecting the organism and protein/gene names using the entity tagging software NLProt (Mika and Rost 2004).

**Support vector machines in text categorization.** The problem of detecting whether an abstract contains HPI information can be formulated as a problem of supervised text categorization, with the goal of classifying abstracts into one of the selected categories. In our case, two categories can be naturally defined: (i) abstracts containing HPI information and (ii) abstracts without HPI information. Formally, given a training set of $n$ objects, each represented as a vector of $N$ numerical features, $\boldsymbol{x}^i = (x_1, x_2, \ldots, x_N)$, and their classification into one of the two classes $\boldsymbol{y} \in \{-1,1\}$, the goal is to train a feature-based classifier based on the training set. After the training stage is completed, the classifier can assign a class label from $\boldsymbol{y}$ for any new abstract $\boldsymbol{x}$. In our approach, we use support vector machines (SVM) (Vapnik 1998), a supervised learning method, which is well established in bioinformatics and has been recently applied to identify abstracts containing host-bacteria interaction

information (Yin, Xu et al. 2010). The basic type of support vector machine (SVM) that addresses this problem is a linear classifier defined by its discriminant function:

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b, \quad \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \sum_{i=1}^{N} w_i x_i ,$$

where $\boldsymbol{w}$ is a weight vector (Vapnik 1998). Geometrically, the problem can be described as finding the decision boundary, a hyperplane that separates two sets of points, corresponding to the sets of positive and negative examples. To do that, we maximize the margin defined by the closest to the hyperplane positive and negative examples. An optimal solution can be found by solving a related quadric optimization problem. The problem is further generalized by introducing soft margins, allowing the classifier to misclassify some points. The general optimization problem is often formulated in its dual form:

$$\underset{\alpha}{\text{minimize}} \left[ -\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}^i, \boldsymbol{x}^j \rangle \right]$$

$$\text{subject to: } \sum_{i=1}^{n} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le C \text{ for } i = 1, 2, ..., n$$

and the discriminant function is defined as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} y_i \alpha_i \langle \boldsymbol{x}^i, \boldsymbol{x} \rangle + b .$$

Examples from the training set for which $\alpha_i > 0$ are called support vectors. The formalism can be further extended by introducing non-linear classifiers defined using kernel functions, $K(\boldsymbol{x}, \boldsymbol{x}')$, similarity measures that replace the standard inner product $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$. In our approach, we applied and compared two widely used non-linear kernel functions: the polynomial kernel, $K^P(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + 1)^d$, where $d$ is degree of the polynomial, and Gaussian radial basis function (RBF), $K^G(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2 / c)$. Both kernels are implemented using *libsvm* a freely available SVM software package (Chang and Lin 2001).

**Feature vectors.** One approach to generating a descriptive set of features for an abstract is to calculate the frequencies of occurrences of individual words (unigrams) as well as the word pairs (bigrams) from a biomedical text corpus (Yin, Xu et al. 2010). While these features can provide important information on the word usage, the number of features depends strongly on the size of the corpus and can easily reach thousands of features. In our approach, we propose to use a simpler 12-dimensional feature vector representation, $\boldsymbol{x} = (x_1, x_2, \rightleftharpoons, x_{12})$, focusing on quantifying the information directly related to host-pathogen interaction. Features $x_1$ and $x_2$ quantify the presence of host and pathogen protein or gene names in the abstract and are calculated based on the protein/gene entity tagging obtained by NLProt (Mika and Rost 2004). Each protein is classified as a host or pathogen protein based on the source organisms extracted either from the NLProt tagging results or directly from the abstract by searching against our dictionary of host and pathogen organisms (Table 1). The dictionary was built using the set of organisms extracted from several databases (Winnenburg, Urban et al. 2008; Driscoll, Dyer et al. 2009; Kumar and Nanduri 2010) and by adding generic keywords, such as "pathogen", "host", "plant", etc. Similarly, features $x_3$

and $x_4$ specify the number of occurrences for the host and pathogen organism names. These features are defined using NLProt-based organism annotation and the dictionary of host and pathogen organisms. Binary feature $x_5$ specifies the presence or absence of the general protein-protein interaction keywords in the abstract. It is obtained by scanning the extended abstract against our interaction keyword dictionary (Table 1). Features $x_6$ and $x_7$ describe additional statistics on protein-protein interaction keyword occurrences. The former feature is defined as the percentage of interaction keywords in the total number of words in the abstract. The latter feature is defined as the percentage of sentences containing the interaction keywords in the total number of abstract sentences. Feature $x_8$ is calculated based on the cumulative keyword *typicality* for each abstract. We define the typicality of a keyword as the percentage of abstracts in the training set containing this keyword. Feature $x_8$ is calculated as a sum of typicalities for all protein-protein interactions keywords in a given abstract. Our next feature $x_9$ quantifies the amount of experimental evidence used to support the HPI and is defined as the total number of experimental keywords in the abstract, where each keyword is detected by scanning the abstract against our dictionary experimental keywords (Table 1). Some abstracts report the absence of an interaction between host and pathogen proteins. Determining the absence of interaction in an abstract by a feature-based approach is difficult, since such an abstract is likely to contain the information similar to an abstract describing a true HPI. One of the key differences between these abstracts is the presence of negation keywords present in the former abstract. Feature $x_{10}$ accounts for such keywords and is defined as the percentage of negation keywords in the total number of words in the abstract. Similar to other keywords, these keywords are identified using our dictionary of collected negation keywords (Table XX3). A related feature, $x_{11}$, estimates whether a negation keyword is related specifically to the information on protein-protein interaction in the abstract. The feature is defined as the number of words between the negation keyword and the closest interaction keyword in a sentence. The last feature, $x_{12}$, accounts for the HPI-specific keywords, such as *virulence, effectors, factors, etc*. determined using the corresponding dictionary (Table 1). It is calculated as a percentage of such keywords in the total number of words in the abstract.

**Supervised training and HPI detection using SVM.** The trained SVM classifier is applied in our method twice. First, it is applied to the abstracts to identify those containing HPI information (Task 1). Second, it is applied to the individual sentences to determine those that contain this HPI information (Task 2). When applied to a sentence, we generate a 12-dimensional feature vector solely based on the information in this sentence and use it as an input to the SVM classifier. Once the sentences containing HPI are identified, we use the dictionaries of host and pathogen organisms combined with the protein/gene names to find the pairs of host and pathogen organisms and the corresponding proteins/genes (Task 3). The accuracy of an SVM-based classifier generally can be improved by optimizing a number of parameters during the training stage. The error cost parameter, *C*, controls the tradeoff between allowing training errors and forcing rigid margins. In our approach we select the cost parameter and another parameter, Gamma, by evaluating the accuracies of trained models for Task 1 using leave-one-out cross-validation. The values of *C* range from 2 to 20 and the values for Gamma range from $2^{-10}$ to $2^{1}$. The set of parameters on which the SVM classifier reaches its maximum accuracy is selected as a final model. In addition, we optimize the degree of the polynomial when considering the polynomial kernel.

**Assessment protocols.** To assess the performance of the feature-based approach in abstract classification, we employ two benchmarking protocols. In the first protocol, the SVM model

training is done on the training set and the assessment is performed exclusively on the testing set (Table 2). For the second protocol we use the leave-one-out and 10-fold cross validations on the training set.

| Type | Training | Testing |
|---|---|---|
| Negative | 131 | 44 |
| Positive- Human | 67 | 22 |
| Positive-Non-Human | 64 | 22 |
| Total | 262 | 88 |

Table 2. Testing and training sets of positive (HPI-relevant) and negative (HPI-irrelevant) abstracts. Testing data are used to evaluate all three approaches, and training data are used for SVM learning in the feature-based approach. The abstracts are extracted from then PubMed database and manually curated.

**Performance of the feature based approach.** During the leave-one-out cross-validation, an SVM model with the polynomial kernel of degree 3 and parameter values $C$=2 and Gamma=0.0175 was found to be the most accurate in the abstract classification problem (Table 3). The polynomial kernel was also the most accurate SVM model across both assessment protocols. In addition, this SVM model had the highest recall value, with the precision approaching its highest value. Overall, the performance of all three SVM kernels, across all evaluation protocols, was similar. The performance of the feature-based approach on Task 3 was slightly better than that of the language-based approach in partial predictions: $g_{ORG}$ = 0.39 and $g_{PRT}$ = 0.35. However the performance in complete pair predictions was worse: $f_{ORG}$ = 0.07 and $f_{PRT}$ = 0.07. The SVM classifier was efficient, taking only 0.003 sec. to classify 92 abstracts by an SVM classifier on a 2.66 Ghz Intel Xeon (Quad) workstation. However, the high efficiency of this approach was offset by a significantly slower protein tagging component that was done using NLProt and took ~18 min. on the same workstation to tag proteins in 262 abstracts from the dataset.

| Protocol | $f_{AC}$ | PR | RE | AUC | F-score |
|---|---|---|---|---|---|
| 10-fold | 72% | 73% | 71% | 0.78 | 0.72 |
| Test | 66% | 69% | 60% | 0.72 | 0.64 |
| LOO | 71% | 72% | 72% | 0.78 | 0.71 |

Table 3. Evaluations of the feature-based classifier. LOO and 10-fold denote leave-one-out and 10-fold cross-validation protocols applied to the models that are trained on the set of 262 abstracts. The last protocol corresponds to the evaluation performed only on the testing set of 88 abstracts.

## 5. Conclusion

In this chapter, we discussed a new problem for biomedical literature mining that was concerned with mining molecular interactions between the host and pathogen organisms. Collecting HPI data is one of the very first steps towards studying and fighting infectious diseases. Creating an automated framework for extracting the HPI information from the

biomedical literature, including millions of abstracts publicly available in PubMed database, is instrumental in completing this step. We formulated three key tasks of HPI literature mining and proposed three computational approaches that addressed these tasks: (i) a naïve approach, which was based on the existing protein-protein interaction mining methods, (ii) a language-based approach, which employed the link grammar, and (iii) a feature-based supervised learning approach, which relied on SVM methodology. Both, feature-based and language-based, approaches have been implemented in the PHILM (Pathogen-Host Interaction Literature Mining) web-server, accessible at http://korkinlab.org/philm.html.

Several important conclusions can be drawn from the comparative assessment of all three approaches. First, it became clear that being a new problem in biomedical literature mining (and a more difficult one than mining general protein-protein interactions), HPI text mining required development of new methods tailored to address the specifics of this problem. Indeed, for the first task the naïve approach performed with the disappointingly low accuracy of 53% and f-score of just 13%, while accuracy and f-score of the language-based approach were significantly higher, 65% and 51%, correspondingly; the feature-based method had even higher (10-fold) accuracy and f-score, 72% and 72%, correspondingly. We note that the performance accuracy of both language-based and feature-based approaches even at this early stage were comparable to the state-of-the-art protein-protein interactions mining methods (Krallinger, Leitner et al. 2008). In addition to its poor performance in the abstract classification task, the naïve approach completely failed to detect protein interaction pairs and organism pairs in the third task. The feature-based approach performed significantly better when detecting one of the interacting proteins or organisms, while still failing to accurately detect the complete pairs. It was not surprising that the highest accuracy of detecting both, host-pathogen organism pairs and protein pairs, was achieved by the most sophisticated language-based approach. Second, the analysis of incorrectly classified abstracts and identified pairs of proteins and organisms supported our conclusion that increasing the accuracy of the name tagging system is pivotal to increasing the classification accuracy in both approaches. Finally, both language-based and feature-based approaches demonstrated good performance but in different tasks, which suggests that by integrating these two approaches, one can obtain a system with a more accurate overall performance than either of the individual approaches.

## 6. Acknowledgment

We acknowledge funding from University of Missouri (Mizzou Advantage to DK), National Science Foundation (DBI-0845196 to DK), and Department of Education (GAANN Fellowship to SW).

## 7. References

(2004). WHO, The world health report 2004: changing history. Geneva, World Health Organization.

Abramovitch, R. B., J. C. Anderson, et al. (2006). "Bacterial elicitation and evasion of plant innate immunity." *Nat Rev Mol Cell Biol* 7(8): 601-611.

Ahmed, S. T., D. Chidambaram, et al. (2005). *IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text*. Proceedings of the ACL-ISMB Workshop

on Linking Biological Literature. Ontologies and Databases: Mining Biological Semantics, Detroit, Association for Computational Linguistics.

Anderson, R. M. and R. M. May (1979). "Population biology of infectious diseases: Part I." *Nature* 280(5721): 361-367.

Aranda, B., P. Achuthan, et al. (2010). "The IntAct molecular interaction database in 2010." *Nucleic Acids Research* 38(Database issue): D525-531.

Ayoola, E. A. (1987). "Infectious diseases in Africa." *Infection* 15(3): 153-159.

Bairoch, A., R. Apweiler, et al. (2005). "The Universal Protein Resource (UniProt)." *Nucleic Acids Res* 33(Database issue): D154-159.

Bartlett, J. G. (1997). "Update in infectious diseases." *Annals of internal medicine* 126(1): 48-56.

Blaschke, C. and A. Valencia (2001). "The potential use of SUISEKI as a protein interaction discovery tool." *Genome informatics. International Conference on Genome Informatics* 12: 123-134.

Boeckmann, B., A. Bairoch, et al. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." *Nucleic Acids Research* 31(1): 365.

Bowers, J. H., B. A. Bailey, et al. (2001). "The impact of plant diseases on world chocolate production." *Plant Health Progress*.

Calli, C. (2009). *Prediction of protein-protein interaction relevance of articles using references*. 24th International Symposium on Computer and Information Sciences (ISCIS 2009), Guzelyurt, IEEE.

Chang, C. and C. Lin (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cherkasov, A. and S. J. Jones (2004). "An approach to large scale identification of non-obvious structural similarities between proteins." *BMC Bioinformatics* 5: 61.

Cherry, S. and N. Silverman (2006). "Host-pathogen interactions in drosophila: new tricks from an old friend." *Nat Immunol* 7(9): 911-917.

Chiang, J. H. and H. C. Yu (2003). "MeKE: discovering the functions of gene products from biomedical literature via sentence alignment." *Bioinformatics* 19(11): 1417-1422.

Corney, D. P., B. F. Buxton, et al. (2004). "BioRAT: extracting biological information from full-length papers." *Bioinformatics* 20(17): 3206-3213.

Daraselia, N., A. Yuryev, et al. (2004). "Extracting human protein interactions from MEDLINE using a full-sentence parser." *Bioinformatics* 20(5): 604-611.

Daszak, P., A. A. Cunningham, et al. (2000). "Emerging infectious diseases of wildlife--threats to biodiversity and human health." *Science* 287(5452): 443-449.

Davis, F. P., D. T. Barkan, et al. (2007). "Host pathogen protein interactions predicted by comparative modeling." *Protein Sci* 16(12): 2585-2596.

Ding, J., D. Berleant, et al. (2003). "Extracting biochemical interactions from MEDLINE using a link grammar parser."

Donaldson, I., J. Martin, et al. (2003). "PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine." *BMC Bioinformatics* 4: 11.

Doolittle, J. M. and S. M. Gomez (2011). "Mapping protein interactions between Dengue virus and its human and insect hosts." *PLoS neglected tropical diseases* 5(2): e954.

Driscoll, T., M. D. Dyer, et al. (2009). "PIG--the pathogen interaction gateway." *Nucleic Acids Research* 37(Database issue): D647-650.

Driscoll, T., M. D. Dyer, et al. (2009). "PIG--the pathogen interaction gateway." *Nucleic Acids Res* 37(Database issue): D647-650.

Dyer, M. D., T. M. Murali, et al. (2007). "Computational prediction of host-pathogen protein-protein interactions." *Bioinformatics* 23(13): i159-166.

Dyer, M. D., C. Neff, et al. (2010). "The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis." *PLoS One* 5(8): e12089.

Espinosa, A. and J. R. Alfano (2004). "Disabling surveillance: bacterial type III secretion system effectors that suppress innate immunity." *Cell Microbiol* 6(11): 1027-1040.

Evans, P., W. Dampier, et al. (2009). "Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs." *BMC medical genomics* 2: 27.

Fehr, A., P. Thurmann, et al. (2006). "Editorial: drug development for neglected diseases: a public health challenge." *Trop Med Int Health* 11(9): 1335-1338.

Fellbaum, C. (1998). *WordNet : an electronic lexical database*. Cambridge, USA, MIT Press.

Forst, C. V. (2006). "Host-pathogen systems biology." *Drug Discov Today* 11(5-6): 220-227.

Friedman, C., P. Kra, et al. (2001). "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles." *Bioinformatics* 17(Suppl 1): S74.

Fundel, K., R. Kuffner, et al. (2007). "RelEx--relation extraction using dependency parse trees." *Bioinformatics* 23(3): 365.

Gilbert, D. (2005). "Biomolecular interaction network database." *Brief Bioinform* 6(2): 194-198.

Hao, Y., X. Zhu, et al. (2005). "Discovering patterns to extract protein-protein interactions from the literature: Part II." *Bioinformatics* 21(15): 3294-3300.

Hirschman, L., A. Yeh, et al. (2005). "Overview of BioCreAtIvE: critical assessment of information extraction for biology." *BMC Bioinformatics* 6 Suppl 1: S1.

Hobbs, J. (1978). "Resolving pronoun references." *Lingua* 44(4): 311-338.

Hoffmann, R., M. Krallinger, et al. (2005). "Text mining for metabolic pathways, signaling cascades, and protein networks." *Sci STKE* 2005(283): pe21.

Hoffmann, R. and A. Valencia (2005). "Implementing the iHOP concept for navigation of biomedical literature." *Bioinformatics* 21 Suppl 2: ii252-258.

Hu, Z. Z., M. Narayanaswamy, et al. (2005). "Literature mining and database annotation of protein phosphorylation using a rule-based system." *Bioinformatics* 21(11): 2759-2765.

Huang, M., X. Zhu, et al. (2004). "Discovering patterns to extract protein-protein interactions from full texts." *Bioinformatics* 20(18): 3604-3612.

Jaeger, S., S. Gaudan, et al. (2008). "Integrating protein-protein interactions and text mining for protein function prediction." *BMC Bioinformatics* 9 Suppl 8: S2.

Kahn, R. A., H. Fu, et al. (2002). "Cellular hijacking: a common strategy for microbial infection." *Trends Biochem Sci* 27(6): 308-314.

Kim, S., S. Y. Shin, et al. (2008). "PIE: an online prediction system for protein-protein interactions from text." *Nucleic Acids Research* 36(Web Server issue): W411-415.

Kolchinsky, A., A. Abi-Haidar, et al. (2010). "Classification of protein-protein interaction full-text documents using text and citation network features." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 7(3): 400-411.

Krallinger, M., F. Leitner, et al. (2008). "Overview of the protein-protein interaction annotation extraction task of BioCreative II." *Genome Biology* 9 Suppl 2: S4.

Krallinger, M. and A. Valencia (2005). "Text-mining and information-retrieval services for molecular biology." *Genome Biol* 6(7): 224.

Kumar, R. and B. Nanduri (2010). "HPIDB-a unified resource for host-pathogen interactions." *BMC Bioinformatics* 11(Suppl 6): S16.

Lee, H., G. Yi, et al. (2008). "E3Miner: a text mining tool for ubiquitin-protein ligases." *Nucleic Acids Research* 36(Web Server issue): W416.

Lee, S. A., C. H. Chan, et al. (2008). "Ortholog-based protein-protein interaction prediction and its application to inter-species interactions." *BMC Bioinformatics* 9 Suppl 12: S11.

Lengeling, A., K. Pfeffer, et al. (2001). "The battle of two genomes: genetics of bacterial host/pathogen interactions in mice." *Mamm Genome* 12(4): 261-271.

Leroy, G. and H. Chen (2002). "Filling preposition-based templates to capture information from medical abstracts." *Pac Symp Biocomput*: 350-361.

Leroy, G., H. Chen, et al. (2003). "A shallow parser based on closed-class words to capture relations in biomedical text." *Journal of biomedical informatics* 36(3): 145-158.

Mandell, G. L. and G. C. Townsend (1998). "New and emerging infectious diseases." *Transactions of the American Clinical and Climatological Association* 109: 205-216; discussion 216-207.

Marcotte, E. M., I. Xenarios, et al. (2001). "Mining literature for protein-protein interactions." *Bioinformatics* 17(4): 359-363.

Matthews, L., G. Gopinath, et al. (2009). "Reactome knowledgebase of human biological pathways and processes." *Nucleic Acids Research* 37(Database issue): D619-622.

Maurer, S. M., A. Rai, et al. (2004). "Finding cures for tropical diseases: is open source an answer?" *PLoS Med* 1(3): e56.

Mika, S. and B. Rost (2004). "NLProt: extracting protein names and sequences from papers." *Nucleic Acids Res* 32(Web Server issue): W634-637.

Mika, S. and B. Rost (2004). "Protein names precisely peeled off free text." *Bioinformatics* 20(suppl 1): i241.

Mitchell, J. A., A. R. Aronson, et al. (2003). "Gene indexing: characterization and analysis of NLM's GeneRIFs." *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*: 460-464.

Miyao, Y., K. Sagae, et al. (2009). "Evaluating contributions of natural language parsers to protein-protein interaction extraction." *Bioinformatics* 25(3): 394-400.

Moran, G. J., D. A. Talan, et al. (2008). "Biological terrorism." *Infect Dis Clin North Am* 22(1): 145-187, vii.

Morse, S. S. (1995). "Factors in the emergence of infectious diseases." *Emerg Infect Dis* 1(1): 7-15.

Munter, S., M. Way, et al. (2006). "Signaling during pathogen infection." *Sci STKE* 2006(335): re5.

Park, J. C., H. S. Kim, et al. (2001). "Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*: 396-407.

Pruitt, K. D., T. Tatusova, et al. (2003). "NCBI Reference Sequence project: update and current status." *Nucleic Acids Research* 31(1): 34-37.

Pyysalo, S., F. Ginter, et al. (2004). *Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions*. nternational Workshop on Natural Language

Processing in Biomedicine and its Applications (JNLPBA), Association for Computational Linguistics.

Pyysalo, S., T. Salakoski, et al. (2006). "Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches." *BMC Bioinformatics* 7(Suppl 3): S2.

Rodriguez-Esteban, R. (2009). "Biomedical text mining and its applications." *PLoS Comput Biol* 5(12): e1000597.

Rossi, V. and J. Walker (2005). Assessing the Economic Impact and Costs of Flu Pandemics Originating in Asia. Oxford: Abbey House, Oxford Economic Forecasting Group.

Russell, R. B., F. Alber, et al. (2004). "A structural perspective on protein-protein interactions." *Curr Opin Struct Biol* 14(3): 313-324.

Sansonetti, P. (2002). "Host-pathogen interactions: the seduction of molecular cross talk." *Gut* 50 Suppl 3: III2-8.

Santos, C. and D. Eggle (2005). "Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction." *Bioinformatics* 21(8): 1653.

Seki, K. and J. Mostafa (2005). "A hybrid approach to protein name identification in biomedical texts." *Information Processing & Management* 41(4): 723-743.

Seoud, A., A. Youssef, et al. (2008). *Extraction of protein interaction information from unstructured text using a link grammar parser*, IEEE.

Shapira, S. D., I. Gat-Viks, et al. (2009). "A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection." *Cell* 139(7): 1255-1267.

Shatkay, H., A. Hˆglund, et al. (2007). "SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data." *Bioinformatics* 23(11): 1410.

Shoemaker, B. A. and A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners." *PLoS Comput Biol* 3(4): e43.

Sleator, D. and D. Temperley (1995). *Parsing English with a Link Grammar*. Third International Workshop on Parsing Technologies, ACL/SIGPARSE.

Snyder, E. E., N. Kampanya, et al. (2007). "PATRIC: the VBI PathoSystems Resource Integration Center." *Nucleic Acids Research* 35(Database issue): D401-406.

Stebbins, C. E. (2005). "Structural microbiology at the pathogen-host interface." *Cell Microbiol* 7(9): 1227-1236.

Stephens, M., M. Palakal, et al. (2001). "Detecting gene relations from Medline abstracts." *Pac Symp Biocomput*: 483-495.

Sullivan, D. E., J. L. Gabbard, Jr., et al. (2010). "Data integration for dynamic and sustainable systems biology resources: challenges and lessons learned." *Chemistry & biodiversity* 7(5): 1124-1141.

Tanabe, L., N. Xie, et al. (2005). "GENETAG: a tagged corpus for gene/protein named entity recognition." *BMC Bioinformatics* 6(Suppl 1): S3.

Thomas, J., D. Milward, et al. (2000). "Automatic extraction of protein interactions from scientific abstracts." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*: 541-552.

Thornton, J. M. (2001). "From genome to function." *Science* 292(5524): 2095-2097.

Trouiller, P., P. Olliaro, et al. (2002). "Drug development for neglected diseases: a deficient market and a public-health policy failure." *Lancet* 359(9324): 2188-2194.

Trouiller, P., E. Torreele, et al. (2001). "Drugs for neglected diseases: a failure of the market and a public health failure?" *Trop Med Int Health* 6(11): 945-951.

Tyagi, N., O. Krishnadev, et al. (2009). "Prediction of protein-protein interactions between Helicobacter pylori and a human host." *Molecular bioSystems* 5(12): 1630-1635.

Vapnik, V. N. (1998). *Statistical learning theory*. New York, Wiley.

Wheeler, D., T. Barrett, et al. (2006). "Database resources of the national center for biotechnology information." *Nucleic Acids Research*.

Whitby, S. M. (2001). "The potential use of plant pathogens against crops." *Microbes Infect* 3(1): 73-80.

Williams, E. S., T. Yuill, et al. (2002). "Emerging infectious diseases in wildlife." *Revue scientifique et technique* 21(1): 139-157.

Winnenburg, R., M. Urban, et al. (2008). "PHI-base update: additions to the pathogen host interaction database." *Nucleic Acids Research* 36(Database issue): D572.

Winnenburg, R., M. Urban, et al. (2008). "PHI-base update: additions to the pathogen host interaction database." *Nucleic Acids Research* 36(Database issue): D572-576.

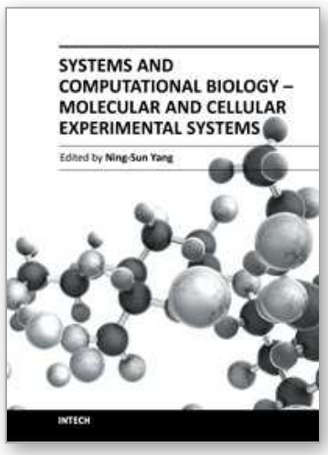
Xenarios, I., E. Fernandez, et al. (2001). "DIP: The Database of Interacting Proteins: 2001 update." *Nucleic Acids Research* 29(1): 239-241.

Yang, Z., H. Lin, et al. (2009). "BioPPIExtractor: A protein-protein interaction extraction system for biomedical literature." *Expert Systems with Applications* 36(2): 2228-2233.

Yin, L., G. Xu, et al. (2010). "Document classification for mining host pathogen protein-protein interactions." *Artificial Intelligence in Medicine* 49(3): 155-160.

Zanzoni, A., L. Montecchi-Palazzi, et al. (2002). "MINT: a Molecular INTeraction database." *FEBS Lett* 513(1): 135-140.

IntechOpen




**Systems and Computational Biology - Molecular and Cellular Experimental Systems**

Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-280-7
Hard cover, 332 pages
**Publisher** InTech
**Published online** 15, September, 2011
**Published in print edition** September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book presents a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

**How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Dmitry Korkin, Thanh Thieu, Sneha Joshi and Samantha Warren (2011). Mining Host-Pathogen Interactions, Systems and Computational Biology - Molecular and Cellular Experimental Systems, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-280-7, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-molecular-and-cellular-experimental-systems/mining-host-pathogen-interactions

INTECH
open science | open minds

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen