Plant Omics Journal POJ

POJ 6(6):388-398 (2013) ISSN:1836-3644

The prediction of protein-protein interaction of *A. thaliana* and *X. campestris* pv. *campestris* based on protein domain and interolog approaches

Nilubon Kurubanjerdjit [1,2], Jeffrey J.P Tsai [1], Chen-Yu Sheu [1, 3], Ka-Lok Ng [1, 4 *]

[1]Department of Biomedical Informatics, Asia University, 41354, Taiwan
[2]School of Information Technology, Mea Fah Luang University, 57100, Thailand
[3]Department of Electrical Engineering and Computer Science, University of California, Irvine, USA
[4]School of Pharmacy, China Medical University, Taichung, Taiwan 40402

***Corresponding author: ppiddi@gmail.com**

Abstract

Plant pathogenic bacteria cause many serious disease symptoms by injecting a variety of effector proteins to reprogram the host defense mechanism. Protein-protein interaction is an essential process playing a crucial role in host-pathogen interactions and pathogenicity. However few have been known for how pathogen bacteria interact with their hosts. In this study, the interactions of *A. thaliana* proteins and Xanthomonas campestris pv.campestris (*Xcc*) pathogen bacteria proteins are identified by two different approaches: the domain-based approach which infers interspecies protein-protein interactions by known domain-domain interactions recorded by various databases; and the interolog approach that identifies protein-protein interactions based on homologous pairs of protein interactions across different organisms. The results from these two methods are integrated and the dense protein interaction regions are specified by clique percolation analysis. In particular, the plant resistance genes (PRG) information and the bacterial effector proteins are studied to provide new insights into the molecular mechanism of plant immunity systems against bacteria. From our findings, we demonstrate that a pathogen employs five strategies to reprogram the host defense mechanism. First, ome *Xcc* proteins tend to interact with *A. thaliana*'s hub proteins or the PRGs. Second, *Xcc* proteins tend to interact with many *A. thaliana* proteins indicating that a pathogen mutates its genes to infect the host. Third, some *Xcc* proteins target a group of *A. thaliana* proteins that are involved in responding to cadmium ions, a significant plant biological process against pathogen. Fourth, many *Xcc* proteins target a few *A. thaliana* proteins which are involved in the plant-pathogen interaction pathways. Finally, the pathogen may make use of a type III effector protein to reprogram the host protein-protein interactions. Host-pathogen interactions remain a largely unexplored area in computational biology. The present work may provide some key information useful for revealing the biological mechanism of a plant's immune system against bacteria. A web-based interface has been established (http://ppi.bioinfo.asia.edu.tw/AtXccPPI) where investigators can pose queries.

Keywords*: Arabidopsis thaliana, Xanthomonas campestris pv campestris*, Protein-protein interaction, Plant-pathogen interactions, Domain-based prediction, Interolog, Clique network clustering, Gene Ontology, Effector protein.
Abbreviations: *A. thaliana_Arabidopsis Thaliana*; *Xcc_Xanthomonas campestris* pv. *campestris;* PRG_Plant Resistance Genes; PPI_Protein-Protein Interaction; GO_Gene Ontology; DDI_ Domain-Domain Interaction; DIP_the Database of Interacting Proteins; TAIR_The Arabidopsis Information Resource; 3DID_3D Interacting Domains database; DAVID_The Database for Annotation, Visualization and Integrated Discovery; PFam_Protein Family; UniProt_Universal Protein Resource; CPM_Clique Percolation Method; KEGG_Kyoto Encyclopedia of Genes and Genomes.

Introduction

Plants are continuously invaded by pathogens including bacteria, fungi, nematodes, viruses and insect pests. Generally, a pathogenic bacterium attacks hosts in many ways including sticking and colonizing host tissues, secreting degradation enzymes and toxins release. Many of such mechanisms involve host-pathogen protein-protein interactions (PPI). PPI is an essential process of living cells (Lin et al., 2004). It also plays a crucial role in some critical interspecies interactions such as host-pathogen interactions and pathogenicity (Casadevall and Pirofski, 2000). Recently high throughput proteomic technology has uncovered a large number of PPI, particularly in interspecies protein interactions of plants and bacteria (Tsao et al., 2011). Therefore, comprehensive knowledge of host-pathogen PPI and interactome analysis can help accelerating protein annotations and elucidate a plant's immune system against bacteria. Several computational methods have been developed to evaluate and predict PPI, such as mRNA-co expression based on the assumption that proteins that are co-expressed are more likely to interact in comparison to proteins that are not co-expressed (Browne et al., 2010). The Gene Ontology (GO) annotation (Wu et al., 2006) implies that proteins found within the same biological process are more likely to interact than proteins from a different biological process. The Interolog approach involves PPI transfer from one organism to another using comparative genomics (Jansen et al., 2003). With the protein domain interaction approach (Ng et al., 2003), PPI could be inferred

by recognizing protein domains and the interaction transfers by known domain-domain interactions (DDI). Also, it was proposed that PPI can be inferred from protein structural information (Ogmen et al., 2005). Among those computational methods, the interolog approach has been broadly used for PPI prediction (Von-Mering et al., 2007). Also, the interolog approach has been justified to be reliable on exploring interaction subnetworks in cancer (Rhodes et al., 2005). The interolog hypothesis is based on the assumption that two proteins (A and B) are predicted to interact if a known interaction between two proteins (A' and B') exists, where A is similar to A' and B is similar to B'. Gallone (Gallone et al., 2011) introduced InterologWalk, a PERL program, to identify putative PPI by implementing the interolog method on fruit fry (Drosophila melanogaster). Wang (Wang et al., 2012) had worked on the prediction of PPI networks in swine based on the interolog, domain-motif interaction and motif-motif interaction approaches. Their work has shown that the interolog approach could yield high accuracy and precision in PPI prediction. Moreover, the work of He (He et al., 2008) implemented the interolog method to construct a predicted PPI map of rice and fungus Magnaportha grisea. They found pathogenicity proteins tend to interact with hub proteins, i.e., proteins with higher numbers of interaction partners. A protein domain is a conserved structural or functional unit of protein sequence which is the key regulator in PPI. The biological mechanism underlying PPI involves protein domains and their interactions. The domain-based method adopts DDI information to infer potential PPI based on the assumption that if proteins *A* and *B* contain an interacting domain pair, it is expected that the two proteins interact with each other. Recently, the domain-based method has received much attention in PPI prediction. Kim (Kim et al., 2008) constructed XooNET which is an integrated database for *Xanthomonas oryzae pathovar oryzae* based on the Protein Structural Interactome MAP (PSIMAP), Protein Experimental Interactome MAP (PEIMAP) and DDI. Furthermore, the interacting domain profile pairs (IDPP) approach was proposed by Wojcik and Schachter (Wojcik and Schachter, 2001). This method uses a combination of sequence similarity search and clustering based on interaction patterns and domain information. Their work proved that the use of domain information has provided better identification than the use of methods which are based on sequence information only. In addition, Li (Li et al., 2011) predicted the host-pathogen PPI between bacterium *R. solanacearum* and *A. thaliana* based on two well-known PPI prediction methods: (1) interolog, and (2) domain-based. Their findings illustrated that the potential targeted proteins of a pathogen are more important in the host PPI network, and the over represented biological functions are different in the two species: The *R. solanacearum* proteins are mainly enriched in transportation, and *A. thaliana* targeted proteins are enriched in the functions that respond to environmental stimuli. It is well-known that *A. thaliana*, a long day plant, is a good model organism for plant science (Mandoli and Olmstead, 2000). *A. thaliana* is chosen as the model system for two reasons: (1) The complete genome sequence has been known since 2000; and (2) There are many molecular tools, such as cDNA, genomic libraries, bacterial artificial chromosomes, microarrays and ESTs, are available for the study of its biological functions (Mandoli and Olmstead, 2000). Only a small number of bacteria are pathogenic on *A. thaliana*, where more than 3,000 proteins are directly related to the plant defense response mechanism (Bishop et al., 2000, Postnikova et al., 2011). In this work, we focus on *Xanthomonas campestris pv campestris* (*Xcc*) which is one of the pathogenic bacteria that cause blights and rots in plants (Tsuji and Somerville, 1988; Tsuji et al., 1991; Tsuji and Somerville, 1992; Buell C. Robin, 2002). Host infections caused by *Xcc* can occur in any stage of the plant life cycle. Symptoms resulted from this pathogen have been reported in many previous research works (Tsuji and Somerville, 1988; Tsuji et al., 1991; Tsuji and Somerville, 1992; Buell C. Robin, 2002; Mayer et al., 2005). Recently, PPI networks in model organisms such as *Saccharomyces* cerevisiae (Uetz et al., 2000; Ito et al., 2000; Ito et al., 2001) and *Escherchia coli* (Arifuzzaman et al., 2006) have been reported. However, few have been known for plant-pathogen interactions. Flor (Flor 1971) proposed a theory on the PPI between pathogen effector protein and a specific receptor in the host plant that results in a hypersensitive response and resistance. Pinzon (Pinzol et al., 2010) purposed a targeted metabolic reconstruction approach for characterizing plant-pathogen interactions by investigating the host metabolic network phenotype during the interaction with pathogens. Protein interactions appear to form a molecular network, which usually contains small circuit patterns called network motifs. The proteins of a network motif are usually involved in similar biological processes, and protein complexes can be identified by clustering the network (Palla et al., 2005; Jonsson et al., 2006). While many computational techniques have been proposed to predict and prioritize PPI (Berggard et al., 2007), we focus here on the interspecies PPI of *A. thaliana* and *Xcc* based on the interolog and the domain based approach. With comprehensive and up to date sets of known PPI and DDI, experimentally confirmed PPI data can be obtained from the Database of Interacting Proteins (DIP) (Xenarios I et al., 2000) which integrates a diverse body of experimental evidences on PPI into a single, easily accessible online database. This resource contains much more PPI information compared to the Arabidopsis Information Resource (TAIR) (Rhee et al., 2003) and BioGrid (Stark et al., 2005). Besides, we adopt the known DDI information obtained from two different up to date resources: the iPFam database (Finn et al., 2005) and the 3D Interacting Domains database (3DID) (Stein et al., 2011). Furthermore, we study the molecular network motifs specified by the clique percolation technique and the enriched biological processes of both *A. thaliana* and *Xcc* appeared in the motifs. In particular, we focus on the plant resistance genes (PRG) information and the effector bacterial protein to provide new insights into the plant pathogen interaction networks. A prediction pipeline has been implemented as an online system, which can be freely accessible at http://ppi.bioinfo.asia.edu.tw/AtXccPPI.

Results and Discussion

The A. thaliana and Xcc Protein-Protein Interaction network

A total of 1,011 possible PPI were predicted, which include 398 *A. thaliana* proteins and 57 *Xcc* proteins, where 241 and 913 predicted PPI were derived from the domain-based and interolog approaches respectively. Figure 1 shows the *A. thaliana-Xcc* PPI networks predicted by the domain-based and interolog approaches. Ten PPI were consistently predicted by both methods. Of the consistently predicted PPI, five PPI predicted by the domain-based approach were confirmed by both the iPFam and 3DID databases. The number of consistent predicted PPI, i.e. 10, found in our work is close to that was predicted in the work of Li (Li et al.,

Table 1. List of top five *Xcc* proteins interact with *A. thaliana* proteins with large number of interaction partners.

Xcc	# *A. thaliana* interacting partner	Predicted approach	Effector protein	Biological process
Q8PAK9 (DnaK)	140	Interolog	Type III effector (predicted by ModLab)	Protein folding, response to stress
Q8PD23 (60 kDa chaperonin)	135	Interolog	N/A	Protein refolding
P63447 (Acyl carrier)	66	Domain-based	N/A	Fatty acid biosynthetic process
Q8PC59 (Elongation factor Tu-A)	48	Interolog	N/A	Protein biosynthesis
Q8PC51 (Elongation factor Tu-B)	48	Interolog	N/A	GTP catabolic process

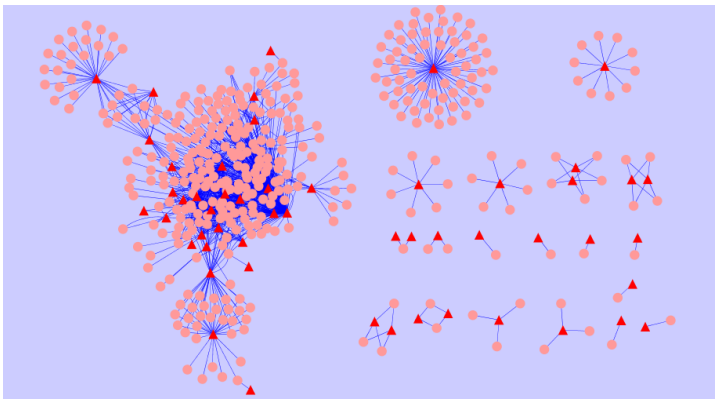

Fig 1. The *A. thaliana-Xcc* PPI network predicted by domain-based and interolog approaches, circle node indicates *A. thaliana* protein, triangle node indicates *Xcc* protein, and the edge indicates PPI.

2011), i.e. 12, which used the interolog and domain-based methods to predict the PPI between the *R. solanacearum* and *A. thaliana* proteins. On average an *Xcc* protein has about 18 *A. thaliana* interacting partners, while an *A. thaliana* protein interacts with around 2.5 *Xcc* proteins. Our result indicates that *Xcc* proteins tend to interact with many *A. thaliana* proteins. Our findings are consistent with a scenario in which a few pathogenic proteins attack the host's proteome (He et al., 2008; Li et al., 2011). In addition, the work of Stahl and Bishop (Stahl and Bishop, 2000) indicates that a pathogen mutates its genes to infect the host. However, the plant defends the attacks by expanding its gene families. Table 1 lists the top five *Xcc* proteins that interact with *A. thaliana* proteins which have a large number of interaction partners (so-called hub proteins). Our result indicates that effectors Q8PAK9 (Dnak), Q8PD23 (60 kDa chaperonin), Q8PC59 (Elongation factor Tu-A) and Q8PC51 (Elongation factor Tu-B) target AT4G20360 (ATRAB8D). It is known that ATRAB8D is involved in plants that are lack of a animal-like adaptive immunity mechanism; therefore it evolves a specific system with multiple layers against invading pathogens recorded by the Kyoto Encyclopedia of Genes and Genomes (KEGG). Furthermore, the *A. thaliana* protein AT3G13860 (heat shock protein 60-3A, HSP60-3A) is one of the targeted proteins of Q8PAK9, Q8PC59 and Q8PC51. This AT3G13860 protein is involved in the process against the pathogen when responding to cadmium ion. Besides, Q8PD23 (60 kDa chaperonin) targets PRG, AT3G48750 (CDC2) and AT1G54270 (EIF4A-2) which are also involved when responding to cadmium ion. Previous studies have suggested that metal ions are required for pathogen virulence and plant defense. Our findings support the work of Fones (Fones et al., 2010) who demonstrated that Zn, Ni or Cd are accumulated when Thlaspi caerule resists to a leaf spot caused by Pseudomonas syringae pv. maculicola (Psm). Thlaspi was grown in soil with a higher concentration of Zn, Ni and Cd. It was found that the degree of these metals' concentration is a direct variation of the intensity of the defensive mechanism of Thlaspi. In addition, there are other metal ions needed for plant defenses. For example, high levels of Fe are needed for the infection process of bacteria (Franza et al., 2005); a high Zn or Mn is important in the mandibles of seed-penetrating larvae (Morgan et al., 2002). Furthermore, our result indicates that Q8PD23, Q8PC59 and Q8PC51 target AT2G25140 (casein lytic proteinase B-M) which is the gene involved in the response to heat, light intensity and hydrogen peroxide. As shown in Table 2, a few *A. thaliana* proteins interact with more than ten *Xcc* pathogen proteins. It is found that AT4G02930 and AT4G20360 (ATRAB8D) which interacts with 15 different *Xcc* proteins are involved in the plant-pathogen interaction pathway recorded by KEGG and they are also consistent with a scenario in which an essential metal ion such as Cd is required for pathogen virulence and plant defense. Furthermore, the TAIR results indicate that the proteins AT5G02490 (ATHSP70-2) and AT5G02500 (AT-HSC70-1) are involved in the defense response to bacterium.

Xcc effector predicted by EffectiveT3 and the type III secretion system effector prediction (ModLab) system

Among 57 *Xcc* proteins involved in our predicted PPI, two (P58892 and Q8PC32) are predicted as bacterial secreted proteins by the EffectiveT3 tool, and two (Q8PBK7 and

Table 2. *A. thaliana* proteins with their *Xcc* interacting partners (more than ten *Xcc* interaction partners).

A. thaliana	# *Xcc* interaction partner	Biological processes (TAIR)	KEGG pathway
AT4G02930	15	Response to cadmium ion	Plant-pathogen interaction
AT4G20360 (ATRAB8D)	15	N/A	Plant-pathogen interaction
AT5G02490 (ATHSP70-2)	14	Defense response to bacterium, response to cadmium ion, response to virus	N/A
AT5G28540 (BIP1)	14	Response to heat	N/A
AT5G02500 (AT-HSC70-1)	14	Defense response to bacterium, defense response to fungus, response to cadmium ion, response to virus	N/A
AT5G42020 (BIP)	14	Response to cadmium ion	N/A
AT4G37910 (MTHSC70-1)	12	Response to cadmium ion	N/A
AT5G49910 (CPHSC70-2)	12	Response to cadmium ion	N/A
AT5G09590 (HSC70-5)	12	Response to cadmium ion, response to heat, response to high light intensity, response to virus	N/A
AT4G24280 (CPHSC70-1)	12	Response to cadmium ion, response to cold	N/A

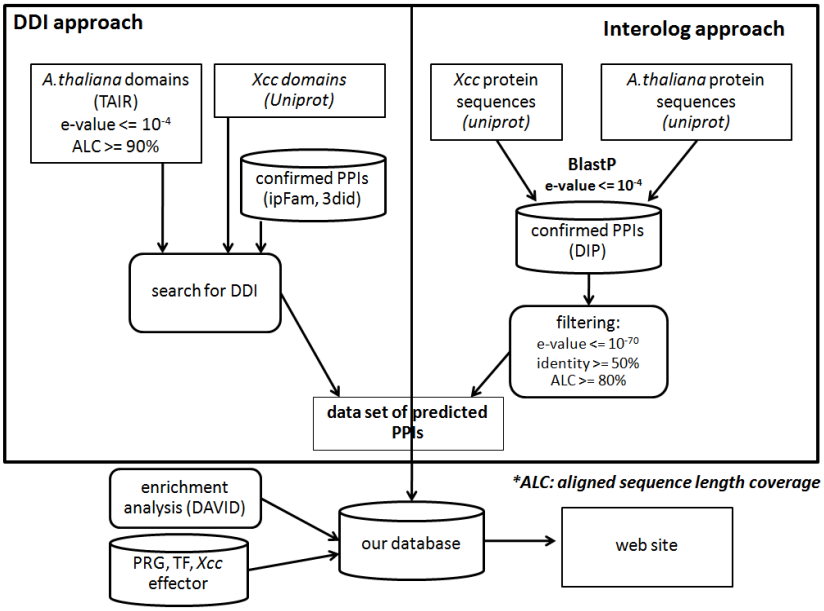

Fig 2. System flowchart of PPI prediction of *A. thaliana* and *Xcc*. Prediction is based on the (i) domain-based and, (ii) interolog approaches.

P22260) are predicted as type III effector proteins. Q8PC32 (dsbB) had been reported in the work of Jiang (Jiang et al., 2008) that a mutation in the dsbB gene can result in ineffective type II and type III secretion systems. The dsbB gene is required for the pathogenesis process of *Xcc* because it is involved in virulence, hypersensitive response and bacterial growth in plants. Besides, the work of Hsiao (Hsiao et al., 2005) demonstrated that P22260 (Clp) up-regulates the transcription of the engA gene encoding a virulence factor in *Xcc* by a direct binding to the upstream tandem Clp sites.

Another four *Xcc* proteins (Q8P7S1, P22260, Q8PAK9 and Q8P815) were predicted as type III effector proteins by the ModLab software. Interestingly, P22260 (CRP-like protein) was identified as a type III effector protein by both predictors and also it was recorded by the Universal Protein Resource (UniProt) as a pathogenesis effector protein, as it undergoes

Table 3. KEGG pathway annotations of the *A. thaliana* proteins and *Xcc* effector interaction.

Xcc	effector type	*# A. thaliana* interaction partner	KEGG pathway
P58892 (predicted by EffectiveT3)	Secretion protein	4	Metabolic pathways ***(4)**
			Cysteine and methionine metabolism ***(3)**
			Biosynthesis of secondary metabolites ***(3)**
			Phenylalanine, tyrosine and tryptophan biosynthesis ***(3)**
			Tyrosine metabolism ***(3)**
			Isoquinoline alkaloid biosynthesis ***(3)***
			Tropane, piperidine and pyridine alkaloid biosynthesis ***(3)***
			Carbon fixation in photosynthetic organisms ***(2)**
			Alanine, aspartate and glutamate metabolism ***(2)**
			Arginine and proline metabolism ***(2)**
			2-oxocarboxylic acid metabolism ***(2)**
Q8PC32 (predicted by EffectiveT3)	Secretion protein	1	Not available
Q8PBK7 (predicted by EffectiveT3)	Type III effector protein	16	Biosynthesis of secondary metabolites ***(4)**
			Metabolic pathways ***(4)**
			Glutathione metabolism ***(4)**
			Pentose phosphate pathway ***(3)**
			Tyrosine metabolism ***(1)**
			Alanine, aspartate and glutamate metabolism ***(1)**
P22260 (predicted by EffectiveT3, ModLab)	Type III effector protein	6	Hepatitis B ***(1)**
Q8P7S1 (predicted by ModLab)	Type III effector protein	10	Protein processing in endoplasmic reticulum ***(7)**
			Protein export ***(3)**
			Endocytosis ***(5)**
			Spliceosome ***(4)**
Q8PAK9 (predicted by ModLab)	Type III effector protein	20	Biosynthesis of secondary metabolites ***(6)**
			Metabolic pathways ***(6)**
			Glycolysis/gluconeogenesis ***(5)**
			Hepatitis B ***(2)**
Q8P815 (predicted by ModLab)	Type III effector protein	6	Biosynthesis of secondary metabolites ***(4)**
			Metabolic pathways ***(4)**
			Cysteine and methionine metabolism ***(4)**
			Plant-pathogen interaction ***(2)**

****(x)**, x indicates number of A. thaliana proteins involve in KEGG pathway*

A.thaliana	-all protein-	Identity(%)	>=50	aligned coverage	>=80
XCC	-all protein-	e-value	<=1E-70		Submit

Protein 1	P1 Interactor	<=PPI=>	P2 Interactor	Protein 2	P1 Identity	P2 Identity	P1 Evalue	P2 Evalue	P1 Align Cov	P2 Align Cov
AT4G02930.1	P23568	<===>	P23568	Q8PC51	64.47	65.4	4e-150	2e-151	86.3436	99.7475
AT4G02930.1	P23568	<===>	P23568	Q8PC59	64.47	65.4	4e-150	2e-151	86.3436	99.7475
AT4G02930.1	P0A6N1	<===>	P0A6F5	Q8PD23	69.85	77.65	8e-156	0	87.4449	96.5201
AT4G02930.1	P0A6N1	<===>	P0A6H5	Q8P552	69.85	66.52	8e-156	3e-165	87.4449	98.022
AT4G02930.1	P0A6N1	<===>	P03004	Q8PEH5	69.85	54.62	8e-156	3e-145	87.4449	99.5475
AT4G02930.1	P0A6N1	<===>	P33138	Q8PBY5	69.85	75.06	8e-156	0	87.4449	99.7664
AT4G02930.1	P0A6N1	<===>	P32168	Q8P552	69.85	66.52	8e-156	3e-165	87.4449	98.022

Fig 3. Output screen of predicted PPI by interolog approach; P1 Interactor: homolog protein of protein 1, P2 Interactor: homolog protein of protein 2, P1 Identify: identity degree of protein1 and its homolog protein, P2 Identity: identity degree of protein 2 and its homolog protein, P1 e-value: e-value of protein 1 and its homolog protein, P2 e-value: e-value of protein 2 and its homolog protein, P1 Align Cov: aligned sequence length coverage of protein 1 and its homolog protein, P2 Align Cov: aligned sequence length coverage of protein 2 and its homolog protein.

specific processes that generate the ability of an organism to cause disease. It is found that Q8P815 targets two *A. thaliana* proteins involved in the plant-pathogen interaction pathway recorded by KEGG. This evidence is consistent with the report of Buell (Buell 2002) indicating that the genes involved in the resistance response can be classified into three classes: (1) R genes which are involved in the recognition of the pathogen; (2) signal transduction genes; and (3) defense response genes which are involved in the suppression of pathogen development. Our predictions demonstrated that the *Xcc* effector tends to interact with *A. thaliana* proteins, which are likely to be involved in the same pathway. For instance, the type III effector protein, Q8P7S1, targets ten *A. thaliana* proteins, where seven of them are involved in the same process in the endoplasmic reticulum pathway and five of them are involved in the endocytosis pathway. More instances are presented in Table 3.

Gene ontology enrichment analysis result

The functional annotation of *Xcc* and their interacting *A. thaliana* proteins involved in the predicted PPI are given in the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang et al., 2009). The over-represented biological processes of *Xcc* proteins involved in the predicted PPI are mainly enriched in (1) the generation of precursor metabolites and energy with $1.16x10^{-4}$ as the p-value; (ii) the protein folding process with $2.82x10^{-4}$ as the p-value; and (3) the carboxylic acid biosynthetic process with $3.27x10^{-4}$ as the p-value. On the other hand, the *A. thaliana* proteins involved in the predicted PPI are mainly enriched in (1) oxidation and reduction with $7.07x10^{-26}$ as the p-value; (ii) protein folding with $8.76x10^{-24}$ as the p-value; and (iii) response to cadmium ion with $1.76x10^{-21}$ as the p-value. Previous study also reported that responding to cadmium ion is a significant strategy adopted by plant defense against pathogen (Fones et al., 2010).

Our result also demonstrated that *Xcc* proteins tend to interact with a group of *A. thaliana* proteins involved in the same biological process. For instance, (1) P63447 interacts with 60 *A. thaliana* proteins all are involved in the oxidation reduction process; and (2) the type III effector Q8PAK9 interacts with 21 proteins of *A. thaliana* proteins all are involved in the process of responding to cadmium ion (Table 4). More examples are presented in Table 4.

Clique network analysis result

The *A. thaliana* interactome set obtained from TAIR and BioGrid was analyzed by CFinder to identify PPI dense regions. A total of 116 three-community were obtained, of which eight communities contain *A. thaliana* proteins targeted by *Xcc*. These eight communities are referred to as pathogen-targeted communities in this study. The over-represented biological processes of *A. thaliana* proteins in these eight communities whose satisfied p-value cutoff was set to 0.05 are: (1) those involving photosynthesis (three out of the eight communities); (2) those responding to steroid hormone stimulus (two out of the eight communities), steroid hormone mediated signaling (two out of the eight communities), inorganic substances (two out of the eight communities), metal ion (two out of the eight communities), regulation of the cell cycle process (two out of the eight communities), brassinosteroid mediated signaling (two out of the eight communities), and the generation of the precursor metabolites and energy (two out of the eight communities). Among the eight communities, it was found that AT4G01370 (ATMPK4) exists in two communities; this protein is recorded by TAIR that it encodes a nuclear and cytoplasmically localized MAP kinase that is involved in mediating the responses to pathogens, the defense responses to bacterium, the defense response to fungus, the jasmonic acid mediated signaling pathway, the regulation of innate immune responses, the responses to bacterium, the responses to cadmium ion, the responses to cold, the responses to fungus, and the responses to jasmonic acid stimulus. The KEGG (http://www.genome.jp/kegg/) also records that AT4G01370 is involved in the plant-pathogen interaction pathway. Besides, one of the *Xcc* proteins found in the eight communities is P22260 (CRP-like protein), which was recorded by UniProt (www.uniprot.org) as a pathogenesis effector protein that undergoes specific processes generating the ability of an organism to cause disease. A 9-community, which is the highest *k* degree, was identified by Cfinder. Figure 5 shows that the 9-community is composed of four *Xcc* effectors (P22260, Q8PD23, Q8P494, Q8PAV3) and six *Xcc* interacting partners (AT3G54180, AT1G66750, AT1G18040, AT4G28980, AT1G76540, AT3G48750). The over-represented biological processes of a group of interacting proteins whose satisfied p-value cutoff was set to 0.05 are: (1) cell cycle (p-value: $1.25x10^{-9}$), (2) cell division

Table 4. Enriched biological process of the *Xcc* protein (including effector) interaction with *A. thaliana* protein.

Xcc	effector type	# *A. thaliana* interaction partner	biological process
P63447	-	60	Oxidation reduction
Q8PD23	-	24	Protein folding
Q8PAK9	Type III effector (predicted by ModLab)	21	Response to cadmium ion
Q8PC59	-	20	Response to abiotic stimulus
Q8PC51	-	13	Response to heat
Q8PAK9	Type III effector (predicted by ModLab)	7	Response to bacterium
Q8PBK7	Type III secreted (predict by EffectiveT3)	4	Oxidation reduction
Q8P815	Type III effector (predict by ModLab)	4	S-adenosylmethionine biosynthetic process

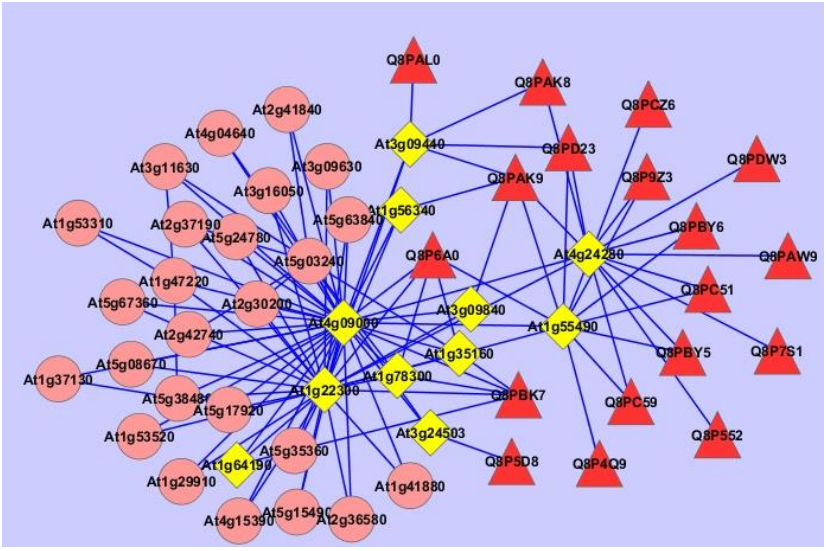

Fig 4. The 3-community PPI network; triangle: *Xcc* pathogen protein, diamond: *A. thaliana* protein interacts to *Xcc* protein, circle: *A. thaliana* protein.

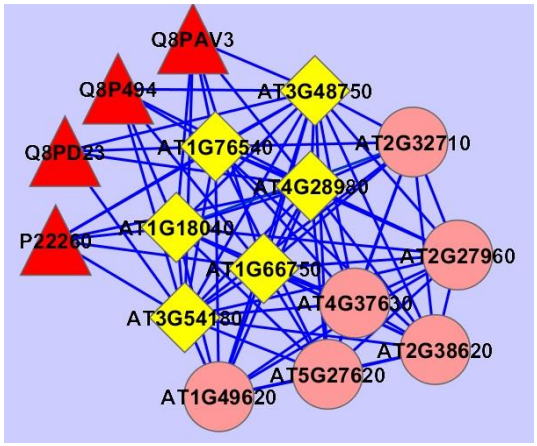

Fig 5. The 9-community PPI network; triangle: *Xcc* pathogen protein, diamond: *A. thaliana* protein interacts to *Xcc* protein, circle: *A. thaliana* protein.

(p-value: $1.04x10^{-7}$), (3) protein amino acid phosphorylation (p-value: $2.52x10^{-6}$), (4) phosphorylation (p-value: $4.19x10^{-6}$), (5) phosphate metabolic process (p-value: $1.68x10^{-6}$), and (6) phosphorus metabolic process (p-value: $6.20x10^{-6}$).

Materials and Methods

Data sources

To implement the domain-based approach, information about a collection of 46,991 PFam domains of *A. thaliana* was downloaded from TAIR (http://www.arabidopsis.org) version 10. From which we filtered out a set of PFam domains which satisfy the BLAST e-value, 10^{-4} , and the aligned sequence length coverage was selected to be 90%. A set of 304 PFam domains of *Xcc* was obtained from UniProt, the Universal Protein Resource (http://www.uniprot.org). The DDI was gathered from iPFam (http://ipfam.sanger.ac.uk) and 3DID (http://3did.irbbarcelona.org). To perform the interolog prediction, a collection of 35,386 *A. thaliana* protein sequences was obtained from TAIR, and a set of 202 *Xcc* protein sequences was gathered from UniProt. Besides, a total of 63,830 experimentally verified PPI were obtained from DIP (http://dip.doe-mbi.ucla.edu/dip/main.cgi). To build the community of interacting *A. thaliana* PPI, a collection of 7,466 experimentally confirmed *A. thaliana* interactomes was obtained from two databases: TAIR and BioGrid (http://www.thebiogrid.org). A total of 1713 putative PRGs were obtained from the PRGdb (Sanseverino et al., 2010) by using the keyword "*A. thaliana*," and the TF data were retrieved from the TRANSFAC® database (Matys et al., 2006).

A. thaliana and Xcc Protein-Protein Interaction prediction

In the present study, the protein interaction between *A. thaliana* and *Xcc* was identified using two pipelines as shown in the system flowchart (Figure 2). The potential PPI between *A. thaliana* and *Xcc* proteins were predicted using the domain-based approach and the interolog approach. Subsequently, the set of predicted PPI from both methods was subjected to enrichment analysis via DAVID (Huang et al., 2009), a web service that provides functional genomic annotations for large gene lists. In addition, information about PRG, TF and *Xcc* effector proteins was integrated into our system.

Protein-Protein Interaction prediction by the domain-based approach

The predicted PPI were identified if a protein pair contains an interacting PFam domain pairs. In other words, interspecies PPI between *A. thaliana* and *Xcc* could be inferred by known DDIs which were recorded by the iPFam and 3DID databases. With this approach three sets of input were adopted: (1) information about a set of 1,555 *A. thaliana* PFam domains, (2) information about a total of 304 *Xcc* PFam domains, and (3) a collection of 7,039 known PFam DDIs recorded by the iPFam and 3DID databases.

Protein-Protein Interaction prediction by the interolog approach

An interolog is a conserved interaction between a pair of proteins which have interacting homologs in another organism. Suppose that *A* and *B* are two different interacting proteins of one organism, and *A'* and *B'* are two different interacting proteins of another organism. Then the interaction between *A* and *B* is an interolog of the interaction between *A'* and *B'*, if all of the following conditions are held: (1) *A* is a homolog of *A'*; (2) *B* is a homolog of *B'*; (3) *A* and *B* interact, and *A'* and *B'* interact, i.e., interologs are homologous pairs of protein interactions across different organisms. We adopted BLAST for searching homolog proteins and then inferred the interacting protein pairs. A total of 35,386 *A. thaliana* protein sequences and 202 *Xcc* protein sequences were blasted against a set of 63,830 known PPI from the DIP database with the e-value cutoff, sequence identity bound and aligned sequence length coverage set to $1.0x10^{-70}$, 50% and 80% respectively. The cutoff value was carried forward from the work of Yu (Yu et al., 2004), stating that interspecies PPI can be transferred when a protein pair satisfies $1.0x10^{-70}$ of e-value or sequence identity bound satisfies 80%. Their work assessed the degree to which interologs can be reliably transferred between *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Helicobacter pylori* as a function of the sequence similarity of the corresponding interacting proteins. Moreover, the threshold adopted in this work is more strict than the work of Li (Li et al., 2011) which predicted the interspecies PPI of bacterium *R. solanacearum* and *A. thaliana* by the interolog approach (0.001 blast e-value, 30% sequence identity and 80% aligned sequence length coverage). Finally, the host-pathogen PPI were identified if a protein pair of *A. thaliana* and *Xcc* has the corresponding homologs in the DIP database.

Xcc effector protein prediction

Bacterial protein secretion is an essential key virulence mechanism of symbiotic and pathogenic bacteria. The effector proteins are transported across their cell membranes and channel these proteins into the eukaryotic host cell. A type III secretion system represents such a mechanism. To identify the *Xcc* effector protein, two efficient prediction tools were employed in our study: EffectiveT3 (http://www.effectors.org/) (Jehl et al., 2011) and the type III secretion system effector prediction system developed by the Molecular Design Laboratory (ModLab, http://gecco.org.chemie.uni-frankfurt.de/index.html) (Lower and Schneider, 2009). The EffectiveT3 system identified putative effectors by a combination of two complementary approaches: (1) independent of the transport mechanism by identifying eukaryotic like protein domains; and (2) detecting the two known types of signal peptides regardless of the presence of well-conserved protein domains. A set of *Xcc* protein sequences was submitted to EffectiveT3, and we used the default parameter setting in the prediction: (1) the organism type was set to gram-negative; (2) the classification module was set to type III effector prediction of the plant set; and (3) the cutoff default setting was set to 0.999; and (4) the domain score was set to 4.0. The type III secretion system effector predicted by ModLab is a prediction system for identifying the existence of type III secretion system (T3SS) signals in amino acid sequences. This predictor is based on known protein effectors compiled from literature and sequence databases, and it served as the training data for artificial neural networks and support vector machine classifiers. A set of *Xcc* protein sequences was input into the system, where the parameters were set to default values: (1) the prediction method was set to "neural network;" (2) the sequence truncation: N and C terminals were set to 1 and 30 respectively; and (3) the neural network threshold was set to 0.4.

Gene ontology enrichment analysis

The functional annotation of the predicted interspecies PPI was given by the Database for Annotation, Visualization and Integrated Discovery, i.e., DAVID (Huang et al., 2009), which accepts batch annotation and conducts GO term enrichment analysis. Sets of *A. thaliana* and *Xcc* proteins (genes) involved in the predicted PPI were submitted to DAVID for clustering of the annotation terms. With such the enriched biological processes related to both gene lists were obtained.

Clique clustering Protein-Protein Interaction network analysis

Clique is a complete graph, i.e., a set of elements where each pair of elements is connected. Entirely connected graphs called cliques have been found to have a high functional significance in cellular processes (Spirin and Mirny, 2003; Yeger-Lotem et al., 2004). A cluster is a densely connected area of proteins (nodes), which is a functional module. The nodes of a cluster are usually involved in similar biological processes, and protein complexes can be identified through the clustering of a PPI network (Palla et al., 2005; Jonsson et al., 2006). To investigate the functional modules in which potential pathogen-targeted *A. thaliana* proteins are involved, a set of 7,466 experimentally confirmed *A. thaliana* interactomes was submitted to the CFinder software (Adamcsek et al., 2006) based on the clique percolation clustering approach, where CFinder is a tool for finding and visualizing overlapping dense groups of nodes in a network based on the clique percolation method (CPM) (Derenyl et al., 2005, Palla et al., 2005). CPM infers large community or cluster, which is also known as percolation community or cluster. Percolation community is a maximal k-clique-connected subgraph, i.e. it is the union of all k-cliques that are k-clique-connected to a particular k-clique. Initially, the 3-community (complete sub-graph of size 3) was preliminary considered to analyze a PPI topological network. The 3-community, which contains at least one predicted *A. thaliana* protein that interacts with *Xcc*, were filtered and kept for enrichment analysis by using DAVID with the e-value cutoff set to 0.05. Communities of higher degrees were also observed in our study.

Website

A web site was set up to provide comprehensive information of our result. It comprises the predicted interspecies PPI of *A. thaliana* and *Xcc* for the study of host-pathogen interactions. This system was constructed by using PHP as a scripting language and MySQL as the database. Given an *A. thaliana* or *Xcc* protein, the system returns useful outputs such as the putative PPI, the prediction method, functional annotations of the predicted protein and 3-community PPI networks. Figure 3 depicts the output screen of a predicted PPI based on the interolog approach. Figure 4 is an example of the output *A. thanliana* and *Xcc* 3-community PPI network. Our web site can be freely accessed at: http://ppi.bioinfo.asia.edu.tw/ AtXccPPI.

Conclusions

In this study, we have identified interspecies PPI of *A. thaliana* and *Xcc* based on the domain-based and interolog approaches. Comprehensive and updated information from various resources was adopted as the input for our prediction in order to obtain reliable and up to date results. The functional annotations of both *A. thaliana* and *Xcc* proteins involved in the predicted PPI were observed. The molecular network motifs were specified by the concept of clique percolation, and the biological processes of both *A. thaliana* and *Xcc* that are involved in the network motifs were observed. The PRG and also the effector bacterial protein were focused in our study. Our result suggested that bacterial proteins could possibly reprogram the host defense mechanism. We have demonstrated that a pathogen employs five strategies to achieve this goal. First, a few *Xcc* proteins tend to interact with *A. thaliana's* hub proteins or the PRG. It is known that disruption of the hub protein may lead to the PPI network disintegration. Second, some of the *Xcc* proteins tend to interact with many *A. thaliana* proteins. This is consistent with the previous work of Stahl and Bishop (Stahl and Bishop, 2000) indicating that a pathogen mutates its genes to infect the host. Third, we found that some *Xcc* proteins target at a group of *A. thaliana* proteins involved in the response to cadmium ions, which is a significant plant biological process against pathogen. This finding is in accordance with the previous works revealing that metals ions (i.e., cadmium ion) are required for pathogen virulence and plant defenses. Fourth, our findings indicated that many *Xcc* proteins target a few *A. thaliana* proteins involved in the plant-pathogen interaction pathways such as the response to the cadmium ion pathway and the defense response to bacterium, fungus and virus pathway. Fifth, the pathogen may make use of a specific kind of protein, i.e., a type III effector protein, to reprogram the host PPI. For instance, the effector Q8P815 targets two *A. thaliana* proteins which are involved in the plant-pathogen interaction pathway. Host-pathogen interactions remain a largely unexplored area in computational biology. The present work may provide some key information about and new insights into the molecular mechanism of the plant immunity system against bacteria attacks. A web-based service based on our study has been provided to support public querying (http://ppi.bioinfo.asia.edu.tw/AtXccPPI.)

Acknowledgements

The work of Ka-Lok Ng and Nilubon Kurubanjerdjit was supported by the National Science Council of Taiwan, under the grants NSC 100-2221-E-468-013 and NSC 101-2221-E-468-027. The work of Ka-Lok Ng, Chen-Yu Sheu and Jeffrey J.P Tsai was supported by the grants NSC 99-2632-E-468-001-MY3. Our gratitude goes to Dr. Timothy Williams, Asia University, for his help in proof reading the manuscript.

References

Adamcsek B, Palla G, Farkas IJ, Derenyi I and Vicsek T (2006) CFinder: locating cliques and overlapping modules in biological networks. BMC Bioinformatics. 22:1021-1023

Arifuzzaman et al (2006) Large-scale identification of protein-protein interaction of *escherichia coli* k-12. Genome Res. 16(5):686-691

Berggard T, Linse S and James P (2007) Methods for the detection and analysis of protein-protein interactions. Proteomics. 7(16):2833-2842

Bishop JG, Dean AM and Mitchell-Olds T (2000) Rapid evolution in plant chitinases: molecular targets of section in plant-pathogen coevolution. P Natl Acad Sci USA. 97(10):5322-5327

Browne F, Zhang HR, Wang HY and Azuaje F (2010) From experimental approaches to computational techniques: a review on the prediction of protein-protein interaction. Adv Art Int. 2010(924529)

Buell CR (2002) Interactions between *xanthomonas* species and *arabidopsis thaliana*. Arabidopsis Book. 1:e0031

Casadevall A and Pirofski LA (2000) Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease. Infect Immun. 68(12): 6511-6518

Derenyi et al (2005) Clique percolation in random networks. Phys Rev Lett. 94, 160–202.

Finn RD, Marshall M and Bateman A (2005) iPfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. BMC Bioinformatics. 21(3):410-412

Flor HH (1971) Current status of the gene-for-gene concept. Annu Rev Phytopathol. 9:275-296

Fones H et al (2010) Metal hyperaccumulation armors plants against disease. PLoS Pathog. 6(9): p1

Franza T, Mahe B and Expert D (2005) Erwinia chrysanthemi requires a second iron transport route dependent of the siderophore achrophore achromobactin for extracellular growth and plant infection. Mol Microbiol. 55: 261-275

Gallone G, Simpson TI, Armstrong JD and Jarman AP (2011) Bio: homology interologwalk-a perl modole to build putative protein-protein interaction networks through interolog mapping. BMC Bioinformatics. 12:289

He F, Zhang Y, Chen H, Zhang Z, and Peng YL (2008) The prediction of protein-protein interaction networks in rice blast fungus. BMC Genomics. 9:519

Hsiao YM et al (2005) Clp upregulates transcription of *engA* gene encoding a virulence factor in xanthomonas campestris by direct binding to the upstream tandem Clp sites. Febs Lett. 579: 3525-3533

Huang DW, Sherman BT and Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4:44–57

Ito T et al (2000) Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci USA. 97(3):1143-1147

Ito T et al (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA. 98(8):4569-4574

Jansen R et al (2003) A bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 302:449-453

Jehl MA, Arnold R and Rattei T (2011) Effective-a database of predicted secreted bacterial proteins. Nucleic Acids Res. 39:D591-595

Jiang BL et al (2008) DsbB is required for the pathogenesis process of xantomonas campestris pv. campestris. Mol Plant Microbe In. 21(8): 1036-45

Jonsson P, Cavanna T, Zicha D and Bates P (2006) Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics. 7:2

Kim JG et al (2008) Predicting the interactome of *xanthomonas oryzae pathovar oryzae* for target selection and db service. BMC Bioinformatics. 9:41

Li ZG, He F, Zhang Z and Peng YL (2011) Prediction of protein-protein interactions between *ralstonia solanacearum* and *arabidopsis thaliana*. Amino Acids. 42(6):2363-2371

Lin N, Wu B, Jansen R, Gerstein M and Zhao H (2004) Information asssesment on predicting protein-protein interactions. BMC Bioinformatics. 5: 154

Lower M and Schneider G (2009) Prediction of type III secretion signals in genomes of gram-negative bacteria. PLoS One. 4(6):1

Mandoli DF and Olmstead R (2000) The importance of emerging model systems in plant biology. J Plant Growth Regul. 19(3):249-252

Matys V et al (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34:D108-D110

Meyer D, Lauber E, Roby D, Arlat M and Kroj T (2005) Optimization of pathogenicity assays to study the *arabidopsis thaliana–xanthomonas campestris pv. campestris* pathosystem. Mol Plant Pathol. 6(3):327-333

Morgan TD, Baker P, Kramer KJ, Basibuyuk HH and Quicke DLJ (2002) Metals in mandibles of stored product insects: do zinc and manganese enhance the ability of larvae to infest seeds?. J Stored Prod Res. 39:65-75

Ng SK, Zhang Z and Tan SH (2003) Integrative approach for computationally inferring protein domain interactions. BMC Bioinformatics. 19(8):923-929

Ogmen U, Keskin O, Aytuna AS, Nussinov R and Gursoy A (2005) Prism: protein interactions by structural matching. Nucleic Acids Res. 33:W331-W336

Palla G, Derenyi I, Farkas I and Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature. 435:814-818

Pinzon A, Rodriguez-R LM, Gonzalez A, Bernal A and Restrepo S (2010) Targeted metabolic reconstruction: a novel approach for the characterization of plant-pathogen interactions. Brief Bioinform. 12(2): 151-162

Postnikova OA, Minakova NY, Boutanaev AM and Nemchinov LG (2011) Clustering of pathogen-response genes in the genome of *arabidopsis thaliana*. J Integr Plant Biol. 53:824-834

Rhee SY et al (2003) The arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to *arabidopsis* biology, research materials and community. Nucleic Acids Res. 31(1):224-228

Rhodes DR et al (2005) Probabilistic model of the human protein-protein interaction network. Nat Biotechnol. 23(8):951-959

Sanseverino W et al (2010) PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic Acids Res. 38:D814–D821

Spirin V and Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA. 100:12123-12128

Stahl EA and Bishop JG (2000) Plant-pathogen arms races at the molecular level. Curr Opin Plant Biol. 3(4):299-304

Stark C et al (2005) BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34: D535-D539

Stein A, Ceol A and Aloy P (2011) 3DID: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 39:D718-D723

Tsao TH, Chan CH, Huang Chi-Yang F and Lee SA (2011) Systems and computational biology-molecular and cellular experimental systems. In: Prof.Ning-Sun Yang (ed) The prediction and Analysis of Inter- and Intra-Species Protein-Protein Interaction, ISBN: 978-953-307-280-7. InTech, China.

Tsuji J and Somerville SC (1988) *Xanthomonas campestris pv. campestris* induced chlorosis in *Arabidopsis thaliana*. Arabidopsis Information Service. 26:1-8

Tsuji J, Somerville SC and Hammerschmidt R (1991) Identification of a gene in *arabidopsis thaliana* that controls resistance to *xanthomonas campestris pv. campestris*. Physiol Mol Plant P. 38:57-65

Tsuji J and Somerville SC (1992) First report of the natural infection of *arabidopsis thaliana* by *xanthomonas campestris pv. campestris*. Plant Dis. 76:539

Uetz P et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature. 403(6770):623-627

Von-Mering C et al (2007) String7-recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. 35:D358-362

Wang F et al (2012) Prediction and characterization of protein-protein interaction networks in swine. Proteome Sci. 10:2

Wojcik J and Schachter V (2001) Protein-protein interaction map inference using interacting domain profile pairs. BMC Bioinformatics. 17:S296-S305

Wu X, Shu L, Guo J, Zhang DY and Lin K (2006) Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations. Nucleic Acids Res. 34(7):2137-2150

Xenarios I et al (2000) DIP: the database of interacting proteins. Nucleic Acids Res. 28(1):289-291

Yeger-Lotem E et al (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. Proc Natl Acad Sci USA. 101:5934-5939

Yu H et al (2004) Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. Genome Res. 14(6):1107-1118