The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Motif based identification of pathogenic patterns for filamentous fungi

Xing-Ming Zhao[1] Weihua Tang[2] Luonan Chen[1,3]

[1]Institute of Systems Biology,
Shanghai University, Shanghai 200444, China
[2]Institute of Plant Physiology and Ecology,
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences,
300 Fenglin Road, Shanghai 200032, China
[3]Department of Electrical Engineering and Electronics,
Osaka Sangyo University, Osaka 574-8530, Japan

Abstract Identifying the unique pathogenic patterns underlying the phytopathogenic fungi is important for plant disease control. This paper presents a novel method for identifying the specific conserved patterns for the pathogenic species. By employing a feature selection technique, the specific patterns that best discriminate the pathogenic species from non-pathogenic ones are identified, which are assumed to be the potential pathogenic patterns that cause pathogenesis. In particular, the specific patterns for *Fusarium graminearum* (*F. graminearum*) were found, which can discriminate *F. graminearum* from other pathogenic species. The identified patterns covered a set of pathogenic genes that have been determined experimentally, which demonstrates the effectiveness of the proposed method.

Keywords Feature selection; Motif; Pathogenic pattern; Phytopathogenic fungi

1 Introduction

Phytopathogenic fungi are responsible for many of plant diseases by invading the plant host. For example, *Fusarium graminearum* (*Gibberella zeae*), is the most common causal agent of Fusarium head blight of wheat and barley, and Fusarium stalk rot of maize [7]. It is estimated that *F. graminearum* causes economical losses of $3 billion in the US between 1991 and 1996 [15]. Another well known phytopathogenic fungus is the rice blast fungus *Magnaporthe oryzae* (previously known as *Magnaporthe grisea*), which is the most destructive pathogen of rice worldwide [3]. Therefore, it is necessary to investigate the mechanism underlying the pathogenetic process and the conserved specific patterns among the phytopathogenic fungi.

Recently, a large number of fungi genomes have been sequenced, including pathogenic and non-pathogenic fungi. The genomic sequences can provide insight into the pathogenic patterns, which in turn help to understand the evolution and function of the pathogenic genes. In literature, phylogenic tree has been constructed based on sequenced fungi genomes [5] to study the evolution of pathogenesis [19]. It has been shown that pathogenic fungi are not necessarily close to each other in the phylogeny tree. Instead, some pathogenic fungi spread throughout all taxonomic groups, and show close evolutionary relationship to non-pathogenic species [11]. Recently, comparative genome analysis has shown that the expressed gene inventories of pathogenic fungi are not more similar to each other than to the non-pathogenic fungi [20]. In other words, it is impossible to find out the unique genes that only belong to the pathogenic species rather than non-pathogenic species, which make it difficult to identify individual genes that are responsible for infecting plants and cause diseases. In this paper, we present a new motif based method for identifying the conserved specific patterns underlying the filamentous phytopathogenic fungi. Especially, the specific patterns identified for *F. graminearum* covered pathogenic genes that have been determined experimentally, which demonstrates the effectiveness of the proposed method.

2 Methods

With the translated amino acid sequences available for the target genomes, the motifs for each sequence are found from the Prosite database [10]. After getting the motifs for the sequences, gene S is expressed as a vector:

$$S = [M_1, M_2, \dots, M_n]; \tag{1}$$

where, n is the total number of motifs for all the genomes studied in this work, M_i is the ith motif and $M_i = 1$ if the ith motif occurs in the gene S, otherwise $M_i = 0$. In this way, each gene is expressed as a vector based on its motif component(s).

After getting the gene vectors for the pathogenic and non-pathogenic fungi, we aim to find out the specific features (i.e. motifs in this case) that best discriminate the pathogenic from non-pathogenic species, which can be formalized as a feature selection problem, where each gene vector is a sample and its motif components are features. In this work, the t-test is utilized to identify those features that are distributed differentially between the two class of species (i.e. pathogenic versus non-pathogenic species), where the identified motifs are assumed to be the conserved patterns underlying the pathogenic species. Let x denote the set of pathogenic genes and y the set of non-pathogenic genes, the score for each feature $i (i = 1, \dots, n)$ can be defined as:

$$t_i = \frac{\bar{x}_i - \bar{y}_i}{s_{xi} + s_{yi}} \tag{2}$$

where $\bar{x}_i$ is the sample mean for the ith feature, s_{xi} is the sample standard deviation for the ith feature, and the same for y. The features are ranked based on their scores, and the differentially distributed features are assumed to be the ones that discriminate the two classes of species.

3 Results

3.1 The conserved patterns underlying the pathogenic fungi

To find out the conserved patterns underlying the pathogenic fungi, we identify the features that best discriminate the pathogenic from non-pathogenic fungi. In this work, we chose three pathogenic fungi families and three non-pathogenic fungi families according to the fungi phylogenetic tree constructed in literature [5]. Table 1 lists the genomic

Table 1: The fungi genomic

	Genomes	Number of proteins
Pathogenic familes	*Fusarium graminearum*	13295
	Fusarium verticillioides	14116
	Magnaporthe grisea	12743
non-Pathogenic familes	*Neurospora crassa*	9791
	Podospora anserina	10573
	Trichoderma reesi	9117

families used in this work. The selected six fungi families belong to the same superfamily in the phelogenetic tree, and thereby the specific patterns discriminating pathogenic families from non-pathogenic ones are possibly the real conserved patterns underlying the pathogenic fungi.

With t-test, some motifs were found distributed differentially in the pathogenic and non-pathogenic families. Especially, those motifs that occurred more frequently in pathogenic families than non-pathogenic families were obtained because these motifs are more possibly the specific patterns that the pathogenic species have. Figure 1 shows the distribution of these motifs in the pathogenic and non-pathogenic species. It can be easily seen from the figure that these motifs occurred more frequently in pathogenic species than non-pathogenic species. Therefore, the identified motifs are more possibly the conserved patterns among pathogenic species and thereby provide insight into the evolution of the pathogenic species.

Table 2 shows the motifs that occurred more frequently in the pathogenic species than non-pathogenic species, where the annotations of the motifs were obtained from the Gene Ontology (GO) database [1], and the “prosite2go" [2] was used to link the GO terms and Prosite motifs. The descriptions were obtained from the Prosite database [10] for the motifs without GO annotation. We can see from the annotations of the identified motifs that the identified patterns have the functions of amino acid metabolism, binding and membrane, which may be responsible for the interactions between the pathogenic fungi and plants. Note that some of the identified patterns are related to pathogenesis while some may be just the essential conserved patterns for the pathogenic species.

3.2 Identification of pathogenic patterns for *Fusarium graminearum*

In this part, our proposed method was used to identify the pathogenic patterns for *F. graminearum*. To find out the specific patterns for *F. graminearum* rather than the other two pathogenic fungi, the samples of *F. graminearum* were compared against the ones of the other two pathogenic fungi species. Consequently, eight motifs were found that occurred more frequently in *F. graminearum* than the other two pathogenic families.

Table 3 shows the identified motifs that discriminate *F. graminearum* from the other two pathogenic fungi, where the patterns are assumed to be the specific pathogenic patterns for *F. graminearum*. From the results, we can see that the identified patterns have the function of phosphorylation site, binding site and transcription regulation, which are related to the pathogenesis while the fungus infect the plant. In particular, the identified motifs covered some pathogenic genes that have been experimentally determined

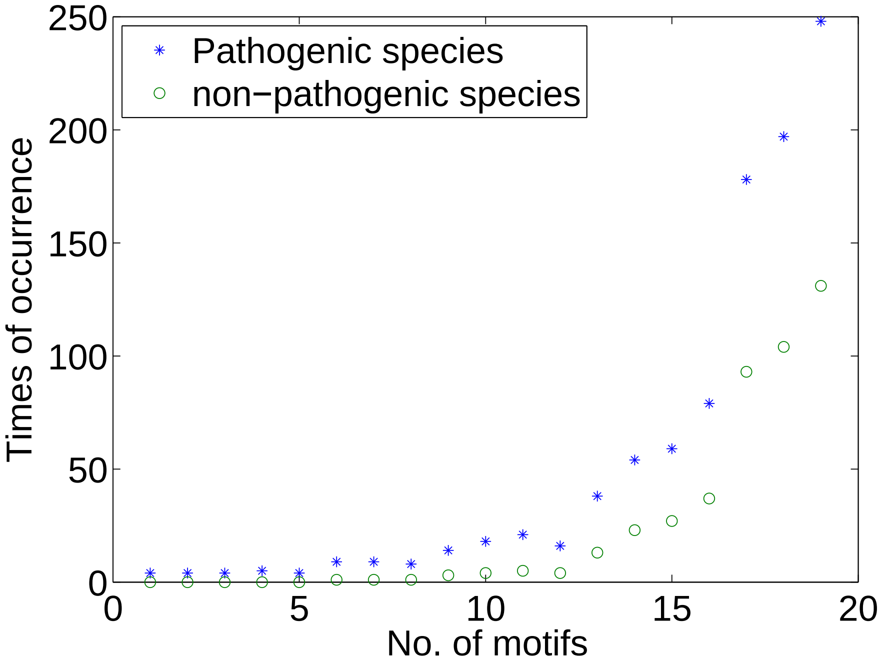

Figure 1: The distribution of the specific motifs that occurred more frequently in pathogenic species than non-pathogenic species

(found in Table 3). The known pathogenic genes make it clear that the patterns PS00001, PS00004, PS00005, PS00006, PS00007, and PS00008 are really related to pathogenesis. The consistency with published results demonstrate the effectiveness of the proposed method.

4 Conclusions

Identifying the specific and conserved patterns underlying the pathogenic fungi can provide insight into the evolution of pathogenesis. In this paper, we present a new method for identifying the specific patterns underlying the phytopathogenic fungi. Utilizing a feature selection technique, the patterns that best discriminate the pathogenic from non-pathogenic fungi were obtained. In particular, the specific patterns of *F. graminearum* were found, and these patterns covered some known pathogenic genes that have been determined in literature, which demonstrate the effectiveness of the proposed method. Note that some of the identified motifs are pathogenic patterns while some may be just the essential factors for the biological process. In this work, only a small set of pathogenic and non-pathogenic fungi in the same superfamily were analyzed, which may reflect some bias toward certain species. More fungi families will be taken into account in the future. In addition, the distribution of the fungi will also be considered in order to avoid the bias for certain fungus.

Table 2: The motifs that occurred more frequently in pathogenic families than non-pathogenic families.

Motif	GO annotation & descriptions from Prosite
PS00917	amino acid metabolic process
PS00609	hydrolase activity, hydrolyzing O-glycosyl compounds; carbohydrate metabolic process
PS00977	glycerol-3-phosphate dehydrogenase activity; glycerol-3-phosphate metabolic process; glycerol-3-phosphate dehydrogenase complex
PS00570	Bacterial ring hydroxylating dioxygenases alpha-subunit signature
PS01090	endodeoxyribonuclease activity, producing 5'-phosphomonoesters
PS00560	serine carboxypeptidase activity; proteolysis
PS00214	transporter activity; lipid binding; transport
PS00144	amino acid metabolic process
PS00589	sugar:hydrogen symporter activity; phosphoenolpyruvate-dependent sugar phosphotransferase system
PS01164	copper ion binding;amine oxidase activity; quinone binding; amine metabolic process
PS00026	chitin binding;
PS00155	cutinase activity; extracellular region
PS00775	hydrolase activity, hydrolyzing O-glycosyl compounds; carbohydrate metabolic process
PS00137	subtilase activity; proteolysis
PS00122	Carboxylesterases type-B signatures
PS00430	TonB-dependent receptor proteins signatures
PS00216	transporter activity; transport; membrane
PS00217	transporter activity; transport;membrane
PS00086	monooxygenase activity; iron ion binding electron carrier activity; heme binding

Acknowledgments.

This work was partly supported by the National High Technology Research and Development Program of China (2006AA02Z309), and JSPS-NSFC collaboration project.

References

[1] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT. Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, J. E Richardson, M Ringwald, GM Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *NatGenet*, 25:25–29, 2000.

[2] E Camon, M Magrane, D Barrell, D Binns, W Fleischmann, P Kersey, N Mulder, T Oinn, J Maslen, A Cox, and R Apweiler. The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, 13(4):662–672, 2003.

Table 3: Pathogenic patterns for *F. graminearum*

Motif	GO annotation& description from Prosite	Pathogenic genes in literature
PS00463	transcription factor activity;zinc ion binding; regulation of transcription, DNA dependent; nucleus	
PS00007	Tyrosine kinase phosphorylation site	FGSG_10313 [8] FGSG_10464 [6] FGSG_02395 [13] FGSG_08208 [6] FGSG_10548 [6] FGSG_01790 [6] FGSG_08795 [6] FGSG_06631 [14] FGSG_03537 [16] [17] [4]
PS00017	ATP/GTP-binding site motif A (P-loop)	FGSG_10464 [6] FGSG_08795 [6]
PS00001	N-glycosylation site	FGSG_10313 [8] FGSG_06385 [12] FGSG_02395 [13] FGSG_02397 [13] FGSG_08208 [6] FGSG_10548 [6] FGSG_10464 [6] FGSG_01790 [6] FGSG_08795 [6] FGSG_06631 [14] FGSG_05906 [21] [9] FGSG_01665 [18] FGSG_03537 [16] [17] [4]
PS00004	cAMP- and cGMP-dependent protein kinase phosphorylation site	FGSG_10313 [8] FGSG_06385 [12] FGSG_08208 [6] FGSG_10548 [6] FGSG_10464 [6] FGSG_01790 [6] FGSG_08795 [6] FGSG_06631 [14] FGSG_05906 [21] [9] FGSG_01665 [18]
PS00006	Casein kinase II phosphorylation site	FGSG_10313 [8] FGSG_06385 [12] FGSG_02395 [13] FGSG_02397 [13] FGSG_08208 [6] FGSG_10548 [6] FGSG_10464 [6] FGSG_01790 [6] FGSG_08795 [6] FGSG_06631 [14] FGSG_05906 [21] [9] FGSG_01665 [18] FGSG_03537 [16] [17] [4]
PS00005	Protein kinase C phosphorylation site	FGSG_10313 [8] FGSG_06385 [12] FGSG_02395 [13] FGSG_02397 [13] FGSG_08208 [6] FGSG_10548 [6] FGSG_10464 [6] FGSG_01790 [6] FGSG_08795 [6] FGSG_06631 [14] FGSG_05906 [21] [9] FGSG_01665 [18] FGSG_03537 [16] [17] [4]
PS00008	N-myristoylation site	FGSG_10313 [8] FGSG_06385 [12] FGSG_02395 [13] FGSG_02397 [13] FGSG_08208 [6] FGSG_10548 [6] FGSG_10464 [6] FGSG_01790 [6] FGSG_08795 [6] FGSG_06631 [14] FGSG_05906 [21] [9] FGSG_01665 [18] FGSG_03537 [16] [17] [4]

[3] BC Couch and LM Kohn. A multilocus gene genealogy concordant with host preference indicates segregation of a new species, magnaporthe oryzae, from M. grisea. *Mycologia*, 94(4):683–693, 2002.

[4] A Desjardins, G Bai, RD Plattner, and RH Proctor. Analysis of aberrant virulence of Gibberella zeae following transformation-mediated complementation of a trichothecene-deficient (Tri5) mutant. *Microbiology*, 146(8):2059–2068, 2000.

[5] D Fitzpatrick, M Logue, J Stajich, and G Butler. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology*, 6(1):99, 2006.

[6] I Gaffoor, D Brown, R Plattner, R Proctor, W Qi, and F Trail. Functional analysis of the polyketide synthase genes in the filamentous fungus gibberella zeae (anamorph fusarium graminearum). *Eukaryotic Cell*, 4(11):1926–1933, 2005.

[7] RS Goswami and HC Kistler. Heading for disaster: Fusarium graminearum on cereal crops. *Molecular Plant Pathology*, 5(6):515–525, 2004.

[8] Z Hou, C Xue, Y Peng, T Katan, HC Kistler, and JR Xu. A mitogen-activated protein kinase gene (mgv1) in fusarium graminearum is required for female fertility, heterokaryon formation, and plant infection. *Molecular Plant-Microbe Interactions*, 15(11):1119–1127, 2002.

[9] G Hu, A deHart, Y Li, C Ustach, V Handley, R Navarre, C Hwang, B Aegerter, V Williamson, and B Baker. EDS1 in tomato is required for resistance mediated by TIR-class R genes and the receptor-like R gene Ve. *The Plant Journal*, 42(3):376–391, 2005.

[10] N Hulo, A Bairoch, V Bulliard, L Cerutti, B Cuche, E Castro, C Lachaize, P Langendijk-Genevaux, and C Sigrist. The 20 years of PROSITE. *Nucl. Acids Res.*, page gkm977, 2007.

[11] TY James, F Kauff, CL Schoch, PB Matheny, V Hofstetter, CJ Cox, G Celio, C Gueidan, E Fraker, J Miadlikowska, HT Lumbsch, A Rauhut, V Reeb, AE Arnold, A Amtoft, JE Stajich, K Hosaka, GH Sung, D Johnson, B O'Rourke, M Crockett, M Binder, JM Curtis, JC Slot, Z Wang, AW Wilson, A Schussler, JE Longcore, K O'Donnell, S Mozley-Standridge, D Porter, PM Letcher, MJ Powell, JW Taylor, MM White, GW Griffith, DR Davies, RA Humber, JB Morton, J Sugiyama, AY Rossman, JD Rogers, DH Pfister, D Hewitt, K Hansen, S Hambleton, RA Shoemaker, J Kohlmeyer, B Volkmann-Kohlmeyer, RA Spotts, M Serdani, PW Crous, KW Hughes, K Matsuura, E Langer, G Langer, WA Untereiner, R Lucking, B Budel, DM Geiser, A Aptroot, P Diederich, I Schmitt, M Schultz, R Yahr, DS Hibbett, F Lutzoni, DJ McLaughlin, JW Spatafora, and R Vilgalys. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, 443(7113):818–822, 2006.

[12] NJ Jenczmionka, FJ Maier, AP Lösch, and W Schöfer. Mating, conidiation and pathogenicity of Fusarium graminearum, the main causal agent of the head-blight disease of wheat, are regulated by the MAP kinase gpmk1. *Curr Genet.*, 43(2):87–95, 2003.

[13] Y Kim, Y Lee, J Jin, K Han, H Kim, J Kim, T Lee, S Yun, and Y Lee. Two different polyketide synthase genes are required for synthesis of zearalenone in Gibberella zeae. *Molecular Microbiology*, 58(4):1102–1113, 2005.

[14] S Lu, S Kroken, B Lee, B Robbertse, A Churchill, O Yoder, and B Turgeon. A novel class of gene controlling virulence in plant pathogenic ascomycete fungi. *Proceedings of the National Academy of Sciences*, 100(10):5980–5985, 2003.

[15] FG Priest and I Campbell, editors. *Brewing Microbiology*, volume 3. Springer, 2002.

[16] RH Proctor, TM Hohn, and SP McCormick. Reduced virulence of Gibberella zeae caused by disruption of a trichothecene toxin biosynthetic gene. *Molecular Plant-Microbe Interactions*, 8(4):593–601, 1995.

[17] RH Proctor, TM Hohn, and SP McCormick. Restoration of wild-type virulence to Tri5 disruption mutants of Gibberella zeae via gene reversion and mutant complementation. *Microbiology*, 143(8):2583–2591, 1997.

[18] W Shim, U Sagaram, Y Choi, J So, H Wilkinson, and Y Lee. FSR1 is essential for virulence and female fertility in Fusarium verticillioides and F. graminearum. *Molecular Plant-Microbe Interactions*, 19(7):725–733, 2006.

[19] DM Soanes, TA Richards, and NJ Talbot. Insights from sequencing fungal and oomycete genomes: What can we learn about plant disease and the evolution of pathogenicity? *Plant Cell*, 19(11):3318–3326, 2007.

[20] DM Soanes and NJ Talbot. Comparative genomic analysis of phytopathogenic fungi using expressed sequence tag (EST) collections. *Molecular Plant Pathology*, 7(1):61–70, 2006.

[21] C Voigt, W Schafer, and S Salomon. A secreted lipase of fusarium graminearum is a virulence factor required for infection of cereals. *The Plant Journal*, 42(3):364–375, 2005.