Bioinformation Volume 5

open access

www.bioinformation.net Issue 5 **Hypothesis**

Fungome: Annotating proteins implicated in fungal pathogenesis

Ranganath Gudimella[1], Sivaramaiah Nallapeta[1], Pritish Varadwaj[2], Prashanth Suravajhala[3,*]

[1] Onan Biotech, Malkajgiri, Secunderabad 500047, AP, India; [2]Bioinformatics division, Indian Institute of Information Technology, Allahabad 211012, UP, India; [3]Department of Science, Systems and Models, Roskilde University, Universitetsvej 1, 4000 Roskilde, Denmark; Prashanth Suravajhala - Email: prash@bioclues.org [*]Corresponding author.

Received June 29, 2010; Accepted August 25, 2010; Published November 1, 2010

Abstract:
Sequencing genomes of different pathogenic fungi produced plethora of genetic information. This “omics” data might be of great interest to probe strain diversity, identify virulence factors and complementary genes in other fungal species, and importantly in predicting the role of proteins specific to pathogenesis in humans. We propose a component called “fungome” for those fungal proteins implicated in pathogenesis which, we believe, will allow researchers to improve the annotation of fungal proteins.

Keywords: Fungome, omics, pathogenesis

Introduction:
Fundamental biological processes can now be studied by applying the full range of omics technologies viz genomics, transcriptomics, proteomics, metabolomics. Till the time a biological sample is prepared for use in a specific omics assay, its description is inherently technology independent. Wide array of assays including high-throughput methods like MS/MS, Yeast two hybrids and pull down assays are preferentially used to navigate them to applications. However, an accurate analysis of these high throughput methods to remove redundancy and false-positive data still remains a challenge. A pre-requisite to decrease redundancy of data bringing a significant degree of harmonization across omics data is critical to understand if the data comes from different sources **[13]**. In the recent past, global analysis of proteins in fungal organisms, has contributed greatly to our understanding of gene function. Several bioinformatics applications characteristic for fungal proteomics research is needed to overcome these challenges especially fungal proteins specific to human pathogenesis **[22]**. Though, fungi are very useful in the study of genetics, they can also be used as model systems for studying higher and more complex organisms. Furthermore, their easy growth conditions and simple nutritional requirements make them model systems for complex studies in the biology. Hence there is a need to understand and regularize the concepts of molecular fungal pathogenesis in correlation with complex organisms. A standard example can be derived from the experiments of Beadle and Tautum who isolated a number of mutants of Neurospora crassa **[2]**. The infectious fungi prove to be a serious problem as different modes of entering into the body defines long lasting and detrimental effects to human health. Several molecular studies of virulence in pathogenic fungi reveal a complex interaction between each fungal species and the human host while it is known that the fungi that cause invasive disease differ considerably in their inherent pathogenicity **[15]**. Certain fungal virulence factors reveal increase or decrease of expression of specific gene with the host pathogenic studies. The fungal genomic data pouring into the web repositories are to be complemented with the functomic studies while for expression profiling experiments, most bioinformatics tools and resources that have been applied in functional genomics studies in S. cerevisiae can be applied to other fungal pathogens. However, more advanced functional genomics tools involving high-throughput data **[1]** are important as microarrays are used to identify transcription factors and regulatory elements. With most of the proteomic studies in pathogenic fungi have been limited to 2-D gel analysis and mass spectrometry; powerful proteomics methods like MALDI-TOF have been developed for genome wide analyses of protein expression, protein localization, and protein-protein interactions in fungi.

Although human pathogenic fungal genome sequences of Aspergillus fumigatus **[14]** Candida glabrata; Cryptococcus neoformans; Encephalitozoon cuniculi **[12, 6]** are available, new sequence information is generated from time to time has increased the gap between the existing information and knowledge about the protein function in fungal organisms. A variety of functionality analyses including phylogenetic profiling has allowed the researchers to annotate the fungal genomes. Furthermore, the profiling suggests an understanding besides validating high-confidence predictions for interacting pairs in the genome. Today, the primary techniques employed for the identification of peptides and proteins from biological sources is tandem mass spectrometry (MS/MS) in the form of whole-protein analysis (“top-down" proteomics) or analysis of enzymatically produced peptides ("bottom-up" approach). Recently, the complete sequencing of the Magnaporthe oryzae, a causal agent of blast disease of rice has shown that multi locus genes are concordant with host preference indicating segregation of a new species of Magnaporthe oryzae from Magnoporthe grisea **[5]**. This example implicates us to comprehend human pathogenic fungi having common virulence factors that point toward horizontal transfer of gene clusters with host preference. Although the Gene Ontology (GO) annotation provides valuable means for identifying such proteins or assigning functionality, the functional assignments are to be cross-validated with manually reviewed data, conserved domains, or data determined by wet lab experiments. Furthermore, as novel genetics and genomics-based strategies are gaining importance to speed up discovery of novel drug targets **[4, 5]**, we envisage retrieving information containing fungal proteins involved in pathogenesis allows us to understand the pathological manifestation of various fungal pathogens that take place before or after the entry into animal or human body. Two methods like yeast two hybrid and affinity purification of complex by MS approach will complement the interactome approach **[3]**. The yeast two hybrids (Y2H) is the most powerful tool in systems biology to understand the complexities of proteins and their interactions which cause a metabolic flux to accelerate the pathways. Now Y2H provides cellular proteome for screening transcriptional active proteins which are localized in different sub-cellular systems. There are many enzymes with multiple catalytic activities being identified, hypothesized and

Bioinformation Volume 5

open access

www.bioinformation.net Issue 5 **Hypothesis**

characterized in various fungal genomes. This will allow us to find human pathogenic fungal proteins with combinations of protein motifs that might be present or absent in other plant, animal or fungal genomes.

Methods:

Similarity searches

A defined set of pathogenic fungi (Figure 1) were collected and all these genomes were examined by bidirectional similarity searches. Most virulence influenced genes like DNA repair, metabolites, cancer genes, GST (glutathione S-transferase), cell wall biogenesis, vegetative growth and sporulation were chosen for bidirectional best hits (Table 1 See Supplementary material). The homology was cross checked with reference genome dataset of Neurospora crassa OR74A. All these proteins were subsequently predicted for sub-cellular location using Ptarget [8]. The Neurospora crassa gene and transcripts have been downloaded from http://www.broadinstitute.org/annotation/genome/neurospora/Home.html.

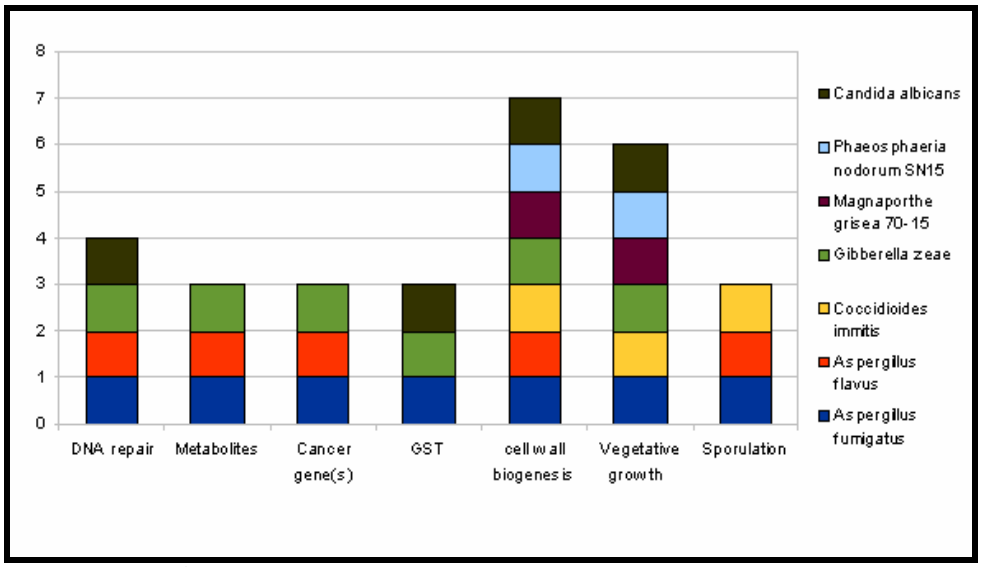

Figure 1: Pathogenic fungi with important proteins.

Interolog mapping

Fasta Pearson version 2.5 **[17]** was used to find the similarities between the Uniprot(www.uniprot.org) accessions across the most selective fungal organisms of taxonomical interest and those containing the important proteins and the String database **[11]** containing all fasta sequences. The similar sequences that match the e value less than 0 were considered for further protein-protein interaction mapping.

Prediction of protein interactions

Protein-protein interactions (PPI) studies for different proteins specific to few fungal families have been done. We used a tool viz DIMA **[16]**, a Domain Interaction MAp that finds functional and physical interactions among conserved protein-domains. All the accessions were queried using the list of protein family (Pfam) identifiers. The integration of evidence from different sources was carried out by the server which involves analyses using the domain phylogenetic profiling, domain-pair exclusion method for predicting domain interactions from experimentally demonstrated protein-protein interactions using IntAct **[20]**, STRING **[11]** and domain contacts from crystal structures using iPFAM **[7]**.

Results and Discussion:

We found an approximate 30% of the predicted genes sequenced till date has no significant homologs in other organisms (Non redundant search). Therefore, it was a daunting task to identify sets of fungal-specific genes that were associated with human carcinogenesis. Leaving aside the vast genomic data available for non-human pathogenic fungi the data for pathogenic fungi is comparatively limited. Thus, a more precise analysis of pathogenic fungal genomes was followed to find sets of proteins with similar homology and similar functions, similar homology and different functions and different homology and similar functions and as well their biological networks. When compared with genomes of 13 other fungi, 3340 yeast genes had homologs in at least 12 of them while only 17 of the common fungal genes had no significant homologs in other organisms **[10]**. Out of 82 proteins in Aspergillus fumigatus, we found as any as 37 putative proteins known to have significant interactors as per our STRING analyses **(Figure 2)**. We obtained substantial number of domains across the fungal proteins as the most likely domain-domain interactions from experimentally supported protein-protein interactions **[18]** were found by DIMA. On the other hand, domain phylogenetic profiling was done indicating the presence ('1') or absence ('0') of a domain in the selected genomes **(Table 1 See Supplementary material)**. The rationale behind protein phylogenetic profiling is that protein(function)s/domains that depend on each other for an important cellular function generally need to be present together or not at all in a given genome/proteome. Now that a certain genome has been profiled, in the near future this phylogenetic profiling constituting the domains per se MutS would allow us to have a broad selection of organisms or a specific group. In conclusion, it is like all these protein domains are similar in very closely related organisms while the converse is that it is a good idea to use a large number of organisms from the widest phylogenetic spectrum possible and crosscheck whether or not the organism proteins harboring the domains are similar. Assessing which of the protein (domain) s have few or no predicted neighbors would allow us to understand which of these proteins are essential to a specific fungal genome. Lower the entropy, least is the chance for the domain to yield a signal suggesting that positive predictions provide good hints for a domain-domain relationship. However, careful assessment and performance of computing entire networks using conserved gene neighbors' and Rosetta stone method, co-expression using gene expression omnibus (GEO), gene fusion and gene neighbors' methods are being carried out to establish interactors in silico.

Our phylogenic studies reveal (Figure 3) that unique genes are responsible for pathogenesis. The interactors include DNA repair genes, vegetative and sporulation. Organisms like budding yeast and Candida albicans which are known to be highly similar, hyphae development and sporulation are most important characteristic features of fungi which are shared between pathogenic and non pathogenic fungi (He, F et al). A study of these interactions has revealed pathogenic pathways specific to cancer in human.

Bioinformation Volume 5

open access

www.bioinformation.net Issue 5

Hypothesis

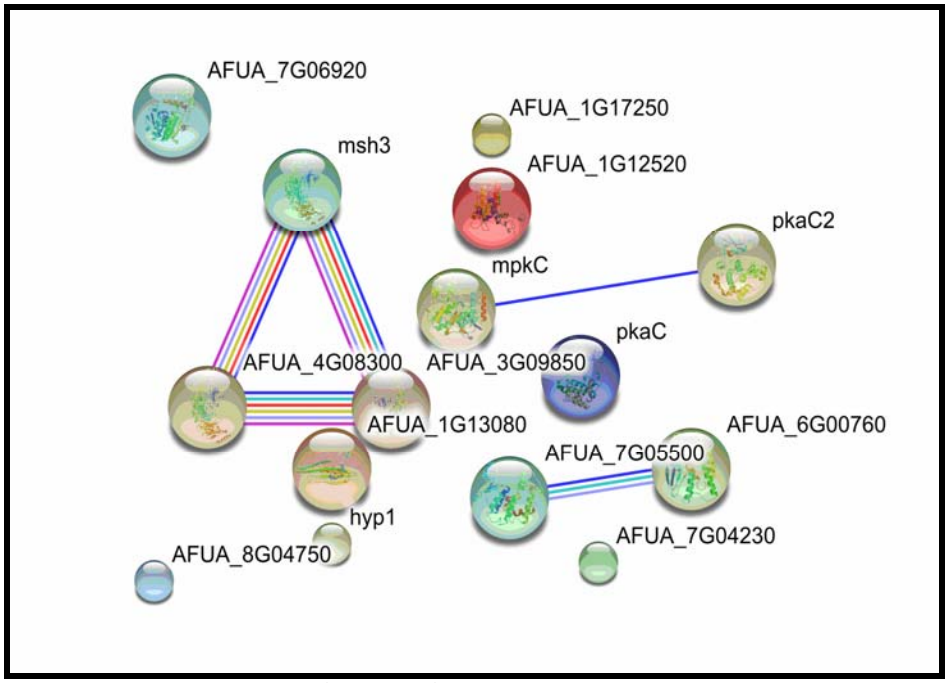

Figure 2: Protein-protein interactions of Aspergillus fumigatus build using STRING analysis.

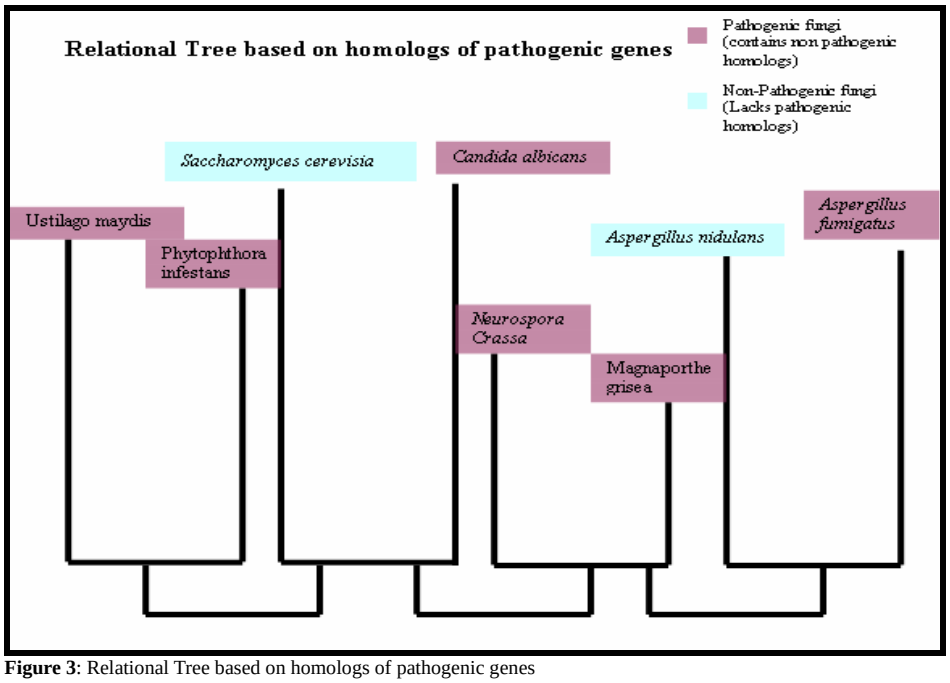

Figure 3: Relational Tree based on homologs of pathogenic genes

Existing programs for automated gene prediction and annotation are not perfect and need to be improved. Follow-up manual annotation also is necessary to improve the accuracy of automated annotation especially proteins specific to fungal pathogenesis. A comprehensive fungal genome database similar to yeast (http://www.yeastgenome.org) would overcome the limitation of searching umpteen proteins through redundant databases. Using comparative genomics, we addressed evolutionary and phylogenetic questions in closely related pathogenic fungi even as our annotation was perused to identify candidate genes involved in host range determination, infection-related morphogenesis, and virulence. Therefore with an open source repository of all fungal pathogenic proteins we believe, would allow the researchers to focus on functional characterization of candidate genes specific to fungal pathogenesis. We show that the fungome integrates genes specific to human cancers across diverse strains by elucidating the functionality and sub-cellular localization of selected proteins related to signaling cascades to understand a particular disease. There needs to understand the interactions between components of the pathogens (Figure 4) and specific cells when an infection occurs. We believe this will

Bioinformation Volume 5

open access

www.bioinformation.net Issue 5 **Hypothesis**

facilitate knowledge on the genes that are controlled during the host-pathogen interaction ex- vivo while providing a critical detail of pathobiology of the fungus. Could a model be developed with a pathogenic fungus based on the available information? With an approximate 4000 genes shared by all the sequenced yeast genomes, is there a gain of genes by horizontal gene transfer? According to Ronald Gdanski, fungal control of a mass of human cells can explain all known observations about cancer thereby proving the cause of cancer. In the process, some important but less investigated pathogens will be further neglected or under investigated. One well known example is fungal infections triggering down-stream signaling which de-regulates cell cycle and might activate carcinogenic events. Some wet-lab methods like protein polymorphisms, electrophoretic karyotyping, PCR based printing, RFLP, AFLP and DNA sequencing mostly focus on the gene specific molecular markers, nevertheless could detect such changes in the fungal genome or pattern of gene expressions With over 25 fungal genomes sequenced and over three genome sequence yet to be released, we highlight the need for 'fungome' – a FUNGal protein interactOME specific to pathogenic organisms. Given the magnitude of data available and the bio-information leveraged, we believe fungome comprehends the omics-es of fungal oncogenes, specific to pathogenesis.

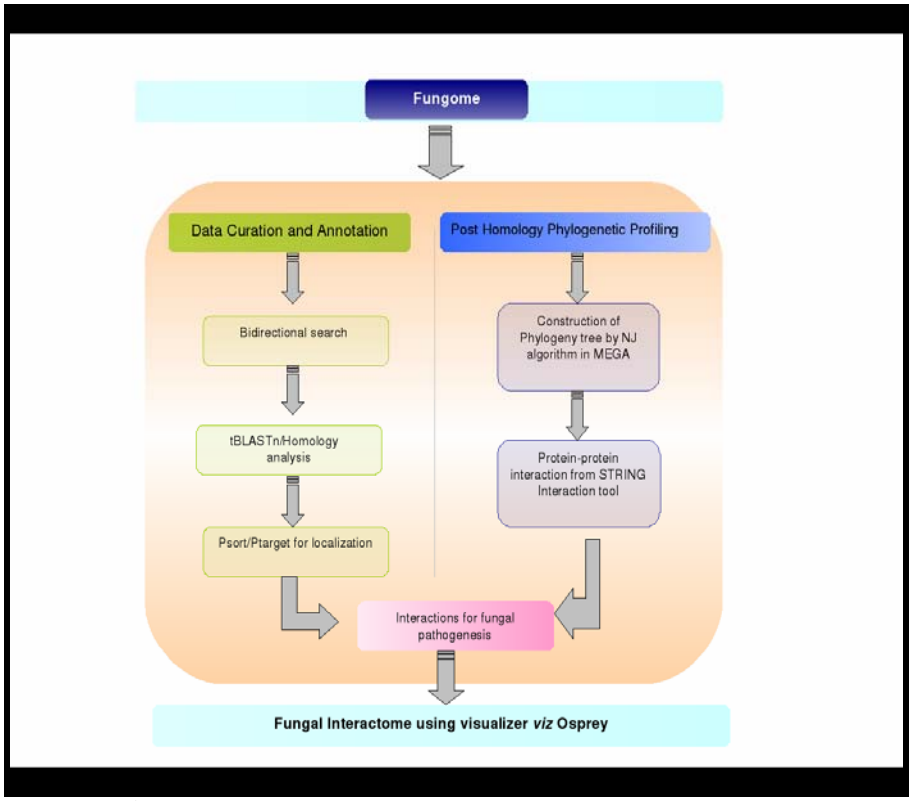

Figure 4: Workflow that could be complemented to study protein-protein interactions

Competing interests:
The authors claim no competing interests. This is an open source project hosted virtually through Bioclues.org and the work will have had received no funding.

Acknowledgement:
The authors would like to thank Dr. Krishna Mohan for constructive comments.

References:

[1] JS Bader *et al. Nature Biotechnology* **22**: 78 (2003) [PMID: 14704708]
[2] GW Beadle & EL Tatum. *Proc Natl Acad Sci U S A.* **27**(11): 499 (1941) [PMID: 1658849]
[3] A Brückner *et al. Int J Mol Sci* **10**(6): 2763 (2009) [PMID: 19582228]
[4] MD De Backer *et al. Am J Pharmacogenomics* **2**(2): 113 (2002) [PMID: 12083946]
[5] MD De Backer & P Van Dijck. *Trends Microbiol.* **11**: 470 (2003) [PMID: 14557030]
[6] B Dujon *et al. Nature* **430**(6995): 35 (2004) [PMID: 15229592]
[7] RD Finn *et al. Bioinformatics* **21**(3): 410 (2005) [PMID: 15353450]
[8] C Guda & S Subramaniam. *Bioinformatics* **21**: 3963 (2005) [PMID: 16844995]
[9] F He *et al. BMC Genomics* **9**: 519 (2008) [PMID: 18976500]
[10] T Hsiang & DL Baillie. *J Mol Evol* **60**(4): 475 (2005) [PMID: 15883882]
[11] LJ Jensen *et al. Nucleic Acids Res* **37**(Database issue): D412 (2009) [PMID: 18940858]
[12] MD Katinka *et al. Nature* **414**(6862): 450 (2001) [PMID: 11719806]
[13] N Morrison *et al. OMICS Summer* **10**(2): 127 (2006) [PMID: 16901217]
[14] WC Nierman *et al. Nature* **438**(7071): 1151 (2005) [PMID: 16372009]
[15] FC Odds *et al. Genome Biol* **2**(3): REVIEWS1009 (2001) [PMID: 11276429]
[16] P Pagel *et al. J. Mol. Biol* **344**(5): 1331 (2004) [PMID: 15561146]
[17] W Pearson & D Lipman. *Proc. Natl. Acad. Sci.* USA **85**: 2444 (1988) [PMID: 3162770]
[18] ML Riley *et al. Nucleic Acids Res* **33**(Database issue): D308 (2005) [PMID: 15608204]
[19] Ron Gdanski. Cancer, Cause, Cure and Cover-Up, New Century Pr, 3 edition (2000).
[20] S Kerrien *et al. Nucleic Acids Res.* **35**(Database issue): D561 (2006) [PMID: 17145710]

[21] DR Scannell *et al. Yeast* **24**(11): 929 (2007) [PMID: 17621376]
[22] DP Thomas et al. Infect Disord Drug Targets. 6(4): 335 (2006) [PMID:17168799]

Edited by P. Kangueane
Citation: Prashanth *et al.* Bioinformation 5(5): 202-207 (2010)

Bioinformation Volume 5

open access

www.bioinformation.net Issue 5 **Hypothesis**

Supplementary material:

Table 1: Presence and absence of important genes in select fungal organisms; and the annotation methods employed

Organism	DNA repair	Metabolites	Cancer gene(s)	GST	cell wall biogenesis	Vegetative growth	Sporulation
Aspergillus fumigatus	1	1	1	1	1	1	1
Aspergillus flavus	1	1	1	0	1	0	1
Coccidioides immitis	0	0	0	0	1	1	1
Gibberella zeae	1	1	1	1	1	1	0
Magnaporthe grisea 70-15	0	0	0	0	1	1	0
Phaeosphaeria nodorum SN15	0	0	0	0	1	1	0
Candida albicans	1	0	0	1	1	1	0