2011 International Conference on Bioscience, Biochemistry and Bioinformatics
IPCBEE vol.5 (2011) © (2011) IACSIT Press, Singapore

An Integrated in *Silico* Approach to Deduce Interacting Domains of Primary Immunodeficiency Disease Causing Genes

Suresh Kumar Ramadoss [1], Akhilesh Pandey[2], Osamu Ohara[3], and Sujatha Mohan[1*]
[1]Research Unit for Immunoinformatics, Research Center for Allergy and Immunology (RCAI), RIKEN Yokohama Institute, Japan.
[2]McKusick-Nathans Institute of Genetic Medicine and Departments of Biological Chemistry, Pathology and Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA & Institute of Bioinformatics, International Technology Park, Bangalore 560 066, India,
[3]Laboratory for Immunogenomics, Research Center for Allergy and Immunology, RIKEN Yokohama Institute, Kanagawa 230-0045, Japan & Laboratory of Genome Technology, Department of Human Genome Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan.
[*]To whom correspondence should be addressed: Sujatha Mohan, Ph.D E-mail: sujatha@rcai.riken.jp

***Abstract*—Primary immunodeficiency diseases (PIDs) are complex and intrinsic genetic disorders that leads to immune dysfunction. We have recently developed "Resource of Asian Primary Immunodeficiency Diseases", an open access database on PIDs. We propose a heuristic approach of PID gene mutation data analysis based on the functional domain interactions. We identified functionally significant domains that disrupts PID genes' protein-protein interactions (PPI) associated with disease mutation We have also prioritized the domains to be associated with immune diseases and function based on observed PID gene mutations.**

Keywords-domain-domain interaction; disease mutation; HitPredict; DIMA

I. INTRODUCTION

Primary immunodeficiency diseases (PIDs) are genetic disorders resulting in abnormalities in the development and maintenance of immune system. Patients with these intrinsic defects have common and overlapping manifestations which pose daunting task to clinicians in providing definitive diagnosis based on determined sequence variations with observed phenotype. We have recently developed an open access database on PID designated as "Resource of Asian Primary Immunodeficiency Diseases (RAPID)", a web-based compendium of molecular alterations and gene expression at the mRNA and protein levels of all PID genes reported from PID patients in the public literature. The database also includes other pertinent information about protein-protein interactions, mouse studies and microarray gene expression profiles in various organs and cells of the immune system and it can be freely accessible at http://rapid.rcai.riken.jp [1].

Several studies indicate the importance of interacting domains and its disease association [2-7]. It is apparent that protein sequence-specific functional domains play an important role in key biological events including protein-protein interactions, post-translational modifications (PTMs) and so on. We understand that any mutations occurring in such functional domains would cause varying effects in the subsequent biological events. With this collective knowledge, we propose a heuristic integrated *in silico* approach to analyze PID gene-specific mutations observed in the spanning region of the functional domains. For this analysis, we have obtained PID gene-specific available mutation data from RAPID. This kind of systematic study on the reported mutations occurring in PID gene interacting domains will shed light on the domain function and genotype-phenotype correlation of PIDs thereby further enhancing our current knowledge of PID pathogenesis.

II. METHODS

DATA SOURCES

A. Identification and Characterization Of Pid Gene Interacting Domains

We have used BioMarts' MartView [8] to obtain defined domain spanning region for PID genes and its interacting partners along with the GO terms and also selected Pfam [9] source signature database as a unique domain data source. To introduce stringency in the generated list of InterPro [10] domains, we have considered active site, binding site, conserved site, domain, family and post-translational modifications (PTMs). The compiled results of domain spanning regions for PID gene and its interactors are derived from BioMart by selecting Uniprot/Swiss-Prot [11] as source protein database with its UniprotKB accession IDs as input data.

B. Protein-Protein Interaction (PPI)

For our analysis, we have generated the list of high confidence binary interaction data that are directly observed through experimental studies as reported in HitPredict, a comprehensive resource of high confidence protein-protein interactions [12]. To generate human interaction data sets, we have sorted and consolidated a list of non-redundant

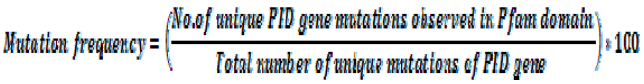
interaction pairs of known PID genes along with the binary interaction data as available in HPRD [13].

C. *Domain-Domain Interaction (DDI)*

Domain Interaction Map (DIMA) is a comprehensive resource of functional and physical interactions among conserved protein domains [14]. We used this resource for mapping the available domain-domain interaction pairs representing PID genes and its interactors. We obtained available DDI data for Homo sapiens with the default parameters as given in DIMA tool.

III. Implementation And Results

To understand and dissect-out disease causing mutations observed in the defined functional domain spanning region of PID gene that disrupt the intensity of interaction, we mapped the PID gene mutations and its interacting domains. We have used Python scripts to extract and analyze the data from various open-accessible resources. The workflow of our approach is depicted in Fig 1. First, we have obtained the Pfam domain spanning region along with its domain function from MartView by submitting the Uniprot IDs of all PID genes and its interactors and for all those missed entries, necessary details have been collected from available published literature. Next, we processed PPI data and obtained 2603 non-redundant interaction pairs for all available PID genes. This is followed by generation of DDI data for the interaction pairs by mapping HitPredict and DIMA. In this step, we have considered data if both the Pfam IDs of interacting domains in DIMA matches with both PID gene and its interactors' Pfam IDs to confirm its interacting partners. In other words, we have excluded the interaction pairs if either one of the matching entries was not found in the same. This stringent criterion has been implemented throughout this analysis in order to control and eliminate ambiguous entries thereby presenting the best possible results. Among PID gene interaction pairs, we have identified only 341 pairs having the associated DDI evidence from DIMA. Subsequently, we scanned the RAPID mutation data which are mapped to the protein-coding regions and obtained 105 PID genes. In this analysis, mutation frequency has been calculated using the following formula:



Moreover, we also observed that varied frequency of mutations has been occurred in the interacting domains of individual PID genes. But, we considered only the genes having the mutation frequency of 80% and above to prioritize its respective domains. Using this approach, we could filter 39 PID genes having frequency of 80% and above for the observed mutation in its interacting domains as well as total of 33 prioritized functional domains. The overall statistics of interacting domain analysis and the list of prioritized domains are shown in Tables 1 and 2 respectively. The compiled results are given in a supplementary file and it can be accessed at http://rapid.rcai.riken.jp/RAPID/mut/domain_results.xls.

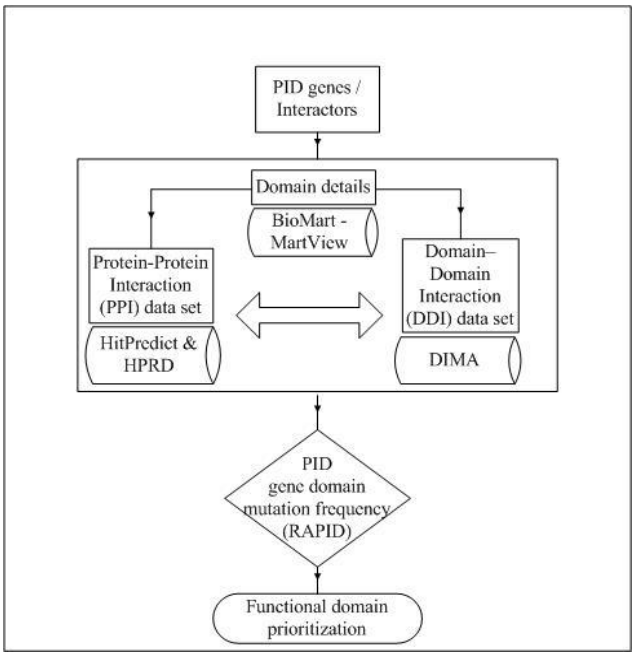


Figure 1. Overview of in silico approach to prioritize interacting domain associated high frequency PID gene-specific mutations. Domain spanning region for PID gene and its interactors are retrieved from BioMart. PPI and DDI data sets are referred from HitPredict & HPRD and DIMA resources respectively. Functional domain annotation and prioritization are performed using UniProt and RAPID respectively to gather PID gene-specific mutations with more than 80% frequency along with integration of available PPI and DDI data set.

TABLE I. Statistics Of PID Gene-Specific Mutation-Based Interacting Domain Analysis

Type of analyzed data set	Total number of analyzed data set
PID gene interaction pairs	2603
Domain-domain interaction pairs obtained from mapping PPI and DDI data sets	341
PID genes' domain-domain interaction pairs with reported mutations	199
PID genes with domain-domain interaction data	105
PID genes with observed mutation frequency (≥ 80%) in the functional domain	39
Prioritized Pfam domains having mutation frequency (≥ 80%)	33

IV. DISCUSSION

Our results suggest that the prioritized domains have a significant role in immune function. For *in silico* evaluation of our analyzed results, we screened a few selected high frequency mutation entries for thorough description of PID gene-specific functional domain annotation as mentioned below:

Lectin_C domain in C-type lectin domain family 7, member A (CLEC7A) is mainly responsible for recognizing pathogens and immune regulation [15]. Serine proteases including trypsin domain involve in innate immune response [16]. Also, ELANE gene encodes Elastase 2, neutrophil protein containing trypsin domain interacts with LPA having reported mutation frequency of 93.65% (as per the data shown in the supplementary file) that causes characteristic neutropenia in the diagnosed patients. The Collagen domain in human Surfactant protein A plays a major role in innate immune defense mechanism [17]. These studies clearly illustrated the potential role of prioritized domains (Table 2) in human immune system. The list of domains highlighted in bold letters (in Table 2) are having defined UniProt sequence feature description viz., phosphorylation site, sites of disulfide bond, active site, binding site and other regions of interest. Thus, these annotated sequence features serve as added evidences for its functional importance.

Earlier our team published a mutation evaluation tool [18] using SIFT program [19]. This study clearly demonstrated a new dimension towards identification and description of PID causing mutations in its functional domains in human disease pathogenesis.

In a similar way, our integrated approach is expected to provide an essential insight on functionally defined domains based on the frequency of observed mutations. Such mutation based domain analysis in the PID genes should help further in demonstrating genotype-phenotype correlation wherever functional studies are not available in the published literature. Although, there are many factors involve in deducing the effect of a particular mutation, we consider the conserved functional domain as our primary aspect in analyzing the mutation data. Based on this collective domain interaction data, it has been confirmed that about 56% (in average) of observed mutations are found in the interacting domains of PID genes. However, for individual PID gene, the frequency of mutations in interacting domain varies from 2.17% to 100%. This is clearly illustrated that more functional annotation of domains in disease causing genes can be identified through such *in silico* analysis. In general, these domain regions are considered to be functionally involved in protein interactions and have biological significance in any immune signaling pathways. PID causing through the genetic defects must involve a disruptive interaction in the functional pathway leading to expression of specific phenotypic traits. Thus the domains obtained through our results should provide further clue for screening the PID candidate genes having similar mode of DDI pairs involved in a disease-specific pathway. We propose to apply similar approach with orthologous species' domain and interaction data sets to enhance our results.

It is also apparent that disrupted interaction in a signaling pathway leads to a phenotype trait [20-22]. To delve further, we will incorporate and integrate these results in our ongoing project in the development and construction of PID specific immune signaling pathways.

V. CONCLUSION

With successful implementation of our integrated approach, it should assist in deriving genotype and phenotype correlation thereby improving phenotype-based genetic analysis of PID genes. Moreover, this kind of mutation analysis should augment well with the understanding of domain specific interaction and its impact on the disease pathogenesis. Eventually, it would facilitate clinicians in confirming early PID diagnosis and proper therapeutic interventions.

TABLE II. LIST OF PID GENE DOMAINS REPORTED WITH DISEASE-CAUSING MUTATIONS* FROM RAPID

Domain name	Pfam ID	Molecular function class	Reported PID gene	Distribution of mutations (average mutation frequency in %)
7tm_1	PF00001	G-protein coupled receptor protein signaling pathway	FPR1	100
ABC_membrane	**PF00664**	ATPase activity, coupled to transmembrane movement of substances	**TAP2**	100
Actin	PF00022	Protein binding	ACTB	100
ApoL	PF05461	Lipoprotein metabolic process	APOL1	100
C1q	PF00386	Complement activation	C1QA	100
Collagen	PF01391	Connective tissue formation	MBL2	100

CUB	PF00431	Complement activation; Cell signaling; Tissue repair	MASP2	100
Cytochrom_B558a	PF05038	Heme binding	CYBA	100
Death	PF00531	Signal transduction	FADD	100
DUF1650	PF07856	Mediation of CRAC channel activity	ORAI1	100
FAT	PF02259	Protein binding	PRKDC	100
Fibrinogen_C	PF00147	Signal transduction; Receptor binding	FCN3	100
IL6Ra-bind	PF09240	Cytokine binding	CSF2RA	100
Interfer-bind	PF09294	Ligand binding	IL10RB	100
Lectin_C	PF00059	Carbohydrate-binding activity	CLEC7A	100
MACPF	PF01823	Transmembrane channel formation	C8A	100
Peptidase_C14	PF00656	Cysteine-type endopeptidase activity	CASP10; CASP8	100
PX	PF00787	Cell communication; Phosphoinositide binding	NCF4	100
RAG2	PF03089	DNA recombination; DNA binding	RAG2	100
Sushi	PF00084	Complement control protein	CFHR1	100
Tetraspannin	PF00335	Signal transduction	CD81	100
UPAR_LY6	PF00021	Membrane attack complex inhibition factor	CD59	100
V_ATPase_I	PF01496	Proton-transporting two-sector ATPase complex, proton-transporting domain	TCIRG1	100
V-set	PF07686	Cell-surface receptor	CD8A; CD79B	100
WD40	PF00400	Signal transduction; Transcription regulation	CORO1A	100
A_deaminase	PF00962	Deaminase activity; Purine ribonucleoside monophosphate biosynthetic process	ADA	98.08
Trypsin	**PF00089**	Serine-type endopeptidase activity; Proteolysis	CFD; **ELANE**	96.83
MFS_1	PF07690	Transmembrane transport	SLC37A4; SLC46A1	96.66
Ras	**PF00071**	Guanine nucleotide exchange factors interaction	NRAS; RAC2; **RAB27A**	94.44
An_peroxidase	PF03098	Heme binding; Peroxidase activity	MPO	90.91
PNP_UDP_1	PF01048	catalytic activity; nucleoside metabolic process	NP	90.48
Serpin	PF00079	Serine-type endopeptidase inhibitor activity	SERPING1	83.77
Cobalamin_bind	PF01122	Cobalamin binding	TCN2	80

*PID gene domains having 80% and above percentage of mutations are considered. Note that bold entries have mutations in functional sites as described in Uniprot's sequence features

VI. Future Perspectives

Future efforts will aim towards integrated data analyses, thereby leading to a more comprehensive view of the biology of PID, a prerequisite for identification of diagnostic and prognostic markers and improved patients' therapeutic modalities.

Acknowledgment

The authors thank the research team at Institute of Bioinformatics, India for their collaboration in developing RAPID. Also, we thank all PID physicians involved in the PID Japan project as well as RAPID – PID experts and Dr. Ashwini Patel, Human Genome Center, Institute of Medical Science, University of Tokyo for their valuable inputs and suggestions.

References

[1] Keerthikumar, S., R. Raju, et al. (2009a). "RAPID: Resource of Asian Primary Immunodeficiency Diseases." Nucleic Acids Res 37(Database issue): D863-7

[2] Argentaro, A., J. C. Yang, et al. (2007). "Structural consequences of disease-causing mutations in the ATRX-DNMT3-DNMT3L (ADD) domain of the chromatin-associated protein ATRX." Proc Natl Acad Sci U S A 104(29): 11939-44

[3] Pippal, J. B., Y. Yao, F. M. Rogerson, P. J Fuller. (2009). "Structural and functional characterization of the interdomain interaction in the mineralocorticoid receptor." Mol Endocrinol 23(9): 1360-70

[4] Perreau, V. M., S. Orchard, et al. "A domain level interaction network of amyloid precursor protein and Abeta of Alzheimer's disease." Proteomics 10(12): 2377-95

[5] Limviphuvadh, V., S. Tanaka, S. Goto, K. Ueda, M. Kanehisa. (2007). "The commonality of protein interaction networks determined in neurodegenerative disorders (NDDs)." Bioinformatics 23(16): 2129-38

[6] George, R. A., J. Y. Liu, et al. (2006). "Analysis of protein sequence and interaction data for candidate disease gene prediction." Nucleic Acids Res 34(19): e130

[7] Gandhi, P. N., S. G. Chen, A. L. Wilson-Delfosse (2009). "Leucine-rich repeat kinase 2 (LRRK2): a key player in the pathogenesis of Parkinson's disease." J Neurosci Res 87(6): 1283-95

[8] Haider, S., B. Ballester, et al. (2009). "BioMart Central Portal--unified access to biological data." Nucleic Acids Res 37(Web Server issue): W23-7

[9] Finn, R. D., J. Mistry, et al. (2010). "The Pfam protein families database." Nucleic Acids Res 38(Database issue): D211-22

[10] Hunter, S., R. Apweiler, et al. (2009). "InterPro: the integrative protein signature database." Nucleic Acids Res 37(Database issue): D211-5

[11] UniProt consortium. "The Universal Protein Resource (UniProt) in 2010." Nucleic Acids Res 38(Database issue): D142-8

[12] Patil, A., K. Nakai, H. Nakamura et al. (2011). "HitPredict: a database of quality assessed protein-protein interactions in nine species." Nucleic Acids Res. (in press)

[13] Keshava Prasad, T. S., R. Goel, et al. (2009). "Human Protein Reference Database--2009 update." Nucleic Acids Res 37(Database issue): D767-72

[14] Luo, Q., P. Pagel, B. Vilne, D. Frishman et al. (2010). "DIMA 3.0: Domain Interaction Map." Nucleic Acids Res. (in press)

[15] Cambi, A. and C. G. Figdor (2003). "Dual function of C-type lectin-like receptors in the immune system." Curr Opin Cell Biol 15(5): 539-46

[16] Dale, C. and N. Vergnolle (2008). "Protease signaling to G protein-coupled receptors: implications for inflammation and pain." J Recept Signal Transduct Res 28(1-2): 29-37

[17] Garcia-Verdugo, I., G. Wang, J. Floros, C. Casals. (2002). "Structural analysis and lipid-binding properties of recombinant human surfactant protein a derived from one or both genes." Biochemistry 41(47): 14041-53.

[18] Hijikata, A., R. Raju, et al. "Mutation@A Glance: an integrative web application for analysing mutations from human genetic diseases." DNA Res 17(3): 197-208

[19] Ng, P. C. and S. Henikoff (2003). "SIFT: Predicting amino acid changes that affect protein function." Nucleic Acids Res 31(13): 3812-4

[20] Gong, Z., Y. W. Cho, J. E. Kim, K. Ge, J. Chen. (2009). "Accumulation of Pax2 transactivation domain interaction protein (PTIP) at sites of DNA breaks via RNF8-dependent pathway is required for cell survival after DNA damage." J Biol Chem 284(11): 7284-93

[21] Jiang, W., R. Sordella, et al. (2005). "An FF domain-dependent protein interaction mediates a signaling pathway for growth factor-induced gene expression." Mol Cell 17(1): 23-35

[22] Kann, M. G. (2007). "Protein interactions and disease: computational approaches to uncover the etiology of diseases." Brief Bioinform 8(5): 333-46.