DAPD: A Knowledgebase for Diabetes Associated Proteins

Krishnasamy Gopinath, Ramaraj Jayakumararaj, and Muthusamy Karthikeyan

Abstract—Recent advancements in genomics and proteomics provide a solid foundation for understanding the pathogenesis of diabetes. Proteomics of diabetes associated pathways help to identify the most potent target for the management of diabetes. The relevant datasets are scattered in various prominent sources which takes much time to select the therapeutic target for the clinical management of diabetes. However, additional information about target proteins is needed for validation. This lacuna may be resolved by linking diabetes associated genes, pathways and proteins and it will provide a strong base for the treatment and planning management strategies of diabetes. Thus, a web source “Diabetes Associated Proteins Database (DAPD)” has been developed to link the diabetes associated genes, pathways and proteins using PHP, MySQL. The current version of DAPD has been built with proteins associated with different types of diabetes. In addition, DAPD has been linked to external sources to gain the access to more participatory proteins and their pathway network. DAPD will reduce the time and it is expected to pave the way for the discovery of novel anti-diabetic leads using computational drug designing for diabetes management. DAPD is open accessed via following url www.mkarthikeyan.bioinfoau.org/dapd.

1 INTRODUCTION

Diabetes is a metabolic disorder in both the developed and developing countries. It is estimated that about 371 million people have diabetes worldwide and this number has been projected to rise to 552 million by 2030. The rate of prevalence of diabetes is high in Asian countries. About 92.3 and 63 million people in China and India suffer from this metabolic disorder respectively [1], [2]. Diabetes (T2D) reduces the life expectancy by 5-10 years [3]. Prevalence of diabetes is rapidly increasing worldwide due to interactions between environmental factors, genetic factors and lifestyle. Furthermore, factors such as population growth, aging, urbanization, and increasing prevalence of obesity and physical inactivity are significantly contributing to the increasing global burden of diabetes [4], [5].

Deranged glucose homeostasis have been identified as the core aspect that leads to the onset of diabetes, mainly due to impairment of the biochemical pathways involved in insulin secretion, insulin action, glucose metabolism and autoimmune destruction of Beta-cells [6], [7]. Furthermore, complex interplay of environmental and genetic factors has been reported to hamper insulin signal transduction coupled with changes in the expression level of genes and proteins involved in the glucose metabolism [8]. This brings significant metabolic abnormalities in glucose transport in liver, pancreas, skeletal muscle at sub cellular cells and tissues level [9].

It has been well established that interaction between the cellular proteins are quite complex. Further, they are highly related to upstream, downstream signaling and metabolic pathway(s). In many diseases, proteins are regarded as molecular targets for drug discovery. Nevertheless, an exact relationship between the proteins and the disease has not yet been set up [10]. On the contrary, analysis of genomic data indicates that the proteins outnumber the genes involved in a particular pathway. This complexity is due to alternative splicing and post-translation modifications [11]. Therefore, holistic understanding of diabetes pathogenesis is expected to surmount the obstacles in diabetes management strategies [12].

A quantum of information is available in open source databases that have a significant impact on pathway based drug discovery by in silico approaches. Information on signaling and metabolic pathway(s) provides a basic understanding towards the role of cellular proteins in pathogenesis of diabetes to identify the potential target for therapeutics and to validate the selected target [13]. If, the target protein is not responding to the drug, an alternative target from disease associated biochemical pathway(s) might address this issue. [14]. The scattered information pertinent to biochemical pathways and related proteins in literatures and online biological databases make it extremely difficult to mine the required data. Even though, the researcher collects relevant data for a selected target protein from several informants, this is done piecemeal, and hence needs a long time to set up a perfect relationship.

Thus, information pertinent to the biochemical pathways and its associated proteins if available on a single open access webpage is expected to speed up the rate of molecular diagnosis and formulate strategies for individualized
diabetes care with more honest reason. An effort is made to evade the present day laborious process of collecting relevant information about various types of diabetes.

At present, there are only a few online sources that link the proteomic information with diabetes, such as T1Dbase [15], T2D-Db [16], The Human Diabetes Proteome Project (HDPP) [17] and T2D@ZJU [18]. But, they focus on a specific type of diabetes. Thus, the current version of DAPD has been developed to focus protein and pathways of various types of diabetes. It is expected that DAPD will serve as the platform for researchers and the scientific community to gain in-depth protein information from the single site.

2 METHODS

2.1 Integration of Pathway and Gene Information with Protein

Information pertinent to diabetes and its associated pathway(s) were collected from Reactome [19], KEGG [20] and Pathway Interaction Databases (PID) [21] with respect to the gene and protein data. Reactome provides a list of proteins involved in signaling and metabolic pathways, where users can retrieve proteins with their respective UniProt IDs. KEGG and PID databases provide gene list(s). Genes involved in diabetes pathways were related to KEGG GeneIDs and subsequently matched to their corresponding UniProt IDs using an identifier mapping tool [22], [23].

KEGG PATHWAY and KEGG DISEASE include pathway(s) relevant to human diseases such as cancer, immune, neurodegenerative, cardiovascular, endocrine and metabolic and infectious diseases. Metabolic network for Type I, II diabetes mellitus, Maturity Onset Diabetes of Young (MODY), Permanent Neonatal Diabetes Mellitus (PNDM), Transient Neonatal Diabetes Mellitus (TNDM), Ketosis-prone Diabetes Mellitus (KPD), Insulin Resistant diabetes mellitus with Acanthosis NIGRICANS (IRAN) and Rabson-Mendenhall Syndrome are available under Endocrine and Metabolic Diseases in KEGG PATHWAYS and KEGG DISEASE [20]. Metabolic network in KEGG relates the participating gene molecule to its respective pathways. The KEGG ID of diabetes related genes was linked to its product (protein) using the UniProt ID mapping tool. The Proteins associated with all diabetes categories were retrieved from UniProt using a batch query on the UniProt site. The entire collected proteins list from diabetes types and diabetes associated pathways were stored in a single comma delimited (comma separated value (CSV)) file with their respective UniProt IDs.

2.2 Retrieval of Information about Individual Protein

Proteins involved in the diabetes types and diabetes associated pathways were merged together and the repeats were removed. The protein related information (general, gene and sequence) were collected from UniProt. Gene related information (chromosomal location, gene symbol, complete name and gene summary) were collected from Entrez Gene [24]. Furthermore, protein sequence was used to compute physico-chemical properties (molecular weight, theoretical pi, amino acid composition, atomic composition, estimated half life, instability index, aliphatic index and grand average of hydropathicity) using the ExPASy ProtParam tool [25]. In order to provide comprehensive information for each of protein, the respective entry was linked to dbSNP [26], Reactome, KEGG Gene, Ensembl [27], GeneCards [28], HGNC [29], OMIM [30], NCBI nucleotide sequence Graphical view [31], PharmGKB [32] and PubMed (www.ncbi.nlm.nih.gov/pubmed).

2.3 Creation of DAPD

DAPD is built using WAMP server, that includes PHP 5.3, MySQL 5.5 and Apache 2.2 server [33]. This WAMP server supports open source MySQL database along with the scripting environment with Apache server.

A MySQL database (bioinfoa_dapd) was created with tables to store the entire data set. Information about the diabetes associated proteins was compiled into various comma separated value files. The CSV file was imported into various tables of MySQL database. DAPD data set mainly focuses on the proteins, which are linked with the diabetes associated pathways and diabetes types. Thus, the table having pathway and diabetes type information was considered as the main table. In addition, it was linked to protein table through pathway ID and protein ID assigned to all proteins in the protein information table. Uniprot ID is specified as the protein ID to integrate data in protein tables. Generally, sequence information and physico-chemical properties of proteins and gene information were stored in separate tables and linked to Uniprot ID. To avoid technical complication Uniprot ID was used as a unique identifier in these tables. Citation information was added in the reference table and it has links to other tables through reference ID. In addition, one table was added to store feedback (Fig. 1). Input for the feedback form will be given by the user. A separate php file was designed with form action to receive user comments/suggestions/complaints. The user input from the feedback form will be updated automatically in the feedback table and the DAPD administrator will be notified by email. A copy of the feedback will be sent to the user’s mail ID.

The web interface was developed in HTML-CSS and server side scripting was encoded in PHP to link front and back ends of DAPD for the dynamic execution of MySQL queries. Additionally, a search page has been developed with HTML and JavaScript. Browser files will pass the user selected values to the web server. The web server receives the selected user value as variables and executes the requested PHP file and it passes a request to the PHP ‘engine’ to execute them. PHP runs through the sequence of instructions in the php file. MySQL server acts in the back-end and executes commands to retrieve the stored information. It checks the user requested values in bioinfoa_dapd tables and fetches all the information from the row of respective tables and linked tables in the form of a query written in SQL. A set of variables was assigned to store the values fetched from bioinfoa_dapd for user queries. Finally, the web server responds to the user with the requested results.

3 RESULTS

3.1 Data Integration

DAPD includes 361 proteins from six diabetes associated pathways and eight diabetes types (Table 1). All the
information was gathered from the various available sources (Reactome, KEGG and PID). Each of the entries includes selected information that is collected from the UniProt and Entrez Gene. Moreover, the database includes physico-chemical properties of the proteins that can be used to identify similar proteins elsewhere to gain more information. DAPD has been linked to external databases where the user community can get additional information such as SNP details from NCBI dbSNP; Pharmacogenomics knowledge with clinical information from PharmGKB; Orthology, Pathway, Disease, Drug Target, Motif etc from KEGG; Gene sequence in graphical view from NCBI Nuclotide’s Graphical View; transcripts details from Ensembl and all genetic features from OMIM; Approved gene name and symbol from HGNC; Gene related transcriptomic, genetic, proteomic, functional and disease information from GeneCards; Biomedical literatures from PubMed. Thus, DAPD is expected to serve as a useful tool to gain an in-depth understanding of human diabetes associated proteins for selecting a suitable target for the management of diabetes. The overall architecture of DAPD with all its contents is given as an illustration in Fig. 2a, and 2b.

### 3.2 User Interface

DAPD has been launched to select a suitable therapeutic target based on the biochemical pathway approach. Here, a user can retrieve protein details by (i) diabetes type (ii) diabetes associated pathway (iii) selected candidate protein (Fig. 3a). The diabetes associated pathways and its sub networks appear in the dropdown menu (Fig. 3b), by selecting a pathway of interest one can get a list of participating proteins and a short summary of that selected pathway. Furthermore, all proteins are hyperlinked to an HTML implemented PHP file, which executes MySQL queries to provide comprehensive information about each entry (Fig. 3e). In a similar way a user can obtain protein details and its individual information from Diabetes types upon selection from the dropdown menu (Fig. 3c). Third, a browser page is created with reported proteins involved in diabetes associated pathways and diabetes types (Fig. 3d). Apart from all the search pages, DAPD also includes a home page with current versions summary and latest updates. In addition, the diabetes informative website has been linked with DAPD.

### 3.3 Implementation

DAPD has been developed and tested with a WAMP server. DAPD is hosted on a Linux based server and it is accessible.
worldwide. DAPD is accessible via www.mkarthikeyan-bioinfoau.org/dapd.

3.4 How DAPD Can Help?

Target-identification is significant in drug discovery as the world is in need of new therapeutic strategies for the management of diabetes [34]. Linking structurally and functionally characterized targets with the disease still remains a challenge [35]. Thus, DAPD is mainly focused to assist the identification of therapeutic target for diabetes treatment and management. Accumulating information at multiple levels ultimately helps to focus and better understand the therapeutic targets. Thereby, it may help to identify novel therapeutic targets that have not previously focused for diabetes management also new users can be aware of existing molecular targets. The current version of DAPD is the collection of protein list from different pathways and diabetes types. It also has many proteins that have many emerging therapeutic targets. For example: Perforin-1 is a protein found in cytotoxic cells, which plays a role in the pathogenesis of Type 1 diabetes by destructing the beta cell [36]. Findings of Young et al. (1989) support the involvement of perforin-1 in the pathogenesis of diabetes [37]. However, targeting perforin-1 for preventing beta cell damage has not been well explained. Only a few literature sources are available to support perforin-1 as the drug target for the prevention of Type I diabetes. This protein has included in DAPD protein list of type I diabetes (Fig. 4). As mentioned in Fig. 2, DAPD provides elaborated information for the user at a single site, as it allows the user to get more information from the external sources, thereby helping the user to acquire more information in less time.

Fig. 2. (a) Illustration of Search forms in DAPD. (b) Overall architecture of DAPD with all its contents.

Fig. 3. (a) Search interfaces in DAPD. (b) Interface to search diabetes associated proteins from pathways; (c) Search interface for protein participating in diabetes. (d) Search interface provides participating protein to know the in-depth information about selected protein. (e) Retrieved information for query from pathway search interface, diabetes type search interface also will provide the information in a similar way.
The DAPD supports computer aided drug design for selecting a suitable target and to identify a potent lead molecule by high throughput virtual screening for the clinical management of diabetes. Further, in vitro and in vivo studies are required to validate the selected drugs.

4 Discussion and Future Prospects

Open source biological databases are crucial to assist life science researchers to access the latest information, which is generated using various available scattered literatures and databases [38], in an easily understandable summarized form. It has been well established that altered gene expression and its products (proteins) lead to impairment in the biochemical pathways that are linked to the insulin synthesis and processing, glucose metabolism, insulin signaling etc., [39] that ultimately ends-up in diabetes. Considerable effort using fine mapping and functional approaches is now being made to characterize pathophysiological mechanisms involved [40]. Therefore, the mechanism of pathogenesis in relation to protein is of prime importance while planning for its treatment and management [41] by identifying key regulatory genes and protein. Identification of a novel diabetes drug target and the study of related proteins is a continuous process. Updates of proteins related to diabetes are available in literature and databases; however, it may not be holistic in nature. Linking diabetes associated pathways, genes and proteins may provide a strong base for diabetes management.

At present, T1Dbase and T2D-Db, 1000-HDPP and T2D@ZJU are active serving knowledgebase for Type I and Type II diabetes. T1Dbase supports the type 1 diabetes community with genetics and genomics of type 1 diabetes susceptibility (T1D); T2D-Db database provides an integrated platform for the better molecular level understanding of type 2 diabetes mellitus and its pathogenesis; 1000-HDPP is generating a list of proteins that are of central interest in the condition of diabetes and T2D@ZJU contains heterogeneous connections associated with type 2 diabetes. However, apart from these two major types of diabetes, information on other types viz. MODY, PNDM, TNDM, KPD, IRAN, Rabson-Mendenhall Syndrome are less available in the form of databases, even if they are incomplete. To fill this gap an attempt has been made to aggregate disease associated pathways, gene and proteins by developing diabetes protein knowledgebase from the primary sources. Thus, DAPD has been developed to provide details about diabetes associated proteins. DAPD provides reliable scientific information
as the data has been collected from the leading online databases. The current version of DAPD is expected to support anyone in research community that keenly looking for a potential diabetes target to design an effective drug using computational drug design for diabetes management. DAPD may be an effective platform as it links the biochemical pathway involved in different types of diabetes with its participating proteins. Furthermore, these proteins are linked to the candidate genes, genetic information and physico-chemical properties.

DAPD consists of proteins from primary databases, which excludes protein details from Gestational diabetes mellitus (GDM) as the information is not available in any primary databases and protein literatures. Current research updates from diabetes researchers suggest that information on diabetes associated proteins and pathways is scattered and it has not been completely digitized to the best of our knowledge. Hence the future updates are mainly focused to include GDM associated proteins and diabetes associated proteins, validated target details from the available literature sources. Addition of structural information, protein–protein interaction is expected to provide a strong base for diabetes management by designing novel inhibitors or inducers for target proteins using computer aided drug design approaches. This lacuna will be fulfilled in the next version of this database.

ACKNOWLEDGMENTS
This work was supported by Post Graduate Diploma in Structural Pharmacogenomics (funded by UGC Innovative Program, UGC, Government of India) and the Department of Bioinformatics, Alagappa University, Karaikudi, India.

REFERENCES


Krishnasamy Gopinath received the master’s degree in bioinformatics from Kongunadu Arts & Science College, Bharathiar University, Coimbatore-29, India, in 2009. He is currently working toward the PhD degree in bioinformatics with Muthusamy Karthikeyan at Alagappa University, Karaikudi, India. His research work includes database creation, phytochemistry, and computer-aided drug discovery.

Ramaraj Jayakumararaj received the PhD degree in molecular genetics from Madurai Kamaraj University, Madurai, India, in 2009. Currently, he is an associate professor in the Department of Botany, Govt. Arts College Melur, Madurai. He has received the Best Teacher Award by TNSCHE, Govt. of Tamil Nadu, India, in 2010. His area of research includes Microbial Genetics and Phytochemistry.

Muthusamy Karthikeyan received PhD degree in biomedical genetics from the University of Madras, Chennai, India. He was a postdoctoral research associate and instructor at the Nova South-eastern University, Florida, in 2007. Then, he joined as a scientist at Nichi-in center for regenerative Medicine, Chennai, India. Since August 2008, he has been an assistant professor at Department of Bioinformatics, Alagappa University, Karaikudi. His area of interest is pharmacogenomics and computer-aided drug discovery.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.