Review Article

DATABASES DEVELOPED IN INDIA FOR BIOLOGICAL SCIENCES

Gitanjali Yadav[1] and Debasisa Mohanty[2*]

[1]*National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi-110067, India*
[2]*National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi-110067, India*

Abstract: The complexity of biological systems requires use of a variety of experimental methods with ever increasing sophistication to probe various cellular processes at molecular and atomic resolution. The availability of technologies for determining nucleic acid sequences of genes and atomic resolution structures of biomolecules prompted development of major biological databases like GenBank and PDB almost four decades ago. India was one of the few countries to realize early, the utility of such databases for progress in modern biology/biotechnology. Department of Biotechnology (DBT), India established Biotechnology Information System (BTIS) network in late eighties. Starting with the genome sequencing revolution at the turn of the century, application of high-throughput sequencing technologies in biology and medicine for analysis of genomes, transcriptomes, epigenomes and microbiomes have generated massive volumes of sequence data. BTIS network has not only provided state of the art computational infrastructure to research institutes and universities for utilizing various biological databases developed abroad in their research, it has also actively promoted research and development (R&D) projects in Bioinformatics to develop a variety of biological databases in diverse areas. It is encouraging to note that, a large number of biological databases or data driven software tools developed in India, have been published in leading peer reviewed international journals like Nucleic Acids Research, Bioinformatics, Database, BMC, PLoS and NPG series publication. Some of these databases are not only unique, they are also highly accessed as reflected in number of citations. Apart from databases developed by individual research groups, BTIS has initiated consortium projects to develop major India centric databases on *Mycobacterium tuberculosis*, Rice and Mango, which can potentially have practical applications in health and agriculture. Many of these biological databases have also helped in development of novel data mining methods, prediction strategies and data driven application software or web servers. In this article, we give an overview of biological databases developed in India and their impact on data driven research in biology. We also provide some suggestions for planning training programs in biological data science for making transitions to big data revolution in biology by combining advanced techniques like Deep Learning with biological big data.

Keywords: Biological big data; Database development; Indian efforts; Biomedical databases; plant databases; BTIS network

1. Introduction

A variety of experimental methods with ever increasing sophistication are being used to probe complex biological systems at molecular and atomic

resolution. Applications of high-throughput sequencing technologies in biology and medicine for analysis of genomes, transcriptomes, epigenomes and microbiomes have generated massive volumes of sequence data. Advances in mass spectrometry based techniques have also helped in generating huge volumes of proteomic data and together they constitute a major component of big data in biology (Li and Chen, 2014). High-throughput studies have also generated

large data sets on interactomes involving protein-protein, protein-DNA, and protein-RNA interactions. Similarly, apart from biomolecular structure data obtained from X-ray crystallography and NMR studies, in recent years developments in cryo-electron microscopy have facilitated structural elucidation of large macromolecular complexes. Data driven research has been used as a novel strategy in modern biology to convert this massive high throughput data into biologically meaningful knowledge for solving practical problems. Data driven research (Figure 1) typically involves, development of databases for organization, storage and retrieval of biological data from experimental studies, development of powerful data analytics or softwares for analysis of the data to generate testable hypothesis and subsequent cycles of experimental studies for validation of hypothesis and generation of new data. Therefore, databases and analytical tools for data management, visualization, integration, analysis, modeling, and prediction are central to modern biology.

As volume of sequence and structural data on biomolecular systems slowly started increasing, India was one of the few countries to realize early, the utility of such databases for progress in modern biology/biotechnology. Department of Biotechnology (DBT), India established Biotechnology Information System (BTIS) network in late eighties to facilitate access to databases and promote use of computational methods in biology(Krishnaswamy and Madhan Mohan, 2016). India is currently making huge strides in the area of generating new biological data as well as development of computational resources for accessing this data towards applications in biology, chemistry, pharmaceuticals, medicine and agriculture. Today development of databases and software for data analysis/visualization is an integral part of modern biology research. All major journals in biological sciences routinely publish articles on development or novel usage of databases and software (Galperin *et al.,* 2017). In this article, we give an overview of biological databases developed in India and their impact on data driven research in biology. We also highlight how these advances in bioinformatics research are opening up new career opportunities for young researchers to become data analysts, deep learning experts, database curators, phylogeneticists, software developers, programmers, chemi-informaticians and much more.

2. Peer reviewed publications describing biological databases developed in India

Since peer reviewed publications describing development of databases have been a common practice in current bioscience literature, we carried out searches in PUBMED for publication having 'India' in Affiliation field and terms 'database', 'repository', 'server' or 'software' in Title field. Interestingly this resulted in 574 database related publications from India till July 2017. Most of these publications started appearing in late nineties and have increased to 50 or more papers per year during last 5-6 years. Table 1 provides the information for comparison of the number of database related publications from India with those from other countries. While number of such publications from India is higher than those from South Korea, Israel and Singapore, countries like China, Japan, UK and USA are publishing significantly higher number of papers compared to India. This may also be related to number of research groups working in bioinformatics or biological data science in different countries. It may also be noted that, this analysis is just to explore the trend in database related publications word wide, the numbers shown may be lower than the actual number of database related publications because of under representation of plant and agriculture related journals in PUBMED. We also analyzed the type of journals in which database related publications from India were published (Table 2). It is encouraging to note that, there are 106 database or web server publications from India in Nucleic Acids Research which regularly publishes articles on all major databases or web servers in the area of bioscience or biotechnology. Similarly a total of 318 database related publications have appeared in other leading high impact journals, thus reflecting high quality of database related publications from India. The analysis of citations of database related publications from India also revealed many of these databases are being extensively used by a variety of research group from India as well as abroad. Since covering all these databases is beyond the scope of any single review article, we briefly describe below some of the representative biological databases developed in India covering broad areas of biological sciences.

3. Biological Databases developed in India

The various biological databases can be broadly divided into two categories, namely primary databases or data repositories and secondary

Table 1
Number of database, repository, software or web server related publications in PUBMED

Country	Number of database or software related publications
USA	7870
UK	2204
JAPAN	1374
CHINA	1264
INDIA	574
SOUTH KOREA	483
ISRAEL	232
SINGAPORE	190

Table 2
Number of biological database or web server publications from India in leading international journals

Country	Number of publications
NAR (Database Issue) 2001-2017	63
NAR (Webserver Issue) 2003-2017	43
PLoS One	53
Database	52
Bioinformatics	20
Scientific Reports	18
BMC Bioinformatics	17
BMC Genomics	15
Plant Methods	16
Molecular Plant	21
Total	318

Figure 1: **Flowchart describing data driven research in biological sciences**

databases or value added databases. The primary databases are essentially repositories of data directly obtained from an experimental study, while value added secondary databases typically obtain primary data from other repositories and store information about results from variety of computational analysis carried out on primary data. Currently India does not have any major primary data repositories like GenBank, UniProt or PDB. Most of the biological databases developed in India will belong to the secondary database category. However, they cater to niche areas in biology and are extremely useful for providing answers to specific biological questions. BTIS network of DBT has been a major facilitator for development of biological databases in India and many of the databases have been developed in BTIS centers. Outside BTIS network individual investigators supported by R&D projects from DBT have also developed several unique and useful databases and web servers. Apart from DBT, which has promoted bioinformatics activity in India for close to three decades, during the last decade other research funding agencies like CSIR, ICMR, ICAR, DIT and Ministry of Environmental have also supported biological database development in several institutes and research laboratories. Here, we describe some of the examples of biological databases developed in India from three broad application domains of biological sciences.

3.1. Databases in the area of biomedical sciences

Some of the widely used Indian resources the area of biomedical sciences are databases on MHC binding peptides (MHCBN) (Lata *et al.*, 2009), B-cell epitopes (Bcipep) (Saha *et al.*, 2005), haptens (HaptenDB) (Singh *et al.*, 2006), antigens from pathogenic species (AntigenDB)(Ansari *et al.*, 2009), genes involved in cervix cancer (CCDB)(Agarwal *et al.*, 2011), tumor homing peptides (TumorHoPe) (Kapoor *et al.*, 2012), cancer drug resistance (CancerDR) (Kumar *et al.*, 2013b), anticancer peptides and proteins (CancerPPD) (Tyagi *et al.*, 2015), cell penetrating peptides (CPPsite) (Agrawal *et al.*, 2016), naturally occurring plant based anticancer compounds and targets (NPACT) and HIV inhibiting siRNAs (HIVsirDB) (Tyagi *et al.*, 2011). These value added databases have been developed at IMTECH, Chandigarh and together with their associated prediction tools, they cater to researchers in the area of immunology, vaccine design and cancer biology. The BTIS center at Pune
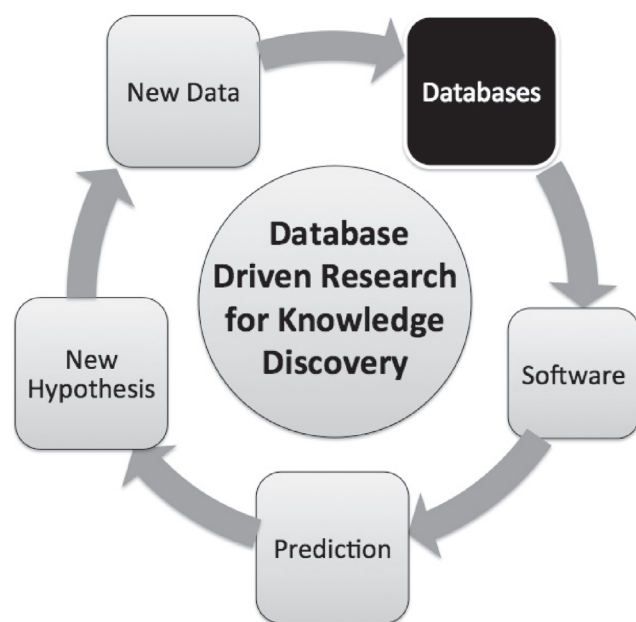
University has developed databases on conformational epitopes (CEP) (Kumar *et al.*, 2013a), antigen-antibody interaction (AgAbDb)(Kulkarni-Kale *et al.*, 2014), viral protein structure (ViPS), viral genome (VirGen) (Kulkarni-Kale *et al.*, 2004), animal virus information system and many other tools, which are extremely useful for researches in virology. Integrated drug target and chemical database for P. falciparum (PfalDB) (Kumar *et al.*, 2014a) and database of genetic polymorphisms in clinical isolates of M. Tb (MTCID) (Bharti *et al.*, 2012) are some of the major databases developed by the BTIS Center at JNU. The translational bioinformatics group at ICGEB, New Delhi has developed database of protozoan virulent proteins (protvirdb) (Ramana and Gupta, 2009), structural models of plasmodium proteins and resources for fungal adhesins. North-East India Helminth Parasite Information Database (NEIHPID) (Biswal *et al.*, 2016) developed by BTIS center at North Eastern Hill University (NEHU) is a major resource dedicated to parasite biology.

3.2. *Databases in the area of plant sciences*

National repositories like the Botanical Survey of India (BSI) and Environmental Information System (ENVIS) have historically been the major source of information on Indian flora, forest wealth, vegetable resources and medicinal plants (Pathak and Bharati, 2014). The onset of molecular and genomic era paved the way for large number of R&D projects in the area of plant molecular biology, stress genomics and proteomics, generating a huge volume of primary data over the past decades, both for crops and model plants. This is evident from a surge of databases on a variety of themes in plant sciences, ranging from agronomy and forestry to stress biology, nutri-genomics and emergent areas like epigenomes and cell-specific proteomes. A large number of databases have been developed for information on plant aromatics and phytochemicals for use in basic research, therapeutics, cosmetics or horticulture, and these may be useful as general information resources like EssOilDB (Kumari *et al.*, 2014) and Phytochemica (Pathania *et al.*, 2015a), or designed specifically for certain regions of India such as the North East (Meetei *et al.*, 2012) or focused on specific plant species (Pathania *et al.*, 2015b), or even ailments (Mangal *et al.*, 2013), (Vetrivel *et al.*, 2009). Taxon-specific databases allows users to gain access to diverse features for a given plant family of interest such as, for example legumes (CGWR

(Misra *et al.*, 2014), CicarVarDB (Doddamani *et al.*, 2015), CicarMiSatDB (Doddamani *et al.*, 2014), PvTFDB (Bonthala *et al.*, 2016)), grains (FMmDB (B *et al.*, 2013), SorghumFDB (Tian *et al.*, 2016)) or herbs (Kumar *et al.*, 2015). Indigenous Rhododendron and Orchid bioresources of North-east India and peninsular India have been documented by Sikkim State Council of Science & Technology and ICAR respectively. In the area of disease diagnostics, databases of genetic stocks and nutrient deficiencies in major Indian fruits and crops have been implemented at TNAU (Asaithambi *et al.*, 2013) and Gujarat University (Chauhan *et al.*, 2014). Many of the leading national centers in these efforts are funded by the BTISNET program of the DBT and there are numerous success stories of peer reviewed databases on plant specific gene families and other molecular level information, which have been published in leading international journals such as Nucleic Acids Research, Database, BMC Bioinformatics, Molecular Plant and Plant Methods. Theme-specific plant databases exist on algal feedstock (dEMBF (Misra *et al.*, 2016)), miRNAs (PmiRexat (Gurjar *et al.*, 2016)), marker genes (ChloroSSRDB (Kapil *et al.*, 2014), ChloroMitoSSRDB (Sablok *et al.*, 2015)), lectins (PlecDom (Shridhar *et al.*, 2009)), genome wide maps (IGMAP (Priya *et al.*, 2015)) and functional genomics (CicerTransDB (Gayali *et al.*, 2016), PpTFDB (Singh *et al.*, 2017), RiceSRTFDB (Priya and Jain, 2013)). Most of these have been developed at leading plant biology centers in India such as NIPGR, TNAU, ICRISAT, NRCPB, IITs and others.

3.3. *Databases on biomolecular structure, function and interaction networks*

BTIS centers and other research groups at IISc, NCBS, IOB, NII and IGIB have developed databases on structure-function of proteins and nucleic acids, protein-protein interactions and biosynthetic pathways. Databases on protein kinases in genomes (KinG) (Krupa, 2004), transmembrane helices in genomes (THGS) (Fernando *et al.*, 2004), protein super families (SUPFAM) (Pandit *et al.*, 2004), phylogeny and alignment of homologous proteins (PALI) (Balaji *et al.*, 2001), structure based sequence alignment (PASS2) (Gandhimathi *et al.*, 2012), length-variant protein domains (LenVarDB) (Mutt *et al.*, 2014) and remote homology of proteins (NrichD) (Mudgal *et al.*, 2015) are widely used for fold, superfamily and function annotation of proteins, while bioinformatics resources like PIC,

MolBridge (Kumar *et al.*, 2014b), ASSP(Kumar and Bansal, 2015), HELANAL-PLUS (Bansal *et al.*, 2000), CADB (Samaya Mohan *et al.*, 2005), 3dSS and SSEP (Shanthi *et al.*, 2003) are extremely useful for analysis of protein structures. Based on analysis of sequence dependent structural polymorphism of DNA using unique tools like NUPARM and NUCGEN-Plus (Bansal *et al.*, 1995), novel methods have been developed for promoter prediction and predicted promoters in different organisms have been cataloged in databases like EcoProm, PromBase and Plantcisprom. PLIC, PocketMatch (Yeturu and Chandra, 2008), MODPROPEP (Kumar and Mohanty, 2007) and modPDZpep (Sain and Mohanty, 2016) are data driven web resources for identification of ligand binding pockets in proteins and protein-peptide interactions. Databases like HPRD (Human Protein Reference Database) (Keshava Prasad *et al.*, 2009), Human Proteinpedia and NetPath (Kandasamy *et al.*, 2010) developed at IOB, Bangalore have been widely used for analysis of human proteome, protein-protein interactions and cell signaling. Databases on secondary metabolite biosynthetic pathways like NRPS-PKS (Ansari *et al.*, 2004), GTrDB (Kamra *et al.*, 2005), SBSPKS(Anand *et al.*, 2010) and RiPPDB (Agrawal *et al.*, 2017) developed at NII, New Delhi are valuable resources for identification of novel natural products by genome mining (Yadav *et al.*, 2009). Databases like LncRBase (Chakraborty *et al.*, 2014), lncRNome (Bhartiya *et al.*, 2013), dbSMR (Hariharan *et al.*, 2009) and miRvar (Bhartiya *et al.*, 2011) are some of the valuable resources developed in India for analysis of long noncoding RNA and microRNA.

3.4. *Major Indian databases developed by consortium projects*

Apart from the databases developed by individual research groups or BTIS centers, DBT has also funded development of three major databases under multi institutional consortium projects (Krishnaswamy and Madhan Mohan, 2016). NRDD, a database on rice germplasm and genomic information for ~15000 accessions is being developed by eight centers, namely NBPGR, NRCPB, CRRI, DRR-Hyderabad, BHU, CSSRI-Karnal, IGKV-Raipur and University of Delhi-South Campus. National database on *Mycobacterium tuberculosis* containing molecular level information on genes, proteins, PPI and PTMs is being developed by NJIL & OMID-Agra, JNU-New Delhi, TRI-Chennai and NTI, IOB, IISc from Bangalore.

Similarly CISH-Lucknow and nine other agricultural research institutes are developing a national database on Mango.

4. Impact of databases on research in biological sciences

Organization of huge volume of data from diverse sources and their analysis with suitable data analytics make hidden patterns in data apparent. Increase of biological databases in India has also facilitated development of novel data driven prediction methods for structure and function of proteins and nucleic acids, identification of epitopes for vaccine research and discovery of natural products by genome mining. Machine learning approaches like ANN and SVM are routinely using data compiled in these databases to train new prediction models and this is reflected in increase in the number of software developed in India as well as number of computational biology publications from India in the last decade. There are several success stories of experimental validation of *in silico* predictions and they have encouraged increasing number of collaborative studies involving integration of computation with wet biology.

5. Need for primary data repositories and long term support for database projects

Even though large number of biological databases developed in India are catering to unique and niche areas in biology, they are essentially secondary databases providing value additions to primary data taken from major international repositories. At present bioscience research community in India is generating huge volume of genome sequence, transcriptome, metabolome and interactome data and clinical research is also generating huge volume of data on population or disease specific variations. For example, Indian structural biology alone is generating more than 250 protein structures per year, while sequencing data will be much higher. Therefore, it is necessary that major data repositories in different areas of biology are set up in India so that Indian researchers play a major role in defining data standards. Development of data repositories in India will also facilitate greater international visibility of closely integrated secondary databases and data driven softwares developed in India. In contrast to only two or three databases which have been developed in India under consortium projects, majority of the

biological databases in India have been developed by individual research groups. Hence, there are often difficulties in keeping these information resources updated for long periods of time. Therefore, it is necessary that suitable action plans are formulated so that many of the databases remain up to date and are considered 'Gold Standard' databases in respective domains.

6. Conclusion and Future Prospects

Data driven knowledge discovery is fundamental for life science research across the globe. In this article we have made an attempt to assess the role played by Indian bioinformatics community towards development of cutting edge data resources, and to highlight the support of DBT in establishment of high-quality computational infrastructure at various centers across the country, resulting in creation of trained manpower for data driven research in biology. We have limited our observations primarily to databases instead of web-servers or online computational tools, many of which have indeed been developed in India, and are being used worldwide.

Overall, we found a large number of research centers engaged in resource development at all levels of system complexity. There is indeed a greater emphasis on functional genomics based databases in both biomedical and plant sciences, but such a trend is crucial for us to make our presence felt internationally. Most of these databases have been published in leading peer-reviewed journals in biological sciences. Our survey also revealed how collaborations between plant and biomedical scientists have led to emergence of databases on plant based remedies to human diseases through convergence of disciplines like nutrition and genomics in the era of genomic medicine. Similarly integration of data from genomics and metabolomics using tools from bioinformatics and chemo-informatics have resulted in novel resources for analysis of biosynthetic pathways. Further - omics based efforts, like metabolomics and microbiomics, are expected to contribute to multi disciplinary approaches for better understanding of small molecular based therapeutics. Many of the current databases on biosynthetic pathways or Protein-Protein interactions etc., have been developed through mining of published literature followed by manual curation. This task is time consuming and manpower intensive, which puts a limit on growth of databases. Developing databases by crowd sourcing with active involvement of experimental research groups involved in data generation should be explored as an alternate strategy.

Data science is a multi disciplinary subject. Ever increasing size of biological big data requires efficient storage and retrieval mechanisms, while powerful data mining techniques like deep learning are essential for knowledge discovery (Angermueller *et al.*, 2016). Development of user-friendly visualization tools using computer graphics is equally important for converting data to knowledge. Therefore, it is anticipated that, in the coming years biological big data will have tremendous impact on bioinformatics service industry by opening up new career opportunities for young professionals to become software developers (virtualization and cheap cloud database services), hardware engineers (SSDs, massively multi-core CPUs, GPUs), data analysts, deep learning experts, database curators, phylogeneticists, programmers and chemi-informaticians etc.

References

Agarwal, S.M., Raghav, D., Singh, H., and Raghava, G.P.S. (2011). CCDB: A curated database of genes involved in cervix cancer. Nucleic Acids Res. *39*, D975-D979.

Agrawal, P., Bhalla, S., Usmani, S.S., Singh, S., Chaudhary, K., Raghava, G.P.S., and Gautam, A. (2016). CPPsite 2.0: A repository of experimentally validated cell-penetrating peptides. Nucleic Acids Res. *44*, D1098–D1103.

Agrawal, P., Khater, S., Gupta, M., Sain, N., and Mohanty, D. (2017). RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. Nucleic Acids Res. *45*, W80–W88.

Anand, S., Prasad, M.V.R., Yadav, G., Kumar, N., Shehara, J., Ansari, M.Z., and Mohanty, D. (2010). SBSPKS: Structure based sequence analysis of polyketide synthases. Nucleic Acids Res. *38*, W487-W496.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol. *12*, 878.

Ansari, H.R., Flower, D.R., and Raghava, G.P.S. (2009). AntigenDB: An immunoinformatics database of

pathogen antigens. Nucleic Acids Res. *38*, D847-D853.

Ansari, M.Z., Yadav, G., Gokhale, R.S., and Mohanty, D. (2004). NRPS-PKS: A knowledge-based resource for analysis of NRPS-PKS megasynthases. Nucleic Acids Res. *32*,W405-W413

Asaithambi, D., Natesan, S., Venkatesan, V., Muthuraja, R., Muthusamy, K., and Vinayagam, P. (2013). MAGICdb – Mango Genetic stocks Identification and Characterisation database. Bioinformation *9*, 838–839.

B, V.S., Muthamilarasan, M., Misra, G., and Prasad, M. (2013). FmMDb: A Versatile Database of Foxtail Millet Markers for Millets and Bioenergy Grasses Research. PLoS One *8*, e71418.

Balaji, S., Sujatha, S., Kumar, S.S.C., and Srinivasan, N. (2001). PALI—a database of Phylogeny and ALIgnment of homologous protein structures. Nucleic Acids Res. *29*, 61–65.

Bansal, M., Bhattacharyya, D., and Ravi, B. (1995). NUPARM and NUCGEN: Software for analysis and generation of sequence dependent nucleic acid structures. Bioinformatics *11*, 281–287.

Bansal, M., Kumar, S., and Velavan, R. (2000). HELANAL: a program to characterize helix geometry in proteins. J. Biomol. Struct. Dyn. *17*, 811–819.

Bharti, R., Das, R., Sharma, P., Katoch, K., and Bhattacharya, A. (2012). MTCID: A database of genetic polymorphisms in clinical isolates of Mycobacterium tuberculosis. Tuberculosis *92*, 166–172.

Bhartiya, D., Laddha, S. V., Mukhopadhyay, A., and Scaria, V. (2011). miRvar: A comprehensive database for genomic variations in microRNAs. Hum. Mutat. *32*, E2226-2245.

Bhartiya, D., Pal, K., Ghosh, S., Kapoor, S., Jalali, S., Panwar, B., Jain, S., Sati, S., Sengupta, S., Sachidanandan, C., *et al.* (2013). LncRNome: A comprehensive knowledgebase of human long noncoding RNAs. Database *2013*, bat034.

Biswal, D.K., Debnath, M., Kharumnuid, G., Thongnibah, W., and Tandon, V. (2016). Northeast India Helminth Parasite Information Database (NEIHPID): Knowledge Base for Helminth Parasites. PLoS One *11*, e0157459.

Bonthala, V.S., Gajula, M.N. V, and Others (2016). PvTFDB: a Phaseolus vulgaris transcription factors database for expediting functional genomics in legumes. Database *2016*, baw114

Chakraborty, S., Deb, A., Maji, R.K., Saha, S., and Ghosh, Z. (2014). LncRBase: An enriched resource for lncRNA information. PLoS One *9*, e108010.

Chauhan, R., Jasrai, Y., Pandya, H., Chaudhari, S., and Samota, C.M. (2014). FCDD: A Database for Fruit Crops Diseases. Bioinformation *10*, 595.

Doddamani, D., Katta, M.A., Khan, A.W., Agarwal, G., Shah, T.M., and Varshney, R.K. (2014). CicArMiSatDB: the chickpea microsatellite database. BMC Bioinformatics *15*, 212.

Doddamani, D., Khan, A.W., Katta, M.A.V.S.K., Agarwal, G., Thudi, M., Ruperao, P., Edwards, D., and Varshney, R.K. (2015). CicArVarDB: SNP and InDel database for advancing genetics research and breeding applications in chickpea. Database *2015*, bav078.

Fernando, S. a, Selvarani, P., Das, S., Kumar, C.K., Mondal, S., Ramakumar, S., and Sekar, K. (2004). THGS: a web-based database of Transmembrane Helices in Genome Sequences. Nucleic Acids Res. *32*, D125-8.

Galperin, M.Y., Fernández-Suárez, X.M., and Rigden, D.J. (2017). The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic Acids Res. *45*, 5627.

Gandhimathi, A., Nair, A.G., and Sowdhamini, R. (2012). PASS2 version 4: An update to the database of structure-based sequence alignments of structural domain superfamilies. Nucleic Acids Res. *40*, D531-4.

Gayali, S., Acharya, S., Lande, N.V., Pandey, A., Chakraborty, S., and Chakraborty, N. (2016). CicerTransDB 1.0: a resource for expression and functional study of chickpea transcription factors. BMC Plant Biol. *16*, 169.

Gurjar, A.K.S., Panwar, A.S., Gupta, R., and Mantri, S.S. (2016). PmiRExAt: Plant miRNA expression atlas database and web applications. Database *2016*, baw060.

Hariharan, M., Scaria, V., and Brahmachari, S.K. (2009). dbSMR: a novel resource of genome-wide SNPs affecting microRNA mediated regulation. BMC Bioinformatics *10*, 108.

Kamra, P., Gokhale, R.S., and Mohanty, D. (2005). SEARCHGTr: a program for analysis of glycosyltransferases involved in glycosylation of secondary metabolites. Nucleic Acids Res. *33*, W220-5.

Kandasamy, K., Mohan, S.S., Raju, R., Keerthikumar, S., Kumar, G.S.S., Venugopal, A.K., Telikicherla, D., Navarro, J.D., Mathivanan, S., Pecquet, C., *et al.* (2010). NetPath: a Public Resource of Curated Signal Transduction Pathways. Genome Biol. *11*, R3.

Kapil, A., ant Rai, P.K., and Shanker, A. (2014). ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence repeats (cpSSRs) of green plants. Database (Oxford). *2014*, bau107.

Kapoor, P., Singh, H., Gautam, A., Chaudhary, K., Kumar, R., and Raghava, G.P.S. (2012). Tumorhope: A database of tumor homing peptides. PLoS One *7*, e35187.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.* (2009). Human Protein Reference Database - 2009 update. Nucleic Acids Res. *37*, D767-72.

Krishnaswamy, S., and Madhan Mohan, T. (2016). The largest distributed network of bioinformatics centres in the world: Biotechnology Information System Network (DBT-BTISNET). Curr. Sci. *110*, 556–561.

Krupa, A. (2004). KinG: a database of protein kinases in genomes. Nucleic Acids Res. *32*, 153D–155.

Kulkarni-Kale, U., Bhosle, S., Manjari, G.S., and Kolaskar,

a S. (2004). VirGen: a comprehensive viral genome resource. Nucleic Acids Res *32*, D289–92.

Kulkarni-Kale, U., Raskar-Renuse, S., Natekar-Kalantre, G., and Saxena, S.A. (2014). Antigen-Antibody Interaction Database (AgAbDb): A compendium of antigen-antibody interactions. Methods Mol. Biol. *1184*, 149–164.

Kumar, N., and Mohanty, D. (2007). MODPROPEP: A program for knowledge-based modeling of protein-peptide complexes. Nucleic Acids Res. *35*, W549-55.

Kumar, P., and Bansal, M. (2015). Identification of local variations within secondary structures of proteins. Acta Crystallogr. D. Biol. Crystallogr. *71*, 1077–1086.

Kumar, A., Rajendran, V., Sethumadhavan, R., and Purohit, R. (2013a). CEP proteins: The knights of centrosome dynasty. Protoplasma *250*, 965–983.

Kumar, A., Agarwal, N., Pant, L., Singh, J.P., Ghosh, I., and Subbarao, N. (2014a). Pfaldb: An integrated drug target and chemical database for Plasmodium falciparum. Curr. Drug Targets *15*, 1089–1093.

Kumar, A., Chetia, H., Sharma, S., Kabiraj, D., Talukdar, N.C., and Bora, U. (2015). Curcumin Resource Database. Database *2015*, bav070.

Kumar, P., Kailasam, S., Chakraborty, S., and Bansal, M. (2014b). MolBridge: A program for identifying nonbonded interactions in small molecules and biomolecular structures. J. Appl. Crystallogr. *47*, 1772–1776.

Kumar, R., Chaudhary, K., Gupta, S., Singh, H., Kumar, S., Gautam, A., Kapoor, P., and Raghava, G.P.S. (2013b). CancerDR: Cancer Drug Resistance Database. Sci. Rep. *3*, 1445.

Kumari, S., Pundhir, S., Priya, P., Jeena, G., Punetha, A., Chawla, K., Jafaree, Z.F., Mondal, S., and Yadav, G. (2014). EssOilDB: A database of essential oils reflecting terpene composition and variability in the plant kingdom. Database *2014*, bau120.

Lata, S., Bhasin, M., and Raghava, G.P. (2009). MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. BMC Res. Notes *2*, 61.

Li, Y., and Chen, L. (2014). Big Biological Data: Challenges and Opportunities. Genomics. Proteomics Bioinformatics *12*, 187–189.

Mangal, M., Sagar, P., Singh, H., Raghava, G.P.S., and Agarwal, S.M. (2013). NPACT: Naturally occurring plant-based anti-cancer compound-activity-target database. Nucleic Acids Res. *41*, D1124-9.

Meetei, P.A., Singh, P., Nongdam, P., Prabhu, N.P., Rathore, R.S., and Vindal, V. (2012). NeMedPlant: a database of therapeutic applications and chemical constituents of medicinal plants from north-east region of India. Bioinformation *8*, 209.

Misra, G., Priya, P., Bandhiwal, N., Bareja, N., Jain, M., Bhatia, S., Chattopadhyay, D., Tyagi, A.K., and Yadav, G. (2014). The chickpea genomic web resource: visualization and analysis of the desi-type Cicer arietinum nuclear genome for comparative exploration of legumes. BMC Plant Biol. *14*, 315.

Misra, N., Panda, P.K., Parida, B.K., and Mishra, B.K. (2016). dEMBF: A Comprehensive Database of Enzymes of Microalgal Biofuel Feedstock. PLoS One *11*, e0146158.

Mudgal, R., Sandhya, S., Kumar, G., Sowdhamini, R., Chandra, N.R., and Srinivasan, N. (2015). NrichD database: Sequence databases enriched with computationally designed protein-like sequences aid in remote homology detection. Nucleic Acids Res. *43*, D300–D305.

Mutt, E., Mathew, O.K., and Sowdhamini, R. (2014). LenVarDB: Database of length-variant protein domains. Nucleic Acids Res. *42*, D246-50.

Pandit, S.B., Bhadra, R., Gowri, V.S., Balaji, S., Anand, B., and Srinivasan, N. (2004). SUPFAM: a database of sequence superfamilies of protein domains. BMC Bioinformatics *5*, 28.

Pathak, M., and Bharati, K.A. (2014). Botanical survey of india (1971-2010): A scientometric analysis. Curr. Sci. *106*, 964–971.

Pathania, S., Ramakrishnan, S.M., and Bagler, G. (2015a). Phytochemica: A platform to explore phytochemicals of medicinal plants. Database *2015*, bav075.

Pathania, S., Ramakrishnan, S.M., Randhawa, V., and Bagler, G. (2015b). SerpentinaDB: a database of plant-derived molecules of Rauvolfia serpentina. BMC Complement. Altern. Med. *15*, 262.

Priya, P., and Jain, M. (2013). RiceSRTFDB: A database of rice transcription factors containing comprehensive expression, cis-regulatory element and mutant information to facilitate gene function analysis. Database *2013*, bat027.

Priya, P., Bandhiwal, N., Misra, G., Mondal, S., and Yadav, G. (2015). IGMAP: an interactive mapping and clustering platform for plants. Mol. Plant *8*, 818–821.

Ramana, J., and Gupta, D. (2009). ProtVirDB: A database of protozoan virulent proteins. Bioinformatics *25*, 1568–1569.

Sablok, G., Padma Raju, G. V, Mudunuri, S.B., Prabha, R., Singh, D.P., Baev, V., Yahubyan, G., Ralph, P.J., and Porta, N. La (2015). ChloroMitoSSRDB 2.00: more genomes, more repeats, unifying SSRs search patterns and repeat detection. Database *2015*, bav084.

Saha, S., Bhasin, M., and Raghava, G.P.S. (2005). Bcipep: a database of B-cell epitopes. BMC Genomics *6*, 79–85.

Sain, N., and Mohanty, D. (2016). modPDZpep: a web resource for structure based analysis of human PDZ-mediated interaction networks. Biol. Direct *11*, 48.

Samaya Mohan, K., Sheik, S.S., Ramesh, J., Balamurugan, B., Jeyasimhan, M., Mayilarasi, C., and Sekar, K. (2005). CADB-2.0: Conformation angles database. Acta Crystallogr. Sect. D Biol. Crystallogr. *61*, 637–639.

Shanthi, V., Selvarani, P., Kumar, C.K., Mohire, C.S., and Sekar, K. (2003). SSEP: Secondary structural elements of proteins. Nucleic Acids Res. *31*, 3404–3405.

Shridhar, S., Chattopadhyay, D., and Yadav, G. (2009). PLecDom: A program for identification and analysis of plant lectin domains. Nucleic Acids Res. *37*, W452-

8.

Singh, A., Sharma, A.K., Singh, N.K., and Sharma, T.R. (2017). PpTFDB: A pigeonpea transcription factor database for exploring functional genomics in legumes. PLoS One *12*, e0179736.

Singh, M.K., Srivastava, S., Raghava, G.P.S., and Varshney, G.C. (2006). HaptenDB: A comprehensive database of haptens, carrier proteins and anti-hapten antibodies. Bioinformatics *22*, 253–255.

Tian, T., You, Q., Zhang, L., Yi, X., Yan, H., Xu, W., and Su, Z. (2016). SorghumFDB: sorghum functional genomics database with multidimensional network analysis. Database *2016*, baw099.

Tyagi, A., Ahmed, F., Thakur, N., Sharma, A., Raghava, G.P.S., and Kumar, M. (2011). HIVsirDB: A database of HIV inhibiting siRNAs. PLoS One *6*, e25917.

Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., Joshi, A., Singh, S., Gautam, A., and Raghava, G.P.S. (2015). CancerPPD: A database of anticancer peptides and proteins. Nucleic Acids Res. *43*, D837–D843.

Vetrivel, U., Subramanian, N., and Pilla, K. (2009). InPACdb Indian plant anticancer compounds database. Bioinformation *4*, 71.

Yadav, G., Gokhale, R.S., and Mohanty, D. (2009). Towards prediction of metabolic products of polyketide synthases: an in silico analysis. PLoS Comput. Biol. *5*, e1000351.

Yeturu, K., and Chandra, N. (2008). PocketMatch: A new algorithm to compare binding sites in protein structures. BMC Bioinformatics *9*, 543.