doi: 10.1038/nindia.2015.118

Protein databases from India

Harsha Gowda[1,2], Aditi Chatterjee[1,2] & T. S. Keshava Prasad[1,2,3]

Harsha Gowda

Aditi Chatterjee

T. S. Keshava Prasad

In recent years, there has been an exponential surge in biological data. The advent of high-resolution instruments with better sensitivities and new-age software suites have made it easier for data analysts and researchers to use omics approaches in biomedical investigations. Using such approaches, researchers are generating vast amounts of molecular data that provide a platform for several novel hypotheses.

Biological databases play an important role in assimilating large amounts of data and enabling users to access this information using dynamic query systems. These diverse datasets also serve as a medium to identify enzyme-substrate- and metabolic- networks involved in various biological processes, which can be helpful in molecular diagnostics and therapeutics for human diseases; and genomic selection of better biological traits in crop plants and dairy animals. Databases also provide a scaffold for the execution of regular updates, international data exchange and global meta-analysis.

Databases of protein features, resources for signaling and metabolic pathways that drive specific biological processes, and repositories of disease-specific molecular level alterations will help researchers apply systems biology approaches to unearth mechanisms or leads with greater biological significance.

Large protein databases

Leading the Indian proteomic database revolution is Bangalore-based Institute of Bioinformatics (IOB; http://ibioinformatics.org) with its flagship Human Protein Reference Database (HPRD) and other noteworthy platforms – the Human Proteinpedia, Human Proteome Map, a database on molecular alterations reported in pancreatic cancers and several human signaling pathways.

The HPRD (http://www.hprd.org/) has highly curated information on a non-redundant set of 30,047 human proteins, which include 40, 042 protein-protein interactions and 1, 09, 518 post-translational modifications[1]. A number of biomedical scientists use HPRD either directly by downloading data or indirectly by using the information available in RefSeq and Entrez Gene databases of the National Center for Biotechnology Information (NCBI) and University of California Santa Cruz (UCSC) genome browser.

The Human Proteinpedia (http://www.humanproteinpedia.org) provided a platform for over 200 proteomic laboratories for storing, sharing and dissemination of multidimensional proteomic datasets, even before publication[2, 3]. Data from Human Proteinpedia is also made available to the larger scientific community through HPRD.

Human Proteome Map (http://humanproteomemap.org/) is another interactive web portal, which represents the largest mass spectrometry- derived label-free quantitative proteomic data from 30 different human tissues[4]. The Plasma Proteome Database (PPD, http://plasmaproteomedatabase.org/), initiated as a part of the Human Plasma Proteome Project of Human Proteome Organization, contains information on 10, 546 proteins detected in human serum/plasma[5].

The institute has also created a highly curated database on molecular alterations reported in pancreatic cancers, initially published as a compendium of overexpressed proteins, and followed up with the database of molecular alteration at mRNA, protein and miRNA[6].

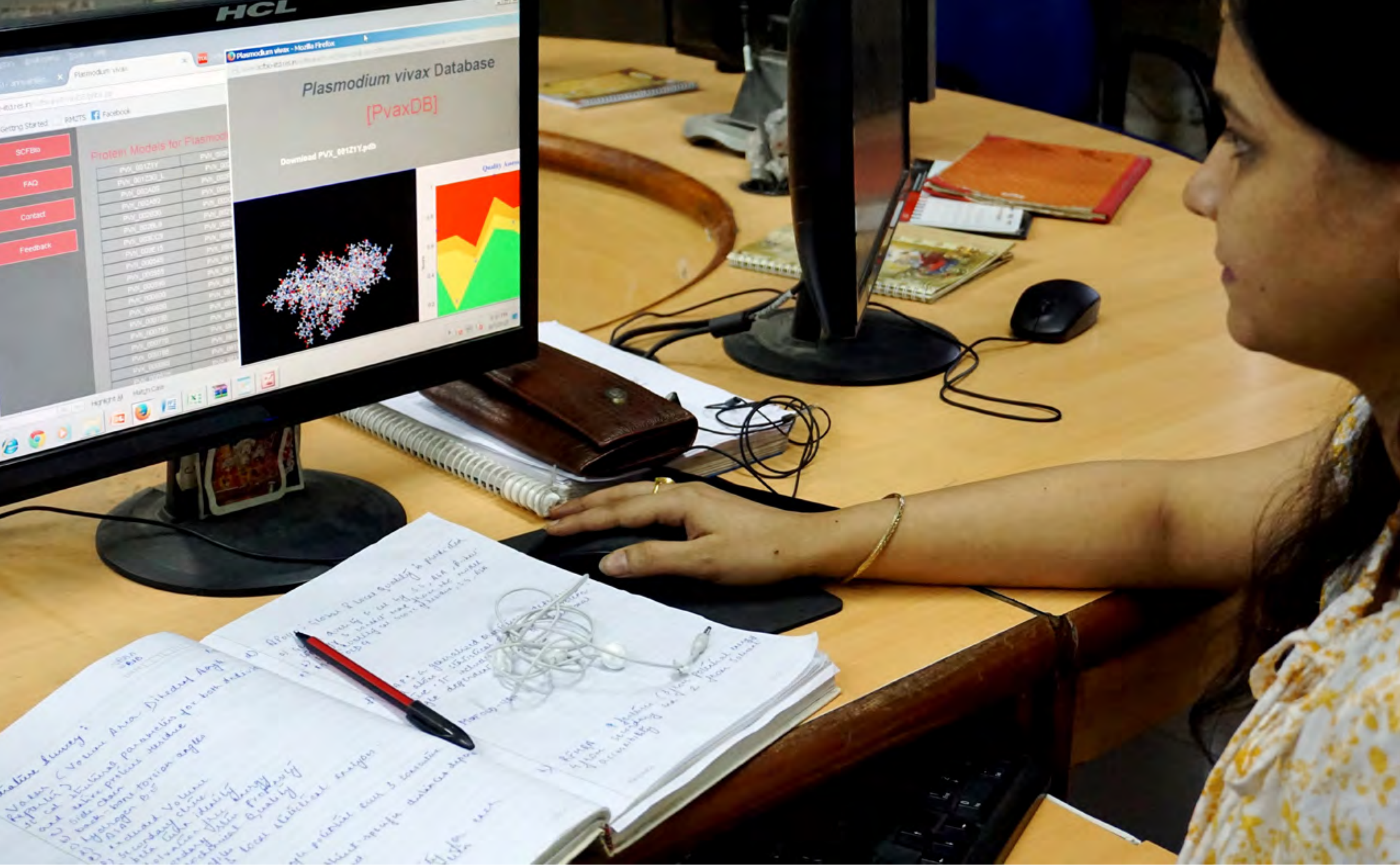
IOB's centralised resource for human signaling pathways is called NetPath (http://www.netpath.org). NetPath contains manually curated data for 36 signaling pathways including prolactin[7], gastrin[8], corticotropin-releasing hormone (CRH)[9], fibroblast growth factor-1 (FGF1)[10], interleukins (IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-9, and IL-11), brain-derived neurotrophic factor (BDNF)[11], leptin[12], oncostatin-M[13] and RANKL[14], among others. IOB scientists have curated several signaling pathways such as Delta-Notch, EGFR1, Hedgehog, TNF-alpha and Wnt for Cancer Cell Map (http://cancer.cellmap.org/cellmap/), which is a database of human cancer focused pathways developed by Memorial Sloan-Kettering Cancer Center in New York, USA.

A slimmer version of signaling pathways annotated in NetPath were gathered to form NetSlim (http://www.netpath.org/netslim/), which comprises a graphical network of core signaling reactions[15].

The Bioinformatics Centre at Institute of Microbial Technology in Chandigarh has created many web servers and protein databases to perform structure and function of proteins based on their amino acid sequences, potential MHC class I and II binding regions in antigens, subcellular localisation and classification of eukaryotic and prokaryotic proteins, identification of bacterial toxins and several analytical tools for pattern finding in genome annotation. The IMT group has created a curated database of proteins associated with cervix cancer – CCDB[16]; a database of anticancer peptides and proteins called CancerPPD[17], and a database of hemolytic and non-hemolytic peptides - Hemolytik.

The National Centre of Biological Sciences (NCBS) in Bangalore has created a number of publicly available databases of protein

[1]Institute of Bioinformatics, International Technology Park, Bangalore, India (harsha@ibioinformatics.org, aditi@ibioinformatics.org, keshav@ibioinformatics.org). [2]YU-IOB Center for Systems Biology and Molecular Medicine, Yenepoya University, Mangalore, India. [3]NIMHANS-IOB Proteomics and Bioinformatics Laboratory, Neurobiology Research Centre, National Institute of Mental Health and Neurosciences, Bangalore, India.



family trees by analysis of protein motifs and domains. These resources include SMOS.2, LenVarDB, 3PFDB, SUPFAM; GenDiS[18], MegaMotifbase, iMOTdb and PASS2[19]. A database on disulphide bonds (DSDBASE)[20], A database of olfactory receptors (DOR)[21] and a resource for transcription factors responding to stress in Arabidopsis thaliana (STIFDB)[22] were also developed by the team.

NCBS has also developed the Database of Quantitative Cellular Signaling (DOQCS), which represents the collection of basic models of signaling pathways[23, 24, 25, 26].

Researchers at the Indian Institute of Science have created a number of protein databases related to structure and function of protein kinases. These include 'KinG', a database of kinases[27], NrichD28, PALI, PRODOC and MulPSSM[29], resources of protein domains and alignments.

Databases for analysis of secondary structures of proteins, which include Conformation Angles DataBase of proteins (CADB), a web-based database of Transmembrane Helices in Genome Sequences (THGS)[30] and Secondary Structural Elements of Proteins (SSEP)[31] have also been received well by researchers worldwide.

At the Center of Bioinformatics in Pondicherry University, researchers have developed a number of protein databases including Peptide Binding Protein Database, Immune Epitope Prediction Database & Tools, Structural Epitope Database (SEDB)[32], Clostridium-DT(DB)[33], a comprehensive database for potential drug targets of *Clostridium difficile* and Viral Protein Structural Database (VPDB)[34].

A manually curated database of rice proteins (http://www.genomeindia.org/biocuration) was developed by a team from the University of Delhi South Campus and is an important plant protein database from India[35].

Future outlook

India is uniquely positioned to take a lead in developing and maintaining world class biological databases. A large base of human resource in biological sciences as well as software technology serves as an advantage. Although there are independent efforts from different research labs in India to build and maintain biological databases, there hasn't been a dedicated effort to do it at a scale that is required to build and maintain world class databases much like how NCBI and EBI are doing for several years.

There are several companies in India that are offering annotation and database services to industries, which clearly demonstrates existence of such capability in India. Much of the data that is being generated within India is also not organised for the immediate use by fellow researchers.

Dedicated funding from government and corporates for academic research groups with demonstrated capability in developing and maintaining world class databases could be a good starting point. Such resources will become a necessity in the future as the amount of data being generated continues to grow.

References

1. Prasad, T. S .K. *et al*. Human Protein Reference Database 2009 update. *Nucleic Acids Res.* **37**, D767-772 (2009).
2. Kandasamy, K. *et al*. Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res.* **37**, D773-781 (2009).
3. Mathivanan, S. *et al*. Human Proteinpedia enables sharing of human protein data. *Nat. Biotechnol.* **26**, 164-167 (2008).
4. Kim, M. S. *et al*. A draft map of the human proteome. *Nature* **509**, 575-581 (2014).
5. Nanjappa, V. *et al*. Plasma Proteome Database as a resource for proteomics research: 2014 update. *Nucleic Acids Res.* **42**, D959-965 (2014).

6. Thomas, J. K. *et al.* Pancreatic Cancer Database: an integrative resource for pancreatic cancer. *Cancer Biol. Ther.* **15**, 963-967 (2014).

7. Radhakrishnan, A. *et al.* A pathway map of prolactin signaling. *J. Cell Commun. Signal* **6**, 169-173 (2012).

8. Subbannayya, Y. *et al.* A network map of the gastrin signaling pathway. *J. Cell Commun. Signal.* **8**, 165-170 (2014).

9. Subbannayya, T. *et al.* An integrated map of corticotropin-releasing hormone signaling pathway. *J. Cell Commun. Signal.* **7**, 295-300 (2013).

10. Raju, R. *et al.* A network map of FGF-1/FGFR signaling system. *J. Signal Transduct.* 962962 (2014).

11. Sandhya, V. K. *et al.* A network map of BDNF/TRKB and BDNF/p75NTR signaling system. *J. Cell Commun. Signal.* **7**, 301-307 (2013).

12. Nanjappa, V. *et al.* A comprehensive curated reaction map of leptin signaling pathway. *J. Proteomics Bioinformatics.* **4**, 184-189 (2011).

13. Dey, G. *et al.* Signaling network of Oncostatin M pathway. *J. Cell Commun. Signal.* **7**, 103-108 (2013).

14. Raju, R. *et al.* A comprehensive manually curated reaction map of RANKL/RANK-signaling pathway. *Database (Oxford).* 2011, bar021 (2011).

15. Raju, R. *et al.* NetSlim: high-confidence curated signaling maps. *Database (Oxford).* 2011, bar032 (2011).

16. Agarwal, S. M. *et al.* CCDB: a curated database of genes involved in cervix cancer. *Nucleic Acids Res.* **39**, D975-979 (2011).

17. Tyagi, A. *et al.* CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res.* **43**, D837-843 (2015).

18. Pugalenthi, G. *et al.* GenDiS: Genomic distribution of protein structural domain superfamilies. *Nucleic Acids Res.* **33**, D252-255 (2005).

19. Gandhimathi, A. *et al.* PASS2 version 4: an update to the database of structure-based sequence alignments of structural domain superfamilies. *Nucleic Acids Res.* **40**, D531-534 (2012).

20. Vinayagam, A. *et al.* DSDBASE: a consortium of native and modelled disulphide bonds in proteins. *Nucleic Acids Res.* **32**, D200-202 (2004).

21. Nagarathnam, B. *et al.* DOR - a Database of olfactory receptors — integrated repository for sequence and secondary structural information of olfactory receptors in selected eukaryotic renomes. *Bioinform. Biol. Insights.* **8**, 147-158 (2014).

22. Shameer, K. *et al.* STIFDB-Arabidopsis stress responsive transcription factor database. *Int. J. Plant Genomics.* 583429 (2009).

23. Bhalla, U. S. & Iyengar, R. Robustness of the bistable behavior of a biological signaling feedback loop. *Chaos.* **11**, 221-226 (2001).

24. Bhalla, U. S. & Iyengar, R. Functional modules in biological signalling networks. *Novartis Found. Symp.* **239**, 4-13; discussion 13-15, 45-51 (2001).

25. Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science.* **283**, 381-387 (1999).

26. Sivakumaran, S. *et al.* The database of quantitative cellular signaling: management and analysis of chemical kinetic models of signaling networks. *Bioinformatics.* **19**, 408-415 (2003).

27. Krupa, A. *et al.* KinG: a database of protein kinases in genomes. *Nucleic Acids Res.* **32**, D153-155 (2004).

28. Mudgal, R. *et al.* NrichD database: sequence databases enriched with computationally designed protein-like sequences aid in remote homology detection. *Nucleic Acids Res.* **43**, D300-305 (2015).

29. Gowri, V. S. *et al.* MulPSSM: a database of multiple position-specific scoring matrices of protein domain families. *Nucleic Acids Res.* **34**, D243-246 (2006).

30. Fernando, S. A. *et al.* THGS: a web-based database of transmembrane helices in genome sequences. *Nucleic Acids Res.* **32**, D125-128 (2004).

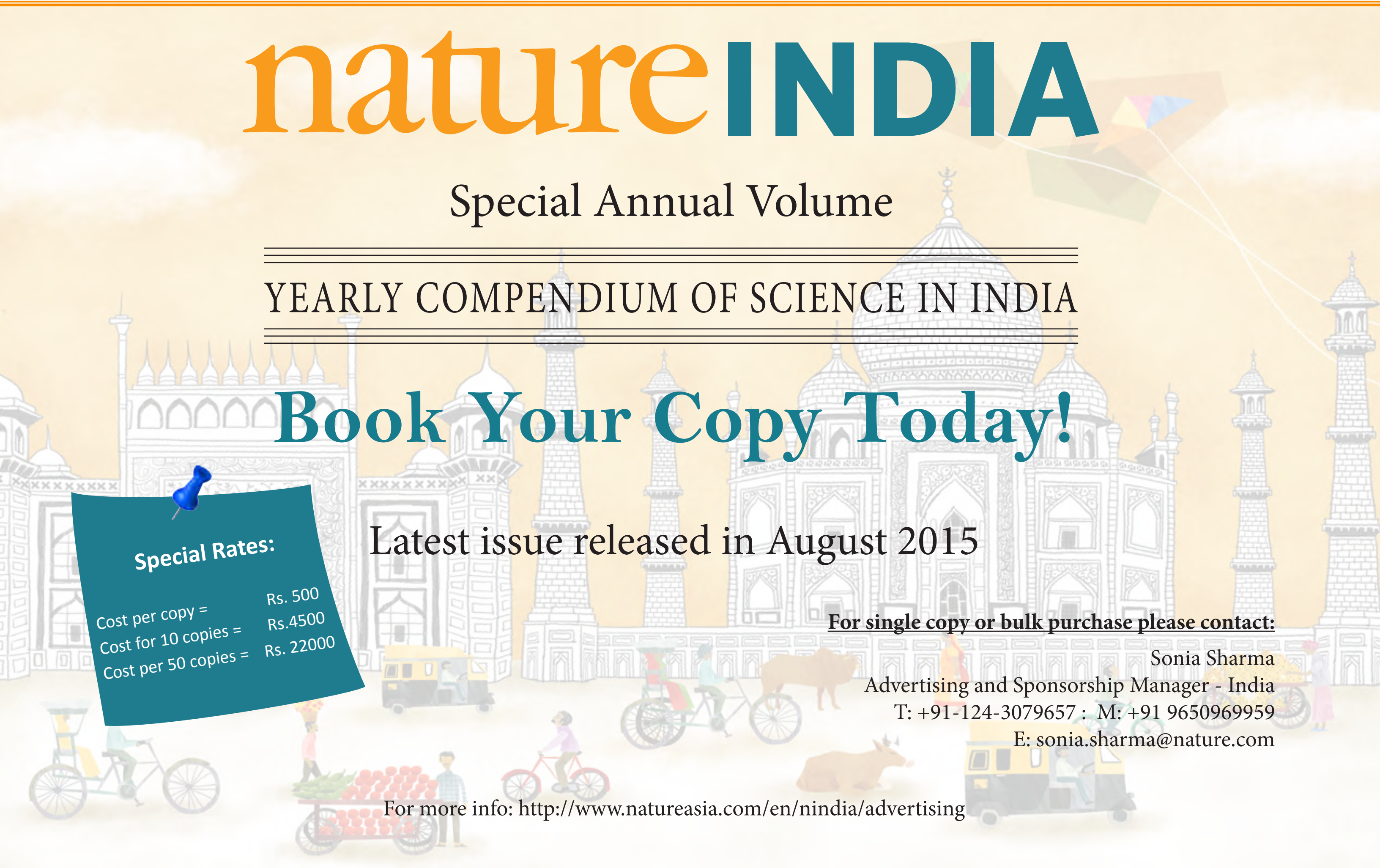
31. Balamurugan, B. *et al.* SSEP-2.0: Secondary structural elements of proteins. *Acta Crystallogr. D Biol. Crystallogr.* **61**, 634-636 (2005).

32. Sharma, O. P. *et al.* Structural Epitope Database (SEDB): A web-based database for the epitope, and its intermolecular interaction along with the tertiary structure information. *J. Proteomics Bioinformatics.* **5**, 084-089 (2012).

33. Jadhav, A. *et al.* Clostridium-DT(DB): a comprehensive database for potential drug targets of *Clostridium difficile*. *Comput. Biol. Med.* **43**, 362-367 (2013).

34. Sharma, O. P. *et al.* VPDB: Viral Protein Structural Database. *Bioinformation.* **6**, 324-326 (2011).

35. Gour, P. *et al.* Manually curated database of rice proteins. *Nucleic Acids Res.* **42**, D1214-1221 (2014).



natureINDIA

Special Annual Volume

YEARLY COMPENDIUM OF SCIENCE IN INDIA

Book Your Copy Today!

Special Rates:

Cost per copy = Rs. 500
Cost for 10 copies = Rs.4500
Cost per 50 copies = Rs. 22000

Latest issue released in August 2015

For single copy or bulk purchase please contact:

Sonia Sharma
Advertising and Sponsorship Manager - India
T: +91-124-3079657 : M: +91 9650969959
E: sonia.sharma@nature.com

For more info: http://www.natureasia.com/en/nindia/advertising