The Andhra Agric. J 62(3):491-498, 2015

Invited Article

Development of Bio-computing Portal, Tools and Algorithms for Biological Data analysis in Agriculture – ICAR-IASRI perspective

Dr. Anil Rai
Head & Principal Scientist,Centre for Agricultural Bioinformatics
ICAR- Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012

1. Preamble

Agriculture in India is an assemblage of diverse physical, chemical and biological components which when synchronized in a patterned array, results in greater productivity of crops to feed the growing population of the country. Present day agriculture is moving very fast from ***green revolution*** to ***ever-green revolution*** which not only benefits the human civilization but also the native natural biotic associates of agriculture, ecology of soil, water and other habitats. Also, there is a need for ***smart farming for small farmers***.

In India, research efforts are being made in the field of agricultural biotechnology, molecular biology, genomics and allied sciences in the last two decades, but, large population of the country is still waiting for significant and desirable output of these efforts. Data-driven multi-omics research is now inevitable for all the biological disciplines related to agricultural sciences. In present days, biological data from the phenotypic and genotypic analysis of crop plants, livestock, microbes, insects, habitats and other interrelated entities are being generated at a very fast pace. In order to extract information and knowledge for understanding the biological process from this massive data generation from the "omics" research (genomics, proteomics, metabolomics etc.), computational biology and bioinformatics have evolved across the globe. This also helps to solve the problems of analysis, prediction, storage, management, pattern recognition, submission, retrieval and storage of the data to find out a meaningful value. Either complete or in-progress whole genome sequencing projects on plants, livestock, insects, fish and microbes have generated huge informatics, resources and forced scientists/researchers across different disciplines to join hands with each other. In order to look into the in-depth of hidden knowledge in the cells to decipher their agriculturally important traits, responses towards biotic and abiotic stresses in the environment and interactions within or between or among the organisms, is essential.This is possible when the traits connect with their genes, alleles, proteins, value-added metabolic products, phenological properties, adaptation behavior, to achieve this evolutionary pattern and possible interactions, scientist from different disciplines of biological sciences along with the scientist from computational sciences need to work with a cohesive team framework. In the multi-omics biological era, every trait and property of living entities is being signified with the "ome", like phenome, genome, epigenome, transcriptome, proteome, metabolome, reactome, localizome, interactome, rhizome, biome, microbiome, trichome, vacuome etc. to represent the functions in their fullness and completeness. The size of data generated on each and every aspect of the organisms are too large to decipher, analyze, storage, manage and retrieve in biologically significant pattern. Therefore, the need of bioinformatics and its integration in the present day of agricultural research and education with high end computer applications, tools and software, database development and management, computational biology, biotechnology and bio-statistics is inevitable. Research in this field is highly interdisciplinary in nature as none of individual discipline can solve the mystery of these biological processes alone. In order to develop effective biotechnological products/commodities for farmers almost all disciplines of agricultural sciences need to be integrated including plant breeding, agronomy, plant protection etc.

The development of related computational tools, has united global efforts and brought revolutionary changes to the research of biology

during the last decade. Today, biologists work globally in association with scientists from a broad spectrum of disciplines to unravel the functions of complex biological systems. Genetic engineering and genomic approaches have opened new vistas for increasing the productivity and quality attributes of biosystems. Genomic databases contain huge amounts of information that are not amenable to traditional analytical approaches. Therefore, bioinformatics and computational biology has emerged as an inter-disciplinary programme which links computational and mathematical sciences with life sciences. Computational biology and agricultural bioinformatics aims to bring the biologists, statisticians and computer scientists together from the point of view of system biology approach and effective problem solving. Huge efforts have been made in agricultural biotechnological research in the country during last two decades, but the real impact of these efforts is still awaited at farm level. This is mainly due to lack of biological computing infrastructure in the country. The computational infrastructures are needed to bridge the gap between genomic information and knowledge, utilizing statistical and computational sciences. These were also required for establishment of large genomic databases, data warehouse, software and tools, algorithms, genome browsers with high-end computational power to extract information and knowledge from cross-species genomic resources. The infrastructures are also needed to open up new vistas for downstream research in bioinformatics ranging from modelling of cellular function, genetic networks, metabolic pathways, validation of drug targets to understand gene function and culminating in the development of improved varieties and breeds for enhancing agricultural productivity to many folds. Further, mitigation of challenges of climate change as well human health through development of varieties/ breeds (i) resistance to biotic (diseases) and abiotic stresses such as salinity, temperature, drought etc. (ii) emit less carbon dioxide (ii) reduction in residue of environmental pollutants such as agricultural chemicals, are also important to be addressed through agricultural biotechnological research in the country. This will not only help in development of globally competitive agricultural products which may increase our agricultural exports but also in protection of Intellectual Property Rights (IPR) and patents. It is expected to lead to next evergreen revolution ensuring our nutritional, food and livelihood security in the country. The genome sequencing of a number of organisms has led to unraveling of many fascination facts. Today, the globe feels the need of this discipline to save resources and time.

Basic task of bioinformatics and computational biology in agricultural and biological science is to answer biological queries that can be utilized for the improvement of agriculture production and productivity and sustainability of the agro-ecosystem.

2. Three basic task components of the field

1 The creation of databases for the storage and management of large biological datasets

2 The development of novel algorithms and statistical programs, mathematical simulation models, machine learning techniques, high-end software, customized computer programs and languages to determine relationships among units of large datasets and,

3 The use of tools to find out data integration, sharing, interpretation and analysis and interconnections among different types of biological data entities.

One way, bioinformatics applies principles of information technologies to make the vast, diverse and complex biological data more readable, understandable and usable while, computational biology uses algorithms, mathematical models and computational approaches to address experimental and theoretical queries. In this way, apart from being distinct in functions and approaches, there is a significant overlap in their activities to bridge the interface of the science of any biology discipline with the informatics.

One of the main aims is to bring the biologists, statisticians and computer scientists together from the point of view of system biology approach and effective problem solving. This approach assist in the development of partnerships at various levels among national and international organizations. The functional linkage among researchers and scientists in the field of agricultural bioinformatics, computational biology and related

fields is to be established. The activities related to bioinformatics initiated at different ICAR institutions at small scale in isolated mode were recently uplifted and encouraged at national level in the field of agriculture. Consolidated efforts are made for collection, compilation, storage and knowledge mining of indigenous agricultural genomic resources. In order to keep pace with the research and developments in agricultural bioinformatics at global level, country needs expertise and exposure in this area of research through ***National Agricultural Bioinformatics Grid*** established by Indian Council of Agricultural Research, New Delhi, with its databases, data warehouse, software and tools, algorithms, genome browsers which are being developed. Also high-end computing facility in the form of first supercomputer for Indian Agriculture (ASHOKA) has been made available to the agricultural researchers through systematic and integrated approach. Various genomic analyses from different ICAR institutions are being carried out at this facility. In long term, information and knowledge generated through inter-disciplinary research from the genomic knowledge base will start flowing downward and experiments in different sectors of agriculture will be able to evolve internationally superior competitive varieties/breeds and commodities in agriculture. Also, the sub-project on ***" National Agricultural Bioinformatics Grid (NABG)"*** under " ***National Agricultural Innovation Project (NAIP)*** of Indian Council of Agricultural Research, New Delhi provided platform for inter-disciplinary research in cross species genomics which led to many collaborative research projects and quality publications across various institutes of ICAR.

3. ASHOKA

The first supercomputing hub for Indian Agriculture ***ASHOKA*** (Advanced Super-computing Hub for OMICS Knowledge in Agriculture) has been established at Centre for Agricultural Bioinformatics (CABin), Indian Agricultural Statistics Research Institute, New Delhi, India. The facility is set up in a state-of-art Data Centre and two super-computers of this hub are listed at rank 11 and 24 in the list of top super-computers of India http://topsupercomputers-india.iisc.ernet.in/jsps/june2013/index.html.

This super-computing hub consists with hybrid architecture with high performance computing having (i) 256 nodes Linux cluster with two masters with 3072 cores and 38 Tera Flops computing, (ii) 16 nodes windows cluster with one master, (ii) 16 nodes GPU cluster with one master with192 CPUs + 8192 GPUs and (iv) SMP based machine with 1.5 TB RAM. Also, this hub has

approximately 1.5 Peta Byte storage divided in to three different types of storage architecture i.e. Network Attached Storage (NAS), Parallel File System (PFS) and Archival. This hub also consists of super-commuting systems (16 node Linux cluster with one master with 40 TB storage) at ICAR-National Bureaux of Plant Genetic Resources (NBPGR) New Delhi, ICAR-National Bureaux of Animal Genetic Resources (NBAGR) Karnal, ICAR-National Bureaux of Fish Genetic Resources (NBFGR) Lucknow, ICAR-National Bureaux of Agriculturally Important Microbes (NBAIM) Mau and ICAR-National Bureaux of Agricultural Insects Resources (NBAIR), Bangalore which forms a National Agricultural Bioinformatics Grid in the country. Number of computational biology and agricultural bioinformatics software/workflow/pipelines along with National Biological Computing Portal have been developed, which provide seamless access to these biological computing resources to the biological researchers across the country.

4. Development of Biological Databases

The development of biological databases relevant to agriculture is one of the important activities in the field of agricultural bioinformatics. Tomato MicroSatellite Database (TomSatDb), http://webapp.cabgrid.res.in/tomsatdb/ , the first whole genome based microsatellite DNA marker database of tomato, houses a total of 146602 STR markers (Iquebal*et al.* 2013). In order to cater the customized needs of wet lab, automated primer designing tool is added. Tom Sat DB is a user-friendly and freely accessible tool which offers chromosome wise as well as location wise search of primers. These markers are expected to pave the way of germplasm management over abiotic and biotic stress as well as improvement through molecular breeding, leading to increased tomato productivity in various parts of the world. Apart from abiotic stress there are more than 200 tomato diseases caused by pathogenic fungi, bacteria, viruses and nematodes affecting tomato productivity as biotic stresses. In order to manage the germplasm in abiotic and biotic stress with desired productivity such closely linked DNA markers are imperatively needed. Further economically and commercially important genes can be used for markers assisted introgression especially in new variety development programme. These findings can be of immense use in tomato genomics research in endeavor of tomato improvement and variety management at global level. Similarly, the micro satellite databases of Pigeonpea (PIPEMicroDB), http://webapp.cabgrid.res.in/pigeonpea/, (Sarika *et al.* 2013)has been developed. These markers will be of immense use in marker assisted selection which would help to overcome approximately 50% loss in pigeonpea productivity due to biotic and abiotic stress in India as well as many parts of the world. Buffalo MicroSatellite Database with 910529 microsatellite markers http://webapp.cabgrid.res.in/buffsatdb/ (Sarika *et al*. 2013) has been made available to international community. This database has been further appended with Primer3 for primer designing of the selected markers enabling researchers to select markers of choice at desired interval over the chromosome. The unique add-on of degenerate bases further helps in resolving presence of degenerate bases in current buffalo assembly. Being first buffalo STR database in the world, this would not only pave the way in resolving current assembly problem but shall be of immense use for global community in QTL/gene mapping which is critically required to increase knowledge in the endeavour to increase buffalo productivity, especially for third world country, where, rural economy is significantly dependent on buffalo productivity. These markers can be used for parentage testing, breed identification, population structuring and admixture analysis. They can also be used for germplasm identification especially in germplasm exchange or issues related to trans-border movement of germplasm.A Goat Microsatellite Database (GoSatDb), http://webapp.cabgrid.res.in/goat/ , has also been developed with 865210 microsatellite markers present in the whole genome sequence of goat.

Halophilicarchaea/bacteria adapt to different salt concentration, namely extreme, moderate and low. This type of adaptations may occur as a result of modification of protein structure and other changes in different cell organelles. Thus, proteins may play an important role in the adaptation of halophilicarchaea/bacteria to saline conditions. The Halophile protein database (HProtDB) is a systematic attempt to document

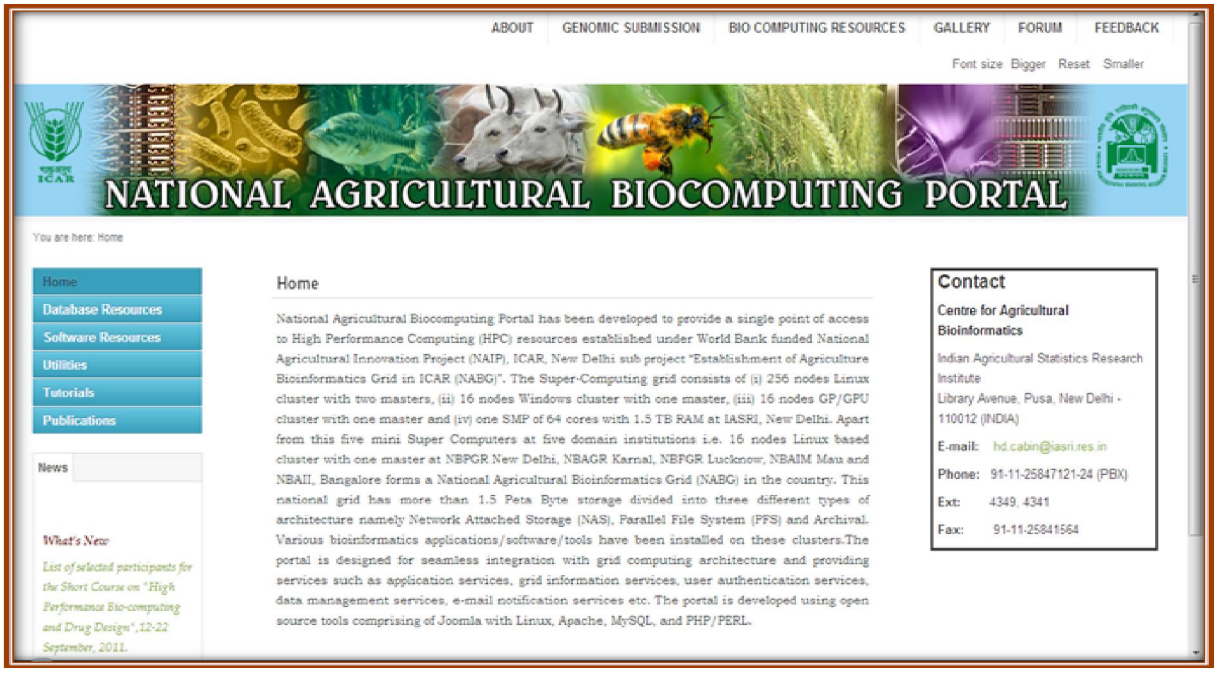
the biochemical and biophysical properties of proteins from halophilicarchaea/bacteria which may be involved in adaptation of these organisms to saline conditions. In this database, various physicochemical properties such as molecular weight, amino acid composition, atomic composition, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (Gravy) have been listed. These physicochemical properties play an important role in identifying the protein structure, bonding pattern and function of the specific proteins. This database is comprehensive, manually curated, non-redundant catalogue of proteins. The database currently contains 59 897 protein properties extracted from 21 different strains of halophilicarchaea/bacteria. The database can be accessed through link. Database URL: http://webapp.cabgrid.res.in/protein/(Naveen *et al.* 2014).

Epigenetics refers to the changes in gene expression that do not involve changes in the DNA sequence. This concept implies that, once established, a new genetic state can be stably propagated through mitosis or meiosis independently of the inducible signal, yet can still revert to its original state. The information related to the epigenetic mechanisms in livestock species is not available at one place. Moreover, analysis of epigenetic information is required for improvement in production traits and controlling diseases in livestock. A web-based "Livestock Epigenetic Information System" has been developed (http://bioinformatics.iasri.res.in/edil/) with MySQL database as bottom layer, PHP as server side application-middle layer and HTML, CSS and JavaScript at top layer.

A web interface has been developed using php scripts and MySQL database to fetch orthologous annotated data of buffalo genes on cattle genome. A browser using light weight genome viewer tool has been developed for mapping buffalo genes on to cattle genome. Annotation files required for construction of genome have been prepared. Also, the information on different functional elements of buffalo genome has been parsed and populated. Mapping of these elements on to buffalo genome has been done. The mapped information has been displayed through light weight genome browser tool. A website has been set up with buffalo genome database and browser tool. Integration and mapping of buffalo gene information with cattle and buffalo genomes has also been done for the users.

5. Development of Bio-Computing Portal and tools:

A National Agricultural Biocomputing portal (http://webapp.cabgrid.res.in/biocomp/) provides a single point of access to High Performance Computing (HPC) resources. The portal provides an environment to carry out the bioinformatics



tasks. The portal facilitates the user in submitting and managing application specific jobs. The jobs submitted through the portal are scheduled and resources are allocated through the Resource Manager. The Resource Manager deals with the access, allocation and management of resources and the execution of tasks. The users can manage the input and output data through the portal interfaces. The users have to login into the portal using the User's login to access the facilities of the portal such as job submission, job status tracking and view output/error data. It provides a Web interface for submitting and monitoring jobs with the specific parameters. The sequential and parallel application jobs can be submitted through portal. The grid provides its users with an environment possessing enormous computing power and large volumes of data comprising of a large number of systems or resources. The resource directories are used to retrieve the information about the state, configuration and status of the grid resources. Round the clock helpdesk support is also made available to address the issues related to operational management of this facility. Different sets of automation tools are configured to manage these computational resource. This will support computational requirements of the biotechnological research in the country. This will also bridge the gap between genomic information and knowledge, utilizing statistical and computational sciences. This will further help in establishing the large genomic databases, data warehouse, software tools, algorithms, genomic browsers with high end computational power to extract information and knowledge from cross-species genomic resources.

5.1 Sequence Submission Portal:

Various studies conducted by agricultural scientists, generate massive data related to biological information of plants, animals, insects, microbes and fisheries. They are dependent on NCBI, EMBL, DDBJ and other portals for their sequence submissions. Due to various limitations imposed on these sites and the poor connectivity problem prevents them to conduct their studies on these open domain databases. A secured sequence submission portal has been developed with a backend database following standard database management concepts http://webapp.cabgrid.res.in/dnadb/ (La*et. al.* 2013). This initiative has been taken to build indigenous genome database and analysis platform in the country. Advanced hardware resources and parallel computing facilities have been installed for high speed information processing and knowledge extraction from this database. The database design has been made generic to integrate numerous genomic databases developed by agricultural scientists working in the area of bioinformatics. This portal is now open for users for submitting their sequences. Auto-curation program is being developed for management of data quality. There is need to encourage our researchers to make use of this portal by building confidence through providing security about their data, sharing these data for extracting knowledge for improving the technology and agricultural productivity.

5.2 Goat Breed Identification server:

A web based server for goat breed identification using microsatellite DNA marker has been developed. This is based on Bayesian Networks classifier with accuracy of 98.7% using 51850 reference allele data generated by 25 microsatellite loci on 22 goat breed population of Indiahttp://webapp.cabgrid.res.in/gomi/(Iquebal *et al.* 2014).Normally, breed descriptor has been developed to identify breed but such descriptors cover only "pure breed" type animals excluding undefined or admixture population. Moreover, in case of semen, ova, embryo and breed product, the breed cannot be identified due to lack of visible descriptors. Therefore, advent of molecular markers like microsatellite and SNP can be easily used for breed identification from even small biological tissue or germplasm. This server is likely to reduce the cost with computational ease. This methodology would become a model for all flora and fauna as a valuable tool for conservation, breed improvement programmes, sovereignty issues in trans-border germplasm movement and management.

5.3 Cattle Breed Identification Server:

In order to overcome challenges for identification of cattle breed, due to lack of phenotypic description especially in ova, semen, embryos and breed products, determination of the degree of admixture and non-descript animals, lack of reference molecular data for identification of

breeds *etc.,* a Web server was developed for maintaining reference data and cattle breed identificationhttp://webapp.cabgrid.res.in/biscattle. The reference data used for developing prediction model were obtained from 8 cattle breeds and 18 microsatellite DNA markers yielding 18000 allele data. Various algorithms were used for reducing number of loci or for identification of important loci. Minimization up to 5 loci was achieved using memory-based learning algorithm without compromising with accuracy of 95%.This model approach and methodology can play immense role in all domestic animal species across globe in breed identification and conservation programme.

5.4 Workflow pipelines and tools for biological data analysis:

Antimicrobial peptides (AMPs) are the defence molecules and are natural alternative to chemical antibiotics. Machine learning techniques is quite useful for understanding hidden patterns in large biological data. It was found that performance of SVM based models for *in-silico* prediction/ identification of AMPs of cattle is superior than ANN. A total of 99 AMPs related to cattle are collected from various databases and published literature. N-terminus residues, C-terminus residues and full sequences were used for SVM model development and identification/prediction (Sarika *et al.* 2015). These SVM models were implemented on web server and made available to users at http:/ /cabin.iasri.res.in/ amp/ for classification/prediction of novel AMPs of cattle.

A parallel framework for workflow and pipeline gene prediction, phylogenetic analysis and SSR-primer has been developed through integration of various tools available in public domains. Web based software for codon usage analysis for gene expression identification has been implemented. Prediction of trait (e.g. abiotic or biotic stress) associated genes is very useful for the researchers of biological sciences. The gene expression data can be helpful in this but it requires specialized analytical and computational support. Trait Associated Genes Prediction Tool (TAGPT) which is a user friendly web based analytical solutionare developedby the scientists. TAGPT implements the proposed algorithm based on sound statistical principles and requires gene expressions data as input. Presently, scientists are in the process of development of a Web based tool for modeling gene regulatory network (GRN). This online tool will facilitate pre-processing Next Generation Sequencing (NGS)/Microarray data, constructing GRN via different modeling formalisms and visualization of network. The program computes gene expression from microarray and NGS data, which can be used further for reconstruction of regulatory networks and thereby subjected for visualization.Protein Structure Comparison (PSC) is an important task for understanding the evolutionary relationships among proteins, predicting structure and function of proteins. Structures are compared to find homologous proteins, for their functional classification, and for the discovery of structural motifs. Various methods have been proposed for comparing protein structures, each method optimizing its own scoring scheme. A Web based tool for comparison of protein structure based on efficient algorithm has been developed for the users.The quantitative comparison of protein 3D structures is an important and fundamental task in structural biology to study evolutionary and structural related issues with other proteins. This helps biologists to understand various aspects of function, evolution from structures and identify its structural neighbors. Now, database of three-dimensional protein structures are becoming large, hence fast and precise search tools and comparison methods are essentially required. Structure comparison may play a key role in understanding the diversity of structure by analyzing and searching structures to derive interesting scientific insights. Based on literature survey, graph theoretic approach can be used for quantifying 3D protein structure and pair wise comparison. Scientists are in process of development of efficient algorithm along with a tool for quantifying 3D protein structure and pair wise comparison using graph theoretic approach.

It is expected that soon, sufficient trained manpower for undertaking research in the field of bioinformatics will be available to support research and development in agricultural biotechnology. A Central Genomic Data Warehousing (CGDW) and data mining facilities with high end computational capability to provide knowledge extraction from these genomic data and the data repository will be

developed in the country. We are in the process of integration of other institutions of NARES to this national grid in phased manner depending on priority of agricultural research. This will help us to eradicate hunger not only from the country but also from globe.

LITERATURE CITED

Iquebal MA, Sarika, Arora Vasu, Verma Nidhi, Rai Anil and Kumar Dinesh 2013 First whole genome based microsatellite DNA marker database of tomato for mapping and variety identification. *BMC Plant Biology* 2013, 13:197.doi:10.1186/1471-2229-13-197.

Iquebal M A, Ansari M S, Sarika Dixit S P, Verma N K, Aggarwal R A K, Jayakumar S, Rai A and Kumar D 2014 Locus minimization in breed prediction using artificial neural network approach. *Animal Genetics,* 45(6), 898–902. NAAS Score: 8.21

Lal S B, Pandey P K, Rai P K, Rai A, Sharma A, Chaturvedi K K 2013 Design and Development of Portal for Biological Database in Agriculture. *Bioinformation,* 9(11): 588-598.

Naveen Sharma, Mohammad Samir Farooqi, Krishna Kumar Chaturvedi, Shashi Bhushan Lal, Monendra Grover, Anil Rai, and Pankaj Pandeyv 2014 The Halophile Protein Database. Database (Oxford): *The Journal of Biological Databases and Curation,* doi: 10.1093/database/bau114.

Sarika Arora, Vasu, Iquebal M A, Rai Anil and Kumar Dinesh 2013 In silico mining of putative microsatellite markers from whole genome sequence of water buffalo (Bubalusbubalis) and development of first BuffS at DB.BMC Genomics. 14, 43 doi:10.1186/1471-2164-14-43.

Sarika Arora, Vasu, Iquebal M A, Rai Anil and Kumar Dinesh 2013PIPEMicroDB: Microsatellite database and primer generation tool for pigeonpea genome. Database: The Journal of Biological Databases and Curation. Vol. 2013, Article ID bas054, doi:10.1093/database/bas054.

Sarika Iquebal MA, Arora Vasu, Rai Anil and Kumar Dinesh 2015 Species specific approach for development of web-based antimicrobial peptides prediction tool. *Computer and Electronics in Agriculture,* 111, 55-61. NAAS Score 7.49.

(Received on 20.06.2015 and revised on 19.08.2015)