Journal of
Information Technology & Software Engineering

Chakraborty et al., J Inform Tech Softw Eng 2017, 7:4
DOI: 10.4172/2165-7866.1000207

Review Article Open Access

Artificial Intelligence in Biological Data

Indrajeet Chakraborty[1], Amarendranath Choudhury[2]*, Tuhin Subhra Banerjee[3]

[1]*Department of Bioinformatics, Karunya University, Coimbatore, Tamil Nadu, India*
[2]*Independent Researcher, Jaipal Homes, Flat No. 101, Kondapur, Hyderabad-500084, India*
[3]*Assistant Teacher of Life Sciences (H/PG), Satpalsa High School, P.O.- Satpalsa, Birbhum, West Bengal, India*

Abstract

Artificial Intelligence (AI) or Machine learning in present era, serves as the primary choice for data mining and big data analysis. With effective learning and adaptation model, it provides solutions to several engineering applications. These include techniques such as Artificial Neural Network modelling, Reasoning based decision algorithms, Simulation models, DNA computing and Quantum computing among several others. With the application of AI in Biomedical research, the fuzziness and randomness in handling such type of data has significantly reduced. Rapid technological advancements have helped AI techniques evolve in manner which promotes handling such fuzzy data effectively and much more conveniently. The review presents a comprehensive view of machine learning and AI computing models, advanced data analytics and optimisation approaches used in Bioengineering such as Drug Designing and Analysis, Medical imaging, biologically inspired learning and adaption for analytics, etc.

Keywords: Artificial intelligence; Machine learning; Bioengineering; Big data

Introduction

With computers becoming the necessity of modern era, it becomes imperative for machine to adapt to the recent trend in the consumer industry. There has been a steady growth in demand for machines that are intelligent and can autonomously react to situations and clearly explain the reasons or the logic behind it. Therefore in layman terms, Artificial Intelligence (AI) can naturally be explained as an action a machine performs which otherwise would have been done by a human using his intelligence [1,2].

With recent advancements in AI technology, many have already grown accustomed to talking and interacting with their gadgets at home. AI technology dominates most of the fictional literature works and cinema around, presenting a popular but scary picture of the coming future [3]. It has already begun changing our lives however most of it is yet to happen. Almost all of the major IT firms are spending millions on developing and implementing AI considering it critical for its future state. Providing personalised relationship with machines is the recent trend in product based industries and is believed to flourish even more lately [3,4].

Artificial Intelligence/Machine Learning

Machine Learning is the form of AI that enables machine to learn without being specifically programmed for each instance [5,6]. The fundamental aim in this context is to make decisions. At the root level, more than one neuron (the fundamental unit of a learning system) group together to form a network also called as a neural network is responsible for the Learning process [7-9]. The algorithm provides the guidelines for rules that are to be followed in the learning process. The target here is to look for a solution for network parameters that yield optimised cost function [10,11]. Training is performed by providing the algorithm with complete set of training examples presented and processed once [12-14]. Thereafter, the neural network is presented a complex relationship with the ability to classify input data. The complexity depends of the number of operation simultaneously being carried out [9]. These learning methods are usually classified into three sub-categories namely.

Supervised learning

Supervised Learning is a closed loop feedback system wherein network parameters are adjusted by comparing the actual and desired output of the system. Labelled set of training data is mapped onto its output using a general rule or mapping pattern which acts as an input function [15,16]. Differences between these values are considered as the error measure and are used to control the learning process. Learning process is repeated until the error measure becomes sufficiently small or the process meets a failure [17,18]. Gradient descent algorithm is used to minimise the error measure [19-21].

Unsupervised learning

Unsupervised Learning is implemented without any defined output for a given input set. The pattern or rule used for classification is learned by the algorithm itself while training. The task here is to generate a hypothesis for the input data and then obtain the output as per the postulates. In this process, all possible hypotheses are evaluated however the output is obtained using the optimal one of them all. Final hypothesis determination governs sub-classification criteria from unsupervised learning techniques [12].

Reinforcement learning

A form of reinforcement based learning technique that identifies general patterns or classification rule in a training dataset and then apply the experience and learning upon another dataset. The classification rule is thus based entire upon the training provided. The method uses two different datasets that are passed on to the learning algorithm [18,22,23]. One of the datasets is analysed and all possible hypothesis are tested on it. The optimal of which is then applied onto the other dataset [1,24].

Big Data

Big data is what transforms case-based studies to large-scale, data-driven research works. The characteristics of big data are defined by three major features namely Volume, Variety, and Velocity [25,26].

***Corresponding author:** Amarendranath Choudhury, Independent Researcher, Jaipal Homes, Flat No. 101, Kondapur, Hyderabad-500084, India, tel: +91 7003017920; Email: anc.au@hotmail.com

Received August 03, 2017; **Accepted** August 29, 2017; **Published** September 06, 2017

Citation: Chakraborty I, Choudhury A, Banerjee TS (2017) Artificial Intelligence in Biological Data. J Inform Tech Softw Eng 7: 207. doi: 10.4172/2165-7866.1000207

Over the years volume of biological data has grown exponentially. This is evident from the fact that ProteomicsDB covers 92% (18,097 of 19,629) of known human genes that are annotated in the Swiss-Prot database. Millions of patient's data have already been collected and stored electronically in databases worldwide. Analysis of these accumulated data would not only enhance health-care services but also bring about major breakthroughs in research [27-29]. Medical imaging also produces vast amounts of data with even more complex features and broader dimensions. The Visible Human Project has archived 39 GB of female datasets [30,31].

Next in line is the variety among data types and its structure. Biological data includes several levels of data sources thus providing a rich array of data for researchers. From genomics, proteomics, metabolomics to protein interactions all of these are unstructured challenges for novel investigations [30-33].

Lastly, velocity refers to producing and processing data. The next generation of sequencing technologies (NGS) [34,35] enables reading billions of DNA sequence data each day at relatively lower costs. Faster speeds are needed for gene sequencing along with faster technologies to process them. In medicine, big data technology is providing faster tools for discovering new patterns among large datasets.

Biological Big Data

With the advent of enhanced computing and storage capabilities, the level of analysis for biological data has shifted from sequence based to a molecular level. This switch has been on account of massive rise in demand for personalised medicine. We are now witnessing an ever increasing demand for producing, storing and analysing huge datasets within a given time frame. Next Generation Sequencing (NGS) has brought about a tide of genomic data and the challenge presently is to store, compute and efficiently manipulate it. Hadoop and MapReduce are the most extensively used currently [25,27,29,36-38].

It is due to lack of universal definition of AI has helped rapid developments in AI. Basically AI can be regarded as an event or activity that provides machine the ability to make decisions and function appropriately and as per their external surroundings. AI focuses not on expanding the scale or the speed rather the emphasis is on making machines autonomous and comprehensive. Till date there has been no match to human intelligence in terms of reasoning ability, perform set tasks, perceive language, handling sensory signals, response to stimuli, artistic and literary works and even gaming. In recent news is AlphaGo - a Narrow AI which defeated its rival the 18-time world champion Lee Sedol, this being the first time in history when a human brain got defeated by an algorithm. This marks the beginning of a new age in computing, where we can expect building imminent computers [39,40]. Computers that are tailor made for neural network based calculations provide better coordination between hardware and software for advanced capabilities and performance [26,41].

AI applications

The genesis of Deep learning approaches dates back to nineteenth century. Having flourished all over the years, showing steady growth however wide stream applications began not before 2012. With the industry investing in the technology along with the advent of high performance computing capabilities, enhanced storage and parallel computing facilities, its applications in our daily lives has made it increasingly important for us. From business to automobile, art to linguistics nothing remains unaffected by its presence. Medical informatics, the application of Information Technology techniques involves examination of patient records and reports and through analysing of huge amounts of such data; complex interactions and correlation in it is revealed [42-46]. Practical application is observed in areas such as oncology, liver pathology, thyroid disease diagnosis, rheumatology, dermatology, cardiology, neuropsychology, gynaecology and perinatology [47-50]. Medical data is now facing serious setbacks as we still have not been able to come up with statistical methods capable of dealing with noisy and missing data [51-53]. Due to this reason the results drawn out of an AI experiment on medical data still faces uncertainty and errors [45,54-56]. Growing trend in the web world has come up with a trending new system called the 'Internet of things' (IoT) wherein several devices are interconnected and keep sharing useful sensory data and commands among them helping devices understand and respond to the external environment. The technique is now creating new opening in a wide range of sectors such as healthcare, retail, banking, manufacturing, smart homes, and personalised user application are some of them [57-59].

AI in genomics

Over the last ten years, machine learning has made tremendous progress in the world of computer science, and still among the fastest growing areas. By 2014, the scientific community had several published research works where machine learning is applied to interpret genome data. Nevertheless, wide scale practical application is something yet to happen. Understanding genes would help transform medicine across the globe [60-62]. The advantage with computers is that you can provide them lots of training. Teach them what is wrong and what is right and meanwhile it keeps learning from its mistakes and eventually starts recognising patterns in data. With large amounts of memory and processing power, computers can learn effectively and continuously for huge amounts of unstructured data. Growing influence of AI over medicine and its worldwide implementation is helping make decision making in machines accurate, personalised and faster. However the current healthcare system remains not capable enough to implement the rapid advancements being made. The implementation of Electronic Health Records has advanced the clinical setup a bit and is being looked into as the maiden step towards revolutionizing modern day healthcare [1,2,5].

The issue here is hardware limitation which is encountered while handling huge amount of data, especially when the training set is huge. Such computation tasks require huge memory and processing capabilities. To overcome this better GPUs with greater amount of memory are being developed such as the ones being developed by companies like Intel and Nvidia. However this still remains a work in progress and currently leaves parallel computing as the only alternative [63,64].

AI in proteomics

Proteins came into the picture ever since it was possible to obtain them in purified form using Mass Spectroscopy and Blotting approaches. Ever since then, the development of high-throughput methods in protein based studies also called as 'proteomics' has been expanding. With more amounts of data available, machine learning has found increased applications in prediction, feature selection, pattern recognition as well as numerous automation works. The major application is in the form of semi-supervised learning techniques where the algorithm learns from large datasets out of which only few are labelled. The technique finds vast applications as researcher are able to handle big data by labelling limited set of examples nevertheless handling huge sets of unlabelled data. In understanding protein sequences, the first step is to generate a profile for the unknown protein based upon its sequence using homology modelling techniques [28,32,33,37].

Multiple local alignment of the query sequence is made with existing databases containing non-redundant records of protein structure and evolutionary information. This helps building a comprehensive representation for the query protein sequence or sequences. Next step is the analysis of this information using the prediction engine which then classifies these into families, superfamilies, folds, clusters etc. based upon the classifier used [37,65-68].

AI in proteome informatics

Protein structure prediction and dynamic analysis of the predicted structures is one of the very first areas to apply machine learning. Later these came to be broadly known as AI and have now currently moved into an even broader interdisciplinary technology called as Deep Learning. As determining protein structures is essential for the understanding of Biological processes and for understand cell functioning [69-72].

Protein structure and fold prediction has had a profound impact in understanding their function. Many new protein sequences have been stored over the past few years in numerous databases globally. Determining the structure and folding of these proteins experimentally would not only be cumbersome but also it would cost a lot of time and money. As experimental verification needs a lot of time and also includes risks of inherent human errors, it is therefore imperative that we develop computational techniques for structure and fold prediction of protein structures based on available sequence data form various public domain databases. The need of the hour is to look for methods that are fast, accurate and automated tools, which would rapidly analyse these sequences and predict its function [73-75].

Understanding how proteins attain their three-dimensional structure is still among some of the raffling mysteries in Biology. Understanding of this would server highly beneficial to medicine especially pharmaceutical sector.

Summarising, machine learning based AI is successfully being applied in protein fold prediction and structure prediction applications. In coming days, application of it are expected to spread to even more dominantly on areas such as disease based genome modification prediction, protein-protein binding site prediction, protein-protein network prediction as well as other allied research topics within the genre [76-78].

AI in phylogeny

Over the last few years, Computational Biology or Bioinformatics as we know it has grown by leaps and bounds. An event mainly caused by massive explosion in the amount of available data and developments of tools for automated analysis of this data. These methods have now become the "workhorse" of Bioinformatics. With simple techniques such as decision tree builder we are able to select among data and subset using limits. These algorithms are an efficient mean for computational time reduction showing clear advantage over others. However, logic behind a particular outcome may not always be clearly justified. Among the most popular ones are the k-nearest neighbour, Bayes theorem, neural network, decision trees [79-82].

Phylogeny is explained through trees wherein the roots are the origin of evolution, the leaves are the species/an organism or a genomic sequence, the branches are the relationship among the leaves and the branch length represents the evolutionary time taken for a particular evolution. A cladogram however is that form of representation which does not take into account time considerations. Construction of this involves the following steps namely [82-85].

- **Alignment:** Sequence alignment is the first step in evolutionary tree construction. The sequences can be either nucleotide sequences or amino acid sequences that are arranged using multiple alignment algorithms.
- **Alignment check:** Thereafter the alignments of these sequences are checked and are looked for evolutionary similarities.
- **Distance computation:** Once the relationships are established, the next step is to find the evolutionary distance between two nodes.
- **Validation:** Once the tree is built, the final step is to validate our results which are done statistically such as back propagation techniques.

Some popular distance computation methods are Distance Matrix methods which work by constructing a correlation matrix which is done in two similar ways namely Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbour Joining method (NJ). Both work in a similar fashion by pairing two leaves (nodes) based upon shortest distance. Then recalculate the distance matrix using average distance between the newly constructed node and the remaining leaves. The calculation is done iteratively until only two clusters of nodes are left [86-88].

AI in next-generation sequencing

Biological Databases are a huge collection of Biological information collected, curated and stored in a defined schema. These include experimental results, high-throughput experimental results, published literature and computational analysis. These databases include data from a wide variety of field such as proteomics, metabolomics, genomics, microarray data analysis and the latest trending is Next-Generation Sequence (NGS) data. NGS data now accounts for whole genome analysis as well has undoubtedly served many issues of genomics and proteomics. These databases can be broadly classified into structure databases and sequence databases. Nucleic acid and protein sequences are stored in sequence databases also known as primary databases and protein structures are stored in structure databases also known as secondary database. Some popular among these are GenBank [80], SwissProt [89] and PIR [90]. GenBank is a fast growing repository of known genetic sequences. In addition to sequence data, GenBank files contain information like accession numbers and gene names, phylogenetic classification and references. SwissProt is a protein sequence database that provides a high level of integration with other databases and also has a very low level of redundancy. Many software tools have come up over the years to retrieve, analyse and visualise data from such databases. These tools cover a wide range to handy operations such as homology modelling, similarity and functional analysis etc. One such fascination tool presently trending is "MethBank" involving whole genome sequences which provide configurable and interactive data analysis. Working on a Red Hat Enterprise Linux server, Java front-end and MySql based query environment. The interface is web friendly and helps retrieve a variety of diverse information [91-94].

AI in genomic expression profiling

Owing to technological advancements in genomics, we are now able to check the expression analysis of thousands of genes at a time using a microchip. Due to this technology, expression profiles for thousands of genes are now available that has helped greatly in the identification and treatment of a variety of diseases. 'DNA Microarrays' as they are popularly referred to; is a high intensity gene array having thousands of spots that helps examine such huge numbers in one go. Here, health

and control samples can be compared to see the abnormalities during diseased cell state [95-99]. With technology advancing rapidly, more and more researchers are being attracted to work on microarray technology and are an integral part of Molecular Biology and Medicine studies. Gene Expression analysis can easily reveal the finding for a patient by checking the disease-related genes. However the problem arises in classification of the available data. Classification algorithms involve statistical methods such as Support Vector Mechine (SVM), Decision Trees and Bayesian Network have been most popularly used. Nevertheless with the advent of AI techniques, these have been employed extensively for classification [100-106].

Integration of Biological Databases with AI

For decades, we tried building computational models for teaching machines. However, one major setback here is the amount of variation in data collected from difference sources. As we know, we shall be able to achieve optimised results only when we would be able to integrate data from a variety of different sources and then devise an automated learning algorithm to analyse and infer prediction based on previous learning experiences. Machines are thus able to serve request from users as well as from other servers more efficiently utilising minimum computational power [107-114].

Biological databases that once comprised of sequences and structures of compounds have now advanced into storage of more complex and bulk data. Most Microarray Profiling studies are based upon a limited subset of the complete expression dataset. We realise that full potential can only be reached upon integration and unification of all available data. Unification and standardisation of public data provides for deep and more accurate insights in analysis and make it easier as well as accurate for machine to find patterns in data. 'ONCOMINE' is one such tool for rapid interpretation of gene's potential role in a particular disease. Expression sets from multiple sources can be retrieved and analysed along while integrating it with multiple other resources such as gene ontology annotations, target gene data, etc. When searching for a disease of interest list of all differential expression analyses is made available [115-136].

Conclusion

History records, humans have went on adapting to new techniques and developing better technology. As reviewed AI, we could see that it had been developing ever since its introduction in 1943 with McCullouch and Pitts giving the world the concept of artificial neurons. Ever since then, it has been growing rapidly, showing unexpected growth at times with new cutting edge technology coming into play. Compared with traditional methods, machine learning based methods are more accurate, robust and reliable. Conversely, since AI has shortcoming as well, we constantly need to look for improvement in its design and application. In days to come, AI would constantly find extensive applications in lot many unexplored areas. The measure of its success would eventually be measured by the amount of change it causes in people's lives. The ease with which people nowadays are adapting to AI technologies, the future definitely looks good.

References

1. Brunette ES, Flemmer RC, Flemmer CL (2009) A review of artificial intelligence. ICARA 2009 Proceedings of the 4th International Conference on Autonomous Robots and Agents, pp: 385-392.
2. Boden MA (1998) Creativity and artificial intelligence. Artif Intell 103: 347-356.
3. Müller VC, Bostrom N (2014) Future progress in artificial intelligence. AI Matters 1: 9-11.
4. Glover F (1986) Future paths for integer programming and links to artificial intelligence. Comput Oper Res 13: 533-549.
5. Ghahramani Z (2015) Probabilistic machine learning and artificial intelligence. Nature 521: 452-459.
6. Dietterich T (1996) Machine learning. Annual Review of Computer Science.
7. Jaeger H (2016) Artificial intelligence: Deep neural reasoning. Nature 538: 467-468.
8. Wang SC (2003) Artificial neural network. Interdisciplinary Computing in Java Programming 743: 81-100.
9. Jain AK, Mao J, Mohiuddin KM (1996) Artificial neural network: A tutorial. Computer 29: 31-44.
10. Tan CNW (1999) A hybrid financial trading system incorporating chaos theory, statistical and artificial intelligence/soft computing methods. Queensland Finance Conference.
11. Pau LF (1991) Artificial intelligence and financial services. IEEE Trans Knowl Data Eng 3: 137-148.
12. Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. J Glob Optim 39: 459-471.
13. Akay B, Karaboga D (2012) A modified Artificial Bee Colony algorithm for real-parameter optimization. Inf Sci (Ny) 192: 120-142.
14. Karaboga D, Gorkemli B, Ozturk C, Karaboga N (2014) A comprehensive survey: Artificial bee colony (ABC) algorithm and applications. Artif Intell Rev 42: 21-57.
15. Bastanlar Y, Ozuysal M (2013) Introduction to machine learning. Methods in Molecular Biology 1107: 105-128.
16. Jordan MI, Rumelhart DE (1992) Forward models: Supervised learning with a distal teacher. Cogn Sci 16: 307-354.
17. Balkenius C (1994) Biological learning and artificial intelligence. Cogn Sci 30: 1-19.
18. Szepesvári C (2010) Algorithms for reinforcement learning. Synth Lect Artif Intell Mach Learn 4: 1-103.
19. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Networks 5: 157-166.
20. Ahmad F, Isa NAM, Osman MK, Hussain Z (2010) Performance comparison of gradient descent and genetic algorithm based artificial neural networks training. Intelligent Systems Design and Applications (ISDA), 10th International Conference.
21. Motta AD, Barreto S, Anderson CW (2008) Restricted gradient-descent algorithm for value-function approximation in reinforcement learning. Artif Intell 172: 454-482.
22. Sutton RS, Barto AG (2012) Reinforcement learning. Adaptive Computation and Machine Learning 3: 322.
23. Schölkopf B (2015) Artificial intelligence: Learning to see and act. Nature 518: 486-487.
24. Nilsson NJ (1991) Logic and artificial intelligence. Artif Intell 47: 31-56.
25. Wu X, Zhu X, Wu GQ, Ding W (2014) Data mining with big data. Knowl Data Eng IEEE Trans 26: 97-107.
26. Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage 35: 137-144.
27. Li Y, Chen L (2014) Big biological data: Challenges and opportunities expanding volume of the big biological data and its bonanza. Genomics Proteomics Bioinformatics 12: 187-189.
28. No Authors Listed (2011) Big data: The next frontier for innovation, competition, and productivity. McKinsey Glob Inst.
29. Marx V (2013) Biology: The big challenges of big data. Nature 498: 255-260.
30. O'Driscoll A, Daugelaite J, Sleator RD (2013) Big data Hadoop and cloud computing in genomics. J Biomed Inform 46: 774-781.
31. Green ED, Watson JD, Collins FS (2015) Human genome project: Twenty-five years of big biology. Nature 526: 29-31.
32. Ozdemir V, Dove ES, Gursoy UK, Sardas S, Yıldırım A, et al. (2017)

Personalized medicine beyond genomics: alternative futures in big data-proteomics, environtome and the social proteome. J Neural Transm 124: 25-32.

33. Savage N (2015) Proteomics: High-protein research. Nature 527: S6-S7.
34. Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10: 135-151.
35. Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52: 413-435.
36. Demchenko Y, De Laat C, Membrey P (2014) Defining architecture components of the Big Data Ecosystem. 2014 International Conference on Collaboration Technologies and Systems, CTS.
37. Jacobs A (2009) The pathologies of big data. Queue-Data 7: 10.
38. Aronova E, Baker KS, Oreskes N (2010) Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957–Present. Hist Stud Nat Sci 40: 183-224.
39. Silver D, Hassabis D (2016) AlphaGo: Mastering the ancient game of Go with Machine Learning. Google Research Blog.
40. Borowiec S (2016) AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol. The Guardian.
41. Katal A, Wazid M, Goudar RH (2013) Big data: Issues, challenges, tools and Good practices. 2013 6th International Conference on Contemporary Computing, IC3.
42. Costa FF (2014) Big data in biomedicine. Drug Discov Today 19: 433-440.
43. Gligorijevic V, Malod-Dognin N, Przulj N (2016) Integrative methods for analyzing big data in precision medicine. Proteomics 16: 741-758.
44. Dilsizian SE, Siegel EL (2014) Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep 16: 441.
45. Hu H, Wen Y, Chua TS, Li X (2014) Toward scalable systems for big data analytics: A technology tutorial. IEEE Access 2: 652-687.
46. Zhang Y, Guo SL, Han LN, Li TL (2016) Application and exploration of big data mining in clinical medicine. Chin Med J 129: 731-738.
47. Banerjee AK, Arora N, Murty USN (2009) Structural model of the Plasmodium falciparum thioredoxin reductase: a novel target for antimalarial drugs. J Vector Borne Dis 46: 171-183.
48. Murty USN, Banerjee AK, Arora N (2009) Application of Kohonen maps for solving the classification puzzle in AGC kinase protein sequences. Interdiscip Sci 1: 173-178.
49. Banerjee AK, Harikrishna N, Kumar JV, Murty US (2011) Towards classifying organisms based on their protein physicochemical properties using comparative intelligent techniques. Appl Artif Intell 25: 426-439.
50. Murty U, Banerjee A, Arora N (2012) Analyzing a potential drug target N-myristoyltransferase of Plasmodium falciparum through in silico approaches. J Glob Infect Dis 4: 43-54.
51. Banerjee AK, Arora N, Murthy US (2007) Stability of ITS2 secondary structure in anopheles : What lies beneath ? IJIB 1: 232-238.
52. Arora N, Banerjee AK, Murty USN (2010) In silico characterization of Shikimate kinase of Shigella flexneri: A potential drug target. Interdiscip Sci Comput Life Sci 2: 280-290.
53. Banerjee AK, MS, MN, Murty US (2010) Classification and clustering analysis of pyruvate dehydrogenase enzyme based on their physicochemical properties. Bioinformation 4: 456-462.
54. Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, et al. (2013) Some experiences and opportunities for big data in translational research. Genet Med 15: 802-809.
55. Merelli I, Perez-Sanchez H, Gesing S, D'Agostino D (2014) Managing, analysing, and integrating big data in medical bioinformatics: Open problems and future perspectives. Biomed Res Int, pp: 1-13.
56. Frankel F, Reid R (2008) Big data: Distilling meaning from data. Nature 455: 1-30.
57. O'Leary DE (2013) Big data, the internet of things and the internet of signs. Int J Intell Syst Account Financ Manag 20: 53-65.
58. Perera C, Ranjan R, Wang L, Khan SU, Zomaya AY (2015) Big data privacy in the internet of things era. IT Prof 17: 32-39.
59. Ashton K (2009) That 'internet of things' thing. RFiD J.
60. Higuchi N (2013) Three challenges in advanced medicine. Japan Med Assoc J 56: 437-447.
61. World Bank Institute (2007) Building knowledge economies: advanced strategies for development. WBI Development Studies.
62. Karp JM, Langer R (2007) Development and therapeutic applications of advanced biomaterials. Current Opinion in Biotech 18: 454-459.
63. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, et al. (2009) The coming of age of artificial intelligence in medicine. Artif Intell Med 46: 5-17.
64. Poole DL, Mackworth AK (2010) Artificial intelligence - Foundations of computational agents. Artificial Intelligence.
65. Yang S, Njoku M, Mackenzie CF (2014) Big data approaches to trauma outcome prediction and autonomous resuscitation. Br J Hosp Med 75: 637-641.
66. Frey LJ, Lenert L, Lopez-Campos G (2014) EHR big data deep phenotyping. Contribution of the IMIA Genomic Medicine Working Group. Yearb Med Inform 9: 206-211.
67. Krumholz HM (2014) Big data and new knowledge in medicine: The thinking, training and tools needed for a learning health system. Health Aff 33: 1163-1170.
68. Hough LM (1992) The big five personality variables--construct confusion: Description versus prediction. Hum Perform 5: 139-155.
69. Chen X-W, Lin X (2014) Big data deep learning: Challenges and perspectives. IEEE Access 2: 514-525.
70. Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Networks 61: 85-117.
71. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521: 436-444.
72. Cox DR (2015) Big data and precision. Biometrika 102: 712-716.
73. Leff DR, Yang GZ (2015) Big data for precision medicine. Engineering, pp: 277-279.
74. Schmitt R, Dietrich F, Droder K (2016) Big data methods for precision assembly. Procedia CIRP, pp: 91-96.
75. Lv Y, Duan Y, Kang W, Li Z, Wang FY (2014) Traffic flow prediction with big data: A deep learning approach. IEEE Transactions on Intell Trans Sys 16: 865-873.
76. Soding J (2017) Big-data approaches to protein structure prediction. Science 355: 248-249.
77. Suthaharan S (2014) Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Perform Eval Rev 41: 70-73.
78. Higgs PG, Attwood TK (2013) Bioinformatics and molecular evolution.
79. Bergeron B (2003) Bioinformatics computing. Prentice Hall, USA.
80. Pevsner J (2009) Access to sequence data and literature information. Bioinformatics and Functional Genomics (2nd edn.). John Wiley & Sons, USA.
81. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333-340.
82. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10: 249.
83. Podell S, Gaasterland T (2007) DarkHorse: A method for genome-wide prediction of horizontal gene transfer. Genome Biol 8: R16.
84. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462: 1056-1060.
85. Li Y, Xu L (2010) Unweighted multiple group method with arithmetic mean. Proceedings 2010 IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications.
86. Tajima F (1990) A simply graphic method for reconstructing phylogenetic trees

from molecular data. Mol Biol Evol 7: 578-588.

87. McGarvey PB, Huang H, Barker WC, Orcutt BC, Garavelli JS, et al. (2000) PIR: A new resource for bioinformatics. Bioinformatics 16: 290-291.

88. Wu CH, Zhao S, Chen HL (1996) A protein class database organized with ProSite protein groups and PIR superfamilies. J Comput Biol 3: 547-561.

89. Zou D, Sun S, Li R, Liu J, Zhang J, et al. (2015) MethBank: A database integrating next-generation sequencing single-base-resolution DNA methylation programming data. Nucleic Acids Res 43: D54-D58.

90. Lesk AM (2002) Introduction to bioinformatics.

91. Risler JL (2001) Developing bioinformatics computer skills. Computers and Chemistry 5: 553-555.

92. Altman RB, Dugan JM (2005) Defining bioinformatics and structural bioinformatics. Struct Bioinforma. 44: 3-14.

93. Kohane IS, Kho AT, Butte AJ (2003) Microarrays for an integrative genomics. Comput Math with Appl 46: 505-506.

94. Spang R (2004) Diagnostic signatures from microarrays: A bioinformatics concept for personalized medicine. Drug Discov Today 9: S32-36.

95. Malone JH, Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 9: 34.

96. Smyth GK, Ritchie M, Thorne N, Wettenhall J (2005) Limma: Linear models for microarray data BT - bioinformatics and computational biology solutions using R and bioconductor.

97. Kumar C, Mann M (2009) Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Letters. 583: 1703-1712.

98. Byvatov E, Schneider G (2002) Support vector machine applications in bioinformatics. Appl Bioinformatics 2: 67-77.

99. Chang C, Lin C (2013) LIBSVM : A library for support vector machines. ACM Trans Intell Syst Technol 2: 1-39.

100. Cortes C, Vapnik V (1995) Support vector machine. Mach Learn, pp: 1303-1308.

101. Darwiche A (2010) Bayesian networks. Commun ACM 53: 80-90.

102. Cooper GF, Herskovits E (1992) A bayesian method for the induction of probabilistic networks from data. Mach Learn 9: 309-347.

103. Peer D (2005) Bayesian network analysis of signaling networks: A primer. Sci STKE 281: pl4.

104. Wilkinson DJ (2007) Bayesian methods in bioinformatics and computational systems biology. Briefings in Bioinformatics 8: 109-116.

105. Lacroix Z, Ludascher B, Stevens R (2008) Integrating biological databases. Bioinformatics-From Genomes to Therapies, pp: 1525-1571.

106. Stein L (2013) Creating databases for biological information: An introduction. Curr Protoc Bioinforma 9: 1.

107. Zou D, Ma L, Yu J, Zhang Z (2015) Biological databases for human research. Genomics, Proteomics and Bioinformatics, pp: 55-63.

108. Reddy MP, Prasad BE, Reddy PG, Gupta A (1994) A methodology for integration of heterogeneous databases. IEEE Trans Knowl Data Eng 6: 920-933.

109. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al. (2004) ONCOMINE: A cancer microarray database and integrated data-mining platform. Neoplasia 6: 1-6.

110. Domshlak C, Hullermeier E, Kaci S, Prade H (2011) Preferences in AI: An overview. Artificial Intelligence, pp: 1037-1052.

111. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, et al. (2007) Oncomine 3.0: Genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 9: 166-180.

112. Yannakakis GN (2012) Game AI revisited. Proceedings of the 9th Conference on Computing Frontiers.

113. Banerjee AK, Arora N, Pranitha V, Murty USN (2008) Exploring the interplay of sequence and structural features in determining the flexibility of AGC kinase protein family: A bioinformatics approach. J Proteomics Bioinform 1: 77-89.

114. Banerjee AK, Arora N, Murty US (2007) How far is ITS2 reliable as a phylogenetic marker for the mosquito genera? Electro J Biol 3: 61-68.

115. Banerjee AK, Arora N, Murty US (2008) Classification and regression tree (CART) analysis for deriving variable importance of parameters influencing average flexibility of CaMK kinase family. Electro J Biol 4: 27-33.

116. Suryanarayana Murty U, Banerjee AK (2012) Seaweeds: The wealth of oceans. Handbook of Marine Macroalgae: Biotechnol Appl Psycho, pp: 36-44.

117. Banerjee AK, Ravi V, Murty US, Sengupta N, Karuna B (2013) Application of intelligent techniques for classification of bacteria using protein sequence-derived features. Appl Biochem Biotechnol 170: 1263-1281.

118. Banerjee AK, Kiran K, Murty US, Venkateswarlu C (2008) Classification and identification of mosquito species using artificial neural networks. Comput Biol Chem 32: 442-447.

119. Duddela S, Sekhar PN, Padmavati GV, Banerjee AK, Murty US (2010) Probing the structure of human glucose transporter 2 and analysis of protein ligand interactions. Med Chem Res 19: 836-853.

120. Banerjee AK, Manasa BP, Murty US (2010) Assessing the relationship among physicochemical properties of proteins with respect to hydrophobicity: A case study on AGC kinase superfamily. Indian J Biochem Biophys 47: 370-377.

121. Banerjee AK, Ravi V, Murty US, Shanbhag AP, Prasanna VL (2013) Keratin protein property based classification of mammals and non-mammals using machine learning techniques. Comput Biol Med 43: 889-899.

122. Banerjee AK, Arora N, Murty US (2012) Aspartate carbamoyltransferase of Plasmodium falciparum as a potential drug target for designing anti-malarial chemotherapeutic agents. Med Chem Res 21: 2480-2493.

123. Banerjee AK, Arora N, Murty US (2009) Clustering and classification of anopheline spacer sequences using self-organizing maps. Internet J Genomics Proteomics 4: 2.

124. Chen D, Moulin B, Wu J (2014) Analyzing and modeling spatial and temporal dynamics of infectious diseases. John Wiley & Sons, USA.

125. Murty US, Banerjee AK (2011) Bioinformatics with Solutions in Pest Management Science: An Insight into the Evolving Technologies. Pest Pathogens Manag Str, p. 521.

126. Banerjee AK (2015) Computation in analyzing inflammation: A general perspective. Interdiscip J Microinflammation 2: 2.

127. Banerjee AK, Wu J (2014) West Nile Virus: A narrative from bioinformatics and mathematical modeling studies. Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases. John Wiley & Sons, US.

128. Arora N, Narasu ML, Banerjee AK (2016) Shikimate kinase of Yersinia pestis: A sequence, structural and functional analysis. Int J Biomed Data Min 5: 2-11.

129. Shinde SP, Banerjee AK, Arora N, Murty US, Sripathi VR, et al. (2015) Computational approach for elucidating interactions of cross-species miRNAs and their targets in Flaviviruses. J Vector Borne Dis 52: 11.

130. Pal-Bhadra VR, Arora N, Shinde Santosh P, Ray P, Banerjee AK, et al. (2010) Target sites for microRNA expressed in pancreatic islets in Type 2 diabetes mellitus associated genes. Online J Bioinform 11: 224-243.

131. Arora N, Kumar Banerjee A (2012) Looking beyond the obvious: search for novel targets and drugs for reducing the burden of infectious diseases. Mini Rev Med Chem 12: 185-186.

132. Arora N, Banerjee AK (2010) Emerging trends, challenges and prospects in healthcare in India. Electron J Biol 6: 24-25.

133. Arora N, Banerjee AK (2012) Targeting tuberculosis: a glimpse of promising drug targets. Mini reviews in medicinal chemistry 12: 187-201.

134. Arora N, Banerjee AK (2012) New targets, new hope: novel drug targets for curbing malaria. Mini Rev Med Chem 12: 210-226.

135. K Saxena S, Gupta A, Bhagyashree K, Saxena R, Arora N, et al. (2012) Targeting strategies for human immunodeficiency virus: a combinatorial approach. Mini Rev Med Chem 12: 236-254.

136. Arora N, Banerjee AK, Narasu ML (2016) Zika virus: An emerging arboviral disease. Future Virol 26: 395-399.