International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

AI in Bioinformatics

Muhammad Noman Akhtar[a*], Gohar Abbas[b], Mohsin Ali Khan[c]

[a]*Bahaudin Zakeria University Multan*

[b]*Institute of Southern Punjab*

[c]*University of Engineering and Technology Taxila*

[a]*Email: noman.akhter86@gmail.com,* [b]*Email: goharabbas321@gmail.com,* [c]*Email: 17-ee-85@students.uettaxila.edu.pk*

Abstract

In bioinformatics science and computational molecular biology, artificial intelligence (AI) has rapidly gained interest. With the availability of numerous types of AI algorithms, it has become prevalent for researchers to use off-shelf programmes to identify their datasets and mine them. At present, researchers are facing difficulties in selecting the right approach that could be extended to a given data collection, with numerous intelligent approaches available in the literature. Researchers need instruments that present the data in an intuitive manner, annotated with meaning, precision estimates, and description. In the fields of bioinformatics and computational molecular biology (DNA sequencing), this article seeks to review the use of AI. These fields have evolved from the needs of biologists to use the large volumes of data continuously obtained in genomic science and to better understand them. For several approaches to bioinformatics and DNA sequencing, the fundamental impetus is the evolution of species and the difficulty of dealing with incorrect results. The type of software programmes developed by the scientific community to search, identify and mine numerous usable biological databases are also mentioned in this article, simulating biological experiments with and without mistakes. The review of antibody-antigen interactions and their diversity, and the study of epidemiological evidence that can help forecast antibody-antigen interactions and the induction of broadly neutralising antibodies are important questions to be answered in the field of vaccinology.

Keywords: intelligent bioinformatics system; AI tools in bioinformatics.

---

* Corresponding author.

1. Introduction

Bioinformatics is a subdiscipline of science and software engineering in which organic information, most usually DNA and amino corrosive groupings, are obtained, prepared, broke down, and spread. For various applications, bioinformatics utilizes PC programs, including assessing quality and protein functions, framing developmental affiliations, and foreseeing the three-dimensional types of proteins. There has been a lot of development in bioinformatics and artificial intelligence (AI) over the past decade, and there are more ways to consider biological data and challenges. Bioinformatics is an interdisciplinary area of research that uses mathematical equations, statistical methods, and algorithms along with computational powers to solve various biological problems [1]. AI is a computer system's ability to perform various tasks associated with intelligent beings and to simulate human intelligence processes by computer systems [2]. AI is a machine's capacity to abstract, be imaginative and carry out activities dependent on its training [3]. To address various biological questions, bioinformatics methods have been used. The structural bioinformatics tools have AI applications and have powerful methodologies to use the tools that have AI applications to design active novel compounds against neurological disorders and cancer by in silico means. In order to interpret biological data for rational conclusions, bioinformatics methods are used. The whole genome sequencing programmes generate a large volume of biological data and bioinformatics methods help to solve and annotate the data in practical ways [4]. The use of AI approaches along with bioinformatics approaches has solved various biological problems and helps to design efficient algorithms for gene prediction, computational drug design, protein-protein interaction studies, genome-wide association studies, next generation sequences and software creation [5]. With the assistance of biological sequence matching, protein-protein interaction and function-structure analysis, AI in bioinformatics covers both fundamental as well as clinical science. In the creation and discovery of medications as well as complex structures, this study supports [6].

A. Application of AI in Bioinformatics

Applications in bioinformatics are mainly software instruments that help produce and store valuable biological information. But, integrated with AI technologies, bioinformatics software can help produce the knowledge even faster and also help create predictions. AI bioinformatics tools allow advanced biological sequence comparison, information management, and protein-protein interaction methods to be developed. For several bioinformatics applications, machine learning, a subfield of AI, has become an important method. For prediction and pattern recognition based on massive datasets, machine learning algorithms are particularly useful. Within the bioinformatics framework, there are a range of emerging machine learning applications.

B. DNA Sequencing

DNA is a double stranded helical twisted structure consisting of four nitrogenous bases i.e. adenine, thymine, cytosine and guanine. These bases are covalently bonded to each other with double and triple bonds. Specific arrangement of nucleotides on DNA fragment that code a specific protein is called gene. The molecular foundation of different genetic components expressed in our genetics, such as hair colour, is made up of functional molecules encoded by genes. DNA regions encoding genes in bioinformatics are also known as

prediction of the genes, or gene finding.

A mixture of extrinsic and intrinsic searches consists of the method of gene finding. In the extrinsic scan, the target genome is screened for sequences similar to extrinsic data in the shape of previously defined and identifying gene encoding sequences. Anyway a characteristic inquiry is additionally completed where gene prediction calculations plan to arrange pieces of the DNA that may hypothetically have gene encoding successions, thinking about the innate expense and intricacy in gathering outward confirmation for specific genes. The genomic DNA is routinely scanned for protein-coding genes through these predictive models. These algorithms use a mixture of indications, complex sequences, material and statistical characteristics to produce predictions. Different models for machine learning and deep learning, some of which are Random Forest, K-Nearest Neighbor and Perceptron Multilayer, are currently being implemented in this area. An inquiry into classification strategies in metagenomic gene prediction and a gene prediction with deep learning in the two research studies follows, the use and results obtained using the two approaches for the prediction of alien genes were closely analysed.

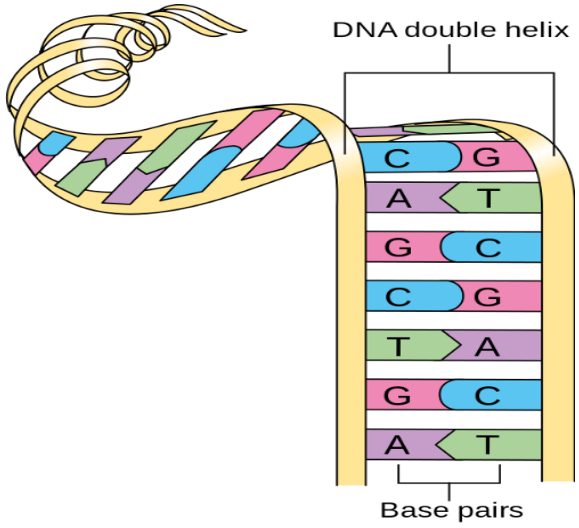

Figure 1: (DNA sequencing)

C. Protein Classification

The "doers" of our cells are proteins, performing several functions that eventually make life. They are answerable for an expansive assortment of capacities within living things, for example, metabolic capacities, boost reaction, cell organizing, sub-atomic vehicle and some more. Our chromosomes codify proteins which form the basis of living tissues. An important step in truly explaining the complexity of the human body is the classification of protein patterns through human cells. With the recent development in high-performance microscopy, cellular images are being produced more rapidly than can be processed and identified by any human team. Thus, AI models, through various AI approaches and PC vision methods, for example, Deep Convolutionary Neural Fields (DCNF), are kicking things off by ordering protein pattern in human cells. We left on an excursion at Optima AI to make a model of protein pattern classifier that would extricate multiclass arrangements dependent on cell pictures created by microscopy of high throughput.

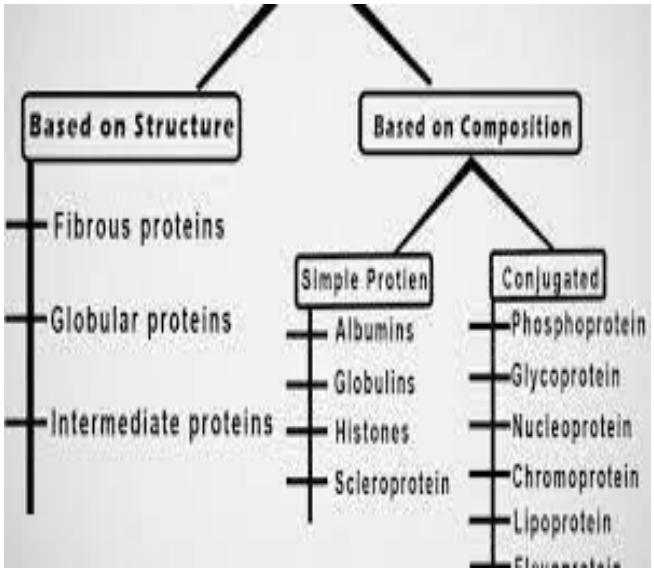

Figure 2: (classification of protein)

D. Analysis of Gene Expressions

DNA Microarray, is a series of DNA spots attached to a solid surface, a form of lab-on-a-chip. Microarrays are utilized to assemble and compute gene articulation levels inside species consequently. A gene articulation is the process by which the preparing of a useful gene product utilizes contribution from a gene. A functional gene product is now and again a protein, or it is a functional RNA for non-protein encodings. In the examination, pattern recognition and characterization of gene articulations, AI is often utilized. Including cutting-edge AI techniques in microarrays and RNA sequencing, the limitations in tumour identifiable proofs and molecular level in the area of disease research have been demonstrated. Therefore it helps doctors to give cancer patients customised care depending on the genetic build-up of a single tumour. Microarray Data will improve cancer detection in AI Approaches applied to DNA. This exploration delineates the perceptions and AI methods used for these problems in classification.. In genetics, cellular biology, cancer treatment and precision medicine, recent developments in the fields of bioinformatics and AI offer various potential outcomes. We are anxious to be at the cutting edge of this innovation at Optima AI and we are intending to work with organizations hoping to utilize AI to make the following tremendous forward leaps in clinical examination and medication [7].

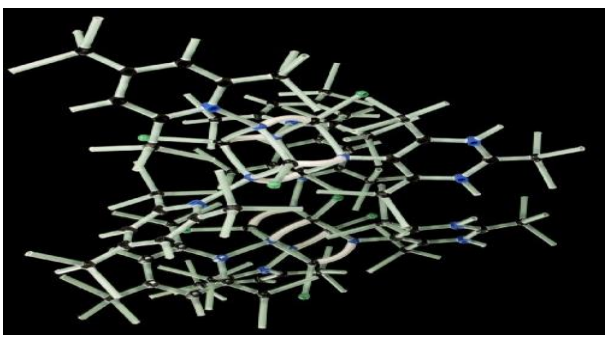

Figure 3: Analysis of gene expression

E. Protein Structure Generative Modelling

New data examples can be provided by generative models that can be used to train AI algorithms for different bioinformatics fields. This models strive to learn the representation of data and generate new instances of data that look close to the original. Generative Adversarial Networks (GANs), which are variants of neural networks that use a generator and discriminator network to produce protein structures, are one of the powerful approaches to generating models. The generator network attempts to produce a natural image, while the discriminator tries to decide if the former network's generated image is false or true. And then to train AI algorithms, the images that are determined as valid can be used. Although the effect of AI on bioinformatics is intriguing, more data is required to strengthen it. And with the sequencing of more patients, it would be feasible. Yeah you got it correctly, since bioinformatics gathers data, every patient is a source of knowledge on their own. And once AI systems have data, there are different ways of converting this knowledge into observations and using them for different advantages. In streamlining diverse analytical workflows into a single research process, the integration of bioinformatics and AI technologies can play an important role. And such structures would allow biological data to be processed and analysed at unparalleled rates. It can definitely be said that the future of bioinformatics lies in the study with AI algorithms of a vast volume of data to provide significant savings in time and resources, which will drive biological science further.

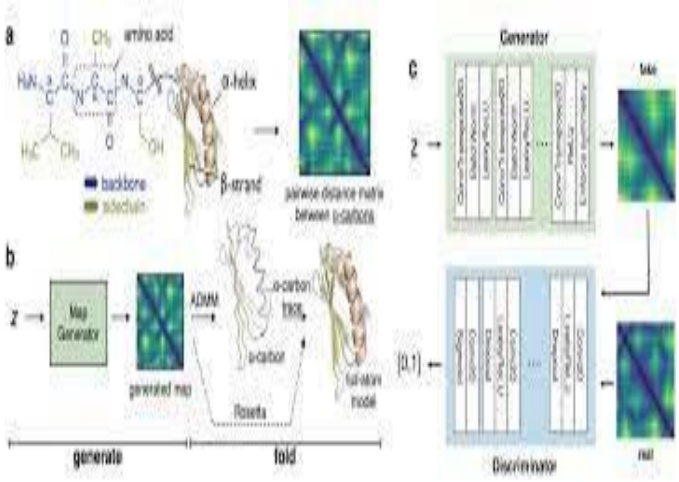

Figure 4: (generative model for protein structure)

F. Genetic Algorithm

A genetic algorithm is a search algorithm focused on heuristics inspired by Charles Darwin's theory of natural evolution. In order to choose the most suitable survival and meaningful output, the genetic algorithm follows the principle of natural selection [8]. The genetic algorithm is intended to execute the tasks in five key phases and to choose the generation's most suitable iteration [9]. Initially, the first population (defined by a set of parameters) contains a set of individuals known as population and each individual has genes. A chromosome forms a combination of a given group of genes. The fitness function determines an individual's efficiency and calculates a score for each person [10]. The determined score determines that an individual's selection corresponds to the

selection of the fittest person based on the progeny's fitness score. Via this process, the parents are chosen to replicate. For each pair of parents, the crossover is known to be the essential step for the random sorting of the genes resulting in the generation of offspring. For the development of new offspring at the mutation stage, there is a chance that the strings of the genes will flip. The algorithm finishes with the repetitive creation of the same generations, which are called the final result [11].

G. The flowchart of the Genetic Algorithm

The Genetic Algorithm is used to improve the arrangement of different sequences efficiently [12]. In order to produce a fitness score based on the similarity and mismatching of the columns, this method uses a population of alignments [13]. Half of the required alignments are copied to the next generation, resulting in the crossover points selecting a cut in the first alignment sequence for the random selection stage, and another cut is made to change the first sequence for the second alignment [14]. The gaps leading to the splicing of another parent are spliced by one parent to add the gaps to ensure continuity of alignment [15]. Compared to other alignment algorithms, the genetic algorithm estimates successful alignment outcomes. AI technologies of bioinformatics play an important part in the statistical resolution of biological problems [10].

H. Intelligent knowledge discovery in Bioinformatics

1) Intelligent knowledge discovery in Bioinformatics

A considerable quantity was obtained by the Human Genome Project [16], as well as other large-scales biological experiments. The bioinformatics company is now engaged in big data and faces issues like sequence, expression, structure and understanding of pathways [17]. AI and heuristic methods are highly important for current and future advances in bioinformatics. It is generally accepted today that these two possible realms overlap [18]. Bioinformatics is a modern, interdisciplinary and strategic research area which integrates, through information technology and informatics, and analyses the complexities of any biological data. This area of study seeks to produce new algorithms and software, data-storage approaches and contemporary computing architectures in order to meet computer requirements [19]. The architecture of the algorithm is a step-by-step method for estimation, data analysis and automatic reasoning (a list of well-defined instructions). In fact, to measure a function, an algorithm is applied. For example, in order to work out the Entscheidungs problem, a partial formalisation of the term was adopted [20]. The Bioinformatics Development and the Implementation of Novel Algorithms basically addresses four areas of analyses, including DNA sequence analysis, protein structure prediction, genomics and proteomics accessibility, and system biology [21]. In the area of bioinformatics, finding solutions to biological problems is where DM approaches can be used effectively. Both DM and bioinformatics are research fields that are fast developing [22]. A significant amount of raw data has been generated by the growth of information storage technology, taking into account two aspects: the advancement of algorithms and the emergence of modern storage facilities. Such raw data contains substantial detail. In the 1990s, in order to derive information from libraries, researchers used knowledge discovery from data (KDD). Discovery of knowledge is the nontrivial retrieval from evidence of tacit, previously unknown, and likely valuable facts [23]. Appropriate time complexity, precision, comprehensibility and valuable outcomes are of course, essential characteristics that should be taken into account in the extraction of new information. DM is

a synonym for KDD [24].

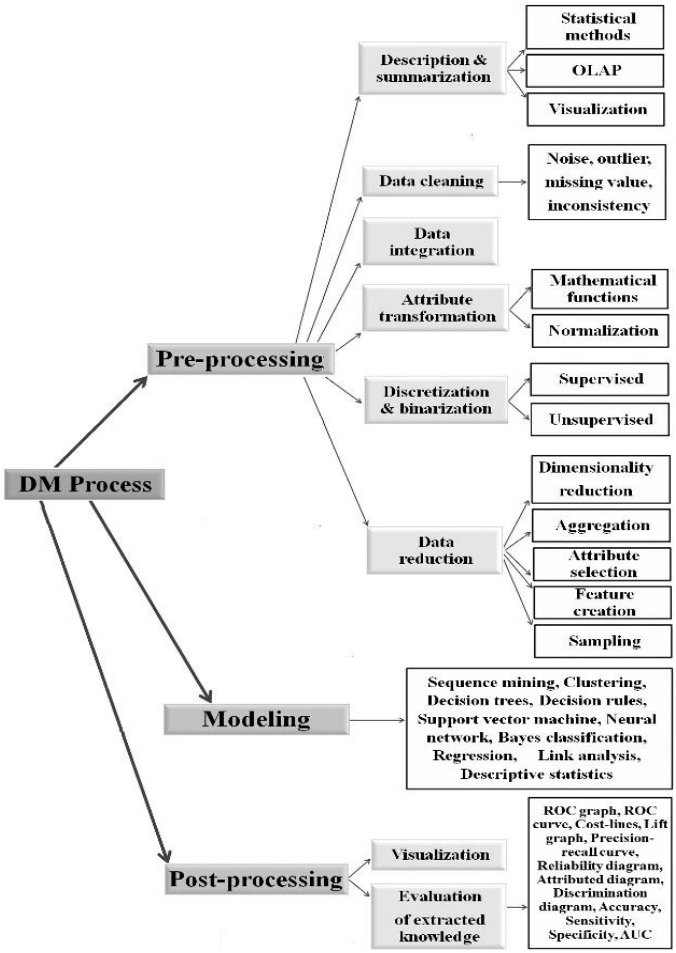

Figure 5: Flow chart (basic concepts of data mining)

The DM technology has three stages, in particular details pre-preparation, data demonstration and data post-treatment. Crude knowledge in the main stage is ready for mining. Due to the widely scattered, unrestricted period and usage of various bio-knowledge, effective and organised bio-database investigations, the information processing, pre-preparation of information and literary consolidation is important [25]. The subsequent stage has to be related to important new trends between various data-bases [26]. The basic requirements of DM are goals and meanings in this way. The measurables (for instance, gathering, relapse, the investigation of time scheduling, appraisal and the like using established qualities) will determine unknown value. To check the derived information, context knowledge may also be used. Schematic description of potential DM method inputs and consequently possible DM algorithm predictions and outputs that exploit several genome-scale datasets. In the upper circles are found various inputs/data sets such as single-nucleotide polymorphisms, organic environments, chromosome preparation, phylogenetic details, gene expression profiles, DNA/RNA/protein structure, and biochemical pathways [27]. The most popular DM algorithms and procedures are shown in the circle. The lower segment of the circle displays different forms of potential DM yields. These findings include the representation of proteins, the representation of datasets and pathways, the depiction of succession DNA and RNA, and the depiction of associations.

I. Data Mining

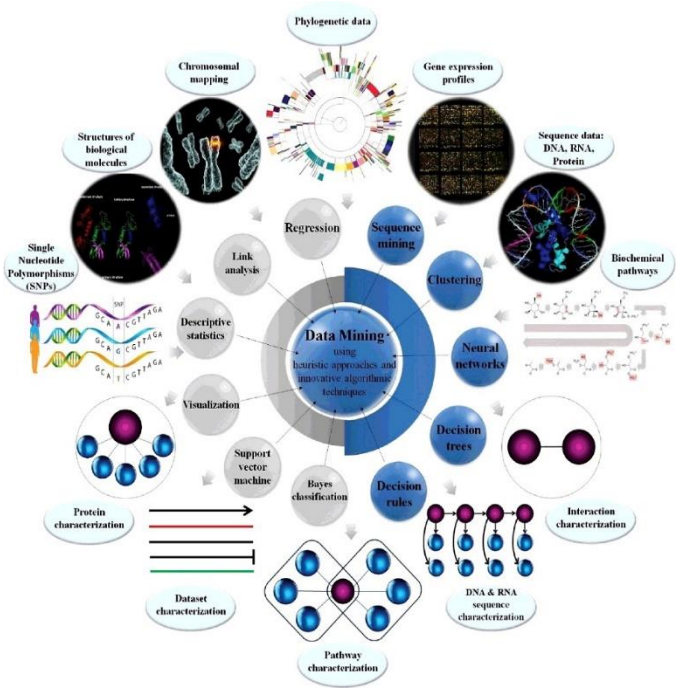

Figure 6: Flow scheme 3 **(**Schematic overview of possible inputs for DM process and subsequently possible predictions and outputs from DM algorithms leveraging many genome-scale datasets).

DM models are categorised according to parameters such as the form of the data source mined, data model, Relational Data, Data Model, Object Dependent Data Model, the mining methodology e.g. machine learning, and genetic algorithms (GA). The degree of user engagement involved in DM can also be included in the classification. In diverse situations and choices, a robust framework should include numerous DM methods that are suitable and reflect different types of user engagement [28]. It is possible to categorise DM methods and strategies into three main classes, i.e. supervised learning techniques, unsupervised learning techniques and others. Classification and estimation functions are involved in the first category. The second group is the mining of clustering and membership law. Again there are no direct or unmonitored learning protocols for such undertakings. They are in the third floor anyway. The DM power rundown, however, is not over [29], for example, sequence mining, sampling, decision treeing, and decision rules (orders) are the Bayes characterization, reciprocating, relation inquiries, spell-bundling insight and representing. DM behaviours need fitting calculations to be examined. The choice of DM and algorithm is based on the structural needs and the algorithm of data-bases highlight as do the limits of the optimal estimation. Many DM tests, such as inspections, arrangement minings, strings, computer analysis and data base hypotheses were actually considered [22].

J. Neural networks

Initially, the term neural network refers to a biological neuron circuit. Nonetheless, its contemporary use is with regards to ANNs, which contain programming arrangements looking like the capacity of counterfeit neurons, or nodes . From neural transmitter dissemination, electrical signalling and different types of flagging emerge.

Accordingly, with the advancement of various biological information bases holding DNA/RNA arrangements, protein structures and successions, and other macromolecular structures, neural organizations are very convoluted and have gotten one of the basic methods in the bioinformatics region [30]. In bioinformatics, forecast is the most oftentimes discovered expertise of neural organizations, particularly in instances of a limited quantity of accessible crude information that can be utilized to extricate the expectation model [16]. In different fields of bioinformatics, AI approaches can be utilized, support vector machine for protein overlay recognition, shrouded Markov model (HMM) for arrangement and profile arrangement, Bayesian organizations for gene networks and ANNs for forecast of protein secondary structure, grouping of illnesses and identification of biomarkers [31]. Network based researches have been generally utilized in malignancy research owing to gene cooperation in useful molecular networks to remember sub-atomic delineation for malignant growth patients, to gauge infection results, to consider tumourigenesis and the component of activity of tumor-prompting infections, to anticipate the cancer-causing nature of substance mixes and to prioritise the hurtful impacts of disease changes [32].

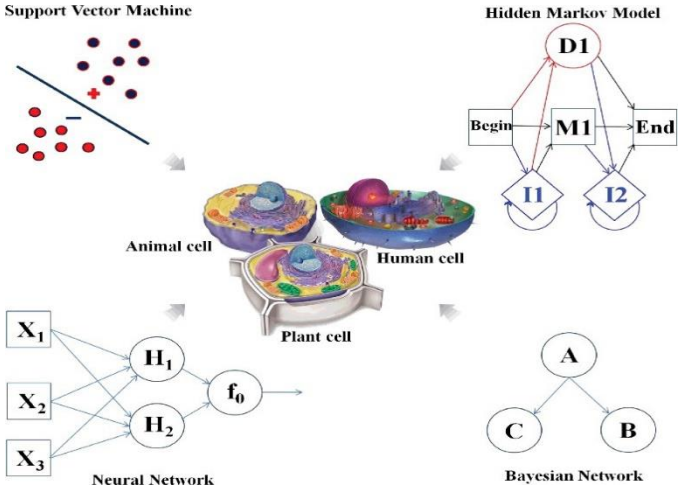

Figure 7: (Schematic overview of machine-learning applications in bioinformatics)

Protein secondary structure information, on the other hand, can help to explain human diseases and to improve therapeutic enzymes and drugs. Therefore for protein secondary structure estimation, different AI techniques are applied. Where there are strongly nonlinear and complex correlations, traditional mathematical methods such as discriminatory regression and generalised linear models have limitations. Actually, computer programming is made possible by machine learning to improve biological data set efficiency [16].

2. Conclusion

Bioinformatics can play an important role in processing large data sets in the future by using AI to save time and money. Biological discoveries, especially in medicine, biomedical fields, and robotic surgery, will also be accelerated. We may see an increased role of AI in bioinformatics, provided such a climate. The review of antibody-antigen interactions and their diversity, and the study of epidemiological evidence that can help forecast antibody-antigen interactions and the induction of broadly neutralising antibodies are important

questions to be answered in the field of vaccinology. Vaccine-deficient infectious diseases such as measles, flaviviruses, tuberculosis, HIV, hepatitis C and others will be gradually analysed in silico, while a limited number of primary confirmation studies will be carried out.

References

[1]. S.A. Sehgal, A.H. Mirza, R.A. Tahir, and A. Mir (2018). "Quick Guideline for Computational Drug Design 2018." Bentham Science Publishers.

[2]. S.A. Sehgal, N.A. Khattak, and A. Mir (2013). "Structural, phylogenetic and docking studies of D-amino acid oxidase activator (DAOA), a candidate schizophrenia gene." Theoretical Biology and Medical Modelling, 2013, vol. 10(1), pp.3.

[3]. S.A. Sehgal, S. Mannan, S. Kanwal, I. Naveed, and A. Mir (2015). "Adaptive evolution and elucidating the potential inhibitor against schizophrenia to target DAOA (G72) isoforms." Drug design, development and therapy, vol. 9, pp. 3471.

[4]. P. K. Donepudi (2018). "AI and Machine learning in pharmacy: Systematic review of related literature." ABC journal of advanced research, Vol.7 (2), pp. 109-112.

[5]. S. A. Sehgal, M. Hassan, and S. Rashid (2014). "Pharmacoinformatics elucidation of potential drug targets against migraine to target ion channel protein KCNK18." Drug design, development and therapy, Vol. 8, pp. 571.

[6]. P. K. Donepudi (2015). "Crossing point of artificial intelligence in cybersecurity". American journal of trade and policy, Vol. 2 (3), 121-128. https://doi.org/10.18034/ajtp.v2i3.493

[7]. H. Golkarieh (2019). "How AI is shaping the future of Bioinformatics." Retrieved November 14, 2020, from https://medium.com/optima-ai/how-ai-is-shaping-the-future-of-bioinformatics-f4aa17bce5a6

[8]. R. Agrawal, and R. Srikant (1994). "Fast algorithms for mining association rules." BMC Bioinformatics, Vol. 3(35), pp. 12-16.

[9]. D. Bhandari, C.A. Murthy, and S.K. Pal (2012). "Variance as a stopping criterion for genetic algorithms with elitist model." Fundamata Informaticae, Vol. 120, pp. 145-164.

[10]. P. K. Donepudi (2016). "Influence of cloud computing in business: are they robust?" Asian journal of applied science and engineering, Vol 5(3), pp. 193-196. DOI: https://doi.org/10.5281/zenodo.4110308

[11]. C. Burge, and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." Journal of Molecular Biology, Vol. 268, pp. 78-94.

[12]. H. Douzono, S. Hara, and Y. Noguchi Y (1998). "An application of genetic algorithm to DNA sequencing by oligonucleotide hybridization." Proceedings of the IEEE international joint symposia on intelligence and systems Rockville, Maryland, USA. 5(34), pp. 92–98

[13]. F. Corpet (1988). "Multiple sequence alignment with hierarchical clustering." Nucleic Acids Research, Vol. 16 (22), pp. 10881- 10890

[14]. Y.C. Chung, and L.H. Randy (2000). "Amplitude and phase adaptive nulling with a genetic algorithm." Journal of Electromagnetic Waves and Applications, Vol. 14, (5), pp. 631-649.

[15]. N. Cannata, M. Schröder, R. Marangoni, and P.A. Romano (1992). "Semantic Web for bioinformatics: goals, tools, systems, applications." BMC Bioinformatics, Vol. 9(4), pp.1.

[16]. L. Hunter, "Artificial intelligence and molecular biology." San Jose (CA), AAAI Press.

[17]. G. Valentini, R. Tagliaferri, and F. Masulli (2009). "Computational intelligence and machine learning in bioinformatics." Artif. Intell. Med., Vol. 45, pp. 91–96.

[18]. J. Pitrat (1996). "Artificial intelligence and heuristic methods." Revue Francaise De Recherche Operationnele, Vol. 10, pp. 137–137.

[19]. S. Kumar, T.W. Banks, and S. Cloutier (2012). "SNP discovery through next-generation sequencing and its applications." Int J Plant Genom, [cited 2017 Feb 10];2012:831460. DOI:10.1155/2012/831460

[20]. D. Hilbert, J.V. Neumann, and L. Nordheim (1928). "Über die grundlagen der quantenmechanik." On the fundamentals of quantum mechanics], Math Ann. Vol. 98, pp.1

[21]. M. Al-Haggar, B. Khair-Allaha, and M. Islam (2013). "Bioinformatics in high throughput sequencing: application in evolving genetic diseases." Jour. Data Mining Genomics Proteomics, vol. 4, 131. DOI: 10.4172/2153-0602.1000131.

[22]. Z. Ezziane (2020). "Applications of artificial intelligence in bioinformatics: A review." Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417405002344

[23]. G. Piatetsky-Shapiro, and W. Frawley (1991). "Knowledge discovery in databases." San Jose (CA), AAAI/MIT Press.

[24]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth (1996). "From data mining to knowledge discovery in databases." AI Mag, 1996, Vol. 17, 37–54.

[25]. J. Han (2002). "How can data mining help bio-data analysis?" Paper presented at: BIOKDD02: Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference), Edmonton (Canada). Available from: https://web.njit.edu/~wangj/publications/biokdd02/01-han.pdf

[26]. N. Esfandiari, M.R. Babavalian, and A.M.E. Moghadam (2014). "Knowledge discovery in medicine: current issue and future trend." Expert Syst Appl., Vol. 41, pp. 4434–4463.

[27]. N. Padhy, P. Mishra, and R. Panigrahi (2012). "The survey of data mining applications and feature scope." International Journal Comp Sci Eng Inf Tech. [cited 2017 Feb 10], Vol. 2, 2. DOI:10.5121/ijcseit.2012.2303.

[28]. G. Piatetsky-Shapiro (2017). "CRISP-DM, still the top methodology for analytics, data mining, or data science projects [Internet]." Available from: http://www.kdnuggets.com/2014/10/ crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[29]. O. Niakšu (2015). "Development and application of data mining methods in medical diagnostics and healthcare management." Dissertation, Vilnius: Vilnius University.

[30]. S.A. Sehgal (2017). "Pharmacoinformatics and molecular docking studies reveal potential novel Proline Dehydrogenase (PRODH) compounds for Schizophrenia inhibition." Medicinal Chemistry Research, 2017, Vol. 26(2), pp. 314-326

[31]. K. Moore, and H. Passley (2020). "Leveraging the Benefits of Artificial Intelligence Technology in Bioinformatics." Retrieved from https://www.bbntimes.com/technology/ leveraging-the- benefits-of-artificial-intelligence-technology-in-bioinformatics

[32]. G.B. Fogel (2008). "Computational intelligence approaches for pattern discovery in biological systems." Briefings in Bioinformatics, 2008, Vol. 9 (4), pp. 307–316.