IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-12 E-ISSN: 2347-2693

Next Generation Sequencing: an Emerging Bioinformatics Field

A. Jiwan[1*], S. Singh[1]

[1*]Dept. of CSE, Punjab Engineering College (Deemed to be University), Chandigarh, India
[1]Dept. of CSE, Punjab Engineering College (Deemed to be University), Chandigarh, India

**Corresponding Author: ankita9506@gmail.com, Tel.: +91-78372-13387*

Available online at: www.ijcseonline.org

Received: 28/Nov/2017, Revised: 03/Dec/2017, Accepted: 22/Dec/2017, Published: 31/Dec/2017

***Abstract*—** DNA sequences are the store house of all the biological information. DNA sequencing focuses on determining the exact order of the nucleotides in a DNA sequence. Many research efforts have been put to the development of cheaper and increasingly higher-throughput sequencing techniques. This lead to development of a massively parallel and efficient method called as Next Generation Sequencing (NGS). The massively parallel NGS technologies have a high throughput with reduced cost. This paper gives a brief working principle of sequencer such as Roche 454 (GS FLX Titanium/GS Junior), Illumina (Genome Analyzer/HiSeq 2000/MiSeq) and Life Technologies (SOLiD/Ion Torrent PGM) along with a their comparison.

***Keywords*—** Next generation sequencing, Roche, Illumina, SOLiD, Sanger

I. INTRODUCTION

Bioinformatics is the science in which mathematical theories and computational technologies are exercised in order to process, relate, interpret and derive inferences from data obtained in molecular biology. Biology, mathematics and computer science technologies collaborate amidst each other with their own richness and limitations emerging together with the intern to interpret the information control and flow within different organisms [1][2].

Bioinformatics is the name given to interaction in which mathematical and computing approaches used to gather deeper understanding of biological processes. Bioinformatics can also be stated as application of the power computer technology for management of biological information. To gather, store, visualize, asses and make deductions about biological and genetic information available computers are used. Therefore, computational power is required to manage, analyse, interpret and manipulate biological data. Essentially, this process can be divided into three components:

Managing biological data: This includes storage and management of large biological data. The data is collected in publicly available databanks, from which data can be submitted and retrieved easily. The nucleotide sequences for all organisms are stored in three databases, DDBJ (DNA Data Bank of Japan) [3], EMBL (European Bioinformatics Institute) [4][5] and GenBank [6]. These are the primary databases. The accumulated data needs to be retrieved in a meaning full way; hence secondary databases have been created catering to a specification interpretation of the biological data. Numerous secondary databases are available, below are a few of these databases with examples. Genome Databases (for example: Human Gene Mutation Database [7], SNPedia [8]), Protein Sequence Database (UniProt [9], Protein Information Resource [10]), Protein Structure Database (Protein DataBank [11], SCOP [12]), RNA Database (Rfam [13], C-It-Loci [14]), Protein-protein and other Molecular Interactions (BIND [15], BioGRID [16]), Gene Expression Database (ArrayExpress [17], GEO [18]).

Analysing biological data: This includes designing methods, algorithms, tools, resources and processes to throw light on biological data in a meaningful manner. After consideration, biological problems can be grouped as:

- Sequence based: it covers DNA and RNA alignment of sequences, comparison between sequences for mutation identification, finding functionally significant religions by pattern discovery, locating coding regions by pattern discovery, finding genetic disease mutations, discovering evolution history from phylogenetic trees.
- Structure based: it includes stable RNA and proteins 3D structure prediction, to determine folding sites which define their stable state, to identify the purpose of the 3D structure.
- Functionality based: it encompasses the living organisms metabolism reactions and its regulatory network.

- Pharmacology based: it is comprised of drug design, drug targets for patients, drug resistance causes.

- Interpreting biological data results: An abundance of additional information is made available by analysis of the biological data. Therefore, in order to tap into this interpretation and visualization tools are needed.

This paper discusses the DNA sequencing by next generation sequencing technology. It further provides comparison amongst the various commercially available platforms like Roche 454 (GS FLX Titanium/GS Junior), Illumina (Genome Analyzer/HiSeq 2000/MiSeq) and Life Technologies (SOLiD/Ion Torrent PGM) for next generation sequencing. Lastly the impact of next generation sequencing in the field of bioinformatics is dicusses

II. APPLICATIONS OF BIOINFORMATICS

Bioinformatics helps in deciphering the process of evolution, inheritance, disease and the nature of life. Although the data and problems dealt in bioinformatics are biological in nature, yet the techniques applied are computational such as databases, data mining, pattern recognition, neural networks, and other machine learning techniques, etc. Bioinformatics has strongly influenced the fields of drug design, forensic data analysis and agricultural sector [19] [20].

Broadly the applications on which research is being carried out in the field of bioinformatics can be grouped into three main categories namely sequence analysis, functional analysis and structure analysis as shown in Fig. 1.

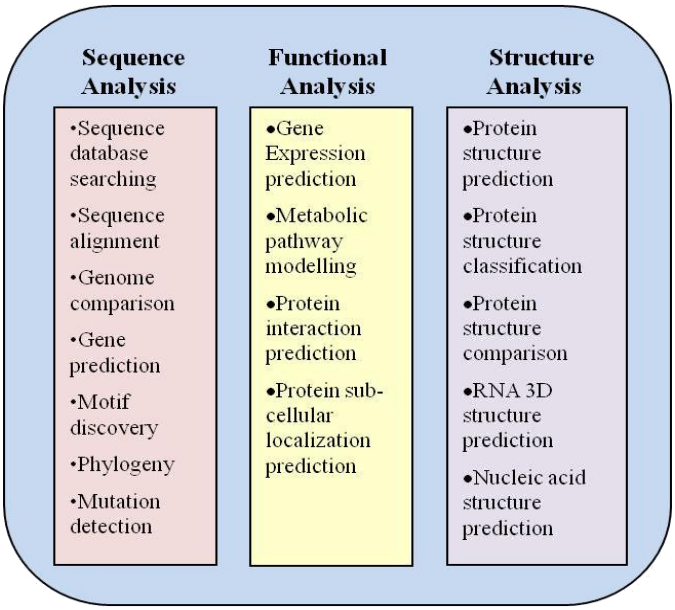

Fig. 1. Applications of bioinformatics

Sequence Analysis:
All the applications that focus on analysing various types of sequence data fall under this category. It includes sequence database searching, sequence alignment, genome comparison, gene prediction, motif discovery and phylogeny.

Functional Analysis:
Functional analysis includes all the applications that aim to discover the functionality encoded inside the sequence and predict the functional interaction that exists between proteins and genes. It includes gene expression prediction, metabolic pathway modeling, protein interaction prediction, protein sub-cellular localization prediction.

Structure Analysis: The structure of an RNA or protein plays a vital role in determining its interactions and functionality. All the applications devoted to predict structure of protein and RNA along with its functionality fall in this category. It includes protein structure prediction, protein structure classification, protein structure comparison, RNA 3D structure prediction and nucleic acid structure prediction.

III. NEXT GENERATION SEQUENCING

A DNA polymerase-based chain-termination sequencing method was developed by Sanger and Coulson in the 1970s for DNA sequencing. This method was the first enzyme-based approach, exclusively carried out with semi automated capillary electrophoresis and is known as Sanger sequencing [22]. The method was able to identify upto 80 nucleotides precisely from the 5-end and the following 50 nucleotides with reduced precision. Over the years the method was optimized but the primary limitation remained of low throughput.

Sanger methods were used in Human Genome Project (HGP); still the overall process was complex and time consuming and thus involved major expenses [23]. This triggered research efforts in development of cheaper and increasingly higher throughput sequencing techniques leading to a massively parallel and efficient method called as Next Generation Sequencing (NGS) in 2005. The sequencing cost was about $10 per base in 1985, and dropped 10,000 times by 2005 [24]. Heaps of DNA strands in millions or billions are sequenced in parallel, resulting in significantly higher throughput and reducing the need for the fragment cloning methods that was used in Sanger sequencing of genomes.

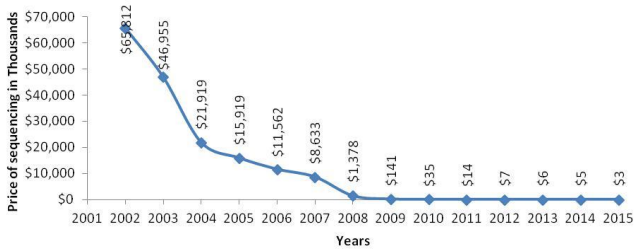

Fig. 2. Cost per Megabyte of DNA Sequencing

The NGS technologies are different from the older Sanger method in aspects that they have massively parallel analysis, high throughput, and reduced cost. This has caused an explosion of generated DNA sequence data and a steady drop in sequencing costs. The Fig. 2. shows how the cost of sequencing has drastically reduced from Sanger based technologies in 2001 to Next Generation Sequencing technology in 2005 and the steady decline in cost ever since [25].

NGS consists of several different platforms each having a distinct set of characteristics. Although the technologies differ in their biochemistry all of them work on the principle of cyclic array sequencing, where DNA features are extracted from a dense array by iterative cycles of enzymatic reactions followed with imaging-based data detection. The main commercially available NGS platforms are: Roche 454 (GS FLX Titanium/GS Junior), Illumina (Genome Analyzer/HiSeq 2000/MiSeq) and Life Technologies (SOLiD/Ion Torrent PGM).

Roche 454 GS FLX was the first NGS based platform developed by 454 Life Sciences in 2004 based on massively parallel sequencing-by-synthesis technology. The sequencing platform uses pyrosequencing method [26] for sequencing and emulsion polymerase chain reaction (PCR) for amplification of the light signal. Roche 454 systems advantages are speed and the read length in comparison to other NGS technologies. The latest 454 Roche system called GS Junior plus allows a read length of approximately 700-800 bases, with an accuracy of 99% accuracy at 700 bases and higher for preceding bases [27].

Illumina HiSeq 2500 sequencing platform was the first "short read" sequencing platform based on widely available high throughput sequencing technology [24][28][29]. The Illumina system works on the same principle as Roche 454 technology of massively parallel 'sequencing-by-synthesis'. It uses reversible terminator-based sequencing chemistry method for sequencing and bridge amplification for better imaging. Illumina systems have a higher error rate but produces massive amounts of data at extremely low price.

SOLiD (Supported Oligonucleotide Ligation and Detection) was introduced in 2006 supporting a version of high throughput sequencing chemistry. This is a short read technology based on ligation developed at Agencourt Personal Genomics [30][31]. SOLiD 4 sequencing system uses sequencing by ligation method for identifying the order of nucleotides in the DNA sequence and emulsion polymerase chain reaction (PCR) for amplification [32].

The comparison of NGS platforms including 454 GS FLX, HiSeq 2500, SOLiD 4 System amongst each other and with Sanger sequencing technique is shown in Table 1. 454 GS FLX sequencing platform uses pyrosequencing method for sequencing and emulsion PCR for amplification, whereas HiSeq 2500 sequencing platform uses sequencing by synthesis method for sequencing and bridge amplification, while SOLiD 4 sequencing system uses sequencing by ligation method for sequencing and emulsion PCR for amplification, whereas Sanger sequencing systems use dideoxy chain termination for sequencing and PCR for amplification.

Table 1 The table shows a comparison between the next generation sequencing platforms and Sanger sequencer

Parameters	454 GS FLX	HiSeq 2500	SOLiD 4 System	Sanger
Sequencing method	Pyrosequencing	Sequencing by synthesis (Reversible dye terminators)	Sequencing by ligation and two base coding	Dideoxy chain termination
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR	PCR
Read lengths	700 bp	2x125 bp in standard mode, 2x150 bp in rapid mode	50x50 bp paired end	400-900 bp
Data output per run	0.7 GB	800 GB in standard mode, 180GB in rapid mode	120 GB	1.9~84 Kb
Time per run	24 hours	3-12 days in standard mode, 1-5 days in rapid mode	7–14 days	20 minutes - 3 hours
Accuracy	99.90%	99.90%	99.94%	99.99%

Read lengths of 454 GS FLX, HiSeq 2500, SOLiD 4 System and Sanger sequencing are 700bp, 2x125bp, 50x50 and 400-900bp respectively. The data generated after every run is 0.7GB, 800GB, 120GB and 1.9-84KB for 454 GS FLX, HiSeq 2500, SOLiD 4 System and Sanger sequencing respectively. From evaluation it is can be deduced that all DNA sequencing systems have high accuracy, but still Sanger has the highest accuracy of 99.99% followed by SOLiD 4 system with 99.94% and then 454 GS FLX and HiSeq 2500 with 99.90% of accuracy.

IV. NEXT GENERATION SEQUENCING AND BIOINFORMATICS

The Next Generation Sequencing analysis process is complex includes multiple analysis steps and involves handling large amounts of heterogeneous data. Numerous bioinformatics techniques are used during the analysis process of NGS data. The analysis process includes quality evaluation of the sequence generated. In this step the accuracy with which the nucleotides are detected by the sequencer is evaluated. To analyse the quality of the generated data, various image processing techniques are used. After selecting sequences with high quality, it is aligned to a reference genome. In this step various algorithms are used for alignment and numerous algorithms are available for the same. The next step is most crucial in which variations in the sequence are identified; these comprise of simple nucleotide variations (SNVs), including single-nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs), structural variations (SVs) and copy number variations (CNVs) [33]. The variant identification step requires computational techniques such as Probabilistic model, Bayesian model, Heuristics and Statistical algorithms to achieve their goal[34]. The last step in the analysis process is validation visualization of the variants identified. This step also requires data processing and a number of web based applications are available for the same [33].

Next Generation Sequencing has numerous bioinformatics applications in disease research, diagnosis and therapy, sequencing of bacterial and viral genomes, understanding the genetic mechanisms underlying human gene expression variation, understanding tissues and organisms, genome-wide profiling of DNA-binding proteins and epigenetic marks, comparative biological studies and many more. NGS has started replacing microarrays in areas of clinical diagnostics and gene expression.

V. CONCLUSION

Bioinformatics is an interdisciplinary field which utilises the power of computational technologies to gain better knowledge of living organisms. It explores new ways to contemplate DNA\RNA sequences, RNA\protein structures, metabolic\regulatory functions and human pathology. The availability of large volumes of data at a low cost has steered the research towards Next Generation Bioinformatics. This has lead to development of new bioinformatics techniques for storage and analysis of the data generated. Various next generation platforms exist, which follow the same basic principle but vary in their internal biochemistry. A comparison among theses platforms is available. A vast number of tools are available for the analysis of next generation sequencing data and this list is ever increasing. The paper discusses the impact of next generation sequencing in the field of bioinformatics. There is immense scope for advancement in the field of Next Generation Bioinformatics.

REFERENCES

[1] A. Jiwan and S. Singh, *"A review on RNA pseudoknot structure prediction techniques,"* in 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET). IEEE, 2012, pp. 975–978.

[2] R. Garc´ıa, *"Prediction of RNA pseudoknotted secondary structure using Stochastic Context Free Grammars (SCFG),"* CLEI Electronic Journal, vol. 9, no. 2, 2006.

[3] Y. Tateno, T. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou,
H. Sugawara, and T. Gojobori, *"DNA Data Bank of Japan (DDBJ) for genome scale research in life science*," Nucleic Acids Res, vol. 30, no. 1, pp. 27–30, 2002.

[4] G. H. Hamm and G. N. Cameron, "*The EMBL data library,"* Nucleic Acids Res, vol. 14, no. 1, pp. 5–9, 1986.

[5] C. Kanz, P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates,
P. Browne, A. van den Broek, M. Castro, G. Cochrane et al., "*The EMBL nucleotide sequence database*," Nucleic Acids Res, vol. 33, no. suppl 1, pp. D29–D33, 2005.

[6] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "*Genbank*," Nucleic Acids Res, vol. 41, no. D1, pp. D36–D42, 2013.

[7] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. Thomas, S. Abeysinghe, M. Krawczak, and D. N. Cooper, "*Human gene mutation database (HGMD R): 2003 update,"* Hum Mutat, vol. 21, no. 6, pp. 577–581, 2003.

[8] M. Cariaso and G. Lennon, "*SNPedia: a wiki supporting personal genome annotation, interpretation and analysis*," Nucleic Acids Res, vol. 40, no. D1, pp. D1308–D1312, 2012.

[9] U. Consortium et al., "*Reorganizing the protein space at the Universal Protein Resource (UniProt),"* Nucleic Acids Res, p. gkr981, 2011.

[10] C. H. Wu, L.-S. L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek et al., *"The protein information resource,"* Nucleic Acids Res, vol. 31, no. 1, pp. 345–347, 2003.

[11] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "*The protein data bank,"* Nucleic Acids Res, vol. 28, no. 1, pp. 235–242, 2000.

[12] A. Andreeva, D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and
A. G. Murzin, "*SCOP database in 2004: refinements integrate structure and sequence family data*," Nucleic Acids Res, vol. 32, no. suppl 1, pp. D226–D229, 2004.

[13] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, *"Rfam: an RNA family database,"* Nucleic Acids Res, vol. 31, no. 1, pp. 439–441, 2003.

[14] T. Weirick, D. John, S. Dimmeler, and S. Uchida, "*C-It-Loci: a knowl-edge database for tissue-enriched loci,"* Bioinformatics, p. btv410, 2015.

[14] G. D. Bader, D. Betel, and C. W. Hogue, "*BIND: the biomolecular interaction network database,"* Nucleic Acids Res, vol. 31, no. 1, pp. 248–250, 2003.

[15] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and

M. Tyers, "*BioGRID: a general repository for interaction datasets,"* Nucleic Acids Res, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[16] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena,
R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk et al., "*ArrayExpressa public database of microarray experiments and gene expression profiles*," Nucleic Acids Res, vol. 35, no. suppl 1, pp. D747–D750, 2007.

[18] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evan-gelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "*NCBI GEO: mining tens of millions of expression profilesdatabase and tools update,"* Nucleic Acids Res, vol. 35, no. suppl 1, pp. D760–D765, 2007.

[19] R. Backofen and D. Gilbert, "*Bioinformatics and constraints,"* CSTR, vol. 6, no. 2-3, pp. 141–156, 2001.

[20] N. M. Luscombe, D. Greenbaum, and M. Gerstein*, "What is bioinfor-matics? an introduction and overview*," Yearb Med Inform, vol. 1, pp. 83–99, 2001.

[21] M. V. Schneider, J. Watson, T. Attwood, K. Rother, A. Budd, J. Mc-Dowall, A. Via, P. Fernandes, T. Nyronen, T. Blicher et al., "*Bioin-formatics training: a review of challenges, actions and support require-ments,"* Briefings in bioinformatics, p. bbq021, 2010.

[22] M. J. Munoz and D. L. Riddle, "*Positive selection of caenorhabditis el-egans mutants with increased stress resistance and longevity,"* Genetics, vol. 163, no. 1, pp. 171–180, 2003.

[23] M. P. Sawicki, G. Samara, M. Hurwitz, and E. Passaro, "*Human genome project*," Am J Surg, vol. 165, no. 2, pp. 258–264, 1993.

[24] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti, "*BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies,"* Nucleic Acids Res, vol. 34, no. 3, p. e22, 2006.

[25] KA Wetterstrand, "*DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)"* National Human Genome Research Institute the cost of sequencing a human genome, 2016.

[26] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen,´ and P. Nyren,´ "*Real-time DNA sequencing using detection of pyrophosphate release*," Anal. Biochem., vol. 242, no. 1, pp. 84–89, 1996.

[27] J. M. Heather, and B. Chain, "*The sequence of sequencers: the history of sequencing DNA*," Genomics, vol. 107.1, pp. 1-8, 2016.

[28] G. Turcatti, A. Romieu, M. Fedurco, and A.-P. Tairi, "*A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis,"* Nucleic Acids Res, vol. 36, no. 4, p. e25, 2008.

[29] C. Adessi, G. Matton, G. Ayala, G. Turcatti, J.-J. Mermod, P. Mayer, and E. Kawashima, "*Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms,"* Nucleic Acids Res, vol. 28, no. 20, p. e87, 2000.

[30] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon,
A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church, "*Accurate multiplex polony sequencing of an evolved bacterial genome*," Science, vol. 309, no. 5741, pp. 1728–1732, 2005.

[31] K. McKernan, A. Blanchard, L. Kotler, and G. Costa, "*Reagents, methods, and libraries for bead-based sequencing,"* Feb. 1 2006.

[32] S. Myllykangas , J. Buenrostro , and H. P. Ji "Bioinformatics for High Throughput Sequencing", Springer, pp 255, 2012 .

[33] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova,
B. Krabichler, M. R. Speicher, J. Zschocke, and Z. Trajanoski, "*A survey of tools for variant analysis of next-generation genome sequencing data*," Briefings Bioinf., vol. 15, no. 2, pp. 256–278, 2014.

[34] R. Bao, L. Huang, J. Andrade, W. Tan, W. A. Kibbe, H. Jiang, and
G. Feng, "*Review of current methods, applications, and data man-agement for the bioinformatics analysis of whole exome sequencing*," Cancer Inform, pp. 67–83, 2014.

Authors Profile

Dr. S Singh is currently working as Professor in Computer Sc. and Engineering Department at Punjab Engineering College (Deemed to be University), Chandigarh, India. His research interest includes bioinformatics, natural language processing, speech technology and soft computing. He has received several awards and recognition in the field. He is member of various professional societies like IEEE, IEEE Computational Intelligence Society, Computer Society of India etc. He has contributed various research papers in International and national journals/Conferences.

Mrs A. Jiwan received the BTech degree in computer science and engineering from B.B.S.B.E.C, Fatehgarh Sahib, India. The ME degree from Punjab Engineering College, Chandigarh, India. She is currently pursuing PhD degree in computer science and engineering from Punjab Engineering College (Deemed to be University), Chandigarh in the area of bioinformatics. Her research interests include bioinformatics and softcomputing.

