Bioinformatics Approaches for Improvement of Biohydrogen Production: A Review

Muhammad Jawed, Wang Jun, Pi Jian, Xu Li, and Yunjun Yan

***Abstract*—Hydrogen (H_2) production from biomass is considered as the main source of renewable energy, is a characteristic feature of prokaryotes. H_2 is believed as the cleanest fuel without evolution of greenhouse gases on combustion. The main biological processes for H_2 production are: biophotolysis of water by algae and cyanobacteria, dark fermentation and photo-fermentation. Since last decades, a lot of work has been carried out for understanding and refining bio-hydrogen production and still it has to overcome some of the serious limitations so that it becomes a viable proposal. The bottlenecks include thermodynamic inefficiency, trouble in using lignocellulosics as feedstock, cost of raw material and low H_2 molar yields (HMY). To get rid of these major problems, the conventional approach is inadequate and people has to dynamically think modern bioinformatics approaches to overcome these factors. The accessibility of enormous sequenced genomes, functional genomic studies, and the progress of in-silico models at the genome level, metabolic pathway reconstruction, and synthetic biology tactics predict engineering strategies to enhance H_2 production in an organism. This review examines the current status and progressions that have been made in the area of biotechnology and bioinformatics, to understand and enhance H_2 evolution to overcome current limitations and make H_2 production from biological means, a reality in the coming future.**

***Index Terms*—Biohydrogen, in-silico metabolic engineering, functional genomics, synthetic biology.**

I. Introduction

Studies on the feasibility and viable production of renewable fuels from biological means as an alternative to fossil fuels has enhanced in recent years and additionally have been increased further by various ecological problems related with fossil fuels, such as greenhouse gas emission, global warming and higher price level hikes with an unstable supply. Biological Hydrogen Production (BHP) plays a vital role since it is supposed as the sparkling fuel with no emission of greenhouse gases on combustion [1].

Apart from wet lab experiments, in silico approaches, functional genomics, metabolic modifications on genomic level and flux balance analysis can be used to improve the hydrogen (H_2) producing abilities [2]. In silico models at genomic level provide a powerful resource for logical engineering of biological systems for improvement in BHP [3], [4]. A precise genome-scale model of an organism helps us in studying the effect of environmental and genetic limitations on it, and hence these facts motivate experiments in the area of metabolic engineering. Ever since the development of the first genome scale model in *Haemophilus influenzae* [5], the obtainable high-throughput biological data have been utilized professionally for structures level modeling approaches. The construction and modeling of biological components, functions and organisms non-existing in nature or redesigning present biological organizations to perform novel functions is termed as Synthetic biology [6]. In this review, we discuss the utilization of bioinformatics approaches and techniques to improve H_2 production to overcome the current difficulties (Fig. 1).

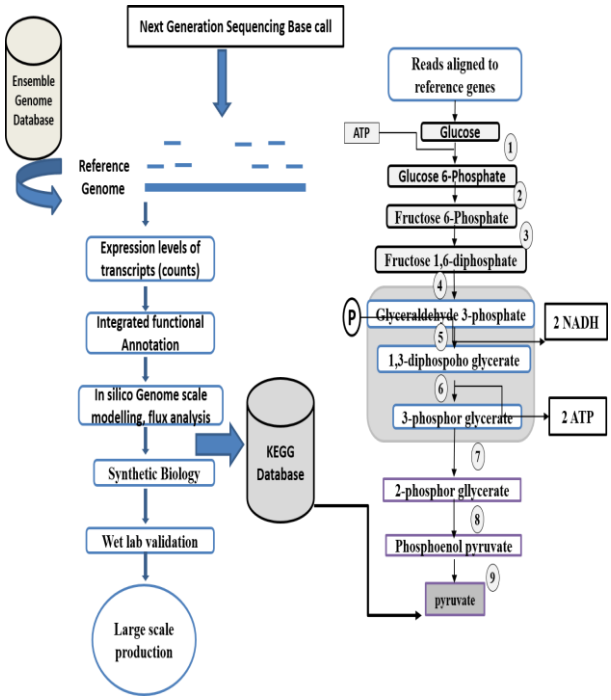


Fig. 1. Scheme represents various approaches for enhancement of H_2 production.

II. Methods for Biohydrogen Production

For H_2 production from biomass, direct or indirect bio-photolysis, photo-fermentation and dark-fermentation methods are often used. Photo-autotrophic microorganisms like cyanobacteria or green microalgae used the method of bio-photolysis that possess chlorophyll A and other pigments and has a capacity to capture sunlight and split water to make H_2. Even though, there are some barriers in this method such as: the O_2 evolved during this method suppresses the activation of hydrogenase subunits, gas mixture formed during the process is combustible, low photosynthetic conversion efficiencies and large surface area requirement

Manuscript received April 22, 2016; revised July 1, 2016.

The authors are with the Key Laboratory of Molecular Biophysics, the Ministry of Education, Huazhong University of Science and Technology, Wuhan 430074, PR China (e-mail: Dr_jawed@hust.edu.cn, yanyunjun@hust.edu.cn, 1018497090@qq.com, xuli@hust.edu.cn, yanyunjun@hust.edu.cn).

[7].

Design of reactor is one of the most important features of H_2 production from dark fermentation. In the meantime, increasing the partial pressures of H_2 during fermentation affects the process performance. Various strategies have been reported to reduce H_2 partial pressure, among which gas sparging and vacuum application are widely used. Although, it is essential to establish an effective method for H_2 production and gain electric energy through fuel cells.

III. Hydrogen Production from Other Renewable Energy Sources

The performance examination of H_2 production from other renewable sources have been investigated in many studies. Solar energy has been found to be the most inexhaustible; H_2 production from solar energy is considered to be the decisive solution for sustainable energy. Solar energy is also used for the technology of splitting water and can be divided into the four main categories: photochemical systems; semiconductor systems; photo biological systems; and hybrid and other systems [8].

IV. Limitations and Important Barriers to H_2 and Biohydrogen Production

The main limitations in the development of H_2 production processes include economic and technological barriers. Economic barriers incorporate the cost of H_2 production and distribution, expenditure of components and materials, and competition with the fossil fuels, while technological limitations include issues like H_2 storage, compressor and distribution networks, absence of strong fuel cell technologies, and incorporation with the existing infrastructure.

V. Various Approaches for Improving H_2 Production

A. Next Generation Sequencing: Application in Bio-hydrogen Production

The exact order of nucleotides exist in any DNA or RNA molecule is determined by the method of nucleic acid sequencing. The demand for cheaper and faster sequencing methods has augmented significantly after the completion of the first human genome sequence, which led to the development of next-generation sequencing (NGS). NGS platforms allows the sequencing of millions of fragments of DNA from a single sample, which facilitates an entire genome to be sequenced in less than one day [9].

The full benefit of NGS will not be achieved until extremely high performance computing and intensive bioinformatics support are able to interpret and utilize raw sequence data. A variety of software tools for NGS data analysis (Table I) are available online e.g. 1) alignment of reads to a reference sequence; 2) de novo assembly; 3) reference-based assembly; 4) base-calling and/or genetic variation detection (such as SNV, Indel); 5) genome annotation; and 6) utilities for data analysis. NGS appears to have almost boundless applications in the field of life sciences including H_2 and biodiesel production. The whole-genome shotgun project of C. perfringens strain JJC containing its assembly and annotation has been deposited at DDBJ/EMBL/GenBank under the accession no. AWRZ00000000 for further applications [10].

TABLE I: Tools for NGS Data Analysis

Tools	Description	Web links
FASTQC Toolkit	A quality control application for high throughput sequence data.	http://www.bioinformatics.babraham.ac.uk/projects/ fastqc/
NGS QC Toolkit	A toolkit for the quality control of NGS data.	http://www.nipgr.res.in/ngsqctoolkit.html
Velvet	De novo Genome Assembly.	http://www.ebi.ac.uk/zerbino/velvet/
ALLPATHS	De novo Genome Assembly.	http://www.broadinstitute.org/software/allpaths-lg/ blog/
SOAPdenovo	De novo Genome Assembly.	http://soap.genomics.org.cn/soapdenovo.html
SSAKE	De novo Genome assembler for short DNA sequence reads.	http://www.bcgsc.ca/platform/bioinfo/software/ssake
Maq	Mapping and assembly with qualities.	http://maq.sourceforge.net/
BWA	Burrows-wheeler alignment tool.	http://bio-bwa.sourceforge.net/bwa.shtml
SSAHA	Sequence Search and Alignment by Hashing Algorithm.	https://www.sanger.ac.uk/resources/software/ssaha/
Scripture	Method for transcriptome reconstruction that relies solely on RNA-Seq reads.	http://www.broadinstitute.org/software/scripture/
Cufflinks	Transcriptome assembly and differential expression analysis for RNA-Seq.	http://cole-trapnell-lab.github.io/cufflinks/
TopHat	Fast splice junction mapper for RNA-Seq reads.	http://ccb.jhu.edu/software/tophat/index.shtml
SpliceMap	De novo splice junction discovery and alignment tool.	http://web.stanford.edu/group/wonglab/SpliceMap/
SISSRs	A novel algorithm for precise identification of binding sites from short reads generated from ChIP-Seq experiments.	http://sissrs.rajajothi.com/
MACS	Model-based analysis for ChIP-Seq.	http://liulab.dfci.harvard.edu/MACS/
PeakSeq	A program for identifying and ranking peak regions in ChIP-Seq experiments.	http://info.gersteinlab.org/PeakSeq
CisGenome COV2HTML	An integrated tool for tiling array, ChIP-seq, genome and cis-regulatory element analysis. A visualization and analysis tool of bacterial next generation sequencing (NGS) data.	http://www.biostat.jhsph.edu/hji/cisgenome/ http://www.ncbi.nlm.nih.gov/pubmed/24512253
Artemis	Genome browser and annotation tool.	http://www.sanger.ac.uk/resources/software/artemis/
Ngsplot	Quick mining and visualization of next-generation sequencing data by integrating genomic databases.	https://code.google.com/p/ngsplot/

Galaxy	Galaxy is an open source, web-based platform for data intensive biomedical research.	https://usegalaxy.org/

DNA sequence analysis depends on the development of novel transcriptome analysis approaches, but still techniques are highly dependent on available genomic sequence. Currently, NGS based transcriptomics method has been applied to species without reference genome sequences as well. To reconstruct metabolic network terpenoid biosynthesis pathway in green alga *Botryococcus braunii* race B [11], to investigate the triacylglyceride accumulation mechanism of the unsequenced oleaginous microalgae of Neochloris oleoabundans and Chlorella vulgaris, and to investigate the transcriptomic profiling during induction of H2 photo-production in the organism *Chlamydomonas moewusii* via RNA-Seq [12]. The innovative transcriptomic analysis technique on *C. moewusii* was reported to be the first to be applied to a potential H_2 producing green alga. In another study, transcriptomic, metabolite and proteomic analyses were carried out in the H_2-producing bacterium *Clostridium butyricum* to investigate the changes at the molecular level that occur when the metabolism shifts to H_2 production [13].

B. Functional Genomics

The main objective of functional genomics is to study the gene/protein functions and interactions at the genome level. The enormous generation of genomic data from the NGS platforms in the recent years is responsible for the accumulation of genome information in databases (Fig. 2).

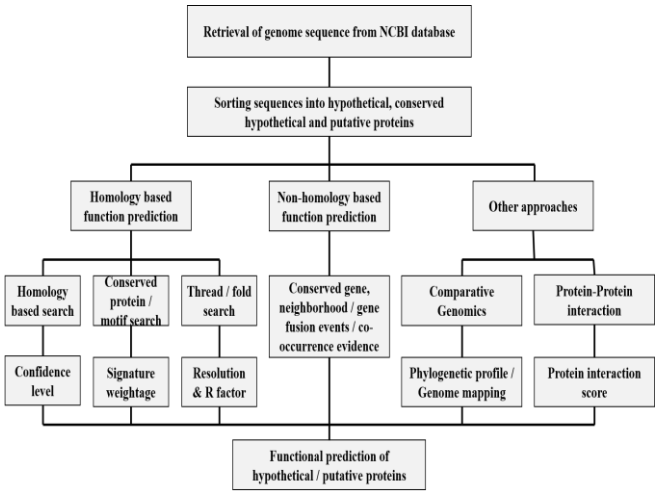


Fig. 2. Scheme characterizes integrated annotation / re-annotation strategy used for functional assignment of complete genome of organisms.

This creates a significant move towards gathering gene information through comparative genomics, proteomics, metabolomics and in particular, functional genomics [14].

"Dynamic biological data fusion method" was used to re-annotate the complete proteome of *E. coli* K-12, for enhanced H_2 production. About 29% of the protein sequences previously annotated as imaginary have now been assigned with clear/known functions. Furthermore, the updated functional information is publicly available as a database, "REC-DB" (http://recdb.bioinfo.aukbc.org.in/recdb/) [15].

C. Genome-Scale in Silico Metabolic Engineering

The growing oil price and environmental concerns have transformed our interest in utilizing biomass for the production of biofuel. Though, it is essential to develop high performance microbes that are capable of producing biofuels with very high efficiency in order to compete with the fossil fuel (Fig. 3). Microbes, capable of producing different biofuels including bioethanol, biobutanol, alkane, biodiesel, and H_2 have been successfully developed by using methods such as systems metabolic engineering [16].

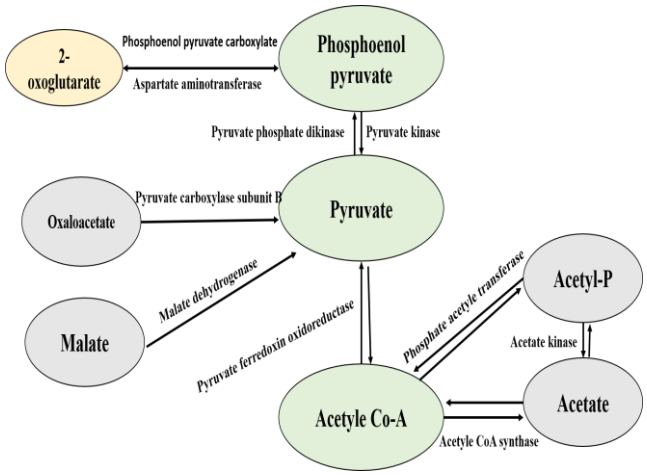


Fig. 3. Scheme demonstrates the reconstructed pyruvate metabolism pathway of *C. hydrogenoformans*. The conversion of Phosphoenol pyruvate to oxaloacetate and conversion of oxaloacetate to 2-oxoglutarate catalyzed by phosphoenolpyruvate carboxylase and aspartate aminotransferase respectively have been newly included in the pathway [12].

Likewise, metabolic flux analysis and in-silico metabolic models study for increased H_2 production are available for a number of H_2 producers. Several organism based pathway databases such as Cyclone, MediCyc and PathCase [17], are available that assimilates functional genomics and metabolic pathway information. Despite the presence of several databases with gene and metabolic pathway information, the pathway, metabolite and enzyme data of complete gene sets are unrevealed [18].

D. Synthetic Biology Approach

Synthetic biology is the engineering of biological molecules in the deliberate redesign and construction of innovative biological systems and organisms that does not exist in nature to perform new functions for useful purposes. Table II shows list of tools used for functional prediction and Table III shows tools used for pathway reconstruction and flux balance analysis. Another potential route for biological H_2 production is the conversion of biomass into formate, which can subsequently be processed into H_2 by *E. coli*., formate is also a widely used commodity chemical [19].

TABLE II: LIST OF TOOLS FOR FUNCTION PREDICTION

Function prediction tools	Prediction method	Web links
BLAST	Similarity based search.	http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
CDD	Conserved proteins domain classification.	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

InterProScan	Protein domain search.	http://www.ebi.ac.uk/Tools/InterProScan/
String	Known and predicted protein interaction.	http://string.embl.de/
Pfam	Conserved region search.	http://pfam.xfam.org/
ProtFun	Ab initio function prediction.	http://www.cbs.dtu.dk/services/ProtFun/
PFP tool	Automated protein function prediction.	http://kiharalab.org/web/pfp.php
ScanProsite	Pattern and profile search.	http://prosite.expasy.org/scanprosite/
PPSearch	Protein motifs search.	http://www.ebi.ac.uk/Tools/ppsearch/
CombFunc	GO based function prediction server.	http://www.sbg.bio.ic.ac.uk/mwass/combfunc/
I-TASSER	Protein structure and function predictions.	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
PSORTb 3.0	Subcellular localization (SCL) based predictor.	http://www.psort.org/psortb/index.html
AutoFACT	Automated protein function prediction.	http://megasun.bch.umontreal.ca/Software/AutoFACT.htm
ConFunc	Gene Ontology based protein function prediction.	http://www.sbg.bio.ic.ac.uk/confunc
ProFunc	Prediction of protein function from 3D structure.	http://www.ebi.ac.uk/thornton-srv/databases/ profunc/
SVMProt	Web-based support vector machine software for functional classification of a protein from its primary sequence.	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi

TABLE III: LIST OF TOOLS FOR PATHWAY RECONSTRUCTION AND FLUX BALANCE ANALYSIS

Tools	Description	Web links
MetRxn	To integrate genome-scale metabolic network reconstructions.	http://www.metrxn.che.psu.edu/
GraphViz	Interaction network visualization software tools.	http://www.graphviz.org/
Systrip	Interaction network visualization software tools..	http://tulip.labri.fr/TulipDrupal/?q¼ systrip
Cytoscape	Interaction network visualization software tools	http://www.cytoscape.org/
CyTargetLinker	To integrate different regulatory interactions into their network analysis approaches.	http://projects.bigcat.unimaas.nl/cytargetlinker
GLAMM	Reaction network visualization tools that incorporate experimental data.	http://glamm.lbl.gov/
Vanted	Reaction network visualization tools that incorporate experimental data.	http://vanted.ipk-gatersleben.de/
NUPACK, Vienna	RNA structure design and prediction.	http://www.nupack.org/
RNA Web suite PathFinder	Tool for the dynamic visualization of metabolic pathways based on annotation data.	http://bibiserv.techfak.uni-bielefeld.de/pathfinder/
CARMEN	Performs in silico reconstruction of metabolic networks to interpret genome data in a functional context.	http://carmen.cebitec.uni-bielefeld.de/cgi-bin/index.
MetNetMaker	Used for the reconstruction of the metabolic networks.	http://www.tomforth.co.uk/metnetmaker/
CellNetAnalyzer	Constraint-based flux analysis tool used in network visualization.	http://www2.mpimagdeburg.mpg.de/projects/cna/ cna.html
Omix	A network visualization tool that is particularly good for visualizing fluxes.	http://www.omix-visualization.com/#sthash. hQJZSx7S.dpbs
OpenFlux	Integrates many powerful C flux analysis tools while using a GUI (Graphical User Interface).	http://openflux.sourceforge.net/
FiatFlux	Performs local flux ratio analysis to determine relative fluxes at branch points.	http://www.imsb.ethz.ch/research/zamboni/resources/fiatflux.html
COBRA	Flux-balance analysis and flux distribution calculation.	http://opencobra.github.io/
Systems Biology Research Tool	Flux balance analysis.	http://www.ieu.uzh.ch/wagner/software/SBRT/

VI. CONCLUSION

In the recent years, substantial research on H_2 producing organisms, have given visions into their metabolism and physiological properties. Before using an organism for H_2 production process on industrial level, there is a need to overcome many challenges. The present review demonstrates that the mainstream of the present challenges can be minimize by a combination of different methods using bioinformatics and biotechnology approaches starting from reactor designs to synthetic biology. Nevertheless, more research devotion is needed to implement these approaches as a vital tactic for the development of microorganisms for large scale applications on industrial level. In silico metabolic engineering at genome-scale and use of synthetic biology are especially likely to be at the front of these developments. Hence, it is possible to achieve approximately 10 mol H_2 per mol of glucose for industrialization by using the versatile bioinformatics approaches trailed by wet lab validation in the near future.

VII. CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

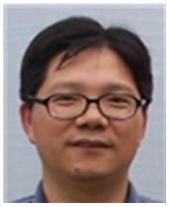
[1] G. Lazzerini, S. Lucchetti, and F. P. Nicese, "Green House Gases (GHG) emissions from the ornamental plant nursery industry: A life cycle assessment (LCA) approach in a nursery district in central Italy," *Journal of Cleaner Production*, 2016, vol. 112, Part 5, pp. 4022-4030.

[2] S. Y. Zhou *et al.*, "In-silico design of a new energetic material—1-Amino-5-nitrotetrazole with high energy and density," *Computational Materials Science*, 2016, vol. 112, pp. 67-74.

[3] J. Mathews and G. Wang, "Metabolic pathway engineering for enhanced biohydrogen production," *International Journal of Hydrogen Energy*, 2009, vol. 34, no. 17, pp. 7404-7416.

[4] M. Jawed *et al*., "Enhanced H production and redirected metabolic flux via overexpression of fhlA and pncB in Klebsiella HQ-3 strain," *Appl Biochem Biotechnol*, 2015.

[5] D. J. Morton *et al*., "The heme-binding protein (HbpA) of Haemophilus influenzae as a virulence determinant," *International Journal of Medical Microbiology*, 2009, vol. 299, no. 7, pp. 479-488.
[6] D. Jullesson *et al*., "Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals," *Biotechnology Advances*, 2015, vol. 33, no. 7, pp. 1395-1402.
[7] J. Wang *et al*., "Effects of increasing the NAD(H) pool on hydrogen production and metabolic flux distribution in Enterobacter aerogenes mutants," *International Journal of Hydrogen Energy*, 2013, vol. 38, no. 30, pp. 13204-13215.
[8] H. Ahmad *et al*., "Hydrogen from photo-catalytic water splitting process: A review," *Renewable and Sustainable Energy Reviews*, 2015, vol. 43, pp. 599-610.
[9] M. Datto and R. L. Lundblad, "DNA, RNA chemical properties (including sequencing and next-generation sequencing)," *Encyclopedia of Cell Biology*, 2016, Academic Press: Waltham, pp. 24-35.
[10] K. A. Hassan *et al*., "Genomic analyses of clostridium perfringens isolates from five toxinotypes," *Research in Microbiology*, 2015, vol. 166, no. 4, pp. 255-263.
[11] Z. Xu *et al*., "Nitrogen deprivation-induced de novo transcriptomic profiling of the oleaginous green alga Botryococcus braunii 779," *Genomics Data*, 2015, vol. 6, pp. 231-233.
[12] G. R. Kumar and N. Chowdhary, "Biotechnological and bioinformatics approaches for augmentation of biohydrogen production: A review," *Renewable and Sustainable Energy Reviews*, 2016, vol. 56, pp. 1194-1206.
[13] C. Li *et al*., "Complete genome sequence of Clostridium butyricum JKY6D1 isolated from the pit mud of a Chinese flavor liquor-making factory," *Journal of Biotechnology*, 2016, vol. 220, pp. 23-24.
[14] M. Fernández-Escobar *et al*., "Use of functional genomics to understand replication deficient poxvirus-host interactions," *Virus Research*, 2016, vol. 216, pp. 1-15.
[15] K. V. Solomon *et al*., "Extracting data from the muck: deriving biological insight from complex microbial communities and non-model organisms with next generation sequencing," *Current Opinion in Biotechnology*, 2014, vol. 28, pp. 103-110.
[16] J. M. Park, T. Y. Kim, and S. Y. Lee, "Constraints-based genome-scale metabolic simulation for systems metabolic engineering," *Biotechnology Advances*, 2009, vol. 27, no. 6, pp. 979-988.
[17] M. A. Nik *et al*., "A comparative study of metamodeling methods for the design optimization of variable stiffness composites," *Composite Structures*, 2014, vol. 107, pp. 494-501.
[18] E. Kırtay, "Recent advances in production of hydrogen from biomass," *Energy Conversion and Management*, 2011, vol. 52, no. 4, pp. 1778-1789.
[19] A. V. Puga, "Photocatalytic production of hydrogen from biomass-derived feedstocks," *Coordination Chemistry Reviews*, 2016, vol. 315, pp. 1-66.



Yan Yunjun was born at Luotian county, Hubei Province, China, in September 1969. He got the B.S. in biology, Central China Normal University, Wuhan city. He got the M.S. in zoology, Central China Normal University, Wuhan City in 1993. He got the Ph.D. in hydrobiology, Institute of Hydrobiology, the Chinese Academy of Sciences, Wuhan City, China in 1995. He was professor of biotechnology and from 1998 to2002. He is the vice dean of College of Life Sciences and Technology, College of Life Sciences and Technology, Huazhong University of Science and Technology; 1037, Luoyu Road, Hongshan District, Wuhan 430074, China from 2012 still now. Prof. Yan is senior committee member of the Chinese Society of Biological Engineering and Chinese Society of Microbiology. He holds a Second-grade Award from China Medical Association in 2009, and from Journal "Chinese Journal of Biotechnology" in 2009 and 2011.

Muhammad Jawed was born in August 1977, obtained the master's degree in biotechnology from Karachi University, Pakistan. He worked as a medical technologist in the Aga Khan University and hospital, which is the premier institute of medical research in Pakistan (2005 - 2007). He was appointed as a research officer in Liaquat University of Medical and Health Sciences, Pakistan, and worked intensively on real time PCR and DNA sequence analyzer for different projects related to viral hepatitis and genetic disorders in Humans (2008 - 2013). Currently he is doing research as a Ph.D. candidate in the field of biochemistry and molecular biology, in Huazhong University of Science and Technology, Wuhan, P.R. China (Sept. 2013 - July 2016).