Vol.53, n. 2: pp.319-325, March-April 2010
ISSN 1516-8913 Printed in Brazil

BRAZILIAN ARCHIVES OF
BIOLOGY AND TECHNOLOGY
AN INTERNATIONAL JOURNAL

Recent Progress in the Methods of Genome Sequencing

Ning-wei Zhao*
School of Biotechnology; Albanova University Center; Royal Institute of Technology (KTH); Roslagstullsbacken 21 SE-106 91 Stockholm - Sweden

ABSTRACT

Genome sequencing is a very important tool for the development of genetic diagnosis, drugs of gene engineering, pharmacogenetics, etc. As the HGP comes into people's ears, there is an emerging need for the genome sequencing. During the recent years, there are two different traditional strategies available for this target: shotgun sequencing and hierarchical sequencing. Besides these, many efforts are pursuing new ideas to facilitate fast and cost-effective genome sequencing, including 454 GS system, polony sequencing, single molecular array, nanopore sequencing, with each having different unique characteristics, but remains to be fully developed.

Key words: Genome sequencing, HGP, shotgun, hierarchical, polony, nanopore

INTRODUCTION

About 12 years ago, the first sequences of complete genomes were published; these were of two bacteria, *Haemophilus influenzae* and *Mycoplasma genitalium*. Although they are fairly small (1.83Mb and 0.58Mb, respectively) in comparison with other bacteria, let alone mammalian or plant genomes, they are much larger than the examples being considered so far. These have been followed by sequences of the genomes of a number of other bacteria, as well as some eukaryotes, including yeast (*Saccharomyces cerevisiae*), and of course, the human genome sequences. The international HapMap Project (Bennett et al, 2005) is a $100 million worldwide collaboration of research groups seeking to define the regions of the human genome that have been inherited together as a block with little genetic shuffling. Ventures such as this will go long way to facilitate the genetic exploration, but, in the absence of new technological progress, the genetic dissection of clinical phenotypes will remain restricted, largely by the overall cost and effort required. Therefore, completing the sequencing of the human genome has been a fantastic achievement that marked the beginning of the genetics chapter in society (Dale et al, 2002). Genome sequencing has markedly changed the nature of biomedical research and medicine. Reductions in the cost, complexity and time required for sequencing the large amounts of DNA, including improvements in the ability to sequence the bacterial and eukaryotic genomes will have significant scientific, economic and cultural impacts. Hitherto, various strategies, or combinations of strategies, can be used for determining the sequences of DNA as long as a complete genome, even a complete chromosome.

The techniques for genome sequencing can be divided into two groups. The first are the traditional techniques, known as hierarchical sequencing and shotgun sequencing, based on fragmentation and reassembly. The second are the techniques invented during recent years, such as Solexa Single Molecule Array (SMA) technology,

* Author for correspondence: znw_sun@sina.com

454 genome sequencer system, accurate multiplex polony sequencing from Harvard Medical School, a Sanger/Pyrosequencing hybrid approach from Craig Venter Institute, Fast DNA sequencing via transverse electronic transport and the latest method called diploid genome reconstruction containing a key step called Gibbs sampling.

Traditional Strategies

In this type of strategy, whole chromosome sequences are reassembled from the sequences of hundreds of thousands of fragments, each typically between 500 and 1000bp in length (Gibson et al, 2004). There are two traditional methods for fragmentation and reassembly, known as hierarchical sequencing and shotgun sequencing (Lander et al, 2001). The first one is preferred by the HGP. In this approach, the genomic DNA is sheared into pieces of about 150 kb and inserted into BAC vectors, then transformed into *E. coli* where they are replicated and stored. The BAC inserts are isolated and mapped to determine the sequence of each cloned 150 kb fragment. This is referred to as the Golden Tiling Path. Each BAC fragment in the Golden Path is fragmented randomly into smaller pieces and each of which is cloned into a plasmid and sequenced on both strands. These sequences are aligned so that identical sequences overlap. These contiguous pieces are then assembled into finished sequence once each strand is sequenced about four times to produce 8× coverage of data with high quality. The advantage to the hierarchical approach is that sequencers are less likely to make mistakes when assembling the shotgun fragments into contigs as long as full chromosomes, because the chromosomal location for each BAC is known, and there are fewer random pieces to assemble. The disadvantage to this method is time and expense (Green, 1997).

The second method was developed and preferred by Celera (Venter et al, 1998). This approach has been developed and preferred for prokaryotic genomes, which are smaller in size and contain less repetitive DNA. Shotgun sequencing randomly shears genomic DNA into small pieces, which are cloned into plasmids and sequenced on both strands, thus eliminating the BAC step from the HGP's approach. Once the sequences are obtained, they are aligned and assembled into finished sequence (Venter et al, 2001). The shotgun method is faster and less expensive, but it is more prone to errors due to incorrect assembly of finished sequence. For example, if a 1000 kb portion of a chromosome is duplicated and each one is cut into 2kb fragments, then it would be difficult to determine where a particular 2 kb piece should be located in the finished sequence since it occurs twice (Weber et al, 1997). Between these two methods, which one is better depends on the size and complexity of the genome.

NEW METHODS

Single molecule arrays (SMA)

Solexa has developed this method to sequence clonally amplified single DNA molecules (Brakmann et al, 2001; Braslavsky et al, 2003).The procedure is as follows. In a single tube reaction, gDNA is fragmented and processed into ss-Oligonucleotide fragments. Hundreds of millions of molecules are deposited and attached to discrete sites on a SMA. Complementary nucleotides base-pair to the first base of each oligonucleotide fragment and are added to the primer by the enzyme. Laser light excites the label on the incorporated nucleotides, which fluoresce. This fluorescence is detected by a CCD that rapidly scans the entire array to identify the incorporated nucleotides on each fragment. This cycle of incorporation, detection and identification is repeated nearly 25 times to determine the first 25 bases in each fragment. These hundreds of millions of sequences are aligned and compared with the reference sequence (Smith, 2004). This approach generates raw data of 100M reads/run, totally 2.5Gbps/run, which is required for genome resequencing of human genomes. The advantages to this method is that it is possible to create arrays of very high site density, around 10^8 sites per cm^2 or more, allowing massively parallel processing, because of only a single molecule at each site (www.illumina.com). Based on these two features, SMA creates a breakthrough in economics and throughput. However, it also has two limits: (1) The short DNA readouts are hard to reassemble; (2) PCR may introduce some copy errors and may be a barrier in decreasing the cost.

454 Genome sequencer

The 454 GS is another alternative for whole-genome sequencing developed by the 454 Life Sciences. Its focus is on the co-development of an emulsion-based method (Margulies et al, 2005) to isolate and amplify the DNA fragments *in vitro*, and of a fabricated substrate and instrument that performs Pyrosequencing (Dressman et al, 2003) in picolitre-sized wells (Thomas et al, 2006). The genome is cut randomly into small fragments, denatured and adaptors are added to each of them in order to link to the beads. One bead has its specific DNA fragment and then administrated the emulsion-based PCR. PCR terminates at different sites to make every bead own different copies of the same template, in which PCR primers are biotinylated. The emulsion is washed, the DNA is denatured again, and then the beads are put with ssDNAs on fiber-optic slide. Finally, sequencing is done by the synthesis simultaneously in open wells of fiber-optic slide using a modified Pyrosequencing protocol that is designed to take advantage of the small scale of the wells, while the read length of this method reaches 100, which is somewhat larger than traditional Pyrosequencing. The obvious advantage to this method is fast, ultra-high throughput (up to 25million bps per 4 hours) (Kaeller et al, 2007). However, it still has some limitations, such as its accuracy is about 99%, which is not satisfactory. It still can't resolve the problem of paired-end sequencing information and faces a bug which is also faced by shotgun sequencing that is "short read length" which is not easily aligned without cloning.

Accurate multiplex polony sequencing

It's a new DNA sequencing technology in which a commonly available, inexpensive epi-fluorescence microscope is converted to rapid non-electrophoretic DNA sequencing automation (Hutchison III, 2007). This technology has been applied to re-sequence an evolved strain of *E. coli* at less than one error per million consensus bases (Dahl et al, 2007). A cell-free, mate-paired library provided the single DNA molecules that were amplified in parallel to 1µm beads by emulsion polymerase chain reaction. Millions of beads were immobilized in a polyacrylamide gel and subjected to automated cycles of sequencing by ligation and four-color imaging. Cost per base was roughly one-ninth as much as that of conventional sequencing. This technology developed by Church's team currently can read a slide with 10 million polonies in about 20 minutes, making it become one of the swiftest DNA sequencing methods available now, while its read length is only “13+13” (Hall, 2007). On the other hand, with the help of this method, one can collect 786 Gbits of image data from which only 60 Mbits of sequence can be gleaned. The sparsity is a ripe avenue for improvement. The natural limit of this direction is single-pixel sequencing, in which the commonplace analogy between bytes and bases will be at its most manifest, and it is only for small-genome-scale genome analysis for the present (Shendure et al, 2005).

A Sanger/Pyrosequencing hybrid approach

This method is offered by Craig Venter Institute and University of New South Wales. This study showed that the 454 GS system was able to sequence regions of a genome for which Sanger sequencing was ineffective (Goldberg et al, 2006). For organisms with a small genome size (<3 Mb) and/or a small number of gaps and/or high levels of repetitive structure, including physical ends, 8×Sanger sequencing could be the most cost-effective approach. For organisms with a large genome size, many sequencing gaps, and/or hard stops, researchers found initial sequencing of 5×Sanger data followed by the addition of two 454 runs to be the most cost-effective approach (Fig 1). The combination of 454 sequencing data at any ratio and Sanger sequencing data is helpful for the final draft genome in terms of coverage, reduction of gaps, and poorly sequenced regions that degrade the value of an assembly, and leading to an 25% reduction in cost compared to the traditional Sanger-only approach (Ahmadian, 2002). However, although the number of sequencing gaps was reduced, the overall reduction was not sufficient to permit additional high-throughput sequencing. Therefore, this technology needs more improvement.

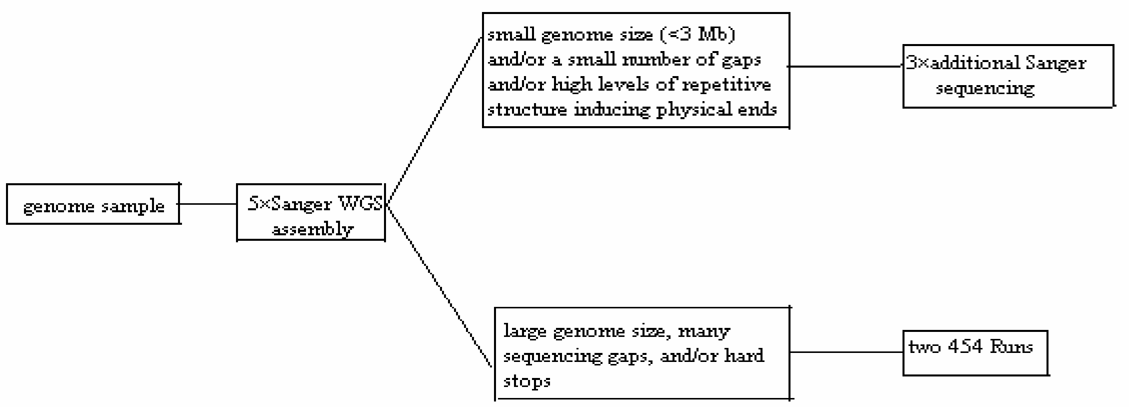
genome sample
5×Sanger WGS assembly
small genome size (<3 Mb) and/or a small number of gaps and/or high levels of repetitive structure inducing physical ends
3×additional Sanger sequencing
large genome size, many sequencing gaps, and/or hard stops
two 454 Runs

Figure 1 - Decision tree for hybrid sequencing strategy.

Nanopore sequencing

This method is based on the distributions of transverse electrical currents of single-stranded DNA while it translocates through a nanopore (Hao et al, 2006; Guiducci et al, 2004). The inventors, the physicists from UCSD, used mathematical calculations and computer modeling of the motions and electrical fluctuations of DNA molecules to determine how to distinguish each of four different bases (A, G, C, T). They built their calculations on a pore about a nanometer in diameter made from silicon nitride, a material that is easy to work with and commonly used in nanostructures surrounded by two pairs of tiny gold electrodes. The electrodes can record the electrical current perpendicular to the DNA strand as the DNA passed through the pore (Zwolak et al, 2005). Each DNA base is structurally and chemically different, thus creating its own distinct electronic signature. Previous attempts to sequence the DNA using nanopores were not successful because the twisting and turning of the DNA strand introduced too much noise into the signal being recorded. This new method takes advantage of the electric field that drives the current perpendicular to the DNA strand to reduce the structural fluctuations of DNA while it moves through the pore, thus minimizing the noise (Ashkenasy et al, 2005). Although this method is cheap and fast, the inventors caution that there are still hurdles to overcome because no one has yet made a nanopore with the required configuration of electrodes (Lagerqvist et al, 2006).

Diploid genome reconstruction

The researchers from Yonsei University and University of Southern California have recently invented a new method to sequence all the chromosomes of one creature. The first researcher, Jong Hyun Kim successfully inferred the haplotype of one kind of marine invertebrate, called *Ciona intestinalis* from published sequence data. This method is based on the high genic variability in creatures (Fig 2). Other creatures with high genic variability are also suitable for this method, such as fish. The core step of this method is a program, called Gibbs sampling, which is often used for probabilistic inference, mainly for the copies of incomplete information. Relying on this method, researchers estimated that the polymorphism rate was 1.2%, or 1.5% (derived from two different algorithms). According to this reconstruction analysis, the estimate of haplotype has been achieved whose accuracy is 97%, so as to construct this method of comparative analytics, which is used to study on conservative DNA regions of vertebrates. However, this method is not suitable for human, because the variability of homo-genome is rather low comparatively. Kim et al (2007) found that this method could be used for sequencing those regions with high variability in homo-genome. On the other hand, this method confirmed that some short fragments of non-coding DNA were rather conservative, which offered an important and indirect evidence that junk DNA could have some unknown functions.

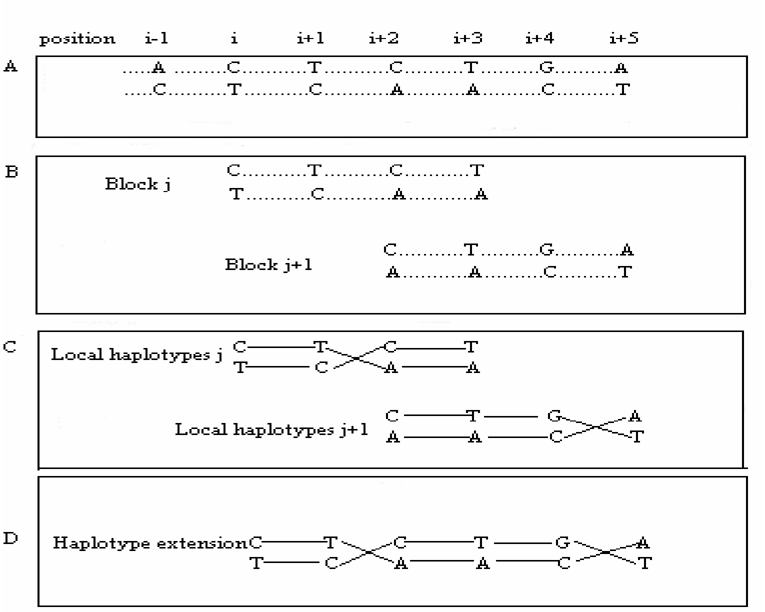
position i-1 i i+1 i+2 i+3 i+4 i+5
A
.....A .........C...........T...........C..........T..........G..........A
.....C..........T...........C...........A..........A..........C..........T
B
Block j
C...........T...........C...........T
T...........C...........A...........A
Block j+1
C...........T..........G...........A
A...........A..........C..........T
C
Local haplotypes j
C T C T
T C A A
Local haplotypes j+1
C T G A
A A C T
D
Haplotype extension
C T C T G A
T C A A C T

Figure 2 - Overview of haplotype reconstruction. (A) Potential polymorphic sites are determined after the potential polymorphism detection step. Dotted lines between strongly potential polymorphisms indicate that their phases are not determined yet. (B) Adjacent blocks, block j and block j+1 share two strongly potential polymorphisms at the (i + 2) and (i + 3) positions. (C) After the Gibbs sampler is applied to each block, the phases are determined. Each solid line indicates a direction of connection. (D) If adjacent short haplotypes overlap and show a consistency, those haplotypes are combined, and then used for haplotype bridging.

DISCUSSION AND CONCLUSION

The genome sequencing has great significance for mankind, which include:

(1) In gene-related disease: Disease-related genes play important roles in genomic integrality of conformation and function. For monogenic diseases, several monogenes corresponding to genetic diseases have been discovered (based on both "positional cloning" and "positional candidate cloning"), which helps in genetic diagnosis and therapy.

(2) In medicine: The treatment based on genomic knowledge such as genetic diagnosis; prevention based on genomic information; the realization of genes susceptible to ailment; life styles of risky persons; intervention of environmental factors.

(3) In biotechnology: the drugs of gene engineering; diagnosis and research chemicals and kits; facilitation of cell, embryo and tissue engineering.

(4) In pharmacy: to select the targets of drugs; initiate the individual treatment: pharmaceutical genomics.

(5) In social economics: transgenic food; transgenic drugs such as diet pills.

(6) In studies on biological evolution: phylogeny.

Completing the draft sequence of the human genome in 2003 has been a fantastic achievement that marked the beginning of the genetics chapter in society (Lander et al, 2001; Venter et al, 2001). On May 31, 2008, Watson, "the father of DNA", got a special present, which was a CD printed with his individual whole genomic map. This showed that there could be an emergence of individual genome sequencing in future. The genetic information behind the individual genomic maps seems to be "life codes". When owning it, people can adopt related countermeasures to reduce the risk of having special disease before children's birth. For example, if a genomic map of a child shows that he has high risk of diabetes, his parents can rigidly control his weight after his birth, in order to reduce his risk of having diabetes before learning how to walk. However, the cost of Watson’s individual genome sequencing is about $1 million, which is not affordable for most people. Hence, present genome sequencing technologies pay most attention on how to reduce time and cost while keeping high throughput and high accuracy. As the sequencing techniques referred above, it is difficult to say which one is better than another one, with each of them having its unique advantages and limits at the same time. More importantly, among those new technologies, there is enough space for improvement. For example, if the 454 GS successfully solves the problem of paired-end sequencing and the "short read length", it could have the biggest potential. The Sanger/Pyrosequencing hybrid approach has shown that the combination of two or more techniques could be a better choice, such as 454 GS and Solexa SMA, coupling multiplex sequencing with paired-end ditags (Ng et al, 2006). On the other hand, for a long time, how to sequence the genome has been only built on fragments sequencing, which is as follows: shear the genomic DNA into small fragments; sequence them separately by different kinds of techniques; collect the data; align them by software finally. Since the introduction of the method of diploid genome reconstruction, one can select sequencing technology in terms of the characteristics of target sequence, which means that when encountering DNA regions with high variability, one can adopt this method. After all, most ailment-related genes are usually located in those regions with high variability. Anyway, human genome sequencing is only one part of genome sequencing. When coping with genome of other creatures, one can adopt different sequencing methods based on the characteristics of both your preferred method and your target creature.

It is expected that the technology capable of achieving the $1000 human genome project would be available by 2016 (Zimmerman 2004). It is predicted that these breakthrough technologies, such as SMA, could exponentially accelerate the rate at which genetic discoveries could be made, validated and transferred from the laboratory into the ordinary life. Therefore, given that recent progress, we can realize this goal in the near future.

ACKNOWLEDGEMENTS

Author thanks the senior researcher of KTH genome center, Dr. Afshin Ahmadian for giving many suggestions and also to the colleagues in Biotechnology at KTH for discussions.

RESUMO

Sequenciação do genoma foi um instrumento muito importante para o desenvolvimento do diagnóstico genético, a droga da engenharia genética, farmacogenética, etc. Como o HGP entrar em ouvidos do povo, há uma necessidade emergente para a sequenciação do genoma. Durante os últimos anos, há duas diferentes estratégias tradicionais disponíveis para este objectivo: seqüenciamento shotgun hierárquico e sequenciação. Além desses, muitos esforços estão a prosseguir novas idéias para facilitar a rápida e eficaz em termos de custos sequenciação do genoma, incluindo 454 GS sistema, polony seqüenciamento, único molecular array, nanopore seqüenciamento, com cada um dos quais com diferentes características únicas, e que resta para ser mais desenvolvido.

REFERENCES

Ahmadian, A.; Lundeberg, J. (2002), A brief history of genetic variation analysis. *Biotechniques*, **32**, 1122-1137.

Ashkenasy N, et al. (2005), Recognizing a single base in an individual DNA strand: a step toward DNA sequencing in nanopores. *Angew. Chem. Int. Ed. Engl.*, **44**(9), 1401-1404.

Bennett, S. T.; et al. (2005), Toward the $1000 human genome. *Pharmacogenomics*, **6**(4), 373-382.

Brakmann, S. et al. (2001), High-density labeling of DNA: preparation and characterization of the target material for single-molecule sequencing. *Angew. Chem. Int. Ed. Engl.*, **40**(8), 1427-1429.

Braslavsky, I. et al. (2003), Sequence information can be obtained from single DNA molecules. *PNAS*, **100**, 3960-3964.

Dahl, F. et al. (2007), Multi-gene amplification and massively parallel sequencing for cancer mutation discovery. *PNAS*, **104**, 9387-9392.

Dale, J. W.; Schantz, M. V. (2002), *From genes to genomes.* John wileysons. Ltd, England, pp. 169-172.

Dressman, D. et al. (2003), Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *PNAS*, **100**(15), 8817-8822.

Gibson, G.; Spencer, V. (2004), *A primer of gene science*. Sinauer associates, Incorporated, USA, pp. 81-93.

Goldberg, S. M. D. et al. (2006), A Sanger/Pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *PNAS*, **103**, 11240-11245.

Green, P. (1997), Against a whole-genome shotgun. *Genome Res.*, **7**, 410-417.

Guiducci, C. et al. (2004), DNA detection by integrable electronics. *Biosens. Bioelectron.*, **19**(8), 781-787.

Hall, N. (2007), Advanced sequencing technologies and their wider impact in microbiology. *J Exp. Biol.*, **210**, 1518-1525.

Hao, X. Y. Bingqian. (2006), *Towards rapid DNA sequencing: detecting single-stranded DNA with a solid-state nanopore*. Small (Weinheim an der Bergstrasse, Germany), **2**(3), 310-312.

Hutchison III C. A., (2007), DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.*, [Epub ahead of print].

Kaeller, M. et al. (2007), Arrayed identification of DNA signatures. *Expert Rev. Mol. Diagr.*, **7**(1), 65-76.

Kim, J. H. et al. (2007), Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res.*, **11**, 1101-1110.

Lagerqvist, J., et al. (2006), Fast DNA sequencing via transverse electronic transport. *Nano Lett.*, **6**(4), 779-782.

Lander, E. et al; International human genome sequencing consortium. (2001), Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Margulies, M. et al. (2005), Genome sequencing in micro fabricated high-density picolitre reactors. *Nature*, **437**(7057), 376-380.

Ng, P. et al. (2006), Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res.*, **34**(12), e84.

Shendure, J. et al. (2005), Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728-1732.

Smith, T. (2004), Whole genome variation analysis using single molecule sequencing. Drug *Discovery Today: Targets*, **3**(3), 112-116.

Thomas, R. K. et al. (2006), Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picolitre reactor sequencing. *Nat. Med.*, **12**(7), 852-855.

Venter, J. C.; Adams, M. et al. (1998), Shotgun sequencing of the human genome. *Science*, **280**, 1540-1542.

Venter, J. C.; et al. (2001), The sequence of the human genome. *Science*, **291**, 1304-1351.

Weber, J. Myers, H. (1997), Human whole genome shotgun sequencing. *Genome Res.*, **7**, 401-409.

Zimmerman, Z. (2004), *The $1000 human genome-implications for life sciences*, Healthcare and IT. IDC Publishers, USA, pp. 1-10.

Zwolak, M. Ventra, M. D. (2005), Electronic signature of DNA nucleotides via transverse transport. *Nano Lett.*, **5**(3), 421-424.

Received: January 09, 2008;
Revised: May 21, 2008;
Accepted: May 20, 2009.