Published in final edited form as:

Chem Commun (Camb). 2011 July 14; 47(26): 7281–7286. doi:10.1039/c1cc11078k.

Sequencing nucleic acids: from chemistry to medicine

Shankar Balasubramanian^{*,a,b,c}

^aDepartment of Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK CB2 1EW

^bCancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, UK CB2 0RE

^cSchool of Clinical Medicine, University of Cambridge, Cambridge, UK CB2 0SP

Abstract

Chemistry has played a vital role in making routine, affordable sequencing of human genomes a reality. This article focuses on the genesis and development of Solexa sequencing that originated in Cambridge, UK. This sequencing approach is helping transform science and offers intriguing prospects for the future of medicine.

Chemistry played a vital role in elucidating the chemical structure of the natural nucleic acids DNA and RNA. This grounding was absolutely essential to enable the discovery of the three-dimensional structure of DNA and the Watson–Crick rules for the molecular recognition of base pairs.¹ In a paper published in 1952, Todd and Brown remarked “There can be no question of finality about any nucleic acid structure at the present time, since it is clear that there is no available method for determining the nucleotide sequence”.² This statement signified the beginnings of a quest for determining the exact sequence of nucleic acids.

A very early attempt at nucleic acid sequencing by Todd and co-workers explored the use of chemistry for the sequential degradation of RNA from its 3' end, by cycles of oxidation followed by β -elimination.³ However, this and other attempts did not lead to a practical sequencing approach for decades to come. It wasn't until the 1970s that two groups developed distinct sequencing chemistries that held considerable promise. At Harvard, Cambridge, USA, Maxam and Gilbert developed an approach that exploited the selective chemical reactivity of reagents towards the nucleobases to induce sequence-specific cleavage patterns from which the sequence could be derived after separating the cleaved DNA fragments, based on their size, by high resolution gel electrophoresis (Fig. 1).⁴ In Cambridge, UK, Sanger and his co-workers developed a very different approach that involved the synthesis of DNA in the presence of nucleotides that terminate synthesis. Specifically, Sanger's sequencing used a DNA polymerase to extend a DNA primer on a DNA template using all four activated 2'-deoxynucleoside triphosphate (dNTP) building blocks (i.e. dATP, dCTP, dGTP and dTTP). Four parallel sequencing reactions would be carried out in which one of each of the four dideoxynucleoside triphosphates (ddATP, ddCTP, ddGTP and ddTTP) would be present. The C-terminator ddCTP would cause partial termination in some of the copies of DNA at each insertion of nucleotide C leading to the generation of a ladder of DNA fragments each of which is terminated at C and similarly for the remaining three termination reactions. The A, C, G and T DNA ladders could then each

be separated by polyacrylamide gel electrophoresis and the sequence obtained by reading off the DNA bands from the longest to the shortest.⁵

Sanger sequencing proved to be the practical method of choice and subsequently opened up the field of genetics and molecular biology. The basic technology invented by Sanger was subjected to an impressive level of continuous improvement that included dye-labelled termination chemistry and advances in the engineering of polymerases and automation during the two decades after its creation. The best-known and arguably most important application of Sanger sequencing was the completion of the first human genome in 2004 by The Human Genome Project.⁶ A copy of the human genome comprises just over 3 billion bases of DNA and was decoded over a period of several years using hundreds of capillary sequencers in a collaborative effort involving laboratories worldwide, at an overall cost of several hundred million US dollars. This important milestone provided the opportunity for humanity to begin to understand our genome on an unprecedented level. It has also stimulated the consideration of alternative approaches to nucleic acid sequencing that would go beyond what has been possible using Sanger's approach.

In the remainder of this article, I will focus on discussing the genesis and development of a massively high throughput nucleic acid sequencing approach called Solexa sequencing that originated in Cambridge, UK and is currently being widely used for the routine sequencing of human genomes. This is not intended as a review; other 'next generation' sequencing approaches have been described elsewhere.⁷ The founding ideas and early proof of concept experiments that underpin Solexa sequencing occurred at Lensfield Road in laboratories led by myself and by my colleague David Klenerman. We then launched a small start-up company called Solexa in 1998 to fully develop and commercialise the approach. Solexa sequencing was reduced to practice into a commercial sequencing system in 2006 and was acquired by Illumina in 2007.

During the mid 1990s my co-workers and I were using chemical and biophysical methods to study the synthesis of DNA by a DNA polymerase enzyme. DNA polymerases are remarkable molecular machines that are used in nature to catalyse the DNA- (or in the case of reverse transcription, RNA-) templated transfer of nucleotide monophosphates to a growing chain of DNA with exquisite accuracy. As this study progressed my need for expertise in laser spectroscopy to complete the study led to the beginning of what was to become a valuable long-term interdisciplinary collaborative relationship with my physical chemistry colleague David Klenerman. Our mutual interests and intuition led us down a pathway to observe the action of a DNA polymerase on a DNA substrate to explore what new insights could be gained using single molecule fluorescence detection methods that had, at that time, only just been made accessible. One of our very early experiments included the observation of single molecules of DNA polymerase binding to substrate DNA (template plus primer) in solution, using a highly sensitive confocal single molecule detection system.⁸ The initial studies progressed towards experiments in which we immobilized discrete single molecules of DNA to a surface⁹ in order to ultimately observe the incorporation of fluorescently labeled deoxynucleotide monophosphates by a DNA polymerase using labeled deoxynucleoside triphosphates as a co-substrate (Fig. 2). These studies were all fundamental in nature and driven solely by our curiosity.

Around this time the International Human Genome Project was at a relatively early stage, in terms of the proportion of genome sequenced, but unmistakably gathering momentum. At the Lensfield Road laboratories we were but a short drive away from the Wellcome Trust Sanger Institute, which co-led the Human Genome Project and an even shorter cycle ride from the MRC laboratories where Sanger-sequencing had been invented. In this environment, we were naturally inclined to think about a world beyond the sequencing of

the first human genome and how our studies and observations might lead us to a new method for sequencing DNA. Together with our postdocs, we had a series of stimulating and creative discussions culminating in a defining and exciting moment in the Panton Arms (our 'local') one August afternoon in 1997 that gave rise to the key concepts that ultimately led to Solexa sequencing technology. At that point we also became intensely inspired by the possibility of sequencing the whole genomes of individuals to enable the comprehensive elucidation of genetic variation and the genetic basis of human disease.¹⁰

A key concept was an adaptation of our earlier studies to enable solid phase DNA sequencing (Fig. 3). Each of the four dNTPs is orthogonally colour-coded with fluorophores to allow the identity of each incorporated nucleobase, and thus the templating base, to be decoded by imaging the surface. During one cycle, all four colour-coded dNTPs are introduced under conditions that allow them to compete such that a DNA polymerase accurately incorporates the correct complementary base opposite the first templating base of a DNA substrate. By chemically 'blocking' the labeled dNTP a subsequent incorporation event could be prevented, thus restricting the cycle to a single nucleotide extension step. The incorporated base is then identified by imaging the labeled DNA at the surface and detecting the colour code. Chemical unblocking of the last incorporated nucleotide and removal of the fluorophore render the system ready for the next cycle. Thus a continuous sequence of n DNA bases can be decoded in n sequencing cycles.

A key feature of solid phase sequencing is that it provides the means to achieve a massively parallel process by having a great many different DNA fragments immobilised on a surface that can be sequenced simultaneously. The theoretical limit of having one sequenceable DNA fragment per diffraction limited site (less than 1 square μm) suggested the potential to build a system with considerably more than a million sequencing features on a surface less than a square cm. At that time (1997) we had in mind an audacious goal of being able to sequence as much as a billion bases of DNA per experiment—comparable in magnitude with the size of a human genome.

Another key consideration was the ability to prepare a massively parallel DNA sample array. We reasoned that a genomic DNA sample could be fragmented and the fragments immobilized at high dilution onto a suitably reactive surface to generate a random array of single molecules that could then each be sequenced by the solid phase sequencing scheme described in Fig. 3. A considerable advantage being that in effect many DNA fragment samples could be prepared in 'one pot' and sequenced on one sample array, rather than having to individually prepare each DNA fragment (Fig. 4). Sequencing a randomly generated fragment of DNA enables the sequence fragment to be identified and located in the context of a genome by realignment to a preexisting 'master' genome sequence.

The 3'-oxygen of the DNA being extended is the essential nucleophile for the nucleotidyl transfer reaction. Therefore for practical implementation of the sequencing chemistry, the most straightforward way to control stepwise incorporation was to design a bioorthogonal protecting group to mask the 3'-oxygen of the modified dNTP. We elected to keep this small in order to minimize steric interference in the polymerase active site. The colour-coding tag (fluorophore) could then be attached to the non-Watson-Crick edge *via* the 5-position (pyrimidines) and the 7-position (purines) by a cleavable linker (Fig. 5).

After an extensive exploration of various chemistries, an adaptation of Staudinger's chemistry¹¹ ultimately prevailed to yield a working solution. A removable azido methyl group was employed to mask the 3'-oxygen of an incorporated nucleotide, whereby deprotection could be carried out cleanly by a water-soluble phosphine after each cycle (Fig. 5).¹² The same chemistry was applied to enable the simultaneous cleavage/removal of the

fluorophore, thereby simplifying and streamlining the process. Besides bioorthogonality, this chemistry also allowed a relatively compact protecting group at the 3'OH of the incoming nucleotide, which was important for being tolerated by the DNA polymerase. Improved compatibility of the sequencing deoxynucleoside triphosphates with the active site of a DNA polymerase was ultimately achieved by re-engineering the polymerase active site by mutagenesis.

Many other important technical challenges were addressed during the developments that ultimately led to a robust, exportable commercial sequencing system. They included the engineering of a suitable surface for sequencing, molecular biology developments for sample preparation along with considerable engineering and computational developments, not least the challenge of how to manage, process and assemble the vast volume of primary/raw data (terabytes) anticipated from the technology. A notable modification from our original vision was the introduction of solid phase DNA amplification¹³ after the formation of the single molecule array to generate many copies of each DNA molecule at each same site. This offered practical advantages compared to the sequencing of single molecules, such as a stronger signal, a less expensive imaging system and reduction of stochastic single molecule errors to improve accuracy. Further developments have enabled the sequencing of one end of the original DNA fragment, followed by the other end (called paired-end sequencing) with read lengths that can each exceed 100 bases to facilitate *de novo* sequencing of genomes in addition to re-sequencing of genomes.

Fig. 6 shows an image taken from one of the early experimental sequencing runs; each spot represents many copies of a single DNA fragment and the colour informs us of the identity of the DNA base being decoded at each site during that sequencing cycle.

The first commercially available sequencing system using this method was called the *Genome Analyser* and was able to accurately sequence a billion bases of DNA (1 gigabase, or 1 G) per experiment in 2006, achieving the goal we had originally set in 1997. The first human African,¹⁴ Asian¹⁵ and cancer¹⁶ genomes plus the first giant panda genome¹⁷ were sequenced on the *Genome Analyser*. Substantial further improvements and innovations have followed with the most recently developed adaptation of this technology (in 2010), called the *HiSeq 2000*, which can sequence greater than 200 billion (200 G) accurate bases per experiment (the equivalent of 2 human genomes with 30-fold average oversampling) in about a week. This represents an approximately million-fold improvement in throughput compared to the state-of-the art at the time the project was conceived. This was largely accomplished as a result of the extent of parallelization achieved, equivalent to a billion sequencing features per experiment. It is noteworthy that the sequencing capacity of this system has already surpassed our prediction of 1 billion bases per experiment by 200-fold suggesting that our original target had actually turned out to be somewhat conservative after all (!). It is noteworthy that there has been an approximately five-fold increase per annum in Solexa-Illumina sequencing capacity per system, since the introduction of the *Genome Analyser*. This gradient of sequencing power has so far exceeded the prophecy of Moore's law on the doubling in the number of transistors per circuit each year.¹⁸ It is also very significant that there has also been a considerable shift in the economics of sequencing with the cost per accurate human genome falling below 10 000 dollars (greater than 10 000-fold cheaper than the first human genome).

While our original vision was to enable routine, low-cost accurate sequencing of genomes, there has been an impressive array of creative applications of massively parallel short read sequencing, most of which we had not foreseen at the early stages of the work. In general, these applications exploit the capacity to accurately analyse a large number of experimentally generated DNA (or RNA) fragments without the need to presume any

sequence information. Furthermore, since every sample molecule of DNA (or RNA) on the sample array leads to a discrete sequence read, the method is digital (*i.e.* one can count the number of sequenced nucleic acid fragments, derived from a particular region in the genome). These include the genome-wide mapping of protein (*e.g.* transcription factors) interactions to the genome (ChIP-seq),¹⁹ genome-wide single-base resolution analysis of cytosine methylation,²⁰ digital RNA sequencing (RNA-seq)²¹ to name just a few. Collectively, such applications of the method are providing fundamental insights into biology on a scale that was hitherto not possible.

A single sequencing system today has a capacity that is far greater than the pooled sequencing capacity of the entire world a decade ago. An important consequence of this is that smaller laboratories are now able to carry out routine genome-scale experimentation that were previously only conceivable as part of large-scale, multi-year, international projects. This has been clearly reflected in the fact that most such instruments have been placed in smaller labs rather than in large genome centres.²² Over 1500 publications²³ have employed Solexa–Illumina sequencing since its general release with most of the work having emerged from smaller laboratories. This democratization of sequencing capability is changing the nature and culture of life sciences and rapidly broadening the impact of sequencing across a number of fields such as plant sciences, environmental sciences, bioenergy and the study of all organisms.

Given that human (and other) genomes can now be sequenced quickly, routinely and at low cost, we are at the beginning of a phase of immense and exponential growth in the acquisition of genome sequence data. The use of high capacity sequencing to comprehensively characterise human genomes (DNA), transcriptomes (RNA) and the epigenomes (epigenetic changes) holds the potential to reveal the genetic basis of disease and disease pre-disposition in very fine detail. This is exemplified by large-scale projects such as The 1000 Genomes Project²⁴ and the International Cancer Genome Project.²⁵ There are already some promising early examples that have demonstrated how genome sequencing and digital RNA sequencing can be used to inform clinical decision making.²⁶ The next decade will reveal how far genome-wide sequence analysis will be able to inform clinical management in various disease areas. The vital role that chemistry has played in making this possible should be noted in this special year of celebration.

This project required the contributions of highly talented and dedicated individuals, too numerous to mention, to take the fundamental science and concepts that initiated this project and progress Solexa–Illumina sequencing to where it is today.

I dedicate this article to our co-workers who helped start the project in the University of Cambridge, the staff and scientists of Solexa and Illumina, the Solexa scientific advisory board, the founding investors and the numerous others who provided invaluable advice, inspiration and support. I thank Tony Smith and David Klenerman for their helpful comments on the manuscript. I would like to acknowledge the Biotechnology and Biological Sciences Research Council (BBSRC) of the UK for its vision in funding the fundamental science that led to the concepts that underpin Solexa–Illumina sequencing.

Biography

Shankar Balasubramanian is the Herchel Smith Professor of Medicinal Chemistry at the University of Cambridge. His research involves exploring the chemical biology of nucleic acids and the genome.

References

1. Watson JD, Crick FH. *Nature*. 1953; 171:737. [PubMed: 13054692]
2. Brown DM, Todd AR. *J. Chem. Soc.* 1952; 52
3. Brown DM, Fried M, Todd AR. *Chem. Ind.* 1953; 352
4. Maxam AM, Gilbert W. *Proc. Natl. Acad. Sci. U. S. A.* 1977; 74:560. [PubMed: 265521]
5. Sanger F, Nicklen F, Coulson AR. *Proc. Natl. Acad. Sci. U. S. A.* 1977; 74:5463. [PubMed: 271968]
6. International Human Genome Consortium. *Nature*. 2004; 431:931. [PubMed: 15496913]
7. Shendure J, Hanlee J. *Nat. Biotechnol.* 2008; 26:1135. [PubMed: 18846087]
8. Osbourne MA, Balasubramanian S, Furey WS, Klenerman D. *J. Phys. Chem. B.* 1998; 102:3160.
9. Osborne MA, Furey WS, Klenerman D, Balasubramanian S. *Anal. Chem.* 2000; 72:3678. [PubMed: 10952560] Osborne MA, Barnes CL, Balasubramanian S, Klenerman D. *J. Phys. Chem. B.* 2001; 105:3120.
10. An example of one of the early Solexa patents is: Balasubramanian S, Klenerman D, Barnes C, Osborne MA. *US Patent* 7,232,656 B2. 2004
11. Staudinger H, Meyer J. *Helv. Chim. Acta.* 1919; 2:635.
12. Milton, J.; Wu, X.; Smith, M.; Brennan, J.; Barnes, C.; Liu, X.; Ruediger, S. World Intellectual Property Organization WO/2004/018497. 2004.
13. Adessi C, Matton G, Ayala G, Turcatti G, Mermod J-J, Mayer P, Kawashima E. *Nucleic Acids Res.* 2000; 28:e87. [PubMed: 11024189]
14. Bentley D, Balasubramanian S, Swerdlow H, et al. *Nature*. 2009; 456:53. [PubMed: 18987734]
15. Wang J, Wang W, Li R, et al. *Nature*. 2009; 456:60. [PubMed: 18987735]
16. Ley TJ, Mardis ER, Ding L, et al. *Nature*. 2009; 456:66. [PubMed: 18987736]
17. Li R, Fei W, Tian G, et al. *Nature*. 2010; 463:311. [PubMed: 20010809]
18. Moore GE. *Electronics*. 1965; 38:114.
19. Barski A, et al. *Cell (Cambridge, Mass.)*. 2007; 129:823. Johnson DS, Mortazavi A, Myers RM, Wold B. *Science*. 2007; 316:1497. [PubMed: 17540862]
20. Lister R, O'Malley RC, Gregory BD, Berry CC, Millar AH, Ecker JR. *Cell (Cambridge, Mass.)*. 2008; 133:523.
21. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. *Nat. Methods*. 2008; 5:621. [PubMed: 18516045]
22. Reported by Illumina CEO in the transcript of 2010 Q4 results. <http://seekingalpha.com/article/251647-illumina-s-ceo-discusses-q4-2010-results-earnings-call-transcript>
23. Tracked and updated on the Illumina website. <http://www.illumina.com/>
24. The International 1000 Genomes Project Consortium. *Nature*. 2010; 467:1061. [PubMed: 20981092]

25. The International Cancer Genome Consortium. *Nature*. 2010; 464:993. [PubMed: 20393554]
26. Jones SM, Laskin J, Li YY, et al. *GenomeBiology*. 2010; 11:R82. [PubMed: 20696054]

Fig. 1.

Left: Maxam and Gilbert sequencing employed sequence-selective chemical fragmentation followed by separation by polyacrylamide gel electrophoresis to de-code DNA. Right: Sanger sequencing used chemistry to selectively terminate DNA synthesis at each of the four bases, followed by polyacrylamide gel electrophoresis to read the DNA sequence.

Fig. 2.
Early experiments to visualize DNA synthesis at the single molecule level.

Fig. 3.
Solid phase DNA synthesis.

Fig. 4.
Creating a clonal single molecule DNA array.

Fig. 5.

Left: reversible terminator dNTP with cleavable coding dye. Right: sequencing dNTP (dTTP shown) used for 1st generation Solexa sequencing.

Fig. 6.
An image of the surface taken during a cycle of an early Solexa sequencing experiment. Each of the spots is a cluster of identical DNA sample fragments and the colour indicates which of the four bases has been incorporated at that particular cycle.