Genome sequencing in the fast lane

Automation has increased the speed of DNA sequencing by established methods by several orders of magnitude. Now, commercial and academic efforts to develop alternative sequencing technologies are trying to push the envelope even further. Laura Bonetta reports.

The sequencing of genomes has become a mainstay of biological research. According to the genome project database of the National Institutes of Health (NIH) National Center for Biotechnology Information, over 300 genomes have been completely sequenced and another 750 are in the works. And genome sequencing projects are not only proliferating but also reaching for increasingly loftier goals. In December of last year, the NIH announced plans to embark on a mission to identify the genomic alterations involved in all types of cancers by using an array of technologies, first among them being large-scale genome sequencing (see **Box 1**).

The primary method of sequencing DNA — referred to as dideoxynucleotide sequencing, chain-termination sequencing or, more commonly, Sanger sequencing, after its inventor, Frederick Sanger[1] — was first developed in 1977. Thanks to many improvements in the equipment and reagents used, the speed at which sequences are read has greatly advanced since then. Today automated sequencers carry out many of the steps that were once done by hand to churn out over two million bases in one day. But a growing demand for even greater speeds and lower costs is pushing the development of new sequencing technologies, which are just starting to make their way into the marketplace.

The PyroMark ID uses Pyrosequencing to read a DNA sequence. (Courtesy of Biotage.)

Traditional sequencing

The Sanger sequencing method is based on the incorporation of 2′,3′-dideoxynucleotide triphosphates (ddNTPs) — similar to the deoxynucleotides (dNTPs) that link up to make DNA, but with a chain-terminating hydrogen atom instead of a hydroxyl group attached to the 3′ carbon — to a growing DNA chain. In a sequencing reaction, a single-stranded DNA fragment is combined with the appropriate sequencing primer; a ddNTP (for example, ddTTP); and the normal dNTPs (dTTP, dCTP, dATP and dGTP), one of which is labeled. When DNA polymerase is added to the mix, it begins to synthesize the corresponding DNA strand. DNA synthesis will stop every time the ddTTP is added, resulting in many labeled DNA fragments of varying lengths but always with a T residue at the end. This reaction is done four times using a different ddNTP in each reaction. After gel electrophoresis and autoradiography, the arrangement of the nucleotides in the DNA can be determined by putting the fragments in the four lanes in order.

Scientists can obtain reagents for Sanger sequencing, including ddNTP mixes and buffers, as well as electrophoresis systems, from several companies, including GE Healthcare (formerly Amersham), USB, Beckman Coulter, Bio-Rad, CBS Scientific, Sigma-Aldrich and others. Increasingly, companies are producing specialized reagents that make it easier to read through difficult regions, such as stretches of DNA that contain the same nucleotide repeated over and over or that have many C and G residues in a row. As an example, Beckman Coulter's new GenomeLab Methods Development Kit contains a set of core reagents plus a choice of nucleotide mixtures: dITP for routine sequencing, and dGTP for sequencing through difficult G-C –rich and polymerase 'hard stop' regions.

Although sample preparation and sequencing reactions are still mostly done by hand, these days automated sequencers take care of loading and running the gels and reading the results.

BOX 1 SEQUENCING CANCER

The National Cancer Institute and the National Human Genome Research Institute launched a comprehensive effort, dubbed 'The Cancer Genome Atlas', to identify the genomic changes involved in all types of cancer. For now, the two institutes have committed $50 million each to a three-year pilot project whose success will determine the feasibility of a full-scale effort. "This is a revolutionary project. I think you will remember this day. I am sure this will be a turning point for cancer research," said National Cancer Institute deputy director Anna Barker, speaking at a press conference in Washington, DC, on 13 December 2005.

For the pilot project, a small number of cancers, of the 200 or so types that exist, will be chosen for study. Hundreds of samples from each cancer type will be characterized by sequencing a subset of genes to identify possible mutations, as well as by finding other types of larger-scale genomic alterations, such as copy-number changes and chromosomal translocations, which contribute to cancer development or progression or both. "It is great to include technologies other than sequencing," says Elaine Mardis of Washington University School of Medicine. "I think it would be useful to also tie in protein-type measurements."

Another component of the pilot project will be to support the development of new methods for genomic analysis. For more information about The Cancer Genome Atlas, visit http://cancergenome.nih.gov/.

Faster sequencing by automation

The key to faster, automated sequencing was switching from a single terminator nucleotide in a sequencing reaction to four terminator nucleotides labeled with four different fluorophores that can be easily distinguished from one another. All four reactions are analyzed in a single lane of a gel rather than in four lanes, thereby increasing throughput. A laser in the automated sequencer constantly scans the bottom of the gel, detecting bands as they move past, producing a complex electropherogram with a colored peak representing each different nucleotide. Although all automated sequencers come with their own software to read the results of the sequencing reactions, several companies provide additional software modules for the analysis of automated or conventional sequencing (see **Box 2**).

Automated sequencers commercially available differ in the speed at which samples are run, the number of samples that can be processed in parallel in a single 'run' of the instrument and, of course, in cost. The market leader is Applied Biosystems (ABI)'s flagship 3730xl sequencer. The machine contains a capillary array — with each capillary not wider than a human hair and equivalent to one slab gel lane — that can run 96 sequencing reactions, each generating some 800 bases, in parallel. "The 3730xl can process up to 3,000 samples a day. You just load samples and walk away," says Phillippe Nore, senior director of strategic planning and business analysis at ABI. Indeed, the machine dunks the capillary array into a sample 96-well plate, denatures and loads the samples, applies the voltage program and analyzes the data.

ABI's first automated sequencer, introduced in the mid-1980s, had a 16-lane slab gel and could sequence up to 6,400 bases in a 24-hour day; two decades later, the 3037xl can 'spell out' two million bases in the same time period and at a fraction of the cost. "The improvements have come from a large number of different areas — for example, how the instrument is configured, what gel matrix is used in the capillaries, and so on. They all have an effect on speed," says Nore.

The capillary array of ABI's 3730xl sequencer. (Courtesy of Applied Biosystems.)

The company plans to release additional improvements to allow users to reach even higher outputs. "When we first introduced the 3730xl system most of our customers were involved in *de novo* sequencing. For that, it was necessary to offer long read lengths of 600 to 800 bases per run or else the assembly [of the complete sequence] is a nightmare," says Nore. "But today more and more activity is shifting to resequencing for which you don't need long read lengths." For resequencing, a researcher will typically use PCR primers to amplify a sequence of interest, typically an exon of a gene, determine its sequence and compare it with a reference sequence to identify any changes. Thus, unlike *de novo* sequencing, resequencing does not require the assembly of hundreds of overlapping fragments of DNA.

A new module that ABI will release this month will allow users to carry out shorter runs, yielding about 400 quality bases of sequence, thereby increasing the number of runs that can be done in one day from about 40 (with the existing software) to 72. "This will enable users to double sample throughput and significantly lower cost per sample," says Nore.

GE Healthcare and Beckman Coulter also sell capillary-based systems. GE Healthcare's Megabace 4500 system, new this year, graduated from a 96-well to a 384-well based apparatus, allowing for much higher throughput than its predecessor. Another improvement has been the implementation of a new matrix for the capillaries that allows longer read lengths. "The application that we focus on most is *de novo* sequencing," says Carl Fuller, vice president for science at GE Healthcare, "the 384 capillaries are better suited for high throughput. Read lengths are regularly above 800 bases. The feedback from customers is that they appreciate the ability to get a lot of data quickly." With the Megabace 4500, a researcher can sequence 2.8 million bases in a 24-hour period.

Beckman Coulter's offerings in the area of automated sequencers are the CEQ 8000 Series genetic analysis systems. "In the early 1990s, many companies started focusing on high throughput. We thought there was still a need for medium throughput applications," says Noreen Galvin, GenomeLab business manager. Indeed researchers use the CEQ 8000 systems for confirmatory sequencing and mutation analysis, as well as other genomic applications such as amplified fragment length polymorphism fingerprinting. "All these genetic analysis functions can be performed with one gel, one array and one software package," adds Galvin.

LI-COR Biosciences is one of the few companies that continue to sell a more traditional gel-based system. The system is semiautomatic, in that the user still must pour and load the gel, but the machine reads the results. Although it involves more labor, the LI-COR sequencer is not as expensive as a capillary-based machine. As a result, the system is popular among researchers studying animal phylogenetics and marine life, "areas where funding is not as high," says Jeff Harford, product marketing manager at LI-COR. Another niche market for the system is education. "Universities are buying them to train undergraduate students," says Harford.

Pyrosequencing

As an alternative to Sanger sequencing, Biotage pioneered a technology called pyrosequencing that reads the DNA sequence as the DNA strand is synthesized. In a pyrosequencing reaction, a primer hybridized to a single-stranded DNA template is incubated with DNA polymerase, ATP sulfurylase, firefly luciferase and a nucleotide-degrading enzyme. A particular dNTP is added to the reaction and if it is incorporated into the growing DNA strand, a signal is produced; unincorporated dNTPs are degraded. DNA synthesis is accompanied by the release of inorganic pyrophospate that is converted to ATP by the ATP sulfurylase. The production of ATP is then sensed by the luciferase. The amount of light produced in the luciferase-catalyzed reaction is measured by a charge-coupled device camera or other instrument.

BOX 2 HELPING TO ANALYZE

Automated sequencers have taken many of the steps of sequencing out of scientists' hands. But researchers still need to design sequencing primers, analyze the output from a sequencer, resolve problem areas in the sequence, run queries against several databases and assemble individual sequences together. A variety of commercial and free software tools helps to simplify these steps.

Software that is commonly used to view and edit results obtained with an automated sequencer include Geospiza's Finch Trace View, which is suitable for both PC and Mac computers. The software displays an entire chromatogram trace in a scalable multipane view, allows the user to view the raw data, can launch BLAST searches and produces reverse-complement sequences and traces. DNASTAR's Lasergene v6 analysis software allows the user to make alignments, assemble contigs, design primers, perform restriction mapping and predict protein structure from DNA sequence. The Chromas package by Technelysium, which works only on PCs, opens chromatogram files, exports sequences in plain text with base numbering, pastes the sequences into other applications, searches for specific sequences, and displays translations. The 4Peaks program by Mekentosj performs similar functions for Mac computers.

Invitrogen's Vector NTI Advance software, now in version 9.1, can 'string' routine operations, such as BLAST searching, primer design and open-reading-frame mapping on multiple sequences, into a seamless pipeline. In addition, the software has robust primer and molecule design capabilities, alignment tools that retain and link annotations, and powerful DNA sequence assembly algorithms.

One of the more popular sequence analysis packages is Sequencer from Gene Codes. It contains powerful algorithms to assemble DNA fragments quickly and accurately based on given parameters. "Sequencer will compare the forward and reverse-complement orientations to assemble the best contig," says marketing manager, Frank White. The program's fast contig assembly is coupled with a set of user-friendly editing tools that allow restriction enzyme mapping; heterozygote detection; conversion of cDNA to genomic DNA sequence; large gap alignment; support for confidence scoring; comparative sequencing; and open reading frame, motif and single-nucleotide polymorphism analysis. "It has been used in applications as diverse as mutation detection and forensics," says White.

In addition, a user can scroll through aligned data or use selection tools to highlight regions of discrepancy or low quality. When working with multiple sequences, Sequencher will 'call' secondary peaks, trim vectors, trim low-quality ends and create consensus sequences.

Different dNTPs are added sequentially, one at a time, to obtain a sequence in real time. "You can either do directed dispensations [of dNTPs], if you know the sequence and are studying single-nucleotide polymorphisms or mutations, for example, or cyclic additions of nucleotides if you don't know the expected sequence. The machine does everything in real time, from the dispensations to displaying the sequence as it is synthesized," says Robert England, global marketing director at Biotage. The company's instrument, PyroMark ID, will sequence 96 samples in less than an hour. "You don't need any gels or labels, so those costs are absent," says England. Biotage has introduced PyroMark sequence analysis kits of reagents optimized for specific applications such as microbial identification, promoter methylation and cancer mutations.

The signal produced during DNA synthesis is quantitative, which means that if two C residues are incorporated, the signal will be twice as intense as it would be if one were incorporated. "In quantification, what makes pyrosequencing different from everything else is its resolution. Instead of 'Yes, No or Half', it tells you, 'This CpG site is 37% methylated', or '5 out of the 8 copies of this gene are mutated'," says England.

One of the limitations of pyrosequencing is difficulty in reading the sequence in homopolymeric regions — stretches of the same base in the DNA. The other potential limitation is that read lengths reach up to only 100 bases, so the PyroMark ID instrument is not suitable for genome-sequencing projects.

Next-generation sequencing

In the past decade, automated sequencers became faster and cheaper, but the stream of improvements seems to have reached a plateau. "Fundamentally we have not reached a barrier but it is becoming harder to justify the costs associated with pushing the limit even further," says ABI's Nore. As a result, newer systems that have just come on the market or are under development use completely different approaches for sequencing DNA. "Conventional sequencing has not seen any real improvements in the last few years," says Marcel Margulies, vice president of engineering at 454 Life Sciences, a company that has adapted pyrosequencing to high-throughput, large-scale projects[2]. "Our view and mission is to democratize sequencing.

The Megabace 4500 system can generate 2.8 million bases in a 24-hour period. (Courtesy of GE Healthcare.)

Right now sequencing is very capital intensive. Our system will allow individual laboratories to generate results extremely rapidly in a more cost-effective way."

This year the company signed an exclusive distribution agreement with Roche Applied Science for the marketing and sales of the Genome Sequencer 20 System. With this system, randomly overlapping segments of DNA are clonally amplified on beads that are 30 micrometers in diameter. After PCR, the beads (each of which will carry 10 million molecules of DNA) are centrifuged in PicoTiterPlates containing 1.6 million wells. The sequencing reactions, based on pyrosequencing technology that was 'tweaked' to routinely reach read lengths of over 100 bases, are carried out in the plates and the results are then read by the instrument. Each run yields at least 20 million bases (for example, 200 thousand reads at an average of 100 bases per read). "If you want to sequence a small bacterial genome, it will take you three and a half days, compared to one month with Sanger," says Marcus Droege, global marketing director for genome sequencing at Roche. "Some things are now possible that were not practical before. For example, you can sequence four or five bacterial genomes and compare them to one another to identify determinants of drug resistance in a couple of weeks," he adds. Elaine Mardis, codirector of the Genome Sequencing Center at Washington University School of Medicine, a facility that has just purchased its second instrument from Roche, agrees that there is a lot of demand for people wanting bacterial sequences quickly.

With the right controls, the 454 platforms can achieve a level of accuracy comparable to that of traditional sequencing. "We have done those comparisons and the numbers are quite similar," says Mardis. "In terms of substitutions and deletions, they are very similar. Homopolymers are more of a problem with the 454 platform. That is really the main difference."

Because read lengths generated by pyrosequencing are not as long as those generated by Sanger sequencing, the new technology is not yet suitable for sequencing large mammalian genomes. In *de novo* sequencing, researchers determine the sequence of overlapping DNA segments and then string them together based on the regions of overlap; the shorter the read lengths, the more redundancy that is created.

Right now the Genome Sequencer 20 System is being marketed for sequencing bacterial genomes, but "there are a lot of other projects for which it works," says Droege. In 2006 the company plans to release reagents and software for a new application for so-called 'ultra deep' or amplicon sequencing. The application will provide a way to analyze exons of genes by PCR to find mutations in them. "It will be particularly valuable for detecting mutations present at low frequencies in a mixed population of samples," says Droege.

The biotech company Solexa is developing a different technology referred to as 'Sequencing-By-Synthesis'—a method that uses proprietary fluorescence-labeled modified nucleotides. These nucleotides, which have a reversible termination property, allow each cycle of the sequencing reaction to occur simultaneously in the presence of all four nucleotides. According to the company, homopolymer repeats are dealt with as any other sequence and with high accuracy; this avoids the problems of measuring intensity and deducing how many bases were present in the repeat. "If all of their promises hold up, Solexa's instrument has a lot of potential," says Mardis. "It may be more of a cost-effective option for sequencing large genomes."

Companies like ABI are also in the run for developing next-generation sequencing technologies. "When they are mature enough to bring to the market we will do it," says Suresh Pisharody, product manager for high-throughput sequencing. "We are not a small start-up that needs to have a product. We will launch when we are ready."

The company is pursuing a platform based on single-molecule sequencing in both internal R&D and collaborative efforts. Earlier this year ABI entered a formal collaboration with VisGen, one of the companies awarded NIH grants to develop techniques for sequencing a human-sized genome for $1,000 or less (see http://www.genome.gov/15015202 for information on this National Human Genome Research Institute grant program). VisGen's platform is based on the engineering of both polymerase and dNTPs to act as molecular sensors of DNA base identity in real time. The company says that once their system is on the market, it will be possible to read one million bases per second per machine.

Other efforts aiming at faster and cheaper sequencing include a method for clonally amplifying short DNA fragments on magnetic beads and then embedding them into a polymer matrix on the surface of microscope slides[3]. The technology, dubbed 'polymerase colony (polony) sequencing', was licensed and is being further developed by Agencourt Biosciences. Another technology, licensed to Lasergen, uses a four-laser system that overcomes some of the limitations of labeling pieces of DNA with four colors of fluorescent dyes. Agilent is developing nanopore sequencing. As a DNA strand passes through tiny channels (called 'nanopores'), different base pairs obstruct the pore to varying degrees, causing measurable variations in the electrical conductance of the pore, producing a unique electronic signature that can be used to infer the DNA sequence. The Affymetrix platform is a sequencing-by-hybridization method that uses a DNA sequence immobilized on a chip or membrane. The degree to which different oligonucleotide probes bind the target DNA can be used to infer a sequence.

These next-generation sequencing technologies are generating great excitement in the life-science community with the offer of high throughput at a lower cost. But, at least for now, Sanger sequencing is not in danger of extinction. "It is the gold standard for many applications that require very high accuracy, such as mutation profiling, or long read lengths, such as *de novo* sequencing," says ABI's Nore. Five years into the future, however, the sequencing landscape will undoubtedly look very different.

1. Sanger, F. *Proc. Natl. Acad. Sciences USA* **74**, 5463–5467 (1977).
2. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
3. Shendure, J. *et al. Science* **309**, 1728–1732(2005).

Laura Bonetta is a freelance writer based in the Washington, DC area (lbonetta@nasw.org).

SUPPLIERS GUIDE: COMPANIES OFFERING EQUIPMENT, REAGENTS AND SOFTWARE FOR DNA SEQUENCING

Company	Web address
Accelrys	http://www.accelrys.com/
Agencourt Bioscience	http://www.agencourt.com/
Applied Biosystems	http://www.appliedbiosystems.com/
Beckman Coulter	http://www.beckmancoulter.com/
Biotage	http://www.biotage.com/
Bio S&T	http://www.biost.com/
Bio-Rad	http://www.bio-rad.com/
Cambrex	http://www.cambrex.com/
CBS Scientific Company, Inc.	http://www.cbssci.com/
DNASTAR, Inc.	http://www.dnastar.com/
GE Healthcare	http://www.gehealthcare.com/
454 Life Sciences	http://www.454.com/
Genamics	http://genamics.com/
Geospiza	http://www.geospiza.com/
Gene Codes Corporation	http://www.genecodes.com/
GeneFlow, Ltd.	http://www.geneflow.co.uk/
Invitrogen	http://www.invitrogen.com/
Mekentosj	http://www.mekentosj.com/
MiraiBio	http://www.miraibio.com/
Mobious Genomics	http://www.mobious.com/
Molecular Cloning Laboratories	http://www.mclab.com/
MWG Biotech	http://www.mwg-biotech.com/
Nanofluidics	http://www.nanofluidics.com/
Network Biosystems	http://www.networkbiosystems.com/news.asp
Nucleics	http://www.nucleics.com/
Parallabs	http://www.parallabs.com/
PerkinElmer	http://www.perkinelmer.com/
Polymorphic DNA Technologies Inc.	http://www.polymorphicdna.com/
Retrogen, Inc.	http://www.retrogen.com
Roche Applied Sciences	http://www.roche-applied-science.com/
Rockland Immunochemicals, Inc.	http://www.rockland-inc.com/
Solexa	http://www.solexa.com/
Stratagene	http://www.stratagene.com/
Technelysium	http://www.technelysium.com.au/chromas.html
Textco	http://www.textco.com/
TimeLogic	http://www.timelogic.com/
TriStar Technology Group	http://www.tristargroup.us
USB	http://www.usbweb.com/
US Genomics	http://www.usgenomics.com/
VisiGen Biotechnologies	http://www.visigenbio.com/