10 DNA sequencing

This exposition is very closely based on the following sources, which are all recommended reading:


10.1 Introduction

The term DNA sequencing refers to sequencing methods for determining the order of the nucleotide bases - adenine, guanine, cytosine, and thymine - in a molecule of DNA.

Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as:

- diagnostic,
- biotechnology,
- forensic biology and
- biological systematics.

The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project.

Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.

10.2 History

Fifteen years elapsed between the discovery of the DNA double helix in 1953 and the first experimental determination of a DNA sequence. This delay was caused by several factors that made the problem intimidating:

1. The chemical properties of different DNA molecules were so similar that it appeared difficult to separate them.
2. The chain length of naturally occurring DNA molecules was much greater than for proteins and made complete sequencing seems unapproachable.
3. The 20 amino acid residues found in proteins have widely varying properties that had proven useful in the separation of peptides. The existence of only four bases in DNA therefore seemed to make sequencing a more difficult problem for DNA than for protein.
4. No base-specific DNAases were known. Protein sequencing had depended upon proteases that cleave adjacent to certain amino acids.
10.2.1 Early RNA sequencing

Some RNA molecules did not share all of these drawbacks with DNA.

In particular, transfer RNA molecules were small and individual types could be purified. RNAases with base specificity were known so methods analogous to those used in protein sequencing could be developed. Consequently Escherichia coli alanine tRNA was the first nucleic acid molecule to be sequenced by Holley and coworkers in 1965:

![Structure of an alanine RNA](source: Holley et al, 1965)

Unlike amino acid sequences, which were not interpretable until 3D protein structures were determined by X-ray crystallography, models for the structure of tRNAs could be deduced by assuming base pairing analogous to that found in the DNA double helix:

![Structure of a tRNA](source: Holley et al, 1965)

10.2.2 Restriction enzymes

An important event for DNA sequencing was the discovery of type II restriction enzymes by Hamilton Smith and coworkers (1970).

These enzymes recognize and cleave DNA at specific short nucleotide sequences, typically 4 – 6 bp in length. The restriction enzymes provided a general method for cutting a large DNA molecule into a number of smaller pieces that could be separated by size using gel electrophoresis. These pieces have specific ends that could function as starting points for the sequencing methods that developed over the next few years.

10.2.3 Early DNA sequencing techniques

In 1975, Sanger introduced the plus and minus method for DNA sequencing (Sanger and Coulson, 1975).

This was a critical transition technique leading to the modern generation of methods that have com-
pletely dominated sequencing over the past 30 years. The key to this advance was the use of polyacrylamide gels to separate the products of primed synthesis by DNA polymerase in order of increasing chain length.

Maxam and Gilbert (1977) developed a DNA sequencing method that was similar to the Sanger and Coulson method in using polyacrylamide gels to resolve bands that terminated at each base throughout the target sequence, but very different in the way that products ending in a specific base were generated.

### 10.2.4 Chain-terminator sequencing

Both the plus-minus and Maxam–Gilbert methods were rapidly replaced by the *chain-terminator* method (Sanger, Nicklen, Coulson, 1977), also known as *Sanger sequencing* (after its developer Frederick Sanger) or *dideoxy sequencing*, as it was more efficient and used fewer toxic chemicals and lower amounts of radioactivity.

The key principle of the Sanger method is the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-terminator method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, radioactively or fluorescently labeled nucleotides, and modified nucleotides that terminate DNA strand elongation.

The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase.

To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length:

The newly synthesized and labeled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image.

In the image above, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths.

A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP).

The relative positions of the different bands among the four lanes are then used to read the DNA sequence.

The obtainable sequence length is \( \approx 100 \) nucleotides.
10.2.5 Dye-terminator sequencing

In the dye-terminator variant of Sanger sequencing, each of the four dideoxynucleotide chain terminators is labelled with a different fluorescent dye, permitting sequencing in a single reaction rather than four:

![Dye-terminator sequencing diagram](source: Wikipedia)

In 1986 the laboratory of Leroy Hood at Caltech, in collaboration with Applied Biosystems (ABI), published the first report of automation of DNA sequencing (Smith et al, 1986), which established the dye-terminator variant of Sanger sequencing. This initial report showed that sequencing data could be collected directly to a computer.

Sanger sequencing using dye-terminators became the dominant sequencing technique until the introduction of so-called next-generation sequencing technologies beginning in 2005.

The obtainable sequence length is now $\approx 1000$ nucleotides.

10.2.6 Automated sequencing

The ABI 370A DNA sequencer appeared very shortly, and was first used in 1987 to determine the sequence of a gene by Craig Venter and colleagues at NIH (National Institutes of Health in Bethesda, MD). At NIH, Venter set up a sequencing facility with six automated sequencers and two ABI Catalyst robots.

In 1992 Venter established The Institute for Genomic Research (TIGR) to expand his sequencing operation to a facility with 30 ABI 373A automated sequencers and 17 ABI Catalyst 800 robots.

The organizational differences between this facility and the manual sequencing projects that preceded it were as important as the use of automation. This was a real factory with teams dedicated to different steps in the sequencing process such as template preparation, gel pouring and sequencer operation. Data analysis was integrated into the process so that problems in earlier steps could be detected and corrected as soon as possible.

By contrast, in the sequencing of the 237 kb human cytomegalovirus genome in 1991, each participating investigator sequenced a region of the genome single-handedly.

An early demonstration of the power of automated sequencing was the development of the expressed sequence tag (EST) approach to gene discovery.

In this approach cDNA copies of messenger RNA are cloned at random and subjected to automated sequencing.

In the first report from Venter and colleagues in 1991, 337 new human genes were reported, 48 homologous to genes from other organisms (Adams et al, 1991).

This initial study was expanded to an analysis of 83 million nucleotides of cDNA sequence that identified fragments of more than 87000 human cDNA sequences, more than 77000 of which were previously unknown.
This approach was adopted by many genome projects. Today the EST database contains over 43 million ESTs from over 1300 different organisms.

In 1993 the Sanger Centre, later renamed the Wellcome Trust Sanger Institute, was established by the Wellcome Trust and the Medical Research Council in Great Britain.

This is arguably the most important of the sequencing centers established to participate in the worldwide public effort to sequence the human genome.

Located in Hinxton, near Cambridge, “the Sanger” is an outstanding example of a modern sequencing center. The facility had produced $\approx 3.4 \times 10^9$ bases of finished sequence by the 30th anniversary of the dideoxy sequencing method.

10.3 Cellular genomes

Until 1995 the only completely sequenced DNA molecules were viral and organelle genomes.

That year Craig Venter’s group at TIGR, and their collaborators, reported complete genome sequences of two bacterial species, *Haemophilus influenzae* (Fleischmann et al, 1995) and *Mycoplasma genitalium* (Frazer et al, 1995). The *H. influenzae* sequence gave the first glimpse of the complete instruction set for a living organism.

*H. influenzae*, is a non-motile Gram-negative rod-shaped bacterium first described in 1892 by Richard Pfeiffer during an influenza pandemic. It is a member of the *Pasteurellaceae* family.

*H. influenzae* was mistakenly considered to be the cause of influenza until 1933, when the viral etiology of the flu became apparent. Still, *H. influenzae* is responsible for a wide range of clinical diseases.

The genome consists of 1,830,140 base pairs of DNA in a single circular chromosome that contains 1740 protein-coding genes, 58 transfer RNA genes tRNA, and 18 other RNA genes.

The *M. genitalium* sequence showed an approximation to the minimal set of genes required for cellular life.

*Mycoplasma genitalium* is a small parasitic bacterium which lives on certain cells of the primate genital and respiratory tracts. It is the smallest known free-living bacterium, and the second-smallest bacterium after the recently-discovered endosymbiont *Carsonella ruddii*. Until the discovery of *Nanoarchaeum* in 2002, *M. genitalium* was also considered to be the organism with the smallest genome ($\approx 580$ kb).

The methods used to obtain these sequences were as important for subsequent events as the biological insights they revealed.

Sequencing of *H. influenzae* introduced the whole genome shotgun (WGS) method for sequencing cellular genomes.

In this method, genomic DNA is fragmented randomly and cloned to produce a random library in *E. coli.*
Clones are sequenced at random and the results are assembled to produce the complete genome sequence by a computer program that compares all of the sequence reads and aligns matching sequences. Sanger and colleagues used this general strategy to sequence the lambda phage genome (48.5 kb), published in 1982. However, no larger genome was shotgun sequenced until \textit{H. influenzae} (1.83 Mb).

Outline of the basic WGS strategy:

Another critical factor in the application of shotgun sequencing to cellular genomes was the TIGR assembler (Sutton et al, 1995). Previous assembly programs were not designed to handle thousands of sequence reads involved in even the smallest cellular genome projects. However, the TIGR assembler that had been designed to assemble vast amounts of EST data was adequate for the job.

Once these initial sequences were reported the floodgates were open and a steady stream of completed genome sequences has been appearing ever since.

Some of the most important milestones were:

- The bacteria \textit{E. coli} and \textit{Bacillus subtilis},
- the yeast \textit{Saccharomyces cerevisiae},
- the nematode \textit{C. elegans},
- the fruit fly \textit{Drosophila melanogaster},
- the plant \textit{Arabidopsis thaliana}, and
- the human genome.

10.3.1 \textbf{Sequencing of E. coli}

Because of its position as the pre-eminent model organism of molecular biology, sequencing of the genome of \textit{E. coli} (4.6 Mb) was already proposed in 1983.

Sequencing proceeded as sequencing technology improved, starting with manual methods and finishing in 1997 with automated sequencers.

This \textit{E. coli} genome sequence, along with several other subsequently sequenced strains, has yielded a wealth of information about bacterial evolution and pathogenicity.
10.3.2 Sequencing of \textit{S. cervisiae}

Meanwhile, another model for large-scale genome sequencing projects had emerged; the \textit{international consortium}.

The first genome sequence to be completed by this approach was the yeast \textit{S. cerevisiae} (12.0 Mb), in late 1996. This was the also the first eukaryotic organism to be sequenced. 

The project involved about 600 scientists in Europe, North America and Japan. The participants included both academic laboratories and large sequencing centers.

10.3.3 Sequencing of \textit{B. subtilis}

The next success of the consortium approach was the genome of the bacterium \textit{B. subtilis} (4.2 Mb), in 1997.

The project began in 1990 with the participation of five European laboratories. The project finally became a consortium of 25 laboratories in six European countries coordinated at the Institut Pasteur in France. A consortium of seven Japanese laboratories, coordinated at the Nara Institute of Science and Technology, Japan, also participated, as well as one Korean and two US laboratories.
10.3.4 Sequencing of *C. elegans*

The first animal genome sequenced was that of “the worm” *C. elegans* (97 Mb), in 1998. The authorship of this work was simply *The C. elegans Sequencing Consortium*, which was a collaboration between the Washington University Genome Sequencing Center in the United States and the Sanger Centre in UK.

*C. elegans* is studied as a model organism for a variety of reasons. It is a multicellular eukaryotic organism that is simple enough to be studied in great detail. Strains are cheap to breed and can be frozen. When subsequently thawed they remain viable, allowing long-term storage.

10.3.5 Sequencing of *Arabidopsis thaliana*

The flowering plant *Arabidopsis thaliana* is an important model system for identifying genes and determining their functions.

Its genome sequence was published in 2000 by the *Arabidopsis Genome Initiative* and it was the first plant to be sequenced.

The sequenced regions cover 115.4 Mb of the 125 Mb genome and extend into centromeric regions.

The evolution of *Arabidopsis* involved a whole-genome duplication, followed by subsequent gene loss and extensive local gene duplications. The genome contains 25,498 genes encoding proteins from 11,000 families, similar to the functional diversity of *Drosophila* and *C. elegans*. 
10.3.6 The Celera sequencing factory

In 1996, ABI introduced the first commercial DNA sequencer that used capillary electrophoresis rather than a slab gel (the ABI Prism 310), and in 1998 the ABI Prism 3700 with 96 capillaries was announced. For the first time DNA sequencing was truly automated. The considerable labor of pouring slab gels was replaced with automated reloading of the capillaries with polymer matrix. Samples for electrophoresis were automatically loaded from 96-well plates rather than manually loaded as the previous generation of sequencers had been.

Celera Genomics was found by Applera Corporation (the parent company of ABI) and Craig Venter in May 1998 to exploit these new machines by applying Venter’s methods for WGS sequencing to the human genome, in direct competition with the publicly funded Human Genome Project.

Celera acquired 300 of the machines, each capable of producing $1.6 \times 10^5$ bases of sequence data per day, for a total theoretical capacity of $\approx 5 \times 10^7$ bases of raw sequence data per day.

The Celera sequencing factory:

10.3.7 Sequencing of D. melanogaster

Celera chose the D. melanogaster genome to test the applicability of the WGS approach to a complex eukaryotic genome.

This involved a scientific collaboration between the scientists at Celera and those of the Berkeley and European Drosophila Genome Projects. These projects had finished 29 Mb of the 120 Mb of euchromatic portion of the genome. (About one-third of the 180 Mb Drosophila genome is centromeric heterochromatin.)

Using the WGS approach, data was collected over a 4-month period that provided more than $12 \times$ coverage of the euchromatic portion of the genome. The results validated the data produced by the ABI 3700s, the applicability of the WGS approach to eukaryotic genomes, and the assembly methods developed at Celera.

This was a nearly ideal test case because the WGS data could be analyzed separately and then portions of it could be compared with finished sequence already produced by the Drosophila Genome Projects. At the same time the sequence information provided a valuable resource for Drosophila genetics. More than 40 scientists at an “Annotation Jamboree” did initial annotation of the sequence.
These scientists, mainly drawn from the Drosophila research community, met at Celera for a 2-week period to identify genes, predict functions, and begin a global synthesis of the genome sequence information.


10.4 The human genome

Eventual sequencing of the human genome became an imaginable goal at the outset of the sequencing era 30 years ago. Formal discussions of the idea began in 1985 at meetings in the USA.

Discussions continued and in 1988 reports recommending a concerted genome research program were issued by committees of the congressional Office of Technology Assessment and the National Research Council. In 1990 a 5-year US Human Genome Project plan was presented to the US Congress. It was estimated that the project would take 15 years and cost 3 billion USD.

The US Human Genome Project established goals of mapping, and in some cases sequencing, several model organisms as well as humans. These included E. coli, S. cerevisiae, C. elegans, D. melanogaster and mouse (laboratory strains of Mus domesticus).

So, several of the sequences discussed above received funding from the Human Genome Project. The publicly funded effort became an international collaboration between a number of sequencing centers in the United States, Europe and Japan. Each center focused sequencing efforts on particular regions of the genome, necessitating detailed mapping as a first step.

In 1994, a detailed genetic map of the human genome was published including 5840 mapped loci with a mean spacing of 0.7 cM (1 centiMorgan = \( \approx 10^6 \) bp).

In 1998 the public project, now in a race with Celera, also adopted the new ABI Prism 3700 capillary sequencers. In 1999 the Human Genome Project celebrated passing the billion base-pair mark, and the first complete sequence of a human chromosome was reported (chromosome 22).

At the same time at Celera, human genome sequencing was underway using the WGS strategy. Human genome sequencing began in September 1999 and continued until June 2000, when data collection was completed and an initial assembly was achieved.

The Celera data provided approximately 5-fold coverage of the genome. An additional 3-fold coverage of unordered and unoriented BAC sequences from the public effort was included in the assembly. The power of the WGS strategy was amply demonstrated.

Bioinformaticians led by Gene Myers at Celera developed the first computer program capable of assembling a mammalian-sized genome and thus proved that the computational challenges posed by such a huge dataset can be overcome.

(An article in “Die Zeit” asks whether Gene Myers is the one that deserves a Nobel prize: “Es wird sich noch zeigen, ob der legendäre Genom-Kartierer Craig Venter den Nobelpreis verdient hat oder doch sein Chefmathematiker Gene Myers.”. [http://www.zeit.de/2004/50/C-Mathematik])

On 25 June 2000 at the White House, President Clinton with Francis Collins (NIH) and Craig Venter publicly announced draft versions of the human genome sequence from both the publicly funded project and from Celera. In February 2001 the Celera and the public draft human genome sequences
were published the same week in Science and Nature.

The race was officially a tie, but it was clear to all that the entry of Celera had speeded the process by several years. Both projects ended up needing the other to make the progress that was made. The Celera assembly benefited from data produced in the public project and the public project quickly adopted some of Celera’s methods, in particular the paired-end strategy. Celera’s basic methods have now been adopted by all publicly funded genome projects.

10.5 Next generation sequencing

Since 2005 new methods are emerging that challenge the supremacy of the dideoxy method.

The common feature of these methods is that they are massively parallel, meaning that the number of sequence reads from a single experiment is vastly greater than the 96 obtained with modern capillary electrophoresis-based Sanger sequencers.

At present this very high throughput is achieved with sacrifices in length and accuracy of the individual reads when compared to Sanger sequencing. Nonetheless, assemblies of such data can be highly accurate because of the high degree of sequence coverage obtainable.

The methods are designed for projects that employ the WGS approach. They are most readily applied to resequencing, in which sequence data is aligned with a reference genome sequence in order to look for differences from that reference.

A few examples of specific instruments that employ massively parallel strategies are discussed below. Other technologies are under development, and all of these methods will undoubtedly continue to improve.

10.5.1 Roche 454 sequencing

The first of the massively parallel methods to become commercially available was developed by 454 Life Sciences and is based on the pyrosequencing technique (Margulies et al, 2005).

This system allows shotgun sequencing of whole genomes without cloning in E. coli or any host cell. First DNA is randomly sheared and ligated to linker sequences that permit individual molecules captured on the surface of a bead to be amplified while isolated within an emulsion droplet.

A very large collection of such beads is arrayed in the 1.6 million wells of a fiber-optic slide. As with the Sanger method, sequencing is carried out using primed synthesis by DNA polymerase. The array is presented with each of the four dNTPs, sequentially, and the amount of incorporation is monitored by luminometric detection of the pyrophosphate released (hence the name “pyrosequencing”).

A CCD imager coupled to the fiber-optic array collects the data. In sequencing across a homopolymer run, the run length is estimated from the amount of pyrophosphate released, which is proportional to the number of residues incorporated. Errors that result from misjudging the length of homopolymer runs result in single-base insertions and deletions (indels). These constitute the major source of errors in 454 data.
The 454 Genome Sequencer FLX is reportedly able to produce 100 Mb of sequence with 99.5% accuracy for individual reads averaging over 250 bases in length.

Emulsion-based clonal amplification:

Deposit DNA beads into a PicoTiterPlate:
Parallel sequencing of fragments:

Base-calling is performed by analyzing a series of images produced in the flow cycles:

10.5.2 Illumina (Solexa) sequencing

The second next-generation sequencing technology to be released (in 2006) was Illumina (Solexa) sequencing (Bennett, 2004).

A key difference between this method and the 454 is that it uses chain-terminating nucleotides. The fluorescent label on the terminating base can be removed to leave an unblocked 3’ terminus, making chain termination a reversible process. The method reads each base in a homopolymer run in a separate step and therefore does not produce as many indels within such runs as the 454.

Because the reversible dye terminator nucleotides are not incorporated efficiently, the read length of the Solexa method is less than for 454. Also more base-substitution errors are observed due to the use of modified polymerase and dye terminator nucleotides. The method sequences clusters of DNA molecules amplified from individual fragments attached randomly on the surface of a flow cell.

Because of the very high densities of clusters that can be analyzed, the machine can reportedly produce more than one billion bases (1 Gb) of 75 base (paired) reads in a single run.
Produce fragments and add adapters:

1. Attach fragments to surface:

   - Adapter
   - DNA fragment
   - Dense lawn of primers
   - Adapter

   Attach DNA to surface
   Bind single stranded fragments randomly to the inside surface of the flow cell channels.
Add nucleotides for amplification:

Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Make fragments double stranded:

Attached terminus
Free terminus

Fragments become double stranded
Denature double strands:

[Diagram of denatured DNA with attached molecules]

Repeat and complete amplification:

[Diagram of clusters forming, text: On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.]
Add all four labeled nucleotides:

Measure intensities of nucleotide colors:
Add next nucleotides:

Repeat measurement:

Image of second chemistry cycle is captured by the instrument
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Repeat cycles to sequence bases:

The next instrument to be released by Illumina, the HiSeq 2000, will be able to sequence 200 GB in 12 days.

The BGI (formerly Beijing Genomics Institute) has ordered 128 instruments and will thus be able to generate \( \approx 2 \text{ Tera bases per day.} \)

The Weigel lab at the MPI for Developmental Biology here in Tübingen has two Illumina sequencers.
10.5.3 ABI’s SOLiD sequencing

Applied Biosystems has also developed a massively parallel sequencer, its Supported Oligonucleotide Ligation and Detection system (SOLiD), released in 2008.

The technology is based on a hybridization-ligation chemistry (Shendure et al, 2005). The sample preparation aspect of this technology including library preparation, clonal amplification of the target DNA by emulsion PCR on beads is very similar to the 454 processes in principle.

However, the size of the beads used for emPCR (1 µm versus 26 µm) and the array format (random versus ordered) are different. These differences afford the SOLiD technology the potential of generating a significantly higher density sequencing array (potentially over a few hundred fold higher), as well as more flexibility in terms of sample input format.

The sequence interrogation is done through repeated cycles of hybridization of a mixture of sequencing primers and fluorescently labeled probes, followed by ligation of the sequencing primers and the probes, then the detection of the fluorescent signals on the probes which encode the bases that are being interrogated. Although it has a short read length of about 25 – 35, it can generate $\approx 2 \text{ – } 3 \text{ Gb}$ of sequence per run.

A small company in Tübingen called CeGAT uses a SOLiD sequencer for contract sequencing and research.

10.5.4 Pacific Biosciences SMRT DNA sequencing

According to the Pacific Biosciences website:

“SMRT™ DNA Sequencing offers very long reads, ultra-fast cycle times, and the flexibility to cost-effectively perform small or large projects.

Because the SMRT DNA sequencing system provides temporal information about every base incorporation event, it can measure the kinetics of the enzyme independently for each base in a DNA sequence. The kinetics of incorporation are sensitive to, for example, the methylation status of the DNA template being sequenced.

SMRT sequencing will unify the formerly separate applications of real-time, single-molecule DNA sequencing and methylation sequencing. This creates the potential to visualize methylation status and other epigenomic markers as a by-product of ordinary DNA sequencing with no change in sample preparation or run conditions.

SMRT sequencing will unify the formerly separate applications of DNA sequencing and methylation sequencing. In addition, elimination of the bisulfite conversion step will save time and money as well as avoid the deleterious effects of conversion.”

The first instruments are to be shipped in 2010.

http://www.pacificbiosciences.com/

10.6 Genomic medicine

All disease has a genetic basis, whether in genes inherited by the affected individual, environmentally induced genetic changes that produce a cancer, or the genes of a pathogen and their interaction with
those of the infected individual.
The sequencing of the human genome as well as all major pathogens is beginning to have a major impact on the diagnosis, treatment and prevention of diseases. Genome sequences have provided potential targets for drug therapy as well as vaccine candidates.
An era of personalized medicine, informed by information concerning the patient’s genotype, has been widely predicted.
The Holy Grail in this field is the “$1000 human genome” sequence. To stimulate work in this area the X Prize Foundation has established the $10 million Archon X Prize for Genomics (http://genomics.xprize.org/).

10.7 Metagenomics

Only a very small fraction of the microbes found in nature have been grown in pure culture. Consequently we lack a comprehensive view of the genetic diversity to be found on Earth.

An approach to this problem has emerged called *metagenomics* or *environmental genomics*. DNA is isolated directly from environmental samples and sequenced, without attempting to culture the organisms from which it comes.

Early studies using manual DNA sequencing methods focused on 16S rRNA genes as a measure of phylogenetic diversity (Schmidt et al, 1991).

High-throughput sequencing methods make shotgun sequencing of the whole metagenome from environmental samples informative. A study of the DNA isolated from an acid-mine biofilm yielded nearly complete genome sequences for two organisms and partial sequences for three others from just 76.2 Mb of data (Tyson et al, 2005).

Studies of the more complex environment of the oceans have shed new light on the diversity of life on Earth. A total of more than 1.6 Gb of sequence from Sargasso Sea samples yielded 1.2 million previously unknown gene sequences, only about one-third of which were recognizable homologs of previously known genes (Venter et al, 2004). Before analysis of the Sargasso Sea data the NCBI non-redundant amino acid (nr) dataset contained some 1.5 million peptides, about 630,000 of which were classified as bacterial.

In an expanded study, ocean samples were collected during an expedition from the North Atlantic through the Panama Canal and into the South Pacific. Shotgun sequencing of these samples produced a dataset of 7.7 million sequence reads yielding 6.3 Gb of sequence (Yutin et al, 2007). More than 6 million proteins are predicted from this global ocean sampling (GOS) data.

These include members of almost all previously known prokaryotic protein families. When these GOS data were clustered with all previously known protein sequences, about 4000 out of a total of 17,000, medium and large clusters contained only GOS sequences.

The metagenomic approach is being applied to study microbial populations in many environments, for example the human gut. (See also: http://nihroadmap.nih.gov/hmp/.)

A difficulty with current metagenomic sequence datasets from complex microbial communities is that the vast majority of the data cannot be assembled. This is mainly due to the fact that the cost of
acquiring enough sequencing data for the full assembly of all or most of the organisms present in any given sample is still prohibitively high using the state-of-the-art Sanger-based sequencing technology. The development of strategies to apply the new massively parallel sequencing methods may solve this problem if the sequence read length and the accuracy of these technologies can be dramatically improved.

In the meantime, another approach is the sequencing of individual cells isolated from the environment. Substantial progress toward this goal has been achieved.

It is clear that the metagenomic approach is becoming a major new tool for understanding our environment. It will also provide a vast array of new genes for application in the emerging field of synthetic biology.

### 10.8 What next?

The amount of nucleotide sequence in the databases has increased logarithmically by nine orders of magnitude over the 40-year period from 1965 to 2005:

![Sequence Database Size Graph](image)

This amounts to an average doubling time of about 16 months.

In 1977 it was possible to imagine a benchtop machine that could sequence the *E. coli* genome in a few days (although it could only be seen as a black box).

The *sequencing factory* would have been inconceivable because it was at odds with the whole culture of biological research. But the sequencing factory did come to pass, and with it came vast amounts of data ideally suited for computational analysis.

Taken at face value, the 16-month doubling time for the sequence database is substantially faster than the 24-month doubling in computer power predicted by Moore’s law (see [http://www.intel.com/technology/mooreslaw/](http://www.intel.com/technology/mooreslaw/)).

It appears possible that methods for collecting sequence data could soon outstrip our capacity to adequately analyze that data.

A prime example is all-against-all sequence comparison, which requires a computation time that increases with the square of the amount of sequence data.
At present a sizable bioinformatics core is an essential part of a sequencing center.

The currently popular vision that an investigator with a single benchtop machine could replace a large sequencing center can only be realized with increases in the productivity of computers and bioinformaticians even more dramatic than that expected for sequencers.

It appears that for our individual $1000 genome sequences to be truly useful, fundamental advances in computation and bioinformatics will be essential.

The obvious importance of computational analysis of sequence data has led to a greater overall appreciation of the role of theory in biology.

A relationship between theory and experiment, not unlike that found in 20th century physics, seems to be taking shape.