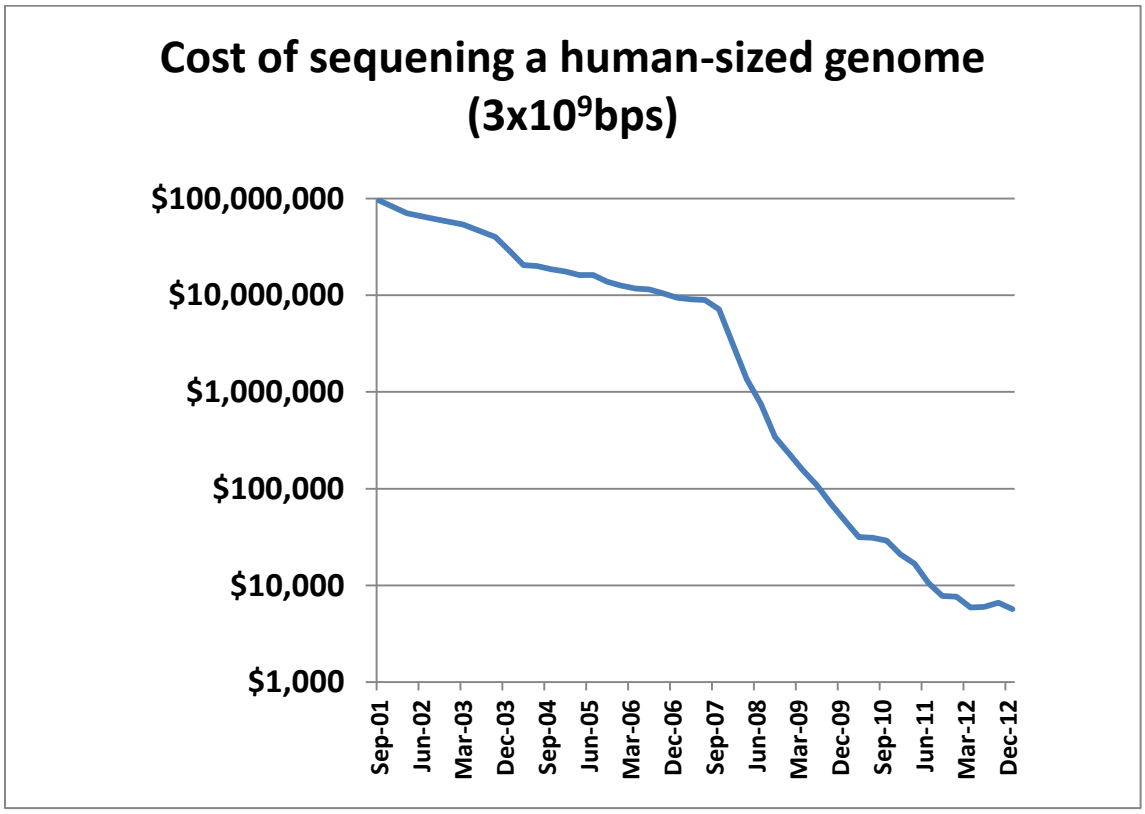
1 DNA Sequencing and Bioinformatics

DNA sequencing technologies have evolved rapidly. Between 2001 and 2007 the sequencing cost was roughly decreased by half every two years, but since July 2007 the cost has been reduced by half every 7 months. The steep decrease in cost since July 2007 is due to the implementation of the next-generation sequencing technologies.



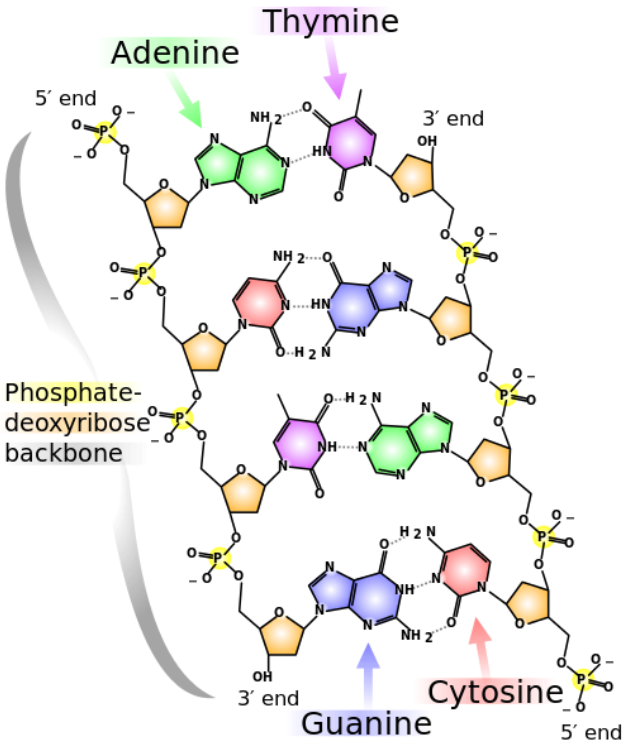
Data obtained from the National Human Genome Research Institute

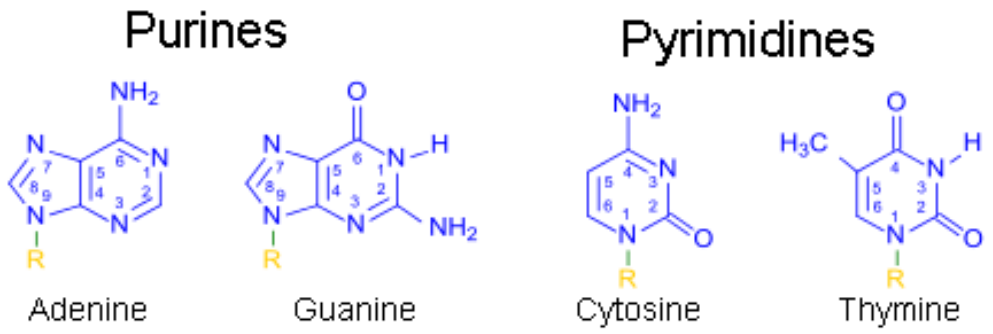
A brief historical overview of some of the past and current sequencing technologies is first presented. The second section discusses the underlying principle for the computer-assisted alignment of DNA sequences with an emphasis on the Basic Local Alignment Search Tool (BLAST) that is by far the most commonly used alignment approach in genomic research. A step-by-step illustration of the analysis of a given sequencing result with BLAST will be also presented.

1.1 DNA Sequencing Methods

1.1.1 Maxim and Gilbert's Method: Chemical Cleavage

One should be familiar with the chemical structure of DNA to appreciate the principle of Maxim and Gilbert's sequencing method. You may first refer to this interactive 3D display of double-stranded DNA. Make sure you can identify the ribose phosphate backbone, the C3' and C5' phosphorylation positions on the ribose ring, and the pairing of the nucleic bases. Also make sure you can distinguish between purine pyrimidine nucleic bases.





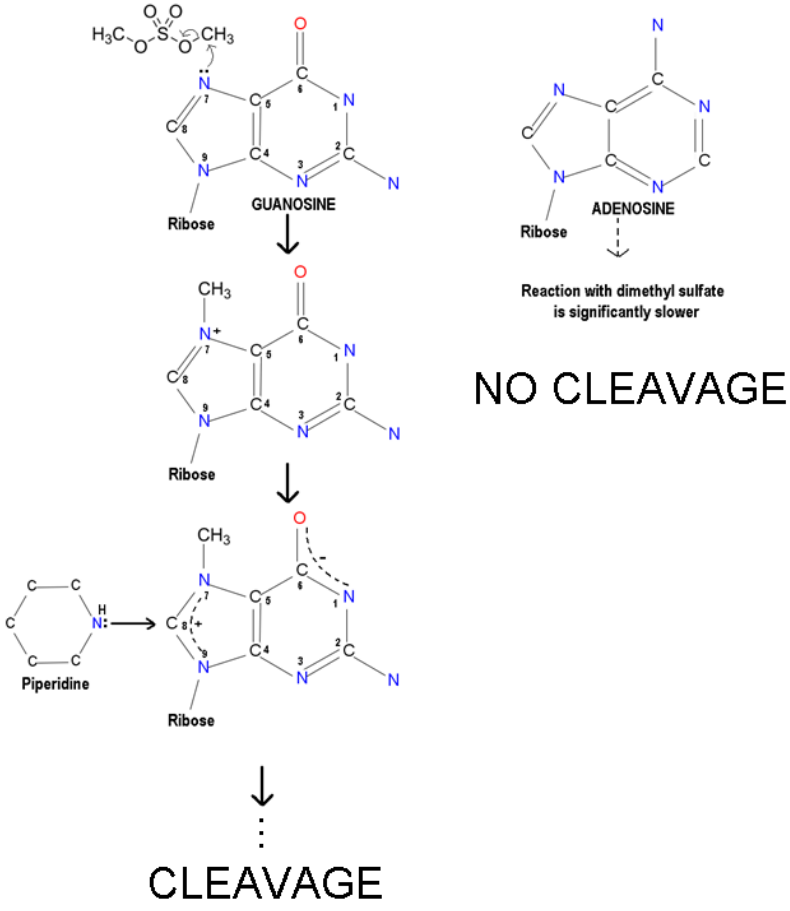
Maxam and Gilbert sequencing uses four different chemical reactions that each partially cleaves DNA at one or two specific nucleotides (PNAS, 1977). Note that mild reaction conditions are used to favor partial cleavage of end-labeled DNA strands (labeling is commonly done with radioactive ^{32}P attached at the 5' end of a DNA strand). Partial cleavage is necessary to ensure

that only a small fraction of the target nucleotides are cleaved leaving a series of DNA fragments of different lengths.

The table below summarizes the four chemical reactions used in Maxam and Gilbert sequencing along with the reaction mechanism 1 selectively cleaving the G residues.

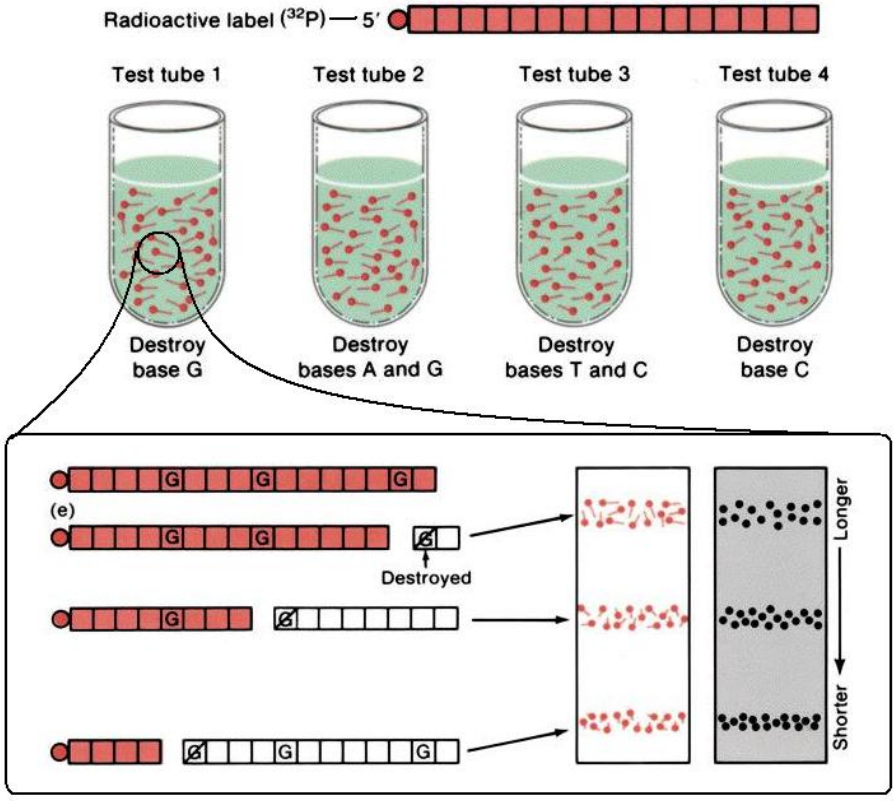
	Base modification	**Strand scission**	**Nucleotide(s) removed**
Reaction 1	Dimethylsulfate	Piperidine	G
Reaction 2	Acid	Acid	C and G
Reaction 3	Hydrazine (NH_2-NH_2)	Piperidine	C and T
Reaction 4	Hydrazine + salt	Piperidine	C

Molecular mechanism for reaction 1: Partial cleavage at some of the G residues

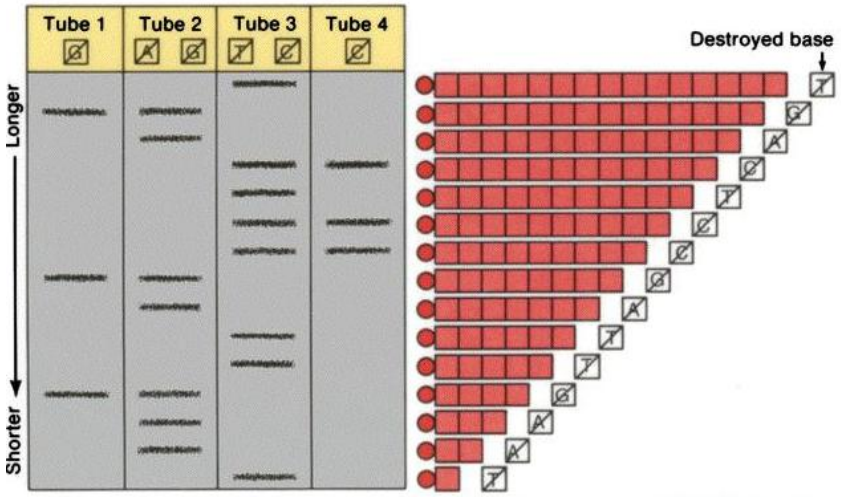


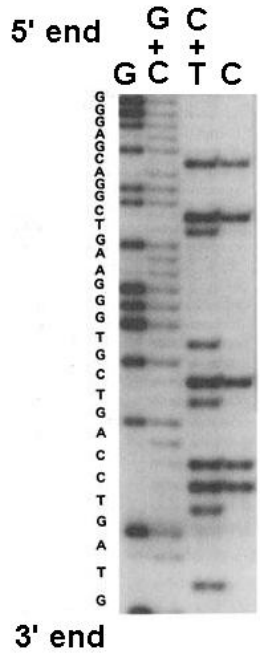
Once the four reactions have been separately completed the mixture of fragments of different lengths - remember that mild conditions are used so that cleavage happens at a small fraction of the targeted nucleic acids – are separated onto a gel with smaller fragments running faster and showing at lower positions onto the gel. Up to 100 base pairs (bps) can be read per sequencing reaction. A summary of Maxam and Gilbert sequencing is illustrated on next page.

Step 1: Labeling and partial cleavage at specific base(s): four reaction tubes are needed.



Step 2: Resolution of labeled fragments by gel electrophoresis (left) and base-calling of nucleotides (right).





Get ready for final exam

In 2013 the Maxam and Gilbert sequencing technology is obsolete and you will likely never be required to sequence a gene through this approach. The questions are intending to focus your attention on some important limitations of Maxam and Gilbert approach so you can better appreciate the efficacy of the newer sequencing technologies.

How many reaction tubes are needed to sequence a short DNA fragment of 50bps?

How many reaction tubes would be necessary to sequence a DNA fragment with 1000bps?

What is the labeling system used in Maxam and Gilbert sequencing? How safe is it for lab staff members?

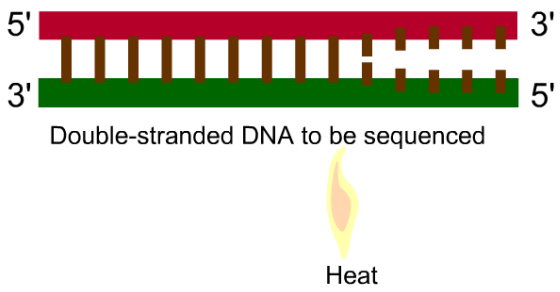
How easy is it to read and analyze the sequencing signal?

1.1.2 First Generation: Sanger Sequencing

The Sanger method of sequencing was introduced in 1975 and further refined in 1977 (PNAS, 1977). Note that both Sanger and Maxam-Gilbert methods were published the same year, 1977. Sanger's innovative approach takes advantage of DNA polymerase to synthesize a DNA strand that is complementary to a single-stranded DNA template. The first automated DNA sequencers designed by Caltech were based on Sanger method and commercially released in 1986 (Nature, 1986). The automated Sanger method has dominated the sequencing industry for almost two decades and it was used for sequencing the human genome. Small research laboratories still use Sanger sequencing, though it has been gradually replaced by Next-Generation Sequencing Technologies (NGST) since 2007, especially in larger genomics research institutes.

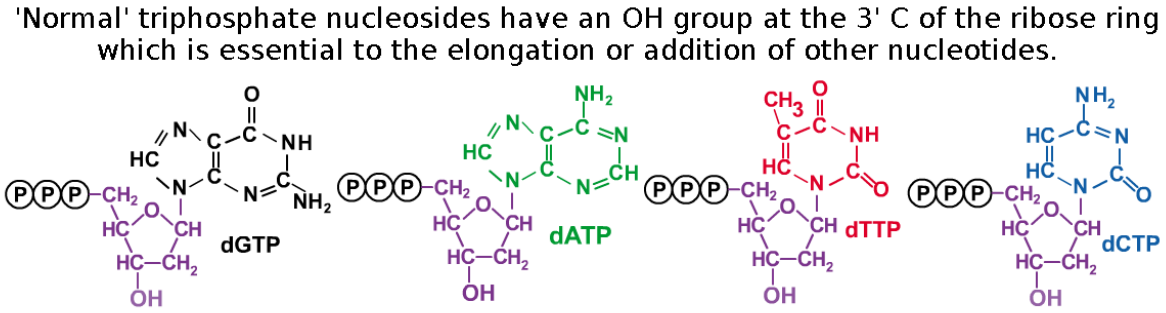
The distinctive features of Sanger sequencing method are listed below. You can also see an explanatory animation of Sanger sequencing method.

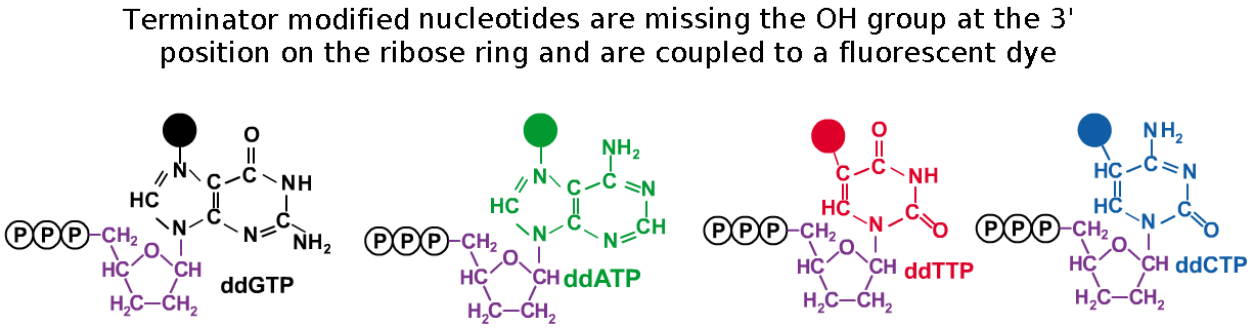
- The sequencing reaction mixture contains a DNA polymerase, the DNA template to be sequenced, a primer that can hybridize near the 5'end of the DNA template, and a mixture of free nucleotides.
- The sequencing template is single-stranded. Heat can be used to denature double-stranded DNA and obtain single-stranded DNA.



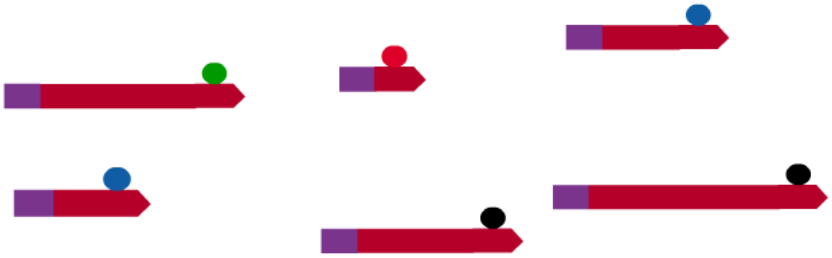
- A short oligonucleotide or a primer is necessary to prime addition of nucleotides at its 3’ end – remember DNA polymerase can only add free nucleotides, one at a time, at the 3’ end of an elongating fragment.

- The mixture of nucleotides contains mainly ‘regular’, which are ATPs, CTPs, GTPs and TTPs, along with a tiny amount of fluorescently labeled terminator nucleotides or dye terminator nucleotides. Modified labeled dideonucleotides can be accepted by the DNA polymerase, but their incorporation blocks any further elongation as they lack the usual OH group at the 3’ carbon onto the ribose ring. This explains why it`s only the 5’ terminal nucleotide that is labeled with a fluorescent dye. Each dye terminator dideoxynucleotide is labeled with a different fluorescent dye; there are four different dyes, one for each of the four nucleotides.

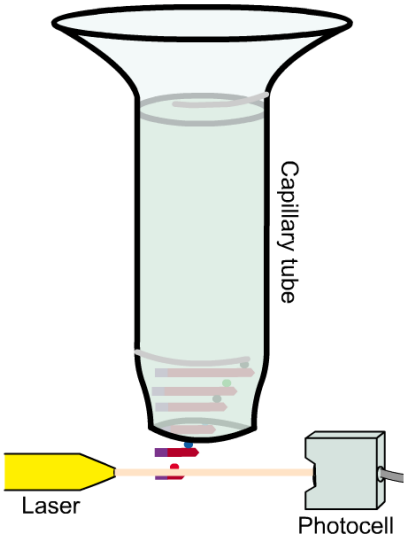


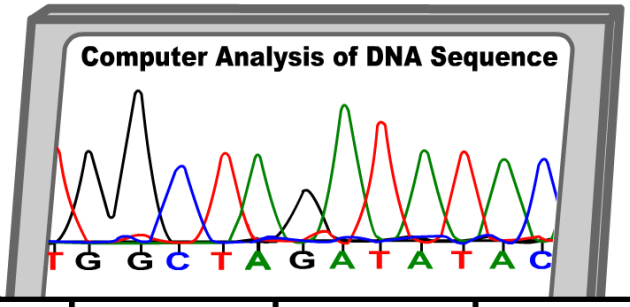


- Dye terminator nucleotides are used to (1) identify the identity of the terminal nucleotide and (2) prevent any further elongation or addition of nucleotides.
- At the end of a sequencing reaction a mix of fragments of different lengths is obtained and each fragment has its 3’ terminal nucleotide labeled with a fluorescent dye.



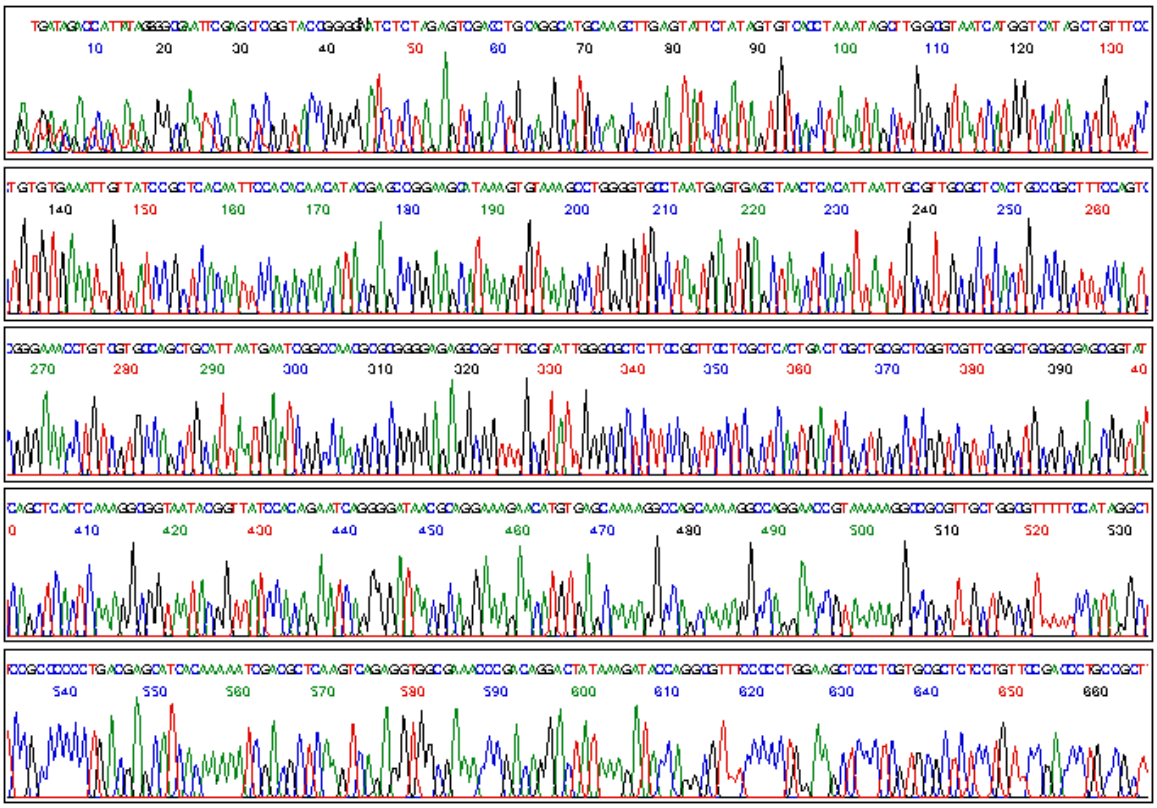
- The sequencing reaction mixture is resolved by capillary electrophoresis, which is more or less like a sieving chromatography column allowing smaller fragments to move faster and elute first. A laser-assisted detector is used to 'read the color' of the terminal nucleotide identifying its identity, that is an A, C, G or T.





Terminator	Fluorescence Wavelength	Fluorescence Color	Color Displayed by Computer
G	540 nm	green	black
A	570 nm	green-yellow	green
T	595 nm	yellow-orange	red
C	620 nm	orange-red	blue

A typical sequencing reaction based on Sanger method can read up to 1000 nucleotides, though the signal tends to be weaker near the end with higher error rates – the peaks just get too low to be accurately read.



Sanger sequencing is relatively simple as only one reaction tube is needed – all four nucleotides can be combined into one single tube as four different dyes are used - compared to four tubes for the Maxam and Gilbert sequencing. Sanger sequencing was therefore easier to adapt for automated high-throughput sequencing as needed for the Human Genome Project.

You may look at the sequencing facility at the Centre for Advanced Research in Environmental Genomics at Faculty of Science (http://www.careg.uottawa.ca/core-facilities.html).

Get ready for final exam

How many reaction tubes are needed to sequence a short DNA fragment of 50bps according to Sanger sequencing?

How many reaction tubes would be necessary to sequence a DNA fragment with 1000bps according to Sanger sequencing?

What is the labeling system used in Sanger sequencing? How safe is it for lab staff members?

How difficult is it to read and analyze the result signal in Sanger sequencing?

1.1.3 Next Generation(s) of Sequencing

The automated Sanger method is considered as a 'first-generation' technology and newer methods are referred to Next-Generation Sequencing (NGS). The major advance offered by NGS is the ability to produce an enormous volume of data cheaply – up to 1 billion short reads per instrument run!

1.1.3.1 Template Preparation

A common theme among NGS technologies is that the template DNA to be sequenced is immobilized to a support. The attachment of separated DNA templates allows thousands to billions of sequencing reactions to be performed simultaneously – it's not anymore needed to have a distinct tube for each fragment to be sequenced. In the **emulsion PCR technique** beads covered by a universal primer are suspended within an emulsion. The DNA to be sequenced is first split into fragments of 40-100 nucleotides that are further coupled or ligated to a short oligonucleotide complementary to the universal primer. The short ligated DNA fragments are then added to the emulsion. The DNA to bead ratio is kept low to ensure that most beads have either only one DNA template molecule hybridized onto one of the DNA primers coating a bead. A PCR reaction is then performed to amplify by several times the number of copies of the DNA molecule to be sequenced. The mixture of DNA beads can be further added and immobilized onto a glass slide. Each glass slide typically contains 100-200 million beads!

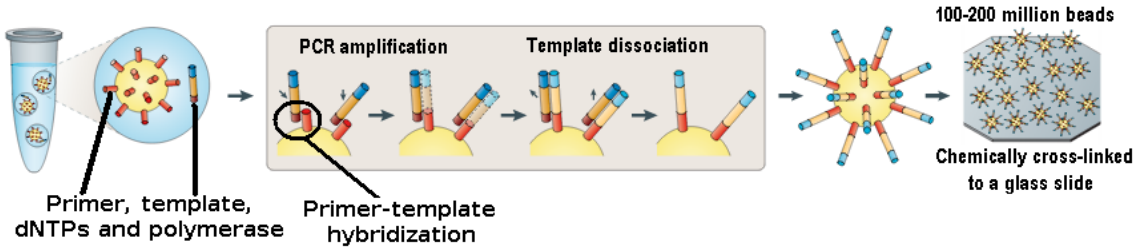
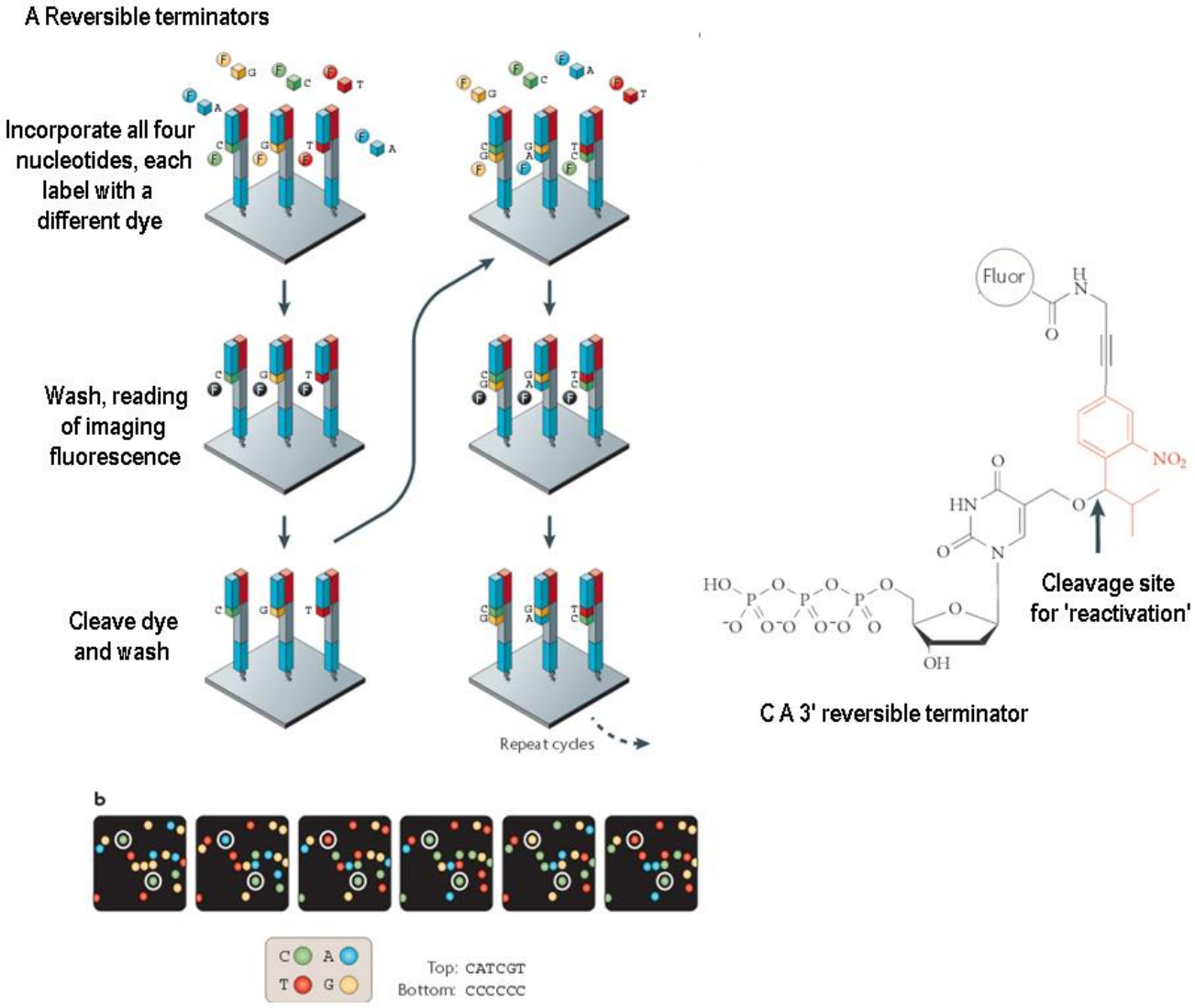


Figure from Nature Reviews/Genetics, 1010

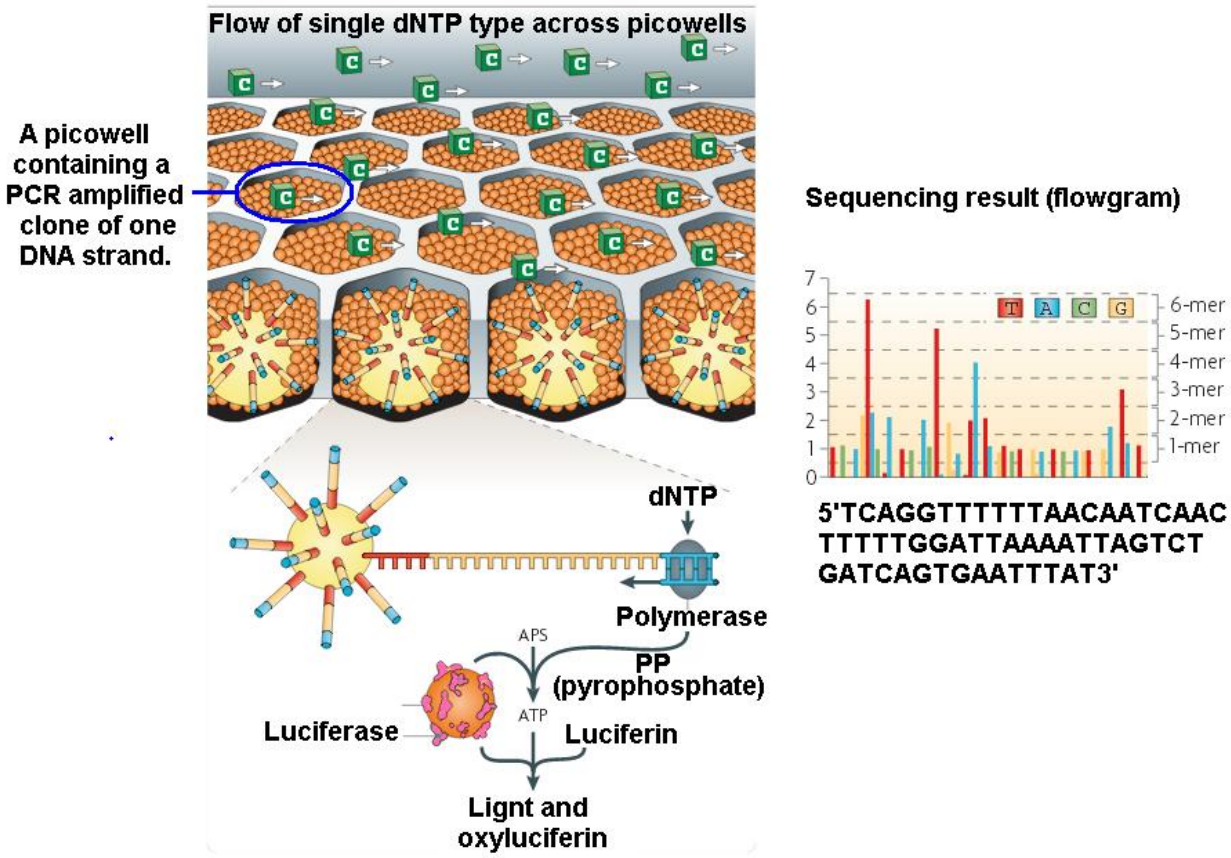
1.1.3.2 Sequencing and Imaging Technologies

In Sanger sequencing the addition of a dye terminator nucleotide irreversibly prevents any further elongation. A large number of DNA strands is needed for a sequencing reaction because the labeled strands cannot be recycled for further nucleotide reads. Some NGS technologies take advantage of reversible dye terminator nucleotides. Reversible termination is possible owing to modified nucleotides that can temporarily block any further incorporation of further nucleotides until the last added nucleotide can be read by fluorescence imaging. The fluorophore attached to the dye terminator nucleotide is then further cleaved to permit the addition of the next dye terminator nucleotide. The use of reversible terminators makes it possible to sequentially determine all nucleotides of a single DNA molecule, but one at a time.



Four-colour cyclic reversible termination method. The four reversible dye terminator nucleotides can be immobilized onto single-stranded DNA templates. Following reading of fluorescence color to identify the identity of the terminal nucleotide, a cleavage step removes the fluorescent. The four-colour images displayed in b highlight the sequencing data from two clonally amplified templates. Figure modified from Nature Reviews/Genetics, 1010

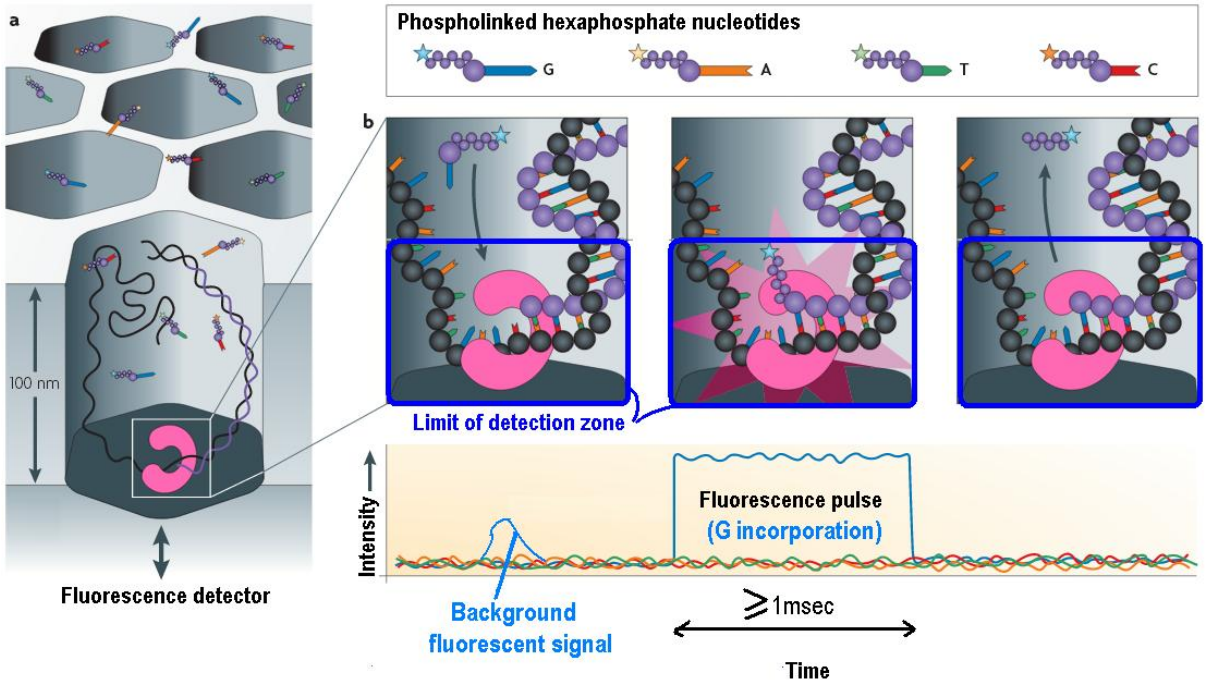
It is possible to avoid using terminator dye nucleotides and to directly sequence incorporated nucleotides by monitoring the release of pyrophosphate, which is the by-product of nucleotide incorporation. The incorporation of an adenosine residue, for instance, requires the input of an adenosine triphosphate (ATP) that is converted into adenosine monophosphate, the incorporated nucleic acid base, and pyrophosphate (P-P), the by-product. The sequencing technology is referred as pyrosequencing, that is sequencing based on detection of pyrophosphate (P-P) release. The figure below illustrates the pyrosequencing sequencing technology.



Pyrosequencing: following loading of the DNA-amplified beads into picowells, additional beads coated with luciferase are added. In this example, a single type of nucleotide, cytidine triphosphate or CTP, is shown flowing across the wells. A sensitive detector device is located just underneath the picowell plate and can read luciferin fluorescence. The picowell plate is exposed to only one nucleotide at a time and washes are performed between each cycle. A complete sequencing cycle involves the sequential exposure of a plate with dTTP's dATP's, dCTP's, and dGTP's, and several cycles are performed until the whole length of the DNA templates can be read. The result displayed at the right indicates that fluorescence intensity is directly related with the number of nucleotides incorporated. Figure modified from Nature Reviews/Genetics, 2010.

The next sequencing technology to be commercialized is likely to be real-time sequencing. Unlike reversible terminators, real-time sequencing relies on nucleotides that do not halt the process of DNA elongation. Pyrosequencing can be considered as a real-time technology, though only one type of nucleotides can be added at a time and this delays the elongation and sequencing process. Ongoing technological developments are intended to develop a real-time sequencing technology involving imaging the continuous incorporation of dye-labelled nucleotides during DNA synthesis. One real-time technology, which is displayed below, involves the attachment of one single DNA polymerase molecule to the bottom of nanocells with a fluorescence detector underneath each nanocells. The detector is designed to selectively read fluorescence within a small zone located in the immediate surroundings – tiny observation volume - of the DNA polymerase molecule. The overall setup reduces the number of stray fluorescently labelled nucleotides that enter the detection zone for a given period. The residence time of phospholabelled dNTP's that are incorporated at the 3' elongating end is

typically 1 millisecond; if a stray nucleotide transiently enters the detection zone for less than 1 millisecond the signal is considered as undesirable background and can be ignored.



The company that is aiming to commercialize this platform, Pacific Biosciences, reported a reading accuracy of 99.999% with read lengths of nearly 1000 nucleotides (similar to the first generation Sanger sequencing method). This is truly exceptional considering that fluorescence detection is based on single molecule event detection. Figure modified from Nature Reviews/Genetics, 1010.

.

Get ready for final exam

What is the underlying principle for pyrosequencing? Use a diagram to explain the detection signal that is read for sequencing nucleotides in the pyrosequencing method.

What is real-time sequencing? Use a diagram to explain the detection signal that is read for sequencing nucleotides in the real-time sequencing method discussed in class.

Get ready for final exam

At the beginning of this section on NGS it is mentioned that up to 1 billion of short reads can be obtained per instrument run. How is this possible to simultaneously sequence such a large number of fragments?

How many reaction tubes would be necessary to sequence a DNA fragment with 1000bps if NGS were used?

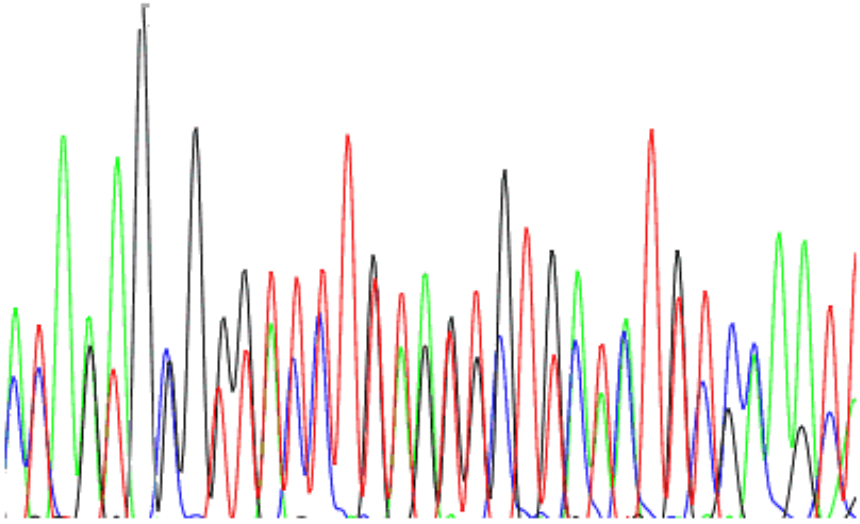
The emulsion PCR technique is used in NGS. Use a diagram to illustrate the underlying principle of this technique and explain how it contributes to the overall efficacy of NGS.

NGS uses reversible dye terminators. How do reversible dye terminators differ from irreversible dye terminators? How can reversible dye terminators improve the efficacy of NGS?

Do all sequencing technologies rely on DNA polymerase? If not, identify which ones (or one) do and which ones (or one) do not.

Maxam Gilbert sequencing is based on mild reaction conditions leading to partial DNA cleavage at specific nucleotide(s). Explain what would happen if harsh reaction conditions leading to complete cleavage were used. Would it still possible to sequence DNA? If not, explain why. If yes, explain how the sequencing results might be affected.

- **Sequencing troubleshooting**: (1) Identify the problem with the sequencing trace – electropherogram – that is provided below and that was obtained through Sanger sequencing. (2) Explain what are the sequencing conditions that may have caused the identified problem. What might happened during the sequencing reaction?



1.2 Bioinformatics or Computational Analysis of DNA Sequencing

The advent of rapid sequencing technology has led to an information explosion that continues unabated today. GenBank integrates data from the major DNA and protein sequence databases and contains publicly available nucleotide sequences for almost 260 000 formally described species (Nucleic Acids Res, 2013). Sequence data accumulate at an exponential rate with nearly a 10-fold increase every three years (see below).

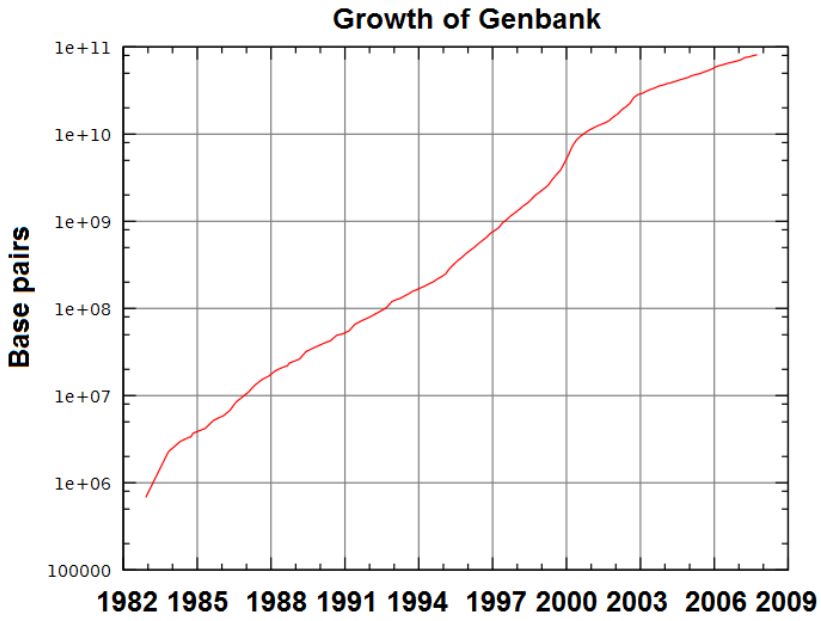


Figure adapted from Wikipedia

	Release 191 (8/2012)	Annual increase (%)[a]
Functional divisions		
Transcriptome shotgun data	5 759 588 580	207.3%
Whole-genome shotgun data	308 196 411 905	47.9%
Patented sequences	12 118 622 726	8.6%
Genome survey sequences	21 947 780 105	5.7%
Expressed sequence tags	40 888 051 100	4.8%
High-throughput genomic	24 359 210 558	0.1%
Sequence tagged sites	636 262 446	0.1%
High-throughput cDNA	639 165 410	−3.5%
All GenBank sequences	451 278 177 138	33.1%

Growth of Genbank divisions (base pairs). Modified from Nucleic Acids Res, 2013.

Database	
Posted date	Oct 12, 2013 4:14 AM
Number of letters	50,694,274,412
Number of sequences	20,064,200
Entrez query	none

Figure from BLAST statistics

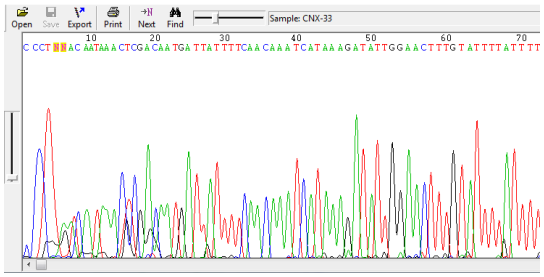
One major goal in bioinformatics is to analyze DNA or protein sequences through computational algorithms to decipher the sequence information found in biomolecules. Remember that according to the Central Dogma in molecular biology biomolecules, mainly DNA, RNA and proteins, are used by living organisms to encode important information in sequences of monomers. This section illustrates, in a very simple way, the process by which computers are used to read and analyze DNA sequencing results.

1.2.1 Reading and Analysis of Sequencing Data Generated through Sanger Method

For conventional Sanger sequencing a raw sequencing file is essentially the flurorescence signal recorded over time at the output of resolving column. The fluorescence trace should be converted into a nucleotide sequence, and this process is referred as the **base-calling process**. There are many programs that can be used for base-calling, including Chromas. This application is definitely not the most powerful one, but it can be freely downloaded and it is easy to use – you won`t have to invest hours to figure out how it works.

Take-home Exercise

Install Chromas on your computer than access to a sequencing file. The sequencing file should appear as belowhe file should be automatically displayed as below. Use the option Edit/Copy Sequence/Plain Text to copy the DNA sequencing results to be used for the BLAST alignment results described below.



Get ready for final exam

Explore the different Chromas options. How long was the sequencing results ? Chromas automatically analyzes the fluorescence trace and reports the DNA sequence. What`s the meaning of an N ? Why do you think the peaks near the 3' end of the trace are smaller (see nucleotide positions 500-600) ? Some nucleotides are identified as N`s, especially near the 3` end. What does an N mean ? While examining the 3`end of the trace adjust the Y-scale by sliding up or down the zooming button at the very left. Could you identify properly some of the nucleotides identified with an N ?

.

1.2.2 Analysis of Sequencing Results through BLAST Alignment

DNA sequencing yields a series of A's, C's, T's and G's, but this crude information needs to be further analyzed to deduce the identity and function of the DNA fragment under assessment. The extraction of biological knowledge out of a given stretch of nucleotides requires the availability of a reference database of information such as Genbank. Bioinformatics, which can be defined as a computer-assisted processing of sequential biological information, doesn`t create any new information, but simply takes profit of genomic databases. For instance, BLAST can associate the sequence gaacctgaggagccccaacaacttcctgtcctactaccgc to the cyclin D1 human gene, which is related to breast cancer, only because this information was previously characterized, annotated and stored into the Genbank database. All a BLAST request does is to simply probe the Genbank database to identify and return the sequences that are the most similar to a submitted query sequence.

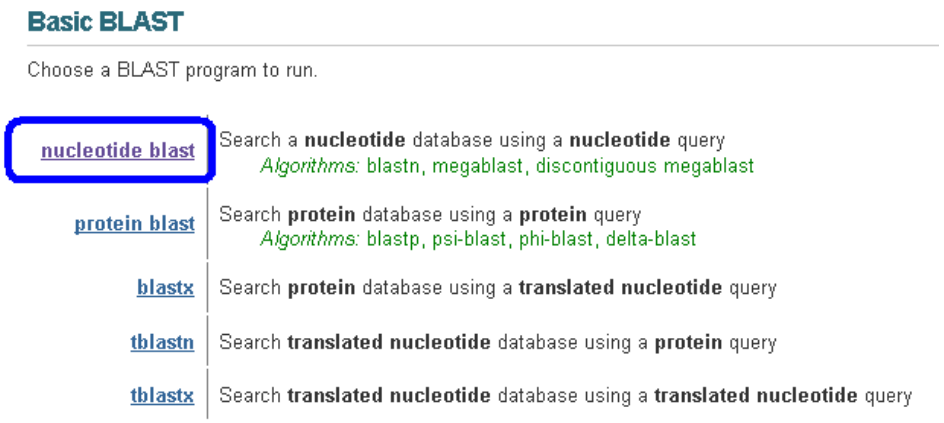
The steps below illustrate the process to perform a BLAST analysis with the DNA sequencing data from the take-home exercise above.

- First highlight and copy the raw sequencing data

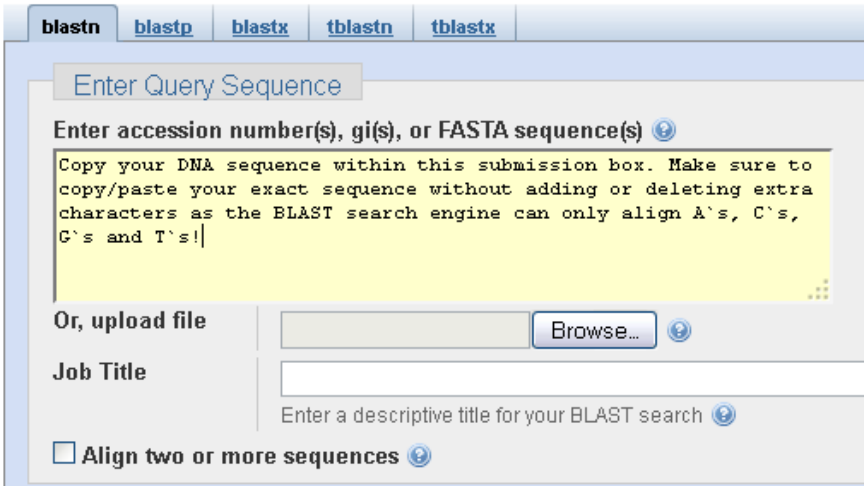
  CCCTNNACAATAAACTCGACAATGATTATTTTCAACAAATCATAAAGATATTGGAACTTTGTATTTTATTTTTGGA
  GCATGAGCCGGAATAGTTGGAACATCTTTAAGAATTTTAATTCGAGCAGAATTAGGTCACCCAGGAGCCTTAAT
  TGGAGATGATCAAATTTATAATGTAATTGTTACAGCTCATGCTTTTATTATAATTTTTTTTATAGTTATACCTATTA
  TAATTGGAGGATTTGGAAATTGATTAGTTCCTTTAATATTGGGAGCCCCAGATATGGCCTTTCCCGAATAAATAA
  TATAAGTTTTTGACTTCTTCCTCCAGCTTTATCCCTTCTTCTAGTCAGTAGTATAGTGGAAAATGGGGCCGGAACA
  GGGTGAACTGTTTACCCTCCCCTATCGTCAGGAATTGCACATGGTGGGGCTTCTGTTGACTTAGCAATTTTCTCTT
  TACATTTAGCCGGAATTCTTCAATTTTAGGGGCTGTAAATTTTATTACAACTGTAATTAATNTACGATCATCAGGA
  ATTACCTTTNAATCGAAATNCCCTTTAATTNGGTATGGGTCNNGNNGGTAAATTAACNNGCNTNNAATNA

TAAACTNNCTAATCCTTTCCCCANGNTC

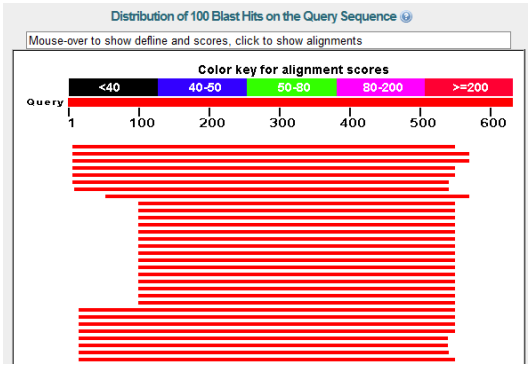
- Access to the BLAST alignment tool from the National Center for Biological Information website – look at the very bottom within the list of popular links.
- Access to the option ``Nucleotide BLAST``.



- Copy/paste the sequencing result within the submission box and submit your request.



- You should obtain a screen similar to the one below. On this graphical display each line represents an alignment hit.



- You can scroll further down to obtain the alignment statistics for each hit. Statistics for the first three hits are displayed below. One can easily access to more biological

information about the best alignment results by clicking on the links. Note that all alignment results refer to the same gene, NADH dehydrogenase, but too different species. The alignment hits are displayed from the most similar to the least similar alignments as quantified by the ``SCORE`` value. The E value is like the P value in statistical tests and it represents the probability that the alignment could be randomly obtained. Smaller E values can be interpreted as statistical evidence that the reported alignment hits are not random alignments, and therefore have true biological value. For the first few alignment hits, BLAST reports E values of 0.0 and this means that the actual values are less than $1x10^{-180}$. The ``Query cover`` values indicate the percentage of nucleotides that are identical between a hit and the full-length query sequence, though the ``Ident`` values are the identity percentage within an aligned area only – an aligned area doesn`t necessarily the full length of a query sequence.

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Drosophila neohypocausta voucher 109402 NADH dehydrogenase subunit 2 (ND2) gene, partial cds; tRNA-Trp, tRNA-Cys, and tRNA-Tyr gene	946	946	85%	0.0	98%	EU493590.1
☐	Drosophila nasuta voucher 103957 NADH dehydrogenase subunit 2 (ND2) gene, partial cds; tRNA-Trp, tRNA-Cys, and tRNA-Tyr genes, comp	933	933	89%	0.0	97%	EU493589.1
☐	Drosophila albomicans voucher 109408 NADH dehydrogenase subunit 2 (ND2) gene, partial cds; tRNA-Trp, tRNA-Cys, and tRNA-Tyr genes,	933	933	89%	0.0	97%	EU493584.1

- Keep scrolling still further down to visualize the alignment results for each of the identified hits. Alignment between the submitted or query sequence and the first hit is shown below.

Drosophila neohypocausta voucher 109402 NADH dehydrogenase subunit 2 (ND2) gene and cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial

Sequence ID: gb|EU493590.1| **Length:** 2018 **Number of Matches:** 1

Range 1: 475 to 1015 GenBank Graphics Next Match Previous

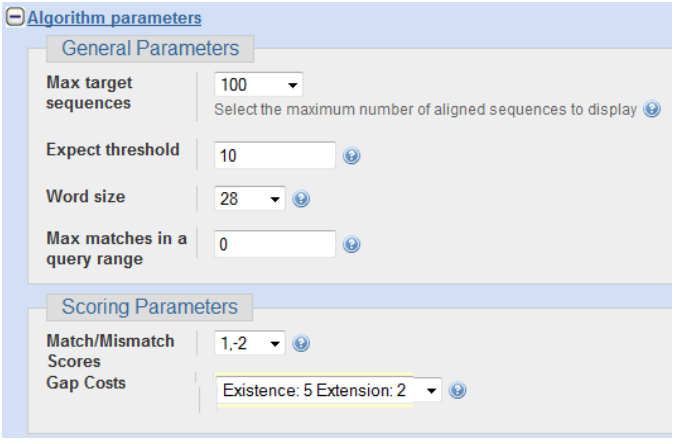
Score	Expect	Identities	Gaps	Strand
946 bits(512)	0.0	532/542(98%)	3/542(0%)	Plus/Plus

```
Query  8     CAATAAACTCGACAATGATTATTTTCAACAAATCATAAAGATATTGGAACTTTGTATTTT  67
             ||||||||                                                    
Sbjct  475   CAATAAACGCGACAATGATTATTTTCAACAAATCATAAAGATATTGGAACTTTGTATTTT  534

Query  68    ATTTTTGGAGCATGAGCCGGAATAGTTGGAACATCTTTAAGAATTTTAATTCGAGCAGAA  127
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  535   ATTTTTGGAGCATGAGCCGGAATAGTTGGAACATCTTTAAGAATTTTAATTCGAGCAGAA  594

Query  128   TTAGGTCACCCAGGAGCCTTAATTGGAGATGATCAAATTTATAATGTAATTGTTACAGCT  187
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  595   TTAGGTCACCCAGGAGCCTTAATTGGAGATGATCAAATTTATAATGTAATTGTTACAGCT  654

Query  188   CATGCTTTTATTATAAtttttttATAGTTATACCTATTATAATTGGAGGATTTGGAAAT  247
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  655   CATGCTTTTATTATAATTTTTTTTATAGTTATACCTATTATAATTGGAGGATTTGGAAAT  714

Query  248   TGATTAGTTCCTTTAATATTGGGAGCCCCAGATATGGCCTTTCC-CGAATAAATAATATA  306
             |||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct  715   TGATTAGTTCCTTTAATATTGGGAGCCCCAGATATGGCCTTTCCTCGAATAAATAATATA  774

Query  307   AGTTTTTGACTTCTTCCTCCAGCTTTATCCCTTCTTCTAGTCAGTAGTATAGTGGAAAAT  366
             ||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  775   AGTTTTTGACTTCTTCCTCCAGCTTTATCCCTTCTTCTTGTCAGTAGTATAGTGGAAAAT  834

Query  367   GGGGCCGGAACAGGGTGAACTGTTTACCCTCCCCTATCGTCAGGAATTGCACATGGTGGG  426
             |||||||||||||||||||||||||||||| ||||||||||||||||||||||||||||
Sbjct  835   GGGGCCGGAACAGGGTGAACTGTTTACCCTCCTCTATCGTCAGGAATTGCACATGGTGGG  894

Query  427   GCTTCTGTTGACTTAGCAATTTTCTCTTTACATTTAGCCGGAA-TTCTTCAATTTTAGGG  485
             |||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct  895   GCTTCTGTTGACTTAGCAATTTTCTCTTTACATTTAGCCGGAATTTCTTCAATTTTAGGA  954

Query  486   GCTGTAAATTTTATTACAACTGTAATTAATNTACGATCATCAGGAATTACCTTTNAATCG  545
             |||||||||||||||||||||||||||||| |||||||||||||||||| |||  ||||
Sbjct  955   GCTGTAAATTTTATTACAACTGTAATTAATATACGATCATCAGGAATTAC-TTTAGATCG  1013

Query  546   AA  547
             ||
Sbjct  1014  AA  1015
```

- Return to the very top of the alignment results and click ``Distance tree of results``. This option takes profit of all alignment hits to analyze their relative sequence homology and construct their phylogenetic tree.

Get ready for final exam

BLAST alignment can be personalized by modifying general and scoring parameters. Explain the meaning of each of six parameters indicated below. Also indicate the expected effect on the number of alignment hits for (1) an increase and (2) a decrease of each of the parameters.



1.2.3 Underlying Principles of BLAST Alignment

As illustrated in the previous section, sequence analysis can be used to identify and understand the biological functions and evolutionary origins of DNA sequences – a similar approach is also possible with sequences of amino acids. Not only DNA alignment tools can reveal useful biological information, but this can be achieved almost instantaneously. The underlying principles of two alignment processes will be discussed in the present section.

The first one was initially proposed in 1970 by Needleman and Wunsch (J Mol Biol, 1970), though this multistep process was later put aside as it didn`t allow for high throughput analysis of long sequences. Needleman and Wunsch`s alignment process is relatively tedious as it involves comparing every nucleotide with every other. Another alternative is BLAST, which stands for Basic Local Alignment Search Tools. BLAST uses a simpler alignment algorithm with shortcuts making it possible to process quickly larger sequences. BLAST, which was designed by Altschul et al. (J Mol Biol, 1990). Details of these two alignment algorithms are explained in appendix I.

Get ready for final exam

Refer to the Needleman-Wunsch algorithm to fill the alignment matrix provided below using matching scores of +1 for a match, -1 for a mismatch, and -1 for a gap (these are the default alignment score values initially used by Needleman and Wunsch).

	H	E	L	L	O
H					
O					
L					
A					

The BLAST alignment algorith makes use of seeding ‘words’. What is the specific meaning of a seeding ‘word’ ? Also explain how a seeding ‘word’ contribute to a higher alignment performance compare to the Needleman-Wunsch algorithm.

.a

Appendix I : Underlying Principle of Two Alignment Algorithms

The present appendix corresponds to a Nature Education paper written by Ingrid Lobo (Nature Education, 2008)

Basic Local Alignment Search Tool (BLAST)

Awash in a sea of data, how do scientists identify the function of a newly cloned gene? Online resources like the Basic Local Alignment Search Tool (BLAST) provide a helping hand.

Since the discovery of the genetic code, biological research has undergone a sea change in the way it is performed. Until the early twentieth century, biology focused on the processes of living organisms and almost always involved experiments in laboratories and in the field. The growth of molecular biology in the twentieth century moved research into the test tube, where biological systems could be painstakingly dissected and reassembled. Then, beginning in the 1970s, scientists began to accumulate DNA and protein sequence data at an exponential rate; in fact, researchers currently have approximately 97 billion bases sequenced and over 93 million records. Amazingly, this sequence data doubles every 18 months!

But how do investigators search through, organize, and make sense of this massive amount of data? And how can they identify the functions of newly cloned genes? Is it possible to estimate the evolutionary relationships between genes or proteins just by examining their nucleotide or amino acid sequences? The answer to this question is yes. The relationships between organisms can be teased out as different species are connected via descent from a common ancestor. Thus, sequence similarity can be helpful in inferring function and evolutionary relationships. One common way to examine a new gene is to search for similarities between newly sequenced DNA and databases of gene sequences that have already been described. By identifying a related gene or gene family with a known function, scientists can infer the function and evolutionary relationships of newly cloned genes or even whole genomes. If genes have similar sequence regions, then the genes may share similar functions.

As gene and protein sequence databases grew at the end of the twentieth century, scientists turned to computers to help analyze the abundant and ever-growing amounts of data. Today, one of the most commonly used tools to examine DNA and protein sequences is the Basic Local Alignment Search Tool, also known as BLAST (Altschul *et al*., 1990). BLAST is a computer algorithm that is available for use online at the National Center for Biotechnology Information (NCBI) website and many other sites. BLAST can rapidly align and compare a query DNA sequence with a database of sequences, making it a critical tool to ongoing genomic research. In fact, the initial paper describing the program, titled "Basic Local Alignment Search Tool" and published in the *Journal of Molecular Biology*, was the most highly cited publication of the 1990s (Taubs, 2000). The parallel development of large-scale sequence projects and bioinformatic tools like BLAST has enabled scientists to study the genetic blueprint of life across many species and has helped bridge the gap between biology and computer science in the maturing field of bioinformatics.

Alignment Theory

While the computer science principles behind BLAST have been around for some time, prior to BLAST, they had not been applied to biology. Before BLAST, alignment programs used dynamic programming algorithms, such as the Needleman-Wunsch and Smith-Waterman algorithms, that required long processing times and the use of a supercomputer or parallel computer processors (Collins & Coulson, 1984; Gotoh & Tagashira, 1986; Smith & Waterman, 1981).

Figure 1A depicts a Needleman-Wunsch alignment of the words "PELICAN" and "COELACANTH." The search space of the alignment is shown using a Cartesian grid and is proportional to the length of the sequences being compared plus one extra row and column.

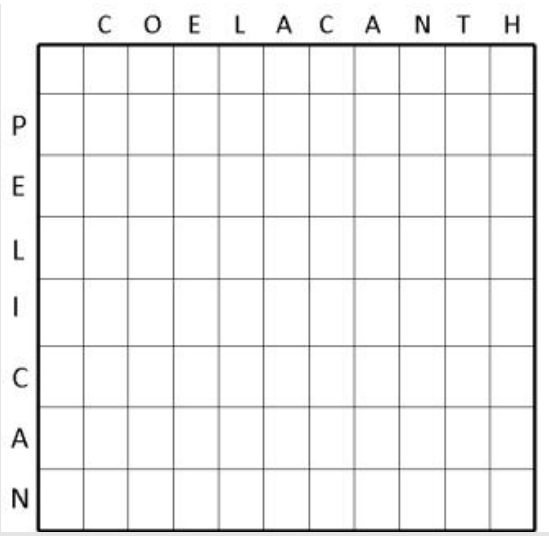


Figure 1A: Initialization of the alignment matrix.

Next, the alignment matrix is initialized with a zero in the upper left corner. For each letter of the word being aligned, a point is deducted so that each letter has a progressively more negative score. Why does the algorithm subtract a point? In an alignment, the diagonal is read from the upper left to the lower right, and when the analysis moves vertically or horizontally, it indicates a gap in the sequence. Thus, each time the program moves straight up or down, a gap penalty is applied that takes away points from the alignment score. Finally, a little arrow, or pointer, is added to indicate which direction to follow the alignment (Figure 1B).

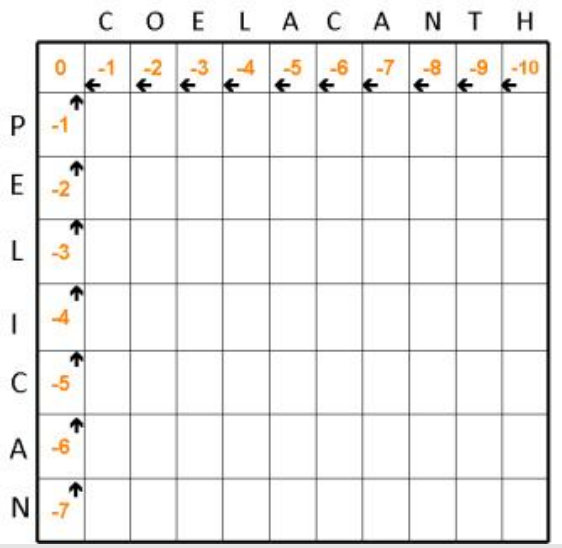


Figure 1B: Filling the axes of the alignment matrix.

When filling the axes of the alignment matrix, start in the upper left corner and set it to 0. Next, assign a score for each letter in the row or column. Note that there is a penalty for gaps, and that the arrow should point toward the origin of the alignment.

In the third stage, the algorithm starts to actually build and score the alignment in a step called fill or induction. In this example, the analysis begins by aligning the C to the P and calculating a score. In Figure 1C, one point is added if two letters match, and one point is subtracted if they do not. This calculation is carried out three times, once for each square to the left (dark blue), above (green), and upper left (brown). Using a value from either the upper or left square, the final score is -2 (-1 + -1). Using the 0 score in the upper left diagonal square, the final score is -1 (0 + -1). Because -1 is the highest score, this score is jotted down in the alignment matrix, and because the upper left square was the one leading to the best score, an arrow is inserted in the box pointing toward this square (light blue, Figure 1C).

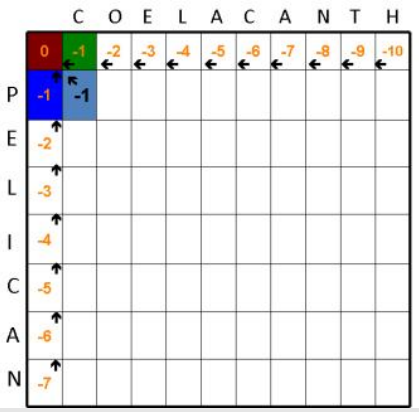


Figure 1C: Induction or filling of the alignment matrix, part I.

One point is added if two letters match, and one point is subtracted if they do not. Using a value from either the upper (green) or left (dark blue) square, the final score is -2; however, using the value from the upper left (brown) square, the final score is -1. Because this is the highest score, it is recorded in the alignment matrix along with an arrow pointing to the upper left square.

This same process continues, calculating two scores for every square in the matrix (Figures 1D and 1E).

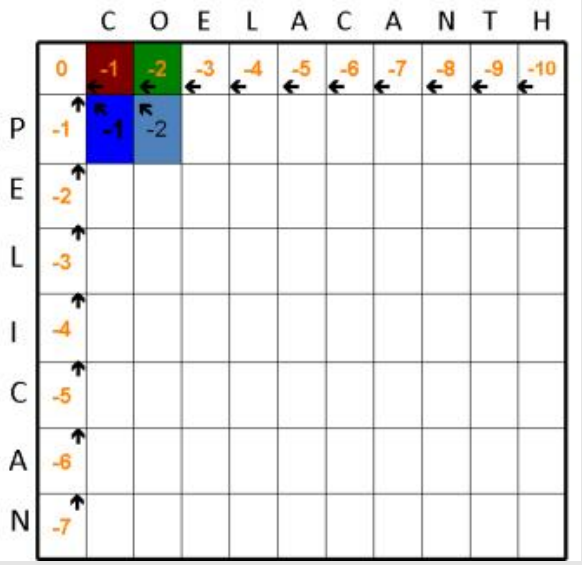


Figure 1D: Induction or filling in of the alignment matrix, part II.

The same process is carried out for the next square in the alignment. Here, using the value in upper left (brown) square yields a sum of -2, using the value in the upper (green) square yields a sum of -3, and using the value in the left (dark blue) square yields a sum of -2. Because -2 is the highest score and was initially calculated using the upper left square, -2 is recorded in the matrix along with an arrow pointing toward the brown square.

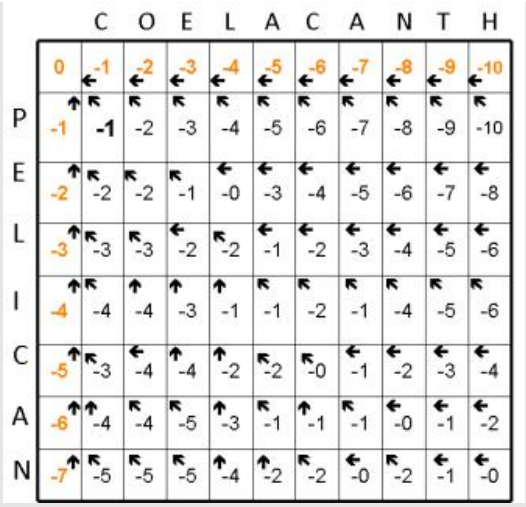


Figure 1E: Induction or filling in of the alignment matrix, part III.

The rest of the matrix is completed using the same method.

Once the matrix is completed, the optimal alignment is found through a process called traceback. The traceback starts in the lower right of the matrix and follows the pointers to adjacent boxes. By definition, traceback involves determining the best scoring path through the alignment (Figure 1F).

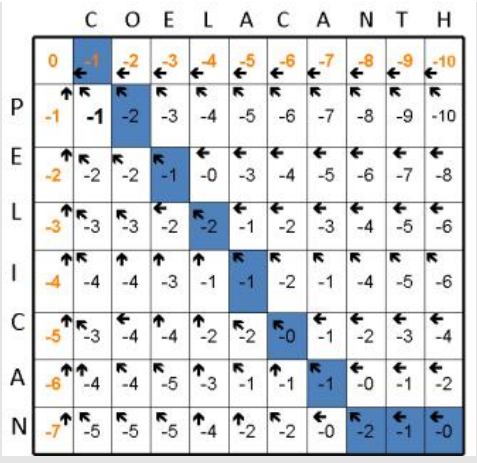


Figure 1F: Traceback of the optimal complete alignment.

During traceback, following the best scoring path reveals the best alignment.

Although this sort of dynamic programming did a complete job of comparing every single residue of one sequence to every single residue of a second sequence and kept track of how well the sequences aligned at every step, these algorithms required a considerable amount of computer memory and processing time. Computing speed was an especially important concern, because these exhaustive programs had to search databases that continued to grow at exponential rates. Moreover, most regions of the search space did not score very well and therefore probably could have been skipped during the

calculation process. Finally, these programs required powerful computing hardware that was expensive, rare, and ultimately impractical for most scientists and labs.

Researcher Stephen Altschul and colleagues wanted to bypass these challenges and develop a way for databases to be searched quickly on routinely used computers. In order to increase the speed of alignment, the BLAST algorithm was designed to approximate the results of an alignment algorithm created by Smith and Waterman (1981), but to do so without comparing every residue against every other (Altschul *et al*., 1990). BLAST is therefore heuristic in nature, meaning it has "smart shortcuts" that allow it to run more quickly (Madden, 2005). However, in this trade-off for increased speed, the accuracy of the algorithm is slightly decreased.

The BLAST Heuristic

BLAST increases the speed of alignment by decreasing the search space or number of comparisons it makes. Instead of comparing every residue against every other, BLAST uses short "word" (*w*) segments to create alignment "seeds." BLAST is designed to create a word list from the query sequence with words of a specific length, as defined by the user (Figure 2). Requiring three residues to match in order to seed an alignment means that fewer sequence regions need to be compared. Larger word sizes usually mean that there are even fewer regions to evaluate (Figure 3A versus Figure 3B). Once an alignment is seeded, BLAST extends the alignment according to a threshold (*T*) that is set by the user. When performing a BLAST query, the computer extends words with a neighborhood score greater than *T* (Figure 3C). A cutoff score (*S*) is used to select alignments over the cutoff, which means the sequences share significant homologies. If a hit is detected, then the algorithm checks whether *w* is contained within a longer aligned segment pair that has a cutoff score greater than or equal to *S* (Altschul *et al*., 1990). When an alignment score starts to decrease past a lower threshold score (*X*), the alignment is terminated (Figure 3C). These and many other variables can be adjusted to either increase the speed of the algorithm or emphasize its sensitivity.

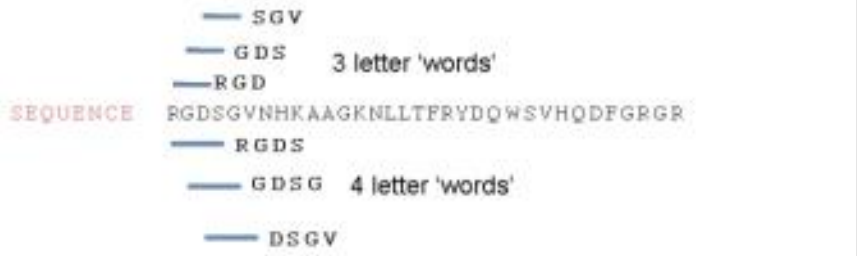


Figure 2: Generating word seeds.

Instead of comparing every residue against each other, BLAST uses short “word” (w) segments to create alignment "seeds." BLAST is designed to create a word list from the query sequence with words of a specific length, as defined by the user.

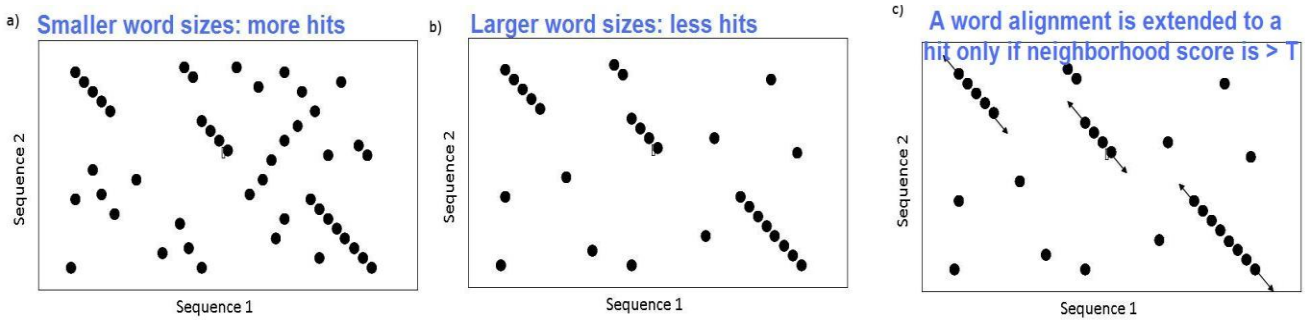


Figure 3 Longer words result into smaller seeding hits (b compare to a), but a word alignment is extended to an alignment hit only if the neighboring residues provide a score superior to the set threshold T. In BLAST, the T value is indirectly determined via the expect value (E) with is a measure of the tolerance to the risk of alignment error. An expect value of 10, for instance, means that there is a 10% risk that each hit represents a random match. Lower is the E value is higher is the likelihood that the alignment indicates a biologically relevant alignment reflecting related evolution.

Testing the BLAST Algorithm

Altschul and colleagues tested the BLAST algorithm on a database of randomly generated sequences, and they examined the output resulting from different *w* and *T* parameters. If *T* is set to be a lower threshold, then the algorithm detects more word pairs and requires a longer processing time (Altschul *et al*., 1990). Choosing the value for *T* was a major decision because the researchers wanted to reach a compromise between the algorithm's sensitivity and its processing time.

Next, Altschul and colleagues tested BLAST on a database of real sequences, and they found it was successful in quickly identifying alignments with high scores. In searching the globin gene family, for example, they found that BLAST identified 88 of the 89 globin alignments that scored above 80. Other gene families, including the immunoglobulins, protein kinases, and cytochrome c genes, were then examined to measure the number of alignments detected when using different *T* and *S* values. BLAST was also able to detect similar regions within pairs of long sequences. These tests showed that BLAST was fast, sensitive, and accurate as a tool for analyzing sequence alignments (Altschul *et al*., 1990).

Bringing Mathematical Rigor to Alignment

One of the most notable innovations of BLAST is that the program calculates the statistical significance for each sequence alignment result. This is known as the expect value (E-value) or probability value (P-value), and it is calculated for each alignment. The E-value describes how many hits you can expect to see by chance when searching a database of a certain size, whereas the P-value describes the probability that the alignment you are observing is due to chance. In general, the lower the E- or P-value is, the more likely it is that an alignment is significant. Below the common 10^{-5} score, P and E are roughly equivalent (Madden, 2005).

The addition of statistical rigor to sequence alignment has been controversial. Some researchers rely too much on significance values to include or exclude sequences despite poorly chosen parameters, while others overinterpret "insignificant" results because the results "look" right. While all scientific results are subject to interpretation, BLAST scores and statistics bring much-needed objectivity to sequence comparisons, and the debate about them has helped improve methods for determining significance.

The BLAST Family

Since 1990, many variants of BLAST have been developed, each with specialized features. Early on, the original BLAST was split into two adaptations: NCBI BLAST and Washington University BLAST (WU BLAST). Both BLASTs have program variations. For instance, BLASTN can be used to compare a nucleotide sequence with a nucleotide database; BLASTP can be used to compare a protein sequence with a database of protein sequences; and BLASTX can take a nucleotide sequence, translate it, and query it versus a protein database in one step (Gish & States, 1993). TBLASTN compares a protein query sequence to all six possible reading frames of a database and is often used to identify proteins in new, undescribed genomes. Finally, TBLASTX compares all six reading frames of a query sequence to all six reading frames of a database—an intensive algorithmic feat that can bring even modern computers to a grinding halt if not used properly.

In addition, NCBI has some of its own specialized variants of BLAST. For example, MEGABLAST is a program that can rapidly complete searches for sequences with only minor variations and can more efficiently manage queries with longer sequences (Altschul *et al*., 1994). PSI- and PHI- are other powerful BLAST tools that allow more complex and evolutionary divergent proteins to be aligned (Altschul *et al*., 1997). These and other programs, as well as genomic BLAST databases, are all available on the NCBI BLAST website.

Since its creation, BLAST has become an essential bioinformatics tool for biologists. Its speed and sensitivity allow scientists to compare both nucleotide and protein sequences to single sequences and to large databases. Most importantly, BLAST has helped democratize bioinformatics analysis and make it accessible to any researcher over the Internet. It is rare to read a modern molecular biological paper that does not refer to a BLAST alignment, and this information has permitted scientists to predict the functions of genes and proteins in whole genomes, answering questions *in silico* that could never be answered at a bench or in the field