CS123A Midterm #2 Study Guide

By: Zayd Hammoudeh

The Big Jaw

Big Jaw – Constraint that inhibited brain growth

Powerful masticatory muscles: Found in most primates including chimpanzees and gorillas

Human Masticatory Muscles – Much smaller compared to other animals in the Homo genus.

Myosin Heavy Chain (MYH): Gene expressed in masticatory muscles. **Inactivated in humans** by a frameshift mutation after the lineage of humans and chimpanzees diverged.

Mutation **removed a barrier for the remodelling of the hominid cranium** which consequently **allowed for an increase in the size of the brain**.

Viruses

Virus – Small living particles that can infect cells and change how the cells function. The effect on the cell's function depends on the type of virus and the cells that are infected. **Surrounded by protein case.**

Retrovirus – Single stranded RNA virus that employs a double standed DNA (dsDNA) intermediate for replication.

Pathogen: A disease product. It can include both infectious organisms (bacteria, fungi, etc.) as well as viruses.

Virulence: Ability of an infectious agent (i.e. pathogen) to cause a disease.

Many viruses are virulent sometimes and asymptomatic at other times.

Immunodeficiency – The result when the immune system is unable to protect the host from disease causing agents or from malignant cells.

Acquired Immunodeficiency: Loss of immune function because the genetic or development deficiency was not acquired at birth. It results from exposure to various agents.

Virus – A single stranded RNA virus that employs a **double stranded DNA** (**dsDNA**) intermediate for replication.

Reverse Transcriptase: Turns viral RNA into DNA. It turns the RNA strand into DNA. It then uses the DNA to make it complementary strand.

cDNA – Complementary DNA made from mRNA by reverse transcriptase.

Capsid – Surrounds mRNA in virus particle. It's the outer protein shell.

Viral DNA is integrated into the DNA of the host cell.

Virion – Entire virus particle including the capsid (protein shell) and the inner core of nucleic acid.

HIV

HIV – Human Immunodeficiency Virus

Type of retrovirus.

Inherited from:

- **Chimpanzees**
- **Mangabyes**

Transmitted through **bodily fluids** (e.g. blood, semen) when the virus of an infected individual contacts the **mucous membrane** or **enters the blood stream** of an uninfected individual.

Cells Affected by HIV:

- **Macrophages** - Large immune cells that devours invading pathogens and other intruders. Stimulates other immune cells by presenting them small pieces of the invaders.
- **CD4+ T Cells (aka T-Helper Cells)** – White blood cells that **orchestrate the immune response**. They signal other cells to perform their special functions.

Lentivirus – "**Slow viruses**" where the period between initial infection and the onset of serious symptoms is long.

Other Lentiviruses:

- **FIV** – Feline Immunodeficiency Virus
- **SIV** – Simian Immunodeficiency Virus (Infects monkeys and nonhuman primtates)

Body's Immune Response to HIV

Destroys the **virions** floating in the bloodstream before they can infect new cells.

Destroys the infected CD4 helper T-cells depleting the body's ability to fight disease. This causes an immune system collapse. This leads to **AIDS** (**Acquired immunodeficiency Syndrome**).

Three Types of Proteins Involved in Viral (Virion) Replication

- **GAG** – Encodes for **core proteins** and structural virion components.
- **POL** – Encodes for **reverse transcriptase, integrase**, and **protease**.
- **Env** – Encodes for the **structural protein components** that surrounds the virus. Needed for the virus to leave the cell.

Miscellaneous HIV Notes

Many subgroups of HIV-1 exist.

Within a single subtype and in a single infected person, the virus changes constantly.

Transmission from chimp to humans happened multiple times in the past.

Sanger Sequencing

Developed by Frederick Sanger in **1977**. Most widely used technology for ~25 years.

Replaced by "Next Generation Sequencing" techniques.

Still widely used for **smaller scale projects** and for **long contiguous DNA sequences** (>500 nucleotides).

Dideoxynucleotides – Nucleotides where the OH molecule on the 3' carbon of the sugar is modified to simply an –H making a subsequent phosphodiester linkage impossible.

These are floating in the gel and sometimes DNA polymerase selects a normal nucleotide and other times the **dideoxy analog** which terminate the sequence. **This sugar can fluoresce**.

DNA polymerase makes a complement of a partial sequence within a DNA molecule. **Synthesis is primed from a chemically synthesized fragment** (i.e. **primer**) that is complementary to a part of the DNA sequence known from other studies.

DNA polymerase **builds strand from 5' to 3'**.

Pyrosequencing

Developed in **1996**.

Based off the detecting of released **pyrophosphate** during DNA synthesis. This detection is through the **detection of light**.

Sequences a single strand of DNA by synthesizing the strand's complement.

Benefits of Pyrosequencing:

- "**Sequencing by synthesis**"
- Accurate
- Simple and robust
- No labels or gels
- Real time results.

Nucleotides are dispensed sequentially and then removed from the reaction (this is done by **apyrase**). Light is only produced when the solution complements the first unpaired base of the template strand.

A mini strand with a **magnetic bead** for DNA polymerase serves as a **primer**.

Example Pryosequencing Instruments

PSQ96

- 500 samples per hour
- 4500 samples per day.
- Includes CCD camera.

PSQHS96A

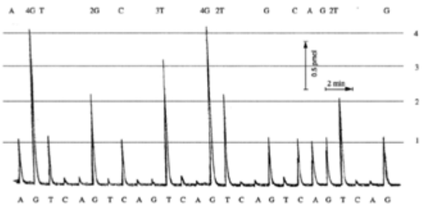
- 10,000 samples per day
- 30,000 samples per day with triplex analysis.

Procedure:

1. Prepare samples
2. Insert samples in PSQ96
3. Insert **reagent cartridge**
4. Start run.

96 represents the number of wells.

Reading a **Pyrogram**



Bases released sequentially. Depending on intensity of light you can determine the number of sequential bases.

Single Nucleotide Polymorphism

- Occurs every 500 to 1000 bases in DNA.
- Most common cause of inter-individual variation.

HPV – Human Papillomavirus

- Sexually transmitted infection (STI)
- Usually does not cause health problems but can cause cancer of the vulva, vagina, penis, and anus as well as in the back of throat.
- Different primers used to detect the specific strain of HPV infection.

Wild Type – Original, non-mutated version of a gene. For bacteria, it would be the original non-drug resistant version.

Pyrosequencing begins with a primer that binds to the DNA sequence. Primer has a "**general primer site**."

Primers

Primer Design: Required step before beginning pyrosequencing. This includes running PCR (polymerase chain reaction).

General Primer – Will anneal with all alleles.

Have a **magnetic bead** at **5' end**.

Polymerase Chain Reaction

Used to amplify a specific DNA sequence. Exponentially increases number of copies of a DNA sequence.

Step #1: Denaturing – Heating the DNA sequence to render it single stranded (Double helix is also removed). Example time: 1 minute at 94C.
Step #2: Annealing – Forward and reverse primers bind to the appropriate complementary strands.
Step #3: Extension – DNA polymerase extends the primers.

These three steps are repeated **30-40 times**. First couple of PCR copying does not actually created double stranded DNA of the right gene. This is eventually achieved through the primer.

Primer Characteristics

- Lack of **secondary priming sites** (uniqueness)
- **Absence of hairpin** formation (bends in single strand). Caused by intermolecular interaction within the primer.

Uniqueness: There should be only one place the primer can bind in the template DNA. There should be no possible contaminant binding sites either (e.g. from other animals such as human, rat, mouse, etc.)

Length – Related to uniqueness and melting/annealing temperatures. **The longer the primer, the more likely to be unique and the higher the melting/annealing temperatures.**

Minimum Length: 15 bases
Ideal Length: 17-28 bases

Base Composition – **Random base composition is best**. Best to avoid long A/T and G/C chains.

50-60% G+C content leads to the right annealing/melting temperatures.

Melting Temperature

Temperature at which half the DNA strands are single stranded and half are double stranded.

More G/C nucleotides in a strand means higher melting temperature since more hydrogen bonds.

Notation: T_m
Target: 52C to 65C

Annealing Temperature

Temperature at the primer anneals (bonds) to the DNA stand.

Calculated as:

$$T_{anneal} = T_{m_{primer}} - 4^oC$$

Internal Structure

Primers can anneal to themselves or to other primers.

Hairpin: Primer bending back to **bind to itself**.
Self Dimer: Primer bonding to **another of the same primer**
Dimer: Primer binding to a **complementary strand primer**.

Stability at 5' end of the primer is critical.

Primers **work in pairs**. Two types of primers:

- **Forward Primer**
- **Reverse Primer**

Annealing temperatures of the two strands must be compatible (maximum 3 degrees) of each other.

Primer Design

Generally done best by computers.

Example primer design tools: **Primer3** (tool used in class from MIT), **BioTools**, **GCG**, **Oligo**

Adjustable Features in Primer Design Tools:

- **Primer Length**
- **Melting Temperature**
- **(G+C)%**

Forward and Reverse Primer

```
                               Reverse (Right) Primer
                              3'<-------GGAA---- 5'
                  Plus Strand           ||||
5'-----ATCG--------=========------------CCTT---- 3'
       ||||           Target            ||||
3'-----TAGC--------=========------------GGAA---- 5'
       ||||       Minus Strand
   5'--ATCG----> 3'
Forward (Left) Primer
```

Multiplex PCR
Multiple primer pairs are added together in PCR. **This allows for the amplifying of multiple sites.**

Challenges of this Approach:

- Different melting temperatures.
- Ensuring no **dimer** formation

Most primers are designed to **amplify a single product.**

Universal Primer: A single primer that can be used to amplify multiple products.

Semi-universal Primer: A single primer can be used to amplify a subset of template sequences from a large group of similar sequences.

Guessmer
In some cases, DNA sequences are unavailable or difficult to align. Example: **Back-translated a protein** which is degenerate so the nucleotide sequence cannot be known.

Procedural Differences in Primer Design:

- Length: Primer should be **longer than normal at about 30 bases** to offset decreased hybridization.
- Set higher annealing temperature to **increase primer annealing stringency**.

Homework #3 Keywords

Hydrophobic Amino Acid – Does not react with water. Tends to be **buried within** a protein surrounded by other hydrophobic amino acids.

Polar, hydrophilic Amino Acid – Typically surrounded by water molecules. Generally **buried inside protein surrounded** by other oppositely charged hydrophilic amino acids.

Protein Domain – Conserved part of protein where the protein has folded into these discrete structural units. Core of each domain is α-helixes, β-sheets, or a mixture of both.

Tetramer – Protein that consists of **four monomer subunits**. Example: **Hemoglobin**.

SQL – Structured Query Language
HTML – Hypertext Markup Language
XHTML – Extended Hypertext Markup Language

Flat File Database – Rely on basic text files to store information. Still in use due to their simplicity and relatively low cost.

Annotation – Non-sequence information in a database entry. It can include **interpretation of data**, **relevant research citations**, and **related entries in other databases**.

Non-redundant database – Database with no duplicate entries.

GI Number – Unique identifier given to a sequence. If a sequence ever changes, a new GI is assigned.

Accession Number – Unique number assigned to a database entry, and it never changes.

Ontology – Set of field-specific descriptors that enable the sharing of the same concepts and definitions for specific terms.

Gene Ontology – Collaborative project that provides a controlled vocabulary that describes genes and gene-associated information.

Hypothetical Gene – A gene identified in a protein sequence purely through computational methods. No experimental data supports this as a gene. May be correct or incorrect.

Swiss-Prot – Protein database that does not use computer based annotation. All curating is done manually by experts.

Three Types of Nucleotide Sequences in Databases

Raw Genomic Sequence – Represents chromosomal DNA and includes non-coding regions (e.g. introns, control regions, UTR) as well as coding regions (i.e. exons).

cDNA (Complementary DNA) – Result **of reverse transcription from RNA to DNA**. Does not include anything beyond the coding sequence.

Expressed Sequence Tag (EST) – **A partial cDNA sequence** that is around 300 bases in length.

Pseudogene – Sequences in genomic DNA that is **similar to known coding genes but do not produce functional proteins**. Mutate faster than normal genes since no longer under selection pressure. Possibly due to gene duplication. Up to 20,000 pseudogenes in human genome.

Window Size – Number of consecutive sequence objects used for comparison in a dot plot.

Stringency – Number of exact matches in a dot plot window that must be identical to be considered a match.

Protein versus Nucleotide Sequence Comparison

- Proteins have more 20 amino acids versus only 4 nucleotide bases. Hence, **one character has more information in a protein**.
- Genetic code is redundant so **insignificant variations are filtered out when looking at proteins**.
- Structure and function of protein is entirely dependent on amino acid sequence so **amino acid sequences tend to change less over time**.

RCSB Protein Database – Tool used to view protein structures.

Gene Locus Format: {Chromosome #}.{P/Q}{Chromosome Region} **Example:** 17.p21 is on the P-arm of chromosome 17 in region 21.

Nonsense Mutation– A point mutation in a DNA sequence that results in a **premature stop codon** (or **nonsense codon**). The mRNA is **truncated**, **incomplete**, and usually **non-functional**.

Missense Mutation – A point mutation in which a single nucleotide change results in a codon that codes for a different amino acid.

Homework #4 Keywords

Purine: Adenine and Guanine – Two carbon rings (one pentagonal and one hexagonal)
Pyrimidine: Cytosine and Thymine – Single hexagonal ring.

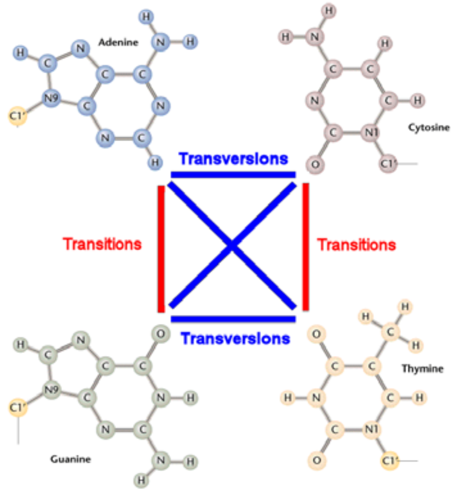
Transition: Nucleotide substitution of:

- Purine to Purine
- Pyrimidine to Pyrimidine

Transversion: Nucleotide substitution of:

- Purine to Pyrimidine
- Pyrimidine to Purine

Transitions are more common than transversions since transitions tend to have less effect on the protein sequence.



Silent Mutation – Point mutation that does not affect the amino acid sequence of a protein. Can **occur in both coding and non-coding regions**.

Synonymous Mutation: Point mutation that does not affect the amino acid sequence of the gene. Limited to those **substitutions in the coding region** (i.e. exons).

"misc_feature"– Notation in a Genbank record to indicate a noteworthy feature in a sequence. **Example: "polymorphic (TAAA)n"** is a repeat sequence of "TAAA" multiple times.

Missense Mutation – A point mutation in which a single nucleotide change results in a codon that codes for a different amino acid.

Needleman-Wunsch – **Global sequence alignment** dynamic programming algorithm.

Most **rigorous** and **guaranteed to return the best alignment**.

Smith-Waterman – **Local sequence alignment** dynamic programming algorithm.

Most **rigorous** and **guaranteed to return the best alignment**.

SSEARCH – Alignment program built off of Smith-Waterman.

Twilight Zone – Region **between 20% and 30% identity in amino acid** sequences. Homology may exist between the proteins but cannot be reliably assumed in the absence of other experimental data.

Midnight Zone – Region where there is **less than 20% identity** between two amino acid sequences. Very unlikely the two sequences are homologous.

Gap Penalty – Penalty that is subtracted from the alignment score anytime a gap is inserted in the alignment.

Gap Extension Penalty – Penalty to extend an existing gap. This penalty is generally less than when a new gap is inserted.

Gap Penalty Differences Based on Amino Acids – Some amino acids tend to be more important for a protein's protein (e.g. **Tryptophan**). Hence, depending on the amino acid, the gap penalty may vary.

FASTA and **BLAST:** Heuristic based sequence alignment programs. Use **heuristic to filter possible matches then runs dynamic program on those that passed the heuristic test**.

FASTA is a fast database-search method based on **matching short identical sequences**.

BLAST based on finding **very similar short segments**.

Phenotype: A composite of an organism's observable characteristics or traits. It is related to but not entirely dependent on the genotype as two organisms may have the same genotype but different physical characteristics due to the environmental factors.

Using Different Substitution Matrices – A single substitution score matrix is not ideal for all cases. Differences can include:

- Degree of similarity between the sequences
- Looking for closely related sequences or very distantly related evolution relationships

Blastp – Compares a query protein sequence against a protein database.

Blastn – Compares a query nucleotide sequence against a nucleotide database.

Blastx – Compares a nucleotide sequence translated into all six reading frames against a protein database.

E-Value – Expectation value. Statistical measure for estimating the significance of alignments.

The smaller the E-value, the more likely that the two sequences are homologous.

Closely related sequences have an E-value of **less than 10^{-20}**.

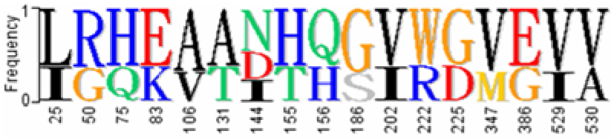
Low Complexity Region – Sequence segments (both nucleotide and amino acid) that have only a few types of bases or amino acids.

Often removed from protein sequences before a database search as they can lead to misleading hits.

Motif – A conserved element of a sequence alignment.

Constructed by the "**consensus method**" where **multiple sequences are aligned** and the **most conserved regions are used to construct a pattern**.

Logo – Visual representation of a set of aligned sequences. For each position in the sequence, a letter or set of letter is shown with larger letters indicating more conservation.



ClustalW and ClustalOmega

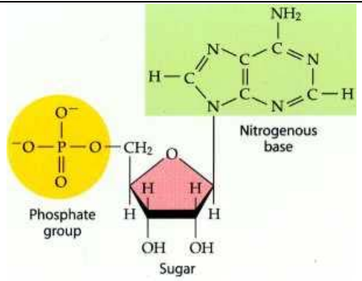
Tools used for Multiple Sequence alignment.

Can illustrate both transitions and transversions.

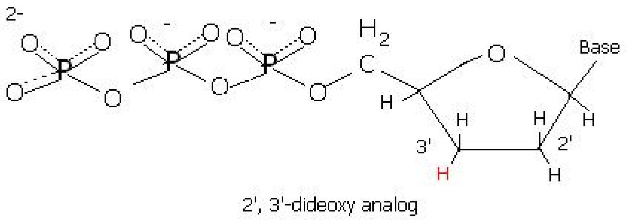
Indels (insertions and deletions) are indicated with a "-".

Example CLUSTAL OMEGA Output

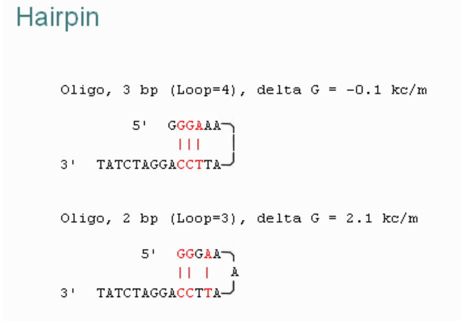
```
gi|37222316|gb|AY350716.1|  ACAAGCTGATGACCACCCTCCACAGCACTGCACCCCATTTTGTCCGCTGT
gi|37222318|gb|AY350717.1|  ACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222320|gb|AY350718.1|  ACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222322|gb|AY350719.1|  ACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222324|gb|AY350720.1|  ACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222326|gb|AY350721.1|  ACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|37222328|gb|AY350722.1|  ACAAGCTGATGACCACCCTCCATAGCACCGCACCCCATTTTGTCCGCTGT
gi|45752610:36298-36383     ACAAGCTGATGACCACCCTCCATAG--CCGCACCCCATTTTGTCCGCTGT
                            ********************** **   * *********************
```



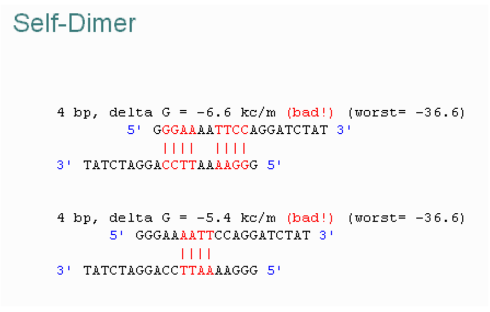
RNA Nucleotide: Note –OH molecule on 3' carbon in RNA not –H as in DNA.



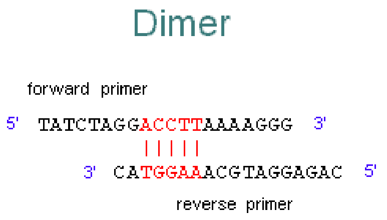
-OH on 3' carbon is changed to only a –H.


Hairpin

```
Oligo, 3 bp (Loop=4), delta G = -0.1 kc/m

        5'   GGGAAA┐
               |||  │
3'   TATCTAGGACCTTA┘

Oligo, 2 bp (Loop=3), delta G = 2.1 kc/m

         5'   GGGAA┐
              || |   A
3'   TATCTAGGACCTTA┘
```

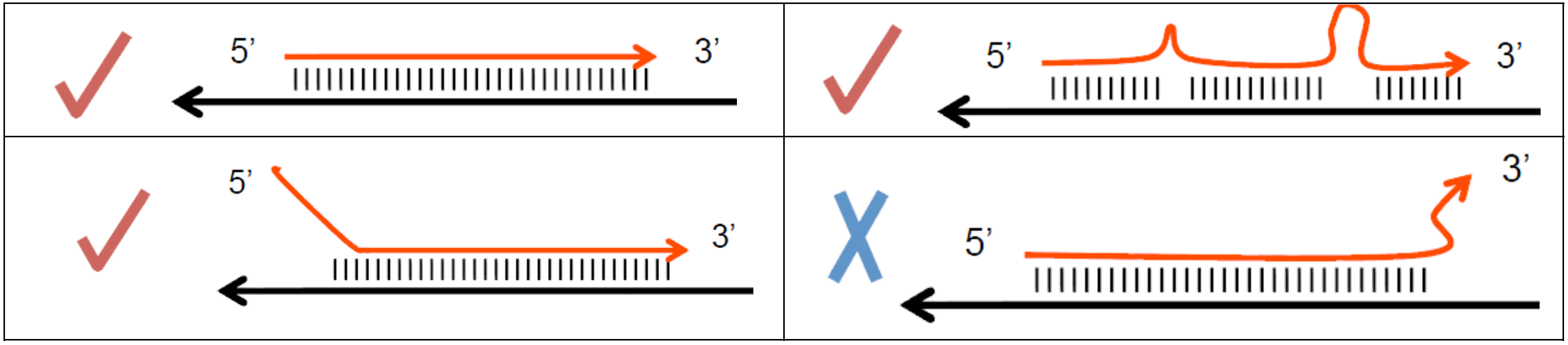

Self-Dimer

```
4 bp, delta G = -6.6 kc/m (bad!)  (worst= -36.6)
        5' GGGAAAATTCCAGGATCTAT 3'
                |||| ||||
3' TATCTAGGACCTTAAAAGGG 5'

4 bp, delta G = -5.4 kc/m (bad!)  (worst= -36.6)
      5' GGGAAAATTCCAGGATCTAT 3'
                ||||
3' TATCTAGGACCTTAAAAGGG 5'
```


Dimer

```
forward  primer

5'  TATCTAGGACCTTAAAAGGG  3'
              |||||
       3'  CATGGAAACGTAGGAGAC  5'

                 reverse  primer
```

Acceptable and Unacceptable Primer Bonding

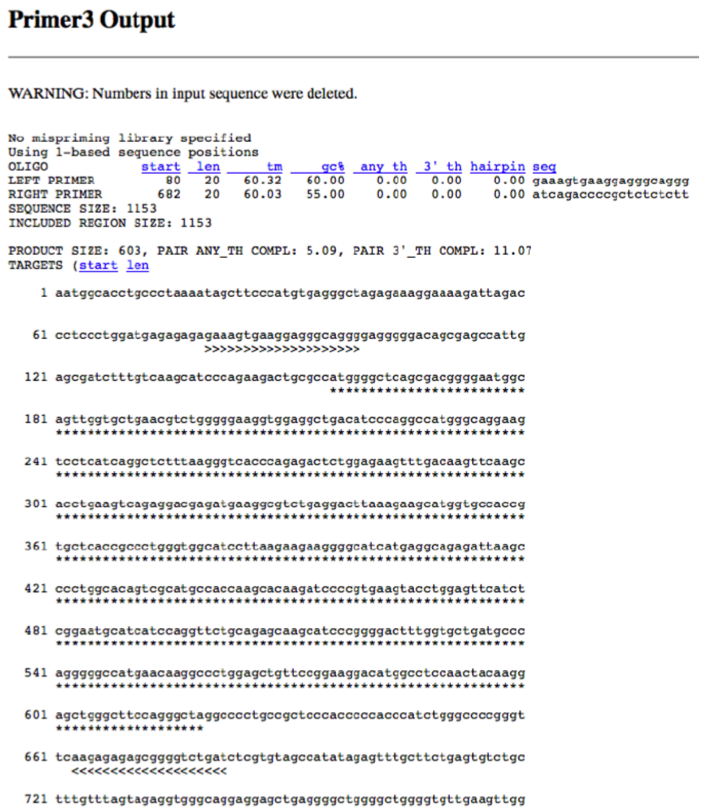


FASTA Format – File format for specifying sequence information.

Techniques to Reduce the Number of Search Hits

- **Provide a maximum E-Value.**
- **Search only newest elements in the database.**

Rigorous Alignment Method – Example: Dynamic Programming. Guaranteed to find the optimal alignment.



Primer3 Output

WARNING: Numbers in input sequence were deleted.

```
No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len      tm     gc%  any th  3' th hairpin seq
LEFT PRIMER         80   20   60.32   60.00    0.00   0.00    0.00 gaaagtgaaggagggcaggg
RIGHT PRIMER       682   20   60.03   55.00    0.00   0.00    0.00 atcagaccccgctctctctt
SEQUENCE SIZE: 1153
INCLUDED REGION SIZE: 1153

PRODUCT SIZE: 603, PAIR ANY_TH COMPL: 5.09, PAIR 3'_TH COMPL: 11.07
TARGETS (start len

    1 aatggcacctgccctaaaatagcttcccatgtgagggctagagaaaggaaaagattagac


   61 cctccctggatgagagagagaaagtgaaggagggcaggggaggggacagcgagccattg
                         >>>>>>>>>>>>>>>>>>>>

  121 agcgatctttgtcaagcatcccagaagactgcgccatggggctcagcgacggggaatggc
                                         *************************

  181 agttggtgctgaacgtctgggggaaggtggaggctgacatcccaggccatgggcaggaag
      ************************************************************

  241 tcctcatcaggctctttaagggtcacccagagactctggagaagtttgacaagttcaagc
      ************************************************************

  301 acctgaagtcagaggacgagatgaaggcgtctgaggacttaaagaagcatggtgccaccg
      ************************************************************

  361 tgctcaccgccctgggtggcatccttaagaagaaggggcatcatgaggcagagattaagc
      ************************************************************

  421 ccctggcacagtcgcatgccaccaagcacaagatccccgtgaagtacctggagttcatct
      ************************************************************

  481 cggaatgcatcatccaggttctgcagagcaagcatcccggggactttggtgctgatgccc
      ************************************************************

  541 aggggggccatgaacaaggccctggagctgttccggaaggacatggcctccaactacaagg
      ************************************************************

  601 agctgggcttccagggctaggcccctgccgctcccaccccccacccatctgggccccgggt
      *******************

  661 tcaagagagagcggggtctgatctcgtgtagccatatagagtttgcttctgagtgtctgc
        <<<<<<<<<<<<<<<<<<<<

  721 tttgtttagtagaggtgggcaggaggagctgaggggctggggctggggtgttgaagttgg
```

Primer3 Primer Builder Output

Created by MIT.

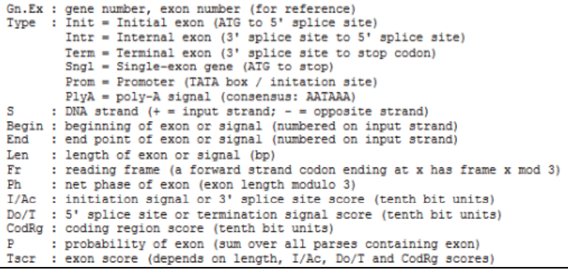
- >>> - Left primer (5' end of the sequence)
- <<< - Right primer (3' end of the sequence)
- *** Coding sequence (begins with AUG for methionine)
- tm – Melting point (Ideally 52C-65C)
- gc% - Percent content of Cytosine/Guanine (ideally between 50-60%)
- len – Primer length (i.e. number of nucleotides)

Specifying a range for primer design:

- Initial Number – Starting Nucleotide
- Second Number – Terminal Nucleotide

GENSCAN – Gene predictor

Genscan Codes:



```
Gn.Ex : gene number, exon number (for reference)
Type  : Init = Initial exon (ATG to 5' splice site)
        Intr = Internal exon (3' splice site to 5' splice site)
        Term = Terminal exon (3' splice site to stop codon)
        Sngl = Single-exon gene (ATG to stop)
        Prom = Promoter (TATA box / initation site)
        PlyA = poly-A signal (consensus: AATAAA)
S     : DNA strand (+ = input strand; - = opposite strand)
Begin : beginning of exon or signal (numbered on input strand)
End   : end point of exon or signal (numbered on input strand)
Len   : length of exon or signal (bp)
Fr    : reading frame (a forward strand codon ending at x has frame x mod 3)
Ph    : net phase of exon (exon length modulo 3)
I/Ac  : initiation signal or 3' splice site score (tenth bit units)
Do/T  : 5' splice site or termination signal score (tenth bit units)
CodRg : coding region score (tenth bit units)
P     : probability of exon (sum over all parses containing exon)
Tscr  : exon score (depends on length, I/Ac, Do/T and CodRg scores)
```

GENSCAN Output – P is probability of the exon being correctly classified.

```
Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..

----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Init +    151    242   92  0  2  103   77   133 0.987  13.71

 1.02 Intr +    373    595  223  1  1  100   96   217 0.999  20.91

 1.03 Term +   1446   1574  129  2  0  116   43   119 0.969   7.40

 1.04 PlyA +   1682   1687    6                                1.05
```

In mutation planning, the **first codon (i.e. methionine) is codon 0**.

In Expasy, an open reading frame (start codon to stop codon inclusive of introns) is shown in **red**.

bl2seq – Blasting two sequences against each other. It is a type of pairwise alignment.

BRCA1 – Gene associated with an increased risk of contracting breast cancer. **It is a transcription factor and tumor suppressor**.

Steps in Cell Lifetime

1. G1 – Growth
2. S – DNA Synthesis
3. G2 – Growth and preparation for division
4. **M – Mitosis**

Reading a Clustal Omega Output

- ***** (**Asterisk**) - Identical alignment
- **:** (**Colon**) - Strongly conserved alignment
- **.** (**Period**)– Weakly conserved alignment
- **Blank** – No meaningful alignment