**Lecture 16 : Applications of Polymerases - DNA Sequencing**

**DNA Sequencing (Sanger Method):** DNA sequencing determines the order of the bases. To sequence the DNA we will begin with a primer-template junction and then use a DNA polymerase to add new bases. We will determine the **order** at which each base is added by the polymerase. This will give the sequence of the DNA that is complementary to the template strand – or the sequence of the strand containing the primer.



In the above example, the order of addition of bases by the polymerase would be:

1st G 2nd C 3rd G 4th A 5th T …….

Determining the order of bases is accomplished:

1. By generating DNA molecules that end in a known base.
2. Determining the position of that base by measuring the length of the DNA.



**1. Known Base:** the DNA fragments that are generated will end with a known, **colored**, base. This is accomplished by including a small amount (1%) of a dideoxy nucleoside triphosphate in the reaction with normal dNTPs. Each type of dideoxy (A,G,C,T) has a different color that comes from special fluorescent properties of the dideoxybase.



*What is the consequence of missing a 3’-OH on the dideoxynucleotide*? *Can it be incorporated into the growing chain?*

**Example**: Consider elongation of a collection of five primer-templates, assuming a ratio of dNTP to ddNTP of 4:1 – in this case the chance of termination by a ddNTP is ~20% - one in four additions of a base will terminate.

**5’GATTCA** **5’GATTCA**G3’OH **5’GATTCA**GC3’OH **5’GATTCA**GCG3’OH **5’GATTCA**GCGA3’OH

3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT--

**5’GATTCA** **5’GATTCA**G3’OH **5’GATTCA**GC3’OH **5’GATTCA**GCG3’OH **5’GATTCA**GCG**A3’H**

3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT--

**5’GATTCA** **5’GATTCA**G3’OH **5’GATTCA**GC3’OH **5’GATTCA**GC**G3’H**

3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT--

**5’GATTCA** **5’GATTCA**G3’OH **5’GATTCAGC3’H**

3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT--

**5’GATTCA** 5’GATTCA**G3’H**

3’—CCTAAGTCGCTATT-- 3’—CCTAAGTCGCTATT--

After completion of synthesis all of the following fragments would be generated.

**5’GATTCAG3’H 5’GATTCAGC3’H 5’GATTCA**GC**G3’H 5’GATTCA**GCG**A3’H 5’GATTCA**GCGA**T3’H 5’GATTCA**GCGAT**A3’H ...**

**2. Determining the Position:** Electrophoresis – separation by size using an electric field.



**Translation:** DNA Sequence to Protein Sequence.

*A codon* is a series of three nucleotide bases that encode a single amino acid.

**Codon Table:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **5' Base** | **Middle Base** | | | | **3'** |
|  | **T** | **C** | **A** | **G** |  |
| **T** | Phe | Ser | Tyr | Cys | **T** |
|  | Phe | Ser | Tyr | Cys | **C** |
|  | Leu | Ser | **Term** | **Term** | **A** |
|  | Leu | Ser | **Term** | Trp | **G** |
| **C** | Leu | Pro | His | Arg | **T** |
|  | Leu | Pro | His | Arg | **C** |
|  | Leu | Pro | Gln | Arg | **A** |
|  | Leu | Pro | Gln | Arg | **G** |
| **A** | Ile | Thr | Asn | Ser | **T** |
|  | Ile | Thr | Asn | Ser | **C** |
|  | Ile | Thr | Lys | Arg | **A** |
|  | **Met** | Thr | Lys | Arg | **G** |
| **G** | Val | Ala | Asp | Gly | **T** |
|  | Val | Ala | Asp | Gly | **C** |
|  | Val | Ala | Glu | Gly | **A** |
|  | Val | Ala | Glu | Gly | **G** |

1. Three DNA bases specify a single amino acid. These are called a 'codon'. For example, the following codon is translated as follows:

TGG =

1. The first codon in all genes that encode proteins start with ATG (AUG in the RNA), or the amino acid methionine *(HIV protease DNA sequence does not start with Met because it was produced by cleavage of a longer immature protein).*
2. Special codons (termination codons) indicate the end of the protein. These are UAA, UAG, UGA. (*The HIV protease DNA sequence lacks a stop codon because its carboxy terminus is produced by proteolysis).*
3. The "reading frame" must be defined during the translation of the mRNA to protein. The reading frame is the base that is taken to be the first base of the codon. The rest of the codons are obtained by taking 3 bases at a time. Without knowledge of the reading frame a sequence could be punctuated in any one of the following three ways, giving three completely different sequences.

*Frame 1 Frame 2 Frame 3*

**C**CT CAG ATC ***or*** C **C**TC AGA TC ***or*** CC **T**CA GAT C

Pro Gln Ile Leu Arg Ser Ser Asp-

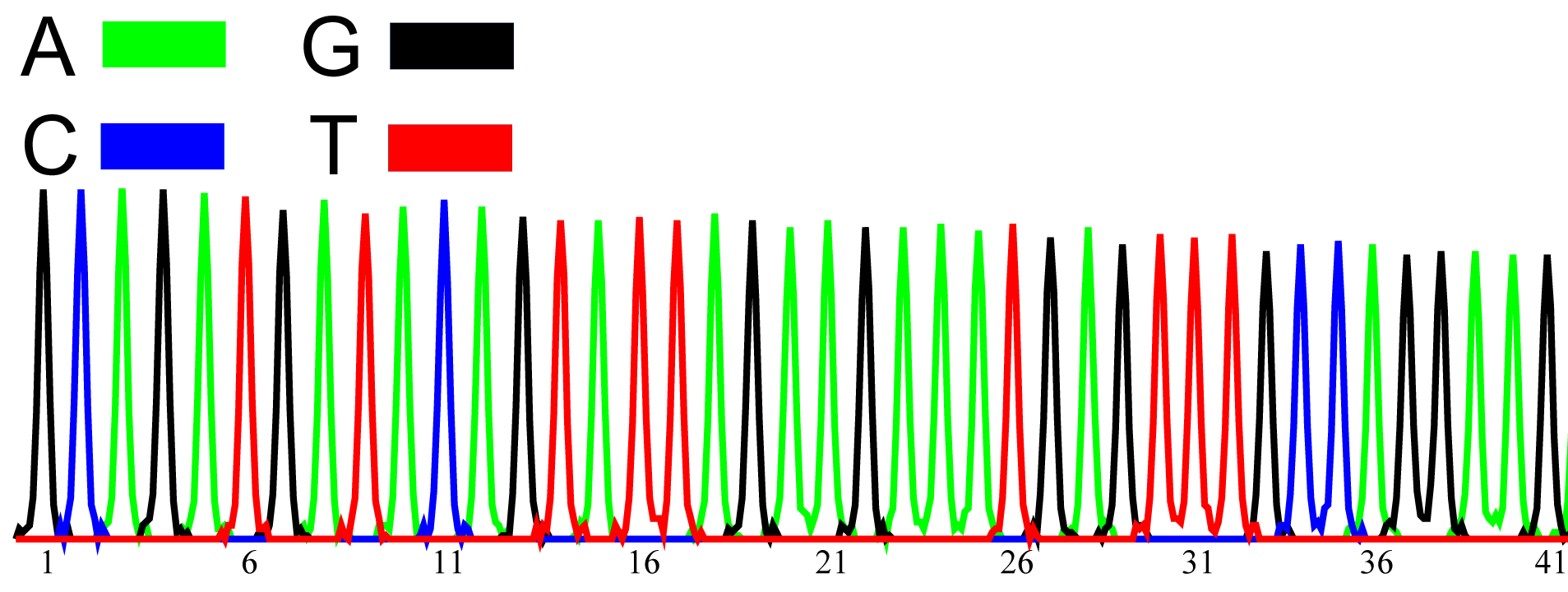
The reading frame from a DNA sequencing experiment can only be established by:

* Comparing the predicted protein sequence to the protein sequence determined by chemical methods.
* A comparison of the DNA sequence to the previously determined reading frame, as will be done in the example below.

***There is only one correct reading frame – the one that gives the observed amino acid sequence of the protein.***

**Example 1** – Sequencing the HIV Protease Gene – using the known reading frame.

|  |
| --- |
| **Region of HIV DNA Coding for HIV protease.**  5'-ggagccgatagacaaggaactgtatcctttaacttccctcagatcactctttggcaa57  **ProGlnIleThrLeuTrpGln7**  58cgacccctcgtcacaataaAgataggggggcaactaaaggaagctctatta**gat**acagga117  **8ArgProLeuValThrIleLysIleGlyGlyGlnLeuLysGluAlaLeuLeuAspThrGly27**  118gcagatgatacagtattagaagaaatgaGtttgccaggaaGatggaaaccaaaaatgata177  **28AlaAspAspThrValLeuGluGluMetSerLeuProGlyArgTrpLysProLysMetIle47**  178gggggaattggaggttttatcaaagtaagacagtaTgatcagatacTCAtagaaatctgt237  **48GlyGlyIleGlyGlyPheIleLysValArgGlnTyrAspGlnIleLeuIleGluIleCys67**  238ggacataaagctataggtacagtattagtaggacctacacctgtcaacataattggaaga297  **68GlyHisLysAlaIleGlyThrValLeuValGlyProThrProValAsnIleIleGlyArg87**  298aatctgttgactcagattggttgCactttaaatttTcccattagccctattgagact354-3'  **88AsnLeuLeuThrGlnIleGlyCysThrLeuAsnPhe** |



**AGATGATACAGTATTAGAAGAAAT…**

**AspAspThrValLeuGluGluMet**

**Example 2** – Finding Mutations in the HIV protease gene.

a) What is the change in the DNA sequence (Hint – start at 118 in the DNA sequence)?

**Codon Table:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **5' Base** | **Middle Base** | | | | **3'** |
|  | **T** | **C** | **A** | **G** |  |
| **T** | Phe | Ser | Tyr | Cys | **T** |
|  | Phe | Ser | Tyr | Cys | **C** |
|  | Leu | Ser | **Term** | **Term** | **A** |
|  | Leu | Ser | **Term** | Trp | **G** |
| **C** | Leu | Pro | His | Arg | **T** |
|  | Leu | Pro | His | Arg | **C** |
|  | Leu | Pro | Gln | Arg | **A** |
|  | Leu | Pro | Gln | Arg | **G** |
| **A** | Ile | Thr | Asn | Ser | **T** |
|  | Ile | Thr | Asn | Ser | **C** |
|  | Ile | Thr | Lys | Arg | **A** |
|  | **Met** | Thr | Lys | Arg | **G** |
| **G** | Val | Ala | Asp | Gly | **T** |
|  | Val | Ala | Asp | Gly | **C** |
|  | Val | Ala | Glu | Gly | **A** |
|  | Val | Ala | Glu | Gly | **G** |

b) What is the change in the amino acid sequence (Find the correct reading frame)?

c) Give the last 6 bases of the primer was used to produce these sequences?

1. Read both sequences, beginning a few bases before where they are different.

Wild-type sequence: TATTAGAAGAAA

Mutant sequence: TACTCGAGGAAA

2. Find the sequence that you just read in the known sequence of HIV protease.

118gcagatgatacag**tattagaagaaa**tg…

**28AlaAspAspThrValLeuGluGluMet……**

3. I happened to start reading at the second base of a codon, so I can punctuate the DNA sequences as follows:

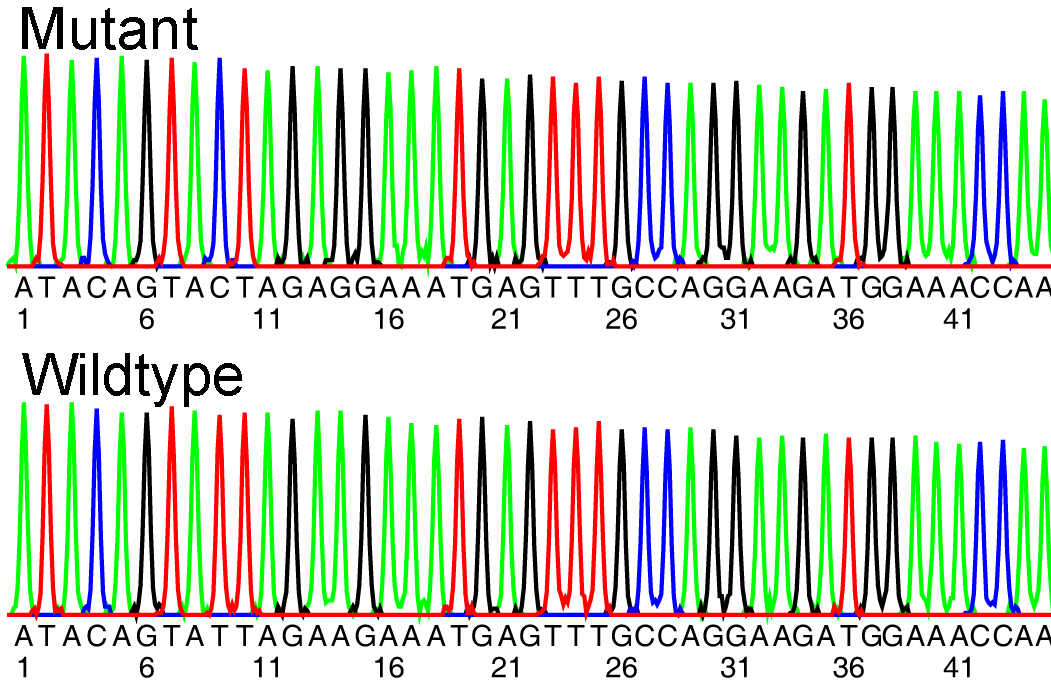
4. Translation

Wild-type sequence: TA TTA GAA GAA A

Val Leu Glu Glu Met

Mutant sequence: TA CTC GAG GAA A

Val



|  |
| --- |
| **Region of HIV DNA Coding for HIV protease.**  5'-ggagccgatagacaaggaactgtatcctttaacttccctcagatcactctttggcaa57  **ProGlnIleThrLeuTrpGln7**  58cgacccctcgtcacaataaAgataggggggcaactaaaggaagctctatta**gat**acagga117  **8ArgProLeuValThrIleLysIleGlyGlyGlnLeuLysGluAlaLeuLeuAspThrGly27**  118gcagatgatacag**tattagaagaaa**tgaGtttgccaggaaGatggaaaccaaaaatgata177  **28AlaAspAspThrValLeuGluGluMetSerLeuProGlyArgTrpLysProLysMetIle47**  178gggggaattggaggttttatcaaagtaagacagtaTgatcagatacTCAtagaaatctgt237  **48GlyGlyIleGlyGlyPheIleLysValArgGlnTyrAspGlnIleLeuIleGluIleCys67**  238ggacataaagctataggtacagtattagtaggacctacacctgtcaacataattggaaga297  **68GlyHisLysAlaIleGlyThrValLeuValGlyProThrProValAsnIleIleGlyArg87**  298aatctgttgactcagattggttgCactttaaatttTcccattagccctattgagact354-3'  **88AsnLeuLeuThrGlnIleGlyCysThrLeuAsnPhe** |

**Underlined** = complete sequence contained in chromatogram.

**Bold** = region where sequences are different between the mutant and the wild-type.

**Box** = primer that was used to generate these sequences (same primer used for both DNAs)

Computational Translation:

Translation by hand using the codon table is tedious – it is easy to automate using a computer. There are a number of web servers available. One is the EXPASY server: <http://web.expasy.org/translate/>

Submitting the sequence data for the wild-type: Expasy translates all six frames (three forward, three from the other strand), here are the three forward. It gives the one letter abbreviation for the amino acids, a “-“ indicates a stop codon.

[5'3' Frame 1](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.8990,1)

atacagtattagaagaaatgagtttgccaggaagatggaaaccaa

I Q Y - K K - V C Q E D G N Q

[5'3' Frame 2](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.8990,2)

atacagtattagaagaaatgagtttgccaggaagatggaaaccaa

Y S I R R N E F A R K M E T

[5'3' Frame 3](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.8990,3)

atacagtattagaagaaatgagtttgccaggaagatggaaaccaa

T V L E E M S L P G R W K P

We still need to select the correct reading frame.

Frame 1 – Contains two stop codons, so that is incorrect

Frame 2 – Sequence is not found in the HIV protease amino acid sequence, so that is incorrect.

Frame 3 – Is correct because we can find Thr31Val32Leu33Glu34Glu35 in the protease sequence.

Now translate the mutant sequence:

[5'3' Frame 1](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.19685,1)

atacagtactagaagaaatgagtttgccaggaagatggaaaccaa

I Q Y - K K - V C Q E D G N Q

[5'3' Frame 2](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.19685,2)

atacagtactagaagaaatgagtttgccaggaagatggaaaccaa

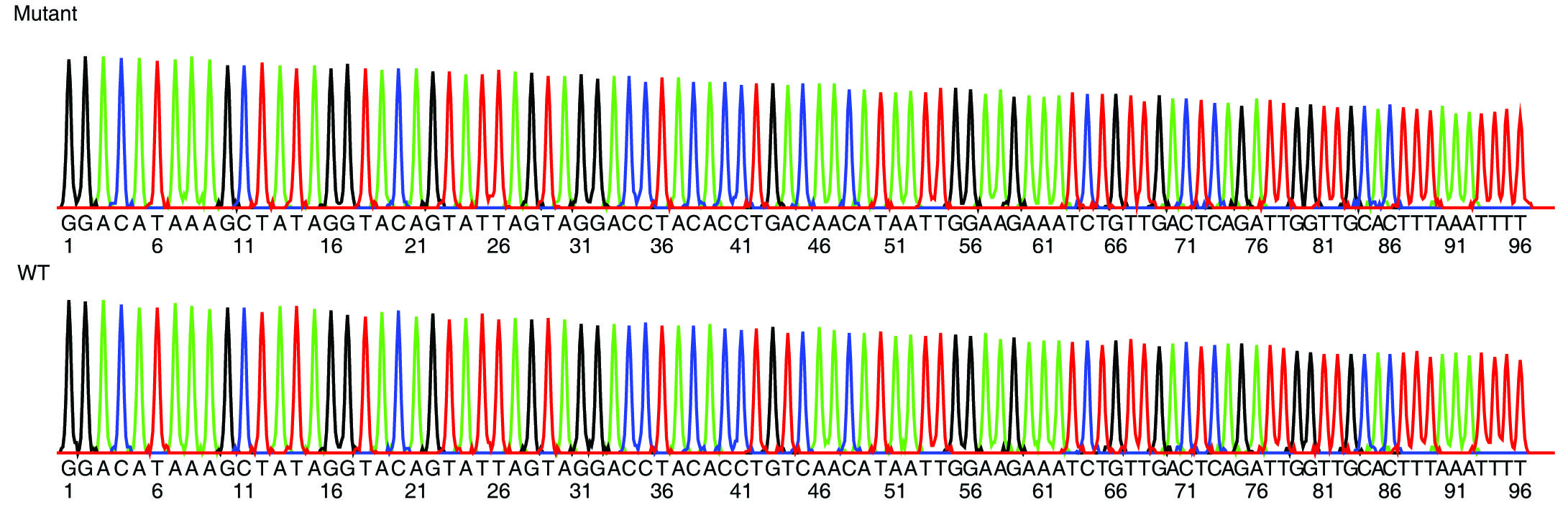
Y S T R R N E F A R K M E T

[5'3' Frame 3](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.19685,3)

atacagtactagaagaaatgagtttgccaggaagatggaaaccaa

T V L E E M S L P G R W K P

Since both sequences started at the same point, frame 3 is the correct one to use. We see no difference in the amino acid sequence – which is what we found when we translated by hand.

**Example 3:**

Submitting the wild-type sequence to Expasy gives:

[5'3' Frame 1](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.25526,1)

ggacataaagctataggtacagtattagtaggacctacacctgtcaacataattggaaga

G H K A I G T V L V G P T P V N I I G R

Aatctgttgactcagattggttgcactttaaatttt

N L L T Q I G C T L N F

[5'3' Frame 2](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.25526,2)

ggacataaagctataggtacagtattagtaggacctacacctgtcaacataattggaagaa

D I K L - V Q Y - - D L H L S T - L E E

atctgttgactcagattggttgcactttaaatttt

I C - L R L V A L - I

[5'3' Frame 3](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.25526,3)

ggacataaagctataggtacagtattagtaggacctacacctgtcaacataattggaagaaa

T - S Y R Y S I S R T Y T C Q H N W K K

tctgttgactcagattggttgcactttaaatttt

S V D S D W L H F K F

Only frame 1 can be correct, since it lacks stop codons. The sequence starts at amino acid 68 in the protein sequence – Gly68 His69 Lys70 Ala71 ….. Val82…….Phe99

**Mutant sequence** – it will be frame 1 again if we submit the sequence starting at the same position.

[5'3' Frame 1](http://web.expasy.org/cgi-bin/translate/dna_sequences?/work/expasy/tmp/http/seqdna.14108,1)

ggacataaagctataggtacagtattagtaggacctacacctgacaacataattggaaga

G H K A I G T V L V G P T P D N I I G R

aatctgttgactcagattggttgcactttaaatttt

N L L T Q I G C T L N F

The changed residue is highlighted; valine (Val, V) at 82 has been changed to aspartic acid (Asp, D).