# DATA DESCRIPTION

<https://en.wikipedia.org/wiki/HIV-1_protease>

<https://en.wikipedia.org/wiki/HIV#Replication_cycle>

## Col-1: patient ID

## Col-2: responder status

("1" for patients who improved and "0" otherwise)

## Col-3: Protease nucleotide sequence (virus)

Nucleotides: A nucleotide is the basic building block of nucleic acids (RNA and DNA). A nucleotide consists of a sugar molecule (either ribose in RNA or deoxyribose in DNA) attached to a phosphate group and a nitrogen-containing base. The bases used in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). In RNA, the base uracil (U) takes the place of thymine. DNA and RNA molecules are polymers made up of long chains of nucleotides.

Protease: HIV-1 protease (PR) is an enzyme involved with peptide bond hydrolysis in retroviruses, that is essential for the life-cycle of HIV. It cleaves newly synthesized polyproteins to create the mature protein components of an HIV virion, the infectious form of a virus outside of the host cell. Without effective HIV protease, HIV virions remain uninfectious.

## Col-4: Reverse Transciptase nucleotide sequence (virus)

When viral HIV-RNA enters the cell, it is accompanied by a reverse transcriptase, an integrase, and a mature HIV-1 PR (protease). The reverse transcriptase converts viral RNA into DNA, facilitating the integrase's role in incorporating viral genetic information with the host cell DNA.

## Col-5: viral load at the beginning of therapy (log-10 units)

viral load is the number of viral particles in one mL of blood.  In this dataset the viral load is represented in a log-10 scale.

## Col-6: CD4 count at the beginning of therapy

The CD4+ cell count is an estimate of the number of white-blood-cells in the same 1 mL of blood for the viral load, in a log-10 scale.

The higher the number the more "active" the immune system. Paradoxically higher CD4 counts imply both a healthier individual but also a higher amount of viral reproduction (the virus primarily replicates in CD4 cells).

# DATA PREPROCESSING

<https://iubmb.qmul.ac.uk/misc/naseq.html#201>

## Nucleotide sequences

Table 4. Triplet correspondence for amino acids (standard genetic code)

|  |  |  |
| --- | --- | --- |
| Amino acid | Single-letter code | Triplet (5'-3') |
| Glycine | G | GGN |
| Alanine | A | GCN |
| Valine | V | GTN |
| Leucine | L | *YTN* (CTN and TTR) |
| Isoleucine | I | ATH |
| Proline | P | CCN |
| Phenylalanine | F | TTY |
| Tyrosine | Y | TAY |
| Cysteine | C | TGY |
| Methionine | M | ATG |
| Histidine | H | CAY |
| Lysine | K | AAR |
| Arginine | R | *MGN* (CGN and AGR) |
| Tryptophan | W | TGG |
| Serine | S | *WSN* (TCN and AGY) |
| Threonine | T | ACN |
| Aspartic acid | D | GAY |
| Glutamic acid | E | GAR |
| Asparagine | N | AAY |
| Glutamine | Q | CAR |
| Aspartic acid or asparagine | B | RAY |
| Glutamic acid or glutamine | Z | SAR |
| Terminator | . | *TRR* (TAR and TGA) |
| Unknown | X | NNN |