

BLUFF YOUR WAY IN GENETICS

From this document you will learn basic background information on:

- Cell biology
- Molecular biology
- Genetics

© Fachinformationszentrum Karlsruhe, March 2008

Robert Austin / Dr. Ilka Schindler
Fachinformationszentrum (FIZ) Karlsruhe
Hermann von Helmholtz Platz 1
76344 Eggenstein-Leopoldshafen
Germany

Tel: +49 7247 808 555

Fax: +49 7247 808 131

Email: helpdesk@fiz-karlsruhe.de

Web: www.fiz-karlsruhe.de

BLUFF YOUR WAY IN GENETICS

CONTENTS

Overview	5
Cell Biology	6
Molecular Biology	7
The Alphabet of Proteins	8
The Alphabet of RNA	9
The Alphabet of DNA	10
Double Stranded DNA	11
Transcription	12
Translation	13
Reading Frames	14
Genetics	15
Complementary DNA	16
Viruses	17
The Universal Genetic Code	19
Amino acids	20

Overview

This document is aimed at equipping searchers with the necessary basic scientific knowledge to understand the sequence search features of [REGISTRY](#), [DGENE](#), [USGENE®](#) and [PCTGEN](#).

Searchers who were originally trained as, e.g. organic chemists, may find themselves increasingly called upon by end-users within their organisations to supply intellectual property information for protein and nucleic acid molecules (sequences). The end-user might be a research scientist or a patent attorney; and at first glance this kind of request can be quite a daunting prospect.

Surprising as it may seem, the level of scientific understanding required to search sequences on STN is actually relatively low. The sequence content of REGISTRY, DGENE, USGENE and PCTGEN is readily accessible to anyone trained in any of a broad spectrum of chemical or biological sciences. All that is required is a high-level grasp of the processes involved in the flow of information from an organism's genetic "blue print" into the molecules which govern an organism's structure and function.

This high-level understanding falls broadly into three categories.

- Cell biology
- Molecular biology
- Genetics

Helpful HINT

Find out more about searching sequence databases on STN at:
http://www.stn-international.com/training_center/mat_sea_stn.html#Bioseq

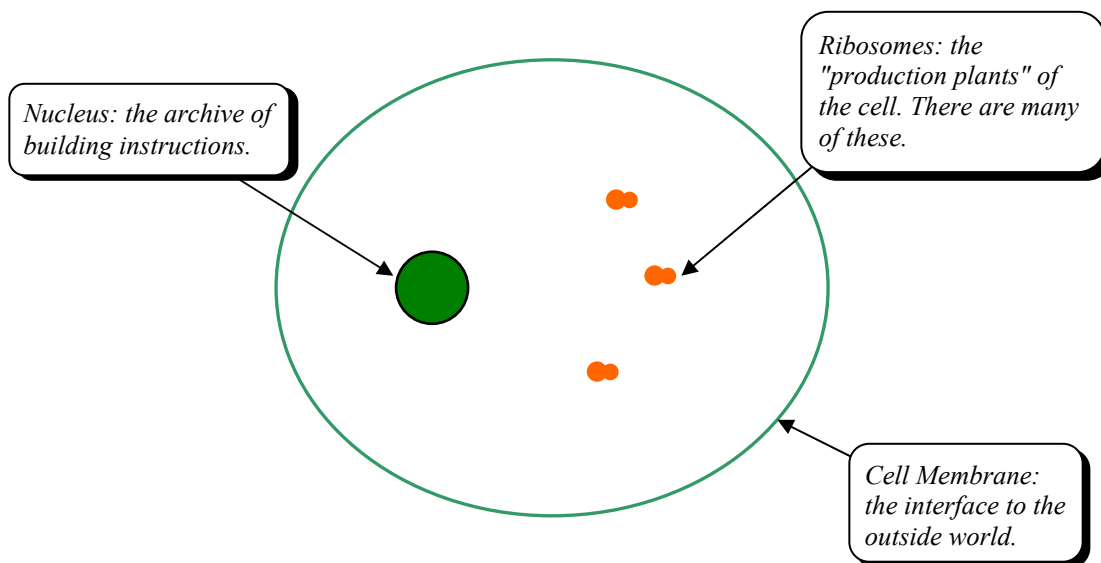
Cell Biology

For the purposes of understanding sequences on STN, all you need to know about *biology* is that all organisms, except viruses, are made up of *cells*. (We will come back to viruses later.)

A cell is essentially a microscopic, spherical lipid *membrane* which encapsulates the cellplasma, a myriad of different interacting biomolecules, both large and small, and functional substructures called *organelles*.

There are only two *organelles* of interest to us in understanding the biological background: the *nucleus*, where the organism's archive of self-building instructions is located, and *ribosomes*, where these building instructions are turned into a biological manufacturing process. It is not necessary to understand the detailed structure of these two organelles, just broadly what they do. The "blue-print" of life is stored in the *Nucleus*, and the various biomolecules which control the structure and function of an organism are manufactured, following the blue-print's instructions, at ribosomes.

The cell membrane itself, is either an interface with surrounding cells or, in the case of a single-celled organism, the surrounding external environment. This extra-cellular interaction is policed by a variety of different biomolecules, which essentially "float" in the lipid membrane, connecting the world outside to the world within.

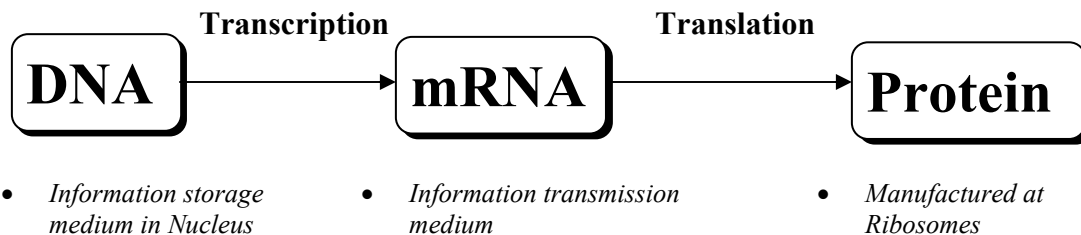


Molecular Biology

There are three broad types of biomolecules which are represented in the STN sequence databases: deoxyribose nucleic acid (DNA), ribose nucleic acid (RNA) and proteins/polypeptides.

Central Dogma of Biology

The information pathway relationship between the three types of biomolecules could be called the *Central Dogma of Biology*. DNA is *transcribed* into mRNA, in the nucleus, and mRNA is *translated* into protein at the ribosomes.



Proteins

Within a cell, and on its membrane surface, there is a highly complex network of (bio)chemical processes at work, primarily controlled and/or catalysed by *proteins*. In higher organisms *proteins* are also important in extra-cellular features, e.g. hair (keratin). Proteins are manufactured at the *ribosomes*, from the production instructions received as *mRNA* molecules. This process is called *translation*. Chemically, proteins are poly-alpha-amino acid biopolymers (or *polypeptides*).

Ribose nucleic acid (RNA)

Messenger RNA (mRNA) is the molecular transmission medium by which information is carried from storage, as *DNA*, to the production plants, the *ribosomes*. mRNA is *translated* into *protein* at the ribosomes. Chemically, RNA is a polysaccharide biopolymer of *nucleotide* monomers (a *polynucleotide*).

Deoxyribose nucleic acid (DNA)

DNA is the molecular storage medium for genetic information, located in the *nucleus* of a cell. DNA is *transcribed* into *mRNA* in the nucleus. Chemically, *DNA* is also a polysaccharide biopolymer of nucleotide monomers (a *polynucleotide*).

note

For the purposes of a simple explanation it is important to focus just on mRNA here. However, note that RNA actually comes in three basic forms: messenger (mRNA), ribosomal (rRNA) and transfer (tRNA). Each has separate function, and all are included in STN sequence databases.

The Alphabet of Proteins

Viewed from a simple chemical perspective, proteins are sequential polyamide copolymers made from up to 20 naturally occurring alpha-amino acid monomers. The precise order of the amino acid monomers along the length of the polymer corresponds to a defined sequence of nucleotides in the *mRNA* from which the protein was *translated*.

The alpha-amino acid amide bond in proteins is called a "peptide" bond, so a protein can also be described as a *polypeptide*; literally "a polymer of peptide linkages". However it is important to note that the word *peptide* is frequently used, in both scientific literature and patents, as shorthand to mean *polypeptide*.

Polypeptides are sequential molecules which are often represented as a series of letters, e.g.

...ALKSPRGFHITD...

The twenty naturally occurring alpha-amino acids, or *letters of the protein alphabet*, are given in the table on page 20.

Proteins in 3-D

In reality proteins are three-dimensional, and are formed from one or more polypeptide molecules. During *post-translation* biochemical processes, a protein's polypeptide(s) will twist together and cross-link in a precise manner to form, e.g. biologically active globular structures. Proteins formed in this way can even incorporate non-polypeptide components, e.g. iron in haemoglobin. This is known as a protein's tertiary and quaternary structure. It is not necessary to know the detail of this to search polypeptide sequences on STN.

note

By convention it is normal to represent polypeptide sequences with the terminating amine group to the left, and the terminating carboxyl group to the right. I.e. the convention is read/write from NH₂ (left) to COOH (right), along the polypeptide chain. This is how polypeptides are represented on STN.

The Alphabet of RNA

Ribose nucleic acid (RNA) is a sequential copolymer of four monomers called nucleotides. The precise order of the monomers along the length of the polymer corresponds to a precise sequence of nucleotides in the parent DNA molecule, from which the RNA was *transcribed*.

The four nucleotide monomers, or *letters of the RNA alphabet*, are:

- Adenine
- Guanine
- Cytosine
- Uracil

Like polypeptides, RNA can be represented in shorthand as a *sequential* series of letters, e.g.

... AGCUAAUCGAGCUAAUCG ...

Nucleotides in RNA

A *nucleotide* is made up from three components: a *base*, a *sugar* and a *phosphate* group. To be precise, it is the *bases* in nucleotides which are actually "A, G, C or U", and the sugar component in RNA is ribose. The third component of an RNA nucleotide is a phosphate, which is esterified to the 5'-hydroxyl of the ribose sugar. A nucleotide without this phosphate group is called a *nucleoside*.

Chemical structure of RNA

In RNA each nucleotide is joined via a phosphodiester linkage from the 5' hydroxyl group on one ribose ring, to the 3' hydroxyl group on the next ribose ring (forming a type of polysaccharide). At one end of a RNA molecule, the terminating nucleotide has - conceptually - a free (unesterified) 5'-hydroxyl, and conversely at the other end of the sequence there is a free 3'-hydroxyl group.

An RNA sequence therefore has direction, and the convention is to read from 5' (left) to 3' (right). RNA sequences on STN follow this convention.

The Alphabet of DNA

Like RNA, deoxyribose nucleic acid (DNA) is also a sequential copolymer of four monomers called nucleotides. The precise order of the monomers along the length of the polymer constitutes specific instructions for defining the structure and function of a living organism. DNA is the cellular storage medium for biological inheritance information.

The four nucleotide monomers, or *letters of the DNA alphabet*, are:

- Adenine
- Guanine
- Cytosine
- Thymine

DNA and RNA are chemically very similar, and have closely related alphabets. However, in DNA Thymine (T) is present, instead of Uracil (U). DNA can also be represented as *sequential* series of letters, e.g.

... AGCTAATCGAGCTAATCG ...

Nucleotides in DNA

A nucleotide is made up from three components: a *base*, a *sugar* and a *phosphate* group. To be precise, it is the bases in nucleotides which are actually "A, G, C or T", and the sugar for DNA is 2'-deoxyribose, i.e. where the 2'-hydroxyl is absent from ribose. The third component, a phosphate, is esterified to the 5'-hydroxyl of 2'-deoxyribose. A nucleotide without this phosphate group is called a *nucleoside*.

Chemical structure of DNA

In DNA each nucleotide is joined via a phosphodiester linkage from the 5' hydroxyl group on one 2'-deoxyribose ring, to the 3' hydroxyl group on the next 2'-deoxyribose ring. This means that – conceptually - at one end of a DNA molecule the terminating nucleotide has a free (unesterified) 5'-hydroxyl, and at the other end of the sequence there is a free 3'-hydroxyl group.

A DNA sequence therefore has direction, and the convention is also to read from 5' (left) to 3' (right). DNA sequences on STN follow this convention.

Double Stranded DNA

So far in this explanation DNA and mRNA have been represented as single stranded molecules. By and large this is true for mRNA in its natural state. The same is not the case for DNA which is stored in the nuclei of most organisms in an elegant *double-stranded* helical structure.

In double stranded DNA a nucleotide in one strand corresponds to a chemically complementary nucleotide in the second DNA strand. The pairs of nucleotides are held together by hydrogen-bonds. Three dimensionally the two strands appear to wind around one another, forming a shape known as the double helix.

The pairings in double stranded DNA are as follows:

- A to T
- G to C
- T to A
- C to G

It is important to note that in double stranded DNA the second strand runs in the opposite direction from the first DNA strand, i.e. one runs from 5' to 3' and the other runs from 3' to 5'. The 5' to 3' strand is known as the *coding strand* and the 3' to 5' is known as the *non-coding strand*. The coding strand is where the information about the organism is actually stored.

For example:

DNA **5'...AGCTAAAG...3'**
DNA **3'...TCGATTTC...5'**

On STN, as elsewhere, the convention is to represent all polynucleotides, including non-coding DNA sequences, from 5' to 3'. So on STN the second strand would be written:

DNA **5'...CTTTAGCT...3'**

Three types of DNA

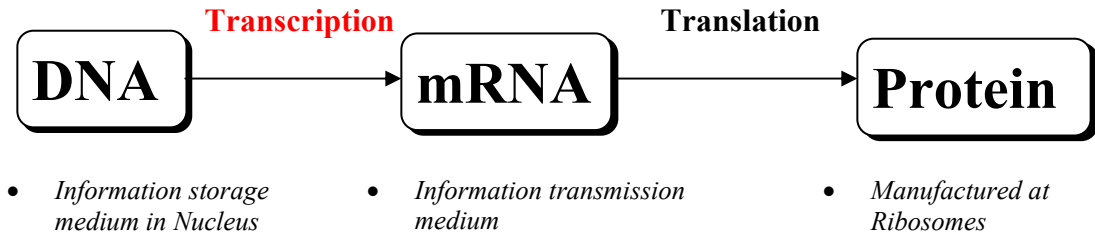
There are three important classifications of DNA molecules to consider at this point:

- Double stranded (ds) DNA
- Single stranded (ss) DNA – coding
- Single stranded (ss) DNA – non-coding

Transcription

Central Dogma of Biology

The information pathway relationship between DNA, mRNA and proteins could be called the *Central Dogma of Biology*. DNA is *transcribed* into mRNA, in the nucleus, and mRNA is *translated* into protein at the ribosomes.

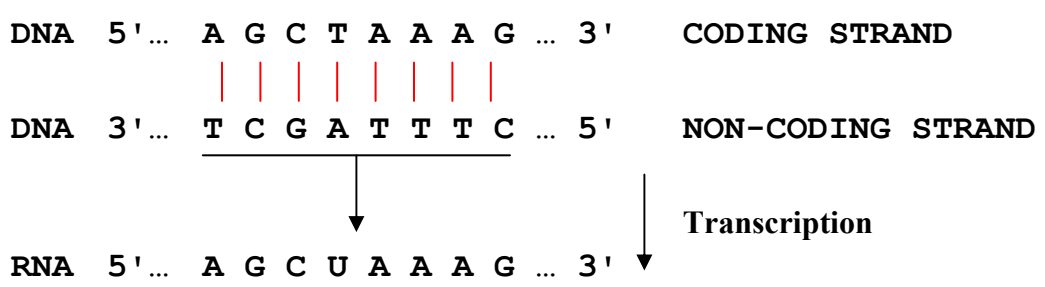


Transcription

DNA is transcribed to RNA in the nucleus. The key high-level concept is that one letter in the DNA template corresponds to a chemically complementary letter in the RNA transcript. The base pairing of both partners (via hydrogen bonds), i.e. leading from DNA template to RNA transcript is as follows.

- A to U
- G to C
- T to A
- C to G

For example:

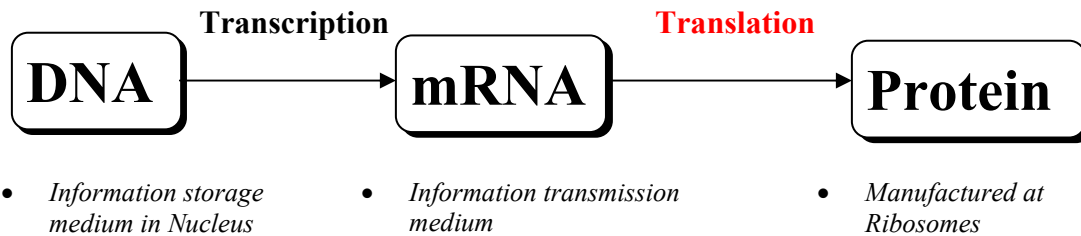


RNA is derived by complementary base pairing with the *non-coding strand* of DNA. This ensures that it has the corresponding sequence to the *coding strand*, with the same direction 5' (left) to 3' (right). In reality DNA is "unzipped" before this can happen. However it is not necessary to understand how this works to use REGISTRY, DGENE or PCTGEN.

Translation

Central Dogma of Biology

The information pathway relationship between DNA, mRNA and proteins could be called the *Central Dogma of Biology*. DNA is *transcribed* into mRNA, in the nucleus, and mRNA is *translated* into protein at the ribosomes.



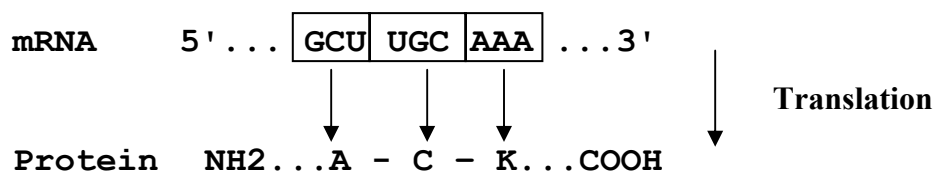
Translation

Proteins are not chemically similar to RNA or DNA. One alphabet of letters is converted to an entirely different set for proteins. This is why the process is called *translation*. The alphabet for protein sequences comprises 20 naturally occurring alpha-amino acids, i.e. a 4-letter alphabet based message is converted into a 20 letter alphabet based product.

The key high-level concept to grasp is that for every three nucleotides in the mRNA message, there is only one amino acid placed in the corresponding protein sequence at the ribosome.

The nucleotide *triplets* involved are called *Codons*.

For example:



Where: A is Alanine, C is Cysteine, and K is Lysine

Translation from 5' to 3' mRNA generates a *corresponding* polypeptide sequence, which reads from amine terminal group to carboxyl terminal group, left to right.

The Universal Genetic Code

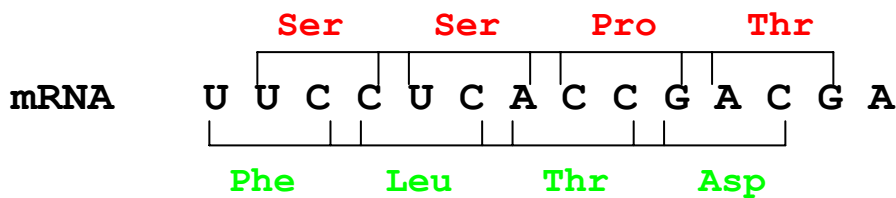
The vast majority of living organisms share exactly the same *codon* to amino acid relationships for translating mRNA into protein structure. The whole scheme for all 20 amino acids is therefore known as the *Universal Genetic Code*. In order to intellectually convert a nucleotide sequence into a peptide sequence, as in the example above, a Universal Genetic Code conversion table is required (see page 19).

Reading Frames

As described on page 13, mRNA is *translated* into a polypeptide based on its sequential order of nucleotide triplets (or *codons*), where each triplet encodes an amino acid in the polypeptide. However, if the same sequence is viewed slightly differently, by shifting the frame of reference by one nucleotide along the sequence, a completely different sequence of codons can be seen. Clearly, if the mRNA were to be translated this way, an entirely different polypeptide sequence would be produced.

Each such view point on a nucleic acid sequence is known as a *reading frame*, and moving between reading frames is known as *frame shift*. Since there are 3 nucleotides per codon, there are 3 reading frames to consider.

For example (showing two of the three reading frames):



To search sequences on STN it is not necessary to understand how organisms define the correct reading frame for translation. However for the "translated polypeptide" homology (similarity) search options in DGENE, REGISTRY and PCTGEN a simple understanding of what *reading frames* are, and that there are 3 of them, is required.

Helpful HINT

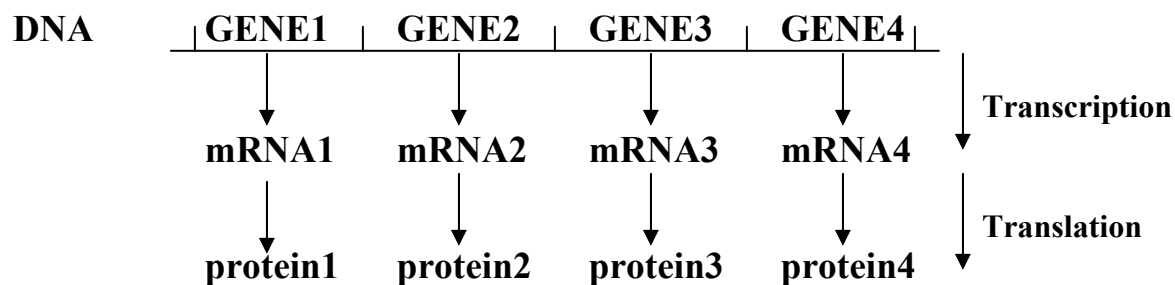
In the example above, what is the amino acid sequence defined by the third reading frame? (The universal genetic code table is given on page 19.)

Genetics

Genetics is a science which is conceptual step-up from the study of DNA, RNA and proteins as interacting macromolecules. Genetics looks at the bigger picture, i.e. the road map, rather than just the roads. Viewing DNA at this higher level we are by definition interested in *genes*; specifically their inter-relationships, hierarchies, variation within and between species, mutations, and their inheritance from one generation to the next.

However, very little knowledge is actually required about the subject of genetics *per se* to search sequences on STN. The only high-level concept to grasp about genes, is that they are discrete regions within DNA sequences, which correspond, via their mRNA transcripts, to proteins/polypeptides manufactured at ribosomes. Humans, and other organisms, have many thousands of Genes. Consequently protein and mRNA sequences are substantially shorter in length than DNA.

The concept is shown in the simple diagram below.



The comparative difference in size is reflected by the fact that mRNA and proteins can journey relatively freely around a cell, whilst substantially larger DNA molecules remain immobile in the nucleus.

Complementary DNA

Turning the logic of transcription around, it is possible to intellectually derive the DNA sequence which gave rise to an mRNA transcript. The complementary nucleotide relationships given on page 11 are simply used in reverse to accomplish this idea.

Of course, as it turns out, experts in the field have already turned this concept into reality in the laboratory. The product of this *reverse-transcription* process is known as complementary DNA (cDNA). In other words: cDNA is DNA which was reverse-transcribed artificially *from* a corresponding mRNA transcript.

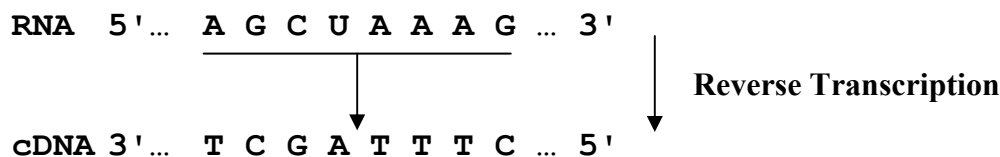
Thus DNA sequences fall in *four* basic categories:

- Double stranded (ds) DNA
- Single stranded (ss) DNA - coding
- Single stranded (ss) DNA - non-coding
- Complementary DNA (cDNA)

For cDNA each nucleotide corresponds to a complementary nucleotide in the RNA sequence from which it was reverse-transcribed. The pairings are analogous to the way in which the mRNA transcript complements the corresponding DNA during transcription (page 12), only in reverse.

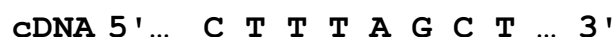
- A to T
- G to C
- U to A
- C to G

For example:



The cDNA runs in the opposite direction from the RNA, i.e. 3' to 5'. On STN, as elsewhere, the convention is to represent all polynucleotides, including cDNA sequences, from 5' to 3'.

So the above example would be found in a database as follows:



Viruses

To complete this brief *bluffer's guide*, it is necessary to touch on the subject of Viruses, as they are fundamentally different from cellular-based forms of life.

Viruses are not cellular organisms, but are basically just a nucleic acid core wrapped in a protein coat. Essentially they are just very large molecular systems. In order to reproduce (*replicate*) viruses invade cells of other organisms and "hijack" the *gene expression* machinery of the cell, so that it produces viral proteins. The viral proteins are then used to assemble new virus particles (*virions*) facilitating the production of the next generation of the virus. After a period of time, the invaded cells die and burst, releasing the new virions into the surrounding environment to move on and infect more cells.

There are two types:

- DNA viruses
- RNA viruses

DNA viruses

DNA viruses have DNA at their core, surrounded by the protein coat, and follow the general principle of the central dogma of biology: DNA to mRNA to protein. However they reproduce using "hijacked" ribosomes (as described above).

Two kinds of DNA viruses are distinguished:

- Double stranded DNA viruses with classes papovavirus (warts, cervical cancer), herpesvirus (cold sores, chickenpox), poxvirus (smallpox, cowpox), polyomavirus, adenovirus (some tumours)
- Single stranded DNA viruses (e.g. roseola)

RNA viruses

RNA viruses have RNA at their core, normally also surrounded by a protein coat, and as such are an exception to the central dogma of biology. Genetic information about the virus is stored as *RNA*, not as DNA. Probably the most infamous example of this kind of virus is HIV.

Once inside the host cell, invading viral RNA takes over the host cell's *gene expression* process, just as DNA viruses do, and produce viral proteins for the next generation of the virus.

RNA viruses fall into four classes:

- Double stranded (+/-) RNA viruses
e.g. reovirus, diarrhoea
- Single stranded positive sense (+) RNA viruses - acting as mRNA
with classes picornavirus (polio, common cold), togavirus (rubella, yellow fever)
- Single stranded negative sense (-) RNA viruses- acting as an mRNA synthesis template
with classes rhabdovirus (rabies), paramyxovirus (measles, mumps), orthomyxovirus (influenza viruses)
- Single stranded RNA viruses - acting as DNA synthesis template
e.g. retroviruses (HIV, Leukaemia viruses, Rous Sarcoma virus,)

note

→ All viral sequences are included in REGISTRY/DGENE/PCTGEN.

The Universal Genetic Code

5' -terminal	Middle				3' -terminal
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met*	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val*	Ala	Glu	Gly	G

Initiation Codons:

*AUG (Met_i) in eukaryotes and in some prokaryote genes.

*GUG (Val_i) in some prokaryotes.

Termination Codons:

UAA, UGA or UAG in prokaryotes and eukaryotes.

These termination codons are sometimes referred to as:

“Stop codons”, “Nonsense codons”

or by the trivial names “amber” (= UAG), “ochre” (= UAA) and “opal” (= UGA).

Degenerate Codons:

There are 61 codons to represent 20 naturally occurring amino acids. All of the amino acids, *except for* Met (= AUG only) and Trp (= UGG only) are represented by more than one codon, i.e. there is degeneracy in the genetic code. The synonym or “degenerate” codons usually form groups in which the base at the third position (3'-terminal) has the least meaning.

With the notable exception of Arg (one of only 3 amino acids, the other two being Ser and Leu, to have 6 codons), the number of codons for each amino acid correlates with its frequency of occurrence in proteins.

Amino acids

The following table lists the 1- and 3-letter codes for the 20 naturally occurring alpha-amino acids found in proteins.

1-Letter Code	3-Letter Code	Name
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

Uncommon amino acids

Alpha-amino acids can exist in two stereoisomeric forms: L or D. Amino acids are always in the natural L-form in unmodified proteins. However it is relatively common for synthetic or semi-synthetic polypeptide sequences to incorporate D-form amino acids. There are also many amino acids which are not part of the list of the 20 above, but which can be incorporated into polypeptide sequences. These are referred to as *uncommon amino acids* and possess different naming conventions.

© Fachinformationszentrum Karlsruhe, March 2008

Robert Austin / Dr. Ilka Schindler
Fachinformationszentrum (FIZ) Karlsruhe
Hermann von Helmholtz Platz 1
76344 Eggenstein-Leopoldshafen
Germany

Tel: +49 7247 808 555

Fax: +49 7247 808 131

Email: helpdesk@fiz-karlsruhe.de

Web: www.fiz-karlsruhe.de