

Lezione 10

Bioinformatica

Mauro Ceccanti[‡] e Alberto Paoluzzi[†]

[†]Dip. Informatica e Automazione – Università “Roma Tre”

[‡]Dip. Medicina Clinica – Università “La Sapienza”



Lezione 10: Sintesi proteica

Synthesis of proteins

Central dogma: DNA makes RNA makes proteins

Genetic code



Sommario

Lezione 10: Sintesi proteica

Synthesis of proteins

Central dogma: DNA makes RNA makes proteins

Genetic code



Synthesis of proteins¹

The key molecular process that makes modern life possible is protein synthesis, since proteins are used in nearly every aspect of living

- ▶ The synthesis of proteins requires a tightly integrated sequence of reactions, most of which are themselves performed by proteins

¹From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins¹

The key molecular process that makes modern life possible is protein synthesis, since proteins are used in nearly every aspect of living

- ▶ The synthesis of proteins requires a tightly integrated sequence of reactions, most of which are themselves performed by proteins
- ▶ (Thus posing one of the unanswered riddles of biochemistry: which came first, proteins or protein synthesis? If proteins are needed to make proteins, how did the whole thing get started?)

¹From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins²

Each different protein is made according to a blueprint

- ▶ The unique linear sequence of amino acids in a protein is encoded in the linear sequence of nucleotides in DNA

²From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins²

Each different protein is made according to a blueprint

- ▶ The unique linear sequence of amino acids in a protein is encoded in the linear sequence of nucleotides in DNA
- ▶ Because DNA is composed of only four types of nucleotides, compared to the twenty types of amino acids in protein, there cannot be a one-to-one correspondence of amino acid to nucleotide

²From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins²

Each different protein is made according to a blueprint

- ▶ The unique linear sequence of amino acids in a protein is encoded in the linear sequence of nucleotides in DNA
- ▶ Because DNA is composed of only four types of nucleotides, compared to the twenty types of amino acids in protein, there cannot be a one-to-one correspondence of amino acid to nucleotide
- ▶ Cells resolve this problem with the most conservative possible coding: a triplet of nucleotides, three in a row, is used to specify one amino acid

²From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins²

Each different protein is made according to a blueprint

- ▶ The unique linear sequence of amino acids in a protein is encoded in the linear sequence of nucleotides in DNA
- ▶ Because DNA is composed of only four types of nucleotides, compared to the twenty types of amino acids in protein, there cannot be a one-to-one correspondence of amino acid to nucleotide
- ▶ Cells resolve this problem with the most conservative possible coding: a triplet of nucleotides, three in a row, is used to specify one amino acid
- ▶ Each position in the triplet can be occupied by one of the four types of nucleotide, so each triplet could potentially specify up to sixty-four amino acids

²From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins²

Each different protein is made according to a blueprint

- ▶ The unique linear sequence of amino acids in a protein is encoded in the linear sequence of nucleotides in DNA
- ▶ Because DNA is composed of only four types of nucleotides, compared to the twenty types of amino acids in protein, there cannot be a one-to-one correspondence of amino acid to nucleotide
- ▶ Cells resolve this problem with the most conservative possible coding: a triplet of nucleotides, three in a row, is used to specify one amino acid
- ▶ Each position in the triplet can be occupied by one of the four types of nucleotide, so each triplet could potentially specify up to sixty-four amino acids
- ▶ This is more than enough to specify the twenty amino acids actually used by cells, along with some special triplet codes for starting and stopping

²From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins²

Each different protein is made according to a blueprint

- ▶ The unique linear sequence of amino acids in a protein is encoded in the linear sequence of nucleotides in DNA
- ▶ Because DNA is composed of only four types of nucleotides, compared to the twenty types of amino acids in protein, there cannot be a one-to-one correspondence of amino acid to nucleotide
- ▶ Cells resolve this problem with the most conservative possible coding: a triplet of nucleotides, three in a row, is used to specify one amino acid
- ▶ Each position in the triplet can be occupied by one of the four types of nucleotide, so each triplet could potentially specify up to sixty-four amino acids
- ▶ This is more than enough to specify the twenty amino acids actually used by cells, along with some special triplet codes for starting and stopping
- ▶ Proteins are built by reading the sequence of nucleotide triplets in DNA and using the information to link amino acids in the proper order.

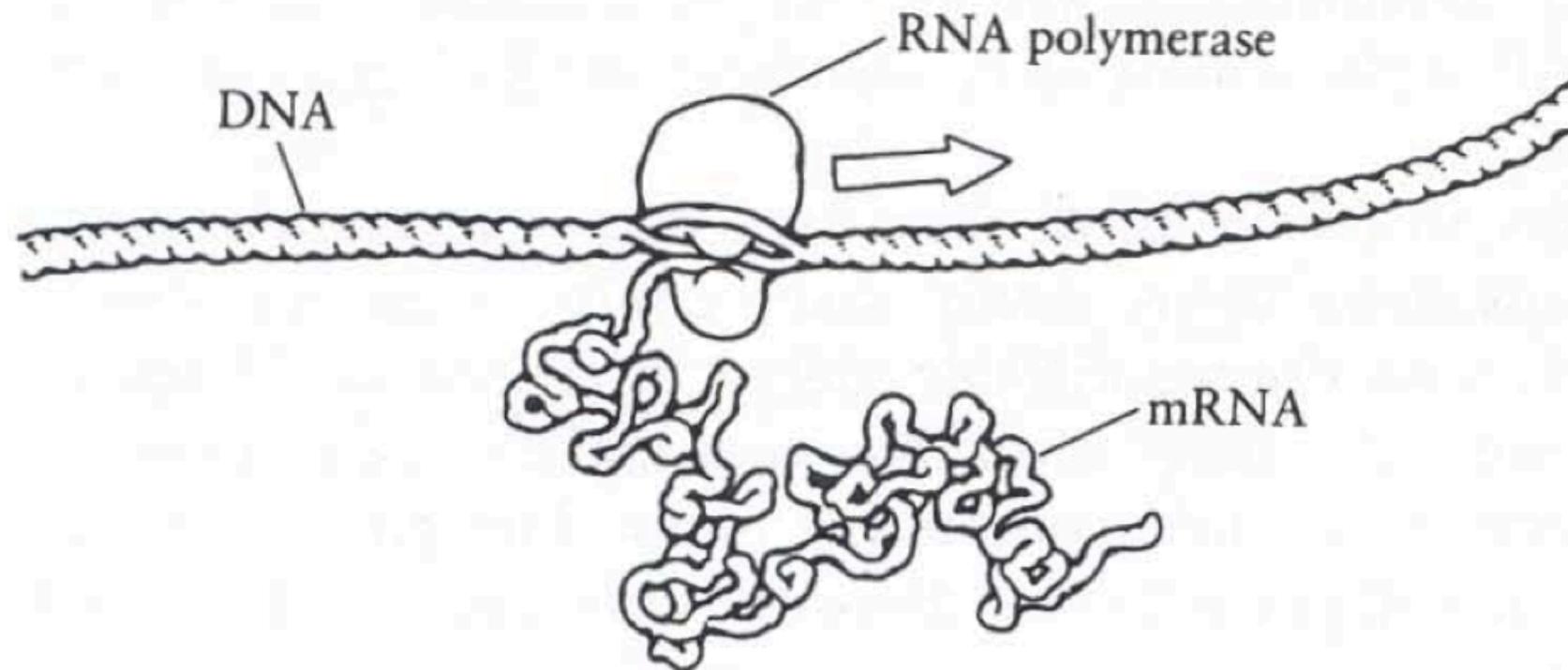
²From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins³

Cells build proteins in two steps, using an intermediary messenger molecule between DNA and a new protein (m-RNA)

Proteins are made in two steps: first the information in DNA is transcribed into mRNA, a messenger molecule, by RNA polymerase, as shown below



³From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁴

In the first step, **transcription**, the messenger molecule is made according to the information stored in DNA

- ▶ The enzyme RNA polymerase unrolls a section of the DNA double helix and, at a rate of about thirty nucleotides per second, builds a strand of RNA complementary to it

⁴From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁴

In the first step, **transcription**, the messenger molecule is made according to the information stored in DNA

- ▶ The enzyme RNA polymerase unrolls a section of the DNA double helix and, at a rate of about thirty nucleotides per second, builds a strand of RNA complementary to it
- ▶ When finished, the DNA winds back to its stable, double-helical form

⁴From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁴

In the first step, **transcription**, the messenger molecule is made according to the information stored in DNA

- ▶ The enzyme RNA polymerase unrolls a section of the DNA double helix and, at a rate of about thirty nucleotides per second, builds a strand of RNA complementary to it
- ▶ When finished, the DNA winds back to its stable, double-helical form
- ▶ The strand of RNA, known as mRNA ("messenger" RNA), contains exactly the same information as the segment of DNA copied, still in a sequence of nucleotides

⁴From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁴

In the first step, **transcription**, the messenger molecule is made according to the information stored in DNA

- ▶ The enzyme RNA polymerase unrolls a section of the DNA double helix and, at a rate of about thirty nucleotides per second, builds a strand of RNA complementary to it
- ▶ When finished, the DNA winds back to its stable, double-helical form
- ▶ The strand of RNA, known as mRNA ("messenger" RNA), contains exactly the same information as the segment of DNA copied, still in a sequence of nucleotides
- ▶ But it is a throw-away molecule, to be used and then discarded.

⁴From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁵

In the second step, translation, the sequence of nucleotides in mRNA is read and used to link amino acids in the proper order to form a new protein

- ▶ Translation requires the combined efforts of over fifty different molecular machines

⁵From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁵

In the second step, translation, the sequence of nucleotides in mRNA is read and used to link amino acids in the proper order to form a new protein

- ▶ Translation requires the combined efforts of over fifty different molecular machines
- ▶ The actual physical matching of each nucleotide triplet with its proper amino acid is performed by another type of RNA, known as tRNA ("transfer" RNA)

⁵From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁵

In the second step, translation, the sequence of nucleotides in mRNA is read and used to link amino acids in the proper order to form a new protein

- ▶ Translation requires the combined efforts of over fifty different molecular machines
- ▶ The actual physical matching of each nucleotide triplet with its proper amino acid is performed by another type of RNA, known as tRNA ("transfer" RNA)
- ▶ Transfer RNA is made in twenty varieties, one for each amino acid

⁵From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁵

In the second step, translation, the sequence of nucleotides in mRNA is read and used to link amino acids in the proper order to form a new protein

- ▶ Translation requires the combined efforts of over fifty different molecular machines
- ▶ The actual physical matching of each nucleotide triplet with its proper amino acid is performed by another type of RNA, known as tRNA ("transfer" RNA)
- ▶ Transfer RNA is made in twenty varieties, one for each amino acid
- ▶ They are L-shaped, with the proper triplet of nucleotides at one end and the amino acid attached to the other end

⁵From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁵

In the second step, translation, the sequence of nucleotides in mRNA is read and used to link amino acids in the proper order to form a new protein

- ▶ Translation requires the combined efforts of over fifty different molecular machines
- ▶ The actual physical matching of each nucleotide triplet with its proper amino acid is performed by another type of RNA, known as tRNA ("transfer" RNA)
- ▶ Transfer RNA is made in twenty varieties, one for each amino acid
- ▶ They are L-shaped, with the proper triplet of nucleotides at one end and the amino acid attached to the other end
- ▶ A separate set of twenty different enzymes (amino-acyl tRNA synthetases) load the proper amino acid onto each type of tRNA

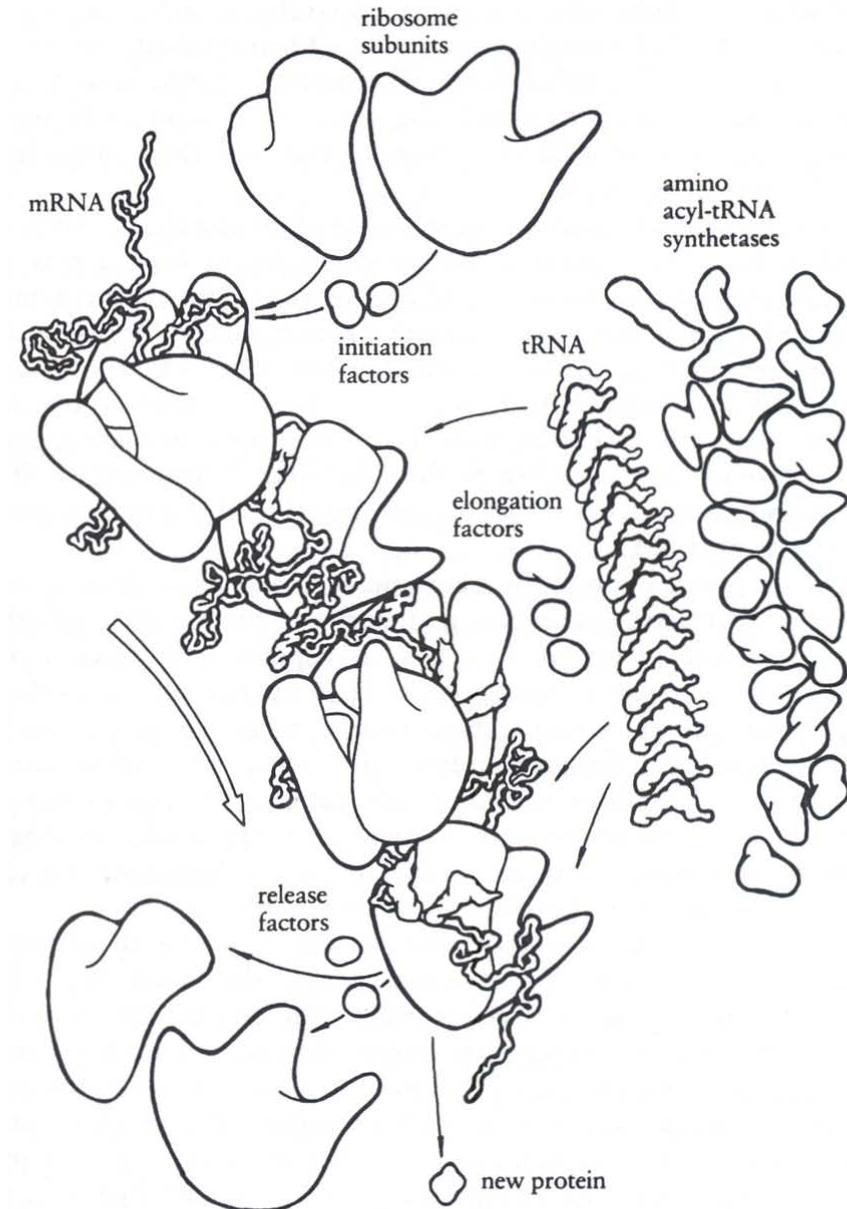
⁵From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁶

In the second step, **translation**, the sequence of nucleotides in mRNA is read and used to link amino acids in the proper order to form a new protein

The information in m-RNA is then translated into a sequence of amino acids in a new protein by the combined effort of over fifty molecular machines, as shown here (1,000,000 x)



⁶From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁷

Proteins are physically built by ribosomes, the engines of protein synthesis

- ▶ Chaperoned by proteins that initiate and terminate the process, and other proteins that inject the energy for each step, ribosomes walk down a strand of mRNA, align tRNA adapters alongside, and link up the amino acids they carry

⁷From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁷

Proteins are physically built by ribosomes, the engines of protein synthesis

- ▶ Chaperoned by proteins that initiate and terminate the process, and other proteins that inject the energy for each step, ribosomes walk down a strand of mRNA, align tRNA adapters alongside, and link up the amino acids they carry
- ▶ At a rate of about twenty amino acids per second, an average protein takes about twenty seconds to build

⁷From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁷

Proteins are physically built by ribosomes, the engines of protein synthesis

- ▶ Chaperoned by proteins that initiate and terminate the process, and other proteins that inject the energy for each step, ribosomes walk down a strand of mRNA, align tRNA adapters alongside, and link up the amino acids they carry
- ▶ At a rate of about twenty amino acids per second, an average protein takes about twenty seconds to build
- ▶ Over fifty individual protein chains and three long RNA chains combine to form these large molecular factories

⁷From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁷

Proteins are physically built by ribosomes, the engines of protein synthesis

- ▶ Chaperoned by proteins that initiate and terminate the process, and other proteins that inject the energy for each step, ribosomes walk down a strand of mRNA, align tRNA adapters alongside, and link up the amino acids they carry
- ▶ At a rate of about twenty amino acids per second, an average protein takes about twenty seconds to build
- ▶ Over fifty individual protein chains and three long RNA chains combine to form these large molecular factories
- ▶ That ribosomes are composed of both RNA and protein is provocative-perhaps a relic from the earliest cells

⁷From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Synthesis of proteins⁷

Proteins are physically built by ribosomes, the engines of protein synthesis

- ▶ Chaperoned by proteins that initiate and terminate the process, and other proteins that inject the energy for each step, ribosomes walk down a strand of mRNA, align tRNA adapters alongside, and link up the amino acids they carry
- ▶ At a rate of about twenty amino acids per second, an average protein takes about twenty seconds to build
- ▶ Over fifty individual protein chains and three long RNA chains combine to form these large molecular factories
- ▶ That ribosomes are composed of both RNA and protein is provocative-perhaps a relic from the earliest cells
- ▶ Ribosomes perform the central task of life, so they have probably remained essentially unchanged over the billions of years of evolution

⁷From: David S. Goodsell, *The machinery of life*, Springer, 1998.



Sommario

Lezione 10: Sintesi proteica

Synthesis of proteins

Central dogma: DNA makes RNA makes proteins

Genetic code



Central dogma: DNA makes RNA makes proteins

The information archive within each organism -the blueprint of potential development and activity-is the genetic material, DNA or, in some viruses, RNA

- ▶ DNA and RNA molecules are long, linear, chain molecules containing a message in a fourletter alphabet



Central dogma: DNA makes RNA makes proteins

The information archive within each organism -the blueprint of potential development and activity-is the genetic material, DNA or, in some viruses, RNA

- ▶ DNA and RNA molecules are long, linear, chain molecules containing a message in a fourletter alphabet
- ▶ Even for microorganisms the message is long, typically 10^6 characters
- ▶ Implicit in the structure of the DNA are mechanisms for self-replication and for translation of genes into proteins



Central dogma: DNA makes RNA makes proteins

The four naturally occurring nucleotides in DNA (RNA)

a	adenine
g	guanine
c	cytosine
t	thymine
(u)	(uracil)

Why DNA has thymine instead of uracil (RNA)?

Current consensus seems to indicate the liability of cytosine to easily degrade into uracil: with the use of thymine in DNA, any uracil is easily recognized as a damaged cytosine and repaired



Central dogma: DNA makes RNA makes proteins

The 20 naturally occurring amino acids in proteins⁸

Non-polar amino acids

G	glycine	A	alanine	P	proline	V	valine
I	isoleucine	L	leucine	F	phenylalanine	M	methionine

Polar amino acids

S	serine	C	cysteine	T	threonine	N	asparagine
Q	glutamine	H	histidine	Y	tyrosine	W	tryptophan

Charged amino acids

D	aspartic acid	E	glutamic acid	K	lysine	R	arginine
---	---------------	---	---------------	---	--------	---	----------

⁸The rare amino acid *selenocysteine* has the three-letter abbreviation Sec and the one-letter code U



Central dogma: DNA makes RNA makes protein

Amino acid names are frequently abbreviated to their first three letters, except for isoleucine, asparagine, glutamine and tryptophan, using Ile, Asn, Gln and Trp

Name	Symbol	Mass (-H ₂ O)	Side Chain	Occurrence (%)
<u>Alanine</u>	A, Ala	71.079	CH ₃ -	7.49
<u>Arginine</u>	R, Arg	156.188	HN=C(NH ₂)-NH- (CH ₂) ₃ -	5.22
<u>Asparagine</u>	N, Asn	114.104	H ₂ N-CO-CH ₂ -	4.53
<u>Aspartic acid</u>	D, Asp	115.089	HOOC-CH ₂ -	5.22
<u>Cysteine</u>	C, Cys	103.145	HS-CH ₂ -	1.82
<u>Glutamine</u>	Q, Gln	128.131	H ₂ N-CO-(CH ₂) ₂ -	4.11
<u>Glutamic acid</u>	E, Glu	129.116	HOOC-(CH ₂) ₂ -	6.26
<u>Glycine</u>	G, Gly	57.052	H-	7.10
<u>Histidine</u>	H, His	137.141	N=CH-NH-CH=C-CH ₂ - _____	2.23
<u>Isoleucine</u>	I, Ile	113.160	CH ₃ -CH ₂ -CH(CH ₃)-	5.45
<u>Leucine</u>	L, Leu	113.160	(CH ₃) ₂ -CH-CH ₂ -	9.06
<u>Lysine</u>	K, Lys	128.17	H ₂ N-(CH ₂) ₄ -	5.82
<u>Methionine</u>	M, Met	131.199	CH ₃ -S-(CH ₂) ₂ -	2.27
<u>Phenylalanine</u>	F, Phe	147.177	Phenyl-CH ₂ -	3.91
<u>Proline</u>	P, Pro	97.117	-N-(CH ₂) ₃ -CH- _____	5.12
<u>Serine</u>	S, Ser	87.078	HO-CH ₂ -	7.34
<u>Threonine</u>	T, Thr	101.105	CH ₃ -CH(OH)-	5.96
<u>Tryptophan</u>	W, Trp	186.213	Phenyl-NH-CH=C- CH ₂ - _____	1.32
<u>Tyrosine</u>	Y, Tyr	163.176	4-OH-Phenyl-CH ₂ -	3.25
<u>Valine</u>	V, Val	99.133	CH ₃ -CH(CH ₂)-	6.48



Sommario

Lezione 10: Sintesi proteica

Synthesis of proteins

Central dogma: DNA makes RNA makes proteins

Genetic code



Coding $\equiv \text{map} : \text{'Name'} \mapsto \text{tuple}(\text{'C'}, \text{'Cod'})$

```
1 aacode = {
2     'Alanine' : ( 'A', 'Ala' ),
3     'Arginine' : ( 'R', 'Arg' ),
4     'Asparagine' : ( 'N', 'Asn' ),
5     'AsparticAcid' : ( 'D', 'Asp' ),
6     'Cysteine' : ( 'C', 'Cys' ),
7     'Glutamine' : ( 'Q', 'Gln' ),
8     'GlutamicAcid' : ( 'E', 'Glu' ),
9     'Glycine' : ( 'G', 'Gly' ),
10    'Histidine' : ( 'H', 'His' ),
11    'Isoleucine' : ( 'I', 'Ile' ),
12    'Leucine' : ( 'L', 'Leu' ),
13    'Lysine' : ( 'K', 'Lys' ),
14    'Methionine' : ( 'M', 'Met' ),
15    'Phenylalanine' : ( 'F', 'Phe' ),
16    'Proline' : ( 'P', 'Pro' ),
17    'Serine' : ( 'S', 'Ser' ),
18    'Threonine' : ( 'T', 'Thr' ),
19    'Tryptophan' : ( 'W', 'Trp' ),
20    'Tyrosine' : ( 'Y', 'Tyr' ),
21    'Valine' : ( 'V', 'Val' ) }
```



Inverse coding

```
1 aacid = {  
2     'A' : 'Alanine',  
3     'R' : 'Arginine',  
4     'N' : 'Asparagine',  
5     'D' : 'AsparticAcid',  
6     'C' : 'Cysteine',  
7     'Q' : 'Glutamine',  
8     'E' : 'GlutamicAcid',  
9     'G' : 'Glycine',  
10    'H' : 'Histidine',  
11    'I' : 'Isoleucine',  
12    'L' : 'Leucine',  
13    'K' : 'Lysine',  
14    'M' : 'Methionine',  
15    'F' : 'Phenylalanine',  
16    'P' : 'Proline',  
17    'S' : 'Serine',  
18    'T' : 'Threonine',  
19    'W' : 'Tryptophan',  
20    'Y' : 'Tyrosine',  
21    'V' : 'Valine' }
```

```
1 aminoacid = {  
2     'Ala' : 'Alanine',  
3     'Arg' : 'Arginine',  
4     'Asn' : 'Asparagine',  
5     'Asp' : 'AsparticAcid',  
6     'Cys' : 'Cysteine',  
7     'Gln' : 'Glutamine',  
8     'Glu' : 'GlutamicAcid',  
9     'Gly' : 'Glycine',  
10    'His' : 'Histidine',  
11    'Ile' : 'Isoleucine',  
12    'Leu' : 'Leucine',  
13    'Lys' : 'Lysine',  
14    'Met' : 'Methionine',  
15    'Phe' : 'Phenylalanine',  
16    'Pro' : 'Proline',  
17    'Ser' : 'Serine',  
18    'Thr' : 'Threonine',  
19    'Trp' : 'Tryptophan',  
20    'Tyr' : 'Tyrosine',  
21    'Val' : 'Valine' }
```



Coding use

It is conventional to write nucleotides in lower case and amino acids in upper case.
Thus **atg** = adenine-thymine-guanine and **ATG** = alanine-threonine-glycine

```
1 In [9]: aacode['Phenylalanine']
2 Out[9]: ('F', 'Phe')
3
4 In [10]: aacode['Asparagine']
5 Out[10]: ('N', 'Asn')
6
7 In [11]: aacode['Phenylalanine'][0]
8 Out[11]: 'F'
9
10 In [12]: aacode['Phenylalanine'][1]
11 Out[12]: 'Phe'
12
13 In [13]: aacid['E']
14 Out[13]: 'GlutamicAcid'
15
16 In [14]: aminoacid['Cys']
17 Out[14]: 'Cysteine'
```



Genetic code

The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA) is translated into proteins (amino acid sequences) by living cells⁹

- ▶ A more precise term for the concept might be "genetic cipher"

⁹From *Wikipedia*



Genetic code

The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA) is translated into proteins (amino acid sequences) by living cells⁹

- ▶ A more precise term for the concept might be "genetic cipher"
- ▶ The code defines a mapping between tri-nucleotide sequences, called codons, and amino acids

⁹From *Wikipedia*



Genetic code

The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA) is translated into proteins (amino acid sequences) by living cells⁹

- ▶ A more precise term for the concept might be "genetic cipher"
- ▶ The code defines a mapping between tri-nucleotide sequences, called codons, and amino acids
- ▶ A triplet codon in a nucleic acid sequence usually specifies a single amino acid (though in some cases the same codon triplet in different locations can code unambiguously for two different amino acids, the correct choice at each location being determined by context)

⁹From *Wikipedia*



Genetic code

The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA) is translated into proteins (amino acid sequences) by living cells⁹

- ▶ A more precise term for the concept might be "genetic cipher"
- ▶ The code defines a mapping between tri-nucleotide sequences, called codons, and amino acids
- ▶ A triplet codon in a nucleic acid sequence usually specifies a single amino acid (though in some cases the same codon triplet in different locations can code unambiguously for two different amino acids, the correct choice at each location being determined by context)
- ▶ Because the vast majority of genes are encoded with exactly the same code (see the RNA codon table), this particular code is often referred to as the canonical or standard genetic code, or simply the genetic code, though in fact there are many variant codes

⁹From *Wikipedia*



Genetic code

The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA) is translated into proteins (amino acid sequences) by living cells⁹

- ▶ A more precise term for the concept might be "genetic cipher"
- ▶ The code defines a mapping between tri-nucleotide sequences, called codons, and amino acids
- ▶ A triplet codon in a nucleic acid sequence usually specifies a single amino acid (though in some cases the same codon triplet in different locations can code unambiguously for two different amino acids, the correct choice at each location being determined by context)
- ▶ Because the vast majority of genes are encoded with exactly the same code (see the RNA codon table), this particular code is often referred to as the canonical or standard genetic code, or simply the genetic code, though in fact there are many variant codes
- ▶ Thus the canonical genetic code is not universal

⁹From *Wikipedia*



Genetic code

The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA) is translated into proteins (amino acid sequences) by living cells⁹

- ▶ A more precise term for the concept might be "genetic cipher"
- ▶ The code defines a mapping between tri-nucleotide sequences, called codons, and amino acids
- ▶ A triplet codon in a nucleic acid sequence usually specifies a single amino acid (though in some cases the same codon triplet in different locations can code unambiguously for two different amino acids, the correct choice at each location being determined by context)
- ▶ Because the vast majority of genes are encoded with exactly the same code (see the RNA codon table), this particular code is often referred to as the canonical or standard genetic code, or simply the genetic code, though in fact there are many variant codes
- ▶ Thus the canonical genetic code is not universal
- ▶ For example, in humans, protein synthesis in mitochondria relies on a genetic code that varies from the canonical code.

⁹From *Wikipedia*

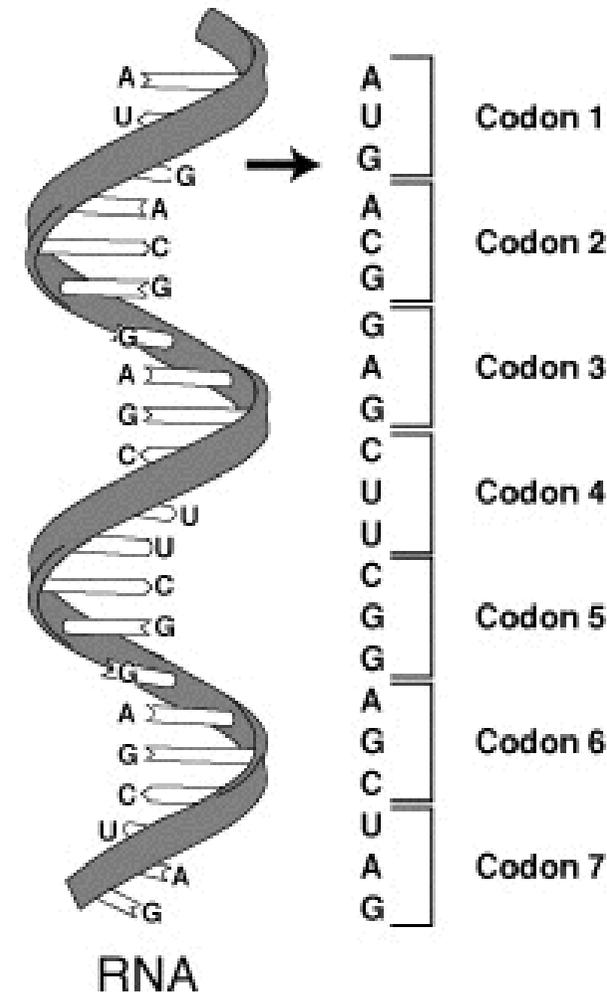


Genetic code

Those genes that code for proteins are composed of tri-nucleotide units called **codons**, each coding for a single amino acid

George Gamow postulated that a **three-letter** code must be employed to encode the 20 standard amino acids used by living cells to encode proteins

(because 3 is the smallest integer n such that 4^n is at least 20)



RNA codon table

nonpolar polar basic acidic (stop codon)

The table shows the 64 codons.

		2nd base			
		U	C	A	G
1st base	U	UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine
		UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine
		UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (Stop)	UGA Opal (Stop)
		UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (Stop)	UGG (Trp/W) Tryptophan
	C	CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine
		CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine
		CUA (Leu/L) Leucine	CCA (Pro/P) Proline	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine
		CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine
	A	AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine
		AUC (Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine
		AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
		AUG ^[A] (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
	G	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine
		GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine
		GUA (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine
		GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine



Standard genetic code (codon \mapsto Cod)

ttt	Phe	tct	Ser	tat	Tyr	tgt	Cys
ttc	Phe	tcc	Ser	tac	Tyr	tgc	Cys
tta	Leu	tca	Ser	taa	STOP	tga	STOP
ttg	Leu	tcg	Ser	tag	STOP	tgg	Trp
ctt	Leu	cct	Pro	cat	His	cgt	Arg
ctc	Leu	ccc	Pro	cac	His	cgc	Arg
cta	Leu	cca	Pro	caa	Gln	cga	Arg
ctg	Leu	ccg	Pro	cag	Gln	cgg	Arg
att	Ile	act	Thr	aat	Asn	agt	Ser
atc	Ile	acc	Thr	aac	Asn	agc	Ser
ata	Ile	aca	Thr	aaa	Lys	aga	Arg
atg	Met	acg	Thr	aag	Lys	agg	Arg
gtt	Val	gct	Ala	gat	Asp	ggt	Gly
gtc	Val	gcc	Ala	gac	Asp	ggc	Gly
gta	Val	gca	Ala	gaa	Glu	gga	Gly
gtg	Val	gcg	Ala	gag	Glu	ggg	Gly



Standard genetic code (Python dict)

```
1 genetic_code = { 'ttt': 'Phe', 'tct': 'Ser', 'tat': 'Tyr',  
    'tgt': 'Cys', 'ttc': 'Phe', 'tcc': 'Ser', 'tac': 'Tyr', '  
tgc': 'Cys', 'tta': 'Leu', 'tca': 'Ser', 'taa': 'STOP', '  
tga': 'STOP', 'ttg': 'Leu', 'tcg': 'Ser', 'tag': 'STOP', '  
'tgg': 'Trp', 'ctt': 'Leu', 'cct': 'Pro', 'cat': 'His', '  
cgt': 'Arg', 'ctc': 'Leu', 'ccc': 'Pro', 'cac': 'His', '  
cgc': 'Arg', 'cta': 'Leu', 'cca': 'Pro', 'caa': 'Gln', '  
cga': 'Arg', 'ctg': 'Leu', 'ccg': 'Pro', 'cag': 'Gln', '  
cgg': 'Arg', 'att': 'Ile', 'act': 'Thr', 'aat': 'Asn', '  
agt': 'Ser', 'atc': 'Ile', 'acc': 'Thr', 'aac': 'Asn', '  
agc': 'Ser', 'ata': 'Ile', 'aca': 'Thr', 'aaa': 'Lys', '  
aga': 'Arg', 'atg': 'Met', 'acg': 'Thr', 'aag': 'Lys', '  
agg': 'Arg', 'gtt': 'Val', 'gct': 'Ala', 'gat': 'Asp', '  
ggt': 'Gly', 'gtc': 'Val', 'gcc': 'Ala', 'gac': 'Asp', '  
ggc': 'Gly', 'gta': 'Val', 'gca': 'Ala', 'gaa': 'Glu', '  
gga': 'Gly', 'gtg': 'Val', 'gcg': 'Ala', 'gag': 'Glu', '  
ggg': 'Gly' }
```



Example of translation

```
1 In [145]: RNA_strand = 'atgcatccctttaat'
2
3 In [146]: RNA_strand = array(list(RNA_strand))
4
5 In [147]: RNA_strand.shape
6 Out [147]: (15,)
7
8 In [148]: RNA_strand.size
9 Out [148]: 15
10
11 In [149]: RNA_strand = RNA_strand.reshape(RNA_strand.
12         size/3, 3)
13 Out [149]:
14 array([[ 'a', 't', 'g'],
15        [ 'c', 'a', 't'],
16        [ 'c', 'c', 'c'],
17        [ 't', 't', 't'],
18        [ 'a', 'a', 't']],
19        dtype='<S1')
```



Example of translation

Let us define yet another dictionary, allowing for conversion from the 3-character code to the 1-character code for amino acids

```
1 from numpy import *
2 code = { 'Ala':'A', 'Arg':'R', 'Asn':'N', 'Asp':'D', '
          Cys':'C', 'Gln':'Q', 'Glu':'E', 'Gly':'G', 'His':'H',
          'Ile':'I', 'Leu':'L', 'Lys':'K', 'Met':'M', 'Phe':'F',
          'Pro':'P', 'Ser':'S', 'Thr':'T', 'Trp':'W', 'Tyr':
          'Y', 'Val':'V' }
```

therefore we have

```
1 genetic_code['atg'] ≡ 'Met'
2
3 code[genetic_code['atg']] ≡ 'M'
```



Preparatory work

Remove the uracil and set the sequence to lower case

```
1 def rna2dna (nucleotideList):  
2     return [n if n != 'u' else 't' for n in  
3             nucleotideList]  
4 rna2dna(list('augaaaugaau')) ≡ ['a', 't', 'g', 'a', 'a',  
   , 'a', 'a', 't', 'g', 'a', 'a', 't']
```

```
1 'ATGAAAATGAAT'.lower() ≡ 'atgaaaatgaat'
```

```
1 '1234567890'[:10/3*3] ≡ '123456789'
```

We need to make some **curation of the input**, in order to:

- ▶ transform from 'u' to 't' (as in the standard genetic code)
- ▶ transform nucleotides from UPPER to lower case
- ▶ truncate the nucleotide sequence at the (maximum) multiple of 3



Translation function

works like a ribosome!!... :o)

```
1 def translation (strand):
2     def curation (strand):
3         strand = strand[:len(strand)/3*3]
4         return array(rna2dna(list(strand.lower())))
5     strand = curation(strand)
6     strand = strand.reshape(strand.size/3,3)
7     codons = map(''.join, strand)
8     return [genetic_code[c] for c in codons]
9
10 def polypeptide (DNAstrand):
11     return ''.join([code[peptide]
12     for peptide in translation(DNAstrand)])
13
14 strand = 'atgaaaatgaataaaagtctcatcgtcc\
15 tctgtttatcagcagggttactggcaagc'
16 translation(strand) ≡ ['Met', 'Lys', 'Met', 'Asn', 'Lys',
17     , 'Ser', 'Leu', 'Ile', 'Val', 'Leu', 'Cys', 'Leu', '
18     Ser', 'Ala', 'Gly', 'Leu', 'Leu', 'Ala', 'Ser']
19
20 polypeptide(strand) ≡ 'MKMNKSLIVLCLSAGLLAS'
```



Example of translation

Take a quite common virus, in this period ...

strand = dna of H1N1 virus

```
1 polypeptide (strand) ≡ '  
2 TVTHSVNLLDKHNGKLCKLRGVAPLHLGKCNIAGWILGNPECESLSTASSWS  
3 YIVETSSSDNGTCYPGDFIDYEELREQLSSVSSFERFEIFPKTSSWPNHDSNK  
4 GVTAAACPHAGAKSFYKNLIWLVKKGNYPKLSKSYINDKGKEVLVLWGIHHPS  
5 TSADQQSLYQNADAYVFBVGTSRYSKKFKPEIAIRPKVRDQEGRMNYYWTLVEP  
6 GDKITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTS  
7 LPFQNIHPITIGKCPKYVKSTKLRLATGLRNVP SIQSRGLFGAIAGFIEGGWT  
8 GMVDGWYGYHHQNEQGSGYAADLKSTQNAIDEITNKVNSVIEKMNTQFTAVGK  
9 EFNHLEKRIENLNKKVDDGFLDIWTYNAELLVLLLENERTLDYHDSNVKKLYEK  
10 VRSQLKNNAKEIGNGCFEFYHKCDNTCMESVKNGTYDYPKYSEEAKLNREEID  
11 GVKLESTRIYQILAIYSTVASSLVLVSLGAI SFWM'
```

