# Lezione 10

## Bioinformatica

Mauro Ceccanti[‡] e Alberto Paoluzzi[†]

[†]Dip. Informatica e Automazione – Università "Roma Tre"
[‡]Dip. Medicina Clinica – Università "La Sapienza"



## Sommario



## Synthesis of proteins[1]

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?)

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

# Synthesis of proteins[2]

Each different protein is made according to a blueprint

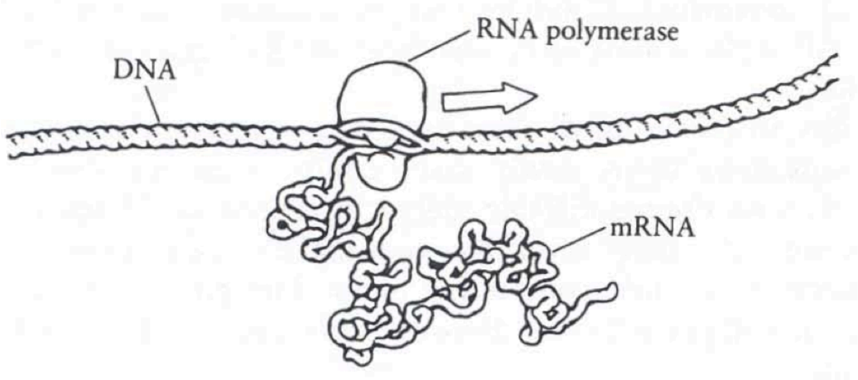
- The unique linear sequence of amino acids in a protein is encoded in the linear sequence of nucleotides in DNA
- Because DNA is composed of only four types of nucleotides, compared to the twenty types of amino acids in protein, there cannot be a one-to-one correspondence of amino acid to nucleotide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

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

# Synthesis of proteins[3]

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




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

# Synthesis of proteins[4]

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.

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

# Synthesis of proteins[5]

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

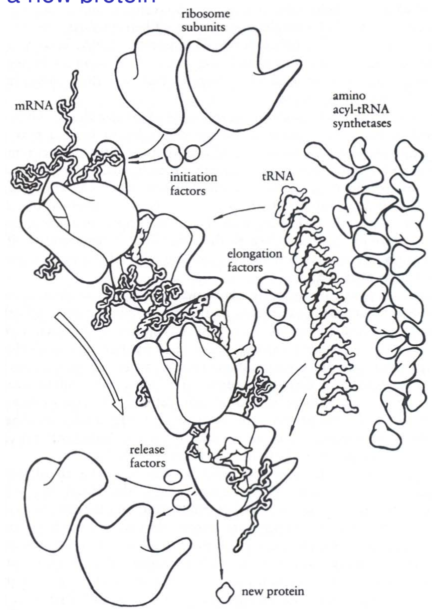
- Translation requires the combined efforts of over fifty different molecular machines
- The actual physical matching of each nucleotide triplet with its proper amino acid is performed by another type of RNA, known as tRNA ("transfer" RNA)
- Transfer RNA is made in twenty varieties, one for each amino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

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

# Synthesis of proteins[6]

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)




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

# Synthesis of proteins[7]

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

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

# Sommario



# 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
- The double helix, and its internal self-complementarity providing for accurate replication, are well known
- Near perfect replication is essential for stability of inheritance; but some imperfect replication, or mechanism for import of foreign genetic material, is also essential, else evolution could not take place in asexual organisms.

# 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 proteins[8]

**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 |

[8]The 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 (-$H_2O$) | Side Chain | Occurence (%) |
|---|---|---|---|---|
| Alanine | A, Ala | 71.079 | CH3- | 7.49 |
| Arginine | R, Arg | 156.188 | HN=C(NH2)-NH-(CH2)3- | 5.22 |
| Asparagine | N, Asn | 114.104 | H2N-CO-CH2- | 4.53 |
| Aspartic acid | D, Asp | 115.089 | HOOC-CH2- | 5.22 |
| Cysteine | C, Cys | 103.145 | HS-CH2- | 1.82 |
| Glutamine | Q, Gln | 128.131 | H2N-CO-(CH2)2- | 4.11 |
| Glutamic acid | E, Glu | 129.116 | HOOC-(CH2)2- | 6.26 |
| Glycine | G, Gly | 57.052 | H- | 7.10 |
| Histidine | H, His | 137.141 | N=CH-NH-CH=C-CH2-<br>\|__________\| | 2.23 |
| Isoleucine | I, Ile | 113.160 | CH3-CH2-CH(CH3)- | 5.45 |
| Leucine | L, Leu | 113.160 | (CH3)2-CH-CH2- | 9.06 |
| Lysine | K, Lys | 128.17 | H2N-(CH2)4- | 5.82 |
| Methionine | M, Met | 131.199 | CH3-S-(CH2)2- | 2.27 |
| Phenylalanine | F, Phe | 147.177 | Phenyl-CH2- | 3.91 |
| Proline | P, Pro | 97.117 | -N-(CH2)3-CH-<br>\|_________\| | 5.12 |
| Serine | S, Ser | 87.078 | HO-CH2- | 7.34 |
| Threonine | T, Thr | 101.105 | CH3-CH(OH)- | 5.96 |
| Tryptophan | W, Trp | 186.213 | Phenyl-NH-CH=C-CH2-<br>\|__________\| | 1.32 |
| Tyrosine | Y, Tyr | 163.176 | 4-OH-Phenyl-CH2- | 3.25 |
| Valine | V, Val | 99.133 | CH3-CH(CH2)- | 6.48 |

# Sommario

## Coding $\equiv$ *map* : *'Name'* $\mapsto$ *tuple('C', 'Cod')*

```
aacode = {
    'Alanine' : ( 'A', 'Ala' ),
    'Arginine' : ( 'R', 'Arg' ),
    'Asparagine' : ( 'N', 'Asn' ),
    'AsparticAcid' : ( 'D', 'Asp' ),
    'Cysteine' : ( 'C', 'Cys' ),
    'Glutamine' : ( 'Q', 'Gln' ),
    'GlutamicAcid' : ( 'E', 'Glu' ),
    'Glycine' : ( 'G', 'Gly' ),
    'Histidine' : ( 'H', 'His' ),
    'Isoleucine' : ( 'I', 'Ile' ),
    'Leucine' : ( 'L', 'Leu' ),
    'Lysine' : ( 'K', 'Lys' ),
    'Methionine' : ( 'M', 'Met' ),
    'Phenylalanine' : ( 'F', 'Phe' ),
    'Proline' : ( 'P', 'Pro' ),
    'Serine' : ( 'S', 'Ser' ),
    'Threonine' : ( 'T', 'Thr' ),
    'Tryptophan' : ( 'W', 'Trp' ),
    'Tyrosine' : ( 'Y', 'Tyr' ),
    'Valine' : ( 'V', 'Val' )   }
```

## Inverse coding

```
aacid = {
    'A' : 'Alanine',
    'R' : 'Arginine',
    'N' : 'Asparagine',
    'D' : 'AsparticAcid',
    'C' : 'Cysteine',
    'Q' : 'Glutamine',
    'E' : 'GlutamicAcid',
    'G' : 'Glycine',
    'H' : 'Histidine',
    'I' : 'Isoleucine',
    'L' : 'Leucine',
    'K' : 'Lysine',
    'M' : 'Methionine',
    'F' : 'Phenylalanine',
    'P' : 'Proline',
    'S' : 'Serine',
    'T' : 'Threonine',
    'W' : 'Tryptophan',
    'Y' : 'Tyrosine',
    'V' : 'Valine'   }
```

```
aminoacid = {
    'Ala' : 'Alanine',
    'Arg' : 'Arginine',
    'Asn' : 'Asparagine',
    'Asp' : 'AsparticAcid',
    'Cys' : 'Cysteine',
    'Gln' : 'Glutamine',
    'Glu' : 'GlutamicAcid',
    'Gly' : 'Glycine',
    'His' : 'Histidine',
    'Ile' : 'Isoleucine',
    'Leu' : 'Leucine',
    'Lys' : 'Lysine',
    'Met' : 'Methionine',
    'Phe' : 'Phenylalanine',
    'Pro' : 'Proline',
    'Ser' : 'Serine',
    'Thr' : 'Threonine',
    'Trp' : 'Tryptophan',
    'Tyr' : 'Tyrosine',
    '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

```
In [9]: aacode['Phenylalanine']
Out[9]: ('F', 'Phe')

In [10]: aacode['Asparagine']
Out[10]: ('N', 'Asn')

In [11]: aacode['Phenylalanine'][0]
Out[11]: 'F'

In [12]: aacode['Phenylalanine'][1]
Out[12]: 'Phe'

In [13]: aacid['E']
Out[13]: 'GlutamicAcid'

In [14]: aminoacid['Cys']
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 cells[9]

- 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.

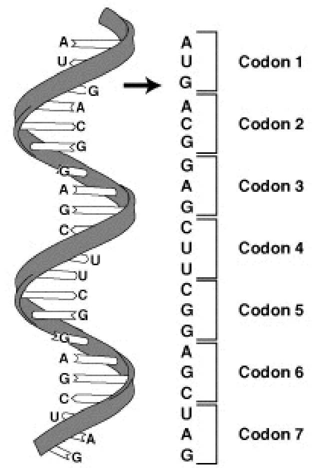
[9] From *Wikipedia*

# 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

Ribonucleic acid

# RNA codon table

nonpolar | polar | basic | acidic | (stop codon)

The table shows the 64 codons.

| 1st base | | 2nd base: U | C | A | G |
|---|---|---|---|---|---|
| | U | UUU (Phe/F) Phenylalanine | UCU (Ser/S) Serine | UAU (Tyr/Y) Tyrosine | UGU (Cys/C) Cysteine |
| | U | UUC (Phe/F) Phenylalanine | UCC (Ser/S) Serine | UAC (Tyr/Y) Tyrosine | UGC (Cys/C) Cysteine |
| | U | UUA (Leu/L) Leucine | UCA (Ser/S) Serine | UAA Ochre (*Stop*) | UGA Opal (*Stop*) |
| | U | 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 |
| | C | CUC (Leu/L) Leucine | CCC (Pro/P) Proline | CAC (His/H) Histidine | CGC (Arg/R) Arginine |
| | C | CUA (Leu/L) Leucine | CCA (Pro/P) Proline | CAA (Gln/Q) Glutamine | CGA (Arg/R) Arginine |
| | C | 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 |
| | A | AUC (Ile/I) Isoleucine | ACC (Thr/T) Threonine | AAC (Asn/N) Asparagine | AGC (Ser/S) Serine |
| | A | AUA (Ile/I) Isoleucine | ACA (Thr/T) Threonine | AAA (Lys/K) Lysine | AGA (Arg/R) Arginine |
| | A | 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 |
| | G | GUC (Val/V) Valine | GCC (Ala/A) Alanine | GAC (Asp/D) Aspartic acid | GGC (Gly/G) Glycine |
| | G | GUA (Val/V) Valine | GCA (Ala/A) Alanine | GAA (Glu/E) Glutamic acid | GGA (Gly/G) Glycine |
| | G | 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`)

```
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

```
In [145]: RNA_strand = 'atgcatccctttaat'

In [146]: RNA_strand = array(list(RNA_strand))

In [147]: RNA_strand.shape
Out[147]: (15,)

In [148]: RNA_strand.size
Out[148]: 15

In [149]: RNA_strand = RNA_strand.reshape(RNA_strand.size/3,3)
Out[149]:
array([['a', 't', 'g'],
       ['c', 'a', 't'],
       ['c', 'c', 'c'],
       ['t', 't', 't'],
       ['a', 'a', 't']],
      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

```
from numpy import *
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

```
genetic_code['atg'] ≡ 'Met'

code[genetic_code['atg']] ≡ 'M'
```

# Preparatory work

Remove the uracil and set the sequence to lower case

```
def rna2dna (nucleotideList):
    return [n if n != 'u' else 't' for n in nucleotideList]

rna2dna(list('augaaaaugaau')) ≡ ['a', 't', 'g', 'a', 'a', 'a', 'a', 't', 'g', 'a', 'a', 't']
```

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

```
'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)

```
def translation (strand):
    def curation (strand):
        strand = strand[:len(strand)/3*3]
        return array(rna2dna(list(strand.lower())))
    strand = curation(strand)
    strand = strand.reshape(strand.size/3,3)
    codons = map(''.join, strand)
    return [genetic_code[c] for c in codons]

def polypeptide (DNAstrand):
    return ''.join([code[peptide]
    for peptide in translation(DNAstrand)])

strand = 'atgaaaatgaataaaagtctcatcgtcc\
tctgtttatcagcagggtactggcaagc'
translation(strand) ≡ ['Met', 'Lys', 'Met', 'Asn', 'Lys', 'Ser', 'Leu', 'Ile', 'Val', 'Leu', 'Cys', 'Leu', 'Ser', 'Ala', 'Gly', 'Leu', 'Leu', 'Ala', 'Ser']

polypeptide(strand) ≡ 'MKMNKSLIVLCLSAGLLAS'
```

# Example of translation

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

strand = dna of H1N1 virus

```
polypedtide(strand) ≡ '
TVTHSVNLLEDKHNGKLCKLRGVAPLHLGKCNIAGWILGNPECESLSTASSWS
YIVETSSSDNGTCYPGDFIDYEELREQLSSVSSFERFEIFPKTSSWPNHDSNK
GVTAACPHAGAKSFYKNLIWLVKKGNSYPKLSKSYINDKGKEVLVLWGIHHPS
TSADQQSLYQNADAYVFVGTSRYSKKFKPEIAIRPKVRDQEGRMNYYWTLVEP
GDKITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTS
LPFQNIHPITIGKCPKYVKSTKLRLATGLRNVPSIQSRGLFGAIAGFIEGGWT
GMVDGWYGYHHQNEQGSGYAADLKSTQNAIDEITNKVNSVIEKMNTQFTAVGK
EFNHLEKRIENLNKKVDDGFLDIWTYNAELLVLLENERTLDYHDSNVKKLYEK
VRSQLKNNAKEIGNGCFEFYHKCDNTCMESVKNGTYDYPKYSEEAKLNREEID
GVKLESTRIYQILAIYSTVASSLVLVVSLGAISFWM'
```