Algorithms for Computational Biology

Zsuzsanna Lipták

Masters in Molecular and Medical Biotechnology
a.a. 2015/16, fall term

Strings and Sequences in Biology

Strings in molecular biology

Strings are finite sequences over an alphabet Σ (also called *sequences*).

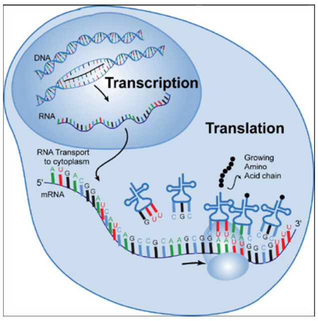
- DNA (characters: nucleotides) $\Sigma = \{\texttt{A,C,G,T}\}$
- RNA (characters: nucleotides) $\Sigma = \{\texttt{A,C,G,U}\}$
- proteins (characters: peptides) $\Sigma = \{\text{A,C,D,E,F,}\ldots\text{,W,Y}\}$
- many other problems in molecular biology can be modelled by strings (e.g. gene order, SNPs, haplotypes, ...)

DNA: nucleotides

```
5' ...AACAGTACCATGCTAGGTCAATCGA...3'
3' ...TTGTCATGGTACGATCCAGTTAGCT...5'
```

- 4 characters: `A C G T`: adenine, cytosine, guanine, thymine (bases, nucleotides)
- orientation (read from 5' to 3' end)
- length measured in bp (base pairs)
- double stranded, the two strands are *antiparallel*
- `A - T` and `C - G` complementary (Watson-Crick pairs)
- reverse complement: $(\texttt{ACCTG})^{rc} = \texttt{CAGGT}$

The central dogma of molecular biology



source: Wonderwikikids.com

DNA: nucleotides

```
5' ...AACAGTACCATGCTAGGTCAATCGA...3'
3' ...TTGTCATGGTACGATCCAGTTAGCT...5'
```

- during transcription, one strand is copied into mRNA (messenger RNA), except all T's are replaced by U's
- the strand which is identical to the mRNA is called *coding* strand
- the other strand (the one which is used for the transcription) is called *template* strand
- Both strands can be used as coding strands (for different genes).
- Some DNA strings are circular: bacterial DNA, mitochondrial DNA.

RNA: nucleotides

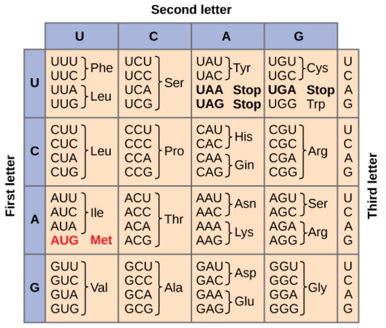
- like DNA, except:
- 4 characters: `A C U G`: adenine, cytosine, uracil, guanine (U instead of T)
- RNA is single-stranded
- builds double stranded hybrids with DNA
- RNA folds upon itself (makes complex 3-dim structures), using the Watson-Crick pairs and other bonds (RNA folding)

Protein: Amino acids

There are 20 common amino acids (aa's); two systems of abbreviations are used: 3-letter-code and 1-letter-code. We usually use the 1-letter-code.

alanine	Ala	A	leucine	Leu	L
arginine	Arg	R	lysine	Lys	K
asparagine	Asn	N	methionine	Met	M
aspartic acid	Asp	D	phenylalanine	Phe	F
cysteine	Cys	C	proline	Pro	P
glutamine	Gln	Q	serine	Ser	S
glutamic acid	Glu	E	threonine	Thr	T
glycine	Gly	G	tryptophan	Trp	W
histidine	His	H	tyrosine	Tyr	Y
isoleucine	Ile	I	valine	Val	V

The genetic code



Second letter

First letter	U	C	A	G	Third letter
U	UUU UUC } Phe UUA UUG } Leu	UCU UCC UCA UCG } Ser	UAU UAC } Tyr **UAA Stop** **UAG Stop**	UGU UGC } Cys **UGA Stop** UGG Trp	U C A G
C	CUU CUC CUA CUG } Leu	CCU CCC CCA CCG } Pro	CAU CAC } His CAA CAG } Gln	CGU CGC CGA CGG } Arg	U C A G
A	AUU AUC AUA } Ile **AUG Met**	ACU ACC ACA ACG } Thr	AAU AAC } Asn AAA AAG } Lys	AGU AGC } Ser AGA AGG } Arg	U C A G
G	GUU GUC GUA GUG } Val	GCU GCC GCA GCG } Ala	GAU GAC } Asp GAA GAG } Glu	GGU GGC GGA GGG } Gly	U C A G

source: Wikimedia commons

The genetic code

- standard genetic code (some organisms use a different one)
- 3 different reading frames for translation: The DNA sequence

  `5' ...TATTCGAATCGGC...3'`

  can be translated in 3 different ways, leading to different aa sequences.
- *degeneracy of the genetic code*
- silent mutations

The genetic code

- standard genetic code (some organisms use a different one)
- 3 different reading frames for translation: The DNA sequence

  `5' ...TATTCGAATCGGC...3'`

  can be translated in 3 different ways, leading to different aa sequences.
- *degeneracy of the genetic code*: 64 codons but only 20 aa's plus stop codon
- silent mutations

The genetic code

- standard genetic code (some organisms use a different one)
- 3 different reading frames for translation: The DNA sequence

  `5' ...TATTCGAATCGGC...3'`

  can be translated in 3 different ways, leading to different aa sequences.
- *degeneracy of the genetic code*: 64 codons but only 20 aa's plus stop codon
- silent mutations: if third position mutates, this often does not alter the aa

The genetic code

Exercise:
Translate this DNA sequence according to the 3 different reading frames:

`5' ...TATTCGAATCGGC...3'`