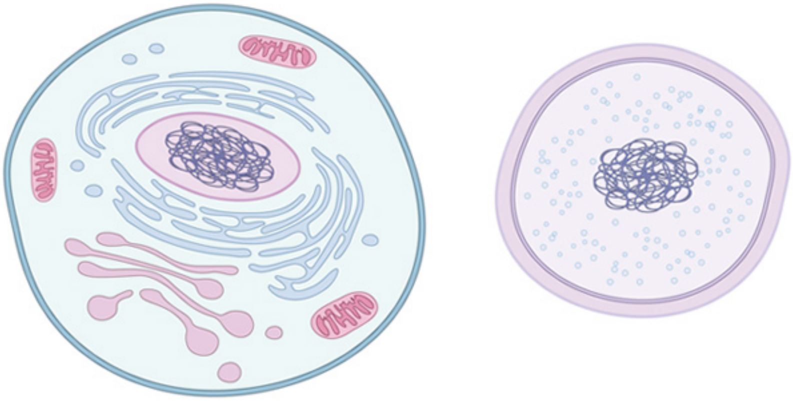
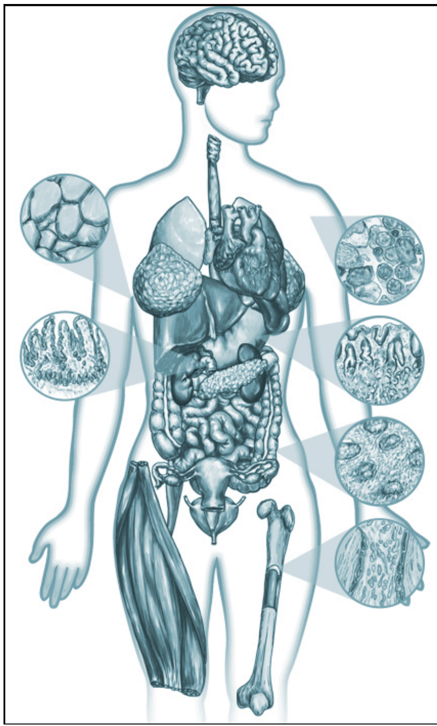
“Modern biology is undergoing an historical transformation, becoming – among other things – increasingly data driven. A combination of statistical, computational, and biological methods has become the norm in modern genomic research. Of course this is at odds with the standard organization of university curricula, which typically focus on only one of these three subjects. Yet, the importance of the algorithms typical of this field can only be appreciated within their biological context, their results can only be interpreted within a statistical framework, and a basic knowledge of all three areas is a necessary condition for any research project.”

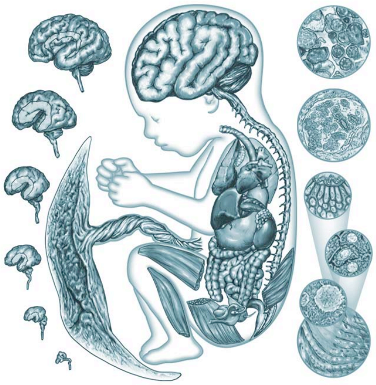
-- Nello Cristianini

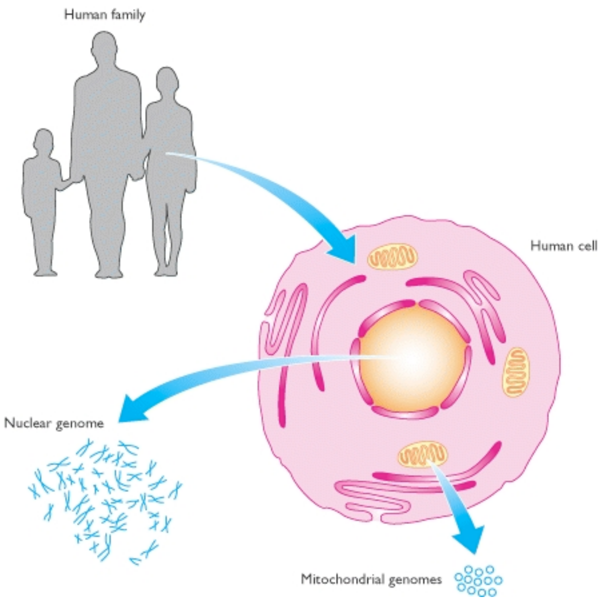
How do cells work?





Every cell contains essentially the same DNA, but is distinct



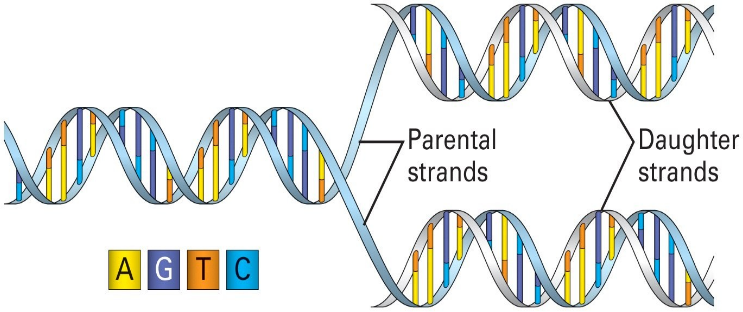


DNA (Deoxyribonucleic Acid)

- DNA holds your specific code for every part of your body. It is the collection of recipe books.
- A gene is made of a long strand of DNA.
- There are about 30,000 genes in your DNA.

DNA

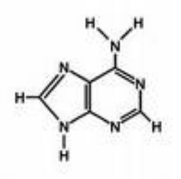
DNA: The Code of Life



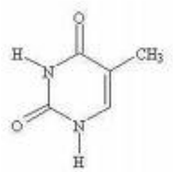
- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.

Bases are Important!

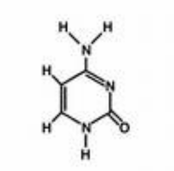
- There are four bases:



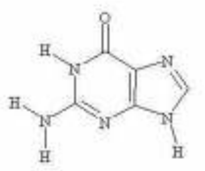
Adenine
A



Thymine
T



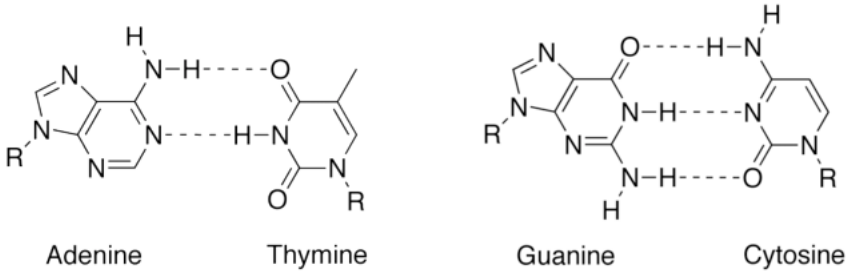
Cytosine
C

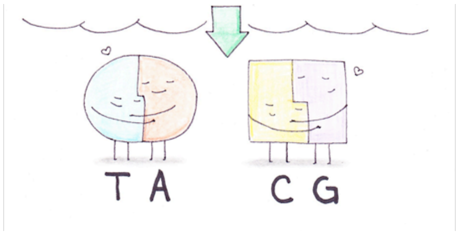


Guanine
G

- The order of these bases along a strand of DNA codes for life.

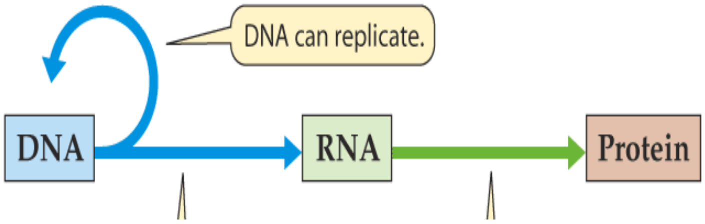
Pairing is stereotypic





The Flow of Information

Replication



Transcription

Translation

A DNA sequence: as a FASTA file

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1)
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCG
CCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACC
ACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGA
CGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACA
AGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCC
GAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCG
TTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGT
ACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC
```

An RNA sequence: as a FASTA file

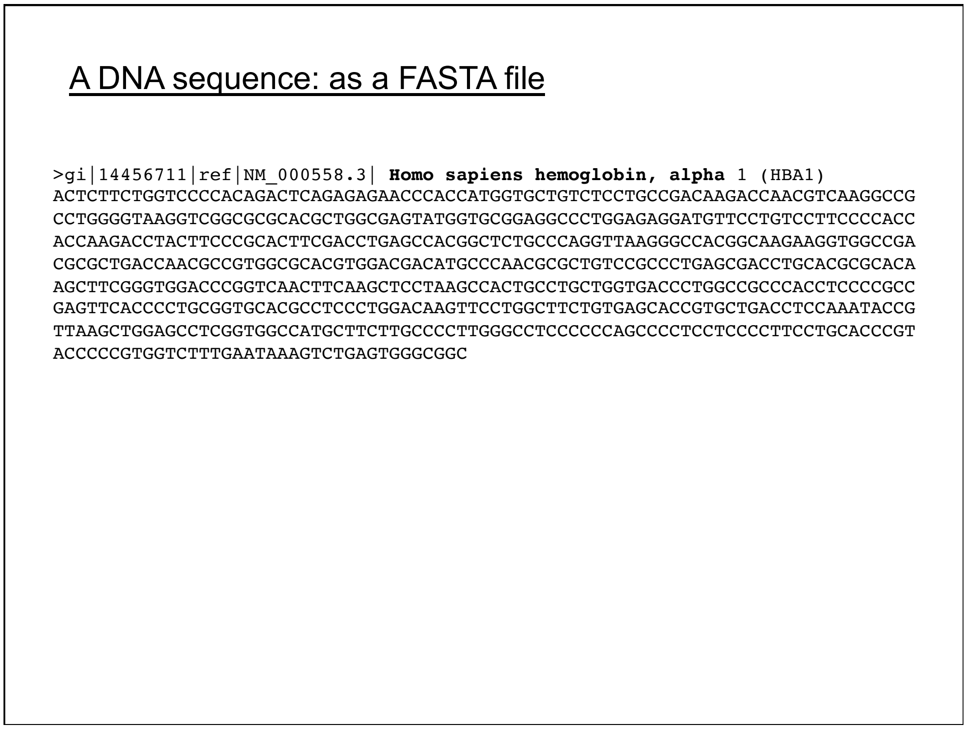
```
> HBA1, mRNA
ACUCUUCUGGUCCCCACAGACUCAGAGAGAACCCACCAUGGUGCUGUCUCCUGCCGACAAGACCAACGUCAAGGCCG
CCUGGGGUAAGGUCGGCGCGCACGCUGGCGAGUAUGGUGCGGAGGCCCUGGAGAGGAUGUUCCUGUCCUUCCCCACC
ACCAAGACCUACUUCCCGCACUUCGACCUGAGCCACGGCUCUGCCCAGGUUAAGGGCCACGGCAAGAAGGUGGCCGA
CGCGCUGACCAACGCCGUGGCGCACGUGGACGACAUGCCCAACGCGCUGUCCGCCCUGAGCGACCUGCACGCGCACA
AGCUUCGGGUGGACCCGGUCAACUUCAAGCUCCUAAGCCACUGCCUGCUGGUGACCCUGGCCGCCCACCUCCCCGCC
GAGUUCACCCCUGCGGUGCACGCCUCCCUGGACAAGUUCCUGGCUUCUGUGAGCACCGUGCUGACCUCCAAAUACCG
UUAAGCUGGAGCCUCGGUGGCCAUGCUUCUUGCCCCUUGGGCCUCCCCCCAGCCCCUCCUCCCCUUCCUGCACCCGU
ACCCCCGUGGUCUUUGAAUAAAGUCUGAGUGGGCGGC
```

Protein sequence: as a FASTA file

```
>gi|4504347|ref|NP_000549.1| alpha 1 globin [Homo sapiens]
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAH
VDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

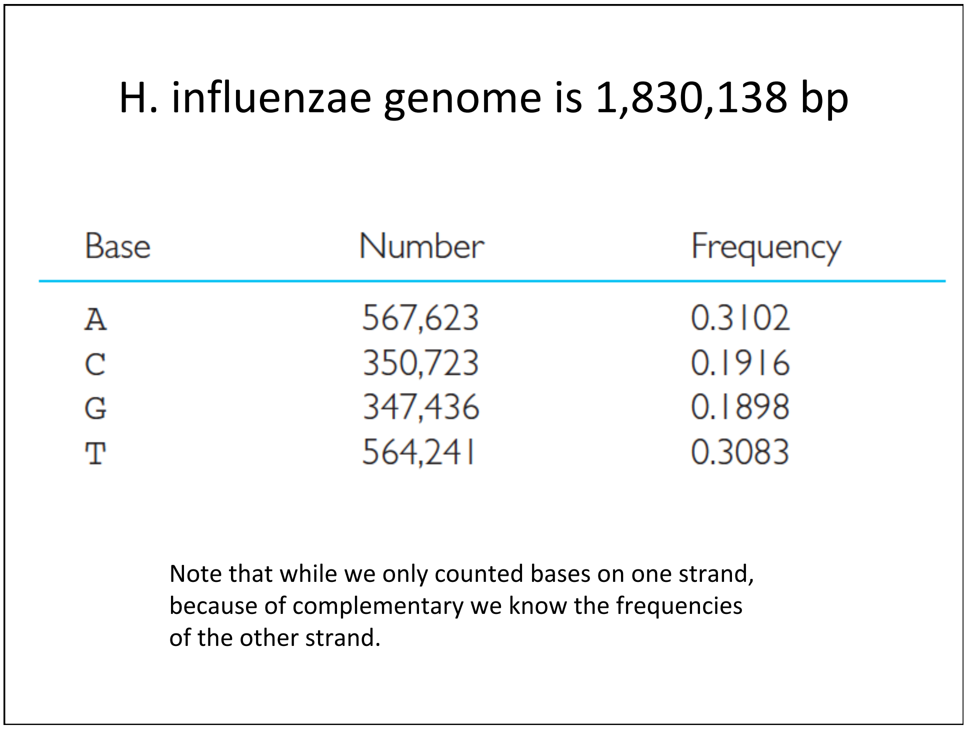
DNA sequence: as a FASTA file

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCC
GCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCAC
CACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCG
ACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCAC
AAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGC
CGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACC
GTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCC
GTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC
```



A DNA sequence: as a FASTA file

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1)
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCG
CCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACC
ACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGA
CGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACA
AGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCC
GAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCG
TTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGT
ACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC
```



H. influenzae genome is 1,830,138 bp

Base	Number	Frequency
A	567,623	0.3102
C	350,723	0.1916
G	347,436	0.1898
T	564,241	0.3083

Note that while we only counted bases on one strand, because of complementary we know the frequencies of the other strand.