Einführung in die Bioinformatik

Wintersemester 2012/13
16:00-16:45 Hörsaal N100 B3

Peter Güntert

Literatur

- *Jean-Michel Claverie, Cedric Notredame: Bioinformatics for Dummies, 2nd ed. (2007)*
- Arthur M. Lesk: Introduction to Bioinformatics (2008)
- Marketa Zvelebil, Jeremy O. Baum: Understanding Bioinformatics (2008)
- Jonathan Pevsner: Bioinformatics and Functional Genomics (2009)
- Michael S. Waterman: Introduction to Computational Biology (1995)
- R. Durbin, S. Eddy, A. Krogh, G. Mitchison: Biological Sequence Analysis (1998)

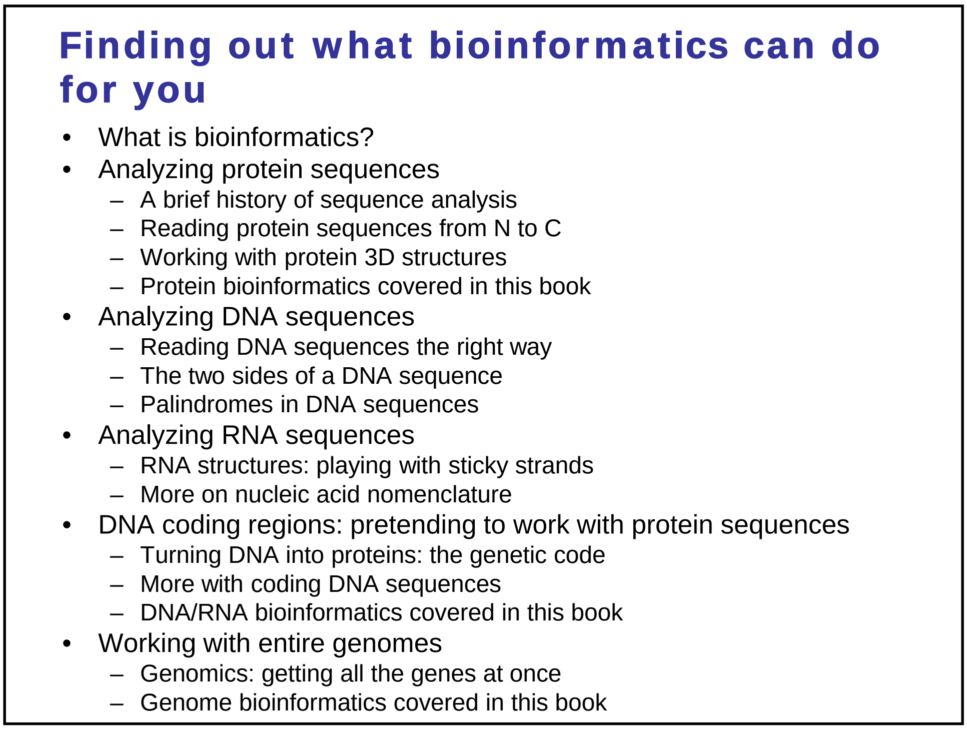
BCDS Seminar

Biochemische Datenbanken und Software

biokemika.uni-frankfurt.de/wiki/Portal:Seminare/BCDS-Seminar
(wird noch aktualisiert)

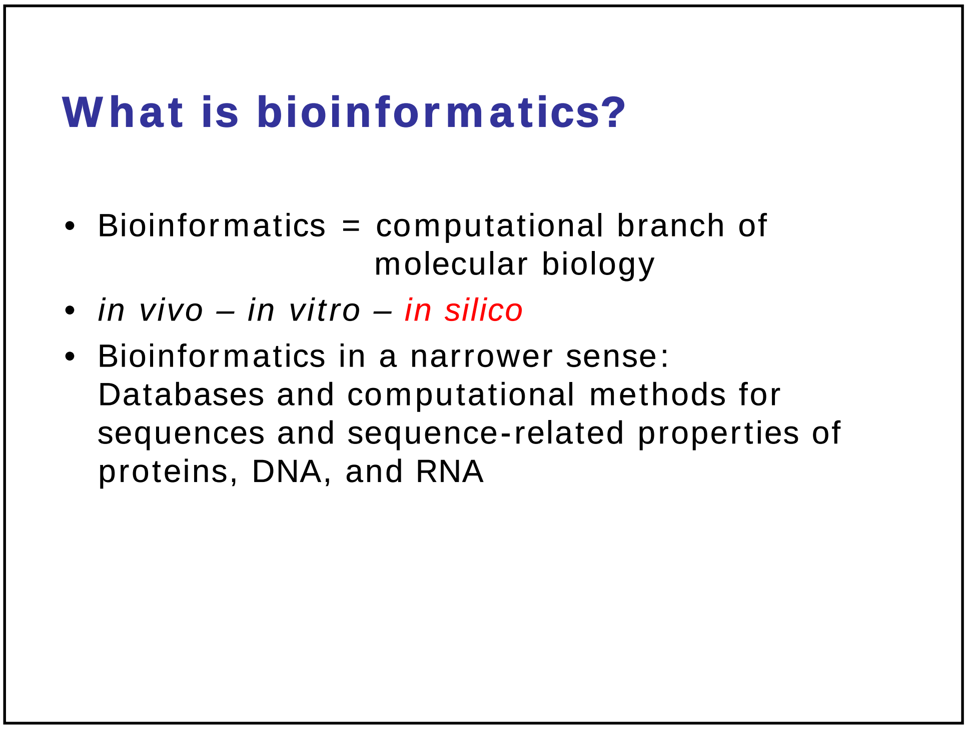
17.11. 2012– 8.12.2012, jeweils samstags, 9-18 Uhr
Beilstein Zentrum, Raum C

Anmeldung ab Montag 22.10.2012



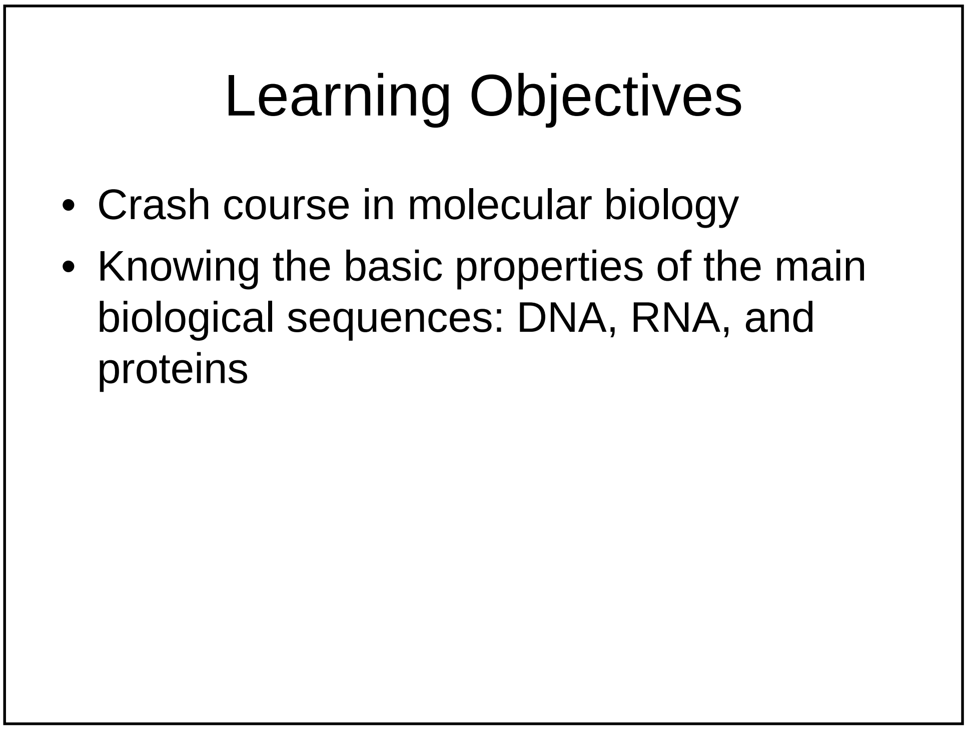
Finding out what bioinformatics can do for you

- What is bioinformatics?
- Analyzing protein sequences
  - A brief history of sequence analysis
  - Reading protein sequences from N to C
  - Working with protein 3D structures
  - Protein bioinformatics covered in this book
- Analyzing DNA sequences
  - Reading DNA sequences the right way
  - The two sides of a DNA sequence
  - Palindromes in DNA sequences
- Analyzing RNA sequences
  - RNA structures: playing with sticky strands
  - More on nucleic acid nomenclature
- DNA coding regions: pretending to work with protein sequences
  - Turning DNA into proteins: the genetic code
  - More with coding DNA sequences
  - DNA/RNA bioinformatics covered in this book
- Working with entire genomes
  - Genomics: getting all the genes at once
  - Genome bioinformatics covered in this book



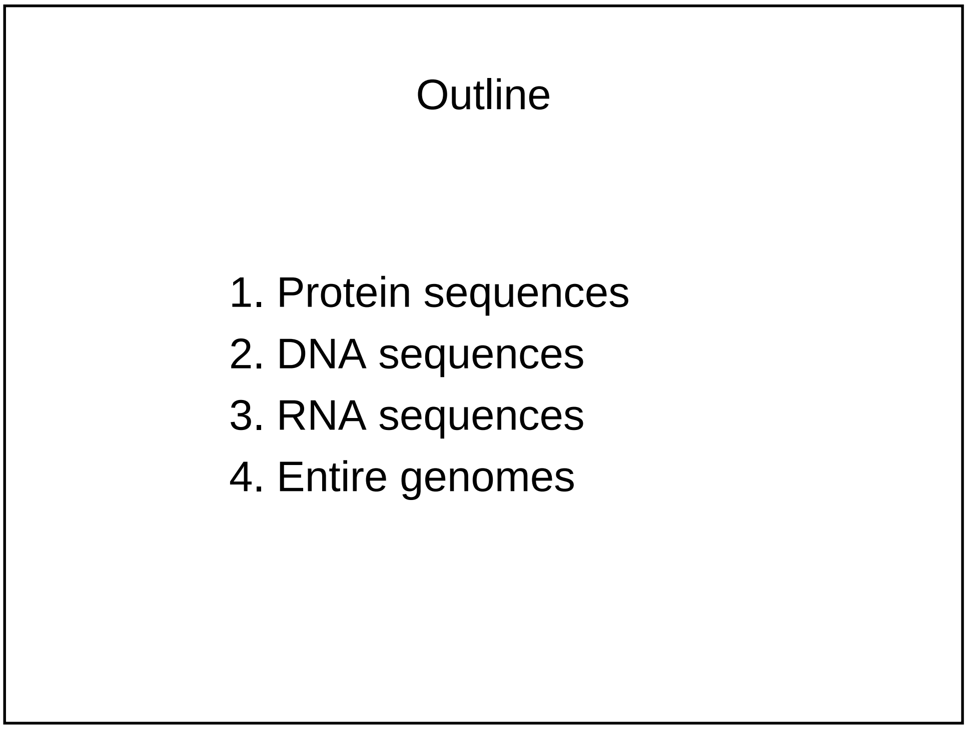
What is bioinformatics?

- Bioinformatics = computational branch of molecular biology
- *in vivo – in vitro – in silico*
- Bioinformatics in a narrower sense: Databases and computational methods for sequences and sequence-related properties of proteins, DNA, and RNA



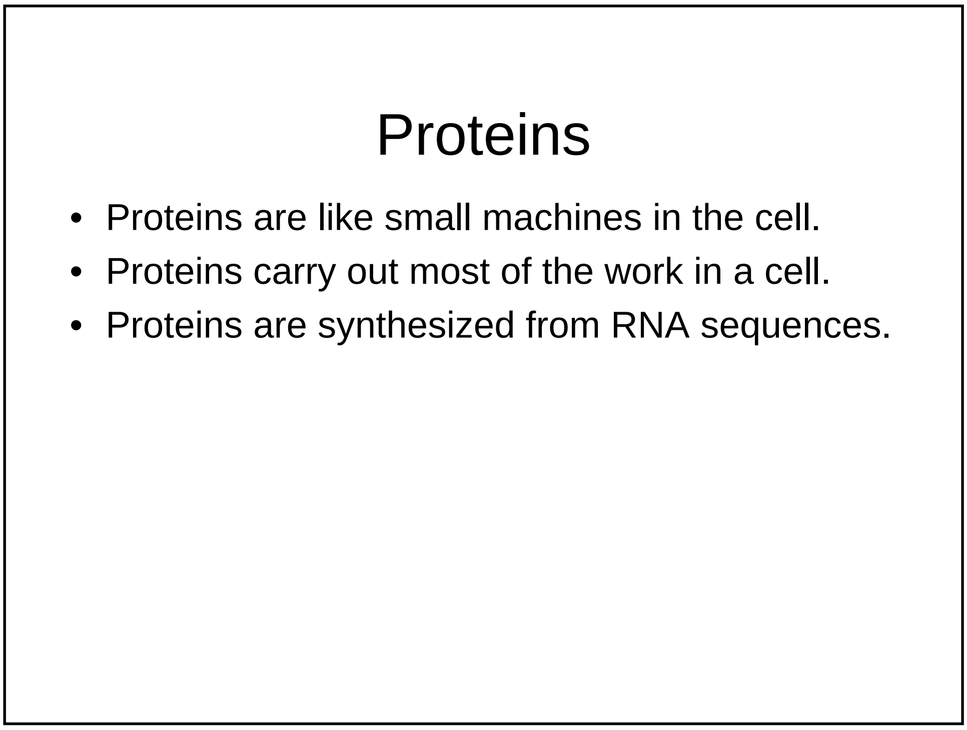
Learning Objectives

- Crash course in molecular biology
- Knowing the basic properties of the main biological sequences: DNA, RNA, and proteins



Outline

1. Protein sequences
2. DNA sequences
3. RNA sequences
4. Entire genomes

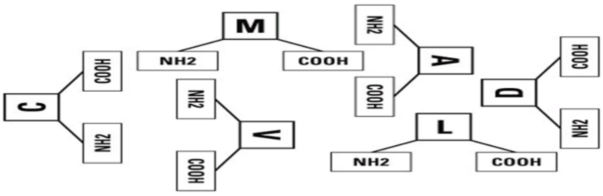


Proteins

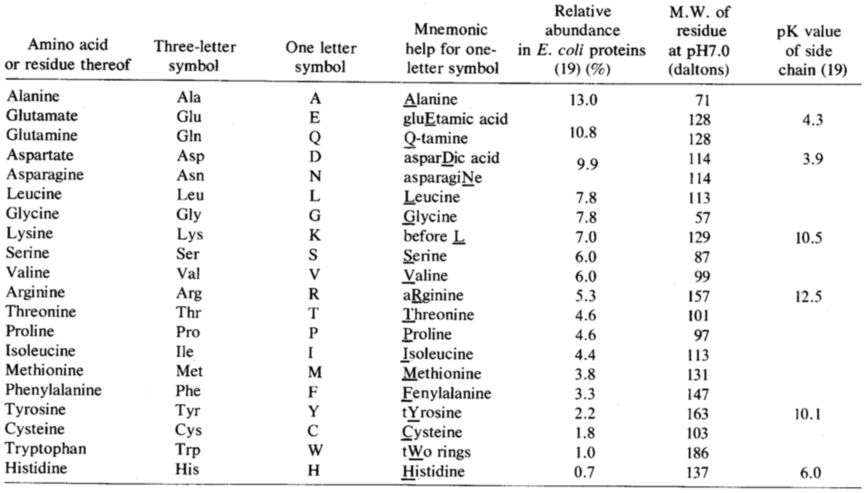
- Proteins are like small machines in the cell.
- Proteins carry out most of the work in a cell.
- Proteins are synthesized from RNA sequences.

Amino Acids

- Proteins are made of 20 amino acids.
- Each amino acid is small molecule made up of fewer than 100 atoms.
- The 20 amino acids have similar terminations; they can be chained to one another like Lego bricks.

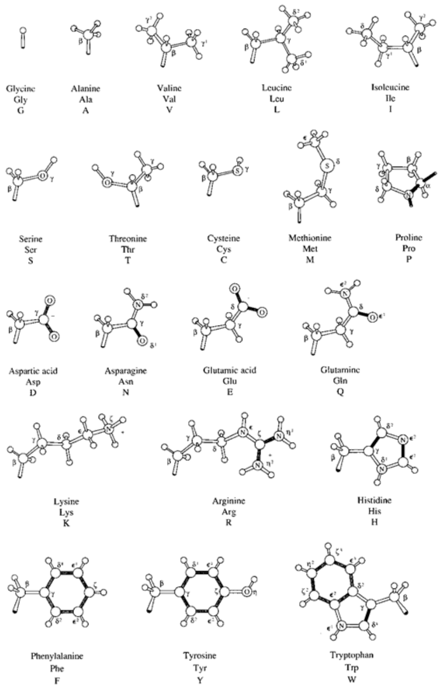


Amino acid names and symbols



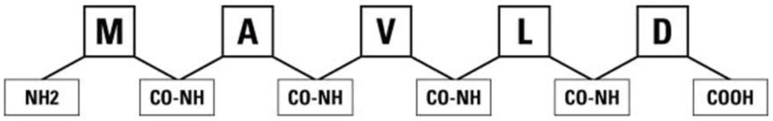
Amino acid or residue thereof	Three-letter symbol	One letter symbol	Mnemonic help for one-letter symbol	Relative abundance in *E. coli* proteins (19) (%)	M.W. of residue at pH7.0 (daltons)	pK value of side chain (19)
Alanine	Ala	A	Alanine	13.0	71	
Glutamate	Glu	E	gluEtamic acid	10.8	128	4.3
Glutamine	Gln	Q	Q-tamine		128	
Aspartate	Asp	D	asparDic acid	9.9	114	3.9
Asparagine	Asn	N	asparagiNe		114	
Leucine	Leu	L	Leucine	7.8	113	
Glycine	Gly	G	Glycine	7.8	57	
Lysine	Lys	K	before L	7.0	129	10.5
Serine	Ser	S	Serine	6.0	87	
Valine	Val	V	Valine	6.0	99	
Arginine	Arg	R	aRginine	5.3	157	12.5
Threonine	Thr	T	Threonine	4.6	101	
Proline	Pro	P	Proline	4.6	97	
Isoleucine	Ile	I	Isoleucine	4.4	113	
Methionine	Met	M	Methionine	3.8	131	
Phenylalanine	Phe	F	Fenylalanine	3.3	147	
Tyrosine	Tyr	Y	tYrosine	2.2	163	10.1
Cysteine	Cys	C	Cysteine	1.8	103	
Tryptophan	Trp	W	tWo rings	1.0	186	
Histidine	His	H	Histidine	0.7	137	6.0

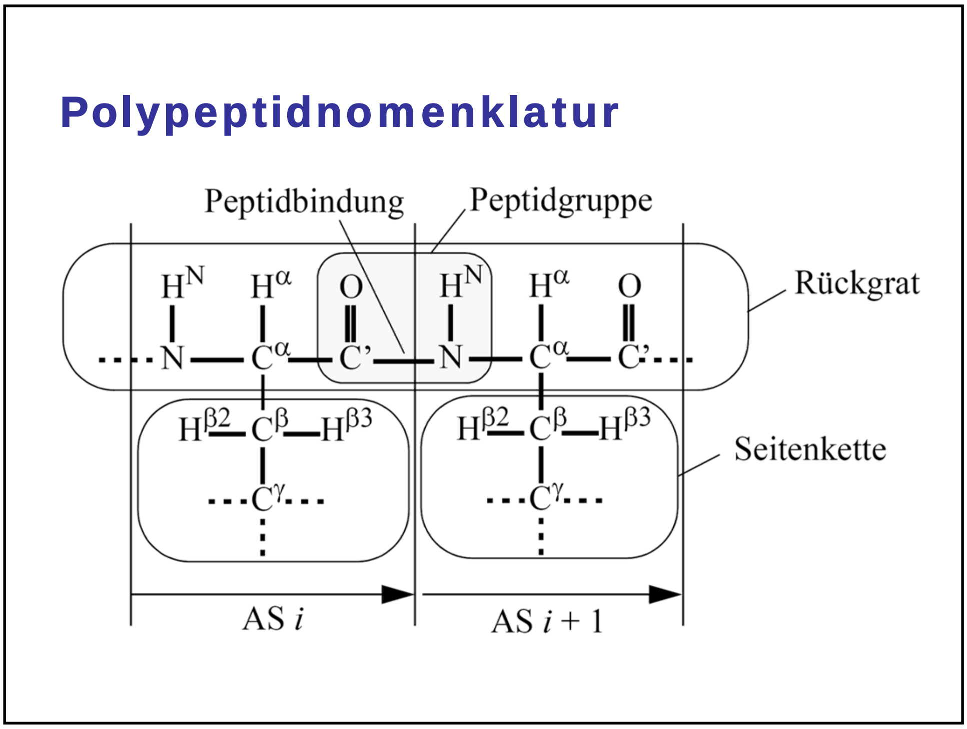
Amino acid side chains



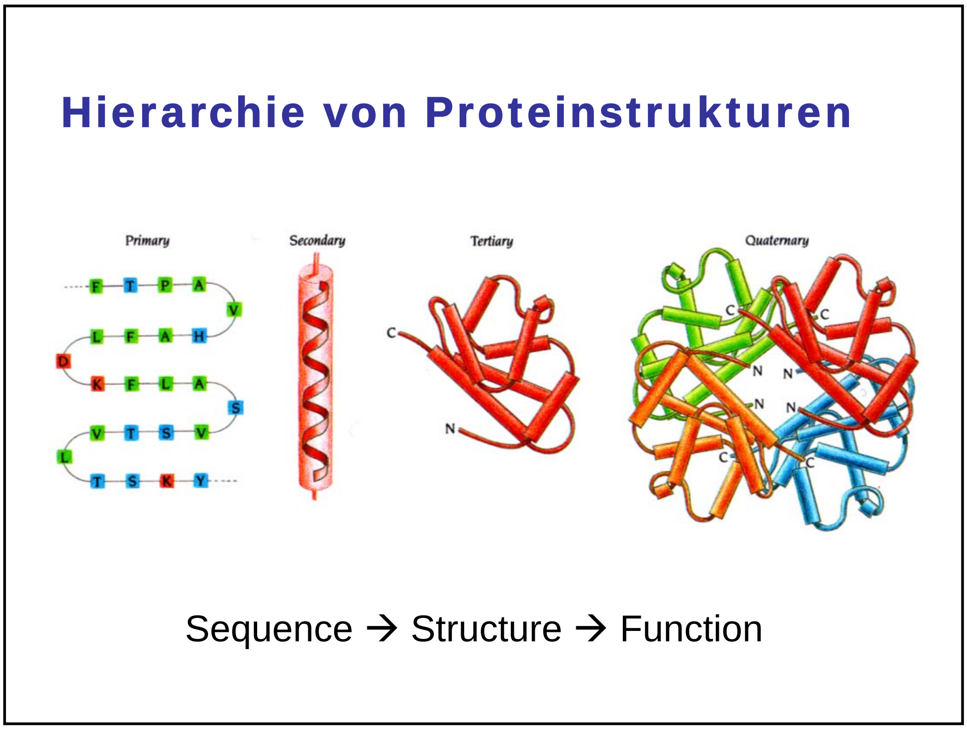
Protein Sequences

- Proteins are made of amino acids chained by peptide bonds.
- Protein sequences are written from the N to the C-terminus.
- Your average protein is 400 amino acids long.
- The longest protein is 30,000 amino acids long.





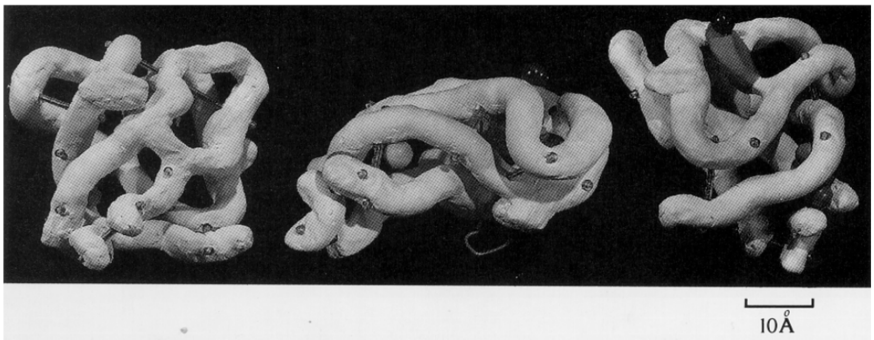
Polypeptidnomenklatur



Hierarchie von Proteinstrukturen

Sequence → Structure → Function

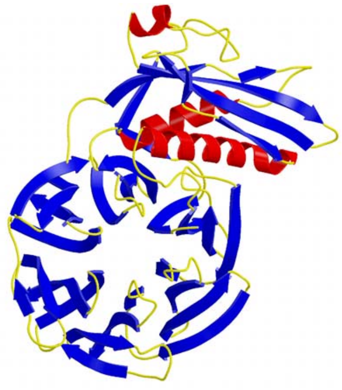
Myoglobin Struktur



“*Vielleicht die bemerkenswerteste Eigenschaft des Moleküls ist seine Komplexität und die Abwesenheit von Symmetrie. Der Anordnung scheinen die Regelmässigkeiten, die man instinktiv erwartet, fast völlig zu fehlen, und sie ist komplizierter als von irgendeiner Theorie der Proteinstruktur vorhergesagt.*” — John Kendrew, 1958

Protein Structures

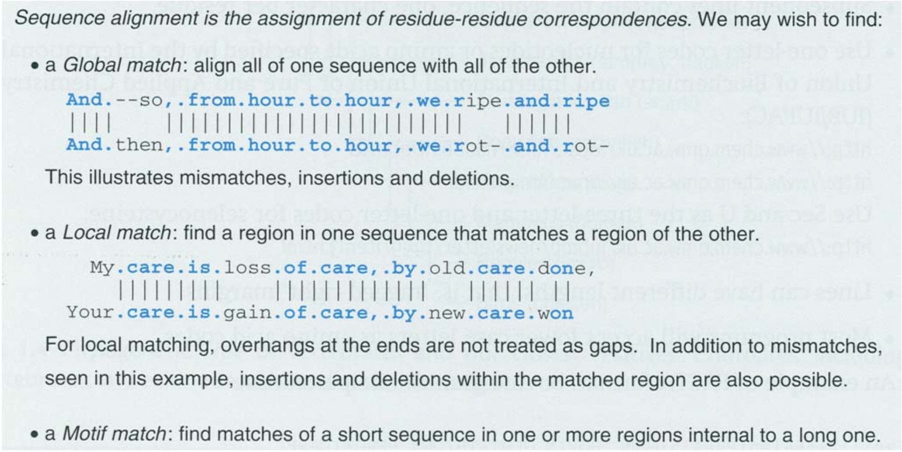
- Proteins have well-defined 3-dimensional structures.
- Hydrophobic amino acids are in the protein’s core.
- Hydrophilic amino acids are on the protein’s surface.



Techniques for Bioinformatic Analysis of Proteins

- Retrieving protein sequences from databases (Ch. 2, 3, 4)
- Computing amino acid composition, molecular weight, isolelectric point, etc. (Ch. 6)
- Computing how hydrophobic or hydrophilic a protein is, predicting antigenic sites, locating membrane-spanning segments (Ch. 6)
- Predicting elements of secondary structure (Ch. 6, 11)
- Predicting the domain organization of proteins (Ch. 6, 7, 9, 11)
- Visualizing protein structures in 3D (Ch. 11)
- Predicting a protein's 3D structure from its sequence (Ch. 11)
- Finding all proteins that share a similar sequence (Ch. 7)
- Classifying proteins into families (Ch. 7, 8, 9)
- Finding the best alignment between two or more proteins (Ch. 8, 9)
- Finding evolutionary relationships between proteins, drawing proteins' family trees (Ch. 7, 9, 11, 13)

Sequence Alignment



Sequence alignment is the assignment of residue-residue correspondences. We may wish to find:

- a *Global match*: align all of one sequence with all of the other.

```
And.--so,.from.hour.to.hour,.we.ripe.and.ripe
||||    |||||||||||||||||||||||||   ||||||
And.then,.from.hour.to.hour,.we.rot-.and.rot-
```

This illustrates mismatches, insertions and deletions.

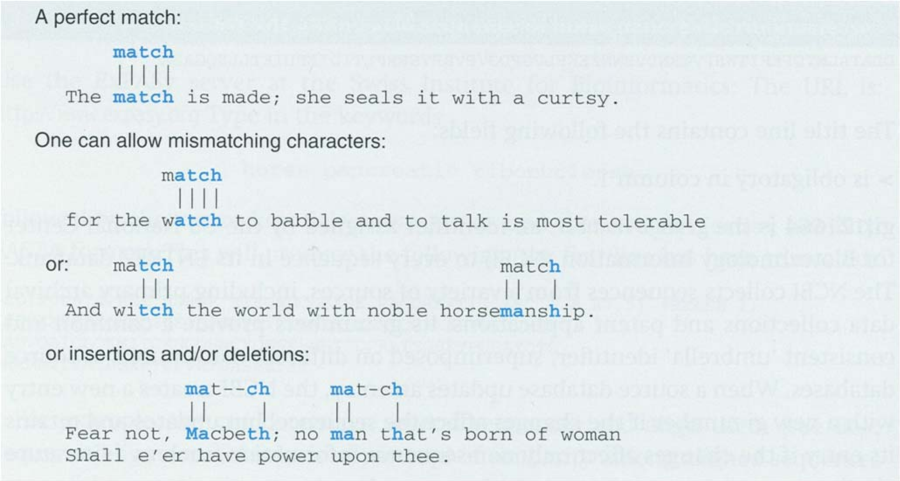
- a *Local match*: find a region in one sequence that matches a region of the other.

```
  My.care.is.loss.of.care,.by.old.care.done,
    |||||||||    |||||||||||||   |||||| ||
Your.care.is.gain.of.care,.by.new.care.won
```

For local matching, overhangs at the ends are not treated as gaps. In addition to mismatches, seen in this example, insertions and deletions within the matched region are also possible.

- a *Motif match*: find matches of a short sequence in one or more regions internal to a long one.

Sequence Alignment



A perfect match:

```
    match
    |||||
The match is made; she seals it with a curtsy.
```

One can allow mismatching characters:

```
        match
         ||||
for the watch to babble and to talk is most tolerable
```

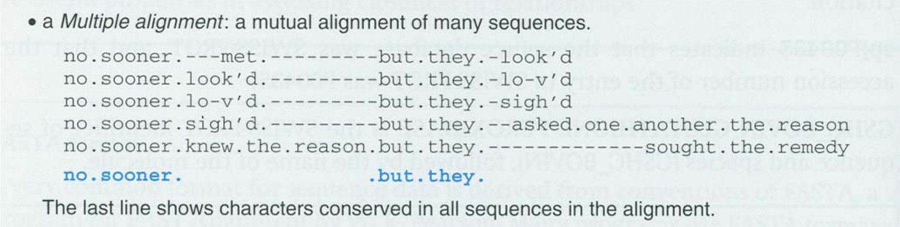
or:

```
    match                           match
      |||                           ||  |
And witch the world with noble horsemanship.
```

or insertions and/or deletions:

```
          mat--ch     mat-ch
          ||    |     ||   |
Fear not, Macbeth; no man that's born of woman
Shall e'er have power upon thee.
```

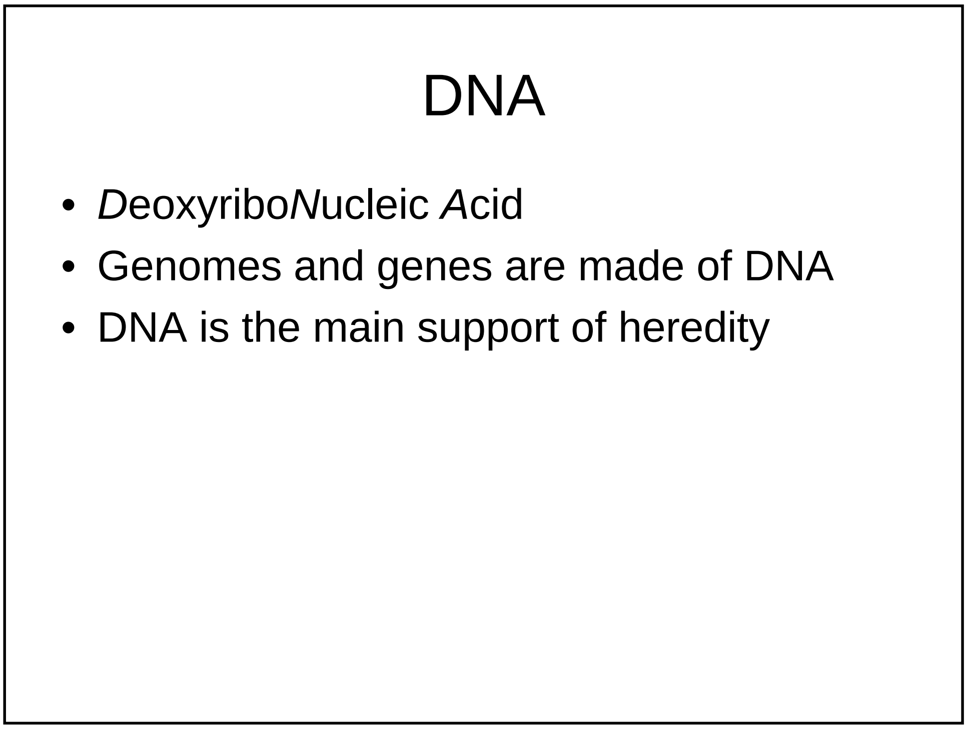
Multiple sequence Alignment



- a *Multiple alignment*: a mutual alignment of many sequences.

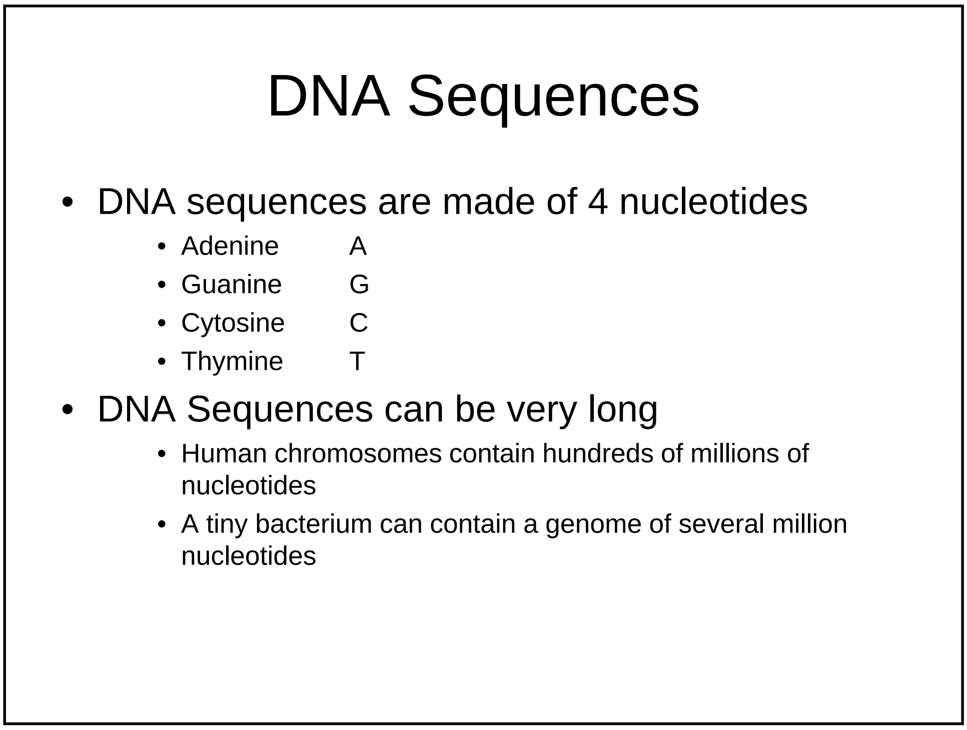
```
no.sooner.---met.---------but.they.-look'd
no.sooner.look'd.---------but.they.-lo-v'd
no.sooner.lo-v'd.---------but.they.-sigh'd
no.sooner.sigh'd.---------but.they.--asked.one.another.the.reason
no.sooner.knew.the.reason.but.they.-------------sought.the.remedy
no.sooner.               .but.they.
```

The last line shows characters conserved in all sequences in the alignment.



DNA

- *D*eoxyribo*N*ucleic *A*cid
- Genomes and genes are made of DNA
- DNA is the main support of heredity

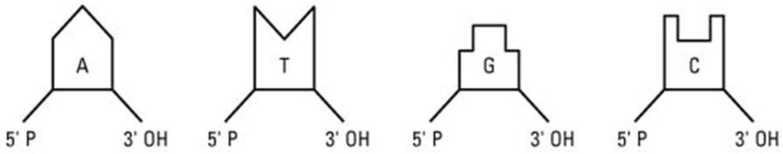


DNA Sequences

- DNA sequences are made of 4 nucleotides
  - Adenine A
  - Guanine G
  - Cytosine C
  - Thymine T
- DNA Sequences can be very long
  - Human chromosomes contain hundreds of millions of nucleotides
  - A tiny bacterium can contain a genome of several million nucleotides

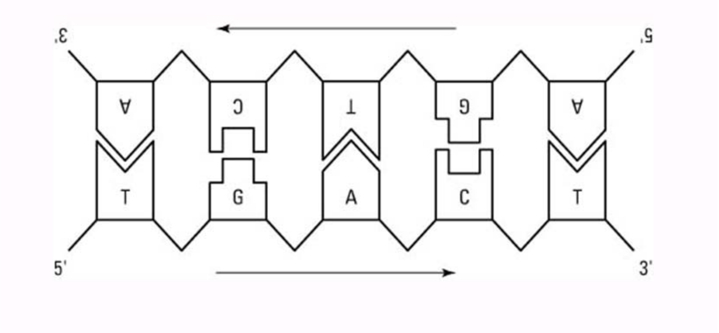
Nucleotides

- Nucleotides have similar terminations.
- Nucleotides are meant to be chained like Lego bricks.
- Nucleotides can interact with each other:
    - Adenine with thymine (A with T)
    - Guanine with cytosine (G with C)



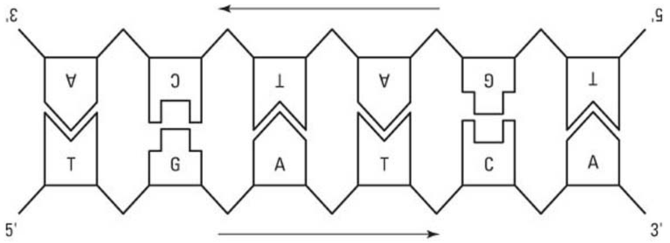
Double-strand DNA

- DNA sequences always come in two strands.
- The strands are complementary and opposite in orientation.
- By convention, sequences are written in 5’ → 3’ direction.
- Most database-search programs search both strands automatically.



Palindromic DNA

- Regions of DNA may correspond to sequences that are identical when read from the two complementary strands.
- Example: TGATCA



- Palindromic sequences play important biological roles:
  - Most restriction enzymes recognize palindromic target sequences.
  - Binding sites for regulatory proteins are often palindromic.
  - Strong influence on 3D structure of DNA (and RNA).

RNA

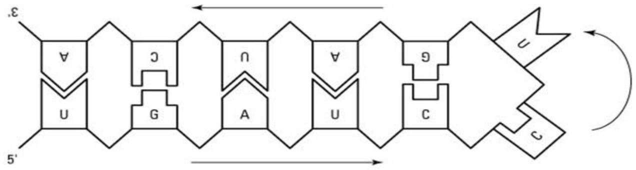
- *R*ibo*N*ucleic *A*cid
- RNA is a close relative of DNA
- RNA has many functions
  - Provides coding for proteins
  - Helps synthesize proteins
  - Helps many basic processes in the cell
- RNA is not very stable
  - RNA is synthesized and very often degraded
  - DNA, by contrast, is very stable

The RNA Sequence

- RNA contains 4 nucleotides:
  - A, G, C, U
  - U is Uracil
- RNA does not contain Thymine (T)
- Uracil replaces Thymine in RNA
- RNA is single-stranded

RNA Secondary Structures

- RNA can make secondary structures
- RNA can make 1 strand with itself as a secondary structure
- Secondary structures are made of stems and loops

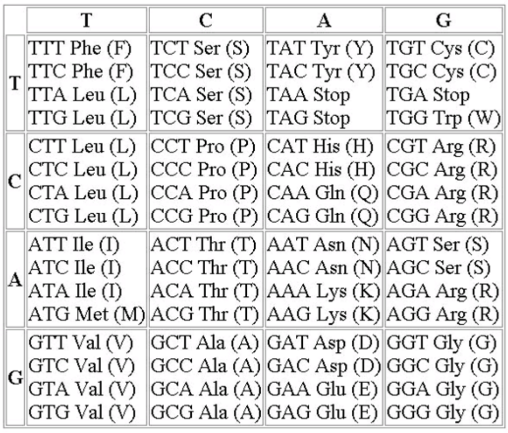


What Is the Length of My Sequence ?

- Protein sizes are expressed in amino acids or in Daltons
  - 115 Daltons ~ 1 amino acid
- DNA and RNA sequences length are expressed in
  - Base-pairs (bp)
  - One Kbp or Kb: 1 thousand base pairs
  - One Mbp or Mb: 1 million base pairs
  - One Gbp or Gb: 1 billion base pairs
- The following terms often have the same meaning:
  - Base
  - Base-pair (bp)
  - Nucleotide (nt)
  - Positions, nucleotides, residues

Turning DNA into Proteins: The Genetic Code

- DNA gets transcribed into RNA using nucleotide complementarity.
- RNA gets translated into proteins using the genetic code:
  - UCU UAU GCG UAA
  - SER-TYR-ALA-STOP



	T	C	A	G
T	TTT Phe (F) TTC Phe (F) TTA Leu (L) TTG Leu (L)	TCT Ser (S) TCC Ser (S) TCA Ser (S) TCG Ser (S)	TAT Tyr (Y) TAC Tyr (Y) TAA Stop TAG Stop	TGT Cys (C) TGC Cys (C) TGA Stop TGG Trp (W)
C	CTT Leu (L) CTC Leu (L) CTA Leu (L) CTG Leu (L)	CCT Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAT His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGT Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)
A	ATT Ile (I) ATC Ile (I) ATA Ile (I) ATG Met (M)	ACT Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAT Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGT Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)
G	GTT Val (V) GTC Val (V) GTA Val (V) GTG Val (V)	GCT Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAT Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGT Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)

Coding DNA sequences

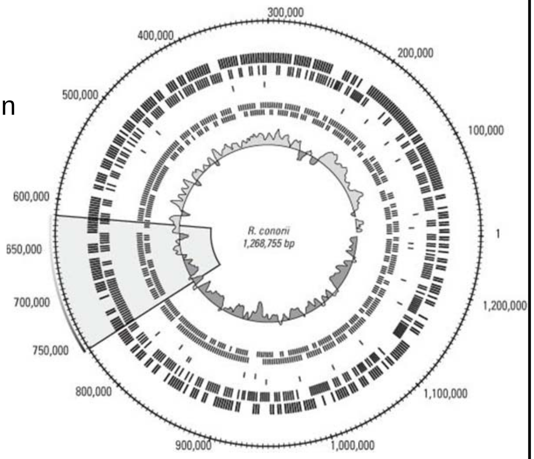
- Base triplets are translated into amino acids.
- Example DNA sequence:
  ATGGAAGTATTTAAAGCGCCACCTATTGGGATATAAG…
- Decompose into successive triplets:
  ATG GAA GTA TTT AAA GCG CCA CCT ATT GGG ATA TAA G…
- Translate each triplet into the corresponding amino acid:
  M E V F K A P P I G I stop
- Other reading frames:
  A TGG AAG TAT TTA AAG CGC CAC CTA TTG GGA TAT AAG…
  W K Y L K R H L L G Y K
  AT GGA AGT ATT TAA AGC GCC ACC TAT TGG GAT ATA AG…
  G S I STOP
- Together with the complementary strand there are 6 possible reading frames. In nature usually only one of these is translated into a protein.
- Open reading frame (ORF): interval of DNA sequence without stop codons.
- Eukaryotic genes can be interrupted by non-coding intervals (introns).
- Locating protein-coding regions in DNA is an important part of bioinformatics.

Techniques for Bioinformatic Analysis of DNA/RNA

- Retrieving DNA sequences from databases (Ch. 2, 3)
- Computing nucleotide compositions (Ch. 5)
- Identifying restriction sites (Ch. 5)
- Designing polymerase chain-reaction (PCR) primers (Ch. 5)
- Identifying open reading frames (ORFs) (Ch. 5)
- Predicting elements of DNA/RNA secondary structure (Ch. 12)
- Finding repeats (Ch. 5)
- Computing the optimal alignment between two or more DNA sequences (Ch. 7, 8, and 9)
- Finding polymorphic sites in genes (single nucleotide polymorphisms, SNPs) (Ch. 3)
- Assembling sequence fragments (Ch. 5)

Bioinformatics applications for entire genomes

- Finding which genomes are available (Ch. 3)
- Analyzing sequences in relation to specific genomes (Ch. 3, 7)
- Displaying genomes (Ch. 3)
- Parsing a microbial genome sequence: ORFing (Ch. 5)
- Parsing a eukaryotic genome sequence: GenScan (Ch. 5)
- Finding orthologous and paralogous genes (Ch. 3)
- Finding repeats (Ch. 5)



Amount of data in bioinformatics

- The amount of the data of bioinformatics is very large.
- Not only are the individual data banks large, but their sizes are increasing at a very high rate.

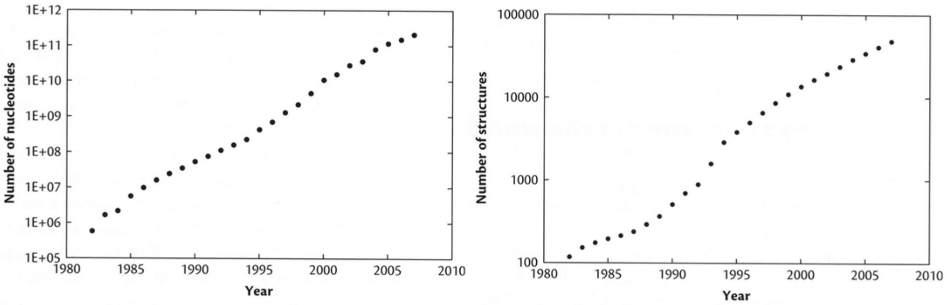


Fig. 1.1 (a) Growth of the International Nucleotide Sequence Database Collection. (b) Growth of the world-wide Protein Data Bank, archive of three-dimensional biological macromolecular structures, from the wwPDB, a collaboration between groups in the US, Europe and Japan. Note log scale on y-axes.

Genome sizes

Size of human genome

= 1 huge ("human genome equivalent")

= 3 x 10^9 bases

= number of characters in 6 complete years of issues of the New York Times

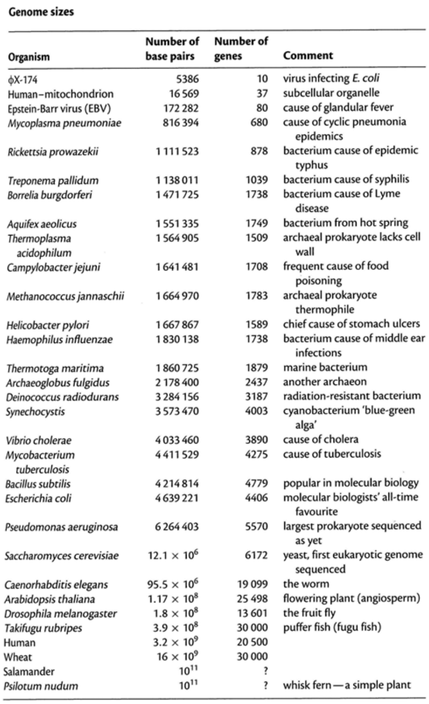
Size of *E. coli* genome

= 4.6 x 10^6 bases = 0.0015 huges

= number of characters in Shakespeare's plays

Size of nucleotide sequence databanks (2007 est.)

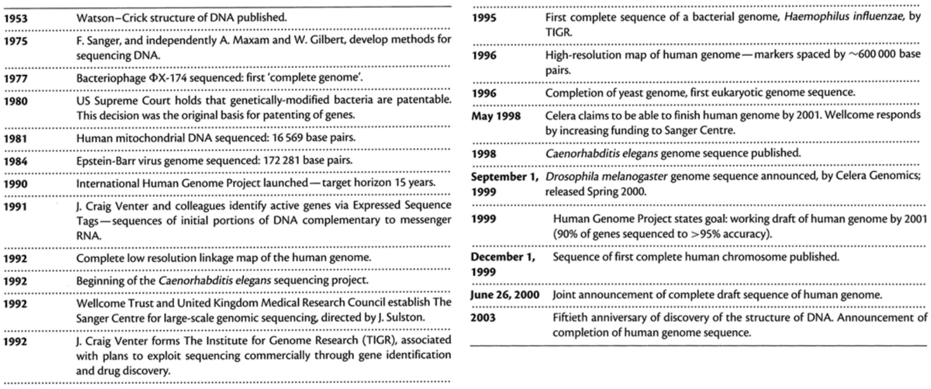
= 1.7 x 10^{12} bases = 567 huges



Genome sizes

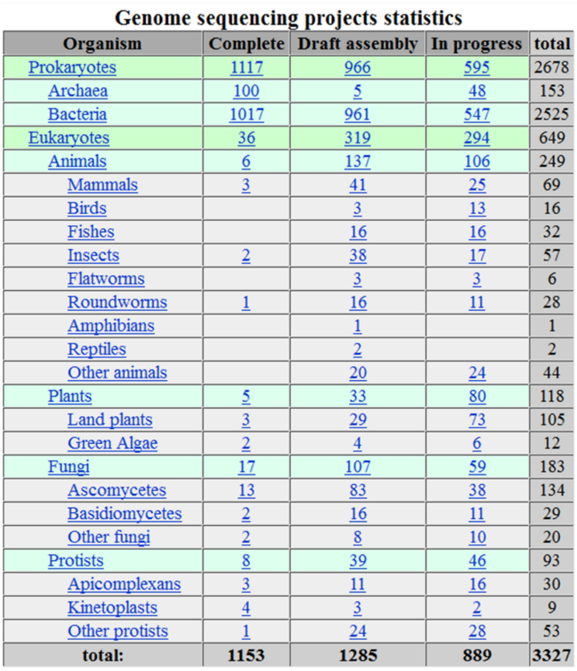
Organism	Number of base pairs	Number of genes	Comment
ϕX-174	5386	10	virus infecting *E. coli*
Human–mitochondrion	16 569	37	subcellular organelle
Epstein-Barr virus (EBV)	172 282	80	cause of glandular fever
Mycoplasma pneumoniae	816 394	680	cause of cyclic pneumonia epidemics
Rickettsia prowazekii	1 111 523	878	bacterium cause of epidemic typhus
Treponema pallidum	1 138 011	1039	bacterium cause of syphilis
Borrelia burgdorferi	1 471 725	1738	bacterium cause of Lyme disease
Aquifex aeolicus	1 551 335	1749	bacterium from hot spring
Thermoplasma acidophilum	1 564 905	1509	archaeal prokaryote lacks cell wall
Campylobacter jejuni	1 641 481	1708	frequent cause of food poisoning
Methanococcus jannaschii	1 664 970	1783	archaeal prokaryote thermophile
Helicobacter pylori	1 667 867	1589	chief cause of stomach ulcers
Haemophilus influenzae	1 830 138	1738	bacterium cause of middle ear infections
Thermotoga maritima	1 860 725	1879	marine bacterium
Archaeoglobus fulgidus	2 178 400	2437	another archaeon
Deinococcus radiodurans	3 284 156	3187	radiation-resistant bacterium
Synechocystis	3 573 470	4003	cyanobacterium 'blue-green alga'
Vibrio cholerae	4 033 460	3890	cause of cholera
Mycobacterium tuberculosis	4 411 529	4275	cause of tuberculosis
Bacillus subtilis	4 214 814	4779	popular in molecular biology
Escherichia coli	4 639 221	4406	molecular biologists' all-time favourite
Pseudomonas aeruginosa	6 264 403	5570	largest prokaryote sequenced as yet
Saccharomyces cerevisiae	12.1×10^6	6172	yeast, first eukaryotic genome sequenced
Caenorhabditis elegans	95.5×10^6	19 099	the worm
Arabidopsis thaliana	1.17×10^8	25 498	flowering plant (angiosperm)
Drosophila melanogaster	1.8×10^8	13 601	the fruit fly
Takifugu rubripes	3.9×10^8	30 000	puffer fish (fugu fish)
Human	3.2×10^9	20 500	
Wheat	16×10^9	30 000	
Salamander	10^{11}	?	
Psilotum nudum	10^{11}	?	whisk fern—a simple plant

Landmarks in the Human Genome Project



1953	Watson–Crick structure of DNA published.
1975	F. Sanger, and independently A. Maxam and W. Gilbert, develop methods for sequencing DNA.
1977	Bacteriophage ΦX-174 sequenced: first 'complete genome'.
1980	US Supreme Court holds that genetically-modified bacteria are patentable. This decision was the original basis for patenting of genes.
1981	Human mitochondrial DNA sequenced: 16 569 base pairs.
1984	Epstein-Barr virus genome sequenced: 172 281 base pairs.
1990	International Human Genome Project launched—target horizon 15 years.
1991	J. Craig Venter and colleagues identify active genes via Expressed Sequence Tags—sequences of initial portions of DNA complementary to messenger RNA.
1992	Complete low resolution linkage map of the human genome.
1992	Beginning of the *Caenorhabditis elegans* sequencing project.
1992	Wellcome Trust and United Kingdom Medical Research Council establish The Sanger Centre for large-scale genomic sequencing, directed by J. Sulston.
1992	J. Craig Venter forms The Institute for Genome Research (TIGR), associated with plans to exploit sequencing commercially through gene identification and drug discovery.
1995	First complete sequence of a bacterial genome, *Haemophilus influenzae*, by TIGR.
1996	High-resolution map of human genome—markers spaced by ~600 000 base pairs.
1996	Completion of yeast genome, first eukaryotic genome sequence.
May 1998	Celera claims to be able to finish human genome by 2001. Wellcome responds by increasing funding to Sanger Centre.
1998	*Caenorhabditis elegans* genome sequence published.
September 1, 1999	*Drosophila melanogaster* genome sequence announced, by Celera Genomics; released Spring 2000.
1999	Human Genome Project states goal: working draft of human genome by 2001 (90% of genes sequenced to >95% accuracy).
December 1, 1999	Sequence of first complete human chromosome published.
June 26, 2000	Joint announcement of complete draft sequence of human genome.
2003	Fiftieth anniversary of discovery of the structure of DNA. Announcement of completion of human genome sequence.

Genome sequencing projects



Genome sequencing projects statistics

Organism	Complete	Draft assembly	In progress	total
Prokaryotes	1117	966	595	2678
Archaea	100	5	48	153
Bacteria	1017	961	547	2525
Eukaryotes	36	319	294	649
Animals	6	137	106	249
Mammals	3	41	25	69
Birds		3	13	16
Fishes		16	16	32
Insects	2	38	17	57
Flatworms		3	3	6
Roundworms	1	16	11	28
Amphibians		1		1
Reptiles		2		2
Other animals		20	24	44
Plants	5	33	80	118
Land plants	3	29	73	105
Green Algae	2	4	6	12
Fungi	17	107	59	183
Ascomycetes	13	83	38	134
Basidiomycetes	2	16	11	29
Other fungi	2	8	10	20
Protists	8	39	46	93
Apicomplexans	3	11	16	30
Kinetoplasts	4	3	2	9
Other protists	1	24	28	53
total:	**1153**	**1285**	**889**	**3327**

http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html (Feb 16, 2012)