CMSC423: Bioinformatic Algorithms, Databases and Tools

Dr. Todd Treangen

Lecture #2

9/4/2012

Molecular Biology primer

Admin…

- Review website
  - Recommend books
  - Supplemental reading material
- Have you tried your glue accounts?
  - Part of HW1 will require you to use glue system
    - Let's try it out
- Issues/concerns/questions about class and policies?
  - Attendance policy important

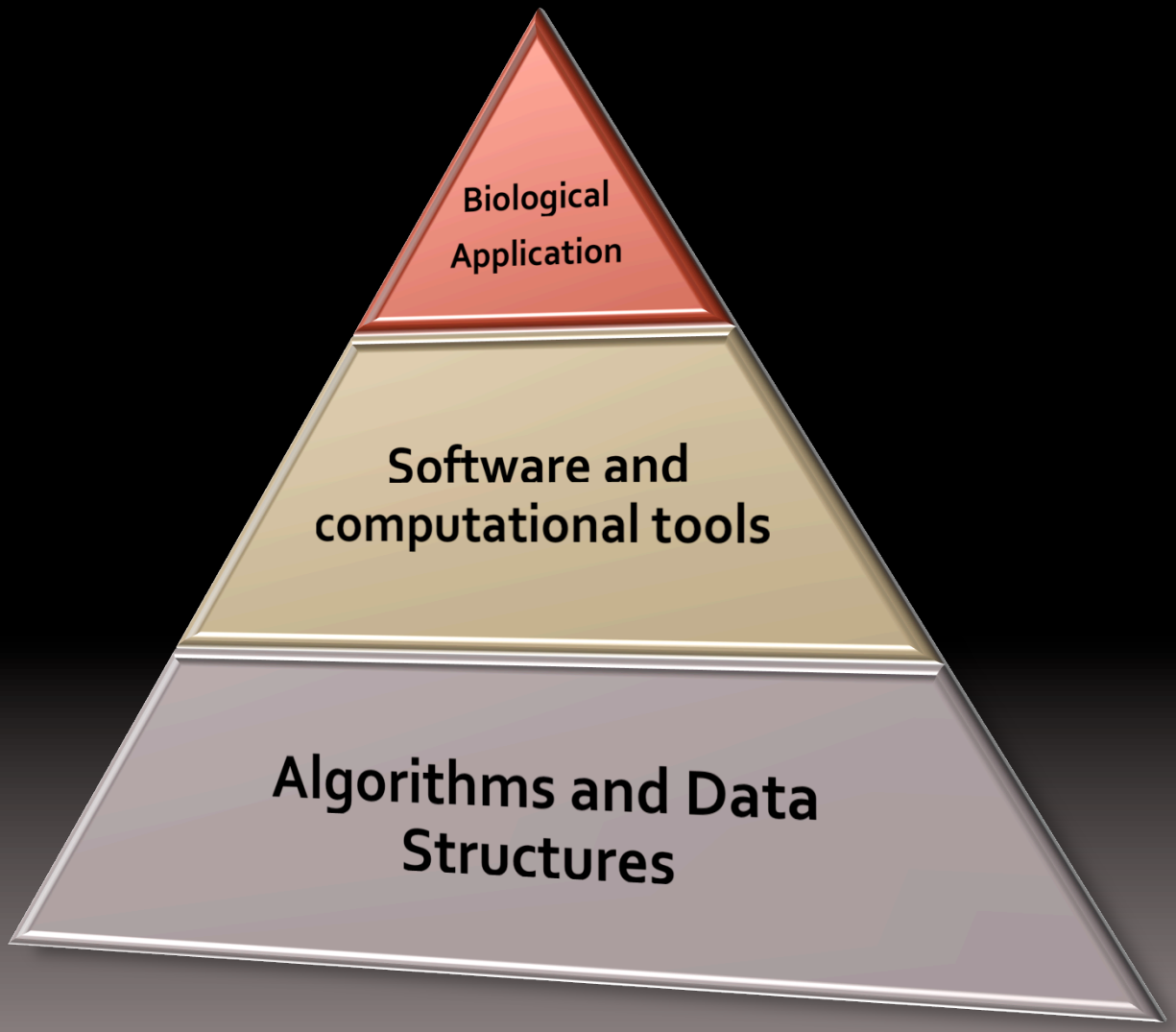
Admin continued…

- TA
  - Milad Gholami
    - Mondays & Wednesdays 9-11am
  - Everyone know where the TA office is located?

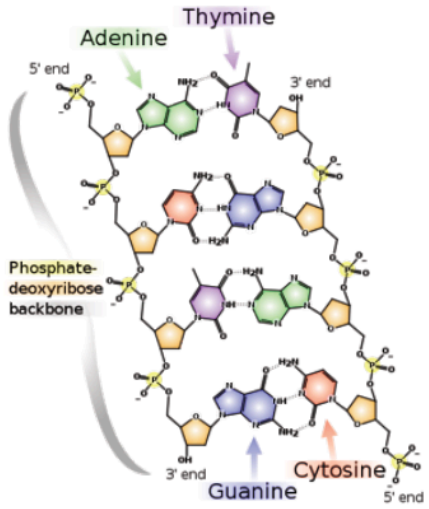
  - Grades will be available through the grade server grades.cs.umd.edu

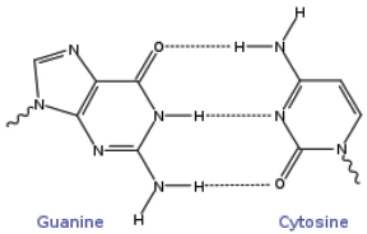
  - **Frontiers in Genomics symposium (free!)**
    - **Excellent speakers**
    - **October 18th, from 8:45am to 3pm**
    - **RSVP to igs-event@som.umaryland.edu**

Bioinformatics

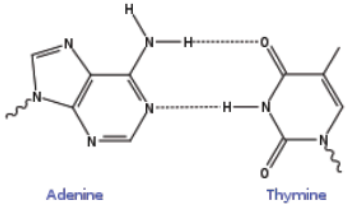


DNA


Adenine
Thymine
5' end
3' end
Phosphate-
deoxyribose
backbone
3' end
Guanine
Cytosine
5' end


Guanine
Cytosine

G C


Adenine
Thymine

A T

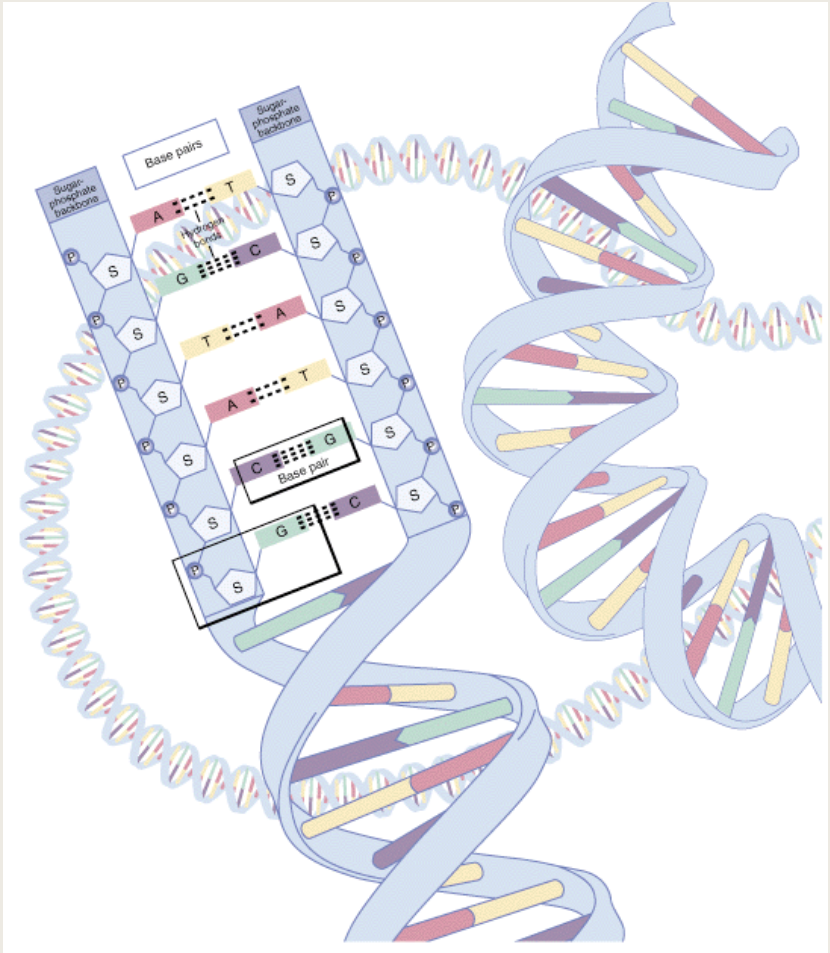
Documents of evolutionary history

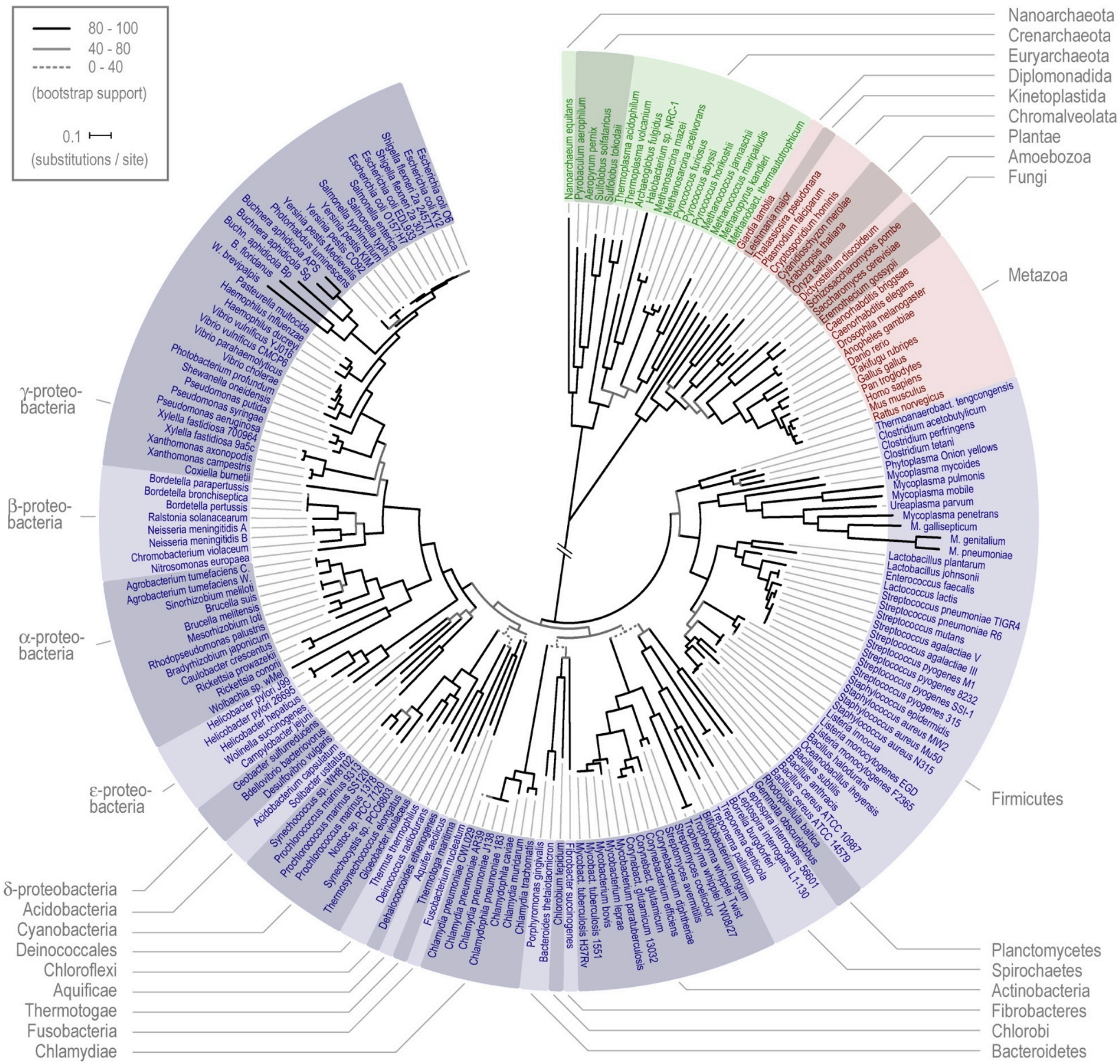
The four nucleotides in DNA contain the bases: guanine (**G**), cytosine (**C**), adenine (**A**), and uracil (**U**), read as thymine (**T**) in DNA.

The **genome** of an organism contains its hereditary information and is encoded in its DNA .

In 1965, Zuckerandl and Pauling described an organisms DNA as "Documents of Evolutionary History"

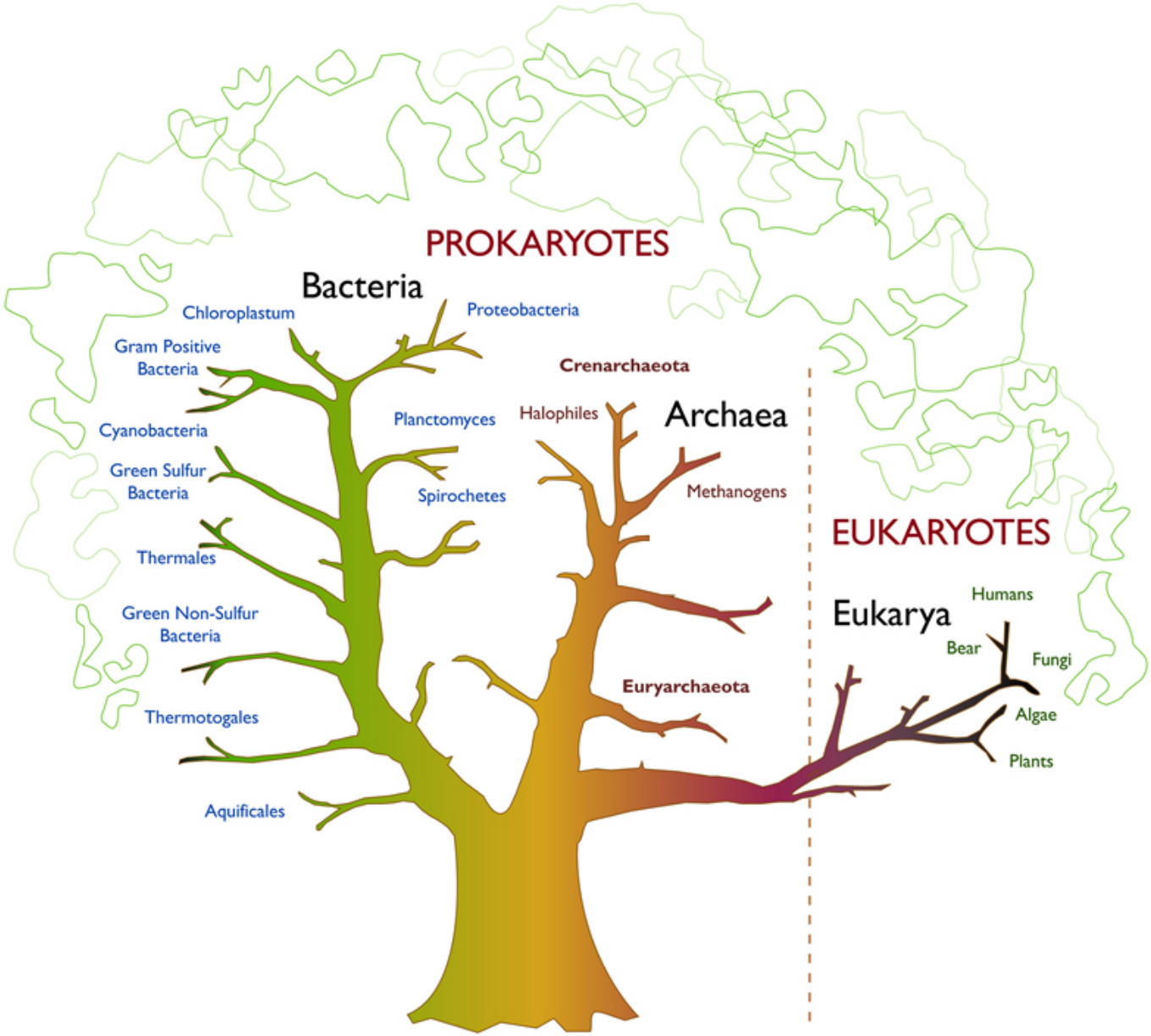
See reading material !





80 - 100
40 - 80
0 - 40
(bootstrap support)
0.1
(substitutions / site)
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Diplomonadida
Kinetoplastida
Chromalveolata
Plantae
Amoebozoa
Fungi
Metazoa
Firmicutes
Planctomycetes
Spirochaetes
Actinobacteria
Fibrobacteres
Chlorobi
Bacteroidetes
Chlamydiae
Fusobacteria
Thermotogae
Aquificae
Chloroflexi
Deinococcales
Cyanobacteria
Acidobacteria
δ-proteobacteria
ε-proteo-bacteria
α-proteo-bacteria
β-proteo-bacteria
γ-proteo-bacteria
Nanoarchaeum equitans
Pyrobaculum aerophilum
Aeropyrum pernix
Sulfolobus solfataricus
Sulfolobus tokodaii
Thermoplasma acidophilum
Thermoplasma volcanium
Archaeoglobus fulgidus
Halobacterium sp. NRC-1
Methanosarcina mazei
Methanosarcina acetivorans
Pyrococcus furiosus
Pyrococcus abyssi
Pyrococcus horikoshii
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanobact. thermautotrophicum
Giardia lamblia
Leishmania major
Thalassiosira pseudonana
Plasmodium falciparum
Cryptosporidium hominis
Cyanidioschyzon merolae
Arabidopsis thaliana
Oryza sativa
Dictyostelium discoideum
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Eremothecium gossypii
Caenorhabditis briggsae
Caenorhabditis elegans
Drosophila melanogaster
Anopheles gambiae
Danio rerio
Takifugu rubripes
Gallus gallus
Pan troglodytes
Homo sapiens
Mus musculus
Rattus norvegicus
Thermoanaerobact. tengcongensis
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Phytoplasma Onion yellows
Mycoplasma mycoides
Mycoplasma pulmonis
Mycoplasma mobile
Ureaplasma parvum
Mycoplasma penetrans
M. gallisepticum
M. genitalium
M. pneumoniae
Lactobacillus plantarum
Lactobacillus johnsonii
Enterococcus faecalis
Lactococcus lactis
Streptococcus pneumoniae TIGR4
Streptococcus pneumoniae R6
Streptococcus mutans
Streptococcus agalactiae V
Streptococcus agalactiae III
Streptococcus pyogenes M1
Streptococcus pyogenes 8232
Streptococcus pyogenes SSI-1
Streptococcus pyogenes 315
Staphylococcus epidermidis
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Listeria innocua
Listeria monocytogenes EGD
Listeria monocytogenes F2365
Bacillus halodurans
Oceanobacillus iheyensis
Bacillus subtilis
Bacillus anthracis
Bacillus cereus ATCC 10987
Bacillus cereus ATCC 14579
Gemmata obscuriglobus
Rhodopirellula baltica
Leptospira interrogans 56601
Leptospira interrogans L1-130
Borrelia burgdorferi
Treponema denticola
Treponema pallidum
Bifidobacterium longum
Tropheryma whipplei Twist
Tropheryma whipplei TW08/27
Streptomyces coelicolor
Streptomyces avermitilis
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebact. glutamicum
Corynebact. glutamicum 13032
Mycobacterium paratuberculosis
Mycobacterium leprae
Mycobacterium bovis
Mycobact. tuberculosis 1551
Mycobact. tuberculosis H37Rv
Fibrobacter succinogenes
Chlorobium tepidum
Bacteroides thetaiotaomicron
Porphyromonas gingivalis
Chlamydia trachomatis
Chlamydia muridarum
Chlamydophila caviae
Chlamydia pneumoniae J138
Chlamydophila pneumoniae AR39
Chlamydia pneumoniae CWL029
Fusobacterium nucleatum
Thermotoga maritima
Aquifex aeolicus
Dehalococcoides ethenogenes
Deinococcus radiodurans
Thermus thermophilus
Gloeobacter violaceus
Thermosynechococcus elongatus
Synechocystis sp. PCC6803
Nostoc sp. PCC 7120
Synechococcus sp. WH8102
Solibacter usitatus
Acidobacterium capsulatum
Desulfovibrio vulgaris
Bdellovibrio bacteriovorus
Geobacter sulfurreducens
Campylobacter jejuni
Wolinella succinogenes
Helicobacter hepaticus
Helicobacter pylori 26695
Helicobacter pylori J99
Wolbachia sp. wMel
Rickettsia conorii
Rickettsia prowazekii
Caulobacter crescentus
Bradyrhizobium japonicum
Rhodopseudomonas palustris
Mesorhizobium loti
Brucella melitensis
Brucella suis
Sinorhizobium meliloti
Agrobacterium tumefaciens W.
Agrobacterium tumefaciens C.
Nitrosomonas europaea
Chromobacterium violaceum
Neisseria meningitidis B
Neisseria meningitidis A
Ralstonia solanacearum
Bordetella pertussis
Bordetella bronchiseptica
Bordetella parapertussis
Coxiella burnetii
Xanthomonas campestris
Xanthomonas axonopodis
Xylella fastidiosa 9a5c
Xylella fastidiosa 700964
Pseudomonas aeruginosa
Pseudomonas syringae
Pseudomonas putida
Shewanella oneidensis
Photobacterium profundum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Haemophilus ducreyi
Haemophilus influenzae
Pasteurella multocida
W. brevipalpis
B. floridanus
Buchn. aphidicola Bp
Buchnera aphidicola Sg
Buchnera aphidicola APS
Photorhabdus luminescens
Yersinia pestis Medievalis
Yersinia pestis KIM
Yersinia pestis CO92
Salmonella typhimurium
Salmonella enterica
Salmonella typhi
Escherichia coli O157:H7
Escherichia coli EDL933
Shigella flexneri 2a 2457T
Shigella flexneri 2a 301
Escherichia coli K12
Escherichia coli O6

Tree of Life

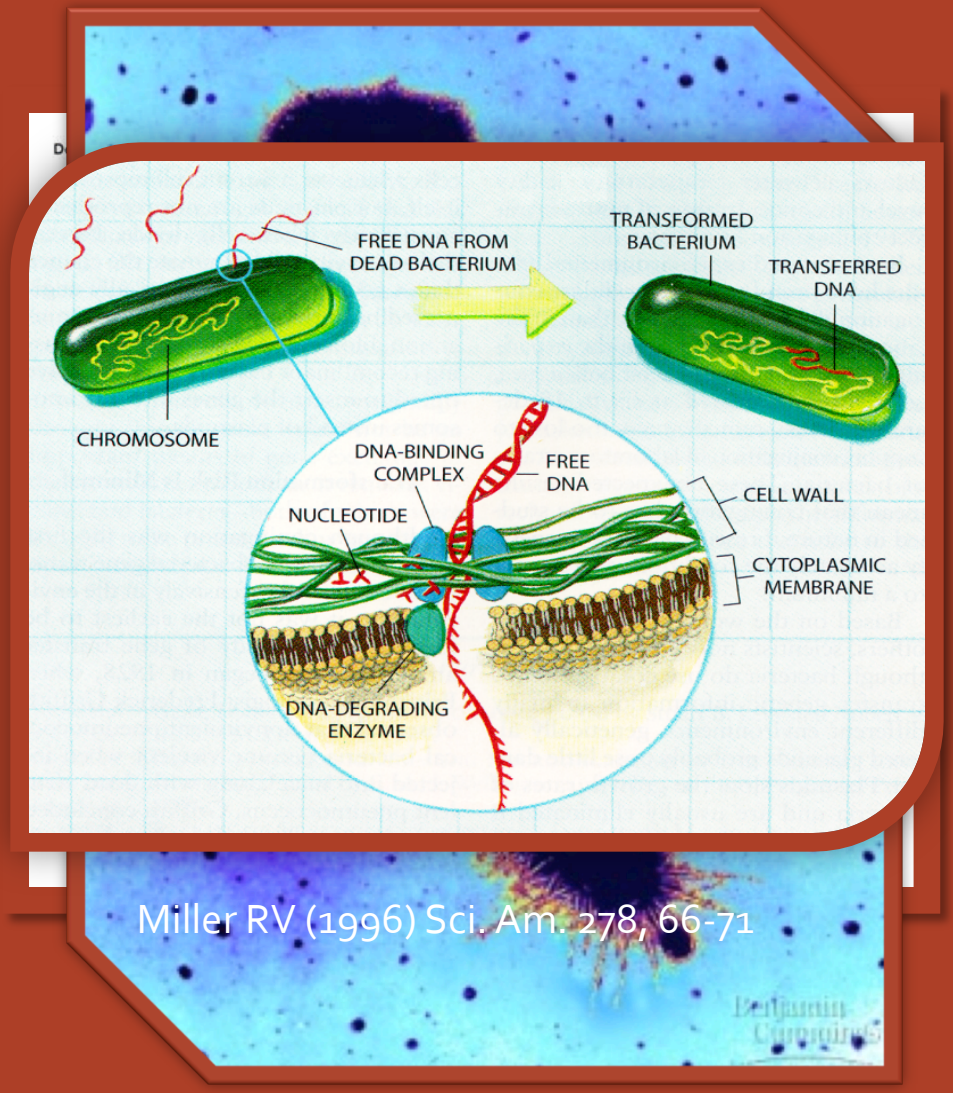

PROKARYOTES
Bacteria
Chloroplastum
Proteobacteria
Gram Positive Bacteria
Crenarchaeota
Planctomyces
Halophiles
Archaea
Cyanobacteria
Green Sulfur Bacteria
Spirochetes
Methanogens
EUKARYOTES
Thermales
Green Non-Sulfur Bacteria
Eukarya
Humans
Bear
Fungi
Euryarchaeota
Thermotogales
Algae
Plants
Aquificales

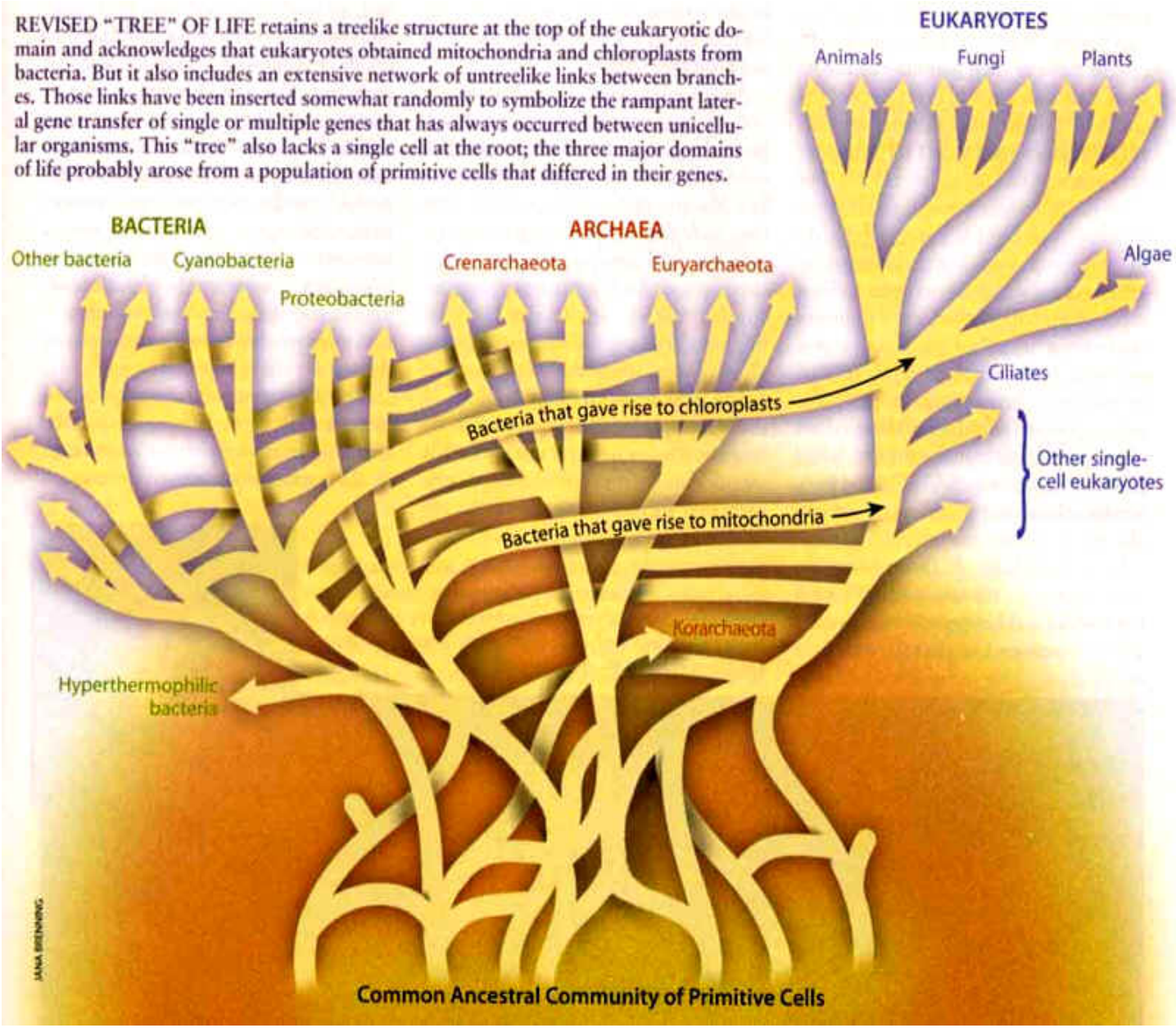
However, there are some exceptions

Genetic Exchange in bacteria

There are three main ways which a bacteria can gain new genetic material (DNA) from donor bacterium:

- **1.** Conjugation: Transfer from one bacteria to another via a pilus. (Lederberg and Tatum, 1946)
- 2. Transduction: Transfer of DNA is mediated by a bacteriophage . (Zinder and Linderberg, 1952)
- 3. **Transformation: DNA is taken up from the environment**
- (Griffith 1928)

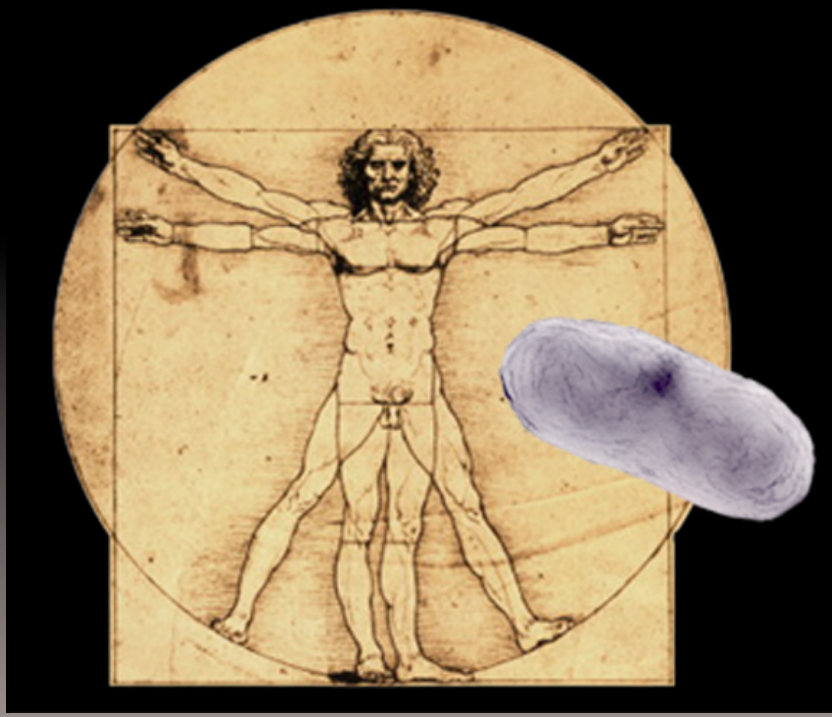




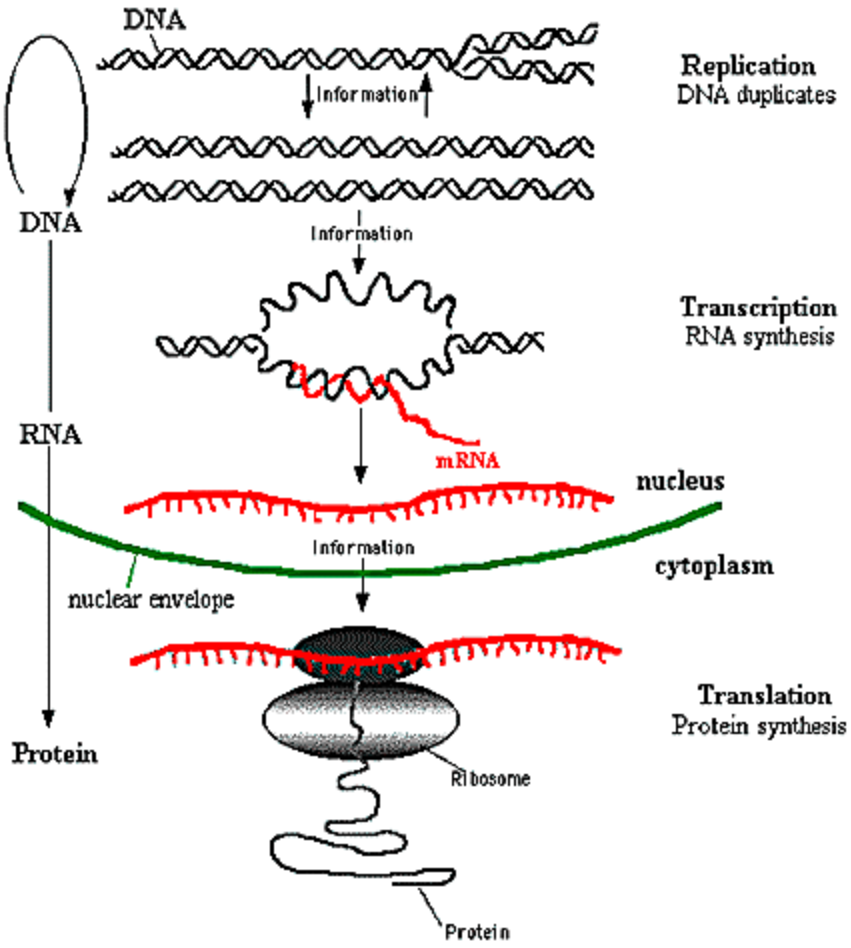
http://sandwalk.blogspot.ca/2007/05/theme-three-domain-hypothesis.html

Extended view of ourselves as a lifeform

- We are composite of species: a ‘supra-organism’
- Our microbial census exceeds the total number of our own human cells by ~10 fold
- Our largest collection of microbes resides in the intestine (~10-100 trillion organisms)
- The aggregate genomes of these gut species (microbiome) may contain >100 fold more genes than our ‘own’ genome
- The microbiome is an integral part of our genetic landscape (‘human metagenome’) and of our genetic evolution

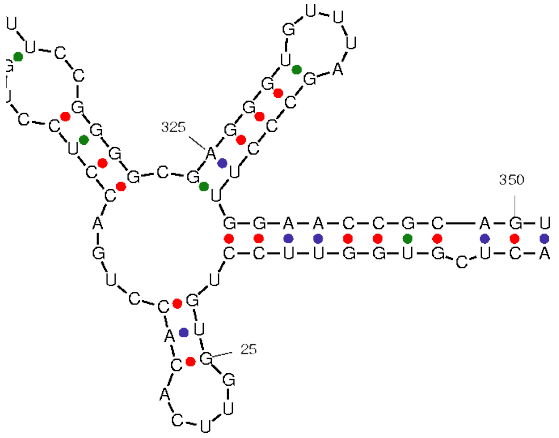


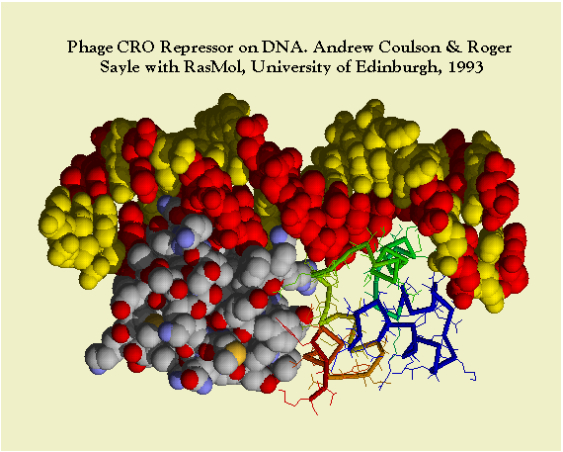
Central dogma



The Central Dogma of Molecular Biology

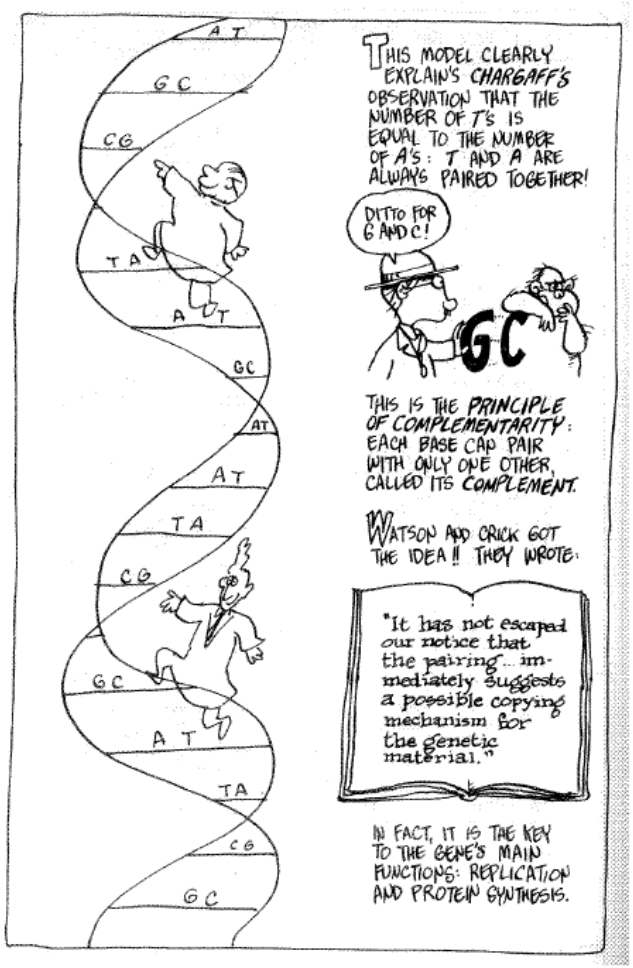
AGGTACGCGTACCTGACAGG

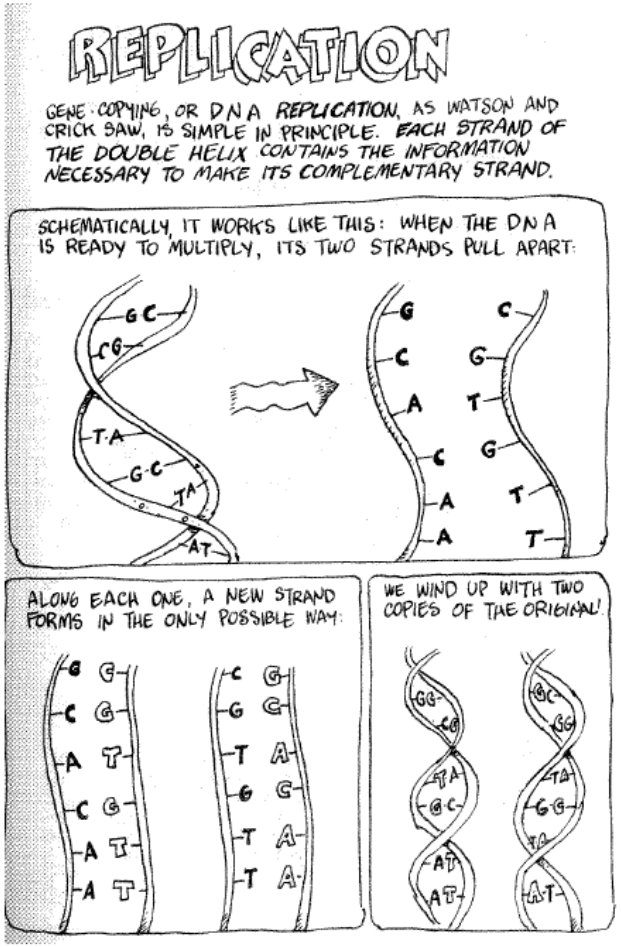




http://www.accessexcellence.org/RC/VL/GG/central.html

DNA Replication





The Cartoon Guide to Genetics
Larry Gonick & Mark Wheelis, 1983

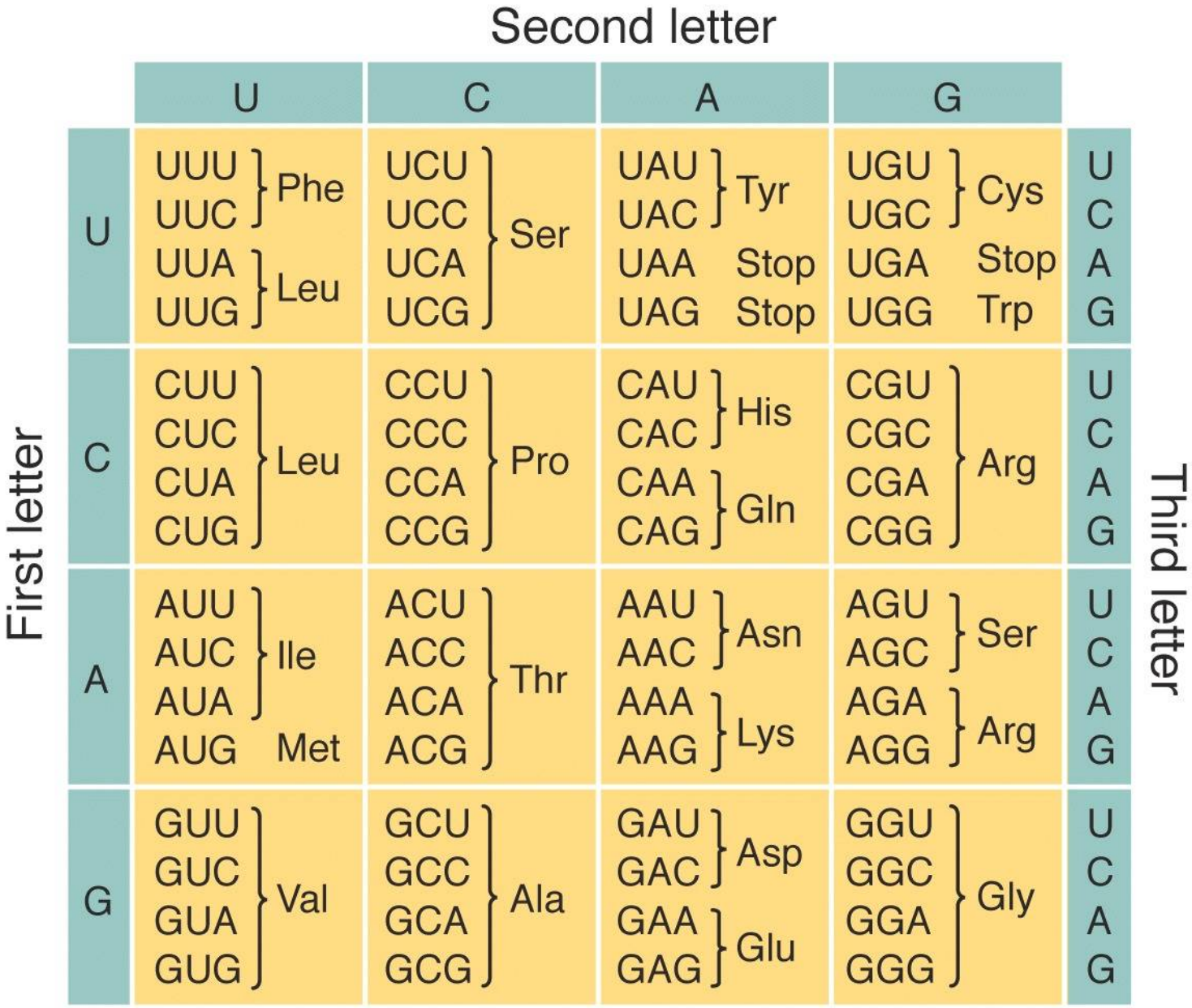
Genes, transcription, translation

- DNA – RNA - Thymine replaced by Uracil (T-U)
- The transcribed segments are called genes

ACCGUACC**AUG**UUA...AUAGGC**UGA**GCA

- AUG – start codon (also amino-acid Methionine)
- UAA, UAG, UGA – stop codons
- Genes are read in sets of 3 nucleotides during translation – $4^3 = 64$ possible combinations
- Each combination codes for one of 20 amino-acids – the building blocks for proteins

Amino-acid translation table



DNA in the computer

- FASTA/multi-FASTA file format

```
>gi|110227054|gb|AE004091.2| Pseudomonas aeruginosa PAO1, complete
   genome
TTTAAAGAGACCGGCGATTCTAGTGAAATCGAACGGGCAGGTCAATTTCCAACCAGCGATGACGTAATAG
ATAGATACAAGGAAGTCATTTTTCTTTTAAAGGATAGAAACGGTTAATGCTCTTGGGACGGCGCTTTTCT
GTGCATAACTCGATGAAGCCCAGCAATTGCGTGTTTCTCCGGCAGGCAAAAGGTTGTCGAGAACCGGTGT
CGAGGCTGTTTCCTTCCTGAGCGAAGCCTGGGGATGAACGAGATGGTTATCCACAGCGGTTTTTTCCACA
CGGCTGTGCGCAGGGATGTACCCCCTTCAAAGCAAGGGTTATCCACAAAGTCCAGGACGACCGTCCGTCG
```

- Parsers easy to write, also available in a variety of software libraries

Genes/proteins in the computer

- >gi|15596155|ref|NP_249649.1| basic amino acid,
- MKVMKWSAIALAVSAGSTQFAVADAFVSDQAEAKGFIEDSSLDLLLRNYYFNRDGKSGSGDRVDWTQGFL
- TTYESGFTQGTVGFGVDAFGYLGLKLDGTSDKTGTGNLPVMNDGKPRDDYSRAGGAVKVRISKTMLKWGE
- MQPTAPVFAAGGSRLFPQTATGFQLQSSEFEGLDLEAGHFTEGKEPTTVKSRGELYATYAGETAKSADFI
- GGRYAITDNLSASLYGAELEDIYRQYYLNSNYTIPLASDQSLGFDFNIYRTNDEGKAKAGDISNTTWSLA
- AAYTLDAHTFTLAYQKVHGDQPFDYIGFGRNGSGAGGDSIFLANSVQYSDFNGPGEKSWQARYDLNLASY
- GVPGLTFMVRYINGKDIDGTKMSDNNVGYKNYGYGEDGKHHETNLEAKYVVQSGPAKDLSFRIRQAWHRA
- NADQGEGDQNEFRLIVDYPLSIL

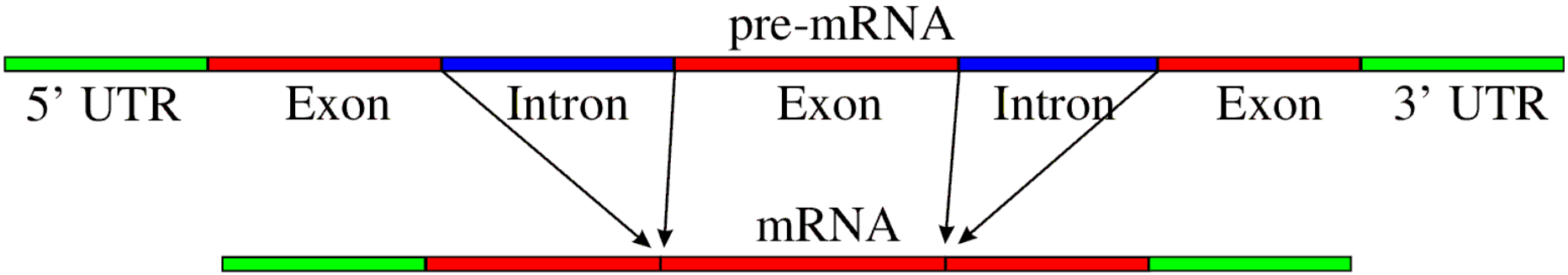
- Same FASTA/multi-FASTA but with bigger alphabet

Genes/proteins in the computer

```
     gene            complement(1043983..1045314)
                     /gene="oprD"
                     /locus_tag="PA0958"
     CDS             complement(1043983..1045314)
                     /gene="oprD"
                     /locus_tag="PA0958"
                     /note="Product name confidence: Class 1 (Function
                     experimentally demonstrated in P. aeruginosa)"
                     /codon_start=1
                     /transl_table=11
                     /product="Basic amino acid, basic peptide and
                     imipenem outer membrane porin OprD precursor"
                     /protein_id="AAG04347.1"
                     /db_xref="GI:9946864"
```

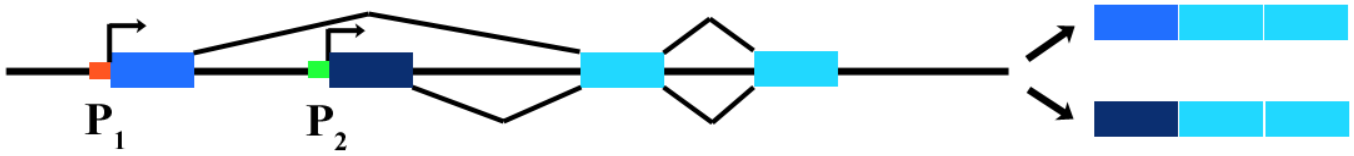
- GenBank file format

Translation – complications

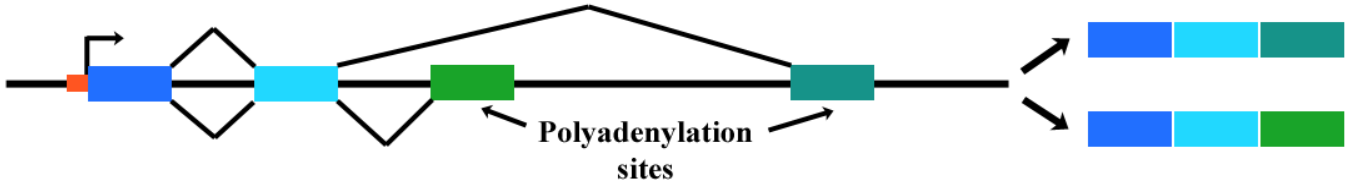


Alternative splicing examples

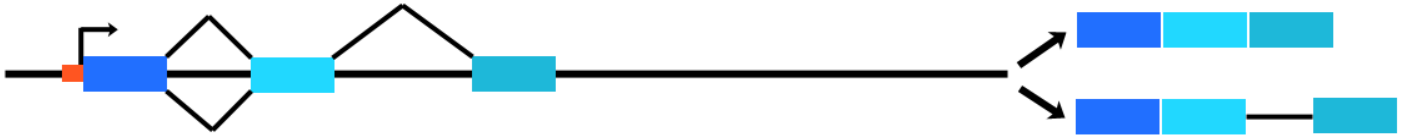
(a) Alternative selection of promoters (e.g., *myosin* primary transcript)



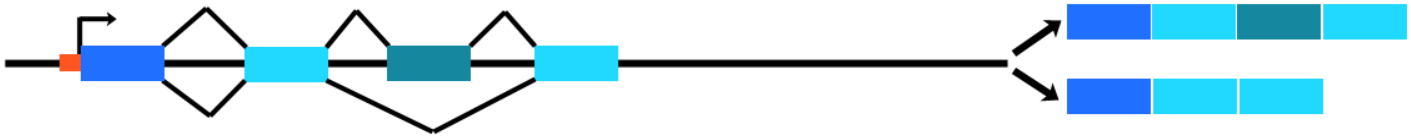
(b) Alternative selection of cleavage/polyadenylation sites (e.g., tropo*myosin* transcript)



(c) Intron retaining mode (e.g., *transposase* primary transcript)



(d) Exon cassette mode (e.g., *troponin* primary transcript)



RECAP

- DNA is a string formed with letters A, C, T, G (called nucleotides or bases)
- DNA is double-stranded – allows replication: transfer of genetic "code" from parents to offspring
- DNA is naturally oriented from 5′ to 3′ and the two strands are anti-parallel
- If you know the sequence of one strand, you can obtain the sequence of the other by reverse-complementation

```
5′ AGACCTAGTGCACGGCTACTACC 3′

5′ CCATCATCGGCACGTGATCCAGA 3′   Reverse

5′ GGTAGTAGCCGTGCACTAGGTCT 3′   Complement
```

RECAP

- Central Dogma of molecular biology:
  - DNA – RNA (transcription)
  - RNA – Protein (translation)
- The transcribed segments of DNA are called "genes"
- Translation occurs in sets of 3 nucleotides – codons
- Each codon encodes one of 20 amino-acids and 3 stop-codons
- In eukaryotes the genes may be split into multiple exons, separated by introns: DNA segments that will not get translated
- The protein is translated from an RNA representing the concatenation of the exons of the gene

The "new" biology

- DNA is not the only heritable information
  - Epigenetic information: RNA molecules, DNA methylation patterns (affects coiling on DNA on histones)
- Complex regulation patterns
  - Genes turn on other genes
  - Genes inhibit other genes
  - RNA interference – small RNA molecules can destroy specific transcripts (down-regulate production)

Playing with DNA

- Biologists can:
- Cut the DNA – restriction enzymes (often palindromes) (Nobel prize – Arber, Nathans, Smith)

```
5'GAATTC          5'---G      AATTC---3'
3'CTTAAG          3'---CTTAA        G---5'
```

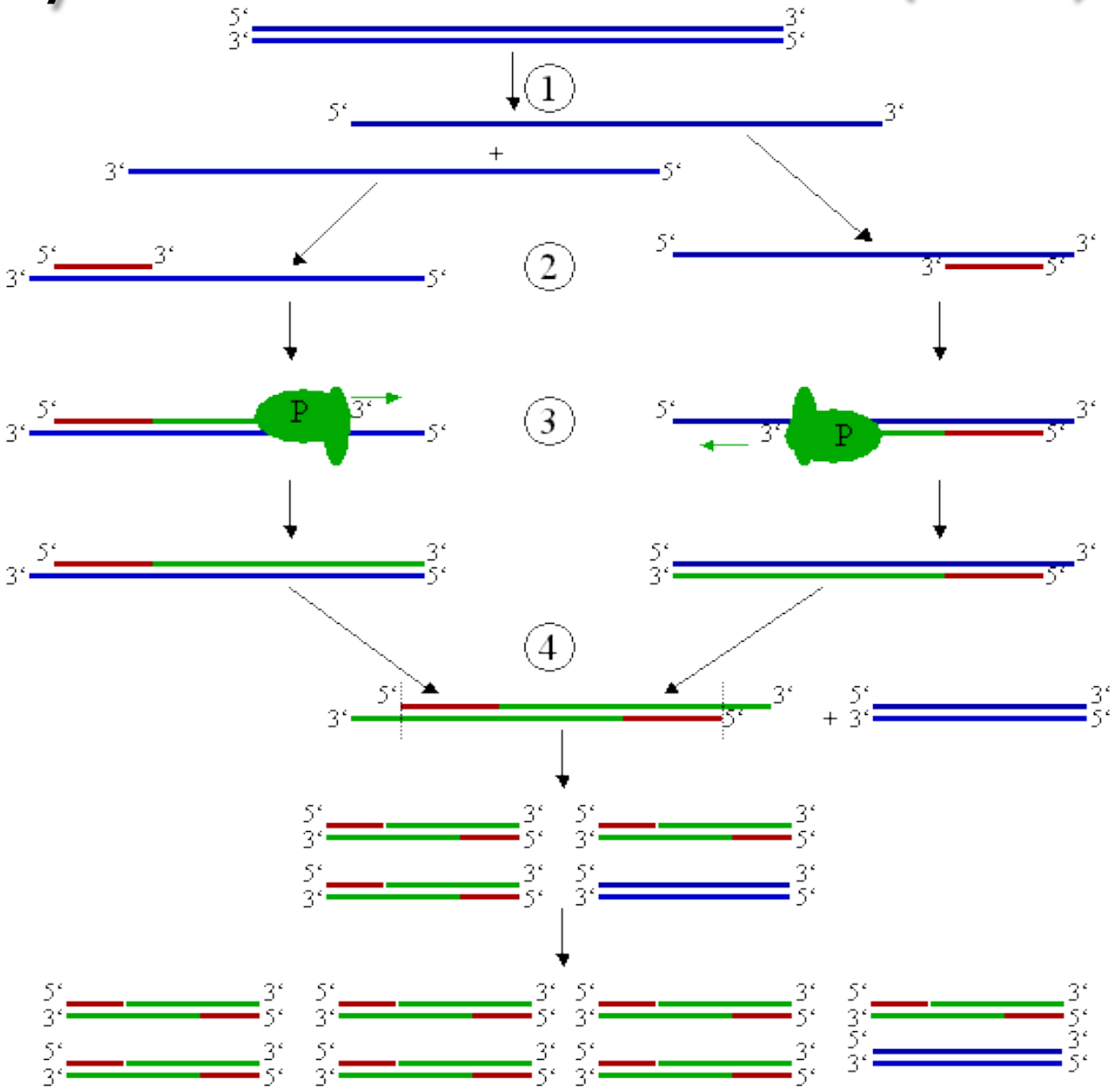
- Attach "things" to DNA (either single or double-strand)

```
TAGGCACGTTGCAACTACGGC

      TGCAACGT
```

- "Amplify" DNA – Polymerase Chain Reaction (Nobel prize – Mullis)

Polymerase chain reaction (PCR)



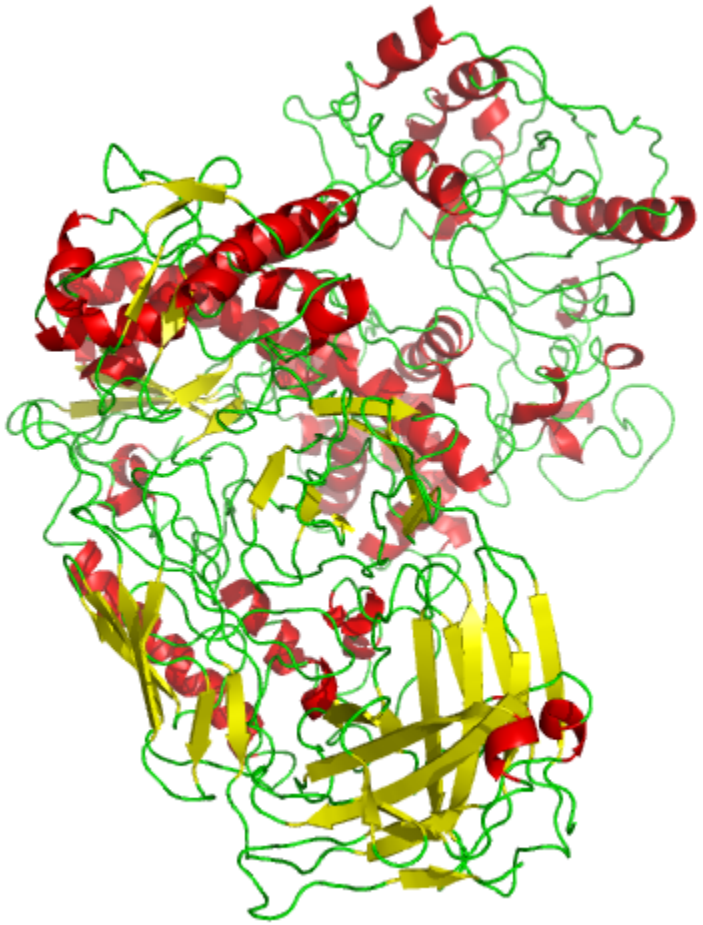
1. Denature

2. Anneal (attach primer)

3. Extend

4. Repeat

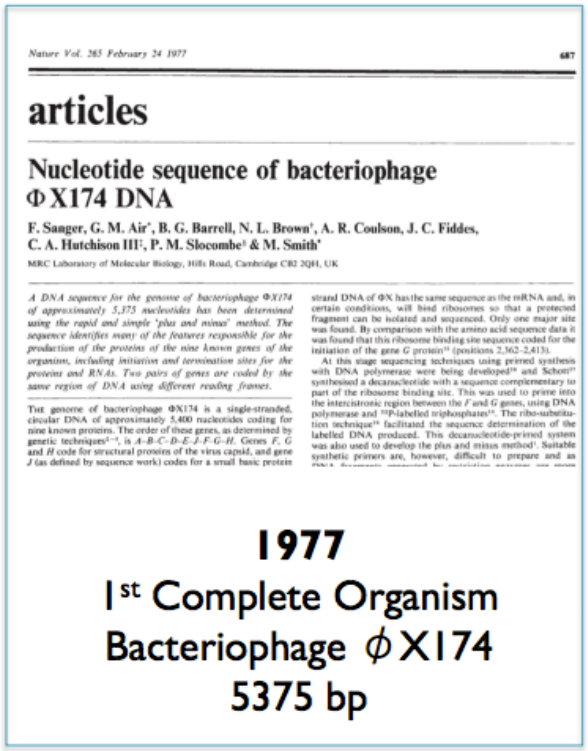
Taq polimerase



DNA sequencing

- Most techniques "trick" the polymerase into revealing the sequence
- The traditional method – Sanger sequencing – based on "terminator" bases – prevent the polymerase from extending the DNA
- Sanger sequencing is essentially PCR + terminator bases
- Other methods "spy" on the polymerase as it incorporates nucleotides

Milestones in Molecular Biology



Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe‡ & M. Smith*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

The genome of bacteriophage ΦX174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques[1–3], is *A–B–C–D–E–J–F–G–H*. Genes *F*, *G* and *H* code for structural proteins of the virus capsid, and gene *J* (as defined by sequence work) codes for a small basic protein

strand DNA of ΦX has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene *G* protein[14] (positions 2,362–2,413).

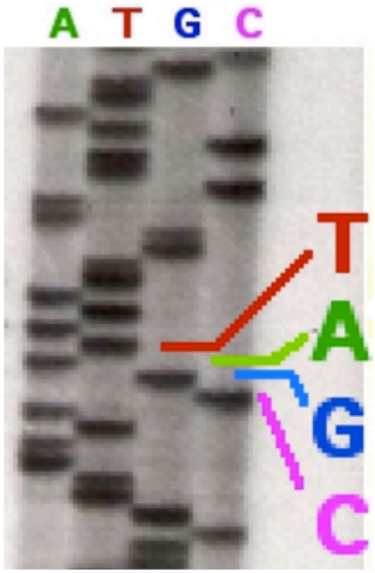
At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[15] and Schott[17] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *F* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[15]. The ribo-substitution technique[16] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[1]. Suitable synthetic primers are, however, difficult to prepare and as

1977

1st Complete Organism

Bacteriophage ϕX174

5375 bp



Radioactive Chain Termination

5000bp / week / person

http://en.wikipedia.org/wiki/File:Sequencing.jpg

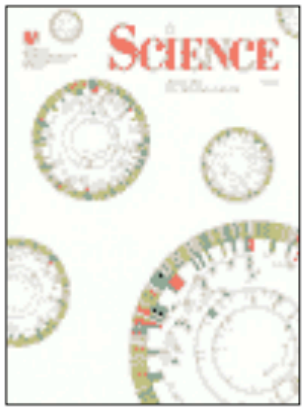
http://www.answers.com/topic/automated-sequencer

Nucleotide sequence of bacteriophage ϕX174 DNA

Sanger, F. et al. (1977) *Nature. 265: 687 - 695*

http://schatzlab.cshl.edu/teaching

Milestones in Molecular Biology



1995
Fleischmann *et al.*
1st Free Living Organism
TIGR Assembler. 1.8Mbp

2000
Myers *et al.*
1st Large WGS Assembly.
Celera Assembler. 116 Mbp

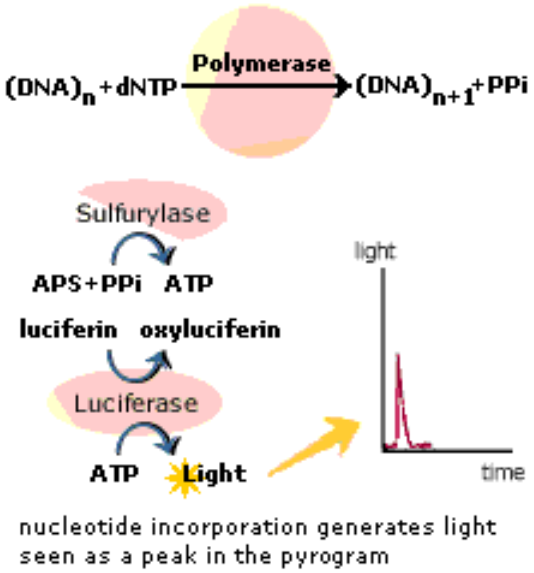
2001
Venter *et al.* / IHGSC
Human Genome
Celera Assembler. 2.9 Gbp

ABI 3700: 500 bp reads x 768 samples / day = 384,000 bp / day.
"The machine was so revolutionary that it could decode in a single day the same amount of genetic material that most DNA labs could produce in a year." J. Craig Venter

The future of sequencing

- Single molecule sequencing - current technology requires many copies of DNA being sequenced - requires DNA amplification
- Massively-parallel sequencing - 100k sequencing reactions occuring at the same time

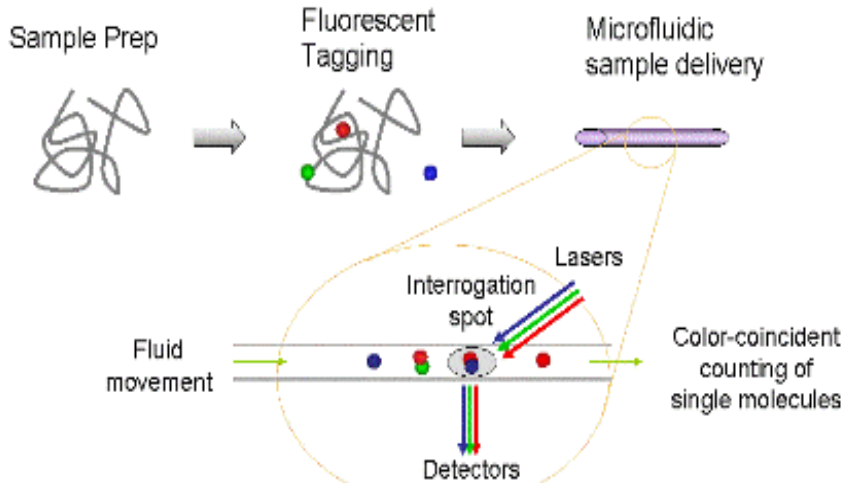
Sequencing by synthesis



TCTAATAGA
AGATTATCTAACAGCTACCCTTCCATCA

http://www.genetics.ucla.edu/sequencing/pyro.php

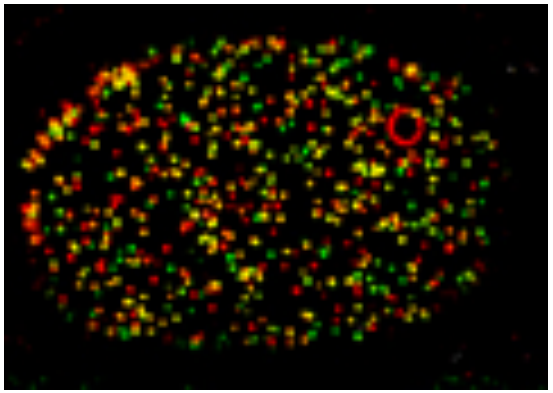
Micro-fluidics



http://www.usgenomics.com

The future of sequencing

Massively parallel sequencing



http://arep.med.harvard.edu/

- each spot is a molecule or amplified from one molecule
- image processing used to track molecules during sequencing by synthesis
- often micro-fluidics/lab-on-a-chip used

- More on this in two weeks!

The evolution of DNA sequencing

Since	Technology	Read length	Throughput/run	Throughput/ hour	cost/run
1977-	Sanger sequencing	> 1000bp	4hr 400-500 kbp	100 kbp	$200
2005-	454 pyrosequencing	250-400bp	4hr 100-500 Mbp	25-100 Mbp	$13,000
2006-	Illumina/Solexa	50-100bp	3 days 2-3 Gbp	25-40 Mbp	$3,000
2007-	ABI SOLiD	35-50bp	3 days 6-20 Gbp	75-250 Mbp	est. $3-5,000
2010-	Illumina HiSeq 2000	2X100 bp	11 days 600 Gbp	2200 Mbp	~$30,000
2010-	Pacific Biosciences **single molecule**	1-10 kbp	1 day 2.4 Gbp	100 Mbp	$1,000
2012-	**Oxford Nanopore**	80-100 kbp	?	?	?

In the news (more info on website)

Published Online August 30 2012

Science DOI: 10.1126/science.1224344

< *Science* Express Index

Read Full Text to Comment (0)

RESEARCH ARTICLE

A High-Coverage Genome Sequence from an Archaic Denisovan Individual

Matthias Meyer[1,*,‡], Martin Kircher[1,*,†], Marie-Theres Gansauge[1], Heng Li[2], Fernando Racimo[1], Swapan Mallick[2,3], Joshua G. Schraiber[4], Flora Jay[4], Kay Prüfer[1], Cesare de Filippo[1], Peter H. Sudmant[6], Can Alkan[4,5], Qiaomei Fu[1,7], Ron Do[2], Nadin Rohland[2,3], Arti Tandon[2,3], Michael Siebauer[1], Richard E. Green[8], Katarzyna Bryc[3], Adrian W. Briggs[3], Udo Stenzel[1], Jesse Dabney[1], Jay Shendure[6], Jacob Kitzman[6], Michael F. Hammer[9], Michael V. Shunkov[10], Anatoli P. Derevianko[10], Nick Patterson[2], Aida M. Andrés[1], Evan E. Eichler[6,11], Montgomery Slatkin[4], David Reich[2,3,*,‡], Janet Kelso[1], Svante Pääbo[1,‡]

± Author Affiliations

± Author Notes

‡To whom correspondence should be addressed. E-mail: mmeyer@eva.mpg.de (M.M.); reich@genetics.med.harvard.edu (D.R.); paabo@eva.mpg.de (S.P.)

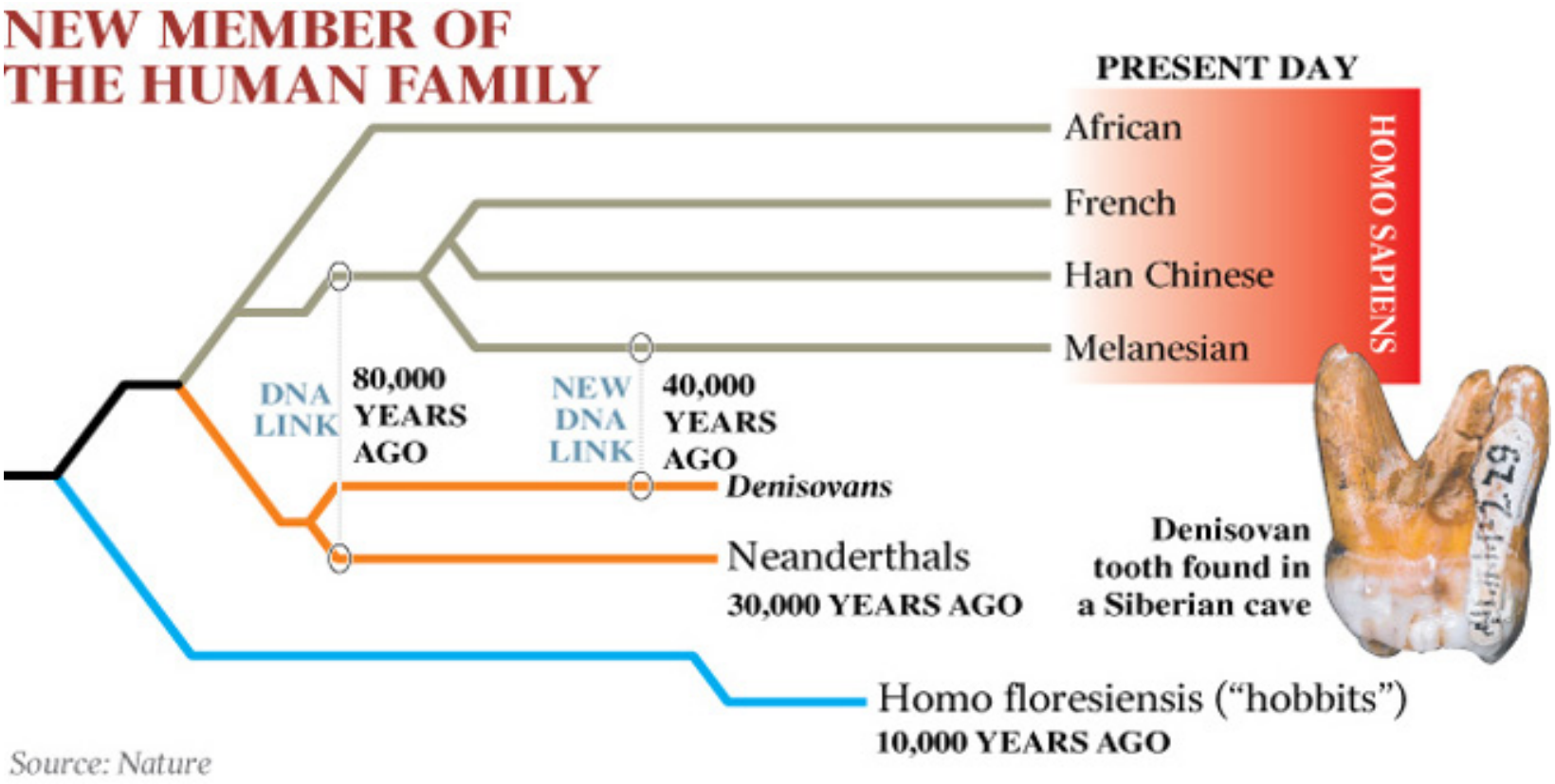
* These authors contributed equally to this work.

ABSTRACT

We present a DNA library preparation method that has allowed us to reconstruct a high-coverage (30X) genome sequence of a Denisovan, an extinct relative of Neandertals. The quality of this genome allows a direct estimation of Denisovan heterozygosity, indicating that genetic diversity in these archaic hominins was extremely low. It also allows tentative dating of the specimen on the basis of "missing evolution" in its genome, detailed measurements of Denisovan and Neandertal admixture into present-day human populations, and the generation of a near-complete catalog of genetic changes that swept to high frequency in modern humans since their divergence from Denisovans.

DNA sequenced from bone is 30,000 to 50,000 years old!

Documents of evolutionary history



Recap

- Central dogma of biology: DNA -> RNA -> Proteins
  - DNA encodes genes, most of which encode for proteins (via the genetic code)
  - Proteins perform much of the work of the cell.
  - RNA acts as an intermediate step

    (it also has other functions as well)
- Huge amount of data now available, need algorithms to make sense of it.

Notes & next lecure

- Homework #1 is posted
- Slides will be posted shortly
- Reading material is posted

- Next lecture:
  - Bioinformatics programming, DBs
  - Go over project specification
  - Might start string comparison