CSE 397-497:
Computational Issues in Molecular Biology

Lecture 2

Spring 2004

Important points to remember

- Class attendance and participation are big part of grade.
- Do assigned readings in advance – be prepared to discuss. (Lecture notes and readings will be posted on Blackboard.)
- Each student enrolled must prepare and deliver one lecture.
- Use of PowerPoint for slides is encouraged, but not required. (See http://www.openoffice.org for open-source version.)
- Final project or paper due at end of course.
- Final exam in event seminar format is not successful.

Procedure for student lectures

By 5:00 pm on Thursday, you give me ranked list of your top 3 topics in order of preference:

order we will cover topics in course ↓

- sequence comparison & alignment (pairwise & multiple),
- sequencing and sequence assembly,
- physical mapping of DNA,
- phylogenetic trees,
- genome rearrangements,
- RNA and protein structure prediction,
- DNA microarrays,
- DNA computing.

For each topic, there is a folder in Blackboard. When making your choice, look at these papers and also your textbook.

Procedure for student lectures

By 5:00 pm on Friday, I will assign your topic and lecture date. (If you know you have a conflict for a specific date, tell me in advance – i.e., when you send me your ranked list of topics.)

Two weeks before your lecture, you meet with me to discuss the material you plan to present (15 minutes).

One week before your lecture, you show me your near-final lecture and a list of discussion topics ($\leq$ 2 hours).

One day before your lecture, we do a run-through ($\leq$ 2 hours).

Course grading

Class attendence / participation = 100 points.

Lecture = 25 points (preparation) + 100 points (delivery).

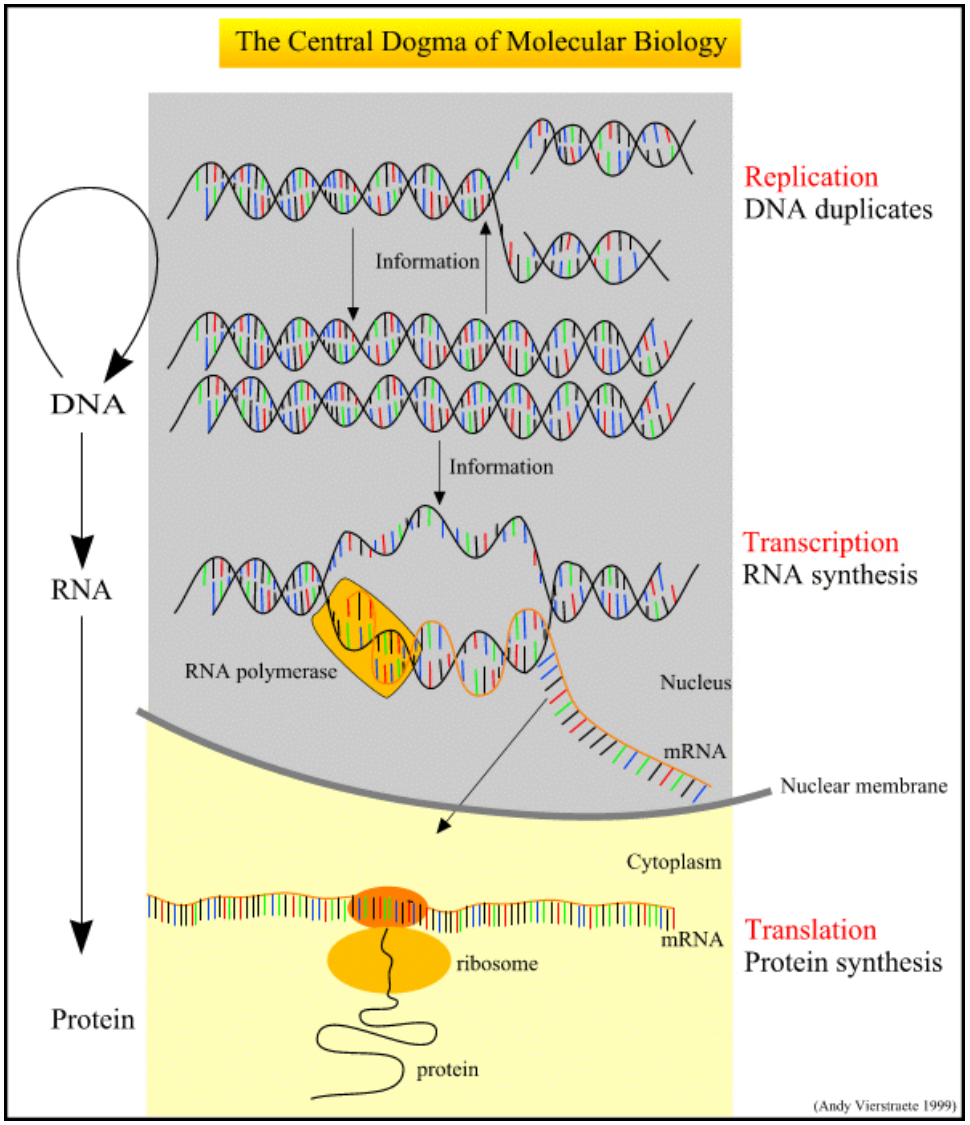
Final project or paper = 100 points.

Scribe (if taking CSE 497)* = 25 points.

Final exam (if we need it) = 100 points.

** Note that CSE 397 and CSE 497 point totals will be different and each will be curved separately.*

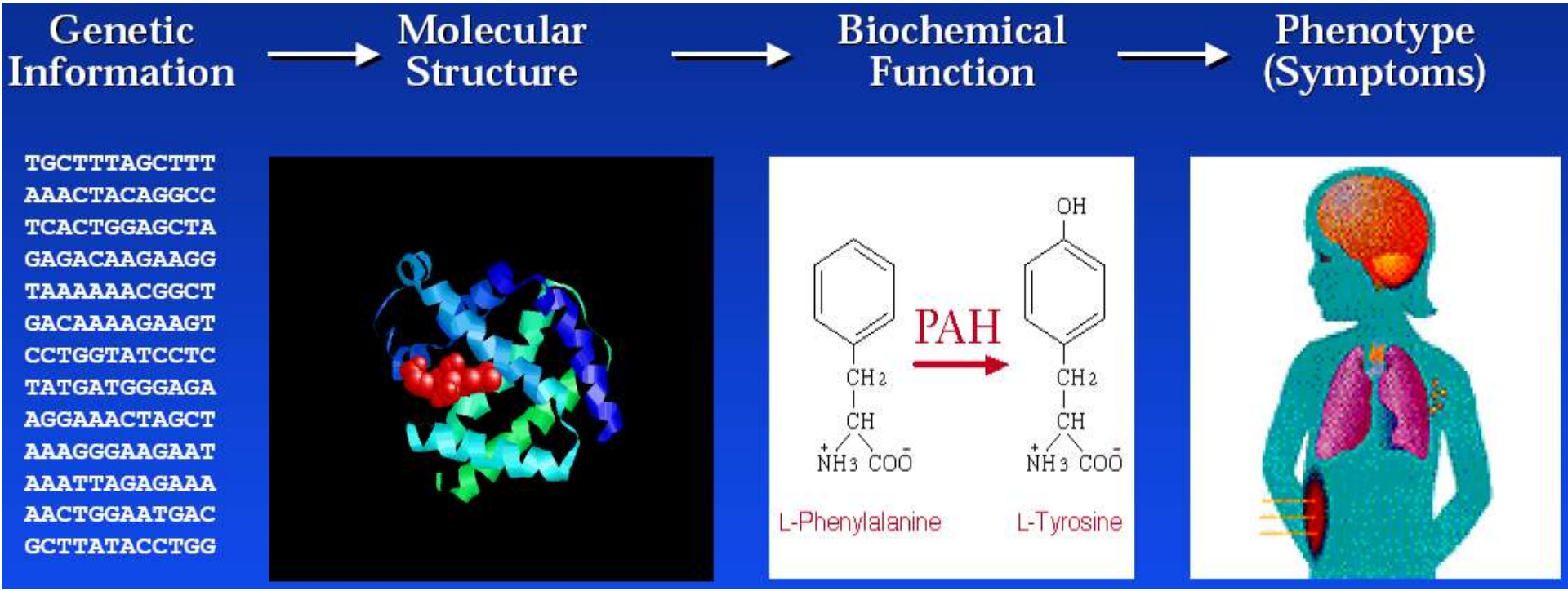
The Central Dogma of Molecular Biology



1. DNA copies its information in process involving many enzymes (replication).
2. DNA codes for production of mRNA during transcription.
3. mRNA migrates from nucleus to cytoplasm.
4. mRNA carries coded information to ribosomes which "read" it and use it for protein synthesis (translation).

http://allserv.rug.ac.be/~avierstr/principles/centraldogma.html

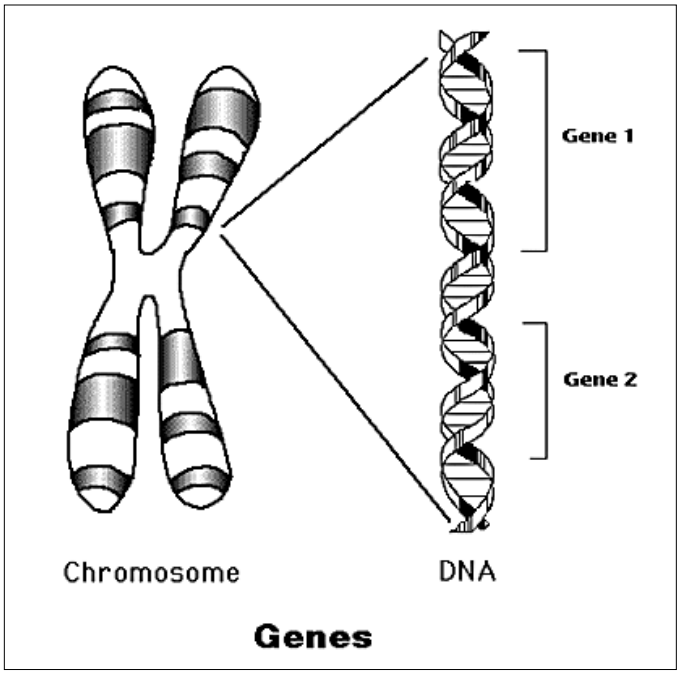
The Central Paradigm of Bioinformatics



By developing techniques for analyzing sequence data and the structures that result, we can attempt to understand the genetic nature of diseases.

http://cmgm.stanford.edu/biochem218/

"Junk" DNA



Recall that genes are contiguous stretches along a chromosome.

At this point in time, <10% of the DNA in the human genome can be associated with genes.

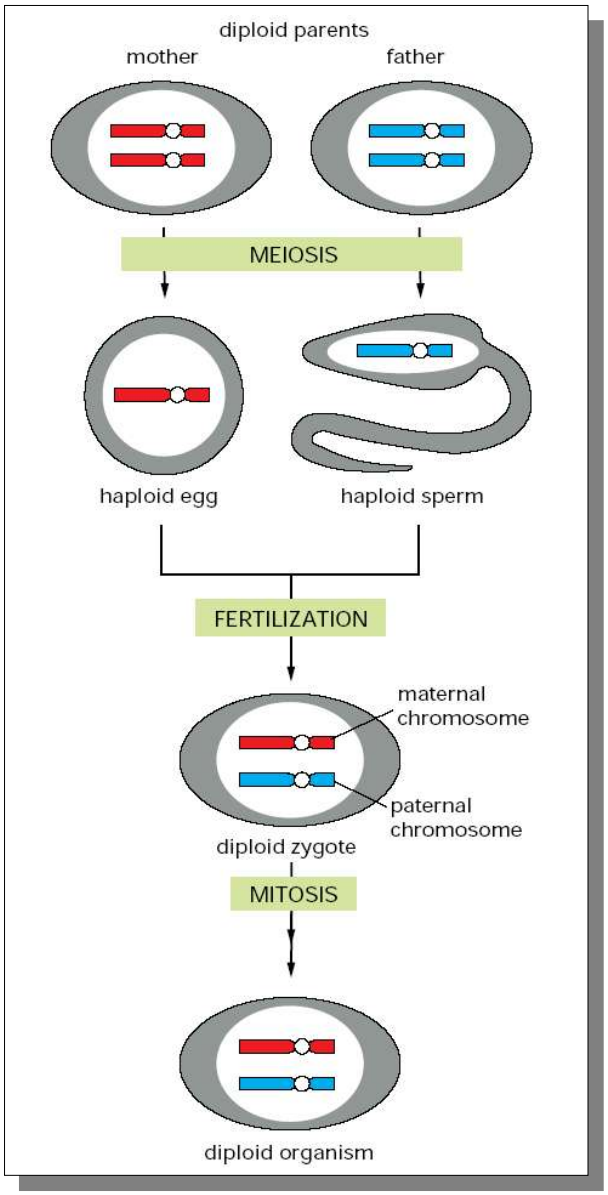
The remainder is known as *junk DNA* because it has no apparent function.

However, recent studies are showing that non-coding DNA may play an important role in regulating gene expression (enhancing or suppressing expression of proximal genes).

It's also used in forensic analysis as mutations are more likely in non-coding DNA regions than within genes (why?).

http://www.accessexcellence.org/AB/GG/genes.html
http://www.psrast.org/junkdna.htm

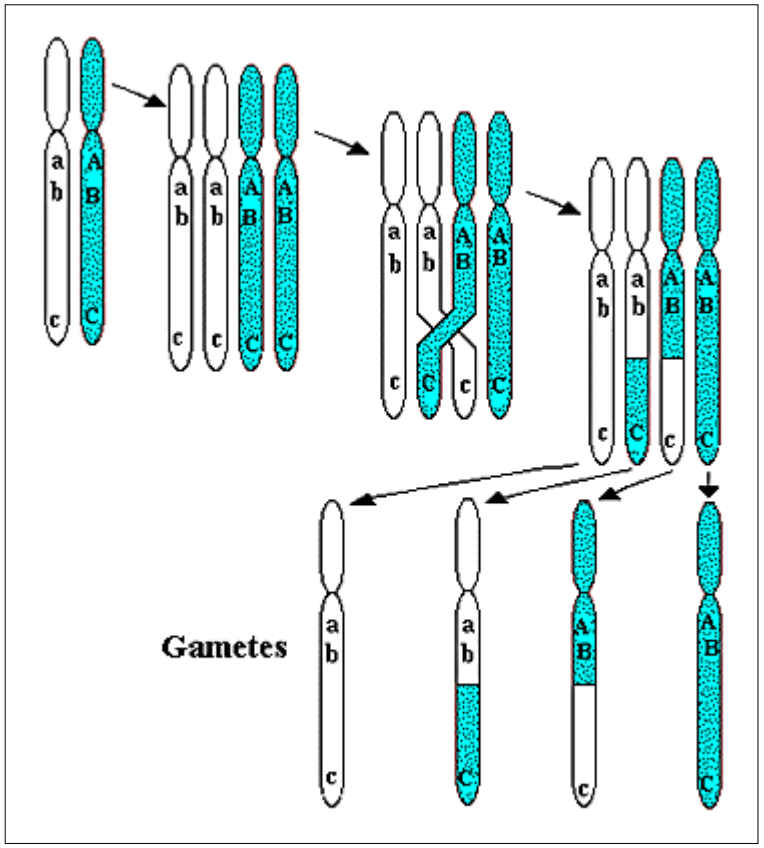
Genetic inheritence



1. Cells in mother and father both contain paired sets of chromosomes (diploid).
2. Through *meiosis*, *gametes* (sex cells) contain only one chromosome from each pair (haploid).
3. Fertilized egg cell (*zygote*) receives one chromosome from mother, one from father.
4. Zygote splits and reproduces through *mitosis* to yield multi-cellular diploid organism.

http://www.accessexcellence.org/AB/GG/hapDIP.html

Crossing over (recombination)



The two chromosomes that form a pair are called *homologous*.

During meiosis, homologous chromosomes may *cross over* (*recombine*) forming chromosomes that mix genes from each parent.

Note that liklihood of recombination is function of distance between two genes. This observation is used in creating genetic linkage maps.

Here we see recombination of gene c/C which appears in two forms (*alleles*). Genes ab (AB) are unlikely to recombine.

http://www.accessexcellence.org/AB/GG/comeiosis.html

Genomes

Complete set of chromosomes that determines an organism is known as a *genome*.

Sizes of some genomes

Mus musculus

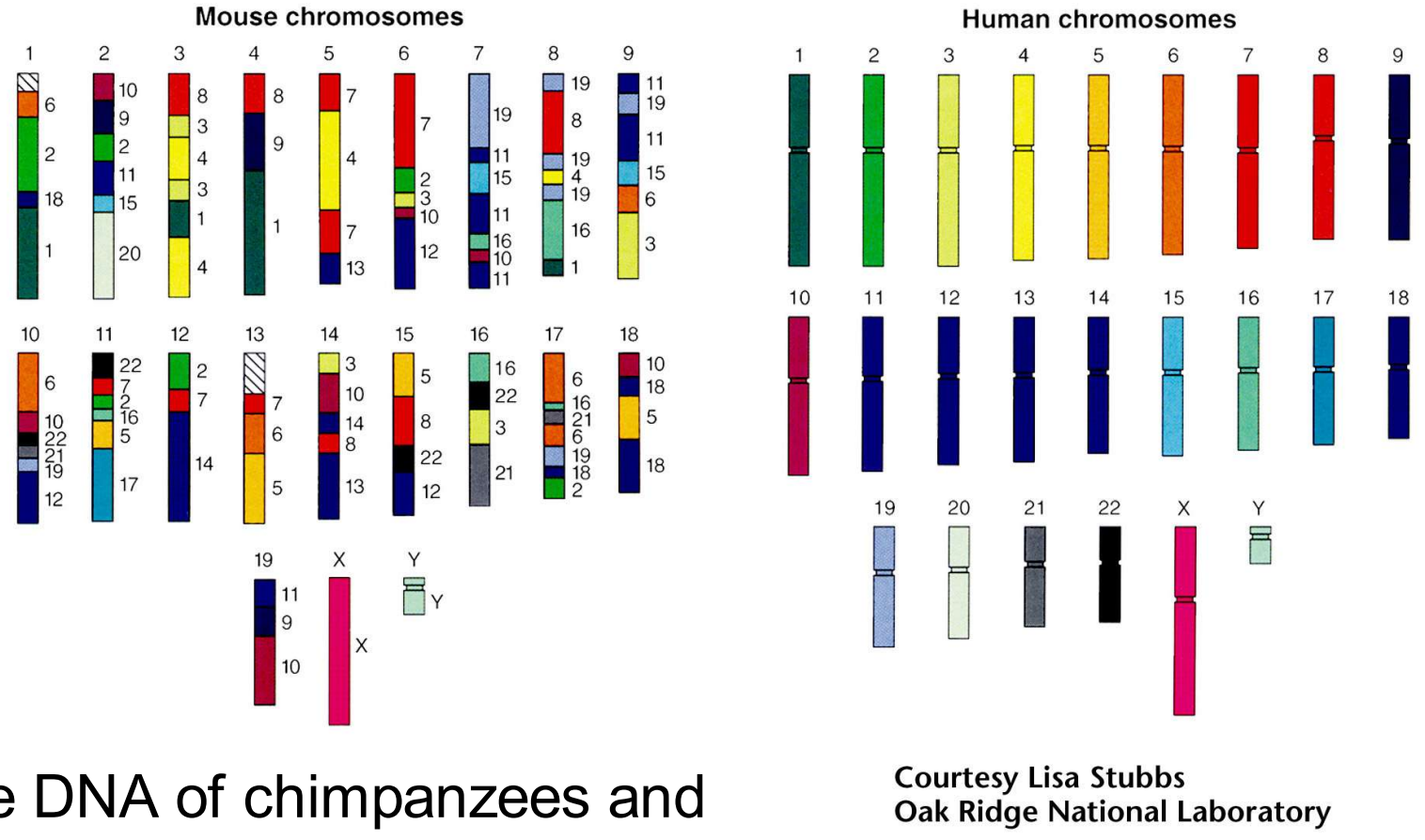
GenBank Release 121.0 — December 15, 2000

Species	Haploid genome size	Bases	Entries
Homo sapiens	3,400,000,000	6,702,881,570	3,918,724
Mus musculus	3,454,200,000	1,291,602,139	2,456,194
Drosophila melanogaster	180,000,000	487,561,384	166,554
Arabidopsis thaliana	100,000,000	242,674,129	181,388
Caenorhabditis elegans	100,000,000	203,544,197	114,553
Tetraodon nigroviridis	350,000,000	165,539,271	188,993
Oryza sativa	400,000,000	125,948,974	151,411
Rattus norvegicus	2,900,000,000	106,344,366	218,598
Bos taurus	3,651,500,000	71,215,626	159,473
Glycine max	1,115,000,000	62,817,102	141,802
Medicago truncatula	400,000,000	50,991,920	104,535
Trypanosoma brucei	35,000,000	49,855,996	91,334
Lycopersicon esculentum	655,000,000	49,415,566	97,112
Giardia intestinalis	12,000,000	47,639,714	54,328
Strongylocentrotus purpur	900,000,000	47,590,936	77,532
Entamoeba histolytica	—	44,522,016	49,938
Hordeum vulgare	—	44,489,692	57,779
Danio rerio	1,900,000,000	40,906,902	83,726
Zea mays	5,000,000,000	36,885,212	77,506
Saccharomyces cerevisiae	12,067,280	32,779,082	18,361

Note that each cell in an organism contains its entire genome!

http://www.cbs.dtu.dk/databases/DOGS/ http://www.nsrl.ttu.edu/tmot1/mus_musc.htm http://www.oardc.ohio-state.edu/seedid/single.asp?strID=324

We're more similar than you might think



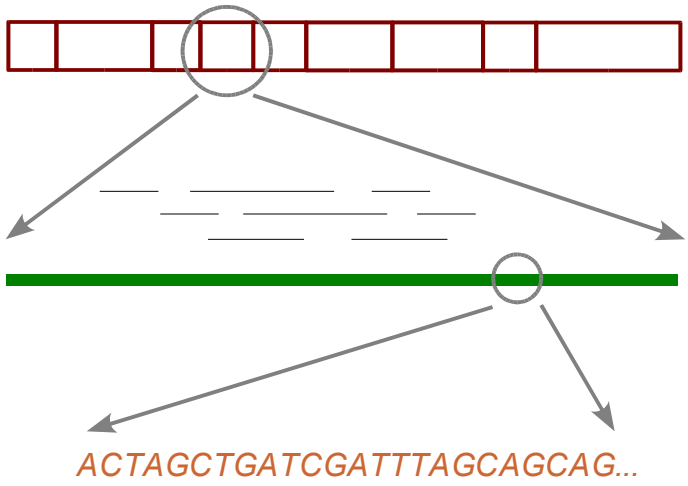
(The DNA of chimpanzees and humans is ~99% similar.)

http://www.ornl.gov/sci/techresources/Human_Genome/graphics/slides/ttmousehuman.shtml
http://www.news.cornell.edu/releases/Dec03/chimp.life.hrs.html

Studying a genome

Most genomes are enormous (e.g., 10^8 base pairs in case of human). Current sequencing technology, on the other hand, only allows biologists to determine ~10^3 base pairs at a time.

This disparatey leads to some of the most interesting problems in computational biology.

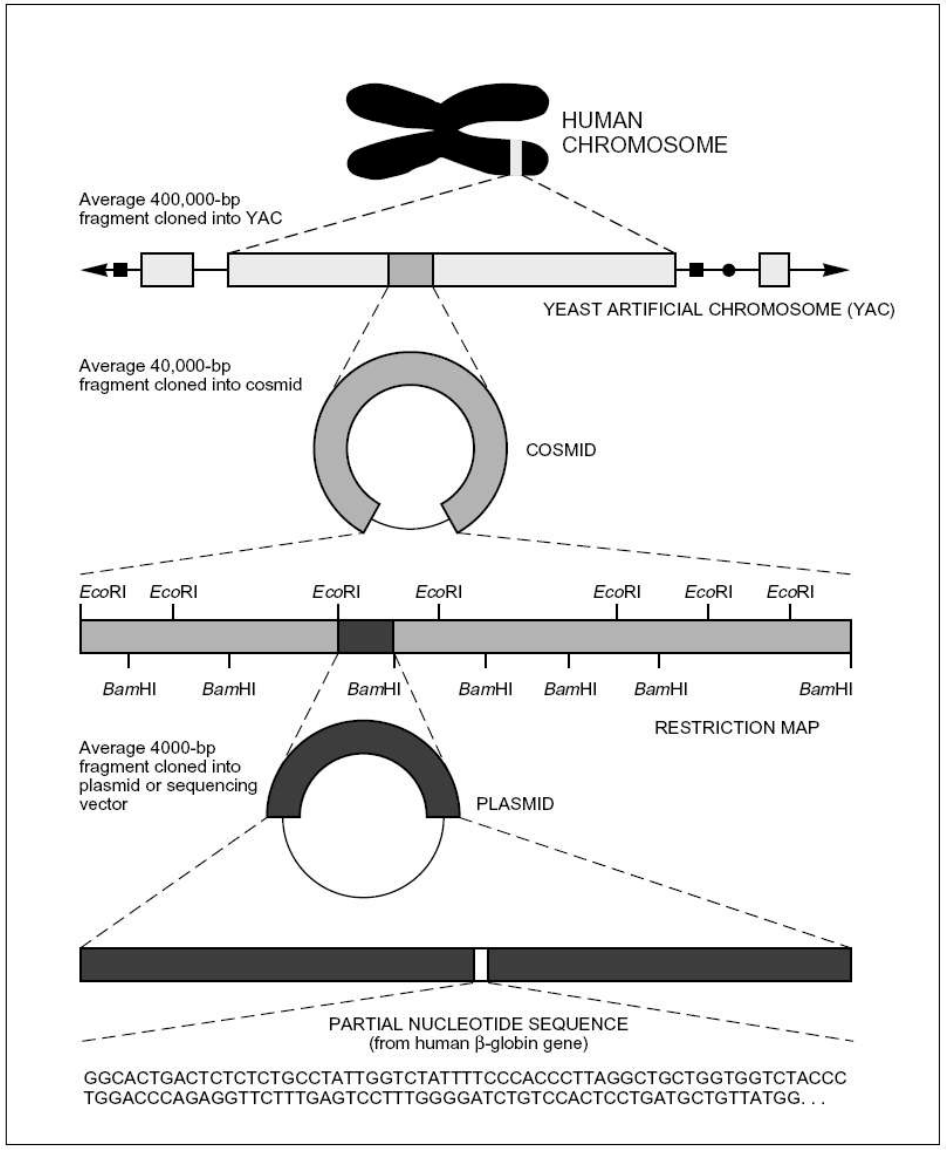


Genetic linkage map
(10^7 – 10^8 base pairs)

Physical map
(10^5 – 10^6 base pairs)

Sequencing
(10^3 – 10^4 base pairs)

Studying a genome



http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/primer.pdf

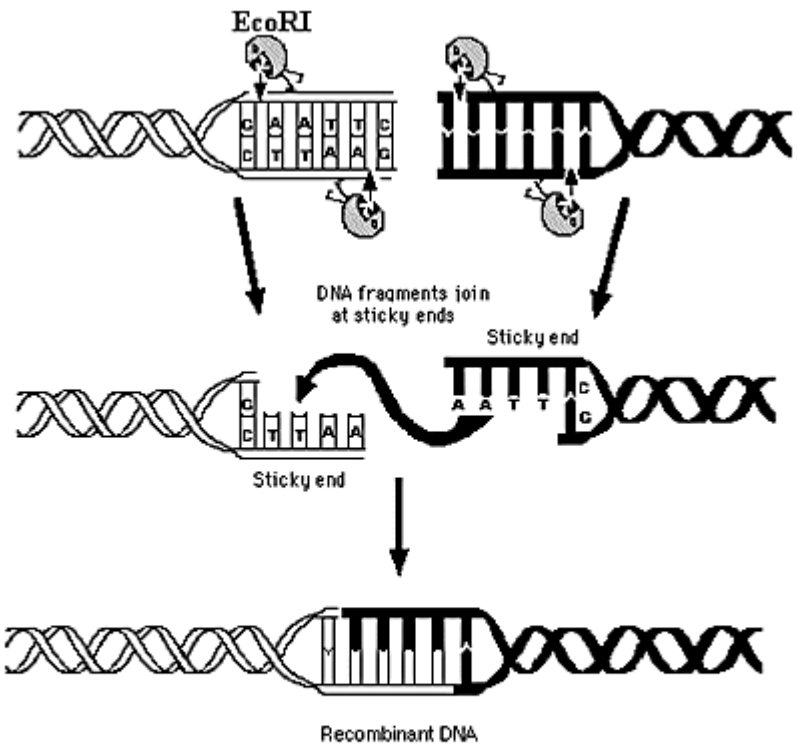
Cloned DNA molecules are made progressively smaller and fragments subcloned to obtain pieces small enough to sequence directly.

These results are compiled to provide sequence across a chromosome.

Yeast artificial chromosome (YAC) is designed to "fool" yeast replication mechanism.

Cosmids and *plasmids* are vectors that can be cloned in bacteria.

Cutting DNA using restriction enzymes



Restriction Enzyme Action of EcoRI

A *restriction enzyme* surrounds DNA molecule at specific point, called *restriction site* (sequence GAATTC in this case). It cuts one strand of DNA double helix at one point and second strand at a different, complementary point (between G and A base).

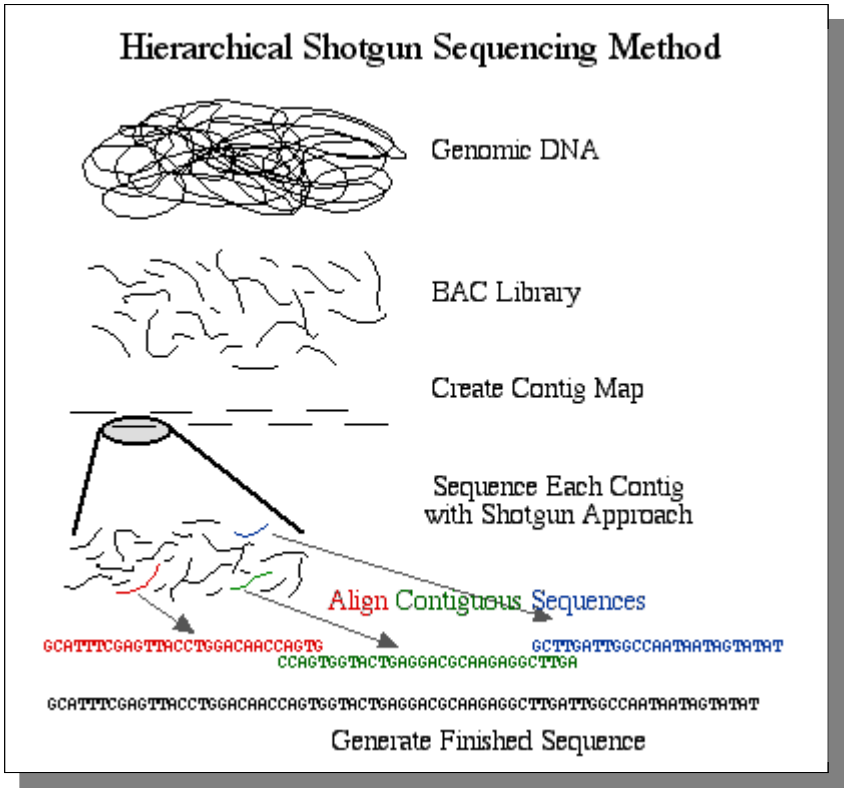
The separated pieces have single-stranded *sticky ends*, which allow complementary pieces to combine.

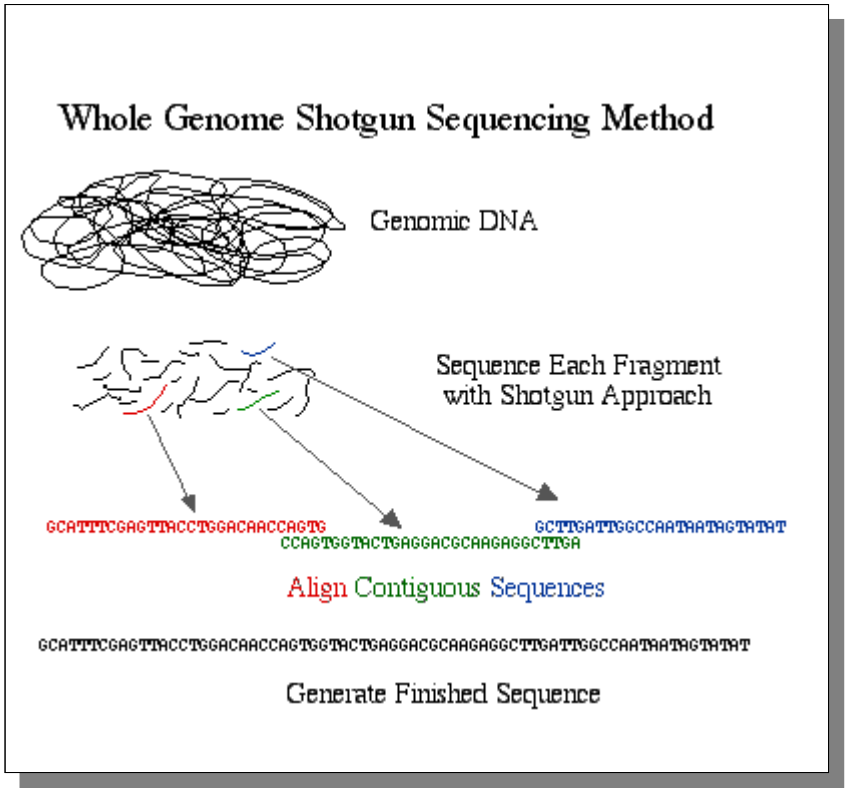
Note that $\overline{\text{GAATTC}} \rightarrow \text{CTTAAG} \rightarrow \text{GAATTC}$ (i.e., palindrome).

http://www.accessexcellence.org/AB/GG/restriction.html

Breaking DNA

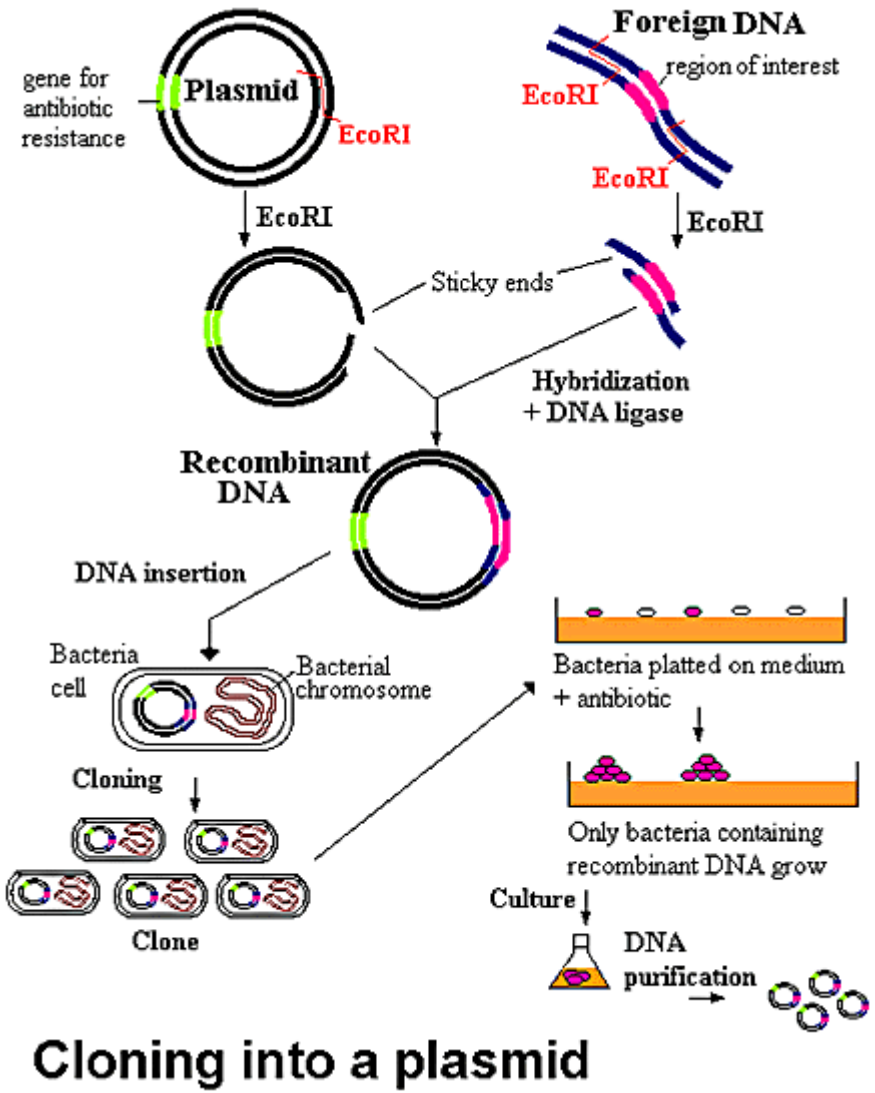
DNA can also be broken in random places through mechanical means (e.g., vibration). This is typically the first step in *shotgun sequencing*.





http://occawlonline.pearsoned.com/bookbind/pubbooks/bc_mcampbell_genomics_1/medialib/method/shotgun.html

Copying DNA



Cloning into a plasmid

http://www.accessexcellence.org/AB/GG/plasmid.html

Most analytic procedures in the lab require a quantity of the DNA under study. The process of copying DNA is known as *amplification*.

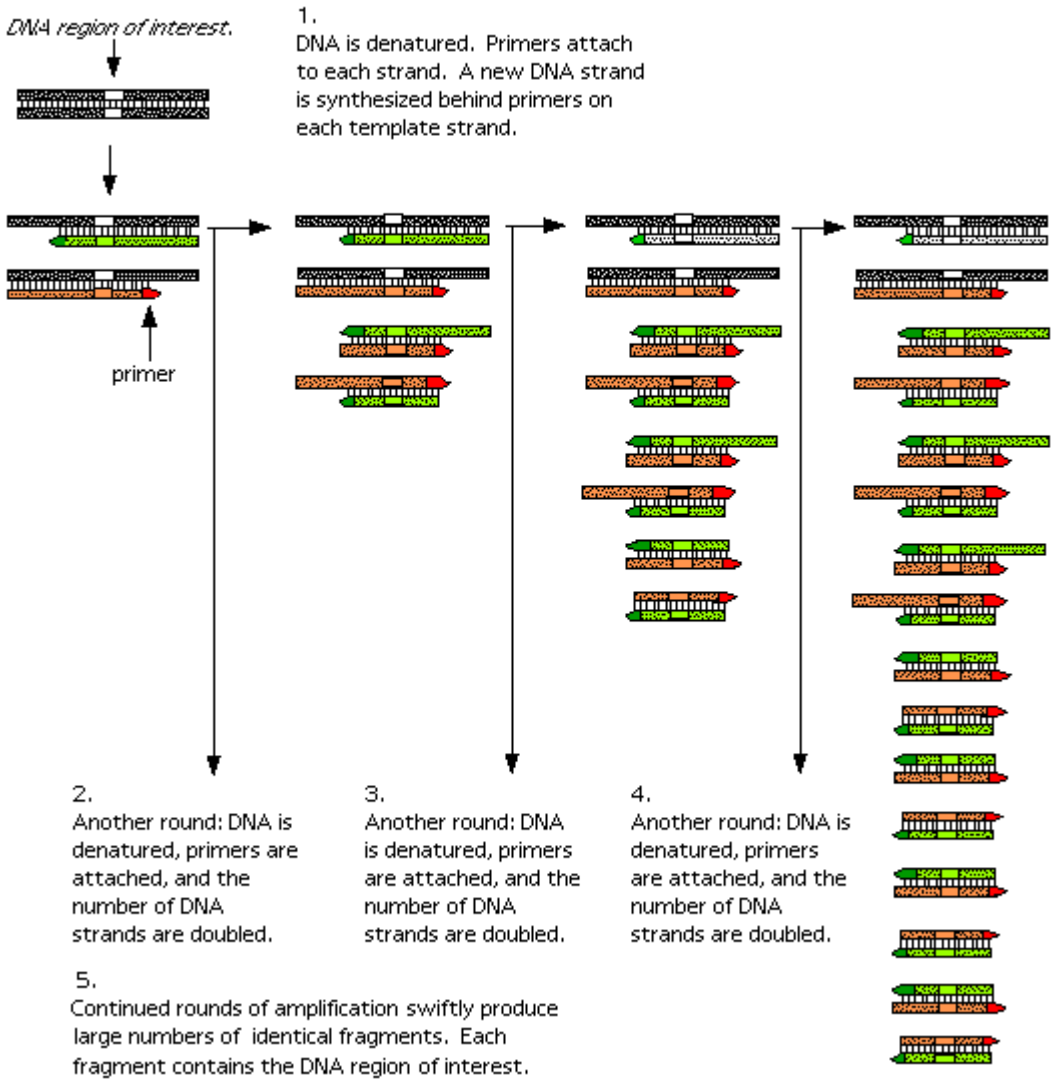
As we have seen, one possible approach is to use nature: insert the DNA of interest into the genome of a host (or vector) and let the organism multiply itself. This is called *recombinant DNA*.

Polymerase Chain Reaction

Another way to amplify DNA is *polymerase chain reaction* (*PCR*).

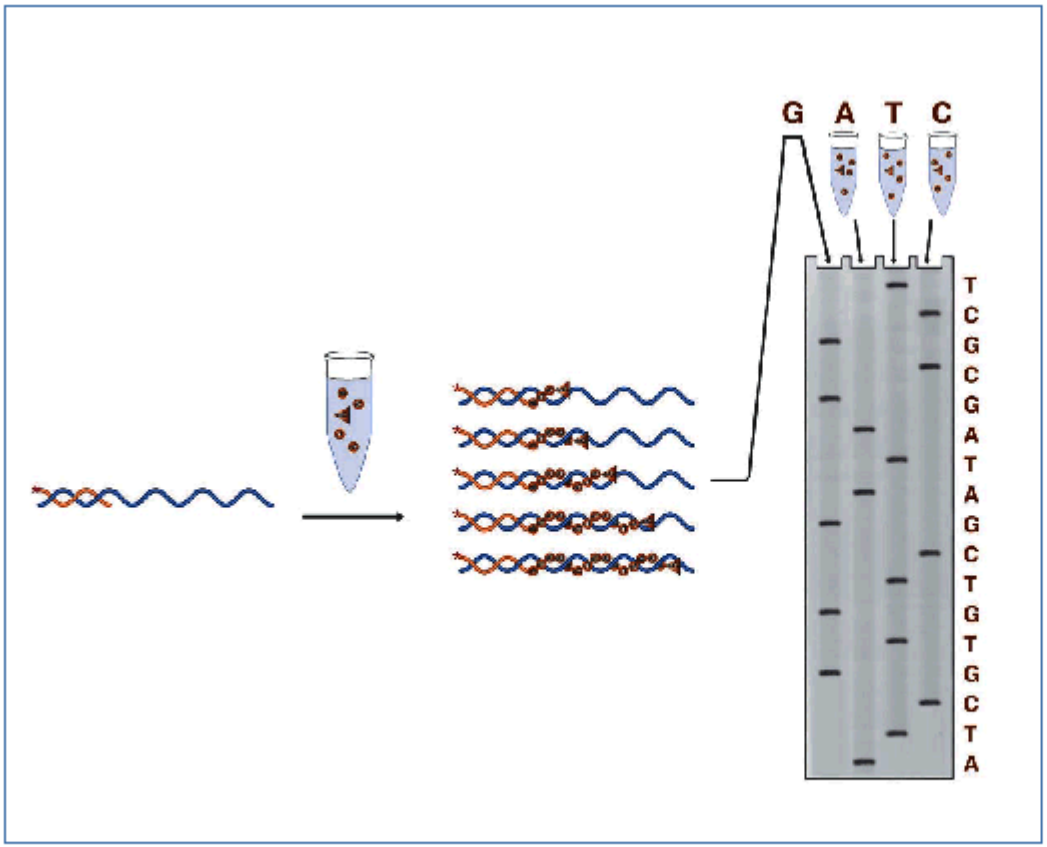
PCR alternates two phases: separate DNA into single strands using heat; convert into double strands using *primer* and polymerase reaction.

PCR rapidly amplifies a single DNA molecule into billions of molecules.



http://www.accessexcellence.org/AB/GG/polymerase.html
http://www.iupui.edu/~wellsctr/MMIA/htm/animations.htm

Reading DNA



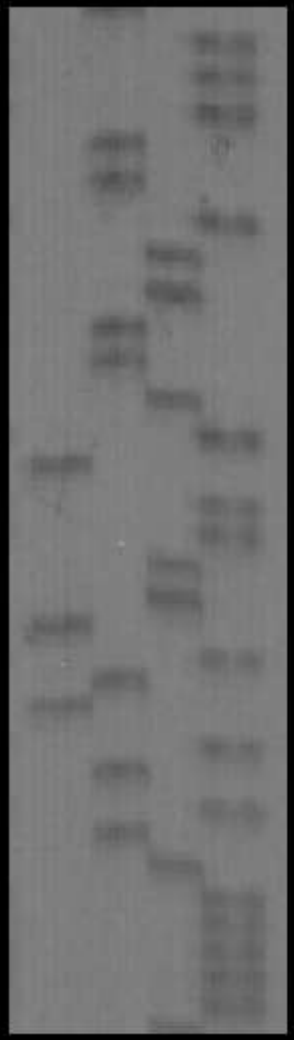
Gel electrophoresis is process of separating a mixture of molecules in a gel media by application of an electric field. In general, DNA molecules with similar lengths will migrate same distance.

First cut DNA at each base: A, C, G, T. Then run gel and read off sequence: TCGCGA ...

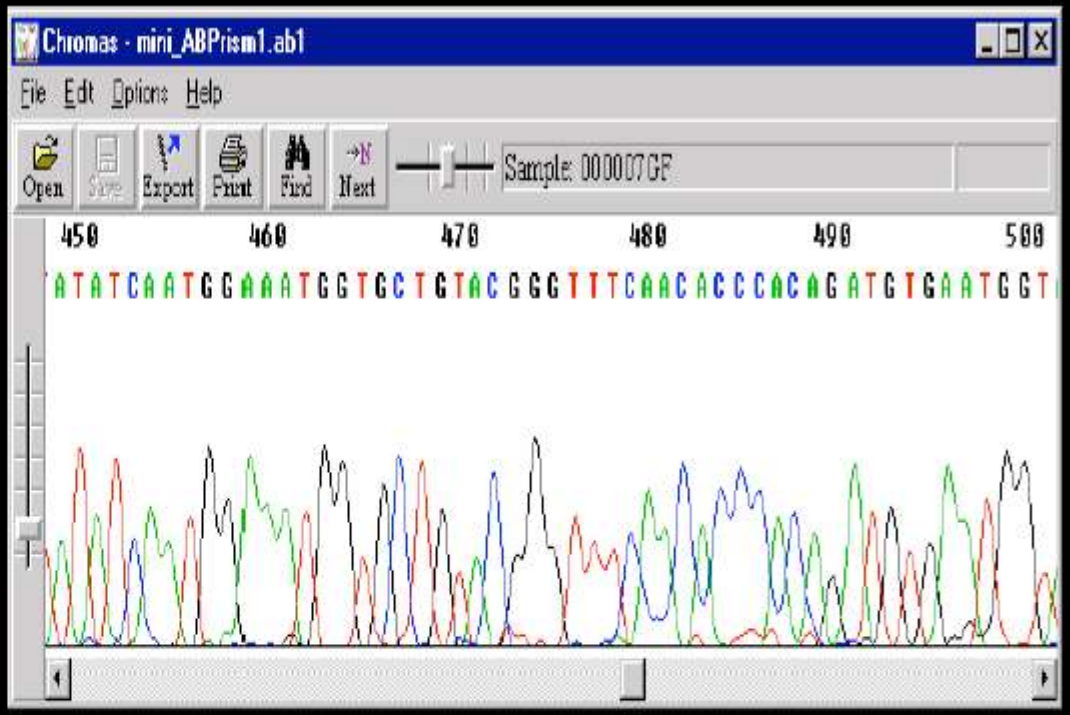
This is known as *Sanger sequencing*.

http://www.apelex.fr/anglais/applications/sommaire2/sanger.htm
http://www.iupui.edu/~wellsctr/MMIA/htm/animations.htm

DNA sequencing



Manual Automatic



Final output in human / machine readable format

A = Adenine, T = Thymine
G = Guanine, C = Cytosine

http://www.bii.a-star.edu.sg/docs/sbg/notes/n1/hu-genome%20lecture2.pdf

Discussion: Chapter 2

What is a bipartite graph?

What is the difference between an Eulerian path (or cycle) and a Hamiltonian path (or cycle)?

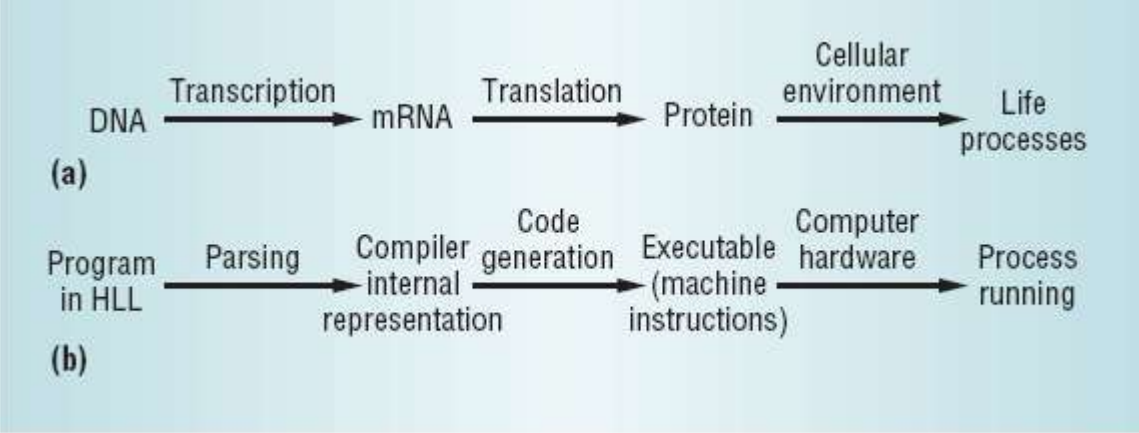
What is a minimum spanning tree?

Discussion: "The Blueprint for Life"

What viewpoint(s) are Feitelson and Treinin criticizing?

What viewpoint(s) are they advocating?

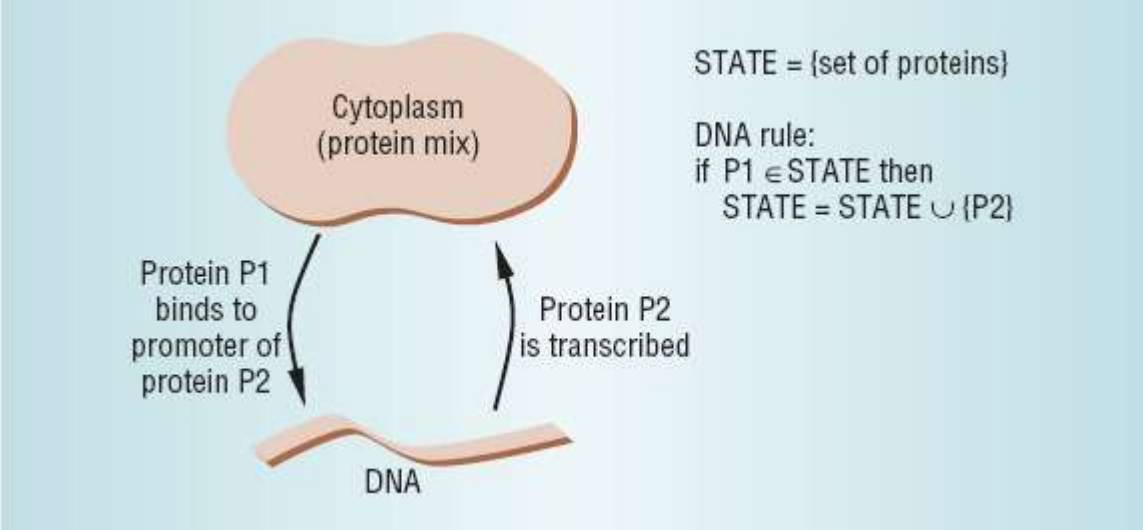
How does the following figure relate to their argument?



"The Blueprint for Life?" by D.G. Feitelson and M. Treinin, *IEEE Computer*, July 2002, pp. 34-40.

Discussion: "The Blueprint for Life"

Do you agree with their "cell as a finite state machine" concept?

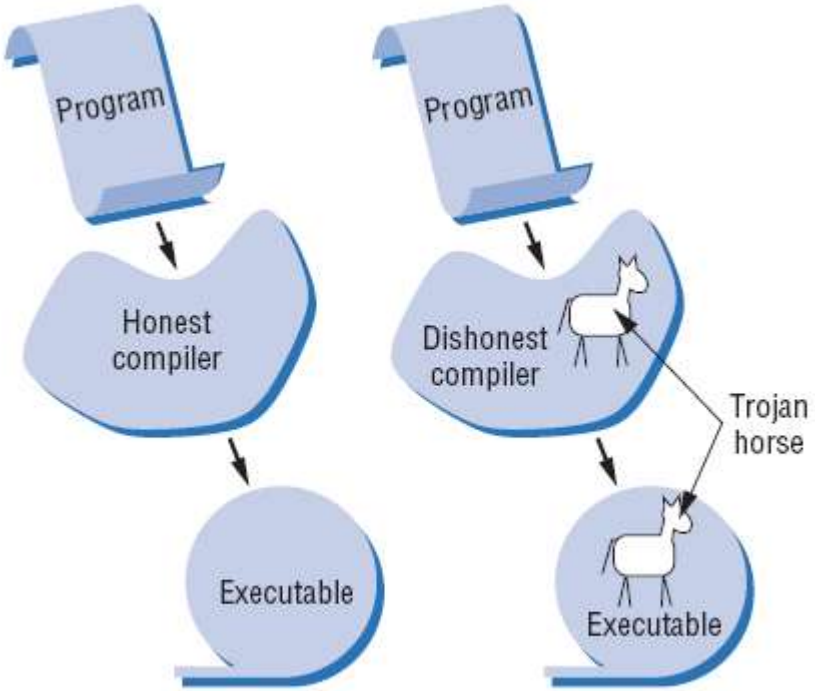


What are the implications of this viewpoint?

"The Blueprint for Life?" by D.G. Feitelson and M. Treinin, *IEEE Computer*, July 2002, pp. 34-40.

Discussion: "The Blueprint for Life"

How does concept of dishonest compiler relate to discussion?

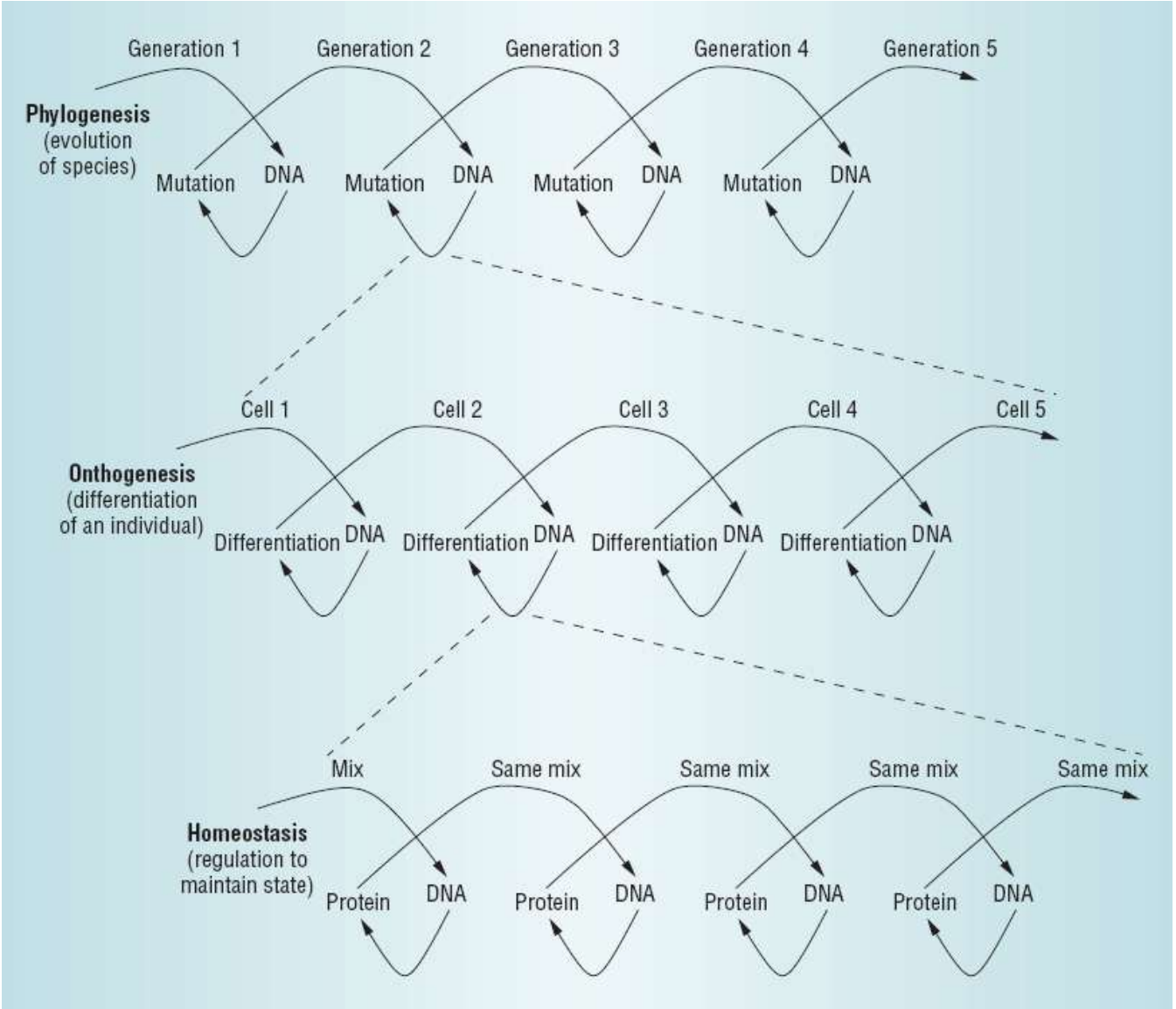


(Traits through cytoplasm vs. genes = epigenetic inheritance.)

"The Blueprint for Life?" by D.G. Feitelson and M. Treinin, *IEEE Computer*, July 2002, pp. 34-40.

Discussion: "The Blueprint for Life"

What is the interpretation for their "spiral of life"?



"The Blueprint for Life?" by D.G. Feitelson and M. Treinin, *IEEE Computer*, July 2002, pp. 34-40.

Wrap-up

Readings for next time:

- Sections 3.1-3.3 in your textbook.
- "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins" by S. B. Needleman and C. D. Wunsch.*
- "The String-to-String Correction Problem" by R. A. Wagner and M. J. Fischer.*

Remember:

- Come to class prepared to discuss what you have read.
- Check Blackboard regularly for updates.

** Original sources to be available on Blackboard soon.*