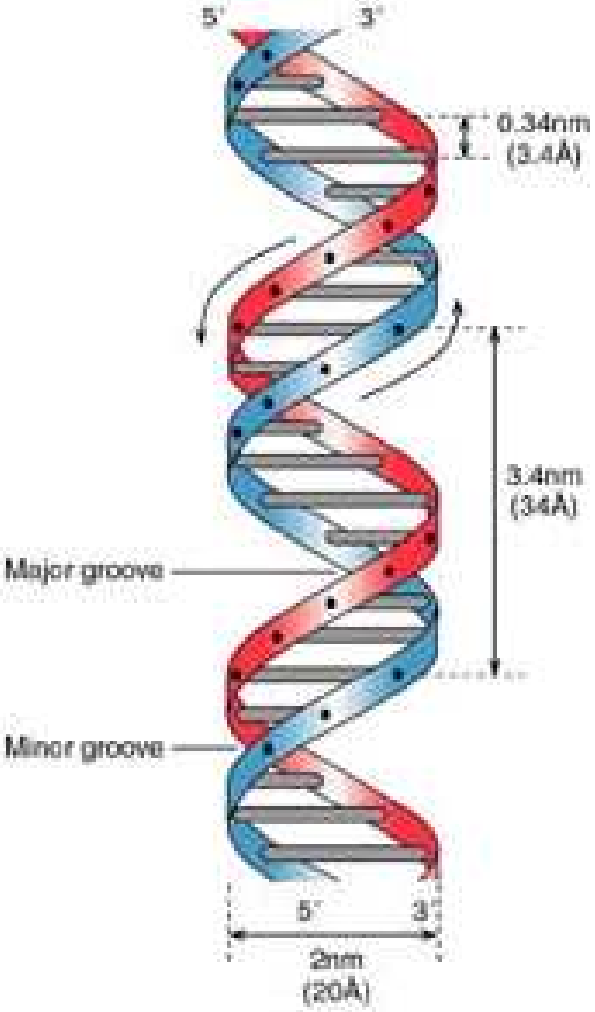
An Introductory Course on BIOINFORMATICS

Liviu Ciortuz



Plan

1 What is bioinformatics?
Why should we study it?

2 Bibliography

3 A molecular biology primer
3.1 The cell
3.2 The DNA
3.3 The Central Dogma of molecular biology
3.4 Model organisms

4 Exemplifying genetic diseases:
4.1 Thalassemia
4.2 Cystic Fibrosis

5 What you should know; Discovery question

6 Special thanks

$\boxed{1}$ What is Bioinformatics?

Bioinformatics is a pluri-disciplinary science focussing on
the applications of
computational methods and mathematical statistics
to molecular biology

Bioinformatics is also called
Computational Biology (USA)
Computational Molecular Biology
Computational Genomics

The related *...ics* family of subdomains:
Genomics, Proteomics, Phylogenetics, Pharmacogenetics,
...

Why should I teach/study bioinformatics?

Because bioinformatics is

an opportunity to use some of the most interesting computatonal techniques...

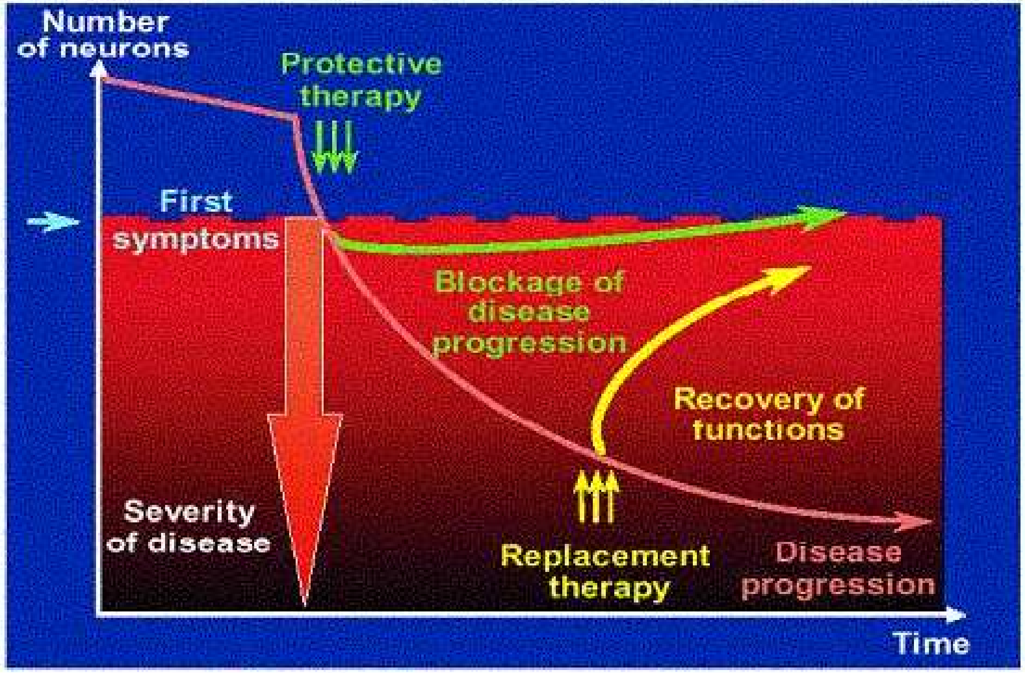
to understand some of the deep mysteries of life and diseases

and hopefully to contribute to cure some of the diseases that affect people.

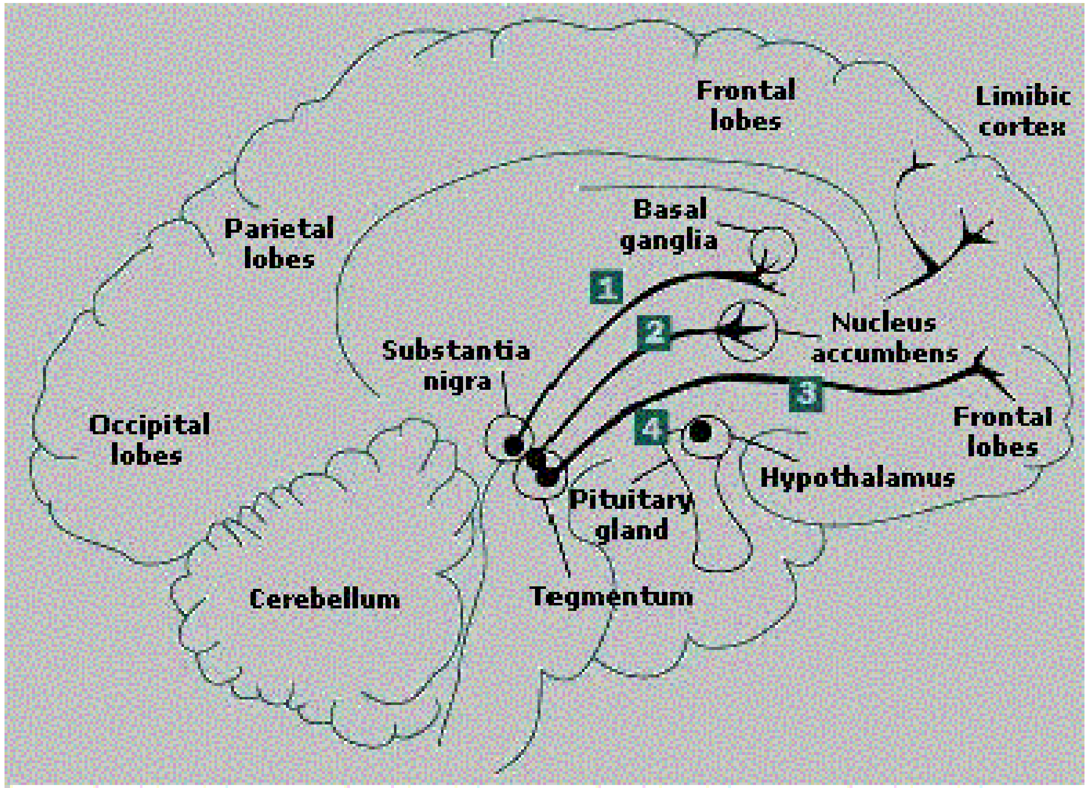
Note: The next 3 slides are from Thomas Nordahl Petersen, University of Copenhagen

Example: Parkinson’s disease

a degenerative central nervous disorder
due to the loss of brain cells which produce dopamine,
a protein important for the initiation of movement



Muhammed Ali, Pope John-Paul II died from Parkinson..., my father too



Dopamine produced by cells in Substantia nigra activates neurons in Striatum/Basal ganglia

Is there a cure for Parkinson's disease?

Parkinson disease may be cured provided that new dopamine producing cells replace the dead ones.
As a medical experiment, dopamine producing brain cells from aborted foetuses have been operated into the brain of Parkinson patients and in some cases cured the disease. Brain tissue from approx. 6 foetuses were needed. Major ethical problems!

Search for a protein drug is the only valid option.
The genes producing dopamine are still unknown. Until now, only genes involved in the dopamine transport were identified.

2 Bibliography for this course

○ Essential Cell Biology, ch. 1, and 5–7
Alberts, Bray, Hopkin, Johnson, Lewis, Raff, Roberts, Walter
Garland Science, 2010

● Biological sequence analysis:
Probabilistic models of proteins and nucleic acids
R. Durbin, S. Eddy, A. Krogh, G. Mitchison,
Cambridge University Press, 1998

● Problems and solutions in Biological sequence analysis
Mark Borodovsky, Svetlana Ekisheva
Cambridge University Press, 2006

“Biological Sequence Analysis” Contents

1. Introduction
3. Hidden Markov Models

2. Alignment of pairs of DNA/protein sequences
4. Alignment of pairs of DNA/protein seq. using HMMs

5. Multiple alignment of DNA/protein sequences
6. Multiple alignment of DNA/protein seq. using HMMs

7–8. Philogenetics; probabilistic models

9. Probabilistic CFGs
10. Alignment of RNA sequences using PCFGs

11. Background on probability

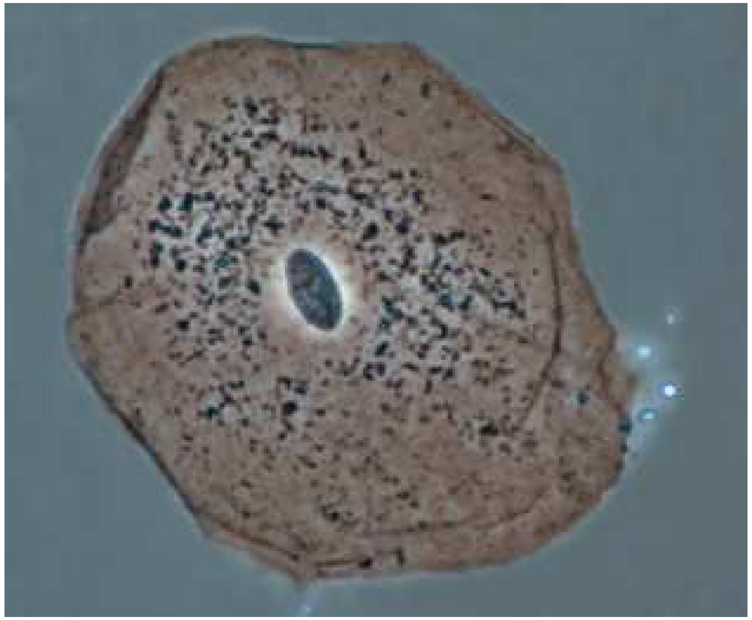
3 A Molecular Biology Primer

3.1 The Cell

The cell is the fundamental working unit of every organism.

Instead of having brains, cells make decisions trough complex networks of chemical reactions called pathways:

- synthesize new materials
- break other materials down for spare parts
- signal to eat, replicate or die



There are two different types of cells/organisms: Prokariotes and Eukariotes.

Life depends on 3 critical molecules

DNAs — made of A,C,G,T nucleotides (“bases”)

hold information on how a cell works

RNAs — made of A,C,G,U nucleotides

provide templates to synthesize amino-acids into proteins
transfer short pieces of information to different parts of the cell

Proteins — made of (20) amino acids

form enzymes that send signals to other cells and regulate gene activity
make up the cellular structure
form body’s major components (e.g. hair, skin, etc.)

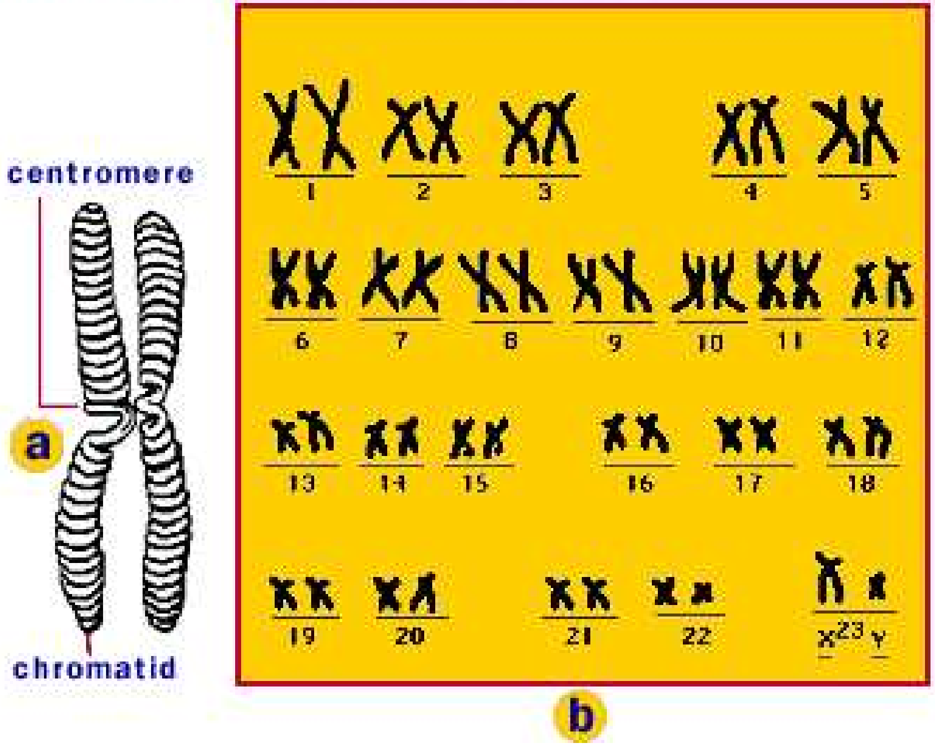
Some basic terminology

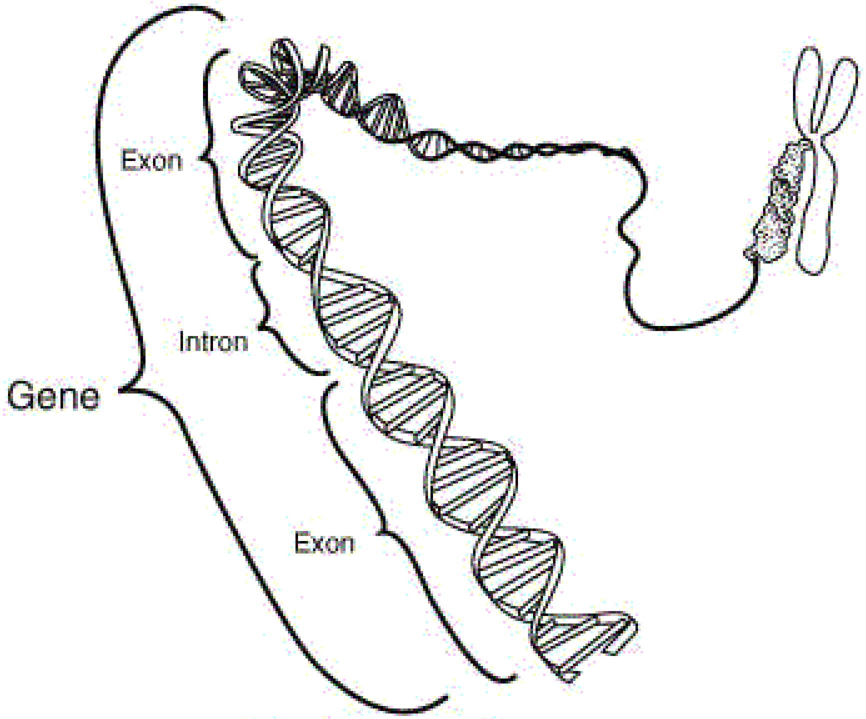
Genome: the complete set of one organism's DNA

- a bacteria contains approx. 600,000 base pairs
- human: approx. 3 billion, on 23 pairs of **chromosomes**
- each chromosome contains many genes

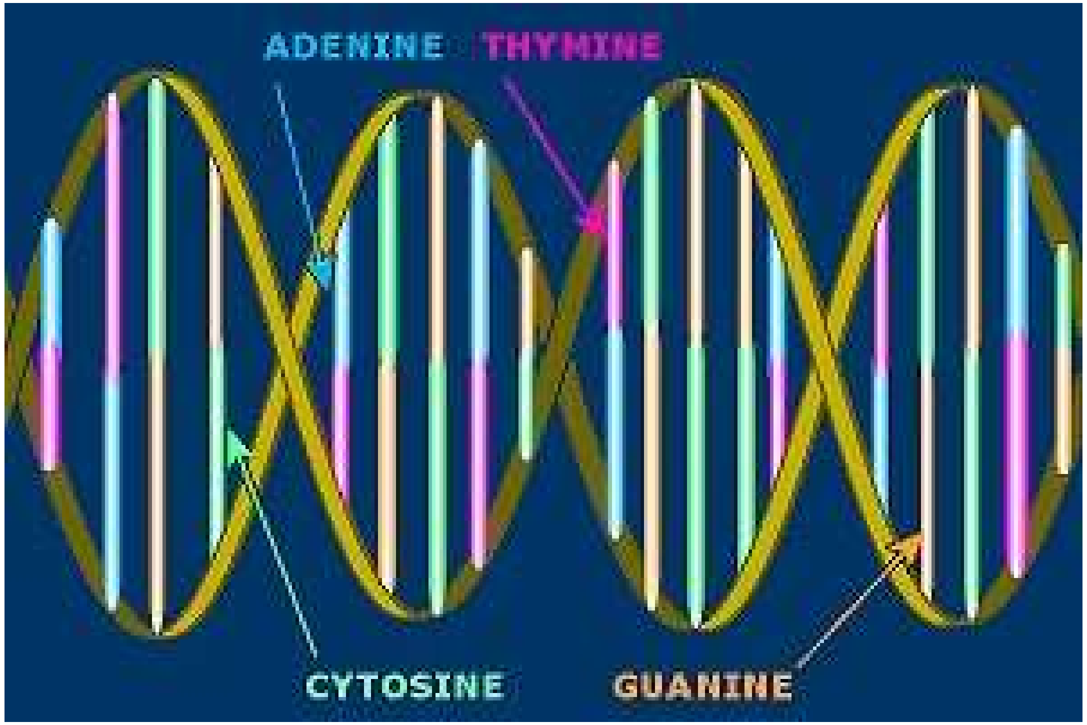
Gene: the basic functional and physical unit of heredity,
a specific sequence of bases that encode instructions on how to make proteins

Human chromosomes!





3.2 The DNA Helix



Discovered in 1953

(following hints by Erwin Chargaff and Rosalind Franklin) by

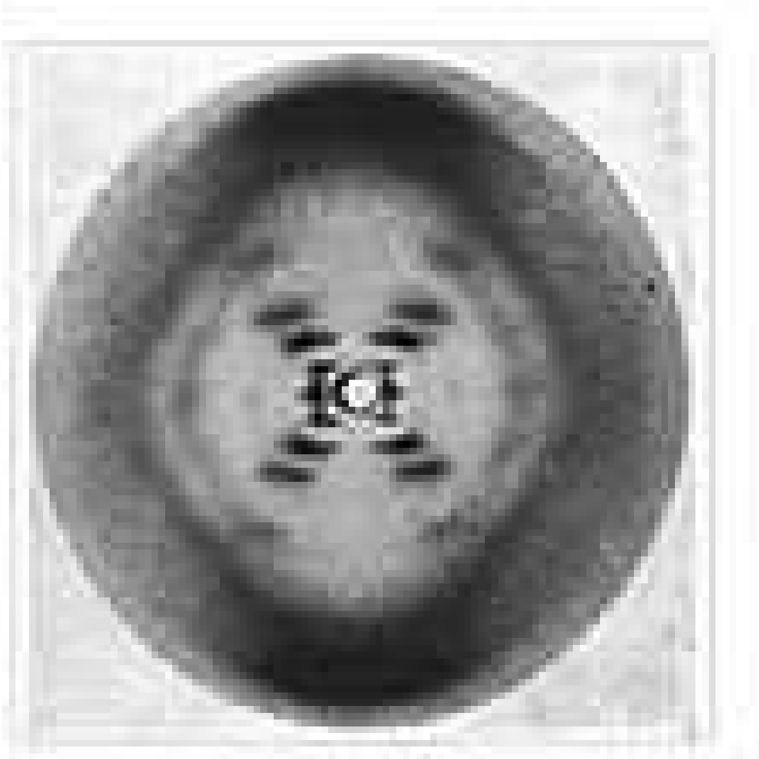
James Watson (biologist), and Francis Crick (phisicist, PhD std.)

James Watson (1928-),
and
Francis Crick (1916-2005)

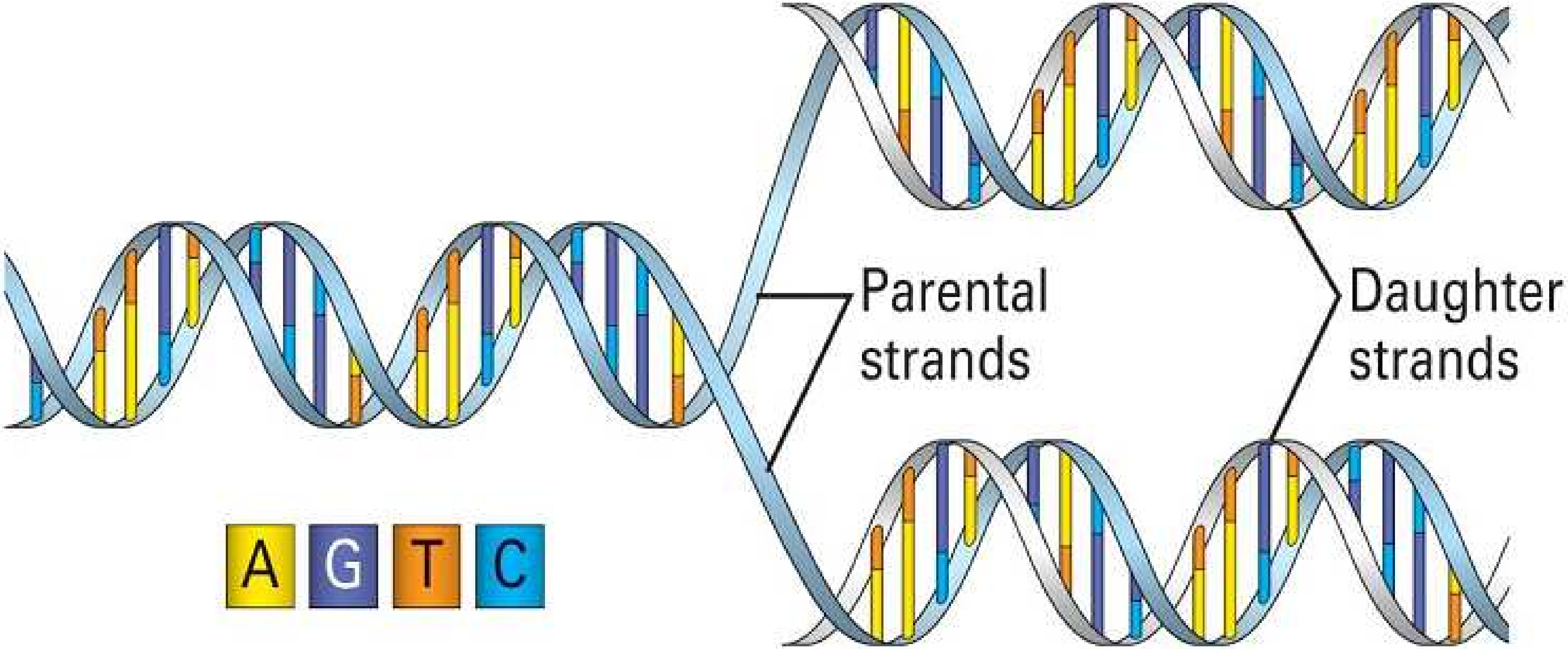
Nobel Prize 1962

Rosalind Franklin
1920-1958

The X-ray image
of a DNA molecule

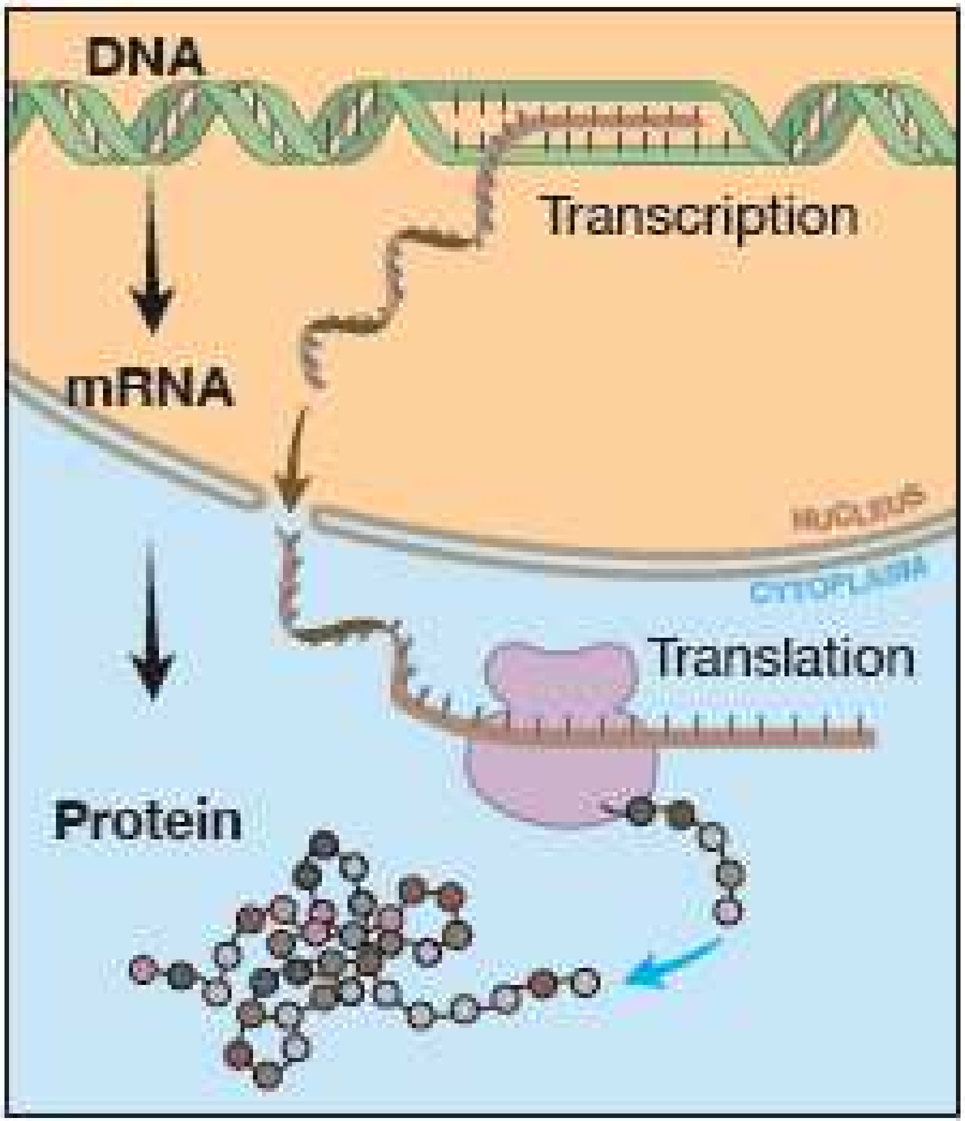


DNA copied/“replicated”



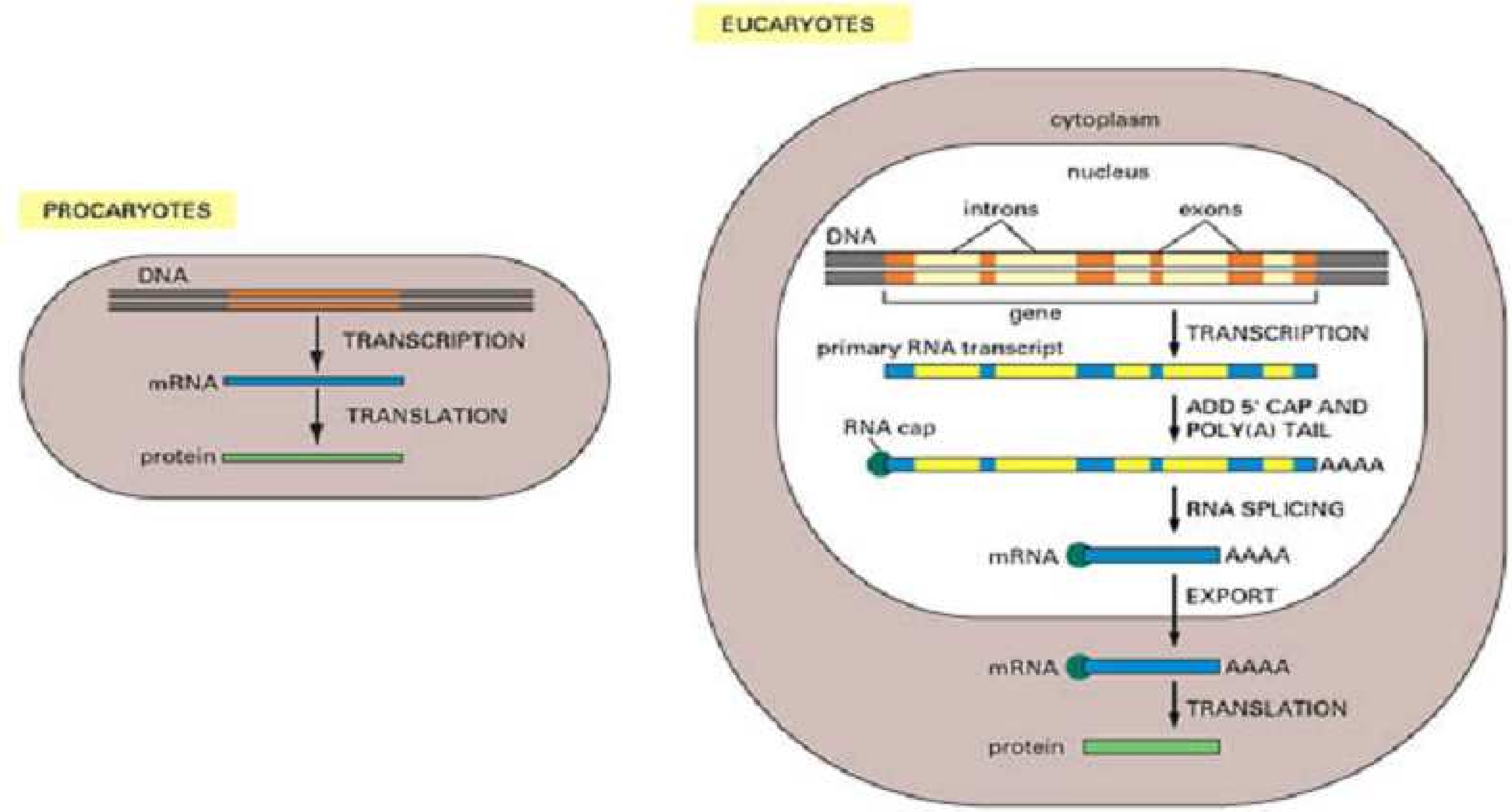
3.3 The Central Dogma of Molecular Biology

DNA → RNA → proteins



The Central Dogma of Molecular Biology

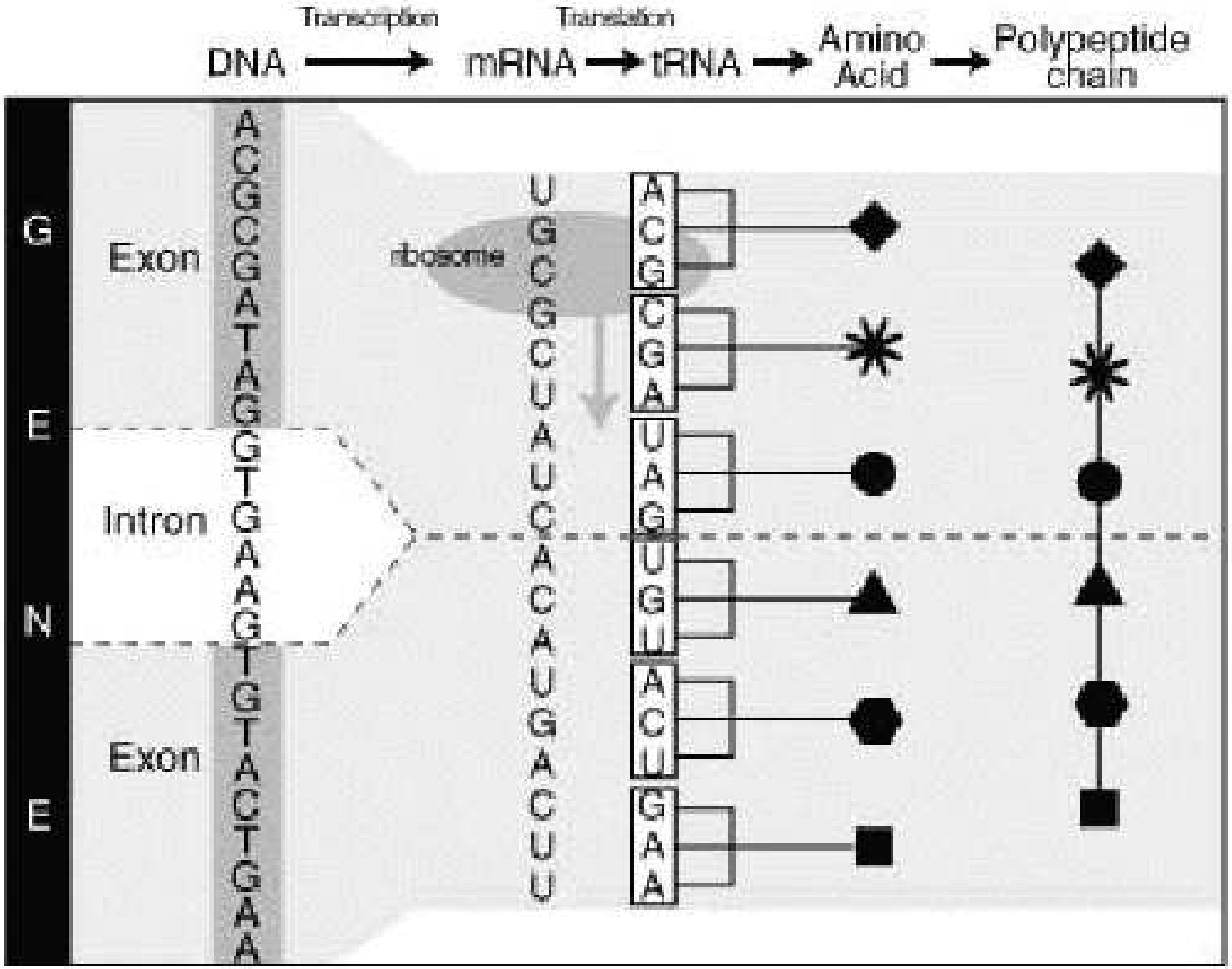
Prokariotes vs. Eukariotes



The Central Dogma of Molecular Biology

DNA → RNA → proteins

in Eukariotes



RNA to Amino Acid Coding Table

Each codon (triplet of DNA nucleotides) correponds to one of the 20 amino acids.

Among the 64 codons there are a start codon and three stop codons.

The redundancy in the table — one amino acid may be encoded by several different codons — is a kind of defence against mutations...

Second letter

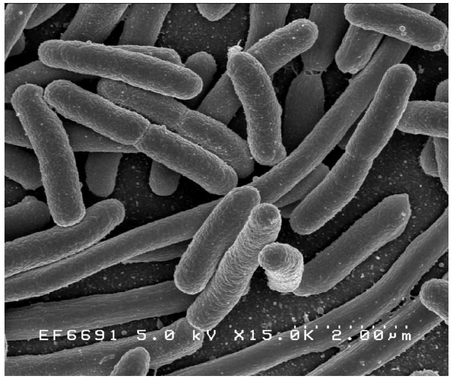
First letter	U	C	A	G
U	UUU UUC Phenil–alanine F UUA UUG Leucine L	UCU UCC UCA UCG Serine S	UAU UAC Thyrosine Y UAA STOP codon UAG STOP codon	UGU UGC Cysteine C UGA STOP codon UGG Trypto–phan W
C	CUU CUC CUA CUG Leucine L	CCU CCC CCA CCG Proline P	CAU CAC Histidine H CAA CAG Glutamine Q	CGU CGC CGA CGG Arginine R
A	AUU AUC AUA Isoleucine I AUG Methionine; START codon M	ACU ACC ACA ACG Threonine T	AAU AAC Asparagine N AAA AAG Lysine K	AGU AGC Serine S AGA AGG Arginine R
G	GUU GUC GUA GUG Valine V	GCU GCC GCA GCG Alanine A	GAU GAC Aspartic acid D GAA GAG Glutamic acid E	GGU GGC GGA GGG Glycine G

Third letter

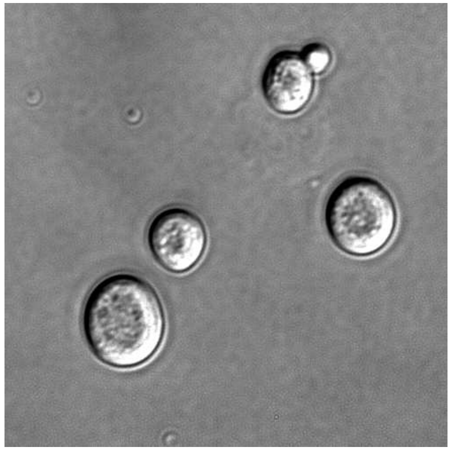
A Romanian won the Nobel Prize in molecular biology

George Emil Palade (1912–2008) showed in 1956 that
the site of protein manufacturing in the cytoplasm is made of RNA organelles called ribozomes.

3.4 Model organisms



Escherichia coli



Saccharomyces cerevisiae

Arabidopsis thaliana

Caenorhabditis elegans

Drosophila melanogaster

Mus musculusi

$\boxed{4}$ Examples of genetic diseases

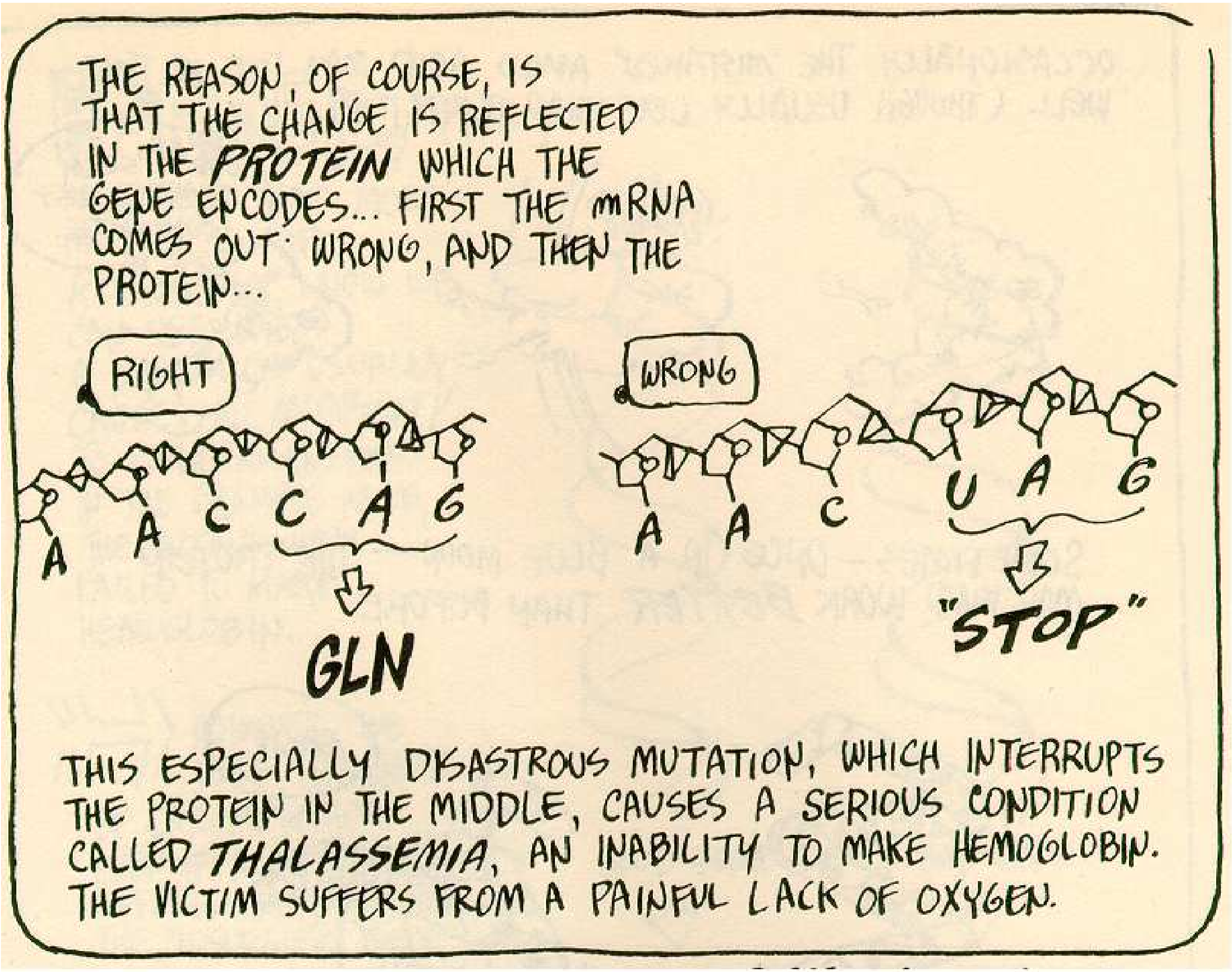
4.1 Thalassemia — a genetic disease due to faulty DNA replication

A **mutation** in a gene is a change in the DNA's sequence of nucleotides.

Sometimes even a mistake of *just one position* can have a profound effect.

Here is a small but devastating mutation in the gene for **hemoglobin**, the protein which carries oxygen in the blood.

good gene:	AACCAG
mutant gene:	AACTAG



THE REASON, OF COURSE, IS THAT THE CHANGE IS REFLECTED IN THE *PROTEIN* WHICH THE GENE ENCODES... FIRST THE mRNA COMES OUT WRONG, AND THEN THE PROTEIN...
RIGHT
A A C C A G
GLN
WRONG
A A C U A G
"STOP"
THIS ESPECIALLY DISASTROUS MUTATION, WHICH INTERRUPTS THE PROTEIN IN THE MIDDLE, CAUSES A SERIOUS CONDITION CALLED *THALASSEMIA*, AN INABILITY TO MAKE HEMOGLOBIN. THE VICTIM SUFFERS FROM A PAINFUL LACK OF OXYGEN.

from "The Cartoon Guide to Genetics", Larry Gomick, Mark Wheelis

Note

In Cyprus, a screening policy — including pre-natal screening and abortion — introduced since 1970s to reduce the incidence of thalassemia,

has reduced the number of children born with the hereditary blood desease from 1 out of every 158 births to almost 0.

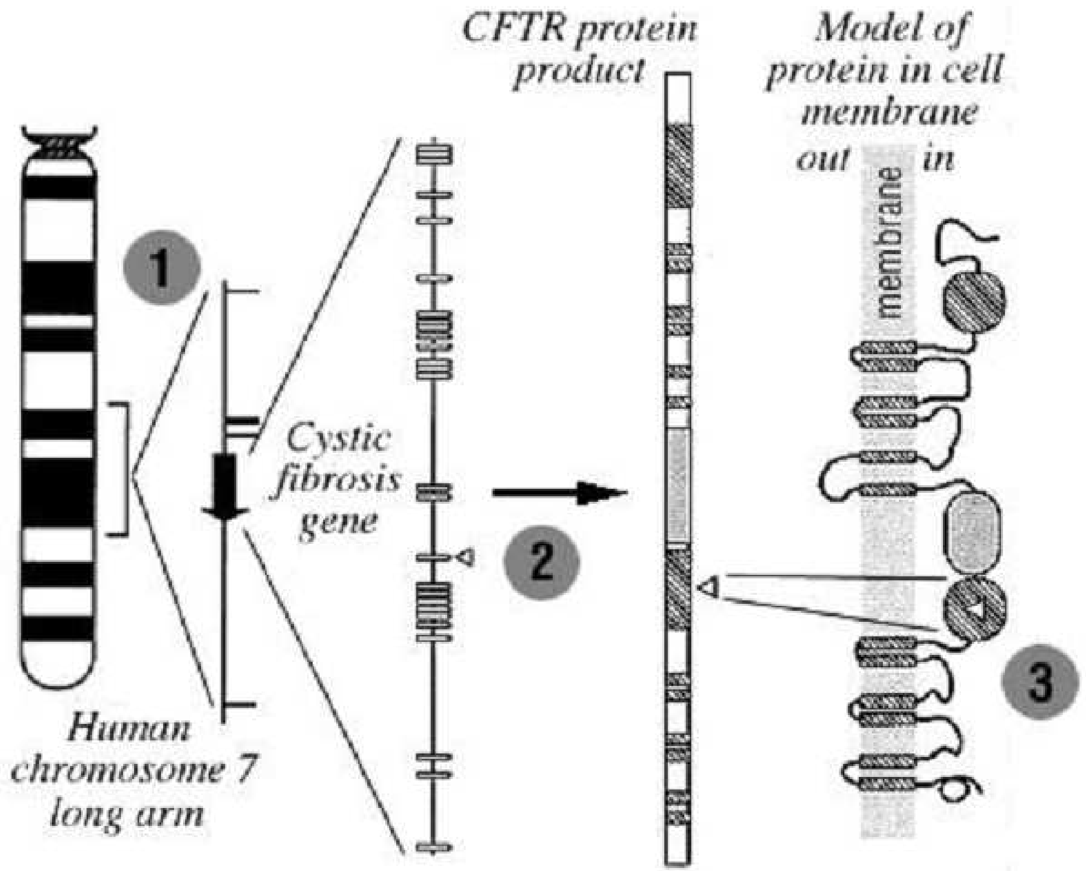
4.2 Cystic Fibrosis — a genetic disease due to deletion of a triplet in the CFTR gene

The cystic fibrosis disease is characterised by an abnormally high content of sodium in the mucus in lungs, that is life threatening for children.

The cystic fibrosis transport regulator (CFTR) gene adjusts the “waterness” of fluids secreted by the cell.

Due to the deletion of a single triplet in the CFTR gene, the mucus ends up being too thick.

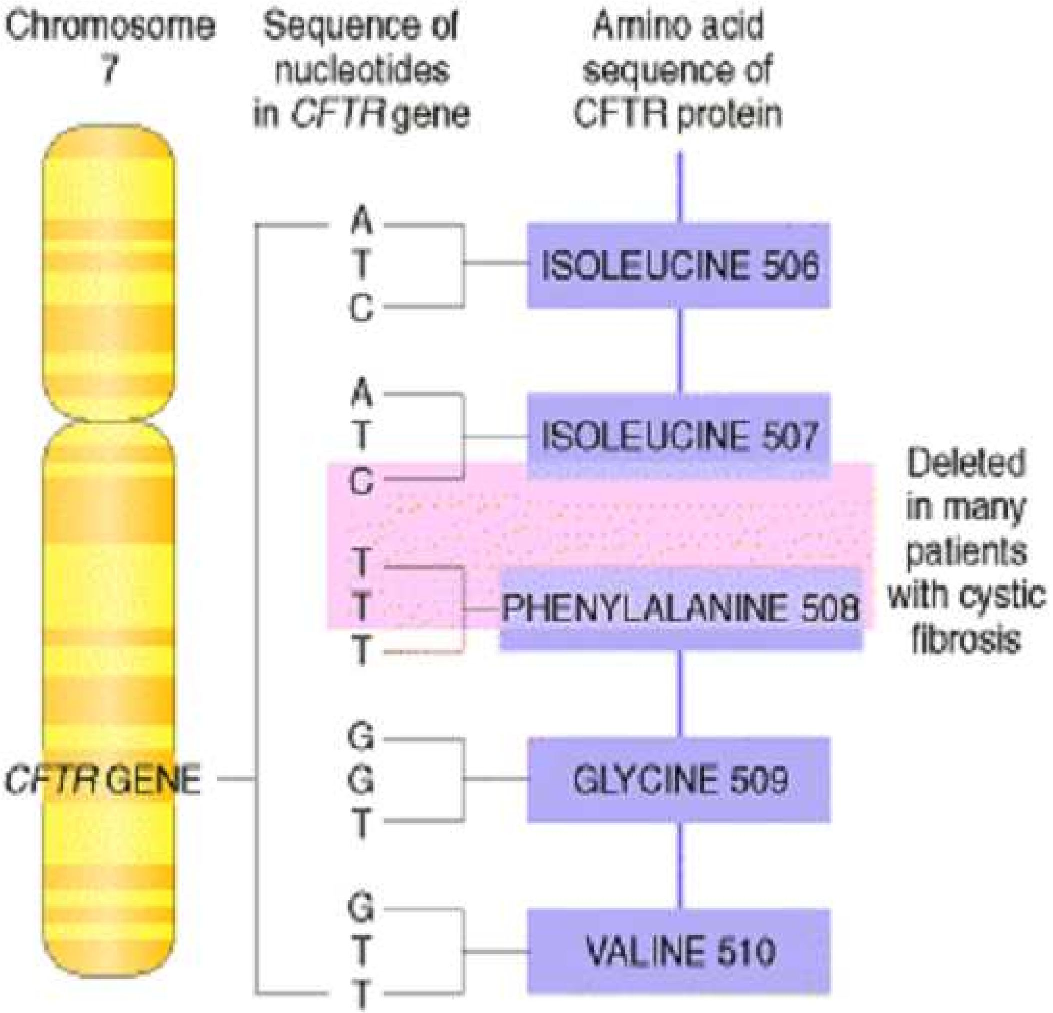
Cystic Fibrosis Transport Regulator (CFTR)


CFTR protein product
Model of protein in cell membrane
out
in
membrane
1
Cystic fibrosis gene
2
3
Human chromosome 7 long arm

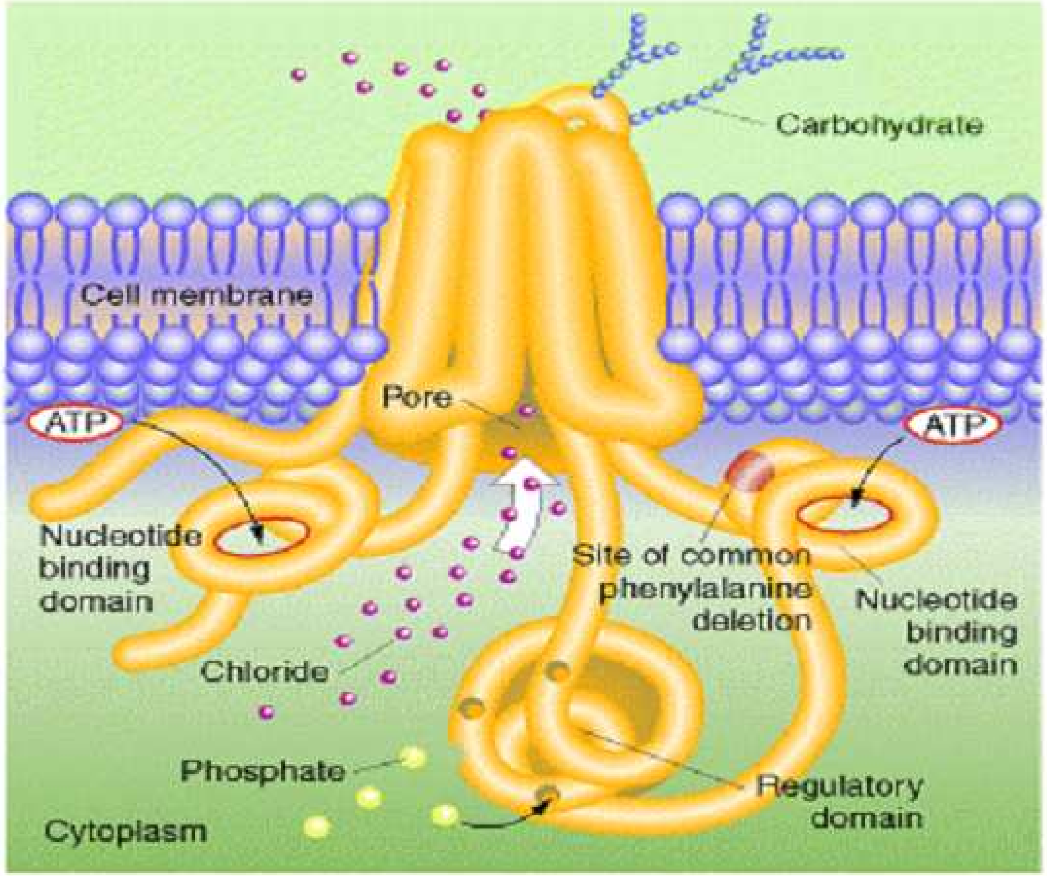
Francis Collins

Acknowledgement: this and the next two slides are from Jones & Pevzner

A fatal mutation in the Cystic Fibrosis Transport Regulator (CFTR) gene


Chromosome 7
Sequence of nucleotides in *CFTR* gene
Amino acid sequence of CFTR protein
CFTR GENE
A
T
C
ISOLEUCINE 506
A
T
C
ISOLEUCINE 507
T
T
T
PHENYLALANINE 508
G
G
T
GLYCINE 509
G
T
T
VALINE 510
Deleted in many patients with cystic fibrosis

The Cystic Fibrosis Transport Regulator (CFTR) Protein


Carbohydrate
Cell membrane
Pore
ATP
ATP
Nucleotide binding domain
Site of common phenylalanine deletion
Nucleotide binding domain
Chloride
Phosphate
Regulatory domain
Cytoplasm

5 What you should know

- What is the “Central Dogma” of molecular biology?
- What is the difference between transcription and translation of the DNA message?
- What is a codon?
- Why it is necessary to have a three-letter code?
- How would you define a gene?
- Why can there be more than one possible mRNA sequence for a DNA sequence?
- What is the difference between an intron and an exon?
- What is DNA sequencing?
- What are the positive results of DNA mutations?

Discovery Question:

How do we read DNA sequences?

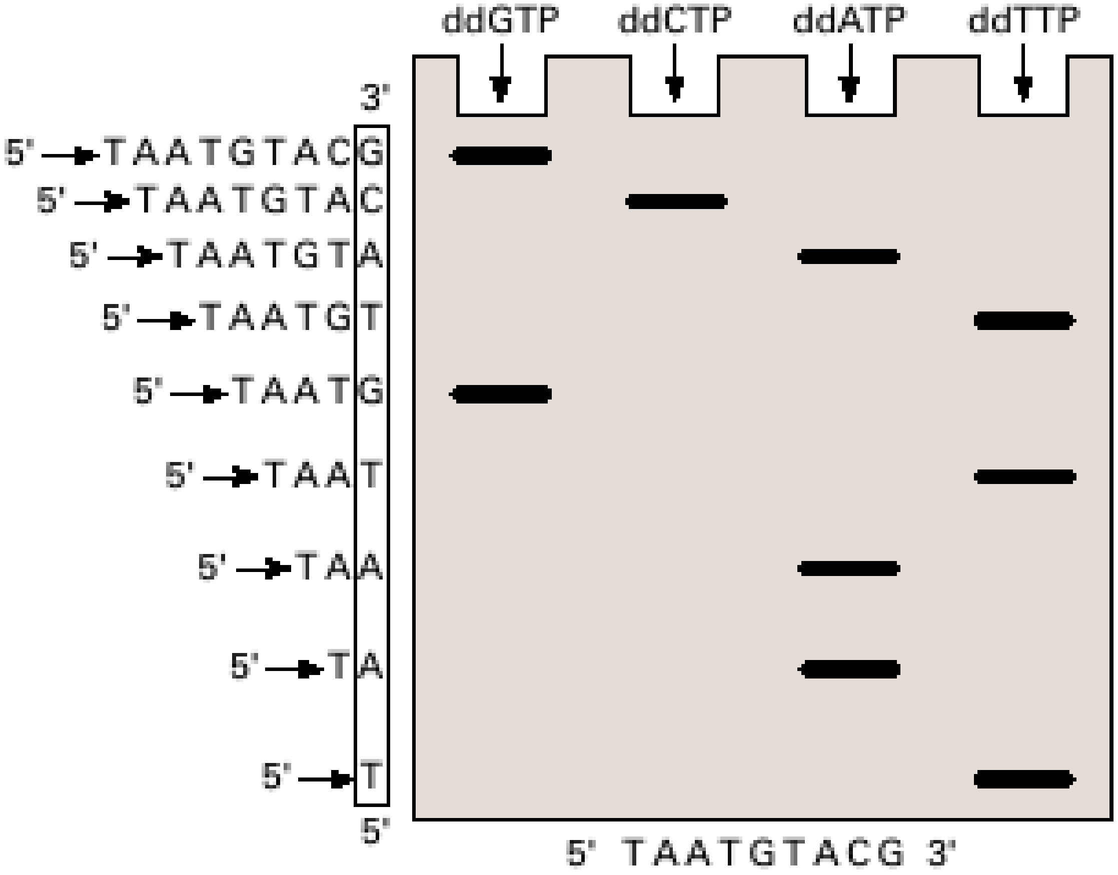
Knowing how DNA replication works,
and assuming that you can get the molecular mass of any given DNA fragment,

design a strategy to get the "reading" of the base composition of an unknown DNA sequence (i.e. the output should be a string over the alphabet $\{A, C, G, T\}$).

What if, due to physical limitations, only fragments of relatively short length (500-700 bases) can be treated in the above way, but the genome that you want to "read" is much larger (10^6 or more)?

Short answer:
Fred Sanger's Method, Nobel Prize, 1980

In 1977 Sanger sequenced the DNA of the FX 174 Phage virus (5386 nucleotides).



From *Discovering Genomics, Proteomics, and Bioinformatics*, Campbell and Hayer, 2006

Scaling up Sanger's method to whole genome sequencing

Problems:

- limited size of the *reads:* 500–700 nucleotides
- genomes are much larger (human: 3 $\times 10^9$), and contain lots of *repeats* (human: more than 50%)
- sequencing errors: 1-3%

Solutions:

- use overlaping reads, then assemble them
- BAC-by-BAC sequencing
- using tandem reads to cope with repeats

Recommened reading:

Bioinformatic Algorithms, Jones & Pevzner, Ch. 8.

6 Special Thanks

This bioinformatics course would not have been possible without the help of

- the BSc students who took my AI labs on bioinformatics, during the spring 2004 semester:
  Ioana Brudaru, Cristian Prisecariu, Lăcrămioara Aştefănoaiei, ...
- the MSc students, the fall 2005 semester:
  Marta Gîrdea, Oana Răţoi, ...
- MSc students, the fall 2006 semester:
  Sergiu Dumitriu, Diana Popovici, ...
- the BSc students, who took my Bioinformatics course during the spring 2007 semester:
  Ioana Boureanu, Anca Luca, Ştefana Munteanu, Irina Ghiorghiţă, Cristian Rotaru, ...
- a former student and colleague of mine who provided me copies of some very good bioinformatics books: Dr. Liliana Ibănescu.

Former students of ours who did or are currently doing PhD's in bioinformatics

- Raluca Gordân, 2005, Duke University, USA
- Raluca Uricaru, 2005, Université de Monpellier, France
- Marta Gîrdea, 2005, Université de Lille, France
- Luminiţa Moruz, 2005, University of Stockholm, Sweden
- Irina Mohorianu, 2008, University of East Anglia, UK
- Alina Sîrbu, 2008, University of Dublin, UK
- Irina Roznovăţ, 2008, University of Dublin, UK
- Florin Chelaru, 2008, University of Maryland, USA

○ [Călin-Rareş Turliuc, 2010, Imperial College of London, UK]

- Alina Munteanu, 2011, University of Iaşi, Romania
- Bogdan Luca, 2012, University of East Anglia, UK
- Claudia Păuleţ (Paicu), 2013, University of East Anglia, UK

Published Papers

- D. Pasailă, I. Mohorianu, A. Sucilă, Şt. Panţiru, L. Ciortuz, *MicroRNA recognition with the yasMiR system: The quest for further improvements.* In “Software Tools and Algorithms for Biological Systems”, volume in the “Advances in Experimental Medicine and Biology” series, Springer Verlag, New York, USA, 2011.

- D. Pasailă, I. Mohorianu, A. Sucilă, Şt. Panţiru, L. Ciortuz, *Yet another SVM for microRNA recognition: yasMiR.* Technical Report (TR-10-01), Faculty of Computer Science, University of Iasi, Romania, 2010, 13 pages.

- D. Pasailă, I. Mohorianu, L. Ciortuz, *Using base pairing probabilities for MiRNA recognition.* In Proceedings of SYNASC 2008, The 9th international symposium on Symbolic and Numeric Algorithms for Scientific Computing, Timişoara, Romania, IEEE Computer Society CPS, 2008, pages 519–525.

- L. Ciortuz, *Support vector machines for microRNAs classification.* In Proceedings of EHB’07, The Workshop on E-Health and Bio-Engineering, Revista Medico-chirurgicală a Universităţii de Medicină “Gr. T. Popa”, Iaşi, Romania, 2007, pages 60–63.

Published Papers (cont'd)

- **A.-L. Ioniţă, L. Ciortuz, *Pre-miRNA features for automated classification.* In Proceedings of The 4th International Workshop on Soft Computing Applications (SOFA), Arad, Romania, 2010. ISBN: 978-1-4244-7985-6, IEEE Catalog Number: CFP1028D-CDR, pages 125–130.**

- **C.-R. Turliuc, L. Ciortuz, *Gaussian Processes for Classification on Cancer and MicroRNA Datasets. Comparison with Support Vector Machines.* In Proceedings of The 7th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB), Palermo, Italy, 2010.**

- **M. Gîrdea, L. Ciortuz, *A hybrid genetic programming and boosting technique for learning kernel functions from training data.* In Proceedings of SYNASC 2007, The 9th international symposium on Symbolic and Numeric Algorithms for Scientific Computing, Timişoara, Romania, IEEE Computer Society CPS, 2007, pages 395–402.**

- **R. Uricaru, L. Ciortuz, *Genic interaction extraction from* MEDLINE *abstracts — A case study.* In Scientific Annals of the "Al.I. Cuza" University of Iasi, Romania, Computer Science Series, 2005, pages 137–152.**

Additional Bibliography (I)

- Algorithms on Strings, Trees, and Sequences
  Computer Science and Computational Biology
  Dan Gusfield
  Cambridge University Press, 1997
- Computational Molecular Biology: An Algorithmic Approach
  Pavel Pevzner
  MIT Press, 2000
- Statistical Methods in Bioinformatics: An Introduction
  Warren Ewens, Gregory Grant
  Springer, 2001
- ○ Introduction to Computational Genomics: A Case Studies Approach
  Nello Cristianini, Matthew Hahn
  Cambridge University Press, 2006
- An Introduction to Bioinformatics Algorithms
  Neil Jones, Pavel Pevzner
  MIT Press, 2004

Additional Bibliography (II), more “Bio...”

- Essential Cell Biology, (2nd ed.)
  B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. Watson
  Garlands, 2005
- Discovering Genomics, Proteomics, and Bioinformatics, (2nd ed.)
  Malcolm Campbell, Laurie Hayer
  Benjamin Cummings, 2006
- Introduction to Bioinformatics
  Arthur Lesk
  Oxfrod University Press, 2002
- Bioinformatics
  David Mount
  Cold Spring Harbor Laboratory Press, 2001
- Fundamental Concepts of Bioinformatics
  Dan Krane, Michael Raymer
  Benjamin Cummings, 2003

Additional Bibliography (III), more "...informatics"

- Machine Learning Approaches to Bioinformatics
  Zheng Rong Yang
  MIT Press, 2010
- Bioinformatics: The Machine Learning Approach
  Pierre Baldi, Søren Brunak
  MIT Press, 2001
- Flexible Pattern Matching in Strings:
  Practical on-line search algorithms for texts and biological sequences
  Gonzalo Navarro, Mathieu Raffinot
  Cambridge University Press, 2002
- Jewels of Stringology
  M. Crochemore and W. Rytter
  World Scientific Press, 2002
- Parallel Computing for Bioinformatics and Computational Biology
  Alber Zomaya (ed.); Wiley, 2006

Recommended bibliography for laboratory

- Bioinformatics and Computational Biology Solutions using R and Bioconductor
  Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit
  Springer, 2005

- Beginning Perl for Bioinformatics
  James Tisdall
  O'Reilly, 2001

- Mastering Perl for Bioinformatics
  James Tisdall
  O'Reilly, 2003