

Probabilistische modellen in de bio-informatica

Yves Moreau
3de jr. Burg. Ir. Elektrotechniek
Dataverwerking & Automatisatie
2001-2002

Overzicht

- Wat is bio-informatica?
- Waarom bio-informatica?
- Planning van de cursus

- Elementen van de moleculaire biologie

Practische regelingen

Uurrooster

- Les
  - `Week 39-40`
  - `Donderdag 8:25-10:25`
  - `Lokaal 00.57`
- Geen les op 4 october en op 1 november
- Oefenzittingen
  - `Vr 23 nov 2001 14.00 91.33`
  - `Di 27 nov 2001 10.30 91.33`
  - `Di 04 dec 2001 10.30 91.33`
  - `Vr 14 dec 2001 14.00 91.33`
  - `Ma 17 dec 2001 10.30 91.33`

Kernreferenties

- Hoofdreferentie: R. Durbin, A. Krogh, S. Eddy, G. Mitchinson, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Oxford University Press, 199x.

- B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*, Garland Publishing, 1998 (+ Interactive CD).
- P. Baldi, S. Brunak, *Bioinformatics: The Machine Learning Approach*, MIT Press, 2001.
- A. Baxevanis, B. Ouellette, *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, Wiley-Interscience, 2001.

- Twee kopieën zullen beschikbaar zijn in de bibliotheek TELE

Cursusoverzicht

- Inleiding tot de moleculaire biologie
- Aligneren van biologische sequenties
  - Globale aligneringsmethoden
  - Locale aligneringsmethoden
- Motieven vinden in sequenties I
  - Meerdere aligneringen
  - Consensussequentie
  - Positie-gewichtmatrices
- Motieven vinden in sequenties II
  - Verborgen Markovketens
  - Leeralgoritmes

Cursusoverzicht

- Genvoorspelling
  - Prokaryoten
  - Eukaryoten
- Analyse van genuitdrukking
  - Microroostertechnologieen
  - Clusteranalyse
- Ontdekking van regulatiemotieven
  - Frequentiemethoden
  - Expectation-Maximization
  - Gibbs sampling

Overzicht van de oefenzittingen

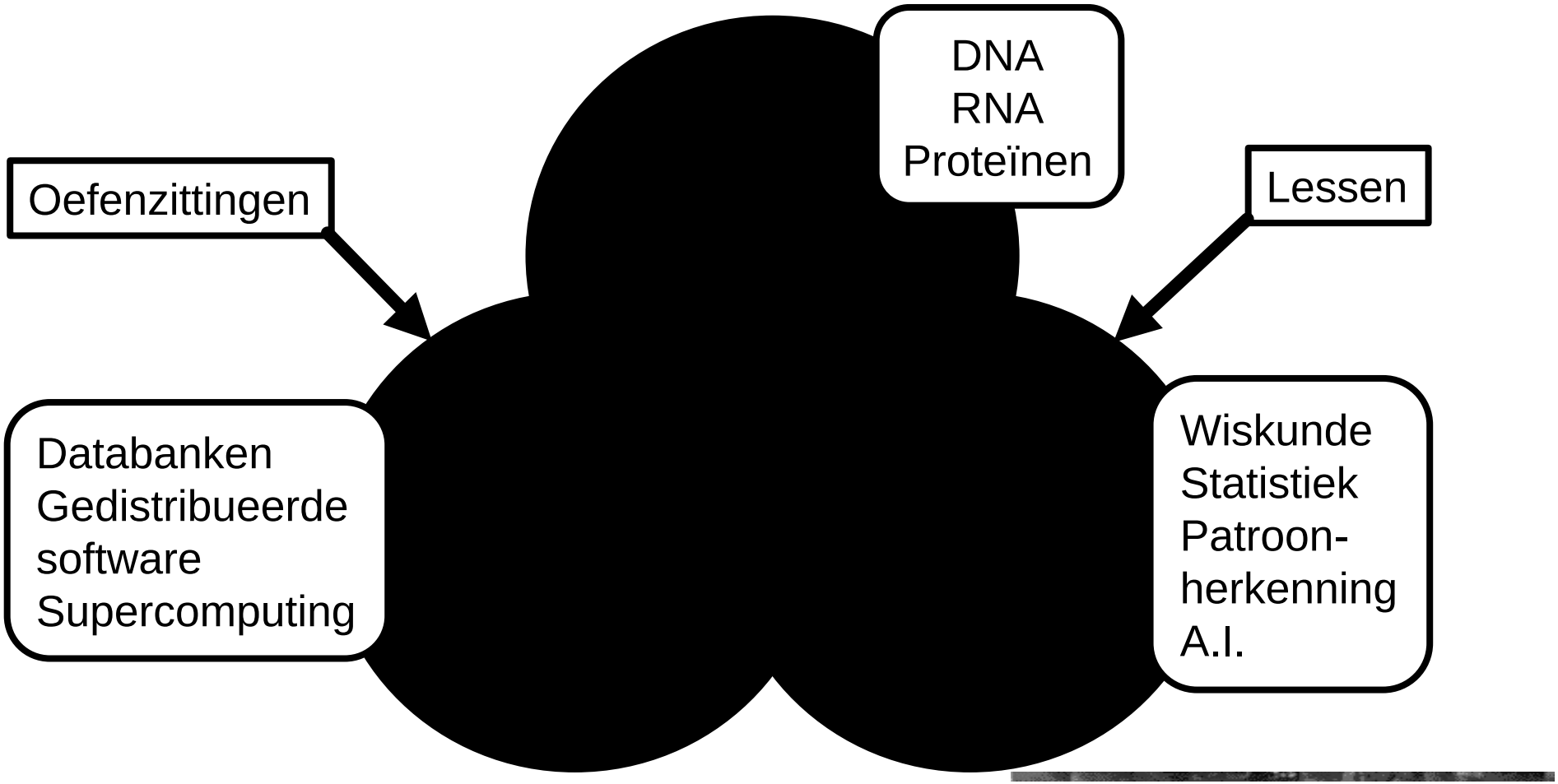
- Oefenzitting 1
  - Web resources, BLAST, CLUSTALW
- Oefenzitting 2
  - Ontwikkeling van een verborgen Markovketen
- Oefenzitting 3
  - Genvoorspelling
- Oefenzitting 4
  - Microroosteranalyse
- Oefenzitting 5
  - Ontdekking van motieven

Bio-informatica

Wat is bio-informatica?

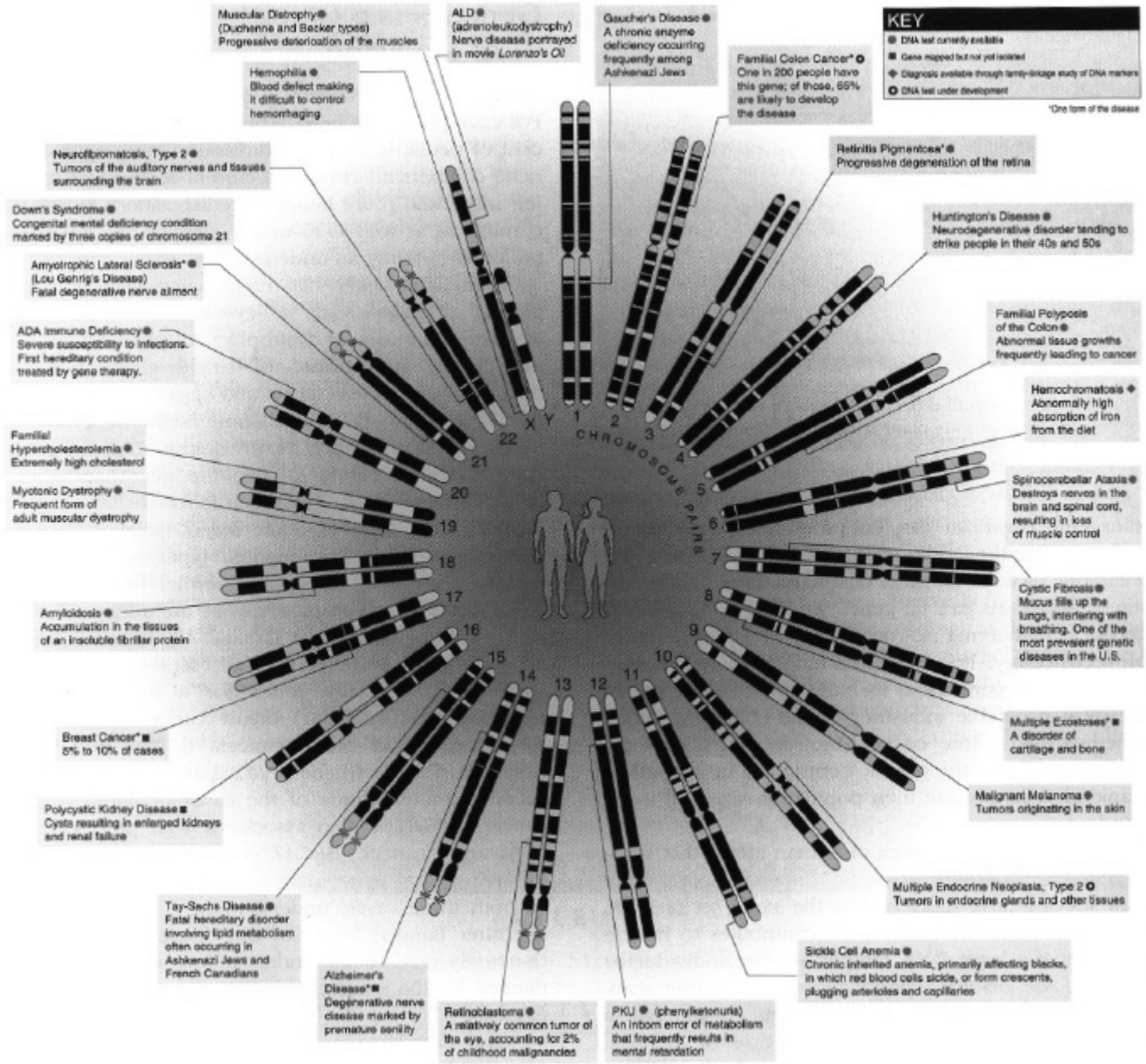
- Computers in de biologie en de geneeskunde
  - Medische informatica
  - Computationele biologie
  - Chemo-informatica
  - Bio-informatica

Wat is bio-informatica?



Waarom bio-informatica?

- Human Genome Project
  - Menselijk genoom: 3 miljard basenparen
  - Celera: ‘shotgun’ – grootste burgerlijke computercluster
  - Post-genomica
- Pharma & biotech
  - Genomica levert massale hoeveelheden data
  - Pharma:
    - Kostprijs per drug: ~ $500.000.000
    - Return per drug daalt
    - Pijplijn moet gestroomlijnd worden
  - Biotech: onderzoekspijplijn voor
    - Transgenische gewassen (Gentse specialiteit)
    - ‘Nutraceuticals’ (e.g., Procter & Gamble, kanker-beschermende broccoli)
    - Productie van medicamenten en moleculen met hoge toegevoegde waarde (e.g., Pharming)

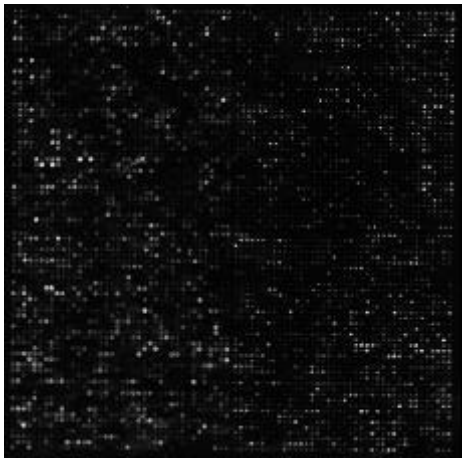

KEY
DNA test currently available
Gene mapped but not yet isolated
Diagnosis available through family-linkage study of DNA markers
DNA test under development
*One form of the disease
Muscular Dystrophy
(Duchenne and Becker types)
Progressive deterioration of the muscles
ALD
(adrenoleukodystrophy)
Nerve disease portrayed in movie Lorenzo's Oil
Gaucher's Disease
A chronic enzyme deficiency occurring frequently among Ashkenazi Jews
Familial Colon Cancer*
One in 200 people have this gene; of those, 85% are likely to develop the disease
Hemophilia
Blood defect making it difficult to control hemorrhaging
Retinitis Pigmentosa*
Progressive degeneration of the retina
Neurofibromatosis, Type 2
Tumors of the auditory nerves and tissues surrounding the brain
Huntington's Disease
Neurodegenerative disorder tending to strike people in their 40s and 50s
Down's Syndrome
Congenital mental deficiency condition marked by three copies of chromosome 21
Familial Polyposis of the Colon
Abnormal tissue growths frequently leading to cancer
Amyotrophic Lateral Sclerosis*
(Lou Gehrig's Disease)
Fatal degenerative nerve ailment
Hemochromatosis
Abnormally high absorption of iron from the diet
ADA Immune Deficiency
Severe susceptibility to infections. First hereditary condition treated by gene therapy.
Spinocerebellar Ataxia
Destroys nerves in the brain and spinal cord, resulting in loss of muscle control
Familial Hypercholesterolemia
Extremely high cholesterol
Myotonic Dystrophy
Frequent form of adult muscular dystrophy
Cystic Fibrosis
Mucus fills up the lungs, interfering with breathing. One of the most prevalent genetic diseases in the U.S.
Amyloidosis
Accumulation in the tissues of an insoluble fibrillar protein
Multiple Exostoses*
A disorder of cartilage and bone
Breast Cancer*
5% to 10% of cases
Malignant Melanoma
Tumors originating in the skin
Polycystic Kidney Disease
Cysts resulting in enlarged kidneys and renal failure
Multiple Endocrine Neoplasia, Type 2
Tumors in endocrine glands and other tissues
Tay-Sachs Disease
Fatal hereditary disorder involving lipid metabolism often occuring in Ashkenazi Jews and French Canadians
Sickle Cell Anemia
Chronic inherited anemia, primarily affecting blacks, in which red blood cells sickle, or form crescents, plugging arterioles and capillaries
Alzheimer's Disease*
Degenerative nerve disease marked by premature senility
Retinoblastoma
A relatively common tumor of the eye, accounting for 2% of childhood malignancies
PKU (phenylketonuria)
An inborn error of metabolism that frequently results in mental retardation
CHROMOSOME PAIRS
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X Y

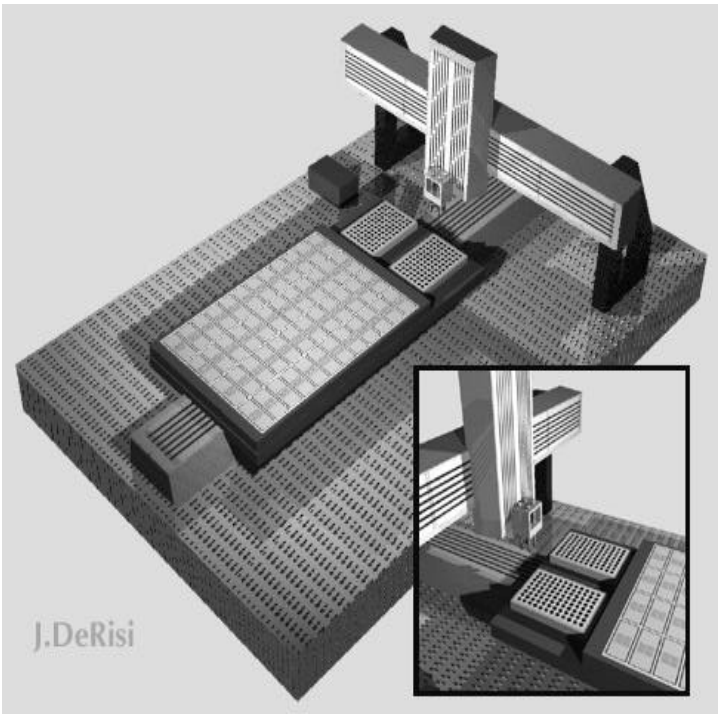
Pharma/biotech in Belgïe

- Janssen Pharmaceutica
- GlaxoSmithKline
- Aventis Crop Science
- Tibotec-Virco
- DevGen
- CropDesign

Bio-informatica op ESAT

- Onderzoeksteam van 12 mensen
  - Elekt. ir., CW, statistiek, physica, wiskunde
  - Bioingenieurs, arts
- Microroostergegevensanalyse





Moleculaire machines

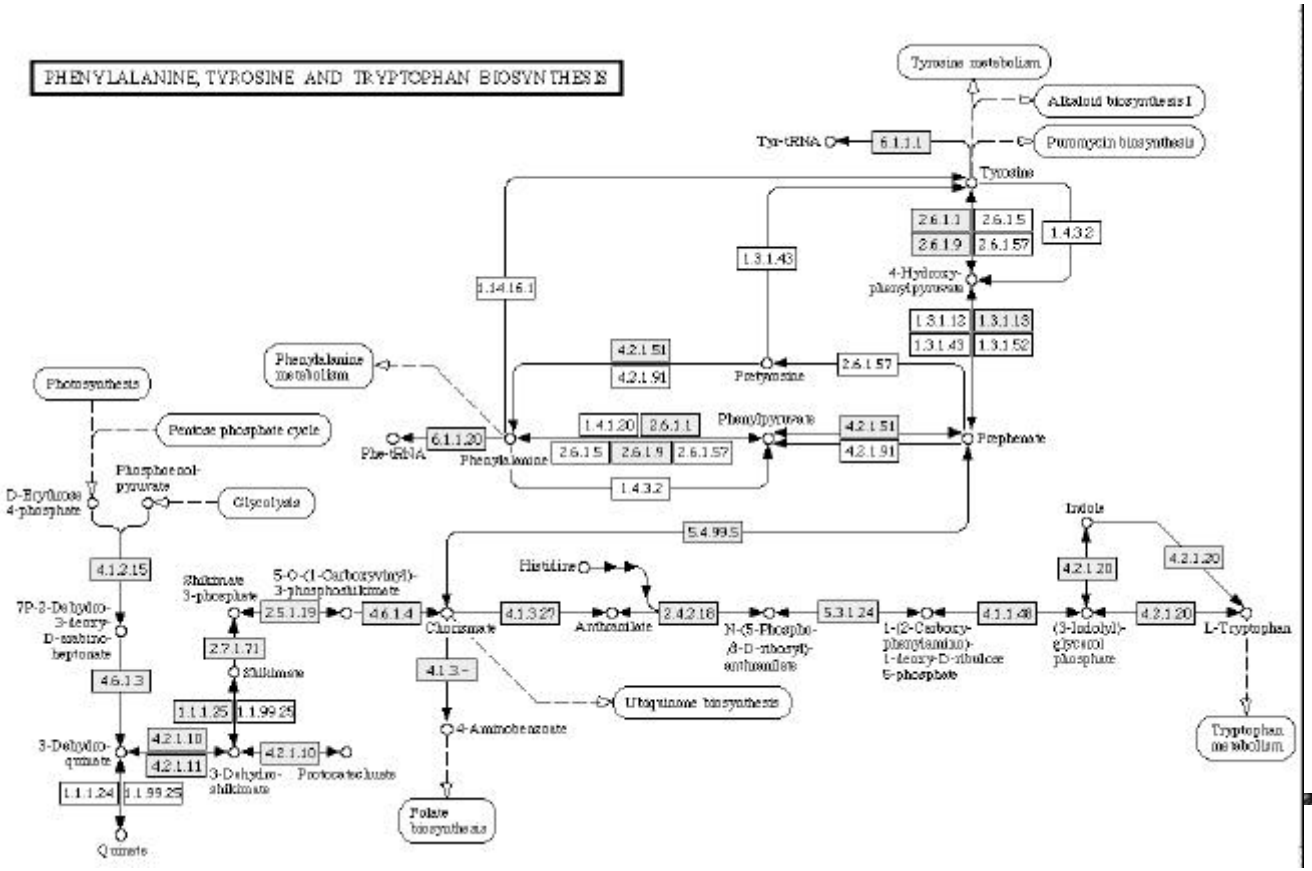
- “DNA replication fork”
- “mRNA translation”
- “Safecrackers”
  - Uit “Essential Cell Biology – Interactive CD”, Alberts et al., Garland Publishing, 1999

Complexiteit in de moleculaire biologie

- Kyoto Encyclopedia of Genomes and Genes
- Evolutie van Genbank en Swissprot
- Publieke databanken van genomische informatie
- MEDLINE query (cancer, tp53)

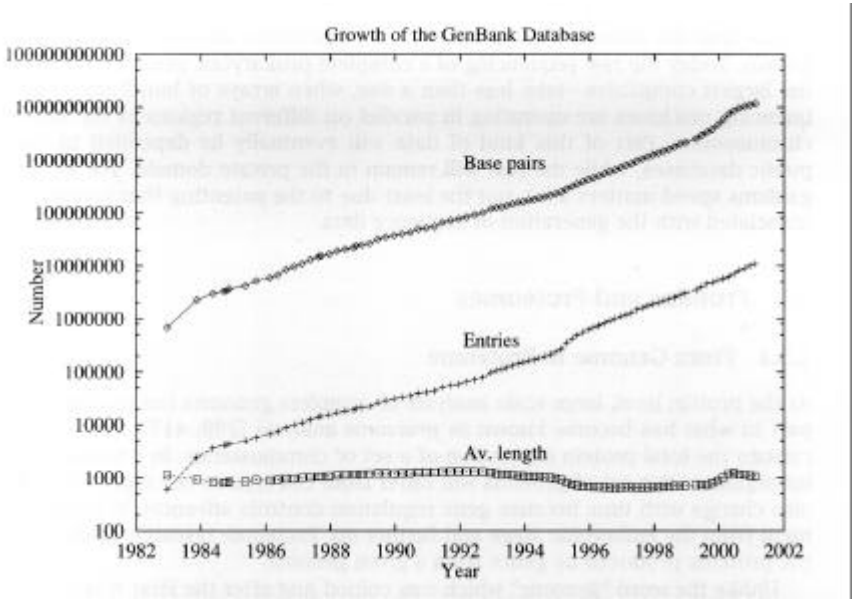
Moleculaire paden

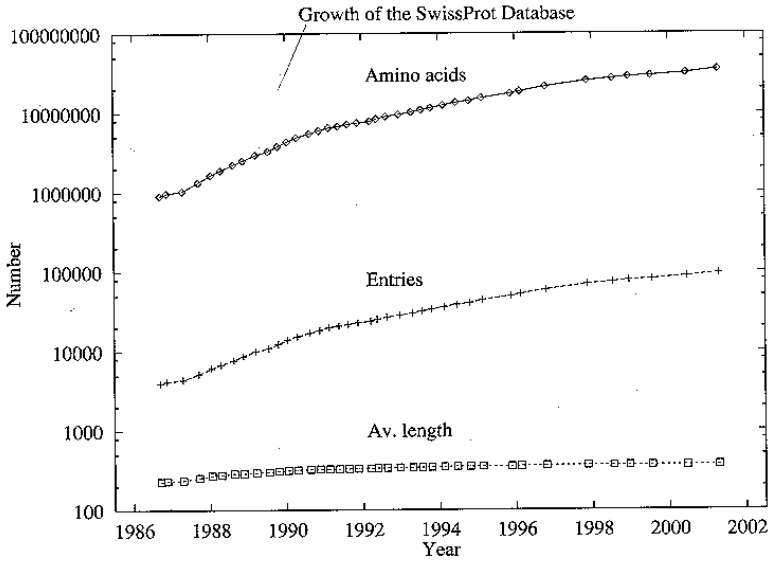
- Kyoto Encyclopedia of Genomes and Genes - Biosynthesepad



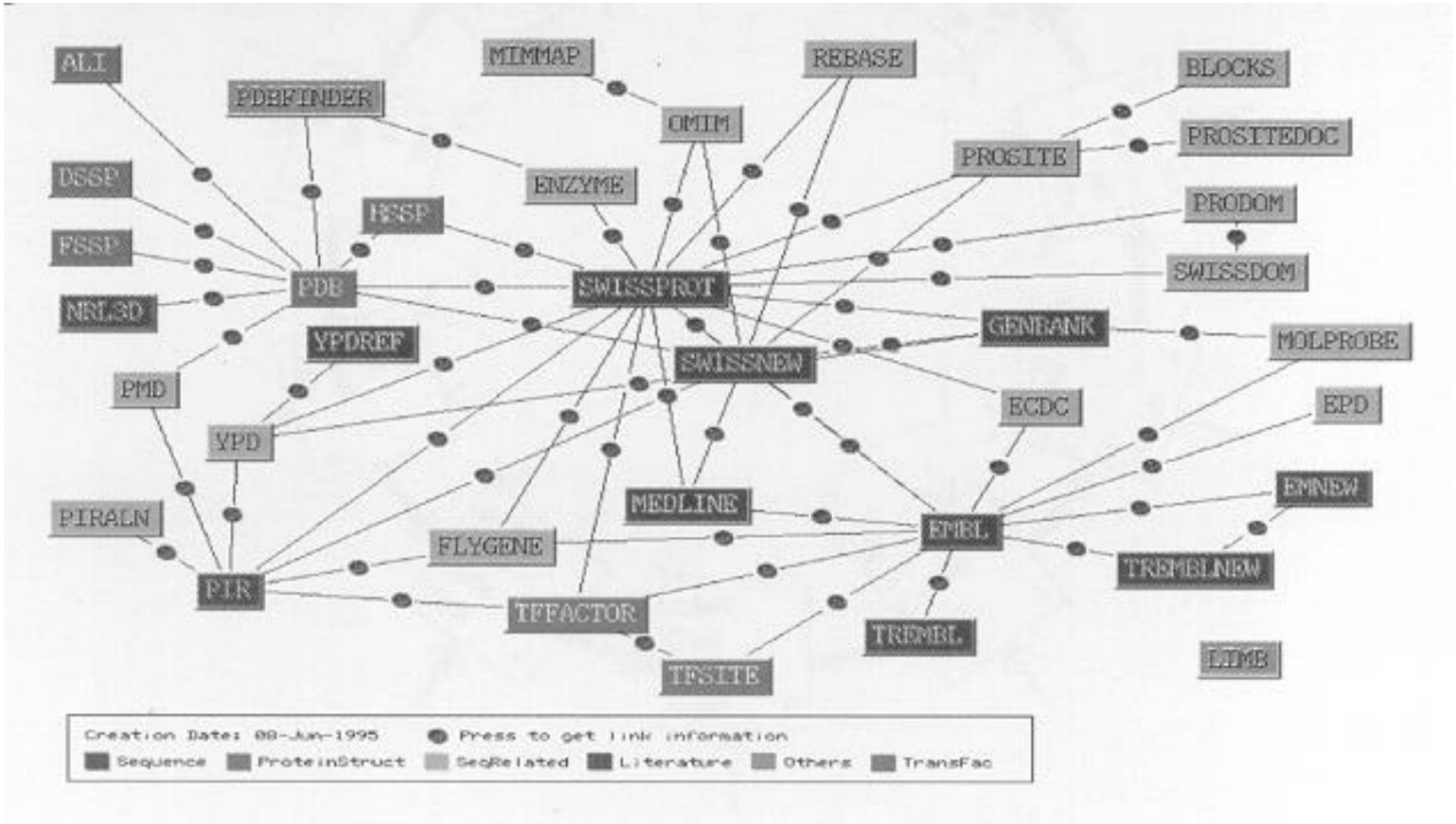
Evolutie van sequentiedatabanken

- Genbank
- SWISSProt

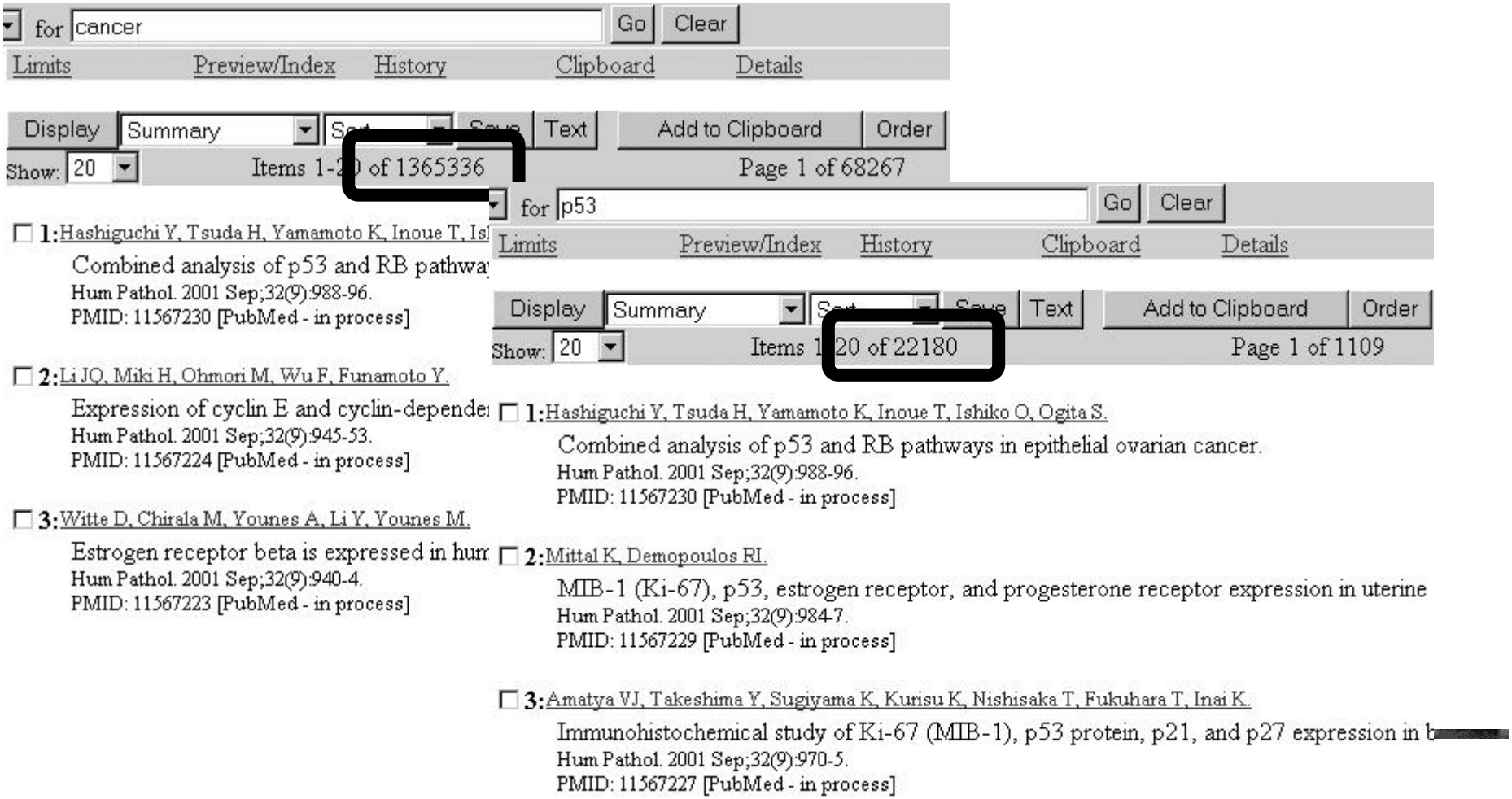




Publieke databanken van genomische informatie



MEDLINE


for cancer
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Display
Summary
Text
Add to Clipboard
Order
Show: 20
Items 1-20 of 1365336
Page 1 of 68267
1: Hashiguchi Y, Tsuda H, Yamamoto K, Inoue T, Is
Combined analysis of p53 and RB pathwa
Hum Pathol. 2001 Sep;32(9):988-96.
PMID: 11567230 [PubMed - in process]
2: Li JQ, Miki H, Ohmori M, Wu F, Funamoto Y.
Expression of cyclin E and cyclin-depende
Hum Pathol. 2001 Sep;32(9):945-53.
PMID: 11567224 [PubMed - in process]
3: Witte D, Chirala M, Younes A, Li Y, Younes M.
Estrogen receptor beta is expressed in hum
Hum Pathol. 2001 Sep;32(9):940-4.
PMID: 11567223 [PubMed - in process]
for p53
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Display
Summary
Text
Add to Clipboard
Order
Show: 20
Items 1 20 of 22180
Page 1 of 1109
1: Hashiguchi Y, Tsuda H, Yamamoto K, Inoue T, Ishiko O, Ogita S.
Combined analysis of p53 and RB pathways in epithelial ovarian cancer.
Hum Pathol. 2001 Sep;32(9):988-96.
PMID: 11567230 [PubMed - in process]
2: Mittal K, Demopoulos RI.
MIB-1 (Ki-67), p53, estrogen receptor, and progesterone receptor expression in uterine
Hum Pathol. 2001 Sep;32(9):984-7.
PMID: 11567229 [PubMed - in process]
3: Amatya VJ, Takeshima Y, Sugiyama K, Kurisu K, Nishisaka T, Fukuhara T, Inai K.
Immunohistochemical study of Ki-67 (MIB-1), p53 protein, p21, and p27 expression in b
Hum Pathol. 2001 Sep;32(9):970-5.
PMID: 11567227 [PubMed - in process]

Elementen van de moleculaire biologie

Illustraties uit *Molecular Biology of the Cell* en
Essential Cell Biology + interactive CD

Tree of Life

- Eukaryoten / prokaryoten (bacteriën + archaeabacteriën)
- Phylogenie

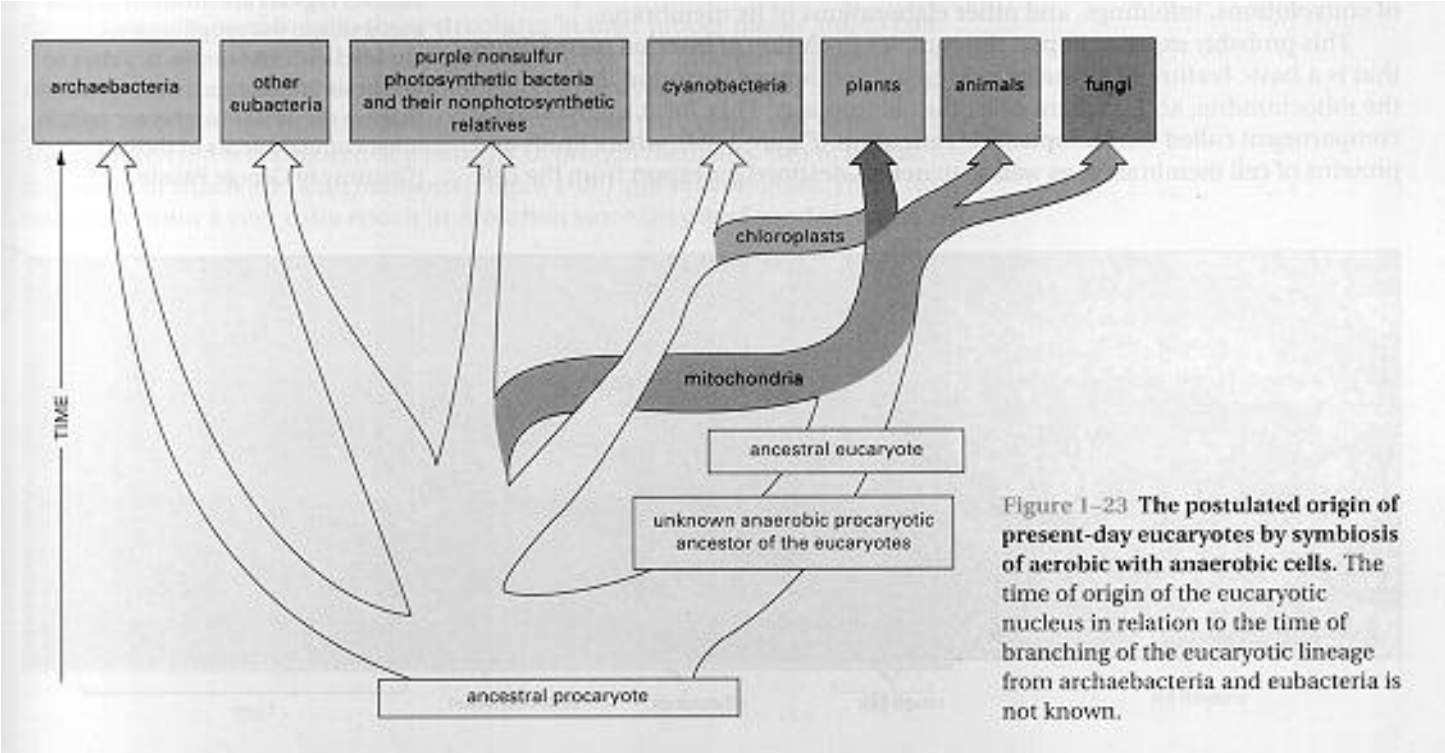


Figure 1–23 **The postulated origin of present-day eucaryotes by symbiosis of aerobic with anaerobic cells.** The time of origin of the eucaryotic nucleus in relation to the time of branching of the eucaryotic lineage from archaebacteria and eubacteria is not known.

Modelorganismen

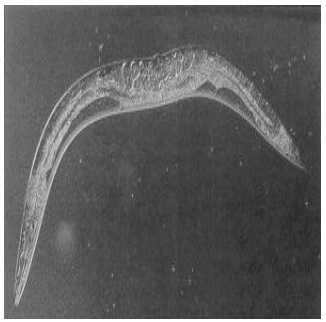
E. coli (3 μm)



Arabidopsis thaliana (20 cm)



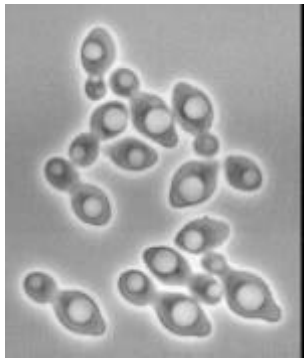
Caenorrabdhitis elegans (1mm)



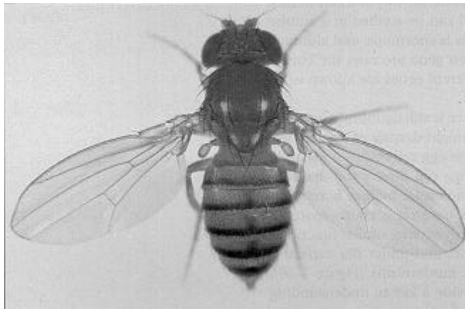
Muis (*Mus musculus*)



Gist (10 μm) (*Saccharomyces Cerevisae*)



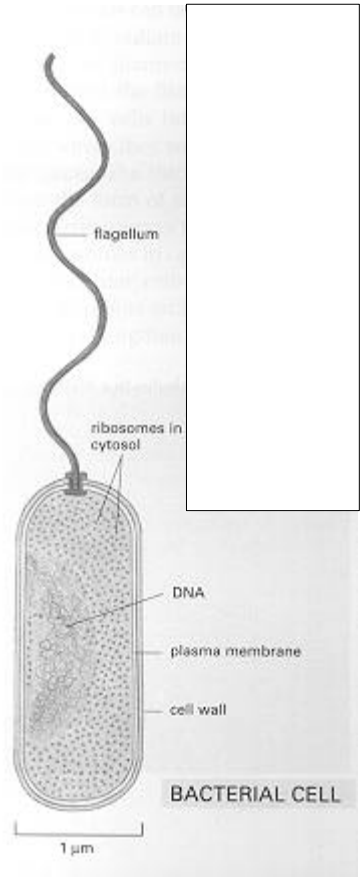
Fruitvlieg (*Drosophila melanogaster*)



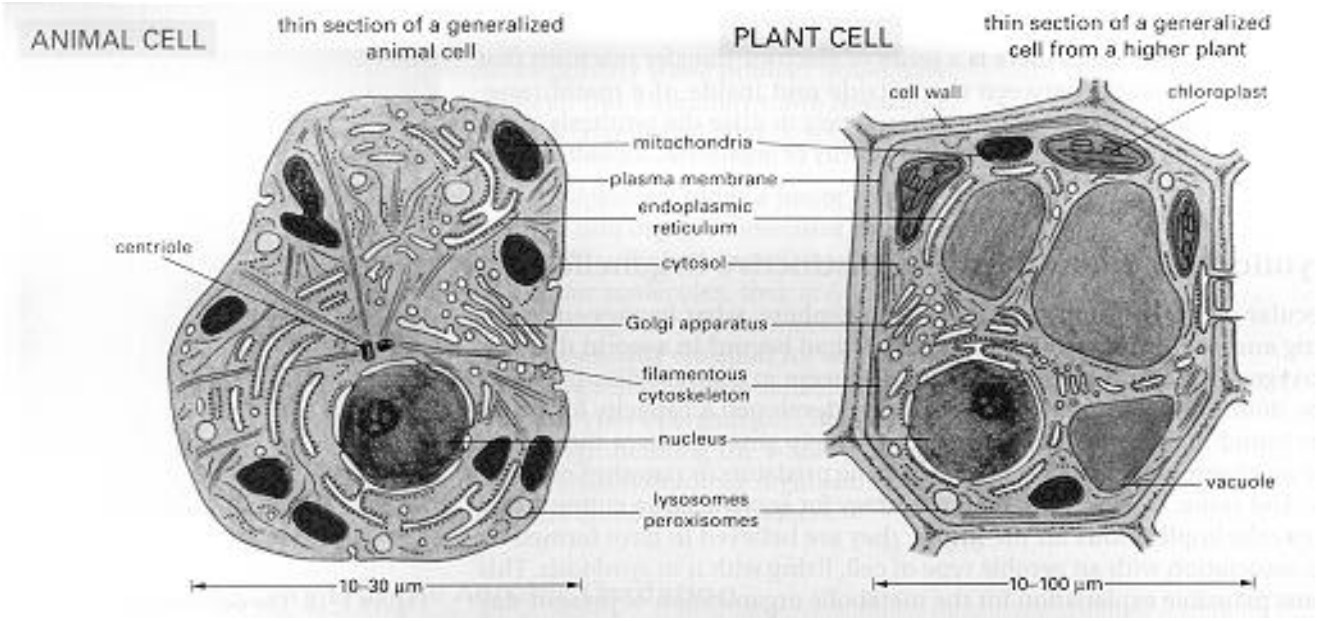
Mens (*Homo sapiens*)



De cel



- Bacteriën/prokaryoten hebben geen celkern om hun genoom te beschermen
- Eukaryoten hebben een celkern



DNA makes RNA makes proteins

- Centraal dogma
- Verschillende mechanisme voor prokaryoten en eukaryoten

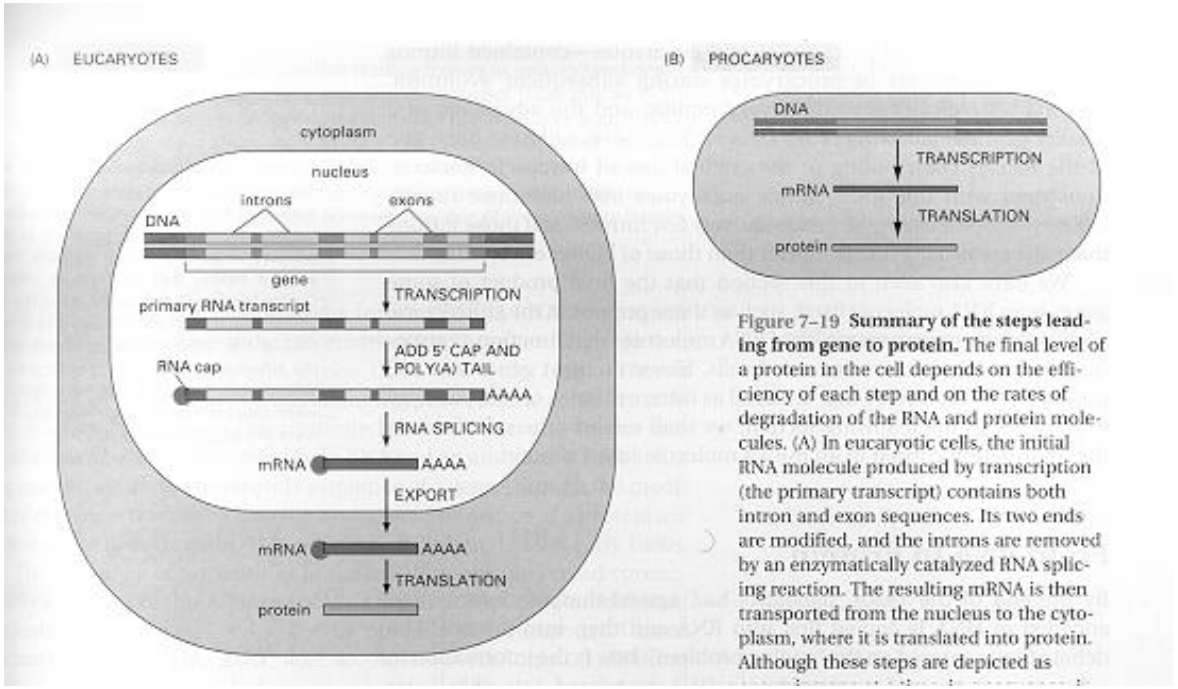
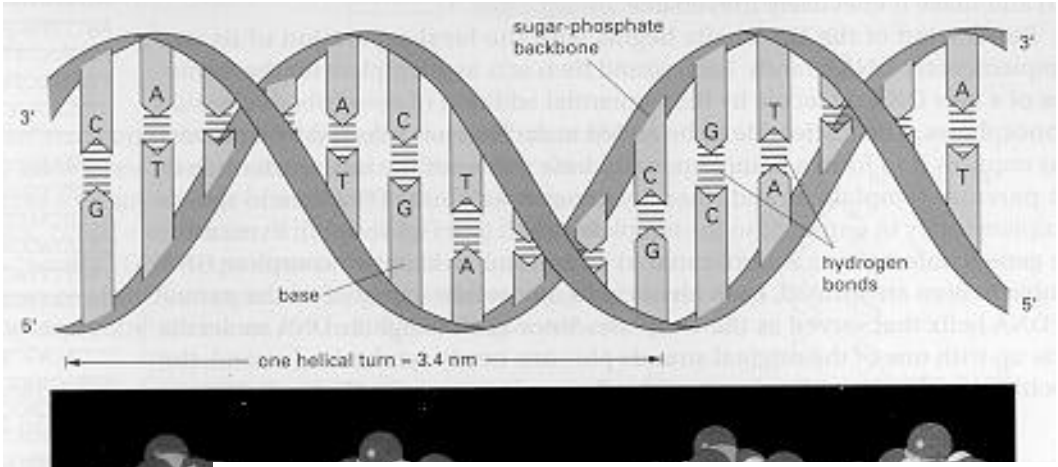
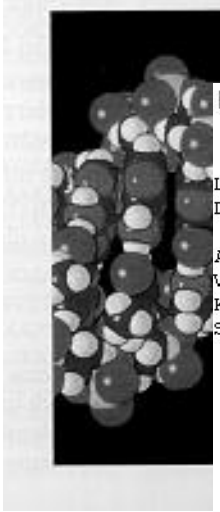


Figure 7–19 **Summary of the steps leading from gene to protein.** The final level of a protein in the cell depends on the efficiency of each step and on the rates of degradation of the RNA and protein molecules. (A) In eucaryotic cells, the initial RNA molecule produced by transcription (the primary transcript) contains both intron and exon sequences. Its two ends are modified, and the introns are removed by an enzymatically catalyzed RNA splicing reaction. The resulting mRNA is then transported from the nucleus to the cytoplasm, where it is translated into protein. Although these steps are depicted as

DNA

- DNA helix (6.1)
- Complementaire strengen (A-T, G-C)
- Draait ‘clockwise’, 10 nucleotiden per omwenteling



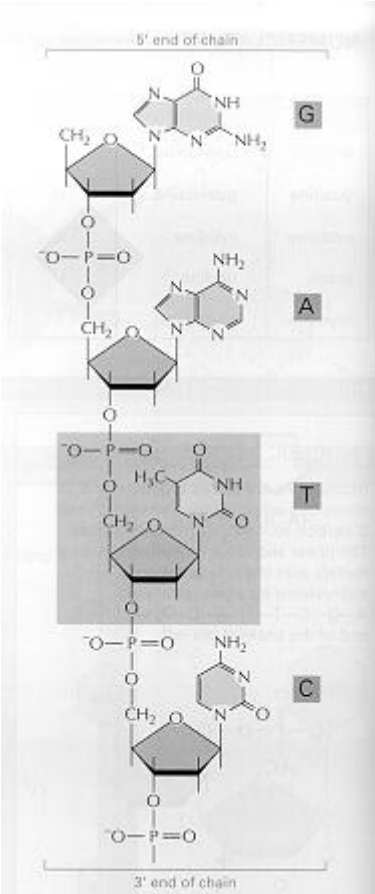


1: NM_007295. Homo sapiens brea...[gi:6552300]

```
LOCUS       NM_007295    7365 bp    mRNA            PRI       03-FEB-2001
DEFINITION  Homo sapiens breast cancer 1, early onset (BRCA1), transcript
            variant BRCA1b, mRNA.
ACCESSION   NM_007295
VERSION     NM_007295.1  GI:6552300
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
 BASE COUNT     2458 a   1434 c   1672 g   1801 t
ORIGIN
        1 ggcagtttgt aggtcgcgag ggaagcgctg aggatcagga agggggcact gagtgtccgt
       61 gggggaatcc tcgtgatagg aactggaata tgccttgagg gggacactat gtctttaaaa
      121 acgtcggctg gtcatgaggt caggagttcc agaccagcct gaccaacgtg gtgaaactcc
      181 gtctctacta aaaatacaaa aattagccgg gcgtggtgcc gctccagcta ctcaggaggc
      241 tgaggcagga gaatcgctag aacccgggag gcggaggttg cagtgagccg agatcgcgcc
      301 attgcactcc agcctgggcg acagagcgag actgtctcaa aacaaaacaa aacaaaacaa
      361 aacaaaaaac accggctgtt cattggaaca gaaagaaatg gatttatctg ctcttcgcgt
      421 tgaagaagta caaaatgtca ttaatgctat gcagaaaatc ttagagtgtc ccatctgtct
```

Chemische structuur van het DNA

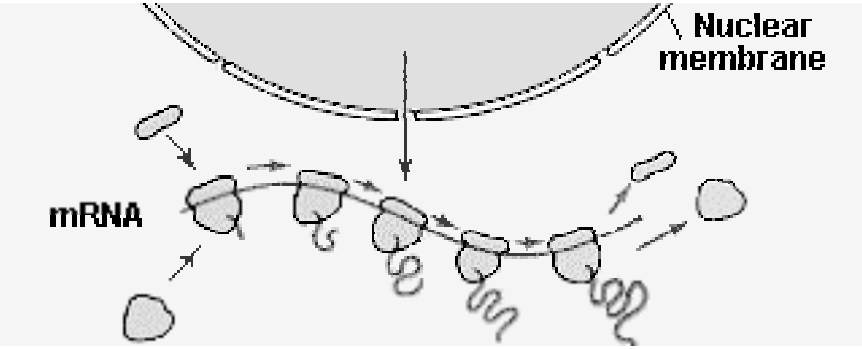
- Deoxyribonucleïnezuur
  - Adenine
  - Thymine
  - Cytosine
  - Guanine
- Structuur
  - Suiker (ribose)
  - Phosphaatgroep
  - Ringverbinding
- Basecomplementariteit via waterstofverbinding
- Orientatie: van 5’ naar 3’



RNA

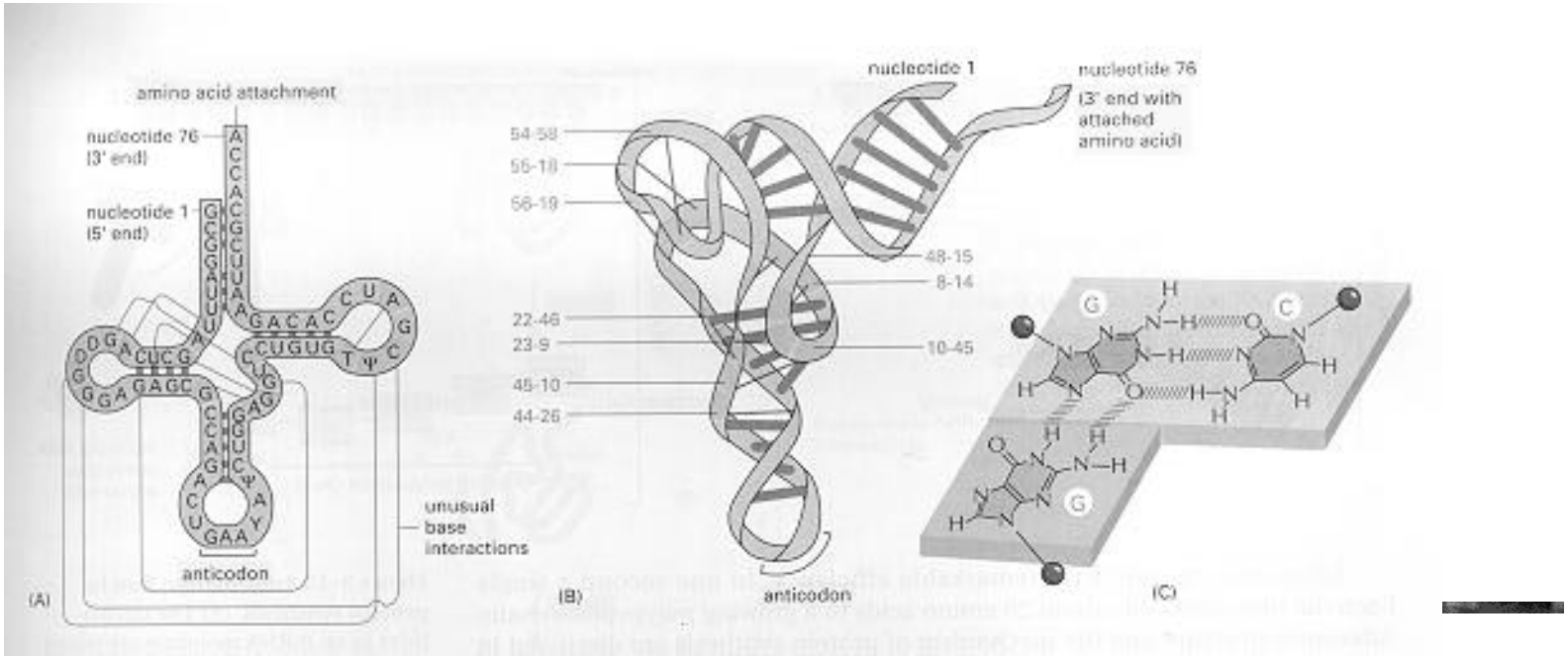
- Adenine – Uracil (vs. Thymine), Guanine – Cytosine

- Enkelvoudige streng (A-U paar zwakker dan A-T)
- Waarschijnlijk voorouder van DNA
- Dient vooral als informatiedrager van DNA naar proteïne



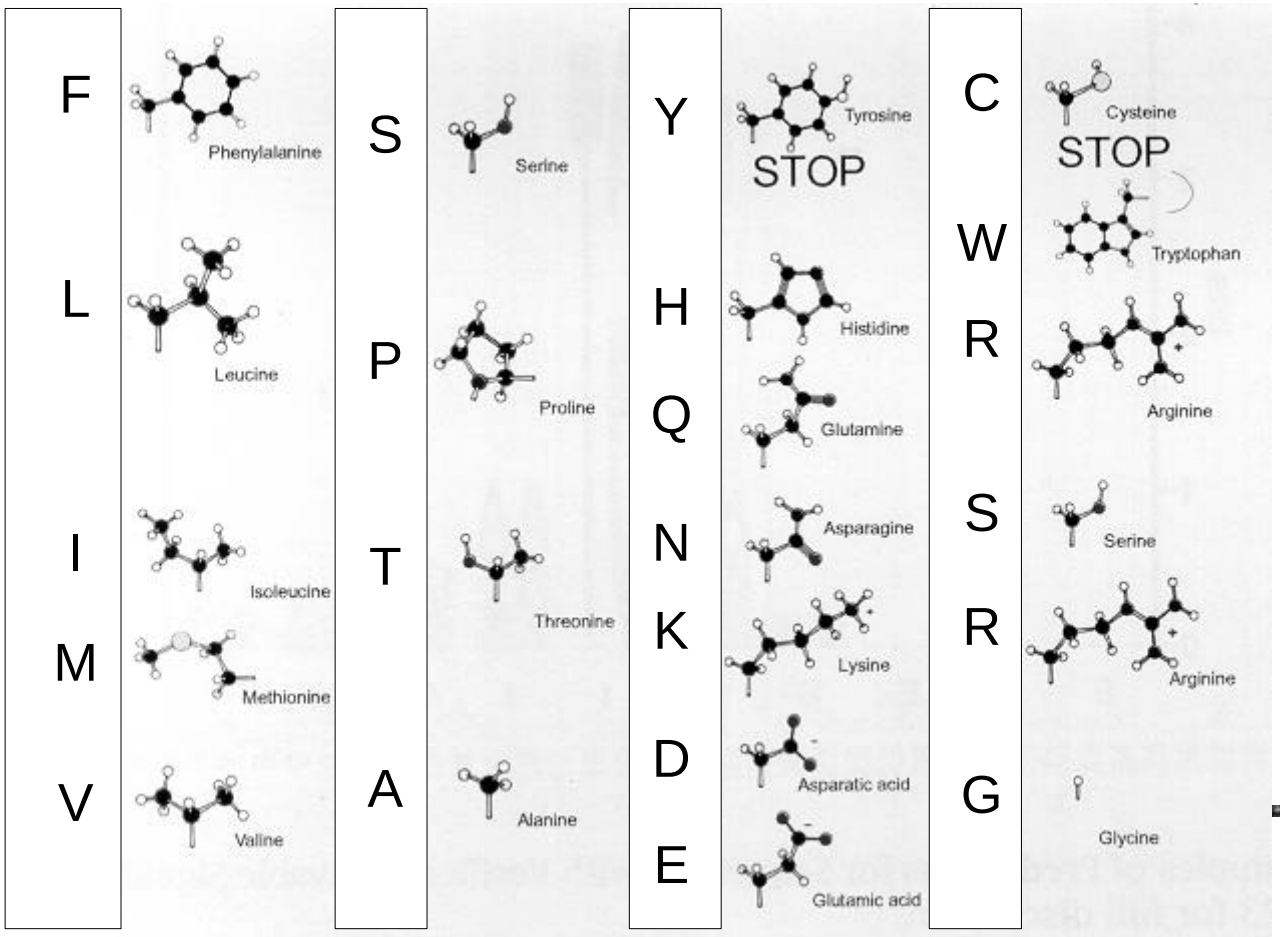
Functioneel RNA

- RNA heeft ook belangrijke en complexe functionele rollen
- Functionele RNAs hebben een complexe structuur
- RNA ribosomal subunit 30S



Proteïnen

- Grote polymeren van 20 aminozuren



Tridimensionale structuur van proteïnen

- Basis tridimensionale structuren
  - Alpha helix (ECBI 5.2)
  - Beta sheet (ECBI 5.3)
  - Loop
- Voorbeelden van proteïnen
  - Antilichaam (ECBI 5.4)
  - Neuramidase (ECBI 5.5)
  - Elongation factor EF-Tu (ECBI 5.
- Computationele biologie

SWISSPROT:P53_HUMAN

```
ID   P53_HUMAN      STANDARD;      PRT;   393 AA.
AC   P04637; Q16848; Q9UBI2;
DT   13-AUG-1987 (Rel. 05, Created)
DT   01-MAR-1989 (Rel. 10, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53) (PHOSPHOPROTEIN
DE   P53).
GN   TP53 OR P53.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI TaxID=9606;
SQ   SEQUENCE   393 AA;  43653 MW;  AD5C149FD8106131 CRC64;
     MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP
     DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ KTYQGSYGFR LGFLHSGTAK
     SVTCTYSPAL NKMFCQLAKT CPVQLWVDST PPPGTRVRAM AIYKQSQHMT EVVRRCPHHE
     RCSDSDGLAP PQHLIRVEGN LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS
     SCMGGMNRRP ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP
     PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL KDAQAGKEPG
     GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD
//
```

Proteïnerollen

- Proteïne hebben veel functies in cellen
  - Katalyse (enzymen)
  - Bouwstructuur
  - Vervoer
  - Motor
  - Moleculaire opslag
  - Signalisatie
  - Receptoren
  - Regulatie
- Voorspelling van proteïnefamilies



Kleine moleculen

- Allerhande kleine organische moleculen zijn nodig voor de cel en zijn betrokken bij het metabolisme

- Enzymen zijn proteïne die de chemische transformatie van deze moleculen catalyseren.

- Glucose (ECBI 2.1)
- Adenosine triphospaat (ECBI 2.3)

Energieproductie

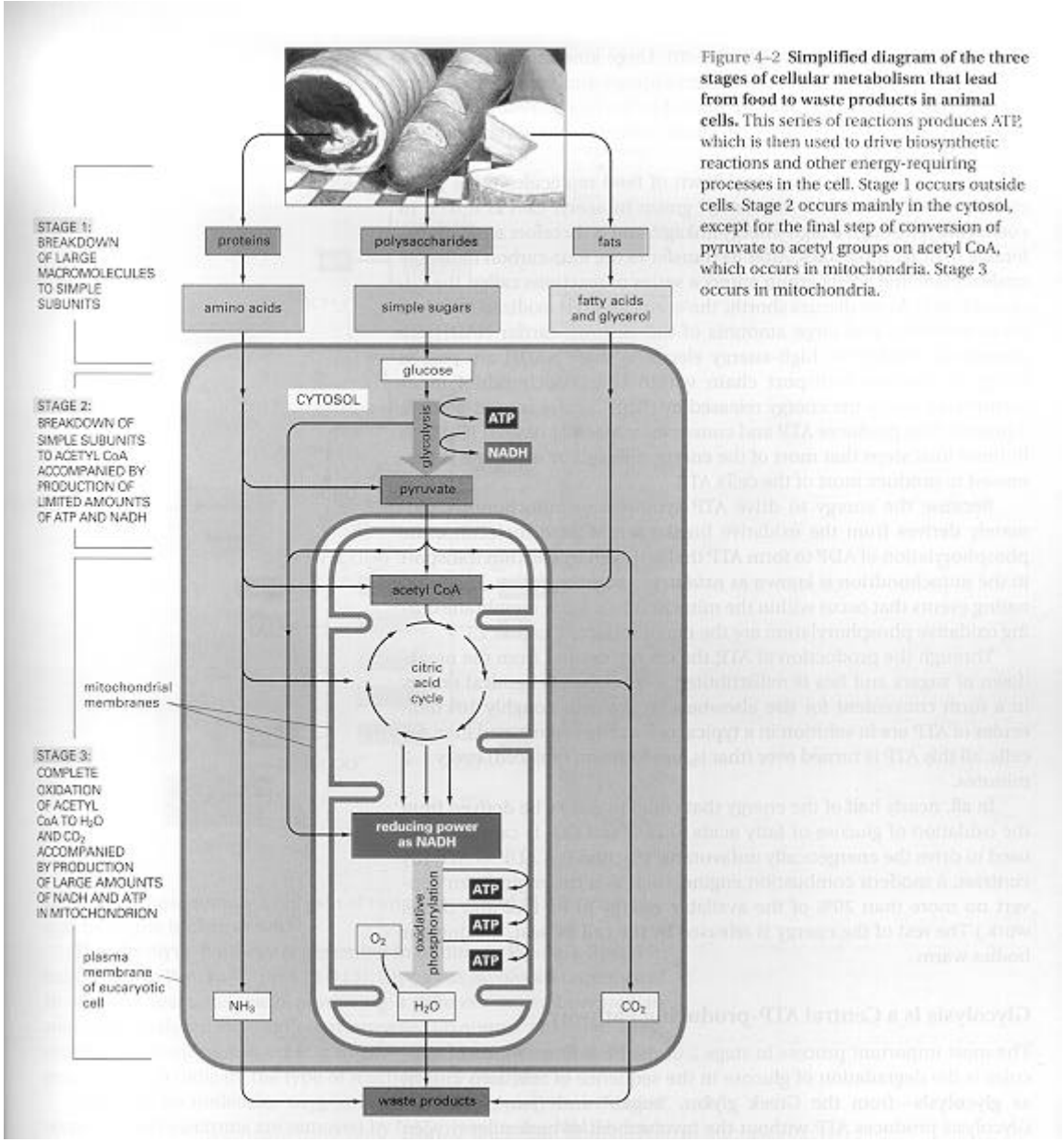


Figure 4–2 **Simplified diagram of the three stages of cellular metabolism that lead from food to waste products in animal cells.** This series of reactions produces ATP, which is then used to drive biosynthetic reactions and other energy-requiring processes in the cell. Stage 1 occurs outside cells. Stage 2 occurs mainly in the cytosol, except for the final step of conversion of pyruvate to acetyl groups on acetyl CoA, which occurs in mitochondria. Stage 3 occurs in mitochondria.

Glycolyse

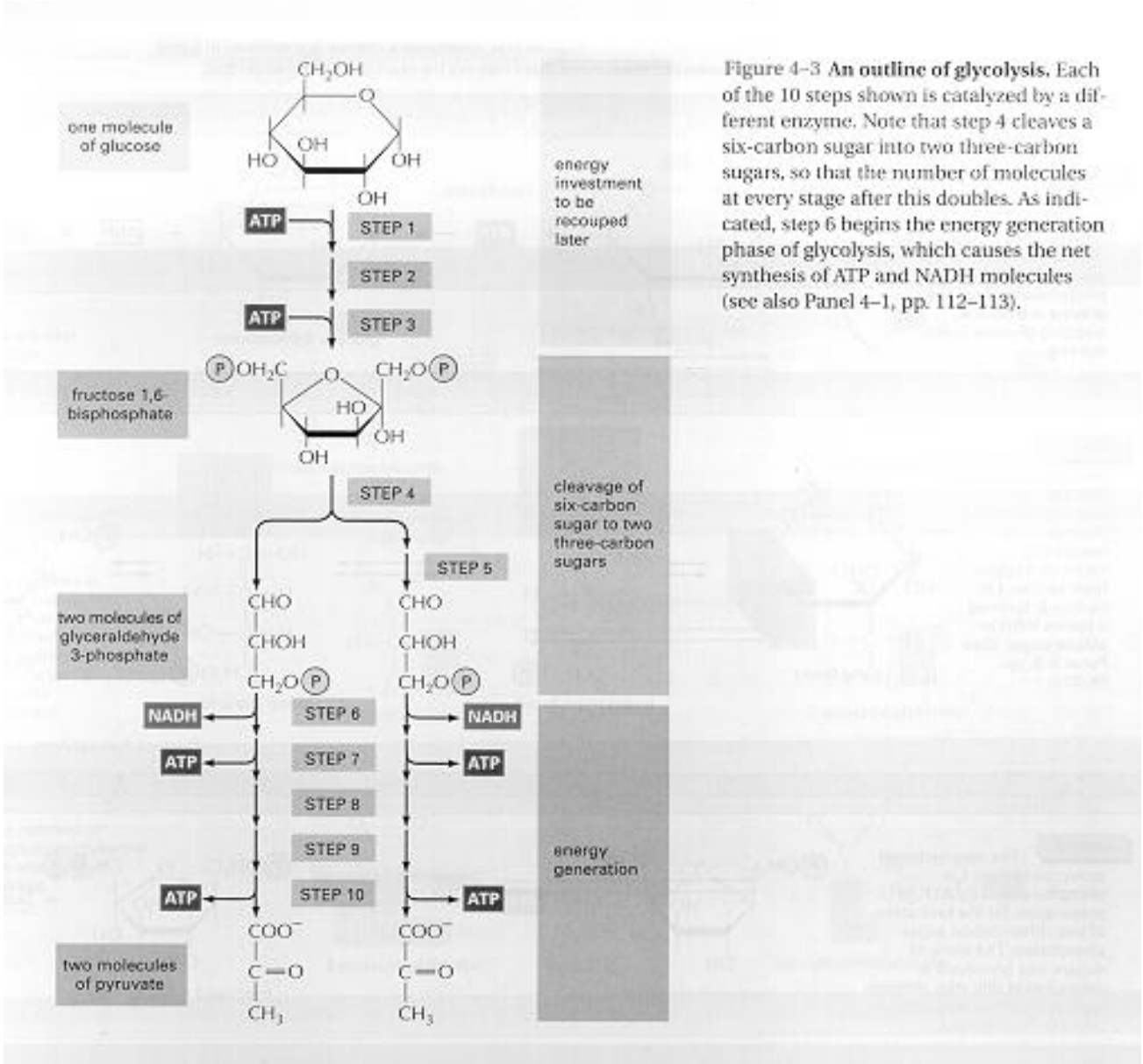


Figure 4–3 **An outline of glycolysis.** Each of the 10 steps shown is catalyzed by a different enzyme. Note that step 4 cleaves a six-carbon sugar into two three-carbon sugars, so that the number of molecules at every stage after this doubles. As indicated, step 6 begins the energy generation phase of glycolysis, which causes the net synthesis of ATP and NADH molecules (see also Panel 4–1, pp. 112–113).

Metabolisme

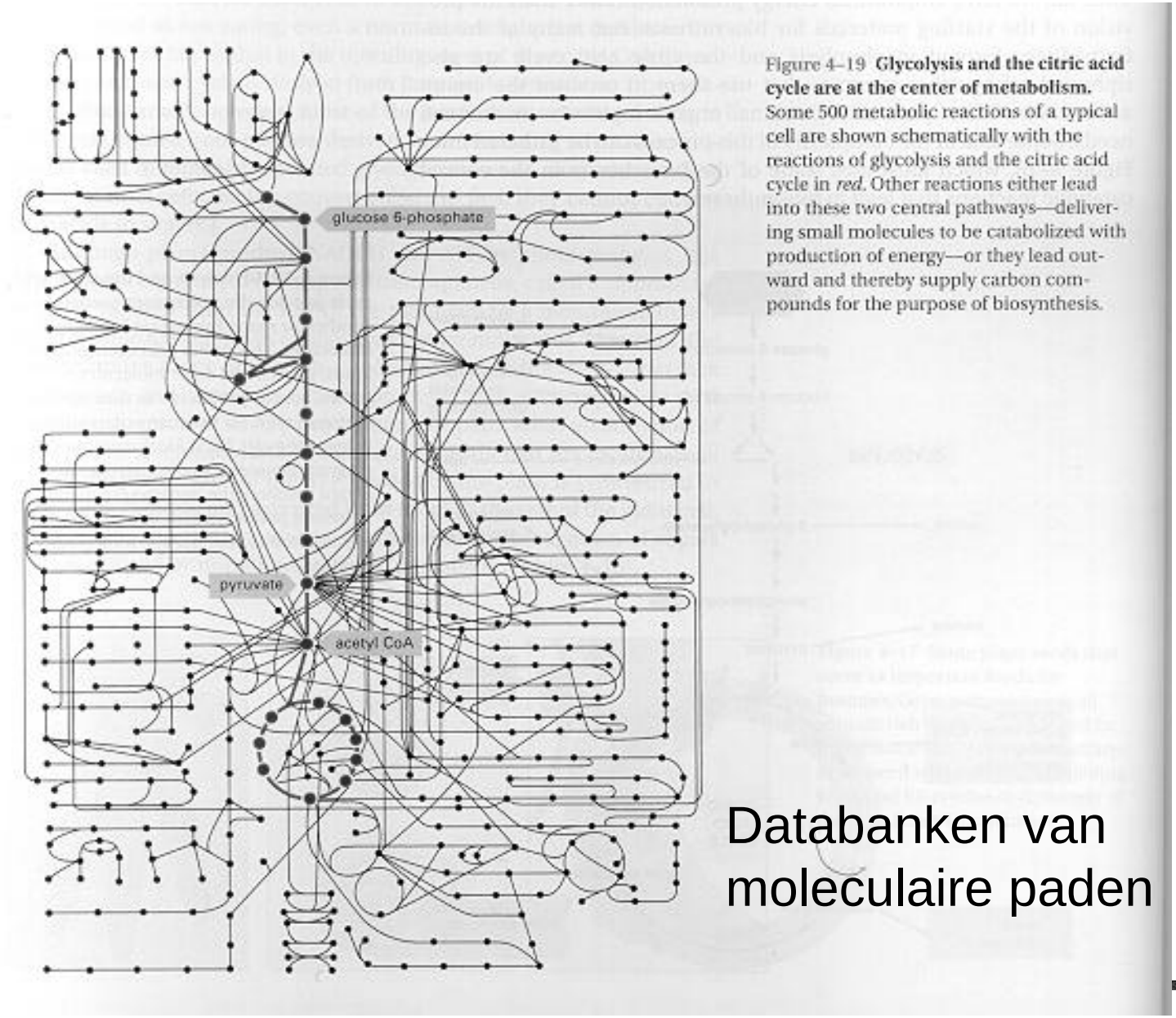


Figure 4–19 **Glycolysis and the citric acid cycle are at the center of metabolism.** Some 500 metabolic reactions of a typical cell are shown schematically with the reactions of glycolysis and the citric acid cycle in *red*. Other reactions either lead into these two central pathways—delivering small molecules to be catabolized with production of energy—or they lead outward and thereby supply carbon compounds for the purpose of biosynthesis.

Databanken van moleculaire paden

DNA makes RNA makes proteins

- Genen dragen de informatie voor de productie van proteïnen
  - Transcriptie (van DNA naar mRNA) door RNA polymerase
  - Translatie (van mRNA naar proteïne) door ribosomen

- Genvoorspelling

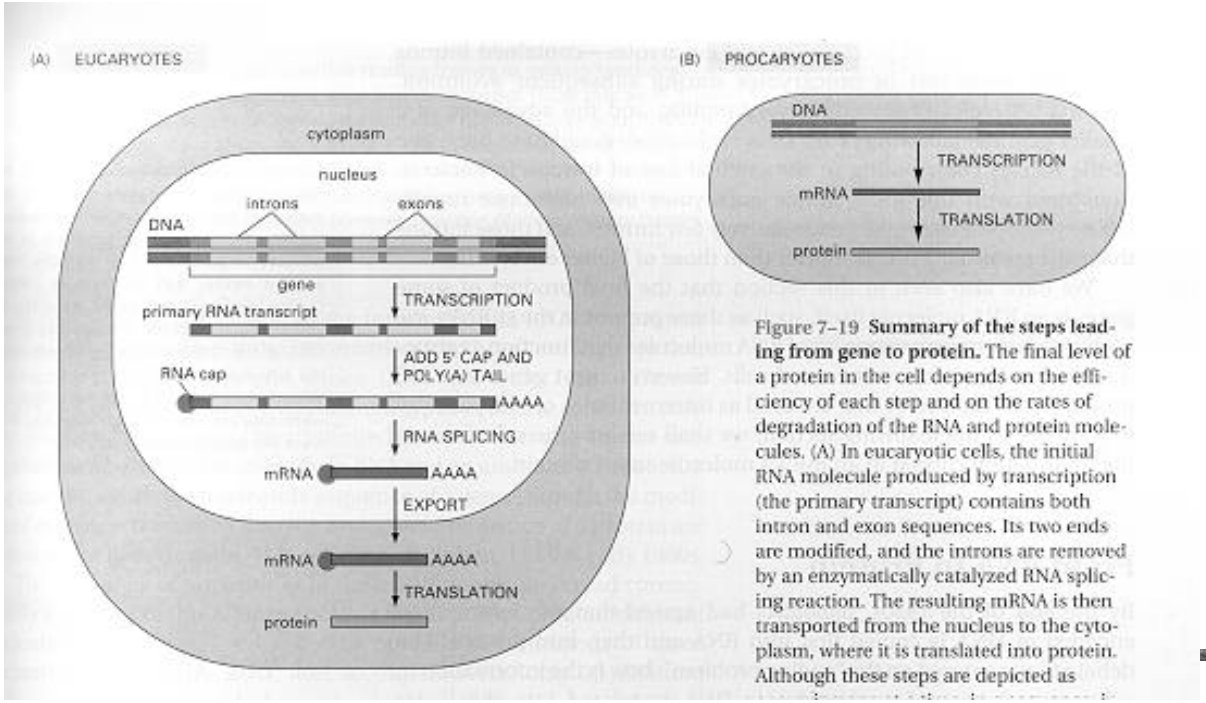
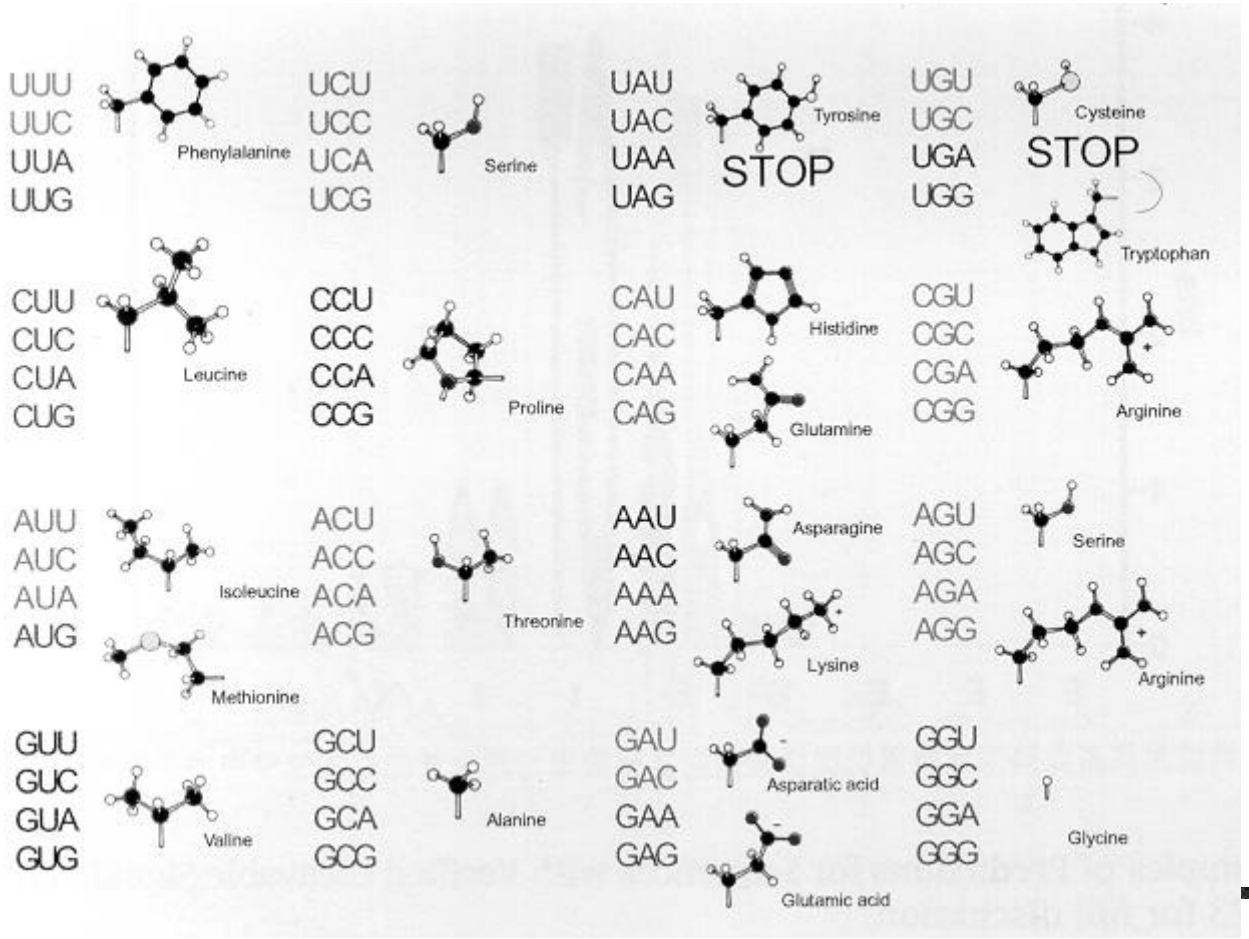


Figure 7–19 **Summary of the steps leading from gene to protein.** The final level of a protein in the cell depends on the efficiency of each step and on the rates of degradation of the RNA and protein molecules. (A) In eucaryotic cells, the initial RNA molecule produced by transcription (the primary transcript) contains both intron and exon sequences. Its two ends are modified, and the introns are removed by an enzymatically catalyzed RNA splicing reaction. The resulting mRNA is then transported from the nucleus to the cytoplasm, where it is translated into protein. Although these steps are depicted as

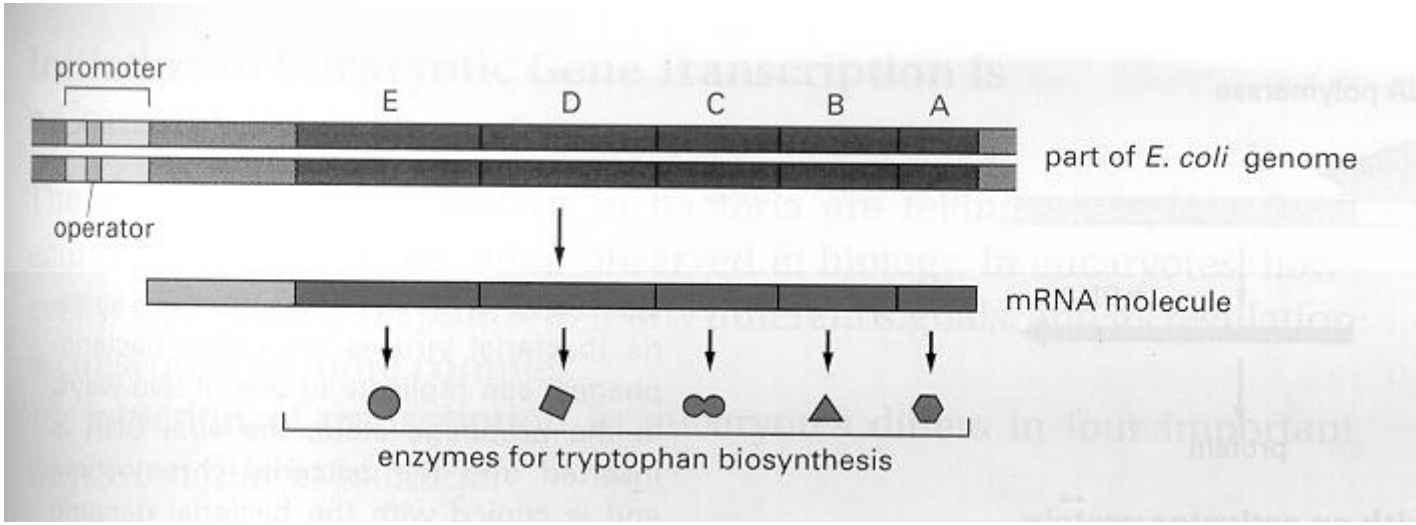
Aminozuren en de genetische code

- 64 codons
- Start & Stop codon
- 6 reading frames



Operons in prokaryoten

- Prokaryoten hebben operons waar meerdere genen tegelijk worden getranscribeerd.



RNA splitsing in eukaryoten

- In eukaryoten wordt het gen getranscribeerd naar pre-mRNA, na welke *intronen* verwijderd worden langs de *splice sites* om een boodschapper mRNA van *exonen* te vormen
- De mRNA wordt getranslateerd van de START codon tot de STOP codon (de 5’ en 3’ UnTranslated Regions worden niet getranslateerd)

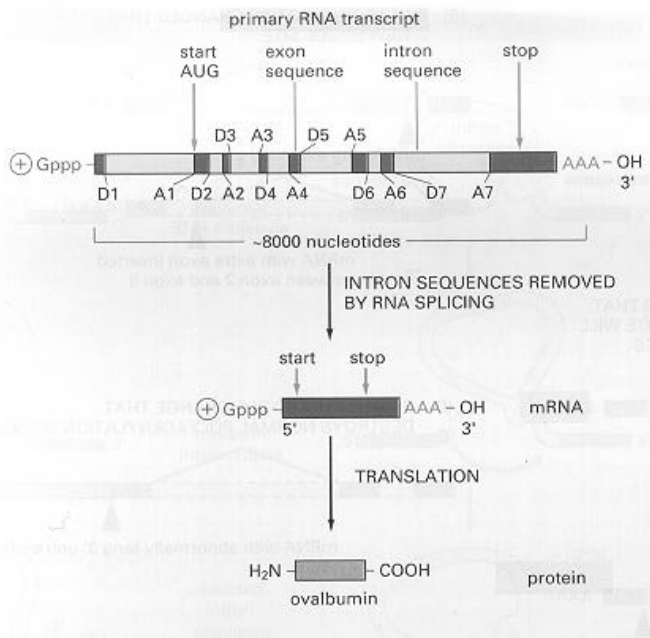


Figure 8–56 Splicing the primary RNA transcript from the chicken ovalbumin gene. The drawing shows the organized removal of seven introns required to obtain a functional mRNA molecule. The 5′ splice sites (donor sites) are denoted by D, and 3′ splice sites (acceptor sites) are denoted by A.

Alternatieve splitsing

- Alternatieve splitsing genereert meerdere proteïnenvarianten uit één gen

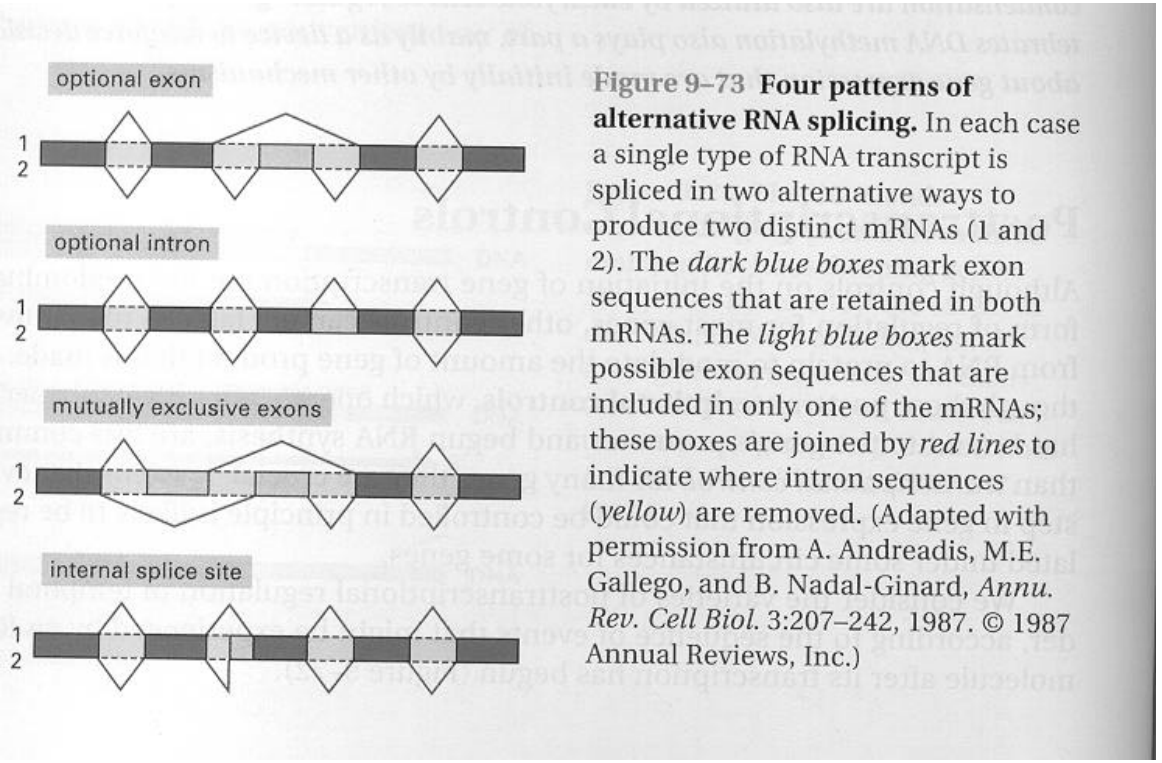


Figure 9–73 Four patterns of alternative RNA splicing. In each case a single type of RNA transcript is spliced in two alternative ways to produce two distinct mRNAs (1 and 2). The *dark blue boxes* mark exon sequences that are retained in both mRNAs. The *light blue boxes* mark possible exon sequences that are included in only one of the mRNAs; these boxes are joined by *red lines* to indicate where intron sequences (*yellow*) are removed. (Adapted with permission from A. Andreadis, M.E. Gallego, and B. Nadal-Ginard, *Annu. Rev. Cell Biol.* 3:207–242, 1987. © 1987 Annual Reviews, Inc.)

Voorbeeld van alternatieve splicing

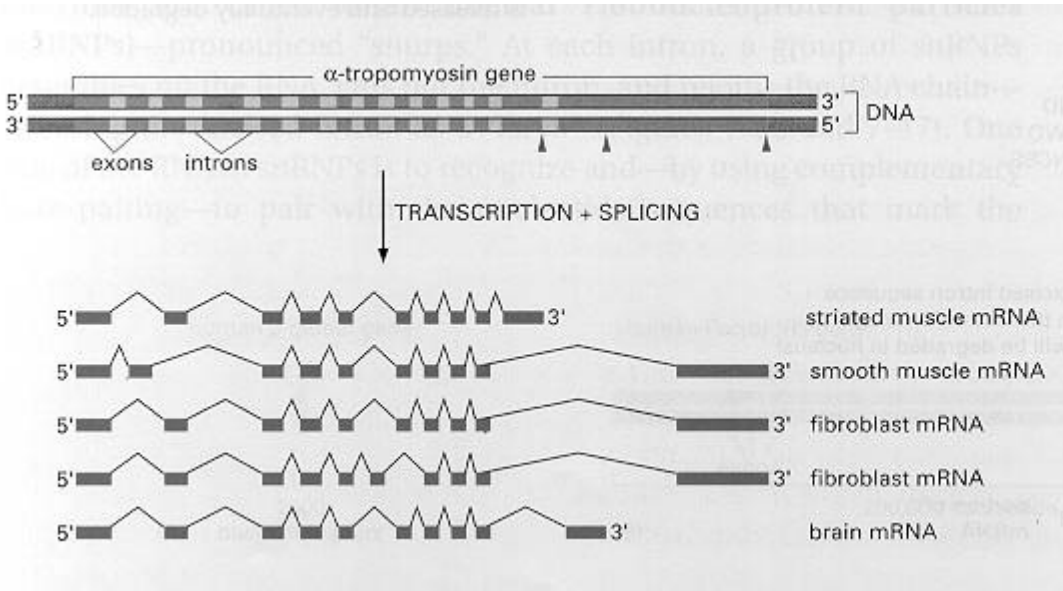


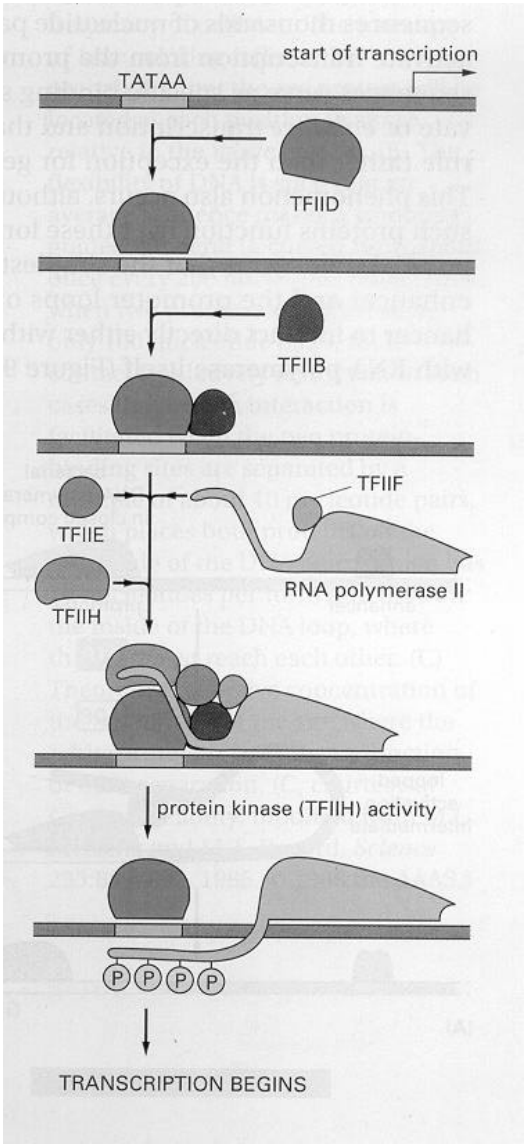
Figure 7–18 Alternative splicing of the α-tropomyosin gene from rat.
α-Tropomyosin is a coiled-coil protein (see Figure 5–11) that regulates contraction in muscle cells. Its role in other cells is not well understood. Once the primary transcript is made, it can be spliced in different ways, as indicated in the figure, to produce distinct mRNAs that then give rise to variant proteins. Some of the splicing patterns are specific for certain types of cells. For example, the α-tropomyosin made in striated muscle is different from that made in smooth muscle. The arrowheads in the top part of the figure represent sites where cleavage and poly(A) addition can occur.

Controle van regulatie

- Om transcriptie te kunnen beginnen moet een uitgebreide complex van proteïnen aanwezig zijn

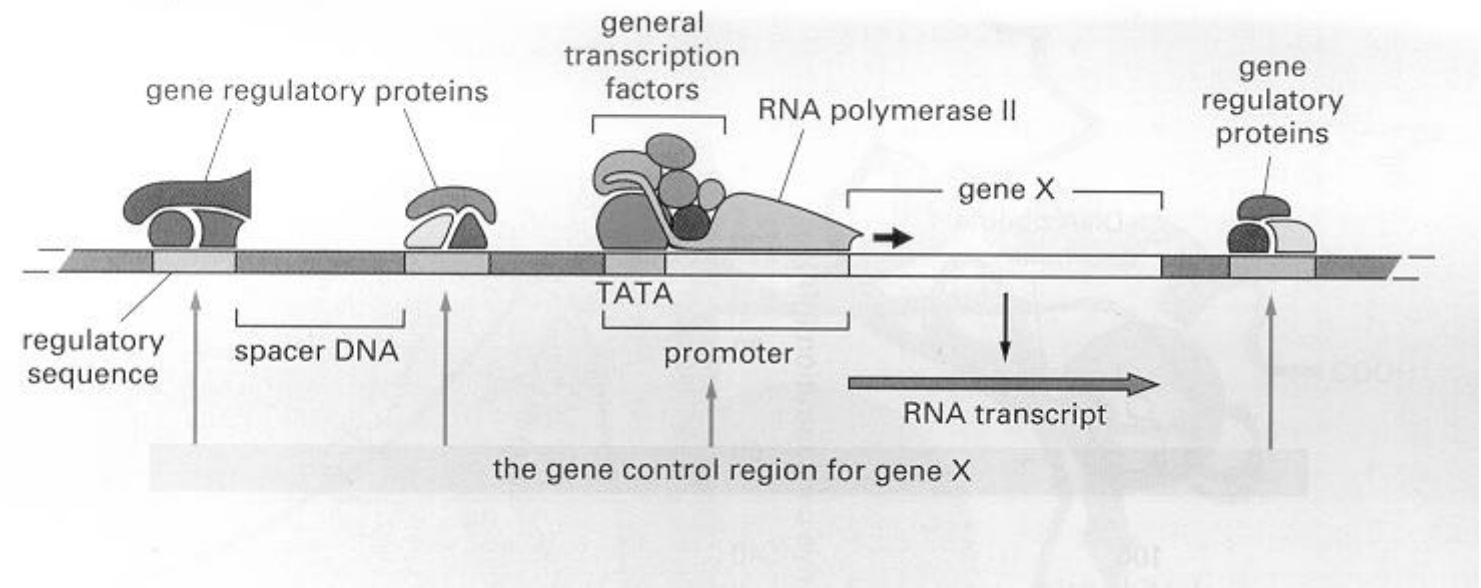
- De controle van regulatie gebeurdt via de promotor
  - Bindingplaats
  - Bindingproteïne / transcriptie factor

- Promotorvoorspelling en motif finding



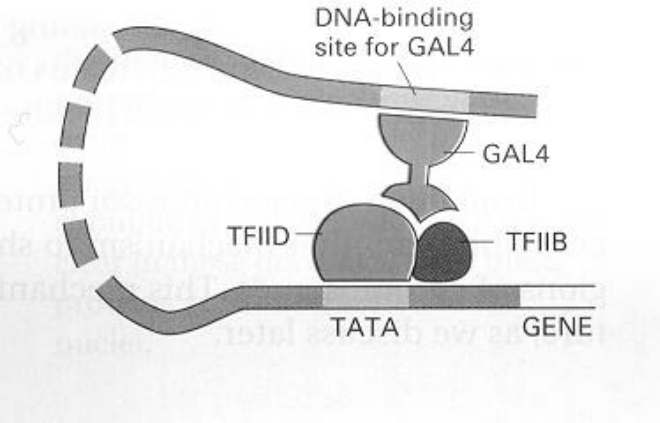
Promotor-architectuur

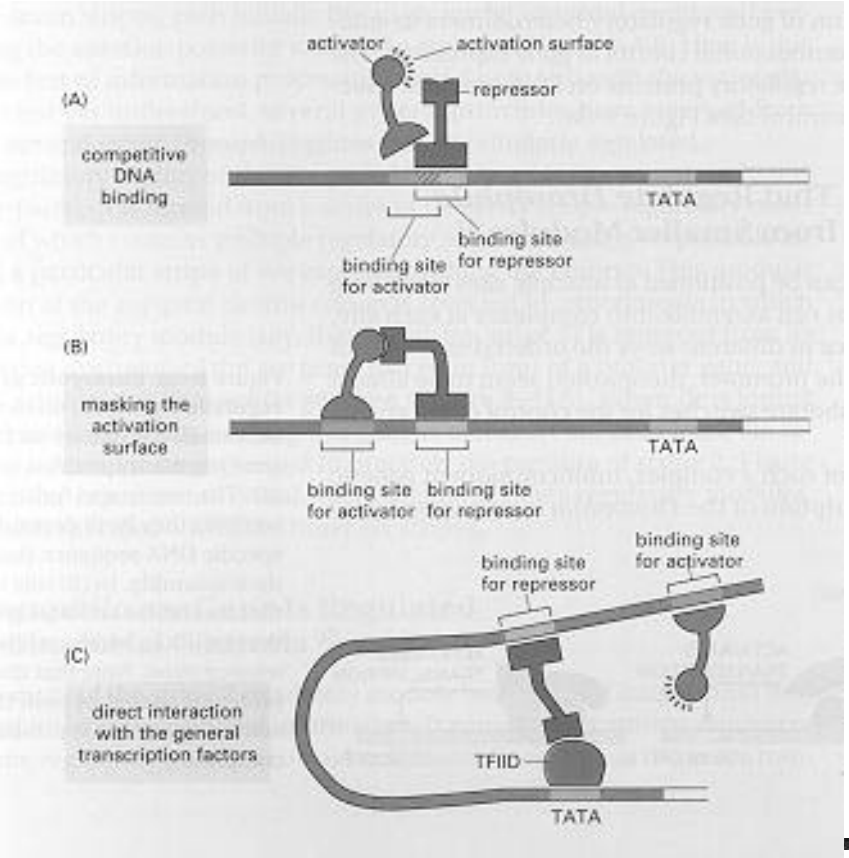
- Kernpromotor: TATA box – TATA binding protein, RNA polymerase II, …
- Distale promoter: enhancers, repressoren



Enhancers en repressoren

- Homeodomain (8.1)
- Zinc finger motif (8.2)
- Leucine zipper motif (8.3)
- TATA-binding protein (8.4)





Combinatoriale controle

- Complexe integratie van signalen bepaalt de genactiviteit
- Microroostergegevensanalyse van coexpressie-experimenten

