

Bioinformatics

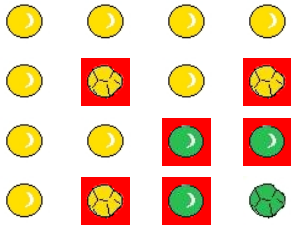

(Warm Up + Cracking the Genetic Code)

Marc Hellmuth

LECTURE 1

Basic Problem: Understand Inheritance

- 1860's Mendel (abstract essentially math. model for "inheritance unit")
- 1869 Miescher discovered DNA
- 1944 Oswald T. Avery, Colin M. MacLeod und Maclyn McCarty: (first clear suggestion that DNA carries genetic information)
- 1949 Erwin Chargaff "bases come in pairs"
- 1952 Herschey and Chase (confirmed results of Miescher)
- 1952 Rosalind Franklin (Photo 51 Xray)
- 1953 Watson and Crick (double helical structure of DNA)
- 2003 Human genome is sequenced



He mentioned that biological variations are inherited from parent organism as specific discrete traits.

1896 Friedrich Miescher discovered DNA

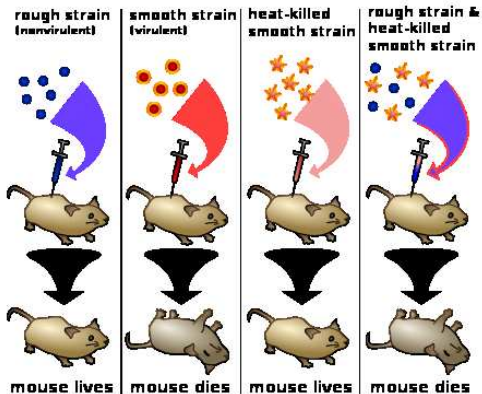
- FM worked with with blood cells
- White blood cells are a major component of pus (dt. Eiter) in infections. So he collected a lot of pus from bandages at local hospitals
- He extracted the nuclei (by adding weak alkaline solution to the white blood cells)
- From the nuclei, he isolated various phosphate-rich chemicals, which he called nuclein (now nucleic acids)
- He raised the idea that the nucleic acids could be involved in heredity

1944 Oswald T. Avery, Colin M. MacLeod und Maclyn McCarty

Experiments on *Streptococcus pneumoniae* - two strains:

S-strain covered itself with a polysaccharide capsule that protected it from the host's immune system, resulting in the death of the host

R-strain didn't have that protective capsule and was defeated by the host's immune system.



S-trains have been found in blood samples from dead mouse. Thus, the ability to build capsules was transferred from dead *S*-strains to living *R*-strains.

1949 Erwin Chargaff

He discovered that DNA was made of the basis Adenin, Guanin, Thymin and Cytosin. Moreover he observed:

DNA-source	Adenin	Thymin	Guanine	Cytosin
Calf Thymus	1,7	1,6	1,2	1,0
Beef Spleen	1,6	1,5	1,3	1,0
Yeast	1,8	1,9	1,0	1,0
Tubercle Bacillus	1,1	1,0	2,6	2,4

Any Idea?

1949 Erwin Chargaff

He discovered that DNA was made of the basis Adenin, Guanin, Thymin and Cytosin. Moreover he observed:

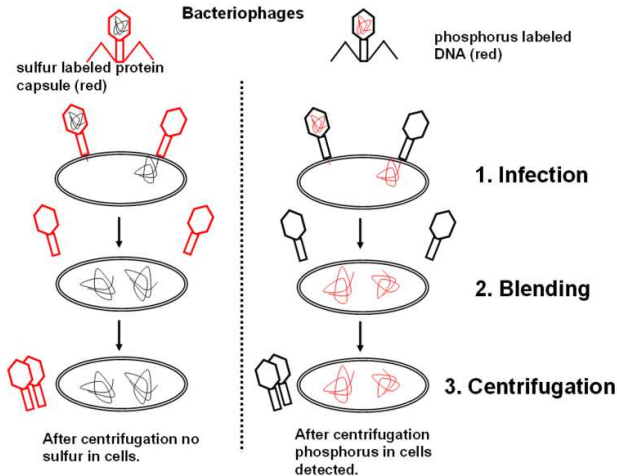
DNA-source	Adenin	Thymin	Guanine	Cytosin
Calf Thymus	1,7	1,6	1,2	1,0
Beef Spleen	1,6	1,5	1,3	1,0
Yeast	1,8	1,9	1,0	1,0
Tubercle Bacillus	1,1	1,0	2,6	2,4

Any Idea?

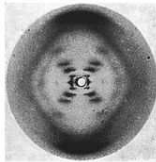
Chargaff's rules: Amounts of adenine and thymine in DNA were roughly the same, as were the amounts of cytosine and guanine. Thus, he assumed that the bases always occure as pairs.

1952 Alfred Hershey und Martha Chase

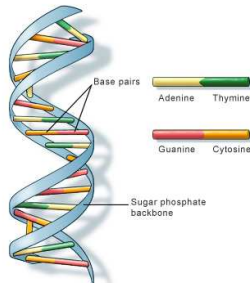
showed that DNA encodes the genetic information.



1952 Rosalind Franklin



1953 Francis Crick and James D. Watson discovered the famous double helical structure of DNA

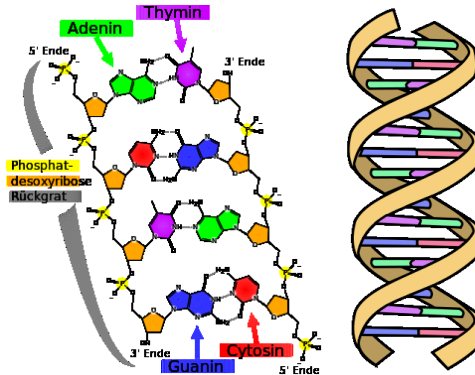


2003 Human Genome Project sequenced the human genome.

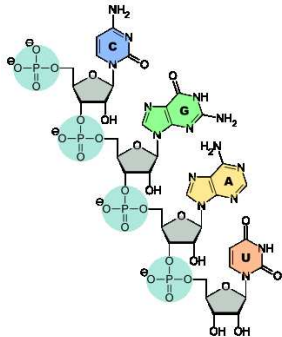
Basic Molecules

- DNA
carries genetic information
- RNA
 - mRNA: convey genetic information from DNA to the ribosome
 - tRNA: linking codons to aminoacids
 - snRNA: splicing
 - microRNA: regulation of gene expression
 - RNA can act as genome (virus)
 - ...
- proteins
perform a vast array of functions within living organisms, including catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another.

DNA (Deoxyribonucleic acid)



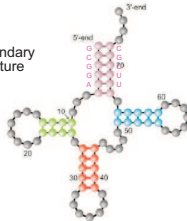
RNA (Ribonucleic acid)



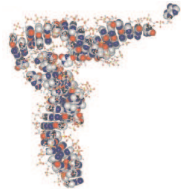
primary structure

5'-end **GCGGAU**UUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAU**CGGAGGUC**CUGUGUUCGAUCCACAGAAU**UUGC**ACCA 3'-end

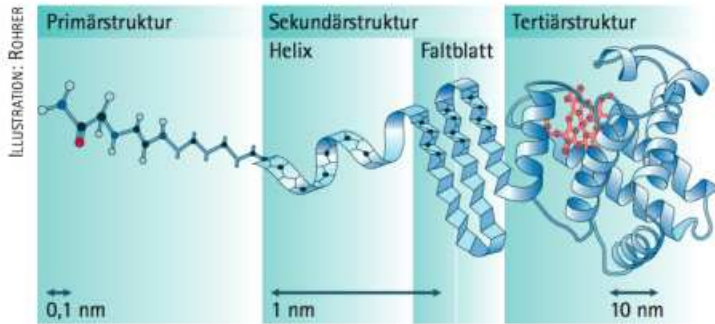
secondary structure



tertiary structure



Proteins



Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)
- RNA (Ribonucleic acid)
- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)
 - double-stranded helices of two polymers
- RNA (Ribonucleic acid)
- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)
 - double-stranded helices of two polymers
 - polymer made of nucleotides+backbone
- RNA (Ribonucleic acid)
- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)
 - double-stranded helices of two polymers
 - polymer made of **nucleotides**+backbone
 - **guanine (G), adenine (A), thymine (T), cytosine (C)**
- RNA (Ribonucleic acid)
- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)
 - double-stranded helices of two polymers
 - polymer made of nucleotides+backbone
 - guanine (G), adenine (A), thymine (T), cytosine (C)
 - alternating sugar (deoxyribose) and phosphat groups (related to phosphoric acid)
nucleotides are attached to sugar
- RNA (Ribonucleic acid)
- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)
 - double-stranded helices of two polymers
 - polymer made of nucleotides+backbone
 - guanine (G), adenine (A), thymine (T), cytosine (C)
 - alternating sugar (deoxyribose) and phosphat groups (related to phosphoric acid)
nucleotides are attached to sugar
 - the nucleotides of two polymers can bind (A-T, C-G)
- RNA (Ribonucleic acid)
- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if

$XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if

$XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

- single-stranded polymer

- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if

$XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

- single-stranded polymer
- polymer made of nucleotides+backbone

- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if

$XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

- single-stranded polymer
- polymer made of nucleotides+backbone
- guanine (G), adenine (A), uracil (U), cytosine (C)

- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if
 $XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

- single-stranded polymer
- polymer made of nucleotides+backbone
- guanine (G), adenine (A), uracil (U), cytosine (C)
- alternating sugar (ribose) and
phospat groups (related to phosphoric acid)
nucleotides are attached to sugar

- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if
 $XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

- single-stranded polymer
- polymer made of nucleotides+backbone
- guanine (G), adenine (A), uracil (U), cytosine (C)
- alternating sugar (ribose) and
phosphat groups (related to phosphoric acid)
nucleotides are attached to sugar
- the nucleotides of polymer can bind (A-U, C-G, G-U)

- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if
 $XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

RNA = single sequence s over the alphabet

$\mathbb{A} = \{A, C, G, U\}$, where $X \in s$ can bind with $Y \in s$ if
 $XY \in \mathbb{B} = \{AU, UA, GC, CG, GU, UG\}$

- Protein

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if
 $XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

RNA = single sequence s over the alphabet

$\mathbb{A} = \{A, C, G, U\}$, where $X \in s$ can bind with $Y \in s$ if
 $XY \in \mathbb{B} = \{AU, UA, GC, CG, GU, UG\}$

- Protein

- large molecule made of amino acids

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if
 $XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

RNA = single sequence s over the alphabet

$\mathbb{A} = \{A, C, G, U\}$, where $X \in s$ can bind with $Y \in s$ if
 $XY \in \mathbb{B} = \{AU, UA, GC, CG, GU, UG\}$

- Protein

- large molecule made of amino acids
- order of amino acids determined by order of genes

Understand Inheritance - Math. Framework

- **DNA (Deoxyribonucleic acid)**

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if
 $XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- **RNA (Ribonucleic acid)**

RNA = single sequence s over the alphabet

$\mathbb{A} = \{A, C, G, U\}$, where $X \in s$ can bind with $Y \in s$ if
 $XY \in \mathbb{B} = \{AU, UA, GC, CG, GU, UG\}$

- **Protein**

- large molecule made of amino acids
- order of amino acids determined by order of genes
- in general, genetic code specifies 20 standard amino acids

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if
 $XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

RNA = single sequence s over the alphabet

$\mathbb{A} = \{A, C, G, U\}$, where $X \in s$ can bind with $Y \in s$ if
 $XY \in \mathbb{B} = \{AU, UA, GC, CG, GU, UG\}$

- Protein

Protein = sequence over the alphabet \mathbb{A} = set of 20 aminoacids

Understand Inheritance - Math. Framework

- DNA (Deoxyribonucleic acid)

DNA = two sequences s_1, s_2 over the alphabet

$\mathbb{A} = \{A, C, G, T\}$, where $X \in s_1$ can bind with $Y \in s_2$ if
 $XY \in \mathbb{B} = \{AT, TA, GC, CG\}$ (base pairing rules)

- RNA (Ribonucleic acid)

RNA = single sequence s over the alphabet

$\mathbb{A} = \{A, C, G, U\}$, where $X \in s$ can bind with $Y \in s$ if
 $XY \in \mathbb{B} = \{AU, UA, GC, CG, GU, UG\}$

- Protein

Protein = sequence over the alphabet \mathbb{A} = set of 20 aminoacids

What is the genetic code?

How is the information on DNA used to code proteins?

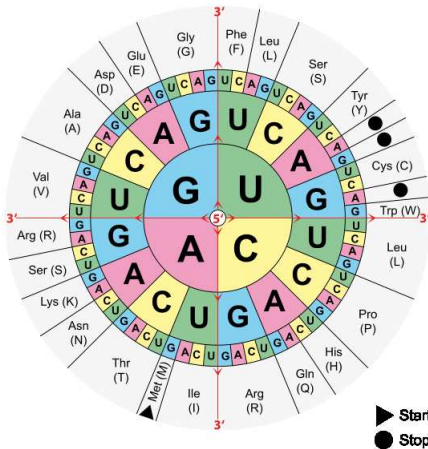
Some More History: Cracking the genetic code - The magic number 20

Question: How can a 4-letter alphabet code for 20 aminoacids?

- Garmov - Diamond Code
- Crick - Non-Overlapping Commafree Code
- Nirenberg - Matthaei - Experiment

→ board

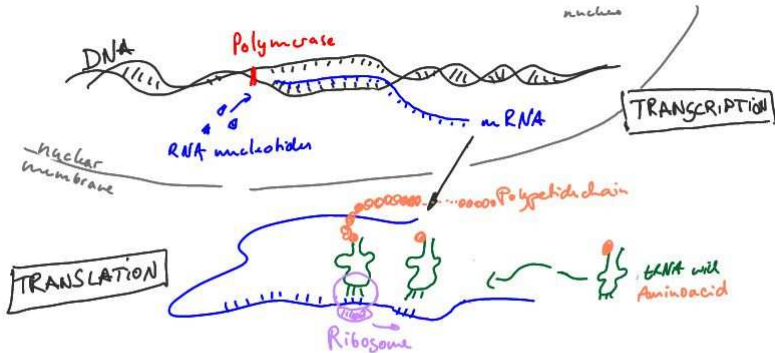
Genetic code is simply a map $f : C \rightarrow A$ where,
 $C = \{(x_1 x_2 x_3) \mid x_i \in \{A, C, G, U\}\}$ and
 $A =$ set of aminoacids and start/termination codon.



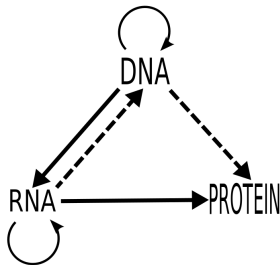
Amino acid	Genetic code	Abbr
Alanine	GCA GCC GCG GCU	Ala
Arginine	AGA AGG CGA CGC CGG CGU	Arg
Asparagine	AAC AAU	Asn
Aspartic acid	GAC GAU	Asp
Cysteine	UGC UGU	Cys
Glutamine	CAA CAG	Gln
Glutamic acid	GAA GAG	Glu
Glycine	GGA GGC GGG GGU	Gly
Histidine	CAC CAU	His
Isoleucine	AUA AUC AUU	Ile
Leucine	CUA CUC CUG CUU UUA UUG	Leu
Lysine	AAA AAG	Lys
Methionine	AUG	Met
Phenylalanine	UUC UUU	Phe
Proline	CCA CCC CCG CCU	Pro
Serine	AGC AGU UCA UCC UCG UCU	Ser
Threonine	ACA ACC ACG ACU	Thr
Tryptophan	UGG	Try
Tyrosine	UAC UAU	Tyr
Valine	GUA GUC GUG GUU	Val
STOP sign	UAA UAG UGA	

¹ picture taken from http://commons.wikimedia.org/wiki/File:Aminoacids_table.svg

Protein Synthesis - what we know now nowadays



Central Dogma - what we know now nowadays²



DNA->DNA

DNA Replication

DNA->RNA

Transcription

RNA->Protein

Translation

RNA->DNA

Reverse Transcription

(e.g. eukaryotes or
retroviruses (as HIV))

RNA->RNA

RNA replication

(e.g. in many viruses)

DNA->Protein

Direct Translation

(in vitro)

²picture taken from

http://commons.wikimedia.org/wiki/File:Crick's_1958_central_dogma.svg

What is Bioinformatics?

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to study and process biological data.

(Wikipedia)

Huge amount of data (genomic material) available, but

Data \neq Knowledge

Question: How to analyse data, how to integrate data, how to get information out of data and which information?

What is Bioinformatics?

Exmpl: Human genome is a string of length $\simeq 3.200.000.000$

However, (parts of) this sequence must be interpreted to get a biological meaning.

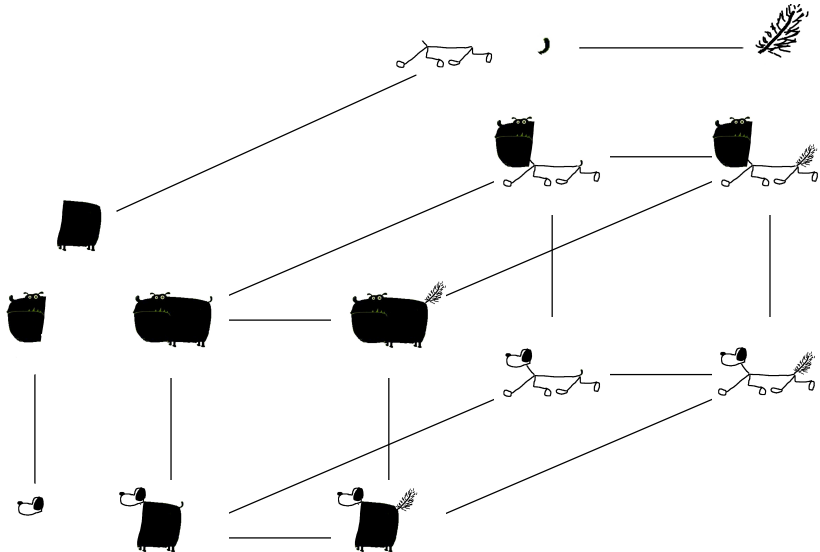
- Find out the sequence of genomes and what does it tell us?
Which parts code for proteins or enzymes?
- Predict structure of RNA or proteins (and thus, determine their function).
- Find out the differences between the genes of different species e.g. to reconstruct the evolutionary history of species based on genomic data or to determine “real” characters that describe a phenotype.
- What does the difference of the “activity” of genes between healthy and unhealthy patients tell us about diseases and possible treatments?
- How to efficiently store and recover all that information from databases?

Bioinformatics

Following outline follows isn't fixed nor it is clear if we have enough time to treat all the things! But we will try it!

1. Warm Up + Cracking the genetic code
2. Basics
 - 2.1 Shotgun sequencing and Factor-Of-Four Approx. Alg.
 - 2.2 String Finding Methods
 - 2.3 Alignments
 - 2.4 Graphs
3. RNA structures
4. Phenotypespaces and Graph Products
5. Phylogenomics

Phenotypespaces and Graphproducts



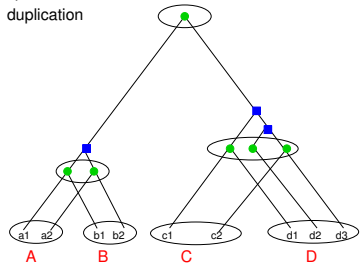
5'-end **GCGGAU**UUAG**GCUC**AGUUGG**GAGCG****CCAGAC**UGAAGA**UCUGG**AGGUC**CUGUG**UUCGAUC**CACAGA****AUUCGCACCA** 3'-end



Phylogenetics - The Evolutionary History of Genes and Species



- speciation
- duplication



Literature

- "Introduction to Computational Biology: Maps, Sequences and Genomes", Michael S. Waterman
- "Understanding Bioinformatics", Marketa J. Zvelebil
- "Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology", Dan Gusfield
- "RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods" Gorodkin, Jan, Ruzzo, Walter L. (Eds.)
- "Phylogenetics", Charles Semple and Mike Steel
- "Handbook of Product Graphs, Second Edition (Discrete Mathematics and Its Applications)", Richard Hammack, Wilfried Imrich and Sandi Klavzar