Selected and Adapted from:
Data Mining:
Concepts and Techniques

— Chapter 8 —

8.4. Mining sequence patterns in biological data

Jiawei Han and Micheline Kamber
Department of Computer Science
University of Illinois at Urbana-Champaign
www.cs.uiuc.edu/~hanj

Mining Sequence Patterns in Biological Data

- A brief introduction to biology and bioinformatics
- Alignment of biological sequences
- GWAS Mining
- Summary

The Structure of DNA

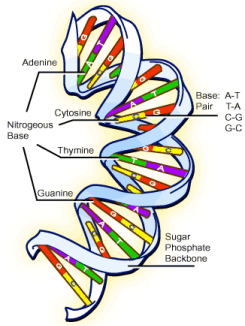
Rosalind Franklin, James D. Watson, Francis Crick (1953)

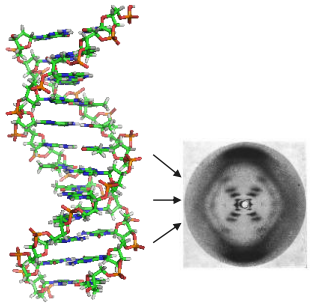
Nucleotides (bases)

- Adenine (A)
- Cytosine (C)
- Guanine (G)
- Thymine (T)

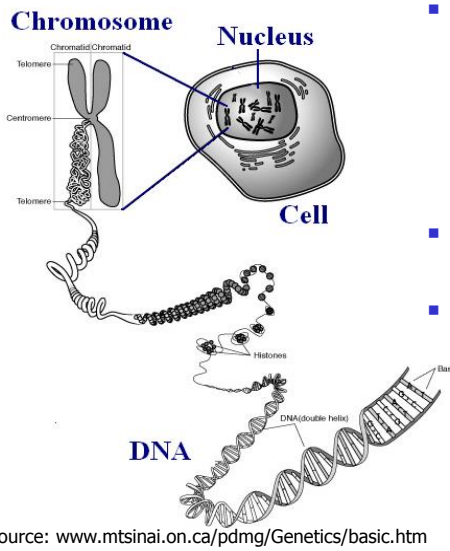
Complementary Binding:

- T – A
- A – T
- C – G
- G - C





Genes (Eukaryote)

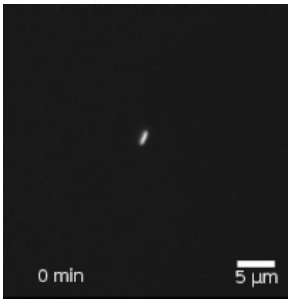


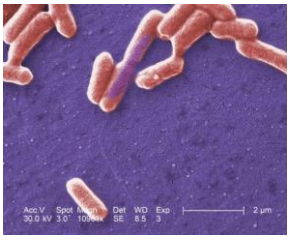
Source: www.mtsinai.on.ca/pdmg/Genetics/basic.htm

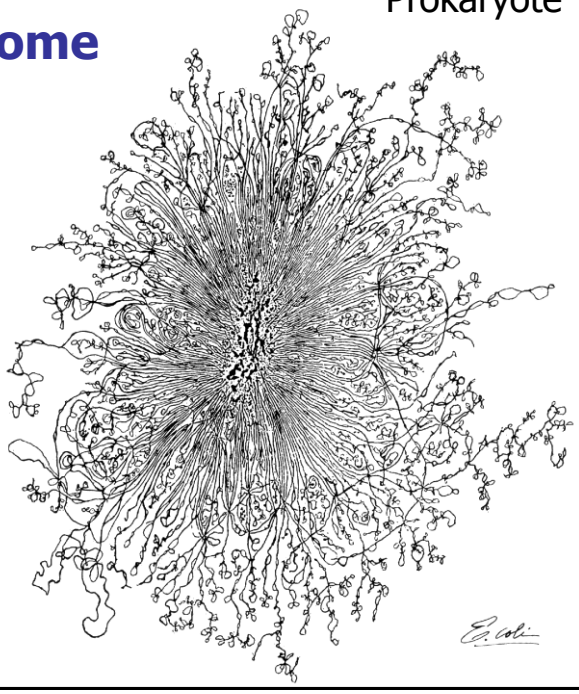
- Gene: Contiguous subparts of single strand DNA that are templates for producing *proteins*. Genes can appear in either of the DNA strand.
  - Chromosomes: compact chains of coiled DNA
- Genome: The *set of all genes* in a given organism.
- *Noncoding* part: The function of DNA material between genes is largely unknown. Certain intergenic regions of DNA are known to play a major role in *cell regulation* (controls the production of proteins and their possible interactions with DNA).

E.Coli Genome

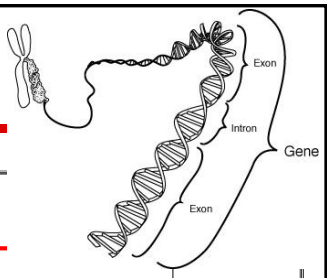
Prokaryote

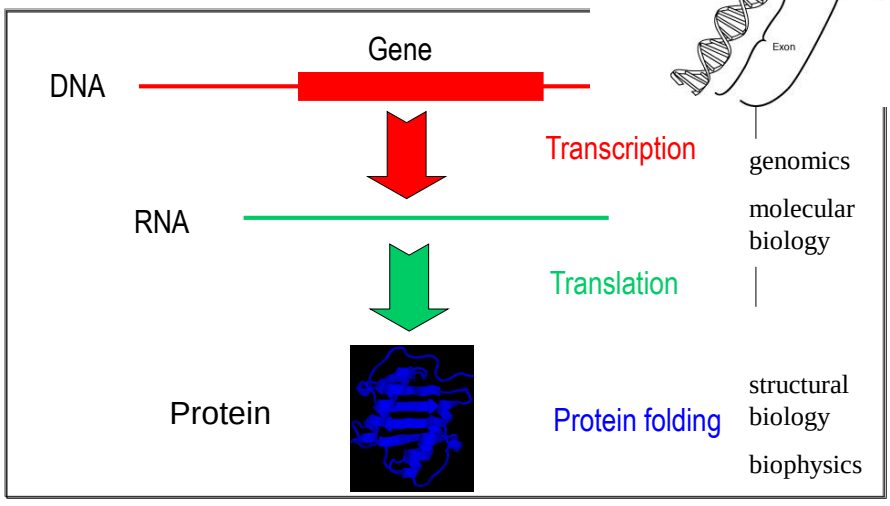






Biological Information: From Genes to Proteins



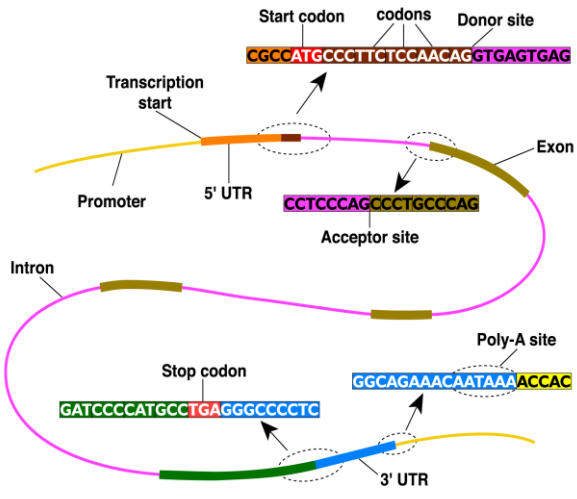


1. Transcription of TCAG to UCAG
2. Splicing (exons, introns)
3. Translation of triplets (codons) to Amino Acids
   Open Reading Frame (ORF): **<start>**<codon><codon>…**<end>** -> KNRS…

Second base

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

Eukaryotes



From DNA to Protein

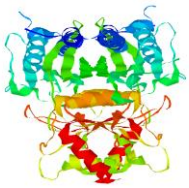
http://www.yourgenome.org/video

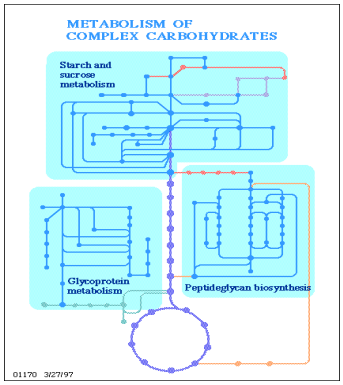
From Amino Acids to Proteins Functions

```
CGCCAGCTGGACGGGCACACC
ATGAGGCTGCTGACCCTCCTG
GGCCTTCTG..
```

↓

```
TDQAAFDTNIVTLTRFVMEQG
RKARGTGEMTQLLNSLCTAVK
AISTAVRKAGIAHLYGIAGST
NVTGDQVKKLDVLSNDLVINV
LKSSFATCVLVTEEDKNAIIV
EPEKRGKYVVCFDPLDGSSNI
DCLVSIGTIFGIYRKNSTDEP
SEKDALQPGRNLVAAGYALYG
SATML
```



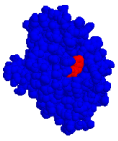


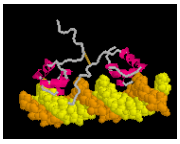
DNA / **amino acid sequence** | **3D structure** | **protein functions**

DNA (gene) →→→ pre-RNA →→→ RNA →→→ Protein

RNA-polymerase *Spliceosome* *Ribosome*

Molecule of the Month www.pdb.org

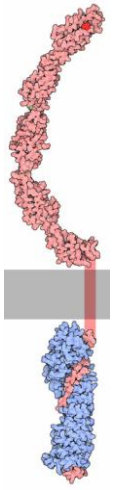




Animated gifs from: proteinexplorer.org

March 2008:

Cadherin
- Adhesive Proteins
- Selective Stickiness: The red tyrosine amino acid will bind to Cadherins on neighbouring cells

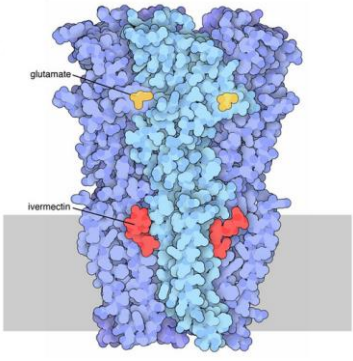


Molecule of the Month www.pdb.org

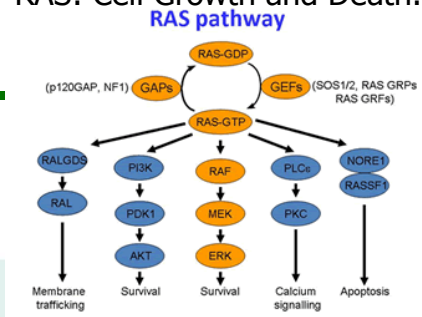
November 2015:

Glutamate-gated Chloride Receptors

- Receptors of Chloride Ion channels in nerve systems of parasites like worms
- Targets for antibiotics, because our own cells don't use them

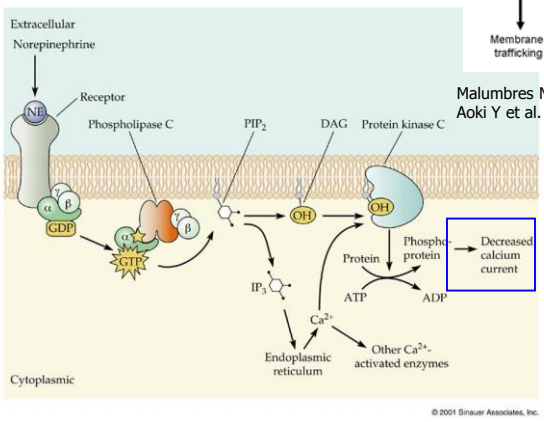


RAS: Cell Growth and Death.

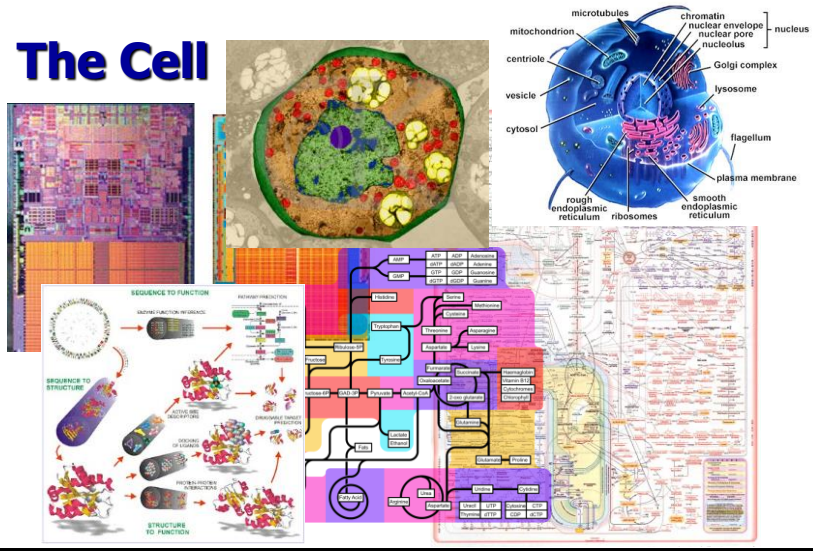


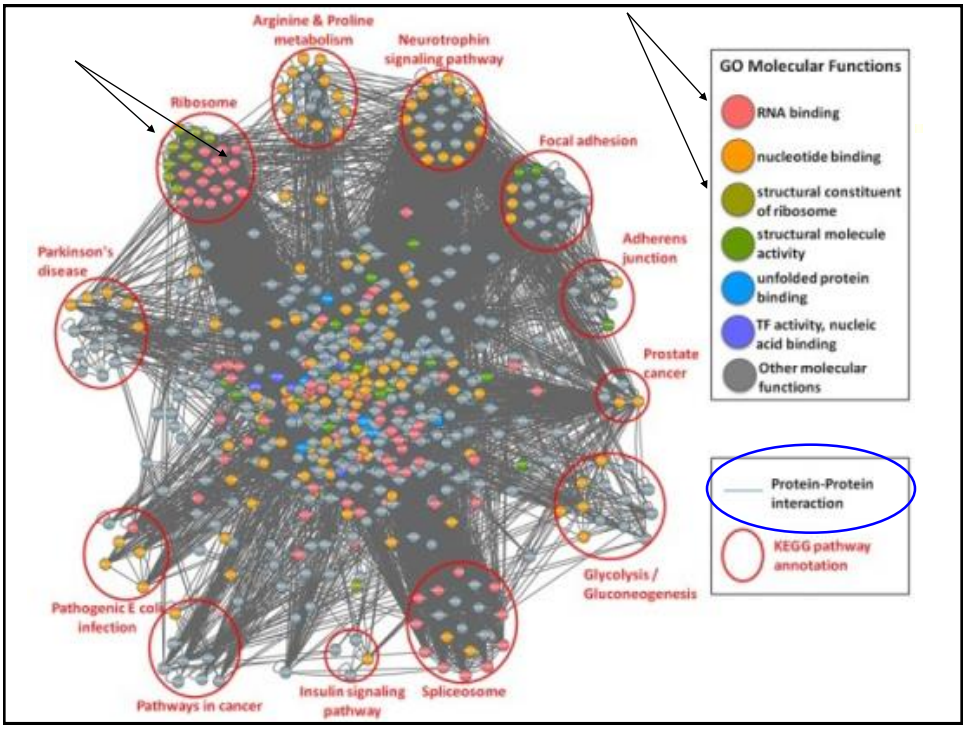
Malumbres M, Barbacid M. Nat Rev Cancer. 6: 459-65, 2003.
Aoki Y et al. Hum Mutat. 29: 992-1006, 2008

Neurosystem



The Cell





Sequencing

1953 Watson and Crick: the structure of the DNA molecule

- DNA carrier of the genetic information, the challenge of reading the DNA sequence became central to biological research.
- methods for DNA sequencing were extremely inefficient, laborious and costly.

1965 Holley: reliably sequencing the yeast gene for $tRNA^{Ala}$ required the equivalent of a full year's work per person per base pair (bp) sequenced (1bp/person year).

1970 Two classical methods for sequencing DNA fragments by Sanger and Gilbert.

Sequencing

1980 In the 1980s methods were augmented by
- partial automation
- the cloning method, which allowed fast and exponential replication of a DNA fragment.

1990 Start of the human genome project: sequencing efficiency reached 200,000 bp/person-year.

2002 End of the human genome project: 50,000,000 bp/person-year. (Total cost summed to $3 billion.)

Note: - Moore's Law doubling transistor count every 2 years
- Here: doubling of #base pairs/person-year every 1.5 years

Bio-Data Sequencing

Recently (2007 – 2011/2012):
- new sequencing technologies next generation sequencing (NGS) or deep sequencing
  - reliable sequencing of 100×10^9 bp/person-year.
  - NGS now allows compiling the full DNA sequence of a person for ~ $10,000
  - within 3-5 years cost is expected to drop to ~$1000.

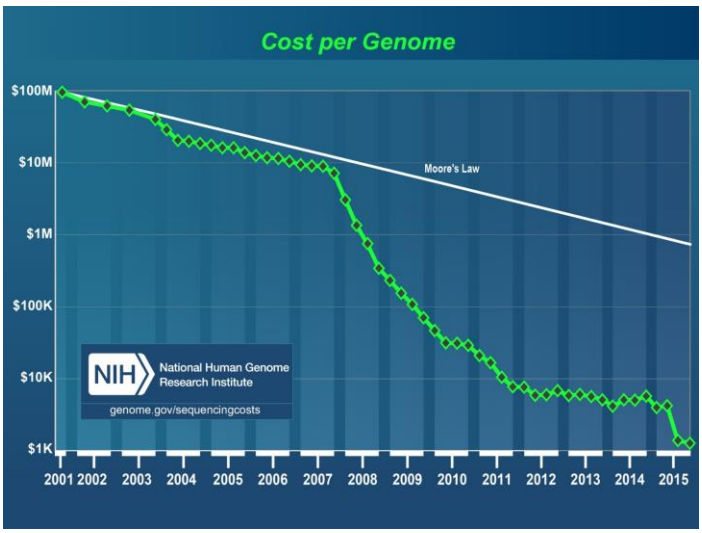
2017:
- *Veritas Genetics* uses a **Illumina HiSeq X Ten system** with an average coverage depth of 30X. Whole genome sequencing in 8-10 weeks at a price of $999.
- (See http://www.nanalyze.com/2016/03/does-full-genome-sequencing-really-cost-1000-now/)

2018: HiSeq's successor NovaSeq 6000 6Tb <2 days (150b reads)

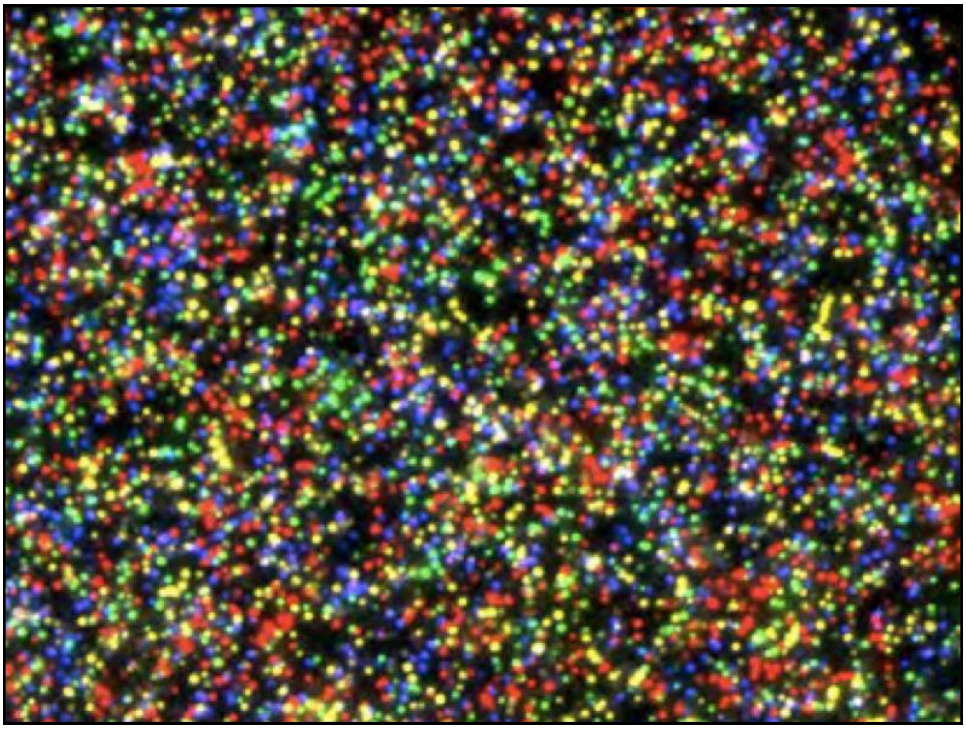
K. Schwarze, J. Buchanan, J.C. Taylor, and S. Wordsworth, Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the Literature, Genetics in Medicine, Volume 20, Number 10, October 2018.

"Thirty-six studies met our inclusion criteria. These publications investigated the use of whole exome sequencing (WES) and whole-genome sequencing (WGS) in a variety of genetic conditions in clinical practice, the most common being neurological or neurodevelopmental disorders. Study sample size varied from a single child to 2,000 patients.
Cost estimates for a single test ranged from
$555 to $5,169 for WES and from
$1,906 to $24,810 for WGS.
Few cost analyses presented data transparently and many publications did not state which components were included in cost estimates."

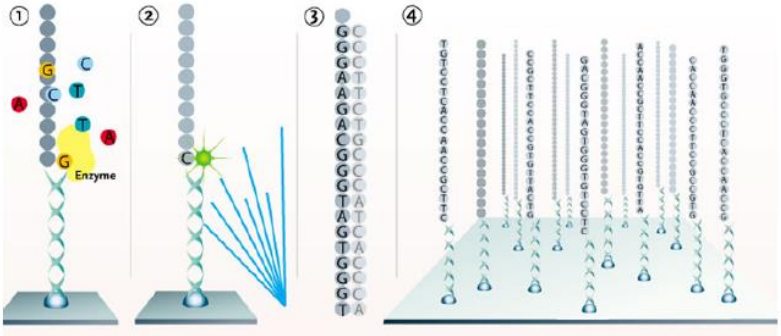
Soon: 100$ Full Genome scan!



From: https://www.genome.gov/sequencingcosts/



Next Generation Sequencing Technologies

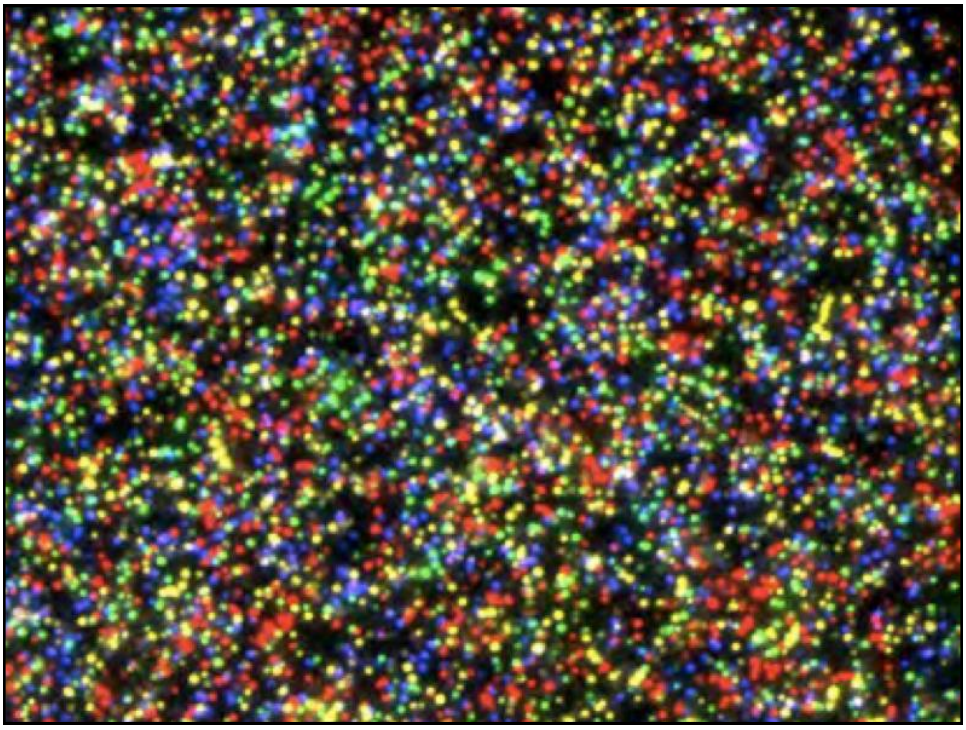


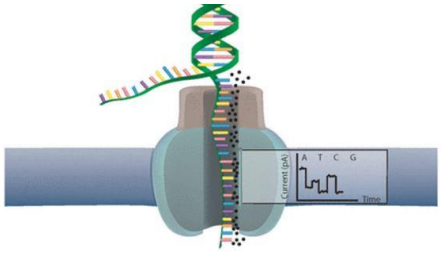
(1) A modified nucleotide is added to the complementary DNA strand by a DNA polymerase enzyme.

(2) A laser is used to obtain a read of the nucleotide just added.

(3) The full sequence of a fragment thus determined through successive iterations of the process.

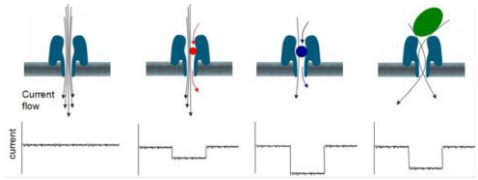
(4) A visualization of the matrix where fragment clusters are attached to flow cells.



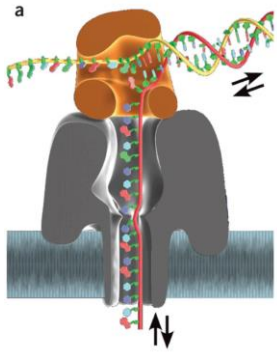

A T C G
Current (pA)
Time


MinION

MinION: nanopores



Phys.org



- Read length ~10kb
- Longest reported 230 – 300 kb
- 500 bps per pore

Oxford NANOPORE Technologies

Coming soon

Flongle

- An adapter for MinION for smaller tests or experiments
- Single-use, on-demand, cost efficient sequencing
- Suitable for quality checks, amplicons, smaller genomes, targeted regions, or those interested in diagnostics/other tests
- MinIT available to support IT/software needs

GridION X5

- Multiple sequencing devices, one compute module
- Use up to five MinION Flow Cells at a time
- Benchtop processor capable of handling high data volumes in real time
- Rapid, real-time applications such as *Read Until* ...

About GridION | Buy GridION

Choose GridION X5 if you:

- would like to offer nanopore sequencing as a service
- want the choice to invest from a CapEx or consumable budget
- work on larger sequencing projects (50–100Gb per 48 hrs)
- would like on-device basecalling – no local infrastructure requirement.

PromethION

- High-throughput, high-sample number benchtop system
- Modular: Up to 48 flow cells, each with up to 3,000 nanopore channels (total up to 144,000)
- Flow cells may be run individually or concurrently
- Same workflow as MinION at larger scale

About PromethION | Early Access

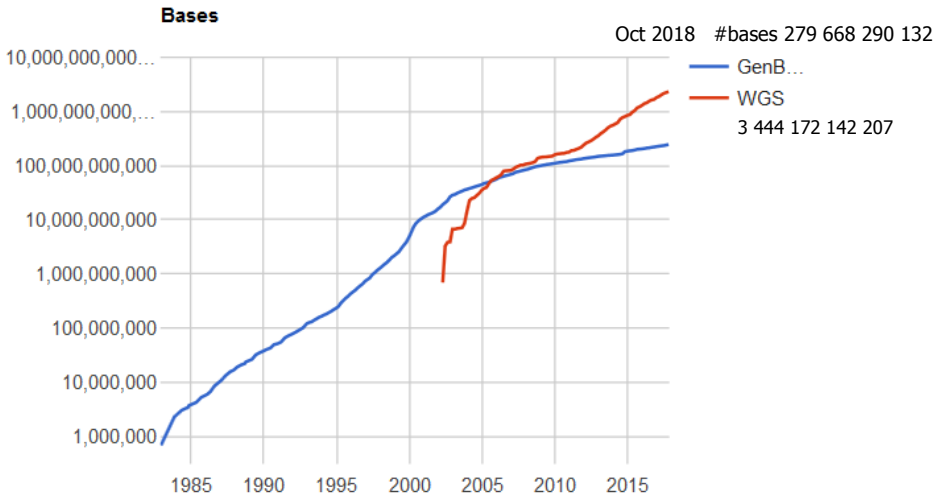
Choose PromethION if you:

- would like to offer nanopore sequencing as a service
- are interested in very large data volumes projects (Tb)
- are seeking on-demand sequencing for large numbers of samples
- would like to avoid CapEx investments.

SmidgION

- Designed to be our smallest sequencing device so far
- Same nanopore sensing technology as MinION and PromethION
- Designed for use with a smartphone in any location

NCBI Genbank



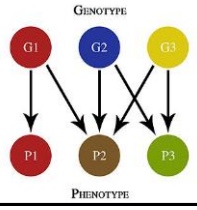
Bio-Sequence Data Mining

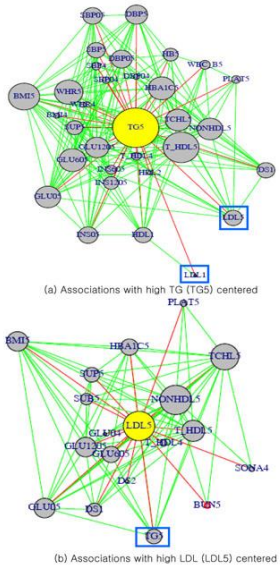
- Sequential Pattern Mining Algorithms: Apriori based approaches for finding **frequent sequences**; SPADE (Zaki, 2001)
- Structured Pattern Mining Algorithms: Pattern Growth methods (>2004); **HMM-based method** VOGUE (Zaki, 2011); Gab-Bide, Gab-Connect (Li, 2012)
- Bio Data Mining Algorithms: Perfect Tandem Repeats, Approximate Tandem Repeats, STAR (Delgrange et al., 2004); Longest Pattern Repetitions; Multiple Sequence Pattern Mining: **Prefix Trees, Hashing**

See: N. Shaji, S. Izudheen, Bio Sequence Data Mining: A Survey, Asian Journal of Computer Science and Information Technology 4: 3(2014) 21 - 24.

S.H. Park et al., A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes, BMC Syst. Biol., 2011.

- **Mine pleiotropic genes**
- Data mining approach to discover patterns of multiple phenotypic associations in 350K variants in 8512 individuals
- analytical scheme for GWAS of multivariate phenotypes defined by the discovered patterns.
- Association rule mining using Apriori (1% support, confidence 10%)
- Rules evaluated with lift.





Trait	:	Description
TCHL	:	Total Cholesterol
HDL	:	High Density Lipoprotein Cholesterol
LDL	:	Low Density Lipoprotein Cholesterol
TG	:	Triglyceride
T_HDL	:	Total Cholesterol/hdl
NONHDL	:	Total Cholesterol-hdl
SBP	:	Systolic Blood Pressure (lying position)
SBP0	:	Systolic Blood Pressure (sitting position)
DBP	:	Diastolic Blood Pressure (lying position)
DBP0	:	Diastolic Blood Pressure (sitting position)
INS0	:	FastingBloodInsulin
INS60	:	BloodInsulin(OGTTafter60min)
INS120	:	BloodInsulin(OGTTafter120min)
GLU0	:	Fasting Blood Glucose
GLU60	:	Blood Glucose (OGTT after 60 min)
GLU120	:	Blood Glucose (OGTT after 120 min)
HBA1C	:	Glycosylated Hemoglobin
HB	:	Hemeglobin (HBi)
WBC_B	:	White Blood Cells
PLAT	:	Platlet
BUN	:	Blood Urea Nitrogen
SONA	:	Sodium in Urine
BMI	:	Body Mass Index
WHR	:	Waist Hip Ratio
DS	:	Distal Radius SOS for bone density
SUB	:	Subscapular
SUP	:	Suprailiac

Bio-Sequence Data Mining

- R. Alves, et al., Gene association analysis: a survey of frequent pattern mining from gene expression data, Briefings in Bioinformatics, Vol II. No 2., 210-224, 2009.
- Q. Zhang et al., AprioriGWAS, a New Pattern Mining Strategy for Detecting Genetic Variants Associated with Disease through Interaction Effects, Computational Biology, 2014.

Bio-Data Data Mining

- Mining Frequent Patterns from Sequences with Wild Cards: MAIL (Fei Xie etr al, 2010)
- **Bio Sequencing in Healthcare:** Genome Wide Association Studies (GWAS) See: N. Shaji, S. Izudheen, Bio Sequence Data Mining: A Survey, Asian Journal of Computer Science and Information Technology 4: 3(2014) 21 - 24.
- Mining Association Rules from Gene Ontology and Protein Networks: Promises and Challenges. P.H. Guzzi et al. ICCS 2014. 14th International Conference on Computational Science
- Genome Ontology Mining: A Survey by K.S. Lakshimi and G. Vadivu Int. J. Pharm. Sci. Rev. Res., 43(1), March - April 2017; Article No. 15, Pages: 65-69

Mining Sequence Patterns in Biological Data

- A brief introduction to biology and bioinformatics
- Alignment of biological sequences
- GWAS Mining
- Summary

Evolution: The Tree of Life

From: wikipedia

Phylogenetic trees:

- Traditional: morphology

REVIEW

RNA Biology 11:3, 176–185; March 2014; © 2014 Landes Bioscience

Molecular phylogenetics before sequences

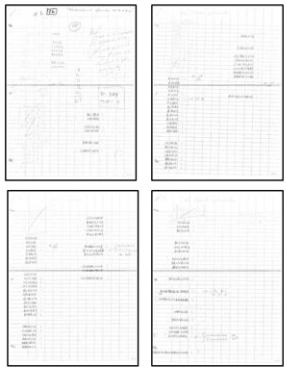
Oligonucleotide catalogs as *k*-mer spectra

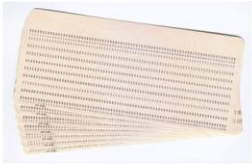
Mark A Ragan*, Guillaume Bernard, and Cheong Xin Chan

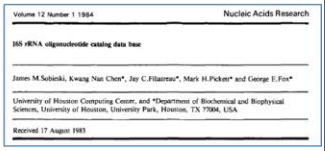
Institute for Molecular Bioscience, and ARC Centre of Excellence in Bioinformatics; The University of Queensland; Brisbane, QLD, Australia

Carl Woese – photo by Ken Luehrsen

Ribosomal RNA (1964)







Volume 12 Number 1 1984 Nucleic Acids Research

16S rRNA oligonucleotide catalog data base

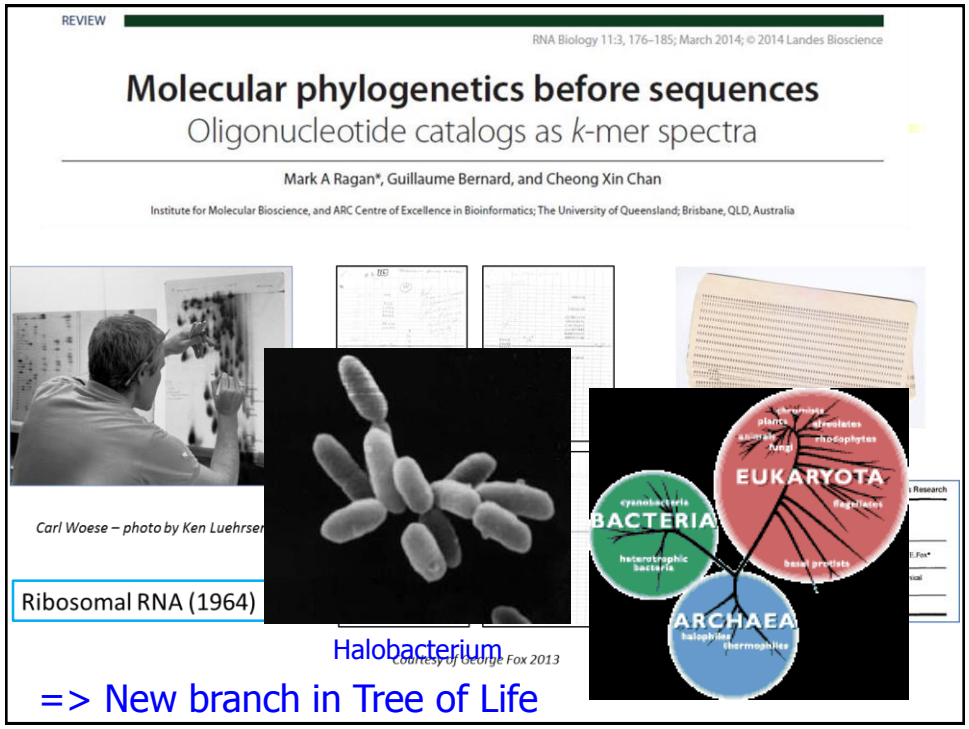
James M.Sobieski, Kwang Nan Chen*, Jay C.Filiatreau*, Mark H.Pickett* and George E.Fox*

University of Houston Computing Centre, and *Department of Biochemical and Biophysical Sciences, University of Houston, University Park, Houston, TX 77004, USA

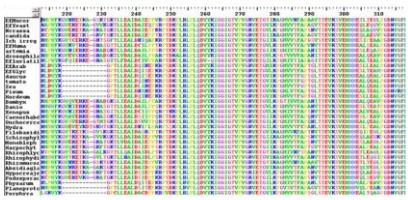
Received 17 August 1983

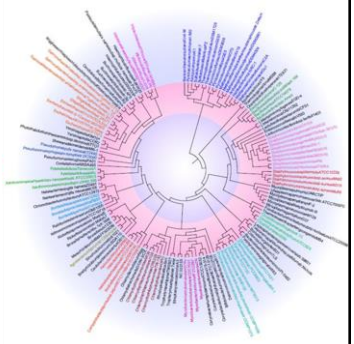
Courtesy of George Fox 2013

=> New branch in Tree of Life



Sequence Alignment: Many different algorithms



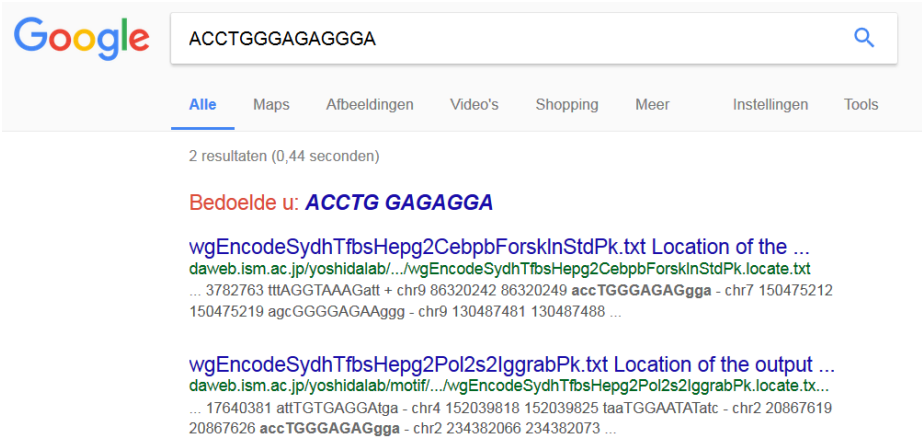


- Optimal Alignment
- BLAST, FASTA (heuristics)
- DNA Sequences
- Protein Sequences
- Multiple Alignment => ancestry of sequences

Sequence Alignment: Problem Definition

- Goal:
  - Given two or more input sequences
  - Identify similar sequences with long conserved subsequences
- Method:
  - Use substitution matrices (probabilities of substitutions of nucleotides or amino-acids and probabilities of insertions and deletions)
  - *Optimal* alignment problem for multiple sequence alignment: NP-hard
  - Heuristic method to find *good* alignments

Search in DNA Sequences



Tekst-Search: Knuth-Morris-Pratt

```
Naive:  Tekst: neeneeneeneeeeneneeeneeneneeneneen
        Patroon: neeneeen
                 neeneeen
                  neeneeen
                   neeneeen
Better: Tekst: neeneeneeneeeeneneeeneeneneeneneen
     Patroon: neeneeen
                  neeneeen
                     neeneeen
```

Errors in text

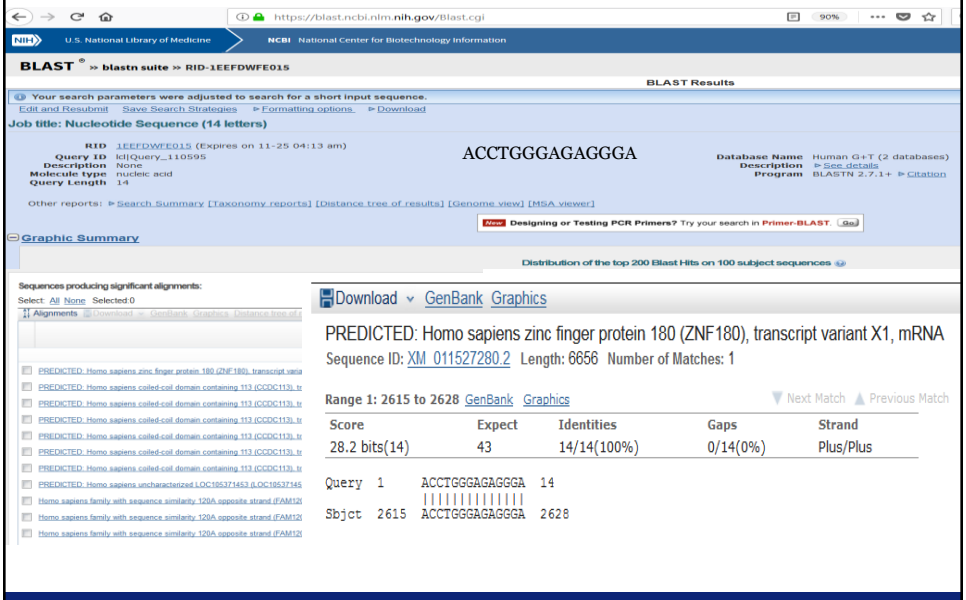
ER IS EEN FOUTJE IN DEZE TEKST GESLOPEN

ER IA EEN GOUTJE IN DEZE TELST GEKLOPEN

ER IS EN FOUTJE IN DE TEKST GESLOPEN

ER IS EEN VOUDJE IN DEZEN TEKST GESLOBEN

ER IS EEN EEN FOUTJE IN DEZE TEKST GESLOPEN



Comparison of DNA Sequences

Input:

S: "ACGCTTTG"

T: "CATGTAT"

Alignments:

S': AC--GCTTTG

T': -CATG-TAT-

S": ACGCTTTG-

T": -CATG-TAT

S"': ACGCTTTG-------

T"': --------CATGTAT

Pair-Wise Sequence Alignment

- Example

```
HEAGAWGHEE
PAWHEAE
```

```
HEAGAWGHE-E
  |   |  ||  |
P-A--W-HEAE
```

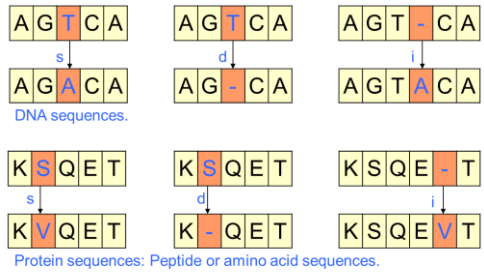
```
HEAGAWGHE-E
     ||  ||  |
--P-AW-HEAE
```

  - Which one is better? → Scoring alignments
- To compare two sequence alignments, calculate a score
  - PAM (Percent Accepted Mutation) or BLOSUM (Blocks Substitution Matrix) (*substitution*) matrices: Calculate matches and mismatches, considering amino acid substitution
  - Gap penalty: Initiating a gap
  - Gap extension penalty: Extending a gap

Comparing Protein Sequences

The BLOSUM62 scoring matrix:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4



s = substitution; d = deletion; i = insertion;

Pair-Wise Sequence Alignment: Scoring Matrix

Scoring Matrix (partly):

	A	E	G	H	W
A	5	-1	0	-2	-3
E	-1	6	-3	0	-3
H	-2	0	-2	10	-3
P	-1	-1	-2	-2	-4
W	-3	-3	-3	-3	15

- Gap penalty*: -8
- Gap extension: -8

```
HEAGAWGHE-E
    ||  ||  |
--P-AW-HEAE
```

```
(-8) + (-8) + (-1) + (-8) + 5 + 15 +
(-8) + 10 + 6 + (-8) + 6 = -41 + 42 = 1
```

Exercise: Calculate for

```
HEAGAWGHE-E
  |   |  ||  |
P-A--W-HEAE
```

*) Here for opening the gap and its first location

45

Formal Description

- *Problem:* **PairSeqAlign**
- *Input:* Two sequences *x, y*
  - Scoring matrix *s*
  - Gap penalty *d*
  - Gap extension penalty *e*
- *Output:* The optimal sequence alignment
- Difficulty:

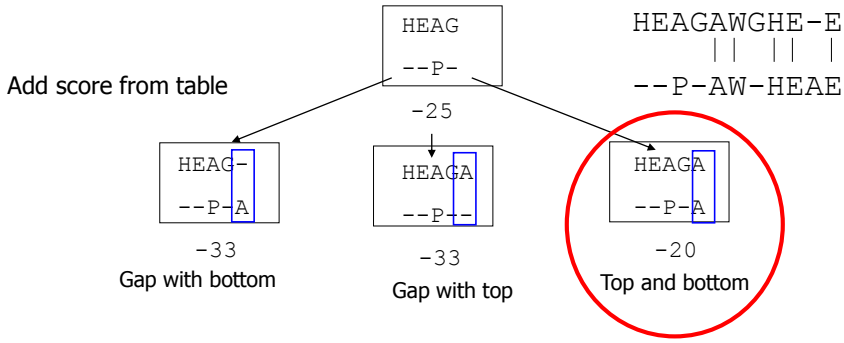
If x, y are of size n then the number of possible global alignments is →

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{(\pi n)}}$$

46

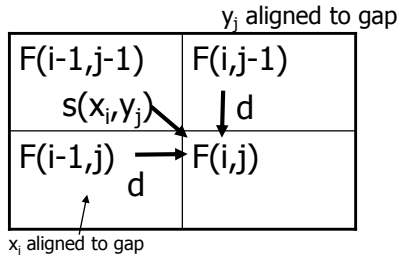
Global Alignment: Needleman-Wunsch

- Needleman-Wunsch Algorithm (1970)
  - Uses weights for the outmost edges that encourage the best overall (global) alignment
- Smith-Waterman
  - Local Alignment algorithm: (favors the contiguity of segments being aligned)
- Idea: Build up optimal alignment from optimal alignments of subsequences



Global Alignment

- Uses recursion to fill in intermediate results table
- Uses O(nm) space and time
  - $O(n^2)$ algorithm
  - Feasible for moderate sized sequences, but not for aligning whole genomes.



While building the table, keep track of where optimal score came from, reverse arrows

Comparison of DNA Sequences

```
C A T G T
| | | | |
A C G T G
```

	-	C	A	T	G	T
-						
A		X				
C			X			
G				X		
T					X	
G						X

Comparison of DNA Sequences

```
C A T G T -
| | | | | |
- A C G T G
```

	-	C	A	T	G	T
-		-				
A			M			
C				X		
G					M	
T						M
G						-

Comparison of DNA Sequences

```
C A T G T - -
| | | | | | |
- - A C G T G
```

	-	C	A	T	G	T
-		-	-			
A				X		
C					X	
G						X
T						-
G						-

Comparison of DNA Sequences

```
- C A T G T
| | | | | |
A C G T G -
```

	-	C	A	T	G	T
-						
A	-					
C		M				
G			X			
T				M		
G					M	-

Comparison of DNA Sequences

```
- - C A T G T
| | | | | | |
A C G T G - -
```

	-	C	A	T	G	T
-						
A	-					
C	-					
G		X				
T			X			
G				X	-	-

Comparison of DNA Sequences

```
C A - - T G T
| | | | | | |
- A C G T G -
```

	-	C	A	T	G	T
-		-				
A			M			
C			-			
G			-			
T				M		
G					M	-

Comparison of DNA Sequences

```
C A T G T - -
| | | | | | |
- - A C G T G
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2			
A				-3		
C					-4	
G						-5
T						-6
G						-7

Comparison of DNA Sequences

```
C A - - T G T
| | | | | | |
- A C G T G -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1				
A			1			
C			0			
G			-1			
T				1		
G					3	2

Comparison of DNA Sequences

```
C A T G T - - - - -
| | | | | | | | | |
- - - - - A C G T G
```

```
- - - - - C A T G T
| | | | | | | | | |
A C G T G - - - - -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1					
C	-2					
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
-1
   C ATGT
   |
   A CGTG
```

```
-2
C- ATGT
||
-A CGTG
```

```
   -C ATGT
-2 ||
   A- CGTG
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1 -2 -2				
C	-2					
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
C A T G T
| |
- A C G T G
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1 -3 -2			
C	-2					
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
C A T T G T
| | |
- A - G T G
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	-3 -4 0		
C	-2					
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
C A - - T G T
| | | | | | |
- A C G T G -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	0		
C	-2					
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
C A - - T G T
| | | | | | |
- A C G T G -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	
C	-2					
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
C A - - T G T
| | | | | | |
- A C G T G -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2					
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
C A - - T G T
| | | | | | |
- A C G T G -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1 -2 -3				
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
C A - - T G T
| | | | | | |
- A C G T G -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
G	-3					
T	-4					
G	-5					

Comparison of DNA Sequences

```
C A - - T G T
| | | | | | |
- A C G T G -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
G	-3	0	0	-1	2	1
T	-4	-1	-1	2	1	4
G	-5	-2				

Comparison of DNA Sequences

```
C A T G T
| | | | |
A C G T G
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
G	-3	0	0	-1	2	1
T	-4	-1	-1	2	1	4
G	-5	-2	-2	1	4	3

Comparison of DNA Sequences

An OPTIMAL solution with Dynamic Programming:

```
- C A T G T
| | | | | |
A C G T G -
```

Costs:
- Gap: -1
- Mismatch: -1
- Match: 2

	-	C	A	T	G	T
-	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
G	-3	0	0	-1	2	1
T	-4	-1	-1	2	1	4
G	-5	-2	-2	1	4	3

Global Alignment
Dynamic Programming

Sequence S -acgctg

Sequence T catg-t-

acgctg-
-ca-tgt

acgctg-
-c-atgt

Sequence S

⟨acgctg
catgt⟩

Sequence T

	-	c	a	t	g	t
-	0	-1	-2	-3	-4	-5
a	-1	-1	1	0	-1	-2
c	-2	1	0	0	-1	-2
g	-3	0	0	-1	2	1
c	-4	-1	-1	-1	1	1
t	-5	-2	-2	1	0	3
g	-6	-3	-3	0	3	2

Trace back to obtain an optimal global alignment.
Note that, here three such optimal global alignments exist.

Pair-Wise Sequence Alignment

Given $s(x_i, y_j)$, d

$F(0,0) = 0$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

Global Alignment:
F(0,0) – F(n,m)

Given $s(x_i, y_j)$, d

$F(0,0) = 0$

$$F(i,j) = \max \begin{cases} 0 \\ F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

Local Alignment: 0 – F(i,j)

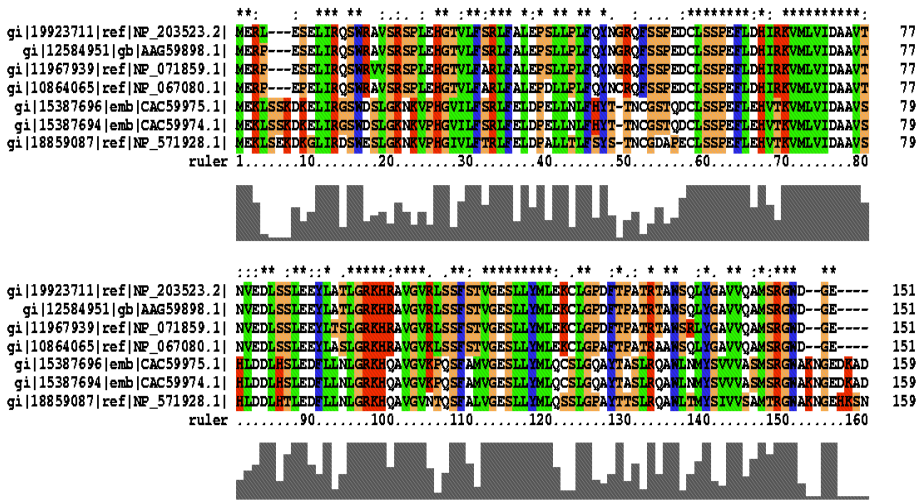
We can vary both the model and the alignment strategies

Heuristic Alignment Algorithms

- Motivation: Complexity of alignment algorithms: O(nm)
  - Protein DB: more than 100 million base pairs
  - Matching each sequence with a 1,000 base pair query takes too long…
- Heuristic algorithms aim at speeding up at the price of possibly missing the best scoring alignment
- Two well known programs
  - BLAST: Basic Local Alignment Search Tool
  - FASTA: Fast Alignment Tool
  - Both find high scoring local alignments between a query sequence and a target database
  - Basic idea: first locate high-scoring short stretches and then extend them

71

Multiple Sequence Alignment



72

Multiple Sequence Alignment: Why?

- Identify highly conserved residues
  - Likely to be essential sites for structure/function
  - More precision from multiple sequences
  - Better structure/function prediction, pairwise alignments
- Building gene/protein families
  - Use conserved regions to guide search
- Basis for phylogenetic analysis
  - Infer evolutionary relationships between genes
- Develop primers & probes
  - Use conserved region to develop
    - Primers for PCR
    - Probes for DNA micro-arrays

Multidimensional Dynamic Programming

Assumptions: (1) columns are independent (2) linear gap cost

$$S(m) = G + \sum_i s(m_i)$$

$$G = \gamma(g) = dg$$

$\alpha_{i1,i2,\ldots,iN}$ **=Maximum score of an alignment up to the subsequences ending with** $x_{i1}^1, x_{i2}^2, \ldots, x_{iN}^N$

$$\alpha_{0,0,\ldots,0} = 0$$

$$\alpha_{i1,i2,\ldots,iN} = \max \begin{cases} \alpha_{i1-1,i2-1,\ldots,iN-1} + S(x_{i1}^1, x_{i2}^2, \ldots, x_{iN}^N) \\ \alpha_{i1,i2-1,\ldots,iN-1} + S(-, x_{i2}^2, \ldots, x_{iN}^N) \\ \alpha_{i1-1,i2,\ldots,iN-1} + S(x_{i1}^1, -, \ldots, x_{iN}^N) \\ \ldots \\ \alpha_{i1,i2,\ldots,iN-1} + S(-, -, \ldots, x_{iN}^N) \\ \ldots \\ \alpha_{i1-1,i2,\ldots,iN} + S(x_{i1}^1, -, \ldots, -) \end{cases}$$

Alignment: 0,0,0…,0---|x¹| , …, |x^N|

We can vary both the model and the alignment strategies

Approximate Algorithms for Multiple Alignment

- Two major methods
  - Reduce a multiple alignment to a series of pairwise alignments and then combine the result (e.g., Feng-Doolittle alignment)
  - Using HMMs (Hidden Markov Models)
- Feng-Doolittle alignment (4 steps)
  - Compute all possible pairwise alignments
  - Convert alignment scores to distances
  - Construct a "guide tree" by clustering
  - Progressive alignment based on the guide tree (bottom up)
- Alignment Free methods
  - K-mers (Carl Woese, …)
  - S. Seo, M. Oh, Y. Park, S. Kim, DeepFam: deep learning based alignment-free method for protein family modeling and prediction, *Bioinformatics*, Volume 34, Issue 13, 1 July 2018

75

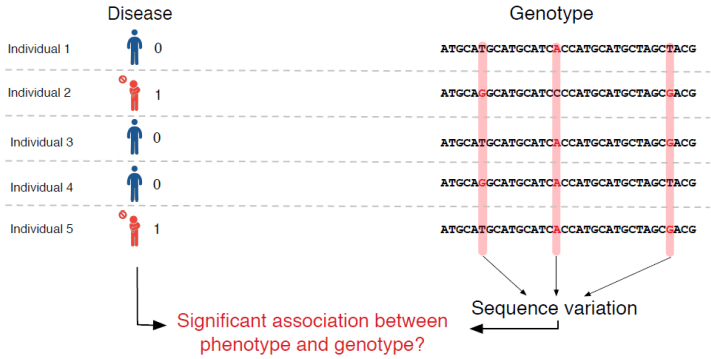
Mining Sequence Patterns in Biological Data

- A brief introduction to biology and bioinformatics
- Alignment of biological sequences
- GWAS Mining
- Summary

76

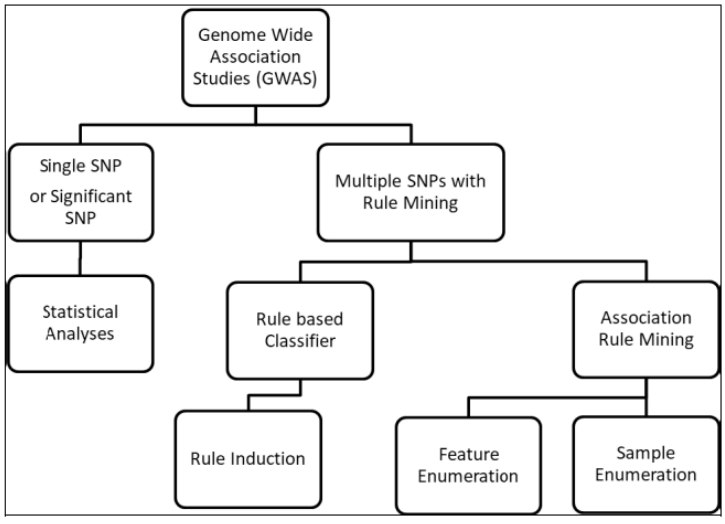
GWAS Mining

S. Mutalib, A. Mohamed, S. Abdul-Rahman, and N. Mustafa, Weighted Frequent Itemset of Single Nucleotide Polymorphisms (SNPs) in Genome Wide Studies (GWAS), *International Journal of Machine Learning and Computing, Vol. 8, No. 4, August 2018*

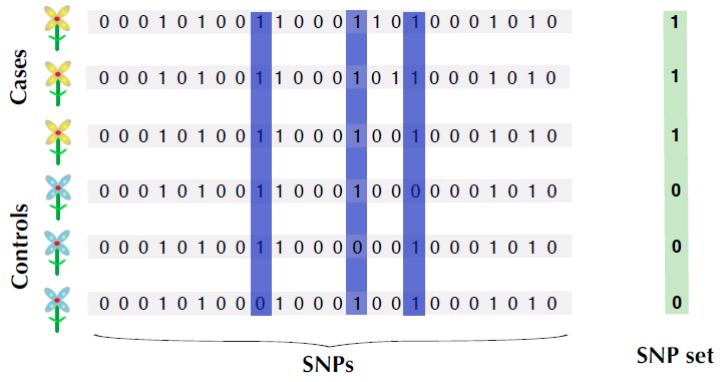


From: K. Borgwardt, Combinatorial Association Mapping, ETH Zurich, Dept. Biosystems, May 2017.

GWAS Mining

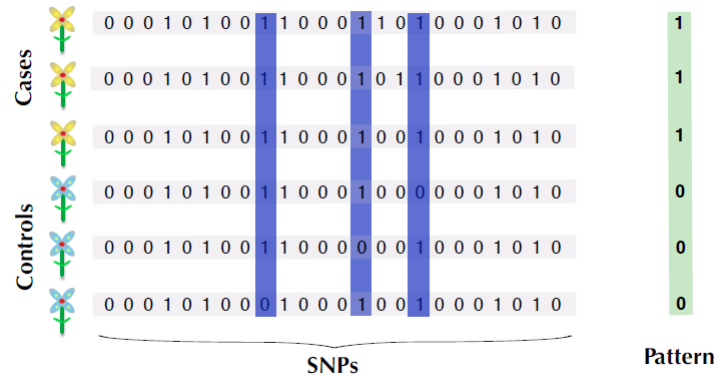


GWAS Mining



From: K. Borgwardt, Combinatorial Association Mapping, ETH Zurich, Dept. Biosystems, May 2017.

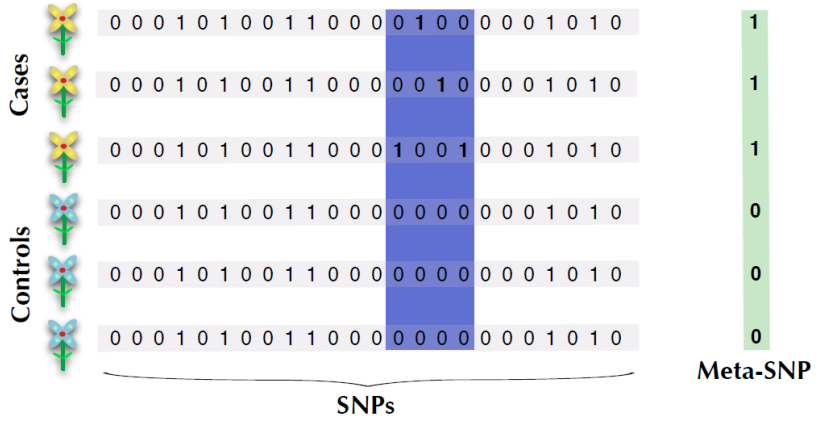
GWAS Mining



Feature selection: patterns.
Use patterns for identifying classes.

From: K. Borgwardt, Combinatorial Association Mapping, ETH Zurich, Dept. Biosystems, May 2017.

GWAS Mining



Intervals of genetic heterogeneity: FAIS (Fast Automatic Interval Search).
See also: Significant Pattern Mining (www.significant-patterns.org)

From: K. Borgwardt, Combinatorial Association Mapping, ETH Zurich, Dept. Biosystems, May 2017.

Weighted Frequent Item Set Mining

Goal: Mining SNPs that may be less frequent but important for diabetes.

"Although conducting association studies based on genetic variants is practical, it is not practical or statistically feasible to genotype and test all SNPs in the genome in conducting association studies."

Many SNPs irrelevant to the study at hand.

=> Feature selection.

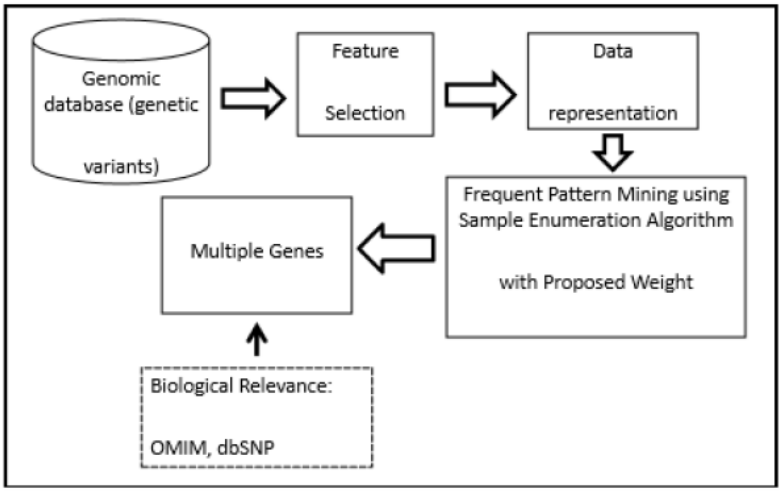
Weighted Frequent Item Set Mining

Goal: Mining SNPs that may be less frequent but important for diabetes.

General setup:

- Reduction of feature space
- Testing with classifiers
- Selection of important SNPs by allelic testing
- Weight assignment for the selected SNPs
- Item set mining
- Gene analysis

Weighted Frequent Item Set Mining



Weighted Frequent Item Set Mining

TABLE VIII: GENE ID INFO FROM THE ITEMSETS

Number	Gene ID	Number	Gene ID
1.	A2BP1	2.	IL21R
3.	AKTIP	4.	KIAA0556
5.	CARHSP1	6.	KIAA1576
7.	CDH13	8.	LITAF
9.	CNGB1	10.	LOC440389
11.	COTL1	12.	LOC727881
13.	CRISPLD2	14.	LYRM1
15.	DEF8	16.	MT4
17.	FTO	18.	MYH11
19.	GPR56	20.	MYLK3
21.	GRIN2A	22.	PLCG2
23.	GSPT1	24.	PRMT7
25.	HERPUD1	26.	SNX29
27.	HSD17B2	28.	SRCAP
29.	TOX3	30.	TMCO7
31.	ZDHHC7	32.	WWOX

- Gene FTO associated with obesity gives highest number of occurrence.

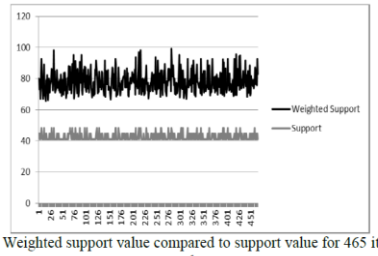


Fig. 5. Weighted support value compared to support value for 465 itemsets only.

Mining Sequence Patterns in Biological Data

- A brief introduction to biology and bioinformatics
- Alignment of biological sequences
- GWAS Mining
- Summary

References

- Lecture notes@M. Craven's website: www.biostat.wisc.edu/~craven
- A. Baxevanis and B. F. F. Ouellette. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (3rd ed.). John Wiley & Sons, 2004
- R.Durbin, S.Eddy, A.Krogh and G.Mitchison. Biological Sequence Analysis: Probability Models of Proteins and Nucleic Acids. Cambridge University Press, 1998
- N. C. Jones and P. A. Pevzner. An Introduction to Bioinformatics Algorithms. MIT Press, 2004
- I. Korf, M. Yandell, and J. Bedell. BLAST. O'Reilly, 2003
- L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE, 77:257--286, 1989
- J. C. Setubal and J. Meidanis. Introduction to Computational Molecular Biology. PWS Pub Co., 1997.
- M. S. Waterman. Introduction to Computational Biology: Maps, Sequences, and Genomes. CRC Press, 1995