ICQB

Introduction to Computational & Quantitative Biology (G4120)
Fall 2022

Oliver Jovanovic, Ph.D.
Columbia University
Department of Microbiology & Immunology

What is bioinformatics and computational biology?

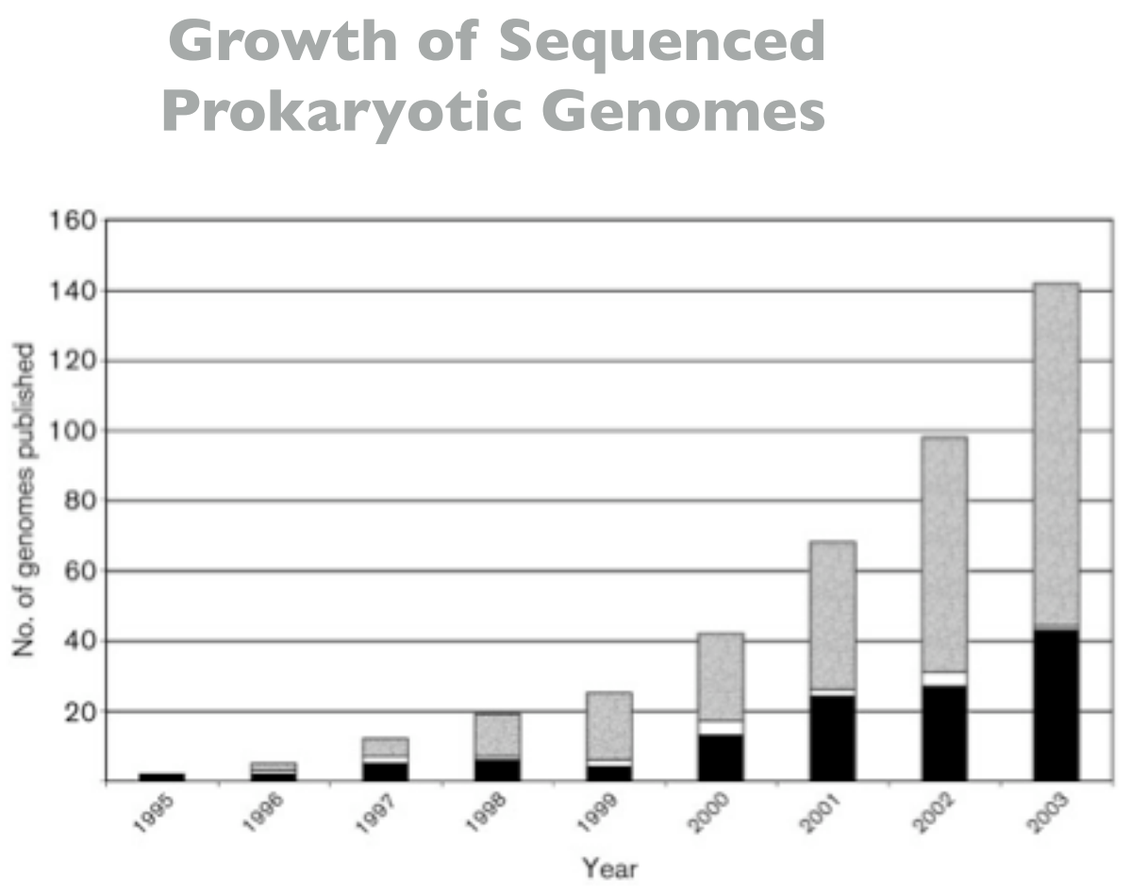
Introduction to Computational Biology
September 13, 2022

What is bioinformatics and computational biology?

“Biologists doing things with computers.”
– Lincoln Stein, CSHL

History of Sequencing

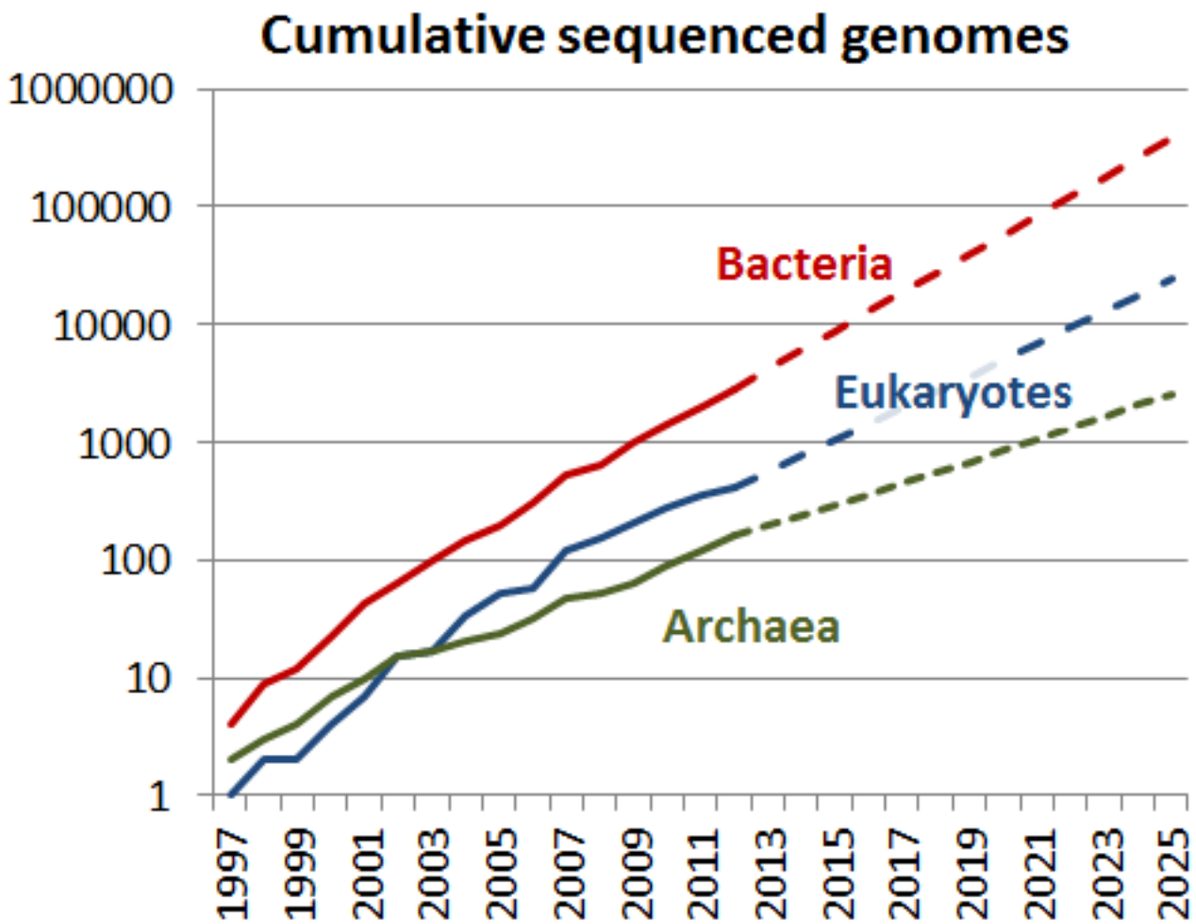
1977	Maxam-Gilbert and Sanger sequencing
1980	øX174 (5,386 bp)
1981	Human mitochondria (16,569 bp)
1981	Poliovirus (7,440 bp)
1990	Human Genome Project
1992	The Institute for Genomic Research
1994	RK2 (60,099 bp)
1995	Haemophilus influenzae (1.8 Mb)
1995	Mycoplasma genitalium (0.58 Mb)
1996	Saccharomyces cerevisiae (12.1 Mb)
1997	Escherichia coli (4.7 Mb)
1998	Celera, Inc.
1998	Caenorhabditis elegans (97 Mb)
2000	Drosophila melanogaster (180 Mb)
2000	Arabidopsis thaliana (115 Mb)
2001	Salmonella typhimurium (4.8 Mb)
2001	Homo sapiens (2.9 Gb)
2002	Mus musculus (2.6 Gb)
2003	Nanoarchaeum equitans (0.49 Mb)
2004	Legionella pneumophila (3.4 Mb)
2005	Pan troglodytes (2.8 Gb)
2006	454 Pyrosequencer
2007	Illumina HiSeq
2010	Ion Torrent
2011	Illumina MiSeq and PacBio RS
2013	PacBio RS II
2014	Illumina NexSeq
2015	Oxford Nanopore MinION and PacBio Sequel
2017	Illumina NovaSeq
2019	Oxford Nano. PromethION and PacBio Sequel



Source: David W. Ussery (2004) Genome Update: 161 prokaryotic genomes sequenced, and counting, *Microbiology* **150:** 261-263.

The Genomics Era

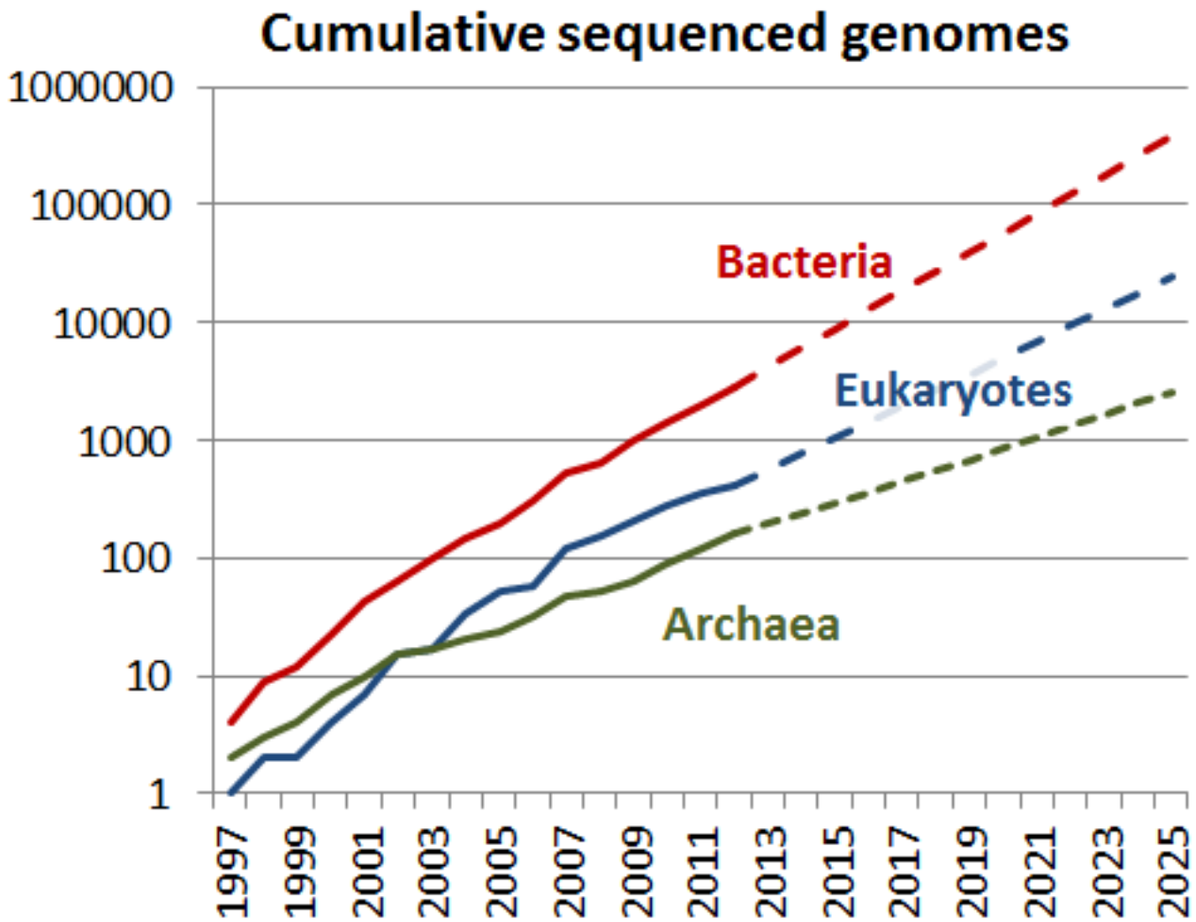
2013 Predictions



Source: GOLD Release v.5 May 28, 2014, genomesonline.org and Su, Andrew (2013) Cumulative sequenced genomes, dx.doi.org/10.6084/m9.figshare.723384

The Genomics Era

2013 Predictions



Source: GOLD Release v.5 May 28, 2014, genomesonline.org and Su, Andrew (2013) Cumulative sequenced genomes, dx.doi.org/10.6084/m9.figshare.723384
https://gold.jgi.doe.gov

2022 Reality

Cumulative sequenced genomes

Bacteria	398,322	**4x**
Eukaryotes	46,481	**5x**
Archaea	4,956	**4x**

Exponential Growth of Biological Data and Computing Power

GenBank

"From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months."

Source: www.ncbi.nlm.nih.gov/genbank/statistics

Exponential Growth of Biological Data and Computing Power

GenBank

"From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months."

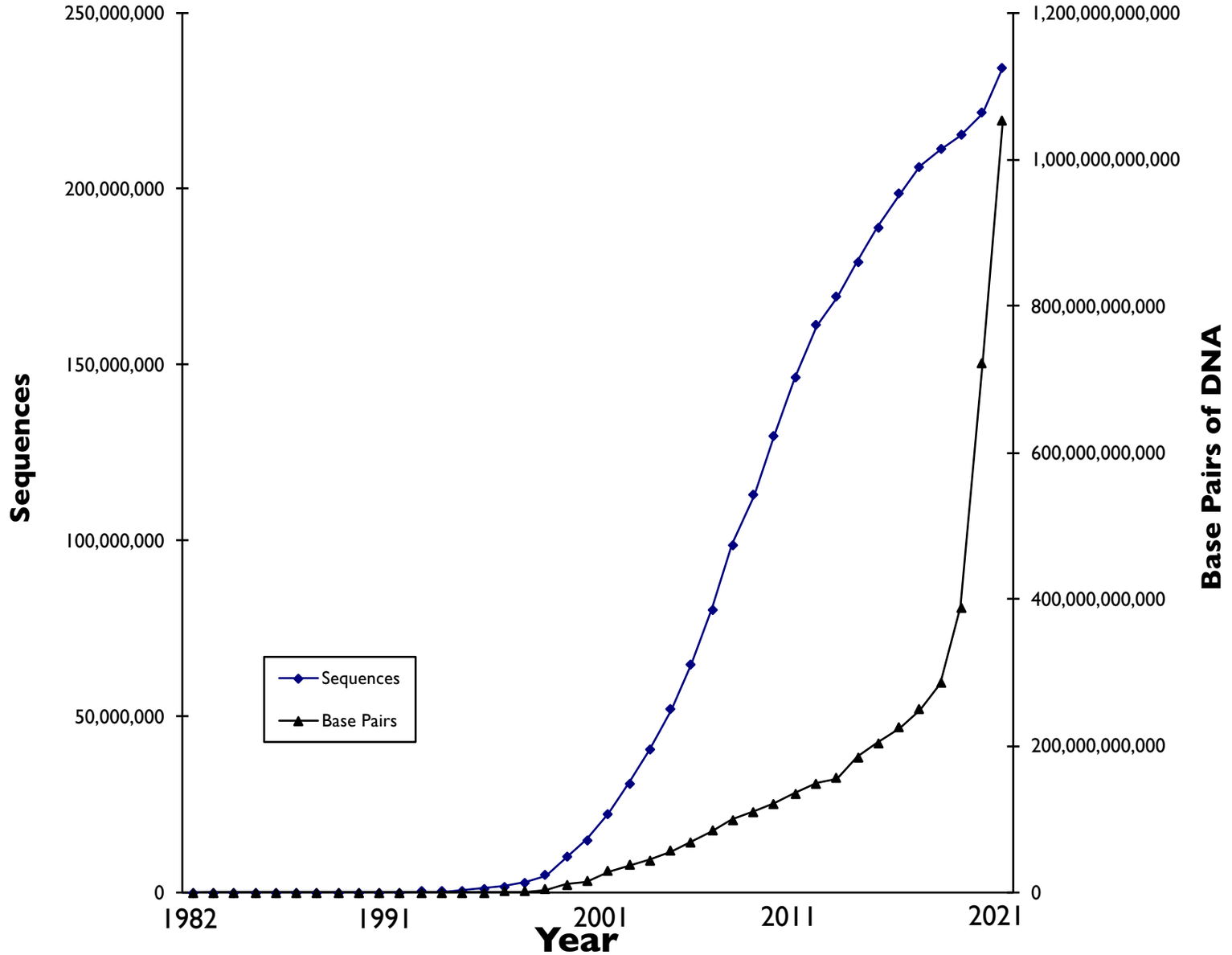
Source: www.ncbi.nlm.nih.gov/genbank/statistics

Moore's Law

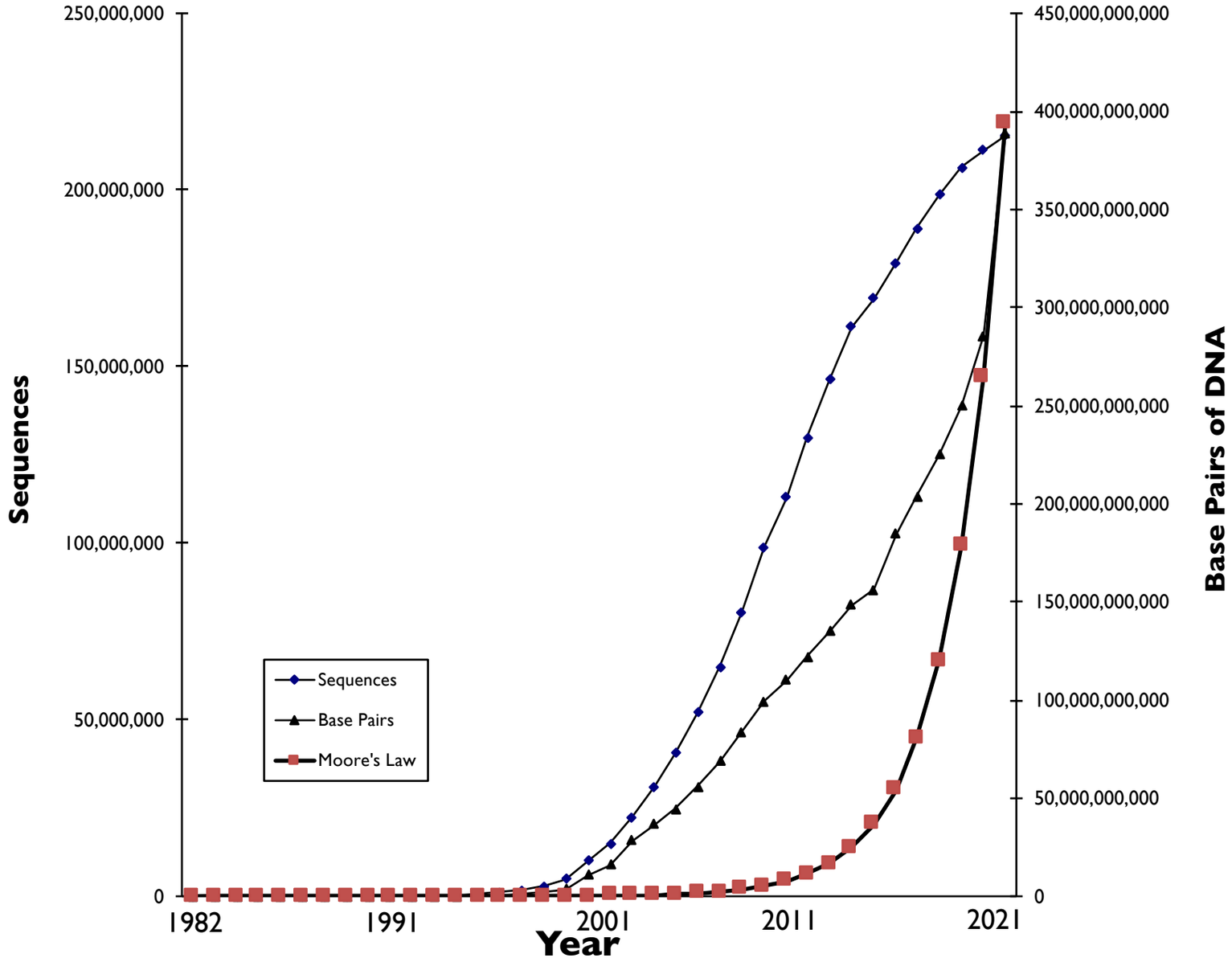
Over the history of computing hardware, the number of transistors in a dense integrated circuit doubles approximately every 18 to 24 months.

Source: Moore, Gordon E. (1965) Cramming more components onto integrated circuits. Electronics: 114-117 (with subsequent adjustments).

Growth of GenBank



Moore's Law

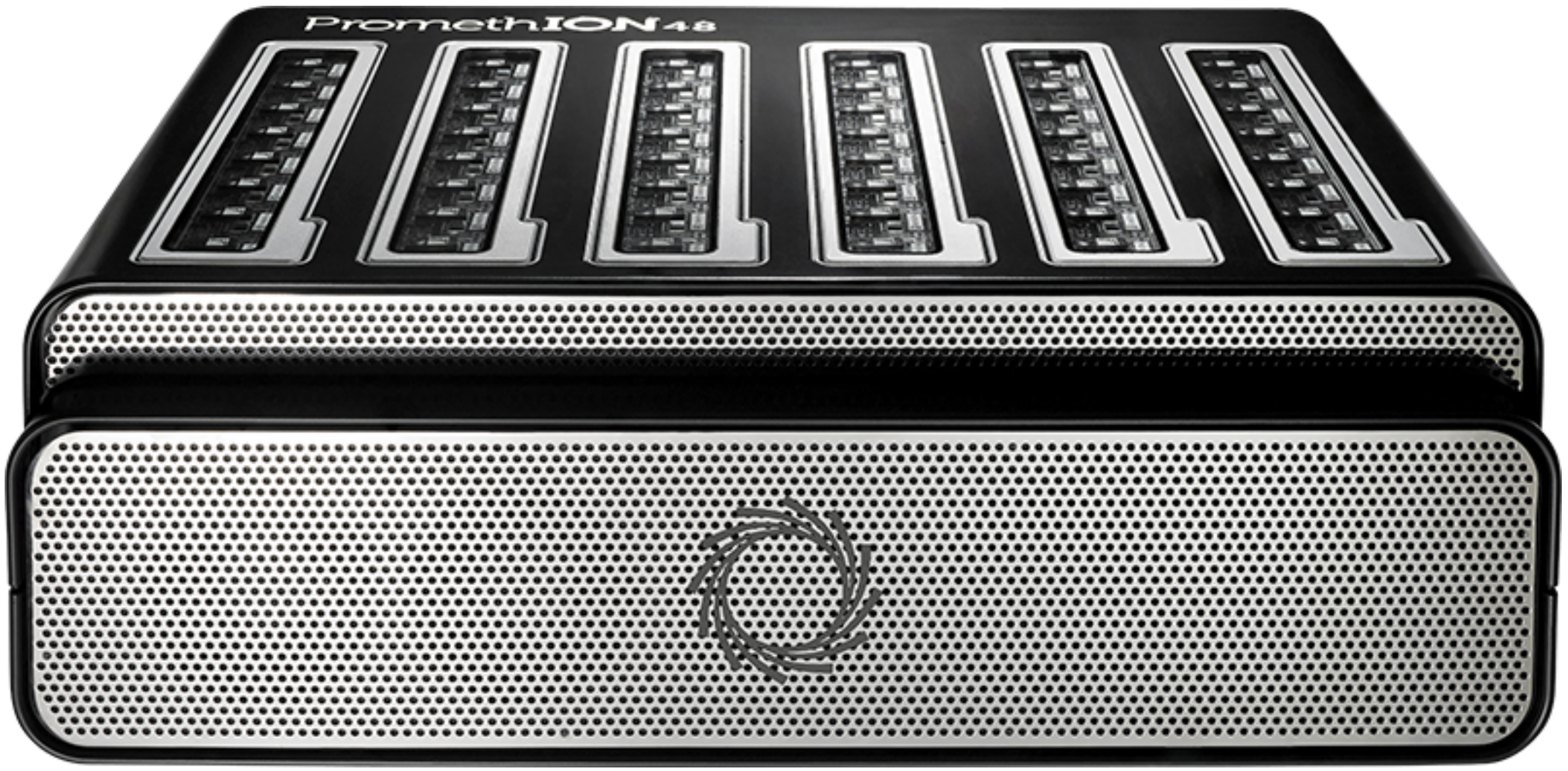

250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
Sequences
450,000,000,000
400,000,000,000
350,000,000,000
300,000,000,000
250,000,000,000
200,000,000,000
150,000,000,000
100,000,000,000
50,000,000,000
0
Base Pairs of DNA
Sequences
Base Pairs
Moore's Law
1982
1991
2001
2011
2021
Year

Third Generation DNA Sequencing?

minION from Oxford Nanopore Technologies

Introduction to Computational Biology
September 13, 2022

Fourth Generation DNA Sequencing



PromethION 48 from Oxford Nanopore Technologies

Introduction to Computational Biology
September 13, 2022

Dealing with exponentially increasing biological data...

...requires assistance.

Introduction to Computational Biology
September 13, 2022

What is the oldest device developed by humans to assist in computation?

Introduction to Computational Biology
September 13, 2022

Introduction to Computational Biology
September 13, 2022

c. 40,000 B.C.

Lebombo Bone Tally Stick

A baboon fibula with 29 notches discovered in the Lebombo Mountains of Africa.

Introduction to Computational Biology
September 13, 2022

Introduction to Computational Biology
September 13, 2022

c. 20,000 B.C.

Ishango Bone Number Stick

A baboon fibula with a sharp piece of quartz embedded in one end and tally marks carved on it in three columns. The left column consists of the prime numbers 19, 17, 13 and 11. The center column consists of the numbers 7, 5, 5, 10, 8, 4, 6 and 3. The right column consists of the numbers 9, 19, 21 and 11. It was discovered in Ishango in central Africa, near one of the headwaters of the Nile.

Introduction to Computational Biology
September 13, 2022

c. 2400 B.C.

The Abacus

Evidence of its use in a simpler form dates back to 2400 B.C. in Sumer. The ancient Akkadian word “abq” means dust. Texts dating to 190 A.D. detail its use in a more sophisticated form in China.

Source: Photo by Dave Fischer depicts a suanpan as used c. 1200 A.D.

Introduction to Computational Biology
September 13, 2022

c. 200 B.C.

The Antikythera Mechanism

An ancient Greek analogue computer consisting of 37 meshed gears that precisely mimicked the movements of the sun and moon, including the phases of the moon, tracked a 19 year Metonic calendar, predicted solar eclipses, calculated the timing of various Panhellenic games, and tracked the position of the five other planets known at the time.

Source: Freeth, T., et al. (2006) Decoding the ancient Greek astronomical calculator known as the Antikythera Mechanism. Nature 444: 587–591.

NVP3D.com

Introduction to Computational Biology
September 13, 2022

History of Early Computing

40,000 BC	Tally systems	Africa & Europe
20,000 BC	Prime system	Africa
2400 BC	Abacus	Sumer & Babylon
200 BC	Antikythera mechanism	Greece
1500	Mechanical calculator	Leonardo da Vinci
1621	Slide rule	William Oughtred
1642	Arithmetic Machine	Blaise Pascal
1822	Difference Engine	Charles Babbage
1831	Computer program	Lady Ada Lovelace
1936	Z1 Computer	Konrad Zuse
1936	Turing Machine	Alan Turing
1938	Boolean Circuits	Claude Shannon
1943	COLOSSUS	Alan Turing
1945	von Neumann Machine	John von Neumann
1946	ENIAC	John Mauchly & J. Presper Eckert
1947	Transistor	William Shockley, John Bardeen & Walter Brattain
1958	Integrated Circuit	Jack Kilby & Robert Noyce

Computational Biology

- Data
  Sequencers, FACS, scanners, microscopes, etc.
- Analysis
  Software, scripting, programming, etc.
- Storage
  Databases, local, network or cloud storage, backup, etc.
- Sharing
  Web, Internet, email, portable or cloud storage, etc.

History of Early Bioinformatics

1869	DNA	Johann Friedrich Miescher
1924	Chromosomal DNA	Robert Feulgen
1928	Transforming principle	Franklin Griffith
1944	DNA transformation	Oswald Avery, Maclyn McCarty & Colin MacLeod
1948	Information Theory	Claude Shannon
1949	Chargaff's Rule	Erwin Chargaff
1953	Double helix	James Watson & Francis Crick
1955	Protein sequencing	Fred Sanger
1961	Codons	Sidney Brenner & Francis Crick
1966	Genetic code	Marshall Nirenberg, Robert Holley & Har Khorana
1970	Restriction enzyme	Hamilton Smith, Johns Hopkins
1970	Needleman-Wunsch	S. Needleman & C. Wunsch
1971	MEDLINE	NIH/NLM
1977	DNA sequencing	Allan Maxam & Walter Gilbert/Frederick Sanger
1977	Staden programs	Roger Staden
1981	Smith-Waterman	Temple Smith & Michael Waterman
1982	GenBank	LANL/EMBL/NCBI
1988	NCBI	NIH/NLM
1988	FASTA	William Pearson & David Lipman
1988	DNA Strider	Christian Marck
1990	BLAST	Stephen Altschul & David Lipman, NCBI
1994	DNA computer	Leonard Adelman
1997	PubMed	NCBI

Sequence Analysis

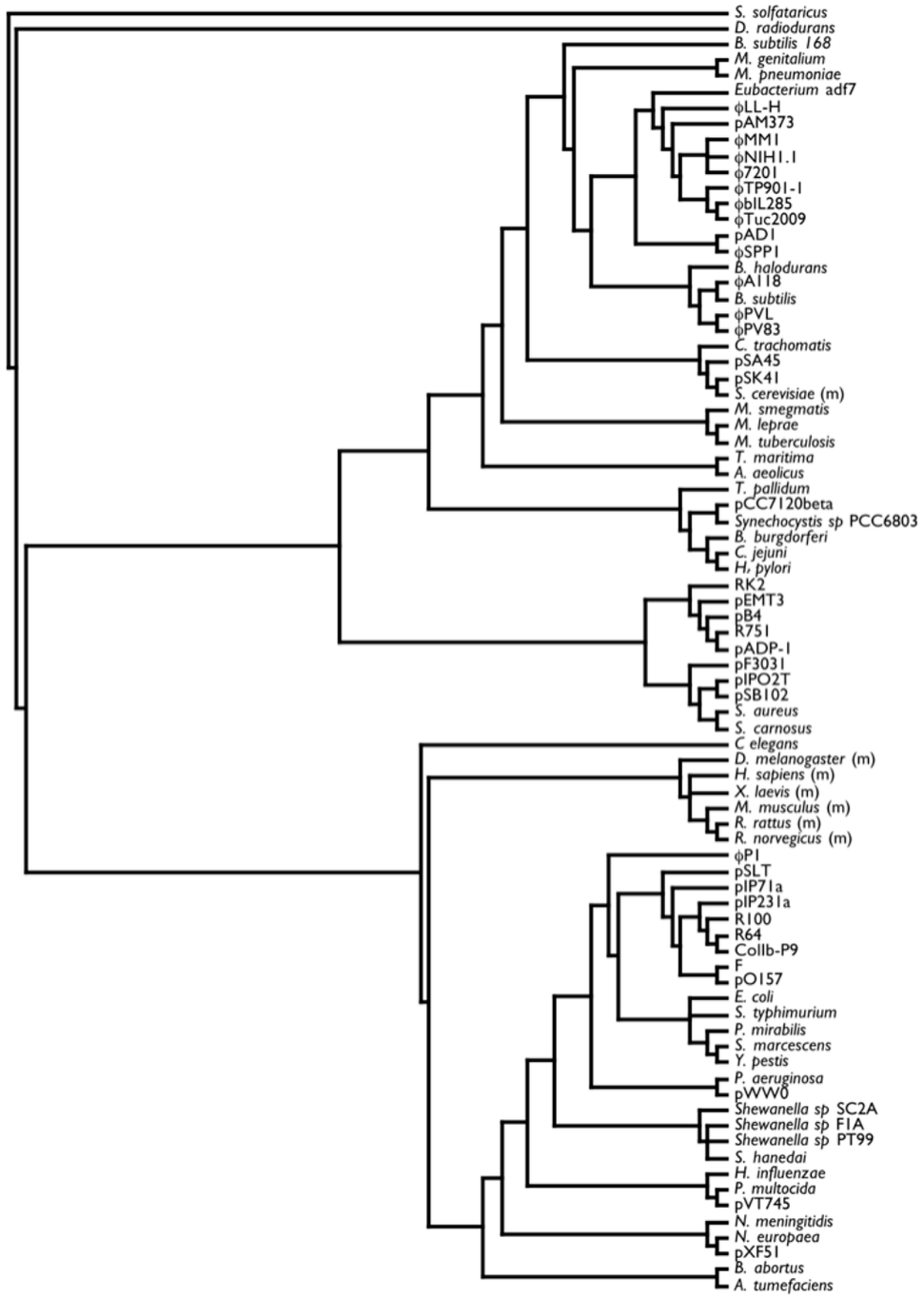
Position	Score	Predicted promoter sequence (-35 < gap> -10)	Name
2313	52.94	GTTAATTGCTTTTCGA <10> TTAGCTAAACTTTC	
3075	55.29	CGACATTGCTTGACCC <11> GCGTGTTCAATTCG	*korE*
3772	55.29	GCGTCATGCTTGAAAA <11> TGGCGTGCAATCAG	
5552	51.17	AAAAGGATCTTCACCT <10> AAATTAAAAATGAA	
5585	53.52	AAATGAAGTTTTAAAT <15> ATGAGTAAACTTGG	
5695	52.35	CATAGTTGCCTGACTC <12> ATAACTACGATACG	
6133	54.11	ACTGCATAATTCTCTT <11> ATCCGTAAGATGCT	
6478	56.47	ACGGAAATGTTGAATA <11> CCTTTTTCAATATT	
6511	53.52	TCAATATTATTGAAGC <12> TATTGTCTCATGAG	
6532	52.35	TCAGGGTTATTGTCTC <11> CATATTTGAATGTA	
6618	57.05	AAGTGCCACCTGACGT <10> ATTATTATCATGAC	
7292	59.41	ACGATTTAATGGACAC <11> CGTTTTACTATGTC	*tnpA*
8080	51.17	ATGAAGCAATTGAACG <11> TGAACGATTTTGGC	
8248	50.58	CCTTTGTCCTTGCATG <11> CGTTGGATGATGCA	
8482	58.23	AAAAGCTGGCTGAAAT <11> TCCGTGAAATTGCC	
11698	51.76	ATTATGCCCTAGCCTG <10> TTAGCTAAACTATG	
13937	60.58	TTGTCATGCTTGACAC <12> AAACATAATATGTC	*tetA*
17967	53.52	ACCGCTATATCGAAAA <10> CTTGTTAGAATTGC	
17999	57.64	AGAATTGCCATGACGT <11> ACGGGTAAGATTAC	*trbA*
18729	53.52	ATTAGCTGTTTGTCTT <14> TTCGGTATATCGTT	*trbB*
26625	50.58	ACCAGGCGTTTGACTA <9> AGGAGTAACTTATG	
35054	43.52	CTCGCGCTGTAGCCTC <11> TGTGCTAATGTGGT	*parD*
38324	64.70	ATCGTGGCGTTGACAA <11> CTGGCTACACTATG	*aphA*
40006	54.70	TCGTAGTTCTTGCCGA <11> TTCTCAAAGATGCC	
47326	52.35	ATCAGTTGCTTGATGC <11> TTGCTGACGTTGCG	
48938	54.70	CAAACGGTTTTGGCTT <12> TTTCGTCCAATGCG	
51306	57.64	GAAAAAGGATGGATAT <9> ATCGCTATAATGAC	*traK*
59051	54.70	TGTTTTTTCTTGGCGT <11> TTCCGGACGATGTA	
2375c	66.47	CTAAAGGTGTTGACGA <12> TTAGCTAAACTTCT	*klaA*
3711c	52.94	ATTCTTGTTTTGAGGC <11> CCAGGTCAATTACC	
3745c	67.05	TAAAATTGCTTGACAA <12> TGCCCTATTCTTGT	*kleC*
3777c	52.94	GACGCCTCGCTGAATC <11> TTAGCTAAAATTGC	
4400c	65.88	TAAATTTCCTTGACTA <12> TGCCCTAATATAGC	*kleA*
5562c	53.52	AGATCCTTTTTGATAA <14> TCCCTTAACGTGAG	
5619c	51.76	CATATATACTTTAGAT <12> TCATTTTTAATTTA	
5647c	54.70	CATTGGTAACTGTCAG <11> TCATATATACTTTA	

SeqMatrix *E. coli* promoter output:

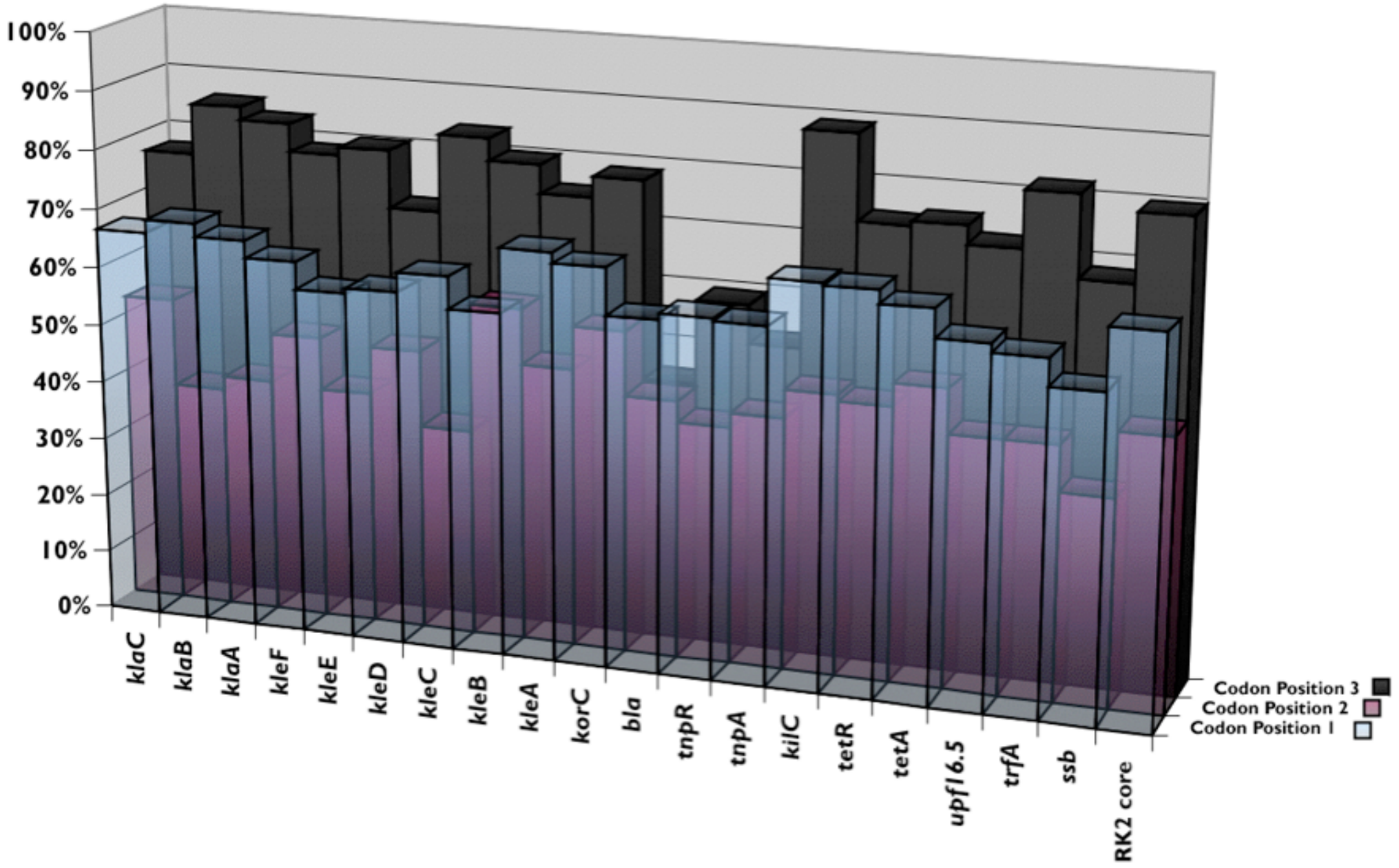
DNA Location: 3,075
Spacer Length: 11
Similarity Score: 55.29

CGACATTGCTTGACCC <11> GCGTGTTCAATTCG

Phylogeny


S. solfataricus
D. radiodurans
B. subtilis 168
M. genitalium
M. pneumoniae
Eubacterium adf7
ϕLL-H
pAM373
ϕMM1
ϕNIH1.1
ϕ7201
ϕTP901-1
ϕbIL285
ϕTuc2009
pAD1
ϕSPP1
B. halodurans
ϕA118
B. subtilis
ϕPVL
ϕPV83
C. trachomatis
pSA45
pSK41
S. cerevisiae (m)
M. smegmatis
M. leprae
M. tuberculosis
T. maritima
A. aeolicus
T. pallidum
pCC7120beta
Synechocystis sp PCC6803
B. burgdorferi
C. jejuni
H. pylori
RK2
pEMT3
pB4
R751
pADP-1
pF3031
pIPO2T
pSB102
S. aureus
S. carnosus
C elegans
D. melanogaster (m)
H. sapiens (m)
X. laevis (m)
M. musculus (m)
R. rattus (m)
R. norvegicus (m)
ϕP1
pSLT
pIP71a
pIP231a
R100
R64
ColIb-P9
F
pO157
E. coli
S. typhimurium
P. mirabilis
S. marcescens
Y. pestis
P. aeruginosa
pWW0
Shewanella sp SC2A
Shewanella sp F1A
Shewanella sp PT99
S. hanedai
H. influenzae
P. multocida
pVT745
N. meningitidis
N. europaea
pXF51
B. abortus
A. tumefaciens

Data Visualization



Introduction to Computational Biology
September 13, 2022

Multimedia

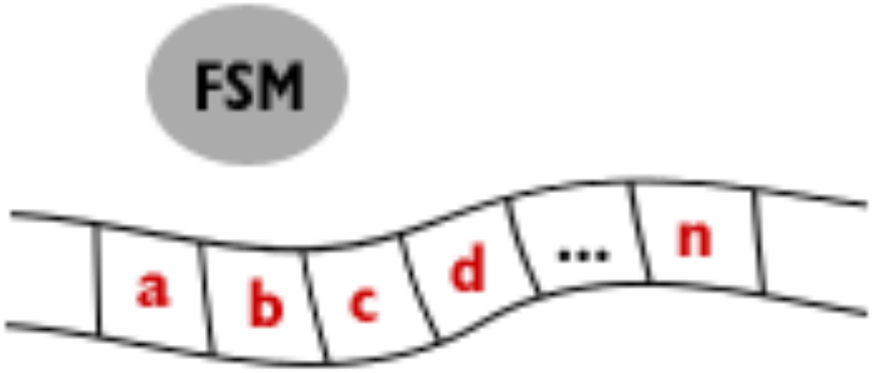
```
L27758. Birmingham IncP-a...[gi:508311]    Related Sequences, PubMed, Taxonomy

LOCUS       BIACOMGEN              60099 bp    DNA     linear   BCT 08-JUL-1994
DEFINITION  Birmingham IncP-alpha plasmid (R18, R68, RK2, RP1, RP4) complete
            genome.
ACCESSION   L27758
VERSION     L27758.1  GI:508311
KEYWORDS    complete genome.
SOURCE      Birmingham IncP-alpha plasmid (plasmid Birmingham IncP-alpha
            plasmid, kingdom Prokaryotae) DNA.
ORGANISM    Birmingham IncP-alpha plasmid
            broad host range plasmids.
REFERENCE   1  (bases 1 to 60099)
AUTHORS     Pansegrau,W., Lanka,E., Barth,P.T., Figurski,D.H., Guiney,D.G.,
            Haas,D., Helinski,D.R., Schwab,H., Stanisich,V.A. and Thomas,C.M.
TITLE       Complete nucleotide sequence of Birmingham IncP-alpha plasmids:
            compilation and comparative analysis
JOURNAL     J. Mol. Biol. 239, 623-663 (1994)
MEDLINE     94285211
FEATURES             Location/Qualifiers
     source          1..60099
                     /organism="Birmingham IncP-alpha plasmid"
                     /plasmid="Birmingham IncP-alpha plasmid"
                     /db_xref="taxon:35419"
BASE COUNT    10839 a  18681 c  18448 g  12131 t
ORIGIN
        1 ttcacccccg aacacgagca cggcacccgc gaccactatg ccaagaatgc ccaaggtaaa
       61 aattgccggc cccgccatga agtccgtgaa tgccccgacg gccgaagtga agggcaggcc
      121 gccacccagg ccgccgccct cactgcccgg cacctggtcg ctgaatgtcg atgccagcac
      181 ctgcggcacg tcaatgcttc cgggcgtcgc gctcgggctg atcgcccatc ccgttactgc
      241 cccgatcccg gcaatggcaa ggactgccag cgccgcgatg aggaagcggg tgccccgctt
      301 cttcatcttc gcgcctcggg cctcgaggcc gcctacctgg gcgaaaacat cggtgtttgt
etc.
```

Binary Computing and DNA

Modern computers are digital machines, which means their basic function involves using discrete symbols from a finite set.

In 1936, Alan Turing proved that a finite state machine (FSM) moving up or down a tape of symbols, reading or writing one symbol at a time, could solve any computable problem, and serve as a universal machine.



Universal Turing Machine

The most basic level of information in nearly all current computers represents only one of two possibilities: **0** (off) or **1** (on)**.** A signal that can carry one of two possible messages (**0** or **1**) is called a binary signal, or a **bit,** so these computers are binary machines.

The Digital Language of Computers

Binary Units

0 or 1 = 1 bit

8 bits = 1 byte

1,024 bits = 1 kilobit

1,024 bytes = 1 kilobyte (K)

1,024 kilobytes = 1 megabyte (M)

1,024 megabytes = 1 gigabyte (G)

1,024 gigabytes = 1 terabyte (T)

1 bit =	0 or 1	= 2 possibilities
2 bits =	0 or 1 0 or 1	= 2 x 2 = 4 possibilities
3 bits =	0 or 1 0 or 1 0 or 1	= 2 x 2 x 2 = 8 possibilities
4 bits =	0 or 1 0 or 1 0 or 1 0 or 1	= 2 x 2 x 2 x 2 = 16 possibilities
5 bits =	0 or 1 0 or 1 0 or 1 0 or 1 0 or 1	= 2 x 2 x 2 x 2 x 2 = 32 possibilities
6 bits =	0 or 1 0 or 1 0 or 1 0 or 1 0 or 1 0 or 1	= 2 x 2 x 2 x 2 x 2 x 2 = 64 possibilities
7 bits =	0 or 1 0 or 1 0 or 1 0 or 1 0 or 1 0 or 1 0 or 1	= 2 x 2 x 2 x 2 x 2 x 2 x 2 = 128 possibilities
8 bits =	0 or 1 0 or 1 0 or 1 0 or 1 0 or 1 0 or 1 0 or 1 0 or 1	= 2 x 2 x 2 x 2 x 2 x 2 x 2 x 2 = 256 possibilities

DNA has only four possibilities (so can be represented by 2 bits)

G = 00

C = 11

A = 01

T = 10

Complementation (with intelligent choice of representation)

G C C A = 00 11 11 01

C G G T = 11 00 00 10

ASCII Coding of DNA

American Standard Code for Information Interchange (ASCII)

- For practical purposes, DNA and RNA is generally represented in ASCII code, using the upper or lower case letters A, C, G, and T or A, C, G and U.
- Each ASCII character occupies one byte, and thus has 256 possibilities, including all upper and lower case letters of the English alphabet, the ten Arabic numerals, punctuation, and special characters, such as @.
- Thus, a kilobase of DNA (1,000 base pairs) occupies just under a kilobyte (1 K = 1,024 bytes) of storage in ASCII. An entire human genome, roughly 3 billion base pairs (3 gigabases), occupies just under 3 gigabytes of storage in ASCII.

Transcription

- Transcription is computationally trivial. One need only substitute a U for a T if dealing with a sense strand, or complement, then transcribe if dealing with the antisense strand.

Translation

- Translation is also computationally trivial. A computer can refer to a species appropriate translation table to translate DNA or RNA into the appropriate protein sequence.

AUA	I	Isoleucine
AUC	I	Isoleucine
AUG	M	Methionine start
AUU	I	Isoleucine
etc.		

Alternate Representation

- Can readily convert an ASCII representation of DNA into other forms, such as graphics, or even music.

Information Content

Uncertainty

Uncertainty can be thought of as the number of yes/no questions required to identify the state something is in. It can be measured in bits.

- A coin toss, with only 2 possibilities, can be identified with a single question (i.e., “Is it heads?”)
- A nucleotide, with 4 possibilities, can be identified with two questions (i.e. “Is it a purine? Is it adenine?”)

Maximum Uncertainty

Maximum Entropy = $\log_2(n)$ where n is the number of possible states

Coin	$\log_2(2) = 1$ bit
DNA	$\log_2(4) = 2$ bits
Protein	$\log_2(20) = 4.32$ bits

Compression algorithms offer one approach to testing the randomicity of a DNA sequence.
A very random DNA sequence will require close to 2 bits per nucleotide to represent it, even when compressed. A sequence of DNA that has repeating patterns, or is otherwise highly structured, should be capable of being represented by less than 2 bits per nucleotide.

Algorithms in Computational Biology

Algorithm

- An algorithm is simply a series of steps used to solve a problem. One of a computer's great strengths is its ability to rapidly and accurately repeat recursive steps in an algorithm.

Consensus

- Early algorithms for searching sequence data depended on consensus sequences. Thus, to find a prokaryotic promoter, one would try to find something that matched a consensus -10 sequence (TATAAT), not too far downstream of a consensus -35 sequence (TTGACA).
- It rapidly became clear that biologically significant sequences rarely perfectly matched a consensus, and more sophisticated approaches were adopted, including the use of matrices, Markov chains and hidden Markov models.

Matrices

- Matrices take into account the distribution of every possible nucleotide (or amino acid) at a position in a set of known sequences. Searching with a matrix is therefore more sensitive than searching with a consensus, and can find biological features that a strict consensus approach would miss.

Markov chains and hidden Markov models (HMMs)

- Markov chains and hidden Markov models are probabilistic models of sequences, and have proven useful in database searching, gene finding and multiple sequence alignment.
- A first-order Markov chain is a finite state automaton (a restricted Turing machine which only moves left to right) with probabilities for each transition to a new state (symbol) based on its current state. Higher order Markov chains take into account one or more previous states.
- A hidden Markov model is a Markov chain in which only the output can be observed (its current state is hidden).

Consensus vs. Matrix

E. coli Promoter Consensus

-35 Region		**-10 Region**
TTGACA	....	TATAAT

E. coli Promoter Matrix

-35 Region

										T	T	G	A	C	A	
A	11	8	8	7	8	7	3	5	5	0	1	0	14	5	9	5
C	3	4	2	4	4	3	5	2	8	1	1	2	3	11	2	5
G	3	2	4	2	4	5	5	5	5	2	1	17	1	2	3	3
T	4	7	7	8	5	6	8	9	3	17	18	2	4	3	7	9

Spacer Region

Length	9	10	11	12	13	14	15
	1	6	14	6	1	1	1

-10 Region

						T	A	T	A	A	T			
A	4	5	3	4	4	0	20	5	12	11	0	7	4	6
C	5	4	5	4	5	2	0	3	3	4	1	2	7	6
G	2	5	5	8	7	2	0	3	3	3	0	6	5	6
T	10	6	8	5	6	17	1	9	3	4	20	6	5	4

Matrix Analysis Example

Position	Score	Predicted promoter sequence (-35 < gap> -10)	Name
2313	52.94	GTTAATTGCTTTTCGA <10> TTAGCTAAACTTTC	
3075	55.29	CGACATTGCTTGACCC <11> GCGTGTTCAATTCG	*korE*
3772	55.29	GCGTCATGCTTGAAAA <11> TGGCGTGCAATCAG	
5552	51.17	AAAAGGATCTTCACCT <10> AAATTAAAAATGAA	
5585	53.52	AAATGAAGTTTTAAAT <15> ATGAGTAAACTTGG	
5695	52.35	CATAGTTGCCTGACTC <12> ATAACTACGATACG	
6133	54.11	ACTGCATAATTCTCTT <11> ATCCGTAAGATGCT	
6478	56.47	ACGGAAATGTTGAATA <11> CCTTTTTCAATATT	
6511	53.52	TCAATATTATTGAAGC <12> TATTGTCTCATGAG	
6532	52.35	TCAGGGTTATTGTCTC <11> CATATTTGAATGTA	
6618	57.05	AAGTGCCACCTGACGT <10> ATTATTATCATGAC	
7292	59.41	ACGATTTAATGGACAC <11> CGTTTTACTATGTC	*tnpA*
8080	51.17	ATGAAGCAATTGAACG <11> TGAACGATTTTGGC	
8248	50.58	CCTTTGTCCTTGCATG <11> CGTTGGATGATGCA	
8482	58.23	AAAAGCTGGCTGAAAT <11> TCCGTGAAATTGCC	
11698	51.76	ATTATGCCCTAGCCTG <10> TTAGCTAAACTATG	
13937	60.58	TTGTCATGCTTGACAC <12> AAACATAATATGTC	*tetA*
17967	53.52	ACCGCTATATCGAAAA <10> CTTGTTAGAATTGC	
17999	57.64	AGAATTGCCATGACGT <11> ACGGGTAAGATTAC	*trbA*
18729	53.52	ATTAGCTGTTTGTCTT <14> TTCGGTATATCGTT	*trbB*
26625	50.58	ACCAGGCGTTTGACTA <9> AGGAGTAACTTATG	
35054	43.52	CTCGCGCTGTAGCCTC <11> TGTGCTAATGTGGT	*parD*
38324	64.70	ATCGTGGCGTTGACAA <11> CTGGCTACACTATG	*aphA*
40006	54.70	TCGTAGTTCTTGCCGA <11> TTCTCAAAGATGCC	
47326	52.35	ATCAGTTGCTTGATGC <11> TTGCTGACGTTGCG	
48938	54.70	CAAACGGTTTTGGCTT <12> TTTCGTCCAATGCG	
51306	57.64	GAAAAAGGATGGATAT <9> ATCGCTATAATGAC	*traK*
59051	54.70	TGTTTTTTCTTGGCGT <11> TTCCGGACGATGTA	
2375c	66.47	CTAAAGGTGTTGACGA <12> TTAGCTAAACTTCT	*klaA*
3711c	52.94	ATTCTTGTTTTGAGGC <11> CCAGGTCAATTACC	
3745c	67.05	TAAAATTGCTTGACAA <12> TGCCCTATTCTTGT	*kleC*
3777c	52.94	GACGCCTCGCTGAATC <11> TTAGCTAAAATTGC	
4400c	65.88	TAAATTTCCTTGACTA <12> TGCCCTAATATAGC	*kleA*
5562c	53.52	AGATCCTTTTTGATAA <14> TCCCTTAACGTGAG	
5619c	51.76	CATATATACTTTAGAT <12> TCATTTTTAATTTA	
5647c	54.70	CATTGGTAACTGTCAG <11> TCATATATACTTTA	

SeqMatrix *E. coli* promoter output:

```
DNA Location: 3,075
Spacer Length: 11
Similarity Score: 55.29

CGACATTGCTTGACCC <11> GCGTGTTCAATTCG
    (TTGACA..............TATAAT)
```

Stochastic Modeling

Stochastic Model

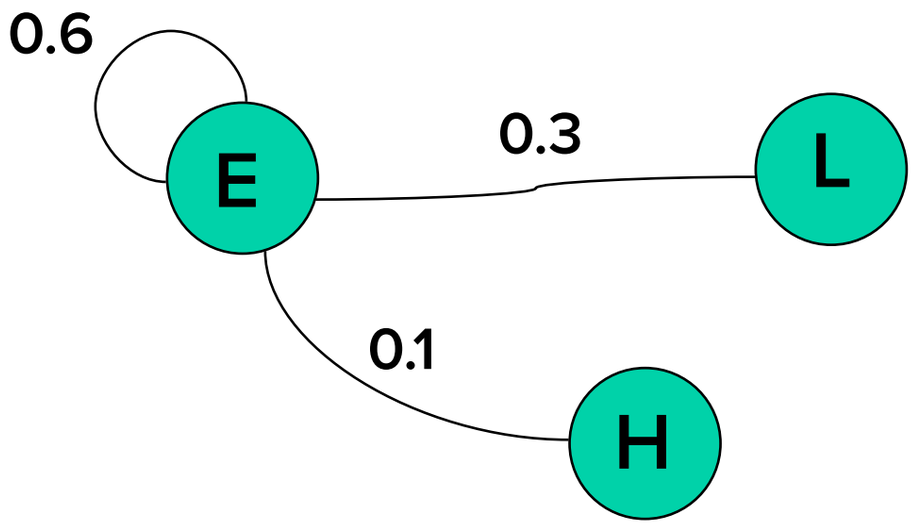
A model involving chance or probability. Markov models are a particular form of stochastic model.

Current Residue	Next Residue A	C	G	T
A	40%	15%	15%	30%
C	25%	25%	25%	25%
G	20%	25%	30%	25%
T	35%	20%	20%	25%

Markov Modeling

Markov State

A Markov state emits a symbol each time you visit it. It connects to other states (and possibly itself), with transition probabilities attached. The sum of the transition probabilities is 1.

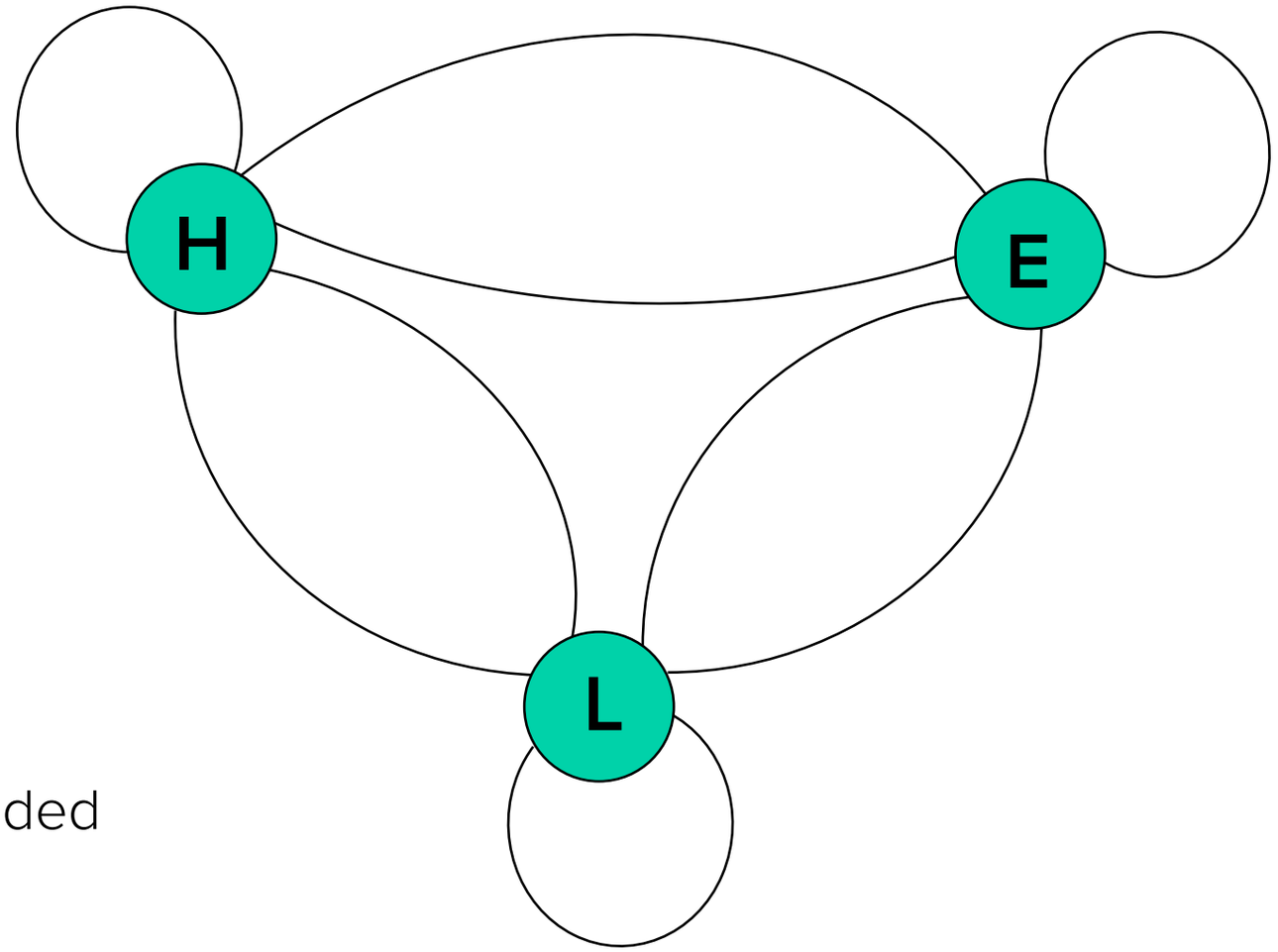


E = Extended
L = Loop
H = Helix

Markov Chains

Markov Chain

A Markov chain is an interlinked chain, or network, of states connected by transition probabilities.

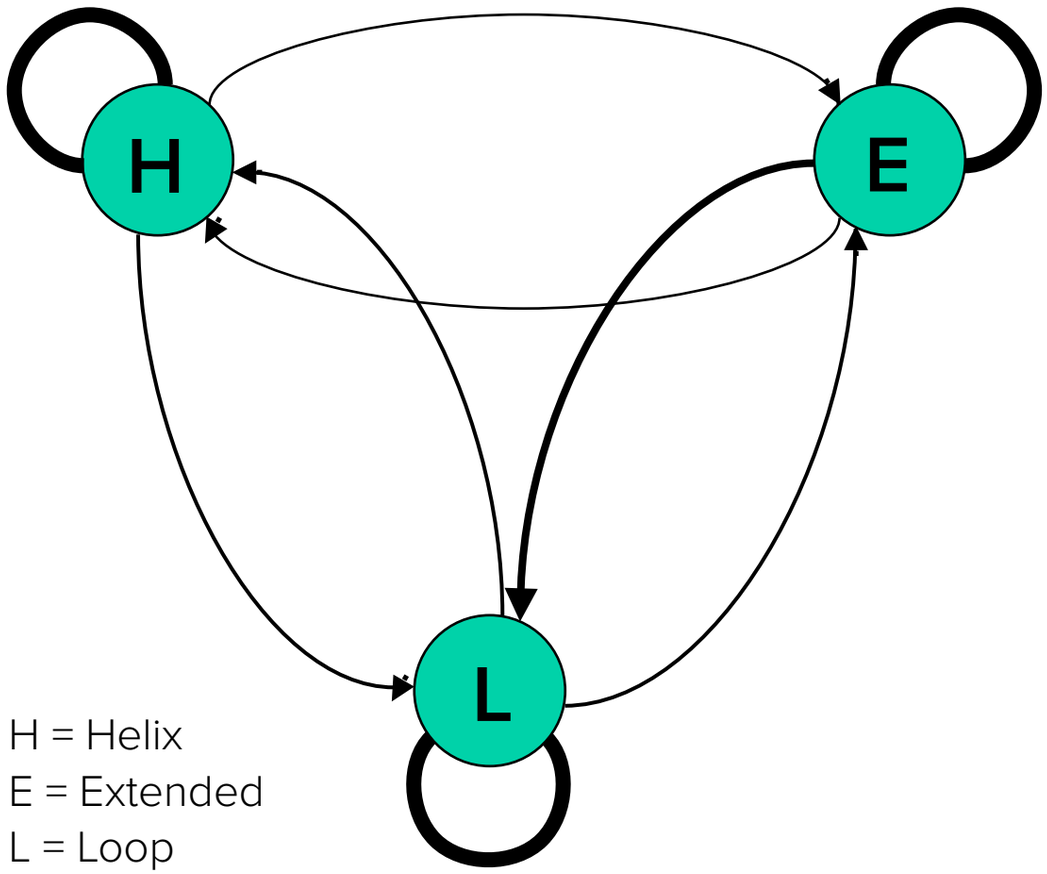


H = Helix
E = Extended
L = Loop

Markov Transition Matrices

Transition Matrix

A transition matrix for a first order Markov chain, the simplest kind. The sum of the transition probabilities from each state is 1.

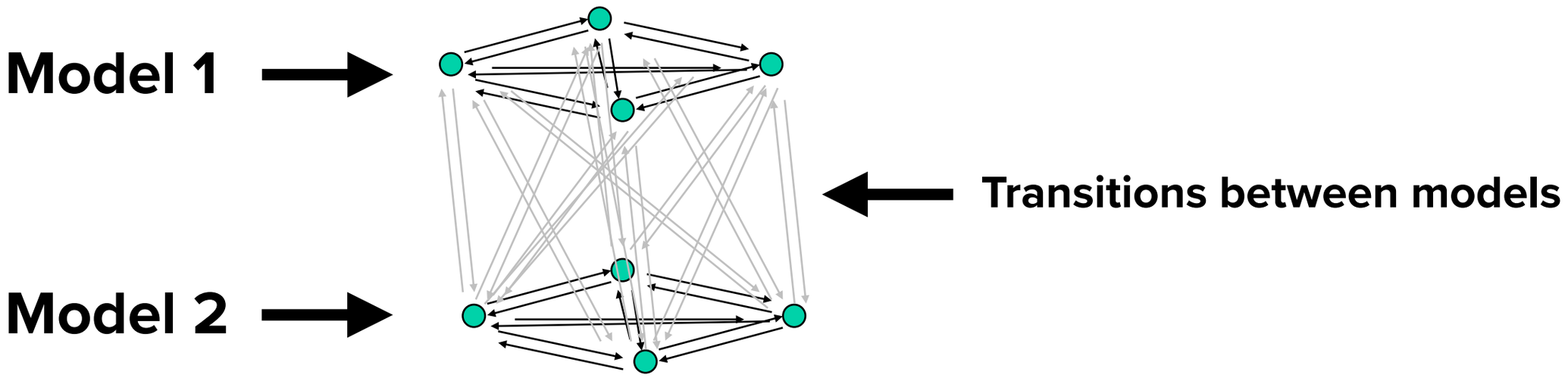


	H	E	L
H	**0.93**	0.01	0.06
E	0.01	**0.80**	**0.19**
L	0.04	0.06	**0.90**

Hidden Markov Models

Hidden Markov Model (HMM)

A hidden Markov model consists of two Markov chains connected such that a one to one correspondence between the state and the emitted symbol no longer exists.



GeneMark

GeneMark and GeneMark.hmm

Mark Borodovsky, Georgia Institute of Technology

http://exon.gatech.edu/GeneMark/

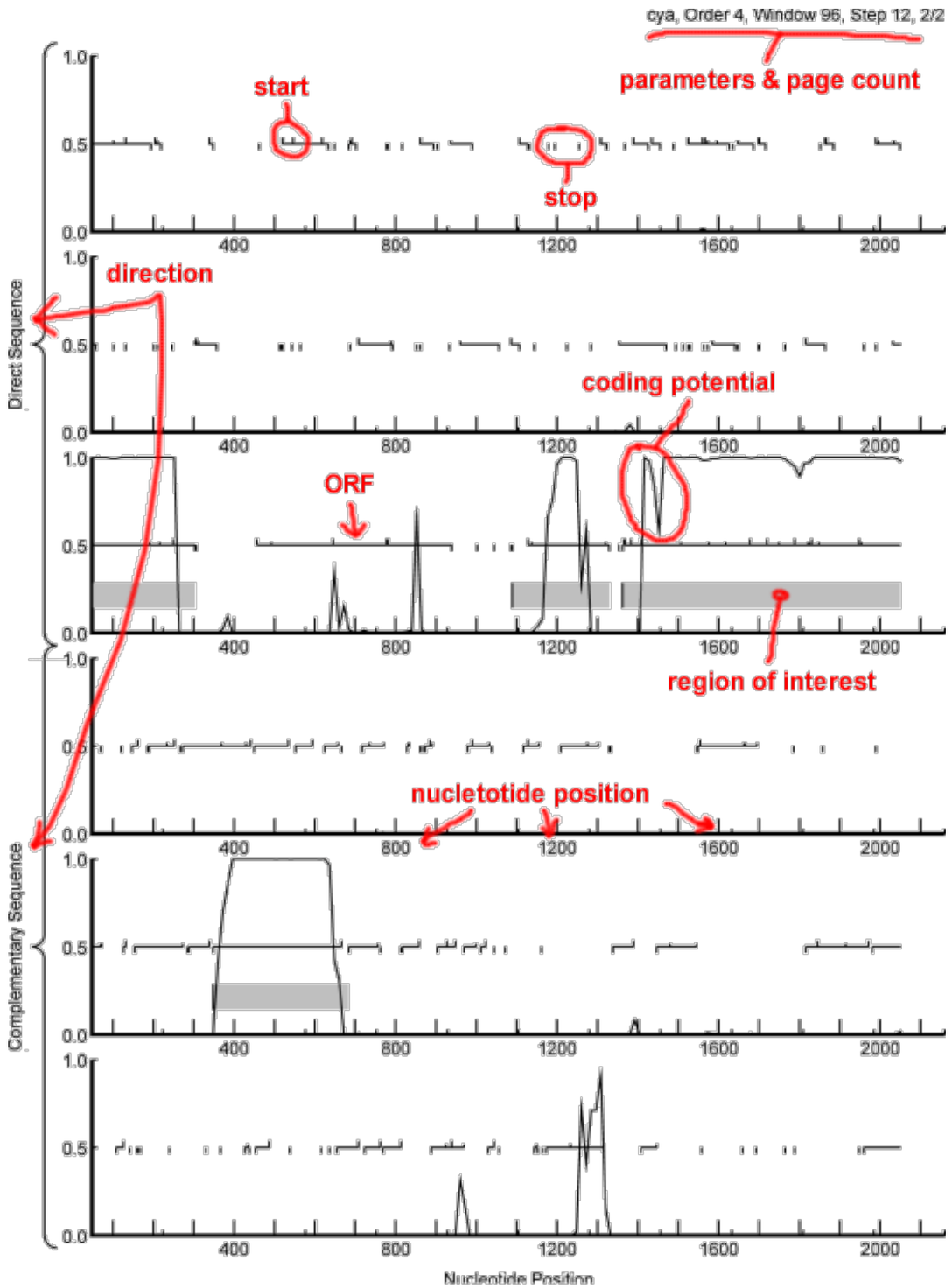
GeneMark

GeneMark evaluates the protein-coding potential of a DNA sequence (within a sliding window) by using Markov models of coding and non-coding regions for various prokaryotic species. This approach is sensitive to local variations of coding potential, and the GeneMark graph shows details of the coding potential distribution along a sequence. It has been used since 1995 to provide automatic gene annotation for the H. influenza, M. jannaschii, B. subtilis and E. coli genomes.

GeneMark.hmm

GeneMark.hmm predicts genes and intergenic regions in a sequence as a whole using hidden Markov models with a hidden state network reflecting the “grammar” of gene organization. It identifies the most likely parse of the whole sequence into protein coding genes (with possible introns) and intergenic regions. It is currently used as a microbial genome annotation tool by the NCBI.

GeneMark Example



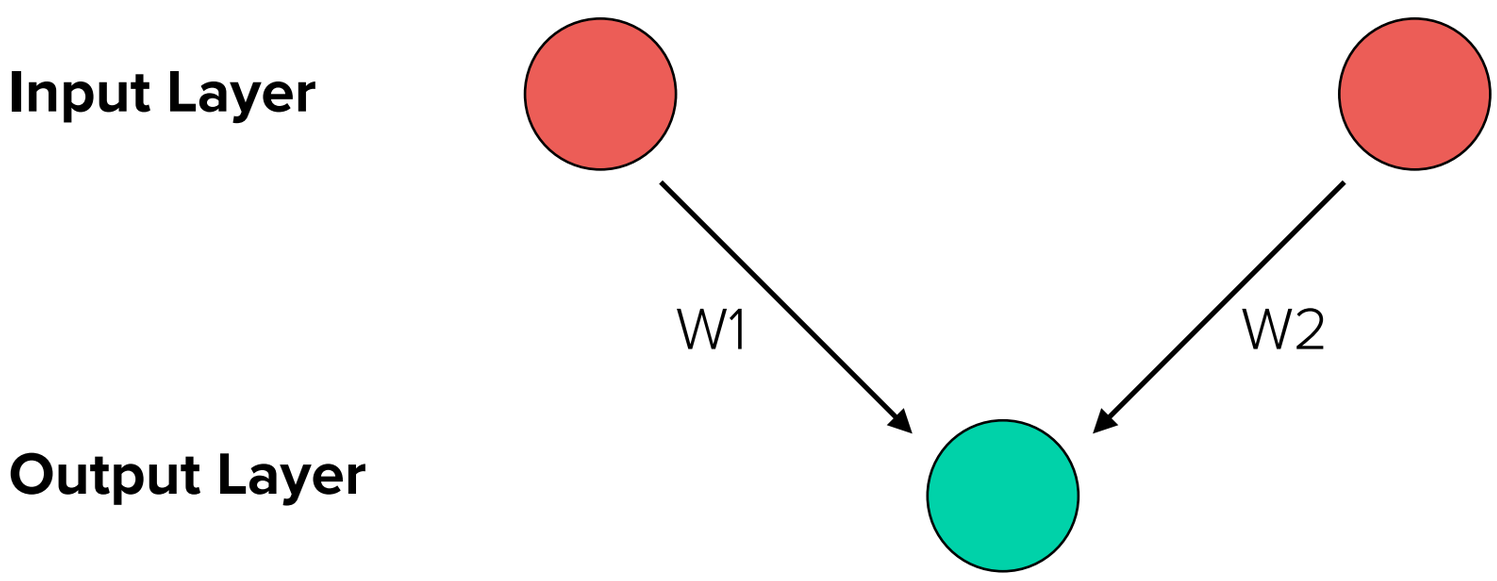
Source

http://bioweb.pasteur.fr/docs/genemark/images/cyay.gif

Neural Networks

Artificial Neural Network

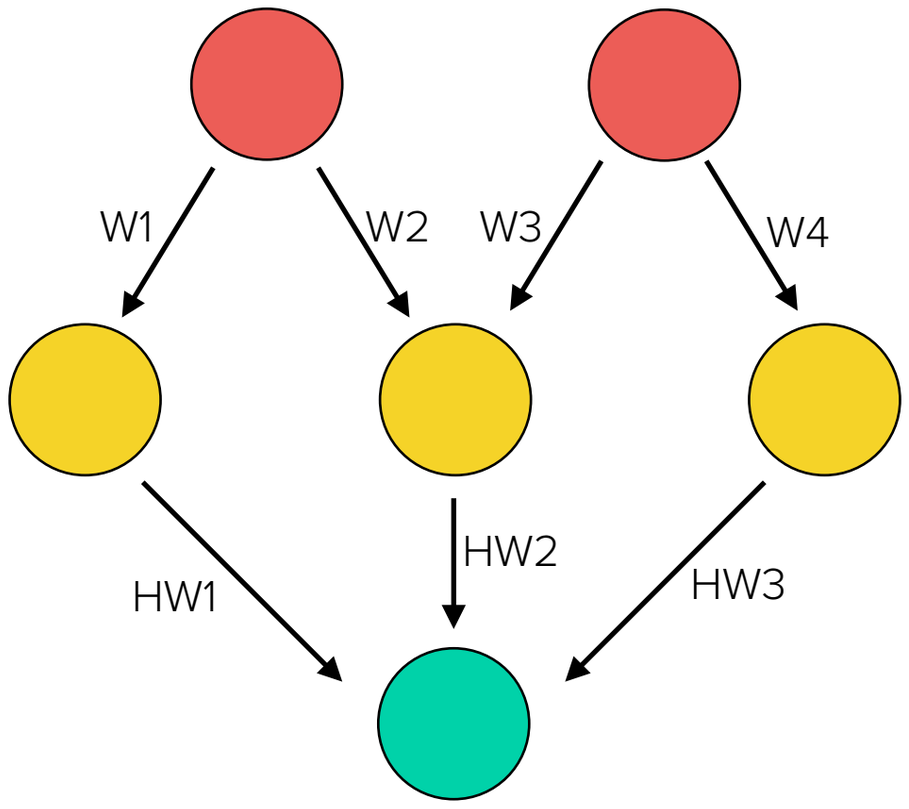
Artificial neural networks (ANN) are computational networks inspired by the connections of neurons in the brain. Artificial neurons are connected in a network that allows the output of some neurons to become the input of others, with weights assigned to each connection. The weights can be adjusted to better perform a particular task. A simple network is limited in the tasks it can perform.



Hidden Layers

Hidden Layers in Artificial Neural Networks

Adding even a single hidden layer to a neural network allows it to perform more complex calculations. Such networks have become widely used in pattern recognition, signal processing and machine learning.



Deep Neural Networks

Deep Neural Networks and Deep Learning

Deep neural networks (DNNs) are artificial neural networks (ANNs) with multiple layers between the input and output layer. They have become widely used in machine learning and performing complex tasks, including cell classification and predicting gene-function relationships in bioinformatics.

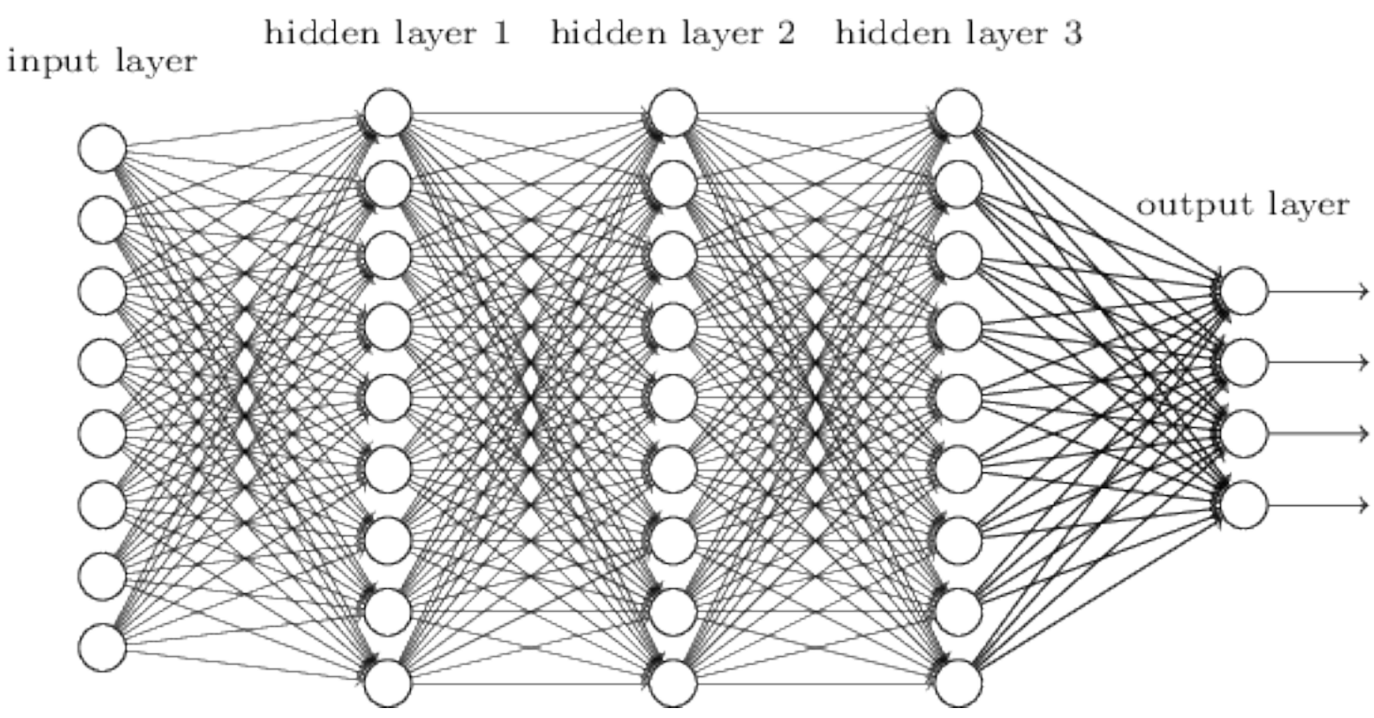


Image Source: Michael A. Nielsen, “Neural Networks and Deep Learning”, Determination Press, 2015.

Programming in Bioinformatics

Computer programming is simply how we instruct computers to perform tasks for us. These are some of the programming techniques and languages commonly used in bioinformatics:

- **Regular expressions (regex)**
- **Shell scripting (e.g bash), pipelines and redirects (Unix) and macros**
- **Structured Query Language (SQL)**
- **Perl practical extraction and reporting programming language**
- **Python programming language**
- **BioPython programming libraries**
- **R statistical computing and graphics language**
- **MATLAB numerical computing environment and language**
- **Java general object oriented programming language**

Regular Expressions

Regular expressions originated in the 1950s, when a mathematician, Stephen Kleene, described regular languages (finite languages that can be described with regular expressions) with a mathematical notation called regular sets. This notation could be used to easily match repeating patterns in strings, and has been widely adapted for this purpose by programmers.

Regular expressions are now a feature of many programming languages, text editors (such as BBEdit) and utilities (such as grep), and can be used in bioinformatics for pattern matching and reformatting text files (commonly known as data munging).

ICQB Course Website

https://microbiology.columbia.edu/icqb

The course website will be the home of all course information, including the syllabus, lecture notes, downloads, and any updates or other news.

ICQB Course Schedule

The course will meet Tuesdays, between 1:00 PM to 2:30 PM in HHSC 1307. A related hands-on session will follow each Thursday, from 4:30 PM to 5:30 PM in HHSC 1307.

Check the syllabus on the course website for the most up to date schedule, but the current schedule is:

September 13th	Introduction to Computational Biology
September 20th	Introduction to Internet Resources and Databases
September 27th	Introduction to Unix and Scripting
October 4th	Introduction to Programming
October 11th	Introduction to Python and BioPython
October 18th	Quantitative Analysis and Presentation of Visual Data
October 25th	Introduction to Statistics
November 1st	Data Visualization with R and RStudio
November 8th	No class (Election Day)
November 15th	Genomics (Anne-Catrin Uhlemann)
November 22nd	No class (Thanksgiving)
November 29th	Introduction to Sequence Analysis and RNA-Seq (Thomas Postler)
December 6th	Sequence Analysis and RNA-Seq (Thomas Postler)