


Elementary Sequence Analysis

Brian Golding, Dick Morton and Wilfried Haerty

Department of Biology
McMaster University
Hamilton, Ontario
L8S 4K1

These notes are in Adobe Acrobat format (they are available upon request in other formats) and they can be obtained from the website <http://helix.biology.mcmaster.ca/courses.html>. Some of the programs that you will be using in this course and which will be run locally can be found at <http://evol.mcmaster.ca/p3S03.html>.

The “blue text” should designate links within this document while the “red text” designate links outside of this document. Clicking on the latter should activate your web browser and load the appropriate page into your browser. If these do not work please check your Acrobat reader setup. The web links are accurate to the best of our knowledge but the web changes quickly and we cannot guarantee that they are still accurate. The links designated next to the JAVA logo, , require that JAVA be installed on your computer.

These notes are used in Biology 3S03. The purpose of this course is to introduce students to the basics of bioinformatics and to give them the opportunity to learn to manipulate and analyze DNA/protein sequences. Of necessity only some of the more simple algorithms will be examined.

The course will hopefully cover ...

- databases of relevance to molecular biology.
- some common network servers/sites that provide access to these databases.
- use of the internet to obtain sequence analysis software and data.
- methods of sequence alignment.
- methods of calculating genetic distance.
- methods of phylogenetic reconstruction.
- codon usage.
- methods for detecting gene coding regions.

The formal part of the course will consist of two approximately one hour lectures each week. Weekly assignments will be provided to practice and explore the lecture material. In addition there will be an optional tutorial to help students with these assignments or other problems. These assignments will be 40% of your grade and three, in class quizzes will make up the remainder.

We would appreciate any comments, corrections or updates regarding these notes.

Golding@McMaster.CA

Morton@McMaster.CA

HaertyW@McMaster.CA

Table of Contents in Brief

In order to speed download, I place here links to the individual chapters in pdf format. The contents of these are shown on the following 'Contents' pages but note that the links will function only for the individual chapter included here.

[Preliminaries](#)
[Basic Unix](#)
[Genomics](#)
[Databases](#)
[Sequence File Formats](#)
[Sequence Alignment](#)
[Distance Measures](#)
[Database Searching](#)
[Reconstructing Phylogenies](#)
[Pattern analysis](#)
[Exon analysis](#)

Contents

1	Preliminaries	1
1.1	Resources	1
1.1.1	Electronic Resources	1
1.1.2	Textbooks	2
1.1.3	Journal sources	7
1.2	Biological preliminaries	10
1.2.1	Some notes on terminology	10
1.2.2	Letter Codes for Sequences	11
2	Computer skills preliminaries	13
2.1	UNIX Operating Systems	13
2.1.1	Logging on/off	14
2.1.2	UNIX File System	14
2.1.3	Commands	17
2.1.4	Help	19
2.1.5	Redirection	20
2.1.6	Shells	20
2.1.7	Special 'hidden' files	21
2.1.8	Background Processes	21
2.1.9	Utilities	22
2.1.10	Editors	22
2.2	Exchange among computers	24
2.2.1	ssh	24
2.2.2	Mail	24
2.3	Scripts-Languages	25
2.4	Obtaining LINUX	25
3	Genomics	27
3.1	Where the data comes from	27
3.2	How DNA is sequenced	27

3.3	First Generation Methods	28
3.4	The reality of sequencing includes errors	32
3.5	From sequence to genome	33
3.6	Second generation sequencing	37
3.7	Paired sequences	42
3.8	Third generation: Yesterday's new sequencing?	44
3.9	Types of sequencing	46
3.9.1	RNA-seq	46
3.9.2	Exome sequencing	46
3.9.3	ChIP-seq	46
3.10	Other kinds of biological data	47
3.10.1	Microarrays	47
3.10.2	Mass spectrometry methods	50
3.10.3	Textual information	52
4	Databases	55
4.1	Introduction	55
4.2	N.C.B.I.	58
4.3	E.M.B.L.	63
4.4	D.D.B.J.	64
4.5	SwissProt	64
4.6	Organization of the entries	67
4.7	Other Major Databases	68
4.8	Remote Database Entry retrieval	71
4.8.1	Entrez	71
4.8.2	NCBI retrieve	74
4.8.3	EMBL get	75
4.8.4	Others	75
4.9	Reliability	76
5	Sequence File Formats	79
5.1	Genbank/EMBL	79
5.2	FASTA	81
5.3	FASTQ	82
5.4	Stockholm format	83
5.5	GDE	85
5.6	NEXUS	87
5.7	PHYLIP	88
5.8	ASN	89

5.9	BSML format	92
5.10	PDB file format	92
6	Sequence Alignment	97
6.1	Dot Plots	97
6.1.1	The Exact Way	97
6.1.2	Identity Blocks	99
6.2	Alignments	106
6.2.1	The Needleman and Wunsch Algorithm	106
6.2.2	The Smith-Waterman Algorithm	109
6.3	Testing Significance	110
6.4	Gaps and Indels	113
6.4.1	“Natural” Gap Weights - Thorne, Kishino & Felsenstein	113
6.5	Multiple Sequence Alignments	114
7	Distance Measures	117
7.1	Nucleotide Distance Measures	117
7.1.1	Simple counts as a distance measure	117
7.1.2	Jukes - Cantor Correction	118
7.1.3	Kimura 2-parameter Correction	120
7.1.4	Tamura - Nei Correction	120
7.1.5	Uneven spatial distribution of substitutions	121
7.1.6	Synonymous - nonsynonymous substitutions	122
7.2	Amino acid distance measures	122
7.2.1	PAM Matrices	123
7.2.2	BLOSUM Matrices	125
7.2.3	GONNET Matrix	126
7.3	Gap Weighting	127
8	Database Searching	129
8.1	Are there homologues in the database?	129
8.1.1	FASTA	129
8.1.2	BLAST	137
8.1.3	MPsrch	144
8.2	BLOCKS	148
8.2.1	BLOCKS output	149
8.2.2	Getting the Block	150
8.3	SSearch	156
8.4	Why you should routinely check your sequence	156

9	Reconstructing Phylogenies	157
9.1	Introduction	157
9.1.1	Purpose	157
9.1.2	Trees of what	157
9.1.3	Terminology	159
9.1.4	Controversy	161
9.2	Distance Methods	161
9.3	Parsimony Methods	163
9.4	Other Methods	166
9.4.1	Compatibility methods	166
9.4.2	Maximum Likelihood methods	166
9.4.3	Method of Invariants	167
9.4.4	Quartet Methods	168
9.5	Consensus Trees	170
9.6	Bootstrap trees	170
9.7	Warnings	174
9.8	Available Packages	175
9.9	PHYLIP	178
9.9.1	PHYLIP Contents	178
10	Pattern Analysis	191
10.1	Base Composition: first order patchiness	191
10.1.1	Genome Patchiness	191
10.2	Dinucleotide Composition: second order patchiness	192
10.3	Strand Asymmetry	193
10.3.1	Chargaff's Rules	193
10.3.2	Replication Asymmetry	194
10.3.3	Transcriptional Asymmetry	195
10.3.4	Codon Selection	196
10.4	Simple Sequence Repeats	196
10.5	Sequence Complexity	196
10.5.1	Information Theory	196
10.5.2	Sequence Window Complexity	198
10.6	Finding Pattern in DNA Sequences	199
10.6.1	Consensus Sequences	199
10.6.2	Matrix Analysis of Sequence Motifs	200
10.6.3	Sequence Conservation and Sequence Logos	201
11	Exon Analysis	205

11.1	Open Reading Frames	205
11.2	Gene Recognition	205
11.2.1	Splice Sites	206
11.2.2	Codon Usage	207
11.2.3	Gene Prediction Software	210
11.2.4	Hidden Markov Models (HMM)	211
11.2.5	Comparison of Programs	211

Chapter 1

Preliminaries

A reminder: if your Acrobat reader is correctly set, the “blue text” should designate links within this document while the “red text” designate links outside of this document. Clicking on the latter should activate your web browser and load the appropriate page into your browser.

1.1 Resources

There are many resources that one can make use of to study bioinformatics and they are becoming increasingly available to the general public. These notes are my attempt at a small contribution toward this growing body of ‘on-line’ literature, software, data and knowledge.

Please note that bioinformatics is inherently a multi-disciplinary field making use of biological, mathematical, statistical and computer science knowledge. As such any resources available for any of these disciplines will be of use in bioinformatics. The more skilled that you are in any one of these areas the better off you will be. But you should have a basic minimum knowledge from each of these fields to study bioinformatics. There is a growing body of information available that is specific for bioinformatics.

1.1.1 Electronic Resources

You should note that there are many other valuable online resources that are available to you. Some older one include the VSNS (Virtual School of Natural Sciences) [Biocomputing](#) course notes from Bielefeld Germany.

- Introduction
- Pairwise Sequence Alignments
- Networking
- Multiple Alignment
- Mathematical Basis of Molecular Phylogenetics
- Genetic Algorithms and Protein Folding

Another VSNS course is [Principles of Protein Structure](#) running out of Birbeck College.

- Overview of Protein Synthesis

- Primary Structure
- Protein Geometry
- Overview of Molecular Forces
- Secondary Structure
- Super-Secondary Structure
- Tertiary Structure
- Protein Folds
- Quaternary Structure
- Protein Interactions

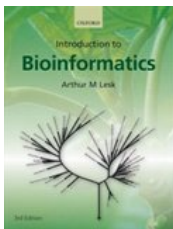
Also excellent is the “lecture notes” site for the [Algorithms in Molecular Biology](#) from the Univ. of Washington. From the Max Planck Institutes in Germany [online-lectures](#), and many more as listed at NYU <http://www.med.nyu.edu/rcr/rcr/btr/complete.html>.

There are many other resources that you should be able to discover on your own (including one that has garnered the domain name bioinformatics.org), As a resource for your future work consider CRdata.org. Old but still very useful web pages at the [MolBiol toolkit](#). Each of these are a fabulous resource and often they are “straight from the horse’s mouth”. You should make frequent use of these resources and others throughout this course (and perhaps you can bring to our attention the ones that you find most valuable that are not listed).

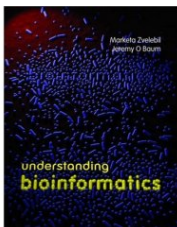
There are also many software packages that provide you with access to a collection of programs that deal with bioinformatics. For example, if you have cash, the famous [MatLab](#) software suite provides a toolbox for [bioinformatics](#). For those with less cash, there are interesting projects – [Biolinux](#), [Bioknoppix](#), [Vigyaan](#) – that provide you with a bootable CD image. Simply burn the CD (it is free) and then boot from the CD. This provides a free computer system with lots of bioinformatic, biomolecular software at your fingertips (nothing to install, nothing to change on your computer, simply remove CD and reboot when done). There are many other software sources that will be explored in this course (and provided through the links of these notes). For our purposes some of the software that will be discussed below has been provided for you at <http://evol.mcmaster.ca/p3S03.html>.

1.1.2 Textbooks

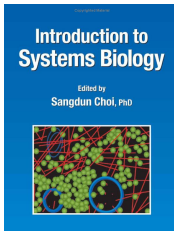
There are now an enormous number of books available that deal with sequence analysis and bioinformatics in biology. A selection of just a few that have been published in the last few years are



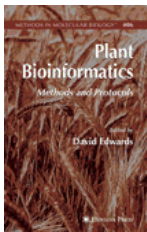
• Introduction to Bioinformatics Introduction to Bioinformatics by A.Lesk. 2008, Oxford Univ. Press



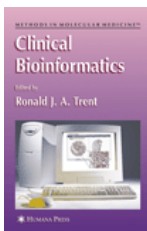
• Understanding Bioinformatics by M.Zvelebil, J.Baum. 2007, Garland Science



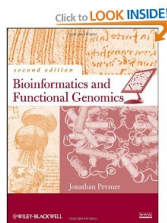
• Introduction to Systems Biology by S.Choi. 2007, Humana Press



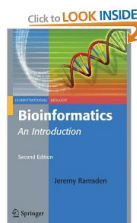
• Plant Bioinformatics edited by D.Edwards. 2007, Humana Press



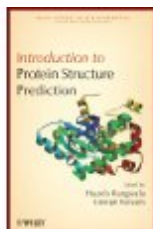
• Clinical Bioinformatics edited by R.Trent. 2007, Humana Press



• Bioinformatics and Functional Genomics by J. Pevsner. 2009, Wiley-Blackwell



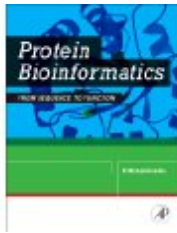
• Bioinformatics an introduction by J. Ramsden. 2009, Wiley-Blackwell



• Protein Structure Methods and Algorithms by H. Rangan and G. Karypis. 2010, Wiley



• Problem Solving Handbook in Computational Biology and Bioinformatics by L.S. Heath and N. Ramakrishnan. 2010, Springer



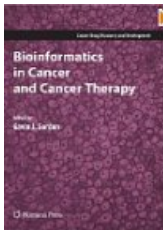
• Protein Bioinformatics: From Sequence to Function by M.M. Gromiha. 2010, Academic Press



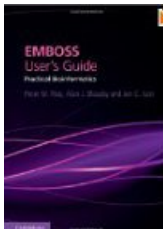
• Knowledge-Based Bioinformatics: From analysis to interpretation by G. Alterovitz and M. Ramoni. 2010, Wiley



• Clustering in Bioinformatics and Drug Discovery by J.D. MacCuish and N.E. MacCuish. 2010, Chapman and Hall



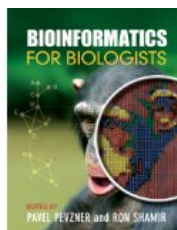
• Bioinformatics in Cancer and Cancer Therapy (Cancer Drug Discovery and Development) by G.J. Gordon. 2011, Humana Press



• EMBOSS User's Guide: Practical Bioinformatics by P.M. Rice et al. 2011, Cambridge Univ Press



• Bioinformatics: High Performance Parallel Computer Architectures (Embedded Multi-Core Systems) by B.Schmidt 2011, CRC Press

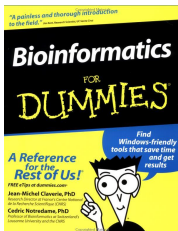


• Bioinformatics for Biologists by P. Pevzner and R. Shamir 2011, Cambridge Univ Press



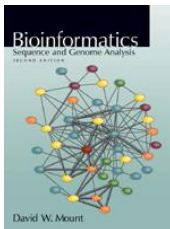
• Bioinformatics for Biomedical Science and Applications by K.-H. Liang, 2012, Biohealthcare Publ

and there is EVEN now a book from the popular “dummy” series

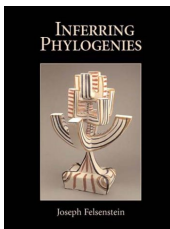


• Bioinformatics for Dummies by J.-M. Claverie and C. Notredame 2003.

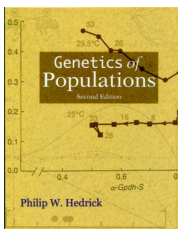
A selection of the more important and recommended texts are



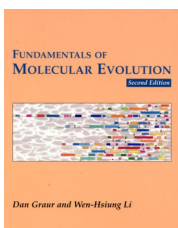
• Bioinformatics: Sequence and Genome Analysis (2nd) by D.W. Mount 2004, Cold Spring Harbor Laboratory Press. (*highly recommended*).



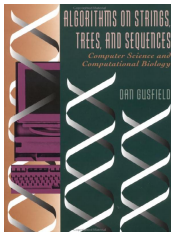
• **Inferring Phylogenies** by J. Felsenstein 2003, Sinauer Associates. (*highly recommended*).



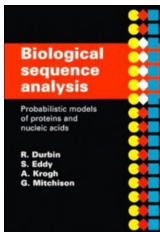
• Genetics of Populations by P.W. Hedrick 2000, Jones and Bartlett.



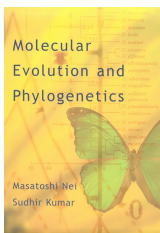
• **Fundamentals of Molecular Evolution** by D. Graur and W.H. Li, 1999, Sinauer.



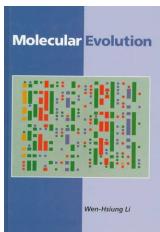
Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology by D.Gusfield 1997, Cambridge University Press (*highly recommended*).



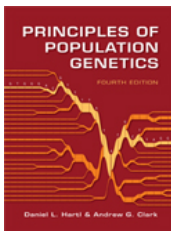
Biological Sequence Analysis by R.Durbin, S.Eddy, A.Krogh and G.Mitchison 1998, Cambridge Univ. Press (*highly recommended*).



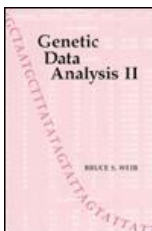
Molecular Evolution and Phylogenetics by M.Nei and S.Kumar, 2000 Oxford University Press.



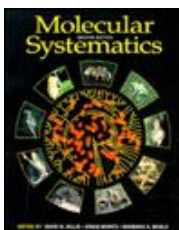
Molecular Evolution by W.H.Li 1997, Sinauer.



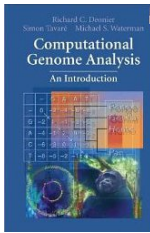
Principles of Population Genetics by D.L.Hartl and A.G.Clark 1997, Sinauer (*highly recommended*).



Genetic Data Analysis II by B.Weir 1996, Sinauer (*highly recommended*).



Molecular Systematics by D.Hillis, C.Moritz and B.Mable 1996, Sinauer.



• **Computational Genome Analysis** by R.C.Deonier, S.Tavare, M.S.Waterman 2005, Springer.

In addition to these there are many texts on evolution, on DNA and on proteins that have useful chapters and sections on sequence analysis.

1.1.3 Journal sources

Again there are many periodicals relevant to sequence analysis. Besides the general science journals such as Nature, PNAS, Science, EMBO, ... there are several which are more specific to molecular evolution, to computers in biology, and to sequence analysis. Some of these journals include ...



• **Applied Bioinformatics**



• **Bioinformatics** (formerly entitled **Computer Applications In The Biosciences : CABIOS**).



• **Briefings in Bioinformatics**



• **BMC Bioinformatics**



• **BMC Evolutionary Biology**



• **Bulletin of Mathematical Biology**



Genome Biology



Genome Biology and Evolution



Genome Research



Genomics

- In Silico Biology



Journal of Computational Biology



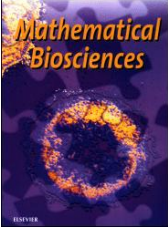
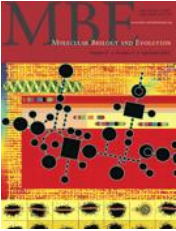
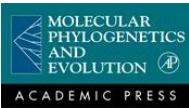
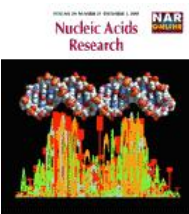


Journal Evolutionary Applications



J. Mathematical Biology



Journal of Molecular Evolution

-  **Mathematical Biosciences**
-  **Molecular Biology and Evolution**
-  **Molecular Phylogenetics and Evolution**
-  **Nucleic Acids Research**
-  **PLOS Computational Biology**
-  **Systematic Biology**

and in the medical sciences there are many (!) more, including

-  **Journal of Biomedical Informatics**
-  **Journal of the American Medical Informatics Association**



• [Medical Informatics and the Internet in Medicine](#)

for a more complete listing of medically related journals see [MedBioWorld](#).

To find individual papers on specific topics there is the Swiss sequence analysis bibliographic database [SeqAnalRef](#) or the more general search engines from N.C.B.I. [Entrez](#) for access to papers in Medline and the more recent [Google Scholar](#) pages.

1.2 Biological preliminaries

We will assume throughout the remainder that some familiarity with basic biology has been obtained. We do not assume any more knowledge of mathematics than can be obtained at a high school level.

1.2.1 Some notes on terminology

There are some terms that will be used here, that are commonly abused. Unfortunately, I too will use some terms that are not precise so you should be aware of the proper definitions (the following are modified from Futuyma 1986, Evolutionary Biology, Sinauer Assoc.).

Homology Contrary to some statements in other bioinformatic texts, homology and similarity are not the same thing. A trait from two different species or taxa are said to be similar if they have some resemblance of one to another. Homology means a great deal more. Two traits from a different species or taxa are homologous if they are derived (with or without modification) from a common ancestor.

In general when working with sequences, one assumes homology if one finds excessive similarity between the two sequences. However, you should be aware that this is an inference that should be consciously made.

Example: The traditional example is that of the wings of birds and bats. Their wings are similar in that they enable flight, have the same name and have similar aerodynamic constraints but they are not homologous. They are not homologous because the common ancestor of both birds and bats did not have wings, rather wings evolved within each group separately.

Mutations A mutation is an error in the replication of a nucleotide sequence. It may encompass one or many nucleotides and in complicated situations may involve disjoint nucleotides. They can be caused by internal errors of metabolism or by external agents such as radiation.

Substitutions Mutations are not substitutions. Substitutions are differences in two sequences (generally the descendant from the ancestral) caused originally by mutations but which have been acted on by selection.

Example: Because substitutions have been exposed to selection, the frequency of occurrence of individual substitutions and mutations are quite different. In general substitutions at the second position of a codon are (almost always) much less frequent than those in the third codon position. This is because a change at the second codon position will alter the amino acid encoded but this is not always the case for changes at the third codon position. By contrast, we expect mutations to occur equally frequently at each of the codon positions.

Replacements The term replacement is suggested to be used when differences between amino acid sequences are observed.

Table 1.1: One letter amino acid codes

Alanine	A	Arginine	R	Asparagine	N
Aspartic acid	D	Cysteine	C	Glutamic acid	E
Glutamine	Q	Glycine	G	Histidine	H
Isoleucine	I	Leucine	L	Lysine	K
Methionine	M	Phenylalanine	F	Proline	P
Serine	S	Threonine	T	Tryptophan	W
Tyrosine	Y	Valine	V	Unknown	X

1.2.2 Letter Codes for Sequences

To store a large amount of data on a computer it would be quite inefficient to store the amino acids as “Glutamic acid” or to store ambiguous nucleotides as “A or G”. For this reason there are standard codes to represent amino acids and nucleotides. Both of these are one letter codes and can be stored on electronic media with reasonable efficiency.

Amino acids have in the past often been designated by a three letter code. This three letter code is not suitable for electronic media and is now largely obsolete. The standard one letter amino acid codes are shown in Table 1.1. Also commonly in use are B to represent either Aspartic acid or Asparagine and Z to represent either Glutamic acid or Glutamine.

There are also standard one letter codes to represent nucleotides. While most people are familiar with the simple codes of A, C, G, T, and U there are more extensive codes to include ambiguities in the nucleotides. The extended one letter code for nucleotides is given in Table 1.2. The complete generality of this code is seldom used. More common is the use of only part of the extended code

R A or G
 Y T or C
 N A,T,C or G
 X unknown

Some programs prefer to store RNA codes rather than DNA codes. In general T and U can often be taken as synonyms.

Table 1.2: One letter nucleotide codes.

Based on Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Molecular Biology and Evolution 3:99-108 (1986).

Guanine	G	G
Adenine	A	A
Thymine	T	T
Cytosine	C	C
Purine	G or A	R
Pyrimidine	T or C	Y
Amino	A or C	M
Keto	G or T	K
Strong (3H bonds)	G or C	S
Weak (2H bonds)	A or T	W
Not G	A or C or T	H
Not A	G or T or C	B
Not T	G or C or A	V
Not C	G or A or T	D
Any	G or C or T or A	N
Unknown	?	X