Molecular biology for bioinformatics - a very short primer

Biology is the science of living things - what they are, how they work, how they interact, and how they evolve.

1 The goals of molecular biology

- Sequencing and comparing full genomes of organisms.
- Identifying the genes and determining the foundations of the proteins they encode.
- Understanding gene expression.
- Understanding genetic diseases.
- Understanding evolution and evolutionary history.
- Understanding proteins, which means predicting the folding of the amino acid sequence, and characterizing the function of the protein based on this folding.
- Constructing synthetic proteins, which means creating amino acid sequences, such that the protein produced from these have a desired function.

2 Polymers

Chemical characteristics of organisms, particularly polymers, can be readily quantified and correlated using logical and statistical methods.

Three types of polymers (DNA, RNA, proteins) play an essential role in biology, either as carriers of information, or as activating molecules of the metabolism.

- DNA sequences are the information-containing molecules and are composed of nucleotides from an alphabet of four letters: a, c, g and t.[1]
  The DNA of an organism plays a central role in its existence. It is arranged in the form of chromosomes. These strings may be millions of nucleotides long, measured in base pairs (bp).
  The entire set of genetic information of an organism is called its genome.

  There are the following genome sizes of certain species[2]:

[1] The meaning of these symbols we will describe below.

[2] Apart from our own species, the organisms listed are important in molecular biology and genetic research

Species	Number of chromosomes (diploid)	Genome Size (haploid) (base pairs)
Bacteriophage λ (viruse)	1	$5 \cdot 10^4$
Escherichia coli (bacterium)	1	$5 \cdot 10^6$
Saccharomyces cerevisiae (yeast)	32	$1 \cdot 10^7$
Caenorhabditis elegans (worm)	12	$1 \cdot 10^8$
Drosophila melanogaster (fruit fly)	8	$2 \cdot 10^8$
Homo sapiens (human)	46	$3 \cdot 10^9$

Fitch [7] gives the following exemplary genome sizes:

Domain	Organism	Size (bp)
Viruses	HIV	$9 \cdot 10^3$
Bacteria	E. coli	$4 \cdot 10^6$
Eukaryotes	mammals	$3 \cdot 10^9$

Roughly speaking, the order of genome size is kbp, Mbp and Gbp for Viruses, Prokarya and Eukarya, respectively.

- Proteins, which are the operational molecules, are composed of chains of amino acids, called polypeptides, each from an alphabet of 20 letters:

	One-letter code	Three-letter code	Name
1	A	ala	alanine
2	C	cys	cysteine
3	D	asp	aspartic acid
4	E	glu	glutamatic acid
5	F	phe	phenylalanine
6	G	gly	glycine
7	H	his	histidine
8	I	ile	isoleucine
9	K	lys	lysine
10	L	leu	leucine
11	M	met	methionine
12	N	asn	asparagine
13	P	pro	proline
14	Q	gln	glutamine
15	R	arg	arginine
16	S	ser	serine
17	T	thr	threonine
18	V	val	valine
19	W	trp	tryptophan
20	Y	tyr	tyrosine

Typical proteins contain about 300 amino acids (aa), but there are proteins with fewer than 100 or as many as 5000 aa.

- RNA sequences, which stand between DNA and protein, are composed of nucleotides from an alphabet of four letters: a, c, g and u.

One-letter code	Name
a	adenine
c	cytosine
g	guanine
t	tymine
u	uracil
r	purine (a or g)
y	pyrimidine (c or t)

The Central Dogma of Molecular Biology describes the interaction of these polymers:

- DNA acts as a template to replicate itself;
- DNA is also transcribed into RNA; and
- RNA is translated into protein.

More precisely,

- Integral form: DNA makes RNA makes protein.
- Differential form: Changed DNA can make changed protein.

This runs in the following steps:

1. Replication of DNA.
   Each strand in a DNA is a chemical "mirror image" of the other. If there is an a on one strand, there will always be a t in the same position on the other strand, and vice versa; if there is a c on the one strand, its "partner" on the other strand will always be a g, and vice versa.
   When a cell divides to form daughter cells, DNA is replicated by untwisting the two strands and using each strand as a template to produce its chemical mirror image.

2. Transcription of DNA.
   DNA also act as a blueprint for RNA, more exactly three main types of RNA: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). They carry information from the genome to the ribosomes, the protein synthesis apparatus in a cell.

3. Translation of mRNA.
   The information in an mRNA will be translated into a sequence of amino acids, creating a polypeptide molecule.[3]

[3]The coding scheme we will discuss below.

3 Proteins

> Organic chemistry is the chemistry of carbon compounds. Biochemistry is the study of carbon compounds that crawl.
>
> Mike Adam

Structural proteins act as tissue building blocks, whereas other proteins known as enzymes act as catalysts of chemical reactions. Proteins are not laid out simply as straight chains of amino acids. The fact that they curl and fold into complex forms plays a crucial role in determining the distinctive biological properties of each protein.

We distinguish the following structural levels for proteins:

1. The primary structure is the amino acid sequence.
2. The secondary structure is the arrangement of the amino acids in space.
3. The tertiary structure is the three-dimensional folding pattern, which is superimposed on the secondary structure.
4. The quarternary structure is the composition of two or more polypeptides.

For instance human insulin is composed by two words (chains):

A: gly ile val glu gln cys cys thr ser ile cys ser leu tyr glu leu glu asn tyr cys asn.

B: phe val asn gln his leu cys gly ser his leu val glu ala leu tyr leu val cys gly glu arg gly phe phe tyr thr pro lys thr.

The function of a protein being a direct consequence of its three-dimensional structure, shortly written by

$$\text{Sequence} \Rightarrow \text{Structure} \Rightarrow \text{Function}.$$

4 Genes

Historically, the heritable factors which determine much of the physical make up of organisms are called genes.

4.1 Genotypes and Phenotypes

Usually there are several different forms one gene can have. These forms are called allels.

A combination of allels describes the make-up of an individual, more exactly:

- The genetic make-up of an individual is its genotype.
- The expression of the genes of an individual is its phenotype.

The AB0 blood groups in humans are determined by a system of three allels A,B and 0. The phenotypes resulting from the various unions of gametic genotypes are shown in the following table:

male/female	A	B	0
A	A	AB	A
B	AB	B	B
0	A	B	0

4.2 The DNA

Genes themselves are composed of a more fundamental molecule called deoxyribonucleid acid or DNA, which has two extremely important properties:

- It contains the information of how organisms should be built;
- It can be replicated, so that these instructions are passed on to successive generations.

DNA (and RNA) are polymer sequences composed of a small number of chemically similar compounds. The individual units are called nucleotides, each made up of three distinct parts: a cyclic base a, c, g or t (or u, respectively), a cyclic sugar deoxyribose (or ribose, respectively), and a phosphate group.
Chargoff's rule says that in a double-stranded DNA there are always equal amounts of a's and t's and also equal amounts of g's and c's, which is an immediate consequence of the pairing of these nucleotides.[4]

The whole genetic information of a organism (a species, all species) is called the genome.

4.3 Mutations

Although the DNA replication is a very accurate system, it does not work correctly on every occasion. Sometimes errors, called mutations, can creep into the process.

There are many different types of mutation:

DNA mutations These point mutations can be placed in the following categories:

1. Transitions occur when a purine nucleotide (a or g) is substituted for another purine; or a pyrimidine (c or t) is replaced by another pyrimidine.
2. Transversions occur when a pyrimidine is substituted for a purine, or vice versa.
3. Indels lead to *in*sertions or *del*etions of nucleotides.
   Indels change the nucleotide sequence such that the grouping of the nucleotides into tripletts during the translation is no longer the same.

[4] Note that the ratio $(a - t) : (c - g)$ base pairs can vary widely from species to species.

Chromosomal mutations These mutations can occur at the chromosomal level, classified in the following way:

1. The number of chromosomes in the cell is altered.
2. An inversion is a break in the chromosome such that the broken part flips end-for-end before rejoining the rest of the chromosome in the reverse direction.
3. In a translocation a part of the broken chromosome may join another chromosome.
4. If breaks occur in the chromosome twice it is called a duplication.
5. A deletion is given if a part of the broken chromosome is lost.

4.4 The genetic code

The crucial concept is that there is a genetic code which specifies how combinations of the four bases encode each of the 20 amino acids. Each amino acid is specified by a consecutive and non-overlapping sequence of three of the bases. Such a three-letter word is called a codon.

	u	c	a	g	
u	phenylalanine	serine	tyrosine	cysteine	u
	phenylalanine	serine	tyrosine	cysteine	c
	leucine	serine	*punctuation*	*punctuation*	a
	leucine	serine	*punctuation*	tryptophan	g
c	leucine	proline	histidine	arginine	u
	leucine	proline	histidine	arginine	c
	leucine	proline	glutamine	arginine	a
	leucine	proline	glutamine	arginine	g
a	isoleucine	threonine	asparagine	serine	u
	isoleucine	threonine	asparagine	serine	c
	isoleucine	threonine	lysine	arginine	a
	methionine	threonine	lysine	arginine	g
g	valine	alanine	aspartic acid	glycine	u
	valine	alanine	aspartic acid	glycine	c
	valine	alanine	glutamic acid	glycine	a
	valine	alanine	glutamic acid	glycine	g

For instance the following messenger RNA sequence is decoded as follows:

1. Read the sequence:

auggcugcuauucccacccacaauaugcccuga

2. Decompose the sequence into successive triples (codons):

   aug gcu gcu auu ccc acc cac aau aug ccc uga

3. Translate each triple in the corresponding amino acid:

   methionine alanine alanine isoleucine proline threonine
   histidine isoleucine methionine proline *punctation*

5 Classifications

Classifications are of great relevance in biology. Here a class is defined as a group of entities which are

- similar, and
- related.

In the book *The System of Nature* Linnaeus introduced a system still in use today. He gave every species two Latinized names; the first for the group it belongs to, the genus; and the second for the particular organism itself. Today we divide life into
- Domain[5];
- Kingdom;
- Phylum;
- Class;
- Order;
- Family;
- Genus;
- Species.

More or less all of these groups are artificial, insofar as their members are categorized according to agreed-upon levels of similarity rather than precise definitions. The exceptions are species, which are defined as a maximal group of individual organisms that are able to interbreed and produce fertile offspring.

For example

group \ species	human	fruit fly
Domain	Eukarya	Eukarya
Kingdom	Animalia	Animalia
Phylum	Chordata	Arthropoda
Class	Mammalia	Insecta
Order	Primata	Diptera
Family	Hominidae	Drosophilidae
Genus	Homo	Drosophila
Species	*sapiens*	*melanogaster*

[5]There are three domains. The first two, Bacteria and Archea, are made up of many microscopic single-celled organisms. The third domain, Eukarya, is diverse.

6 Mendel's laws

A Mendelian population may be considered to be a group of reproducing organisms with a relatively close of genetic relationship. We consider all the gametes produced by a Mendelian population as a hypothetical mixture of genetic units from which the next generation will develop. In such organisms adults produce female and male gametes[6], which fuse to form zygotes, which develop and mature to adulthood. These factors determining various traits are passed through the generations. It is of great interest to describe and to understand this process.

Mendel published the result of his genetic studies on the garden pea in 1866 and thereby laid the foundation of modern genetics. In this paper Mendel proposed some basic genetic principles.

Principle of segregation: From any one parent, only one allelic form of a gene is transmitted through a gamete to the offspring.

Principle of independent assortment: The segregation of one factor pair occurs independently of any other factor pair.

We discuss several specific cases:

- Suppose that there are two and only two alleles A and a that are to be found at a locus. A given individual may then have one of three genotypes: the homozygotes AA or aa or the heterozygote Aa. The allel A may be dominant over a, so that we cannot distinguish between the appearance of AA or Aa. Generation 0 is known as the parental generation ($P = F_0$), and generation n as the nth filial generation (F_n). Then

$$F_0: \quad AA \quad aa \tag{1}$$

is followed by the generation

$$F_1: \quad Aa \quad aA \tag{2}$$

which is uniform. But in the next generation we find

$$F_2: \quad AA \quad Aa \quad aA \quad aa \tag{3}$$

with a ratio of $3:1$ regarding the phenotype of the dominant allel. This leads to the following phenotypes in the next generation:

F_3 :

	AA	Aa	aA	aa
AA	$4A$	$4A$	$4A$	$4A$
Aa	$4A$	$3A+1a$	$3A+1a$	$2A+2a$
aA	$4A$	$3A+1a$	$3A+1a$	$2A+2a$
aa	$4A$	$2A+2a$	$2A+2a$	$4a$
together	$16A$	$12A+4a$	$12A+4a$	$8A+8a$

(4)

[6] for example eggs and sperms in humans

Altogether $48A + 16a = 3A + 1a$. Thus

$$\#A : \#a = 3 : 1. \tag{5}$$

We find the same situation for F_4 and so on.

- Mendel considered peas with pure-bred rounded yellow and wrinkled green phenotypes having genotypes RRYY and WWGG, respectively. Then all of those in F_1 have genotype RWYG and phenotype rounded yellow.
  F_2 is produced by random mating in the following way:

	RY	RG	WY	WG
RY	RY	RY	RY	RY
RG	RY	RG	RY	RG
WY	RY	RY	WY	WY
WG	RY	RG	WY	WG

This shows the phenotype that results from each union of gametic genotypes. Each of these possibilities is equally likely, so that

$$\#RY : \#RG : \#WY : \#WG = 9 : 3 : 3 : 1. \tag{6}$$

7 Darwin's evolution

Biological evolution is part of the general idea that the universe has changed through time.[7] Moreover, Dobzhansky said that "Nothing in biology makes sense except in the light of evolution."

In his fundamental book *The origin of species* [3] Darwin created a theory of evolution, the core of which is described by the following three facts.

- Reproduction;
- Mutation;
- Selection.

(Mayr [16] added a fourth fact: Catalyse.)

Evolution, by definition, is the change in allelic frequencies in populations from generation to generation. Evolution by natural selection depends on five factors:

Excess progeny: More offsprings are produced than can survive to reproduce.

Variability: The characteristics of living entities differ among individuals of the same species.

Heritability: Many differences are the result of heritable genetic differences.

[7] Don't confuse the origin of life itself with evolution, the two are conceptually separate.

Differential adaptedness: Some differences affect how well adapted an organism is.

Differential reproduction: Some differences in the quality adaptation are reflected in the number of offspring successfully reared.

Evolutionary biologists are concerned with both

- The history of life; and
- The processes and mechanisms that produced the tree of life.

Natural selection is evolution's major cause. The principle is simple: Generate a varity of possible solutions, and then pick one that works good for the problem. So the essence of natural selection is

1. Genetic variation within a population,
2. An environmental condition favors some of these variations more than others, and
3. Differential reproduction of the individuals who happen to have the favored variations.

Note that

Natural selection is the "survival of the fit enough";

not the well-described phrase of "survival of the fittest", it is not expected that optimal structures will always be the end result. We will see that "survival of the fittest" can be false, and cannot be a scientific term.

It is crucial to define the term "fitness" for a genotype. We distinguish

Absolute fitness: is defined in terms of its reproductive success.

Relative fitness: is the ratio of its absolute fitness to the absolute fitness of a reference genotype.

Darwinian natural selection and Mendelian genetics came together as scientist recognized the powerful support Mendelian genetics provided to the basic Darwinian model of evolution.[8]

It has become clear that the basic metabolic processes of all living cells are very similar. A number of identical mechanisms, structures, and metabolic pathways are found in all living entities so far observed. In particular

- All cells utilize phosphates, particularly adenosine triphosphate (ATP), for energy transfer.

[8] As it became accepted that evolution was to be understood in terms of Mendelian genetics and Darwinian natural selection, it also became clear that this understanding could not be sought only at a qualitative level. Mathematical methods must to be added.

- The metabolic reactions are catalyzed largely by proteins.
- Proteins are manufactured in the cell by a complete coding process. The sequence of amino acids in each protein is determined by the sequence of nucleotides in its gene, "written" as a DNA.
- The universal genetic code.

That so many things could have originated independently in different organisms by chance is incredible.

This synthetic theory is usually called Neo-Darwinism, and has the following features:

1. The average fitness increases; most of the mutations which are fixed in a population are advantageous.
2. The molecular clock goes faster or slower depending on the population size.

In contrast there is a Neutral Theory, created by Kimura which says:

1. Most of the offspring have disadvantegeous (fatal) genes, few have advantegeous genes.
2. The molecular clock holds.

7.1 Phylogenetic trees

> The holy grail of phylogenetics is the reconstruction of the one true tree of life.
>
> J.T.Thorley and R.D.M.Page

The underlying principle of phylogeny is to try to group "living entities" according to their level of similarity.
In biology for example, such trees ("phylogenies") typically represent the evolutionary history of a collection of extant species or the line of descent of some gene. No two members of a species are exactly the same - each has slight modifications from their parents. As environmental conditions change, nature will favour that branch of a species with some particular modification; as time goes on another mutation of the basic stock will become dominant. In this way, all species are continually evolving. This evolution occurs in a number of ways at the same time: some species die out and some become new species in their own right. This was already seen by Darwin [3]. He recognised that the characteristics which identified the species could indicate a history of descent, that is, a tree of evolution:

> The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of

> extinct species... The limbs divided into great branches, and these into lesser and lesser branches, were themselves once, when the tree was small, budding twigs; and this connexion of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups... From the first growth of the tree, many a limb and branch has decayed and dropped off, and these lost branches of various sizes may represent those whole orders, families, and genera which have now no living representatives, and which are known to us only from having been found in a fossil state... As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all a feebler branch, so by generation I belive it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.

Historically, this was a new idea: The concept of species having a continuity through time was only developed in the late 17th century; higher life forms were no longer thought to transmute into different kinds during the lifetime of an individual. It took over 150 years from the development of this concept before a rooted tree was proposed by Darwin.[9]
The phylogenetic tree can therefore be thought of as a central metaphor for evolution, providing a natural and meaningful way to order data, and with an enormous amount of evolutionary information contained within its branches.
for more information compare [18].

7.2 Looking for LUCA

One of the grand biological ideals is to be able to work out the complete quantitative phylogenetic tree, which means the history of the origin of all living species, back to the very beginning.[10]
Darwin spoke of "descent with modification", which is the central phrase of biological evolution, it refers to a genealogical relationship of species through time. These relationships are described in a phylogentic tree.
Evolution implies that many different species have a common ancestor and that all forms of life probably stem from the same remote beginnings. Hence, one of the tasks evolution sets for biologists is to discover the relationships among the species alive today and to trace the ancestors from which they descended.
For a treatment of several of the major transitions that occurred during the history of life see Maynard Smith and Szathmáry [14], [15].

The underlying principle of phylogenetics is to try to group living entities according to their level of similarity. In this context we assume that the more similar two entities

[9] Note that in Darwin's fundamental book *The origin of species* [3] there is exactly one figure, and this shows the description of the evolutionary history by a tree.

[10] Biologists have had this hope for a long time, modern biology, biochemistry and mathematics now have the actual capabilities of accomplishing it.

are, the closer they are to their common ancestor.

It is a central tenet of modern evolutionary biology that all "living things" trace back to a single common ancestor. Humans and other mammals are descended from shrew-like creatures that lived more than 150 Mya (million years ago); mammals, birds, reptiles and fish share as ancestors aquatic worms that lived 600 Mya; all plants and animals are derived from bacteria-like organisms that originated more than 3000 Mya. If we go back far enough, humans, frogs, bacteria and slime moulds share a common ancestor.
Then in the series of species from the origin of life up to today there must be a last universal common ancestor (LUCA). Note that this proposition does not assert that life arose just once, but that all starting points except one became extinct.[11]

Finding the LUCA for a set of species, or a set of populations, or a collection of genes is a very difficult task. How the LUCA for species can be found is discussed in [24].

Eigen [5] found that the LUCA for genes is an RNA-molecule of length 76 bp and 3.5 - 4 Gya.

7.3 Diversity

The theory of evolution is concerned with the extraordinary diversity of life on Earth. The diversity of the living world is staggering: more than 2 million existing species of plants and animals have been named and described; and many more remain to be discovered - until up to 10 times this number according to some estimates. What is impressive is not just the numbers but also the incredible heterogeneity. These virtually infinite variations of life are the fruit of the evolutionary process.
Taxonomy is the classification of organisms for the first aspect in any view of the life. Each phylogenetic tree is also a classification, but not vice versa.
The classification of animals and plants played an important role as a basis for Darwin's theory of evolution. Moreover, taxonomy is necessary to describe the diversity of living organisms.

The diversity of genomes is twofold:

- The presence of numerous species on Earth; and
- The polymorphism within each species.

There are many reasons why knowledge of the biodiversity is necessary, compare [8], [13] and [21].[12] There are several subquestions:

1. How many species are there?
2. How many have become extinct? In both the past and in the present. How many are lost every year?

[11] For more facts about early (molecular) evolution see Eigen [6].

[12] In particular, there is no successful vaccine to prevent or halt HIV infection. In part, this is because of the high genetic diversity of HIV. For this specific case see Dress and Wetzel [4]. Here, the main question is the prediction of the winning strain (or strains).

3. How long did species typically survive?

4. How much of evolutionary history is knowable?

For the idea of using evolutionary history for describing the biodiversity see Schleifer and Horn [19].

8 Bioinformatics

It is extremely remarkable that the molecules which are the carriers of information and the operational units which make life work are all linear polymers. Such polymers can be written as sequences or words; and exactly these entities are the subjects which can be handled by computers.

Bioinformatics stands for discussing biological questions with a computer, in particular about

- Searching in biological databases, in particular using public databases;
- Comparing sequences, in particular alignment sequences;
- Looking at protein structures;
- Phylogenetic analysis.

It may be of importance here to note that the culture of computational biology differs from the culture of bioinformatics, [11]:

> Sequence analysis plays an important role in both fields, but its methods and goal are understood differently by computational biologists and by bioinformaticians. Computational biology originally attracted a considerable number of practically minded theoretical biologists in the 1970s and 1980s who were both curious about the phenomenon of life and mathematical literate. They wanted to study nucleic acid and protein sequences in order to better understand life itself. In contrast, bioinformatics has attracted a large number of skilled computer enthusiasts with knowledge of computer programs that could serve as tools for laboratory biologists. ... Today's split between computational biology and bioinformatics appears to be a reflection of a profound cultural clash between curiosity-driven attitude of computational scientists and adversarial competitiviness of molecular biology software providers.

The main sources are:

Introduction: - www.molgen.mpg.de
- www.bioinformatik.de
- www.bioinformaticsonline.org

Genebank: - www.ncbi.nlm.nih.gov
- www.ebi.ac.uk
- www.ddbj.nig.ac.ac.jp

Human genome: - www.nhgri.nih.gov

Proteinbank: - www.expasy.ch
- www.embl-heidelberg.de
- www.pdb.bni.gov

Phylogeny: - www.ucmp.berkeley.edu/exhibit/phylogeny.html
- evolution.genetics.washington.edu
- awcmee.massey.ac.nz
- tolweb.org

9 Further reading

About biology - most of them in view of evolution - compare Gould and Keeton [9], Maynard Smith and Szathmáry [14], [15], Mayr [16].

Surveys about "Computational Molecular Biology" (with different approaches) you find by Clote and Backofen [2], Fitch [7], Konopka and Crabbe [11], Setubal and Meidanis [20], Vingron, Lenhof and Mutzel [22], Waterman [23].

Introductions into bioinformatics we can find by Attwood and Parry-Smith [1], Kanehisa [12], Mount [17].

References

[1] T.K. Attwood and D.J. Parry-Smith. *Introduction to bioinformatics*. Prentice Hall, 1999.

[2] P. Clote and R. Backofen. *Computational Molecular Biology*. John Wiley & Sons, 2000.

[3] C. Darwin. *The Origin of Species*. London, 1859.

[4] A. Dress and R. Wetzel. The Human Organism - a Place to Thrive for the Immuno-Deficiency Virus. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, editors, *New Approaches in Classification and Data Analysis*, pages 636–643. Springer Verlag, 1994.

[5] M. Eigen. Das Urgen. Nova Acta Leopoldina 243/52, Deutsche Akademie der Naturforscher Leopoldina, 1980.

[6] M. Eigen. *Stufen zum Leben*. Serie Piper, 1992.

[7] W.M. Fitch. An Introduction to Molecular Biology for Mathematicians and Computer Programmers. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 47:1–31, 1999.

[8] M. Glaubrecht. *Die ganze Welt ist eine Insel*. Hirzel Verlag, 2002.

[9] J.L. Gould and W.T. Keeton. *Biological Sciences*. W.W.Norton and Company, 1996.

[10] D. Graur and W.H. Li. *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., 1999.

[11] A.K. Konopka and M.J.C. Crabbe. *Compact Handbook of Computational Biology*. Marcel Dekker, 2004.

[12] M. Kanehisa. *Post-genome Informatics*. Oxford University Press, 2000.

[13] B. Lomborg. *The sceptical environmentalist*. Cambridge University Press, 2002.

[14] J. Maynard Smith and E. Szathmáry. *The major transitions in evolution*. W.H.Freeman, 1995.

[15] J. Maynard Smith and E. Szathmáry. *Evolution*. Spektrum, 1996.

[16] E. Mayr. *This is Biology*. Harvard University Press, 1997.

[17] D.W. Mount. *Bioinformatics*. Cold Spring Harbor Laboratory Press, 2001.

[18] R.D.M. Page and E.C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, 1998.

[19] K.-H. Schleifer and M. Horn. Mikrobielle Vielfalt - die unsichtbare Biodiversität. *Biologie heute*, 6:1–5, 2000.

[20] J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.

[21] M. Türkey, K. Bachmann, R. Kinzelbach, and E. Stackebrandt. Biodiversität - die Vielfalt in der wir leben. In *Wohin die Reise geht*, pages 72–83. Verband Deutscher Biologen, 2002.

[22] M. Vingron, H.-P. Lenhof, and P. Mutzel. Computational Molecular Biology. In M. Dell'Amico, F. Maffioli, and S. Martello, editors, *Annotated Bibliographies in Combinatorial Optimization*, pages 445–471. John Wiley and Sons, 1997.

[23] M.S. Waterman. *Introduction to Computational Biology*. Chapman & Heil, 1995.

[24] J. Whitfield. Born in a watery commune. *Nature*, 427:674–676, 2004.