Introduction to Genomic Signal Processing with Control

Aniruddha Datta
Edward R. Dougherty

CRC Press
Taylor & Francis Group

Introduction to Genomic Signal Processing with Control

Introduction to Genomic Signal Processing with Control

Aniruddha Datta
Edward R. Dougherty

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-8493-7198-8 (Hardcover)
International Standard Book Number-13: 978-0-8493-7198-1 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Datta, Aniruddha, 1963-
Introduction to genomic signal processing with control / Aniruddha Datta and Edward R. Dougherty.
p. ; cm.
Includes bibliographical references and index.
ISBN-13: 978-0-8493-7198-1 (alk. paper)
ISBN-10: 0-8493-7198-8 (alk. paper)
1. Cellular signal transduction. 2. Genetic regulation. 3. Control theory. I. Dougherty, Edward R. II. Title.
[DNLM: 1. Gene Expression Regulation--physiology. 2. Genomics. 3. Models, Theoretical. 4. Signal Transduction--genetics.]
QU 475 D234i 2007]

QP517.C45D383 2007
571.7'4--dc22 2006024030

Dedication

TO MICHAEL L. BITTNER

We dedicate this book to Michael L. Bittner. A little more than a decade ago, he sat in a bagel shop somewhere in Maryland and explained to Edward Dougherty his belief that engineering methods might be used to intervene in and control genetic regulatory activity with the goal of mitigating the likelihood of disease or its progression. Much of our work in genomic signal processing, in particular the last chapter of this book, has been guided by the vision and active participation of Michael Bittner.

Preface

Recently, the Human Genome Project announced one of the most stunning achievements in the history of science: the generation of a reference sequence of the human genome. This has brought unprecedented interest to the area of genomics, which concerns the study of large sets of genes with the goal of understanding collective function, rather than that of individual genes. Such a study is important since cellular control and its failure in disease result from multivariate activity among cohorts of genes. Very recent research indicates that engineering approaches for prediction, signal processing and control are quite well suited for studying this kind of multivariate interaction. The aim of this book is to provide the readers with an introduction to genomic signal processing, including a state-of-the-art account of the use of control theory to obtain optimal intervention strategies for gene regulatory networks, and to make readers aware of some of the open research challenges.

Successful realization of the full potential of the area of genomics will require the collective skill and creativity of a diverse set of researchers such as biologists (including medical practitioners), statisticians and engineers. This can be possible only when each of these groups is willing and able to venture out of its immediate area of expertise, a task that is not always easy to undertake. We consider ourselves to be mathematically trained engineers who have had the opportunity to learn some basic biology while pursuing engineering research motivated by medical applications. Our main motivation for writing this book is to enable other interested readers with a similar mathematical background to get a reasonably good working knowledge of the genomics-related engineering research while having to expend only a small fraction of the time and effort that we ourselves had to originally invest.

Since the book is targeted primarily at readers whose biology background is practically non-existent, we felt that the book could be made self-contained only by including a substantial amount of material on molecular biology. However, it should be pointed out at the very outset that our molecular biology presentation, which is to a substantial extent gleaned from the classic, *Essential Cell Biology*, by Alberts et. al. [1], is *molecular biology seen through the eyes of engineers* and is the minimum needed for appropriately motivating the genomics-related engineering research presented here. As such, it is our sincere hope that the serious reader, after gaining the skeletal molecular biology knowledge from this text, will be motivated to consult other books such as the one by Alberts et. al. for additional details.

The book provides a tutorial introduction to the current engineering re-

search in genomics. The necessary molecular biology background is presented, and techniques from signal processing and control are used to (i) unearth inter-gene relationships, (ii) carry out gene based classification of disease, (iii) model genetic regulatory networks and (iv) alter (i.e. control) their dynamic behavior. The book can be divided into two parts. In the first eleven chapters, the focus is on building up the necessary molecular biology background. No prior exposure to molecular biology is assumed. In the last six chapters, the focus is on discussing the application of engineering approaches for attacking some of the challenging research problems that arise in genomics-related research.

The book begins with a basic review of organic chemistry leading to the introduction of DNA, RNA and proteins. This is followed by a description of the processes of transcription and translation and the genetic code that is used to carry out the latter. Control of gene expression is also discussed. Genetic engineering tools such as microarrays, PCR, etc., are introduced and cell cycle control and tissue renewal in multi-cellular organisms is discussed. Cancer is introduced as the breakdown of normal cell cycle control. Having covered the basics of genomics, the book proceeds to engineering applications.

The engineering techniques of classification and clustering are shown to be appropriate for carrying out gene-based disease classification. Classification then leads naturally to expression prediction which in turn leads to genetic regulatory networks. Finally, the book concludes with a discussion of control approaches that can be used to alter the behavior of such networks in the hope that this alteration will move the network from undesirable (diseased) states to more desirable (disease-free) ones.

Several people contributed to the writing of this book in many different ways. First and foremost, we would like to thank the authors of *Essential Cell Biology* and Garland Publishers, Inc. for allowing us to use a large number of figures from that book. These figures, which have been identified by proper citations at appropriate places in the text, are so informative that it is hard to imagine how we could have effectively described some of the related material without being able to refer to them. Several of our doctoral students helped at different stages with the preparation of the book. Yufei Xiao helped during the initial stages with the preparation of the first set of class notes that ultimately became this book, Ashish Choudhary was a major player in helping us get the book to its final form, and Ranadip Pal assisted us during the intermediate stages on an as-needed basis.

We would like to thank R. Kishan Baheti, Director of the Power, Control and Adaptive Networks Program and John Cozzens, Director of the Theoretical Foundations Cluster Program at the National Science Foundation for supporting our research(Grant Nos. ECS-0355227 and CCF-0514644). Financial support from the National Cancer Institute (Grant No. CA90301), the National Human Genome Research Institute, the Translational Genomics Research Institute, the University of Texas M. D. Anderson Cancer Center, and the Texas A&M Department of Electrical and Computer Engineering is

also thankfully acknowledged. Last, but not the least, we would like to thank our families without whose understanding and patience, this project could not have been completed.

Aniruddha Datta and Edward R. Dougherty
College Station, Texas
June 2006.

Contents

1

Introduction

In this introductory chapter, we provide a very brief introduction to cells and point out some general characteristics and associated terminology. A cell is the basic unit of life and all living creatures are made of cells. Nothing smaller than a cell can be truly called living, e.g., viruses, which are essentially genetic material encapsulated in a protein coat, have no ability to replicate by themselves, and are therefore nonliving. The only way in which they can replicate is by hijacking the replication machinery of a living cell, and this is what usually happens in a viral infection.

Each cell is typically about $5-20\mu m$ in diameter and this small size means that a cell can only be viewed under a microscope. The word *cell* is due to Robert Hooke, who in 1665 reported examining a piece of cork under a microscope and found it to be composed of a large number of small chambers, which he called *cells*. Fig. 1.1 shows the schematic diagram for a typical plant cell. Enclosing the entire cell is a limiting boundary called the *plasma membrane*. Inside the cell there are a number of compartments or *organelles*. Of these, the most prominent one is the *nucleus*, which serves as the store house of most of the genetic information. The *endoplasmic reticulum* is a continuation of the nuclear envelope and is the site at which many of the cell components are synthesized. The *Golgi apparatus* is the site where components synthesized in the endoplasmic reticulum undergo appropriate modifications before being passed on to their destinations. There are many other smaller organelles in a cell and some of these are shown in Fig. 1.1. The *mitochondria* are the sites at which most of the energy generation for the cell takes place; the *chloroplasts* for a plant cell are the sites where photosynthesis takes place; and *peroxisomes* are the organelles used to compartmentalize cellular reactions that release the dangerous chemical hydrogen peroxide. If one removes all the organelles from the cell, then what remains is called the *cytosol*.

Cells can be grouped into two broad categories depending on whether they contain a nucleus or not. *Procaryotes* (or procaryotic cells) do not contain a nucleus and other organelles. All bacteria are procaryotic cells. *Eucaryotes* (or eucaryotic cells) on the other hand are characterized by the presence of a nucleus. All multicellular organisms, including humans, are made up of eucaryotic cells, while a yeast would be an example of a unicellular eucaryote.

The cytosol in a cell is a concentrated aqueous (watery) gel of large and small molecules. The plasma membrane and the membranes of the other organelles, on the other hand, are made of *lipids*. These molecules have the

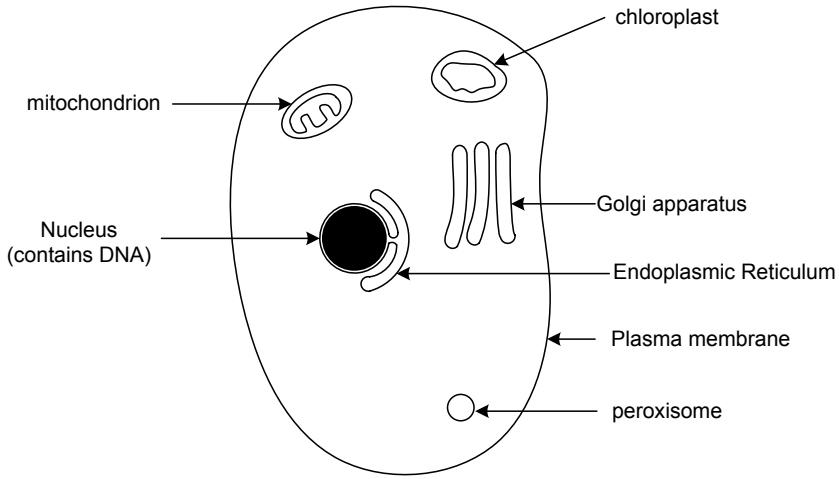


FIGURE 1.1 A typical cell

property that they have a *hydrophilic* (water loving) part and a *hydrophobic* (water hating) part, as shown in Fig. 1.2. So when they come in contact with water, they align appropriately, as shown in the figure, to form a *lipid bilayer*, which is the basis of all cell membranes. The lipid bilayer also facilitates the transportation of hydrophilic materials from the outside of a cell to its cytosol and vice versa, or for that matter from one membrane-bounded organelle to another. All that is required is that the hydrophilic material get encapsulated in a lipid bilayer, forming what is called a *vesicle* which can then be transported across the aqueous environment inside the cell to its appropriate destination organelle. Once the destination is reached, the lipid bilayer of the vesicle "merges" with that of the destination organelle with the result that the aqueous contents of the vesicle are transferred to the interior of the organelle. This vesicular transport is schematically illustrated in Fig. 1.2.

The cytosol of a eucaryotic cell contains what is called a *cytoskeleton*. As the name suggests, this provides structure to the interior of the cell and is made up of three types of proteins: (i) *actin filaments*; (ii) *microtubules*; and (iii) *intermediate filaments*. The cytoskeleton performs important functions such as (i) providing mechanical strength to the cell and its neighbors (intermediate filaments); (ii) generating contractile forces, for instance, in muscle cells (actin filaments); (iii) providing tracks along which different cell components can be moved (microtubules); and (iv) distributing the DNA properly between the daughter cells at cell division (microtubules).

Although cells from different organisms and even from the same organism vary enormously in size and function, they all have the same underlying chem-

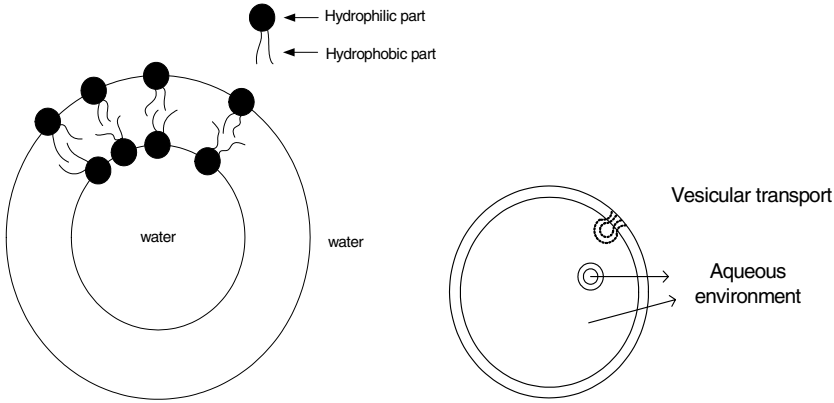


FIGURE 1.2 Lipid bilayer and vesicular transport

istry. As we will see later in this book, the genetic information in the DNA is coded using the same blocks in all organisms from humans all the way down to the simplest unicellular ones. Furthermore, this information is interpreted using essentially the same machinery to produce proteins and different proteins are made up of the same 20 blocks put together in different ways. This greatly simplifies the study of many aspects of molecular biology since they are not organism specific. In order to appreciate Genomic Signal Processing, one must have some basic understanding about genes, proteins and their interactions. Consequently, about one half of this book (the next 10 chapters) is devoted to buiding up this necessary background in sufficient detail. The details about the role of the cytoskeleton and the lipid membranes, although important in their own right, are not crucial to the theme of this book. Consequently, for these topics, only a superficial discussion has been included in this introductory chapter. For a more detailed treatment, the reader is referred to [1].

We conclude this chapter with a mention of *model organisms* that are used for molecular biology studies. These model organisms, which are used for carrying out experiments, must be simple and capable of quickly replicating themselves. For procaryotes, the *E. coli* bacteria, which contain a few genes and reproduce every 20 minutes, is, by far, the most widely used model of choice; for flowering plants, it is the plant *Arabidopsis* which produces thousands of offsprings in 8 to 10 weeks; for insects, it is the *Drosophila* or fruit fly; for mammals, it is rats and mice; and for unicellular eucaryotes, it is the yeast.

2

Review of Organic Chemistry

In this chapter, we provide a brief introduction to organic chemistry. The chemical properties of biological molecules play a crucial role in making all known life possible and so any discussion of molecular biology would have to necessarily include some discussion of organic chemistry. Although the discussion here is far from exhaustive, it is essentially self-contained and should provide a good introduction to anyone who has had some exposure to basic chemistry in the past. For a more detailed treatment, the reader is referred to [2].

Matter is made up of combinations of *elements* — substances such as hydrogen or carbon that cannot be broken down or converted into other substances by chemical means. An *atom* is the smallest particle of an element that still retains its distinctive chemical properties. *Molecules* are formed by two or more atoms of the same element or of different elements combining together in a chemical fashion. An atom is made up of three kinds of subatomic particles: *protons* (mass = 1, charge = +1); *neutrons* (mass = 1, charge = 0); and *electrons* (mass $\simeq$ 0, charge = −1). Protons and neutrons reside in the nucleus of the atom while electrons revolve around the nucleus in certain orbits.

Each element has a fixed number of protons in the nucleus of each of its atoms and this number is referred to as the *atomic number*. We next list a few elements that occur over and over again in organic molecules, along with their atomic numbers: hydrogen (H) has an atomic number of 1; oxygen (O) has an atomic number of 8; carbon (C) has an atomic number of 6; nitrogen (N) has an atomic number of 7; phosphorous (P) has an atomic number of 15; sodium (Na) has an atomic number of 11; and calcium (Ca) has an atomic number of 20.

The total number of protons and neutrons in the nucleus of an atom of a particular element is referred to as its *atomic weight*. The hydrogen atom has only one proton and one electron and so its atomic weight is 1. Thus the atomic weight of an element is the mass of an atom of that element relative to that of a hydrogen atom. The electrons have negligible mass and do not contribute to the atomic weight. However, since the atom as a whole is electrically neutral, the number of protons must be equal to the number of electrons.

The number of neutrons in an atom of a particular element can vary. For instance, carbon usually occurs as atoms possessing 6 protons and 6 neutrons so that its usual atomic weight is 12. However, sometimes carbon atoms can

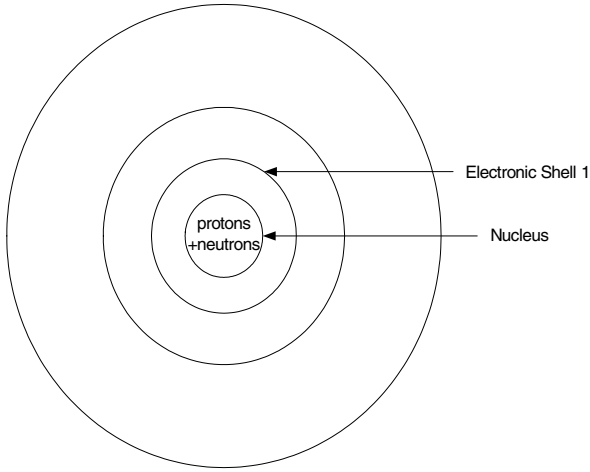


FIGURE 2.1 Schematic diagram of an atom.

have 8 neutrons resulting in a carbon atom with atomic weight 14. These two types of carbon atoms with varying numbers of neutrons are referred to as *isotopes*. Usually, one isotope of an element is the most stable one and the other isotopes radioactively decay towards it with time.

Fig. 2.1 shows the schematic diagram of a typical atom. The protons and neutrons are held together tightly in the nucleus while the electrons revolve around the nucleus in certain discrete orbits called shells. There is a limit to the number of electrons that can be held in each shell. The first shell (the one closest to the nucleus) can accommodate up to 2 electrons, the next one can accommodate up to 8 electrons, the next one up to 8 electrons, and so on. If an atom has its outermost electronic shell completely occupied, then it is unreactive. Helium (He, atomic number = 2), neon (Ne, atomic number = 10), argon (Ar, atomic number = 18) are all unreactive and are, therefore, called inert gases.

2.1 Electrovalent and Covalent Bonds

Atoms that do not have completely filled outer shells are capable of reacting with each other to form compounds. The natural tendency is to acquire completely filled outer shells by donating or accepting electrons (electrovalent or ionic bonds) or by sharing pairs of electrons (covalent bonds). For instance, sodium (Na) has an atomic number of 11 so that its electronic distribution is

$$2 + 8 + 1.$$

Chlorine (Cl), on the other hand, has an atomic number of 17, so that its electronic distribution is

$$2 + 8 + 7.$$

So, the sodium atom has a strong tendency to get rid of the single electron in its outermost shell while chlorine has a strong tendency to acquire one more electron so that it can have a complete outermost shell. Because of this, the sodium atom donates an electron to the chlorine atom to form the compound sodium chloride (or common salt). In the process, the sodium atom becomes a sodium ion with one positive charge (Na^+) while the chlorine atom becomes a chloride ion with one negative charge (Cl^-). The type of bond achieved by this transfer of electrons is called an *electrovalent bond* or *ionic bond*. Such bonds are extremely strong and are held together in place by the forces of electrostatic attraction between the two oppositely charged ions.

Another way in which an atom can achieve a completely filled outermost shell is by sharing pairs of electrons with other atoms. Consider the following three instances:

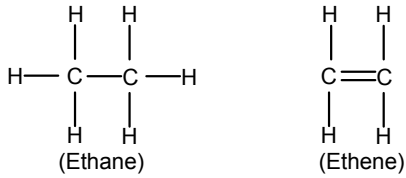
(1) Hydrogen with an atomic number of 1 has only one electron in its outermost shell. Two hydrogen atoms can share their electrons with each other so that each hydrogen atom can have an outer shell with two electrons. The result is the formation of a hydrogen molecule H_2.

(2) Carbon with an atomic number of 6 has the electronic distribution 2+4. Thus it can share its four outermost electrons with the electrons of four different hydrogen atoms. The result is the formation of the gas methane, CH_4.

(3) Oxygen with an atomic number of 8 has the electronic distribution 2+6. Thus, it can share two of its outermost electrons with the outermost electrons of two hydrogen atoms. The result is the formation of water, H_2O.

The bonds in the above examples, achieved by the sharing of pairs of electrons between atoms, are called *covalent bonds*. The number of electrons that an atom of an element must donate/accept/share to form a complete outermost electronic shell is referred to as its *valence*. Thus, from the above examples, hydrogen has a valence of 1, carbon has a valence of 4, oxygen has a valence of 2, sodium and chlorine each has a valence of 1. Since carbon has a valence of 4, each carbon atom can share up to four pairs of electrons with other atoms to achieve a complete outermost shell. Depending on the number of electron pairs that are shared, there can be different types of covalent bonds. A *single bond* is a covalent bond where only one pair of electrons is shared between the participating atoms. An example is ethane (C_2H_6), shown below, where all the participating carbon and hydrogen atoms are linked together by single bonds indicated by the single dashes. Each dash denotes that a pair of electrons is shared between the participating atoms. A *double bond* is formed when two pairs of electrons are shared between the participating atoms. An

example is the gas ethene (C_2H_4), shown below, where the two dashes between the two carbon atoms indicate that two pairs of electrons are being shared.



Double bonds are inherently more rigid. For instance, the single carbon-to-carbon bond in ethane allows free rotation about it for the right and left halves of the molecule. On the other hand, the two halves of the ethene molecule cannot be rotated freely about the carbon-to-carbon double bond without twisting the bond itself. This is referred to as *stearic hindrance* and can have profound consequences in determining the structure of a long biological molecule.

When the atoms joined by a single covalent bond belong to different elements, the two atoms usually attract the shared electrons to different degrees. For example, oxygen and nitrogen atoms attract electrons quite strongly while hydrogen attracts electrons quite weakly. Consequently, in a hydrogen-oxygen covalent bond, the hydrogen atom will tend to acquire a positive charge while the oxygen atom will tend to acquire a negative charge. This makes the overall molecule have an uneven charge distribution and such molecules are called *polar*. For instance, the covalent bond between oxygen and hydrogen $-\mathrm{O}-\mathrm{H}$ is polar. The covalent bond between nitrogen and hydrogen $-\mathrm{N}-\mathrm{H}$ is polar. However, the bond between carbon and hydrogen $-\mathrm{C}-\mathrm{H}$ is non-polar since carbon and hydrogen atoms both attract electrons more or less equally. Polar covalent bonds are extremely important in biology because they allow molecules to interact through electrical forces. Consider a molecule of water, H_2O:

$$\mathrm{H}-\mathrm{O}-\mathrm{H}.$$

Here each hydrogen atom is positively charged while the oxygen atom is negatively charged. Thus, each hydrogen atom in a water molecule could form a weak ionic bond with the oxygen atom of another water molecule. Such a weak ionic bond is called a *hydrogen bond* and the hydrogen bonding between the water molecules is known to be the reason for water being a liquid at room temperatures.

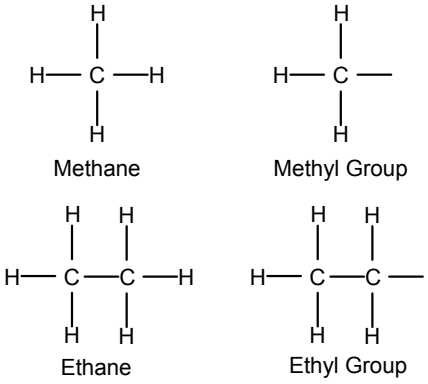
Polar molecules readily dissolve in water because of their electrical interactions with the charges on the water molecules. Such substances are called *hydrophilic* (water loving). On the other hand, non-polar molecules such as the hydrocarbons do not dissolve in water. Such substances are called *hydrophobic* (water hating). A molecule that has both hydrophilic and hydrophobic

parts is called *amphipathic*. From the point of view of molecular biology, it is important to note that water is the most abundant substance inside cells, and amphipathic molecules are crucial building blocks for all cell membranes.

2.2 Some Chemical Bonds and Groups Commonly Encountered in Biological Molecules

C-H Compounds

Compounds containing only carbon and hydrogen are called *hydrocarbons*. Examples are methane (CH_4) and ethane (C_2H_6), which have the structures shown below.



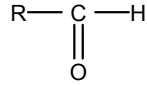
By removing one of the hydrogen atoms from methane, we obtain a highly reactive group called the *methyl group*. The *ethyl group* is similarly derived from ethane.

C−O Compounds

We next consider organic compounds that contain carbon and oxygen in addition to other elements. There are different classes of compounds that fall into this category.

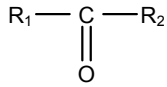
Alcohols are characterized by the presence of the hydroxyl (OH) group and have the general formula R − OH. Here R could be any organic group. When R = CH_3, i.e. the methyl group, the corresponding alcohol is called *methyl alcohol* or *methanol*. When R = C_2H_5 i.e. the ethyl group, the corresponding alcohol is called *ethyl alcohol* or *ethanol*.

Aldehydes are characterized by the general formula



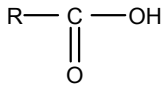
where R is any organic group. When $R = H$, the corresponding aldehyde is called *formaldehyde* and when $R = CH_3$, the corresponding aldehyde is called *acetaldehyde*.

Ketones are characterized by the general formula



where R_1 and R_2 can be any two organic groups. The C = O (C double bonded with O) is called the *carbonyl* group and it is present in both aldehydes and ketones.

Carboxylic Acids are characterized by the general formula

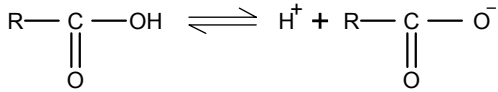


where R is any organic group. When $R = H$, the corresponding acid is called *formic Acid* while when $R = CH_3$, the corresponding acid is called *acetic Acid*. The COOH group present in all carboxylic acids is called a *carboxyl* group.

Recall from inorganic chemistry that substances that release hydrogen ions into solution are called *acids*., e.g.,

$$\underset{\text{hydrochloric acid}}{HCl} \longrightarrow H^+ + Cl^-$$

i.e. hydrochloric acid dissociates in water to yield hydrogen ions and chloride ions. Carboxylic acids are called acids because they *partially* dissociate in solution to yield hydrogen ions:



Because of this partial dissociation, organic acids are called *weak acids* as opposed to say hydrochloric acid, which is called a *strong acid*.

The hydrogen ion concentration of a solution is usually measured according to its pH value, which is defined by $\text{pH} = -\log_{10}[\text{H}^+]$, where $[H^+]$ is the hydrogen ion concentration in moles per liter. For pure water, which is neither acidic nor basic, $[\text{H}^+] = 10^{-7}$ so that the pH of pure water is 7. For acids $pH < 7$ since acids have a higher concentration of hydrogen ions than pure

water. Bases, on the other hand, have a lower concentration of hydrogen ions than pure water and so for bases the $pH > 7$.

In inorganic chemistry, *bases* are defined as substances that reduce the number of hydrogen ions in aqueous solution. For instance, sodium hydroxide (NaOH) dissociates in solution to yield sodium ions and hydroxyl ions:

$$\mathrm{NaOH} \longrightarrow \mathrm{Na^+} + \mathrm{OH^-}.$$

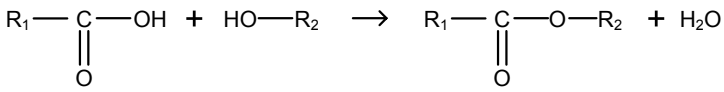
The hydroxyl ions then react with some of the hydrogen ions in the water, thereby reducing the number of hydrogen ions. Thus, NaOH is a base as is the gas ammonia (NH_3) which reacts with a hydrogen ion in solution to produce the ammonium ion (NH_4^+):

$$\mathrm{NH_3} + \mathrm{H^+} \longrightarrow \mathrm{NH_4^+}.$$

Since many bases in inorganic chemistry do contain a hydroxyl group, one could say that alcohols in organic chemistry share some similarity with many of the bases in inorganic chemistry. Indeed, an important fact from inorganic chemistry is that an acid reacts with a base to produce a salt and water. For instance, sodium hydroxide reacts with hydrochloric acid to produce the common salt sodium chloride plus water:

$$\mathrm{NaOH} + \mathrm{HCl} \longrightarrow \mathrm{NaCl} + \mathrm{H_2O}.$$

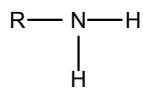
In organic chemistry, when a carboxylic acid reacts with an alcohol, the reaction produces water and a compound called an *ester*. The chemical equation showing how the bonds get rearranged is given below:



Here R_1 and R_2 are the paticular groups associated with the carboxylic acid and alcohol respectively.

C-N Compounds

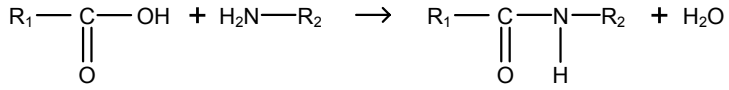
There are many organic compounds that contain carbon and nitrogen, in addition to other elements. In fact, nitrogen occurs in several ring compounds, including important constituents of nucleic acids. Because of their fundamental importance to genomics, these ring compounds will be discussed in some detail in Section 2.3. For the present, we focus on *amines* which are organic compounds characterized by the presence of the amino (NH_2) group. A typical amine will have the general formula



where R is any organic group. An amine can reversibly accept a hydrogen ion from water and, therefore, qualifies as a *weak base*:

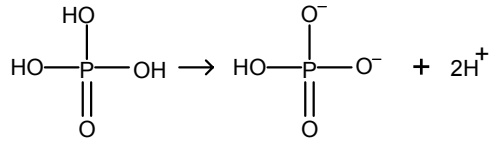
$$\mathrm{R-NH_2} \;+\; \mathrm{H^+} \;\rightleftharpoons\; \mathrm{R-NH_3^+}.$$

Furthermore, an amine can react with a carboxylic acid to produce an *amide* and water. The general reaction is given below:



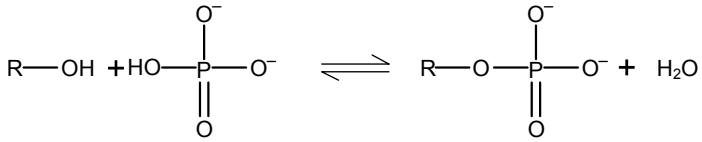
Phosphates

Phosphates are characterized by the presence of the inorganic phosphate ion, which is derived from phosphoric acid (H_3PO_4) by the loss of two hydrogen ions:



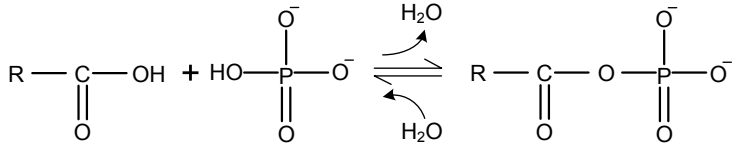
The phosphate ion can take part in many reactions that play an important role in molecular biology. A few instances are listed below.

(i) The inorganic phosphate ion can react with an alcohol to produce a *phosphate ester*:



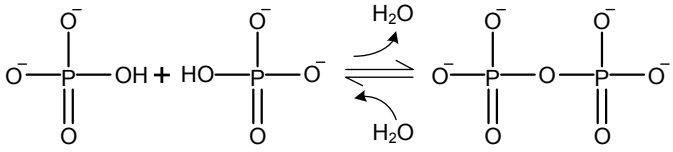
Phosphate ester bonds as shown above are important in the formation of nucleic acids such as DNA and RNA.

(ii) The inorganic phosphate ion can react with the carboxyl group of an organic acid to produce a *carboxylic-phosphoric acid anhydride*.



The compound formed above is called an *anhydride*, since it results from the removal of a water molecule.

(iii) Two inorganic phosphate ions can react together to produce a phosphoanhydride bond:



Phosphoanhydride bonds are high energy bonds and the reversibility of the above reaction makes such bonds suitable for energy storage and transfer. In this connection, we note that each of the above three reactions can be reversed by the addition of a water molecule. In organic chemistry, one encounters many reactions of this type.

2.3 Building Blocks for Common Organic Molecules

If we disregard water, almost all the molecules in a cell are based on *carbon*. Furthermore, most of the molecules are giant *macromolecules*. The small organic molecules constitute a very small fraction of the total cell mass. However, fortunately, even the giant macromolecules are made up of small organic molecules that are bonded together by covalent bonds. Cells contain four major families of small organic molecules, or modular units, that are combined together to form the large macromolecules:

1. *Sugars*, which are the building blocks for more complex sugars and carbohydrates;
2. *Fatty acids*, which are the building blocks for fats, lipids and all cell membranes;
3. *Amino acids*, which are the building blocks for proteins; and
4. *Nucleotides*, which are the building blocks for nucleic acids such as DNA and RNA.

Let us focus on these building blocks one by one.

2.3.1 Sugars

The simplest sugars (monosaccharides) are compounds with the general formula $(CH_2O)_n$ where n is usually $3, 4, 5, 6$ or 7. Sugars are also called *carbohydrates* since their general formula suggests that they are somehow built up from carbon and water. Monosaccharides usually occur as aldehydes or ketones. Five-carbon sugars, called *pentoses* and six-carbon sugars, called *hexoses* are very important in molecular biology. Both the aldehyde and ketone versions of these sugars are shown in Figs. 2.2 and 2.3.

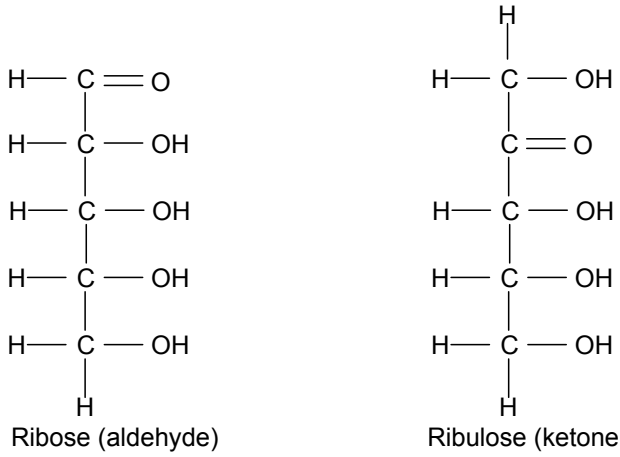


FIGURE 2.2 5-carbon sugars

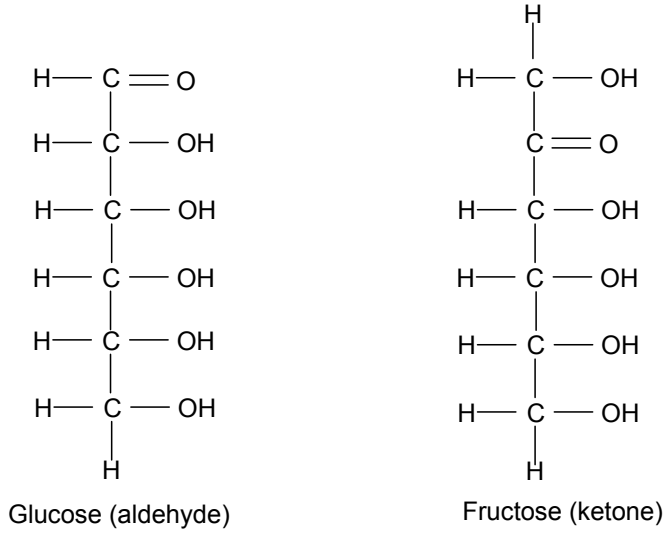


FIGURE 2.3 6-carbon sugars

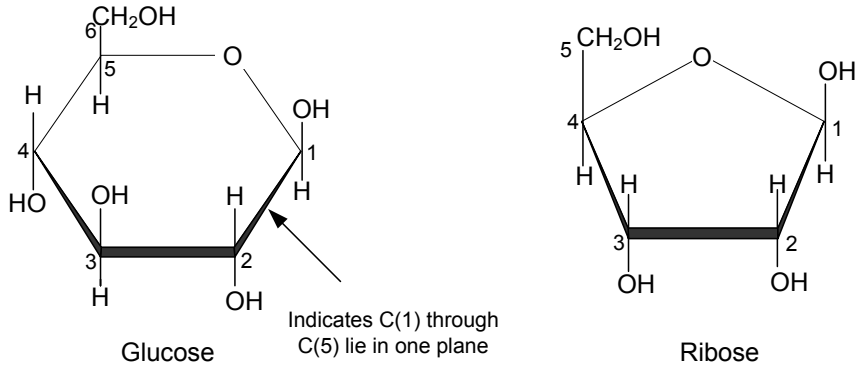


FIGURE 2.4 Ring formation in glucose and ribose

The 5-carbon sugar *ribose* and its derivative *deoxyribose* are important constituents of nucleic acids such as DNA and RNA while the 6-carbon sugar *glucose* serves as an important source of energy. For aldehyde sugars, in the structural representation, the carbon atoms are numbered consecutively with the carbon atom containing the aldehyde group being numbered 1.

In aqueous solution, the aldehyde or ketone group of a sugar molecule tends to react with a hydroxyl group of the same molecule, thereby closing the molecule into a ring. The ring structures for the sugars glucose and ribose are shown in Fig. 2.4. Since the environment inside a cell is aqueous, it is this ring structure that we will be encountering over and over again in this text and, indeed, in any book on molecular biology. If one were to replace the hydroxyl group at carbon number 2 in ribose with hydrogen then one obtains *deoxyribose*, which is the sugar present in DNA.

In inorganic chemistry, the chemical formula of a compound usually determines a unique structural formula and associated properties. In organic chemistry, on the other hand, compounds having the same chemical formula may have totally different structural formulae and properties. For instance, if we interchange the hydrogen and the hydroxyl group at carbon No.4 in glucose we obtain a different sugar called *galactose*. In another instance, if we interchange the hydrogen and the hydroxyl groups at carbon No.2 in glucose, we obtain the sugar *mannose*. Compounds which have the same chemical formula but different structural formulae are called *isomers* of each other. Organic chemistry is filled with instances of isomers whose different properties make them especially well suited for particular biological functions.

The monosaccharide building blocks can be combined together to yield more complex sugars. Disaccharides are made up of two monosaccharides that are linked together in a *condensation* reaction. This reaction, which is accompanied by the removal of a water molecule, is schematically shown in Fig. 2.5. Since there are many hydroxyl groups on each monosaccharide,

two monosaccharides can link together in many different ways. Furthermore, since there are additional hydroxyl groups available on the disaccharide, more monosaccharides can get linked to it, producing chains and branches of various lengths. Short chains are called *oligosaccharides*, while long chains are called *polysaccharides*. An example of a polysaccharide is *glycogen*, which is made up entirely of glucose units linked together. Glycogen serves as an energy storage in animals.

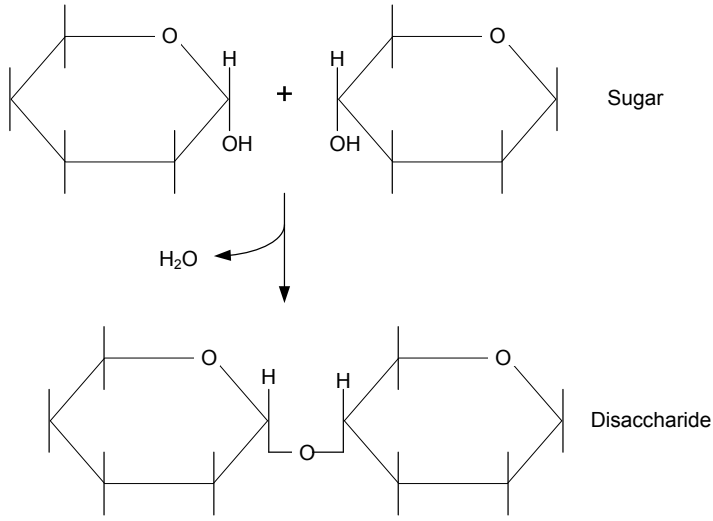


FIGURE 2.5 Condensation reaction showing the formation of a disaccharide.

Condensation reactions are not unique to sugars. In fact, as we will see later in this chapter, they are used to form proteins from amino acids and nucleic acids from nucleotides. The reverse reaction of condensation is called *hydrolysis* (splitting with water) and plays an important role in our digestion of carbohydrates and other foods. We conclude our discussion of sugars by providing some specific examples of disaccharides that can result from linking together two monosaccharide units:

glucose + glucose = maltose;
glucose + galactose = lactose (the sugar found in milk); and
glucose + fructose = sucrose.