# Math 4/5779: Mathematics Clinic
# Pathway Inference: Computational Issues

## *Supplement: Molecular Biology and Chemistry*

Taught by Harvey J. Greenberg

Assisted by Raphael Bar-Or, Lance Lana, Cary Miller,
Tod Morrison and Christiaan van Woudenberg

Sponsored by
DMI BioSciences, Genomica and Tech-X

Spring Semester 2002

Submitted May 23, 2002

This supplement contains a general biology and chemistry background that is designed to provide both some details about the science and a big picture for insights into pathways.

# Contents

# 1 Biochemistry

Fundamentally, life is a collection of biochemical processes or chemical reactions, where a chemical reaction is any process in which molecules interact with one another, leading to chemical or physical change. A chemical reaction that releases some amount of energy as heat is called exergonic, while a chemical reaction that requires an input energy is called endergonic. An enzyme is a molecule that increases the rate of reaction among some other chemical entities; we say the enzyme *catalyzes* the reaction.

## 1.1 The Water Molecule

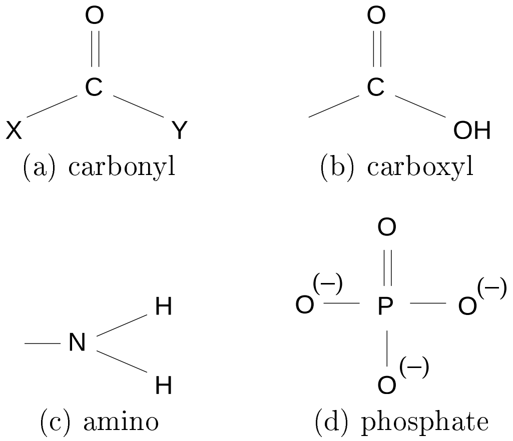
Water is the most important molecule in living systems; it makes life on Earth possible. Most cells are surrounded by water, and themselves are about 70% to 80% water [2]. The basic chemical structure of water is deceptively simple, but due to the orientation of the atoms that comprise it, water is a very special substance. It is the only common substance to exists naturally in all three physical states of matter: solid, liquid, and gas. Its two hydrogen atoms are joined to the oxygen atom by a single *covalent bond*, a strong chemical bond in which two atoms share one pair of electrons in a mutual valence shell. The water molecule is a *polar molecule*, which means the opposite ends of the molecule have opposite charges. This property causes water to be a *solvent*, a dissolving agent of a solution. Polar molecules and ions, which are charged molecules, dissolve completely in water, thus they are termed *solutes*. Non-polar molecules, where opposite ends of the molecule have no charges (for example, oils), do not dissolve in water. *Hydrophilic* substances, or polar molecules, have an affinity for water while *hydrophobic* substances, or non-polar molecules, do not.

## 1.2 Functional Groups

Those compounds of biomolecules that are most commonly involved in chemical reactions are known as *functional groups*. Each functional group behaves consistently from one organic molecule to another. The number and arrangement of the functional groups give each molecule its unique properties. The five functional groups most important in the chemistry of life are the hydroxyl, carbonyl, carboxyl, amino and phosphate groups.

In a ***hydroxyl group*** ($-$OH), a hydrogen atom is bonded to an oxygen atom, which in turn, is bonded to the carbon skeleton of the organic molecule. Organic compounds containing a hydroxyl group are called *alcohol*. The hydroxyl group is polar, so water molecules are attracted to the hydroxyl group and will dissolve organic compounds containing such groups. Sugar, for example, will dissolve in water due to the presence of multiple hydroxyl groups.

Figure 1(a) shows the carbonyl group as a carbon atom joined to an oxygen atom by a double bond. If the carbonyl group is on the end of a carbon skeleton, the organic compound is called an *aldehyde*; otherwise the compound is called a *ketone*.



(a) carbonyl

(b) carboxyl

(c) amino

(d) phosphate

Figure 1: Functional Groups

The carboxyl group consists of an oxygen atom joined to a carbon atom by a double bond that is also bonded to a hydroxyl group. Compounds containing carboxyl group are known as *carboxylic acids*, or organic acids. Vinegar, for example, has a sour taste due to its acidic characteristic.

The amino group consists of a nitrogen atom bonded to two hydrogen atoms and to the carbon skeleton. Amino acids contain this functional group as well as a carboxylic acid group.

The phosphate group is an *anion*, a negatively charged ion, formed by dissociation of an inorganic acid, called phosphoric acid ($H_3PO_4$). Organic compounds that contain a phosphate group have a phosphate ion attached via one of its oxygen atoms to the carbon skeleton. One function of the phosphate group is to transfer energy between organic molecules. As an example, ATP is the most powerful energy biomolecule because it contains three phosphate groups.

## 1.3 Protein

Proteins, the primary components of all living things, are made of the same basic constituents. These basic constituents, the amino acids, share a common basic structure, consisting of a central carbon atom (C), an amino group ($NH_3$) at one end, a carboxylic group (COOH) at the other, and some form of side chain (R), as shown in Figure 2.

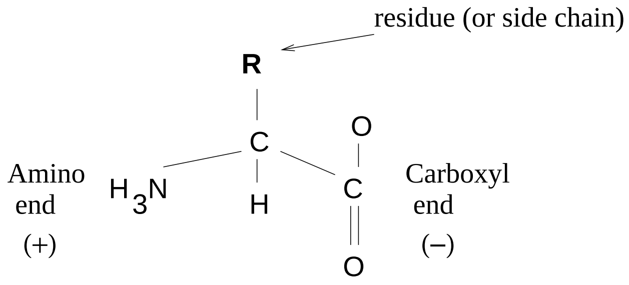


Figure 2: Amino Acid

The chemical properties of an amino acid are determined by its side chain. There are twenty amino acids that cells use to build their thousands of proteins. Amino acids are assembled by a reaction between the nitrogen atom at the amino end of one amino acid and the carbon atom at the carboxyl end of another. This reaction, called a *peptide bond*, links the two amino acids together and releases a molecule of water. By a series of these reactions a long chain of amino acids can be strung together into *polypeptides*. All proteins are polypeptides, where their amino acids are linked together in a linear sequence.

The sequence of amino acid residues that make up a protein is called the protein's *primary structure*. The precise primary structure of a protein is determined not by the random linking of amino acids, but by inherited genetic information. Most proteins have segments of their polypeptide chain repeatedly coiled or folded in patterns that contribute to the protein's overall conformation. These coils and folds, collectively referred to as *secondary structure*, are the result of hydrogen bonds between atoms on the polypeptide backbone. One such secondary structure is the *$\alpha$-helix*, a delicate coil held together by hydrogen bonding between every fourth amino acid. Another fundamental type of secondary structure is the *$\beta$-pleated sheet*, in which two regions of the polypeptide chain lie parallel to each other. Pleated sheets make up many globular proteins in biological systems. The overall three dimensional shape is the protein's *tertiary structure*, consisting of irregular contortions caused by the bonding between the side chains of the various amino acids. Some proteins consist of two or more polypeptide chains aggregated into one functional, large biomolecule. *Quaternary structure* is the overall protein structure, that results from the aggregation of these polypeptide subunits. Hemoglobin, the oxygen-binding protein of red blood cells, is an example of a globular protein with quaternary structure.

*Enzymes* are catalytic proteins; they change rate of reactions without being consumed by them. Enzymes are very selective in the reactions they catalyze and therefore are a large factor in determining which chemical processes will occur in the cell at any particular time. The principles of how enzymes function in a biological system are quite simple: the enzyme binds to its substrate, then a catalytic action of the enzyme converts the substrate to the product, or products, of the reaction.

The activity of an enzyme is determined by:

- Rate of synthesis (transcription)
- Rate of degradation
- Allosteric effects (including competitive inhibition)
- Concentration of reactants/products (negative feedback inhibition)

## 1.4 Nucleic Acids

Nucleic acids are chains, or polymers, of monomers called *nucleotides*. Each nucleotide is composed of three parts: a nitrogenous base, a pentose (five-carbon sugar), and a phosphate group, as shown in Figure 3.

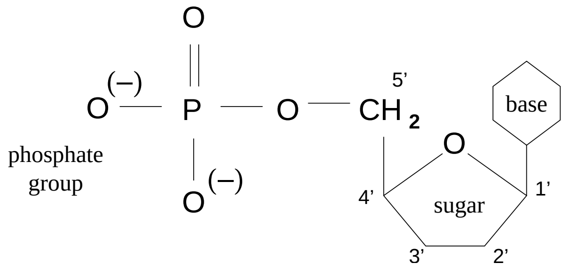


Figure 3: A Nucleotide

In a nucleic acid polymer, or polynucleotide, nucleotides are joined by a covalent bond, called a phosphodiester linkage. The phosphodiester bond links the phosphate group of one nucleotide to the sugar of the next. A *gene*, a unit of inheritance, can be hundreds to thousands of nucleotides long and is encoded in the specific sequence of the four different nitrogenous bases (Adenine(A), Thymine (T), Cytosine(C), and Guanine(G)) that specify the amino acid sequence of a particular protein. There are two types of nucleic acids, *deoxyribonucleic acid*(DNA) and *ribonucleic acid*(RNA). In the nucleotides of RNA the pentose (five-carbon sugar) connected to the nitrogenous base is *ribose*, whereas, in DNA it is *deoxyribose* (see Figure 4).

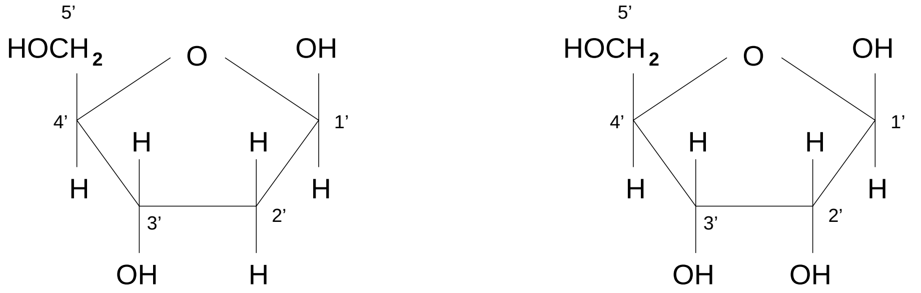


Figure 4: Deoxyribose and Ribose

The only difference between these two sugars is that deoxyribose lacks an oxygen on its $2'$ carbon. The genetic material organisms inherit from their parents is composed of DNA. A DNA molecule is very long, and usually consists of hundreds or thousands of genes. Each DNA molecule directs the synthesis of a type of RNA called messenger RNA (mRNA). The mRNA molecule then interacts with the cell's protein-synthesizing machinery to produce a protein.

## 1.5 Carbohydrates and Lipids

The term *carbohydrate* includes both sugars and their polymers. The simplest carbohydrates are the monosaccharides, a single sugar. Glucose, shown in Figure 5, is the most common monosaccharide. It is of central importance in biological systems because it is the primary nutrient for cells. A *disaccharide* consists of two monosaccharides joined by a *glycosidic linkage*, a covalent bond formed between two monosaccharides by a dehydration reaction (the loss of a water molecule). Polysaccharides are polymers with a few hundred to a few thousand monosaccharides joined by glycosidic linkages. Some polysaccharides serve as storage material (sugar), while others serve as building material for cellular structures, for instance the tough walls of plants.

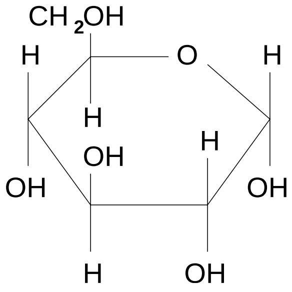


Figure 5: Glucose

The compounds called *lipids* are grouped together because they share one important characteristic: they are hydrophobic molecules. Lipids are the one class of biological molecules that do not form polymers. An important family of lipids is the phospholipids. Figure 6 shows a phospholipid, which is composed of two kinds of smaller molecules, glycerol and fatty acids.

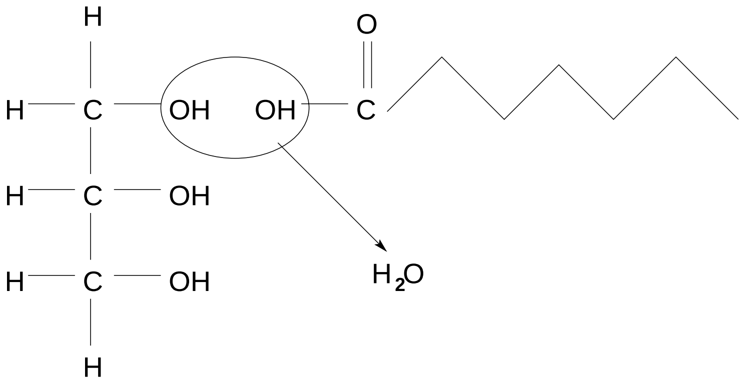


Figure 6: Glycerol Reacting with a Fatty Acid

Glycerol is an alcohol with three carbons, each connected to a hydroxyl group. A *fatty acid* has a long chain skeleton, usually sixteen or eighteen carbon atoms in length. A phospholipid is then assembled from these two components by a *dehydration reaction.*

Phospholipids have both a polar end and a non-polar end. The non-polar ends of distinct phospholipid molecules are oily and are attracted together. The polar end consists of a hydrophilic phosphate group and so is soluble in water. Phospholipids tend to form a *bilayer* when they are in water environment as shown in Figure 7.

Phospholipids are a major component of the cell membrane that forms a boundary between the cell and its external environment. Cell membranes

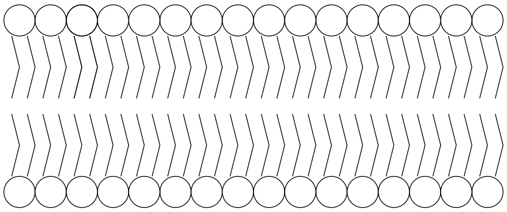


Figure 7: Phospholipid Bilayer

are *semipermeable* which means that water and small non-polar molecules pass through, but large molecules and ions do not pass.

# 2 Cells

Metabolism is what cells do to maintain themselves and reproduce. Every metabolic reaction happens as part of a metabolic pathway. Therefore, we must know something about cells in order to study pathways. Cells have a complex internal structure, with tiny machines and compartments everywhere. For our purposes the most important things about the cell is that it is separated from the rest of the world (by its cell membrane), and its remarkable system for manufacturing proteins.

## 2.1 Nucleus

The nucleus stores a vast amount of information. It is the center for operation of the cell's metabolic and reproductive activities. The nucleus is the largest organelle of a cell. It is surrounded by membranes which serve as a wall between messages and signals traveling in and out of the nucleus. Most of the cell's genes are located in the nucleus. The nucleus contains DNA.

## 2.2 Compartments and Membranes

Membranes separate the cell from the outside world. The membrane is responsible for creating compartments that allow important processes and events within the cell. Like houses are divided into rooms, cells are divided into compartments, each with its own functions and bounded by its own membrane walls.

The membrane contains different proteins and lipids that allow for the functioning of the cell. The membrane is used for protection, regulation of traffic within the cell, and cell recognition; it also provides a passage way for certain molecules and a stable site for binding and catalysis of enzymes [7]. The components that give the membrane its architecture and fluid characteristics are the phospholipid bilayers. The phospholipid bilayer that makes up a cell membrane is a two dimensional fluid. Phospholipids travel freely within the membrane. The membrane also contains embedded proteins which diffuse in the membrane at a slower rate. Membrane proteins may span the membrane or be located in the inner or outer leaf of the membrane. The idea of the membrane as a two dimensional fluid with embedded proteins is known as the fluid mosaic model of biological membranes.

Many membrane proteins are enzymes. The membrane is a location of much of the activity in cellular pathways. The membrane with its different compartments makes it a good location for carrying out enzymatic reactions within the cell.

## 3 Metabolic Pathways

Metabolism represents the sum of the chemical changes that convert nutrients, the raw materials necessary to nourish living organisms into energy and chemically complex finished products of cells. Metabolism consists of hundreds of enzymatic reactions organized into discrete pathways. Metabolic pathways are composed of systems of interacting proteins, which form the basis for nearly every process of living things. Proteins do most of the work of managing the flow of energy, synthesizing, degrading and transporting of materials, sending and receiving signals, exerting forces on the world, and providing structural support. A large proportion of these chemical processes that underlie all of these activities are shared across a very wide range of organisms. These shared processes are collectively referred to as *intermediary metabolism.*

Most of the biochemical processes in intermediary metabolism are catalyzed reactions. Since these processes would hardly take place at normal temperatures and pressures, special proteins, called catalysts or enzymes, are needed to facilitate these reactions. The materials transformed by catalysts are called *substrates.* Unlike the catalysts themselves, the substrates are transformed by these reactions, and the reactions can proceed in one direction. Some reactions are reversible, which means they can proceed in either direction.

The combinations of reactions which accomplish these tasks are called metabolic pathways. The transformations of intermediary metabolism may involve hundreds of catalyzed reactions, so these metabolic pathways may be very complex. Because of the many steps in these pathways and the presence of direct and indirect feedback loops, they can exhibit counterintuitive behaviors. Chemical reactions are also going on in parallel. So a metabolic pathway is a sequence of feasible reactions steps from some set of inputs to some set of outputs. The inputs are called substrates and the outputs are called metabolic products. The pathway flux is the rate at which substrates produce products. The slowest step in a metabolic pathway or series of chemical reactions determines the overall rate of the other reactions in the pathway. In an enzymatic reaction, this rate-limiting step is generally the stage that requires the greatest activation energy.

An example of a metabolic pathway is glycolysis, shown in Figure 8. In this pathway, energy is released from glucose and captured in the form of ATP under anaerobic conditions. ATP is uniquely situated between the very high-energy phosphates synthesized in both the breakdown of fuel molecules and the numerous lower-energy acceptor molecules that are phosphorylated in the course of further metabolic reactions. ADP can accept both phosphates and energy from the higher-energy phosphates, and the ATP thus can donate both phosphates and energy to the lower-energy molecules of metabolism. The ATP/ADP pair is an intermediately placed acceptor/donor system among high-energy phosphates. Thus, ATP functions as a very adaptable but intermediated energy shuttle device that interacts with many different energy-coupling enzymes of metabolism.

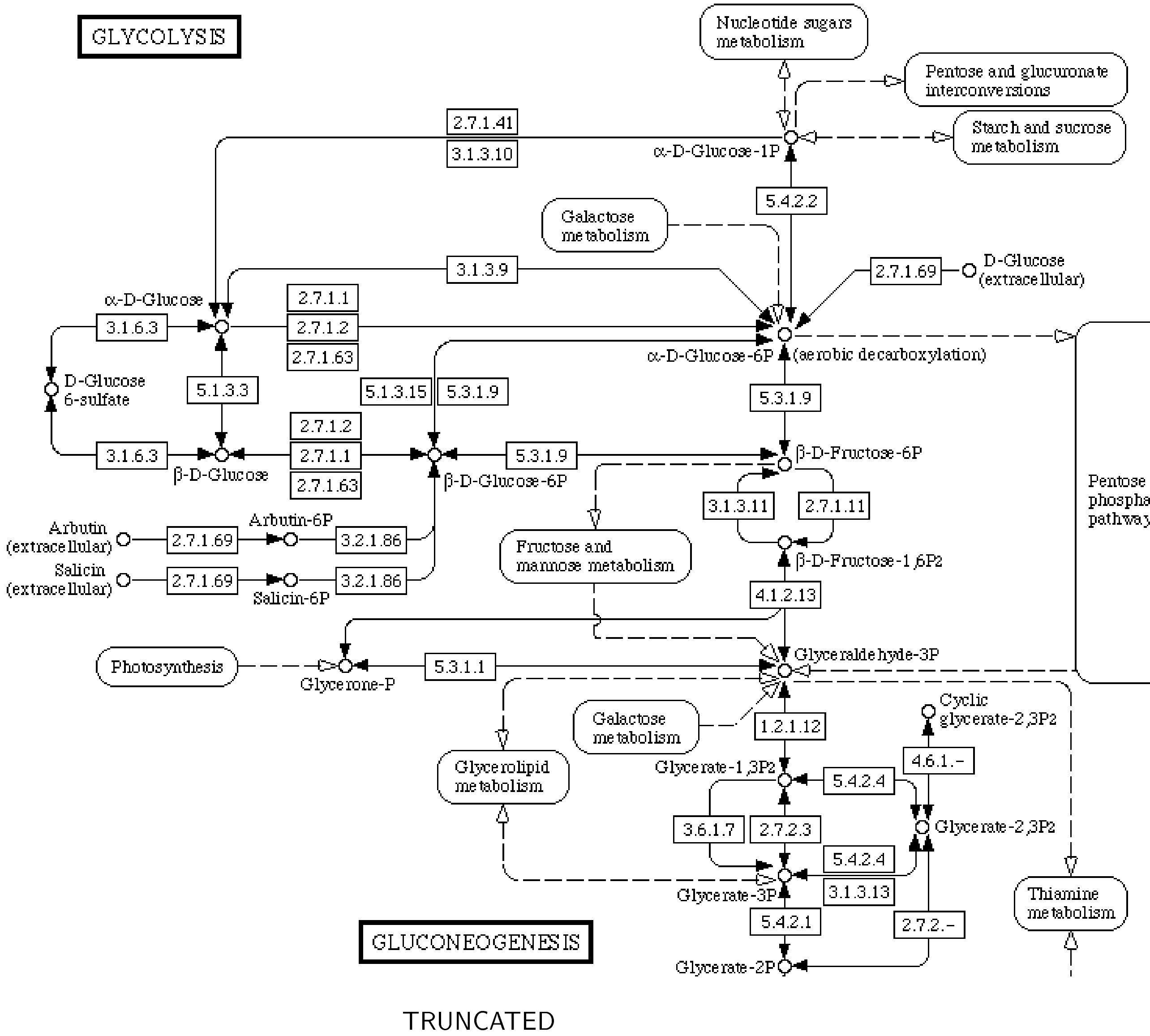


Figure 8: Glycolysis Pathway Map (from KEGG [5])

Overall, the concept or model of a metabolic pathway allows us to draw a map with components of interest and their symbolic names. Metabolic maps are very important and useful in understanding pathways. Maps can be used to portray all of the principal reactions, including the intermediary metabolism of carbohydrates, lipids, amino acids, nucleotides, and their derivatives. Graphical views with arrows modulating flows into and out of components. These maps facilitate a standardization which can be used by practitioners in a similar way to gain major insights and allow the investigator to develop a clear picture of the processes within the system.

## 4 Cell Signaling Pathways

Cell signaling pathways evolve as a necessity to respond to changes in the environment. Signaling transduction can be mediated by kinases, which are enzymes that phosphorylate other enzymes, and these enzymes themselves can activate other enzymes.

The movement of signals can be simple, like small ion movement into our out of a cell. More complex signal transduction involves the coupling of ligand-receptor interactions to many intracellular events, such as phosphorylations by tyrosine kinases. Protein phosphorylations change enzyme activities and protein conformations, resulting in an alteration in cellular activity. A mitogen activated protein kinase (MAPK) is a signaling pathway identified by its activation in response to growth factor stimulation of cells. On the basis of *in vitro* substrates, the MAP kinases have been variously called microtubule associated protein-2 (MAP-2) kinase, myelin basic protein (MBP) kinase, and others.

Two MAPK pathways were shown in the main report: (1) the incomplete MAPK in a human brain (presented by Katheleen Gardiner); (2) Ras-MAPK cascade. Figure 9 shows a short pathway, taken from the Cell Signaling Network Database, from the Epidermal Growth Factor (EGF) receptor to a MAPK cascade.

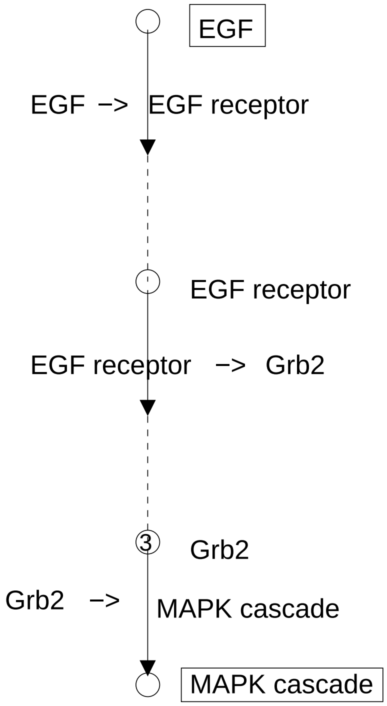


Figure 9: From EGF Receptor to MAP Kinase Pathway (copied from CSN [1])

Cell signaling is among the most fascinating topics, and only recently have we begun to understand its fundamental role in understanding disease and potentials for therapeutics. A comprehensive introduction is given by Krauss [4].

## Bibliography

[1] Cell signaling networks database (CSN). World Wide Web, http://geo.nihs.go.jp/csndb/.

> This is not only a database, but also a knowledgebase for signaling pathways of human cells. It compiles the information on biological molecules, sequences, structures, functions and biological reactions that transfer the cellular signals. The source of the data is from articles in *Nature*, *Science* and *Nature Cell Biology*.

[2] L. Hunter, editor. *Artificial Intelligence and Molecular Biology*, Cambridge, MA, 1993. MIT Press.

> This is now available at http://www.aaai.org//Library/Books/Hunter/hunter.html. Chapter 1, by the editor, provides a good introduction to biology for computer scientists and mathematicians. This is highly recommended for a gentle, informative introduction. Other chapters of direct relevance are by Karp [3] and Mavrovouniotis [6].

[3] P.D. Karp. A qualitative biochemistry and its application to the regulation of the tryptophan operon. In L.E. Hunter, editor, *Artificial Intelligence and Molecular Biology*, pages 289–324, Cambridge, MA, 1993. MIT Press.

> This paper discusses the representation and simulation of biological information so that it can be used in various situations. This is done through the particular example of a bacterial gene regulation system, the tryptophan operon of *E. coli*. The possibility of improving simulation programs that predict the outcome of a gene regulation experiment is explored. The GENISM simulator is held up as a simulator that will efficiently do this. Karp explores its functions, possibilities and limitations. Some results of simulations run through GENISM are presented.
> — Jennifer Phillips

> The focus of this chapter is on the issues of representation and simulation of the gene regulation system of the trytophan operon of *E. coli*. The author presents a model, Gensim, which describes the biochemical reactions that determine the expression of the genes, the reactions by which the genes direct the synthesis of enzymes, and the reactions catalyzed by these enzymes. He then presents a detailed discussion of the implementation of this model.
> — Tod Morrison

[4] G. Krauss. *Biochemistry of Signal Transduction and Regulation*. Wiley-VCH, Weinheim, FRG, 2nd edition, 2001.

> This book is a must for everybody who wants to model signal transduction pathways. It provides the user with necessary facts and principles about signal transduction and regulation. The only negative aspect of the book is that the literature list is not as extensive and almost only restricted to review articles. Of special interest might be chapter 13 (cell cycle) and chapter 15 (apoptosis).
> — Jens Eberlein

[5] Kyoto encyclopedia of genes and genomes (KEGG). World Wide Web, http://www.genome.ad.jp/kegg/.

> This database contains pictures of pathways, hyperlinked to give information about the parts. Basic references to the literature are cited, and some are available as pdf or postscript files.

[6] M.L. Mavrovouniotis. Identification of qualitatively feasible metabolic pathways. In L.E. Hunter, editor, *Artificial Intelligence and Molecular Biology*, pages 325–364, Cambridge, MA, 1993. MIT Press.

> This chapter describes an algorithm for the synthesis of biochemical pathways. Biochemical pathway synthesis is the construction of pathways which produce certain target bioproducts, under partial constraints on the available reactants, allowed by-products, etc. Given a set of stoichiometric constraints and a database of biochemical reactions, this algorithm transforms an initial set of available bioreactions into a final set of pathways by and *iterative* satisfaction of constraints. After explaining the design of the algorithm, the author presents a case study of its application to study of the synthesis of biochemical pathways for the production of lysine from glucose and ammonia.
> — Tod Morrison

> This article discusses the use of an AI method for finding quantitatively feasible metabolic pathways. In order to quantify a pathway's feasibility, the method uses information on the types and amounts of enzymes, ratios of metabolites, and the likelihood of a reactions occurrence in a particular direction within the pathway. The chapter discusses how the AI algorithm works and gives an abstract problem as an example.
> — Rob Wilburn

[7] *MIT Biology Hypertextbook*. Massachusetts Institute of Technology, http://esg-www.mit.edu:8001/esgbio/7001main.html, latest edition, 2002.

> This has undergone maturation since its first posting, and it is a fairly complete introductory resource. The chapters of most direct benefit to understanding pathways are *Chemistry Review, Large Molecules, Cell Biology, Enzyme Biochemistry, Glycolysis and the Krebs Cycle,* and *Prokaryotic Genetics and Gene Expression.* (Of course, the other chapters are also im-

portant.) You can stretch your knowledge by working their "practice problems."