

David S. Goodsell

## **Representation and Misrepresentation in Biomolecular Science**

### **Abstract**

Biological molecules are too small for direct imaging, so synthetic images are used for study and presentation of molecular structure. Three basic types of images are widely used, using different fictional elements to display different aspects of the structure: bond diagrams represent each covalent bond as a single line; spacefilling diagrams represent each atom as a sphere; and ribbon diagrams are schematic representations of the molecular topology. The widespread, and nearly exclusive, use of these three types of diagrams has influenced the course of biomolecular research and the perception of molecules in the popular media.

visual solutions to a visual problem E.

Tufte

### **1. Introduction**

The study of biological molecules, upon reflection, is in many ways a visual problem. Proteins, DNA and RNA are physical objects with defined shapes, sizes, and forms. They are best understood through our most familiar mode of exploring physical objects: through sight. The problem, of course, is that real molecules are small to the level of invisibility. The challenge is to answer the question „What does a molecule look like?“ in a manner that is insightful, and not overly artificial.

Current understanding of atomic structure rests on the unfamiliar world of quantum mechanics. Atoms are described as a probabilistic cloud of electrons surrounding a compact nucleus of protons and neutrons. Under most circumstances, atoms interact weakly when they approach closely, but strongly disfavor overlap with other atoms. Some atoms interact through sharing of electrons (“bonding”) to form molecules of defined stoichiometry and geometry. Biomolecules are large and complex, containing hundreds to millions of atoms, all with a defined spatial relationship to the others.

Diagrams were used from the beginning to classify biomolecular structures and to comprehend them by analogy to real-world objects and processes. The challenge is to simplify this information in a diagram that imparts useful insights, without throwing away essential information. A variety of fictional elements are employed to represent various aspects of the atomic nature and bonding geometry of the molecules. Today, three basic models are most commonly employed (Olson & Goodsell, 1992a,b). Each may be traced to the insight of a scientist who developed a new model to reveal a new aspect of molecular matter.

## 2. Bonding diagrams

Covalent bonding was explained in terms of electron pairing by G. N. Lewis at the end of the nineteenth century, providing a set of empirical rules that revolutionized the study of chemistry. These rules were given more detailed explanation in terms of quantum mechanics in the first part of the twentieth century. In many cases, however, the complexity of quantum mechanics may be discarded, saving only a simple description of the distances between atomic centers, and their relative orientation. These rules are conveniently captured in a bond diagram, where a line is drawn connecting each atom center in a pair of covalently-bonded atoms (Figure 1). Bond diagrams are used throughout chemistry, and are the work-horses of biomolecular science.

Bond diagrams became particularly popular with the advent of interactive molecular graphics. The first hardware was designed for rapidly drawing lines, so the bond diagram, already popular in chemistry, was a perfect fit for the new science of biomolecular structure. A color representation was perfected, with each bond colored according to the atoms connected.

A major problem with bond diagrams is their complexity. Their readability may be improved by replacing each line with a cylinder, and drawing a small sphere at each atomic position. These ball-and-stick models are popular, in part, because of the familiarity of most researchers with physical ball-and-stick models made of wood or plastic, commonly used in chemistry. Use of a shading model

and the occlusion of the rear portions of the molecule add to the realism of the image, making it more interpretable.

### 3. Spacefilling diagrams

Bond diagrams can be misleading. They effectively show the covalent structure of a molecule, but they omit all information displaying its atomic nature. In particular, the repulsive aspect of atoms, disallowing overlap, is missing. Linus Pauling developed the spacefilling representation to restore this information (Figure 2). He placed a sphere around each atom that defined the space occupied by the atom into which other atoms could not easily penetrate. These spheres were used to build physical models of molecules. Many molecular insights were gained, including explanation of barriers to rotation about bonds, and the complex steric properties of alpha helices. Small plastic snap-together models were widely used for chemistry, but became cumbersome when applied to large biomolecular structures.

Computer graphics implementation of spacefilling diagrams has progressed in fits and starts. Until recently, they have been available only to specialists, because of the expense in computation. Today, however, rapid mapping algorithms combined with the power of personal computers and workstations allow calculation of spacefilling diagrams on most machines (although interactivity often suffers).

Many variations on the theme of spacefilling representations have proven useful. The most common is the concept of the solvent-accessible surface of a molecule (Figure 2). This surface is operationally defined by rolling a water molecule (typically a sphere) over the entire surface of the molecule, and saving all of the points at which it will touch. This defines the surface that a molecule presents to the environment, and has important physical properties related to molecular stability and interaction. The surface is composed of two types of surfaces: contact surfaces, where the water molecule touches directly at a point, and reentrant surfaces, which are formed when the water bridges a cleft or valley.

Spacefilling surfaces may be simplified in applications where a gross structure is needed. Spherical harmonics have been used to create a multi-resolution model of molecules (Duncan & Olson, 1993). The process starts with a typical spacefilling surface, and then generates a series of spherical harmonic approximations to the surface. Illustrators do this in a more intuitive manner, often using simple cartoons of molecules, that capture only the gross shape and size.

## 4. Ribbons and ladders

Proteins, DNA, and RNA are all modular molecules, built as a linear chain of subunits. The ways in which these chains fold and interact to form the biologically-relevant structures is critically important to understanding their function. First in the study of DNA, and then in the study of proteins, a schematic representation of the chain provided a truly remarkable clarity.

The report of the double helical structure of DNA by Watson and Crick (Watson & Crick, 1953) was accompanied by the celebrated ladder diagram. The diagram shows the two strands of DNA as helical ribbons, and the nucleotide bases as rungs connecting the ribbons. The ladder diagram may be used to explain DNA replication and transcription in the most simple terms. Jane Richardson refined a similar representation for proteins (Richardson, 1981), creating a diagram with helical ribbons for alpha helices, arrows for beta sheets and simple strands for other regions (Figure 3). As well as lending insight into protein folding and structure, the diagrams are quite beautiful and proved to be an instant success both among scientists and popularizers of science.

## 5. Fictional elements and interpretive conventions

Bond diagrams and spacefilling representations, and even more so in DNA ladders and protein ribbon diagrams, use fictional elements to represent selected properties of the molecules depicted. The lines in a bond diagram are fictional: the interaction between the nuclei connected in the diagram is repulsive and the bond is formed by the cloud of surrounding electrons, which are not depicted in any way. The surfaces in spacefilling diagrams are fictional: the electron distribution and the forces between atoms are continuous, with no such discrete breaks or defined limits. DNA ladders and protein ribbons are even further from the physical reality, using fictional elements to represent entire segments of molecules instead of single atoms. The meaning of these fictional elements must be understood by the viewer; out of context, the diagrams make no sense.

The most useful fictional elements are those that capture a property of the molecule, creating a representation where our intuitive understanding of the diagram matches the physical properties of the molecule. The three examples given here—lines, spheres, and ribbons—are excellent in that respect. The lines capture many of the properties of molecular bonds, including their length, their geometry relative to other bonded atoms, their stability, and the ability for molecules to change conformation by rotating around bonds. Spheres capture

the physical bulk of atoms, presenting the approximate size of the molecule. Ribbons capture the folded structure of the molecules, allowing comprehension of processes involved in folding and unfolding. These representations have been tested through use. They are simple, but not trivial, representations of molecules, imparting maximal information with the minimum of fictional marks.

Appropriate fictional elements have not been found for other molecular properties. Electrostatics remain a severe problem for representation. Biological molecules contain many atoms that carry formal electric charges. These charges interact favorably or unfavorably, often significantly modulating the stability and activity of the molecule. Two common representations of electrostatics were developed by physicists: isopotential surfaces and flux lines. These representations are good for simple systems and are often depicted only in two-dimensional sections. For more complex systems, however, the representations become unmanageable. Computer graphics with interactive manipulation, transparency, and heightened visual cues for depth are necessary to create illustrations of even moderate interpretability.

Dynamics are critically important in biomolecular function, but are a real problem for molecular representation. Animation may be used to show motions over a short period of time, but it is difficult to encapsulate relevant motions into a single picture or concise animated segment. One approach developed for crystal structures of small, organic molecules is the use of thermal ellipsoids. Instead of representing each atom as a sphere, an elongated ellipsoid is used to enclose the space occupied by an atom as it vibrates back-and-forth. These are successful for small vibrational motions, but for the more complex motions of biomolecules, the ellipsoids cause more confusion than clarity. Other similar fictional elements, such as vectors aligned along major avenues of motion and overlapped structures along a dynamic pathway, suffer similar problems when applied to large systems.

## **6. The influence of images on biomolecular science**

Today, many user-friendly computer programs are available for creating molecular illustrations. The wide availability of sophisticated computer graphics tools has shifted the creation of figures away from scientific illustrators and into the laboratory, allowing researchers to create most of their own imagery for publication and presentation. The availability of computer graphics tools carries with it, however, a grave danger. The tools are designed to create a defined set of images from a given set of atomic coordinates, most often including options for basic bond, spacefilling, and ribbon diagrams. Other representations are far more difficult to produce, and are typically the domain of specialists in biomolecular computer graphics. A typical researcher is limited to a few common representational schemes.

These common modes of representation have biased our current biochemical knowledge, coloring our view of molecular structure and function, and perhaps inhibiting some lines of research that are not effectively represented. Perhaps the greatest limitation has been the static nature of molecular imagery. Most molecular illustrations are created from atomic structures solved by x-ray crystallography, which gives a snapshot of the molecule in one static form. In reality, molecules are dynamic, flexing under the thermal motion of their own atoms and the surrounding environment. Many of these motions are crucial for the proper function of the molecules. Advanced graphics tools and animation have been available to specialists, but most researchers rely on static images.

When allosteric motion was first discovered, this static conception of molecules led to similar static theories of allosteric function. Allosteric proteins are complexes of several subunits, which change shape as they perform their biochemical tasks (Figure 4). The first crystallographic structures of these molecules showed “active” and “inactive” states for the complex. Naturally, these were translated into two-state models where the complex switched directly from an “all-inactive” state to an “all-active” state, and back. Although this model was surprisingly successful, later work has shown that the transition is far more complicated, with many subtle intermediates in the process.

The widespread use of ribbon diagrams has led to a chain-centered view of protein structure and function. The secondary structure depicted in these diagrams is given tantamount importance, underplaying the role of sidechains in molecular stability and recognition, and stripping proteins of their rich atomic character. These diagrams are easily misused in improper contexts, misrepresenting the properties they are meant to illustrate. For instance, I attended a lecture where the structure of a new ring-shaped protein was described. To demonstrate the size of this ring, the lecturer showed an illustration of trypsin fit within the ring. Unfortunately, ribbon diagrams were used, which grossly underestimated the bulk of the protein. With a spacefilling diagram, it was later seen that the central cavity was actually far smaller, in fact, too small to be of functional consequence (compare the size difference in Figure 3). Similarly, popularizers of science often use ribbon diagrams inappropriately. Ribbon diagrams serve a specialized purpose, of explaining the folding of a chain into a native structure. They are, however, quite compelling from a design standpoint, so they are commonly used in popular publications to represent what a molecule “looks like.” A spacefilling diagram would be a far better choice, as it retains the concept that molecules are composed of atoms, and more accurately represents the size and shape of the molecule.

Molecular illustration has also fostered a view of biological molecules in splendid isolation (Figure 5). Classical biochemistry typically studies individual biomolecules in pure form, isolated in dilute solution from their cellular environment. X-ray crystallography continues this approach, revealing the

structure of a single molecule or a complex of two or three. The biochemical results and the crystallographic structures are fantastically detailed, but utterly divorced from their cellular context. In cells, molecules are crowded amongst many neighbors. They must compete with other molecules for resources and they must minimize interaction with improper partners. The lack of methods for imaging cells at the molecular level has inhibited study of these processes. Images may be simulated by combining diverse information from protein structure, protein distribution with cells, and cellular ultrastructure, placing biomolecules back into their proper cellular context (Goodsell, 1991).

## **7. Molecular imagery in popular media**

Science has revealed two worlds that are unfamiliar to daily experience: the world of the very large and the world of the very small. The astronomical realm has been a great success, in part due to exciting and accurate visual representation. We can look to the sky and translate in our minds the tiny points of light into entire worlds or fiery suns. Astronomical imagery may be found in technical and popular literature, in fiction and in fact, based in part or in whole on the latest scientific findings. The invisible, inner world of our bodies has not fared as well. For instance, the mention of antibodies will rarely elicit a picture of Y-shaped molecules and mention of viruses will rarely bring to mind their beautiful icosahedral shells. Perhaps this is due to the connection of astronomy with romantic subjects like space exploration and alien intelligence, and the connection of molecular biology with less palatable subjects such as medical treatment for disease and ethical questions raised by biotechnology.

Molecules are a mystery to most of the public, because molecular imagery has remained largely the realm of scientists and armchair scientists. This is a growing problem, given that biomolecular science is playing an ever-increasingly role in our daily lives. We must choose between genetically-engineered products or unmodified products. We must weigh the advantages and the side effects of numerous pharmaceuticals, and decide if we want to be informed of a genetic predisposition for Alzheimer's disease or cancer. Molecular biology now plays a newsworthy role in criminal forensics and archeology. To understand these developments and their ethical implications, an understanding of the molecules and their properties is essential.

Two images have entered popular culture, and probably comprise the basis on which many people understand molecular structure (Figure 6). The first is the planetary atom, a pervasive icon symbolizing the atomic age, and the second is the DNA ladder, now an equally pervasive icon for biotechnology. Both of these images are wonderfully compact, and woefully misrepresentative. They may be likened to

a stick man. The stick man shows the basic structure of a human body, but most viewers add a vast quantity of additional experience to a stick man and easily understand its implications. When we see a Keith Haring a stick man, we easily understand what he is doing—running, thinking, jumping, or whatever. Unfortunately, most people come upon the planetary atom or the DNA ladder with no similar experience, so the schematic is the primary source of information. Most people have no intuitive understanding of the size and context of an atom or a DNA double helix, and base their entire knowledge of atoms and DNA on these abstract pictures.

## 8. Conclusions

The most successful representations are those that encapsulate a key aspect of the molecules and present it in a clear, interpretable manner. Bond diagrams, spacefilling diagrams and ribbon diagrams are well-worn solutions, each presenting one facet of the structure. They are representative, however, only in as much as the creator and the viewer each understand the conventions involved. Given that we have no intuitive understanding of atoms and molecules, the possibility for misrepresentation and misinterpretation are great.

## References

- BS Duncan & AJ Olson (1993) Approximation and characterization of molecular surfaces. *Biopolymers* 33, 219-229.
- DS Goodsell (1991) Inside a living cell. *Trends Biochem. Sci.* 16, 203-206.
- DS Goodsell & AJ Olson (1992) Molecular illustration in black and white. *J. Mol. Graphics* 10, 235-240.
- P Kraulis (1991) MolScript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* 24, 946-950.
- AJ Olson & DS Goodsell (1992a) Macromolecular Graphics. *Cun. Opinion Struct. Biol.* 2, 193-201.
- AJ Olson & DS Goodsell (1992b) Visualizing biological molecules. *Sci. Am.* 267(5), 76-81.
- JS Richardson (1981) The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* 34, 167-339.
- MF Sanner, AJ Olson & J-C Spohner (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* 38, 305-320.
- C Upson, TJ Faulhaber, D Kamins, D Laidlaw, D Schlegel, J Vroom, R Gurwitz & A van Dam (1989) The Application Visualization System: a computational environment for scientific visualization. *IEEE Comp. Graphics Appl.* 9, 30-42.
- JD Watson & FHC Crick (1953) Genetic implications of the structure of deoxyribonucleic acid. *Nature* 171, 737-738.



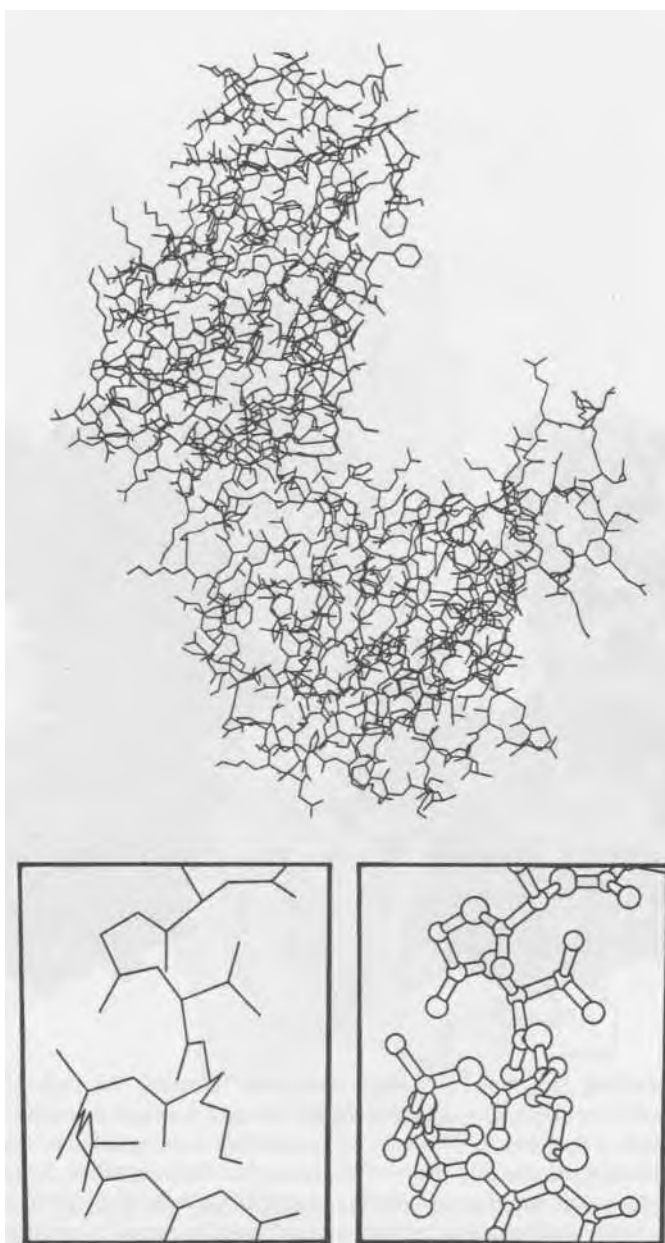


figure 1. Bond Diagrams. Bond diagrams are constructed by drawing a line to connect each pair of covalently-bonded atoms. For proteins, such as phosphoglycerate kinase shown here, bond diagrams can be quite complex. The clarity can be improved, as shown in the details below, in a ball-and-stick version. The illustrations were created with the popular computer program MolScript (Kraulis, 1991) using coordinates from file 3pgk from the Protein Data Bank (<http://www.pdb.org>)

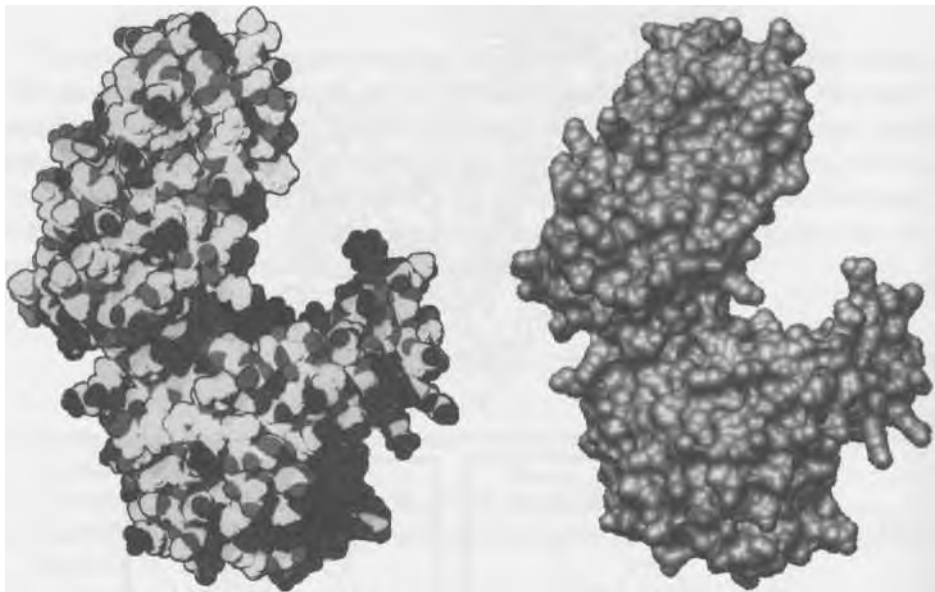


Figure 2. Spacefilling Diagrams. A sphere is used to represent the bulk of each atom in a spacefilling diagram, as shown on the left. At right, a solvent accessible surface displays the surface that may be contacted by surrounding water molecules. Both are excellent for showing the size and shape of biomolecules. The illustration on right was created using the computer programs MSMS (Sanner, Olson & Spehner, 1996) and AVS (Upson et al., 1989), and the one on the right with software developed in my laboratory (Goodseli and Olson, 1992).

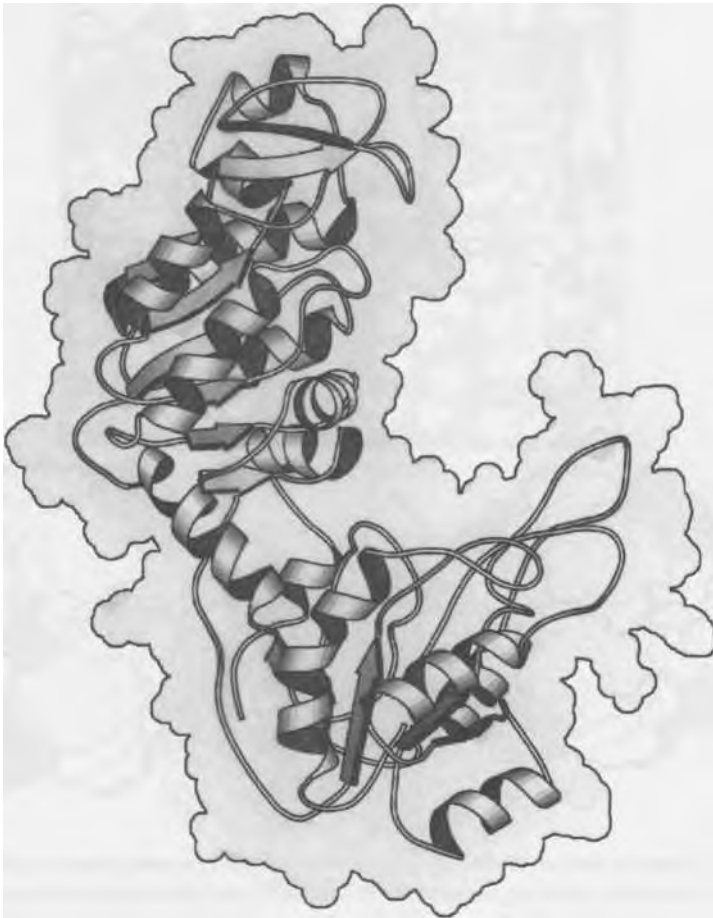


Figure 3. Ribbon Diagram. The ribbon diagram displays the course of the protein chain as it folds and twists into the compact, native structure. Two special forms are used for particularly stable substructures: spring-shaped helices for alpha structure and solid arrows for beta structure. An outline of the spacefilling diagram for this same protein is shown for comparison. Illustration created in MolScript (Kraulis, 1991).

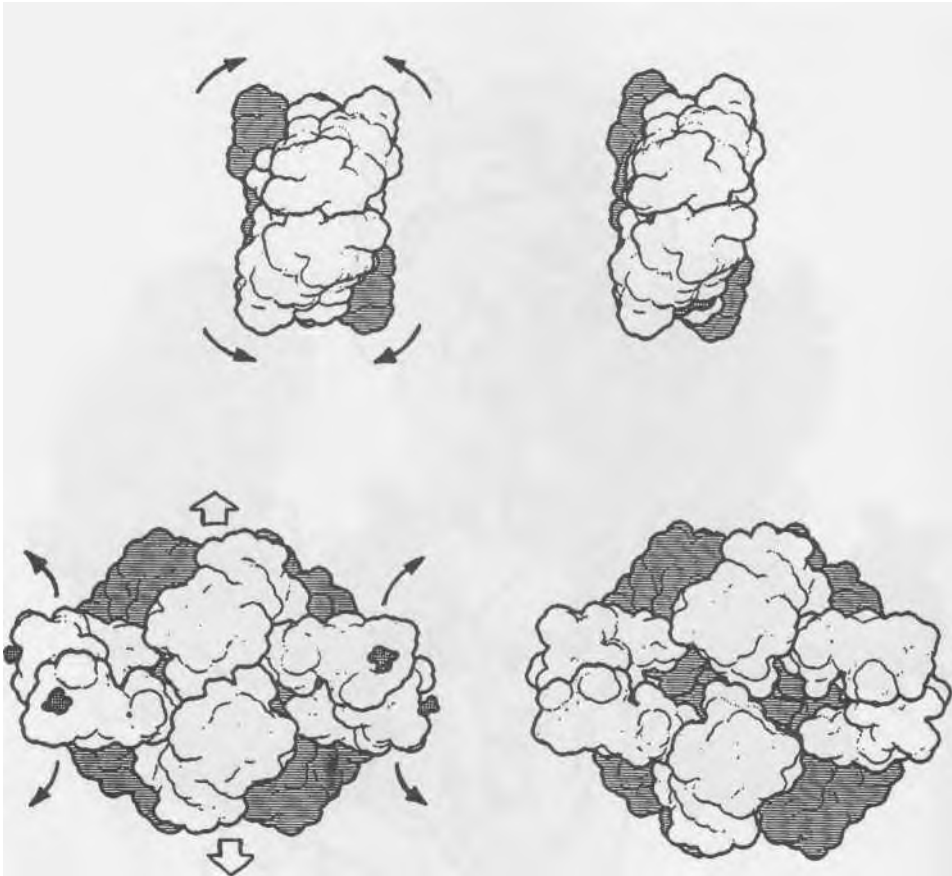


Figure 4. Allosteric Motion. Crystallographic analysis gives a two-state picture of allosteric motions in proteins, providing the structure of fully active and fully inactive forms. Two examples are shown here: fructose-1,6-bisphosphatase at top and aspartate carbamoyltransferase at bottom. Illustrations such as these make this two-state switching mechanism seem like a valid mechanism for the transition, but in reality, the transition is more complex, with many intermediates. The coordinates were taken from entries 4fbp, 5fbp, 4atl and [at] from the Protein Data Bank.

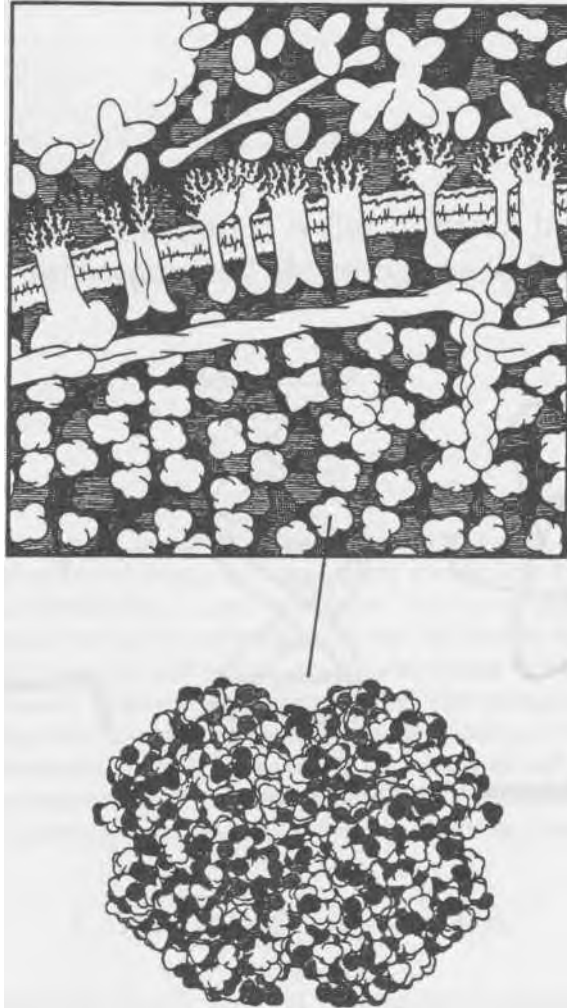


Figure 5. Biomolecules in Context. Biomolecular representation is most often involved with single molecules, shown in atomic detail. The hemoglobin molecule at bottom is an example. Illustrations such as this, motivated by results of biochemistry and molecular biology that look at purified biomolecules, separate the molecules from their cellular context. In the body, hemoglobin is crowded into red blood cells, as shown at the top. The hemoglobin coordinates were taken from entry lhbo from the Protein Data Bank, and the cellular illustration is taken from my „The Machinery of Life,” Springer-Verlag 1993.

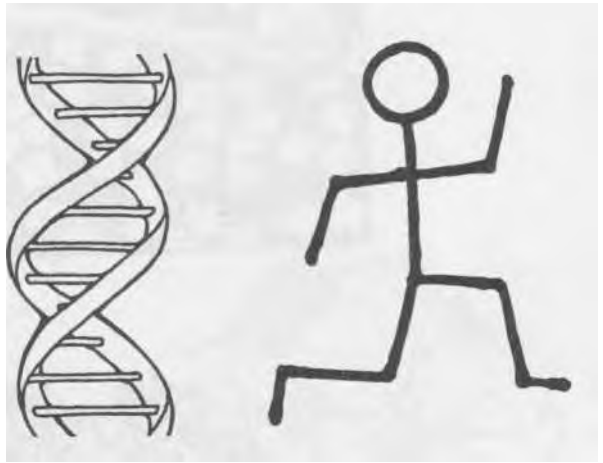
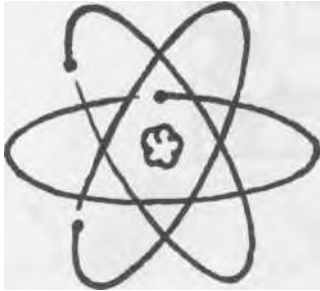


Figure 6. Popular Schematic Representations.