Molecular Behavior in Biological Cells

The Bacterial Cytoplasm as a Model System

Adrian H. Elcock and Andrew S. Thomas

1. Introduction

Given a choice, most, if not all of the molecular machines covered in this book would probably prefer to perform their functions in their intended physiological environment, rather than in a test-tube. This is in contrast to the wishes of most experimental and computational biophysicists, however, for whom there are compelling reasons for choosing to study molecules in vitro, not least of which is the fact that reconstituting interesting behavior with purified components (proteins, DNA etc.) can unambiguously establish the identities of the factors responsible for the observed behavior. But the in vitro environment, while being very well suited to providing clean, unequivocal insights, is in many respects very different from the actual environment encountered by most molecular machines, and this is especially true for those that function inside biological cells. First, there is clearly a vast increase in environmental complexity: a protein in vitro may only be accompanied by a few other types of molecules (water molecules, dissolved salts, and buffer molecules that help to maintain pH), but the same protein inside, for example, the cytoplasm of a bacterial cell may encounter hundreds to thousands of different kinds of protein molecules, RNA molecules, metabolites (glucose, ATP etc.), lipid molecules, and even the bacterial chromosome (DNA). A number of these molecules may be required partners for the protein, e.g. cofactors or substrates of enzymes, or other proteins with which it may form important functional complexes. But the vast majority of them will not be, and each protein must therefore be able to find its intended partners from a very wide array of unsuitable alternatives.

Not only is the compositional complexity vastly greater in vivo, but so is the total concentration of macromolecules. The macromolecular concentration inside a human red blood cell (erythrocyte), for example, is around 350 mg/mL, while that in the cytoplasm of the bacterium Escherichia coli is around 300 mg/mL, of which ~220 mg/mL is protein. For comparison, the protein concentration inside typical crystals (of the type used to solve structures by X-ray crystallography) is of the order of 150 mg/mL; it is therefore interesting that the cellular interior can remain sufficiently like a fluid that macromolecules can still diffuse (albeit somewhat slowly) from one end of the cell to another. This is especially so given that most proteins, when concentrated to
such high levels will “crash out”, i.e. precipitate from solution. As we shall see, this very high concentration of “stuff” can, in principle, have consequences not only for the kinetics of macromolecular processes but also for their thermodynamics.

2. A Simple Analogy for Molecular Behavior in vivo

This chapter will be concerned primarily with outlining the ways in which molecular behavior in vivo might be different from that observed in vitro. As a starting point, we will illustrate potential differences by examining a concrete example: the diffusional behavior of the protein CheY from E. coli; most of the same basic principles should apply, however, to more or less any protein. CheY is a small 128-residue (14 kDa) protein that shuttles back and forth between chemoreceptors (which sense chemical signals present in the exterior environment) and the flagellar motor (which turns the bacterium’s flagellum either counterclockwise, driving the bacterium forward, or clockwise, causing it to tumble).

How are we to think about the diffusional aspect of CheY’s function at a molecular level? To answer this question we will try to construct a real world analogy starting from the moment that the protein becomes phosphorylated and released by the histidine kinase CheA (which is attached to the membrane-bound chemoreceptors) to the moment that it modulates the organism’s swimming behavior by directly binding to the flagellar motor control protein, FliM. One possibility is to think of CheY as a courier, bearing a message that is important, perhaps vital, to the survival of the nation (i.e. organism). We will make life somewhat simpler for ourselves by restricting ourselves to a 2D setting: we will consider the cell as analogous to a large airport terminal. CheY’s job, therefore, is simply to carry its message (phosphate group) from one end of the terminal to the other, where another person — more important than our courier — awaits to receive the message.

Before we even consider adding the rest of the intracellular milieu, it is useful to imagine how things might work if, aside from the chemoreceptors and the flagellar motor, an E. coli cell were to be filled only with water. In terms of our analogy, this would correspond to a situation in which the entire terminal is deserted. We might naively imagine that passing the message to the receiver requires only that the courier walk the length of the terminal. But since our courier and recipient represent molecules, they have neither eyes, ears, noses nor any other senses that might enable them to identify each other at long distances. Furthermore, it is important to remember that since they also do not have brains, they have no conception of what it is that they are “supposed” to do: the courier in our scenario, for example, doesn’t know that he/she is supposed to deliver a message anymore than the recipient knows that he/she is supposed to receive a message. Because of these issues, it is more realistic to assume that our courier will not recognize our recipient until they literally bump — face to face — into each other. We can mimic this in our real world model by applying a blind-fold to both the courier and the recipient, and by assuming that they have no other way of communicating with one another other than by a sense of touch.

It is also important for us to recognize that our courier will not simply move across the platform in a straight line: molecules in aqueous solution do not move as if shot from a cannon (i.e. ballistically); instead, on the kinds of timescales and lengthscales that we are interested in here (nanoseconds and nanometers and beyond), they move diffusively, that is, they are subject to repeated collisions with solvent molecules that continually reorient and redirect them, leading
them to exhibit Brownian motion. A simple way to add this to our real world situation is to imagine that our courier is continually buffeted by sudden, short-lived gusts of wind that come from completely unpredictable directions (clearly, it’s not a very well designed airport terminal). Note that while this “drunken sailor” method of movement may not seem particularly efficient, it may, depending on the time- and length scales over which movement is required, be completely adequate for the purposes of the cell. But in other cases — e.g. when cargo has to be transported from one end of a eukaryotic cell to the other (as, for example, happens in nerve cells in the human body) — diffusion is not efficient enough and motor proteins, moving along pseudo-static “tracks”, are therefore needed (see other chapters in this book). In passing, it is worth noting that we could extend the analogy to account for such motor-driven transport by imagining that the courier, still blind-folded, occasionally stumbles on to the terminal’s moving walkways; note that there is no guarantee that he/she will end up on a walkway moving in the right direction!

Having struggled to set up a crude analogy for how a message might be passed from one protein to another in a biological cell filled only with water, we can now rerun the entire scenario in a way that more realistically mimics intracellular conditions. As noted above, the principal difference is that there are very high concentrations of other macromolecules present; we can account for this by imagining that the platform is now extremely crowded with other people, some the size of small children, some the size of Shrek, but all of them blind-folded. Clearly, the situation will be chaotic: our courier, who was already having a difficult enough time navigating the deserted terminal is now also repeatedly jostled and blocked by the crowd. We can probably imagine that this will slow the courier considerably although the extent to which he/she is slowed may perhaps depend on his/her size: small children, for example, are often quite adept at finding their ways through crowded rooms, whereas we could imagine that Shrek might have a great deal more difficulty. A schematic illustration of these differences is shown in Figure 1. But we can also imagine a more complicated situation. What if the courier encounters a member of the crowd who happens to share some of the physical features of the message’s recipient? In such cases, we must expect that the courier will tend to remain close to the “impostor” for some period of time (perhaps a very long time), with the result that his/her diffusive exploration of the environment might be significantly slowed. In molecular terms, such a situation could occur if CheY encounters macromolecules that have similar electrostatic potentials or similar patches of exposed hydrophobic residues, to those of the flagellar motor protein FliM. In passing, therefore, we can speculate that evolution might have acted to decrease the extent to which such non-specific (and unintended) interactions are likely to occur.

It is hopefully apparent from the above description that we can construct a real world analogy for the situation encountered by CheY in the E. coli cytoplasm, but we have to be prepared to alter it in some fundamental ways to make it work. While it is probably useful at an illustrative level — since it allows us to begin to frame our thinking about what life might be like inside a cell — it should not be thought that it describes all of the possible differences between the in vitro and in vivo environments (for example, Section 3).

3. Macromolecular Crowding Effects

One of the more subtle effects of the highly crowded environment faced by molecules in vivo can only be revealed by using some ideas from statistical thermodynamics. This is the effect of
steric ("excluded-volume") interactions between a protein of interest (again, e.g. CheY) and its environment of surrounding “crowder molecules”. As we will see, such interactions can, in principle, have very significant thermodynamic effects, and this realization has led, in recent years, to the development of an entire sub-field of biophysics devoted to understanding and predicting effects due to “macromolecular crowding”.7,8 An illustrative example of the crowding effect is shown in Figure 2 in which we consider an idealized protein folding equilibrium. In what follows, we will make a number of simplifications in order to ease our calculations, but it is important to bear in mind that the basic argument and conclusions are unaffected by these simplifications.

We consider a simplified “protein” that can exist in one of only two conformations: one, a highly compact native state, and another, an extended unfolded state. These are shown in Figures 2a and 2b respectively. As shown in these figures, we will assume that the protein exists within the confines of a 2D box consisting of 12 x 12 squares. We will further assume that the energies of both protein conformations are equal. This assumption makes our life considerably easier since, in the language of statistical thermodynamics, it means that the two conformations have identical “Boltzmann weights”; all we have to do to estimate the equilibrium constant for the folding reaction, therefore, is take the ratio of the number of possible conformations of the folded protein and the unfolded protein. In the present case, therefore, $K_{eq} = 1/1 = 1$, which in turn means that the free energy of folding of the protein, which is given by $\Delta G = -RT \ln K_{eq} = -RT \ln (1) = 0$.

Figure 1. A schematic illustration of how diffusion might be affected by highly crowded conditions for (a) a small protein, and (b) a large protein. The diffusing protein of interest is shown as a red circle, the other macromolecules comprising the crowded environment are shown as blue circles. The light blue “haloes” around each crowder molecule indicate the volume that is inaccessible to the diffusing protein due to steric interactions: note that the excluded volume experienced by the large protein (b) is considerably larger than that experienced by the small protein (a). The black lines indicate potential diffusional trajectories for the two proteins.
Next we consider what happens when we add an idealized large macromolecule to the system (represented by the large filled circle in Figure 2c). We will assume, again just for the purposes of making the calculations simpler, that the “crowder molecule” can only occupy the larger, dashed squares on the grid (i.e. in the absence of the protein, it would be restricted to $6 \times 6$ possible positions).

**Figure 2.** A simple 2D model of crowding effects on a protein Folding reaction. (a) Folded state structure of a model protein consisting of 16 “amino acids” (circles) connected by peptide bonds (thick lines); the blue and red circles represent, respectively, the N- and C- termini of the protein. (b) The same protein in a putative unfolded conformation. (c) One possible configuration of the same system with a single added crowder molecule (large grey circle); note that for simplicity we assume that the crowder can only be placed within one of the large dashed squares, of which there are 36 in total. (d) One possible configuration of the same system with five added crowder molecules.
We can now assess the effect of this single crowder on the protein’s folding equilibrium simply by asking how many positions can be occupied by the crowder molecule without clashing with the protein; in the language of statistical thermodynamics this means that we assume that any configuration of the system that has a steric interaction between the protein and crowder has an infinitely positive energy, and therefore has a Boltzmann weight of zero, and so can be ignored. Counting up the number of available positions for the crowder molecule, we find that there are 32 and 29, respectively, for the folded and unfolded states of the protein. The equilibrium constant for the folding reaction is now given by 32/29, which in terms of free energy, is
\[ \Delta G = -RT \ln (1.10) = -0.05 \text{ kcal/mol}. \]

What does this mean? It means that the mere presence of the crowder molecule has shifted the protein folding equilibrium in favor of the folded form. This, in essence, is the macromolecular crowding effect: excluded-volume interactions of the crowder molecules with the protein of interest tend to favor the latter, assuming a more compact state. The reason that it is termed a “macromolecular” crowding effect can be demonstrated by considering what would happen if the crowder molecule was a much smaller molecule, e.g. one that occupies only a single square of the system and which therefore could occupy each of the 12 × 12 small, grey squares in the absence of the protein. In this case, we find that the number of possible positions for placing this “micromolecular” crowder molecule is the same for both the folded and unfolded conformations of the protein (128 in both cases). The equilibrium constant for the folding reaction in this case is therefore given by 128/128, which again is 1, the same result that we obtained in the micromolecular crowder molecule’s absence. In other words, there is no formal “crowding effect” induced by very small molecules. It is important to be clear that this does not mean that the “real life” inclusion of high concentrations of small molecules will cause no changes in the observed thermodynamics of processes such as protein folding (they certainly can); it means only that the changes they elicit are not likely to be due to the excluded-volume effect outlined above.

Returning to our macromolecular crowder molecule, we can consider what happens as its concentration increases by adding multiple copies of the molecule to the system. Now the calculations get more complicated because we have to enumerate all of the possible positions that could be occupied simultaneously by the crowder molecules (this is why we made the earlier assumption of only allowing the crowder to occupy 6 × 6 possible positions). An example configuration of the system is shown in Figure 2d. To calculate the number of possible positions, we make use of a familiar equation: the number of distinguishable ways of arranging M identical objects in N positions = N! / {M! (N – M)!}. Using this relation we obtain the numbers of possible positions for placing two crowder molecules in the system as 496 and 406 for the folded and unfolded states respectively. \( K_{eq} \) for the folding reaction is now 496/406 = 1.22. Repeating the calculations for 5, 10 and 15 crowder molecules, respectively, we obtain values of \( K_{eq} = 1.70, 3.22 \) and 7.29, which in free energy terms amount to −0.31, −0.69 and −1.18 kcal/mol at 298 K. Clearly, adding increasing numbers of crowder molecules shifts the protein folding equilibrium increasingly in favor of the folded state. Following this logic, and noting that the 15 crowding molecule situation is qualitatively similar to the total macromolecular concentrations encountered in the E. coli cytoplasm, it seems an inescapable conclusion that proteins will be significantly more stable in vivo than in vitro. Similar lines of thinking can also be used to estimate the likely effects of a highly crowded environment on protein-protein association and aggregation thermodynamics.9
Calculations such as those just described provide a good rough estimate of expected behavior, but it is important to be aware of the simplifications involved. One obvious simplification is that the amino acids of our model protein, and our crowder molecules, are restricted to occupying discrete positions on a 2D grid. This simplification could be set aside if, for example, we used computational methods to enumerate or sample the possible positions of the protein and crowder molecules in 3D space: this is now routinely done in studies aimed at modeling the thermodynamics of protein-related equilibria (e.g. Refs. 10–11). A second obvious simplification concerns our modeling of the protein’s folding equilibrium. In reality, there are far more unfolded conformations than folded conformations, with relative energies that are much less favorable than those of the folded conformations; again, computational modeling methods can address this issue by generating large numbers of reasonable structural models for the unfolded state.12 A third simplification is that our crowding molecule is treated as a circular disc (it could have been any other shape that fitted entirely within one of the larger squares). At the time of writing, most computational studies that examine crowding effects on biomolecular systems use similar, simple geometric shapes (spheres, rods etc.) to model the crowder molecules. This is almost certainly adequate for the purposes of obtaining estimates of the magnitude of crowding effects, but as we will see below, it is now possible to perform similar studies using “real” protein structures, that therefore more closely mimic the structure of intracellular environments.

Two final assumptions of the above calculations are crucial to note. The first is that the model presented explicitly discounted the possibility of favorable interactions occurring between the protein and the crowder molecules and assumed that their only form of interaction is an antagonistic, steric interaction. In reality, of course, molecules such as proteins interact with each other via any or all means at their disposal: oppositely-charged residues can form transient ion pairs (“salt bridges”), unpaired polar groups (-OH, -NH2 etc.) can form hydrogen bonds to one another, and exposed hydrophobic sidechains can “stick” to one another. Again, we shall see later that it is possible to begin adding in the effects of such interactions into calculations; we will also see that these additional terms can lead us to make quite different conclusions about the potential effects of a cellular environment on the thermodynamics of protein folding. Finally, it should be noted that the model presented above includes no solvent; in reality, solvent molecules would occupy all of the apparently unoccupied squares. By leaving them out of the calculations, we have implicitly assumed that the addition of crowder molecules to the system causes no change to the solvent environment. While this is reasonable as a first approximation, it is by no means certain that it will be true of intracellular environments: there is, in fact, continued debate about the extent to which water inside biological cells behaves like “bulk” water (see, e.g. Ref. 13).

4. The Bacterial Cytoplasm as a Model System

The remainder of this chapter will explore one way in which the simple 2D model described above might be replaced by a structurally detailed 3D model of an intracellular environment implemented on a computer. Our ultimate aim is to build a model that allows us, at a molecular level of resolution, to observe directly how macromolecules in vivo negotiate their way through a minefield of competing interactions while performing their intended function; one might eventually hope, in particular, to use such a model to uncover the fundamental ways in which behavior occurring
in vivo might differ from that observed in vitro. We assume, for now, that the factors affecting macromolecular behavior inside bacterial cells will be essentially the same as those that affect behavior in eukaryotic cells. Certainly, there are major differences between the two types of cells: the cytoplasm of a eukaryotic cell, for example, differs considerably in its total macromolecular concentration and in its composition (e.g. its cytoskeletal elements are far more complex), but even with such differences, it is still much more similar to the cytoplasm of a prokaryotic cell than it is to a simple aqueous solution.

Undoubtedly the best characterized of the prokaryotes is *Escherichia coli*, a gram-negative bacterium that forms spherocylindrical cells of approximately 2 µm length and 1 µm diameter. The amount of accumulated knowledge on this organism is so vast that it is impossible to do more than scratch the surface of it in a single chapter; instead, the reader is referred to a number of online databases that serve as excellent portals for accessing much of this information (see for example, http://www.ecocyc.org, http://redpoll.pharmacy.ualberta.ca/CCDB/, and http://www.york.ac.uk/res/thomas/). At a basic level, *E. coli* can be considered to have three primary intracellular aqueous-phase environments: the DNA-dominated nucleoid, which is surrounded by and contiguous with the protein-dominated cytoplasm, and the thin periplasmic layer that is sandwiched between the inner and outer membranes. If we include also the lipid-phase inner- and outer-membranes, both of which are highly enriched in embedded proteins, it is clear that even this (relatively) simple organism provides a wide range of physicochemical environments in which to study macromolecular behavior in vivo. Accordingly, the rest of this chapter will attempt to use the cytoplasm of *E. coli* as a vehicle for illustrating some of the concepts surrounding molecular behavior inside cells.

5. A Structural Model of the Bacterial Cytoplasm

At least two crucial pieces of information are required for the construction of a molecular model of an intracellular environment such as the bacterial cytoplasm. First, we must know, or be able to guess, the relative abundances of each (macro)molecule to be included in the model. Second, we must have, or be able to build, 3D structures of each molecule. For proteins, answers to the first of these problems comes from the experimental techniques of quantitative proteomics, which attempt to quantify the entire complement of proteins (i.e. the proteome) of a given organism under a given set of conditions. Given the extreme compositional complexity of biological cells, this is an enormous challenge. The most recent efforts to quantify proteins in the cytoplasm of *E. coli*, for example, report quantifications of more than 600 proteins, which is a prodigious feat in itself, but given that there are ∼2900 different proteins thought to be resident in the *E. coli* cytoplasm, it is clear that there is still some way to go. Related experimental techniques aim to quantify each of the small molecule metabolites (e.g. glucose, ATP) in an organism; this is the nascent field of metabolomics. The second requirement mentioned above — that of needing 3D structures of the molecules of interest — is addressed by the experimental techniques of structural biology (X-ray crystallography, NMR spectroscopy and cryoelectron microscopy). The progress that has been made in this field over the last 50 years is truly staggering: at the time of writing, the number of individual structural entries in the protein databank (http://www.rcsb.org) is in the tens of thousands, of which approximately 11% are *E. coli* proteins (it should be noted, however, that many of
these are redundant copies). Despite this progress it will be a long time — if ever — before we have high-resolution structures of the entire complement of proteins in the cytoplasm of *E. coli*. For the foreseeable future, therefore, our model will have to contain a mixture of *bona fide* crystal structures of *E. coli* proteins and so-called “homology models” produced by mapping the *E. coli* protein sequence on to the high-resolution structure of an orthologous protein from a different organism.²¹

An illustration of the type of molecular model that can be constructed²² using currently available structural and proteomic data is shown in Figure 3. This model contains ~1000 randomly arranged macromolecules, each of which is of one of 50 different types of molecules. Individual molecule types have been selected according to their relative abundance in proteomics experiments performed with cells growing on minimal media; the most abundant molecule types in the model are the gene products of TufA, which is the translational elongation factor EF-Tu, and of MetE, which is a methyltransferase involved in methionine biosynthesis; the two proteins are present in copy numbers of 181 and 213 respectively. The least abundant molecule in the model is the very large glutamine synthetase, present in only one copy. Also present are tRNAs, enzymes of the glycolytic and tricarboxylic acid cycle pathways, and the two subunits (30S and 50S) of the protein-synthesizing machine, the ribosome. The combined concentration of the macromolecules in the model is 275 mg/mL, which is close to experimental estimates for the *E. coli* cytoplasm.

**Figure 3.** A structural model of the cytoplasm of *Escherichia coli*.²² RNA components are colored green, with their phosphate groups in yellow; protein components are colored arbitrarily by molecule type. Note the 50S ribosomal subunits (blue + green/yellow) on the left-hand side of the image and the 30S subunit (red + green/yellow) on the right-hand side. This image was prepared with the program VMD.³³ This figure has been adapted with permission from McGuffee & Elcock (Ref. 22).
As a point of reference, it is worth noting that the width of this simulation cell amounts to approximately 1/10th of the diameter of a typical *E. coli* cell. Put like that, it sounds impressive, but in terms of volume it sounds a lot less so: the volume of the simulation cell occupies approximately only 1/1000th of the volume of a typical cell. Clearly, therefore, the model shown in Figure 3 represents only the first “baby step” towards modeling entire biological cells. Readers interested in imagining what the interiors of *entire* cells might look like would do well to examine some of the extraordinary and exquisite images made by Dr. David S. Goodsell (see Suggested additional reading material).

6. Some Issues in Simulating Such a System

While Figure 3 is useful in illustrating just how crowded the interior of a typical bacterial cell can be, it is important to note that the true purpose of this particular model is not to produce static images but is instead to provide a dynamic view of the way molecules diffuse, tumble, and interact with each other *in vivo*. There are many computer simulation methods that might be used, in principle, to bring the static model “to life” so it is perhaps worthwhile to very briefly outline some of the factors that might determine one’s choice of method; more detailed examinations of different simulation techniques can be found elsewhere (see, e.g. Refs. 23–24). Conceptually, the most straightforward way to simulate the dynamics of macromolecules in the cytoplasm is to use the technique of molecular dynamics (MD) simulation. In its application to aqueous phase systems, such a simulation method usually employs “explicit solvent”, which essentially means that the spaces surrounding and within macromolecules are filled — as they are expected to be in real life — with water molecules (and dissolved ions). The essence of MD simulation is then to repeatedly solve Newton’s (classical) equations of motion in order to determine the time-dependent trajectory of each individual atom in the system. To do so requires knowledge of the net force acting on *every* atom at each step of the simulation, which, in principle, requires calculating each atom’s interaction with every other atom in the system. Computationally, this can be a very expensive undertaking: when explicit water molecules are added to the system shown in Figure 3, for example, we obtain a model that contains approximately 45 million atoms. When one realizes that a typical current MD simulation of a single protein — in an aqueous system containing perhaps a total of 60 000 atoms — takes approximately 1 month to simulate 1 µs on 64 cpu cores (operating in parallel) it can be seen that performing MD simulations of the bacterial cytoplasm model would be highly non-trivial.

Because of the expense associated with the MD approach, use is often made of more approximate, faster simulation methods such as Langevin dynamics (LD) or Brownian dynamics (BD) techniques. The two techniques share much in common: in both, the explicit water molecules, and often also the solution’s dissolved ions, are simply removed from consideration. As might be expected, this drastically decreases the number of atoms in the modeled system, and therefore, in principle, results in a huge increase in the (computational) speed of the simulations. But of course, this acceleration comes at a price, and both LD and BD methods must seek to add back in the more crucial features of the missing solvent in order to ensure that the thermodynamic and kinetic behaviors of the solutes are not disastrously compromised. In the case of BD, this involves adding random displacements (abrupt “shoves”) to each solute to mimic the Brownian effect that
would normally be exerted by the missing solvent molecules. (Note that we cannot simply remove
the water molecules and continue to use MD techniques and expect the solutes to behave as they
would in solution: in the absence of water, MD will tend to make the solutes move ballistically,
which we will recall from earlier, is not the way that solutes really move in aqueous solution).

Water, of course, does far more than just repeatedly crash into solute molecules. Pure water,
for example, diminishes (i.e. “screens”) long-range electrostatic interactions to approximately
1/80th of the strength that they would have in vacuum, and dissolved salts in solution screen such
interactions even more. Clearly, any attempt to model the interactions between charged groups
such as the positively and negatively charged amino acids and the phosphate groups of RNAs must
take account of this screening effect. Equally importantly, water is also the key ingredient of the
hydrophobic effect, which in turn is thought to provide the primary driving force for protein folding
and protein–protein association events. Again, any attempt to model folding and association
processes as they might occur in vivo must make some attempt to mimic the fact that exposed
hydrophobic groups have a pronounced tendency to associate with one another in aqueous solu-
tion. As is often the case when approximations are involved, there are differing opinions about how
best to implicitly model water’s effects on electrostatic and hydrophobic interactions, and much
current research in computational biophysics is therefore geared toward evaluating and improving
upon the wide range of possible approaches.

One final consequence of omitting explicit water molecules is that the hydrodynamic interac-
tions (HI) that act between macromolecules in aqueous solution are lost. Conceptually, HI account
for the fact that even when two macromolecules are non-interacting with each other in aqueous
solution, the closer they are to one another, the more their motions will tend to become correlated.
A simple way of thinking about the origin of this effect is to imagine what would happen if one
macromolecule were to be displaced in the direction of the other, for example, by the action of an
external force: the displaced macromolecule will displace water molecules that, in turn, will act to
displace the other macromolecule. Obviously, if water molecules are excluded from the simulation
model then this water-mediated effect will be lost. Computational methods for implicitly model-
ing hydrodynamic interactions have been intensively investigated in studies of colloidal systems
(where solutes might have diameters of ~1 µm) (e.g. Ref. 26) and it may be that similar methods
will be needed for modeling macromolecular dynamics in intracellular conditions. Such hydrody-
namic models are, however, expensive to compute, and it is not necessarily clear at the time of
writing how dire the consequences will be for omitting them. In any case, one important point to
note is that in contrast to the effect of water on the thermodynamics of electrostatic and hydropho-
bic interactions, hydrodynamic interactions change only the dynamics, and not the
thermodynamics, of the system: equilibrium properties, such as free energies of protein-protein
associations, should be identical in simulation models that include or exclude HI.

The admittedly rather radical step of omitting all solvent from the model is one way that we
can make simulations of the cytoplasm model shown in Figure 3 more feasible with the kinds of
computational resources routinely available to researchers. But this still leaves us with a system
that contains perhaps 11 million atoms, and that is therefore far from trivial to subject to simula-
tion. A further approximation that one might consider, therefore, is to treat some or all of the
modeled macromolecules as rigid bodies. This simplification enables dynamical simulations to be
accelerated in two ways. First, the forces that act upon rigid models of macromolecules can be
computed much faster since there is no need to calculate or model any of the interactions that occur within the macromolecule itself; this is especially helpful in the case of extremely large members of the model such as the 30S and 50S ribosomal subunits or the GroEL/ES chaperonin complex. Second, depending on the form of the energy function used, the forces that are exerted by rigid models of macromolecules can be computed much faster since they can be precomputed and stored on a 3D grid that surrounds (and moves with) each rigid molecule: the forces acting on atoms of other molecules that approach the rigid molecule can then simply be read from the corresponding region of the 3D grid. Such an approach is obviously only cost-effective if the effort involved in precomputing the 3D grid is expended only once (i.e. before the simulation actually begins); this, in turn, requires that the molecule generating the grid be truly rigid — otherwise, any relative movement of its atoms would necessitate recomputing much, if not all, of the grid. Of course, this rigid body approximation is likely to be more acceptable for some macromolecules than for others: it will definitely not be appropriate for modeling mRNAs or intrinsically unstructured proteins, since their degree of conformational flexibility is likely to be very high.

7. A Dynamic Model of the Bacterial Cytoplasm

With all of the above simplifications in hand, it has become possible, at the time of writing, to conduct dynamic simulations of the cytoplasm model shown in Figure 3 on a timescale of approximately 20 µs; movies showing the resulting behavior can be found at the authors’ website (http://dadiddly.biochem.uiowa.edu). The overall effect is, as might have been anticipated, rather chaotic: macromolecules undergo relatively short periods of free diffusion interspersed with periods during which they become temporarily stuck — through hydrophobic and/or electrostatic interactions — to other members of the ensemble. Smaller macromolecules can be seen to diffuse comparatively rapidly, while the much larger 30S and 50S ribosomal subunits appear to be effectively stationary over the timescale of the simulations. This lack of motion on the part of the largest members of the model immediately tells us that 20 µs is far from being sufficient for the simulations to completely “sample” the system’s possible configurations: in order to achieve proper sampling of the ribosomal subunits, for example, the simulations would probably need to be extended to a timescale of several milliseconds. But that said, for the more mobile and more abundant members of the ensemble, the story is more encouraging: the faster the molecules diffuse, the more rapidly they explore their environment, and the more representatives there are of a particular molecule type, the more complete is our view of this molecule type’s potential behavior. Because of this, even simulations of 20 µs are likely to be sufficient for obtaining reasonable estimates of the behavior of the more abundant members of the system.

8. Quantifying Protein Diffusion in the Bacterial Cytoplasm

While the overall behavior of the simulation model appears to fit crudely with our expectations — i.e. with the airport analogy that we so painstakingly constructed earlier — it is obviously important to try to obtain a more quantitative assessment of the extent to which the observed behavior is realistic. This is an issue that arises, or should arise, with all work that is based on computer simulation; the only real way of addressing it is to find some observable that can be measured
experimentally and examine whether the value of that observable measured from the simulations matches with it. In the particular case of the bacterial cytoplasm, one good choice of observable is the translational diffusion coefficient of the green fluorescent protein (GFP), since this has already been measured by a number of experimental groups (e.g. Refs. 29–30). Such measurements have been based on the technique of \textit{fluorescence recovery after photobleaching} (FRAP), which briefly summarized involves (a) the use of a laser to photobleach (i.e. destroy the fluorescent properties of) GFP molecules in a selected area of the cell, and (b) the measurement of the time required for the fluorescence in that selected area to recover after the laser pulse. Since the recovery of fluorescence — which occurs on a millisecond timescale — results from GFP molecules that were initially \textit{outside} the laser-illuminated region diffusing \textit{into} it, the rate of recovery is dependent on the translational diffusion of the GFP \textit{in vivo}. The process of extracting an effective diffusion coefficient from the experimentally measured data is somewhat involved, but the key finding of these studies is that GFP’s apparent diffusion coefficient \textit{in vivo} is approximately one-tenth of its diffusion coefficient \textit{in vitro}. Clearly, therefore, the crowded cytoplasmic environment strongly affects the abilities of typical macromolecules to freely diffuse.

The translational diffusion coefficient of a molecule in a computer simulation is, at least in principle, one of the easier quantities to measure: one can simply use the Einstein formula, which in 3D reads: \( \langle r^2 \rangle = 6D_{\text{trans}} \Delta t \), to relate the mean squared distance, \( \langle r^2 \rangle \), traveled by the molecule in time \( \Delta t \), to its translational diffusion coefficient, \( D_{\text{trans}} \). In dilute aqueous solution \( D_{\text{trans}} \) is usually found to be essentially independent of \( \Delta t \); in highly crowded environments such as the bacterial cytoplasm, however, the apparent diffusion coefficients of macromolecules may appear to depend significantly on \( \Delta t \). This phenomenon, which is usually termed \textit{anomalous diffusion}, is often a simple manifestation of a macromolecule’s motion being frustrated by the macromolecules that form its immediate neighborhood. On a very short timescale, this hindrance may not be apparent because the protein is still essentially free to diffuse within the “cage” formed by its neighbors; on a longer timescale, however, it would soon be apparent that the macromolecule has been prevented from moving as far as one might expect; its apparent translational diffusion coefficient on this longer time scale must therefore be somewhat decreased. In the case of the BD simulations of the cytoplasm model shown in Figure 3, it appears that the crossover between the short-time and long-time regimes occurs on a timescale of a few microseconds; this is similar to the average lifetime of non-specific encounters observed between the members of the ensemble. Importantly, however, it is possible to show that the simulation model — which it should be remembered attempts to account for (favorable) hydrophobic and electrostatic interactions between neighboring macromolecules — can produce a long-time translational diffusion coefficient for GFP that is in good agreement with experimental FRAP measurements.\textsuperscript{22}

9. Thermodynamic Stability of Proteins in the Bacterial Cytoplasm

Earlier in this chapter we considered an idealized model showing how the thermodynamics of a protein folding reaction might be affected by the addition of high concentrations of macromolecular crowder molecules. A number of the stated simplifications of that model can be overcome using the cytoplasm model described above: in particular, we now have structurally realistic models of the crowder molecules, we have (hopefully) realistic models of the way that the crowders
might be arranged, and we have an energy model that allows for favorable interactions between macromolecules and that seems to be reasonably realistic, at least in so far as it leads to diffusion coefficients of GFP that match well with experiment. In order to use the cytoplasm model to calculate its expected effect on the thermodynamics of a protein folding reaction, the only additional feature that we need is a way of constructing putative models of the protein’s unfolded conformations, which is a problem that others have nicely solved (e.g. Ref. 12). The mechanics of the actual calculations of folding thermodynamics need not concern us here but they essentially involve computing the average interaction energies experienced by unfolded and folded-state conformations of the protein of interest when immersed in the cytoplasm; this is achieved by attempting millions of random insertions of the conformations into the model. Importantly, the success of these calculations can be assessed for at least two proteins for which there are experimental estimates of the folding thermodynamics in vivo; here we will focus on only one of these proteins, namely the 136-residue protein cellular retinoic acid binding protein (CRABP).

Innovative experimental work carried out on CRABP has indicated that its folding free energy is destabilized by $\sim 1.4$ kcal/mol in vivo relative to in vitro conditions. It will be recalled that this experimental result is the exact opposite of what we expected on the basis of our simple 2D crowding model; it is also the exact opposite of what we obtain when we use a strictly steric interaction approach to describe the interaction between CRABP and our more elaborate 3D cytoplasm model. So what is the reason for this qualitative difference between the calculations and experiment? There are several possible explanations. One is that a significant amount of the unfolded protein in vivo might be bound by chaperone systems. A second is that our earlier assumption that the solvent is unaffected by the addition of high concentrations of biological macromolecules might be incorrect. A third possibility — not necessarily mutually exclusive with the first — is that there are favorable interactions between the unfolded protein conformations and the cytoplasm “crowders” that outweigh the excluded volume effect: it will be recalled that we explicitly excluded this possibility when constructing our simple 2D model in order to focus attention solely on the excluded volume effects. Of the three possibilities, the only one that can be directly addressed with the above cytoplasm model is the third; intriguingly, when we allow for such favorable interactions to occur — with the same energy model used in the BD simulations — we find that the calculations predict that CRABP will be destabilized by $\sim 0.9 – \sim 1.8$ kcal/mol, which is in rather good agreement with experiment.

10. Conclusion

This latter result provides an encouraging indication that computational models such as that outlined here might eventually be routinely capable of making quantitative predictions of the thermodynamics of folding and binding reactions in vivo. If that can be achieved then one would have a computational model of an intracellular environment that can meaningfully be said to bridge the biophysical gap that currently exists between in vitro and in vivo experimentation. In order to achieve this, however, much needs to be done, and hopefully the above description has provided at least an idea of the kinds of issues and challenges that must be faced. Most of the key issues relate to the simplifications currently made in modeling such large systems, and ultimately it is probably fair to say that these will eventually be solved when advances in computing power make the simulation of models such as that shown in Figure 3 possible with explicit-solvent MD
methods. Since it will be some time before such power becomes widely available to researchers, however, it is inevitable that some degree of approximation or simplification in the modeling will be involved. We expect that in the coming years significant advances will be made in rapid modeling of hydrodynamic interactions between macromolecules, in modeling of solvation and electrostatic effects in vivo, and in developing flexible molecular models that can be applied on the very large scale demanded of attempts to model intracellular environments. Finally, it is important to note that while we have focused on the bacterial cytoplasm here the same challenges — and potential rewards — are also likely to arise in attempts to model other intracellular environments, be they prokaryotic or eukaryotic in origin.

Acknowledgments

The views expressed in this chapter have been shaped over a period of years by the work of many people including, in particular, the following members of the Elcock group: Dr. Sean R. McGuffee, Dr. Tamara Frembgen-Kesner, Shun Zhu and Eli Musselman. This work was supported by the National Institutes of Health (GM087290).

Suggested Additional Reading Material


References


