Opportunities at the Intersection of Nanoscience, Biology and Computation

Study Leader:
Ellen Williams

Contributors Include:
Henry Abarbanel
Paul Alivisatos
Steve Block
Michael Brenner
Al Despain
Dave Gifford
Bob Grober
Terry Hwa

Herb Levine
Nate Lewis
Paul McEuen
David Nelson
Tim Stearns
John Vesecky
Bob Westervelt

November 2002

JSR-02-300

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 883-6997

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 2002	

4. TITLE AND SUBTITLE

Opportunities at the Intersection of Nanoscience, Biology and Computation

5. FUNDING NUMBERS

13-9020014-DC

6. AUTHOR(S)

E. Williams et al.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
JASON Program Office - W553
7515 Colshire Drive
McLean, Virginia 22102

8. PERFORMING ORGANIZATION REPORT NUMBER

JSR-03-300

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Department of Energy
Biological and Environmental Research SC-70
19901Germantown Road
Germantown, MD 20874-1290

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

JSR-02-300

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited.

12b. DISTRIBUTION CODE

Distribution Statement A

13. ABSTRACT *(Maximum 200 words)*

This report presents background information and recommendations derived from a JASON study carried out during the summer of 2002 at the request of the Department of Energy Office of Science. The charge for the study was to recommend research topics of potentially great impact at the intersection of the scientific disciplines of nanoscience, biology and advanced computation. In performing this study, we shaped our investigations by the context of the DOE missions in these areas. This led to the formulation of two long term goals appropriate to the DOE that we believe can be furthered by appropriate research programs at the intersection of the three chosen areas.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	None	SAR

Contents

1 EXECUTIVE SUMMARY

Research capabilities in nanoscience, molecular biology and computation have advanced to the point where it is possible to define research activities in which the development of nano-bio systems will support major DOE science goals. Specifically, we identify two major long term research goals which can motivate research at the intersection of nanoscience and biology:

1) Development of biological-systems-control for bioremediation, carbon dioxide sequestration and tailored biomaterials fabrication.

2) Development of artificial nanosystems with biomimetic functionality but without biological fragility.

Basic research in support of these goals can be focused by identifying immediate research challenges involving the integration of physical nanostructures and biological nanostructures (i.e. proteins, with a strong emphasis on membrane-bound proteins) in a program of closely correlated theoretical and experimental research. Two specific research challenges that should be pursued are:

1) Fabricate non-trivial assemblies of physical and biological nano-components with linked functionality, and develop carefully designed experiments to directly compare measured behavior to results of systems modeling.

2) Fabricate assemblies of proteins in which protein-protein, protein-lipid or protein-artificial nanocomponent interactions can be tailored, and test/tune computational capabilities to predict changes in their corresponding function.

There are many non-trivial problems that must be overcome in addressing these challenges. Their solutions will draw upon and impact shorter

term research goals aimed at developing improved whole-cell diagnostic tools, new chemical linking methods for assembling biological and physical nano-components, and exploiting the computational data bases resulting from high throughput research approaches.

2 INTRODUCTION

This report presents background information and recommendations derived from a JASON study carried out during the summer of 2002 at the request of the Department of Energy Office of Science. The charge for the study was to recommend research topics of potentially great impact at the intersection of the scientific disciplines of nanoscience, biology and advanced computation. In performing this study, we shaped our investigations by the context of the DOE missions in these areas. This led to the formulation of two long term goals appropriate to the DOE that we believe can be furthered by appropriate research programs at the intersection of the three chosen areas. These long term goals are:

1) Development of biological-systems-control for bioremediation, carbon dioxide sequestration and tailored biomaterials fabrication.

2) Development of artificial nanosystems with biomimetic functionality but without biological fragility.

While no detailed road-map can be laid out for such sweeping goals, it is possible to define intermediate term basic research objectives that will enable the discoveries that will be needed to reach these grand goals. Our philosophy in considering the research areas appropriate for immediate focus was to identify first, from an experimental perspective, the status and potential of studies at the interface of nanoscience and molecular biology. The results of this survey, presented in section 3, clearly demonstrate both the feasibility and the opportunities of developing combined physical and biological nanosystems. Based on this finding, we then considered the potential computational impact of such research. In Section 4 we specifically consider the potential for designing nano-assemblies in the context of clarifying and testing whole-cell simulations. In Section 5, we address the potential impact

of studies using nano-assemblies on molecular computation. In keeping with our initial approach, our recommendations maintain a strong emphasis on experimental validation and interaction with computational studies.

The report contains many sections in which a detailed review of an illustrative topic is presented. To clarify the key conclusions of the individual sections, these are summarized in Section 6, along with the overall findings of the report.

In preparing this report, we were fortunate to have access to advance copies of several DOE reports:

> "Theory and Modeling in Nanoscience," Report of the May 11-12, 2002 Workshop, by Bill McCurdy, Ellen Stechel, Peter Cummings, Bruce Hendrickson, and David Keyes.
>
> "Report of the Workshop on Biomolecular Materials," Jan 13-15, 2002, by Mark D. Alper and Samuel I. Stuupp .
>
> "Genomes to Life Goal 4 Roadmap: Computational Methods and Capabilities," Office of Advanced Scientific Computing Research and Office of Biological and Environmental Research of the U.S. Department of Energy Office of Science.

We are also grateful to the scientists who provided briefings on the topics of the study. During the summer of 2002, we heard from:

A. Christian (LLNL)
M. Colvin (LLNL)
W. A. Goddard (Caltech)
J. Groves (Berkeley)
R. Kelley (DOE-BES)
T. Michalske (Sandia)
H. Wang (Santa Cruz)

We were also fortunate to be able to draw on briefings provided for a previous JASON study on nanoscience and biology, run by Steve Block during the summer of 2001. The scientists who presented briefings for that study were:

R. Austin (Princeton)
R. Colton (NRL)
T. Kenny (Stanford)
C. Mirkin (Northwestern)
M. Roukes (Caltech)
S. Stupp (Northwestern)
D. Tomalia (Dendritic Sciences, Inc.)
M. Wightman (UNC)

3 NANOSCALE COMPONENTS AND SENSING

It is not immediately apparent how to define the interface between nanoscience and biology. The approach in nanoscience stems from the physical sciences, so we refer to the components of nanoscience as physical nanocomponents. The goal of much research in nanoscience is to develop nanocomponents of controlled physical characteristics relevant to physical sensing, e.g. via light, electricity, magnetism, etc. and signal transduction. The development of nanoscience is largely following a bottom-up path, beginning with the development of nano-components. Finding methods for assembly of these components remain a major research challenge.

In contrast, the goal of much research in molecular biology is to extract from the complexity of living cells a basic understanding of the underlying components. These components are primarily proteins. Because of their size and their astonishing diversity of function, it is quite appropriate to refer to proteins as biological nanocomponents, in analogy to the physical nanocomponents of nanoscience. The possibility of isolating and using biological nanocomponents is most readily grasped in considering the motor proteins, illustrated in Figure 1. However, as we will see throughout the report, the possibility of capturing protein function in artificial environments is a common theme.

A final issue in considering the interface of nanoscience and biology is the mechanism of information transfer, or interaction between nanocomponents. As noted above, physical nanocomponents often (although not exclusively) are designed for physical sensing. However, biological nanocomponents almost exclusively interact via chemical transduction. Developing effective methods of interaction between physical and biological nano-components is an important research issue, as is also the development of physical linkages

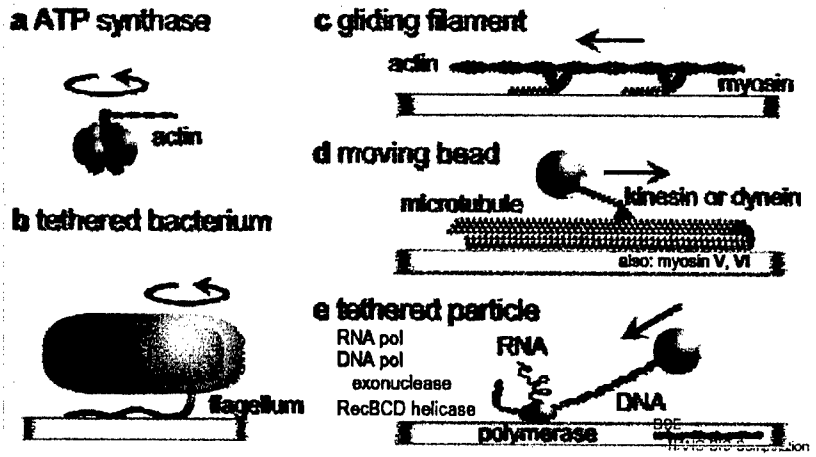


Figure 1: The development of single-molecule assays has made it possible to isolate and characterize the activity of individual motor proteins. Figure courtesy of S. Block, Stanford University.

between biological and physical nanocomponents.

As we will see in the following, the need for improved diagnostics in biotechnology has been an important driver in developing the nanoscience-biology interface. This continues to be an important area of research in its own right, and thus can be expected to continue to help drive developments in "nano-bio". Moreover, the improved information derived from these diagnostics will play an important role in the support of whole cell simulation and molecular computation as discussed in Sections 4 and 5. In Section 3, we will address, some of the issues in understanding cellular function that have begun to drive the development of physical nano-scale sensors. We will then review some examples of nanoscale sensors, and conclude with a discussion of the potential uses of biological structures (membranes and proteins) as nano-components in artificially assembled systems.

3.1 Questions for cellular function at the nanoscale level

Many questions in molecular biology are asked at the level of the behavior of individual proteins, or protein complexes. A typical globular protein is a sphere of <10 nm diameter, and much of what happens in a cell happens through the regulated interaction of proteins. Our ability to discern the location, activity, and interactions of proteins within cells is poor, due to the lack of appropriate tools from the world of biology and chemistry. What do biologists want to know?

Where is a particular protein within a cell? Cells consist of many subcellular organelles and functional domains that are specialized for particular functions. Thus, the location of a protein within the cell is often indicative of the function of that protein. Determining location is currently accomplished by either immunological techniques, such as immunofluorescence, or immuno-electron microscopy, or by constructing a fusion between the protein of interest and a naturally fluorescent protein, such as the green fluorescent protein (GFP) from the jellyfish Aequoria victoria. The cloning of the gene for GFP about ten years ago made possible the genetic tagging of individual proteins in living cells, without the need for purification or chemical modification. Although this technique has revolutionized cell biology, it is limited by the small number of excitation and emission wavelengths that are available from the various modified forms of GFP, by the stability of the GFP protein, and by the fact that GFP is itself a medium-sized protein that sometimes interferes with the function of the protein to which it is fused.

What are the dynamic properties of localization, synthesis and turnover for a protein? The function of most proteins is regulated in part by control of the amount of protein in the cell, and by localization of the protein to a particular compartment of the cell. Each of these properties is

potentially highly dynamic – for example, the half life of a cyclin protein can differ by orders of magnitude during the course of a single cell division cycle, and the localization of transcription factors can change from cytoplasm to nucleus within seconds of activation by an external signal. Although these parameters can be determined by bulk measurements on populations of cells, single cell measurements provide much greater resolution. For example, it is possible to determine the exchange rate of a protein at a particular location in the cell by making use of a technique called "fluorescence recovery after photobleaching" (FRAP). In this technique a focused laser is used to photobleach a fluorescently tagged protein in a limited area of the cell. The time required for proteins in the bleached zone to be replaced by fluorescent copies of the protein from the surrounding area indicates the exchange rate of the protein in that structure. The utility of such techniques is limited by the available fluorescent molecules with which proteins can be tagged. GFP and its derivatives are the most common fluorophores used in these techniques, and have the limitations described above.

For proteins that have a regulated activity, where are the active molecules in the cell, and when are they active? Many of the most important reactions in cells are controlled by transient activation of proteins. For example, one class of receptors that receives signals from the environment is activated by auto-phosphorylation – the receptors phosphorylate themselves upon stimulation, leading to transduction of a signal. There is currently no way to distinguish in living cells the "active" proteins from the pool of mostly "inactive" proteins. This would require reagents that are specific for the active form of a protein, and that could be used in living cells.

What other proteins or molecules are physically associated with a particular protein? Perhaps the most fundamental problem in modern biology is trying to understand the interactions that take place between proteins in the cytoplasm of cells, an environment that is about 100 mg/ml protein concentration. Almost all processes in biology involve the

interaction of multiple proteins, and the identification of specific binding interactions is one of the major goals of the post-genome sequencing effort. There are many techniques for identifying such interactions, however most only work outside of the native environment of the proteins. For example, the two-hybrid system is a powerful method for identifying interactions that makes use of expression of proteins of interest in yeast cells that have been modified for the purpose of detecting interactions. Clearly it would be preferable to have tools that would allow the identification of interactions in situ. A technique that is becoming more commonly used is fluorescence resonance energy transfer (FRET), in which the interaction of two fluorescently tagged proteins is identified by energy transfer between their fluorophores under the appropriate conditions of excitation. Since the fluorophores must be within a few nm to transfer energy, it is only likely occur between two proteins that are in direct physical contact. As for the above cases, this type of analysis is limited by the available fluorescent tags, currently mostly derivatives of GFP.

How does the cellular environment change in response to internal and external stimuli? Cells are responsive to signals from the outside through the action of cell surface receptors, and channels that allow the passage of specific small molecules. They also respond to events that occur inside the cell, such as the completion of the discrete events of the cell cycle (mitosis, cytokinesis, etc.). In many cases, signals are mediated by changes in the concentration of small molecule second messengers, such as Ca^{++} ions, or phospholipids, such as phosphatidylinositol-4,5-bisphosphate. To fully understand signaling in biology, it will be necessary to observe the transient changes in the levels of these second messengers in vivo. As an example, there are several successful techniques for Ca^{++} imaging, using either a natural Ca^{++} sensitive fluorescent protein, or by manipulating GFP so that its fluorescence is Ca^{++} sensitive.

How do changes in the cellular environment impinge upon protein activity? The signals described above must ultimately cause a change in activity of relevant proteins to effect a change in cell function. In most cases it is unclear how this is achieved, due to a lack of suitable reporters for protein conformation and activity in vivo. An example of a technique that has recently been successful is a derivative of the FRET system described above. Rather than place the fluorescent probes on two different proteins whose interaction is being assayed, the probes are placed on two distinct domains of the same protein. If a change in environment causes a change in the structure of the protein such that two domains that were far apart are now close together, then energy can be transferred between the fluorophores.

3.2 Nanoscale Tags for Cellular Interiors

As noted above, the key to understanding cellular function is to map the interactions between biomolecules *in vivo*. To this end, one needs to develop nanoscale *in vivo* probes of cellular function at the level of cellular components. Generically, one uses optical imaging to probe the interior of biological structures, as photons are the only practical probe for which these materials are transparent. Traditional methods of optical microscopy have yielded detailed images of cellular systems. The use of confocal microscopy in particular has yielded three dimensional sectioning of cells at the submicron level [1].

Optical spectroscopy of individual molecular fluors, both organic and inorganic, is now a decade old [2] –[5]. It is well understood that the combined use of spatial and spectral discrimination can be used to isolate individual fluors prepared at very dilute densities on surfaces, in thin film samples and in cells [6]. Whereas conventional molecular spectroscopy is performed on ensembles of many molecules, the great strength of the single

molecule experiment is that it allows the experimentalist to sample inside the molecule ensemble average, obtaining information at the level of the individual molecule. One distinct advantage of this technique is that individual fluors can be located with 10–20 nanometer precision using conventional confocal optics. In particular, the position of the fluor is measured as the centroid of its diffraction limited image, which can be determined with accuracy of order $dx = \Delta x/SNR$, where dx is the width of the spot and SNR is the signal to noise ratio [7, 8, 9, 10].

3.2.1 Optical sensing with functionalized quantum dots

It should be a goal of any nanobiology initiative to push towards nanoscale optical diagnostics of biological systems. A promising path towards this goal is the use of functionalized semiconductor nanocrystals, also known as quantum dots (QDs), combined with the techniques of single molecule imaging. While semiconductor nanocrystals were first developed for application in microelectronics and photonics, they are finding their niche application in biology. Initial interest in these materials focused on the promise of designer photon sources, as the optical emission energy of a QD increases as its size decreases. However, it has recently been realized that QDs have valuable advantages over conventional organic fluors for *in vivo* cellular imaging [11]. First, they are highly photostable (i.e. less likely to bleach), allowing real-time tracking over periods of hours. Their absorption is broad band, allowing QDs of several different emission wavelengths to be photoexcited by a single excitation wavelength. Finally, their emission lifetime is typically tens of nanoseconds, enabling time gated experiments to discriminate QD emission from the cellular autofluorescence background (1–3 nsec).[12]

QDs can be made water soluble by coating with silica/siloxane [13] or with bifunctional ligands, such as mercaptoacetic acid or dithiothreitol [14]. QD surfaces can be functionalized and it has already been demonstrated

that they can be linked to peptides, proteins and DNA [15] as suggested in Figure 2. Finally, their size is typically 1-6 nm and thus comparable to that of biological macromolecules.

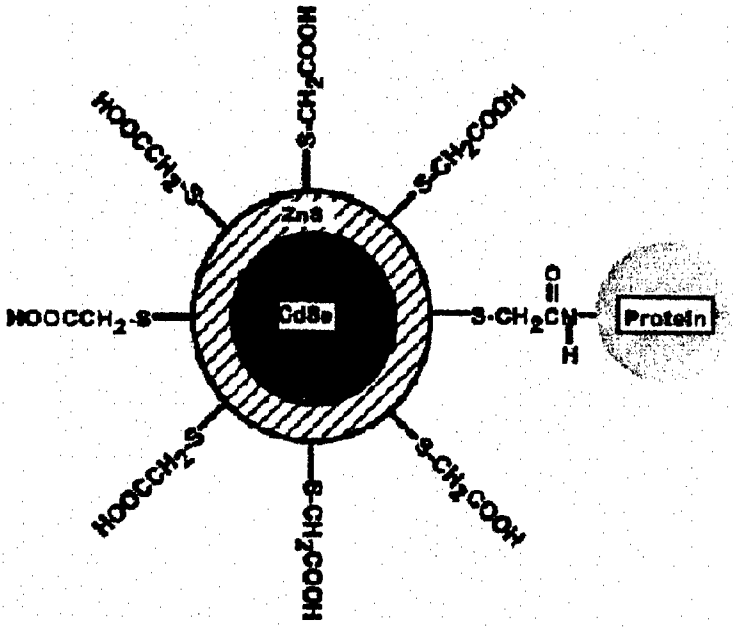


Figure 2: Schematic illustration of a ZnS-capped CdSe Quantum Dot covalently coupled to a protein by a mercaptoacetic acid link. From: W.C.W. Chan and S. Nie, Science 281, 2016 (1998).

We believe that the combined use of QDs and the techniques of single molecule imaging will yield a robust probe of cellular function. Experiments will likely include functionalized QDs that attach preferentially to particular proteins and/or cellular structures, each species of QD fluorescing in a unique emission band as shown in Figure 3. One can then use multicolor imaging techniques to follow these nanoprobes as they diffuse through the cellular environment, tracking the motion with nanometer resolution.

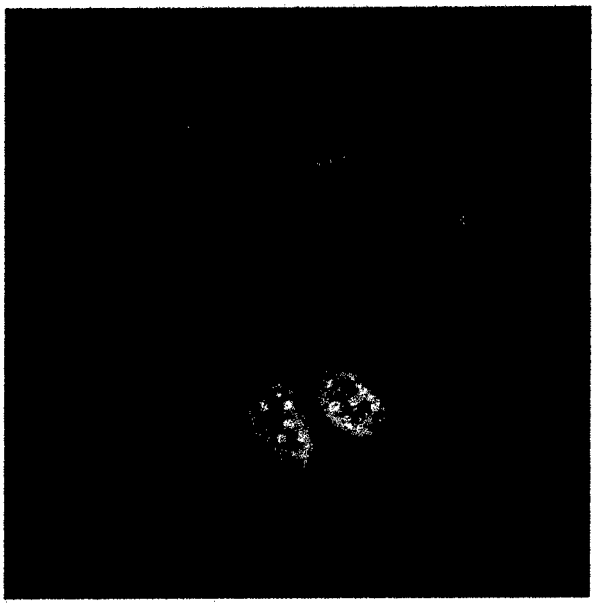


Figure 3: Mouse 3T3 fibroblasts impermeated with biotinylated red nanocrystals and unbiotinylated green nanocrystals and supported on wire grids for imaging. Multi-wavelength imaging with fluorescence confocal microscopy. Actin filaments have bound red nanocrystals preferentially. From: X. Michalet, et al., Single Mol 2, 261 (2001)

3.2.2 Other nano-particle probes

One issue that arises in whole-cell sensing is the toxicity of some potentially useful dyes. Another is interference in the activity of the dyes due to interaction with cell components. The tools of nanoscience offer the possibility of embedding such dyes in a nanoscale matrix of material that is compatible with in-vivo sensing [16]. The matrix can be composed of hydrogel, sol-gel glass, or liquid polymer materials that allow transport of material from the aqueous cellular environment to dye trapped in the matrix. A number of methods for delivering the nanoscale sensors to the cell interior are shown in Figure 4. Such nanoscale matrix-encapsulated probes have been used for sensing a variety of ionic species as well as oxygen. Mixed components can be encapsulated for ratiometric sensing, or for coupling sensitive ionophores

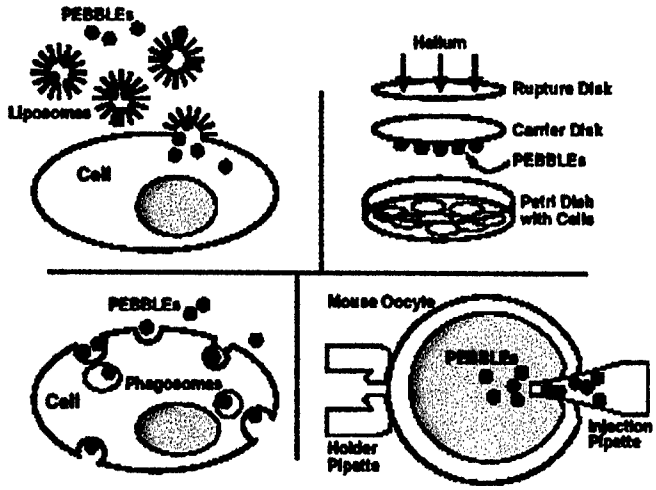


Figure 4: Methods of inserting nanoparticles into a living cell. Clockwise from upper right: gene gun, pico-injection, phagocytosis, and liposomal delivery. Figure from Clark, et al. Sens. and Act. B51, 12 (1998).

with reporter dyes [16]. Another interesting nano-particle application is for controlled absorption of light of a specific wavelength to create a mechanism for controlled local heating. Halas and West [17] have constructed nanoshells of a thin gold layer (d=5 to 20 nm) coating a dielectric core (diameter $\sim$ 60 nm) of Au or silica. These nanoshells have rather specific absorption spectra peaking at about 1100 nm, the near IR, for d=5 nm, and 750 nm, in the visible, for d = 20 nm. The proposed application of these absorbing nansoshells is to embed them in a polymer matrix containing a biochemical which can be released upon heating. Preliminary demonstration showed that the polymeric matrix could be designed to have a release temperature (critical solution temperature) above body temperature. Local heating by illumination of the embedded nano-shells at the resonance wavelength was successful in initiating release of the biochemical.

3.2.3 Magnetic Nanoparticles

Magnetic beads attached to DNA molecules or other biological structures can be manipulated by applying a gradient in magnetic field that pulls the bead to the high field region. Polymer coated magnetite beads with diameters ~ 1 μm are widely used for this purpose. Magnetic nano-particles can be fabricated, and if functionalized (as discussed above for quantum dots, see also section 4.1), could be used for steering nano-components or possibly liposomes in solution.

Micro-electromagnets can be used to manipulate magnetic beads of the type used for DNA in fluids at room temperature. A matrix consisting of two perpendicular arrays of wires can be used to create a maximum in magnetic field magnitude that can trap a magnetic bead and move it continuously along a plane in the fluid with resolution greater than the wire spacing [24]. Using computer control, a large number of magnetic beads can be trapped and separately moved to assemble DNA and other biological molecules and carry out experiments that are more complex than those possible with only one or two beads.

The magnetic field available from a micro-electromagnet matrix is approximately a few hundred Gauss at the location of the particle in fluid. This is enough to make the magnetic pinning energy of a ferromagnetic particle equal to kT at room temperatures for a 10 nm diameter particle, in round numbers. The magnetic field and pinning energy are limited by thermal heating, and the limits are determined by the thermal conductance to a cold plate. The way this scales with wire size is interesting, as there is no characteristic size in Maxwell's equations, and the maximum magnetic field limited by heating turns out to be about constant, independent of size scale, while the force increases proportionately.

3.3 Sensing with Electrical Transduction

An important issue in the use of nanoprobes is the transduction of the signal, either to the outside world for analysis, or to a different internal region of the sensed system to create a response. As discussed above, optical sensing is used to return an external signal that provides information about position or chemical activation. Magnetic sensors may also be used to register the presence of a tagged object. Electronic sensing is desirable because of the potential for directly coupling the output to an integrated circuit for analysis, and possible control of feedback response.

One important goal is to detect biological molecules electrically. If we consider how biological chemicals can be most sensitively detected electrically, we can see that it is desirable to use a biological binding event to alter the conductivity of very small wires, providing an electrical response that can be transmitted outside the measured object. Alternatively, electrostatic interactions can be used to create reversible binding, opening the possibility of direct electrically- controlled feedback of binding.

3.3.1 Nanowire sensors

An innovative new concept for electrical bio-sensors based upon functionalized semiconductor nanowires was recently introduced by Lieber et. al. [20]. According to this scheme, a nanowire decorated with antibodies is contacted electrically with two electrodes as illustrated in Figure 6. When the antigen binds, it perturbs the local charge environment of the nanowire. It is interesting to note that this resembles closely the "ChemFET" idea that was explored extensively over the last fifteen years. However, the ChemFET requires a great many binding events to influence the transport through an entire region of semiconductor. In the nanowire case, even a single binding

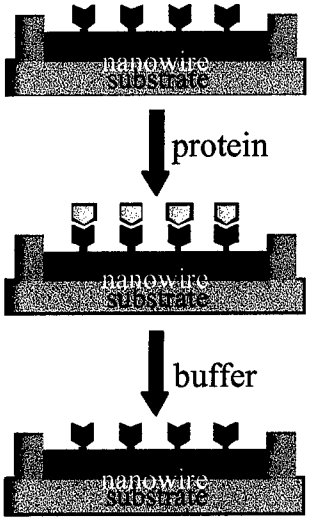


Figure 5: Schematic illustration of a nanowire decorated with antibodies and contacted electrically with two electrodes. When the antigen binds, it perturbs the local charge environment of the nanowire allowing electrical readout. A buffer wash removes the bound antigen for repeated sensing.

event may well be detectable, since all the charges must pass within a few nanometers of where the binding event occurs. Electrical detection removes the need for complex and large optical systems, and potentially enables highly integrated sensors that permit numerous controls and statistically significant sampling to occur in a very small package. To fully realize this goal, computation has a very important role to play. It is possible to calculate the electrical transmission of a nanowire in isolation, as well as of a nanowire in the presence of biological molecules that perturb it, for instance by changes in the local electric field. Such calculations can help to understand the limitations and strengths of this proposed new sensor system.

3.3.2 Nanotube sensors

Carbon nanotubes also offer great potential as biological sensors. There

are three things that distinguish carbon nanotubes for this application. The first is their small size. Single-walled nanotubes (SWNTs) are typically 1 – 3 nm in diameter, significantly smaller than Si nanowires. This transverse dimension is comparable to, or most cases smaller than, biological molecules of interest. The nanotube diameter is thus comparable to the width of the DNA double helix. This matching of the size scale of the sensor to the size scale of the molecule to be detected has many advantages, both in terms of sensitivity and binding as discussed further below.

The second desirable characteristic of nanotubes is their excellent electronic properties [21]. Experiments on semiconducting nanotubes reveal that the carrier mobility and device transconductance can significantly exceed those found in state-of the-art Si devices. This means that these devices will function very well as sensors; a small change in their electrostatic environment will lead to a large measurable current. Using either a micropipette or a microfluidic system these devices have also been shown to operate in water under biologically relevant conditions. Ions or other charge groups in the water dope the tube electrostatically and modify its conductance. Recent measurements indicate that single electronic charge sensitivity should be possible. In other words, if a charged molecule in the fluid changes the number of electrons residing on the tube by one, this should lead to a measurable change in the current. A key point is that the active electronic region of the nanotube is directly in the solution, with no intervening oxide layer, unlike Si nanowires or ChemFETs. This is critical since molecules more than the Debye screening length ($\sim$ 1 nm) away from the active region are heavily screened by counterions in the water and therefore difficult to detect.

The third desirable characteristic of nanotubes is the ability to bind individual molecules to the nanotube for detection. Two approaches are possible. The first is the same as outlined above for Si nanowires; functional groups are bound to the surface that bind the molecule of interest, resulting in a conductance change of the device. For example, the Dai group [22]

has immobilized proteins on the surface of a nanotube using a non-covalent attachment scheme. An alternative approach is to utilize electrostatic interactions between the molecule and the nanotube. The charge on the nanotube can be continuously adjusted by changing the voltage on the nanotube relative to the solution. It should be possible to tune this charge to selectively bind/unbind a large number of biomolecules (such as DNA) based on their charge per unit length. The proper matching of the size of the nanotube to the width of the DNA molecule is critical for this application.

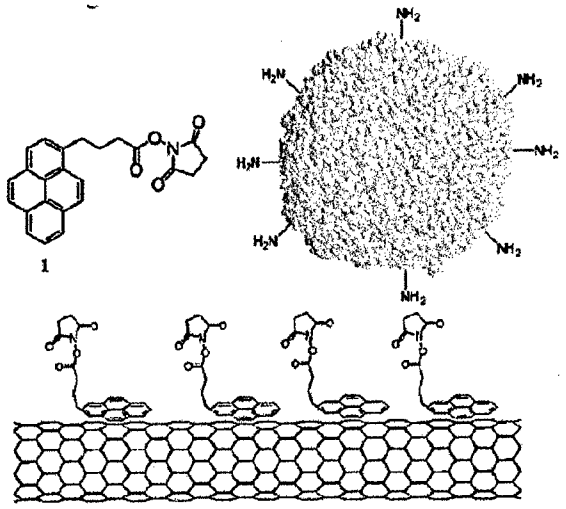


Figure 6: Schematic illustration of binding of a protein through amine group to functionalized carbon nanotube. From Chen, et al. J. Am. Chem Soc **23**. 3838 (2001).

To understand the operation of nanowires and nanotubes as sensors, our understanding of the electrostatic interactions — screening, binding, etc. — between macromolecules at nanometer length scales must be expanded. This is an exceedingly challenging problem due to the complex electrostatic environment (polar water molecules, counterions, etc.) and the other forces (entropic, chemical) that come into play. A great deal of theoretical work, both computational and analytical, is needed in this area. Developing models of macromolecular interactions that keep the essential complexity while

simplifying the overall problem is essential. These sensors may prove to be an important testing ground for these models, since a number of parameters, such as the charge per unit length on a nanotube, can be externally controlled with great precision. Such models would greatly improve our ability to both design new sensors and to understand their operation. Furthermore, studies of interactions of artificial macromolecules (nanotubes, nanowires) and natural macromolecules (DNA, proteins) may improve our understanding of the operation of molecular binding and molecular machinery in living systems.

3.4 Nanopores and Nanoporous Membranes

While much work in nanoscience is focused on developing nanostructures with physical response characteristics, it is also possible to design nanostructures that have a chemical response analogous to that of biological responses. Such chemical responses can be in the controlled catalysis of chemical reactions, or in mediation of chemical transport. Biological systems accomplish controlled transport via the use of channel proteins that contain very small apertures lined with chemical groups tailored to selectively pass different chemical species (see section 5.2). Energetically unfavorable transport, e.g. against a concentration gradient can even be accomplished via coupled reaction with a chemical energy source. Artificial systems based on nanoporous membranes in inorganic materials or fabricated nanopores are being developed to accomplish similar functionality.

3.4.1 Nanoporous membranes

Mesoporous materials, primarily silica and alumina-based materials with regular arrays of pores of typical diameter 2 to 10 nm, and surface areas of approximately 1000 m^2/g were first developed for catalytic applications. The

pores are formed via a chemical reaction under conditions where the reactant is concentrated at regular positions via formation of micelles. Continuing developments in the field have resulted in a variety of synthetic approaches and applications to a wide variety of inorganic materials. The mesoporous (or nanoporous) materials can be fabricated in shapes including films, spheres and fibers.

Selected chemical species can be introduced into the pores either by co-condensation during fabrication of the base material, or by specialized synthesis via linking chemistries that assemble molecular structures into the pores after the base material is fabricated [27]. Pore sizes and internal geometries can be controlled by the assembly of long-chain molecules in various combinations within the pore. The chemical properties of the end groups exposed in the pore can be designed for a variety of functionalities, including mixed functionality within the pore. The pores can then act to sequester a particular species from an external solution via irreversible binding, or a chemical reaction can be catalyzed within the pore with continuous exchange with the outside solution. In addition, functionalized mesoporous materials can be configured to control transport through the pores.

An example of the latter application is shown in Figure 7 [24]. A mesoporous membrane has been functionalized with an antibody that binds selectively to one optical isomer of a chemical being tested for pharmaceutical activity. A mixture of the two isomers in solution is placed on one side of the membrane with the goal of selective diffusion of only one of the enantiomers through the membrane. Ordinarily, antibody binding is strong and irreversible, a feature that would not lead to selective transport. The antibody strength therefore was tuned by the addition of a chemical that decreases the binding strength. The enantiomer that binds to the antibody then is preferentially concentrated into the pores, but with reversible binding, allowing it to be released stochastically across the membrane. Tuning such a process to optimize the separation, while also maintaining a flux sufficient

for practical applications is a continuing research challenge. We also suggest the possibility of incorporating such nanoporous membranes or particles into nano-assemblies designed for interactive functionality as an additional direction of investigation.

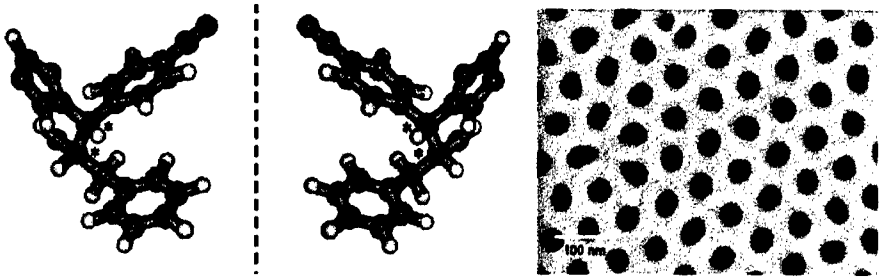


Figure 7: Enantiomeric separation of an organic molecule has been accomplished by transport through alumina membranes lined with silica, and then functionalized with an enzyme that selectively binds one enatiomer. Figure from Lee et al, Science 296, 2198 (2002).

3.4.2 Single nanopores

The ultimate level of selectivity in nanoporous membranes is demonstrated at the cell wall. Here control of chemical transport with chemical specificity and single molecule control is demonstrated by membrane bound proteins. As we will show below, the ability to isolate and use such functional membrace proteins demonstrates the potential of using biological nanocomponents in functional artificial nanosassemblies.

A significant example of the interface between biology and nanoscience is provided by elegant work on the translocation of individual molecules such as single-stranded DNA or RNA through a pore in a membrane [25]–[28]. Eventual applications might include molecular sorting and perhaps even rapid sequencing of entire genomes. Recently, efficient nanopore discrimination between single-stranded polynucleotide molecules has been demon-

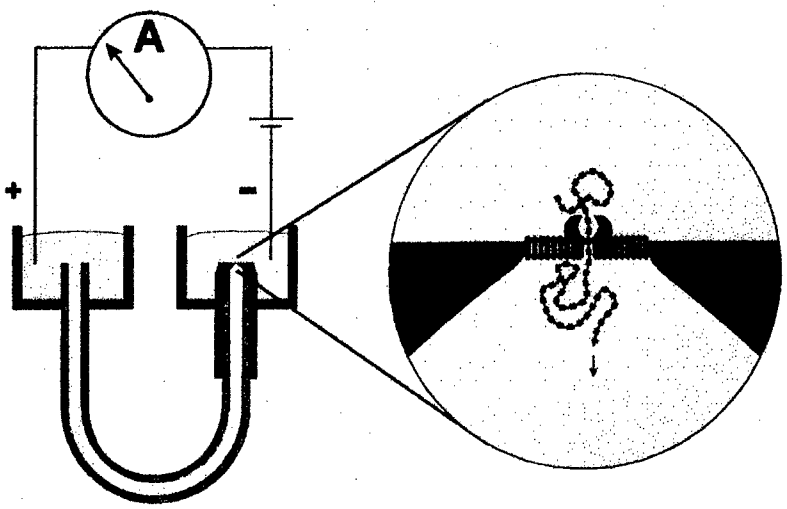


Figure 8: Translocation apparatus for single stranded DNA and RNA (from Ref. [28])

strated in this system.[29] The pore in this case is a naturally occurring biological protein, α-hemolysin, embedded in a flat lipid bilayer.

The protein α-hemolysin is a 33 kD protein with 293 amino acids secreted by the bacteria Staphylococcus areus. It assembles from seven water-soluble components to form a membrane bound pore in target cells such as red blood cells and leukocytes. The resulting water-filled channel triggers osmotic shock and cell lysis, presumably allowing the S. areus to feed on the liberated cell contents. The pore is approximately 1.5 nm in diameter (roughly the width of a *single* DNA strand), and 10 nm long.

The biological pores created by S. areus can be put to use in an artificial environment as indicated in Figure 8 [28]. A single pore is embedded in an artificial lipid (diphytanoyl phosphatidylcholine) bilayer which divides two chambers of an electrolysis cell. Single stranded DNA or RNA is introduced on the cis (negative) side of this device. With appropriate buffers and salt concentrations, these polynucleotides (typically 100's of bases long) ionize to become negatively charged and are drawn through the pore towards the

positively charged anode. Their passage is registered by a drop in the current of smaller ions passing through the core. The current is reduced on a time scale of 100's of microseconds. The voltage drop of order 120 mV is mostly across the hemolysin pore. Double-stranded DNA will *not* pass through such a narrow pore. The single stranded translocation velocities are of order 1 nucleotide every few microseconds. With this speed, if there were a way to deduce precise sequence information from fluctuations in the ionic current during the time the pore is nominally "closed", the entire human genome could be "read" in about 15–30 minutes! Even if such a breakthrough could only occur in the distant future, these pores already show promise as interesting sorting devices.

Figure 9 shows a typical time trace of the current through the nanopore for poly$(dA)_{100}$, i.e., single-stranded DNA composed of 100 adenine nucleotides [29]. The blockage event can be characterized by both a "duration time" t_D (of order 400 μs) and a dimensionless "blockade level" $I_B = < I_{\text{event}} > / < I_{open} >$, where the brackets represent time averaged currents

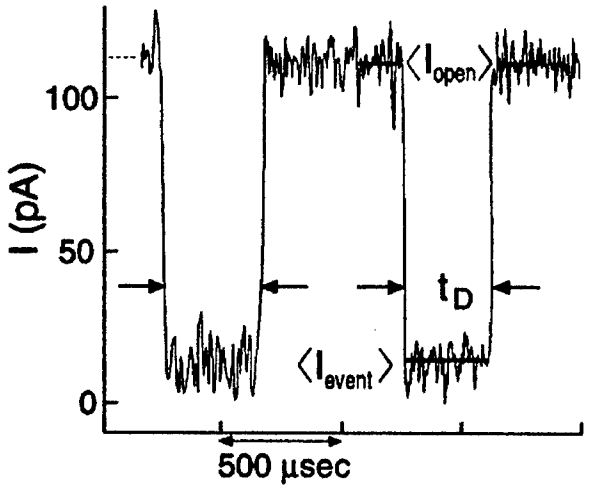


Figure 9: Current through a α-hemolysin nanopore as function of time (figure from ref. [29]).

during the intervals when the pore is closed (with current I_{event}) and open (with current I_{open}). Because the experiment is coupled via the solvent to a heat bath, the translocation is a stochastic process – for low translocation velocities, one might even expect the DNA to back up occasionally as it slides through the pore. For the regimes relevant to these experiments, one can model the translocation by a one dimensional diffusion equation for $P(x,t)$, the probability that the polymer has translocated a distance x after time t [32]. It is easy to show that, if L is the contour length of the polynucleotide sequence, then the mean blockage time for a single homopolymer can be written as

$$t_D = L/v_{\text{eff}}, \tag{3-1}$$

where v_{eff} is an effective translocation velocity which should be proportional to the voltage drop for small voltages. Thermal fluctuations in the environment lead to a dispersion in blockage times given by

$$\Delta t \approx (2D_{\text{eff}}L)^{1/2}/(v_{\text{eff}})^{3/2}, \tag{3-2}$$

where both v_{eff} and D_{eff} depend in a complicated way on the sequence and the degree of tilt. For large barriers and slow translocation velocities, both v_{eff} and D_{eff} are expected to have an approximately Arhenius temperature dependence.

Readily measurable differences in the average blockage duration t_D have been observed for polymers of different effective diameters. This difference appears to be sufficient to distinguish molecules in mixtures with approximately 98% reliability. If this sensitivity could be coupled with a downstream microfluidic switching device, it might be possible create a novel molecular sorting device. For example, an electric field could guide different molecules down the left or right branch of a Y-shaped tube, depending on the translocation time during prior transit through a hemolysin pore.

3.4.2.2 Construction of artificial pores

Inspired by the α-hemolysin example, nanotechnology has now been used to create a 5 nm artificial pore in a silicon-nitride membrane, which was used to record the threading of individual double-stranded DNA molecules [30, 31]. Li et al. [33] have constructed 16 nm artificial pores in silicon nitride (Si_3N_4) using argon ion-beam sputtering with an interesting feedback loop. A lithographically created pore in Si_3N_4 can be made *smaller* by an ion beam "sculpting" process. The process is initiated by lowering the ion intensity so that the incident argon atoms now merely heat the surroundings and facilitate diffusion, rather than etching away additional material. As shown in Figure 10, by monitoring the decrease in the argon current through the hole, one can stop the closing when the desired pore area is reached.

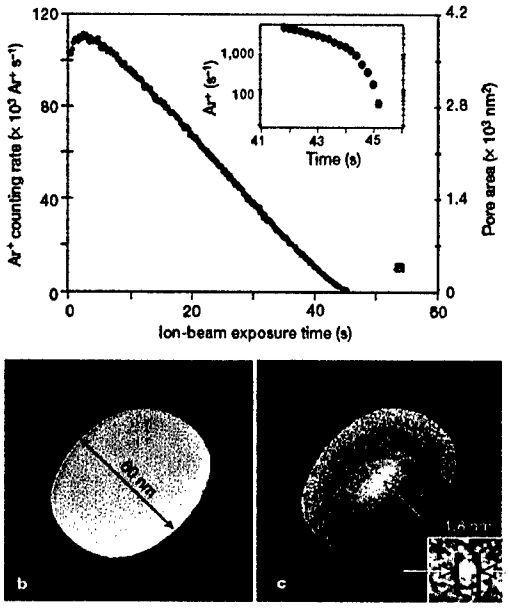


Figure 10: Closure of Si3N4 pore monitored via an Ar+ counter which acts as a feedback circuit allowing the process to be stopped before the hole closes completely (figure from ref. [30])

A 5 nm pore created is this way was used to record the translocation of *double*-stranded DNA [33]. Although still more than three times the size of an α hemolysin pore, a drop in the pore current due to translocation of *double*-stranded DNA through this artificial hole has been observed. Blockages reminiscent of those due to single-stranded DNA passing through α-hemolysin (as in Figure 9) were observed, with durations of order milliseconds and reductions in ion current by 85% or more. Selected Si_3N_4 nanopores are electrically quiet and give rise to blockage signals which meet or surpass the signal from α-hemolysin [34]. Challenges for the future include obtaining nanopores with more reliable characteristics and controlling the geometry of the pore itself.

3.5 Membranes

Biological systems organize structural assemblies using phospholipid membranes. Biological membranes both encapsulate regions of controlled chemical content, and support and organize specialized proteins with sensing and transport functions [4]. The membranes are composed of molecules with polar head groups and organic tails. They self assemble in aqueous solution into a double layer approximately 5 nm thick with a fatty interior and a hydrophilic exterior. Due to the fatty interior, these membranes are impermeable to ions and polar molecules. The membranes can be manipulated into a variety of shapes, including support across a small aperture as mentioned above in section 3.4.2.1, or the formation of spherical cavities (liposomes, shown in Figure 11), or supported on a substrate [1, 37] or on a supporter protein [2].

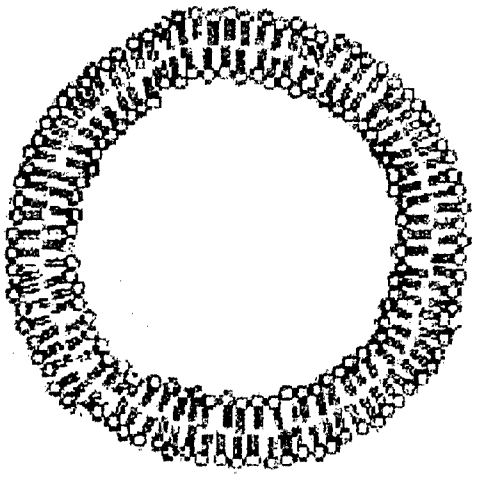


Figure 11: Schematic diagram of phospholipid double layer liposome.

Proteins supported on the membranes can be attached on one side, or can be inserted through the membrane. The latter include ion channels and chemoreceptors, whose functions are respectively transport of material, or information across the membrane. Transmembrane proteins generally include specialized anchoring groups in their structure which bind the protein into the membrane. However, diffusion in the plane of the membrane is facile, and serves an important function in allowing proteins to organize to optimize their function. This is illustrated for instance in Figure 12, which illustrates the evolution of protein conformations during binding between the receptor proteins on the surface of a T-cell (immunological response) and a supported membrane reconstituted from an antigen presenting cell. There are two types of binding pairs present on the two membranes, and these have been labeled with different fluorescent markers. During binding between the cell and the supported membrane, the structurally longer binding pair interacts first, followed by rearrangement of the membrane to allow attachment of the structurally shorter pairs. Subsequent rearrangement of the spatial organization of the proteins occurs to minimize the bending energy of the cell membrane.

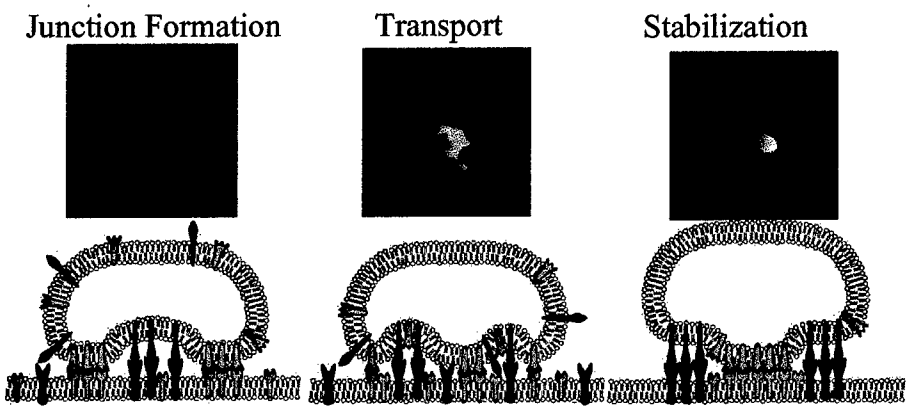


Figure 12: Fluorescent Microscopy imaging of binding of a T-cell to a supported antigen-displaying membrane. Spontaneous reorganization of the receptor complexes lowers bending energy of membrane following binding. Figure provided by J. Groves, University of California, Berkeley, from research discussed in Ref. [37]

.

Membrane proteins can, in some cases, be extracted from their native membranes into a detergent solution, and reinserted into a new membrane (which may be a biological or a synthetic membrane) with their function intact, as shown above for α-hemolysin (see also section 4.2) One interesting example of construction of such an assembly is a recent AFM study in which the membrane support protein apoA-I (a component of high-density lipoprotein) was used to create a template for controlled study of another membrane protein [2]. Thus it is quite possible to envision the creation of artificial assemblies of proteins, either on liposomes or supported membranes, for fundamental studies of their function, or for development of new system behavior. Moreover, it also seems quite feasible to develop synthetic approaches (see section 4.1) to functionalizing physical nanostructures for controlled anchoring in membranes. In this way the variety of functionalities available in both biological and physical nanostructures could be coupled in an exploration of the development of biologically inspired systems that may

be more robust than biological systems, or may incorporate non-biological functions.

3.6 SPM Imaging and Manipulation of Biological Systems

Scanning probe microscopy (SPM) has proven to be a very useful tool to image and to manipulate the properties of nanoscale systems. Using different contact mechanisms a wide variety of phenomena can be imaged including the mechanical, electrical, magnetic and chemical properties of nanoscale objects. In biology, the applications of SPM include both images and measurements of the mechanical properties and adhesion of DNA and other biological molecules. In mixed systems of biological and physical nano-components, SPM can also be applied to detecting magnetic and electrical signals.

New types of SPM cantilevers have been developed to sense different properties of systems under varying conditions with greater sensitivity. Some of the recent developments can be used for biological systems, and collaborations with biologists can produce more sensitive and versatile imagers and manipulators for biological applications. One recent development is the fabrication of small, fast SPM cantilevers. For SPM of biological systems, the cantilever must often operate in a wet environment, so that its motion is heavily damped, unlike the usual situation in air or in a vacuum. Under these conditions, small, high resonant frequency SPM cantilevers can show better sensitivity than conventional units. A typical SPM cantilever has length $\sim 100\mu$m and resonant frequency $\sim$ 100 kHz in vacuum. Nanotechnology can produce much smaller cantilevers with lengths $\sim 10\mu$m and resonant frequencies $>$ 10 MHz in vacuum, that can be read out externally [39] or by using integrated strain sensors [40].

The force sensitivity of a SPM cantilever oscillating in a fluid increases for smaller cantilevers, because thermal noise is associated with viscous drag. The minimum detectable force of a thermally limited measurement is [39]

$$F_{\min} = (4k_B TRB)^{1/2}$$

where R is the coefficient of viscous damping and B is the bandwidth. By fabricating small silicon nitride cantilevers with lengths as small as 9 μm, Viani et al.[39] improved the thermally limited force sensitivity by a factor ~ 5 over a conventional cantilever.

3.7 Nanocomponents Synopsis

In the past few years, there have been substantive advances in the demonstrated capabilities of physical nano-systems. Moreover, the drive for applications in biotechnology has promoted the demonstration of workable linkage schemes allowing physical nano-components to be suspended in solution, or linked directly to biological molecules. Physical nano-components as sensors of biological function are well demonstrated, however the use of physical nano-components as direct mediators of biological function remains an open challenge.

Biological nanocomponents can be extracted from their natural environment and re-assembled for functional applications. The use of a biological membrane is particularly interesting as a potential substrate for supporting assemblies of nano-components including membrane proteins. However, linking chemistries to anchor physical nanocomponents in membrances must be established to make this a reality. Overall, the status of experimental research at the interface of nanoscience and molecular biology is sufficient to support a research program focused on exploration of new functionalities and assemblies.

References

[1] J. B. Pawley, ed., Handbook on Biological Confocal Microscopy, 2nd Edition (New York, Plenum Press, 1995).

[2] Ph. Tamarat, A. Maali, B. Lounis, and M. Orrit, *Ten Years of Single-Molecule Spectroscopy*, J. Phys. Chem **A 104**, 1, 2000.

[3] W. E. Moerner, M. Orrit, *Illuminating Single Molecules in Condensed Matter*, Science **283**, 1670 (1999).

[4] S. Weiss, *Fluorescence Spectroscopy of Single Biomolecules*, Science **283**, 1676 (1999).

[5] A. M. Kelley, X. Michalet, and S. Weiss, *Single-Molecule Spectroscopy Comes of Age*, Science **292**, 1671 (2000).

[6] T. A. Byassee, W.C.W. Chan and S. Nie, *Probing single molecules in single living cells*, Anal. Chem. **72**, 5606 (2000).

[7] N. Bobroff, *Position measurement with resolution and noise-limited instrument*, Rev. Sci. Instrument. **57**, 1152 (1986).

[8] A. M. van Oijen, J. Kohler, J. Schmidt, M. Muller and G.J. Brakenhoff, *Far-field fluorescence microscopy beyond the diffraction limit*, J. Opt. Soc. Am. A 16, 909 (1999).

[9] These centroiding techniques are used routinely in optical tweezer experiments. See for example K. Visscher, S.P. Gross, and S.M. Block, *Construction of multiple-beam optical traps with nanometer-resolution position sensing*, IEEE Journal of Selected Topics in Quantum Electronics **2**, 1066 (1996).

[10] X. Michalet, T.D. Lacoste, and S. Weiss, *Ultrahigh-resolution colocalizatoin of spectrally separable point-like fluorescent probes*, Methods **25**, 87-102 (2001).

[11] M. Bruchez Jr., M. Moronne, P. Gin, S. Weiss, and A.P. Alivisatos, *Semiconductor Nanocrystals as Fluorescent Biological Labels*, Science **281**, 2013 (1998).

[12] M. Dahan, T. Laurence, F. Pinaud, D.S. Chemla, A.P. Alivisatos, M. Sauer, and S. Weiss, *Time-gated biological imaging by use of colloidal quantum dots*, Optics Letters, **26(11)**, 825 (2001).

[13] D. Gerion, F. Pinaud, S.C. Willimas, W.J. Parak, D. Zanchet, S. Weiss, and A.P. Alivisatos, *Synthesis and proberties of biocompatible water-soluble silica-coated CdSe/ZnS semiconductor quantum dots*, J. Phys. Chem **B 105**, 8861-8871 (2001).

[14] W. C W. Chan, S. M. Nie, *Quantum dot bioconjugates for ultrasensitive nonisotopic detection*, Science **281**, 2016-2018 (1998).

[15] W. C. W. Chan, D. J. Maxwell, X. Gao, R. E. Bailey, N, Han, and S. Nie, *Luminescent quantum dots for multiplexed biological detection and imaging*, Current Opinion in Biotechnology **13**, 40–46 (2002).

[16] H. A. Clark et al., "Subcellular optochemical: probes encapsulated by biologically localised embedding (PEBBLEs). Sensors and Actuators" **B51**, 12 (1998).

[17] J. L. West and N.J. Halas, "Applications of nanotechnology to biotechnology," Current Opinion in Biotechnology **11**, 215 (2000).

[18] S. R. Shershen, S.L. Westecott, N.J. Halas, and J.L. West, "Temperature sensitive polymer-nanoshell composites for phothermally modulated drug delivery," J. Biomedical Materials Research **51**, 293 (2000).

[19] C. S. Lee, H. Lee and R.M. Westervelt, "Micro-electromagnets for the control of magnetic nanoparticles", Appl. Phys. Lett., **79**, 3308 (2001).

[20] Y. Cui, Q. Q. Wei, H. K. Park and C. M. Lieber (2001). "Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species." Science 293 (5533): 1289-1292.

[21] P. McEuen, M. Fuhrer, and H. Park, *Single-Walled Carbon Nanotube Electronics*, IEEE Transactions on Nanotechnology **1**, 78 (2002).

[22] R. J. Chen et al., *Noncovalent sidewall functionalization of single-walled carbon nanotubes for protein immobilization*, Journal of the American Chemical Society **123**, 3838 (2001).

[23] J. Liu, Y. Shihn,Z. Nie, J.H. Chang,L-Q. Wang, G. E. Fryxell, W. D. Samuels, and G. J. Exarhos, Molecular Assembly in Ordered Mesoporosity: A new Class of Highly Functional Nanoscale Materials, J. Phys. Chem. 104, 8328 (2000).

[24] S. B. Lee, D.T. Mitchell, L. Trofin, T. K. Nevvanen, H. Soderlund, C.R. Martin, Antibody-Based Bio-Nanotube Membranes for Enantiomeric Drug Separations, Science 296, 2198 (2002).

[25] S. M. Bezrukov, Vodyanoy, I. Parsegian, V. A. Counting polymers moving through a single ion channel. *Nature* **370**, 279-281 (1994).

[26] J. Kasianowicz, Brandin, E., Branton, D. Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl Acad. Sci.* USA **93**, 13770-13773 (1996).

[27] L. Q. Gu, Braha, O., Conlan, S., Cheley, S. Bayley, H. Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adaptor. *Nature* **398**, 686-690 (1999).

[28] M. Akeson, Branton, D., Kasianowicz, J. J., Brandin, E. Deamer, D. W. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. Biophys. J. **77**, 3227-3233 (1999).

[29] A. Meller, Nivon, L., Brandin, E. , Golovchenko, J., and Branton, D., Rapid nanopore discrimination between single polynucleotide molecules, PNAS **97**, 1097-1084 (2000).

[30] Li Jiali, Derek Stein, Ciaran McMullan, Daniel Branton, Michael J. Aziz, Jene A. Golovchenko, Ion-beam sculpting at nanometre length scales, Nature 412, 166 - 169 (2001).

[31] J. Tersoff, Nanotechnology: Less is more, Nature 412, 135 - 136 (2001).

[32] D. K. Lubensky and D. R. Nelson, Driven polymer translocation through a narrow pore, Biophysical Journal **77**, 1824-1838 (1999).

[33] J. Li, D. Stein, C. McMullan, D. Branton, M. J. Aziz and J. A. Golovchenko,"Ion-beam sculpting at nanometre length scales", Nature **412**, 166 (2001).

[34] D. Branton, private communication.

[35] H. Lodish, et al., "Molecular Biology, Scientific American Books, New York, NY 1995.

[36] P. S. Cremer, J. T. Groves, L. A. Kung, and S. G. Boxer, Writing and Erasing Barriers to Lateral Mobility into Fluid Phospholipid Bilayers, Langmuir 15, 3893, 1999.

[37] S. Y. Qi, J. T. Groves, and A. K. Chakraborty, Synaptic pattern formation during cellular recognition, Proceedings of the National Academy of Science 98, 6548 (2001).

[38] T. H. Bayburt and S.G. Sligar, Single-molecule height measurements on microsomal cytochrome P450 in nanometer-scale phospholipd bilayer disks, Proceedings of the National Academy of Science 99, 6725-30 (2002).

[39] M. B. Viani, T. E. Schaffer, G. T. Paloczi, I. Pietrasanta, B. L. Smith, J. B. Thompson, M. Richter, M. Rief, H. E. Gaub, K. W. Plaxco, A. N. Cleland, H. G. Hansma, P. K. Hansma, Fast imaging and fast force spectroscopy of single biopolymers with a new atomic force microscope designed for small cantilevers. Review of Scientific Instruments, vol.70, p.4300-3.

[40] R.G. Beck, M.A. Eriksson, M.A. Topinka, R.M. Westervelt, K.D. Maranowski and A.C. Gossard, "GaAs/AlGaAs Self-Sensing Cantilever for Cryogenic Scanning Probe Microscopy", Appl. Phys. Lett. **73**, 1149 (1998).

4 ASSEMBLY CHALLENGE

Based on the results of the previous section, we conclude that the assembly of functional nano-components is a viable research challenge. However, there must be a significant scientific motivation to drive such a challenge. As noted in the previous section, the development of cellular diagnostics is one motivation that will continue to support research at the interface of nanoscience and biology. However, we also believe that the development of functional nano-assemblies can be used in support of an important basic research goal, that of developing an understanding of the feedback and regulatory mechanisms of the cell. To guide the development of research in this area, we have defined a research challenge requiring a close interaction between experiment and modeling:

> Fabricate non-trivial assemblies of physical and biological nano-components with linked functionality, and develop carefully designed experiments to directly compare measured behavior to results of systems modeling.

Some experimental issues impacting this challenges are illustrated in Figure 13, and discussed in Sections 4.1 and 4.2 below. The theoretical context of this challenge, which is developing controlled systems to quantify analysis of cellular function is described in Section 4.3.

4.1 Chemical Linkages for Nanoassembly

Linking Chemistries for Nano-Bio

A major challenge faced in combining biologically-derived nanostructures (e.g., proteins; lipids, nucleic acids, and carbohydrates) with any physically–

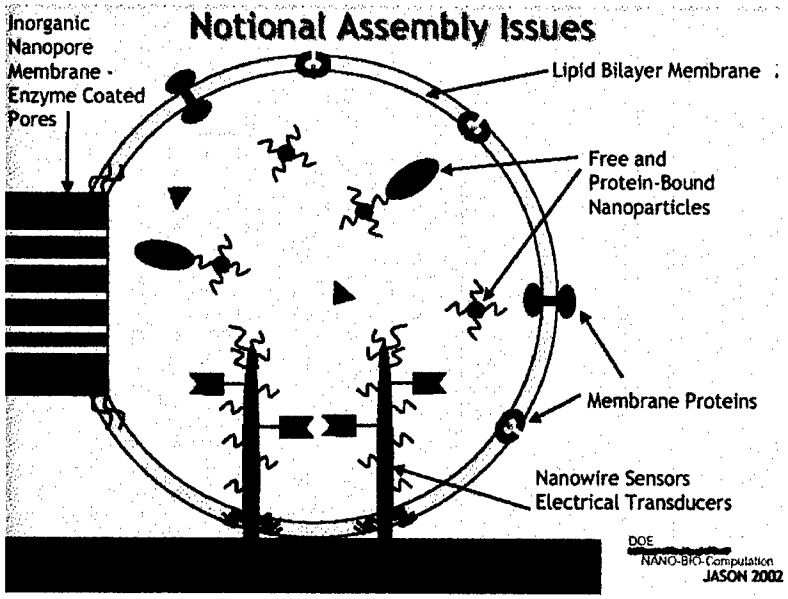


Figure 13: Notional illustration of a nanoassembly, illustrating some of the research issues that must be addressed in design and fabrication. These include: develop linking mechanisms between a broader range of physical and biological nanocomponents, including anchoring linkages for attaching physical nanocomponents to membranes; extract, purify and reassemble biological components in desired configurations; establish effective feedback mechanisms for the interplay of biological and physical nanocomponents; and establish mechanisms for information transfer between the nano-assembly and the outside world.

or chemically-nanofabricated components (e.g., carbon nanotubes, nanolithographed parts, fluorophores, etc.) is to identify appropriate *linking chemistries*. For many of the applications envisioned, and perhaps most, it will be necessary for chemical connections to remain *bio-compatible*, to avoid denaturing or degrading labile biological structures. Broadly speaking, these constraints include the following:

- An aqueous environment (H_2O as the primary solvent)
- Temperature range between 0° and 100°C, optimum generally near 37°C

- pH inside the range of 5–9, optimum generally near pH 7.0
- Approximately normal saline ($\sim$ 150 mOsm)
- Atmospheric pressure, optimum around 1 bar $=10^5$ pascal.

Some deviation from these criteria may prove possible by careful engineering or selection of the biological material, for example through genetic engineering of proteins, *in vitro* evolution/optimization, or the use of extremophilic organisms as source material. And specific exceptions may occur for some small, but useful biomolecules, or isolated ultra-stable biomolecules, such as DNA. However, the room for maneuver is rather limited. This narrow "phase space" of operation rules out a good many synthetic routes currently used in organic chemistry, as well as many common physical techniques (CVD, epitaxial growth, etc.) currently employed in nanofabrication.

Useful linking chemistries in nano-bio must fulfill further criteria as well. Ideally, any connections made would be sufficiently *strong, selective and stereospecific.* Other desiderata include versatility (ability to be used in a variety of contexts), controllability (ability to control the numbers of links and fan-out), and robustness.

The palette of existing linking chemistries for biomolecules, while extraordinarily useful, is really quite limited—due, in part, to many of the very same considerations just outlined. These chemistries fall broadly into the following categories:

Noncovalent attachments:

A whole variety of naturally occurring receptor-ligand pairs have been harnessed to link biological molecules. In general, these pairs consist of a protein-based receptor and a smaller organic or inorganic ligand or peptide moiety that binds tightly to the receptor. The most widely-used of these is the avidin-biotin linkage (Figure 14, topmost illustration) and its close rela-

tives (streptavidin, avidin DM, etc.). Avidin is a protein tetramer, each subunit of which binds a small organic molecule: the water-soluble B-complex vitamin, biotin. Avidin binds so tightly to biotin that the dissociation constant of the complex, K_D, is about 10^{-15} M, which approaches the avidity of a covalent bond (hence the name). Other major noncovalent pairs in widespread

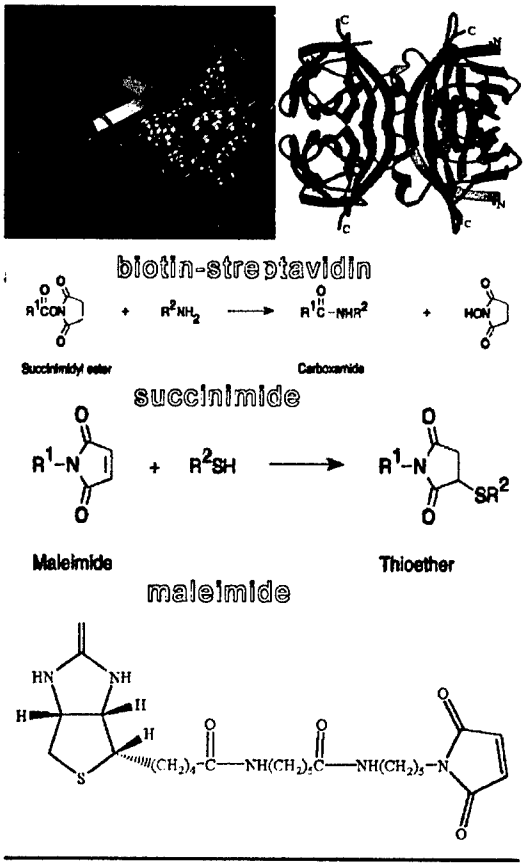


Figure 14: Examples of common biocompatible linking chemistries employed in biotechnology, as indicated. For details, see text.

use in biotechnology include: (1) glutathione-S-transferase (GST), a dimer that binds tightly to its natural substrate, glutathione; (2) nickel-histidine, where a poly-histidine sequence (generally 6 His or more) engineered into a protein or peptide will chelate and bind to nickel ion (held, for example,

in organically coupled Ni-NTA, nitrilotriacetic acid); (3) antibody-epitope linkages, where an immunoglobulin (usually, IgG, but also IgM, IgE, or an FaB fragment) is used to recognize a specific tag, which can be a peptide sequence (examples include c-myc or FLAG or a poly-His) or a convenient small molecule (examples include digoxygenin and dinitrophenol); and (4) Chitin-CBP, where an engineered version of chitin binding protein (CPB) binds tightly to its substrate, chitin (poly-NAG). Other examples of ligand-receptor pairs with somewhat lesser utility include lectin- carbohydrate linkages (e.g., concanavalin A) and protein A-IgG (protein A from S. *aureus* and binds to certain immunoglobulins).

In addition to protein-based receptor-ligand pairs, a great deal of progress has been made in engineering the sequences of comparatively short DNA-or RNA-based sequences, known as *aptamers*, which can serve to bind specific molecules of choice (peptides, small organic molecules, epitopes or moieties) with good specificity and acceptable avidity.

With the exception of biotin-avidin, virtually all receptor-ligand pairs demonstrate comparatively weak binding, and are therefore less well suited for the fabrication of large, multisubunit complexes — or any truly long-lived constructs. Their chief advantage is their reversibility, which facilitates the conditional assembly of intermediates.

Covalent attachments:

Covalent bonds are best for 'permanent' attachments. Covalent bio attachments may be accomplished in biomolecules through *thiol* linkages, for example, via the sulfur-containing amino acids, or through *primary amines* (NH_2-), found on protein termini and certain sidechains. Gold-thiol esters can be used to hook nanogold particles to sulfur-containing sidegroups (for example, using S-biotin primer for DNA). Or, an S-containing cysteine residue, engineered by mutation into a protein subdomain of interest (or occurring naturally), can be linked to an organic molecule using a suitably

derivitized maleimide reagent (Figure 14, where R^1 represents the small organic molecule and R^2 the protein).

Primary amines can be reacted using the related chemistry of a succinimide reagent (Figure 14, where R^{-1} again represents the small organic molecule and R^2 the protein N-terminus or an amino acid sidechain). Other covalent linkages include isothiocyanates (ITCs), and cross-linkers such as glutaraldehyde.

Specialized attachments ("other"):

Recent years have seen the development of additional useful but specialized linking biochemistries, mostly covalent. Many are based on photochemistry, which offers the advantage of speed plus spatial precision at the micro, but not nano, scale. Examples here include a variety of photochemical species, including *photoactivatable* versions of various crosslinkers, succinimidyl, and maleimide reagents. Of related interest are so-called "*caged compounds*", which carry nitrobenzyl or similar leaving groups, and therefore release a biochemically-active molecule when struck by the appropriate wavelength of light. Examples include caged ATP, caged cAMP/cGMP, caged protons, caged neurotransmitters, etc. Finally, self-cleaving and splicing protein reactions, such as those found in the *inteins*, have been harnessed to produce a whole variety of self-excising peptides. Self-cleavage is also a property of certain ribozymes (enzymatically active RNAs), such as the Group I intron from *Tetrahymena*, and engineered ribozymes offer additional opportunities for biofunctional design.

All in all, the 'vocabulary' of available linking chemistries remains quite limited, and this may be a consequence of the many constraints outlined earlier. Most of the available noncovalent technologies have one or more serious drawbacks. They are often non-selective. As discussed, with the exception of avidin/biotin, most noncovalent linkages are also fairly weak. Avidin itself is a tetramer, GST and IgG are dimers, and these multivalent properties

often lead to serious cross-linking difficulties in practice. A good deal of genetic engineering is often required to take advantage of the desired chemistry. This may include preparing and expressing fusion proteins, carrying any of a variety of peptide sequences (myc, or FLAG tags), poly-his (for nickel-his), or biotin carboxyl carrier protein(for biotin), etc. It is also possible to use nucleic acid engineering and selection to create suitable DNA or RNA aptamers.

Covalent chemistry also leaves much to be desired. Thiol chemistry is often oxygen-sensitive. And once again, non-trivial amounts of genetic engineering are often required, such as preparing a Cys-free or "Cys-light" version of a protein so that a single cysteine residue can subsequently be incorporated by mutation at the point of interest. Reagents that react with primary amines often attack much more than the desired target. There are storage issues (ITCs degrade) and often the formation of toxic byproducts.

In considering how to build hybrid nanostructures incorporating both "nano" and "bio" components, it will be critical to take full advantage of existing linking chemistries. But to make real progress, it may well become necessary to go beyond the current, limited repertoire. We have, for example, no convenient ways of attaching each of the 20 specific amino acid sidechains (except for those bearing sulfur or amines). A Grand Challenge for nanobio will therefore be the development of a new suite of linking chemistries better suited to the task.

4.2 Biological Nano-components

Many components of biological systems can be extracted from their natural environment in the cell and inserted into artificial environments without loss of functionality. Membrane proteins are of special interest for research in developing bio-nano-technology, because of the wide range of functional be-

haviors they display and because support on a membrane provides a natural mechanism for organizing systems of nano-components [1, 2]. A simplified environment can be created in this way for research into the functional behavior of combinations of membrane proteins under controlled perturbation. In addition, one can envision incorporating artificial nanostructures, to introduce fields, information transduction or other functionality, by supporting them in or through the membrane, as suggested in Figure 13. Developing appropriate linking chemistries for inserting physical nanocomponents into membrane band structures will be a challenging component of research in this area.

4.2.1 Artificial photosynthetic system

A recent example of an artificially assembled biomimetic assembly provides an example of developing simplified analogs of biological systems [3] in which non-biological components may be mixed with biological components. In an adaptation of the classic experiments that demonstrated the role of the proton pump in ATP synthesis [4], Gust and co-workers have assembled a synthetic photosynthesis system as shown in Figure 15. Their system incorporates a bio-mimetically designed membrane-bound organic compound for photon harvesting along with the membrane protein ATP-synthase.

Biological photosynthetic energy conversion [4] involves the transformation of energy carried by photons to energy stored chemically in the molecule adenosine triphosphate (ATP). Energy storage (and retrieval) involves the reaction

$$ADP + P_i \quad \rightarrow \quad ATP + H_2O$$

where ADP is adenosine diphosphate and P_i is phosphate. The reaction is endothermic by 7.3 kcal/mole (0.3 eV/molecule). The photosynthetic process involves three steps: first trapping light in a molecular excitation, then

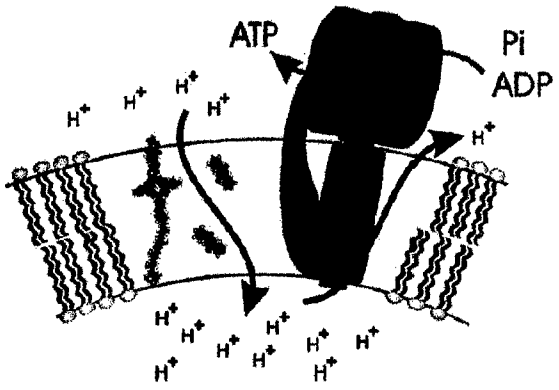


Figure 15: Schematic representation of a lipid (orange) supported light-driven proton pump (green) and an ATP-synthase protein (blue, brown and green). From Gust et al., [3].

charge separation which results in transport of H^+ across a membrane creating a proton gradient, and finally the use of the proton gradient to drive a membrane-bound enzyme protein, ATPase, which catalyzes the formation of ATP. The biochemical light-harvesting molecular complex is of variable complexity depending on the biological system. Even the least complex, in purple bacteria, is a multi-component system of interacting parts designed to meet the biological needs of adaptability, reliability and compatibility with other cellular functions [5]. Both the light-harvesting complex and the enzyme protein are embedded in a membrane. When the light harvesting complex has created an excess of proton concentration on one side of the membrane, proton binding to the ATPase causes a complex molecular response [6]. The protein bends and changes shape as four protons sequentially bind and are moved across the protein. The corresponding changes in shape of the protein result in binding of ADP and catalysis of the reaction between ADP and phosphate to form ATP.

Gust and coworkers [3] created a molecular unit, shown in Figure 16, to recreate the proton pumping action of biological complexes without the complexity required for proton pumps in living cells. The molecule contains a

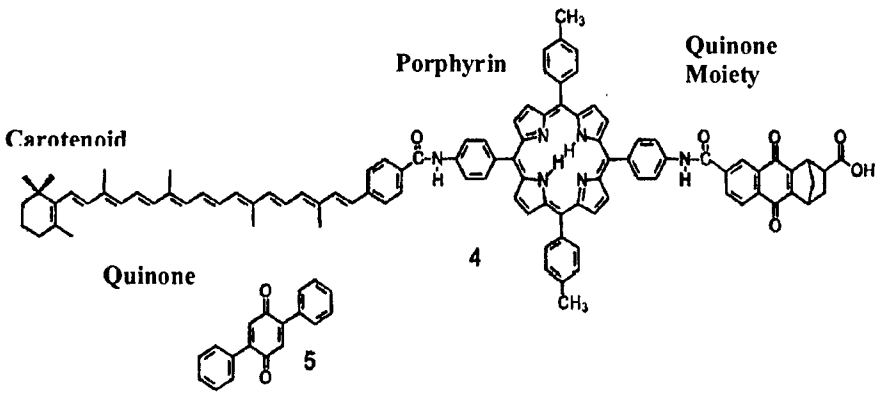


Figure 16: Photo-harvesting molecular unit consists (from left to right) of a carotenoid secondary electron donor, coupled to a prophyrin unit where photon excitation occurs (the chromophore), and a quinone unit that acts as an electron acceptor. The free quinone molecule diffuses through the lipid membrane, diffusively accomplishing proton transport. Figure from Gust, et al. ref [3]

central porphyrin unit that absorbs 1.9 eV photons, creating a singlet excited state. A rapid charge transfer moves a hole to the carotenoid (left) end of the molecule and an electron to the quinone (right) end. The resulting charge separated state has a lifetime on the order of a microsecond.

This molecule, along with quinone molecules, was partitioned from an organic solvent into the membrance of an artificial liposome vesicle. This created preferential alignment of the molecule, with the quinone (Q) end facing outward. Exposure of the liposome vesicles to 1.9 eV photons results in light absorption and charge transfer to the ends of the membrane-spanning molecule. At the outer side, the quinone moiety transfers its excess electron to one of the quinone molecules that diffuse freely in the membrane. The Q^- then accepts a protein from the external aqueous environment, and continues to diffuse randomly through the membrane as a QH radical. When it moves into the vicinity of the positively charged carotenoid end of the

molecular unit, an electron is transferred to the carotenoid end, neutralizing the photon-harvesting molecular unit. The resulting QH^+ immediately releases a proton into the interior of the liposome. Thus proton pumping with a design efficiency of 1 proton/photon can be achieved. Experimental tests of the system demonstrated its successful action in creating a pH differential of 2 across the membrane.

To build a photosynthetic system based on this photon-harvesting molecule, ATP synthase was extracted intact from its biological host into an aqueous detergent and introduced into the medium containing the liposomes. Controlled withdrawal of the detergent resulted in insertion of the ATPase with the catalytic end on the outside of the membrane, as shown in Figure 15. Subsequent, exposure to light when ADP and phosphate are present results in ATP synthesis. A saturation rate of about 100 ATP produced per second per ATPase was observed at an efficiency of 1 ATP per photon. This is about 1/4 of the design efficiency. Losses occur about equally in photon absorption and in the diffusive proton transport process.

The process described above illustrates the possibility of creating simplified versions of biological systems. Without the need for complex regulation of competing signals, or adaptation to changes in environment, simplified nanosystems of focused functionality can be created. Lipid membranes represent the biological equivalent of a circuit board — but one on which signal transfer via physical motion of the components can occur. The biological world offers a broad spectrum of functionality in membrane proteins that can be used as components. These include active and passive chemical transport, mechanical functions, chemical transduction and reaction catalysis. In addition, chemical and solid-state nanostructures can be introduced either inside the fluid or bound into the membrane. These will increase the range of functionality to include electrical, magnetic and optical tagging, information transfer or modification of the working environment of any co-existing membrane proteins.

The design, fabrication and testing of such nano-systems will create opportunities for developing and testing computational tools. It will be necessary to have a strong link with research in modeling to design an appropriate experimental system to challenge predictions of chosen models. Achieving a system with even a relatively straightforward set of linked reaction rates, such as the set described below in section 4.3.1 would be a significant accomplishment experimentally. Moreover, the possibility of isolating the system of interest from other cellular activities, and measuring responses to a variety of stimuli would significantly improve the quantifiation possible in rate equation modeling (see section 4.3.1). Challenges for molecular scale computation (see Section 5) will be generated when model nano-systems are used to generate controlled configurations where protein-protein, protein-membrane and protein-physical nanostructure interactions can be characterized and coordinated with assays of functional response.

4.3 Systems Modeling

The possibility of creating simplified assemblies of biological components, or possibly mixed biological and physical components, may create opportunities for understanding the feedback and regulatory systems of cells.

Thus, this section of our report is devoted to modeling efforts associated with cellular dynamics. Here we discuss some aspects of computational cellular biology, often refered to as in *silico* cell biology (to contrast with in *vivo* and *in vitro*). Not all biologists are in agreement with this direction of studying cellular dynamics. In a recent issue of *Nature*, Diane Gershon [7] quotes John Carson of the University of Connecticut Health Center "many

biologists feel that biology is just too complicated to be dealt with computationally."

Cell biologists are indeed dealing with an amazing *complex system.* From a modeling perspective the cell is **"a system with many degrees of freedom which we do not know how to treat in intricate detail and we do not know yet how to coarse grain to make the problem simpler."** [8] A pragmatic view is that the dynamics within a cell, from genetic control of protein synthesis to protein networks coordinating response to external environments, presents rich dynamical systems with many multiple levels of complexity which need to be approached from different directions with different tools.

Some of us have previously discussed in detail the issues of cell modeling in a previous JASON report[9]. In the following we will present a brief discussion of kinetics modeling, where the possibility of creating controlled experiments in which variables can be limited and controlled could yield significant improvements in understanding. We also add a description of the long term goal of creating an engineering model of cell regulation which ultimately might be applied to the development of artificial biomimatic systems of nano components.

4.3.1 Rate Equation Modeling

Here we give an example of parameter estimation and model development in cellular biology using the modeling approach of incorporating a full description of the kinetics of every chemical reaction involved in the system of interest. While trying to understand full cellular function using such an approach is unlikely to be productive, understanding the behavior of cellular subsystems at this level is an important input to higher level modeling efforts. A description of appropriate fitting techniques for accomplishing the

fit of coupled systems of rate equations is given in Appendix A. A review of publicly available cell modeling routines is presented in Appendix B.

The example presented is the attempt to fit [10] experimental data on the kinetics of the Janus Kinase Signal Transduction and Activation Transduction (JAK-STAT) signaling pathway. This pathway, illustrated in Figure17, is initiated by the binding of erythropoitin (Epo) at an extra

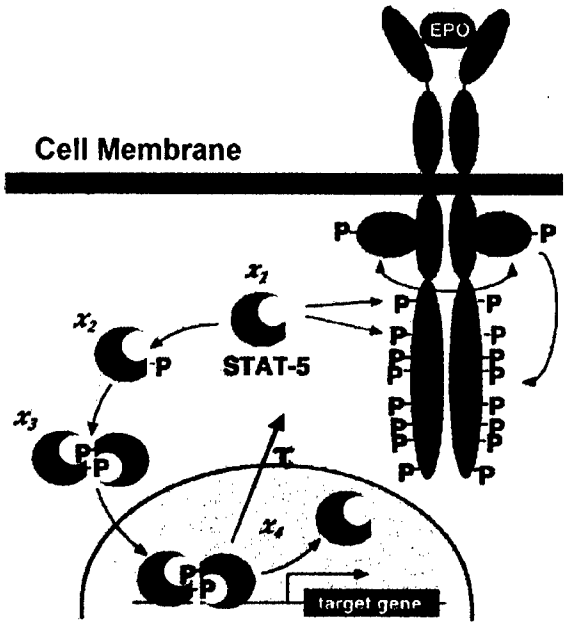


Figure 17: Schematic illustration of the reaction pathway coupling external excitation of the transmembrane receptor by the exciting species EPO, to the desired functional response, which is activation of the target gene to create a protein which helps protect the cell from virus infection. Binding of EPO on the periplasmic side of the membrane causes phosphate release to the signaling molecule, STAT-5, on the cytoplasmic side. Two kinetic models considered to describe the data differed by the inclusion (red arrow) of reappearance of STAT-5 in the cytoplasm with time constant τ. Figure from Timmer et al., [10].

cellular receptor leading to phosphorylation of monomeric STAT-5, a member of the signal transduction and activator of transcription family of transcrip-

tion factors. The phosphorylated STAT-5 forms dimers and those dimers migrate into the nucleus where they bind to promotor regions of DNA. The original model had STAT-5 end its role by dedimerization in the nucleus, but the modeling strongly suggested that STAT-5 reappears in the cytoplasm after a time delay during which some activity takes place in the nucleus which is not observable. Denoting the concentration of monomeric, unphosphorylated STAT-5 as $x_1(t)$, phosphorylated monomeric STAT-5 as $x_2(t)$, phosphorylated dimeric STAT-5 in the cytoplasm by $x_3(t)$, and phosphorylated dimeric STAT-5 in the nucleus by $x_4(t)$, the original model was

$$\begin{aligned}
\frac{dx_1(t)}{dt} &= k_1 x_1(t) f_{\text{EPO}}(t) \\
\frac{dx_2(t)}{dt} &= k_1 x_1(t) f_{\text{EPO}}(t) - k_2 x_2^2(t) \\
\frac{dx_3(t)}{dt} &= -k_3 x_3(t) + \frac{1}{2} k_2 x_2^2(t) \\
\frac{dx_4(t)}{dt} &= k_3 x_3(t)
\end{aligned}$$

and this was fit to experiment using the "cost function" (see appendix A):

$$J(x_1(t=0), k) = \sum_{1=l}^{N} \sum_{j=1}^{2} \frac{|y_j^{\text{Data}}(t_i) - y_j^{\text{Model}}(t_i; x_1(t=0), k)|^2}{\sigma_{ij}^2}$$

From the minimization of this using the observed data on $x_1(t)$ and $x_2(t)$, values for the reaction constants k were determined as was a value for $x_1(t = 0)$. In this formulation f_{EPO} is the time course of the experimental application of EPO to the cell membrane.

The model was found to give a very poor fit to the measured evolution of the concentration of STAT-5 in the cytoplasm. The model then was altered to allow for STAT-5 reappearing in the cytoplasm after a time delay during which it was active, in an unspecified way, in the nucleus. This changed the dynamical equations to

$$\frac{dx_1(t)}{dt} = k_1 x_1(t) f_{\text{EPO}}(t) + 2k_4 x_4(t - \tau)$$

$$\begin{aligned}
\frac{dx_2(t)}{dt} &= k_1 x_1(t) f_{\mathrm{EPO}}(t) - k_2 x_2^2(t) \\
\frac{dx_3(t)}{dt} &= -k_3 x_3(t) + \frac{1}{2} k_2 x_2^2(t) \\
\frac{dx_4(t)}{dt} &= k_3 x_3(t) - k_4 x_4(t-\tau)
\end{aligned}$$

introducing a dwell time in the nucleus and another reaction rate parameter k_4. This model was dramatically successful in reproducing the observables, as shown in Figure 18, and suggested a nuclear dwell time of about 6 minutes.

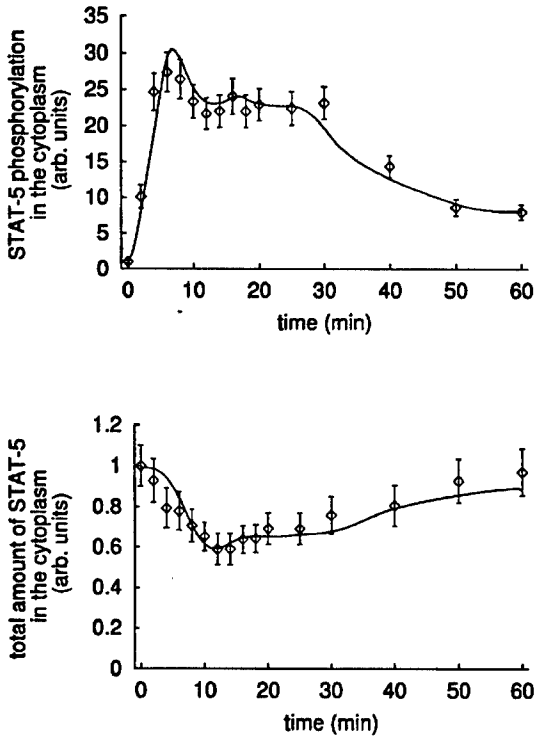


Figure 18: Fit of the kinetic rate equations including a time delay for return of STAT-5 to the cytoplasm, to the measured data for the concentration of STAT-5 and phosphorylated STAT-5 in the cytoplasm. A time constant of 6 minutes was derived from the best fit. Figure from Timmer et. al, ref. [10].

The veracity of such fits is strictly dependent on the choice of the physical model for the kinetic pathway. A good fit is suggestive that the chosen

model may be correct, but not proof, as the solution is generally not unique. Other models with similar numbers of parameters could yield equally good fits. Furthermore, in the highly complex environment of the cell, it is difficult to insure that all of the processes that could be affecting the evolution of the measured quantities have been controlled. The fabrication of controlled assemblies in which ONLY the reactant species postulated in the physical model under test would solve many of the grave difficulties that are encountered in this type of modeling. Specifically, such systems would allow definitive tests of proposed models, by eliminating the possibility of unexpected competing reactions, and by allowing rigorous tests of the model through extensive variation of starting conditions.

4.3.2 High throughput data-bases

Key to DOE's Genomes to Life program is the goal of assembling predictive models of cellular function. Predictive models are the gold-standard in understanding a biological system as they describe the inner workings of a cell at a level that will be useful for engineering cellular systems for DOE applications in energy and environmental control. We feel that it is an important goal for the DOE to assemble a detailed predictive model of transcriptional regulation in an organism of interest to the DOE mission.

Many thought that data from expression microarrays would be the ultimate source of data for models of transcriptional regulation. Others thought that biochemical analysis of transcription factor proteins would be the essential source of understanding. The early reports are in on these individual data sources, and they are not encouraging. Expression microarrrays provide an important view of cellular function, but they do not provide a direct method of understanding the mechanics of transcriptional regulation. Thus, excessive assumptions need to be made when interpreting expression data in a vacuum, and when these assumptions are considered in the light of the inherent noise

in microarray measurements, the models that are generated are simply not comprehensive or predictive. Biochemical studies provide important information regarding the structure and function of individual macromolecules, but they do not provide much information about how these molecules act in concert to bring about the overall behavior of the cell.

The synergistic integration of data from multiple data sources to cellular function has proven to be the best approach for creating reliable predictive models. Although any one witness (such as expression data) may provide at times erroneous evidence, it is possible to build models that consider all evidence in a disciplined manner. For example, Lee et. al [11] describe a model of over 100 transcriptional regulators in Yeast based upon such an integrative approach. Figure 19 shows a portion of the regulatory network described by Lee et. al, along with the automatic resolution of key transcription factors that operate in the yeast cell cycle.

The key to an integrative approach is to have informative witnesses. Expression data can assist in discovering genes that are co-regulated, and sequence data can help to uncover motifs that explain co-regulation. However, direct in-vivo measurement of transcription-factor/genome interactions have turned out to be far more important than either expression or sequence data when one is discovering the mechanisms of transcriptional regulation.

Informative measurements of protein/DNA interactions can be made by creative application of DNA microarray technology. In so called "ChIP to Chip" or "location assays" the binding of a protein of interest can be directly observed under chosen conditions. The technology proceeds in three steps. First, cells with the desired genetic background are grown in a desired condition, and a cross-linking agent is added to cause proteins that are transiently bound to DNA to become covalently attached.

Second, the genome of the cells are sonicated into fragments, and the fragments that contain the protein of interest are immunopurified using stan-

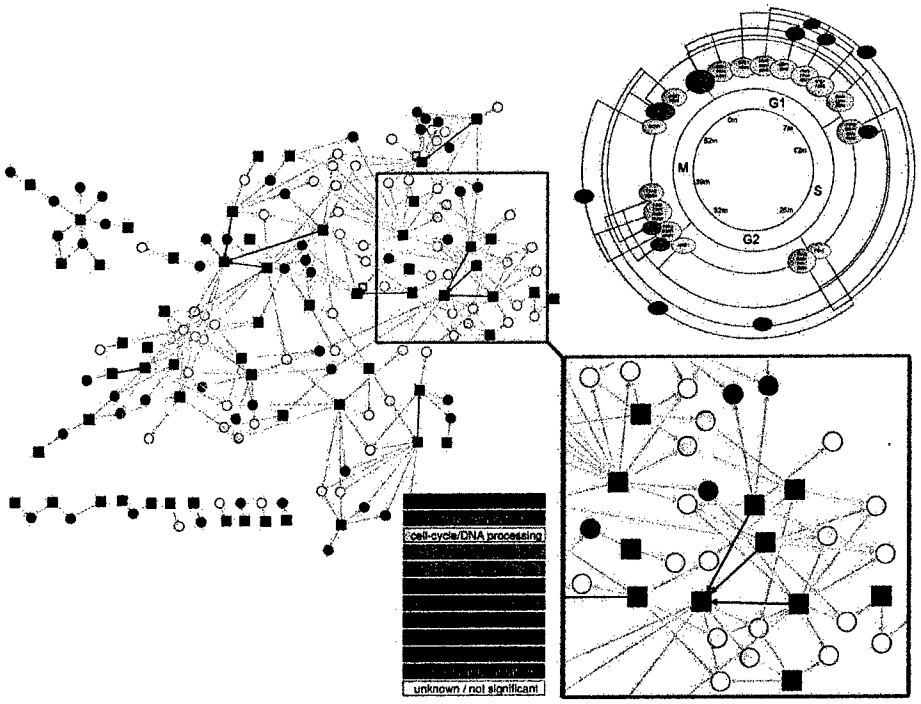


Figure 19: New micro/nano technology will enable a complete predictive model of transcriptional regulation for a selected organism in three years, Figure from Lee, et al., ref [11].

dard chromatin immunoprecipitation techniques and labeled. Finally, the resulting genome fragments are applied to an intergenic DNA microarray to determine which fragments were bound by the protein of interest. The resulting high-throughput data give direct insight into transcription factor / DNA interactions that are directly complementary to expression and sequence data.

In addition to location arrays, there are other new applications of microarray technology that we expect will be of interest to the DOE. For example, "cell array" technology permits the high-throughput examination of the effects of expression constructs on phenotype. Further applications of arrays

include the in-vitro detection of protein / DNA interactions by using tagged protein on a DNA micoarray. All of these microarray technologies provide important witnesses to cellular function.

We recommend that DOE actively address the theoretical and computational challenges of incorporating such multiple data sources, and especially high-throughput data, into cell modeling efforts. A correlated experimental/theoretical effort focused on the challenge of fully characterizing an organism of DOE interest would be an effective and significant research goal.

4.3.3 Bio-engineering Modeling

For centuries man has practiced breeding to provide variations of species to suit his needs. In recent years, much progress has been made in the field of "directed molecular evolution", where proteins of desired properties can be obtained from related natural molecules after just a few rounds of *in vitro* evolution. The future of nano-bio requires not only specific molecular components, but also 'control systems" that coordinate these components to work together in desired fashions. The latter becomes much more difficult to accomplish by directed evolution alone, and system-level modeling becomes an indispensable tool to *guide* the qualitative constructs of control systems which can be subsequently fine-tuned in experiments.

This approach is illustrated by the successful synthesis of two artificial gene networks shown in Figures 20(a) and (b): the bi-stable circuit by Gardner et al. [14] and the oscillator by Elowtize and Leibler [13]. In both cases, one started with simple molecular components derived from nature (e.g., "invertors" derived from the lac–, lambda–, and tet– repressors). Modeling was then used to lay down the connectivity of the network linking these components and estimate the range of critical parameters such as rate constants. The latter in turn provided the desired range and order of molecular

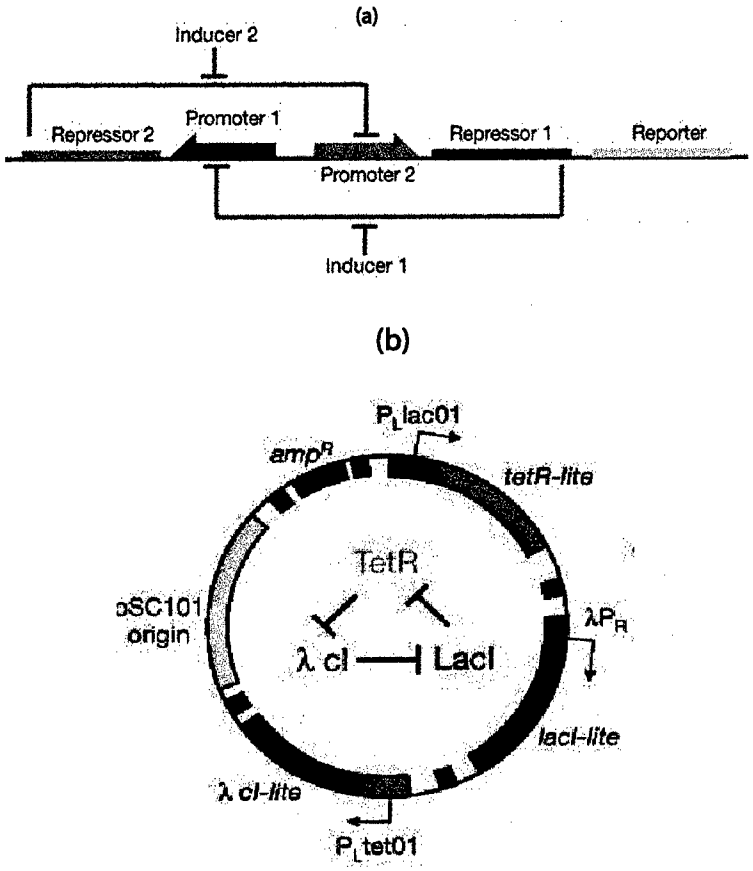


Figure 20: (a): A bi-stable molecular switch: product of gene 1 is a repressor which suppresses the expression of gene 2, while the product of gene 2 is also a repressor which suppresses the expression of gene 1 (Taken from Gardner, Cantor, and Collins, 2000.); (b): An oscillator consisted of 3 genes each coding a repressor, with the product of gene 1 (TetR) repressing gene 2 (λ cI) which represses gene 3 (LacI), and the product of gene 3 repressing gene 1. (Taken from Elowitze and Leibler, 2000).

interaction parameters, e.g., the binding strengths of various promoters, and became the rational starting point in search for the appropriate molecular systems.

More recently, this type of approach has been extended to construct various logic functions, e.g., the NAND gate, by combining and cascading the outputs of the invertors just mentioned (see Figure 21 and ref. [15]). While it might appear to system engineers that these simple gene circuits and gates can be combined to construct very elaborate control systems, just as complex electrical computing devices are routinely made from simple diodes and transistors, it is important to realize that biological systems have a number of specific features making them rather different from the electrical and mechanical devices that system engineers usually encounter. Specifically

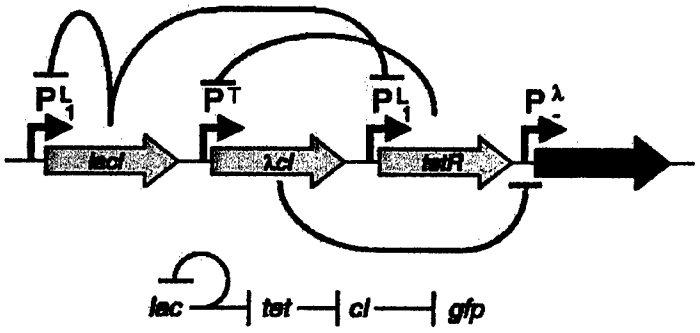


Figure 21: The genes in Figure20(b) can be connected differently to implement desired logic functions whose result is represented by the expression of GFP. (Taken from Guet et al, 2002).

for gene networks, extensive cascade of simple circuits is impractical for a number of reasons: (i) each cascade requires one round of gene transcription and protein synthesis which takes at least several minutes in bacteria and quickly approaches the life time of the organism; (ii) an exponentially large number of genes will need to be devoted to represent intermediate results of a complex computation, a big problem given the limited number of genes in the genome (several hundred transcription factors in total for E. coli); (iii) different cascades working in parallel will need to be carefully *synchronized* to

accomplish desired computations, but a reliable "clock" is difficult to make. [The oscillator of Figure 20b is very noisy [16].]

Solutions to the above problems can be glimpsed from the ways biological organisms perform complex computations, e.g., in the development of body plans. One of the best-characterized cis-regulatory control system is the regulation of the endo16 gene of sea urchin [17]. The complicated control exerted on the expression of the endo16 gene by its 13 inputs was not at all accomplished by gene cascades. In fact, no gene cascade was needed at all, and the control was instead accomplished through the intricate placement of binding sites for the 13 regulatory proteins (the "inputs") in a single regulatory region; see Figure 22. Yuh et al found that the organization of the

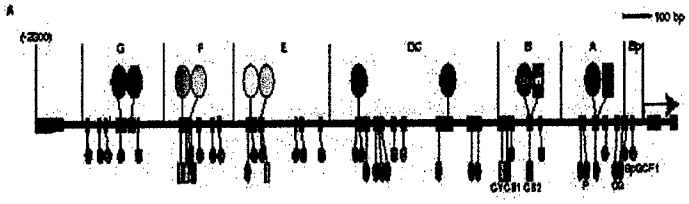


Figure 22: The cis-regulatory region of the endo16 gene: shown are several dozens of binding sites for the 13 different transcription factors controlling the expression of this gene. The regulatory region has a modular organization (denoted by the letters A through G) with each module either activating or repressing transcription. (Taken from Yuh, Bolouri and Davidson, 2001).

regulatory region itself is modular, with each module taking several inputs and exerting either an activating or repressing effect on gene transcription. Thus, gene cascade is effectively accomplished through modularly organized molecular interaction.

For the bioengineering purpose of designing control systems to coordinate artificial molecular components, it is crucial to understand the interaction between the regulatory proteins so that complex regulatory systems can be qualitatively constructed as the simpler circuits of Figures 20 and 21 can be designed today. Towards this end, it is worth investigating in detail the

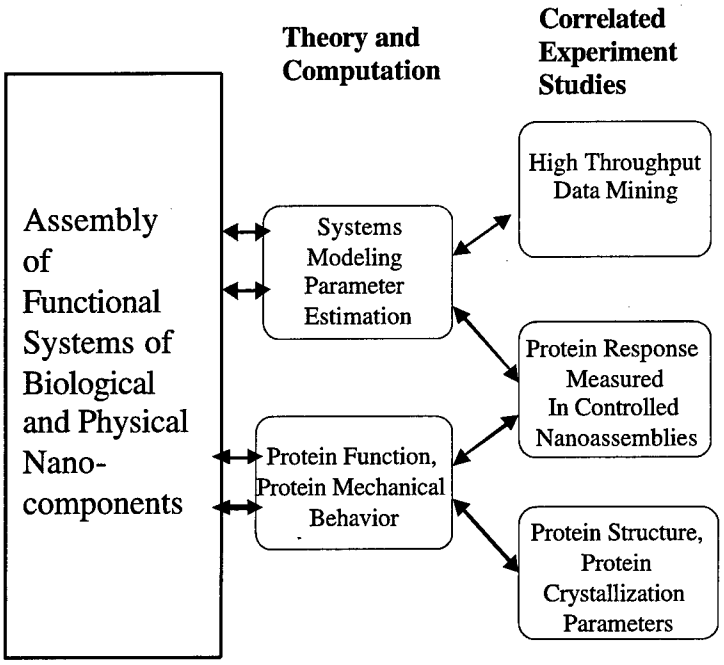
principle of "regulated recruitment" proposed by Mark Ptashne [18], claiming that generic, glue-like protein-protein interaction is sufficient to generate regulatory systems of virtually arbitrary complexity. Whatever the necessary molecular ingredients needed to implement complex regulatory control turn out to be, it is conceivable that these ingredients themselves can be realized by artificially engineering the regulatory proteins, e.g., via directed evolution. The task of putting these components together in a cis-regulatory region to implement desired control functions then depends critically on modeling effort. Even more dependent on modeling are the different ways of linking these control functions together, e.g., via feedback and feedforward loops, to perform complex molecular computations.

4.4 Computational Issues for Bio-Nano Assemblies

In the preceding sections we have addressed the experimental feasibility of fabricating artificial assemblies of proteins as biological nano-components, in possible combination with physical nanocomponents. We have suggested that one way to shape experimental studies on the design of artifical assemblies is in the context of developing improved understanding of cellular dynamics. In the long term, both the experimental techniques for fabrication of artificial assemblies, and the theoretical understanding of cellular regulation will be essential to the goal of creating designer systems.

In the following section, we will discuss another type of experimental motivation, in which the fabrication of artificial nano-assemblies can be used in support of computation of biomolecular properties. The broad problem is the molecular-level understanding of protein function. There are many components to this problem, including computational difficulty, the necessity for high-quality structural input, and the issue of modeling at many length scales. Because the problem of creating designed protein response is

so important, and so far from realization, many directions of research are needed. The interconnections among the general research issues in Sections 4 and 5 are illustrated below.



References

[1] P.S. Cremer, J.T. Groves, L.A. Kung, and S.G. Boxer, Writing and Erasing Barriers to Lateral Mobility into Fluid Phospholipid Bilayers, Langmuir 15, 3893, 1999.

[2] T. H. Bayburt and S.G. Sligar, Single-molecule height measurements on microsomal cytochrome P450 in nanometer-scale phospholipd bilayer disks, Proceedings of the National Academy of Science 99, 6725-30 (2002).

[3] D. Gust, T.A. Moore, and A. L. Moore, Mimicking Photosynthetic Solar Energy Transduction, Accounts of Chemical Research 34, 40-48, 2001.

[4] H. Lodish, et al., "Molecular Biology, Scientific American Books, New York, NY 1995.

[5] T. Ritz, A. Damjanovic, K.Schulten, The Quantum Physics of Photosynthesis, ChemPhysChem 3, 243-8, 2002.

[6] T. Elston, H. Wang and G. Oster, Energy transduction in ATP synthase, Nature 391, 510 (1998).

[7] D. Gershon, Naturejobs,, *Nature* **417**, 4-5. June 2002.

[8] M. Tomita, "Whole-cell simulation: a grand challenge of the 21st Century", *Trends in Biotechnology* **19** 205–210 (2001).

[9] D. Nelson and T. Hwa et al., JASON Report, Biofutures, JSR-00-130, (2001).

[10] I. Swameye, T. g. Muller, J. Timmer, O. Sander and U. Klingmuller, "Identification of Neucleo cytoplasmoic cycling as s remote sensor in cellular signaling," Preprint (2002).

[11] Tong Ihn Lee, et al., Transcriptional Regulatory Networks in Saccharomyces cerevisiae, Science, in press (2002).

[12] E. T. Farinas, T. Bulter and F. H. Arnold. Directed enzyme evolution. *Curr. Opin. Biotechnol.* **12**, 545 (2001).

[13] M. B. Elowitz and S. Leibler, A synthetic oscillatory network of transcriptional regulators, *Nature* **403**, 335 (2000).

[14] T. S. Gardner, C.R. Cantor, J.J. Collins, *Construction of a genetic toggle switch in Escherichia coli*, Nature 403, 339 (2000).

[15] C. C. Guet, M. B. Elowitz, W. H. Hsing, S. Leibler, Combinatorial synthesis of genetic networks. *Science* **296**, 1466-1470 (2002).

[16] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P.S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183-1186 (2002).

[17] C. H. Yuh, H. Bolouri, and E. H. Davidson, Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128**, 617-29 (2001).

[18] M. Ptashne, and A. Gann, *Genes and Signals* (Cold Spring Harbor Press, New York, 2002).

[19] H. H. McAdams and A. Arkin, *Gene regulation: Towards a circuit engineering discipline*, Current Biology 10, R318 (2000).

5 MOLECULAR MODELING CHALLENGES

In the previous section we developed a basic research link between the development of artificial assemblies of physical and biological nano-components and computational approaches to biological systems modeling. In this section, we address issues where experimental development of controlled protein environments can be expected to impact basic research issues in understanding protein function and crystallization at the molecular level. In particular, an important basic research goal is developing computational methods to predict structure-function correlations. As a first step in such a program , we define a research challenge requiring a close interaction between experiment and computation of biomolecular properties:

> Fabricate assemblies of proteins in which protein-protein, protein-lipid or protein-artificial nanocomponent interactions can be tailored, and test/tune computational capabilities to predict changes in their corresponding function.

We focus our discussion of this research challenge on membrane proteins because an important research direction at the nano-bio interface is the development of membrane-embedded proteins as components of intelligent sensor systems. This possibility is based, of course, on the fact that biological systems employ myriad different receptor proteins on their cell membrane; these receptors detect temporal (and sometimes spatial) variations of a variety of ligands (and, sometimes, other signals such as electrical voltage or mechanical stress) and, after "processing" the data in (still poorly understood) biochemical reaction networks, allow for cells to respond to their environment.

In the following sections, we describe the challenges and opportunities in detail, starting at the level of individual proteins, and working our way up to aggregates. In Section 5.1, we outline specific issues in molecular com-

putation, and present an example where computational methods have made significant progress for predicting single protein behavior. We discuss another example illustrating current limitations of such computations in Section 5.2. The following sections then address some of the coarse graining approaches that are currently in use, as well as examples of specific problems which provide a context for studying protein-protein and protein-solvent interactions both experimentally and computationally. The final section discusses the challenges associated with protein crystallization, and outlines a computational and theoretical strategy for controlling crystallization.

5.1 Approaches and Goals in Molecular Computation

As is usually the case in computational approaches to biological systems, the appropriate method is very much a question of what issues are being addressed – there is no such thing as a fully *ab initio* calculation for even the least complex biological system, and thus coarse-graining (e.g. length-scale bridging, or hierarchical modeling) type approximations must be made and then checked for validity by comparison with both experiments and limited-case finer-detailed computations. Appropriately defining the hierarchical level of description is a major challenge, which requires careful correlation with fundamental theoretical principles and with experimental observations. Designer protein assemblies, in which the local environment can be controlled and functional responses evaluated, can be developed and used only in a closely coordinated experimental/ computational program of research.

Some of the issues and capabilities for molecular computation are illustrated in Figure 23, provided by M. Colvin of LLNL. Improved computational capabilities have resulted in dramatic advances in the ability to perform *ab initio* calculations of small molecules under conditions of reaction,

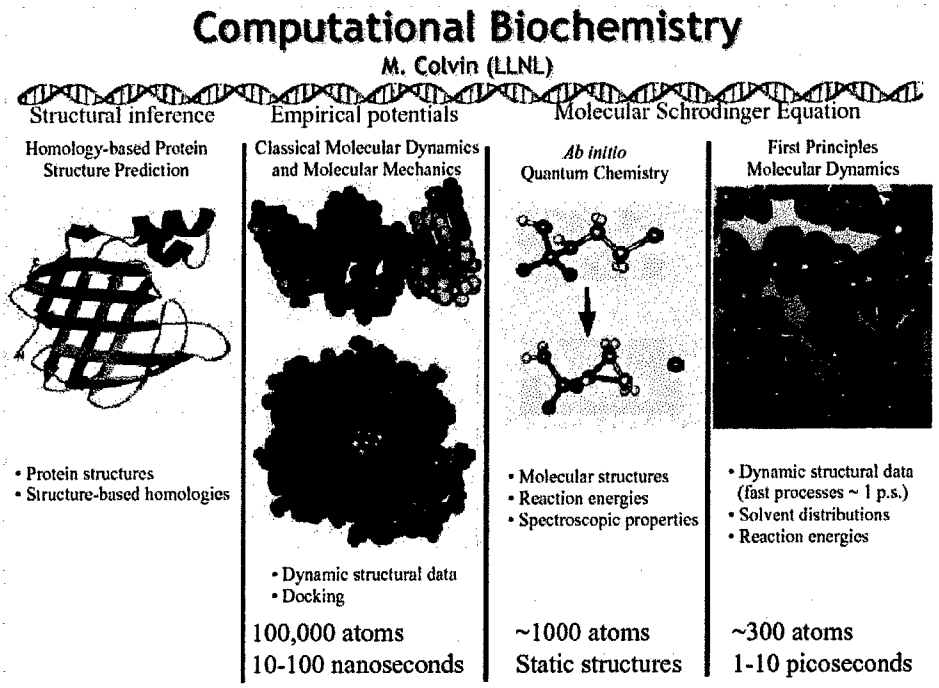


Figure 23: Issues and capabilities in computational biochemistry. Figure provided courtesy of Michael Colvin, Lawrence Livermore National Laboratory.

solvation, etc. This understanding of local bonding and reactivity is used to guide understanding of local chemical properties of biological molecules such as proteins.

However, the multiple length scales of secondary and tertiary structure in proteins cannot be addressed in this way. To deal with the biological molecules, it is necessary to begin with a fairly advanced knowledge of their structure. This can be provided either directly from crystallographic determination of the structure, or by homology, in which known patterns of folding and conformation for specific amino acid sequences are used to deduce major blocks of structure. Even given the structure, addressing biomolecular properties computationally still remains a difficult challenge. To attain computationally tractable modeling, empirical potentials are often used to describe the interactions within the protein and with its environment. These

potentials are tuned to provide a good description of the environments and conditions most commonly encountered, and thus careful consideration must be given to matching empirical potentials to the problems for which they are best suited.

Problems in computation of membrane protein function, and the linked experimental/theoretical problem of protein crystallization are outlined in the paragraphs below. Then a model case of computation tailored to a well-defined question is presented.

5.1.1 Membrane proteins: function and crystallization

Out of the broad spectrum of membrane proteins, receptor proteins provide an excellent example for illustrating the research challenge posed above. Typical receptors include G-protein coupled receptors and receptor tyrosine kinases. Receptor proteins typically detect and transmit signals as independent units, through structural changes of the protein upon ligand binding. Understanding the changes of receptor protein structures in response to either ligand binding and/or environmental changes (e.g. protein-protein interactions) is a crucial step towards using membrane-embedded proteins as a sensing system. The computational challenge here is clear. It will require a significant advance to be able to predict accurately the long range structural changes in a protein in response to a binding event, or to the perturbation (such as charge transfer) induced by a physical nanostructure that has been designed to interact with the protein in an artificial assembly.

An enormous bottleneck for computational studies in this direction is the simple fact that crystallization of membrane proteins has proven extremely difficult, so the number of structures that is known is not large. Given the reliance of computational methodologies on *a priori* knowledge of protein structure, any effort which speeds up the pace of membrane protein

crystallization will have significant impact on the ability of computation to contribute to our knowledge of membrane protein interactions. Moreover, it seems likely that any fundamental understanding of protein crystallization would also impact the more general problem of protein-protein interactions. In recent years there have been substantial efforts in developing experimental techniques of "high throughput crystallography", seeking to reduce the time required for cloning, expression, purification, etc. These approaches have naturally led to an explosion of data documenting the conditions under which specific proteins can crystallize. Developing methods to exploit this data to help develop a fundamental understanding of the process of crystallization of biological macromolecules, especially membrane proteins, is an important goal. Learning how to correlate such large, but indirectly linked information sets to the development of improved computational tools is a significant intellectual and technical challenge.

5.1.2 Illustration of computational application: molecular docking

The multi-level methods that must be used in computational molecular biology, and the strategies for developing effective applications are illustrated in a recent investigation by Goddard and co-workers on the sensitivity of olfactory receptors to different chemical species [1]. Olfactory receptors fall in the class of G-coupled protein receptors. These are membrane-bound proteins, which have binding sites for specific activating chemicals (ligands) on the portions of the protein which lie outside the cell membrane. Upon binding, a structural response causes the release of a G-protein (a signaling molecule) which was initially bound to the portion of the protein that protrudes into the cellular interior. Determining which chemical species bind to the outside of the protein is the so-called "molecular docking" problem, of

interest as a way of identifying target ligands for further testing in drug design [2, 3].

While most molecular docking studies are performed on proteins of known structure (e.g. structure determined by x-ray crystallography), relatively few membrane proteins have been characterized structurally (see section 5.4). Thus determining the protein structure computationally was necessary for addressing the properties of olfactory receptors. This problem was addressed by using computational approaches appropriate to the different length scales of the problem and the known physical properties of similar molecules [1]. Olfactory receptor proteins are known to have a structure in which seven helical coils traverse the thickness of the supporting membrane. The computation was started using structures based on this requirement, with membrane-embedded segments of the protein identified by computational analysis of the hydrophobicity of the amino-acid sequences, and by comparison with known folding properties based on the amino-acid sequence of the segments that fall within the membrane. The structures were then refined using molecular dynamics based on force-fields optimized first for the torsional degrees of freedom of the protein, and then for the interactions with the molecular structure of the phospholipid bilayer of the membrane. Then the segments of protein (the loops) that link the membrane-bound helical segments were added, and the entire structure was optimized using appropriate liquid phase interactions (including counter-ions Na+ and Cl-) for the exterior segments and lipid-phase interactions for the bound segments. To test the computational process, it was blindly applied to the protein bacteriorhodopsin, for which the crystal structure is known. The comparison of the calculated and measured structures is shown in Figure 24.

Clearly there are differences in the large-scale structure. However, the chemical binding properties for small chemical ligands (e.g. epinephrine, salbutamol) depend on the structural configuration over an only the area defining the local binding site. How well this can be predicted using protein

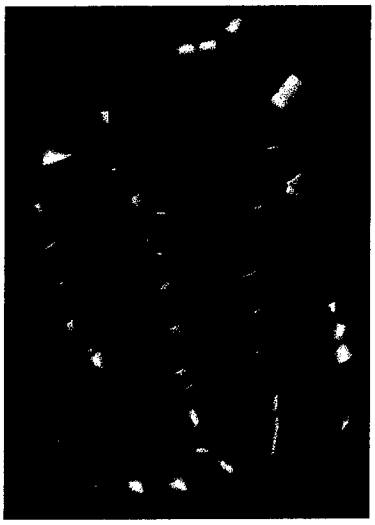


Figure 24: Comparison of calculated (blue) and experimental (red) structure of bacteriorhodopsin. Resolution of the experimental structure determination is 1.5 Å overall and 12 Å for the loops. The rms deviation in the calculated position of the alpha-carbon atoms is 6 Å overall and 8.6 Å for the loops. Figure provided by W. A. Goodard III, California Institute of Technology, from research presented in ref [1].

structures calculated as described above was determined by computational tests of binding of a variety of ligands for chosen olfactory receptor proteins. The results predicted binding sites for alcohol and acids sites, and correctly (and blindly) predicted compounds that are found experimentally to activate the binding site. Similar empirical tests of docking protocols in other cases have also shown a useful capability to identify binding molecules [2, 3].

This application of molecular computation is based on the acceptability of false positives: Clearly molecular binding is a necessary condition for activation of the receptor function, and ligand molecules that do not meet this requirement can be rejected. The sufficient condition for identifying an activating ligand, the observation of the functional response (e.g. release of the G-coupled protein), is not tested computationally. This is, in part, due to the large computational cost of determining the long-range relaxation of the

protein molecule in response to the binding, and in part due to the limitations of the computational accuracy needed to predict the correct molecular response. The nature of the large-scale structural responses involved in olfactory protein response is illustrated in Figure 25. Changes induced by the

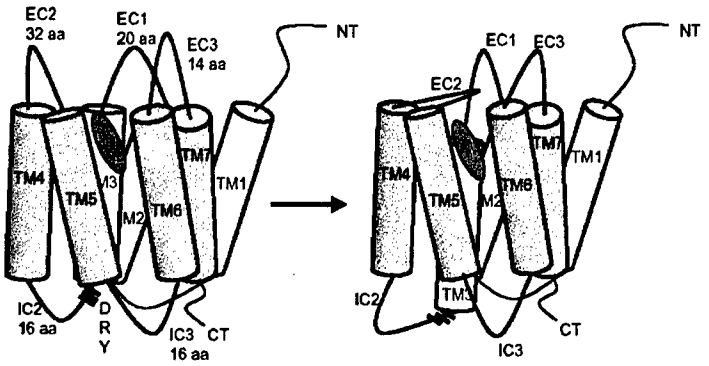


Figure 25: A model for signaling in olfactory receptors (G-coupled protein receptors). Grey and white cylinders indicate thehelical, membrane-embedded segments of the protein, and the exterior and interior loops are indicated by solid black lines. The initial interaction of the activating ligand, indicated as a the pink oval, is followed by structural relaxation of the external loops and coupled conformational changes of the helices and the interior loops, which interact with the G-coupled signaling protein. Figure provided by W.A. Goddard, California Institute of Technology.

local chemical binding of the ligand molecule induce conformational changes in the membrane-crossing portions of the protein, and this in turn results in changes on the segments of the protein protruding into the cytoplasm. Local shifts of atomic positions within the binding area, as well as overall conformational changes effect the unbinding of the signaling protein.

Developing computational methods to predict protein functional response is a serious issue, especially important for long term goals involving the design of biological systems with specialized behavior and biomimetic systems. Predicting functional responses, which are based on the types of subtle and long-range changes in structure described above (and in ensuing

sections) will certainly require the application of multiple scales of modeling, similar in philosophy to those used to model protein structure. A closely correlated interplay of experiment and theory can play an important role in this problem. Specifically, valuable inputs to theory can be derived from experiments in which the protein environment is perturbed, for instance by small changes in chemical structure or by assembly of the protein into artificial configurations, and the *change* in the proteins functional response is determined. Current problems in biology in which such strategies will be important are described in sections 5.3.2 and 5.4. Similar issues in experimental studies of the interactions of proteins with artificial nanostructures will also serve to guide and be guided by computational studies.

5.2 K+ ion Channel Membrane Protein

The example of Section 5.1 shows the success possible when computational capabilities are well matched to the question being addressed. In this section, we illustrate the challenges that can arise in predicting functional behavior, as illustrated by the potassium ion channel. This membrane protein allows selective permeation of potassium ions across the cell membrane. Remarkably, the rate at which potassium flows through the potassium channel is about 10000 times greater than the rate of sodium. The rate at which rubidium (another column 1 element) flows through the potassium channel is about 20 times that of potassium. The ionic radii of the three elements are 1.9 Å for sodium, 2.7 Å for potassium and 3 Å for rubidium).

The mechanism of the ion channel was a mystery until 1998, when MacKinnon and co-workers crystallized the potassium channel and characterized its structure with X-ray diffraction.[1] A cartoon of the channel structure is shown in Figure 26. The main structure of the channel consists of four pairs of alpha helices. The outermost helix in each pair provides structural

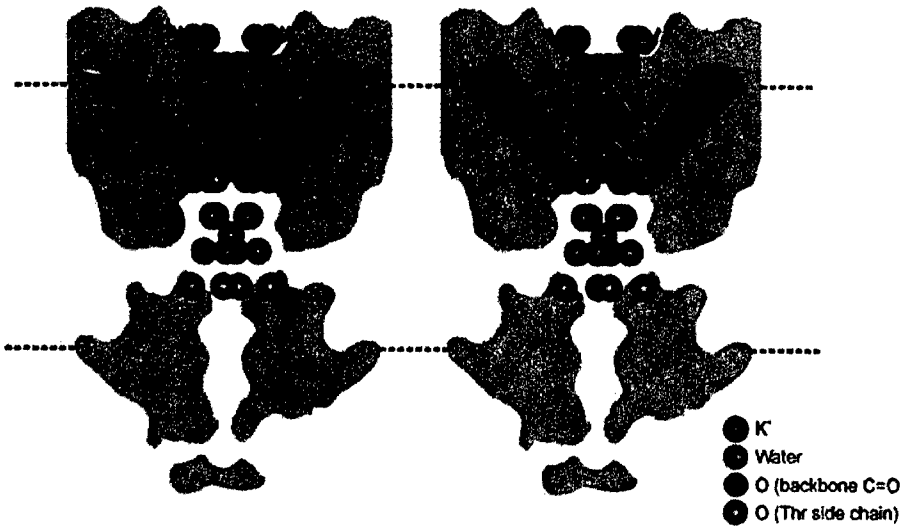


Figure 26: Cartoon of the structure of the Potassium Channel. The pore through the protein contains a central chamber of radius $\sim$ 10 Å and a 12 Å long section of much smaller radius lined with carbonyl oxygens. From Yellen [2].

stability, while the innermost helix is shorter and forms the lining of the outer part of the pore that spans the membrane. The dipole moments of the inner alpha helices are pointed towards the center of the channel, with the negative pole inward. This structure is believed to act as the "selectivity filter", the portion of the membrane protein that is responsible for selecting between the different ionic species. In the original X-ray structure determination, two potassium ions were observed in the selectivity filter, separated by 8 Å. However, recently Mackinnon's group solved for the structure at higher (2.0 Å) resolution, and observed two additional potassium sites.[3] The existence of such states was suggested independently from simulations by Bernéche and Roux [4] (using the lower resolution structure as input) based on the energetics of ions traversing the selectivity filter. This demonstrates that computations can be capable of "filling the gaps" of crystal structures.

However, computations did not fare as well with regard to the conduction mechanism in the selectivity filter. Mackinnon and co-workers have argued that the selectivity filter contains four states where ions are in local equilibriam. The conduction pathway involves the ions going through the pore two by two (Figure 27). An analysis of the measured electron density in

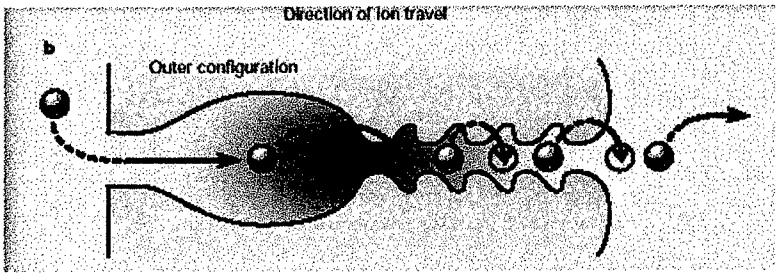


Figure 27: Cartoon of the function of the Potassium Channel, showing transition from the [1010] state to the [0101] state. From C. Miller, Nature, 414, 23 (2001).

the pores, as a function of the concentration of potassium around the crystallized protein, led to the conclusion that the energy difference between the [1010] and [0101] states is zero (the two states have exactly equal energy) whereas for rubidium the energy difference is 5 kBT. Mackinnon identified the zero energy barrier between these two states as the *essential* feature behind selectivity in the potassium channel.

Computational simulations attempting to unravel the operating principles of the selectivity filter disagreed with this result. Different simulations of the energy profile within the selectivity filter have led to quantitatively different results [4, 5] which disagree with those deduced from the high-resolution experiments.[6] One reason for this is that [7] the ion interaction with the selectivity filter is very sensitive to the precise structure: this is why potassium and sodium have such different permeabilities in the first place. Since, the original crystal structure did not resolve several of the side chains in the selectivity filter; the positions of these had to be chosen as part of the modeling

processes. Furthermore, the interactions between the channel, solvent and permeating ions are strong, and there are large compensating terms. This results in a tremendous sensitivity to the exact form of the empirical potential being used. Specifically in this case, the environment in the selectivity filter is not well handled by standard potentials, as discussed by Roux and Bernche.[8] They have shown how modification of the potential to deal with the special environment within the selectivity filter is needed for accurate modeling. The lesson to be learned from this example is that although computation can lead to significant successes, there are also limitations. These may or may not be important in specific cases, depending on how sensitively the function depends on structural details.

We conclude with one final remark regarding current limitations of computational methods. The overarching fundamental question that is posed by the potassium channel is why the structure evolved into its present form (which is known to be highly conserved among disparate species). Is this an optimized structure, or could other possible designs perform the same function as well or better? Lest the reader think that this question is cavalier and/or uninteresting, Figure 28 compares the potassium channel to the chlorine channel, which was also recently crystallized in Mackinnon's laboratory. The chlorine channel has a completely different structure, with both the water filled cavity and the supporting alpha helices having completely different configurations. Why does a negatively charged ion have a completely different channel structure than a positively charged ion?

Such broad questions represent long term goals to which we would like to push the capabilities of computational methods. A carefully coordinated research program focused on understanding structure-function correlations is the first step in bringing such questions within reach. As will be described in more detail in the following sections, experiments in which protein-protein interactions can be controlled and correlated to functional response, should

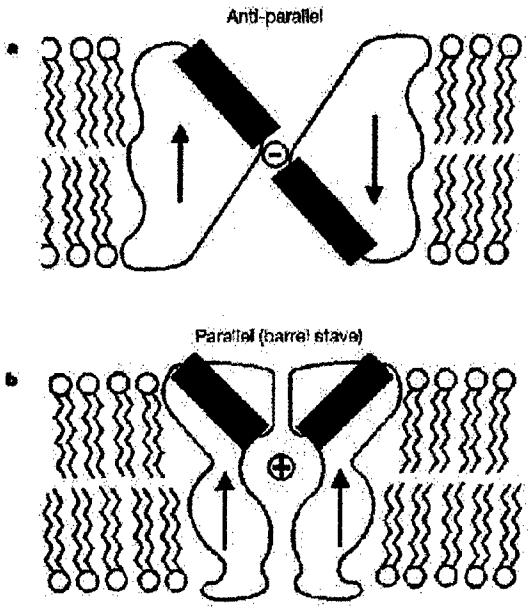


Figure 28: Cartoon comparing the potassium channel with the Chlorine channel. From "X-ray structure of a CIC chloride channel", R. Dutzler, E. Campell, M. Cadene, B. T. Chait and R. MacKinnon, Nature, 415, 287 (2002).

be developed and used to guide improving theoretical and computational approaches to predicting protein function.

5.3 Protein-Protein and Protein-Membrane Interactions

The methodologies for understanding structure-function relations of single proteins cannot practically address protein-protein and protein-membrane interactions. The computation time for such interactions is typically well beyond current capabilities. Since biological detection mechanisms can involve receptor complexes, it is necessary to develop computational capabilities for understanding and simulating such complexes. As above, our goal is to develop the capability for predict changes in function of both individual proteins

and protein complexes) in response to changes in structure of either individual proteins or protein complexes. It should be emphasized that included in this goal is the very real possibility that small changes in the structure of individual proteins can produce important changes in a protein complex. In the following we address generally some of the standard coarse-graining approaches that are used in understanding structure and function in interacting systems. We then describe an example of a biological detection system (the chemotactic response network in E. Coli), where it is believed that the signal detection properties depend on both structural changes of individual proteins, and interactions among proteins in an assembly.

5.3.1 Coarse grained interactions of proteins

Even given the structure of an individual protein, the simulation time attainable by even the fastest supercomputer (nanoseconds scale) pale in comparison to the time scales of interest for many important biological processes. As described above, carefully designed connections between energy surfaces and the kinetic processes of interest must be constructed to address changes in structure in response to binding or other changes in environment.

For dealing with structural interactions between proteins and their environment, another strategy must be adopted. Continuum approaches provide an intermediate scale of computation. For electrostatic interactions, this is afforded by implicit solvent (Poisson-Boltzman equation) approaches. Here, atomic detail of the water and lipid bilayer are replaced by effective media, (e.g. a dielectric with $\epsilon \approx 80$ water, $\epsilon \sim 2$ for the lipid). Continuum calculations for deformations of the lipid bilayer can also be devised. Such approaches have for example successfully addressed the problem of thermodynamic driving force that causes a protein to become inserted in the lipid membrane. Typically, membrane-bound proteins contain stretches of

hydrophobic amino acids; these form an alpha-helix secondary structure that prefer a non aqueous environment. For the simplest cases of an isolated protein with a single trans-membrane helix, there have been reasonably successful attempts to compute this free energy. For example, Kessel [12] studied the interaction of the 20 amino acid peptide, Alamethicin, with a lipid bilayer. They were able to understand typical experimental findings such as the free energy of insertion and the slight deformation of the bilayer due to a mismatch in the width of the peptide hydrophobic region and the equilibrium lipid bilayer width.

On the largest structural scale, a variety of phenomena including lateral segregation of proteins (i.e., clustering), phase transitions of the lipid bilayer structure (e.g. into regions with differing concentration), membrane-budding and/or fusion etc. may occur. It is not feasible nor would it be particularly in formative to study all these processes at the atomic scale of simulation. Instead coarse-grained approaches are both necessary and useful for establishing general physical principles. A typical method here treats the membrane-spanning hydrophobic part of the proton as a rigid rod, and the lipid layer as an elastic layer. The system is typically treated with something like a Helfrich free energy

$$F \simeq F_0 + \frac{1}{2}K(C - C_0)^2 + \frac{1}{2}\lambda(A - A_0)^2$$

where A_0 is the preferred area of the membrane with actual area A and C is the mean curvature, which would prefer at equilibrium to equal the spontaneous curvature C_0. An example of this approach is provided by Ref. [15] which studied how the interaction of peptides with hydrophobic mismatch can induce a phase transition in the lipid state between a bilayer structure and an inverted hexagonal "droplet" phase. If this occurred locally, it might induce some type of budding at that part of the membrane. Finally, we mentioned that this approach can be extended by inclusion of more statistically-based view of the lipid molecules, again treated in a continuum Flory-type framework (Ref. [16]). An important problem in theory and computation is

learning how to predict the coarse-grained parameters from computations of molecular properties.

Once one has a reasonable understanding and control over the structural features of the lipid-protein system, there still is a need for simulations that focus on functionality. One example, discussed in section 5.1, concerns quantum chemistry calculations that try to predict different binding affinities for different receptors/ligands. This approach draws on the fact that the potentials developed for biomolecules are well-tuned for predicting local chemical bonding. The question posed has been chosen to be well-matched to this capability: What is assessed is local binding of the actuator molecule, without the need to assess the functional response of the protein to the binding. In choosing such problems one must also evaluate whether a computational or direct experimental approach is a more direct path. If the goal were to just "tweak" the binding energies by making small changes in the protein, it might be feasible at present to use purely experimental methods (such as directed-evolution approaches [18] which select mutants based on desired chemical properties) than to rely on simulation.

However, if one wishes to assess functional response in such a system, then once the ligand binds (or once a physical perturbation such as a voltage is applied) there must be some method whereby the cytoplasmic side of the protein is altered to pass the signal on. For the case of bacterial chemotaxis (see next section), this method has been speculated to be a very small change (on the order of 1.5 Angstrom) in the position and orientation of the membrane-spanning helix. The change at the cytoplasmic end must be recognized by other molecules, which are then activated or repressed thereby. The continuum solvent Poisson-Boltzmann approach for known protein conformations might play a valuable and important role for such problems. Electrostatics should play a crucial role, as the controlled addition/deletion of extra charges to e.g. the cytoplasmic side of the receptors can be utilized to adapt the sensing system to the chemical environment (see next section).

We feel that there are some significant opportunities in the near future in these directions.

5.3.2 Receptor clustering in bacterial chemotaxis

The bacterial chemotactic system is an excellent example of how biology uses the self-assembly and autocatalytic properties of macromolecules to *detect and process information*, e.g., sense chemicals and direct motion. With many decades of careful experimental studies, this "simple" system involving fewer than 10 different types of well-characterized proteins is among the best-studied biological sub-systems. Yet, much remains to be understood in regard to the mechanism of information processing at the *system level.* Further advances in our understanding of these molecular systems will require the *integration* of the vast amount of existing knowledge on single molecules, molecular interactions, and physiological behavior with the science of phase ordering and self-assembly. New understanding derived from this integration is not only of use to biology, but may also serve as useful guide to the synthesis of artificial nano-systems to detect and report a wide variety of chemical signals of direct relevance to DOE missions, including energy sources and hazardous chemicals. Below we summarize the current understanding of bacterial chemotaxis and highlight problems/areas that will be particularly useful and rewarding to address by the methods of nano technology and large-scale computation.

Much of our knowledge of bacterial chemotaxis is derived from detailed studies of the E. coli chemotactic systems. The key features that emerged from these studies have been found to persist throughout most of the bacteria kingdom.

A bacterium propels itself in the aqueous environment by rotating its flagella. In the presence of spatial gradients of chemical attractants (e.g.,

sugars and certain amino acids) and repellants (e.g., acids and other amino acids), the bacterium generates directed motion by modulating the swimming and tumbling motion according to temporal changes in the concentration of chemical sensed. This results in a biased-diffusion towards the attractants and away from the repellants.

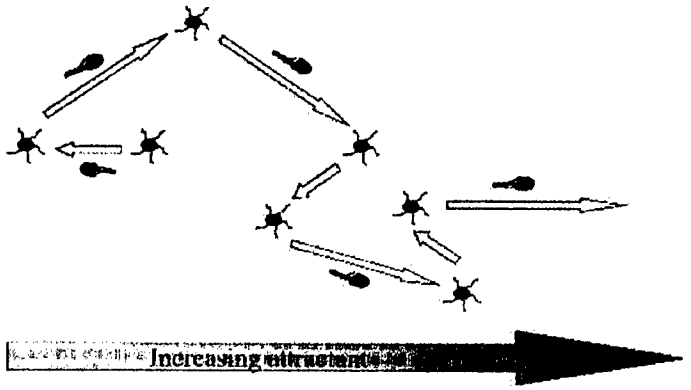


Figure 29: Biased diffusion of a bacterium: If it happens to be swimming up/down an attractant gradient, then its motors are slightly biased to/from rotating in the CCW direction to prolong/shorten the swim phase. This results in a net motion towards the attractant. Figure from reference [26].

The decision of how to modulate motor rotation according to the detected ligand types and concentration is made by the membrane-bound *receptor complex* via the interaction of the receptors with four coordinated proteins (CheA, CheW, CheR, CheB) on the cytoplasmic side of the membrane. E. coli is known to have 5 different families of receptors (Tsr, Tar, Trg, Tap, Aer), each specialized to detect its own spectrum of stimuli, either directly or with the help of periplasmic proteins. In addition to detecting chemicals in the periplasm, these receptors also respond to temperature and pH changes. The receptor molecules number about 7,000 per cell for Tsr and Tar, the "major" receptors, and $\sim$ 10% of those values for each of the other

three types ("minor" receptors), giving a total of about 10,000 receptor monomers per cell.

The four chemical sensing receptors are similar in structure, each consisting of a periplasmic sensing domain, connected across the inner membrane of the bacterium by a linker domain to a cytoplasmic domain which consists of two long anti-parallel a-helices. Two such receptor units homodimerize, forming two symmetric ligand binding sites on the periplasmic side and a four-helix bundle on the cytoplasmic side; see Figure 30 (left panel).

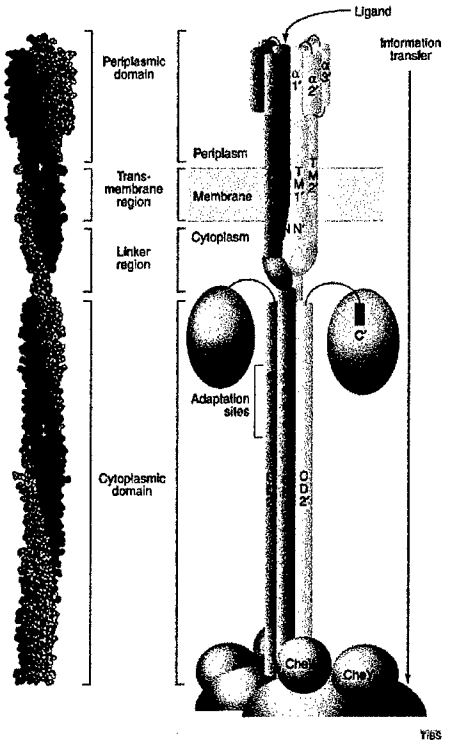


Figure 30: Structure of the transmembrane receptor dimer. Left: crystal structure of the dimer, with each color denoting one of the monomeric subunit. Right: schematic structure, indicating the binding of CheW and CheA to the cytoplasmic end of the dimer, and the binding of CheR and CheB to the C-terminus on the cytoplasmic side. The latter proteins are involved in the methylation and demethylation of the glutamine/glutamate residues (indicated as the "adaptation sites") along the cytoplasmic 4-helix bundles. Figure from reference [25].

Upon binding of a ligand molecule to one of the two binding sites of the receptor dimer, a small asymmetry in the dimer structure is introduced. This asymmetry shuts off further binding to the unoccupied site, and sends a not-yet-understood signal (e.g., a small structural shift) to the cytoplasmic side of the dimer. This signal affects the proteins CheW and CheA that bind to the cytoplasmic end of the dimer (Figure 30, right panel), with the net result that attractants (repellants) decrease (increase) the rate of CheA dimerization and auto-phosphorylation. Phosphorylated CheA then triggers the signal relay that controls the flagella. The range of the possible CheA phosphorylation activity can be 1,000 fold.

The rate of CheA phosphorylation actually depends not only on ligand binding, but can also be modified (e.g., compensated) by putting on or taking off charges at various locations along the cytoplasmic 4-helix bundle. This process is carried out by two enzymes CheR and CheB, which methylates (neutralizes) and demethylates (reinstalls) the negatively charged glutamate residues respectively.

It is traditionally believed that the receptor dimmers function as independent units in the mechanism described above. This view has been challenged by an increasing number of experimental observations, which collectively suggest that receptors form "higher order" structures and these structures may be important for their function as signal processing devices. Among others, the results include:

- The crystal structure of the cytoplasmic domain of Tsr indicates that the four-helix bundles are organized into trimeric clusters [25].
- Soluble constructs of cytoplasmic domains of receptors can form complexes mediated by CheW's and CheA's; the phosphorylation activity of the CheA's in such complexes can be essentially the same as that in fully active complexes with the intact receptor in the membrane [22].

- Complexes formed by adding purified CheW and CheA to Tar and Tsr receptors in E. coli membranes show similar phosphorylation activity as the complex of soluble constructs described above, with a stoichiometry of approximately 6 receptors: 4 CheW: 1 CheA monomers ([26]).

- The Trg receptor has been shown to form a 2D crystalline array in phospholipid membranes with a square unit cell, most likely corresponding to 4 receptor dimers ([20]).

- The major receptors Tsr and Tar are found by immunoelectron micrographs (see e.g., Lybarger and Maddock, 2000) to cluster in vivo into large aggregates and localize at one or both of the cell poles. However the minor receptors Tap and Trg neither cluster nor signal without the major receptors ([28]).

A qualitative picture emerging from these studies is that the receptors tend to form an array mediated by CheW and CheA. An ordered array such as the one proposed in Figure 31 is thought to promote CheA dimerization and phosphorylation, whereas disruption of the array due to, e.g., structural distortion of the ligand-bound receptor dimers or electrostatic repulsion induced by demethylation, will inhibit CheA dimerization and phosphorylation.

While this qualitative picture is very useful in organizing one's thoughts, it is far from predictive. Indeed, even a number of important qualitative questions remain unanswered, e.g.,

- Do the two major receptors form one cluster or do they segregate into two distinct clusters?

- Are the minor receptors interspersed among the cluster(s) formed by the major receptors? If so, why is there still a big effect on the neighboring receptors upon the application of multi-valent ligand?

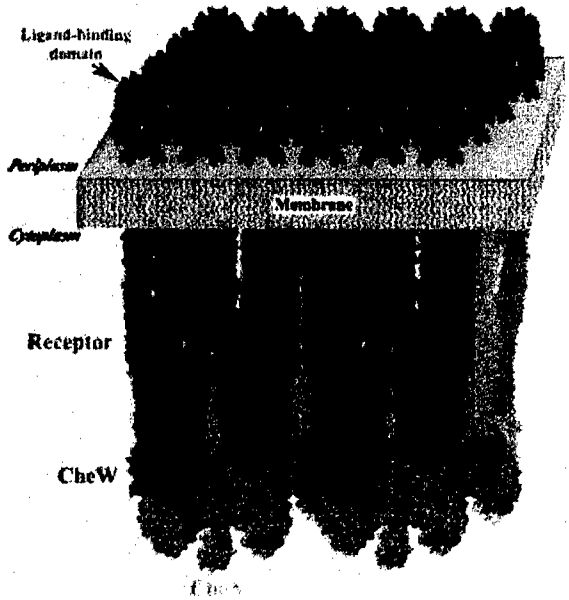


Figure 31: A proposed model of the receptor array mediated by CheW and CheA. In this model, the 4-helix bundles of each receptor dimer (elongated objects in teak) are organized into trimers with the help of the CheW's. Each such trimer then binds to one subunit of CheA. The spacing of the receptor array is such that two neighboring CheA's can dimerize and hence phosphorylate each other. From (Levit, Grebe, Stock, 2002).

- What is the nature of the "disruption" induced by ligand-binding or demethylation? Is the receptor array just locally dilated or does it actually change its order?

The few sample questions posed above concerning the clustering of the receptor-CheW-CheA complexes are very familiar from studies of macromolecular *self-assembly*. What makes the chemotactic system distinct from a conventional self-assembling nano system is that the former is a "smart" medium, whose local structural properties are controlled by the local packing of receptors, through the negative feedback by CheB: If receptors become locally disordered due to attractant binding, the CheAs dissociate and phos-

phorylation rate decreases. But the reduction of CheA phosphorylation also reduces the activity of CheB. The latter leads to a net increase in methylation (charge neutralization) by CheR, which restores the order of the local receptor packing.

The possibility of creating controlled arrays of these transmembrane proteins would open many opportunities for addressing their mechanism. The experimental tools of molecular biology and nano science such as FRET and AFM could be used to probe the segregation/clustering of the different types of receptors, and the local array structure and its stability under different methylation/demethylation conditions. In combination with assays of the functional response of the proteins under these well-defined conditions, the impact of the protein environment on its functional response could be quantified.

This system also lends itself to large-scale computational studies, both at the molecular level and at the systems level. First of all, crystal structures for most of the relevant proteins have been determined. Secondly, there is a great deal of biochemical data on equilibrium binding constants and rate constants for many of the interactions involved, even for some involving mutants. The availability of quantitative macro-scale characterization of the system by the experiments suggested above should provide valuable constraints and feedback for the computational studies. A class of interesting, system-level questions can be *uniquely* addressed by large-scale computation. They are exemplified by the interplay between the microscopic properties of the components and the *emergent* properties of the system, e.g., sensitivity of the array ordering to structural or electrostatic perturbation of the receptor dimers, stability of the array involving different receptor homodimers and also heterodimers, stability of the array with different lipids, effect of the array ordering on the dimerization of CheAs, dependence of structural stabilization on the kinetics of feedback from CheB, etc.

5.4 Protein Crystallization

Protein crystallization is the single most serious and obvious problem in which understanding protein-protein interactions has immediate impact. Although it is straightforward to determine (from, e.g., messenger RNA's extracted from cells) the sequence of amino acids that make up a particular protein, the folded structure that leads to a functioning macromolecule cannot be reliably predicted at the present time. The time honored experimental method for determining protein structure is X-ray crystallography which, under favorable conditions, allows Angstrom-resolution images of the positions of the atoms which form the individual amino acids. Unfortunately, only about 4,000 protein structures have been determined by X-ray diffraction, one crucial bottleneck being the difficulty in obtaining large protein single crystals for experimental analysis.

In the following, we discuss two different aspects of the potential interaction of experimental protein crystallography with computation. On the one hand, the use of combinatoric approaches to crystallization yields a large data base which can be mined to guide development of computational descriptions of protein-protein interactions. On the other hand, understanding the fundamental physics of protein-protein interactions can be used to guide experimental approaches to crystallizing proteins for structure determination.

5.4.1 High-throughput crystallography

Following the acquisition of the complete genome sequence of humans and many other species, there has been a surge of interest in "proteomics", the study of the structure and function of the hundreds of thousands of known proteins. Central to these studies is the determination of protein structure at

atomic resolution, chiefly by X-ray crystallographic methods. Ten years ago it was not unreasonable for an investigator to devote one or even a few years to determine the structure of a single protein. The demands of proteomics, both scientific and commercial, require a 100–1000-fold increase in the rate at which new structures are determined. According to McPherson,[30]

> *"The problem of crystallization ... contains substantial component of trial and error, and draws ... from the collective experience of the past century.... It is much like prospecting for gold".*

As a result, substantial efforts are presently underway in "high-throughput crystallography", seeking to reduce the time required for cloning, expression, purification, crystallization, diffraction data collection, and structure determination.[29] At present the greatest obstacle to achieving the desired level of throughput, especially for a broad range of candidate proteins, is the need to obtain crystals that diffract to high resolution. While there is a reasonable theoretical understanding of the crystallization process, derived mostly from studies in mineralogy and materials science, the process of crystallizing biological macromolecules is largely empirical. One usually adopts a shotgun approach, testing a variety of crystallization conditions that differ with respect to temperature, pH, salt concentration, presence of small-molecule additives, and choice and concentration of precipitant. The shotgun approach is guided by past experience, typically codified as a matrix of favored combinations of the various crystallization parameters.

One of the most significant advances in high-throughput crystallography of the past several years has been the augmentation of the shotgun approach through the use of automated and semi-automated methods. Several academic research consortia (e.g. Structural Genomics Centers, jointly supported by DOE OBER, NIH NIGMS, and NSF) and biotechnology companies (e.g. Syrrx and Structural Genomics Inc.) have applied a combination of microfluidics, robotics, and image analysis to conduct thousands of crys-

tallization trials per day. This "machine gun" approach is proving to be fruitful, although it is not yet clear whether it will extend to highly basic and other atypical proteins that may not be amenable to crystallization.

One consequence of the recent efforts in high-throughput crystallography is that very large databases are being generated concerning crystallization conditions that have or have not proven fruitful. In many cases the atomic-resolution structure of the (ultimately) crystallized protein is available. These data can provide a test bed for theoretical and computational studies of the crystallization process. Those who aim to model the kinetic, thermodynamic, and geometric features of crystal packing and growth should learn how to evaluate and use the spectrum of test data concerning the effects of variable conditions on the success of crystallization. It is reasonable to expect that the resulting improved computational approaches will in turn provide guidance to experimentalists on how best to conduct crystallization screens, as suggested for one specific approach in the following section.

5.4.2 Enhancement of protein crystallization by critical fluctuations

The accumulation of experimental data regarding the conditions for protein crystallization presents a significant theoretical and computational opportunity. Recent ideas inspired by colloid physical chemistry have suggested why protein crystallization is so much more difficult than crystallization in materials science, and moreover suggest how one might create conditions favorable to protein crystallization given a quantitative understanding of protein-protein interactions [31] –[35]. Briefly, the major distinguishing feature of protein crystallization is that the interaction range of proteins is much shorter than the protein size; this causes the usual gas-liquid critical point to become metastable and exist only within a region of two-phase fluid-crystal coexistence. However, supersaturated protein solutions placed

in the vicinity of this metastable critical point are predicted to have crystal nucleation rates enhanced by many orders of magnitude due to critical point fluctuations. However, *subsequent* growth of the crystal nucleus is expected to be slow, suggesting that crystals of high purity might be produced by this technique.

5.4.2.1. Pair Potentials and Phase Diagrams

Controlling crystallization is ultimately based on protein-protein interactions and how they are mediated by the conditions of the solution from which the crystals will be grown. The conditions necessary for easy protein crystallization are illuminated by correlating crystallization rates with the second virial coefficient $B_2(T)$, which determines the leading correction in the protein number density ρ to the osmotic pressure $\pi(\rho, T)$,

$$\pi(\rho, T)/\rho k_B T = 1 + B_2(T)\rho + O(\rho^2). \tag{5-1}$$

In the simplest approximation, if proteins in solution are idealized as approximately 5 nm spheres with an isotropic pair potential $V(r)$ which tends to zero for large r, then the second virial coefficient is given by

$$B_2(T) = 2\pi \int_0 r^2\, dr[1 - \exp(-V(r)/k_B T)] \tag{5-2}$$

A purely repulsive interaction (such as that between hard spheres) leads to a positive correction to the osmotic pressure expected from ideal solution theory. If, however, the potential has an attractive part, then $B_2(T)$ can be negative over a range of temperatures. It turns out that easy protein crystallization is only possible over a narrow range of (slightly negative) $B_2(T)$: solutions with positive B_2 fail to crystallize on an experimental time scale and those with B_2 large and negative lead to a randomly aggregated "gel" state.[36] These observations are consistent with a small temperature window for optimal protein crystallization arising from pair potentials with an attractive part which is quite *short range* on the scale set by the effective protein diameter.

The pair potentials for proteins are quite different from atoms and small molecules, for which the attractive well has a size *comparable* to the effective hard core diameter. For instance, in the familiar Lennard-Jones 6–12 pair potential, for atoms the effective hard core diameter is $r \approx \sigma$, and the attractive part extends out of $r \approx 2r_{\min} = 2^{1/6}\sigma$, so these quantities have the same order of magnitude. Proteins, on the other hand, cannot be represented by a pair potential with a single length scale. The effective hard core diameter is 50 Angstroms or more, while the range of the attractive part of the interaction (assuming good screening of electrostatic interactions) is still only a few atomic diameters. As the result, the phase diagram of proteins in solution is quite different from that of atoms or small molecules as shown in Figure 32.

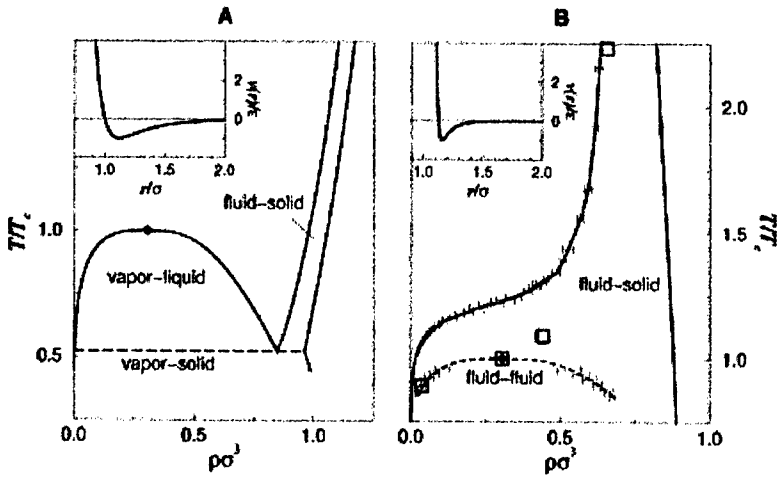


Figure 32: Phase diagrams as function of the number density ρ (scaled using the hard core diameter σ) and temperature T (scaled with the liquid-gas critical temperature T_c) for two different pair potentials. The phase diagram for conventional matter depicted in part A has the familiar solid, liquid and vapor phases with the potential shown in the inset. For the potential with the much shorter range attractive part in B, the vapor-liquid critical point arises as a metastable state in the region of fluid-solid coexistence. Figure from ref. [31].

In the phase diagram for atoms or small molecules (Part A of Figure 32), vapor, liquid and solid phases are typically separated by first order phase transitions (represented here by regions of two-phase coexistence), with a vapor-liquid critical point at the temperature $T = T_c$. Proteins in solution can be modeled by a potential with a very short-range attractive part (Part B of Figure 32). The critical point is then metastable and only occurs in a region of two phase fluid-solid coexistence.

5.4.2.2 Metastable critical points and crystal nucleation

Crystal nucleation under the conditions of Figure 32B has been studied via Monte Carlo simulations by Wolde and Frenkel. [31] Nucleation rates are as much as 10^{13} times large in the vicinity of the metastable vapor-liquid critical point! A qualitative explanation is summarized in Figure 33.

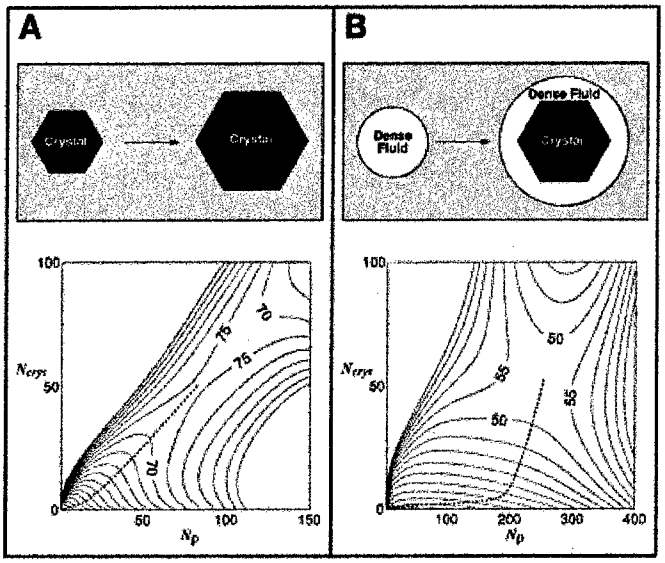


Figure 33: Free energy contours for conventional nucleation (A) and nucleation near a vapor-solid critical point (B). Horizontal axis measures the number of particles in a (roughly spherical) denser critical nucleus which emerges from the fluid, while the vertical axis denotes those particles which sit in crystalline environments. Figure from ref. [31].

Part A shows conventional nucleation, where the path to the saddle point representing the critical crystal nucleation corresponds to *simultaneous* growth of the number of atoms in a dense local region and the number of these atoms which exist in crystalline environments. In Part B, which models a protein solution close to its critical point, the cluster grows along a trajectory to a saddle point by first nucleating a small droplet of dense liquid *before* nucleating a crystal in the center of this droplet. As indicated in the inset, the growing crystal has a "wetting layer" of liquid surrounding it. The extra liquid layer lowers the effective interfacial tension between the crystal and the fluid from which it nucleates. The nucleation barrier to crystallization is given by [33]

$$E_b = 16\pi\sigma^3/[3\rho_s^2(\Delta\mu)^2], \tag{5-3}$$

where ρ_s is the density in the solid phase and $\Delta\mu$ is the chemical potential difference between the fluid and solid. The probability of nucleating a critical cluster is proportional to $\exp(-E_b/k_BT)$. Hence, any mechanism which reduces the surface tension σ can have an enormous effect on the nucleation rate. Critical point fluctuations evidently help the system explore phase space and find the critical nucleation. The resulting growth process is presumably rather slow due to critical slowing down and the diffuse nature of the interface.

5.4.2.3 Link to computational studies of protein-protein interactions

Empirically, one changes the nature of protein-protein interactions by changing solution parameters such as pH, temperature, small molecule additives, or by adding a non-bonding polymer such as polyethylene glycol and exploiting the depletion effect. One would like to systematize this approach by using computational potentials to determine virial coefficients, or more directly to calculate phase diagrams directly under variable conditions

of the solution. As noted in the previous section, combinatorial methods of protein crystallization can provide an extensive data base of information to guide and test development of potentials well suited to such prediction. An intellectually challenging problem is to develop procedures for coupling such information to tuning the tools used in computation. Ultimately, the development of computational tools that can predict protein interactions at this level will feedback into improved methods of crystallization, as well as into understanding of protein function, as described in early sections of this report.

References

[1] D. A. Doyle, R. Mackinon, "Structure of the potassium channel: Molecular Basis of K+ Conduction and Selectivity", Science, 280, 69, 1998.

[2] G. Yellen, "Transmembrance K+ Channel," Nature Structure Biology, **8**, 1011 (2002).

[3] Y. Zhou, J. H. Morais-Cabral, A. Kaufman, R. MacKinnon, "Chemistry of ion coordination and hydration revealed by a K+ channel at 2 A resolution", Nature, **414**, 43 (2001).

[4] S. Berneche, B. Roux,"Energetics of ion conduction through the K+ channel", Nature, **414**, 73 (2001).

[5] J. Aqvist, V. Luzhkov, "Ion permeation mechanism of the potassium channel", Nature, **404**, 881 (2000).

[6] J. H. Morais-Cabral, Y. Zhou and R. MacKinnon, "Energetic optimization of ion conduction rate by the K+ selectivity filter", Nature, **414**,37 (2001).

[7] D. P. Tieleman, P.C. Biggin, G. R. Smit, M. S. P. Sansom, "Simulation approaches to ion channel structure-function relationships", Quart. Rev. Biophys., **34**, 473 (2001).

[8] B. Roux, S. Berneche, "On the Potential Fucntions used in Molecular Dynamics Simulations of Ion Channels", Biophysical Journal 82, 1681 (2002).

[9] W A. Goddard, III, T. Cagin, M. Blanco, N. Vaidehi, S. Dasgupta, W. Floriano, M. Melmares, J. Kua, G. Zamanakos, S. Kashihara, M. Iotov, G. Gao, "Strategies for multiscale modeling and simulation of organic materials: polymers and biopolymers", Computational and Theoretical Polymer Science 11, 329 (2001).

[10] B. K. Schoichet, S. L. McGovern, G. Wei, and J. J. Irwin, "Lead Discovery using Molecular Docking", Current Opinions in Structural Biology 6, 439 (2002).

[11] Y. P. Pang, E. Perola, K. Xu, F.G. Prendergast, Eudoc: " A Computer Program for Identification of Drug Interaction Sites in Macromolecules and Drug Leads from Chemical Databases", Journal of Computational Chemistry, 22 1750 (2001).

[12] A. Kessel, D Cafiso and N. Ben-Tal "Continuum Solvent Model Calculations of Alamethicin-Membrane Interactions and Thermodynamic Aspects", Biophys J. **78**, 571-583 (2000).

[13] D. P. Tieleman, M. S. Sansom and H. J. Berendsen, "Alamethicin helices in a bilayer and in solution", Biophys J. **76**, 40–49 (1999).

[14] W. Helfrich "Elastic properties of Lipid Bilayers: Theory and Possible Experiments", Z. Nature **28**, 693–703 (1973).

[15] S. May and A. Ban-Shaul, "Molecular theory of Lipid-Protein Interaction and the L_α - H_{11} transition", Biophy J. **76**, 751–767 (1999).

[16] D. Dugue, X-J. Li and M. Schick, "Molecular Theory of Hydrophobic Mismatch between Lipids and Peptides", J. Chem Phys **116**, 10478–10484 (2002).

[17] T. L. Yarbrough, T. Lu, J. C. Lee and E. F. Shibeta "Localization of Cardiac Sodium Channels in Caveolin-rich Membrane Domains", Circ. Res. **90**, 443–449 (2002).

[18] J. A. Kolman and W. P. Stemmer, "Directed Evolution of Proteins by Exon Shuffling", Nat Biotech **19**, 423–8 (2001); J. E. Ness, S. B. Cardayre, J. Minshull and W. P. Stemmer, "Molecular Breeding: The Natural Approach to Protein Design", Adv Prot, Chem, **55**, 261–292 (2000).

[19] D. M. Davis, "Assembly of the Immunological Synapse for T Cells and NK Cells", Trends Immun., **23** 356–63 (2002)

[20] A. N. Barnakov, K. H. Downing, G. L. Hazelbauer, "Studies of the structural organization of a bacterial chemoreceptor by electron microscopy", J Struct Biol. 112:117-24 (1994).

[21] X. Feng, J. W. Baumgartner, G. L. Hazelbauer, "High- and low-abundance chemoreceptors in Escherichia coli: differential activities associated with closely related cytoplasmic domains", J Bacteriol. 179: 6714-20 (1997).

[22] N. R. Francis, M. N.Levit, T. R.Shaikh, L. A. Melanson, J. B.Stock, D. J.DeRosier, Subunit Organization in a Soluble Complex of Tar, CheW, and CheA by Electron Microscopy, Journal of Biological Chemistry (2002).

[23] J. E. Gestwicki, L. L. Kiessling, "Inter-receptor communication through arrays of bacterial chemoreceptors", Nature **415**: 81-84 (2002).

[24] G. L. Hazelbauer, C. Park, D. N. Nowlin, "Adaptational "crosstalk" and the crucial role of methylation in chemotactic migration by Escherichia coli", Proceedings of the National Academy of Science 86: 1448-52 (1989).

[25] K. L. Kim, H. Yokota, S. H. Kim, "Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor", Nature 400: 787-92 (1999).

[26] M. N. Levit, T. W. Grebe, J. B. Stock, "Organization of the Receptor-Kinase Signaling Array that Regulates E. coli Chemotaxis", Journal of Biological Chemistry (2002).

[27] Y. Liu, M. N. Levit, R. Lurz, M. G. Surette, J. B. Stock, "Receptor-mediated protein kinase activation and the mechanism of transmembrane signaling in bacterial chemotaxis", European Molecular Biology Organization 16: 7231-7240 (1997).

[28] S. R. Lybarger, J. R. Maddock, "Differences in the polar clustering of the high- and low- abundance chemoreceptors of E. coli", Proceedings of the National Academy of Science, 97: 8057-8062 (2000)

[29] B. Byrne and S. Iwata, "Membrane Protein Complexes", Current Opinion in Structural Biology 12, 239 (2002).

[30] A. McPherson, "Preparation and Analysis of Protein Crystals" (Krieger, Malabar, Florida, 1982).

[31] P. R. ten Wolde and D. Frenkel, Science, "Enhancement of Protein Crystal Nucleation by Critical Density Fluctuations", Science 277, 1975 (1997).

[32] O. Galkin and P. G. Vekilov, "Control of Protein Crystal Nucleation around the Metastable Liquid-Liquid Phase Boundary", Proceedings of the National Academy of Science 97, 6277 (2000).

[33] V. Talanquer and D. Oxtoby, "Crystal Nucleation in the Presence of a Metastable Critical Point", J. Chem. Phys. 109, 223 (1998).

[34] K. G. Soga, J. R. Melrose and R. C. Ball, "Metastable States and the Kinetics of Colloid Phase Separation", J. Chem. Phys. 110, 2280 (1999).

[35] For an account of work at Brandeis, see http://www.elsie.brandeis.edu/.

[36] A. George and W. W. Wilson, Acta Crystallogr. D50, 361 (1994).

6 RECOMMENDATIONS AND SUMMARY

We are impressed with the research opportunities now available due to advances in the fabrication of functional physical nanocomponents and the manipulation and characterization of biological nanocomponents. Continuing research to develop useful interfaces between biological and physical systems can be guided by designing experiments that address theoretical and computational analysis of whole cell systems behavior, and structure-function response of biological molecules.

We recommend a closely linked program of experimental and theoretical work, including development of methods to incorporate the knowledge bases being generated by combinatorial explorations of cell function, protein crystallization, etc. These programs should address DOE's unique focus in biological science on issues such as bioremediation, carbon dioxide sequestration and biomaterials synthesis. To accomplish this requires the exploration of biophysical topics that will not necessarily be addressed in other programs. Specifically, the types of issues that arise will include: developing proteins or protein-like nanocomponents that differ significantly from naturally occurring proteins; and developing modified biological control mechanisms to effect functional responses different from naturally occurring biological systems. Ultimately, one can envision the incorporation of fully integrated non-biological components (e.g. physical nanostructures) in biomimetic systems of designed functionality.

To explore and expand the research landscape that includes these future possibilities we have defined two research challenges on which we believe immediate progress is possible. These challenges are:

1) Fabricate non-trivial assemblies of physical and biological nano-components with linked functionality, and develop carefully designed experiments

to compare measured behavior directly to results of systems modeling.

2) Fabricate assemblies of proteins in which protein-protein, protein-lipid or protein-artificial nanocomponent interactions can be tailored, and test/tune computational capabilities to predict changes in their corresponding function.

As discussed in the text of this report, we explicitly propose in these challenges to shape experimental exploration of issues in advancing cell modeling and computation of biomolecular function. This approach serves the dual purpose of advancing predictive capabilities via computation, while driving the development of new technical capabilities in assembling and exploiting biological and physical nanocomponents.

In the course of preparing this report, we identified many key short-term opportunities and motivators for research at the nano-bio-computational interface. Because these key points are dispersed through the text of the report, we synopsize them here:

- Applications of nano-probes as cellular diagnostics has proven a significant source of innovation in developing techniques for linking physical and biological nanocomponents. A continued emphasis in this area is important both as a driver for continued innovation, and because the improved information available from these diagnostics is essential to improved modeling of cell function. Expanded research emphasis on signal transduction, and the development of sensor feedback mechanisms is needed.

- Both computational and analytical theory is needed to understand the electrostatic interactions between macromolecules and nanosensors in solution.

- Biological (and biomimetic) membranes represent the nano-bio equivalent of a circuit board, but one on which much of the important interac-

tion occurs between components on the membrane (membrance-bound proteins and, potentially, embedded physical structures) and components in the surrounding medium. The development of controlled assemblies of membrane-bound proteins, combined with physical nanostructures provides a clear avenue for developing model nano-bio assemblies.

- The continued development of the chemical linking strategies needed to build hybrid nanostructures in which physical and biological nanocomponents interact in a controlled fashion, also needs a serious research emphasis.

- Cellular modeling tool-kits are increasingly available to front-end users. Validating and establishing standards, both for the numerical tools and in data-bases of parameters, so that non-experts can use these meaningfully and consistently, is an important problem.

- The integration of data from multiple sources to create a meaningful understanding of cell system behavior, protein function, protein crystallization, etc. requires the development of integrative analytical tools that are carefully designed to address both the uncertainties of the data and the input to theoretical description. This is a pressing research issue needed to deal with the effective design and use of high-throughput computational and experimental methods.

- Computation of molecular properties requires a careful match of the computational capabilities to the property of interest. Developing such predictive capabilities for molecular computation of protein functional responses is an important research goal. The development of computational capabilities tuned to meet this goal must be guided by experiments (including high-throughput) in which protein function is assessed in response to changes in environment. The use of nano-assemblies of membrane proteins to create an environment of controlled variability,

and ultimately to incorporate non-biological perturbations such as electrostatic charging, should be used to challenge and guide computation.

A APPENDIX: Multiple Shooting Method

To analyze the output of models of rate equations as described in section 4.3.1, or for example the time series coming from running Virtual Cell (see Appendix B), in order to compare the model with experiment, we need additional tools to those presently in the packages. For example, we may wish to determine, the values of the reaction rate parameters in some model. These models are typically formulated, as differential equations describing the development of the many dimensional state of the system $x(t) = [x_1(t), x_2(t), \cdots, x_N(t)]$ where the $x_a(t)$ are concentrations of various biochemicals, activation variables for voltage gated ion channels, membrance voltages, etc. The differential equations for these reactions have a set of parameters $b = [b_1, b_2 \cdots, b_p]$ that we wish to learn from the data. The nonlinear differential equations

$$\frac{dx(t)}{dt} = F(x(t), b)$$

have solutions which depend on the parameters and the initial conditions. In an experiment one can typically measure only one or a few of the state variables and from that one wishes to determine the b.

Suppose one can measure $V(t) = x_1(t)$, then it is natural to select the parameters b by minimizing the model prediction $x_1(t)$ compared to the observations. Minimize

$$J(b) = \frac{1}{2}\sum_{t=0}^{k-1} |V(t_1) - x_1(t_1, b, x(t=0))|^2$$

with respect to b. The sum is over all observations times $t_1 = t_0 + l\Delta t$.

This familiar procedure has two problems, both serious, and both arising from the intrinsic instabilities of nonlinear systems (often called sensitivity to initial conditions). First, the actual trajectory $x_1(t)$ depends sensitivitely, exponentially sensitively, on $x(t=0)$, and second the function $J(b)$ has many **local minima** and one global minimum. To reach the global minimum,

which is the goal, either requires great luck in selecting starting values for the search in the parameters or sophisticated programs which sometimes can find global minima of functions such as $J(b)$ in P-dimensional space.

A way to avoid this was invented many years ago and codified and analyzed by Bock (1983). The essential idea is that these instabilities develop in time, amplifying any errors in initial conditions $x(t=0)$ say $\Delta x(t=0)$ to $\Delta x(t) = \Delta x(t=0)e^{\lambda t}$; $\lambda > 0$. So if one integrates in small steps δt, such that $e^{\lambda \delta t} < 1$, then the error will be under control.

In detail the method must not only determine the initial conditions at $t = 0$, but must also determine the initial conditions at each integration time interval of length δt. This can be done, and we include below a description of the details. The procedure requires the model output $x_1(t)$ to closely track the observations $V(t)$ and estimates all the other state variables throughout the time interval of observations. The method uses continuity of the state variables in time and the uniqueness of the solutions of the differential equations.

We have already presented one example of rate modeling in section 4.3.1. That example uses minimization of $J(b)$. The following example uses the multiple shooting method described here.

This example is concerned with the particular biochemical reaction where melarsen oxide or MEL binds two cystein residues of a protein. The reaction of interest is of MEL with tri-peptide glutathone (GSH) according to the model

$$\begin{aligned}
\frac{dx_1(t)}{dt} &= -k_1 x_1(t)x_2(t) + k_{-1}x_3(t) \\
\frac{dx_2(t)}{dt} &= -k_1 x_1(t)x_2(t) + k_{-1}x_3(t) - k_2 x_2(t)x_3(t) + k_{-2}x_4(t) \\
\frac{dx_3(t)}{dt} &= k_1 x_1(t)x_2(t) - k_1 x_3(t) - k_2 x_2(t)x_3(t) + k_2 x_4(t) \\
\frac{dx_4(t)}{dt} &= k_2 x_2(t)x_3(t) - k_{-2}x_4(t)
\end{aligned}$$

where $x_1(t)$ is the concentration of MEL, $x_2(t)$, the concentration of GSH, $x_3(t)$ the concentration of the (MEL)(GSH) complex, and $x_4(t)$ the concentration of (MEL)(GSH2) product. The reaction model for this is

$$\begin{aligned} MEL + GSH &\Longleftrightarrow MEL \times GSH \\ MEL \cdot GSH + GSH &\Longleftrightarrow MEL \times GSH2 \end{aligned}$$

and any of the model generation packages described in Appendix B could handle generation of this model quite easily. The experimental data consisted of measurements of many of the concentrations, but their initial values, especially for (MEL) (GSH) and (MEL)(GSH2) were not known and needed to be estimated. Multiple shooting was needed for this because of the instabilities in the nonlinear differential equations describing the reactions. This allowed the accurate determination of the four reaction rate constants.

The message in this example is that to determine the reaction rates required in biochemical processes, one will most likely need tools such as the least squares minimization routines or most likely the multiple shooting methods. Indeed, adopting the latter routinely would allow essentially all the problems described by these system cellular dynamics to be treated accurately.

A.1 Mutiple Shooting Method

Suppose we have a signal $V(t)$ sampled at $t_0, t_1 \cdots t_N$ and we wish to synchronize a D+1 dimensional system (y(t), $z_\alpha(t)$); $\alpha = 1, 2, \cdots D$ to it. The "data" $V(t)$ we imagine comes from a system which we'd like to describe by the D+1 dimensional dynamics on their shared attractor—defined by the motion of $V(t)$.

To start we identify the first component of the system $x_1(t) = y(t)$ as equal to $V(t)$. The other components of the sysem $x(t)$ are $z_\alpha(t); \alpha =$

$1, 2, \cdots D$. We want to integrate forward N times $\Delta t = t_{k+1} - t_k; k = 0, 1, \cdots N-1$ to get in a stepwise fasion from t_0 to t_N.

Since we anticipate that the dynamics may be chaotic, the ability to identify parameters in the vector field for $(y(t), z_\alpha(t))$ will be very tricky if we integrate in one jump from t_0 to t_N. Due to the sensitivity to initial conditions, there may be many minima in a least squares cost function which jumps this large gap. Integrating in small steps feeds information about the external drive into the trajectory in $(y(t), z_\alpha(t))$ space.

The idea is to require the integrated trajectory in $(y(t), z_\alpha(t))$ space to match the values of $V(t_j)$ at each t_j to $y(t_j)$ as an initial condition for integration over the interval of size Δt to arrive at $t_{j+1} = t_j + \Delta t$.

The differential equations for the system are

$$\begin{aligned} \frac{dy(t)}{dt} &= F_y(y(t), z_\alpha(t), b) \\ \frac{dz_\alpha(t)}{dt} &= F_\alpha(y(t), z_\alpha(t), b) \end{aligned}$$

where the b are a set of parameters which we wish to determine. Start at t_0 with initial conditions $y(t_0) = V(t_0)$. This is where the information from the driving force (or data) enters, and $z_\alpha(t_0)$. Integrate from t_0 to $t_0 + \Delta t$. The solutions to the differential equations starting from these initial conditions are written as

$$y(t_0 + \Delta t; z_\alpha(t_0), b)$$

and

$$z_\alpha(t_0 + \Delta t; z_\alpha(t_0), b)$$

$y(t_0+\Delta t; z_\alpha(t_0), b) = V(t_0)$ at $\Delta t = 0$ and $z_\alpha(t_0+\Delta t; z_\alpha(t_0), b) = z_\alpha(t_0)$ at $.\Delta t = 0$

We want to next integrate from t_1 to $t_1 + \Delta t$ starting with initial conditions for this inverval of $y(t_1) = V(t_1)$ and $z_\alpha(t_1)$. This results in the solutions

$$y(t_1 + \Delta t; z_\alpha(t_1), b)$$

and

$$z_\alpha(t_1 + \Delta t; z_\alpha(t_1), b)$$

and so forth until we perform the last integration from t_{N-1} to t_N, giving us

$$y(t_{N-1} + \Delta t; z_\alpha(t_{N-1}), b)$$

and

$$z_\alpha(t_{N-1} + \Delta t; z_\alpha(t_{N-1}), b)$$

These integrations involve the P parameters b and the ND initial conditions

$$z_\alpha(t_0, z_\alpha(t_1), z_\alpha(t_1), z_\alpha(t_2), \cdots, z_\alpha(t_{N-1}).$$

We now impose $N + (N-1)D$ conditions

$$\begin{aligned}
y(t_0 + \Delta t; z_\alpha(t_0), b) &= V(t_1) \\
z_\alpha(t_0 + \Delta t; z_\alpha(t_0), b) &= z_\alpha(t_1) \\
y(t_1 + \Delta t; z_\alpha(t_1), b) &= V(t_2) \\
z_\alpha(t_1 + \Delta t; z_\alpha(t_1), b) &= z_\alpha(t_2)
\end{aligned}$$

$$\begin{aligned}
y(t_{N-2} + \Delta t; z_\alpha(t_{N-2}), b) &= V(t_{N-1}) \\
z_\alpha(t_{N-2} + \Delta t; z_\alpha(t_{N-2}), b) &= z_\alpha(t_{N-1})
\end{aligned}$$

and finally

$$y(t_{N-1} + \Delta t; z_\alpha(t_{N-1}), b) = V(t_N)$$

No condition is placed on the "other" variables $z_\alpha(t)$ at the final integration as we have no information on where those variables go. The tracked variable $y(t)$ is required to match the driving signal $V(t)$ at the end of each stage of the multiple step integration including t_N.

When the number of equality conditions equals the number of parameters plus the number of initial conditions, namely, $N + (N-1)D = P + ND$

or $N = P + D$, we have a well determined system for selecting the P parameters b and the ND initial conditions which allows the computed orbit $y(t)$ and $z_\alpha(t)$ to "track" the input $V(t)$ as accurately as possible.

When $N > P + D$, we have an overdetermined system and can use the pseudo-inverse or equivalently the least squares approximation to the $ND+P$ quantities.

The $N+(N-1)D$ equality constraints are realized in an iterative fashion using Newton's method (or a modern version of that taken from Numerical Recipes, say). The parameters are updated at $t_1 = 0, 1, 2 \cdots, N-1$ from their value at the k^{th} iteration to

$$z_\alpha^{(k+1)}(t_1 = z_\alpha^{(k)}(t_1) + \Delta z_\alpha^{(k)}(t_1)$$

and

$$b_m^{(k+1)} = b_m^{(k)} + \Delta b_m^{(k)}$$

for $m = 1, 2, \cdots, P$.

At $l = 1, 2, \cdots, N$ we have

$$\begin{aligned} &y(t_{l-1} + \Delta t; z_\alpha^{(k)}(t_{l-1}), b_m^{(k)}) - V(t_l) = \\ &-\frac{\partial y(t_{l-1} + \Delta t; z_\alpha^{(k)}(t_{l-1}), b_m^{(k)})}{\partial z_\alpha^{(k)}(t_{l-1})} \cdot \Delta z_\alpha^{(k)}(t_{l-1}) \\ &-\frac{\partial y(t_{l-1} + \Delta t; z_\alpha^{(k)}(t_{l-1}), b^{(k)}m)}{\partial b^{(k)}m} \cdot \Delta b_m^{(k)} \end{aligned}$$

and at $l = 1, 2, \cdots, N-1$ we have

$$\begin{aligned} &z_\alpha(t_{l-1} + \Delta t; z_\alpha^{(k)}(t_{l-1}), b_m^{(k)}) - z_\alpha^{(k)} = \\ &-\frac{\partial z_\alpha(t_{l-1} + \Delta t; z_\alpha^{(k)}(t_{l-1}), b_m^{(k)})}{\partial z_\beta^{(k)}(t_{l-1})} \cdot \Delta z_\beta^{(k)}(t_{l-1}) \\ &-\frac{\partial a_\alpha(t_{l-1} + \Delta t; z_\alpha^{(k)}(t_{l-1}), b^{(k)}m)}{\partial b^{(k)}m} \cdot \Delta b_m^{(k)} \end{aligned}$$

If we collect the parameter changes Δb_m and $\Delta z_\alpha(t_j)$ into one large vector $\Delta\theta_a = (\Delta b_m, \Delta z_\alpha(t_j)); a = 1, 2, \cdots P + ND$, and all the coefficients

into M_{qa} where $q = 1, 2, +(N - 1_D$, the number of conditions, and identify the erors at iteration $kz_\alpha(t_{l-1} + \Delta t; z^{(k)}\alpha(t_{l-1}), b_m^{(k)}) - z_\alpha^{(k)}$ and $y(t_{l-1} + \Delta t; z_\alpha^{(k)}(t_{l-1}), b_m^{(k)}) - V(t_l)$ as a vector H_q, then our update rule is

$$H_q = M_{qa}\, \Delta\theta_a \text{ or } H = M \cdot \Delta\theta$$

with the solution

$$\Delta\theta = (M^T M)^{-1}(M^T H).$$

The update steps make it clear that in addition to solving for $z_\alpha(t_{l-1} + \Delta t; z_\alpha)^{9k}(t_{l-1}), b_m^{(k)})$ and $y(t_{l-1} + \Delta t; z_\alpha^{(k)}(t_{l-1}), b_m^{(k)})$, we will need to integrate equations for

$$\frac{\partial y}{\partial b_k}, \frac{\partial y}{\partial z_\alpha}, \frac{\partial z_\alpha}{\partial b_k}, \text{and} \frac{\partial z_\alpha}{\partial z_\beta}.$$

These equations are directly derivable from the equations for y and z_α.

$$\frac{d\frac{\partial y(t)}{\partial b_k}}{dt} = \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial b_k} + \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial y}\frac{\partial y(t)}{\partial b_k} + \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial z_\beta}\frac{\partial x_\beta(t)}{\partial b_k}$$

$$\frac{d\frac{\partial y(t)}{\partial b_k}}{dt} = \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial y}\frac{\partial y(t)}{\partial z - \alpha(t_l)} + \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial z_\beta}\frac{\partial z_\beta(t)}{\partial z_\alpha(t_l)}$$

$$\frac{d\frac{\partial z_\alpha(t)}{\partial b_k}}{dt} = \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial b_k} + \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial y}\frac{\partial y(t)}{\partial b_k} + \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial z_\beta}\frac{\partial x_\beta(t)}{\partial b_k}$$

$$\frac{d\frac{\partial y(t)}{\partial b_k}}{dt} = \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial a_\alpha(t)}\frac{\partial z_\beta(t_l)}{\partial z - \alpha(t_l} + \frac{\partial F_y(y(t), z_\alpha(t), b)}{\partial z_\beta}\frac{\partial z_\beta(t)}{\partial z_\alpha(t_l)}$$

Natural initial conditions for these are 0 except for $\frac{\partial z_\alpha(t)}{\partial z_\beta(t_l)}|_{t=t_l=\delta_{\alpha\beta}}$.

B APPENDIX: Virtual Cell, E-cell, CellML, and Other "in silico" Cellular Modeling Packages

In the literature and on the web, there are now a number of computational biology toolboxes or engines which permit an easy entre for biologists and others into computing cellular processes. In the web-document "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biological Network Models" (http://www.cds.caltech.edu/erato/sbml/docs/) dated 17 May 2002, there are a variety of sources for computational cellular biology:

Name	Location	Coments
Cellerator	www.aig.jpl.nasa.gov/public/mls/cellerator	Mathematica
DBsolve	Websites.ntl.com/~igor.goryananin/	C++License agreement and permission of author required
E-Cell	e-cell.org	C++ libraries
Gepasi/COPASI	www.gepasi.org/	Has some analysis and optimization methods
Jarnac/Winscamp	www.sys-bio.org/	General numerical program,like Matlab, or XXP, with some biological orientation
ProMoT/DIVA	www.mpi-magdeburg.mpg.de/research/project_a/pro_a4/promot.html	In German "advertisement"for a project
StochSim	www.zoo.cam.ac.uk./comp-cell/StochSim.html	Stochastic simulation of biochemical processes. On "average" gives ODEs
Virtual Cell *	www.nrcam.uchc.edu	Web based computations at remote site. Organized; no analysis
CellMI *	www.cellml.org	Language for unified description and exchange of models

The input to most of these programs is a specification of reactions and parameters, etc which can be translated into mathematical models and used

as parts of differential equations to study the dynamics of the collected reactions. These programs (and announcements) provide a variety of ways for computation both in general (Jarnac) or specifically for cellular biology (Virtual Cell). Their utility depends on the sophistication of the user and, of course, the user's ambition.

We selected Virtual Cell as a subject of study, not because the other programs are not interesting, but because it was the most packaged, graphical, apparently easy to use effort. Furthermore, Virtual Cell is supported by the National Center for Research Resources (NCRR), at the National Institutes of Health (NIH) and thus may be presumed to have some stability as to continuing support. The NRCAM also runs training courses on using the program.

The first tutorial for virtual cell simulates the bleaching of a specified cellular constituent by external light. No details of the process are needed to run the tutorial. One only follows the rules in defining a compartment (interior of a cell or cytoplasm), the constituents in it, and how they react. All this is done graphically and quite neatly. Once the model is defined with its geometry and initial conditions (changeable by the user, with defaults provided), one can run the simulation. The program created by the Virtual Cell interface is hidden from the user (however, the last chapter of the manual discusses how the mathematical framework can be accessed) and is executed on machines at NRCAM. The speed and efficiency of the execution is hard to judge, as one does not know what is being executed, the intensity of use of the network over which one is computing, including other users competing for NRCAM's resources. In addition a large amount of graphics is being prepared by the program. When completed, one is given various options for visualizing the results, including spatial cuts through the data–e.g. density of bleached areas in a slice though the cell, and time series of development of bleaching at a point of the users selection. No indication is given of the meaning of the results or of the changes in the results as a function of initial

conditions, cell geometry, reaction rates, etc.

In our opinion, the flaw with Virtual Cell and the other programs in the list that we were able to examine, is that analysis of the "data" or methods for comparing the data with observations are, on the whole, absent. If one looks at the table, one will see this is not totally true, but in no case was there a systematic discussion of how to compare results with data, to determine the myriad of reaction rates, to establish if the behaviors seen are from a multi-stable system where the attractor is determined by parameters and initial conditions, etc.

There is an exception to this, to a certain extent. In Gepasi the authors are aware of the need for analysis of results and determination of reaction constants and other parameters of the system. In the paper "Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation" Bioinformatics **14 (1)**, 869–883 (1998) the authors address the issue of determining these parameters. The methods they propose are conventional, and often not adequate for the nonlinear problems they pose. The issue is that with nonlinear dynamical systems one typically has very sensitive dependence on initial conditions associated with the intrinsic instabilities in the dynamics itself. In comparing a calculation with observations for purposes of determining parameters of interest, one must also have a way to determine the initial conditions of all dynamical variables, not just those of the observed variables. The multiple shooting methods of Bock (1983) address this and have been described in Appendix A.

B.1 SMBL Consortium

A collection of research groups have banded together under the auspices of John Doyle at Caltech and H. Kitano from Tokyo to encourage a uniform effort in front end languages for creating model for systems cellular biology.

This SBML (Systems Biology Markup Language) consortium identifies the following as the problems they are addressing:

- Users often need to work with complementary resources from multiple simulation/analysis tools in the course of a project. Currently this involves manually reencoding the model in each tool, a time-consuming and error-prone process.

- When simulators are upgraded or no longer supported, models developed in the old systems can become stranded and unusable. This has already happened on a number of occasions, with the resulting loss of usable models to the community. Continued innovation and development of new software tools will only aggravate this problem unless the issue is addressed.

- Models published in peer-reviewed journals are often accompanied by instructions for obtaining the model definitions. However, because each author may use a different modeling environment (and model representation language), such model definitions are often not straightforward to examine, test and reuse.

- Many efforts are underway to construct databases of curated models of biochemical networks. Most of these ventures involve reimplementing published models in a native environment, before reevaluating and annotating the model for future use. Such reimplementation is difficult and error-prone.

The effort to make a uniform language for producing and sharing cell dynamics models within a wide community is precisely in the spirit of successful, and ongoing efforts to make algorithm libraries. The libraries have the advantage, partly because of age and partly because of the user community, of being widely tested and validated. When one used (for instance) a LINPACK routine to perform a **QR** decomposition of a matrix, one could be

rather sure that the operations indicated in the books were being done accurately and efficiently. A concern about the SMBL document quoted above is that the data based models which will be constructed within their agreed-upon framework will be generally assumed to be accurate representations of the cellular dynamics being considered Clearly, some consistent annotation of issues of accuracy and uncertainities must be included in such a data-base to prevent unwarranted reliance on its contents.

The outcome of a successful SBML effort will be a widely adopted common language for formulating cell system models. The output of SBML will be designed to support external software packages which can "read in a model expressed in SBML and translate it into its own internal format for model analysis. For instance, a package might provide differential equations representing the network and then perform numerical integrations on the equations to explore the model's dynamic behavior." Given the variety of possible approaches for solving differential equations, even given the model formulated in a controlled manner, the numerical and graphical output may well differ from investigator to investigator.

A quality control and testing activity added to a universal modeling framework would be a very useful part of the value of a consortium such as the SBML consortium. This is probably a very difficult task for a research oriented organization such as the present SBML group and perhaps should be located at a place such as Argonne or Oak Ridge National Laboratories where some of the excellent mathematical library work of the past has been housed. Assuring model intercomparison on an apples-to-apples basis could well be an important function of these centers, in addition to whatever they may contribute to the cell system research itself. Indeed, this may be a very useful effort for the Department of Energy as part of its overall program in life sciences, in particular its "bio" part of a nanobio program.

DISTRIBUTION LIST

Director of Space and SDI Programs
SAF/AQSC
1060 Air Force Pentagon
Washington, DC 20330-1060

CMDR & Program Executive Officer
U S Army/CSSD-ZA
Strategic Defense Command
PO Box 15280
Arlington, VA 22215-0150

DARPA Library
3701 North Fairfax Drive
Arlington, VA 22203-1714

Assistant Secretary of the Navy
(Research, Development & Acquisition)
1000 Navy Pentagon
Washington, DC 20350-1000

Principal Deputy for Military Application [10]
Defense Programs, DP-12
National Nuclear Security Administration
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585

Superintendent
Code 1424
Attn: Documents Librarian
Naval Postgraduate School
Monterey, CA 93943

DTIC [2]
8725 John Jay Kingman Road
Suite 0944
Fort Belvoir,VA 22060-6218

Strategic Systems Program
Nebraska Avenue Complex
287 Somers Court
Suite 10041
Washington, DC 20393-5446

Headquarters Air Force XON
4A870
1480 Air Force Pentagon
Washington, DC 20330-1480

Defense Threat Reduction Agency
Attn: Dr. Arthur T. Hopkins [12]
6801 Telegraph Road
Alexandria,VA 22310

IC JASON Program [2]
Chief Technical Officer, IC/ITIC
2P0104 NHB
Central Intelligence Agency
Washington, DC 20505-0001

JASON Library [5]
The MITRE Corporation
WA549
7515 Colshire Drive
McLean, VA 22102

U. S. Department of Energy
Chicago Operations Office
Acquisition and Assistance Group
9800 South Cass Avenue
Argonne, IL60439

Dr. Allen Adler
Director
DARPA/TTO
3701 N. Fairfax Drive
Arlington, VA 22203-1714

Dr. Jane Alexander
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5000

Dr. A. Michael Andrews
Director of Technology
SARD-TT
Room 3E480
Research Development Acquisition
103 Army Pentagon
Washington, DC 20310-0103

Dr. William O. Berry
Director, Basic Research
ODUSD(ST/BR)
4015 Wilson Blvd
Suite 209
Arlington, VA 22203

DISTRIBUTION LIST

Dr. Albert Brandenstein
Chief Scientist
Office of Nat'l Drug Control Policy
Executive Office of the President
Washington, DC 20500

Dr. Steve Buchsbaum
DARPA
STO
3701 N. Fairfax Drive
Arlington, VA 22203-1714

Dr. Darrell W. Collier
Chief Scientist
U. S. Army Space & Missile Defense Command
PO Box 15280
Arlington, VA 22215-0280

Dr. James F. Decker
Principal Deputy Director
Office of the Director, SC-1
Room 7B-084
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585

Dr. Patricia M. Dehmer [5]
Associate Director of Science for Basic Energy Sciences, SC-10
Office of Science
U.S. Department of Energy
19901Germantown Road
Germantown, MD 20874

Ms. Shirley Derflinger [15]
Technical Program Specialist
Office of Biological & Environmental Research, SC-70
Office of Science
U.S. Department of Energy
19901Germantown Road
Germantown, MD 20874

Dr. Martin C. Faga
President and Chief Exec Officer
The MITRE Corporation
N640
7515 Colshire Drive
McLean, VA 22102

Mr. Dan Flynn [5]
Program Manager
DI/OTI/SAG
5S49 OHB
Washington, DC 20505

Ms. Nancy Forbes
Senior Analyst
DI/OTI/SAG
5S49 OHB
Washington, DC 20505

Dr. Paris Genalis
Deputy Director
OUSD(A&T)/S&TS/NW
The Pentagon, Room 3D1048
Washington, DC 20301

Mr. Bradley E. Gernand
Institute for Defense Analyses
Technical Information Services
Room 8701
4850 Mark Center Drive
Alexandria, VA 22311-1882

Dr. Lawrence K. Gershwin
NIC/NIO/S&T
2E42, OHB
Washington, DC 20505

General John A. Gordon
National Security Council
National Director for Combating Terrorism and Deputy National Security Advisor
2201 C Street, NW
Washington, DC 20520

Dr. Theodore Hardebeck
STRATCOM/J5B
Offutt AFB, NE68113

Dr. Robert G. Henderson
Director
JASON Program Office
The MITRE Corporation
7515 Colshire Drive
McLean,VA 22102

DISTRIBUTION LIST

Mr. O' Dean P. Judd
Los Alamos National Laboratory
Mailstop F650
Los Alamos, NM 87545

Dr. Bobby R. Junker
Office of Naval Research
Code 31
800 North Quincy Street
Arlington, VA 22217-5660

Dr. Yeongji Kim
IC JASON Program Coordinator
Intelligence Technology Innovation Center
2P0104 NHB
Central Intelligence Agency
Washington, DC 20505-0001

Dr. Anne Matsuura
Army Research Office
4015 Wilson Blvd
Tower 3, Suite 216
Arlington, VA 22203-21939

Dr. Maureen I. Mc Carthy
US DOE
NNSA
1000 Independence Ave, SW
NA-1, Room 7A-199
Washington, DC 20585

Dr. Thomas Meyer
DARPA/ATO
3701 N. Fairfax Drive
Arlington, VA 22203

Dr. Julian C. Nall
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

Dr. C. Edward Oliver [5]
Associate Director of Science for Advanced
Scientific Computing Research, SC-30
U.S. Department of Energy
19901Germantown Road
Germantown, MD 20874

Raymond L. Orbach
Director, Office of Science
U.S. Department of Energy
1000 Independence Avenue, SW
Route Symbol: SC-1
Washington, DC 20585

Dr. Ari Patrinos [5]
Associate Director
Biological and Environmental Research
SC-70
US Department of Energy
19901 Germantown Road
Germantown,MD 20874-1290

Dr. John R. Phillips
Chief Scientist, DST/CS
2P0104 NHB
Central Intelligence Agency
Washington, DC 20505-0001

Records Resource
The MITRE Corporation
Mail Stop W115
7515 Colshire Drive
McLean, VA 22102

Dr. Dan Schuresko
Acting Director
National Security Space Architect
PO Box 222310
Chantilly,VA 20153-2310

Dr. John Schuster
Submarine Warfare Division
Submarine, Security & Tech
Head (N775)
2000 Navy Pentagon, Room 4D534
Washington, DC 20350-2000

Dr. Roanld M. Sega
DDR&E
3030 Defense Pentagon,
Room 3E101
Washington, DC 20301-3030

DISTRIBUTION LIST

Dr. Richard Spinrad
US Naval Observatory
Naval Oceanographers Office
3450 Massachusetts Ave, NW
Building 1
Washington, DC 20392-5421

Dr. Anthony J. Tether
DIRO
DARPA
3701 N. Fairfax Drive
Arlington, VA 22203-1714

Dr. George W. Ullrich [3]
OSD[ODUSD(S&T)]/WS
Director for Weapons Systems
3080 Defense Pentagon
Washington, DC 20301-3080

Dr. Bruce J. West
FAPS
Senior Research Scientist
Army Research Office
P. O. Box 12211
Research Triangle Park,NC 27709-2211

Dr. Linda Zall
Central Intelligence Agency
DS&T/OTS
3Q14, NHB
Washington, DC 20505-0001