Computational challenges in structural and functional genomics

by T. Head-Gordon
J. C. Wooley

The goal of computational biology in the early twenty-first century is to link the various genome sequencing projects to a high-throughput effort in complete structural and functional annotation of whole genomes or biological pathways. It is, in fact, a logical extension of the genome effort to systematically elaborate DNA (deoxyribonucleic acid) sequences into full three-dimensional structures through to functional analysis of cellular networks. The first level of the biological hierarchy is comparative analysis of the rapidly emerging genomic data at the sequence level. However, knowing only the sequence of DNA does not always tell us about the structure or function of the genes, nor does it tell us about the combined action of their protein products, which is the essence of higher order biological function. Complete annotation will include the determination of structure and function of proteins, and a move from analysis of these individual macromolecules to their complex interactions that make up the processes of cellular decisions. This paper represents an effort by a research community to define the hard computational biology problems of the future, to define what mixture of basic research directions and practical algorithmic approaches will be required to achieve our goals, and to outline the directions that will likely be taken in the postgenomic era.

The pace of extraordinary advances in molecular biology has accelerated in the past decade due to discoveries coming from genome projects on human and model organisms. In the next century we can begin to envision the necessary experimental, computational, and theoretical steps necessary to exploit genome sequence information for its medical impact, its contribution to biotechnology, economic competitiveness, and improvements in global environmental quality. Accordingly, a systematic and comprehensive exploration of these sequence data will enable us to shift our view of biology from that of the particular, where it is focused currently, to that of the comprehensive.

The breadth and density of genome data that must be stored, accessed, and annotated immediately suggests that computation will play a central role in the postgenomic phase. As an example, the proposed whole genome "shotgun" strategy at The Institute for Genome Research (TIGR) produced 30 million bases of DNA (deoxyribonucleic acid) per day in January of 1998 that increased to 100 million base pairs per day by midyear. Indeed, the data management issues are and will continue to be important. However, it is the transfer of comprehensive genomic information to the next level of characterization of structure and function that signals a new maturity in biology. The rapid acquisition of quantitative data by experimental techniques (that are themselves being revolutionized) promises to transform much of biology from a descriptive science into a predictive one. The convergence to a quantitative science ensures that these areas of biology are genuinely endowed with computational complexity beyond just data management and organization.

The simultaneous revolutions in genomics and computing will securely establish an unprecedented scientific computing culture in the biological commu-

nity. It will continue to be necessary to step up efforts to train biologists in the quantitative areas of the physical and computer sciences. The possibility that new paradigms in biology might be imaginable when computation becomes firmly entrenched will pose a significant set of new challenges once these "hard" computational biology problems are identified. More importantly, the biology itself will become the central research focus for many scientists with more "traditional" appointments in the physical sciences, mathematics, and computing areas. Perhaps the more difficult task is the adaptation of the physical and computing sciences communities to acquire a deep understanding of biology, because biology in fact will likely dominate science endeavors in the next century, with great mass appeal and public support.

To exploit the inherent genome information derived from knowing the DNA sequences, computational advances, along with related experimental biotechnology, are essential. Knowing the sequence of the DNA does not tell us about the function of the genes, specifically the actions of their protein products—understanding where, when, why, and how the proteins act is the essence of the biological knowledge required. Encoded in the DNA sequence is a protein's three-dimensional topography, which in turn determines function, whereas a protein's function is dependent on the physical state of the cell and other protein interaction partners; uncovering this sequence-structure-function-systems relationship is the core goal of modern structural biology today.

The goal of computational structural and functional genomics of the future is to link the sequencing efforts to a high-throughput program of annotation and modeling of both molecular structures and functional networks. This paper represents a cooperative effort by researchers in government laboratories, universities, and industry (see Acknowledgments section) to outline the current challenges of computational biology. The authors have served both as contributors, by writing pieces, and as editors, by editing contributions from other researchers in the field. There is a dominant emphasis on proteins in this paper, and a more inclusive discussion of nucleic acid modeling can be found elsewhere.[1] In this paper, we identify the computational biology problems of scale, specify the algorithmic issues of the day, and describe the biological computational requirements necessary in order to reach one primary goal: the full integration of genomic scale information needed for understanding the organismal function.

The first step beyond the genome project: High-throughput genome assembly, modeling, and annotation

The first level of the biological hierarchy is a comprehensive genome-based analysis of the rapidly emerging genomic data. With changes in sequencing technology and methods, the rate of acquisition of human and other genome data is ~100 times higher than originally anticipated. Assembling and interpreting these data will require new and emerging levels of coordination and collaboration in the genome research community to develop the necessary computing algorithms and data management and visualization systems.

Annotation—the elucidation and description of biologically relevant features in a sequence—is essential in order for genome data to be useful. The quality of the annotation will have direct impact on the value of the sequence. At a minimum, the data must be annotated to indicate the existence of gene coding regions and control regions. Further annotation activities that add value to a genome include finding simple and complex repeats, characterizing the organization of promoters and gene families, the distribution of guanine-cytosine (G + C) content, and tying together evidence for functional motifs and homologs.

Once the basic structure of genes has been modeled, comparison of new sequences against each other or the existing database is one of the most essential and revealing processes in computational comparative genomics. Such operations relate new sequences to archival sequences that may have meaningful information about patterns in the sequence and its function. Such comparisons are the starting point for the computation of phylogenetic (evolutionary) trees of organisms or genes, pathogenicity studies for public health, polymorphism studies (e.g., of genetic defects), identification of protein motifs, model identification for gene recognition, model identification for organism classification, functional analysis of genomic/protein sequences, and exon identification.

The analyses and inferencing often depend on the quality of the computed multiple sequence alignments (MSAs) used as input. MSAs of biological sequences, e.g., DNA, RNA (ribonucleic acid), or protein sequences, entail the arrangement of many (in some cases thousands) of sequences, so that corresponding positions are aligned in vertical columns, with padding characters (nulls) added to compen-

sate for length variations in some sequences. The most accurate and sensitive alignments must consider gaps in the alignment (insertions and deletions) and are thus rather computationally intensive. The standard algorithm for this is Smith-Waterman,[2] which uses dynamic programming to produce a local optimal alignment between two sequences of length M and N, and scales as $\mathcal{O}(M \times N)$. The simple extension of these algorithms to multiple sequence alignments of K sequences requires time $\mathcal{O}(N^K)$. For sequence lengths in the thousands of nucleotides, this is barely feasible for three sequences, certainly not for thousands of sequences. Hence, common practice is to use "progressive alignments," an inefficient algorithm that adds one sequence at a time to the MSA. This is computationally tractable, but not optimal. It is especially problematic when the sequences are not closely related, e.g., in computing the Tree of Life.

Recently rediscovered hidden Markov models (HMMs)[3,4] and stochastic context-free grammars (SCFGs)[5] offer the prospect of better MSAs, by also modeling higher order structures. The simplest are HMMs, which are stochastic regular grammars. SCFGs are more complex, but permit one to model nested structures, such as the stem and loop structures common in RNA. More elaborate types of grammars permit the modeling of more complex secondary and tertiary structures. To use these models, one must first estimate the many parameters of the model. The resulting model can then be used to "parse" the sequences, and the resulting parses can be transformed into multiple sequence alignments. Iterative estimation of the HMMs entails iterative solution of computations akin to the pair-wise dynamic program sequence comparison computations. At each iteration we must perform M such $\mathcal{O}(N^2)$ computations, one for each of the M sequences being aligned, and sum the results.

Independent computation for each sequence offers a clear target for parallel computation, followed by a logarithmic summation computation. This is particularly true for large sequence collections such as the ribosomal RNA alignments. Some researchers have constructed fine-grained parallel systolic algorithms for the dynamic programming computations, on specialized hardware implementations or single-instruction/multiple-data machines. However, on multiple-instruction/multiple-data machines (with greater costs for interprocessor communication and synchronization), coarser partitioning of the dynamic programming computations appears preferable. Furthermore, these iterative computations often find local optima, requiring multiple computations with different starting states to find the (putative) global optimum.

One difficulty in model estimation for methods like HMM arises from the possibility of over-fitting the very large number of parameters in these models (several per sequence position). Bayesian methods have been adopted to smooth these parameter estimates. Bayesian methods have traditionally been difficult to compute.[6] Several researchers have resorted to Gibbs sampling methods to estimate the posterior probability distribution.[7] These methods entail the construction and simulation of a Markov chain whose equilibrium probability distribution is equal to the target posterior distribution. The Gibbs sampling computations should be amenable to parallelization, assuming that independent parallel random number generators are available. This is a subject of research activity in the Monte Carlo computation community,[8] and code implementations are available from several research groups.

The goal of fold assignment and comparative modeling is to assign, using computational methods, each new genome sequence to the known protein fold or structure that it most closely resembles.

From genome annotation to protein folds: Comparative modeling and fold assignment

The key to understanding the inner workings of cells is to learn the three-dimensional (3-D) atomic structures of the proteins that form their architecture and carry out their metabolism. These three-dimensional structures are encoded in the blueprint of the DNA genome. Within cells, the DNA blueprint is translated into protein structures through exquisitely complex machinery—itself composed of proteins. The experimental process of deciphering the atomic structures of the majority of cellular proteins is expected to take a century at the present rate of work. New developments in comparative modeling and fold recognition will short-circuit this process, that is, we can learn to translate the DNA message by computer.

The goal of fold assignment and comparative modeling is to assign each new genome sequence to the known protein fold or structure that it most closely resembles, using computational methods. Fold assignment and comparative modeling techniques can then be helpful in proposing and testing hypotheses in molecular biology, such as in inferring biological function, predicting the location and properties of ligand binding sites, in designing drugs, and testing remote protein-protein relationships. It can also provide starting models in X-ray crystallography and NMR (nuclear magnetic resonance) spectroscopy.

The success of these methods rests on a fundamental experimental discovery of structural biology: the 3-D structures of proteins have been better conserved during evolution than their genome sequences. When the similarity of a target sequence to another sequence with known structure is above a certain threshold, comparative modeling methods can often provide quantitatively accurate protein structure predictions, since a small change in the protein sequence usually results in a small change in its 3-D structure. Even when the percentage identity of a target sequence falls below this level, then at least information about the overall fold topology can often be predicted. In cases where sequence identity dips below ~25 percent, the so-called "twilight zone," fold assignment algorithms can often be successful in determining the fold class for a new sequence (see Case Study 1).

Several fundamental issues remain to amplify the effectiveness of fold assignment and comparative modeling. A primary issue in fold assignment is the determination of better multipositional compatibility or scoring, functions that will extend fold assignment further into the twilight zone of sequence homology. In both fold assignment and comparative modeling, better alignment algorithms that deal with multipositional compatibility functions are needed. A move toward detailed empirical energy functions and increasingly sophisticated optimization approaches in comparative modeling will also occur in the future.

Fold assignment or threading. Recent work on fold assignment (or "threading") involves two main approaches: developing potentials for fold assignment[9–12] and hidden Markov models (HMMs) or profile methods that are descended from sequence alignment methods.[13,14] The potentials can be contact potentials (potentials of mean force) or they can be more complex semiempirical potentials, involving atomic areas and other properties. Fold assignment approaches further subdivide into two categories: (1) unipositional methods that consider probability distributions of amino acids at single sites and (2) those that consider distributions on pairs (or even triples) of amino acids within a contact distance in a given structure. Hidden Markov models to date have considered single site probability distributions. We discuss these approaches next.

Contact potentials for threading. Unipositional fold assignment approaches score each residue position in a template structure using local 3-D environmental information such as secondary structure propensity, degree of environmental polarity, and the fraction of the residue surface buried and inaccessible to solvent. The 3-D environmental information for each residue then becomes a one-dimensional profile of the tertiary structure or fold, and the compatibility of the 20 common amino acids is evaluated for each position in the 1-D profile.[10] Optimal 1-D alignments of a probe sequence to a given structure can be determined by dynamic programming,[15,16] and the subsequent score of the aligned sequence against the template is determined by the global characteristics of the sequence-environment fit, thereby tolerating locally poor scores.

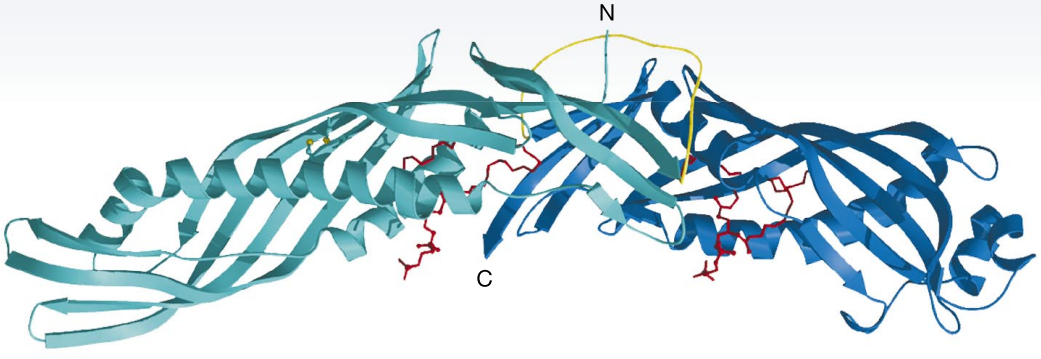
Unipositional methods demonstrated impressive ability for determining similar topological folds for proteins with less than 25 percent sequence identity in some cases. However, only 25 percent of genome sequences recognize their 3-D-protein fold with a sufficient threshold of confidence to be considered a successful fold assignment. One reason for this modest success rate of threading is that the repertoire of folds is thought to be incomplete. This repertoire (or library) is growing modestly through the current efforts of structural biologists, and a strong Structural Genomics Initiative will give a major boost to fold assignment in the future. It has been estimated that the success rate of fold assignment algorithms will increase to roughly 50 percent once these missing folds are identified structurally. For the remaining 50 percent of genome sequences to be assigned to folds, there must be advances in the directions discussed here.

The first advance is to move to multipositional compatibility functions. Pair-wise threading potentials typically consider the propensity of two amino acids to be within a specified distance using a score function compiled from a database of structures. Additional features can be used in addition to the identity of the amino acids, such as the secondary

Case Study 1

Assigning genome sequences to a known 3-D protein structure

Bactericidal permeability-increasing protein (BPI, see figure) from human white blood cells is a potent antimicrobial protein of 456 amino acid residues. Its structure, determined by X-ray crystallography, was found to be a new fold: an elongated, boomerang-shaped molecule, unlike any previously known structure.[1] Prior to the publication of the 3-D structure of BPI, its amino acid sequence was submitted to the second meeting on Critical Assessment of Protein Structure Prediction methods (CASP2). Several methods of fold assignment correctly concluded that the BPI sequence was incompatible with any protein fold then in the database of known protein structures. With the 3-D structure of BPI available, it became possible to search databases of genome sequences to learn which other protein sequences are compatible with the BPI structure. In other words, it was then possible to assign other sequences to the BPI fold.[2] This search uncovered 13 distant relatives of BPI in a diverse set of eurkaryotes, including rat, chicken, worm, and biomphalaria galbrata. The 13 new proteins share only 13–19 percent sequence identity with BPI, below the "twilight zone" of marginal identification by sequence comparison methods. The significance of this case study is that advanced computational methods can assign numerous genome sequences to the 3-D structures by methods of fold assignment, short-circuiting the laborious experimental determination of 3-D structures.



A ribbon diagram of human BPI. The N-terminal domain is aqua, and the C-terminal domain is blue. A proline-rich linker of residues 230–250, which connects the two domains, is shown in yellow. The highly conserved disulfide bonds between Cys 135 and Cys 175 are shown as ball-and-stick atoms.

David Eisenberg
University of California, Los Angeles

References

1. L. J. Beamer, S. F. Carroll, and D. Eisenberg, "The Crystal Structure of Human BPI and Two Bound Phospholipids at 2.4 Å Resolution," *Science* **276**, 1861–1864 (1997).

2. L. J. Beamer, D. Fischer, and D. Eisenberg, "Detecting Distant Relatives of Mammalian LPS Binding and Lipid Transport Proteins," *Protein Science* **7**, 1643–1646 (1998).

structure type, relative exposure to water, relative position, and local atomic density.

Some attempts have been made to go beyond pairwise potentials, but determination of higher order probability distributions is limited by the data available from the present number of structures. New structures made available from the Structural Genomics Initiative would provide some assistance in this regard; however, the number of new structures is unlikely to dramatically increase the order of probability distributions that can be reliably estimated. Therefore, further improvements in pairwise and other potentials of mean force will rest on better identification of the relevant physical effects determining the relation of sequence to structure, and on improved algorithms to extract information about these effects from limited data.

Hidden Markov models. Hidden Markov models consider single site probability distributions for amino acids, but have the added feature of a Markovian transition matrix between "hidden" states.[3] The hidden states effectively perform a choice among a set of position-dependent amino acid probability distributions. In contrast to threading methods, HMMs do not use an explicit scoring function to score the match of an amino acid with its environment, nor do they typically consider pair-wise interactions. HMMs rely heavily on position-specific scoring functions that, in combination with the hidden Markov states, match appropriate probability distributions to sequence positions. Prior knowledge about amino acid probability distributions can be incorporated in a Bayesian framework for HMMs using "Dirichlet prior" probability distributions.[17]

HMMs can be used for fold identification by performing a standard sequence-based homology search using the probe sequence to generate homologous sequences. These sequences can be used to construct an HMM based on the probe, and then the sequences from a library of folds can be matched against the HMM. Similarly, one can construct separate HMMs for each member of a library of folds, and then score the probe sequence against each model. Construction of HMMs is typically an iterative process involving successive periods of model building, searching with the given model, and model refinement. Alignment to an HMM can be performed in an efficient recursive manner, similar to dynamic programming. The results of a typical HMM are illustrated in Case Study 2.

Methods for comparative modeling. Comparative modeling remains the only method at present that can provide models with an RMS (root-mean-square) error lower than 2 Å (angstroms).[18–22] All current comparative modeling methods consist of four sequential steps. The first step is to identify the proteins with known 3-D structures that are related to the target sequence. The second step is to align them with the target sequence and to pick those known structures that will be used as templates. The third step is to build the model for the target sequence given its alignment with the template structures. In the fourth step, the model is evaluated using a variety of criteria. If necessary, the alignment and model building are repeated until a satisfactory model is obtained. The main difference between the different comparative modeling methods is in how the 3-D model is calculated from a given alignment (step three above).

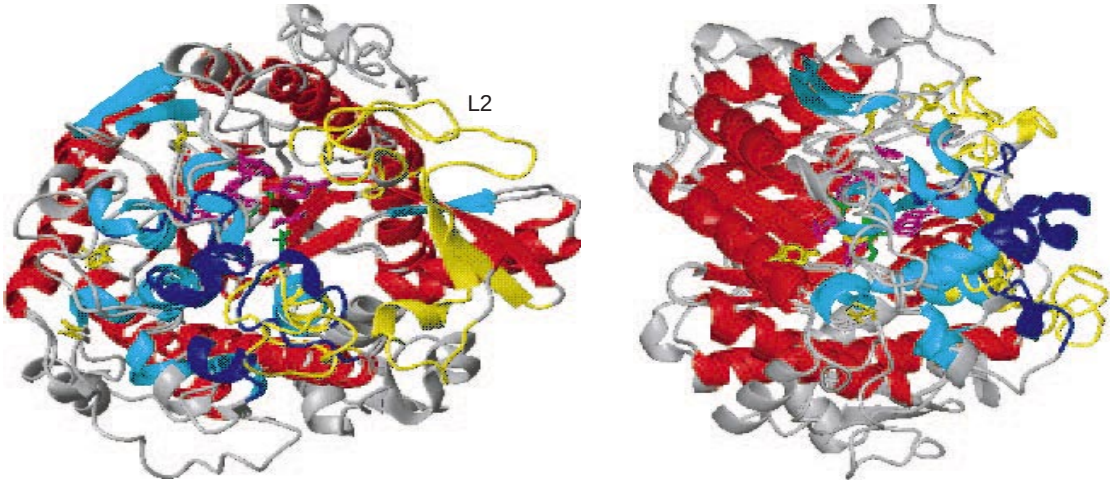
The original and still widely used method is modeling by rigid body assembly. The method constructs the model from a few core regions, and loops and side chains, which are obtained from dissected related structures. This assembly involves fitting the rigid bodies on the framework, which is defined as the average of the C_α atoms in the conserved regions of the fold. Another family of methods, modeling by segment matching, relies on approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms. This is achieved by the use of a database of short segments of protein structure, energy or geometry rules, or some combination of these criteria. The third group of methods, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure. In addition to the methods for modeling the whole fold, numerous other techniques for predicting loops and side chains on a given backbone have also been described. These methods can often be used in combination with each other and with comparative modeling techniques.

Perhaps the most promising comparative model-building technique is the comparative modeling by satisfaction of spatial restraints. The reason is that this approach is based only on optimization of an objective function, and it thus allows an efficient exploration of various representations of protein structure, methods of optimization, and objective function forms. The computational complexity of this

Case Study 2

Sequence comparisons against model protein families to understand human pathology

Hidden Markov model (HMM) -based search methods have been shown to improve both the sensitivity and selectivity of database searches by employing position-dependent scores to characterize and build a model for an entire family of sequences. HMMs have been used to analyze proteins using two complementary strategies. In the first, a sequence is used to search a collection of protein families, such as Pfam, to find which of the families it matches. In the second approach, an HMM for a family is used to search a primary sequence database to identify additional members of the family. The latter approach has yielded insights into protein involved in both normal and abnormal human pathology, such as Fanconi Anaemia A, Gaucher disease, Krabbe disease, polymyositis scleroderma, and disaccharide intolerance II. HMM-based analysis of the Werner syndrome protein sequence (WRN) suggested it possessed exonuclease activity, and subsequent experiments confirmed the prediction.[1] Like WRN, mutation of the protein encoded by the Klotho gene leads to a syndrome with features resembling aging. However, Klotho is predicted to be a member of the family 1 glycosidase (see figure). Eventually, large-scale sequence comparisons against HMMs for protein families will require enormous computational resources to find these sequence-function correlations over genome-scale size databases.



The similarities and differences between two plant and archeal members of a family of glycosidases that includes a protein implicated in aging. Ribbons correspond to the beta-strands and alpha-helices of the underlying TIM barrel (red) and family 1 glycosidase domain (cyan). Amino acid side chains drawn in magenta, yellow, and green are important for structure and/or function. The loop in yellow denotes a region proposed to be important for substrate recognition.

Reprinted with permission from S. Mian, "Sequence, Structural, Functional, and Phylogenetic Analyses of Three Glycosidase Families," *Blood Cells, Molecules and Diseases* **24**, No. 2, 83–100 (June 1998).

Saira Mian
Lawrence Berkeley National Laboratory

Reference

1. S. Huang, B. Li, S. Mian, M. D. Gray, J. Oshima, and J. Campisi, "The Premature Aging Syndrome Protein WRN Is a 3' to 5' Exonuclease," *Nature Genetics* **20**, 114–116 (1998).

approach is directly tied to methods such as global optimization, described later in this paper. This flexibility is essential for improving comparative protein modeling. It will also facilitate simultaneous use of different sources of information when calculating a model of a given protein. For example, a model may be constructed that is consistent with the template structures, potentials of mean force, NMR restraints, cross-linking experiments, site-directed mutagenesis data, etc.

The best comparative techniques can generally produce models with good stereochemistry and overall structural accuracy that is slightly higher than the similarity between the template and the actual target structures, when the modeling alignment is correct. The errors in comparative models can be divided into five categories: (1) side-chain packing errors, (2) distortions and rigid body changes in regions that are aligned correctly (e.g., loops, helices), (3) distortions and rigid body changes in insertions (e.g., loops), (4) distortions in incorrectly aligned regions (loops and longer segments with low sequence identity to the templates), and (5) incorrect fold resulting from an incorrect choice of a template.

The combined consequence of these errors is that the comparative method can result in models with a main-chain RMS error as low as 1 Å for 90 percent of the main-chain residues, if a sequence is at least 40 percent identical to one or more of the templates. In this range of sequence similarity, the alignment is mostly straightforward to construct, there are not many gaps, and structural differences between the proteins are usually limited to loops and side chains. When sequence identity is between 30 and 40 percent, the structural differences become larger, and the gaps in the alignment are more frequent and longer. As a result, the main-chain RMS error rises to ~1.5 Å for about 80 percent of residues. The rest of the residues are modeled with large errors because the methods generally cannot model structural distortions and rigid body shifts, and they cannot recover from misalignments. Insertions longer than about eight residues usually cannot be modeled accurately at this time. Model evaluation methods are frequently successful in identifying the inaccurately modeled regions of a protein. To put the errors into perspective, we list the differences among experimentally determined structures of the same protein: the 1.0Å accuracy of main-chain atom positions corresponds to X-ray structures defined at a low resolution of about 2.5 Å and with an R factor of about 25 percent, as well as to medium-resolution NMR structures determined from 10 interproton distance restraints per residue.

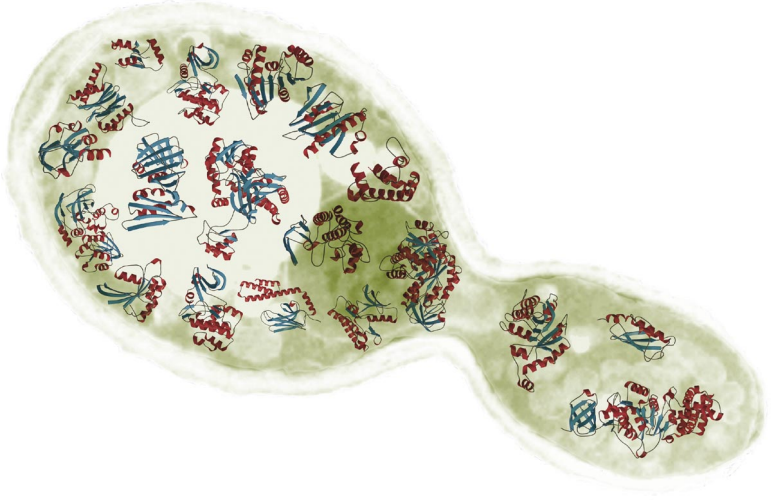
Future improvements of comparative modeling should aim to (1) model proteins with lower similarities to known structures (e.g., less than 30 percent sequence identity), (2) increase the accuracy of the models, and (3) make modeling fully automated (see Case Study 3). The improvements are likely to include simultaneous optimization of side-chain and backbone conformations in side-chain modeling, simultaneous optimization of a loop and its environment in loop modeling, and simultaneous optimization of the alignment and the model. At the same time, better potential functions and possibly better optimizers are needed. The potential function should guide the model away from the templates in the direction toward the correct structure. An addition of atomic or residue-based potentials of mean force to the homology-derived scoring could be one way of achieving this goal. This is a difficult problem, as illustrated by the fact that no present force field or potential of mean force can produce a model with a main-chain RMSD (root-mean-square deviation) from the X-ray structure smaller than about 1 Å, even when the starting conformation is the X-ray structure itself. For example, molecular dynamics simulations in solvent generally have a main-chain RMSD of more than 1 Å, and the most detailed lattice folding simulations result in models with an RMS error larger than 2 Å. Since most of the main-chain atoms in two homologs with at least 40 percent sequence identity usually superpose with an RMSD of about 1 Å, it is currently better to aim to reproduce the template structures as closely as possible rather than to venture away from the templates in the search for a better model.

The major factor that limits the use of comparative modeling in the cases of less than 30 percent sequence identity is the alignment problem, as discussed in the fold recognition problem. In principle, the alignment can be derived by any of the sequence or sequence/structure alignment methods, but in practice even careful manual editing frequently results in significant alignment errors. At 30 percent sequence identity, the fraction of incorrectly aligned residues is about 20 percent and this number rises sharply with further decrease in sequence similarity. This limits the usefulness of comparative modeling, because no current modeling technique can recover from an incorrect input alignment. It would appear that fold recognition methods are a natural solution to the alignment problem in comparative

Case Study 3

Large-scale comparative modeling of protein structures of the yeast genome

Recently, a large-scale comparative protein structure modeling of the yeast genome was performed.[1] Fold assignment, comparative protein structure modeling, and model evaluation were completely automated. As an illustration, the method was applied to the proteins in the Saccharomyces cerevisiae (baker's yeast) genome. It resulted in all-atom 3-D models for substantial segments of 1071 (17 percent) of the yeast proteins, only 40 of which have had their 3-D structure determined experimentally. Of the 1071 modeled yeast proteins, 236 were related clearly to a protein of known structure for the first time; 41 of these have not been previously characterized at all. Many of the models are sufficiently accurate to facilitate interpretation of the existing functional data as well as to aid in the construction of mutants and chimeric proteins for testing new functional hypotheses. This study shows that comparative modeling efficiently increases the value of sequence information from the genome projects, although it is not yet possible to model all proteins with useful accuracy. The main bottlenecks are the absence of structurally defined members in many protein families and the difficulties in detection of weak similarities, both for fold recognition and sequence-structure alignment. However, while only 400 out of a few thousand domain folds are known, the structure of most globular folds is likely to be determined in less than ten years. Thus, comparative modeling will conceivably be applicable to most of the globular protein domains close to the completion of the Human Genome Project.



Large-scale protein structure modeling. A small sample of the 1100 comparative models calculated for the proteins in the yeast genome is displayed over an image of a yeast cell.

Reprinted with permission from the following editorial, "Arise, Go Forth, and Solve Structures," *Nature Structural Biology* **5**, No. 12, 1019–1020 (1998).

Andre Sali
Rockefeller University

Reference

1. R. Sanchez and A. Sali, "Large-Scale Protein Structure Modeling of the Saccharomyces Cerevisiae Genome," *Proceedings of the National Academy of Sciences (USA)* **95** (1998), pp. 13597–13602.

modeling. However, while these techniques are successful in identifying related folds, they appear to be somewhat less successful in generating correct alignments, although improvements in alignment for fold recognition is a goal of future work. To reduce the errors in the model stemming from the alignment errors, iterative changes in the alignment during the calculation of the model are needed. Provided the objective function is capable of distinguishing a good model from a bad one, the iterative realignment and reselection of templates will minimize the effect of errors in the initial alignment and selection of templates.

Finding the best alignment. Even if perfect fold assignment can be achieved, there remains a computational bottleneck in providing a predicted 3-D structure that must begin with proper alignment of the sequence to the structure. As discussed, significant error is present in comparative modeling unless a successful alignment onto a target is realized.

Alignment with unipositional compatibility functions, which adds the independent contributions of a single position of a test sequence to a single position in the target fold, offers the advantage of using well-established dynamic-programming algorithms to find the optimal alignment, although poor gap and insertion penalty parameters can render this optimum somewhat arbitrary. HMMs offer effective position-dependent insertion/deletion penalties as well as an efficient alignment procedure, but ignore more than single site probabilities, as do other unipositional compatibility functions.

Alignment of a genome sequence to 3-D structure using pair-wise potentials is more difficult than using unipositional potentials. Branch and bound algorithms have been shown to yield the optimal alignment when they converge,[23,24] but since the general threading problem for multipositional potentials is NP-complete,[25] branch and bound algorithms will not converge in all cases. Nevertheless, they are often extremely useful. Approximations can also be employed, such as the frozen approximation,[26] in which one assumes that the interaction of test sequence position j with the amino acid k' of the target structure would be similar to k in the sequence. Once the sequence is optimally aligned using the frozen approximation, the multipositional compatibility function is used to score the sequence-structure match. Allowing only a limited number of gaps between secondary structure elements, and exhaustively enumerating all possible resulting threadings, has been implemented for multipositional compatibility functions and has been successful for a subset of interesting cases.

Low-resolution folds to structures with biochemical relevance: Toward accurate structure, dynamics, and thermodynamics

As we move to an era of genetic information at the level of complete genomes, classifying the fold topology of each sequence in the genome is a vital first step toward understanding gene function. However, the ultimate limitation in fold recognition is that these algorithms only provide "low-resolution" structures. It is crucial to enhance and develop methods that permit a quantitative description of protein structure, dynamics, and thermodynamics, in order to relate specific sequence changes to structural changes, and structural changes to associated functional/phenotypic change.

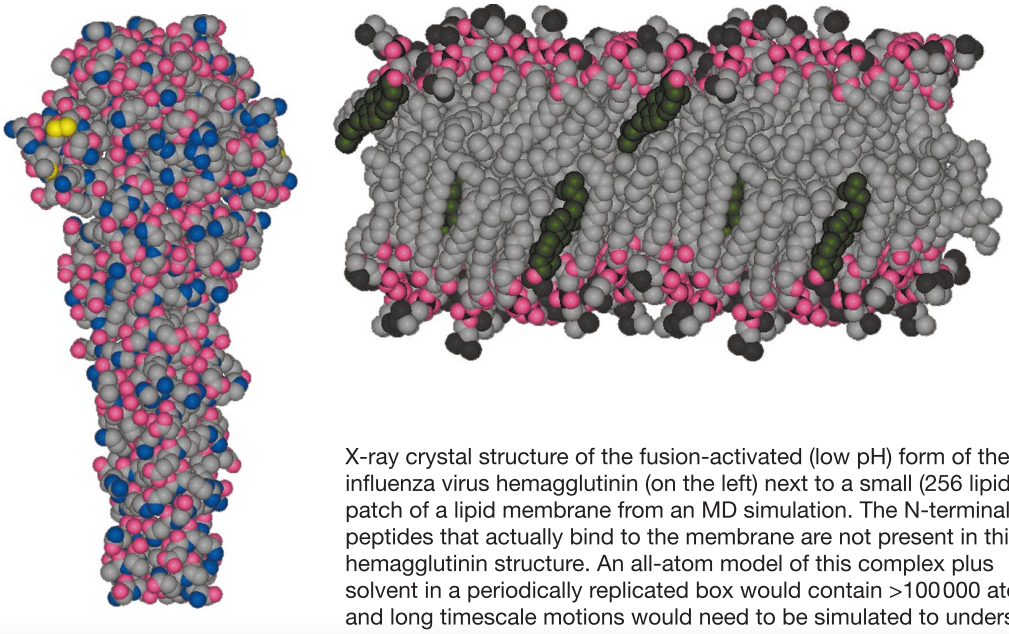
These more accurate approaches will greatly improve our ability to modify proteins for novel uses such as to change the catalytic specificity of enzymes and have them degrade harmful waste products. While often the tertiary fold of proteins involved in disease change dramatically upon mutation, those whose fold remains invariant may have quantitative differences in structure that can have important macroscopic effects on function that can be manifested as disease. More accurate screening for new drug targets that bind tightly to specific protein receptors for inhibition will require quantitative modeling of protein/drug interactions. Therefore, the next step is the quantitative determination of protein structure starting from the fold prediction, and ultimately directly from sequence. Bottlenecks in time- and size-scales are the primary difficulty in predicting and folding protein structure[27–29] and simulating protein function and thermodynamics (see Case Study 4), and we discuss these issues in this section.

Empirical force fields. The translation of protein sequence to protein structure rests upon a central dogma of biology: *proteins adopt their lowest free energy conformation as their functional state*. Thus, key to the success in any computational method that aims to provide sequence to structure predictions, or refinements, is that the energy function model used to represent the biological system yields the functional structure of known proteins as its lowest free energy state. Empirical protein force fields, which have formed the major component of all computational studies of protein structure, function, and dynamics

Case Study 4

Simulation of the molecular mechanism of membrane binding of viral envelope proteins

The first step of infection by enveloped virus proteins involves the attachment of a viral envelope glycoprotein to the membrane of the host cell. This attachment leads to fusion of viral and host cell membranes, and subsequent deposition of viral genetic material into the cell. It is known that the envelope protein, hemagglutinin, exists under normal conditions in a "retracted" state where the peptide that actually binds to the sialylated cell surface receptor is buried some 100 Å from the distal tip of the protein, and is therefore not capable of binding to the cell membrane. However, at low pH a major conformational change occurs whereby the fusion peptide is delivered (~100 Å!) via a "spring-loaded mechanism" to the distal tip where it is available for binding to the cell membrane. As many aspects of this process are likely prototypical of a class of enveloped viruses including HIV, a detailed understanding of the molecular mechanism would be beneficial in guiding efforts to intervene before the viral infection is completed. Atomistic simulations of the process of viral binding to cell membranes are extremely demanding computationally.[1] Taken together, the hemagglutinin, lipid membrane, and sufficient water molecules to solvate the system in a periodically replicable box add up to more than 100 000 atoms, an order of magnitude larger than what is presently considered a large biomolecular system. Furthermore, the large-scale conformational changes in the protein, and displacements of lipids as the protein binds the membrane, involve severe timescale bottlenecks as well.



X-ray crystal structure of the fusion-activated (low pH) form of the influenza virus hemagglutinin (on the left) next to a small (256 lipids) patch of a lipid membrane from an MD simulation. The N-terminal peptides that actually bind to the membrane are not present in this hemagglutinin structure. An all-atom model of this complex plus solvent in a periodically replicated box would contain >100 000 atoms, and long timescale motions would need to be simulated to understand this step in the mechanism of viral infection.

Doug Tobias
University of California, Irvine

Reference

1. D. J. Tobias, K. C. Tu, and M. L. Klein, "Atomic-Scale Molecular Dynamics Simulations of Lipid Membranes," *Current Opinion in Colloid and Interface Science* **2**, 15–26 (1997).

to date, give encouraging results in this regard. However, this conclusion is only qualitatively true for a handful of known proteins, and we need to explore the various ways in which protein surfaces can be modeled according to increasing levels of sophistication depending on the quantitative need.

Empirical protein force fields represent bonds and angles as harmonic distortions, dihedrals by a truncated Fourier series, and pair-wise nonbonded interactions via Lennard-Jones 6–12 terms and Coulomb's Law for electrostatic interactions between point charges. V_{MM} denotes this empirical function and is usually given as:

$$
\begin{aligned}
V_{MM} = & \sum_{i}^{\#\ Bonds} k_b(b_i - b_o)^2 + \sum_{i}^{\#\ Angles} k_\theta(\theta_i - \theta_o)^2 \\
& + \sum_{i}^{\#\ Impropers} k_\tau(\tau_i - \tau_o)^2 \\
& + \sum_{i}^{\#\ dihedrals} k_\phi[1 + \cos(n\phi + \delta)] \\
& + \sum_{i}^{\#\ atoms} \sum_{i<j}^{\#\ atoms} \left\{ \frac{q_i q_j}{r_{ij}} + \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} \\
& + \sum_{i}^{\#\ atoms} \Delta\sigma A \qquad (1)
\end{aligned}
$$

There are several protein force fields of this type in use,[30–32] and it is clear that they are improving in their ability to represent protein conformations near the native fold, but they have not been fully tested outside of this local region on the energy surface. If water is included, and long-range electrostatic effects with methods such as Particle Mesh Ewald[33,34] are used, a simulation will conserve the protein backbone to within 1–2Å RMS of the crystal or NMR structure.[35] This is roughly in experimental error since solution NMR and X-ray crystal structures often differ in backbone RMS by about 1 Å.[36] Some testing indicates that they do not always perform well outside of this local region, and therefore their usefulness in protein structure prediction and folding, which requires a good nonlocal description of the surface, is uncertain.[37]

Beyond the empirical force fields for proteins is the problem of describing a solvent environment and its influence on the protein's conformational behavior. The importance of hydration as a major contributor to protein stability and driving force for folding is widely accepted; in particular it is the hydrophobic interaction that is thought to be dominant, as has been originally pointed out by Kauzmann.[38] Simple models of hydration have been added to empirical protein force fields to attempt a better balance between computational cost and accuracy.[39,40] One functional form of a simple model is to use a generalized Poisson-Boltzmann treatment for the electrostatics, and to include a solvent-accessible surface area term to describe the free energy attributable to the hydrophobic effect.[41,42]

Development of a new set of implicit water potentials for biomolecular simulations is also an important direction.[43] The purpose is to incorporate the statistical properties of water into solute-solute interactions and thereby avoid the computational limitations of simulating the polar solvent, which can be a major obstacle to biomolecular simulations in some cases. Success in developing the implicit solute-solute potentials should lead to future peptide and protein simulations without explicit simulation of the water molecules, with their devastating spatial and temporal scales. These implicit potentials can be fit to a convenient functional form and used in simulations of proteins to describe the important structural influence of aqueous hydration on protein conformations. They are physically complex but can be evaluated with reasonable computational cost, and they are commensurate with both folding studies and protein structure prediction approaches using optimization, since they constitute a well-defined continuous force field, unlike the simpler descriptions above.

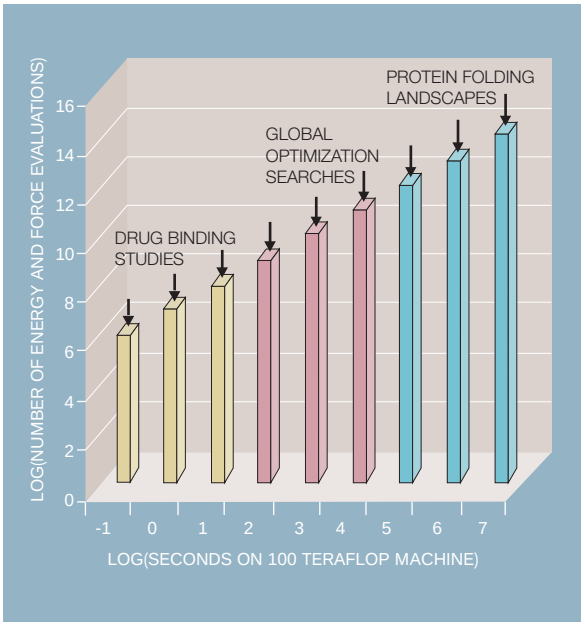
Explicit inclusion of molecular water is the most unambiguous way to describe a solvent environment around a protein, and has in fact been used in many molecular dynamics simulations. Empirical water force fields have been developed for the neat liquid over the last several decades,[44–48] and more recently including explicit polarization,[49–53] and some do quite a reasonable job of reproducing a large number of molecular properties, such as the partial radial distribution functions, thermodynamic, and transport properties.[45] While further improvement in the interface between water and protein force fields is warranted, explicit solvent calculations are important for quantitative studies of interaction of water with the protein.

A fundamental difficulty encountered in the simulation of biomolecular systems is the need to evaluate long-range Coulombic forces, the first term in the double sum in Equation 1. The conventional evaluation of the Coulombic energy and Cartesian derivatives at a given protein configuration requires on the order of N^2 FLOPS (floating-point operations per second), where N is the number of atom centers. The proper accounting of long-range forces is introduced through the Ewald summation.[33,34,54,55] Typical protein in water simulations periodically replicate the system in three spatial dimensions, and divide the long-range Coulombic interactions into a short-range part that is evaluated in real space (as a direct sum over atomic positions) and a long-range part evaluated in reciprocal space. Smooth Particle Mesh Ewald (SPME) employs a Cardinal B-spline-based approximation to the atomic charge density, which can be generated using Fast Fourier Transforms (FFTs), to calculate the Coulombic forces in reciprocal space in order $N\log N$.[34]

For system sizes beyond 10^3 atoms, these algorithms have largely reached their crossover to $N\log N$ scaling. Figure 1 gives an estimate of the number of updates of Equation 1 and its Cartesian derivatives that can be accomplished with $N\log N$ scaling algorithms on a 100 teraflop computer for a 10000 atom system (equivalent to a 100 amino acid protein and 3000 water molecules). In addition, these Ewald-type calculations can be parallelized readily.[55,56] The real space part of the sum can be treated using a standard domain decomposition strategy while the reciprocal space part of the sum requires the efficient parallelization of a three-dimensional FFT with cutoffs in both reciprocal and real space. The force field represented in Equation 1 effectively incorporates many-bodied effects such as polarization effects through the parameters, but only requires the evaluation of two body forces. In particular, the molecular charge distribution is represented by partial charges of fixed magnitude assigned to atoms or sites, and the molecular charge distribution does not respond to the environment explicitly. More explicit inclusion of many-body polarization and many-body dispersion requires the evaluation of many-bodied forces, and can be accomplished through the use of a Drude model.

In a Drude model, the electronic degrees of freedom associated with a site/atom are treated by two noninteracting charges of opposite sign tethered together by a harmonic spring.[57–60] The negative charge has a small mass, and the positive charge is fixed on the site/atom. Charges associated with different sites interact via Coulomb's Law. A classical treatment, minimization of the energy with respect to the light particle positions, yields many-body polarization up to dipole order. A quantum mechanical treatment yields dipole and higher many-body polarization as well as many-body dispersion. Modern path integral techniques can be used to simulate quantum Drude models efficiently.[61,62] Finally, the Drude model scales like $N\log N$ although the computational overhead is larger than a fixed charged treatment.

Figure 1 An estimate of the number of updates of Equation 1 and its Cartesian derivatives that can be accomplished with $N\log N$ scaling algorithms on a 100 teraflop computer for a 10000 atom system (equivalent to a 100 amino acid protein and 3000 water molecules)



Simulation methodologies for dynamics and thermodynamics. A dynamical description of how protein atoms and water molecules evolve in time can be determined by solving Newton's equations of motion. A typical simulation is initiated by inserting a large biomolecule into a box of solvent, removing the overlapping solvent molecules, equilibrating and performing a long simulation. New positions and velocities are determined numerically using various finite difference algorithms, and the propagated error in the updated quantities is proportional to the

power of the time step. Extended system equations of motion and associated numerical integrators have been developed that allow extensions from microcanonical ensemble dynamics to sampling of states in the canonical ensemble, as well as the isothermal-isobaric ensembles.[63–65]

The stability of these finite difference numerical integrator algorithms is dependent upon a time step that is commensurate with the fastest timescale in the system. Bond vibrations have an amplitude of 0.01 Å and therefore limit the use of the central difference equations to time steps on the order of 1 femtosecond (10^{-15} seconds). Constraint dynamics that effectively project out the force along bonds (SHAKE or RATTLE) can increase the time step to 2 femtoseconds (fs), so that the next fastest timescales arise from bond angle distortions. However, freezing bond angles has a significant adverse effect on developed structural and timescale properties of protein dynamics so that "Shaking" bond lengths and the 2 fs timescale resolution was part of the scaling behavior for simulations of protein-water and protein-protein interactions.

Recent advances in modern numerical integrators can now separate out the natural timescales of motions that depend on the strength of forces associated with each term in Equation 1.[66–71] Based on a factorization of the evolution operator used in quantum statistical mechanics, a formal decomposition of the integration time step allows bonds to be updated more frequently than angle bends, and angle bends more often than short-range forces, and short-range forces more often than long-range forces. This formally correct multiple time step integration has been shown to generate about an order of magnitude improvement in computational efficiency in biomolecular systems, although resonance artifacts can reduce this gain in efficiency in practice. The decrease in computer time results from the fact that the most expensive terms in Equation 1, the double sum over atoms, need to be updated less often than local interactions, i.e., the single sum terms in Equation 1. Calculations performed using multiple time step integration methods in isothermal or isobaric-isothermal ensembles are very scalable. Each time step results in a collective "move" and parallelization can proceed using standard domain decomposition paradigms.

It is a difficult task to efficiently sample the conformational space of large complex single domain proteins. Large proteins relax on timescales of an order of a second in solution, a benchmark atomistic simulations cannot approach at present. It would be useful, therefore, to extend the practical range of polypeptide sizes that can be simulated and can be said with reasonable confidence to have achieved conformational equilibrium. Although improved numerical integration and equations of motion have helped, several orders of magnitude improvement in efficiency need to be obtained.

In umbrella sampling, a series of simulations are performed using not the true direct potential energy function but the true potential energy function plus a biasing potential.[72,73] The biasing potential is characterized by a set of parameters that serve to adjust the strength of the bias along a "reaction coordinate." The biasing potential forces the dynamics to sample regions of the "reaction coordinate" that may not otherwise be explored extensively. Thus, a series of calculations at selected parameter sets can be performed to "drag" the system through its configuration space, along the reaction coordinate "in the shade of the umbrella" formed by the biasing potential. It is possible to achieve large reductions of computational effort if the reaction coordinate contains the rate limiting pathway(s) through configuration space. In general, umbrella sampling creates a simple level of parallelism because calculations employing a different set of biasing parameters can be performed independently. Postprocessing using the weighted histogram method[74] can be used to eliminate the systematic bias in a formally exact manner.

The next level of complexity involves borrowing methods from the path integrals molecular dynamics literature.[61,75] Specifically noncanonical, order N variable transformations increase the sampling efficiency of Gaussian random coil calculations by a factor of over 200. Most of the increase in efficiency can be ascribed to the noncanonical variable transformation that permits the long wavelength fluctuations of the coil to occur on a fast timescale. Applying such transformations with umbrella sampling has allowed the efficient and accurate determination of the hinge bending free energy surface of the mutant T4 lysozyme (*in vacuo*) and in computer water solution.[75]

Finally, proteins have a much more complex configuration space than random coils, hinge bending modes, or intermediate size peptides. It is therefore useful to implement yet another class of variable transformation borrowed from the Monte Carlo

literature that can, in principle, make true protein/polypeptides "resemble" a random coil model by analytically eliminating torsional barriers along the peptide backbone. The fast random coil methodology can then be applied "on top" of this first transformation. The idea behind the new technique is to create through the use of noncanonical variable transformations a smooth effective energy landscape without a concomitant modification of the potential energy surface itself.[75] In contrast, standard torsional dynamics schemes seek to eliminate motion in directions tangential to the backbone dihedrals to promote the use of larger time steps and barrier crossing events but do alter the heights of the barriers in the coordinate space. The new method should also be contrasted to importance sampling schemes where barriers are cut by an *ad hoc* modification of the potential energy surface and must be "regrown" by an (*a posteriori*) reweighing of the trajectories. In the new method, the reweighing occurs dynamically, through the properties of the noncanonical variable transformation. Preliminary results are promising.

Quantum mechanical algorithms. Modeling protein structure or enzyme reactions at higher levels of accuracy involves two simulation approaches at present. One is semiempirical or *ab initio* quantum mechanical (QM) methods that possess sufficient or even high accuracy for various biochemical properties of interest, but are currently limited to relatively small chemical systems and nondynamic simulations. Another is classical molecular dynamics (MD), which simulates the motions of atoms in their chemical context for relatively large systems and long timescales, but with empirical force fields that often have insufficient accuracy, and altogether fail to treat the breaking and forming of bonds, that is especially important for enzymatic reactions. We have outlined much of the methodological and computing kernels of classical MD algorithms in the previous section, and focus this section on QM methods.

The next generation of quantum chemistry algorithms will exploit new theories and technology that will reduce the scaling requirements of the HF, DFT, and MP2 methods.[76] The simplest level of *ab initio* QM simulation is the Hartree-Fock (HF) method. This method produces very accurate bond lengths and angles and reasonable reaction energies. Potentially more accurate structures and reaction energies can be determined with density functional theory (DFT) that shares HF's favorable scaling properties. Promising new algorithms, such as the MP2 method, should allow for very accurate energetic calculations on the chemically significant segments of many biochemical reactions. However, for certain properties, such as reaction barriers that are particularly important in nonequilibrium biochemical processes, more sophisticated QM methods, such as the coupled-cluster (CC) theory including all single and double excitations (CCSD) and with perturbative triples (CCSD(T)), may be required. If we consider the series of theoretical models, HF or DFT methods, MP2, CCSD, CCSD(T), for a given size basis set, and varying molecular size M, then in the simplest analysis their computational requirements scale as M^4, M^5, M^6, and M^7, respectively. However, recent research has contributed to a rapid breaking down of the computational bottlenecks in HF, DFT, and MP2 calculations.

Two steps are involved in one HF/DFT energy and derivative calculation. The first step is the construction of the effective one-electron Hamiltonian matrix, usually termed the Fock matrix, given a density matrix. The second is the evaluation of a new density matrix, usually via the generation of new molecular orbitals or Kohn-Sham orbitals. That HF and DFT methods naively scale as the fourth power of molecular size arises because of the evaluation of electron-electron interactions via *four* center two-electron integrals. However, the number of nonnegligible two-electron integrals does not grow quartically with the size of the molecule, but grows as M^2 when the molecular size is large enough (i.e., the two atomic orbitals (AOs) comprising each pair must overlap in order to make a distribution containing non-negligible charge). This realization, together with advances in the speed of two-electron integral evaluation[77] (integrals are generated as they are needed rather than stored), combine to permit routine calculations on systems approaching the 100 heavy (i.e., nonhydrogen) atom range.

Linear scaling in the assembly of the Fock matrix follows directly from the collectivization of distant electron-electron interactions via multipole expansions with controlled error bars known as fast multipole methods.[78,79] In the face of linear scaling methods for electron integral evaluation, the generation of a new density matrix via diagonalization that scales as M^3 will eventually become dominant for large molecular sizes. Current effort has been directed toward methods for updating the density and/or orbitals without explicit diagonalization, taking advantage of the fact that for most molecules, such as proteins, the density matrix is spatially localized.[80,81]

It is important to emphasize that DFT and existing functionals capture only certain types of electron correlation, and therefore the quality of DFT calculations are still under debate. We note that the development of new DFT functionals is an active area of research.[82–84] A potentially feasible alternative is the MP2 method that is the simplest wavefunction-based theory of electron correlation. In most current quantum chemistry program packages MP2 scales as M^5; the M^5 scaling is a consequence of the formulation of MP2 using delocalized MOs (molecular orbitals), which arise from standard HF calculations. However, the MOs can be localized, and there has been some preliminary progress toward developing versions of MP2 theory based on localized orbitals. The "local-MP2" method scales only quadratically with molecular size, and comes to within a few percent of reproducing the exact MP2 energy with a given basis.[85]

The simulation of certain enzyme catalyzed reaction mechanisms involving homolytic bond breaking (i.e., the breaking of electron pairs) or, if transition metal atoms are involved in the active site, will require more accurate electron-correlated quantum chemical methods, e.g., coupled cluster (CCSD),[86] that presently scale as M^6–M^7. For example, a CCSD energy calculation should be feasible for a 40-atom system on a teraflop computer, which is sufficiently large to include a typical enzyme substrate and several catalytic amino acid residues.

Despite the great value of the static properties that can be calculated using QM methods, many biological processes are inherently dynamical. Such problems include processive reactions (DNA or protein synthesis) and processes such as macromolecular conformational changes (DNA unwinding and allosteric enzyme regulation). Empirical force fields, without the inclusion of electrostatic polarization, cannot accurately describe the solvation of highly charged biomolecules and such force fields are inherently unable to treat bond-making and -breaking reactions. Improvements will be made to these classical force fields, but a shift to quantum mechanical force fields (*vide infra*) will be required to achieve quantitatively accurate enzymatic simulations.

The primary advancement will be the merging of the QM and molecular dynamics methods to allow so-called first principles MD, where quantum mechanical forces will be used to drive the classical motions of the atoms.[87–89] Extension to dynamics simulations requires considerable methods development; the DFT force calculation must be converted to a linear-scaling method, and the entire molecular dynamics simulation must be implemented on a massively parallel computer. Even with these improvements, first principles MD will not yet be feasible for long time scales and large molecular sizes such as that outlined for empirical force fields (see Case Study 5). However, this capability will allow the solving of a large number of fundamental biophysical problems that have been inconclusively addressed by existing classical MD methods. These problems include the determination of the hydration structure of the DNA nucleoside bases; the energetic factors leading to DNA base pairing; the hydration of the DNA backbone and basic sites; and the role of polarization in the stability of protein α-helices, for example.

Optimization and search strategies to construct biochemically relevant protein structures. The quantitative determination of protein structure will be critical in extending the information that emerges from fold prediction to structures that are relevant for investigation of biochemical questions. It is well understood that current *de novo* and fold analysis techniques provide structural information that is "low resolution," i.e., structures are typically 4–8Å RMS deviation from structural models emerging from NMR or X-ray structure determinations (these structures are typically precise to a level of 0.25 Å–0.75 Å). To extend the resolution of such structures to levels analogous to that from experimental structure determination methods, and hence to biochemically relevant levels, requires further refinement employing the types of force fields used in the folding free energy mapping calculations noted above. Only once we have achieved such resolution can we be confident in the use of these structural models as starting points in drug discovery and functional assessment methods.

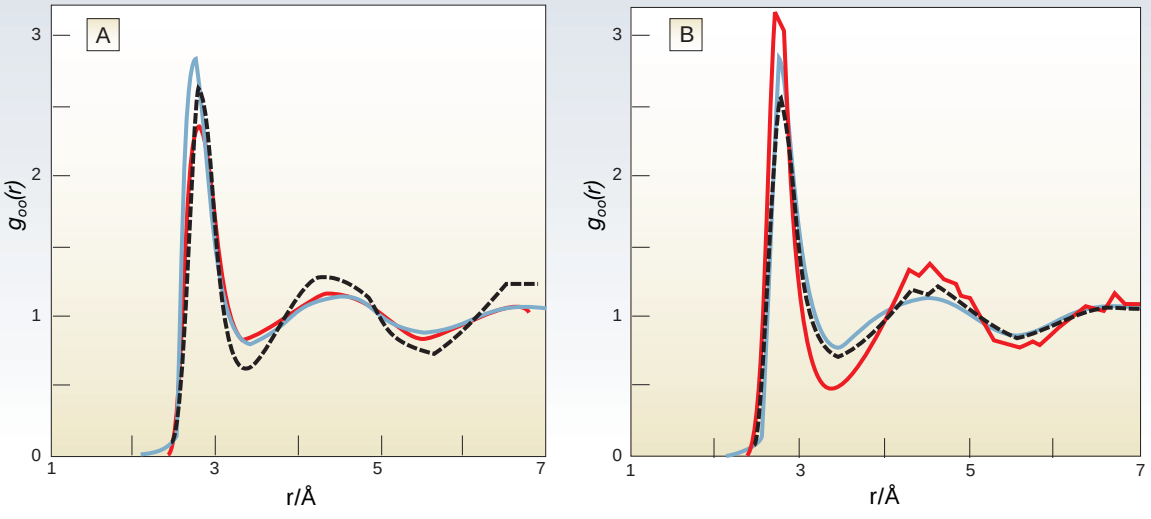
The problem of determining the full three-dimensional arrangement of the protein molecule in its most pragmatic guise is to ignore timescale bottlenecks for simulating the kinetics and mechanisms for how proteins fold, and instead determine effective ways of moving on the surface by walking through barriers. The conformational space of a protein is very high in dimensionality and complexity, so both local and global minima are of interest. The complexity of real protein surfaces rule out exhaustive enumeration of minima, so that sophisticated conformational searches and/or global optimization approaches are necessary to rapidly access the relevant regions of the energy surface. A large number of con-

Case Study 5

Ab initio molecular dynamics simulations

We have recently performed a new X-ray diffraction study of liquid water under ambient conditions at the Advanced Light Source (ALS) at Lawrence Berkeley National Laboratory that takes advantage of various state-of-the-art features of a modern-day experiment,[1] including quantitative characterization of the X-ray source together with the use of a more sophisticated Charge Coupled Device (CCD) area detector. We presented a $g_{oo}(r)$ for water consistent with our recent experimental data gathered at the ALS, and we outlined what features of $g_{oo}(r)$ should be reproduced by a simulation method.[2]

We evaluated the performance of some of the first fully quantum treatments of electronic structure in aqueous simulations,[3-5] all of which rely on a local density approximation to density functional theory. The differences between experiment and the *ab initio* results arise from several sources that typically have been investigated and overcome in the classical simulation literature, including dependence on initial conditions, length of the simulations, variation in system properties that arise with temperature, density, and finite size effects. Given the current computational expense that prohibits box sizes typically used in empirical force field simulations at present, these are largely technical limitations that will clearly diminish over time, and we would expect quantitative agreement to improve in the future.



Comparison of ALS experimental $g_{oo}(r)$ (gray line) with *ab initio* molecular dynamics simulations. (A) Silvestrelli and Parrinello[3] *ab initio* simulation of 10ps for 64 water molecules, average ionic temperature of 318K (red line), Sprik et al.[4] *ab initio* simulation of 5ps for 32 water molecules, average ionic temperature of 303K (black dashed line). (B) Schwegler et al.[5] *ab initio* simulation of 2ps for 54 water molecules, average ionic temperature of ~300K (red line); more recent *ab initio* simulation by Schwegler et al. 5ps for 54 water molecules, average ionic temperature of ~294K (black dashed line).

Teresa Head-Gordon
University of California, Berkeley

References

1. G. Hura, J. M. Sorenson, R. M. Glaeser, and T. Head-Gordon, *Journal of Chemical Physics* **113,** 9140–9148 (2000).

2. See Reference 1, pp. 9149–9161.

3. P. L. Silvestrelli and M. Parrinello, "Structural, Electronic, and Bonding Properties of Liquid Water from First Principles," *Journal of Chemical Physics* **111**, 3572–3580 (1999).

4. M. Sprik, J. Hutter, and M. Parrinello, "*Ab Initio* Molecular Dynamics Simulation of Liquid Water-Comparison Three Gradient-Corrected Density Functionals,"*Journal of Chemical Physics* **105**, 1142 (1996).

5. E. Schwegler, G. Galli, and F. Gygi, "Water Under Pressure," *Physical Review Letters* **84**, 2429–2432 (2000).

formational search or optimization strategies have been developed to tackle protein structure prediction.

Simulated annealing, genetic algorithms, "nonlocal" dihedral angle Monte Carlo, and various mathematical optimization methods attempt to search more globally than local minimization algorithms.[90–102] Simulated annealing is based on statistical mechanical theories for freezing in which the system is artificially "heated" to a high temperature and slowly cooled to "crystallize" to the lowest energy minimum. The correct cooling protocol and schedule is vital since a too-rapid descent in temperature can result in trapping into metastable minima; advances in computing can allow the cooling rate to be a few orders of magnitude slower. Genetic algorithms define a set of genes composed of structural variables and their connection to a potential energy function; new genes are evolved by genetic crossover and random mutation, and genes that are unfit are eliminated from the population. Eventually, a population of genes (variables) is left, which in principle generates the lowest energy value. Nonlocal Monte Carlo methods have been developed where large moves are made with reasonably high acceptance when a small number of dihedral angles (backbone torsions ϕ or ψ, or side-chain torsions) are varied according to probability maps of their amplitude derived from a representative set of proteins.

Mathematical optimization research is a more general approach for obtaining solutions to large nonlinear systems with numerous local minima, with protein folding being a recent example. Constrained optimization methods rely on the availability of sufficiently well-defined constraints so that the desired solution is the only available minimum, or one of few available minima, in the optimization phase of the algorithm. Alternatively, global optimization and conformational search techniques attempt to systematically search the potential energy surface to find all low-lying minima including the global energy minimum.

The technique of "diffusion smoothing" is based on analogies to diffusion and heat conduction. A smoothing operator is applied to the potential energy surface to remove shallow minima and "absorb" them in nonshallow minima that become even deeper. Stochastic/perturbation is a global optimization algorithm that consists of two phases.[103] In the first phase, a set of initial configurations is either designed or randomly generated, and each is used as a starting point for a local minimization. The best of the resulting local minimizers forms a pool used in the next phase. The second phase consists of repeatedly selecting conformations and modifying them using a small-dimensional global optimization probabilistic algorithm of Rinnooy Kan.

The useful application of these conformational search and optimization strategies is itself computationally intensive since it typically requires $\sim 10^5$–10^6 evaluations of an energy function and its derivative such as that given in Equation 1. These optimization approaches are useful in many contexts, including atomic-level structure prediction, for use in comparative modeling, docking of small ligands into protein active sites, or finding optimal protein-protein interaction geometries.

Advancing biotechnology research: *In silico* drug design

Only a small percentage of a pool of viable drug candidates actually lead to the identification of a clinically useful compound, with typically over $200 million spent in research costs to successfully bring it to market. On average, a period of 12 years elapses between the identification and FDA (Food and Drug Administration) approval of a successful drug, with the major bottleneck being the generation of novel, high-quality drug candidates. While rational, computer-based methods represent a quantum leap forward for identifying drug candidates, substantial increases in compute power are needed to allow for both greater selection sensitivity and genome-scale modeling of future drugs.

Table 1 gives an estimate of the method and *current* computational requirements to complete a binding affinity calculation for a given drug library size. Going down the table for a given model complexity is a level of computational accuracy. The benefits of improved model accuracy must be offset against the cost of evaluating a model over the ever-increasing size of the drug compound library brought about by combinatorial synthesis, and further exacerbated by the high throughput efforts of the genome project and structural annotation of new protein targets.

Models and algorithms for *in silico* drug design. Docking methods are computational algorithms developed to both predict the three-dimensional structures of ligand-receptor complexes, and to evaluate the relative affinity or free energy of binding for these bound ligands or drugs.[104–109] The need for improved

Table 1 Estimates of current computational requirements to complete a binding affinity calculation for a given drug library size

Modeling Complexity	Method	Size of Library	Required Computing Time
Molecular mechanics	SPECITOPE	140000	~1 hour
Rigid ligand/target	LUDI	30000	1–4 hours
	CLIX	30000	33 hours
Molecular mechanics	Hammerhead	80000	3–4 days
Partially flexible ligand	DOCK	17000	3–4 days
Rigid target	DOCK	53000	14 days
Molecular mechanics/ fully flexible ligand, rigid target	ICM	50000	21 days
Molecular mechanics/ free energy perturbation	AMBER, CHARMM	1	~several days
QM active site and MM protein	Gaussian, Q-Chem	1	>several weeks

docking and scoring methods is now especially acute given the future direction toward high-throughput annotation of genomes to generate new protein structural targets, combined with revolutionary advances in combinatorial synthesis of small molecule docking candidates.

The current combinatorial library paradigm is to design diverse drug libraries aimed at multiple but unknown targets or directed ligand libraries aimed at optimizing hits against individual targets. An inverted procedure is possible in which one or many libraries are screened on the computer against many targets to determine which libraries have the most desirable characteristics for which targets. This general approach can also include an "optimization" cycle where augmented libraries are scored against the best targets.

The fundamental attractiveness of this approach is that potential targets for all compounds can be addressed at a much earlier time and at much lower cost per compound, and is consistent with genome-scale drug design efforts. The basic challenge is how to improve the accuracy of the fundamental docking algorithms themselves while rapidly screening increasingly larger drug databases both in-house and in the public domain.

Docking methods for geometric optimization of a candidate drug into a target active site is a solved problem when both the ligand and the target are treated as rigid objects. In some cases, limited flexibility is introduced by dividing the ligand into several rigid fragments that are docked separately. In either case, these binding complexes are then evaluated with an empirical scoring function, which we discuss later. This represents the level of sophistication that is currently available from commercially available software packages. This approach is likely to identify, at best, one weakly binding compound per database of 100000 chemicals, because there are too many false negatives generated.

Introduction of full flexibility of at least the ligand for docking into a rigid target in order to refine the binding geometry has been shown to lead to better binding energetics, and therefore finding better drug leads in general. Flexible ligand and rigid target represents the upper limits of what can be attempted with current computational resources. When the peptide backbone and side chains of the target molecule are also treated as flexible, allowing the molecule to undergo locally induced conformational changes upon ligand binding, the resulting induced fit seems to be essential to understand ligand specificity. Large-scale screening with full flexibility of both ligand and localized areas of the target is well beyond reach with current computational resources, since only a few compounds can be screened within a practical time interval. Essentially increased use of geometric refinement with at least full ligand flexibility, and ideally at least localized target flexibility, via standard optimization techniques for large libraries of drug compounds, is an accessible and desirable goal in using future teraflop computing. Once a drug is geometrically docked, scoring of the binding affinity of a drug-receptor complex ranges from

statistical multivariate equations that correlate X-ray crystal structural data of ligand-receptor complexes with experimental free energies of binding, to physically-based molecular mechanics approaches, to computationally intensive free energy perturbation methods.

Overall the rapidly calculated multivariate functions perform as well as computationally intensive free energy perturbation calculations, with estimated relative binding free energies of about 2 kcal/mole, which corresponds to a binding affinity error ($\sim 10^{\Delta G \text{ error}/1.4}$) of about 30-fold. The quality of these scoring functions provides a qualitative filter for ordering the binding affinity of drugs in large databases, but is not a reliable predictor of the most active drug molecules. Ranking drug affinity among many ligands for a given target is where multivariate functions work well, but it is unlikely that these existing scoring functions could determine specificity of a single drug against different receptors.

Both molecular mechanics and free energy perturbation methods have the advantage of being based on physical interactions, so that alternative problems can be treated by the same approach. Molecular mechanics functions with solvent-accessible surface area descriptions for solvation have on average performed significantly worse than multivariate functions in the past, with binding free energy errors typically being 3 kcal/mole/. However, quite good correlation of the enzyme inhibitor activity of 33 inhibitors of HIV-1 proteases were determined by a purely molecular mechanics scoring function, and recent reported results for some new empirical force fields perform as well as correlated MP2 QM methods.

Based on these classical approaches, the cost of evaluating a 100000 compound library against 100000 gene products would take on the order of a full year on a dedicated teraflop computer. Certain large pharmaceutical companies have in-house databases approaching 500000 drugs, and revolutionary combinatorial synthesis approaches are going to expand these databases even further. If we add on top of that additional modeling accuracy requirements to better screen drugs, i.e., better empirical force fields, longer refinement stages in the calculation, and greater target flexibility, the search for better drug candidates will utilize well teraflop capabilities on a sustained basis, and is inherently scalable beyond this projected computing goal.

State-of-the-art quantum chemistry algorithms are also worth consideration in the future since we can expand the applicability of quantum mechanics/molecular mechanics (QM/MM) methods to simulate a greatly expanded QM subsystem for enzymatic studies, or even estimation of drug binding affinities. An estimate of the cost of using QM methods for evaluating a single energy and force evaluation system of 10^4 heavy atoms would require resources that can handle $\sim 10^{16}$ FLOPS. Parallel versions of these methods have been implemented and are available at a number of universities and government laboratories.[110–113] On current generation teraflop platforms, QM calculations of unprecedented size are now possible, allowing HF optimizations on systems with over 1000 atoms and MP2 energies on hundreds of atoms. The presently available traditional-scaling ($\sim N^3$) first principles molecular dynamics code is running efficiently on serial platforms, including high-end workstations and vector supercomputers. The ultimate goal is to develop linear scaling quantum molecular dynamics code. It will be important to adapt these codes to the parallel architectures, requiring rewriting parts of existing programs and developing linear scaling algorithms.

Linking structural genomics to systems modeling: Modeling the cellular program

The cellular program that governs the growth, development, environmental response, and evolutionary context of an organism does so robustly in the face of a fluctuating environment and energy sources. It integrates numerous signals about events the cell must track in order to determine which reactions to turn on, off, or slow down and speed up. These signals, which are derived both from internal processes, other cells, and changes in the extracellular medium, arrive asynchronously, and are multivalued in meaning. The cellular program also has memory of its own particular history as written in the complement and concentrations of chemicals contained in the cell at any instant. The circuitry that implements the working of a cell and/or collection of cells is a network of interconnected biochemical, genetic reactions, and other reaction types.

The experimental task of mapping genetic regulatory networks using genetic footprinting and two-hybrid techniques is well underway, and the kinetics of these networks is being generated at an astounding rate. Technology derivatives of genome data such as gene expression micro-arrays and *in vivo* fluorescent tagging of proteins through genetic fusion with

Case Study 6

Modeling the reaction pathway for the glycolytic biochemical system

A novel gene expression time series analysis algorithm known as the Correlation Metric Construction (CMC) uses a time-lagged correlation metric as a measure of distance between reacting species. The constructed matrix R is then converted to a Euclidean distance matrix D, and multidimensional scaling, MDS, is used to allow the visualization of the configuration of points in high dimensional space as a two-dimensional stick and ball diagram. The goal of this algorithm is to deduce the reaction pathway underlying the response dynamics, and was used on the first few steps of the glycolytic pathway determined by experiment.[1]

The reconstituted reaction system of the glycolytic pathway, containing eight enzymes and 14 metabolic intermediates, was kept away from equilibrium in a continuous-flow, stirred-tank reactor. Input concentrations of adenosine monophosphate and citrate were externally varied over time, and their concentrations in the reactor and the response of eight other species were measured. The CMC algorithm showed a good prediction of the reaction pathway from the measurements in this much-studied biochemical system. Both the MDS diagram itself and the predicted reaction pathway resemble the classically determined reaction pathway (see Figures 3A, 3B, and 5 in Reference 1). In addition, CMC measurements yield information about the underlying kinetics of the network. For example, species connected by small numbers of fast reactions were predicted to have smaller distances between them than species connected by a slow reaction.

Adam Arkin
University of California, Berkeley

Reference

1. A. Arkin, P. D. Shen, and J. Ross, "A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements," *Science* **277**, 1275–1279 (1997).

the green fluorescent protein (GFP) can be used as a probe for network interaction and dynamics. If the promise of the genome projects and the structural genomics effort is to be fully realized, then predictive simulation methods must be developed to make sense of these emerging experimental data.[114–121] First is the problem of modeling the network structure, i.e., the nodes and connectivity defined by sets of reactions among proteins, small molecules, and DNA. Second is the functional analysis of a network using simulation models built up from "functional units" describing the kinetics of the interactions.

Prediction of networks from genomic data can be approached from several directions. If the function of a gene can be predicted from homology, then prior knowledge of the pathways in which that function is found in other organisms can be used to predict the possible biochemical networks in which the protein participates. Homology approaches based on protein structural data or functional data for a protein previously characterized can be used to predict the type of kinetic behavior of a new enzyme. Thus structural prediction methods that can predict the fold of the protein product of a given gene are fundamental to the deduction of the network structure.

The nonlinearity of the biochemical and genetic reactions, along with the high degree of connectivity among substrates, products, and effectors in these reactions, make the qualitative analysis of their behavior as a network difficult (see Case Study 6). Furthermore, the small numbers of molecules involved in biochemical reactions (typical concentrations of 100 molecules/cell) ensure that thermal fluctuations in reaction rates are expected to become significant compared to the average behavior at such low concentrations. Since genetic control generally involves only one or two copies of the relevant promoters and genes per cell, this noise is expected to be even worse for genetic reactions. The inherent randomness and discreteness of these reactions common inside liv-

ing cells can have significant macroscopic consequences.

There are three bottlenecks in the numerical analysis of biochemical reaction networks. The first is the multiple timescales involved. Since the time between biochemical reactions decreases exponentially with the total probability of a reaction per unit time, the number of computational steps to simulate a unit of biological time increases roughly exponentially as reactions are added to the system or rate constants are increased. The second bottleneck derives from the necessity to collect sufficient statistics from many runs of the Monte Carlo simulation to predict the phenomenon of interest. The third bottleneck is a practical one of model building and testing: hypothesis exploration, sensitivity analyses, and back calculations will also be computationally intensive.

Methods and models for deducing genetic and biochemical network structures. First we describe the problem of modeling the network connectivity using time series analysis. Most of the time series analysis techniques that have been applied to gene expression data fall into the category of statistical, distance-based methods. The idea is to define a distance metric on the space of species concentration that associates smaller distances with directly interacting species, larger distances with indirectly relating species, and very large distances with species that do not interact at all. Once a distance matrix has been constructed—an assignment of a number to each pair of species under consideration—analysis techniques such as clustering can be used to draw further meaning from the distance matrix and to represent putative interspecies relationships graphically.

The simplest distance-based technique for analyzing gene expression time series is that of simple correlation. The species are treated as random variables and a correlation coefficient is calculated for each pair of species and used as a measure of distance between chemical species. Simple correlation reveals linear, simultaneous relationships between variables. If two mRNA concentrations co-vary linearly, either positively or negatively, with time and/or perturbation values, this covariance will be reflected in a correlation distance measure. However, nonlinear relationships between variables are not measured by correlation coefficients, nor are time shifted linear relationships. Since gene regulation networks are thought to follow a logic best described by nonlinear hybrid algebraic differential equations, such a measure would seem to be lacking. However, the application of such a simple distance measure combined with clustering techniques has resulted in valuable and unexpected insights (see Case Study 6).

The computational cost of evaluating the correlation distance matrix with a simple correlation distance metric is $NM^2/2$, where M is the number of genes being monitored over N time points. Since there are an estimated 100 000 genes along the human genome, calculating a distance matrix over 2000 observational time points spanning embryonic development would cost 10^{13} operations. Once a distance matrix has been constructed, analysis and visualization techniques must be applied in order to derive meaning from the distance matrix that adds additional computational overhead to the cost of the initial matrix construction.

The next-simplest distance-based techniques for analyzing gene expression time series use time-delayed correlations between variables at different time lags in order to construct a distance matrix. For every pair of species, a correlation coefficient is calculated for the pair at all possible time lags. In its simplest version, the distance between the two species is then taken to be the maximum correlation coefficient calculated, or some function of this maximum.

Time-shifted correlations reveal linear, potentially time-lagged relationships between variables. Being able to capture time-shifted relationships between species is an important feature for a gene expression distance metric to have, because it allows detection of cascade-like regulation mechanisms—fairly common transcription-level gene expression control structures. The simple no-lag correlation metric can miss such relationships altogether. As with the simple correlation metric described previously, nonlinear relationships between variables are not measured by time-shifted correlation coefficients. Though this is a serious limitation, time-shifted correlation metrics can be considered a valuable step up in the representational hierarchy from simple correlation, because they are able to capture linear, time-invariant system dynamics. Because correlations must be calculated at all possible time lags between variable pairs, constructing a time-shifted correlation matrix is more expensive than constructing the simpler metric.

If all the interactions in a network were linear, then multivariate linear regression would provide the best estimate of the dependence of one variable in the system on the others. However, the dependence on

the activity (or concentration) of one component as a function of the others is most often very nonlinear. In this case, linear dependency measures must be discarded in favor of general measures of dependency such as the transinformation. The transinformation is defined in terms of the joint probability distributions of sets of variables. There are a number of analyses that exploit this measure to produce and test network hypotheses against multivariate, often time-resolved, data.

In order to estimate the dependence of one variate on another, we must calculate conditional probabilities, that is, the probability that one variable is in one state (concentration range) given that another variate is in another state. Enough data must be collected so that the deduced relationships among variables can be deemed statistically significant. For these analyses this amount in data can be estimated via the χ^2 statistic. If we assume that every chemical variable in our system can take on only Q different biologically significant states, then the data constraint states that for credible analysis the minimum number of data, d, (where each data element represents the observation of *all* N variables) is governed by:

$$d \geq 5Q^N$$

Thus, over 5000 observations must be made for a system of ten binary variables. Obviously, this data constraint is extremely harsh for biochemical systems in which the number of biologically significant concentrations can be relatively large and the number of variables orders of magnitude greater than ten. Therefore methods must be developed for breaking large biochemical networks into smaller subnetworks, which can be probed using this method.

From the time-series data a statistical analysis must predict the most probable network of interactions between chemical species that produced the observed system dynamics. To do so, the method must effectively check every possible network of connections among the measured species. While the number of such network structures rises exponentially with the number of variables composing the system, practically, the number of possible networks is greatly reduced with constraints on the solution by inserting chemical and genetic knowledge into the analysis, and to simply assume limited dependencies within the network.

Limiting the number of variables that can directly cause variations in an observation severely reduces the model space that it is necessary to test. For each variable, j, one finds the strength of the relationship between j and all other pairs of (perhaps time-lagged) variables. If the strength of the interaction is statistically significant, then retain that pair in the dependency set for the variable j. If after testing all pairs the dependency set is empty, conclude that j does not depend on any other variable in the system. Otherwise, conclude that all variables in the dependency set are causative factors for j.

This is an $N(N-1)/2$ step algorithm (each step is composed of calculating the transinformation for each pair of variables). Each of these steps involves a three variable by M data point evaluation of a distribution estimation algorithm. All of these operations are repeated for each of the N variables, thus the scaling law is on the order of N^3M. However, the number of data points necessary to estimate the joint distributions in the transinformation for variables with Q states is of the order Q^N. The final scaling law for the estimation becomes approximately N^3Q^N. Actually, there is some redundancy in the distribution estimation steps that might be exploited to slightly reduce this Q^N dependency.

However, the assumptions behind this algorithm, that three-way transinformations are enough to predict interactions of order greater than three, can lead to errors of omission in eukaryotic systems, in particular. Given that eukaryotic systems can have many multiprotein complexes containing four or more proteins, this heuristic may have to be extended to cover at least four- and five-way interactions, $N^4Q^N - N^5Q^N$ scaling.

Methods and models for cellular network analysis. The nonlinearity of the biochemical and genetic reactions, along with the high degree of connection (sharing of substrates, products, and effectors) among these reactions, make the qualitative analysis of their behavior as a network difficult. Furthermore, the small numbers of molecules involved in biochemical reactions (typical concentrations of 100 molecules/cell) ensure that thermal fluctuations in reaction rates are expected to become significant compared to the average behavior at such low concentrations. Since genetic control generally involves only one or two copies of the relevant promoters and genes per cell, this noise is expected to be even worse for genetic reactions. The inherent randomness and discreteness of these reactions can have significant macroscopic consequences such as that common inside living cells.

Chemical systems evolve with time because of changes in their constituent molecules when those molecules collide and react. Since naturally occurring molecular collisions are *random*, the temporal evolution of any chemically reacting system is *stochastic*. Elementary kinetic theory shows that, under conditions in which reactive molecular collisions are separated by many nonreactive molecular collisions, the temporal evolution of the system's state, $\mathbf{X}(t)$, constitutes a jump Markov process. That is, $\mathbf{X}(t)$ performs a "random walk" in real time over the N-dimensional integer lattice space, hopping from one lattice point to another as successive reactions occur.

An algorithm simulating jump Markov processes has been rigorously derived from the same premises that lead to the master equation (ME). The ME defines evolution of $\mathbf{X}(t)$ s probability function $P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$, while the simulation generates sample trajectories or "realizations" of $\mathbf{X}(t)$. The heart of the simulation is a procedure for randomly deciding, at any time t when the system's state $\mathbf{X}(t)$ is known, at what time $t + \tau$ the next reaction in the system will occur and which reaction, R_μ, will be the next reaction.

Using a mathematically exact procedure for generating random values for τ and μ, the simulation moves the system forward in time from one reactive collision to the next, continually updating the chemical species population levels in accordance with the outcomes of the selected reactions. The *statistical* properties of the system behavior are estimated using statistics from multiple simulations under identical conditions.

From a modeling standpoint, the simulation has two advantages over the ME: it is straightforward to apply even to complicated coupled chemical reaction schemes, and the results of the simulation are directly comparable with experimental results obtained on real systems. The primary computational bottleneck of a simulation approach to the master equation is that it can be expensive to model behavior of systems with many reacting species over extended time intervals.

There are three bottlenecks in the numerical analysis of biochemical reaction networks; the first two pertain to using the ME approach. The first is the multiple timescales involved. Since the time between biochemical reactions decreases exponentially with the total probability of a reaction per unit time, the number of computational steps to simulate a unit of biological time increases roughly exponentially as reactions are added to the system or rate constants are increased.

The second bottleneck derives from the necessity to collect sufficient statistics from many runs of the Monte Carlo simulation to predict the phenomenon of interest. Often, such phenomena as phase-variation of coat-proteins in pathogenic virus and bacteria, oncogenesis or DNA mutation, occur at very low frequencies. Many runs of the Monte Carlo algorithm are necessary to properly estimate the probabilities of these events if they are to be analyzed via a stochastic simulation.

The third bottleneck is a practical one of model building and testing: hypothesis exploration, sensitivity analyses, and back-calculations will also be computationally intensive before master equation approaches can be applied to learn about the behavior of a proposed network model.

One approach to the first bottleneck problem is to develop a mode-switching algorithm that can change to numerical methods more efficient than the ME when certain conditions are met. A promising approach is to develop a simulation algorithm that "interpolates" between the master equation simulation and the standard ordinary differential equations (ODEs) used for described deterministic chemical kinetics. By first approximating the master equation by a Fokker-Planck equation (FPE), a subsequent step allows the generalization of the FPE to determine an approximate Langevin equation (LE). An algorithm would provide a decision as to how the dynamics should be propagated—using the Monte Carlo ME, the LE, or the ODE method. The decision would be based on criteria such as its current concentration, the concentration of those species with which it directly interacts, and the rate of the reactions in which it participates. As a simulation progresses, the mode of propagation for each species may switch, but switching between regimes must not introduce biased errors in the integration, neglect of the fluctuations should not lead to ablation of certain system behaviors, and some estimate must be made of the error introduced by adding heuristic submodels. This will require multiple exploratory simulations to be performed wherein the effect of varying parameters in the heuristic models is measured.

The above discussion has focused on spatially homogeneous chemical systems where rapid mixing

prevents the formation of persistent concentration gradients. Treating spatially inhomogeneous systems is more difficult. The usual deterministic approach is to convert the ordinary differential reaction rate equations into *partial* differential equations that incorporate Fick's macroscopic diffusion law. For a stochastic treatment, one has to subdivide the system volume into approximately homogeneous spatial subvolumes, and then allow diffusive exchanges of molecules between adjacent subvolumes. Stochastic simulation becomes even more computationally expensive in the spatially inhomogeneous case because of the many "diffusive exchange reactions" that must be simulated in addition to the chemical reactions.

Some stochastic simulations on spatially inhomogeneous systems have been reported, but there are unresolved technical issues associated with the choice of the subvolume size and the form of the probability rates for diffusive molecular transfers. An accelerated stochastic simulation algorithm for homogenous systems will inevitably be useful in accelerating the inhomogeneous case, since inhomogeneous systems are diffusively interacting assemblages of homogeneous subsystems.

In order to simulate the necessary statistics for a given chemical system, many runs of the Monte Carlo algorithm must be executed in parallel. About $4(1 - p)/f_e^2 p$ samples are required to estimate the probability, p, of a binary random event with 95 percent confidence where f_e is the desired maximum fractional error in p. Thus, a low probability event that occurs in one cell 1 percent of the time, and is to be predicted within ± 0.05 percent, would require ~ 10000 simulations. Many genetic and biochemical processes are composed of tens of genes, hundreds of proteins, complexes, and small molecules. In these cases the computational load is restrictive.

Eventually the largest system of genes and proteins that will probably have to be simulated is on the order of 100 genes and regulatory elements and 500 proteins, complexes, and small molecules and maybe 10 cellular compartments or locals. This is currently well beyond the scaling laws of current simulation algorithms and the 0.1 teraflop computing of today. The issues that must be addressed in this area are the disparate timescales requiring new mode-switching algorithms, and the gathering of the necessary statistics to quantify event likelihood. While the second issue unambiguously benefits from greater teraflop machines, the former does too since algorithm switching will likely only realize an order of magnitude savings in simulation time.

In addition, various computational and experimental data are vital to restricting the network structure and analysis space, i.e., more generally integrating domain knowledge into time-series analysis is required to be computationally feasible. Sources of information include not only experimental, but computational data such as large-scale sequence comparisons, phylogeny information, protein fold recognition, folding and prediction of structure, enzymology, and evaluation of ligand-receptor affinities and multiprotein interactions. These areas are compute-bound in their own right, and their connection to modeling of the cellular program is a natural outgrowth of the ambition of the computational biology effort of the future.

Other simulation issues for computational biology

The advanced computational biology simulations described in this paper will require computer performance well beyond what is currently available, but computational speed alone will not ensure that the computer is useful for any specific simulation method. Several other factors are critical including the size of primary system memory, also referred to as random access memory (RAM), the size and speed of secondary storage, and the overall architecture of the computer. Many parallel processing computer architectures have been developed over the years, but the dominant parallel architecture that has emerged is the distributed-memory, multiple-instruction multiple-data (MIMD) architecture, which consists of a set of independent processors with their own local RAM memory interconnected by some sort of communication channel. Such an architecture is characterized by the topology and speed of the interconnection network, and by the speed and memory size of the individual processors.

Just as with the computer hardware, there have been a large number of software programming paradigms developed for parallel computers. A great deal of research has gone into developing software tools to assist in parallel programming or even to automatically parallelize existing single-processor software. Selected parallelization and debugging tools can assist and new programming models such as object-oriented programming (using C++ or FORTRAN90) can help hide the details of the underlying computer architecture. At the current time, however, efficiently

using massively parallel computers primarily involves redesigning and rewriting software by hand. This is complicated by the facts that the best serial (single processor) algorithms are often not the best suited for parallel computers and the optimal choice of algorithm often depends on the details of the computer hardware.

The different simulation methods presently used have different requirements of parallel computer hardware. Simulation methods that involve calculating averaged properties from a large number of smaller calculations that can be individually run on gigaflop class processors are most ideally suited for parallelism. These methods include classical and quantum Monte Carlo simulation. In these simulations a minimal amount of initial data can be sent to each processor, which then independently calculates a result that is communicated back to a single processor. By choosing an appropriate size of problem for each single processor (problem granularity), these algorithms will work efficiently on virtually any MIMD computer, including separate computer workstations linked by local-area networks.

The quantum chemical and molecular dynamics methods, or certain optimization algorithms, in which all processors are applied to the calculation of a single chemical wave function or trajectory, involve much greater challenges to parallelization and involve greater constraints on the parallel computer architecture. Since all processors are involved in a single computation, interprocessor communication must occur. It is the rate of this communication, characterized in terms of raw speed (bandwidth) and initialization time (latency) that usually limits the efficient use of parallel computers. The minimal necessary communication rate depends exquisitely on the simulation type, choice of algorithm, and problem size. Generally, it is essential software design criteria that, as the problem scales to larger size, the ratio of computational operations per communication decrease (or at least remain constant), so that for some problem size, the communication rate will not constitute a bottleneck. Moreover, it is important that the work per processor, or "load balance," scale evenly so that no processors end up with much larger computational loads and become bottlenecks. In a broad sense, the nature of computational biology simulations—in particular the physical principle that interactions attenuate with distance—will ensure that scalable parallel algorithms can be developed, albeit at some effort.

Even given the very broad range of simulation methods required by computational biology, it is possible to provide some guidelines for the most efficient computer architectures. Regarding the size of primary memory, it is usually most efficient if a copy of the $(6 \times N)$ set of coordinates describing a time step of a molecular dynamics simulation or the $(N \times N)$ matrices describing the quantum chemical wavefunction, can be stored on each processing element. For the biological systems of the sort described in this paper, this corresponds to a minimum of several hundred megabytes of RAM per processor. Moreover, since many of the simulation methods involve the repeated calculation of quantities that could be stored and reused (e.g., two electron integrals in quantum chemistry or interaction lists in molecular dynamics), memory can often be traded for computer operations so that larger memory size will permit even larger simulations.

Similarly, general estimates can be made for the minimal interprocessor communication rates. Since the goal of parallel processing is to distribute the effort of a calculation, for tightly coupled methods such as quantum chemical simulations, it is essential that the time to communicate a partial result be less than the time to simply recalculate it. For example, the quantum chemical two-electron integrals require 10–100 floating point operations to calculate, so that they can be usefully sent to other processors only if that requires less than ~100 cycles to communicate to send the 8- or 16-byte result. Assuming gigaflop speeds for individual processing elements in the parallel computers, this translates roughly to gigabyte/sec interprocessor communication speeds. (Note that many partial results involve vastly more operations, so that they place a much weaker constraint on the communication rate.)

Information technologies and database management. "Biology *is* an Information Science"[122] and the field is poised to put into practice new information science and data management technologies directly. Two major conferences are emerging within the field of computational biology (ISMB—Intelligent Systems for Molecular Biology; and RECOMB—Research in Computational Biology). Each year, associated workshops focus on how to push new techniques from computer science into use in computational biology. For example, at ISMB-94, a workshop focused on problems involved in integrating biological databases; follow-up workshops in 1995 and 1996 explored Common Object Request Broker Architecture** (CORBA**) and Java** as methods to be used toward

integration solutions. In 1997, a postconference workshop focused on issues in accurate, usable annotations of genomes. In 2000, a preconference workshop explored how to use text-processing and machine-translation methods for building ontologies to support cross-linking between databases about organisms. All of these criteria require ultra-high-speed networks to interconnect students, experimental biologists, and computational biologists and publicly funded data repositories. This community will, for example, benefit directly from every new distributed networking data exchange tool that develops as a result of Internet2 and the high-speed Energy Sciences Network.

Data warehousing addresses the need to transparently query and analyze data from multiple heterogeneous sources.

Data warehousing addresses a fundamental data management issue: the need to transparently query and analyze data from multiple heterogeneous sources distributed across an enterprise or a scientific community. Warehousing techniques have been successfully applied to a multitude of business applications in the commercial world. Although the need for this capability is as vital in the sciences as in business, functional warehouses tailored for specific scientific needs are few and far between. A key technical reason for this discrepancy is that our understanding of the concepts being explored in an evolving scientific domain change constantly, leading to rapid changes in data representation. When the format of source data changes, the warehouse must be updated to read that source or it will not function properly. The bulk of these modifications involve extremely tedious, low-level translation and integration tasks that typically require the full attention of both database and domain experts. Given the lack of the ability to automate this work, warehouse maintenance costs are prohibitive, and warehouse "up-times" severely restricted. This is the major roadblock to a successful warehouse solution for scientific data domains. Regardless of whether the scientific domain is genome, combustion, high-energy physics, or climate modeling, the underlying challenges for data management are similar and present in varying degrees for any warehouse. We need to move toward the automation of these scientific tasks. Research will play a vital role in achieving that goal and in scaling warehousing approaches to dynamic scientific domains. Warehouse implementations have an equally important role; they allow one to exercise design decisions, and provide a test-bed that stimulates the research to follow more functional and robust paths.

Ensuring scalability on parallel architectures. In all of the research areas, demonstrably successful parallel implementations must be able to exploit each new generation of computer architectures that will rely on an increased number of processors to realize multiple teraflop computing. We see the use of software support tools as an important component of developing effective parallelization strategies that can fully exploit the increased number of processors that will comprise a 100 teraflop computing resource. However, these software support libraries have largely been developed on model problems at a finer level of granularity than "real life" computational problems. But it is the "real life" problems that involve complexity in length scales, timescales, and severe scalings of algorithmic kernels that are in need of the next and future generations of multiple teraflop computing. The problems described in this paper provide for a more realistic level of granularity to investigate the improved use of software support tools for parallel implementations.

Even when kernels can be identified and parallel algorithms can be designed to solve them, often the implementation does not scale well with the number of processors. Since multiple teraflop computing will only be possible on parallel architectures, problems in scalability are a severe limitation in realizing the computational biology goals outlined here. Lack of scalability often arises from straightforward parallel implementations where the algorithm is controlled by a central scheduler and for which communication among processors is wired to be synchronous. In the computer science community, various paradigms and software library support modules exist for exploring better parallel implementations. These can decompose the requirements of a problem domain into high-level modules, each of which is efficiently and portably supported in a software library that can address issues involving communication, embedding, mapping, etc., for a scientific application of interest.

These tools allow different decompositions of a parallel implementation to be rapidly explored, by handling all of the low-level communication for a given platform. However, these paradigms and their supporting software libraries usually are developed in the context of model mathematical applications, which are at a finer level of granularity than "real life" computational science problems. A unique opportunity exists to use some of the computational science problems described in the previous chapters to refine or redefine the current parallel paradigms currently in use. The outcome of this direction could be broader than the particular scientific application, and may provide insight on how to improve the use of parallel computing resources in general.

Conclusions

The Human Genome Project was undertaken with the goal to advance fundamental biological understanding and provide the basis for future advances in biotechnology, agriculture, environmental remediation and quality, and health and medical practice. The successes in the analysis of entire genomes have dramatically changed how the biochemistry of living cells is viewed, and provide clear directions for the future of mathematical and computational modeling of molecular, cellular, developmental, and physiological behavior, work that, in turn, will open new experimental horizons.

Modeling multiple levels of biological complexity is well beyond even the next generation of supercomputers, but each increment in the computing infrastructure makes it possible to move up the biological complexity ladder and tackle problems that could not be previously solved. Along with the maturation of the biosciences, an equally dramatic explosion in computer and information science technology has also taken place. As a consequence, there is exceptional synergism to be gained from exploiting these twin scientific revolutions.

For two decades, each advance in computing power has brought a new level of realism to simulation of biodynamics and has increased the scale of problems that physical instrumentation can address within biology. A challenge for the computational biology field arising from these new opportunities is to expand training and educational opportunities at the interface of biology with computer and information science, bioengineering, and the physical sciences. Computational science is now poised to be a partner in the armament of biological tools, maintaining an essential triangle of theory, computation, and experimentation seen in other scientific areas, but now coming into full fruition for structural and functional genomics of the future.

Acknowledgments

We would like to thank all of the following individuals who contributed to the intellectual content of this document. The following contributors provided written text and should be acknowledged as "virtual" coauthors for their section: "The first step beyond the genome project: High-throughput genome assembly, modeling, and annotation"—Phil LaCascio, Saira Mian, Richard Mural, Frank Olken, Jay Snoddy, Sylvia Spengler, David States, Ed Uberbacher, and Manfred Zorn; "From genome annotation to protein folds: Comparative modeling and fold assignment"—Alan Lapedes, David Eisenberg, and Andrej Sali; "Low-resolution folds to structures with biochemical relevance: Toward accurate structure, dynamics, and thermodynamics"—Charles Brooks, Mike Colvin, Yong Duan, Volkhard Helms, Peter Kollman, Glenn Martyna, and Doug Tobias; "Advancing biotechnology research: *In silico* drug design"—Ruben Abagyan and Jeff Blaney; "Linking structural genomics to systems modeling: Modeling the cellular program"—Adam Arkin and Denise Wolf; "Other simulation issues for computational biology"—Mike Colvin, Terry Gaasterland, and Charles Musick.

We thank the following people for providing us with editorial comments and opinions or proofreading of the early manuscript: Fred Cohen, Silvia Crivelli, Krystof Fidelis, Barry Honig, Peter Karp, Tack Kuntz, Ken Merz, Andy McCammon, Julia Rice, Jeff Skolnick, and Jon Sorenson. Dr. Head-Gordon acknowledges the support from the U.S. Department of Energy through the MICS and LDRD programs, contract number DE-AC-03-76SF00098, as well as from research start-up monies from the University of California, Berkeley.

Cited references

1. T. Schlick, "Computational Molecular Biophysics Today: A Confluence of Methodological Advances and Complex Biomolecular Applications," *Journal of Computational Physics* **151**, No. 1, 1–8 (1999).
2. T. F. Smith and M. S. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology* **147**, 195–197 (1981).

3. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov Models in Computational Biology: Applications to Protein Modeling," *Journal of Molecular Biology* **235**, No. 5, 1501–1531 (1994).
4. L. Parida, A. Floratos, and I. Rigoutsos, "An Approximation Algorithm for Alignment of Multiple Sequences Using Motif Discovery," *Journal of Combinatorial Optimization* **3**, 247–275 (1999).
5. Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjolander, R. C. Underwood, and D. Haussler, "Stochastic Context-Free Grammars for T-RNA Modeling," *Nucleic Acids Research* **22**, 5112–5120 (1994).
6. J. S. Liu and C. E. Lawrence, "Bayesian Inference on Biopolymer Models," *Bioinformatics* **15**, 38–52 (1999).
7. L. Grate, M. Herbster, R. Hughey, I. S. Mian, H. Noller, and D. Haussler, "RNA Modeling Using Gibbs Sampling and Stochastic Context-Free Grammars," *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology* (1994), pp. 138–146.
8. M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics*, Oxford University Press, London (1999).
9. D. Fischer, D. Rice, J. U. Bowie, and D. Eisenberg, "Assigning Amino Acid Sequences to 3-Dimensional Protein Folds," *FASEB Journal* **10**, 126–136 (1996).
10. J. U. Bowie, R. Luthy, and D. Eisenberg, "A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure," *Science* **253**, 164–170 (1991).
11. M. J. Sippl, "Calculation of Conformational Ensembles from Potentials of Mean Force: An Approach to the Knowledge-Based Prediction of Local Structures in Globular Proteins," *Journal of Molecular Biology* **213**, 859–883 (1990).
12. N. N. Alexandrov, R. Nussinov, and R. M. Zimmer, "Fast Protein Fold Recognition via Sequence to Structure Alignment and Contact Capacity Potentials," *Pacific Symposium on Biocomputing*, L. Hunter and T. E. Klein, Editors (1996), pp. 53–72.
13. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge (1998).
14. S. R. Eddy, "Profile Hidden Markov Models," *Bioinformatics* **14**, 755–763 (1998).
15. D. Sankoff and J. B. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley Publishing Co., Reading, MA (1983).
16. A. Sali and T. L. Blundell, "Definition of General Topological Equivalence in Protein Structures: A Procedure Involving Comparison of Properties and Relationships Through Simulated Annealing and Dynamic Programming," *Journal of Molecular Biology* **212**, 403–428 (1990).
17. K. Sjölander, K. Karplus, M. P. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler, "Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology," *Computer Applications in the Biosciences* **12**, 327–345 (1996).
18. A. Martin, M. MacArthur, and J. Thornton, "Assessment of Comparative Modelling in CASP2," *Proteins: Structure, Functions, and Genetics* **29**, No. S1, 14–28 (1997).
19. L. Holm and C. Sander, "Fast and Simple Monte Carlo Algorithm for Side Chain Optimization in Proteins: Application to Model Building by Homology," *Proteins: Structure, Functions, and Genetics* **14**, 213 (1992).
20. K. Fidelis, P. Stern, D. Bacon, and J. Moult, "Comparison of Systematic Search and Database Methods for Constructing Segments of Protein Structure," *Protein Engineering* **7**, 953–960 (1994).
21. R. Samudrala and J. Moult, "Determinants of Side Chain Conformational Preferences in Protein Structures," *Protein Engineering* **11**, 991–997 (1998).
22. M. A. Martí-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali, "Comparative Protein Structure Modeling of Genes and Genomes," *Annual Review of Biophysics and Biomolecular Structure* **29**, 291–325 (2000).
23. P. Horton, "A Branch and Bound Algorithm for Local Multiple Alignment," *Proceedings of the Pacific Symposium on Biocomputing*, Hawaii, USA, World Scientific Publishing Co., Singapore (1996), pp. 368–383.
24. R. H. Lathrop and T. F. Smith, "Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Potentials," *Journal of Molecular Biology* **255**, 641–665 (1996).
25. R. H. Lathrop, "The Protein Threading Problem with Sequence Amino Acid Interaction Preferences Is NP-Complete," *Protein Engineering* **7**, 1059–1068 (1994).
26. H. Flöckner, M. Braxenthaler, P. Lackner, M. Jaritz, M. Ortner, and M. J. Sippl, "Progress in Fold Recognition," *Proteins* **23**, 376–386 (1995).
27. Y. Duan and P. A. Kollman, "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution," *Science* **282**, 740–744 (1998).
28. E. Boczko and C. Brooks, "First-Principles Calculation of the Folding Free Energy of a Three-Helix Bundle Protein," *Science* **269**, 393–396 (1995).
29. C. Brooks, "Simulations of Protein Folding and Unfolding," *Current Opinion in Structural Biology* **8**, 222–226 (1998).
30. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *Journal of the American Chemical Society* **117**, 5179 (1995).
31. A. D. MacKerell, Jr., B. Brooks, C. L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, "CHARMM: The Energy Function and Its Parameterization with an Overview of the Program," *The Encyclopedia of Computational Chemistry* **1**, P. v. R. Schleyer et al., Editors, John Wiley & Sons, Chichester (1998), pp. 271–277.
32. G. Nemethy, M. S. Pottle, and H. A. Scheraga, "Energy Parameters in Polypeptides, 9. Updating of Geometrical Parameters, Non-Bonding Interactions and Hydrogen Bonding Interactions for Naturally Occurring Amino Acids," *Journal of Physical Chemistry* **87**, 1883–1887 (1983).
33. D. M. York, A. Wlodawer, L. G. Pedersen, and T. A. Darden, "Atomic-Level Accuracy in Simulations of Large Protein Crystals," *Proceedings of the National Academy of Sciences (USA)* **91**, 8715–8718 (1994).
34. P. Procacci, T. Darden, and M. Marchi, "A Very Fast Molecular Dynamics Method to Simulate Biomolecular Systems with Realistic Electrostatic Interactions," *Journal of Physical Chemistry* **100**, 10464–10468 (1996).
35. V. Daggett and M. Levitt, "Protein Folding/Unfolding Dynamics," *Current Opinion in Structural Biology* **4**, 291–295 (1994).
36. A. Li and V. Daggett, "Investigation of the Solution Structure of Chymotrypsin Inhibitor 2 Using Molecular Dynamics: Comparison to X-ray Crystallographic and NMR Data," *Protein Engineering* **8**, 1117–1128 (1995).
37. J. Moult, "Comparison of Database Potentials and Molec-

ular Mechanics Force Fields," *Current Opinion in Structural Biology* **7**, 194–199 (1997).
38. W. Kauzmann, "Some Factors in the Interpretation of Protein Denaturation," *Advances in Protein Chemistry* **14**, 1–63 (1959).
39. L. Wesson and D. Eisenberg, "Atomic Solvation Parameters Applied to Molecular Dynamics of Proteins in Solution," *Protein Science* **1**, 227 (1992).
40. J. Vila, R. L. Williams, M. Velasquez, and H. A. Scheraga, "Empirical Solvation Models Can Be Used to Differentiate Native from Near-Native Conformations of Bovine Pancreatic Trypsin Inhibitor," *Proteins* **10**, 199 (1991).
41. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics," *Journal of the American Chemical Society* **112**, 6127 (1990).
42. B. Honig, K. A. Sharp, and A.-S. Yang, "Macroscopic Models of Aqueous Solutions: Biological and Chemical Applications," *Journal of Physical Chemistry* **97**, 1101 (1993).
43. J. M. Sorenson, G. Hura, A. K. Soper, A. Pertsemlidis, and T. Head-Gordon, "Determining the Role of Hydration Forces in Protein Folding," invited feature article for *Journal of Physical Chemistry B* **103**, 5413–5426 (1999).
44. A. Rahman and F. H. Stillinger, "Improved Simulation of Liquid Water by Molecular Dynamics," *Journal of the American Chemical Society* **95**, 7943 (1973).
45. See J. Sorenson, G. Hura, R. M. Glaeser, and T. Head-Gordon, "What Can X-ray Scattering Tell Us About the Radial Distribution Functions of Water?" submitted to *Journal of Chemical Physics* (2000) and references therein.
46. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, "Interaction Models for Water in Relation to Protein Hydration," *Intermolecular Forces*, B. Pullman, Editor, D. Reidel Publishing Company, Dordrecht (1981), p. 331.
47. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of Simple Potential Functions for Simulating Liquid Water," *Journal of Chemical Physics* **79**, 926 (1983).
48. M. W. Mahoney and W. L. Jorgensen, "A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions," *Journal of Chemical Physics* **112**, 8910 (2000).
49. Y.-P. Liu, K. Kim, B. J. Berne, R. A. Friesner, and S. W. Rick, "Constructing *Ab Initio* Force Fields for Molecular Dynamics Simulations," *Journal of Chemical Physics* **108**, 4739 (1996).
50. A. A. Chialvo and P. T. Cummings, "Engineering a Simple Polarizable Model for the Molecular Simulation of Water Applicable over Wide Ranges of State Conditions," *Journal of Chemical Physics* **105**, 8274 (1996).
51. G. Corongiu and E. Clementi, "Liquid Water with an *Ab Initio* Potential-X-ray and Neutron Scattering from 238K to 368K," *Journal of Physical Chemistry* **97**, 2030–2038 (1992).
52. I. M. Svishchev, P. G. Kusalik, J. Wang, and R. J. Boyd, "Polarizable Point-Charge Model for Water-Results under Normal and Extreme Conditions," *Journal of Chemical Physics* **105**, 4742 (1996).
53. B. Chen, J. Xing, and J. I. Siepmann, "Development of Polarizable Water Force Fields for Phase Equilibrium Calculations," *Journal of Physical Chemistry B* **104**, 2391 (2000).
54. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford (1987).
55. M. F. Crowley, T. A. Darden, T. E. Cheatham III, and D. Deerfield, "Fine- and Coarse-Grain Parallel AMBER and Particle Mesh Ewald on MPP's," *Parallel Computing for Industrial and Scientific Applications*, J. Jenness, Editor, Morgan Kaufmann Publishers, San Francisco (1998).
56. T. E. Cheatham III and B. R. Brooks, "Recent Advances in Molecular Dynamics Simulation Towards Realistic Representation of Biomolecules in Solution," *Theoretical Chemistry Accounts* **99,** 279–288 (1998).
57. J. Hoye and G. Stell, "Dielectric Theory for Polar Molecules with Fluctuating Polarizability," *Journal of Chemical Physics* **73**, 461–468 (1980).
58. L. R. Pratt, "Effective Field of a Dipole in Non-Polar Polarizable Fluids," *Molecular Physics* **40**, 347–360 (1980).
59. J. S. Cao and B. J. Berne, "Theory and Simulation of Polar and Nonpolar Polarizable Fluids," *Journal of Chemical Physics* **99**, 6998 (1993).
60. S. J. Stuart and B. J. Berne, "Effects of Polarizability on the Hydration of the Chloride Ion," *Journal of Physical Chemistry* **100**, 11934–11943 (1999).
61. G. J. Martyna, "Adiabatic Path Integral Molecular Dynamics Methods. I. Theory," *Journal of Chemical Physics* **104**, 2018 (1996).
62. J. S. Cao and G. J. Martyna, "Adiabatic Path Integral Molecular Dynamics Methods. II. Algorithms," *Journal of Chemical Physics* **104**, 2028 (1996).
63. H. C. Andersen, "Molecular Dynamics Simulations at Constant Pressure and/or Temperature," *Journal of Chemical Physics* **72**, 2384–2393 (1980).
64. R. Car and M. Parrinello, "Unified Approach for Molecular Dynamics and Density-Functional Theory," *Physical Review Letters* **55**, 2471 (1985).
65. G. J. Martyna, D. J. Tobias, and M. L. Klein, "Constant Pressure Molecular Dynamics Algorithms," *Journal of Chemical Physics* **101**, 4177 (1994).
66. M. E. Tuckerman, G. Martyna, D. J. Tobias, and M. L. Klein, "Explicit Reversible Integrators for Extended Systems Dynamics," *Molecular Physics* **87**, 1117 (1996).
67. S. Samuelson, D. J. Tobias, and G. Martyna, "Modern Computational Methodology Applied to the Simulation of Blocked Trialanine Peptide in Vacuo, Water Clusters, and Bulk Water," *Journal of Physical Chemistry* **101**, 7592 (1997).
68. M. E. Tuckerman, G. J. Martyna, and B. J. Berne, "Reversible Multiple Time Scale Molecular Dynamics," *Journal of Chemical Physics* **97**, 1990–2001 (1992).
69. T. Schlick, E. Barth, and M. Mandziuk, "Biomolecular Dynamics at Long Timesteps: Bridging the Timescale Gap Between Simulation and Experimentation," *Annual Review of Biophysics and Biomolecular Structure* **26**, 179–220 (1997).
70. E. Barth and T. Schlick, "Overcoming Stability Limitations in Biomolecular Dynamics: I. Combining Force Splitting via Extrapolation with Langevin Dynamics in LN," *Journal of Chemical Physics* **109**, 1617–1632 (1998).
71. E. Barth and T. Schlick, "Extrapolation Versus Impulse in Multiple-Timestepping Schemes: II. Linear Analysis and Applications to Newtonian and Langevin Dynamics," *Journal of Chemical Physics* **109**, 1632–1642 (1998).
72. C. Bartels and M. Karplus, "Multidimensional Adaptive Umbrella Sampling: Applications to Main Chain and Side Chain Peptide Conformations," *Journal of Computational Chemistry* **18**, 1450–1462 (1997).
73. S. O. Samuelson and G. J. Martyna, "Two Dimensional Umbrella Sampling Techniques for the Computer Simulation Study of Helical Peptides at Thermal Equilibrium: The 3k (i) Peptide in Vacuo and Solution," *Journal of Chemical Physics* **109**, 11061–11073 (1998).
74. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen,

and P. A. Kollman, "Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method," *Journal of Computational Chemistry* **16**, 1339–1350 (1995).
75. M. E. Tuckerman and G. J. Martyna, "Understanding Modern Molecular Dynamics: Techniques and Applications," *Journal of Physical Chemistry B* **104**, 159–178 (2000).
76. M. Head-Gordon, "Quantum Chemistry and Molecular Processes," *Journal of Physical Chemistry* **100**, 13213–13225 (1999).
77. P. M. W. Gill, "Molecular Integrals over Gaussian Basis Functions," *Advances in Quantum Chemistry* **25**, 141–163 (1994).
78. C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, "Linear Scaling Density Functional Calculations via the Continuous Fast Multipole Method," *Chemical Physics Letters* **253**, 268–278 (1996).
79. L. Greengard and V. Rokhlin, "A Fast Algorithm for Particle Simulations," *Journal of Computational Physics* **135**, 280–292 (1997).
80. X.-P. Li, R. W. Nunes, and D. Vanderbilt, "Density-Matrix Electronic-Structure Method with Linear System-Size Scaling," *Physical Review B* **47**, 10891 (1993).
81. G. Galli and M. Parrinello, "Large-scale Electronic Structure Calculations," *Physical Review Letters* **69**, 3547 (1992).
82. A. D. Becke, "Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior," *Physical Review A* **38**, 3098–3100 (1988).
83. J. P. Perdew, J. A. Chaevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, "Atoms, Molecules, Solids, and Surfaces—Applications of the Generalized Gradient Approximation for Exchange and Correlation," *Physical Review B* **46**, 6671 (1992).
84. C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density," *Physical Review B* **37**, 785–789 (1988).
85. S. Saeb and P. Pulay, "Local Treatment of Electron Correlation," *Annual Review of Physical Chemistry* **44**, 213 (1993).
86. R. J. Bartlett, "Coupled-Cluster Approach to Molecular Structure and Spectra: A Step Toward Predictive Quantum Chemistry," *Journal of Physical Chemistry* **93**, 1697 (1989).
87. R. Car and M. Parrinello, "Unified Approach for Molecular Dynamics and Density-Functional Theory," *Physical Review Letters* **55**, 2471–2474 (1985).
88. K. Laasonen and M. L. Klein, "*Ab Initio* Molecular Dynamics Study of Hydrochloric Acid in Water," *Journal of the American Chemical Society* **116**, 11620 (1994).
89. K. Laasonen, M. Sprik, M. Parrinello, and R. Car, "*Ab Initio* Liquid Water," *Journal of Chemical Physics* **99**, 9080 (1993).
90. A. Monge, E. J. P. Lathrop, J. R. Gunn, P. S. Shenkin, and R. A. Friesner, "Computer Modeling of Protein Folding: Conformational and Energetic Analysis of Reduced and Detailed Protein Models," *Journal of Molecular Biology* **239**, 995–1012 (1995).
91. K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, "A Test of Lattice Protein Folding Algorithms," *Proceedings of the National Academy of Sciences (USA)* **92**, 325–329 (1995).
92. F. H. Stillinger, "Diffusion Smoothing," *Physical Review B* **32**, 3134–3141 (1985).
93. J. Kostrowicki and H. A. Scheraga, "Application of the Diffusion Equation for Global Optimization to Oligopeptides," *Journal of Physical Chemistry* **96**, 7442–7449 (1992).
94. D. Shalloway, "Application of the Renormalization Group to Deterministic Global Minimization of Molecular Conformation Energy Functions," *Journal of Global Optimization* **2**, 281–311 (1992).
95. A. Roitberg and R. Elber, "Modeling Sidechains in Peptides and Proteins: Application of the Locally Enhanced Sampling and the Simulated Annealing Methods to Find Minimum Energy Conformations," *Journal of Chemical Physics* **95**, 9277–9287 (1991).
96. C. A. Laughton, "A Study of Simulated Annealing Protocols for Use with Molecular Dynamics in Protein Structure Prediction," *Protein Engineering* **7**, 235–241 (1994).
97. S. M. Legrand and K. M. Merz, "The Genetic Algorithm and the Conformational Search of Polypeptides and Proteins," *Molecular Simulation* **13**, 299–320 (1994).
98. Z. Li and H. A. Scheraga, "Monte-Carlo-Minimization Approach to the Multiple Minima Problem in Protein Folding," *Proceedings of the National Academy of Sciences (USA)* **84**, 6611–6615 (1987).
99. R. E. Bruccoleri, "Application of Systematic Conformational Search to Protein Modeling," *Molecular Simulations* **10**, 151–174 (1993).
100. R. Abagyan and M. Totrov, "Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins," *Journal of Molecular Biology* **235**, 938–1002 (1994).
101. T. Head-Gordon, J. Arrecis, and F. H. Stillinger, "A Strategy for Finding Classes of Minima on a Hypersurface: Implications for Approaches to the Protein Folding Problem," *Proceedings of the National Academy of Sciences (USA)* **88**, 11076–11080 (1991).
102. T. Head-Gordon and F. H. Stillinger, "Predicting Polypeptide and Protein Structures from Amino Acid Sequence: Antlion Method Applied to Melittin," *Biopolymers* **33**, 293–303 (1993).
103. S. Crivelli, T. M. Philip, R. Byrd, E. Eskow, R. Schnabe, R. C. Yu, and T. Head-Gordon, "A Global Optimization Strategy for Predicting α-Helical Protein Tertiary Structure," *Computers & Chemistry* **24,** Nos. 3–4, 489–497 (2000).
104. J. M. Blaney and J. S. Dixon, "A Good Ligand Is Hard to Find: Automated Docking Method," *Perspectives in Drug Discovery and Design* **1**, 301–319 (1993).
105. I. D. Kuntz, E. C. Meng, and B. K. Shoichet, "Structure-Based Molecular Design," *Accounts of Chemical Research* **27**, 117–123 (1994).
106. T. P. Lybrand, "Ligand-Protein Docking and Rational Drug Design," *Current Opinion in Structural Biology* **5**, 224–228 (1995).
107. M. Totrov and R. Abagyan, "Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates," *Proteins: Structure, Function, and Genetics*, Supplement **1**, 215–220 (1997).
108. T. J. A. Ewing and I. D. Kuntz, "Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening," *Journal of Computational Chemistry* **18**, 1175–1189 (1997).
109. D. M. Lorber and B. K. Shoichet, "Flexible Ligand Docking Using Conformational Ensembles," *Protein Science* **7**, 938–950 (1998).
110. A. Canning, A. De Vita, G. Galli, F. Gygi, F. Mauri, and R. Car, "Quantum Molecular Dynamics on Massively Parallel Computers," *Cray User Group Fall Proceedings*, Tours, France (1994), p. 18.
111. *NWChem: A Computational Chemistry Package for Parallel Computers*, Version 3.3.1, High Performance Computational

Chemistry Group, Pacific Northwest National Laboratory, Richland, WA (1998).
112. G. Kresse and J. Furthmüller, "Efficiency of *Ab-Initio* Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set," *Computational Material Science* **6**, 15–50 (1996).
113. G. Kresse and J. Furthmüller, "Efficient Iterative Schemes for *Ab Initio* Total-Energy Calculations Using a Plane-Wave Basis Set," *Physical Review B* **54**, 11169–11186 (1996).
114. A. Arkin, J. Ross, and H. H. McAdams, "Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage-Infected Escherichia Coli Cells," *Genetics* **149**, 1633–1648 (1998).
115. H. H. McAdams and A. Arkin, "It's a Noisy Business! Genetic Regulation at the Nanomolar Scale," *Trends in Genetics* **15**, 65–69 (1999).
116. H. H. McAdams and A. Arkin, "Towards a Circuit Engineering Discipline," *Current Biology* **10**, 318–320 (2000).
117. H. H. McAdams and A. Arkin, "Simulation of Prokaryotic Genetic Circuits," *Annual Review of Biophysics and Biomolecular Structure* **27**, 199–224 (1998).
118. M. Ehlde and G. Zacchi, "Mist: A User-Friendly Metabolic Simulator," *Computer Applications in the Biosciences* (1995).
119. P. Mendes, "GEPASI: A Software Package for Modelling the Dynamics, Steady States and Control of Biochemical and Other Systems," *Computer Applications in the Biosciences* **9**, 563–571 (1993).
120. H. H. McAdams and L. Shapiro, "Circuit Simulation of Genetic Networks," *Science* **269**, 650–656 (1995).
121. R. Somogyi and C. A. Sniegoski, "Modeling the Complexity of Genetic Networks: Understanding Multigenic and Pleiotropic Regulation," *Complexity* **1**, 45–63 (1996).
122. T. Gaasterland, "Structural Genomics: Bioinformatics in the Driver's Seat," *Nature Biotechnology* **16**, 625–627 (1998).

Accepted for publication September 22, 2000.

Teresa Head-Gordon *Department of Bioengineering, University of California, Berkeley, California 94720-1762 (electronic mail: TLHead-Gordon@lbl.gov).* Dr. Head-Gordon is an assistant professor in the Department of Bioengineering, University of California Berkeley, and a faculty scientist at Lawrence Berkeley National Laboratory. She received her Ph.D. degree in 1989 from Carnegie Mellon University, and was a postdoctoral researcher at AT&T Bell Laboratories during 1990–1992. Her research program encompasses simulation, theory, and experimentation in the area of protein folding, structure prediction, and aqueous hydration of biological systems.

John C. Wooley *University of California, San Diego, Mail Code 0043, 9500 Gilman Drive, La Jolla, California 92037-0043 (electronic mail: jwooley@ucsd.edu).* Dr. Wooley is Associate Vice Chancellor for Research at the University of California, San Diego (UCSD), an adjunct professor in pharmacology and in chemistry and biochemistry, and a Senior Fellow of the San Diego Supercomputer Center. He received his Ph.D. degree in 1975 at The University of Chicago, working with Al Crewe and Robert Uretz in biological physics. Dr. Wooley created the first programs within the U.S. federal government for funding research in bioinformatics and in computational biology, and has been involved in strengthening the interface between computing and biology for more than a decade. His current research involves bioinformatics and structural genomics, and his principal objectives at the UCSD are to stimulate new research initiatives, especially in the application of computational and information technology to science.