Preface

Basic understanding in the life sciences is being greatly advanced by the computationally intensive analysis of biological molecules, organisms, and ecosystems. Combined with the new molecular-based approach to the understanding of biological processes, such analysis is not only advancing research in the life sciences, but is also changing the nature of business in this field. The results will touch the lives of all of us.

This issue of the *IBM Journal of Research and Development* and a companion issue of the *IBM Systems Journal* (Volume 40, Number 2) contain papers that provide a view of some of the key aspects of this computationally intensive, molecular-based approach to the life sciences. The papers have been divided, keeping in mind the following: The *IBM Systems Journal* traditionally covers the fields of computer science, software systems, and architecture; emphasis in the *IBM Journal of Research and Development* is on the chemical and physical sciences, hardware systems, and the engineering of information technology.

A very important application of the computationally intensive, molecular-based approach is in the diagnosis and treatment of human disease. One disease which is currently the subject of considerable study is cystic fibrosis (CF). Work on this disease serves as a useful illustration of how multiple disciplines and techniques can function together to further our understanding. The disease is being studied using gene expression analysis, X-ray crystallographic methods, computational chemistry, rational drug design, combinatorial chemistry, and high-throughput screening techniques, all of which are represented in the papers of these journal issues.

The cover depicts ribbon diagrams pertaining to portions of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) protein, produced using information from the Protein Data Bank[1] and software developed in IBM. Defects in CFTR cause cystic fibrosis, one of the most common fatal genetic diseases, currently affecting approximately 55 000 people in the United States and Europe. Affected individuals have abnormally thick lung secretions and frequently succumb to respiratory infections. Despite continued progress in medical treatment, average survival remains at about 31 years of age.

The identification of the gene that codes for CFTR and recognition of its responsibility for CF were announced in 1989 by a large group of collaborators led by Lap-Chee Tsui and John R. Riordan of the Hospital for Sick Children in Toronto and by Francis S. Collins, then at the University of Michigan and now Director of the Human Genome Project at the National Institutes of Health.[2]

The CFTR protein is a *transmembrane* protein, embedded in the cell membrane of certain types of cells, where it mediates the transport of chloride ions and water molecules. Finding a cure for CF may rest partially on the full characterization of the structure, folding mechanism, and molecular function of the CFTR protein. However, transmembrane proteins are nearly impossible to crystallize, precluding the use of X-ray crystallography to determine their three-dimensional structures. Structure-prediction methods based on the similarity of a protein's amino acid sequence to those with known structures are only partially revealing, since the database of known protein structures contains very few examples of membrane-bound proteins.

The most common CFTR defects associated with CF occur within a 214-amino-acid subregion of the full 1480-amino-acid protein. This subregion is known as the first nucleotide binding domain. The largest of the three structures shown on the cover is a *theoretical* prediction of the structure of this domain, reported in 1996 by a group using database-intensive and computationally intensive methods.[3] The most common defect observed in people who have CF results in the absence of a phenylalanine amino acid at the 508 position (ΔF508), indicated in red. It is believed that this absence may affect the *rate of formation* as well as the *structure* of the final folded state.

The structure of the first nucleotide binding domain is in dispute. Nuclear magnetic resonance (NMR) experiments performed on 25- and 26-amino-acid subregions of the domain that span position 508 have yielded experimental structures for these peptides both with and without the phenylalanine, and in two different solvents.[4] The helical structures shown on the cover were produced using the solvent trifluorethanol (TFE), which is known to promote helical structure in short peptide chains.

The green, red, and yellow portions of the ribbon diagrams on the cover indicate corresponding regions. In contrast to the theoretical prediction, the above experimental results suggest that position 508 is in the middle of a helical region. However, it is quite common for a short sequence to adopt one structure in a protein, and a very different one when removed from the protein. (Solvent effects also play a key role in determining the shape of the folded structure.) These facts may explain

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucl. Acids Res.* **28,** 235–242 (2000). Data on the structures used for producing the cover of this issue of the *IBM Journal of Research and Development* (from left to right) are available from the Protein Data Bank under entries PDB ID: 1CKW, 1CKX, and 1NBD.

[2] J. R. Riordan, J. M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J.-L. Chou, M. L. Drumm, M. C. Iannuzzi, F. S. Collins, and L.-C. Tsui, "Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA," *Science* **245,** 1066–1073 (1989).

[3] Flip J. Hoedemaeker, Alan R. Davidson, and David R. Rose, "A Model for the Nucleotide-Binding Domains of ABC Transporters Based on the Large Domain of Aspartate Aminotransferase," *Proteins: Struct. Funct. Genet.* **30,** 275–286 (1998).

[4] Michael A. Massiah, Young-Hee Ko, Peter L. Pedersen, and Albert S. Mildvan, "Cystic Fibrosis Transmembrane Conductance Regulator: Solution Structures of Peptides Based on the Phe508 Region, the Most Common Site of Disease-Causing F508 Mutation," *Biochemistry* **38,** 7453–7461 (1999).

the discrepancies between the experimental and theoretical structures. It is very important not only to determine the actual structure, but to improve and gain confidence in our ability to make theoretical structural predictions.

Although it has been characterized from a genomics perspective, it is not yet known exactly how the ΔF508 defect causes the disease, nor is it known how to correct for the defect. A detailed molecular-level understanding of the function of the protein is still lacking. Treatments for CF today generally address symptoms and side effects, rather than the underlying protein defect and cell functioning. Advances in therapeutics have been hampered to some extent by the absence of detailed structural data on the CFTR protein. Also, an understanding of the protein folding *mechanism* may provide a key to the problem, but the CFTR protein is very large, and study by today's methods presents too great a challenge. Nonetheless, the intense medical research being applied to CF, the greater understanding of it brought by the general trend toward a molecular-based understanding of life's processes, together with the improved computational, data-measurement, and other information tools available, should result in continuous advances in the treatment of the disease.

A very innovative scientific and business relationship was announced in February 2001 between the Cystic Fibrosis Foundation, a nonprofit organization committed to research on CF, and Structural GenomiX involving $11 million to fund a five-year project to produce a full three-dimensional structure for the CFTR protein.[5]

Work in the life sciences cuts across many disciplines, as described in the overview paper by Swope, "Deep computing in the life sciences," in the companion issue of the *IBM Systems Journal*. That paper presents a short description of the biological processes and information flow that lead from DNA to gene expression, to protein folding, and finally to drug design. It also indicates how the papers in the issues of both journals contribute to an understanding of the overall biological process. Hence, in this preface we mention only briefly the papers which appear here. The papers fall into four categories: simulation and study of very large molecular systems, DNA or protein sequence analysis, protein structure, and drug design.

The paper by Morokuma et al., "Model studies of the structures, reactivities, and reaction mechanisms of metalloenzymes," describes how the techniques of quantum chemistry can be used to efficiently model the reactivity of very large molecules such as the metalloenzymes by using different levels of sophistication for different parts of complex molecular systems. Density-functional theory (DFT) is a quantum-mechanics-based approach used by physical chemists to model molecules and solid-state systems. In "DFT-based molecular dynamics as a new tool for computational biology: First applications and perspective," Andreoni et al. show how to apply the method to the study of proteins and enzymatic processes. This includes the use of molecular dynamics driven by DFT gradients as well as new approaches that include a partitioning of the system under study into classical and quantum-mechanical domains.

The use of X-ray techniques to generate electron densities for the computation of molecular properties of large molecules is described in "Quantum crystallography, a developing area of computational chemistry extending to macromolecules" by Huang et al. The paper describes how electron densities of very large molecular systems can be obtained from those of molecular fragments.

Computational methods have been applied to structures much larger than even large molecules. Large and important cellular structures are now being modeled. To characterize and simulate biological membranes, the techniques of classical molecular dynamics are used in "Interfacing molecular dynamics with continuum dynamics in computer simulation: Toward an application to biological membranes" by Ayton et al. Similarly, Baker et al. simulate a hollow cylindrical cellular structure called a microtubule in their paper, "The adaptive multilevel finite element solution of the Poisson–Boltzmann equation on massively parallel computers." That work was carried out on a massively parallel IBM RS/6000® SP supercomputer.

Three papers deal with genomic or proteomic analysis. "Brute force estimation of the number of human genes using EST clustering as a measure" by Davison and Burke uses expressed sequence tags (ESTs) to estimate the number of possible human genes. The machine learning technique of hidden Markov models, which has been applied to speech recognition, is applied to DNA and protein sequence analysis in "Hidden Markov models in biological sequence analysis" by Birney. "DELPHI: A pattern-based method for detecting sequence similarity" by Floratos et al. describes an approach to searching a database of proteins for those with amino acid sequences similar to a query sequence. Their procedure makes use of sequence patterns that are identified in the underlying database during a preprocessing step.

Four papers describe work related to the prediction of the structure of the folded state of proteins. "Evaluating protein structure-prediction schemes using energy landscape theory" by Eastwood et al. discusses the funnel-like nature of energy functions used for protein structure prediction as a means to assess the usefulness of scoring such functions. "Protein flexibility and electrostatic interactions" by Kumar et al. analyzes the implications of salt bridges and electrostatic interactions for protein

[5] An interesting and recent account of the relationships forming in the business world to address the cystic fibrosis disease may be found at the Forbes web site: *http://www.forbes.com/forbes/2001/0402/080_print.html.*

folding. "A hierarchical, building-block-based computational scheme for protein structure prediction" by Tsai et al. describes a procedure for predicting protein structure that is based on the identification of local building block elements which drive the conformational assembly process. The paper "Determination of optimal Chebyshev-expanded hydrophobic discrimination function for globular proteins" by Fain et al. describes the design of a potential energy function which can be used to evaluate, or score, hypothetical protein structures.

The last two papers in the issue describe mathematical techniques which are useful in designing drugs. "QSAR in grossly underdetermined systems: Opportunities and issues" by Platt et al. describes an approach to quantitative structure activity relationships (QSARs) using regression analysis as a method for drug design. In "Multiobjective optimization of combinatorial libraries," Agrafiotis describes powerful combinatorial chemistry library-design techniques for designing drugs more efficiently.

In summary, the papers in this issue of the *IBM Journal of Research and Development* present some of the leading techniques of chemistry, physics, and mathematics as they apply to biomolecular structures and processes important in the life sciences.

We are grateful to Sharon Nunes, IBM Life Sciences Solutions, for her contributions both to this issue and the companion issue of the *IBM Systems Journal,* and to Frank Suits, IBM Research Division, for his help with the cover. He produced the ribbon diagrams on the cover using information from the Protein Data Bank[6] and software he developed.

William C. Swope
IBM Almaden Research Center, California

James M. Coffin
IBM Life Sciences Solutions, Dallas, Texas

Barry Robson
IBM Thomas J. Watson Research Center, New York

Guest Editors

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucl. Acids Res.* **28,** 235–242 (2000).