Journal of Computer-Aided Molecular Design, 1 (1987) 283–290
ESCOM

J-CAMD 024

Future in biomolecular computation

E. Wimmer

Cray Research Inc., 1333 Northland Drive, Mendota Heights, MN 55120, U.S.A.

Key words: Biomolecular computation; Quantum mechanical calculations; Force-field calculations; Protein folding; Parallel computers; Distributed processing

SUMMARY

Large-scale computations for biomolecules are dominated by three levels of theory: rigorous quantum mechanical calculations for molecules with up to about 30 atoms, semi-empirical quantum mechanical calculations for systems with up to several hundred atoms, and force-field molecular dynamics studies of biomacromolecules with 10,000 atoms and more including surrounding solvent molecules. It can be anticipated that increased computational power will allow the treatment of larger systems of ever growing complexity. Due to the scaling of the computational requirements with increasing number of atoms, the force-field approaches will benefit the most from increased computational power. On the other hand, progress in methodologies such as density functional theory will enable us to treat larger systems on a fully quantum mechanical level and a combination of molecular dynamics and quantum mechanics can be envisioned. One of the greatest challenges in biomolecular computation is the protein folding problem. It is unclear at this point, if an approach with current methodologies will lead to a satisfactory answer or if unconventional, new approaches will be necessary. In any event, due to the complexity of biomolecular systems, a hierarchy of approaches will have to be established and used in order to capture the wide ranges of length-scales and time-scales involved in biological processes. In terms of hardware development, speed and power of computers will increase while the price/performance ratio will become more and more favorable. Parallelism can be anticipated to become an integral architectural feature in a range of computers. It is unclear at this point, how fast massively parallel systems will become easy enough to use so that new methodological developments can be pursued on such computers. Current trends show that distributed processing such as the combination of convenient graphics workstations and powerful general-purpose supercomputers will lead to a new style of computing in which the calculations are monitored and manipulated as they proceed. The combination of a numeric approach with artificial-intelligence approaches can be expected to open up entirely new possibilities. Ultimately, the most exciding aspect of the future in biomolecular computing will be the unexpected discoveries.

I. INTRODUCTION

It is always exciting to look into the future, especially in a young field such as computation in molecular biology. At the same time, it is hard or almost impossible to make accurate predictions in this rapidly evolving field as to where it is going in the future. Therefore, the nature of this contribution will be rather speculative and extrapolations will be made on the basis of today's interests and approaches that may well change entirely as the field develops. Similarly, the prediction

of future computational hardware, software, and network environments as they will influence computation in biology will be extrapolated from the trends we have been witnessing over the past decade or so. Again, the actual development may be quite different, although at closer look, hardware developments based on silicon technology have been following a rather straightforward path over the last two decades and this trend may well continue. On the other hand, new materials such as GaAs, ceramic superconductors, or optical switching devices could change hardware performance in a radical way. A major impact on the way we use this hardware will come from new generations of software. To this end, we just have to think how multiple-window interactive environments compare with punch-card batch processing that we did not all that long ago.

The present paper has been stimulated by the discussion of a panel consisting of David A. Case (Research Institute of Scripps Clinic), William A. Goddard III (California Institute of Technology), Cyrus Levinthal (Columbia University), Michael L. Liebman (Mount Sinai School of Medicine), Gerald M. Maggiora (Upjohn Company), J. Andrew McCammon (University of Houston), and U. Chandra Singh (Research Institute of Scripps Clinic). The theme of this panel discussion, which was chaired by Erich Wimmer (Cray Research Inc.), was 'Future in Biomolecular Computation' and represented a part of the conference 'Supercomputing in Biology', held at the University of Minnesota, September 13-16, 1987. This contribution should not be considered as a detailed report on the actual panel discussion, but rather as a reflection on some of the thoughts that have been brought up during the discussion.

II. CURRENT STATUS

We are currently witnessing the creation of a new branch of science: the computational approach is establishing itself as truly new discipline besides the traditional branches of experimentation and theoretical/analytical theory. This was first seen in areas such as mechanics, fluid dynamics, climatology, chemistry, and solid state physics. In these fields, large-scale computer simulations, relying on classical and quantum mechanics, allow the study of complex systems at an unprecedented level of realism. The goal of all these efforts is a quantitative simulation of complex 'real-world' systems in terms of better resolution in space and time, and in terms of encompassing more and more of the environment. This goal is fundamentally different from the analytic approach that tries to isolate, decompose, and idealize systems. The computational approach is synthetic in nature: the goal is the simulation of a complex system and the study of its behavior in a realistic environment.

Computational approaches to biological systems pursue the same goals: the modeling of proteins, DNA, membranes, tissues, etc. at an ever growing level of complexity with the inclusion of more and more of the surrounding environment such as water, counterions, and lipid bilayers. Simulations of the dynamics over longer and longer time intervals aim at the exploration of conformational domains of biomacromolecules, thus providing insight into these system at a level of spacial and time resolution that is inaccessible by experiment. Ultimately, new systems can be explored and their behavior and responses studied before they are actually synthesized.

Let us now review the current state of the most important computational approaches that are being employed in the study of biomolecular systems on the atomic and molecular scale. We can discriminate between two kinds of approaches: (i) the phenomenological approach such as DNA sequence analysis and comparison; and (ii) a first-principles approach to the molecular structure,

dynamics, and reactivity of biomolecules. Within the second kind, which will be mostly discussed below, we see three levels of theory. At the most fundamental level we are dealing with rigorous quantum mechanical approaches such as *ab initio* Hartree-Fock theory and its refinements. Current computational hardware, software, and algorithms allow the treatment of isolated molecules with approximately up to 30 atoms if the system has no symmetry elements such as mirror planes. Taking advantage of symmetry, systems with more than 150 atoms have been investigated on the ab initio level [1]. For these systems, determination of ground state conformations and the energetics of deformations (rotational barriers, vibrational properties) as well as the distribution of electronic charges (including quantities such as dipole moments and polarizabilities) can predict experimental values to within 1% for structural properties and to within 10% for vibrational frequencies. (Usually, this error is systematic and can be accounted for in an ad hoc manner to reproduce experimental data within a few percent.) A rather annoying numerical property of present-day algorithms stands in the way of moving on to larger systems: if N characterizes the size of a molecule the computational requirements grow with N^4 or even a higher power than 4. Thus, a 1000-fold increase in computational power [2] allows us only to increase the size of the molecule by a factor of about 6. Furthermore, if we are interested in the ground-state conformation of a molecule, direct energy minimizations via quantum mechanical calculations are likely to yield just a local minimum and the scanning of the entire conformational space becomes an added difficulty. Despite these difficulties, ab initio calculations are now becoming possible on organic molecules with specific biological activity. For example, many molecules used as drugs, herbicides, and pesticides and very small fragments of bio-polymers can now be treated on the ab initio level. Experience on large molecules shows [1] that the actually observed growth in computational requirement scales, on the Hartree-Fock level, with N^3 rather than with N^4 because of vanishing integrals associated with distant atoms in the molecule.

The situation is similar if we employ semiempirical quantum mechanical approaches. The scaling in the molecular size is reduced to a third power, but the problem of the conformational space remains. At this level of theory, today's technologies allow calculations on molecules with maybe several hundred atoms. This means that most of the commercially used organic molecules with biological activity can be readily calculated and the major residues of active sites in enzymes can be described on this quantum mechanical level. Due to the computational efficiency of semiempirical methods, these calculations can be carried out on supercomputers in an interactive mode with typical times of several minutes to optimize the geometry of molecules with 30 atoms.

Based on developments in solid state theory, density functional theory [3] is emerging as a new approach for molecular calculations. Present experience shows [4] that the accuracy of local density functional (LDF) calculations matches and sometimes surpasses single-determinant Hartree-Fock calculations. Importantly, LDF calculations scale with a third power in the number of basis functions and hence make molecules with hundred atoms and more accessible to first-principles quantum mechanical calculations. At present, the maturity of LDF computer programs for molecules does not match those of Hartree-Fock based codes. In particular, the analytic evaluation of first and second derivates of the total energy with respect to nuclear displacements as needed for geometry optimizations and vibrational frequency analysis, have yet to be implemented.

One of the most widely used methods in the computational approach to biomolecules is the force-field approach. All quantum mechanical interactions between atoms are incorporated in an analytic expression of the total energy of a molecule as a function of the nuclear positions. Given

such a description, the dynamics of molecules can be readily studied by solving Newton's equations of motion. The evaluation of pair-wise interactions between non-bonded atoms becomes the computational bottleneck. From this fact one might expect that molecular dynamics algorithms scale as N^2 in the number of atoms. This gives rise to some optimism as computational hardware become more powerful. Actually, this scaling looks even better if we employ a cut-off to the interactions between distant atoms. Thus, the scaling approaches a linear behavior for very large systems. Today, systems with 10 000 atoms and more are amenable to molecular dynamics studies over time-scales of 100 ps and more.

In addition, a thermodynamic-cycle free-energy perturbation approach [5] within molecular dynamics provides an elegant and powerful tool for the study of modifications of enzyme inhibitors and changes of amino acids in proteins. The application of this method to biomolecules is new and first results [5] are most encouraging.

So far, only numerical approaches in biomolecular computing have been discussed. It is, however, an open question if this approach will play the same dominant role as it does for example in fluid dynamics, or if non-numeric ('non-traditional') approaches will emerge as more appropriate for biological systems. Such approaches may be based on symbolic processing rather than numeric processing and may lead to the creation of rule-based expert systems as is currently being pursued by several research groups [6]. Related to this aspect is the question of the most appropriate computer languages which we will have to address.

Today, a range of computational hardware is being used for computations on biomolecules. This range goes from personal computers, workstations, departmental minicomputers and mini-supercomputers to the largest supercomputers available. In order to monitor the progress in hardware over the last decade, let us look at the high-end of the computational spectrum since it is reasonable to assume that all other computational machinery will develop accordingly. Today, the fastest general-purpose supercomputers perform scalar, vector, and parallel processing with a clock cycle time of about 4 ns. If we take the clock cycle as a rough measure of a processor performance, we observe that progress made during the last decade has advanced the cycle time by only about a factor of 3. This fact is one of the driving forces to increase the speed by other means such as parallel processing. Thus, this dimension is being explored in the hardware design by introducing parallel processors ranging from a few to thousands. In contrast to the cycle speed, main memory has been expanded by more than two orders of magnitude during the last decade. This development cause a re-thinking of many algorithms that have been used on small memory machines.

III. THE GRAND CHALLENGE AND THE FUTURE

One of the most urgent problems in biomolecular computing is the protein folding problem. Given the sequence of the amino acids in a protein, what secondary and tertiary structure does the protein assume in its final form? At this point, no closed solution has been formulated and it is a matter of controversy which approach should be taken. Can the 'rational' approach such as molecular dynamics or variations thereof lead to success or do we need radically different ways? The answer is open at this point. Furthermore, is the folding of the entire, uncoiled, solvated protein really capturing all aspects or do we have to pay close attention to the ribosome and follow the path of the protein synthesis?

Is the computational technology in place so that we can attempt a direct approach to the pro-

tein folding using molecular dynamics? And even if we succeed for a special case, what have we learned and gained? It is not clear at this point whether the answers to these questions are imminent or afar in the future.

As computational power increases, there is an obvious development that can be anticipated: we will do the current things bigger and better. In terms of molecular dynamics, the immediate development is clearly to move to larger systems, maybe entire viruses, and to study them for longer periods of time, i.e., moving into the nano-second time scale. (Incidentally, within the next few years the time-scale of molecular dynamics simulations will exceed the clock period of the fastest supercomputers, viz., 1 ns.) Furthermore, there is an obvious need to improve the force-field parameters, and also to include effects such as electronic polarization. To this end, the quantum mechanical approach and the quasi-classical formulation have to brought closer together.

The need for linking quantum mechanics and molecular dynamics as it is done already today, for example, in the QUEST program [7], points to a development which gains high priority: biological systems are characterized by an enormously high degree of complexity. Hence, no single approach will be able to cover all aspects and we need an integrated hierarchy of approaches: accurate quantum mechanics in the regions where bonds are made and broken, an atomistic quasi-classical description for particle sizes up to the order of 100 000, joined by a continuum-like theory to capture the larger structures.

Another theoretical possibility is given by an intimate combination of quantum mechanics and molecular dynamics as formulated by Car and Parrinello [8]. This approach has been applied recently to SiO_2 crystals [9] and it would be most interesting to see if this methodology can be transferred to biomacromolecules.

Similarly to the length-scale, we need such a hierarchy also for the time-scale, which may be the intellectually even more challenging task. The length-scale of biological systems ranges from 10^{-10} to 1m (i.e., eleven orders of magnitude) whereas the time scale goes from 10^{-15} s for describing the motions and vibrations of atoms to minutes for protein folding (and years for biological live-cycles). Hence, a time scale covering about 18 orders of magnitude awaits unification.

Today, the focus is on isolated peptides and DNA macromolecules, sometimes surrounded by water molecules and ions. Living systems, though, are characterized by enormously complex cooperative phenomena. As methods for the treatment of individual biomacromolecules become mature, it can be anticipated that the interest will shift to these cooperative phenomena including processes on various biological interfaces. In practical applications such as drug design, the scope will broaden from the active site, where a key-lock mechanism is the most widely used concept, to the entire cycle of a drug within a human body including resorption, delivery, action, decomposition, and removal. A massive task for modeling is ahead of us and many parameters in such a cycle have yet to be captured quantitatively and brought into the form of a coherent theoretical description that lends itself to the algorithmic implementation on computer hardware.

There is general agreement that real progress in biomolecular computing will come from the development of novel approaches and methodologies. Although necessary, increase in computational power alone will not be sufficient. Furthermore, the computational approach to the understanding and manipulation of biomolecules has to go hand in hand with the experimental approach. In fact, there is an urgent need for scientists who understand, or even better, pursue both approaches, the experimental as well as the computational. This can be done only if the computational tools are becoming more readily accessible and easier to use which brings us back to future developments in hardware, software, and languages.

IV. FUTURE HARDWARE, SOFTWARE, AND COMPUTATIONAL TOOLS

Large-scale scientific computing in the 1970s was dominated by a batch mode: large programs, usually written in FORTRAN, were submitted to a mainframe. Without any possibility by the user to interact with the program during execution, the program read a (usually complex and cumbersome) input deck that specified the entire path for this run. After many hours (or days), the run reached completion and left long output files that were then inspected and analyzed. Graphics was used mostly in the postprocessing phase on devices such as pen plotters.

Molecular scientists were among the first to appreciate the value of graphical output as well as input. Thus, in the early 1980s, the Evans & Sutherland picture systems became intimately linked to molecular modeling. This development was enhanced by the availability of convenient and fairly powerful minicomputers such as the ubiquitous VAX. A number of excellent molecular modeling packages became available on the market that made use of this concept. In terms of software, computing in biology was dominated by FORTRAN as language and VMS as operating system. At the same time, powerful general-purpose supercomputers became available to a number of research groups which triggered the extensive use of molecular dynamics and accurate ab initio calculations on relatively large molecules. Still, the operational mode was of a batch character, but with enhanced use of graphical pre- and post-processing. Today, this is probably still the most common use of large-scale computing in molecular biology. A major change in the use of high-speed computational machines can be observed: there is a trend from background batch processing to interactive distributed processing. Several components were necessary to facilitate this development:

(1) the speed of the supercomputers is becoming sufficient to perform calculations on large systems in a time frame that is acceptable for interactive work.

(2) General-purpose graphics workstations such as APOLLO, SUN, and Silicon Graphics IRIS provide the necessary graphics capabilities as well as a convenient, multiwindow environment for carrying out a number of textual tasks. Furthermore, networking concepts and protocols such as TCP/IP are becoming available that allow convenient distribution and communication of processes.

(3) Interactive operating systems and high bandwidth communication channels are becoming standard on high-end computational machines. The supercomputer thus becomes a partner on the network rather than a batch-background processor.

(4) UNIX is evolving as a unifying operating system on workstations, minicomputers, mini-supercomputers, and supercomputers.

(5) Languages such as C are starting to compete with FORTRAN in newly written application packages.

It is safe to assume that this trend of interactive, distributed processing will continue in the future. Presently, we are just at the beginning. In terms of computational hardware, we can expect an almost continuous spectrum (and price range) from personal computers to supercomputers. Scalar speed will continue to grow, albeit progress is slow as we are starting to approach the physical limits in silicon-based device technology. Introduction of other device materials such as GaAs will lead to improvements, yet no quantum leaps can be expected. Vector processing will become standard in machines of all sizes. The way to enhanced hardware performance is parallel processing. Time will tell how we can work around Amdahl's law [10] which tells us that a seemingly small

non-parallel fraction of an algorithm very quickly becomes the bottleneck in a parallel-processing environment. It certainly would be naive to assume that 10 000 processors in parallel automatically give us 10 000-fold capability. Very hard work will be necessary to develop advanced algorithms that can make effective use of such architectures. Nevertheless, there is reason for hope since nature is showing us in the form of the eye, for example, that high-level parallelism with relatively slow single processors can lead to high system performance. In fact, visualization tools may well be the first category of hardware where this kind of parallelism will pay off.

Today, FORTRAN, C, and PASCAL languages dominate computing in molecular biology. The development of sophisticated compilers and precompilers will help to maintain the importance of these languages for computationally demanding tasks. In the future, another layer, on top of these languages could be developed. The major characteristics of such a language would be that it would be closer to the scientific-mathematical notation and thinking of the scientist [11]. For example, there is no deep reason why a computer language should not have as its character set the entire set of mathematical symbols with all its richness. The bottleneck so far were the limitations of ASCII terminals. General-purpose graphics workstations are certainly capable of handling a much richer character set. In addition, such a language could actually be an expert system, knowing, for example, all the current mathematical facts and helping the scientist in choosing the best algorithmic implementation. In such a scenario, this superlanguage could produce as output an usual FORTRAN or C code which can be mixed with existing code and which can then benefit from the sophistication of existing FORTRAN or C compilers. Furthermore, the superlanguage expert system could be taught to choose the best algorithmic choice for a given hardware architecture. At this stage, the FORTRAN or C source code would become a throwaway object much like the object codes of today's compilers. The scientific productivity would be enhanced enormously because the tedious work on the FORTRAN source-code level would become obsolete. We should not underestimate the gigantic intellectual effort that would have to go into such a project and it may take quite a while before these software tools are becoming routine.

Earlier, we discussed the change from the batch mode to the interactive mode. Such a trend would be highly beneficial for computing in biology as it would allow us to 'watch the calculations taking place' and, more importantly, interact with the calculations and change the course as they are executed. On the other hand, the batch mode will not disappear, because there will always be problems, maybe the most important ones, that will challenge the capabilities and capacities of the most powerful computing systems available as well as the patience and endurance of the scientists carrying out these computations.

V. SUMMARY AND CONCLUSIONS

In summary, large-scale computation in molecular biology is emerging as a new branch in the life sciences and we are just at the beginning. Three types of molecular approaches are currently being pursued, ab initio quantum mechanical calculations on molecules with up to about 30-40 atoms, semiempirical quantum mechanical calculations for systems with up to several hundred atoms, and force-field calculations for molecular dynamics and thermodynamic cycle free energy calculations for biomacromolecules, including surrounding solvent molecules, for systems with up to 50 000 atoms. Due to the favorable scaling of the computational requirements within the force-field methods, it can be expected that enhanced computational power will have its largest

impact on this type of approach. On the other hand, methodological advances in the quantum mechanical treatment of molecules may well lead to important breakthroughs also on this level of theory.

It can be expected that methodologies for the treatment of single protein or DNA macromolecules will mature and the focus will shift to more complex interface problems such as membranes and protein/DNA interactions. Finally, using a hierarchy of a variety of approaches will enable us to aim at the simulation of entire living systems.

In terms of hardware, speed and power will continue to improve at a constantly increasing performance-to-price ratio for all classes of machines, from personal computers to supercomputers. Important changes can be expected due to distributed processing such as the intimate connection between general-purpose workstations and powerful supercomputers. We will watch the calculations taking place and will interact with the molecular system as its simulation is being carried out.

Advances in system software, high-level programming languages, and network protocols will make the systems easier to use, particularly for the mainly experimentally oriented scientist. As this new branch of molecular biology evolves, it can be hoped that it will gain recognition as a full partner to the experimental approach, thus enhancing our knowledge of the most exciting, but also most complex systems, namely living organisms. By its very nature, the future cannot be anticipated and the most intriguing promises lie in the unexpected. As we venture further into the beautifully complex world of living systems, we should never get carried away by 'hybris', but rather deepen our respect for the delicate balance and complexity of life.

REFERENCES

1 Almhof, J. and Luthi, H.P., In Jensen, K.F. and Truhlar D.G., (Eds.) Supercomputer Research in Chemistry and Chemical Engineering, ACS Symposium Series 353, American Chemical Society, Washington, D.C., 1987, p.35.

2 Recently (Cray Research Inc. press release November 18, 1987), Seymour Cray projected the performance of his CRAY-4, which is now in its initial planning stage, to be 1000 times that of the CRAY-1 delivered in 1976. If we assume a typical development cycle of 5 years, a factor of 1000 in throughput performance would have then been achieved within about two decades.

3 a. Hohenberg, P. and Kohn, W., Phys. Rev. B 136 (1964) 864–871.
b. Kohn, W. and Sham, L.J., Phys. Rev. A, 140 (1965) 1133–1138.

4 Wimmer, E., Freeman, A.J., Fu, C.-L., Cao, P.-L., Chou, S.-H. and Delley B., In Jensen, K.F. and Truhlar, D.G. (Eds.) Supercomputer Research in Chemistry and Chemical Engineering, ACS Symposium Series 353, American Chemical Society, Washington, D.C., 1987, p.49.

5 McCammon, J.A. and Harvey, S.C., Dynamics of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1987 (and references therein).

6 a. Wipke, W.T. and Hahn, M.A., Pierce T.H. and Hohne, B.A. (Eds.) Artificial Intelligence Applications in Chemistry. ACS Symposium Series 306, American Chemical Society, Washington, D.C., 1986, p. 136.
b. Klein, T.E., Huang, C., Ferrin, T.E., Langridge, R. and Hansch, C., ibid. p. 147.

7 Singh, U.C., QUEST 1.0 Users Manual, Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143, 1986.

8 Car, R. and Parrinello, M., Phys. Rev. Lett., 55 (1985) 2471–2474.

9 Allan, D.C. and Teter, M.P., Phys. Rev. Lett. 59 (1987) 1136–1139.

10 Amdahl, G., American Federation of Information Processing Society Conference Proceedings, Spring Joint Computer Conference 3 (1967) 483.

11 Wilson, K.G., In Matsen F.A. and Tajima, T. (Eds.) Supercomputers – Algorithms, Architectures, and Scientific Computing, University of Texas Press, Austin, 1986, p.431.