The birth of computational structural biology

Michael Levitt

Like Sydney Altman[1], I too was initially rejected by the renowned Medical Research Council (MRC) Laboratory of Molecular Biology in Cambridge, England. The year was 1967 and I was then in my final year of a B.Sc. degree in Physics at Kings College in London. Enthralled by John Kendrew's BBC 1964 television series "The Thread of Life", I wanted desperately to do my Ph.D. at the MRC in Cambridge. Alas there was no room for any new postgraduate students in 1967!

After some negotiations, I was accepted for the following year. More importantly, John Kendrew said that I should spend the intervening period at the Weizmann Institute in Israel with Shneior Lifson. Kendrew had just heard of Lifson's initial ideas[2] on the consistent force field (CFF), which was an attempt to simulate the properties of any molecular system from a simple potential energy function. He believed that these methods should be applied to protein and nucleic acid macromolecules. I arrived in Israel in October, 1967 and set to work programming the consistent force field under the supervision of Lifson and his Ph.D. student Arieh Warshel. At that time, computing at the Weizmann Institute was amongst the best in the world; in 1963 computer engineers there had built their own machine, appropriately known as the Golem, after the Jewish folklore automaton.

In a few short months we had a program called CFF that allowed us to calculate the energy, forces (energy first derivatives with respect to atomic positions) and curvature (energy second derivatives with respect to atomic positions) of any molecular system. Warshel went on to use the program to calculate structural, thermodynamic and spectroscopic properties of small organic molecules[3], while I followed Kendrew's dictum and applied these same programs to proteins. This led to the first energy minimization of an entire protein structure (in fact we did two, myoglobin and lysozyme) in a process that became known as energy refinement[4].

I began my Ph.D. at the MRC in Cambridge in September, 1968 and was immediately immersed in the annual tradition of Lab Talks. These talks by members of the three divisions at the Laboratory of Molecular Biology at that time (Structural Studies Division under Kendrew, the Cell Biology Division under Sydney Brenner and Francis Crick, and Protein and Nucleic Acid Chemistry Division under Fred Sanger) were a treat for newcomers to the Lab. The 'Molecule of the Year' was tRNA, which had been predicted to exist by Francis Crick 10 years before[5] and was now the subject of intense structural and genetic interest. I decided to try to build a model of tRNA and started off playing with CPK space-filling models at home. Transfer RNA has almost 2,000 atoms and a space-filling model weighs over 100 pounds. My most vivid memory is lowering the tRNA CPK model from the first floor window of our terrace cottage in Newnham, while my somewhat pregnant wife was having a hard time controlling her laughter. The model, which was then rebuilt from brass components, towered over me as I measured all atomic positions with a plumb line (a pointed metal weight hanging from a string onto graph paper) so that the model could be energy refined. Modeling tRNA led me to interact closely with both Crick and Aaron Klug and so I was exposed to the wonders of molecular and structural biology.

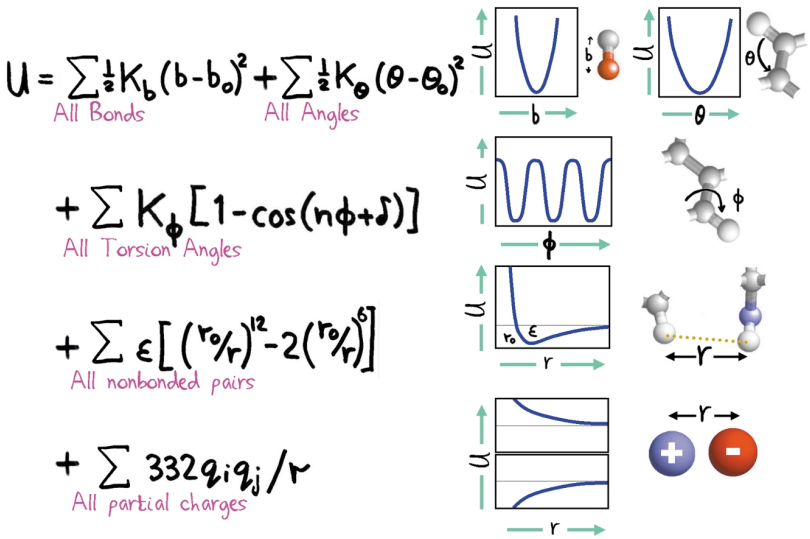

Fig. 1 The total potential energy of any molecule is the sum of simple allowing for bond stretching, bond angle bending, bond twisting, van der Waals interactions and electrostatics. Many properties of a biomolecules can be simulated with such an empirical energy function.

The model was published in 1969 (ref. 6) and I settled down to work on my thesis entitled "Conformation Analysis of Proteins"[7]. This was entirely devoted to computational biology and included chapters entitled "Energy Parameters from Proteins", "Interpreting Problematic Regions of Electron Density Maps Using Convergent Energy Refinement", "Energy Refinement of Enzyme/Substrate Complexes: Lysozyme and Hexa-N-Acetylglucosamine" and "Energy Refinement of Tertiary Structure Changes Caused by Oxygenation of Horse Haemoglobin".

Work on nucleic acids was not neglected and at that time it seemed that RNA folding would be easier to tackle than protein folding[8]. Computational work on protein folding began in 1973 during my postdoctoral research with Shneior Lifson back at the Weizmann Institute. Arieh Warshel had returned from his postdoc at Harvard and we started to work together again on both protein folding and enzyme reactions. Each project led to novel simulations[9,10] that became the basis for a great

deal of future work, with much still to be done a quarter of a century later.

I returned to a staff position at the MRC in Cambridge in October, 1974 and Warshel joined me there as a visitor. Warshel focused his attention on quantum mechanics in biology and published a model of the initial steps in the visual process, based on a molecular dynamics simulation[11]. Meanwhile I worked with Cyrus Chothia on the classification and analysis of protein architecture[12] and with Tony Jack, who passed away in 1978, on the refinement of large structures by simultaneous minimization of the molecular energy and crystallographic R-factor[13]. Both papers were to lead to significant future science: Chothia went on to develop the first web database, SCOP[14] and Axel Brünger based his wonderfully useful X-PLOR program[15] on Jack's work with me.

While Warshel and I were travelling around the world, our computer program, CFF, had wings of its own. Arieh Warshel took the program with him on his postdoctoral visit to Martin Karplus' lab at Harvard in 1969. In 1971, Bruce Gelin, who was just released from the US army, began working with Warshel and started writing a new version of the code. This rewrite was essential as I had learned my programming from an IBM FORTRAN II manual, whereas Bruce Gelin was much better trained. I can still recall my excitement when I saw his version of the program — many of the variable names were those I had invented but the code was so much more elegant!

Bruce Gelin's code led to his pioneering work with Andy McCammon and Martin Karplus on the simulation of protein dynamics[16]. This work, published in 1977, marks the start of the next phase of computational structural biology in that it signaled the linking of computational chemistry with biology. Work in the field was becoming much more widespread; the original program that I wrote with Arieh Warshel went on through Bruce Gelin's rewrite to form the basis of the next generation of programs including CHARMM (Chemistry at HARvard Molecular Mechanics) from Karplus' group at Harvard, AMBER from Peter Kollmann's group at UCSF and Discover from Arnold Hagler's company, Biosym.

Looking back to that period, it is much easier to appreciate who were the key contributors. Shneior Lifson, who passed away on 22 January, 2001, really started it all by defining the form of the empirical potential energy function still in use today (Fig. 1). In particular, he was the first to realize that the hydrogen bond could be described by simple electrostatic interaction of partial charges. With Warshel, he also set up a consistent procedure for deriving the energy parameters.

Sequence analysis, which forms such a key part of modern computational biology, was born in that same 1969–1977 period. In 1969, analysis of tRNA sequences revealed a correlated base change[6] (two bases not in a helical stem change together to maintain function, thereby indicating a possible interaction); in 1971, Needleman and Wunsch applied the computer science method of dynamic programming to sequence alignment[17]; and in 1977, Sanger and coworkers started genome-scale DNA sequencing with the φX-174 bacteriophage sequence[18].

I still remember with much chagrin that day in 1976 when Bart Barrell approached me to help analyze the φX DNA sequence only to be rebuffed; I felt that structure was just so much more interesting than sequence. Having confessed what may be the greatest misjudgment of my career, I would like to conclude with a few words about the future of computational biology.

Computers were made for biology: biology would never have advanced as it did without the dramatic increase in computer power and availability. One day we would like to be able to simulate complicated biological processes, perhaps even going from the genomic sequence to a full simulation of the organism's phenotype. In thinking about how to do this, it is interesting to compare Nature with simulated biology. Some things that are very difficult in Nature are trivial for computers: consider how much cellular machinery is needed to transcribe DNA sequence to RNA sequence — in the computer all one needs to do is change 'T' to 'U'. Translating RNA sequence to protein sequence is even more difficult in the cell, but in a computer one just applies the genetic code table. Other things that appear very easy for Nature are almost impossibly hard for computers: once synthesized a protein sequence spontaneously folds into the native structure, whereas simulating even a part of this process is still well beyond our computational capabilities. Computational structural biology will remain very challenging well into the 21st century.

Michael Levitt is in the Department of Structural Biology, Stanford University, Stanford, Calfornia 94305-5400, USA. email: michael.levitt@stanford.edu

1. Altman, S. *Nature Structural Biology* **7**, 827–828 (2000).
2. Bixon, M. & Lifson, S. *Tetrahedron* **23**, 769–784 (1967).
3. Lifson, S. & Warshel, A. *J. Chem. Phys.* **49**, 5116–5129 (1968)
4. Levitt, M. & Lifson, S. *J. Mol. Biol.* **46**, 269–279 (1969).
5 Crick, F.H.C. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
6. Levitt, M. *Nature* **224**, 759–763 (1969).
7. Levitt, M. Ph. D. Thesis *Conformation analysis of proteins* (Cambridge University, Cambridge, UK; 1971); *http://csb.stanford.edu/levitt/Levitt_Thesis_1971/Levitt_Thesis_1971.html.*
8. Levitt, M. In *Polymerization in Biological Systems Ciba Foundation Symposium* **7**, 146–171 (Eds Wolstenholme, G.E.W. & O'Connor, M., Elsevier, Amsterdam; 1972).
9. Levitt, M. & Warshel, A.*Nature* **253**, 694–698 (1975).
10. Warshel, A. & Levitt, M. *J. Mol. Biol.* **103**, 227–249 (1976).
11. Warshel, A. *Nature* **260**, 679–683 (1976).
12. Levitt, M. & Chothia, C. *Nature* **261**, 552–558 (1976).
13. Jack, A. & Levitt, M. *Acta Crystallogr. A* **34**, 931–935 (1978).
14. Murzin, A.G, Brenner, S.E., Hubbard, T. & Chothia C. *J. Mol. Biol.* **247**, 536–540 (1995).
15. Brünger, A.T., Karplus, M. & Petsko G.A. *Acta Crystallogr. A* **45**, 50–61 (1989).
16. McCammon, J.A., Gelin, B.R. & Karplus, M. *Nature* **267**, 585–590 (1977).
17. Needleman, S.B. & Wunsch, C.D. *J. Mol. Biol.* **48**, 443–453 (1970)
18. Sanger,F. *et al. Nature* **265**, 687–695 (1977).