# Theoretical technologies in an "experimental" setting: empirical modeling of proteinic objects and simulation of their dynamics within scientific collaborations around a supercomputer.

Frédéric Wieber (Archives Henri Poincaré – LHSP, Nancy Université, France)

frederic.wieber@wanadoo.fr

*Paper proposed for thematic session "The nature and epistemological status of the phenomena established through experimental practices", Meeting of the Society for Philosophy of Science in Practice, Minneapolis, June 18-21, 2009*

In this talk, I want to present an historical case study in the field of protein chemistry. This case study focuses on modeling and simulating practices within this field during the 1960's and 1970's. My aims are, first, to analyze the nature of the models that protein scientists have constructed during this period. I will specify here the scientists' epistemic aims for such a construction, and the theoretical, empirical and technological resources available to them. Secondly, I will briefly examine how a particular simulation method termed "Molecular Dynamics" (MD), which has been elaborated in statistical physics, has been adapted to proteins. Here, I will emphasize and discuss the collaboration between physicists and protein scientists that led to this adaptation. Globally, my point will be to show the deep interconnection of these modeling practices with empirical data, and the impact of computer technology on the evolving form of these models and on the way of collaborating adopted by scientists to adapt "Molecular Dynamics" simulation method to proteins. But before that, we have to remind us, schematically, what kind of molecular object proteins are, and what was the "in practice" epistemological situation of the theoretical approaches to proteins properties in the 1960's and 1970's.

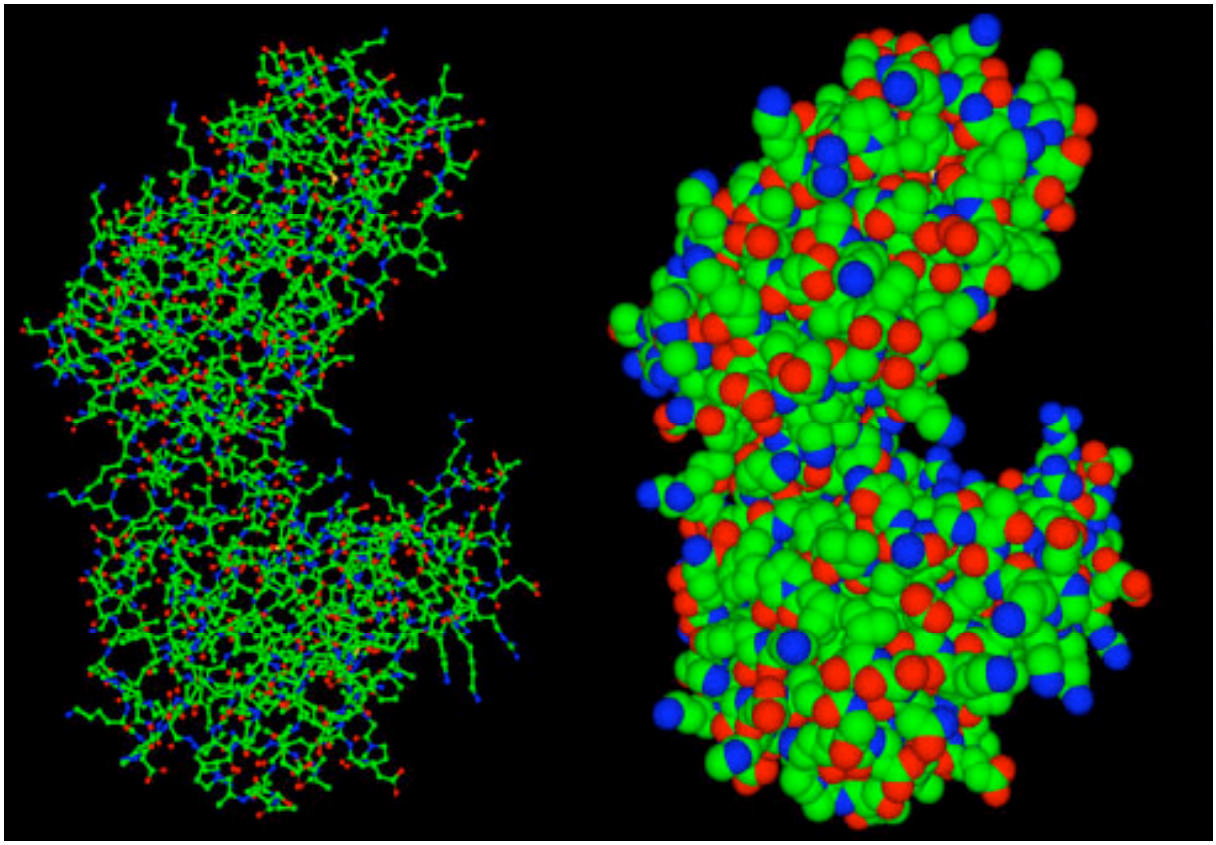
## 1/ The complexity of proteinic molecular objects and the epistemological situation of the theoretical approaches.

I will obviously not follow here any scientific textbook to present precisely all the numerous properties of proteins. For the purpose of this talk, it is probably sufficient to remind that proteins are biological *macro*molecules that play different fundamental functions within cells. These molecules are biopolymers, composed of repeating structural units named amino acids. Finally, most proteins have the property of naturally folding into a precisely defined three-

dimensional structure. (Scientists speak then of the native *conformation* of a protein. A conformation of a molecule is a particular three-dimensional arrangement of its atoms. For one protein, different conformations are theoretically possible. The native conformation is then one among the astronomical number of theoretically possible conformations of a protein).

What is really important here is to emphasize that proteins are typically made of thousands of atoms. This very huge number of atoms can be seen on the type of representations you can watch here ***[see figure 1].*** These representations show how complex the three-dimensional native structure of a protein is. It can be noted here that this structural complexity has immediately been recognized by the scientists (John Kendrew and Max Perutz) who, using X-rays scattering experiments, were able to propose in 1960 the first structure at atomic resolution of a protein.

***Figure 1: Two representations of a protein, namely phosphoglycreate kinase.***



This molecular complexity of proteinic objects (huge number of atoms, intricacy of the folded structure) helps us understanding the epistemological situation of theoretical approaches in

protein chemistry during the 1960's and 70's. If, in order to grasp and deal with the complexity of the structures experimentally produced, some theoretical approaches were needed, and called for by scientists, the theory that *in principle* governs the properties of proteins, just as for any other molecular objects, was nevertheless not applicable *in practice* because of computational intractability. As shown by many philosophers and historians of chemistry interested in the question of the possible reduction of chemistry to physics, the application of quantum mechanics to molecular systems has always been problematic and has led to increasingly complex and laborious computations. That explains the central character of computers in the culture and practices of quantum chemistry after World War II. As quantum theory was already very difficult to apply to simple molecular systems, its use, even in conjunction with the specific theoretical descriptions and computational procedures developed in quantum chemistry, was clearly seen, by scientists, as impracticable for proteins. So, the modeling practices that I want now to discuss have to be understood within this theoretical context.

## 2/ "Empirical modeling" of proteinic objects during the 1960's.

Before the 1960's, a relatively long tradition of modeling structure and possible conformations of proteins already existed, but it was, within this tradition, *material* molecular models that were constructed. Although material models were still used in the 1960's and 1970's, practices of *theoretical* modeling also emerge during the 1960's. As for all processes of emergence of a scientific practice, several factors can be put forward to understand that one. I will only mention here the scientists' epistemic aims that led to such an emergence. From the scientists' point of view, as noted above, the first need was the development of tools to analyze the great intricacy of the first structures experimentally obtained, and to test and refine the structures that were constructed by processing and interpreting X-rays data. But, secondly, there was also a hope: if sufficiently good theoretical models of proteins could be devised on the basis of structural experimental data already obtained, then it would be possible to predict the native conformation of proteins on the unique basis of a knowledge of their amino acids sequence. This would have potentially led to avoid the really laborious work to experimentally determine the structure of proteins. Moreover, this specific epistemic aim fitted in with the then current agenda of Molecular Biology. Molecular Biology was interested, within the so-called "central dogma", in an understanding of genetic information flow from the one-dimensional structure of DNA (the sequence of bases) to the three-dimensional structure of proteins.

So far, we have seen what epistemic aims led scientists to construct theoretical models of protein conformations. But what resources could they exploit for such a construction? As noted above, since the use of theoretical formulations from quantum mechanics lead to non-manageable equations, protein scientists have to find others theoretical resources. As the goal of constructing models of proteins was to gain knowledge of protein's conformations stability, protein scientists used a very simple theoretical formulation that has been proposed at the end of the 1940's, in order to understand the stereochemistry of organic compounds within the field of physical organic chemistry. It is not the place here to precisely discuss the origins, uses, transformations and diffusion of this theoretical formulation. It seems more interesting, for this talk, to write down this formulation in order to understand the characteristics exhibited by models of molecules based on this formulation. The formulation defines a potential energy for a molecule for every set of positions of the atoms, that is for every conformation, as follows:

$$\mathbf{E = \Sigma\ [u_0(r_0/r)^{12} - 2u_0(r_0/r)^6] + \Sigma\ \tfrac{1}{2}\ k_s(l - l_0)^2 + \Sigma\ \tfrac{1}{2}\ k_b(\theta - \theta_0)^2}$$

where l is a chemical bond distance, $\theta$ a bond angle, $k_s$ and $k_b$ are force constants, $l_0$ and $\theta_0$ are equilibrium values of the bond length and angle, r is the distance between two interacting atoms, and $-u_0$ is the minimum value of the interaction energy (at $r = r_0$). The sum, for the first term, is made over all pairs of non-bonded atoms. For the second and third terms, the sums are made, respectively, over all pairs of bonded atoms and over all bond angles.

So, this simple formulation, at the heart of the models of protein that were constructed, implies a particular representation of matter: molecules are constituted of valence-bonded 'atoms' (and not of nucleus and electrons as in quantum mechanics), and are roughly speaking represented by a system of balls connected by springs. This particular idealization shows that the question of the very accuracy of that type of protein models is not a priority for scientists. They obviously know that this representation is not accurate, but they adopt it precisely because it is useful, because it is the unique representation at hand that can lead to computationally manageable models. The validity of the models constructed on the basis of this formulation is thus above all pragmatic.

There is a second interesting and, for scientists, fundamental characteristic of this formulation. In order to construct a model of a protein (the theoretical formulation is obviously not, by itself, a model of protein), one has inevitably to fix the values of parameters appearing in the theoretical formulation. And the number of parameters is really important,

because a protein is made of different types of atoms and of chemical bonds. As the parameters used are *empirical parameters*, we have thus to note here, firstly, that this modeling strategy is called by scientists "empirical modeling", and, secondly and more importantly, that this modeling strategy is very dependent on the availability of the empirical data required. Different types of such data are needed: infrared spectroscopic data, crystallographic data, thermodynamic data *etc*… Thus, scientists who want to construct a model for a particular molecule must find what data are available for that molecule. Of course, all the data needed are never at hand for the particular molecule of interest, and they are then estimated and adjusted from the data available for other molecules. It is important here to stress that such modeling practice of molecular objects couldn't have been developed without the revolution of physical instrumentation in chemistry since the 1930's. But it is equally necessary to remind here the complexity of proteinic objects. Since these objects are constituted of a huge number of atoms, the use of physical instrumentation to obtain typical data for these molecules was very difficult; hence the amount of data necessary to parameterize a model of protein was really thin. The work of estimating and adjusting empirical data was thus more extensive in protein chemistry than in other sub-fields of chemistry. To conclude this point, we can stress that for constructing models of molecules within this modeling strategy, scientists had to exploit creatively some empirical resources. The parameterization is the central stage in the process of modeling, and it demands a good knowledge of empirical results in the field, and specific partially tacit epistemic skills to make and justify the choices and adjustments of data. Finally, different research teams, depending on their own research context, made these choices locally.

These local choices were nevertheless diffused in the 1970's. But to understand that point, it is necessary to turn now to the third resource that protein scientists used when constructing their "empirical models". This third resource is a technological one, namely computers. So, if a protein scientist has made the choice of using the theoretical formulation we have seen above, and has constructed a model for a molecule by choosing, estimating, adjusting different types of empirical data, he can now use this model to study the stability of some conformations. But to do all that, it is necessary to calculate the potential energy of one or several conformations. When the modeling strategy was used in organic chemistry in the 1950's without the help of computers, the task of calculating all the chemical bonds geometries and energies, and all the interactions between all pairs of non-bonded atoms, was still complex and really laborious. But a pencil and paper application of the method to bigger

organic molecules and, of course, to proteins was out of reach. The development and spread out of that type of modeling practices in protein chemistry (and more generally in chemistry) has thus been fully dependent of the use of computers. These practices would not have been efficient if these technological instruments of computation had not been available. But if computers were needed for that efficiency, the use of these calculating machines altered modeling practices in turn. So, to define precisely the characteristics of each atom inside a protein according to their molecular surrounding, scientists were able to use an increasingly large number of parameters stored in databanks, which the computer program could access quickly. A mode of calculation based on pencil and paper would not have allowed such increase in the number of parameters used for modeling, because it would not have been manageable.

But the "by now" computational nature of the modeling practices has also allowed a type of crystallization and spreading of the choices made locally concerning the empirical parameters, thanks to the construction and dissemination of computer programs packages useful for non-specialists. This modeling strategy is very local for it is suitable only for certain types of molecules when one has chosen and estimated locally empirical parameters. Yet, its computerization has allowed an increase in its generality, because of the large number of parameters suitable for more and more molecules that has been stored in computer programs. The use of computers has also allowed to improve the models constructed. With more and more effective calculations executed, scientists have been able to increasingly test the results produced, against empirical data, in order to optimize the parameters chosen for modeling.

So, the computational nature of these modeling practices is really fundamental. The computer, as a technological instrument, has influenced in a major way the form of the models that have been constructed; and that modeling practices has led to the emergence of a theoretical knowledge about proteins structure and stability. Therefore, I refer to these modeling practices as "theoretical technologies". If models can be seen as "epistemic tools", as some philosophers of science have put it, my expression emphasizes the significance of the concrete nature of those tools. The computational and hence technological nature of the tools presented here is obviously important: the technological characteristics of computers – the way they function, the limitations of their processing power, their accessibility for scientists – impacts the evolving epistemic status of that kind of modeling practices.

I turn now to the case of the adaptation of "Molecular Dynamics" simulation method to proteins. I will emphasize, in particular, how computer technology is an important factor for understanding the way scientists collaborated.

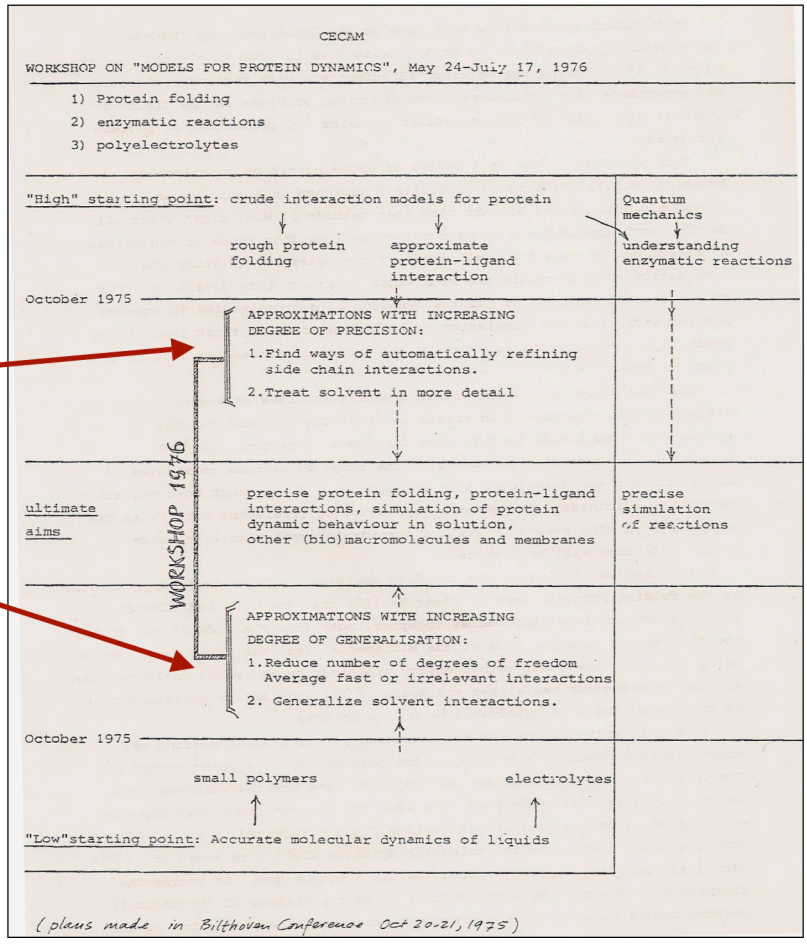
## 3/ Scientists collaboration around a supercomputer in order to simulate protein dynamics.

The aim of "Molecular Dynamics" simulations of proteins is to simulate the actual dynamics of that objects. This method was adapted to proteins in the late 1970's. The simulation process, which was used, was constructed before in statistical physics, in order to understand the behavior of liquids by calculating exactly, on the computers, the dynamics of the biggest possible number of interacting particles. Within this application scheme of Newtonian theory, one of the capital elements is indeed the modeling of interactions between particles. When this method was adapted to proteins, the models of these molecules, which had been constructed previously, was used again. So, that work of adaptation has constituted an enrichment of previous modeling practices.

As calculability in practice impacts fundamentally models that can be simulated, a certain number of approximations was introduced in order to apply the MD method to proteins. The situation is, here, really complicated. Thus, the realization of a simulation process could be assimilated to a complex hierarchy made of various theoretical representations that are useful for certain aims. As a result, introducing an approximation affecting the model is going to alter the subsequent levels and could call, for example, for the development of special algorithms. In this case, there are complex relations between the limitations of the processing power of the machine, the degree of complexity of the modelized and simulated object, the types of simplifications of the model that the scientists consider relevant, the theoretical expressions useful to write down the equations of motion, and the algorithms available and needed. Thus, in order to adapt the MD simulation to proteins, one needed to adjust the complex balance between all these factors. Such a complex adjustment was made possible by the convergence of specialists in the field of statistical physics and of specialists of "empirical modeling" in protein chemistry. A representation, made by one of the actors, of this convergence is reproduced on **figure 2**. We can see here what were, from the scientists' point of view, the different skills and aims ("approximations with increasing degree of precision" or "generalization") of the two groups of actors.

***Figure 2***

***Representation, made by one of the actors, of the convergence of protein scientists (up/top) and of statistical physicists (bottom/down)***




If computer technology affected the way MD simulation was adapted to proteins, its impact can equally be felt on the collaboration process of those actors. In order to understand this process, I studied the work done within a particular institution. It is that work that has led to the first dynamical simulation of a very little protein in 1976. This institution is CECAM, which stands for "Centre Européen de Calcul Atomique et Moléculaire". During the 1970's, when supercomputers were relatively rare (except for military purposes or in high energy physics), CECAM had relatively easy access to them. This center could thus attract computational scientists, thanks to its calculation instruments. In this context, long workshops were organized (it is the planning of such a workshop that is reproduced on figure 2). During these workshops, skills and tacit knowledge needed for performing simulations could be exchanged, as well as simulation software. I quote here briefly the testimony of one actor (the Dutch physicist Herman Berendsen): "The workshop was a worthwhile experience for all concerned. Anees [Aneesur Rahman, one of the founding fathers of MD simulation] shared his programs with us, patiently explained the details and helped us along with his many

practical hints and computational tricks". So, even in a computational science, where all knowledge and skills could perhaps be made explicit (it's fundamentally a science of algorithms!), we find tacit aspects of practices. But we have of course to remind us that these calculation methods were only being developed, at that time.

The process of collaboration of scientists as seen through the activities of this center must be understood according to the *evolutions* in computer technology. So, as computer became widespread (university computing centers mushroomed) and networking developed, scientific cooperation within the CECAM was altered. When simulation protocols became more stable and simulations spread in the scientific community, people started exchanging only ideas, during shorter workshops, instead of skills. This little story of workshops within CECAM clearly shows the localized/delocalized character of simulation practices, which has been emphasized by some historians and philosophers of science.

## 4/ Concluding remarks

I would like to conclude now with a few remarks. It is probably unnecessary to insist on the special technological nature of the theoretical tools developed within modeling and simulating practices in protein chemistry. Yet, it could be interesting to better clarify how these theoretical technologies have to be understood in an "experimental" setting. I have talked about a special theoretical context within which modeling practices have been developed. But, as we have seen, a good knowledge of the available empirical data and the specific epistemic skills to decide and justify the choices of data were necessary to develop "empirical" models of proteins. Moreover, because of the complexity of proteins, few of the data needed for modeling were, on the one hand, available. On the other hand, the intricacy of structural X-rays data called for the development of tools to analyze and test that data. The growth of modeling practices has thus to be understood in a special "experimental" context. Now, concerning simulating practices, the accessibility of supercomputers was a tricky problem for scientists. This situation can probably be compared to the problem of accessing big instruments. But within such centralized organizations, a merging of different epistemic cultures can be achieved, as it was the case with the collaboration and convergence that we have described. Here, we find the localized side of simulation practices. On the delocalized side nevertheless, simulation protocols became more stable and were partly black-boxed as computer software. This leads us to point out the possible comparison between exchange and spread of software, and circulation of experimental systems and instruments. Finally, concerning experimental practices in protein chemistry, the technological tools constructed

have been integrated, thanks to their computational nature, to the classical toolbox used by experimenters for processing and interpretating empirical data of molecular structure. This has led to a shift in the way proteins were regarded, from a static structure to a dynamic one.