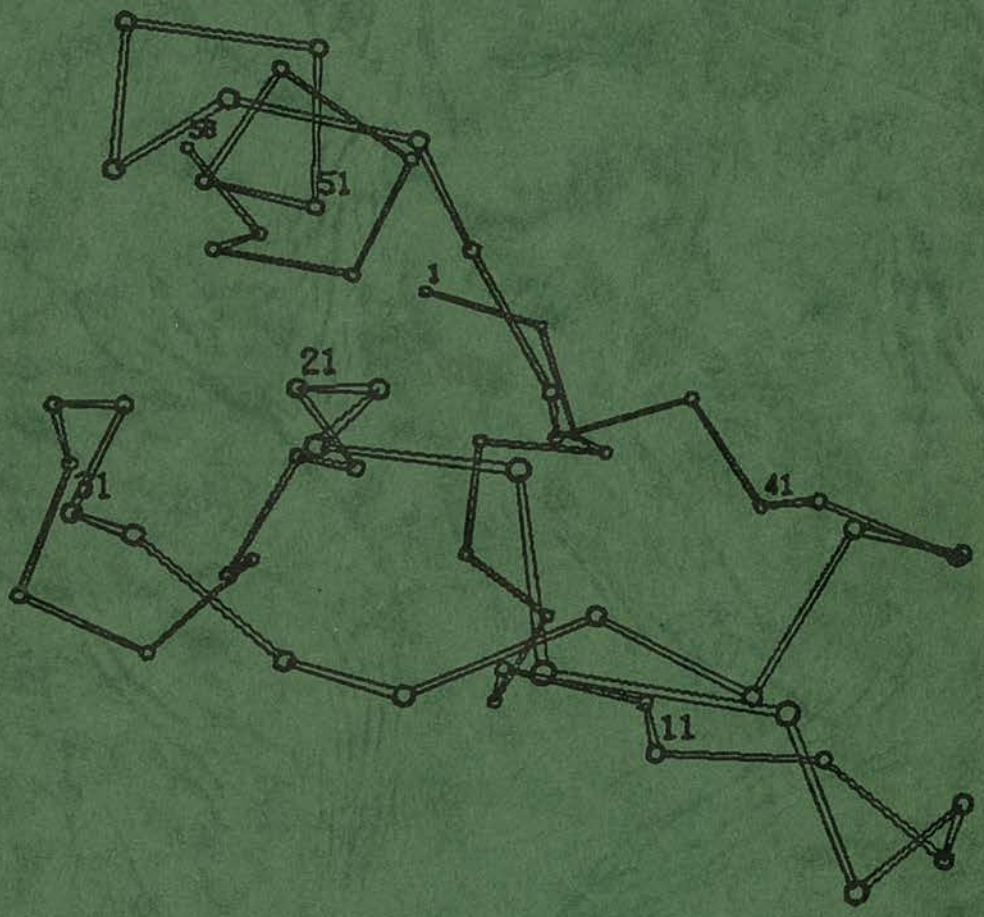
CECAM WORKSHOP

'MODELS FOR PROTEIN DYNAMICS'



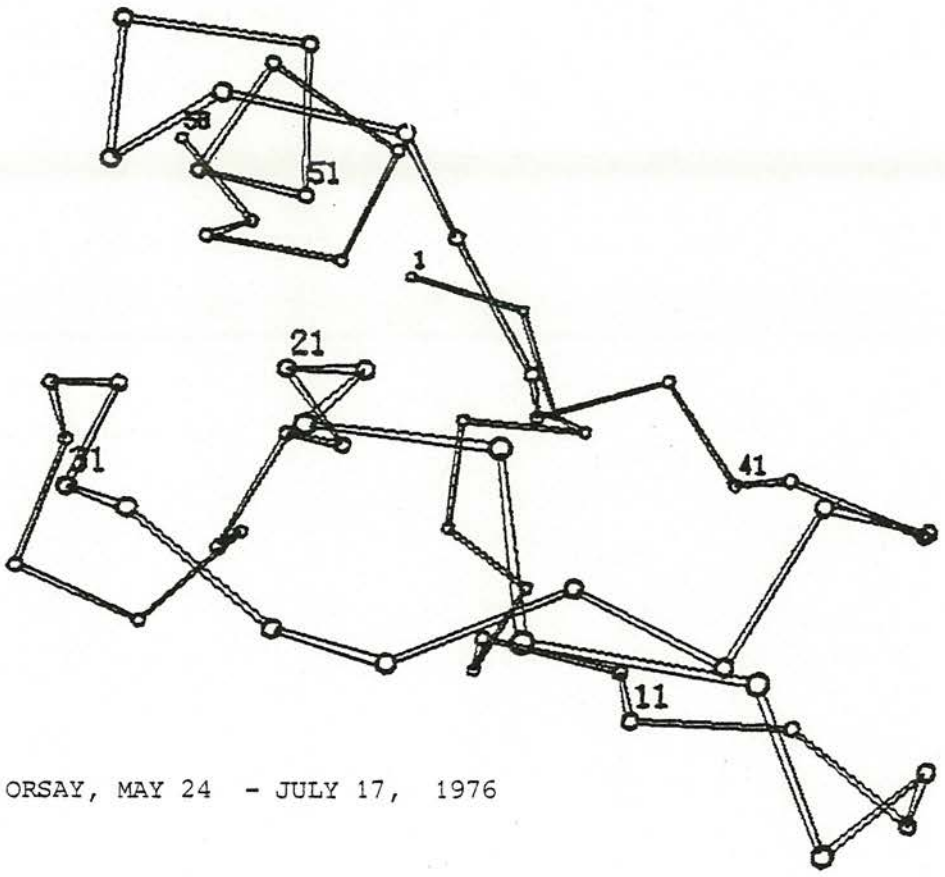
Orsay , May 24 . July 17 , 1976

CENTRE EUROPÉEN
DE CALCUL ATOMIQUE ET MOLÉCULAIRE

Bâtiment 506
UNIVERSITÉ DE PARIS XI
91405 ORSAY, FRANCE

CECAM WORKSHOP

'MODELS FOR PROTEIN DYNAMICS'



ORSAY, MAY 24 - JULY 17, 1976

LIST OF PARTICIPANTS

C.Bennett, IBM Watson Research Center, Yorktown Heights, New York,

H.J.C.Berendsen, Lab.of Physical Chemistry, the University of Groningen, the Netherlands,

G.Careri *, Istituto di Fisica "G.Marconi", Roma,

G.Ciccotti, Istituto di Fisica "G.Marconi", Roma,

C.Chothia, Institut Pasteur, Paris,

D.Elkkoubi, Lab. d'Electrochimie, Université P. et M. Curie, Paris,

A.Englert, Chimie générale, Université Libre, Bruxelles,

D.L.Ermak, Lawrence Livermore Lab., Livermore, California,

*D.R.Ferro**, Istituto di Chimica della Macromolecule, Milano, Italia,

W.F.van Gunsteren, Lab.of Physical Chemistry, the University of Groningen, the Netherlands,

J.Hermans, Dpt. of Biochemistry, University of North Carolina, Chapel Hill, North Carolina,

*M.Karplus***, Dpt. of Chemistry, Harvard University, Cambridge, Mass.,

M.Leclerc, Chimie générale, Université Libre, Bruxelles,

*M.Levitt**, MRC Lab. of Molecular Biology, Cambridge, England,

B.Maigret, Institut de Biologie Physico-Chimique, Paris V,

J.A.Mc.Cammon, Harvard University, Cambridge, Mass.,

K.Nagano, University of Tokyo, Japan,

J.Orban, Faculté des sciences, Université Libre, Bruxelles,

S.Prémilat, Lab. de Biophysique, Nancy, France,

A.Rahman, Argonne Natl. Lab., Argonne, Illinois,

P.Rossky, Harvard University, Cambridge, Mass.,

J.P.Rijckaert, Faculté des sciences, Université Libre, Bruxelles,

P.Turq, Lab. d'Electrochimie, Université P. et M. Curie, Paris,

S.Wodak, Dept. of Chemical Biology, Université Libre, Bruxelles.

*for part of the workshop

**visitor

Note. Many reports contain some work done in a few months after the workshop ended.

In some cases this has involved cooperation with non-participants of the workshop, who are then listed as coauthors of the report.

CONTENTS

INTRODUCTION

H.J.C.Berendsen

This report results from work done at the CECAM workshop "Models for Protein Dynamics",held at Orsay during 8 weeks from May 24 to July 17, 1976. During the workshop 23 scientists with various background and experience have collaborated and discussed problems in the field of computer simulation of macromolecular dynamics and prediction of protein structure.

This workshop is one in a series of CECAM workshops on Molecular Dynamics (M.D.), starting in 1972 with a workshop on liquid water. We vaguely anticipated at that time that methods of M.D. might eventually develop into applications to macromolecules including those of biological interest. But we were also well aware of the necessity to study the interaction of a macromolecule with water in appropriate detail in order to get trustworthy results for the dynamics of macromolecules in aqueous environments. Thus the simulation of liquid water was a first topic to be studied. The application to proteins was then not foreseen in five or ten years to come.

The simulation of macromolecular dynamics involves an enormous difficulty not encountered in simple liquids: The relevant motions spread over time scales ranging from fast bond vibrations (10^{-14}s) to folding and unfolding processes in the range of seconds to minutes. A small workshop was devoted to this problem in 1974. Although bond vibrations can now be eliminated, the general separation of irrelevant details in the simulation of slow events is still a formidable problem to which much future attention will be directed.

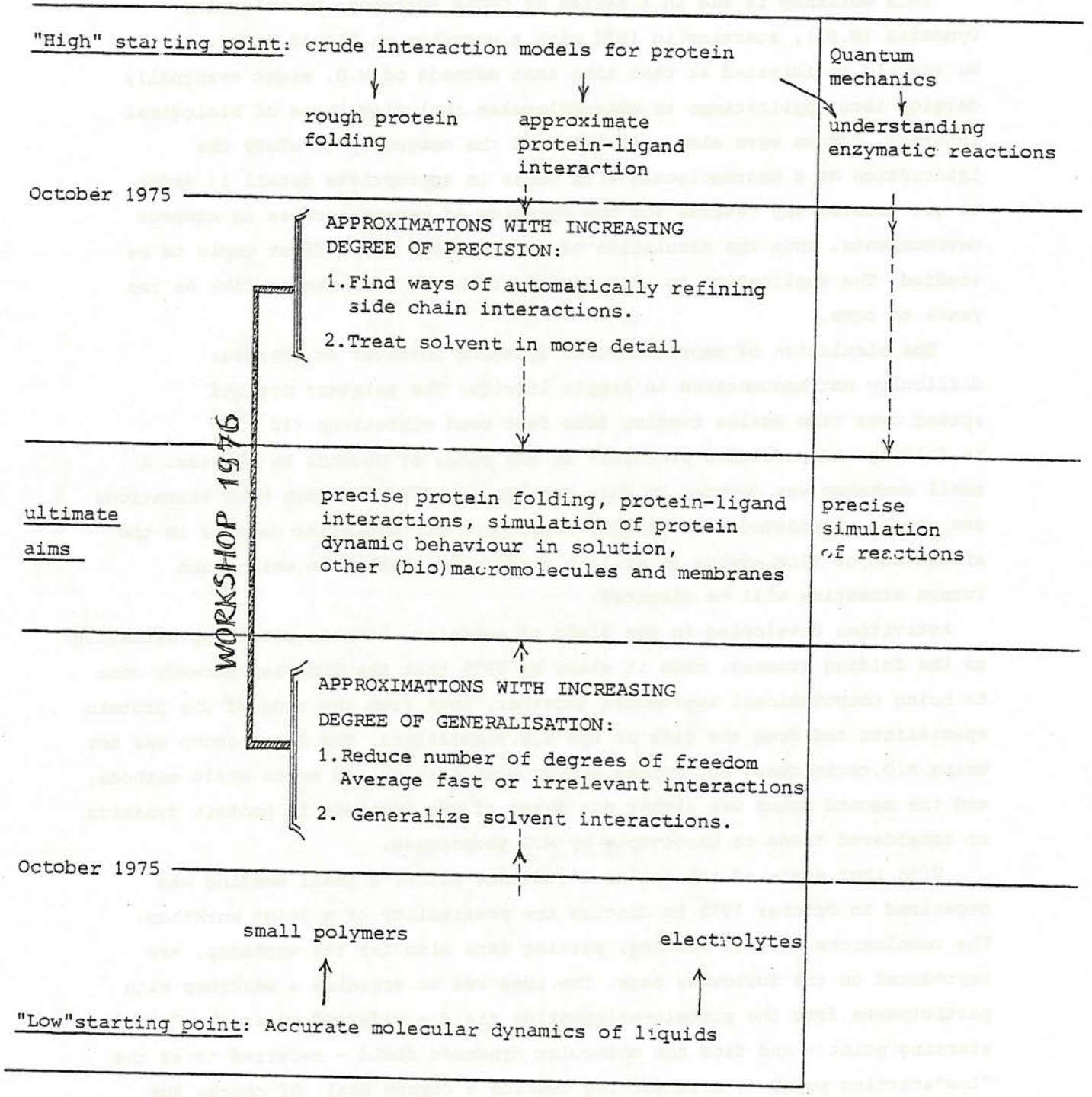
Activities developing in the field of proteins, such as crude approximations to the folding process, made it clear by 1975 that the time had already come to bring computational approaches together, both from the side of the protein specialists and from the side of the M.D.specialists. The first group was not using M.D.techniques, but rather energy minimization and monte carlo methods, and the second group was either not aware of the problems in protein dynamics or considered those as unsolvable by M.D.techniques.

With that state of the art as a starting point, a small meeting was organized in October 1975 to discuss the possibility of a joint workshop. The conclusions of this meeting, setting some aims for the workshop, are reproduced on the following page. The idea was to organize a workshop with participants from the protein-polypeptide field - referred to as the "high" starting point - and from the molecular dynamics field - referred to as the "low"starting point -, both working towards a common goal. Of course the ultimate aim, the precise simulation of dynamic processes of (biological) macromolecules in an aqueous environment, including the prediction of the

CECAM

WORKSHOP ON "MODELS FOR PROTEIN DYNAMICS", May 24-July 17, 1976

1) Protein folding
2) enzymatic reactions
3) polyelectrolytes



(plans made in Bilthoven Conference Oct 20-21, 1975)

folding process, could not possibly be attained. Nevertheless, enough enthousiasm was generated to make a workshop on "Models for Protein Dynamics" viable.

The diversity of approaches at the workshop is reflected by the contents of this report. The contributions have been grouped in sections of related interest.

Section I (Stochastic Dynamics) is seemingly unrelated to problems of protein dynamics. This is not the case, however. If aqueous macromolecules are studied, the interaction with solvent molecules is essential for structure, dynamics and function of the macromolecule. Precise simulations including all details of a large number of water molecules will certainly be too wasteful if the interest is focussed on slower events in the macromolecule. Almost surely the development will be to replace interactions with the solvent by stochastic forces, similar to methods that have already been developed for aquous electrolyte solutions. The applicability of stochastic treatments is not limited to generalization of solvent interactions. It is also possible to treat certain intermolecular degrees of freedom in a stochastic way, thereby reducing the dynamics to relevant degrees of freedom and generalizing others. The field of stochastic dynamics is only in a very early state of development and the reports of section I cover only a limited view. For protein dynamics, however, developments in stochastic dynamics will probably turn out to become essential.

Section II brings together applications of accurate M.D. to macromolecular systems, ranging from a rather small molecule in water including all degrees of freedom, to simulations of the protein *Pancreatic Trypsin Inhibitor* (PTI, 58 amino acids) and simulation of water molecules in a crystal of PTI. The dynamical simulations do not present any basic difficulties, apart from their complexity, but the time span that can be studied is limited in practice to 10^{-10}s, or 10^{-11}s in the case of water simulations. Details on molecular motions have been obtained. The PTI molecule shows a great deal of internal flexibility. It is also apparent from the results that the accessible times are not sufficient to obtain an overall insight into the ensemble of relevant structural states; particular configurations show a rather long persistence. In the case of hydration of the PTI crystal, the details of the initial configuration have not faded out in the time span studied. The feasibility of macromolecular dynamics, even including the conservation of bond constraints, is quite clear, however. There is every reason to pursue the further development of the method.

Section III covers a large number of studies on proteins and polypeptides, again with emphasis on the PTI molecule. The problems treated range from Monte Carlo studies of polypeptides (including the adrenocorticotropic hormone), displaying a distribution of conformations, to the study of protein folding and of protein-protein interactions. Necessarily the interaction functions have to be chosen in an approximate form, reducing the number of degrees of freedom to only a few per amino acid. Compared to full-scale dynamics the efficiency of the simulation increases thousandfold at the expense of accuracy and reliability. In these studies the accurate representation of dynamic events is not the main purpose; one is rather interested in the structural aspects of conformational changes and in the characteristics of the paths by which such changes are realized. Monte Carlo methods are also suitable to achieve such goals, with the main challenge being the faithful prediction of native protein structures.

The search through multidimensional configurational space to reach low-lying native states is a formidable one indeed. It is complicated not only by the gigantic dimensionality but also by energy barriers between valleys in the unpredictable and crooked mountain ranges of configurational space. The investigator is like a blindfolded pedestrian who can only feel the slope of the path on which he is standing and measure his altitude, while striving for the lowest point. It is clear that any a priori information on the low-lying states that is available could be used to speed up the search process. In this connection the use of statistical prediction techniques was considered as well, and their usefulness as a bias in Monte Carlo searches was investigated.

Studies of interactions of secondary structure elements and of complete protein faces conclude Section III. The complicated nature of such contacts adds a new problem: that of man-machine communication. In the field of simulation of biomacromolecular structure, dynamics and function, one becomes increasingly aware that the presentation of results in a comprehensible form ceases to be a trivial matter. The future will undoubtedly show a penetration of advanced graphic-display techniques for this purpose.

Finally, in *Section IV* an addendum is given on potential functions that are in use for the study of interactions in proteins. I wish to thank Dr.Dino Ferro for responding to our request to compile this very useful addendum after the workshop was completed.

The workshop has brought people from very different fields together and has very much stimulated cooperative research, in many cases extending after the workshop. I wish to thank all participants for their dedication and their

contribution to the open-minded and unrestricted atmosphere that is so essential for the success of scientific enterprises like the CECAM Workshops. I particularly wish to mention Aneesur Rahman, who has played a central role in all CECAM Workshops on molecular dynamics. But above all we are grateful to Carl Moser who has stimulated the activities not only during the workshop, but in all phases leading towards its realisation. His sharp judgement of important problem areas has much influenced the course of the developments.

I

STOCHASTIC DYNAMICS

1

STOCHASTIC DYNAMICS

I.1

STOCHASTIC APPROACH TO THE DYNAMICS OF LARGE MOLECULES IN A SOLVENT

G.Ciccotti[1]
J.Orban[2]
J.P.Ryckaert[2]

[1]Istituto di Fisica " G. MARCONI ", Piazzalle delle Scienze 5, 00185 - ROME (Italie).

[2]Université Libre de Bruxelles, Faculté des Sciences, Campus de la Plaine (Code 223), Boulevard du Triomphe, 1050 Bruxelles (Belgique).

I. INTRODUCTION

The theory of random processes could provide a useful approach to the study of the dynamics of complicated systems as, for instance, a long polymer chain in solution. We discuss the general idea underlying such a method, and some preliminary results obtained for a very special case.

The idea is essentially the following : consider a long chain-molecule in a solvent; performing the Molecular Dynamics (MD) calculation for such a system is hopeless, due to the long characteristic times of the chain. One can however write the equations of motion of the chain as

$$m_i \ddot{\underline{r}}_i = \sum_{j \neq i} \underline{\tilde{F}}_{ij} + \underline{F}_i(t) \quad , \quad i = 1 \ldots n \qquad (1)$$

where n is the number of particles in the chain

$\underline{\tilde{F}}_{ij}$ is the force exerted on particle i by particle j, both of them belonging to the chain.

$\underline{F}_i(t)$ is the force on point i, at time t, due to the solvent.

It is clear that, if $\underline{F}_i(t)$ can be computed as a random force, the problem of molecular dynamics could be simplified by some orders of magnitude.

The idea is thus to consider $\{\underline{F}_i(t)\}$ as a vectorial stochastic process ; if we can find a way to construct realisations of this process, we shall avoid the most time-consuming part of mo-

lecular dynamics, that is, the calculation of the forces. This of course implies the knowledge of the characteristics of the process. In particular, if the process is (or can be considered as) a stationary gaussian one, it is completely defined by the correlation matrix

$$\mathcal{R}_{ij}(t) = \langle \underline{F}_i(0) \cdot \underline{F}_j(t) \rangle \qquad (2)$$

Clearly, $\mathcal{R}_{ij}(t)$ will go to zero for t sufficiently large ; moreover, it seems reasonable to suppose it will go to zero for i and j sufficiently apart from each other. This has of course to be checked ; if it is true, it would enable us to determine $\mathcal{R}_{ij}(t)$ for small chains, and then use it for longer ones.

This report will deal with preliminary work in this field ; it is limited to the study of a model much simpler than chain-molecules, that is, the Lennard-Jones fluid. In other words, we look at the possibility of constructing a stochastic dynamics for one atom in a fluid of similar ones. Equation (2) is then replaced by

$$\mathcal{R}(t) = \langle F^{(\alpha)}(0)\, F^{(\alpha)}(t) \rangle \qquad (3)$$
$$\alpha = x, y, z$$

where R(t) is easily obtained by a simple molecular dynamics.

This a-priori tautological model is nevertheless interesting in order to check

(a) under which conditions the process can be regarded as gaussian

(b) the validity of the method, that is, the statistical equivalence of the stochastic dynamics and the exact one.

II. THEORY [1]

A set of random variables $\{\xi(t)\}$, $t \in R$ is said to be a stationary process if for any subset $t_1 \dots t_n$ of values of t, and for any τ

$$\varphi^{(n)}\left(\xi(t_1) \cdots \xi(t_n)\right) = \varphi^{(n)}\left(\xi(t_1+\tau) \cdots \xi(t_n+\tau)\right) \tag{4}$$

where $\varphi^{(n)}$ is the conjoint probability density distribution function of the variables between brackets. Then clearly $\varphi^{(1)}$ is time-independent, $\varphi^{(2)}$ depends only on $t_1 - t_2$, and so on. Moreover, the conditioned probability density distribution function of $\xi(t_n)$, given $\xi(t_{n-1}) \cdots \xi(t_1)$, writes

$$\varphi\left(\xi(t_n) \,\middle|\, \xi(t_{n-1}) \cdots \xi(t_1)\right) = \frac{\varphi^{(n)}\left(\xi(t_n) \cdots \xi(t_1)\right)}{\varphi^{(n-1)}\left(\xi(t_{n-1}) \cdots \xi(t_1)\right)} \tag{5}$$

and the conditional expectation value and conditional variance are :

$$E_c = E\left(\xi(t_n) \,\middle|\, \xi(t_{n-1}) \cdots \xi(t_1)\right) = \int d\xi_n \; \xi_n \, \varphi\left(\xi_n \,\middle|\, \xi_{n-1} \cdots \xi_1\right)$$

$$\sigma_c^2 = D^2\left(\xi(t_n) \,\middle|\, \xi(t_{n-1}) \cdots \xi(t_1)\right) = \int d\xi_n \left(\xi_n - E_c\right)^2 \varphi\left(\xi_n \,\middle|\, \xi_{n-1} \cdots \xi_1\right) \tag{6}$$

If the process is stationary and gaussian, it is entirely defined by

$$E(\xi_i, \xi_j) = R(t_i - t_j) = R_{ij} \quad \text{(all i, j)} \qquad (7)$$

(assuming $E(\xi_i) = 0$) . In this case, we have

$$\varphi^{(n)}(\xi_1 \cdots \xi_n) = \frac{1}{(2\pi)^{1/2}\,|R|} \exp\left(-\frac{1}{2}\sum_{ij} Q_{ij}\,\xi_i\,\xi_j\right) \qquad (8)$$

where $|R|$ is the determinant of the correlation matrix (7) and $Q = R^{-1}$; R is a positive-definite matrix, and

$$E(\xi_n \mid \xi_{n-1} \cdots \xi_1) = \sum_{k=1}^{n-1} c_k^{(n)} \xi_{n-k}$$

$$D^2(\xi_n \mid \xi_{n-1} \cdots \xi_1) = R(0) - \sum_{k=1}^{n-1} c_k^{(n)} R(n-k) \qquad (9)$$

where the $c_k^{(n)}$'s are defined by the system

$$R(s) + \sum_{k=1}^{n-1} c_k^{(n)}\, R(s-k) = 0 \qquad (10)$$

$$s = 1, \cdots n-1$$

The way of sampling realizations $\{\xi_1 \cdots \xi_n\}$ of the process is now obvious : ξ_1 is a gaussian random variable, with expectation

value zero and variance R(o), and is sampled in a straightforward way ; ξ_2, given ξ_1 , is sampled from a gaussian with expectation value and variance

$$E(\xi_2|\xi_1) = c_1^{(1)}\,\xi_1 \quad , \quad D^2(\xi_2|\xi_1) = R(o) - c_1^{(1)} R(1)$$

and $c_1^{(1)}$ is obtained from (10) ; one proceeds in that way up to ξ_n

If, for some $N < n$, $c_k^{(n)}$ $(k > N)$ are zero (or sufficiently close to zero), the "memory" of the system can be considered as finite and a limited number of ξ's have to be retained in (9). Also (10) needs no more to be solved for $n > N$. This behaviour is of course expected if R(t) has short range.

III. THE LENNARD-JONES FLUID

A. Stochastic properties

We solved "exactly" the equations of motion of a L.J. fluid near its triple-point by the standard technique of molecular dynamics. This provides us with the necessary data. The random process is here the sequence of forces acting on a particle, say particle 1 :

$$F_1(t_0),\ F_1(t_0+\Delta t),\ \cdots\ F_1(t_0+n\Delta t)$$

The process is of course stationary, as we study a fluid at equilibrium.

With these data, we constructed some histograms, corresponding to some distribution functions. Fig. 1 gives for instance $\varphi^{(1)}(F)$,

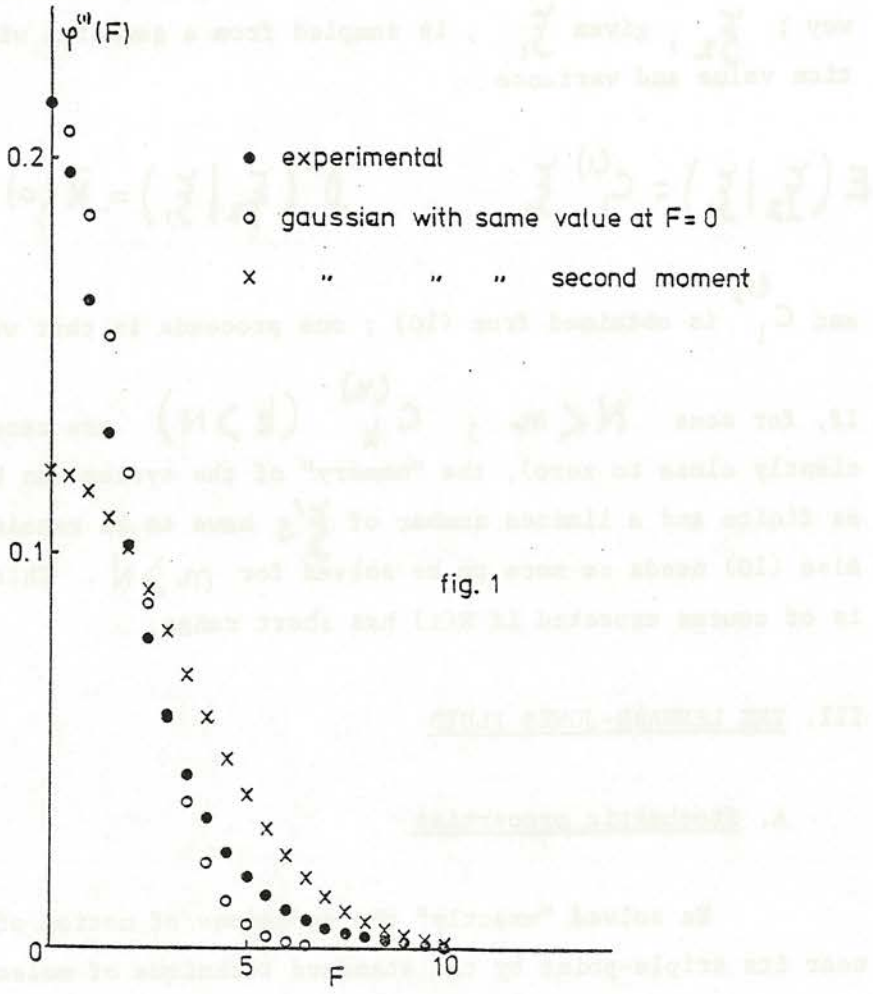


fig. 1

Fig. 1/ Distribution function $\varphi^{(1)}(F)$

- experimental (from the molecular dynamics)
- x gaussian with the same second moment
- O gaussian with the same value at F = 0.

together with the gaussian of same mean value $(\langle F \rangle = 0)$ and variance ; it is clear that the force is far from being gaussian.

This could be related to the long hydrodynamic tail of the $\langle F(0)\, F(t) \rangle$ autocorrelation function. It would be better to look at the distribution properties of a force

$$F_R = F - F_{Hydr} .$$

where F is the total force, and $F_{Hydr.}$ is given in the generalized Langevin equation

$$m \ddot{x} = F_R - \int_{t_0}^{t} \zeta_{Hydr.}(t-t')\, dt' \qquad (11)$$

(see ref. 2).

If the force F_R proved to be nearly gaussian, eq. (11) should be used ; this would in fact suppress the problem of the force-velocity correlation, because

$$\langle v(0)\, F_R(t) \rangle \cong 0$$

while

$$\langle v(0)\, F(t) \rangle \neq 0$$

We are presently working in that direction.

B. Construction of the stochastic dynamics

We believe however that the gaussian hypothesis on the total force F can give rather good results, because

(a) the physical properties of a system are given by low order correlation functions ; a gaussian could then contain enough information
(b) one can hope that the intrinsic properties of a system in solution are fairly well decoupled from the detailed behaviour of the medium.

That is the reason why we have constructed a stochastic dynamics under this hypothesis.

Because of the force-velocity correlation, we had to apply the procedure described before not only to the force, but also to the velocity ; that means that we have a vectorial processs :

$$\underline{\xi}(t) = \left(v(t), F(t) \right)$$

giving rise to not only c_k's but also d_k's ; (10) is then replaced by

$$R_{VF}(s) + \sum_{k}^{n-1} R_{VV}(s-k)\, c_k^{(2n-2)} + \sum_{k}^{n-1} R_{VF}(s-k)\, d_k^{(2n-2)} = 0$$

$$R_{FF}(s) + \sum_{k}^{n-1} R_{FV}(s-k)\, c_k^{(2n-2)} + \sum_{k}^{n-1} R_{FF}(s-k)\, d_k^{(2n-2)} = 0 \qquad (12)$$

where of course (see fig. 2)

$$R_{vF}(t) = \langle v(0) F(t) \rangle \qquad R_{Fv}(t) = \langle F(0) v(t) \rangle$$

$$R_{vv}(t) = \langle v(0) v(t) \rangle \qquad R_{FF}(t) = \langle F(0) F(t) \rangle$$

As the correlation functions do not go to zero before a hundred integration time-steps, the matrix to invert for solving (12) becomes very large. We were not yet able to find the region where, for $N < n$ and any $k > N$

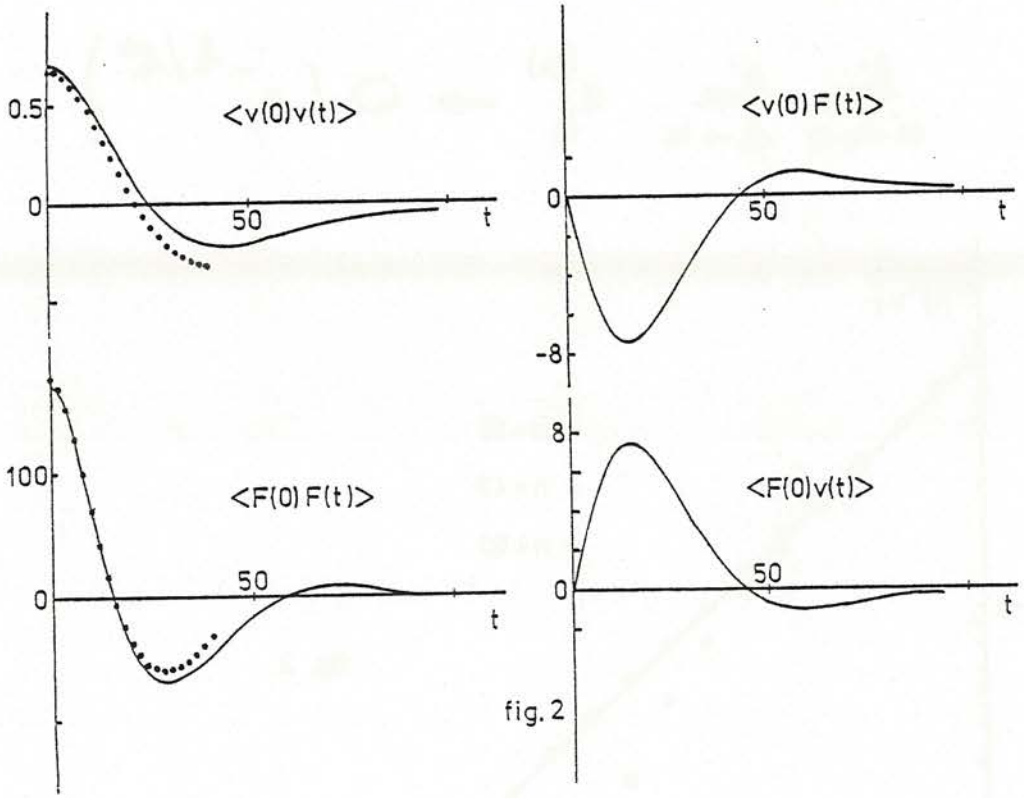


fig. 2

Fig. 2/ Correlation function $\langle v(0) v(t) \rangle$, $\langle F(0) F(t) \rangle$ $\langle v(0) F(t) \rangle$, $\langle F(0) v(t) \rangle$ for the L.J. fluid.

——— "exact" molecular dynamics

o stochastic dynamics

$$c_k^{(2n-2)} \approx d_k^{(2n-2)} \approx 0$$

However, we tested that idea in a very rough way, taking a very approximate picture of the force-force correlation. That correlation is known from the exact dynamics as

$$\langle F(t_0)\, F(t_0 + m\,\Delta t)\rangle \qquad m = 0, \dots 100$$

We retained m = 0, 20, 40, 60, 80 only, putting

$$\langle F(t_0)\, F(t_0 + m\,\Delta t)\rangle = 0 \quad \text{for } m \geq 100$$

One can see from fig. 3 that in this case

$$\lim_{m\to\infty}\ \lim_{k\to n}\ c_k^{(n)} \to 0\left(e^{-k/k^*}\right)$$

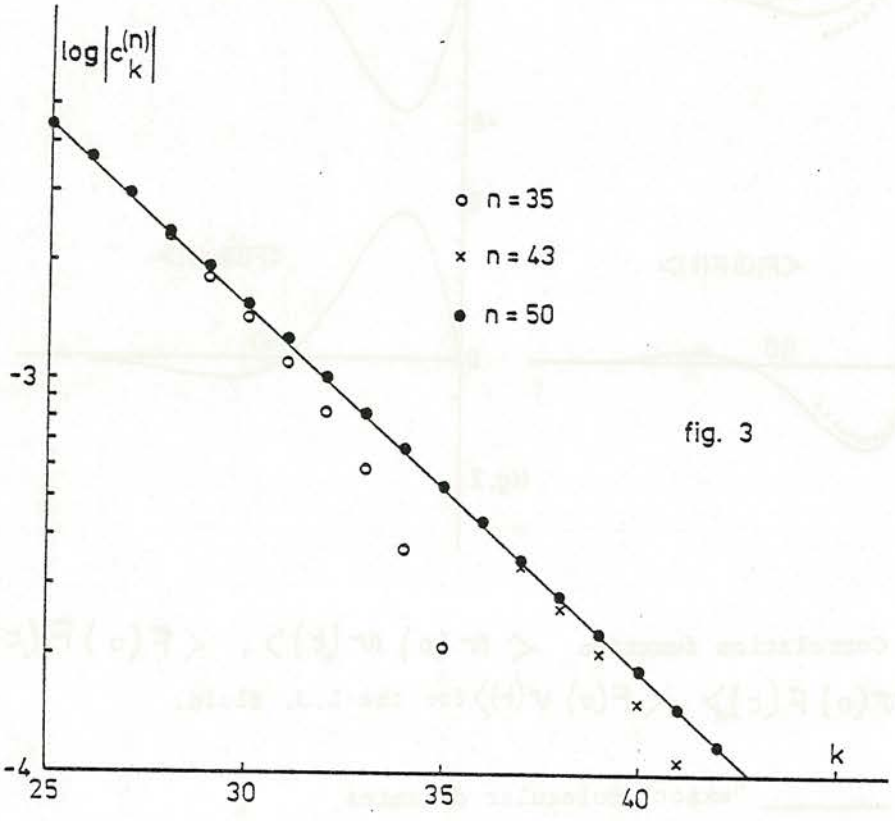


fig. 3

Fig. 3/ Behaviour of $C_k^{(n)}$ for various n (see section III B)

On the other hand, we constructed a sample of stochastic trajectories, each of 50 integration steps (in order to avoid for the moment the problems related to the solution of (11) when n becomes too large). We recalculated with them the known correlation functions, which are fairly well reproduced by this technique (see fig. 2).

IV. CONCLUSION

We think that good results obtained with the gaussian hypothesis - which we know to be far from correct - are very encouraging. We are now engaged in

(a) studying the properties of the force F_R (cfr. section III A) to see whether it behaves more gaussianly than the total force.
(b) trying to simplify the calculation of the C_k's
(c) obtaining the correlation functions involved in the chain problem.

ACKNOWLEDGEMENT

We wish to thank Drs. A. Rahman, Karplus, G. Jona-Lasinio, G. Gallavotti, A. Bellemans, P. Résibois, M. De Leener for valuable discussion, and particularly M. Cassandro for various suggestions.

We wish also to thank Drs. H.C. Berendsen and C. Moser for the organization of the workshop and the kind hospitality at CECAM.

REFERENCES

1) Yaglom, An Introduction to the Theory of Stationary Random Functions

2) M. Nelkin, Phys. Fluids 15, 1685 (1972)
T.S. Chow, J. Chem. Phys. 56, 3150 (1972)
A. Widom, Phys. Rev. A3, 1394 (1971).

I.2

APPLICATION OF HNC APPROXIMATION TO SYSTEMS OF HIGHLY CHARGED HARD SPHERES (MICELLES)

D.Elkoubi[1]
P.Turq[1]
J.P.Hansen[2]

[1]Laboratoire d'Electrochimie, Université Pierre et Marie Curie, Paris.

[2]Laboratoire de Physique Théorique des Liquides, Université Pierre et Marie Curie, Paris.

The approach of macroscopic thermodynamical and transport properties of complex systems by numerical methods of statistical mechanics is considerably simplified by the consideration of limiting physical models.

For charged systems, the two most interesting limiting models are the infinite line with an uniform or periodical charge density and the uniformeously charged sphere. The former can be considered as the schematic representation of linear polyelectrolytes such as polysaccharides and DNA. The latter as a crude draft for globular polyelectrolytes (coil form), micelles and even as a first approximation for some particular living cells such as blood cells and some bacteria .

What results from the high charge density in both linear and spherical cases is that even for the lowest concentrations the ideal behaviour and also the Debye-Hückel limiting laws cannot represent the actual experimental properties of such systems. In the linear case, more convenient limiting laws were derived by MANNING for both thermodynamical and transport properties.

The case of big highly charged particles presents particular properties coming from the great dissymmetry between the big particles and the small surrounding counter ions.

The electrostatic part of the interaction can never be separated from the geometrical part, since for a given charged big particle the strength of the coulombic interaction at its surface will depend directly on its radius.

It is therefore understandable that the ordinary Debye-Hückel limiting laws range of validity will be too far from concentrations of practical interest. Recent adaptations of molecular dynamics calculations to charged solutions such as Brownian dynamics, cannot at the present stage of computationnal techniques, be applied to the simultaneous treatment of the motion of both counter ions and big particles, because of the great dissymmetry in the size of the system.

Integral equations methods of statistical mechanics and especially the HNC approximation, which has been proved to be successful in the treatment of coulombic interactions, can be taken as the starting point of treatments of the thermodynamical properties for model systems consisting of highly charged big particles and surrounding counter ions.

A) The HNC approximation :

I) Introduction.

The ORNSTEIN-ZERNIKE relation gives the first equation between the direct correlation function C(r) and the pair correlation function g(r) :

$$(I) \qquad h(\vec{r}) = C(\vec{r}) + \rho \int C(\vec{r}-\vec{r}') \, h(\vec{r}') \, d\vec{r}'$$

where $h(r) = g(r) - 1$

and ρ is the number density

because of the convolution product between C and h, (I) will be written in the Fourier space :

$$(II) \qquad \hat{h}(k) = \hat{C}(k) + \rho \hat{C}(k) \cdot \hat{h}(k)$$

where $\hat{h}$ and $\hat{C}$ denote the Fourier transforms of h and C.

Another relation between g(r) and C(r) is still needed. Though it is not possible to get another exact equation, one can derive from diagram expansions approximations of g(r).

The first approximation gives the so called PERCUS-YEVICK equations :

$$g(r) = \exp\{-\beta v(r)\} \times (g(r) - C(r))$$

which is a good approximation in the case of non charged hard spheres, but is rather poor when studying a system of charged particles*

The second approximation gives the hypernetted chain with the following g(r) :

* J.C. RASAIAH : J. Chem. Phys. 56, 3071 (1972)

(III) $$g(r) = \exp\left\{ -\beta v(r) + h(r) - C(r) \right\}$$

which is more accurate for charged systems.

II) Method of computation of the HNC equations.

We have to solve the following set of equations :

$$\hat{h} = \hat{C} + \rho \hat{h} . \hat{C}$$
$$h(r) = g(r) - 1$$
$$g(r) = \exp\left[-\beta v(r) + \gamma(r) \right]$$
$$\gamma(r) = h(r) - C(r)$$

We first notice that, in the case of charged particles the interacting potential is the long range coulomb electrostatic field, then problems will arise in the numerical Fourier transforms because of the long tail of this potential.

In order to avoid these problems, most of the functions used in the calculation were split into two parts : long and short range, denoted by an upper script S or 1.

We also made used of the fact that the direct correlation function has the following behaviour :

$$C(r) \underset{r \to \infty}{\longrightarrow} -\beta v(r) = -u(r)$$

We then have the new set of functions :

$$U^S = U - U^1 \qquad C^S = C + u^1$$
$$\gamma^S = \gamma - u^1$$

We must now make a choice for u^1, knowing that u^1 have the same asymptotic expansion as γ, and is finite at zero. It is also more convenient to choose u^1 with an analytical Fourier transform.

Many functions have been used by different authors, we are using :

$$u^1 = \frac{\Gamma}{r} \operatorname{erf}(\alpha r) \qquad u_S = \frac{\Gamma}{r}(1 - \operatorname{erf}(\alpha r))$$

$$\hat{u}^1 = \frac{4\pi}{k^2} \exp -\frac{k^2}{4\alpha^2}$$

where $\operatorname{erf}(x) = \int_0^x e^{-y^2} dy$

and $\Gamma = \frac{e^2}{\varepsilon kT}$

The system of equations to be solved is now :

(1) $\hat{\gamma}^S = \frac{\hat{C}^S \cdot \hat{C} - \hat{u}^1}{1 - \hat{C}}$ (4) $C = C^S - u^1$

(2) $G(r) = \exp \gamma^S - u^S$

(3) $C^S = g(r) - 1 - \gamma^S$

Starting from u^1 and u^S, solutions are obtained by guessing a first C^S (The Debye-Hückel solutions for example) that will give a first $\hat{C}$ and $\hat{C}^S$ by (4), and then generate an iterative process which is repeated until the n^{th} C^S (C^S_n) has the same value, according to the requested precision, than the (n-1)th C^S (C^S_{n-1}).

A test is made at each step to prevent the iterative process from diverging or oscillating and a linear combination of C^S_n and C^S_{n-1} is taken.

In the case of a two component system equation (II) has to be replaced by :

$$\hat{h}_{ij}(k) = \hat{C}_{ij}(k) + \sum_1 \rho_1 \hat{h}_{i1}(k) \hat{C}_{1j}(k)$$

which gives 6 relations between $\hat{h}_{ij}$ and $\hat{C}_{ij}$ the process, however , remains the same.

III) Computation of thermodynamic properties from g(r).

The HNC approximation gives the pair correlation function of the studied system. From this function it is possible to derive most of the thermodynamic parameters of the system.

-a) The excess pressure is given by the following integral :

$$p^{ex}/kT = \frac{2\pi}{3}\rho \sum_{ij} x_i x_j \int g_{ij}(r) \frac{\partial \varphi_{ij}}{\partial r} r^3 \, dr$$

with : $x_i = n_i/N$ $\quad x_j = n_j/N$

n_i, n_j = number of particles of specie i

$\rho = N/V$

φ_{ij} is the pair potentiel for a system of charged hard spheres, as we will study in the beginning, we have :

$$\varphi_{ij} = \frac{e^2 Z_i Z_j}{\varepsilon kT r}$$

Z_i, Z_j = charge of each ion

ε = dielectric constant of the solvant

then :

$$p^{ex}/kT\rho = \frac{2\pi e^2}{3\varepsilon kT} \rho \sum_{ij} x_i x_j Z_i Z_j \int g_{ij}(r) \, 4\pi r dr$$

$$+ \frac{2\pi}{3}\rho \sum_{ij} x_i x_j g_{ij}(r) \sigma_{ij}^3$$

with σ_{ij} is the distance of closest approach.

-b) The Excess Energy :

$$E^{ex}/kT = 2 \frac{e^2}{kT} \rho \sum_{ij} x_i x_j Z_i Z_j \, h_{ij}(r) \, r \, dr$$

with $h_{ij}(r) = g_{ij}(r) - 1$

We can analyze our results more carefully by separating the osmotic coefficient $\phi - 1 = \frac{p^{ex}}{kT\rho}$ from the virial theorem, into

the sum of two terms, when the system of interest consists of charged hard spheres

$$\phi - 1 = \frac{E^{ex}}{3kT} + \frac{2}{3}\pi\rho\sum_{ij} x_i x_j\, g_{ij}(r)\, \sigma^3\, Z_i Z_j$$

The variation of the osmotic coefficient ϕ depends therefore on two opposite contributions.

- the negative excess energy term
- the positive contact term

At low density the excess energy term in $\sqrt{\rho}$ is predominant. For higher concentrations the positive contact term in ρ dominates.

- c) The coordination number from which we can derive the apparent charge of the micelle τ_s given by :

$$N_{ij}(r) = 4\pi\rho_i \int_0^{r'} g_{ij}(r)\, r^2\, dr$$

B) Results and Discussion :

High charge unsymmetrical electrolytes.

We have systematically studied the thermodynamic properties of aqueous solutions of 6-1 electrolytes for various concentrations and ionic diameters.

a) Osmotic coefficients and excess internal energy :

- Case α : Same diameter 4.2 Å. The osmotic coefficient ϕ increases from about 0.5 to 1.11 when the concentration varies. In the same concentration interval, the negative excess energy has only a small variation .

- Case β : $r_+ = 2.1$ Å $r_- = 4.2$ Å. The osmotic coefficient increases rapidly when the concentration increases. Simultaneously the excess negative energy is lowered since the great size of the negative ions has considerably modified the covolume effects. For the highest concentrations these covolume effects are predominant.

b) Radial distribution functions :

On table I are presented the values of $g_{ij}(\sigma_{ij})$ i.e. the values at the contact. For the +- couple this contact value is also the value of the maximum of g_{+-}.

$g_{+-}(\sigma_{+-})$ decreases generally with increasing density. In the caseα $g_{+-}(\sigma_{+-})$ presents a minimum for the concentration 0.5M.

- In the caseβ $g_{++}(\sigma_{++})$ is always larger than 1. This result indicates that many positive ions are in contact, presurably at the surface of a negative one. $g_{++}(\sigma_{++})$ presents in the case α a minimum (larger than 1) for C = 0.5 M. In the case β $g_{++}(\sigma_{++})$ is continuously increasing with concentration and reaches values considerably larger than 1.

- $g_{--}(\sigma_{--})$ has only non zero values for the highest concentrations of the case β.

c) Coordination numbers :

In the introductive section we have seen that the coordination number (number of i ions around aj ion) $N_{ij}(r')$ is a function of the distance r' between i and j. Our purpose is to mainly define the apparent charges of the particles, i.e. the charge of the particle minus the number of particles of opposite sign in contact. We have chosen the following convention for r'_{ij} :

$$r'_{ij} = \frac{\sigma_j}{2} + \sigma_i$$

We take into account all the i particles in a sphere of radius $\frac{\sigma_j}{2} + \sigma_i$, in order to include all the i ions in contact with the reference j ion.

In this convention appear four coordination numbers : N_{++}, N_{--}, N_{+-} and N_{-+}.

The results are presented in table 1.

TABLE I

RESULTS FOR UNSYMMETRICAL 6 -1 ELECTROLYTES

	(1) C	(2) $g_{++}(\sigma_{++})$	(3) $g_{--}(\sigma_{--})$	(4) $g_{+-}(\sigma_{+-})$	(5) N++	(6) N--	(7) N+-	(8) N-+	(9) Zap	(10) ϕ	(11) $-E^{ex}/NkT$
Case α	0.2	1.2147	0.	22.254	0.7335	0.	3.0999	0.51	- 2.90	0.486	4.06
$\mathcal{R}_+$ = 2.1 Å	0.35	1.1873	0.	13.9707	1.210	0.	4.62	0.77	- 1.38	0.517	4.40
$\mathcal{R}_-$ = 2.1 Å	0.5	1.1452	0.	11.1329	1.662	0.	4.669	0.778	- 1.331	0.641	4.58
	0.75	1.2578	0.	8.1563	2.5	0.	6.134	1.022	+ 0.134	0.817	4.85
	1.	1.3241	0.	7.1347	3.338	0.	6.415	1.069	+ 0.415	1.110	5.01
Case β	0.1	0.6914	0.00008	9.2526	0.2289	0.	2.278	0.3797	- 3.722	0.645	2.67
$\mathcal{R}_+$ = 2.1 Å	0.25	0.7334	0.0008	5.5145	0.61	0.165	3.97	0.662	- 2.03	0.807	3.08
$\mathcal{R}_-$ = 4.2 Å	0.5	0.9141	0.007	4.4648	1.39	0.7993	6.4231	1.07	- 0.4281	1.281	3.44
	0.75	1.2738	0.042	4.6334	2.41	2.3	9.44	1.57	+ 3.44	2.105	3.71
	1	1.6572	0.1924	5.6067	3.437	3.85	12.9	2.15	+ 6.9	3.559	3.87

(1) C Concentration molar

(2) (3) (4) Radial distribution functions at the contact

(5) (6) (7) (8) Coordination numbers

(9) Apparent charge of the anion

(10) Osmotic coefficient

(11) Excess energy.

- Case α . N_{+-} varies between 3 and 6.4. The first region below the minimum of g_{++} corresponds almost to a tetracoordinated anion with a negative apparent charge between - 3 and - 1.

The second region beyond the minimum of g_{++} exhibits an increase in the coordination number of the anion tending almost to an octaedric coordinated species of zero net charge. The values of N_{++} increase with the concentration, with a pump after the minimum of g_{++} as required by the two states model. N_{-+} is always near or smaller than 1, showing than the coordinated species are well defined entities with only one anion, as indicated by N_{--} which is still zero.

- Case β . N_{+-} varies between 2.3 and 12.9. Here also appears a transition with increasing concentration. Below 0.5 M (minimum of g_{+-}). We have an almost tetracoordinated or tricoordinated anion. After the minimum of g_{+-}, N_{+-} increases rapidly, but the surrounding cations cannot be attributed to adefinite anion. The values of g_{--}, N_{--} and N_{-+} show clearly that two or more anions have a pool of cations in common. Physically, the repulsive electrostatic forces between the large anions, are not great enough to prevent the formation of such clusters.

d) Discussion :

The application of the HNC approximation to highly dissymetric ionic systems permit the obtention of the thermodynamical and structural properties in the concentration range of aqueous solutions. The studied system (anion of charge 6 and monovalent cations) could be a model for polyphosphates solutions. The extension of this work to more unsymmetrical electrolytes should permit the study of a great variety of systems of physical interest, including globular polyelectrolytes and micelles.

BIBLIOGRAPHY

- J.C. RASAIAH and H.L. FRIEDMAN - J. Chem. Phys., 48, 2742 (1968)
- J.C. RASAIAH and H.L. FRIEDMAN - J. Chem. Phys., 50, 3965 (1969)
- J.C. RASAIAH - J. Chem. Phys., 56, 3071, (1972)
- J.P. HANSEN and I.R. Mc DONALD - Phys. Rev., A11, 2111 (1975)
- J.P. HANSEN and I.R. Mc DONALD - Theory of simple liquids - Academic Press (1976).

I.3

TIME-DEPENDENT IONIC INTERACTIONS IN THE BROWNIAN DYNAMICS OF ELECTROLYTE SOLUTIONS

P.Turq
F.Lantelme

Université P et M.Curie, Laboratoire d'Electrochimie,
Bâtiment F, 8, rue Cuvier, 75 005 Paris (France).

Abstract

The method of Brownian dynamics of electrolyte solutions is applied to a 1M solution of a particular 2-2 electrolyte to generate time dependent configurations.

The result are analyzed phenomenogically in terms of ions pairs and triplets formation.

From the radial distribution functions it is found that the formation of pairs of opposite sign, plays a fundamental part in the structure of the solution.

The persistence of this structure is of about 10^{-12} sec from the time dependent radial distribution $g_{+-}(r,t)$.

The transport properties (self diffusion coefficients and electrical conductance) exhibit a small deviation to the Nernst-Einstein relation which can also be tentatively interpreted in terms of cation-anion clusters.

A more detailed analysis of the time deviation of the different ionic clusters shows that, in fact, there occurs a continuous exchange between the free ions and the ionic clusters and that anion-cation couples have a life span not much longer than that of other types of pairs and triplets.

I Introduction :

"Brownian dynamics" (1) is a kind of molecular dynamics characterized by the use of the Langevin equation in place of the ordinary Newton equation of motion (1) (2) (3).

Brownian dynamics is basically constituted by the numerical integration of a set of Langevin equations for an assembly of interacting particles.

The introduction of a Langevin equation in place of the Newton equation permits the treatment of non-Hamiltonian friction and Random forces which characterize a continuous bath thermalizing the solute particles. Therefore it becomes possible to treat systems where there is great disymetry between particles of particular interest and a greater amount of less interesting or inertial particles.

Brownian dynamics or analogous methods using the Langevin equation and additional specified forces have previously been used by Weiner and Forman (4) to study the motion af an impurity atom in a crystal, by Ermak (5) to study the influence of small ions on the diffusive motion of a polyion, and by Adelman (6), Doll and Dion(7) to study the motion of a particle diffusing on a crystal surface.

As regards to electrolyte solutions, Brownian dynamics is a generalization of the average classical solvent theories of ionic interactions which, since the Debye-Hückel approach of the excess free energy in the early twenties, has known many extentions and developments for the applications to both thermodynamical and transport properties, but has never been changed in the basic physical model.

The electrolyte solution is outlined by a continuous solvent characterized generally by its macroscopic static dielectric constant and shear viscosity , even in the case of interacting microscopic entities.

The solute particles (ions) are discrete charged hard or soft spheres of which electrostatic interactions are divided by the previously mentioned macroscopic dielectric constant, this also goes for the shortest interionic distances.

The only factor which differenciates the different physical models is the short range potential : hard spheres of identical radius (restricted primitive model), of different radius, soft spheres, introduction of supplementary terms in the short range interaction characterizing ionic solvation (Gurney terms) (8).

Beside this relative simplicity the basic physical model (especially for the solvent), a great variety of mathematical treatments were used to approach the problem.

a) Historically the thermodynamic properties were derived first from the solution of the Poisson-Boltzmann equation (9). More rigorous foundations were obtained by adapting the Mayer graphs theory to ionic clusters (10). Most of the modern methods of statistical mechanics (11) were tested on the case of electrolyte solutions.

However these treatments stopped mostly after the well known limiting laws in $\sqrt{C}$ were derived, and a result of this is that there are, at least, thirty different proofs of these limiting laws.

More interesting in practice for the solution chemists is the use of the Hyper-netted chain approximation to get numerical solutions for the ionic radial distribution functions and all corresponding thermodynamical quantities.

This work was developed mainly by Friedman and Rasaiah (12) with a great variety in the choice of the short range interionic potential.

The main interest of this technique is to provide osmotic coefficients which can be compared to the experimental values for a concentration range of interest from 10^{-2}M to 3M.

Another method of practical interest is the Mean Spherical Model first developed for electrolytic solutions by Waisman and Lebowitz (13). Not as accurate as the HNC approximation, it permits in its more recent (14) developements a semi-analytical treatment of the thermodynamical properties of solutions in the range 0 to 1 - .3M. Its principal advantages is to be simple from the numerical point of view and to admit as asymptotic behaviour for dilute solutions, the Debye Hückel limiting laws.

Association phenomena were introduced by Bjerrum (15) as a correction of the Debye Hückel theory. The Bjerrum concept of association is based upon the division of the space around a given ion in separate regions. At short distances (shorter than $q_{+-} = Z+Z-e^2/\varepsilon kT$) * the ions are presumed to be paired . For an interionic distance longer than q_{+-}, normal electrostatic interactions occur.

More rigorous foundations of the Bjerrum space separation in two regions were obtained by Falkenhagen and Justice (16).

For multivalent symmetrical electrolytes or 1-1 electrolytes in low dielectric constant media the Bjerrum concept of ion paring gave a satisfying description of the departures to the Debye Hückel limiting laws.

b) For transport and time dependent properties the variety of approaches is not as great as for thermodynamics.

The hydrodynamical approach, developped by Debye, Onsager, Fuoss and Falkenhagen since the early thirties was the most fruitful and has given the limiting laws in $\sqrt{C}$ for a great number of transport and time dependent phenomena : conductance, self and mutual diffusion, high frequency dependence of relaxation effect.

---

* : where Z+ and Z- are the charge numbers of cations and anions, e the elementary charge, ε the dielectric constant, and kT the Boltzmann factor.

Beyond the limiting law the use of the hydrodynamical approach is tedious and questionable , the physical model being introduced indirectly by the means of boundary conditions in the integration of the hydrodynamics partiol derivative equations (16).

Some statistical mechanical derivations of the limiting laws of the transport coefficients were made, but any extension to higher terms in concentration would be practically impossible (13).

The purpose of Brownian Dynamics (B.D.) is therefore to give, in the continuous solvent model approximation for electrolyte solutions, an evaluation of transport and time dependent phenomena, as made with the H.N.C. integral equation method for the thermodynamic properties.

A first application of the Brownian Dynamics for electrolyte solutions has permitted to derive that all the common features of electrolyte solutions were respected by Brownian dynamics(1)

The thermodynamical properties , the radial distribution function, the osmotic pressure, the excess internal energy, the virial and some other coefficients, were obtained for different concentrations in agreement with other approaches, as with the HNC method.

But the new interesting characteristic of B.D. is to provide time dependent and transport properties. The preliminary results have shown that some correlation functions (velocities, forces...) exhibited a noticeable concentration dependence as required by the experimental observation of the corresponding transport coefficients.

The most interesting result of the first application of B.D. to electrolyte solutions was the observation of a great variation of both thermodynamical and transport properties, with decreasing dielectric constant of the medium for a given electrolyte ;

increase of the pick in the unlike $g_{+-}(r)$ distribution function, decrease of the osmotic coefficient ϕ, and also in the self-diffusion coefficients.

These first results were analyzed phenomenogically in terms of ion pair formation.

In order to rafine and consolidate the previous results the present work will be devoted to the generation and the analysis of a great number of configurations through Brownian dynamics for an electrolyte solution corresponding to an experimental case where either the phenomenological language of ion pair formation or the analysis in terms of electrostatic interactions were used. We have chosen a system equivalent to a 2-2 electrolyte in water at 1 molar concentration. This concentration is in the best range for the application of Brownian dynamics, the analytical theories are not much more valuable at such a concentration and the deviations from ideality (infinite dilution) are significant.

In a first part we will recall the main features of B.D. as applied to electrolyte solutions.

In a second part we will examine the results for equilibrium and transport properties.

The following part will be devoted to the analysis of the time correlations of different types of ionic clusters in solutions (paris, triplets ...)

This work will conclude with the limits and the signification of the description of a solution as a collection of such ionic clusters .

II Brownian dynamics of electrolyte solutions .

The generalized Langevin equation is :

$$\dot{p}(t) = -\int_0^t f(t-s)\, p(s)\, ds + R(t) + X(t) \qquad (1a)$$

Here $p(t) = p_x(t)$ is the x component of the Brownian s particle momentum, R the random force and X the external force.

the mass of the Brownian particle (Brownon) is m.

$\langle \cdots \rangle$ denotes average over an equilibrium ensemble.

$$\langle R(t)\, p(0) \rangle = 0 \quad \text{if} \quad t > 0 \tag{1b}$$

$$\langle R(t) \rangle = 0 \tag{1c}$$

$$\langle p^2 \rangle = m k_B T \tag{1d}$$

where f(t) is the memory function. We have the second fluctuation dissipation theorem (1a) which may be derived from equation (1)

$$\langle R(t)\, R(0) \rangle = \langle p^2 \rangle f(t) \tag{2}$$

We begin by defining correlation times for the momentum and for the random force :

$$t_p = \langle p^2 \rangle^{-1} \int_0^\infty \langle p(t)\, p(0) \rangle \, dt \tag{3a}$$

$$t_R = \langle R^2 \rangle^{-1} \int_0^\infty \langle R(t)\, R(0) \rangle \, dt \tag{3b}$$

these definitions are to be applied quite generally.

In general, the following relation between tp and t_R may be derived from eq (1)

$$t_p\, t_R = \langle p^2 \rangle / \langle R^2 \rangle \tag{4}$$

We choose to simulate the generalized Langevin process for a particular form of the random force ; R is constant for time steps of duration t_S and for each time step a value of R is chosen at random from a distribution $W_1(R)$ whose moments

$$\langle R^n \rangle = \int_{-\infty}^{+\infty} R^n W_1(R)\, dR \tag{5}$$

of all orders are finite. This condition is required because

$\langle R^n \rangle = \langle p^n \rangle$ is a thermodynamic coefficient which exists for non singular potentials (19).

The random force has a time correlation of the form :

$$\langle R(t)\, R(0) \rangle = \langle R^2 \rangle \left\{ (t_S - |t|) / t_S \right\} \theta(t_S - |t|) \tag{6}$$

where θ is the unit step function. For this particular form of the random force we find that the correlation time t_R is

$$t_R = t_{S/2} \tag{7a}$$

and that

$$\langle R^2 \rangle = 2 (k_B T)^2 / D t_S \tag{7b}$$

In order to simulate an electrolyte solution we will consider N cations (index +) and N anions (index -) each in Brownian motion in a volume V. The initial distribution of velocities was Maxwellian. The trajectory in phase space of the system was followed for about 10^4 steps. The previously reported results proved that, without interactions between the Brownian particles, the calculation simulates Langevin's equation reasonably well. (14)

In order to simulate a model with interionic forces we introduce the pair potential (21)

$$U_{jj'}(r) = \left\{ e^2 / n_c (r_j^* + r_{j'}^*) \right\} \left\{ (r_j^* + r_{j'}^*) / r \right\}^{n_c} - e_j e_{j'} / \varepsilon r$$

which is part of the pair potential in a model for electrolyte solutions which was used in the study of thermodynamic properties. Here e_j is the charge, r_j^* the Pauling ionic radius, n_c a parameter and ε the dielectric constant of the solvent. The computations were made with the use of the Ewald methood (22_a) to include the fields from ions outside the cell. The volume of the cell was chosen so that the ion number density was the same as for a 1 M solution. The charge numbers of cations and anions were + 2 and - 2, the dielectric constant ε was 78,54 r_+^* and r_-^* equals to 1.38 Å. The time step of the calculation was $0.8\ 10^{-13}$ sec, the mass of the Brownian particles was 360 and the self diffusion coefficients of cations and anions at infinite dilution equals $2.033\ 10^{-5}$ cm^2 sec^{-1}. n_c was taken equal to 9. The complete symetry of the system will considerably simplify the analysis of the results and improve the statistics of the simulation...

The total number of time steps generated was $1.4\ 10^4$, one set of $8\ 10^3$ and another independent of $6\ 10^3$. We did not see any noticeable discrepancies between the two sets of results.

III - Thermodynamical and transport results.

Thermodynamics :

Interacting solute particles in Brownian dynamics have the behaviour

of a canonical ensemble.

The Langevin equation is limited by the Fokker Planck equation for a great number of particles, and the solution of the infinite time limit of the Fokker Planck equation is the Gibbs distribution $e^{-\beta H}$.

The solute particles are thermalized by the continuous bath of solvent, exchanging energy by the means of the fluctuation-dissipation process.

The radial distribution function $g_{+-}(r)$ has a strong pick at about 3.5 - 4 Å (Fig.1). The radial distribution functions $g_{++}(r)$ and $g_{--}(r)$ exhibit a smooth hump at 6.5 Å.

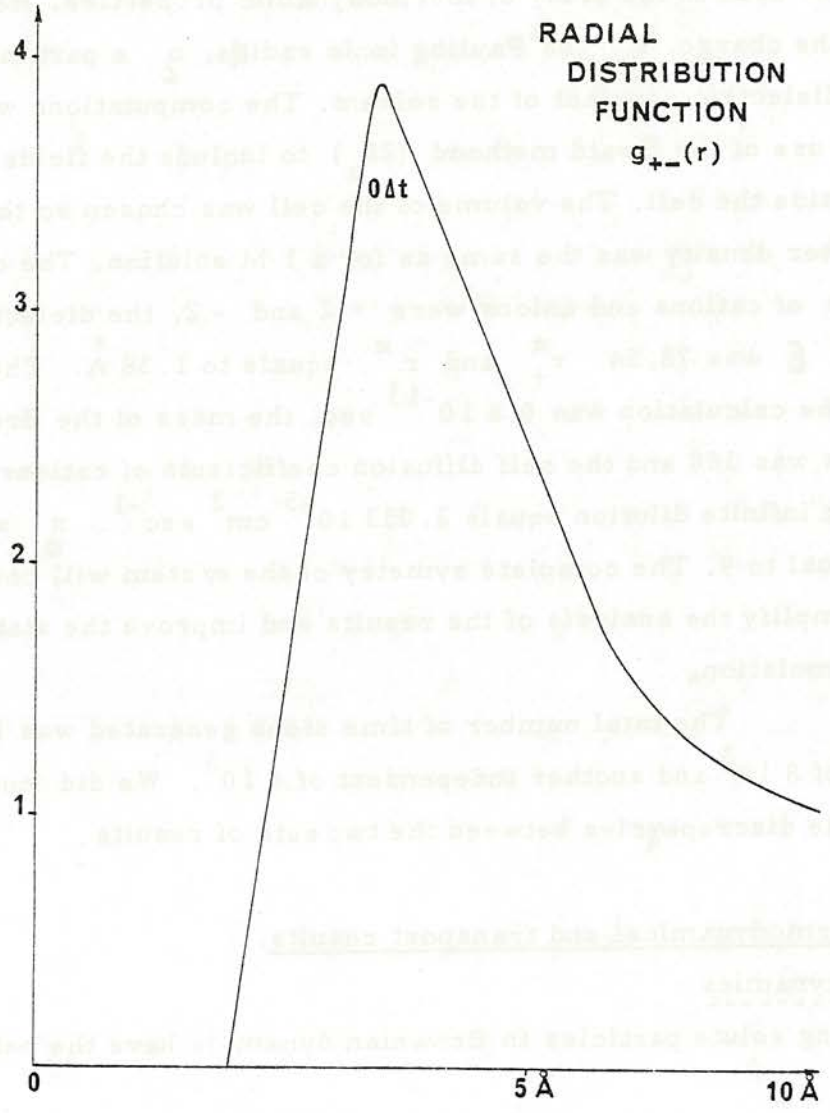


FIG 1

Beyond 7 Å all features characteristic of the short distance structure disappear.

The osmotic coefficient is found to be 0.63 ± 0,01 HNC calculations were made with the same potential and the g(r) and osmotic coefficients results were identical to our of Brownian dynamics findings (15).

Transport results : we have plotted the normalized velocities and current self correlation functions as functions of time coefficients (figure 2) and computed the corresponding self diffusion and conduc-

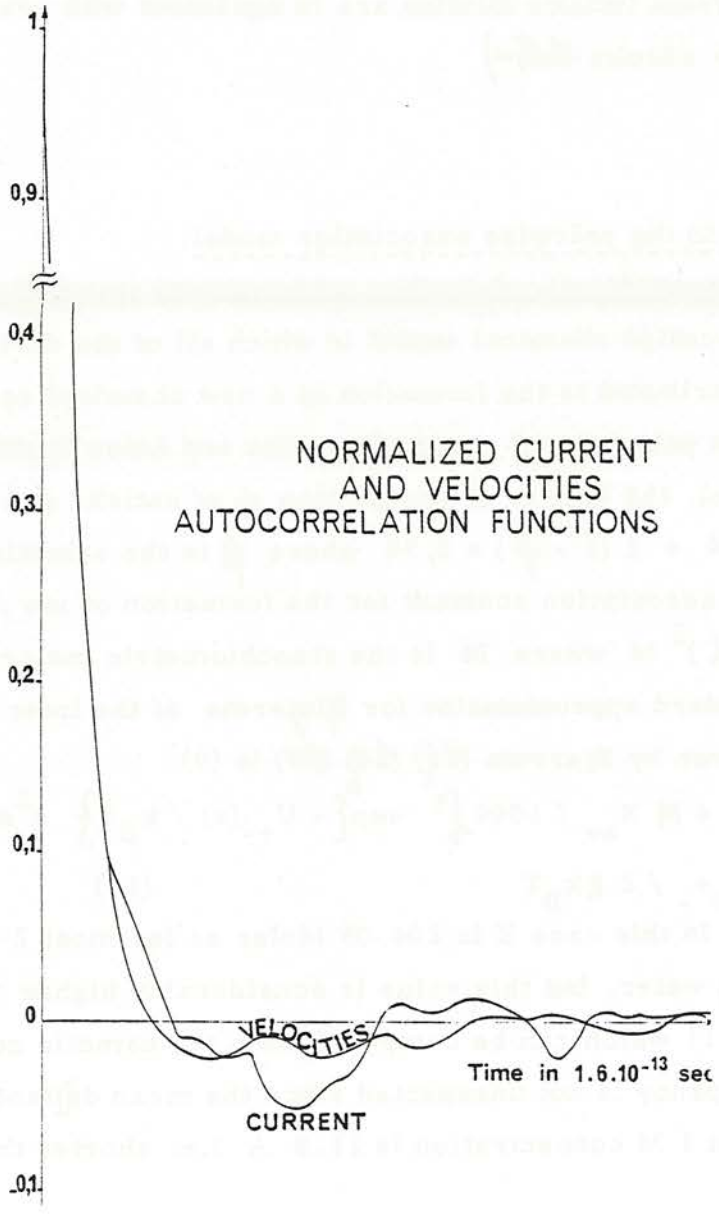


FIG 2

tances. The interesting point is that these functions present a negative part, more pronounced for the current function. The numerical values of the self diffusion coefficients of positive and negative particles are identical as required by the equality of the friction coefficients. The mean value of the self diffusion coefficients is $1.41\ 10^{-5}$ cm^2/sec for a D° of $2.033\ 10^{-5}$ cm^2/sec at infinite dilution. The dispersion of the self diffusion result is very small and less than 1%.

For the conductance results the dispersion is considerably higher and the precision on the conductance value is of 20% to 30%. The result expressed in Ω^{-1} cm^2 is between 74 and 84 for a Λ_o of 152.8 corresponding for example to a CsBr or KCl solution. Both variations of Λ and D from infinite dilution are in agrement with available experimental results (16a)

Interpretation in the pairwise association model.

Tentative interpretations and further comparisons can be made in terms of a so-called chemical model in which all of the deviations from ideality are attributed to the formation of a new chemical species, a complex or ion pair formed from one cation and anion in this case. In such a model, the degree of association α of cations and anions to form ion pairs is $\alpha = 2(1 - \phi) = 0,74$ where ϕ is the osmotic coefficient, and the molar association constant for the formation of ion pairs is $K = \alpha / (1 - \alpha)^2 M$ where M is the stoechiometric molarity of the salt. The standard approximation for K in terms of the inter ionic potential is given by Bjerrum (16) (16) (15) is (9)

$$K = 4\pi N_{av} / 1000 \int_a^b \exp\{- U_{+-}(r) / k_B T\}\ r^2 dr$$

where $q = e_+ e_- / 2 \varepsilon k_B T$ (10)

In this case K is 204.05 Molar as for most 2-2 electrolytes in water, but this value is considerably higher than the value of about 11 which can be computed from the osmotic coefficient. Such a discrepancy is not unexpected since the mean distance between the particles at 1 M concentration is 11.8 Å i.e. shorter than the

Bjerrum distance 14.3 Å. In order to interprete the transport results with the chemical model of pair formation, we assume that the only effect of the ionic interactions upon the diffusion coefficients and the electric conductance is via the formation of neutral (non conducting) + - pairs with diffusion coefficient D_p.

we write

$$\Lambda = \Lambda^\circ (1 - \alpha), \qquad (11)$$

$$D = D^\circ (1 - \alpha) + \alpha D_p \qquad (12)$$

From (11) we get $\alpha_M = 0.52$ and if we introduce this value in (12) we find : $D_p = 0.83\ 10^{-5}\ cm^2/sec.$

If the friction coefficient of the ion pair were the sum of the friction coefficients of the two ions, we would have $D_p = 1.01\ 10^{-5} cm^2/sec$ which is the order of magnitude of the above value. More significant is thedifference between α from thermodynamics (0.74) and from transport properties (O.52). The thermodynamic concept concerning the distance between positive and negative particles (mainly the g+- (r) pick).

The transport concept of ionic association is a little more acute. The ion pair is a definite transport entity if its life span is longer than the characteristic time of the associated transport phenomena. It is well understandable that the proportion of ionic chusters moving together during the characteristic time of the transport process is smaller than that coming from the instantancons picture of the solution. The pair concept in electrolyte solutions depends therefore crucially on the time scale of the considered processes and the next paragraph will be devoted to the time correlations analysis of ionic chusters.

IV - Time correlations.

We will start with the time dependence of the radial distribution function. Fig. III presents g_{+-} (r, t) for different time intervals. The decay of g_{+-} (r, t) is characterized by the peak broadening and displacing itself. The number of neighbours at the

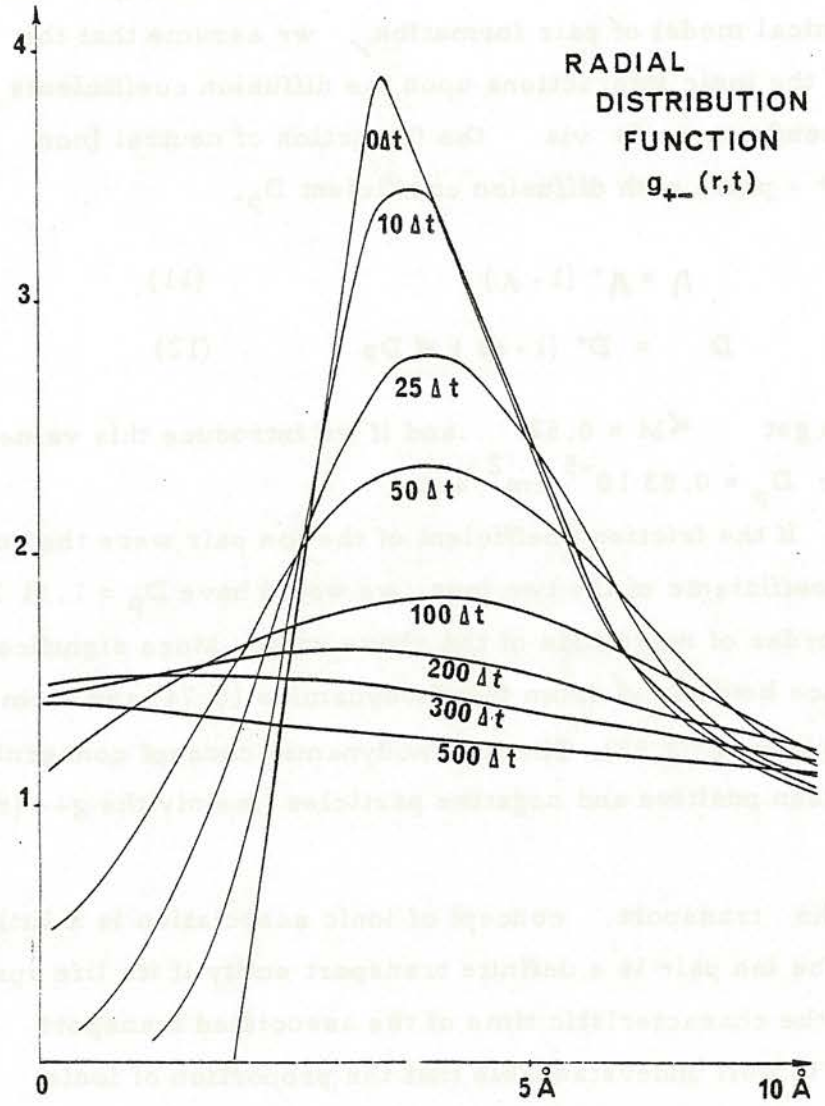


FIG 3

contact, defined by the area of the first pick, is initially between 1 and 2 and remains almost constant during 100 to 200 Δ t. After 300 Δ t ($\Delta t = 0.8\ 10^{-13}$ sec) the pair-wise structure desappears progressively.

In order to separate the displacement (self diffusion) and the decorrelation of the ion pair, we have computed the time-span (in percentage) of an isolated particle, of a group of two particles of the same sign, of a group of two particles of the opposite sign, of a group of three particles of same sign, of a three particles cluster with two of the same sign and one of the opposite.

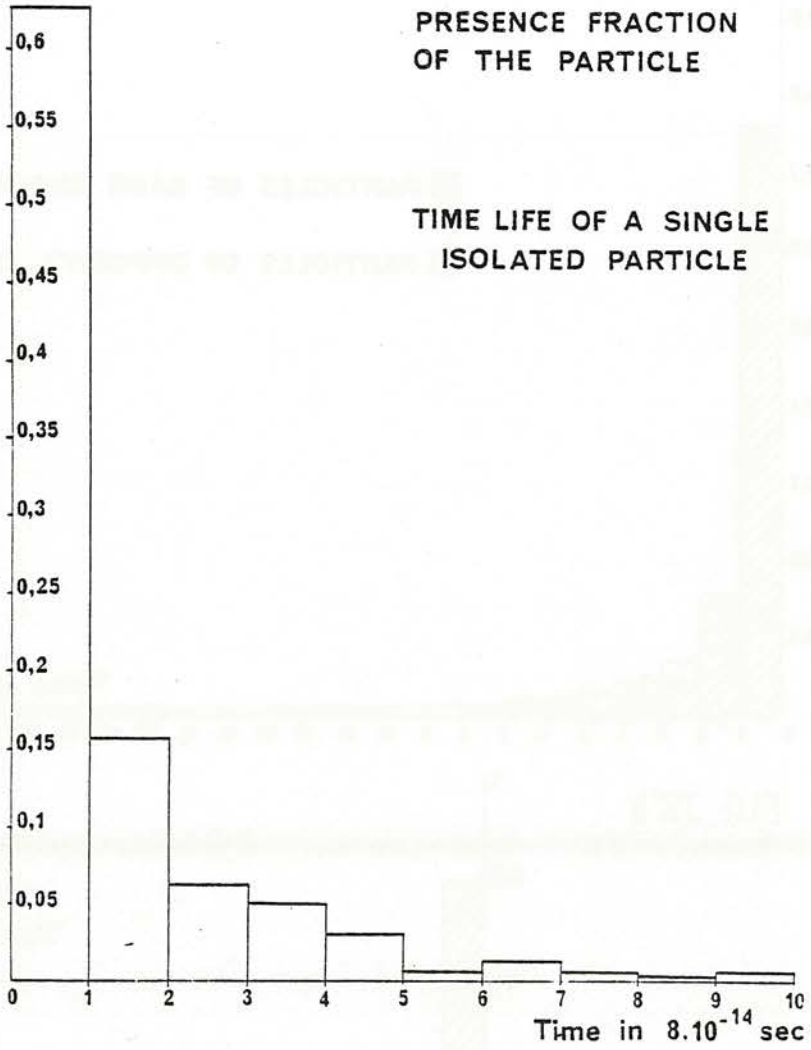


FIG IVa

We have also plotted the results about a four-particles cluster with two particles of one sign and two of the other. All the corresponding plots are presented on figure IVa, IVb, IVc, IVd.

The interesting result for the ion pairs is that the difference between like similar and unlike dissimilar ion pairs occurs only for long times where we observe a small persistency of the opposite sign ion pair.

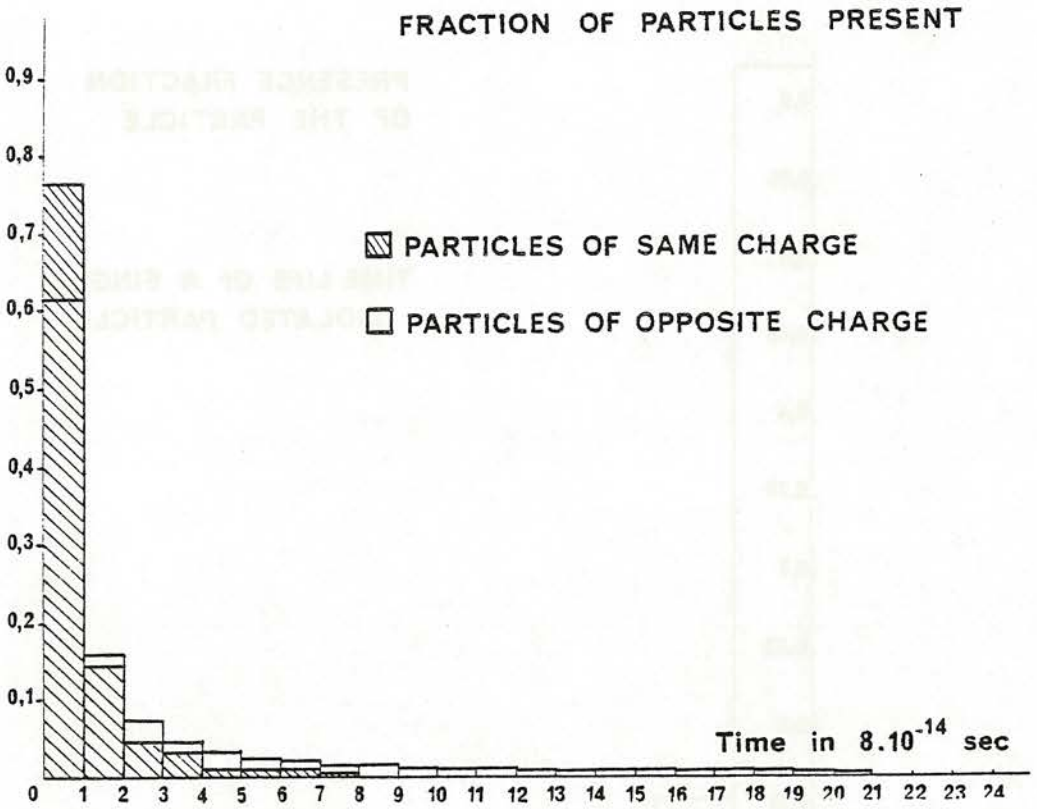


FIG IV b

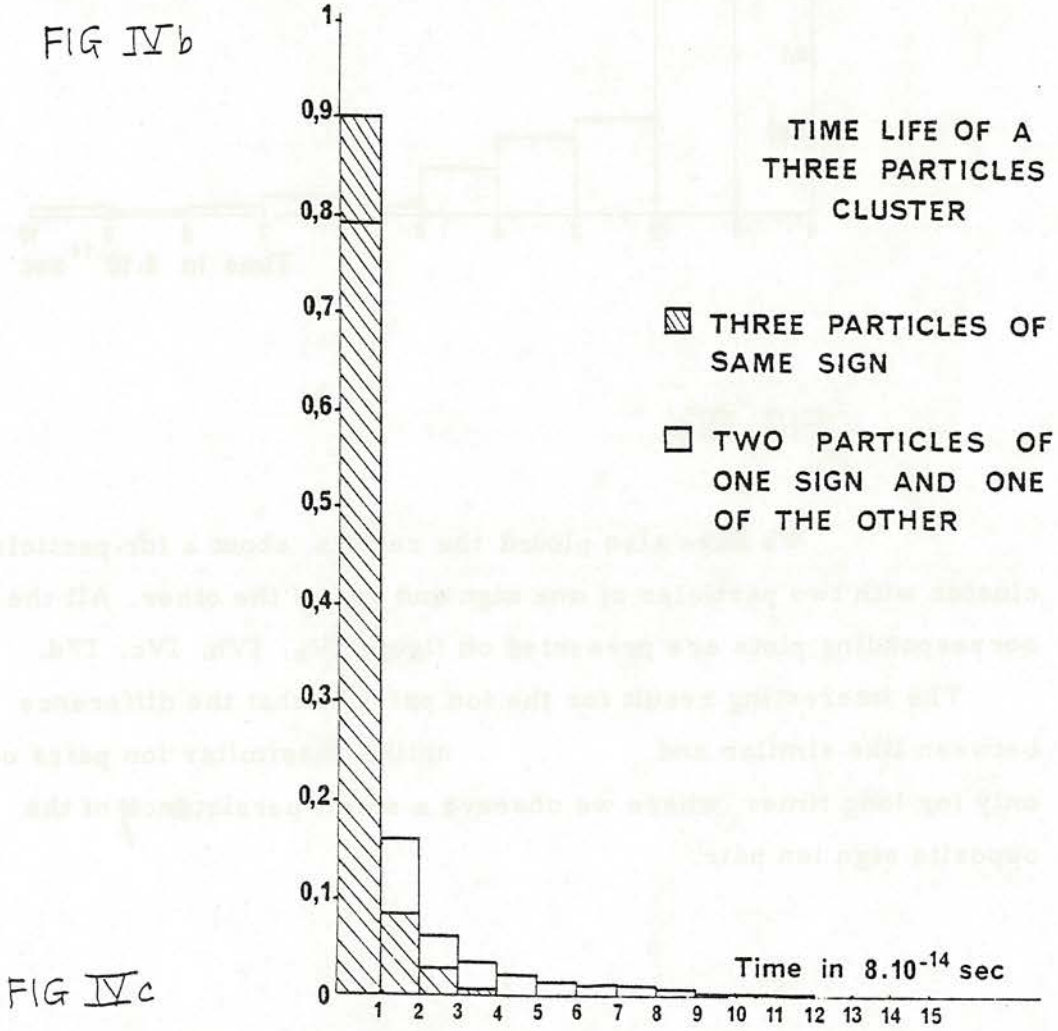


FIG IV c

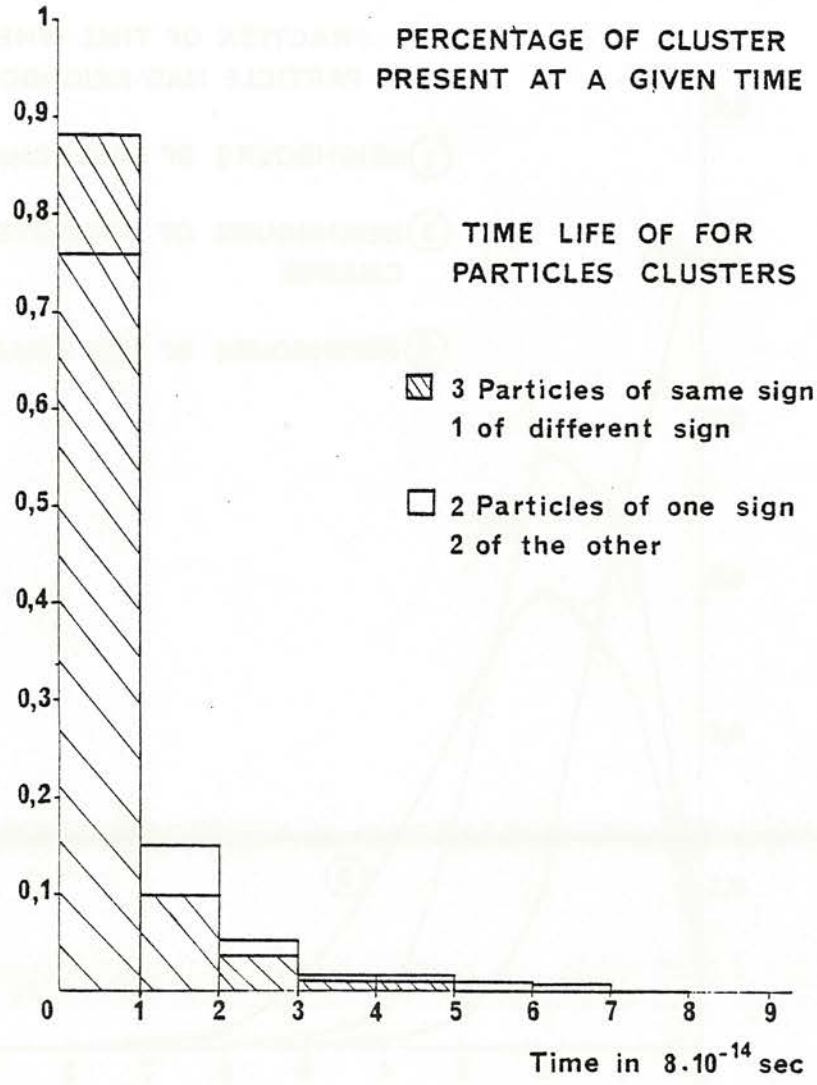


FIG IV d

The triplets ++- or --+ have a time-life of the order of magnitude of that for +- pairs. Even the bipairs clusters $\begin{smallmatrix}+\\-\end{smallmatrix}$ $\begin{smallmatrix}-\\+\end{smallmatrix}$ contribute noticeably to the time structure of the solution.

The last test we did was devoted to the proportion of time where a particle has n neighbours of the same sign, of the opposite sign and of any sign. Each curve is normalized to unity. The results plotted on figure V show that the occurence of dissimular triplets is more frequent than that of different ion pairs.

In all these calculations the cut off distance was taken at 7.5 Å since we have seen from g ++ (r), g -- (r) and g + - (r) that much of the structure of the solution is contained in such a cell.

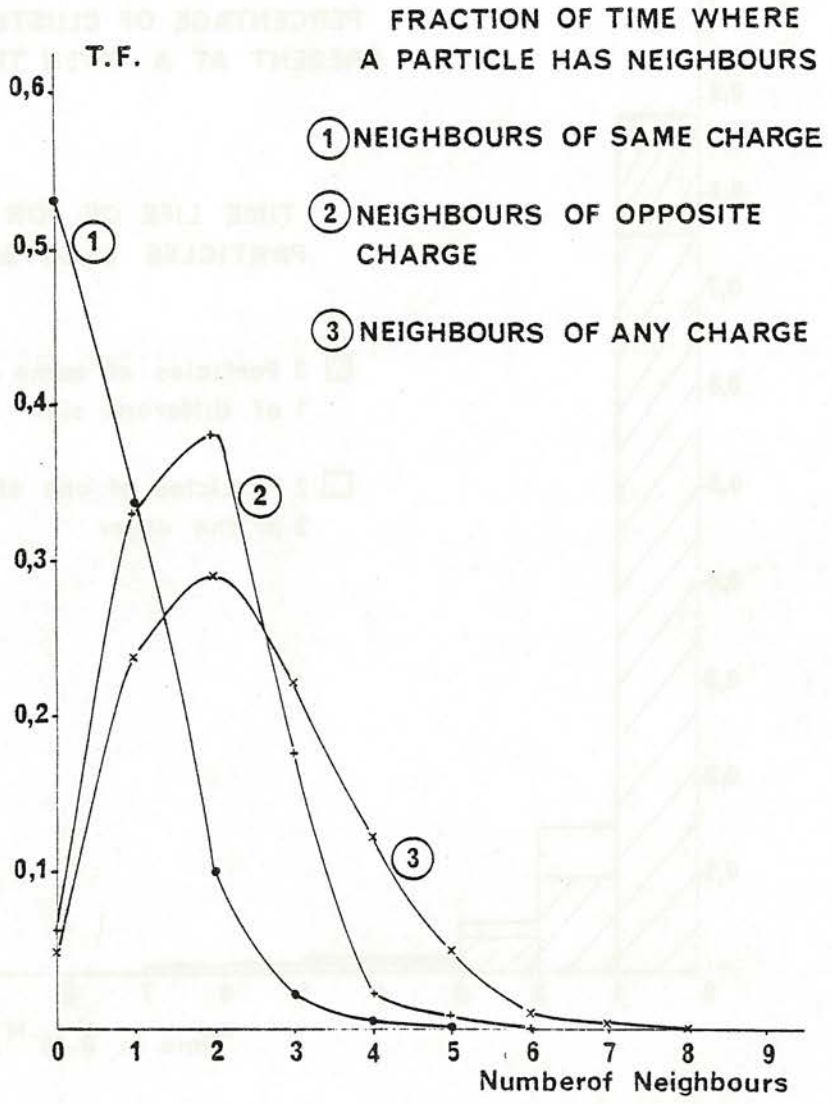


FIG V

V - Discussion.

All these results show that the concept of ionpairs has only a geometrical and thermodynamical signification in the analysis of the instantaneous structure of the electrolyte solution.

The transport properties, self diffusion and electrical conductance concord in the chemical model, as regards to the pair proportion which can be computed. The pair motion in this chemical model is described by a well-defined self-diffusion coefficient, but the pair proportion is considerably smaller than for the thermodynamical properties. A more careful analysis of the time correlations

shows the limits of the pair concept in ionic solutions. Unlike ion pairs don't occur significantly more often than unlike ionic triplets, at least wich regard to the chosen system. The relatively small density of this system corresponding to an 1M solution of a 2-2 electrolyte permits one to predict that in more concentrated, media, solutions or molten salts, individual ion pairs are not well defined entities.

The main conclusion of this work is therefore that ion pairs in solutions are only a convenient language to describe thermodynamical and transport properties, without any real physical significance, except, for dilute solutions where ion pairs can be insulated without any doubt from the surrounding ions. In other cases the correct concept is that of fluctuating clusters depending on the special characteristics of each solution.

Bibliography

1 - P. TURQ, F. LANTELME and H. L. FRIEDMAN "Brownian Dynamics, its application to ionic solutions " to be published.

2 - a) B. Y. ALDER and T. E. WAINWRIGHT J. Chem. Phys. 31, 459 (1959)

b) A. RAHMAN Phys. Rev. A 136 405 (1964)

c) L. VERLET Phys. Rev. A 159 98 (1967)

3 - F. LANTELME, P. TURQ, B. QUENTREC and J. W. E. LEWIS Molecular Physics 28, 1537 (1974)

4 - J. H. WEINER and R. E. FORMAN Phys. Rev B 10 315 (1974)

5 - D. L. ERMAK, J. Chem. Phys. 62, 4189 (1975)

6 - S. A. ADELMAN and J. D. DOLL J. Chem. Phys. 64 724 (1976)

7 - J.D. DOLL and D.R. DION Chem. Phys. Letters 37, 386 (1976)

8 - H.L. FRIEDMAN, C.V. KRISHNAN and L.P. HWANG Chapter in "Structure of water and aqueous solutions". (Verlag Chemie Gmbh 1974) W. Luck editor.

9 - E.A. GUGGENHEIM J. Phys. Chem. 66 , 508 (1962)

10 - H.L. FRIEDMAN "ionic solution theory based on cluster expansion methods". Interscience New-York N.Y. 1962.

11 - a) H.C. ANDERSEN, D. CHANDLER and J.D. WEEKS, J. Chem Phys. 57 2626 (1972)

b) G. STELL and K.C. WU J. Chem. Phys. 63 , 491 (1975)

12 - J.C. RAHMAN and H.L. FRIEDMAN J. Chem. Phys. 48 2742 (1968)

13 - E. WAISMAN and J.L. LEBOWITZ J. Chem. Phys. 56 , 3086 3093 (1972)

14 - R. TRIOLO, J.R. GRIGERA and L. BLUM J. Phys. Chem. 80 1858 (1976)

15 - N. BJERRUM K. DANSKE Videnske Selsk 7n°9 1926

16 - a) H. FALKENHAGEN Théorie der Elektrolyte (Hirzel Leipzig 1971)

b) M.C. JUSTICE and J.C. JUSTICE Journal of Solution Chemistry 5 543 (1976)

17 - H.L. FRIEDMAN J. Chem. Phys. 42 , 450 562 (1965)

18 - R. KUBO Reports of Progress in Physics 29 , 255 (1966)

19 - J.P. HANSEN and I.R. Mac DONALD Phys. Rev 11 A 211 (1975)

I.4

BROWNIAN DYNAMICS TECHNIQUES AND THEIR APPLICATION TO DILUTE SOLUTIONS

D.L.Ermak

Biomedical & Environmental Sciences Division L. 523
Lawrence Livermore Laboratory, Livermore, California 94550 (USA).

ABSTRACT

Monte Carlo techniques are derived for simulating the dynamic behavior of solute particles in solution. Using the Langevin-Fokker-Planck model for interacting Brownian particles, the interaction between a solute particle and the surrounding solvent is described in a statistical manner rather than attempting to treat each solute particle-solvent particle interaction. The methods presented in this paper allow considerable freedom in selecting the size of the time step. Application of these techniques to dilute solutions is discussed.

This work was performed under the auspices of the Centre Europeen de Calcul Atomique et Moleculaire, Centre National de la Recherche Scientifique, France and in part by the U.S. Energy Research and Development Administration under Contract #W-7505-ENG-48.

I. Introduction

The Langevin model of Brownian motion has been used extensively to describe the dynamic behavior of particles in solution under conditions of thermodynamic equilibrium. The model describes the interaction between a solute particle and the surrounding solvent in a statistical manner, rather than attempting to treat each individual solute particle-solvent particle interaction. The force on the solute particle due to the solvent is assumed to be composed of two parts: a frictional force which is proportional to the solute particle velocity; and a randomly fluctuating force which can be described only in terms of its statistical properties. The Langevin equation is obtained by using this force in Newton's equation of motion.

This model has been used recently by Turq, et al. (1) and Ermak (2) to study the excess properties of charged particles in solution. In these works, the Langevin equation is extended to include the Coulomb interaction between the solvent particles. For a system of N solute particles, there are N equations of motion which are coupled together through the Coulomb force. The mathematical solution of these N coupled equations is obtained in terms of the average properties of the solute particles using Monte Carlo simulations.

Several Monte Carlo techniques have been used to simulate the Langevin equation. Turq, et al. (1) and Weiner and Forman (3) use a numerical integration of the Langevin equation to simulate the solute particle trajectories. Consequently, these methods can be used only when the time step between successive displacements is small in comparison to the decay time β^{-1} of the velocity autocorrelation function. Doll and Dion (4) have attempted a more general approach; however, they do not include the interdependence of the velocity change and displacement. The method used by Ermak (2) is derived from the diffusive limit of the Langevin equation and is therefore applicable only when the time step is much longer than the decay time β^{-1}.

The purpose of this work is to present a more general method for simulating the N-body system of interacting Langevin particles. The method is based upon the analytic solution of the stochastic Langevin equation for a single particle in a constant force field. This approach allows the use of intermediate time steps as well as time steps which are either much

shorter or much longer than β^{-1}. The application of these techniques to very dilute systems is also discussed.

2. Langevin Description

The system to be simulated consists of N solute particles immersed in a solvent or heat bath of volume V. The particles are in thermodynamic equilibrium with the solvent at temperature T. In this work only translational motion will be considered. Designating the pair potential between solute particles as $\phi(r_{ij})$, the force on particle "i" due to the remaining N-1 particles is

$$\vec{F}_i = - \sum_{j \neq i}^{N} \nabla_{\vec{r}_i} \phi(r_{ij}) \qquad (1)$$

Generally, periodic boundary conditions are used to keep the number of particles reasonably small while avoiding surface effects. Under this condition, Eq. (1) must be modified to include the images (5).

The Langevin equation of motion for each solute particle is

$$m_i \frac{d\vec{v}_i}{dt} = \vec{F}_i - m_i \beta_i \vec{v}_i + \vec{A}_i \quad ; \; i=1,N \qquad (2a)$$

where t is time, m_i, β_i, and $\vec{v}_i$ are the mass, friction constant, and velocity of particle "i", and $\vec{A}_i$ is a random force. As previously mentioned, the friction force and random force are used to approximate the interaction of the solute particle with the solvent. The random force is assumed to be independent of the particle velocity and to fluctuate rapidly compared to changes in the velocity. The net random force during the time interval t is

$$\vec{B}_i(t) = \int_{t_o}^{t_o+t} \vec{A}_i(t')dt' \qquad (2b)$$

where $\vec{B}_i(t)$ has only statistically defined properties and is independent of the net random force on each of the other particles ($\langle \vec{B}_i(t) \cdot \vec{B}_j(t) \rangle = 0 \; ; \; i \neq j$). The probability of different values of $\vec{B}_i(t)$ is governed by the distribution function

$$w(\vec{B}_i,t) = (4\pi m_i\beta_i kTt)^{-3/2} \exp\left\{\frac{-B_i^2(t)}{4m_i\beta_i kTt}\right\} \qquad (2c)$$

Together Eqs. (2a), (2b), and (2c) completely describe the dynamic behavior of the N particles in the Langevin model.

Expressions for the velocity and position can be obtained by successive integrations of Eq. (2a) and by restricting the size of the time step t such that the solute-solute force $\vec{F}_i$ remains essentially constant during t. In general, this condition on the size of the time step is much less restrictive than the condition $t<<\beta^{-1}$. Dropping the subscripts for convenience, the first integration of Eq. (2a) yields the equation

$$\vec{v}(t) - \vec{v}(o)e^{-\beta t} - \frac{F(o)}{m\beta}\cdot(1 - e^{-\beta t}) = \frac{1}{m}\int_0^t e^{-\beta(t-t')}\vec{A}(t')dt' \qquad (3a)$$

A second integration plus an integration by parts on the $\vec{A}(t)$ term produces the result

$$\vec{r}(t) - \vec{r}(o) - \frac{\vec{v}(o)}{\beta}\cdot(1 - e^{-\beta t}) - \frac{\vec{F}(o)}{m\beta}\cdot\left[t - \frac{1}{\beta}(1 - e^{-\beta t})\right] =$$

$$\frac{1}{m\beta}\int_0^t [1 - e^{-\beta(t-t')}]\vec{A}(t')dt' \qquad (3b)$$

Eqs. (3a) and (3b) cannot be used directly to calculate the velocity and position change due to the statistical nature of $\vec{A}(t)$. One must first calculate the bivariate probability distribution $w(\vec{r},\vec{v},t)$ which governs the probability that a particle, initially located at $\vec{r}_o$ with velocity $\vec{v}_o$ and experiencing the solute force $\vec{F}_o$, will be located at $\vec{r}$ with velocity $\vec{v}$ at time t. For a random force with the properties given by Eq. (2c), Chandrasekar (6) has shown that $w(\vec{r},\vec{v},t)$ is given by

$$w(\vec{r},\vec{v},t) = [4\pi^2(EG-H^2)]^{-3/2} \exp\left\{\frac{-(GR^2-2H\vec{R}\cdot\vec{S}+ES^2)}{2(EG-H^2)}\right\} \qquad (4)$$

where $\vec{R} = \vec{r}-\vec{r}_o - \frac{\vec{v}_o}{\beta}(1 - e^{-\beta t}) - \frac{\vec{F}_o}{m\beta}[t - \frac{1}{\beta}(1 - e^{-\beta t})]$

$$\vec{S} = \vec{v} - \vec{v}_o e^{-\beta t} - \frac{\vec{F}_o}{m\beta}(1 - e^{-\beta t})$$

$$G = \frac{kT}{m}(1 - e^{-2\beta t})$$

$$H = \frac{kT}{m\beta}(1 - e^{-\beta t})^2$$

$$E = \frac{kT}{m\beta^2}(2\beta t - 3 + 4e^{-\beta t} - e^{-2\beta t})$$

Simulation of the particle trajectories through phase space is based entirely upon the bivariate probability distribution. Starting from an initial phase space configuration ($\{\vec{r}_i(o)\}$, $\{\vec{v}_i(o)\}$;i = 1,N), the solute force on each particle ($\{\vec{F}_i(o)\}$;i = 1,N) is calculated. A new set of velocity and position coordinates ($\{\vec{r}_i(t)\}$, $\{\vec{v}_i(t)\}$;i = 1,N) is then chosen for each particle in accordance with the bivariate probability distribution. Methods for making this selection are discussed in Sec. (4). Using the new configuration, the process is repeated for the duration of the simulation.

3. Fokker-Planck Description

An equivalent description of the N-body system of interacting Brownian particles is given by the Fokker-Planck equation (7)

$$\frac{\partial W}{\partial t} + \sum_{i=1}^{N}(\vec{v}_i \cdot \nabla_{\vec{r}_i} W + \frac{1}{m_i}\vec{F}_i \nabla_{\vec{v}_i} W) = \sum_{i=1}^{N} \beta_i(\nabla_{\vec{v}_i} \cdot \vec{v}_i W + kT \nabla^2_{\vec{v}_i} W) \qquad (5)$$

Here $W = W(\{\vec{r}_i\}, \{\vec{v}_i\}, t)$ is the probability distribution function governing the velocity and position of the N solute particles. Eq. (5) neglects hydrodynamic interactions between solute particles as does the Langevin equation (2a).

The equivalence of the Langevin and Fokker-Planck descriptions can be easily shown by transforming Eq. (5) into N coupled equations. Assuming a solution of the form

$$W(\{\vec{r}_i\}, \{\vec{v}_i\}, t) = \prod_{i=1}^{N} w_i(\vec{r}_i, \vec{v}_i, t) \tag{6}$$

one obtains N equations

$$\frac{\partial w_i}{\partial t} + \vec{v}_i \cdot \nabla_{\vec{r}_i} w_i + \frac{1}{m_i} \vec{F}_i \cdot \nabla_{\vec{v}_i} w_i = \beta_i \left(\nabla_{\vec{v}_i} \cdot \vec{v}_i w_i + kT \nabla_{\vec{v}_i}^2 w_i\right) \tag{7}$$

which are coupled together through the solute force $\vec{F}_i$. If t is restricted to be sufficiently short so that $\vec{F}_i$ can be treated as a constant, then the solution to Eq. (7) is readily seen to be given by Eq. (4).

Returning to Eq. (5), the steady state solution to this equation is

$$W(\{\vec{r}_i\}, \{\vec{v}_i\}, \infty) = W_o \exp\left\{-\frac{(K+P)}{kT}\right\} \tag{8}$$

where $K = \frac{1}{2} \sum_{i=1}^{N} m_i v_i^2$

$$P = \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \phi(r_{ij})$$

W_o = a constant

The total energy of the N solute particles can be defined as the sum of the kinetic energy and the potential energy which arises from the solute-solute interaction. According to Eq. (8) the probability of occurrance of the

state described by the phase space coordinates $(\{\vec{r}_i\}, \{\vec{v}_i\})$ is proportional to the negative exponential of the total energy divided by kT. The system of solute particles can therefore be considered to be a member of a canonical ensemble.

4. Equations of Motion

There are numerous methods by which the new phase space coordinates can be chosen from $w(\vec{r},\vec{v},t)$. One method is to first select the new velocity, irrespective of the new position, according to the distribution

$$w_1(\vec{v},t) = \int w(\vec{r},\vec{v},t)d^3r \tag{9}$$

This procedure is equivalent to the velocity equation

$$\vec{v}(t) = \vec{v}(o)e^{-\beta t} + \frac{\vec{F}(o)}{m\beta}(1 - e^{-\beta t}) + \vec{B}_1(t) \tag{10a}$$

where $\vec{B}_1(t)$ is a random velocity change chosen from a gaussian distribution with the properties

$$\langle \vec{B}_1(t) \rangle = 0$$

$$\langle B_1^2(t) \rangle = \frac{3kT}{m}(1 - e^{-2\beta t}) \tag{10c}$$

With the new velocity selected, the new position is chosen from the distribution

$$w_2(\vec{r},t) = \frac{w(\vec{r},\vec{v},\vec{t})}{w_1(\vec{v},t)} \tag{11}$$

This procedure is equivalent to the displacement equation

$$\vec{r}(t) = \vec{r}(o) + \frac{1}{\beta}(\vec{v}(t) + \vec{v}(o) - \frac{2\vec{F}(o)}{m\beta}) \cdot \frac{(1 - e^{-\beta t})}{(1 + e^{-\beta t})} + \frac{\vec{F}(o)}{m\beta} t + \vec{B}_2(t) \tag{12a}$$

where $\vec{B}_2(t)$ is a random displacement chosen from a gaussian distribution with the properties

$$\langle \vec{B}_2(t) \rangle = \langle \vec{B}_1(t) \cdot \vec{B}_2(t) \rangle = 0 \qquad (12b)$$

$$\langle B_2^2(t) \rangle = \frac{6kT}{m\beta^2} \left[\beta t - 2 \frac{(1 - e^{-\beta t})}{(1 + e^{-\beta t})} \right] \qquad (12c)$$

Eqs. (10) and (12) have a linear dependence upon $\vec{r}(o)$, $\vec{v}(o)$ and $\vec{F}(o)$ with non-linear coefficients in time t. The non-linear coefficients will not increase the computing time for problems with fixed time steps as they will have to be calculated only once at the beginning of the simulation. The displacement also has a linear dependence upon the new velocity.

A second set of equations can be obtained by first selecting the new position and then the new velocity. In this case the equations are

$$\vec{r}(t) = \vec{r}(o) + \frac{\vec{v}(o)}{\beta} (1 - e^{-\beta t}) + \frac{\vec{F}(o)}{m\beta} \left[t - \frac{1}{\beta} (1 - e^{-\beta t}) \right] + \vec{B}_1(t) \qquad (13a)$$

$$\vec{v}(t) = \vec{v}(o) \cdot (2\beta t e^{-\beta t} - 1 + e^{-2\beta t})/C(t) + \beta[\vec{r}(t) - \vec{r}(o)] (1 - e^{-\beta t})^2/C(t)$$

$$+ \frac{\vec{F}(o)}{m\beta} \left[\beta t (1 - e^{-2\beta t}) - 2(1 - e^{-\beta t})^2 \right] /C(t) + \vec{B}_2(t) \qquad (14a)$$

where again $\vec{B}_1(t)$ and $\vec{B}_2(t)$ are independent random functions of time chosen from separate gaussian distributions with the properties

$$\langle \vec{B}_1(t) \rangle = \langle \vec{B}_2(t) \rangle = \langle \vec{B}_1(t) \cdot \vec{B}_2(t) \rangle = 0 \qquad (13b, 14b)$$

$$\langle B_1^2(t) \rangle = \frac{3kT}{m\beta^2} C(t) \qquad (13c)$$

$$\langle B_2^2(t) \rangle = \frac{6kT}{m} \left[\beta t (1 - e^{-2\beta t}) - 2(1 - e^{-\beta t})^2 \right] /C(t) \qquad (14c)$$

$$C(t) = 2\beta t - 3 + 4e^{-\beta t} - e^{-2\beta t} \qquad (14d)$$

Comparing Eqs. (10) and (12) with Eqs. (13) and (14), there appears to be two different sets of equations for the motion of the particles. However, both sets of equations are equivalent in the sense that they both select the new phase space coordinates in accordance with the bivariate probability distribution $w(\vec{r}, \vec{v}, t)$. Therefore, the probability of a particular velocity and position change is governed by $w(\vec{r}, \vec{v}, t)$ and Eqs. (10) and (12) and Eqs. (13) and (14) are just different methods of sampling this distribution.

In order to demonstrate the equivalence of these two methods, a number of simulations were conducted on systems of non-interacting particles using both methods. The statistical properties of non-interacting particles can be calculated analytically for any length of time using Eq. (4) and compared with the simulation results. Using the dimensionless variables: $\tau = \beta t$, $\vec{u} = (\frac{m}{3kT})^{\frac{1}{2}} \vec{v}$, and $\vec{x} = \beta(\frac{m}{3kT})^{\frac{1}{2}} \vec{r}$, the analytic results for the calculated properties are

$$\langle \vec{u}(\tau) \cdot \vec{u}(o) \rangle = e^{-\tau} \tag{15a}$$

$$\langle [x(\tau) - x(o)]^2 \rangle = 2(\tau - 1 + e^{-\tau}) \tag{15b}$$

$$\langle u(\tau) \cdot [x(\tau) - x(o)] \rangle = 1 - e^{-\tau} \tag{15c}$$

A number of simulations were conducted using time steps ranging from .01 to 100. Both methods produced trajectories whose average properties were equal to the analytic results. Table 1 shows the results for $\tau = .5$ and using 100,000 trajectories to calculate the averages. The equivalence of the two methods is formally shown in the Appendix.

5. Application to Dilute Solutions

In a Monte Carlo simulation of the N-body solute system, the time step must be short enough to ensure accuracy and yet long enough to be within the practical limitation of allowable computer time. The proper choice of the time step size is a particular problem when studying very dilute solutions. A relatively long time step (and consequently long displacement) is desirable in order to more rapidly sample configuration space. However, a long time

TABLE 1

τ	$\langle u(\tau)\cdot u(0)\rangle$			$\langle u(\tau)\cdot[x(\tau)-x(0)]\rangle$			$\langle[x(\tau)-x(0)]^2\rangle$		
	I	II	III	I	II	III	I	II	III
.0	1.000	1.001	.998	.000	.000	.000	.000	.000	.000
.5	.607	.607	.605	.393	.394	.393	.213	.213	.213
1.0	.368	.368	.366	.632	.633	.632	.736	.736	.736
1.5	.223	.221	.221	.777	.779	.778	1.446	1.448	1.447
2.0	.135	.134	.134	.865	.866	.864	2.271	2.274	2.272
2.5	.082	.080	.079	.918	.917	.914	3.164	3.168	3.163
3.0	.050	.048	.047	.950	.949	.947	4.100	4.103	4.095
3.5	.030	.028	.027	.970	.968	.965	5.060	5.061	5.051
4.0	.018	.019	.018	.982	.983	.979	6.037	6.037	6.025
4.5	.011	.011	.009	.989	.986	.985	7.022	7.022	7.007
5.0	.007	.007	.006	.993	.994	.994	8.013	8.014	7.996
5.5	.004	.004	.005	.996	.999	.999	9.008	9.011	8.992
6.0	.002	.003	.003	.998	1.001	1.000	10.00	10.01	9.993
6.5	.002	.001	.000	.998	.999	.995	11.00	11.01	10.99
7.0	.001	.000	.000	.999	.996	.995	12.00	12.01	11.99
7.5	.001	.001	.001	.999	.994	.987	13.00	13.00	12.98
8.0	.000	.001	.004	1.000	.997	.996	14.00	14.00	13.97

I - Analytic results from Eqs. (15)

II - Simulation results using Eqs. (10) and (12).

III - Simulation results using Eqs. (13) and (14).

step will produce very inaccurate results when two solute particles are close together, a situation which might occur frequently if the force between them is attractive.

This situation might be improved upon through the use of a variable length time step such that a long time step is used when the particle is far from any other particle in the system. When two particles are close together, many short time steps (the sum of which equals the long time step) would be used. The simulation techniques presented in the previous sections could be readily used in this manner.

As a simple illustration, this approach is applied to a model solution containing infinitely long charged rods, charged spheres whose charge is opposite in sign to that of the rods, and solvent in which the two types of ions are immersed. The dynamic behavior of the spheres is described by the Langevin-Fokker-Planck model. The rods are immobile and are distributed throughout the solution in a square lattice of side length L. Consequently, only the trajectories of the spheres is of interest. The parameters used in these calculations are: rod diameter = 2Å; sphere diameter = 2Å, mass m = 65.4×10^{-24}gm friction constant $\beta = 3.1 \times 10^{+13}$ sec^{-1}; dielectric constant $\varepsilon = 78.54$; temperature T = 20°C; L = 167.1 Å.

For simplicity, the minimum image technique (5) is used with only one sphere per cubic cell of side length L. Assuming a uniform surface charge for both the rods and spheres, the force on a sphere located a distance ρ from a rod is

$$\vec{F} = -\frac{2Ne^2}{\varepsilon L \rho}\hat{\rho} \qquad (16)$$

where the product of the sphere charge and the rod charge per unit length is $-Ne^2/L$ with e being the charge of an electron. The Manning ζ parameter (8) is equal to $Ne^2/\varepsilon kTL$.

The length of the time step t was calcualted as a function of the sphere-rod separation distance ρ

$$M = \text{INT}\,[1 + (4750 \cdot \rho)/(128 + \rho^3)] \qquad (17)$$

$$t = 10^{-11}/M \quad \text{(seconds)}$$

where INT(x) equals the largest integer less than or equal to x and ρ is in Å. Thus, the time step ranged from 10^{-13} to 10^{-11} seconds. Discrete values of t were used so that the coefficients of time in the equations of motion would have to be calculated only once. The time step size as a function of the sphere-rod separation distance is shown in Table 2 for several values of ρ. Also shown is the sphere root mean square displacement as a function of time step size, $\langle r^2(t)\rangle^{\frac{1}{2}} = \{ \frac{kT}{m\beta} [t - \frac{1}{\beta} (1 - e^{-\beta t})] \}^{\frac{1}{2}}$.

Table 2

ρ(Å)	t (sec x 10^{-13})	$\langle r^2\rangle^{\frac{1}{2}}$ (Å)
4	1.01	.17
6	1.20	.19
10	2.33	.28
14	4.17	.39
20	8.33	.57
30	16.7	.81
40	33.3	1.15
50	50.0	1.41
70	100.0	2.00

Using this variable time step method, simulations were conducted with the charge parameter N varied from 0 to 28. The sphere mean square velocity, mean square displacement, distribution function relative to the rods, and diffusion coefficient were calculated. These results were compared to those from similar simulation using a constant time step of $t = 10^{-13}$ sec. The results were in agreement to within the statistical error of the calculations.

The simulations using the variable time step were significantly faster than those using the constant time step. This is shown in Table 3 by the parameter R defined as

$$R = \frac{\text{computer time using constant time step method}}{\text{computer time using variable time step method}}$$

The maximum reduction was by a factor of 28 when N = 0. As N is increased the spheres spend more time close to the rods. This is evident from the decrease in the mean sphere-rod separation distance $\langle\rho\rangle$ and the decrease in the sphere diffusion coefficient D.

Table 3

N	$Ne^2/\varepsilon kTL$	$\langle\rho\rangle$(Å)	D(cm^2/sec x 10^5)	R
0	.0	66.	2.0	28.
7	.3	61.	2.0	16.
14	.6	53.	1.7	8.3
21	.9	41.	1.0	4.0
28	1.2	28.	.4	2.3

These calculations were conducted only as an indicator of the potential for reducing computer time using a variable time step method. The results are quite favorable, indicating the possibility of significant time reduction in more complicated problems. When using the variable time step approach on a system with a large number of particles, a method is needed for properly ordering the sequence in which the particle displacements are made. Particles which are close together must undergo their many small displacements simultaneously. Also, better methods for choosing the size of the time step are needed. A possible approach is to use information on the first derivative of the solute force rather than the nearest neighbor distance.

Appendix

A third set of equations for the velocity and displacement changes can be deduced directly from Eq. (4). The equations are

$$\vec{v}(t) = \vec{v}(o)e^{-\beta t} + \frac{F(o)}{m\beta}(1 - e^{-\beta t}) + \vec{B}_1(t) \qquad \text{(Ia)}$$

$$r(\vec{t}) = \vec{r}(o) + \frac{\vec{v}(o)}{\beta}(1 - e^{-\beta t}) + \frac{F(o)}{m\beta}\left[t - \frac{1}{\beta}(1 - e^{-\beta t})\right] + \vec{B}_2(t) \qquad \text{(Ib)}$$

where $\vec{B}_1(t)$ and $\vec{B}_2(t)$ are random functions of time chosen from the bivariate gaussian distribution with the properties

$$\langle \vec{B}_1(t) \rangle = \langle \vec{B}_2(t) \rangle = 0 \qquad \text{(Ic)}$$

$$\langle \vec{B}_1(t) \cdot \vec{B}_2(t) \rangle = 3H \qquad \text{(Id)}$$

$$\langle B_1^2(t) \rangle = 3G \qquad \text{(Ie)}$$

$$\langle B_2^2(t) \rangle = 3E \qquad \text{(If)}$$

and H, G, and E are given by Eq. (4). Eq. (Id) expresses the interdependence of the random functions $\vec{B}_1(t)$ and $\vec{B}_2(t)$. This set of equations can also be obtained by keeping Eq. (10) and using it to replace $\vec{v}(t)$ in Eq. (12) or by keeping Eq. (13) and using it to replace $\vec{r}(t) - \vec{r}(o)$ in Eq. (14).

References

1. P. Turq, F. Lantelme, and H.L. Friedman, J. Chem. Phys., in press.
2. D.L. Ermak, J. Chem. Phys. 62 (1975) 4189,4197.
3. J.H. Weiner and R.E. Forman, Phys. Rev. B 10 (1974) 315.
4. J.D. Doll and D.R. Dion, Chem. Phys. Lett. 74 (1975) 386.
5. S.G. Brush, H.L. Sahlin, and E. Teller, J. Chem. Phys. 45 (1966) 2102.
6. S. Chandrasekhar, Rev. Mod. Phys. 15 (1943) 1.
7. G. Wilemski, J. Stat. Phys. 14 (1976) 153.
8. G.S. Manning, J. Chem. Phys. 51 (1969) 924,934.

II

ACCURATE DYNAMICS ON BIO-MACROMOLECULAR SYSTEMS

II

ACCURATE DYNAMICS ON BIO-MACROMOLECULAR SYSTEMS

II.1

ALGORITHMS FOR MACROMOLECULAR DYNAMICS AND CONSTRAINT DYNAMICS

W.F.van Gunsteren
H.J.C.Berendsen

University of Groningen, Laboratory of Physical Chemistry,
Zernikelaan, Groningen (Pays Bas).

1. *Introduction*

In the field of liquids, until now the method of molecular dynamics (MD) [1] has only been applied to liquids containing not too large molecules, such as argon [2], sodium [3], molten salts [4], water [5], nitrogen [6], n-alkanes [7] etc. Recently, preliminary results of a MD calculation for a small macromolecule, viz. the protein bovine pancreatic trypsin inhibitor (BPTI) have become available [8]. This is a promising first step towards the understanding of the behaviour of macromolecules in terms of interactions on the atomic level. However, the present methods are still not sufficiently reliable or appropriate to simulate many of the significant macromolecular properties: the interaction with the solvent is not yet included, the long-range (Coulomb) interaction is treated incorrectly, and the time step allowed in the computation is too short to permit simulation of relatively slow events such as major conformational changes and protein folding.

Before attacking these major problems, it should be checked whether the algorithms that have been used in MD calculations of simple liquids, will also produce optimum results when applied to macromolecules. A macromolecule differs from a simple liquid by the more complex nature of the interaction potentials and forces and by the presence of a large number of covalent bonds. Since the latter represent the highest components in the frequency distribution of the molecular motion, the introduction of constraints for bond lengths and angles is expected to increase the computational efficiency.

In this paper we investigate the following points:

1. Various algorithms are in use for integrating the equations of motion in a MD calculation. The ones that are most frequently used are those proposed by Verlet [9] and by Gear [10,11]. Recently, Beeman [12] has formulated an algorithm that he claimed to be better than those of Verlet and Gear. Our first point of investigation concerns finding the best presently available algorithm to be used in a MD calculation on a macromolecule. Thereby one should keep in mind that the mathematical treatment of the numerical solution of differential equations is still far from being complete, and that the answer is partially dependent on the properties of the physical system under consideration.

2. Our second point of investigation concerns finding the best algorithms for dynamics in the presence of constraints. The introduction of constraints in a dynamical calculation is only allowed (i.e., will yield physically correct results) if a the frequency components of the motion along the eliminated degrees of freedom are well separated from the other frequencies occurring in the system of particles, and b the coupling between both types of motion is weak. The fulfilment of these conditions depends on the system under consideration: we shall in this investigation assume that the use of constraints is physically justified and evaluate the various algorithms on the basis of time step, accuracy and computer efficiency.

The algorithms for MD without and with constraints are discussed in sections 2 and 3. In section 4 numerical results of the application of these algorithms to BPTI are given. Section 5 contains a summary and conclusions. In an appendix more elaborate data about the best algorithms for macromolecular dynamics are presented.

2. *Algorithms for MD without constraints*

The MD method consists essentially of solving the set of coupled second-order differential equations

$$\frac{d^2\vec{r}_i}{dt^2} = \vec{F}_i\,(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N)/m_i \qquad i = 1, 2, \ldots N \tag{2.1}$$

that are governing the classical dynamical behaviour of a system of N particles, in order to find the positions $\{\vec{r}_i\ (t)\}$, velocities $\{\vec{v}_i\ (t)\}$, etc. of the particles as a function of time t. The initial positions $\{\vec{r}_i\ (t_o)\}$ and velocities $\{\vec{v}_i\ (t_o)\}$ of the particles must be specified. In systems of interest the number of particles N is typically of the order of 1000. The force on the i^{th} particle is denoted by $\vec{F}_i$ and its mass by m_i. The force is derived from a potential V:

$$\vec{F}_i(\vec{r}_1, \vec{r}_2, \ldots \vec{r}_N) = -\vec{\nabla}_i\, V(\vec{r}_1, \vec{r}_2, \ldots \vec{r}_N) \qquad i = 1,2,\ldots N \tag{2.2}$$

For simplicity the force is taken to be conservative, i.e., only position dependent. However, the discussion below is also valid for non-conservative forces.

Mathematically the problem constitutes an initial value problem [10], which can be simply formulated:

$$y'' = f(y) \tag{2.3a}$$

$$y_o = y(t_o),\ y'_o = y'(t_o) \tag{2.3b}$$

where y, y', y" are 3 N-dimensional vectors and their derivatives with respect to t.

The most appropriate mathematical approach to solving such a problem is the use of a difference or step by step method. The solution is approximated by its value at a sequence of discrete points called mesh points. Normally these points are assumed to be equally spaced.

$$t_i = ih, \quad i = 0, 1, 2, \ldots \tag{2.4}$$

where h is the spacing between adjacent points. A step by step method provides a rule or algorithm for computing the approximation y_i at point (or step) k_i to $y(t_i)$ in terms of the values of y at t_{i-1} and possibly at preceding points. A k-step algorithm uses preceding values of y or its subsequent derivatives up to and including t_{i-k}.

The existence and convergence of a solution depends on the properties of the function f and on the specified initial values y_o and y'_o.

In most MD applications the function f-derived from an interaction potential V-obeys the conditions (Lipschitz and differentiability conditions [10]) for existence and convergence. When singularities occur in the interaction potential V, special measures must be taken in order to avoid divergence problems [5].

The next choice to be made is that between the available difference methods. The most important types are:

1. Runge-Kutta methods
2. extrapolation methods
3. multi-value predictor-corrector methods

The most time consuming part in any method of solution is the evaluation of the function f, because it involves at least in principle a double summation over the N particles making up the system. This feature rules out methods 1 and 2 since they require at least several function evaluations per step. Moreover, predictor-corrector methods can easily be applied directly to a higher-order equation, in stead of indirectly, viz. after converting it to a set of first-order equations. The use of a direct method is advantageous, because the conversion of a second-order equation to two first-order ones will roughly double the amount of information to be processed.

In a multi-value predictor-corrector method the calculated values of y or its (successive) derivatives at a number of mesh points are used to aid in the computation at later points. A k-value method uses k previously calculated values of y or its (successive) derivatives. When these k-values are calculated at 1 previous mesh points, the method is called a 1-step method. A multi-value method can be expressed using different representations. The representation is determined by the choice whether, for each mesh point, y or y' or y" , etc. or combinations of them are used in the algorithm. For example in the Nordsieck [13] or N-representation the values

$$y_n, hy'_n, h^2 y''_n/2, \ldots, h^{k-1} y_n^{(k-1)}/(k-1)! \qquad (2.5)$$

are saved and used in the calculation of y_{n+1}, y'_{n+1}, etc.

But in the well-known Adams-methods [10] the values

$$y_n, h\,y'_n, h\,y'_{n-1}, \ldots\ldots, h\,y'_{n-k+2} \qquad (2.6)$$

are saved and used to calculate y_{n+1} and y'_{n+1}.

In the discussion of the existing k-value predictor-corrector algorithms and of the connection between their different representations, the use of a matrix notation [10,11] is very helpful.

Let us write, in the N-representation, the column vector

$$\underline{y}_n(N) \equiv [y_n, h\,y'_n, h^2 y''_n/2, \ldots\ldots, h^{k-1} y_n^{(k-1)}/(k-1)!]^T . \qquad (2.7)$$

The predictor step can be written as

$$\underline{y}_{n+1,(p)} = \underline{A}\,\underline{y}_n , \qquad (2.8)$$

where the matrix $\underline{A}$ is determined by the specific predictor that is used. The corrector equation for a second-order differential equation reads:

$$\underline{y}_{n+1} = \underline{y}_{n+1,(p)} + \underline{a}\,h^2/2!\,[f(y_{n+1,(p)}) - y''_{n+1,(p)}] \qquad (2.9)$$

The second term in eq. (2.9) represents the amount by which the differential equation (2.3a) is not satisfied locally by $\underline{y}_{n+1}{}'_{(p)}$. The column vector $\underline{a}$ characterizes the specific corrector that is used. The corrector process (2.9) may be repeated for a fixed number of iterations or until no further change in $\underline{y}_{n+1}$ is obtained.
Since every iteration of the corrector (2.9) involves one evaluation of the function f, the number of corrector iterations in MD calculations is usually taken equal to 1.

The parameters of the k-value predictor-corrector algorithm (2.8-9) are the coefficients of the matrix $\underline{A}$ and the column vector $\underline{a}$. Their values characterize the algorithms. The usual choice for the predictor is an extrapolation using a Taylor series of polynomial degree (k-1) that satisfies the retained values of y, y', etc., calculated at preceding mesh points. In that case the matrix $\underline{A}$ is equal to the Pascal triangle (see appendix eq. (A.1)). The remaining parameters, viz. the coefficients of $\underline{a}$ can be chosen to achieve optimum stability and accuracy. How this is done has been described by Gear [10,11]. The resulting values for $\underline{a}$ in the case of a second-order equation are quoted in table 1 in the appendix. Only the last (k-2) coefficients of $\underline{a}$ (a_2, a_3, ...) are determined by stability requirements, so that the remaining first two coefficients (a_o, a_1) are chosen to optimize the accuracy of the algorithm.

Gear [10,11] has shown that two different representations of the same k-value predictor-corrector algorithm are equivalent. This means that the one representation can be converted into the other by a (matrix) transformation (of $\underline{A}$ and $\underline{a}$), that does not affect its stability properties or truncation error, but can affect both round-off error properties and the computational efficiency .

This transformation technique is a useful tool to compare the various existing algorithms.

For example the Beeman [12] algorithm is written in the representation

$$\underline{y}_n(F) \equiv [y_n,\ hy'_n,\ h^2y''_n/2,\ .....,\ h^2y''_{n-k+3}/2]^T \qquad (2.10)$$

which we call the force- or F-representation. By transforming it to the N-representation it can be shown that it only differs from the Gear algorithm in the values of (a_o, a_1). Beeman has used a Taylor expansion of degree (k-1) also for the corrector. But, from test calculations we find that the Gear coefficients (a_o, a_1) produce results, at least twice as accurate as the Beeman coefficients (a_o, a_1) do, as is indeed expected from the optimum choice of (a_o, a_1) by Gear.

The algorithm of Verlet [9] consists of the formulae

$$y_{n+1} = 2\ y_n - y_{n-1} + h^2y''_n \qquad (2.11a)$$

$$y'_n = (y_{n+1} - y_{n-1})/2h \qquad (2.11b)$$

After transforming eq. (2.11a) from the representation

$$\underline{y}_n(V) \equiv [y_n,\ h^2y''_n/2,\ y_{n-1}]^T \qquad (2.12)$$

to the N-representation, it turns out that the Verlet algorithm is equivalent to a 3-value predictor-corrector algorithm without evaluating the corrector.

When neglecting round-off errors, the choice of the appropriate representation of an algorithm is determined by practical considerations. Changing the size of the time step during a MD run is most easily done in the N-representation, since only information of the last preceding step is used in the algorithm. Also for changing the degree of the algorithm (changing the k-value) the N-representation is the best one, since in that representation only the column vector $\underline{a}$ is changing, the predictor matrix $\underline{A}$ remaining the same. In other representations the predictor matrix depends on the k-value of the algorithm (see appendix).

From these theoretical considerations the following conclusion can be drawn. The best (most stable, accurate, fast, etc.) presently available algorithm for MD calculations (without constraints) is a k-value predictor-corrector algorithm, written in the N-representation, with the predictor parameters determined by a Taylor expansion and with the corrector parameters derived from stability and accuracy requirements as proposed by Gear [10]. The optimum k-value and step size h are then to be determined by the particular system under consideration.

3. *Algorithms for MD with constraints*

A MD calculation may be speeded up by a reduction of the number of degrees of freedom. For macromolecules one may think of the elimination of the bond stretching vibrations. Such an elimination can be realized by applying constraint dynamics: the bond lengths are kept fixed at a constant value during the MD run. Although the use of generalized coordinates and the Lagrangian equation of motion is well-known for the elimination of internal degrees of freedom, this method is highly unpractical for application to macromolecules. It seems clear that cartesian coordinates have to be used for large molecules, also in the presence of constraints.

In [14] two methods have been proposed for integrating the cartesian equations of motion of a system of particles subject to holonomic constraints. The one that is appropriate for application to large systems, such as macromolecules, is the procedure called "SHAKE". Its essential feature is that at each step of the MD run the constraints are satisfied by adding displacement vectors to the position vectors of the particles that result from a non-constraint time step. The displacement vectors are determined such that the constraints are satisfied at the final positions. Since all constraints may be interdependent, this resetting of the positions of the particles by SHAKE is an interative procedure that considers all constraints in succession. The relative accuracy tol to which the constraints are to be satisfied must be specified. The computing time required by SHAKE depends on tol. In the discussion below we will denote the use of SHAKE by

$$\text{SHAKE } (y_1, y_2, y_3) \tag{3.1}$$

This means that the positions y_2 that result from the non-constraint step will be reset with as result the constrained positions y_3. The direction

of the displacement vectors (y_3-y_2) is determined by the reference positions y_1.

Until now SHAKE had only been used in connection with the Verlet algorithm, because this algorithm only makes use of positions (y_n, y_{n-1}, see eq. (2.11)), which can be reset by SHAKE. Owing to the absence of a corrector step in the algorithm, the effect of the constraint forces need not be incorporated in y". Below we propose a scheme by which the effect of the constraint forces can be incorporated in the values of y". This scheme opens the way to the application of arbitrary predictor-corrector algorithms to MD with constraints. However, the algorithms must be written in a representation containing (besides one value of y') only values of y and/or y" at the mesh points. So constraint dynamics cannot be performed in the N-representation. The most natural representation to choose is the F-representation (2.10), saving present positions and velocities and present and past forces. The predictor matrix $\underline{A}$ and the corrector vector $\underline{a}$ most be transformed from the N-representation to the F-representation. This is done in appendix (A.2). The resulting matrix $\underline{B}$ and vector $\underline{b}$ are given in eqs. (A. 6-7).

The computational scheme for a MD step with constraints is the following.

1. In the predictor step we calculate from

$$\underline{y}_{n+1,\,(p)} = \underline{B}\,\underline{y}_n \tag{3.2}$$

values for $y_{n+1,\,(p)}$ and $y''_{n+1,\,(p)}$.

2. In eq. (2.3a) the function f represents the force. The total force may be thought to be composed of two parts, the constraint force and the rest of the force:

$$f_{tot} = f_{free} + f_{constr} \tag{3.3}$$

The force f_{free} is derived from the interaction potential, from which the interaction along the constrained degrees of freedom is excluded. The symbol f_{constr} denotes the unknown constraint forces. We cannot directly apply the corrector

$$\underline{y}_{n+1,\,(tot)} = \underline{y}_{n+1,\,(p)} + \underline{b}\tfrac{1}{2}h^2[f_{tot}(y_{n+1,\,(p)}) - y''_{n+1\,(p)}] \tag{3.4}$$

since f_{tot} is still unknown.

3. But, we can use the corrector without constraints:

$$\underline{y}_{n+1,\,(free)} = \underline{y}_{n+1,\,(p)} + \underline{b}\tfrac{1}{2}h^2[f_{free}(y_{n+1,\,(p)}) - y''_{n+1,\,(p)}] \tag{3.5}$$

This formula yields the positions $y_{n+1,\,(free)}$ at t_{n+1}, but without the effect of the constraint forces.

4. The latter effect can be incorporated in the obtained positions by resetting these positions with the aid of SHAKE. So we perform

$$\text{SHAKE } (y_{n+1,\,(p)},\ y_{n+1,\,(free)},\ y_{n+1,\,(tot)}) \tag{3.6}$$

5. Now, the forces of constraint can be obtained from

$$f_{constr}(y_{n+1,\,(p)}) = [y_{n+1,\,(tot)} - y_{n+1,\,(free)}]/(b_o \tfrac{1}{2} h^2) \tag{3.7}$$

This equation is obtained by subtracting the first row of eq. (3.5) from that of eq. (3.4). The total force f_{tot} is found from eq. (3.3).

6. The final step of the scheme consists of using f_{tot} in the corrector (3.4).

In order to assure the stability of this calculational procedure two modifications should be incorporated into the scheme.

a. In eq. (3.7) the f_{constr} is calculated as a difference of two positions. The positions $y_{n+1,\,(tot)}$ result from SHAKE, so they satisfy the constraints. In order to get a correct f_{constr} we must be sure that

$$y_2 = y_{n+1,\,(p)} - b_o \tfrac{1}{2} h^2\, y''_{n+1,\,(p)} \tag{3.8}$$

satisfies the constraints too. Therefore an extra step 2a is performed, reading

$$\text{SHAKE } (y_n,\ y_2,\ y_3) \tag{3.9}$$

The resulting reset positions y_3 are then in step 3 added to $b_o \tfrac{1}{2} h^2 f_{free}\,(y_{n+1,\,(p)})$ in order to get $y_{n+1,\,(free)}$.

b. When in step 6 the velocities y'_{n+1} are calculated using eq. (3.4), errors in the preceding y'_n, y'_{n-1}, etc. may propagate. This source of instability can be avoided by calculating y'_{n+1} from the obtained values of y_{n+1} and $y''_{n+1} = f_{tot}(y_{n+1,\,(p)})$ using the formula

$$y'_{n+1} = \{B_{11} y_{n+1} + \sum_{\substack{i=o \\ i\neq 1}}^{k-1} (B_{1i} - B_{11} B_{oi})\, y_{n,\,i} + (b_1 - B_{11} b_o) \tfrac{1}{2} h^2 [f_{tot}(y_{n+1,\,(p)}) - y''_{n+1,\,(p)}]\}/h \tag{3.10}$$

This formula is obtained by eliminating the quantity y'_n from the corrector equations

$$y_{n+1} = \sum_{i=o}^{k-1} B_{oi} y_{n,i} + b_o \tfrac{1}{2} h^2\, [f_{tot}\,(y_{n+1,\,(p)}) - y''_{n+1,\,(p)}] \tag{3.11a}$$

and

$$h\, y'_{n+1} = \sum_{i=o}^{k-1} B_{1i}\, y_{n,i} + b_1 \tfrac{1}{2} h^2\, [f_{tot}(y_{n+1,\,(p)}) - y''_{n+1,\,(p)}] \tag{3.11b}$$

where $y_{n,i}$ denotes the i-th component of the vector $\underline{y}_n$. Eqs. (3.11) are obtained by inserting (3.2) in (3.4).

From test calculations it could be concluded that the scheme given above for MD calculations with constraints works well. The extra computer time that is needed compared to non-constraint MD is almost completely determined by SHAKE. It depends directly on the value chosen for tol, as will be seen in section 4.

From the theoretical considerations of this section the following conclusion can be drawn. The best (most accurate, stable, fast, etc.) presently available algorithm for MD calculations with constraints is a k-value predictor-corrector algorithm, written in the F-representation, with its predictor parameters determined by a Taylor expansion and with the corrector parameters derived from stability and accuracy requirements as proposed by Gear [10]. The constraints should be built in the algorithm following the scheme proposed in this section. The optimum k-value and step size h are then to be determined by the particular system under consideration. In the next section results of calculations with and without constraints will be compared.

4. *Macromolecular dynamics with and without constraints*

In order to test the conclusions drawn in the former sections the MD method was applied to a small (58 residue) protein molecule, viz. bovine pancreatic trypsin inhibitor (BPTI). Only one molecule in vacuum is considered, the solvent effects being neglected.

4.1 Parameters of the calculation

When performing a MD calculation, an interaction potential V (eq. (2.2)) has to be specified. We have used an empirical interaction function developed by Gelin and Karplus [15]. It contains contributions from bond stretching, bond angle bending, dihedral angle twisting, hydrogen bonds, non-bonded (Van der Waals) interactions and electrostatic (Coulomb) interactions. The concept of "extended atoms" is used, so hydrogen atoms are not explicitly considered, but incorporated in the heavy atoms to which they are bound. The number of extended atoms is equal to 458 (454 in BPTI + 4 hydrogen bonded water molecules), making the number of cartesian degrees of freedom equal to 1374. The number of possible bond length constraints is 468 and of possible angle constraints is 626.

The comparison of the various algorithms has to be done for MD runs that start from an equilibrium configuration of BPTI at a temperature of about 300 K. An appropriate starting configuration was obtained as follows.

- A run of 100 steps was performed, starting from the X-ray structure [16] with all velocities equal to zero. The 6-value predictor-corrector algorithm with coefficients of Gear [10] (see appendix) was used. The time step was taken equal to 4.889×10^{-14} sec. In this run the system reached a temperature of about 150 K.
- At this point all velocities were multiplied by a factor 1.74 and 100 more steps were taken, in which the system approached equilibrium at about 300 K.

The final configuration and velocities of this run were used as starting values for the runs that were used to compare the different algorithms. In these runs the analysis was started after 10 steps were performed, otherwise ambiquities in the starting procedures for the different algorithms might have disturbed the analysis.

4.2 Results

The different algorithms are compared for runs of 100 steps. Since a comparison of the solutions of the differential equation (2.1), viz. the trajectories, is too complex a task, we have limited ourselves to an analysis of the root mean square fluctuation of the total energy in relation to the root mean square fluctuation of the kinetic energy. In addition, a least squares fit of the total energy to a linear function of time is performed for each run in order to obtain the drift of the total energy per step. These quantities will depend on the algorithm that is considered, on the k-value, on the time step h and, in the case of constraint dynamics, on the accuracy tol to which the constraints are satisfied. Further valuable information about the algorithms is the required computer time per step and the number of vectors(r, v, etc.) that must be stored per step.

4.2.1. Non-constraint dynamics

In fig. 1 the results of 100 step MD calculations using the Verlet algorithm [9] and k-value predictor-corrector algorithms with Gear coefficients [10] are given as a function of the time step h. The Verlet algorithm requires the storage of three vectors and consumes 1.2 sec. CPU time on a CDC Cyber 74-16 per step. The Gear algorithms require the storage of k+1 vectors and the required CPU time amounts from 1.3 sec/step for k=4 to 1.5 sec/step for k=8.

For $h < 0.02$ the fluctuation in the total energy appears to be more than a factor 100 smaller than that in the kinetic energy. Beyond h=0.02 the algorithms become successively unstable. Below $h \approx 0.03$ the Gear algorithms yield a higher accuracy than the Verlet algorithm does, whereas above that value the Verlet algorithm does better. The latter effect can be understood from the fact that the Verlet algorithm makes use of a polynomial of degree 2, whereas for the Gear algorithms these values equal 3, 4, 5, 6, 7. The higher the degree, the smaller the time step at which the run will start to blow up.

When looking at the accuracy for small time steps as a function of k-value only, we note that the maximum accuracy is reached for the rather large k-value of k=7. This means that the forces are rather predictable, since in case of random forces one would gain no accuracy by going to a higher degree of the algorithm. From the fact that the bond stretching vibrations, which have by far the highest frequencies, are only weakly damped harmonic vibrations, it can be understood that the force is rather predictable.

The drift per step in the total energy shows roughly the same behaviour as a function of h as the fluctuation in the total energy does. Below $h \approx 0.02$ the Gear algorithms perform better, whereas above this value

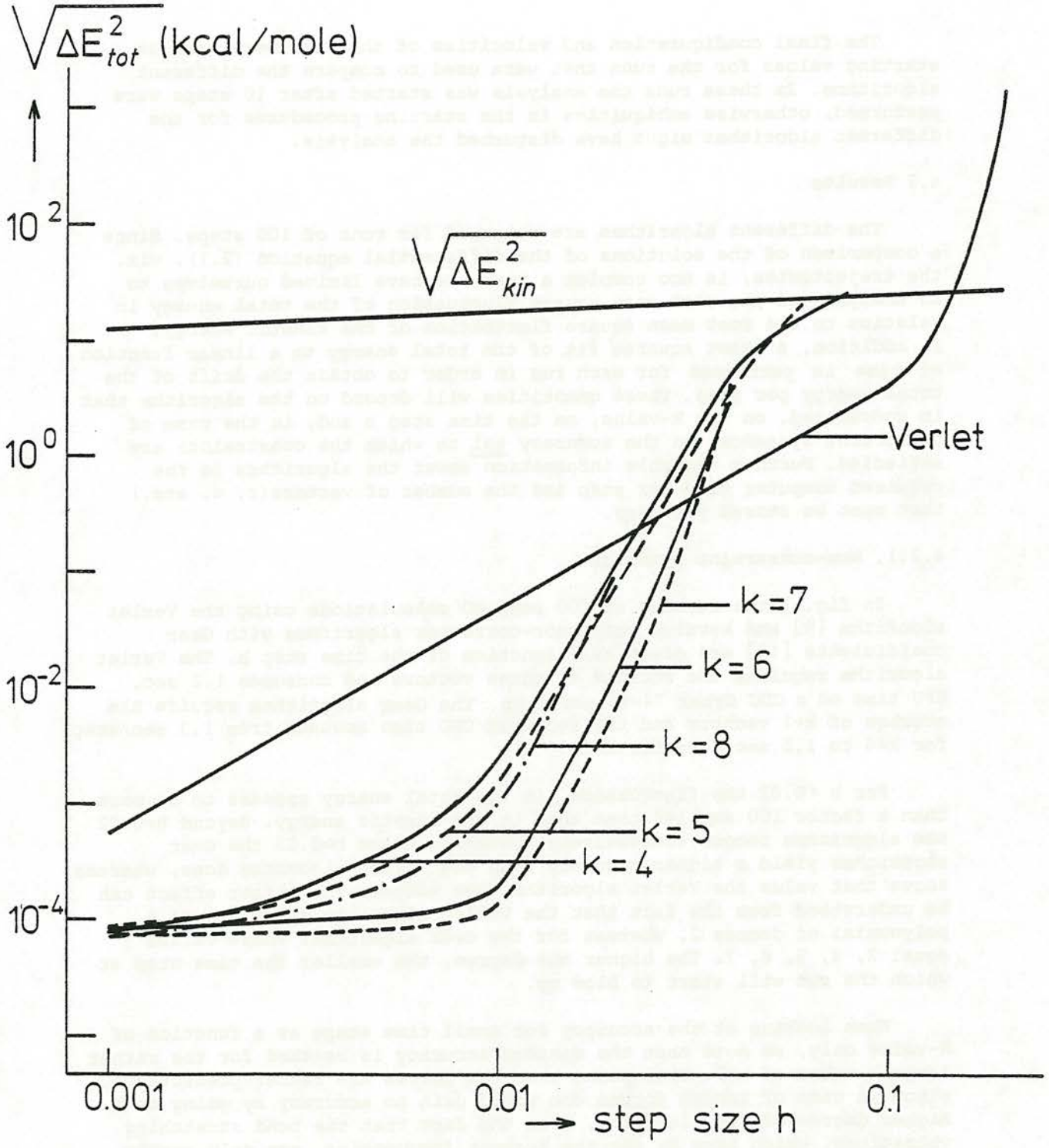


Fig. 1 Results of 100 step MD runs for BPTI (458 extended atoms) without constraints and using the Verlet algorithm [9] and a k-value predictor-corrector algorithm with coefficients proposed by Gear [10]. The root mean square fluctuation of the total energy $\sqrt{\Delta E^2_{tot}}$ (in kcal/mole) is plotted as a function of the time step size h (units: 4.889×10^{-14} sec). The symbol $\sqrt{\Delta E^2_{kin}}$ denotes the root mean square fluctuation of the kinetic energy, i.e. the average value for the different algorithms.

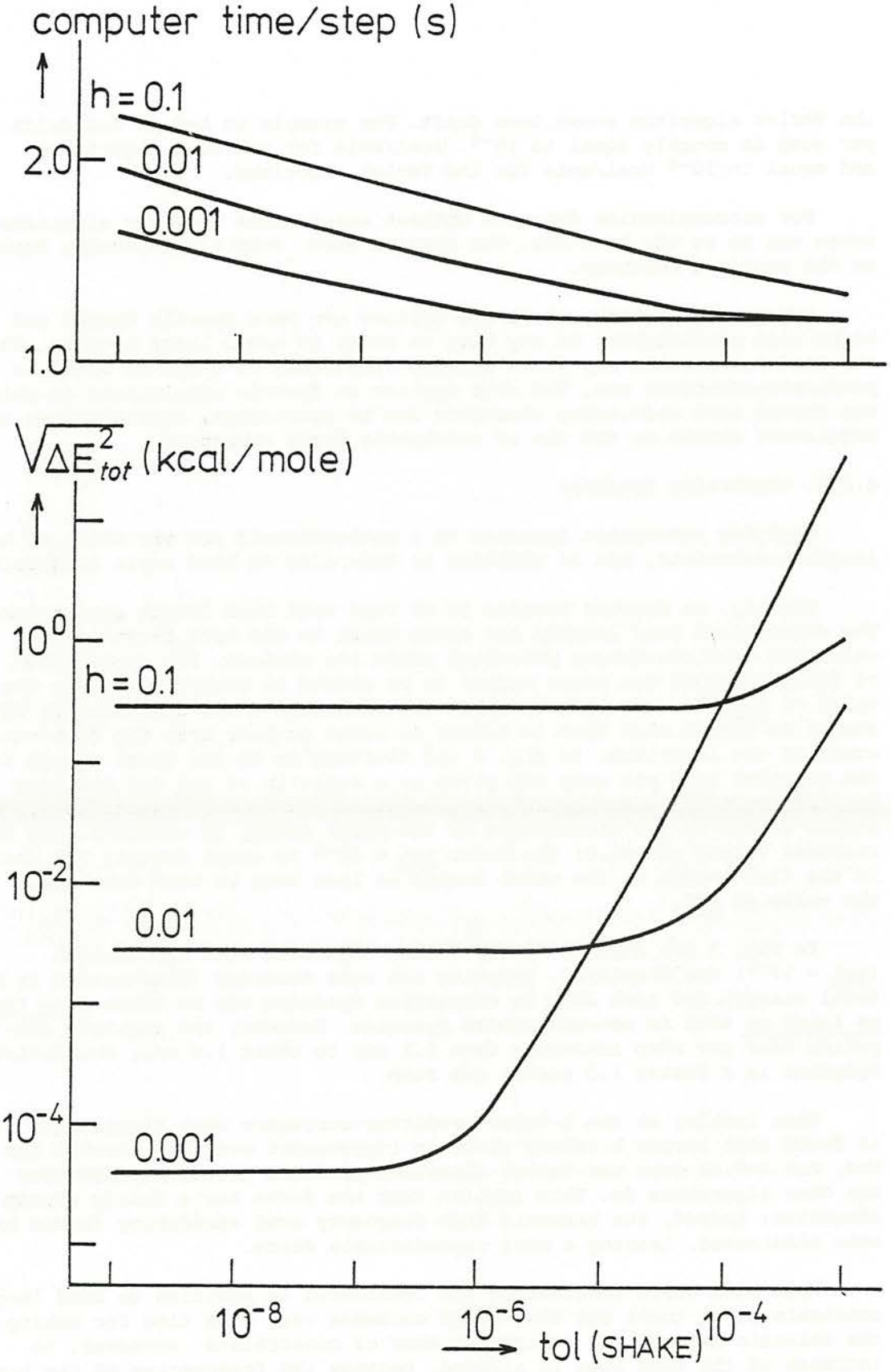


Fig. 2 Results of 100 step MD runs for BPTI with bond length constraints and using the Verlet algorithm [9]. The symbol tol denotes the accuracy to which the constraints are satisfied. The root mean square fluctuation of the total energy $\sqrt{\Delta E^2_{tot}}$ (in kcal/mole) and the required CPU time per MD step (in sec on a CDC Cyber 74-16) are plotted as a function of tol and the time step size h (units: 4.889×10^{-14} sec).

the Verlet algorithm shows less drift. For example at h=0.01 the drift per step is roughly equal to 10^{-6} kcal/mole for the Gear algorithms and equal to 10^{-4} kcal/mole for the Verlet algorithm.

For macromolecular dynamics without constraints the Gear algorithm turns out to be the best one. The optimum time step h and k-value depend on the required accuracy.

For simple liquids, where the motions are more heavily damped and hence less predictable, it may have no sense to use a large k-value. Then the Verlet algorithm may yield results comparable to those of the Gear predictor-corrector one. The same applies to dynamic simulations in which the forces have more noisy character due to truncation, approximation or tabulation errors or the use of stochastic force components.

4.2.2. Constraint dynamics

Applying constraint dynamics to a macromolecule one may think of bond length constraints, but in addition to this also of bond angle constraints.

Firstly, we discuss results of MD runs with bond length constraints. The constrained bond lengths are taken equal to the bond lengths for which the bond stretching potential finds its minimum. The application of SHAKE requires one extra vector to be stored in computer memory. The value of tol, the accuracy by which the constraints are satisfied by SHAKE, should be chosen such that it causes an error smaller than the truncation error of the algorithm. In fig. 2 the fluctuation in the total energy and the required time per step are given as a function of tol and the time step h. The smaller h, the smaller the value of tol beyond which no significant change in the fluctuation of the total energy is observed. For the relevant values of h>0.01 the value tol = 10^{-6} is small enough; the change in the fluctuation of the total energy is less than 1% when decreasing the value of tol.

In fig. 3 the results of the Verlet-algorithm with constraints (tol = 10^{-6}) are displayed. Reaching the same accuracy (fluctuation in the total energy), the time step in constraint dynamics can be taken four times as large as that in non-constraint dynamics. However, the required computing time per step increases from 1.2 sec to about 1.6 sec; constraint dynamics is a factor 1.3 slower per step.

When looking at the k-value predictor-corrector Gear algorithms, it is found that larger k-values yield no improvement over the results for k=4. For h>0.02 even the Verlet algorithm produces better results than the Gear algorithms do. This implies that the force has a fairly strong random character; indeed, the harmonic high frequency bond stretching forces have been eliminated, leaving a more unpredictable force.

When bond angle constraints are considered in addition to bond length constraints, it turns out that SHAKE consumes very much time for making the molecule satisfy the enlarged number of constraints. Moreover, no increase of the time step is allowed, because the frequencies of the bond angle vibrations are not well separated from those of the other vibrations of the macromolecule [8].

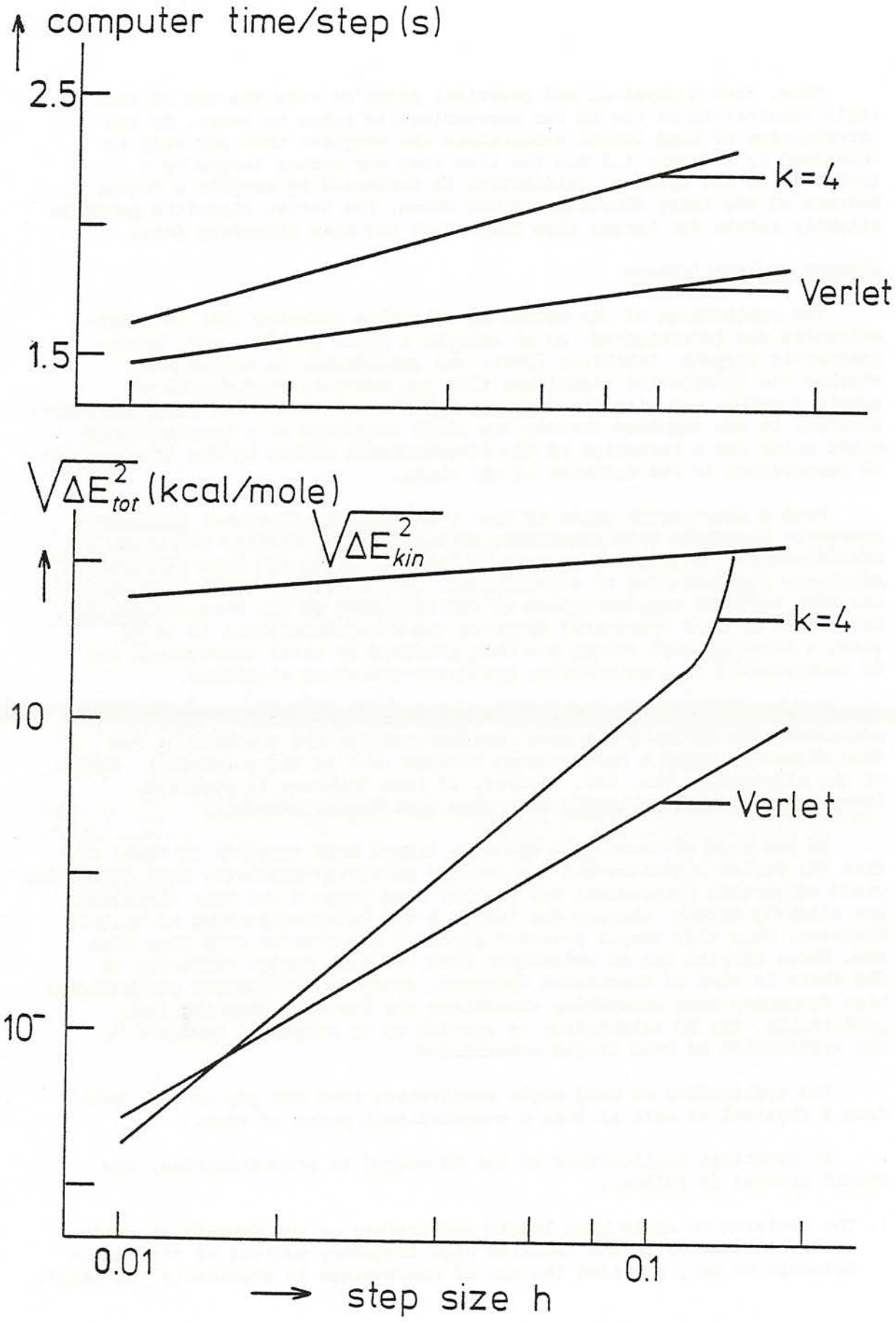


Fig. 3 Results of 100 step MD runs for BPTI with bond length constraints (tol = 10^{-6}). For further explanation see text below figs. 1 and 2.

Thus, from a physical and practical point of view the use of bond angle constraints in the MD for macromolecules makes no sense. By the introduction of bond length constraints the computer time per step is increased by a factor 1.3 but the time step may become longer by a factor 4. So the speed of calculation is increased by roughly a factor 3. Because of the noisy character of the force, the Verlet algorithm performs slightly better for larger time steps than the Gear algorithm does.

5. *Summary and conclusions*

The application of the method of molecular dynamics (MD) to macromolecules was investigated. As an example a small protein, viz. bovine pancreatic trypsin inhibitor (BPTI) was considered. It was checked whether the integration algorithms that are commonly used for MD of simple liquids, are also the most appropriate ones for MD of macromolecules. Besides, it was examined whether the chain structure of a macromolecule might allow for a reduction of the computational effort by the introduction of constraints in the dynamics of the chain.

From a theoretical point of view a multi-value (k-value) predictor-corrector algorithm with parameters obtained from stability and accuracy considerations as proposed by Gear [10] appears to be the best presently available algorithm for MD calculations. In case of non-constraint dynamics the most suitable representation of the algorithm is the N-representation, whereas in case of constraint dynamics the F-representation is to be used. A calculational scheme has been proposed by which constraints can be incorporated in a multi-value predictor-corrector algorithm.

From the simulation of BPTI it can be concluded that in case of non-constraint dynamics the most accurate results are produced by the Gear algorithm using a rather large k-value (k-1 is the polynomial degree of the algorithm), viz. $k \approx 7$. However, if less accuracy is required, lower k-values allow a larger time step than higher k-values.

In the case of constraint dynamics (fixed bond lengths) it turns out that the Verlet algorithm and the k-value predictor-corrector Gear algorithms yield comparable accuracies; for smaller time steps h the Gear algorithms are slightly better, whereas for larger h the Verlet-algorithm is better. Moreover, Gear with larger k-values gives no improvement over Gear with k=4. These results can be understood from the more random character of the force in case of constraint dynamics, since by eliminating the harmonic high frequency bond stretching vibrations the force is becoming less predictable. The MD calculation is speeded up by roughly a factor 3 by the application of bond length constraints.

The application of bond angle constraints does not pay at all, both from a physical as well as from a computational point of view.

In practical applications of the MD method to macromolecules, one should proceed as follows.

1. The decision to apply bond length constraints or not depends on whether one is interested in the detailed high-frequency motions of the macromolecule or not, provided the use of constraints is physically justified.

2. In MD calculations without constraints the Gear k-value predictor-corrector algorithm is to be used. When constraints are applied, the choice between the Gear and the Verlet algorithm depends on the time step to be chosen, so on the desired accuracy.
3. Choose an accuracy, viz. an upper limit for the root mean square fluctuation of the total energy.
4. In case of constraint dynamics, find from test calculations the largest value of tol that is compatible with the chosen accuracy of the total energy.
5. Determine the optimum size of the time step h and the optimum k-value from test calculations.
6. With the obtained tol, h and k-values MD production runs, suitable for analysis, can be done.

Commented copies of the FORTRAN MD and SHAKE subroutines are available on request.

Acknowledgements

We thank Prof. Dr. M. Karplus, Dr. B.R. Gelin and Dr. J.A. McCammon for providing us with force and energy subroutines needed for the application of the MD algorithms to BPTI. We thank Dr. C. Moser for his hospitality at the CECAM workshop. One of us (W.F. v. G.) is grateful to the Stichting voor Fundamenteel Onderzoek der Materie (F.O.M.) for the opportunity of attending the workshop. This work has been financially supported by the Nederlandse Organisatie voor Zuivere Wetenschappelijk Onderzoek (Z.W.O.).

Appendix

A.1. The k-value predictor-corrector algorithm in the N-representation

In the N-representation the predictor matrix $\underline{A}$ is equal to the Pascal triangle ($k \leq 8$):

$$\underline{A} = \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ & & 1 & 3 & 6 & 10 & 15 & 21 \\ & & & 1 & 4 & 10 & 20 & 35 \\ & & & & 1 & 5 & 15 & 35 \\ & & & & & 1 & 6 & 21 \\ & 0 & & & & & 1 & 7 \\ & & & & & & & 1 \end{pmatrix}}_{k} \qquad \text{(A.1)}$$

The corrector column vector $\underline{a}$ that assures optimum stability and accuracy properties for the solution of a second-order differential equation (2.3) is given in table 1 in the N-representation. The values are taken from Gear [10] p 154.

Table 1

Optimum corrector coefficients in the N-representation

k	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7
4	1/6	5/6	1	1/3				
5	19/120	3/4	1	1/2	1/12			
6	3/20	251/360	1	11/18	1/6	1/60		
7	863/6048	665/1008	1	25/36	35/144	1/24	1/360	
8	275/2016	19087/30240	1	137/180	5/16	17/240	1/120	1/2520

Values of the corrector column vector $\underline{a}$ for the Gear k-value predictor-corrector algorithm for second-order differential equations given in the N-representation. The are taken from Gear [10] p. 154.

A.2. The k-value predictor-corrector algorithm in the F-representation

The predictor matrix $\underline{B}$ and the corrector vector $\underline{b}$ in the F-representation can be obtained from the matrix $\underline{A}$ and the vector $\underline{a}$ by performing a transformation T:

$$\underline{B} = \underline{T}\underline{A}\underline{T}^{-1} \tag{A.2.a}$$

$$\underline{b} = \underline{T}\underline{a} \tag{A.2.b}$$

The transformation matrix $\underline{T}$ can be easily derived from the relation

$$\underline{y}_n(F) = \underline{T}\,\underline{y}_n(N) \tag{A.3}$$

One finds ($k \leq 8$):

$$\underline{T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -3 & 6 & -10 & 15 & -21 \\ 0 & 0 & 1 & -6 & 24 & -80 & 240 & -672 \\ 0 & 0 & 1 & -9 & 54 & -270 & 1215 & -5103 \\ 0 & 0 & 1 & -12 & 96 & -640 & 3840 & -21504 \\ 0 & 0 & 1 & -15 & 150 & -1250 & 9375 & -65625 \end{pmatrix} \tag{A.4}$$

When inverting $\underline{T}$ for different values of k, one finds for k = 4

$$\underline{T}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/3 & -1/3 \end{pmatrix} \tag{A.5.a}$$

for k = 5

$$\underline{T}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & -2/3 & 1/6 \\ 0 & 0 & 1/12 & -1/6 & 1/12 \end{pmatrix} \tag{A.5.b}$$

for k = 6

$$\underline{T}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 11/18 & -1 & 1/2 & -1/9 \\ 0 & 0 & 1/6 & -5/12 & 1/3 & -1/12 \\ 0 & 0 & 1/60 & -1/20 & 1/20 & -1/60 \end{pmatrix} \qquad \text{(A.5.c)}$$

for k = 7

$$\underline{T}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 25/36 & -4/3 & 1 & -4/9 & 1/12 \\ 0 & 0 & 35/144 & -13/18 & 19/24 & -7/18 & 11/144 \\ 0 & 0 & 1/24 & -3/20 & 1/5 & -7/60 & 1/40 \\ 0 & 0 & 1/360 & -1/90 & 1/60 & -1/90 & 1/360 \end{pmatrix} \qquad \text{(A.5.d)}$$

and for k = 8

$$\underline{T}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 137/180 & -5/3 & 5/3 & -10/9 & 5/12 & -1/15 \\ 0 & 0 & 5/16 & -77/72 & 107/72 & -13/12 & 61/144 & -5/72 \\ 0 & 0 & 17/240 & -71/240 & 59/120 & -49/120 & 41/240 & -7/240 \\ 0 & 0 & 1/120 & -7/180 & 13/180 & -1/15 & 11/360 & -1/180 \\ 0 & 0 & 1/2520 & -1/504 & 1/252 & -1/252 & 1/504 & -1/2520 \end{pmatrix} \qquad \text{(A.5.e)}$$

Using eq. (A.2.a) one finds for $\underline{B}$

for k = 4

$$\underline{B} = \begin{pmatrix} 1 & 1 & 4/3 & -1/3 \\ 0 & 1 & 3 & -1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad \text{(A.6.a)}$$

for k = 5

$$\underline{B} = \begin{pmatrix} 1 & 1 & 19/12 & -5/6 & 1/4 \\ 0 & 1 & 23/6 & -8/3 & 5/6 \\ 0 & 0 & 3 & -3 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad \text{(A.6.b)}$$

for k = 6

$$\underline{B} = \begin{pmatrix} 1 & 1 & 323/180 & -22/15 & 53/60 & -19/90 \\ 0 & 1 & 55/12 & -59/12 & 37/12 & -3/4 \\ 0 & 0 & 4 & -6 & 4 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad \text{(A.6.c)}$$

for k = 7

$$\underline{B} = \begin{pmatrix} 1 & 1 & 1427/720 & -133/60 & 241/120 & -173/180 & 3/16 \\ 0 & 1 & 1901/360 & -1387/180 & 109/15 & -637/180 & 251/360 \\ 0 & 0 & 5 & -10 & 10 & -5 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad \text{(A.6.d)}$$

and for k = 8

$$\underline{B} = \begin{pmatrix} 1 & 1 & 2713/1260 & -15487/5040 & 1172/315 & -6737/2520 & 263/252 & -863/5040 \\ 0 & 1 & 4277/720 & -2641/240 & 4991/360 & -3649/360 & 959/240 & -95/144 \\ 0 & 0 & 6 & -15 & 20 & -15 & 6 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad \text{(A.6.e)}$$

From eq. (A.2.b) one finds for $\underline{b}$

$$\left.\begin{aligned} b_i &= a_i && i = 0,1,2 \\ b_i &= o && i > 2 \end{aligned}\right\} \qquad \text{(A.7)}$$

It should be noticed that the last (k-3) rows of the matrix $\underline{B}$ contain zero's except for one 1 and that the last (k-3) coefficients of $\underline{b}$ are also equal to zero. As has been said earlier the last (k-2) values of $\underline{b}$ result from stability requirements [10,11]. But, fortunately they happen to be zero (except b_2=1), since otherwise the values of y" calculated at previous steps would change at later steps. This would make the incorporation of SHAKE in the algorithm impossible.

References

[1] B.J. Alder and T.E. Wainwright, J. Chem. Phys. 31 (1959) 459
[2] A. Rahman , Phys. Rev. 136 (1964) A 405
[3] A. Paskin and A. Rahman , Phys. Rev. Lett. 16 (1966) 300
[4] L.V. Woodcock, Chem. Phys. Lett. 10 (1971) 257
[5] A. Rahman and F.H. Stillinger, J. Chem. Phys. 55 (1971) 3336
[6] J. Barojas, D. Levesque and B. Quentrec, Phys. Rev. A 7 (1973) 1092
[7] J.P. Ryckaert and A. Bellemans, Chem. Phys. Lett. 30 (1975) 123
[8] J.A. McCammon, Report of CECAM workshop on Models for Protein Dynamics (CECAM, Orsay, 1976)
[9] L. Verlet, Phys. Rev. 159 (1967) 98
[10] C.W. Gear, Numerical Initial Value Problems in Ordinary Differential Equations (Prentice-Hall, Englewood Cliffs, N.J. 1971)
[11] C.W. Gear, The Numerical Integration of Ordinary Differential Equations of Various Orders (Report ANL 7126, Argonne National Laboratory, Argonne, Ill. 1966)
C.W. Gear, Math. Comp. 21 (1967) 146
[12] D. Beeman, J. Comp. Phys. 20 (1976) 130
[13] A. Nordsieck, Math. Comp. 16 (1962) 22
[14] J.P. Ryckaert, G. Ciccotti and H.J.C. Berendsen, to be published
[15] B.R. Gelin and M. Karplus, Proc. Natl. Acad. Sci. USA 72 (1975) 2002
[16] J.O. Deisenhofer and W. Steigemann, Acta Cryst. B 31 (1975) 238

II.2

MOLECULAR DYNAMICS OF A DIPEPTIDE IN WATER

P.J.Rossky[1]
A.Rahman[2]

[1]Harvard University, Department of Chemistry, Cambridge, Ma.02138 (USA)
[2]Argonne National Laboratory, Solid State Physics Division, Argonne, Illinois 60439 (USA).

I. Introduction

Although it is widely accepted that solvent has an important role in biological systems, relatively little is known, at the molecular level, about the structural and dynamical properties of water when interacting with biochemical molecules. This is a result, in large part, of the relatively subtle effects which are produced by most types of interactions. For example, experimental measurements of rotational diffusion and reorientation times for water near protein surfaces have been able to resolve only three different classes of water molecules.[1] The most rapidly reorienting, which includes the bulk, has a characteristic rotational diffusion time of about 10^{-11} sec. The next most rapid exhibits a rotational reorientation time (τ_r) of about 10^{-9} sec and has been tentatively identified as involving the water strongly bound to ionic groups. The third exhibits a τ_r of about 10^{-6} sec. and these are considered essentially irrotationally bound. It is expected that the fastest time range will be populated by the numerous water molecules which interact relatively more weakly with the protein. These include molecules which form hydrogen bonds to the peptide backbone and those which are influenced by the presence of non-polar groups, i.e., those apparently contributing to the "hydrophobic" effect.[1,2] Similarly, in X-ray studies of protein structure, only those sites which bind water fairly strongly (e.g., through ionic or polar hydrogen bonds) show a density peak sufficiently distinct to be identified.

The properties of the solvated biochemical molecule itself are also of interest. Due to the strong infrared absorption of water, I.R. studies are extremely difficult to perform. Some structural work has been done using other methods,[3] but interpretation of the results is complicated and open to question.

II. System and Method

For these reasons, a molecular dynamics study of a dipeptide in water solution was begun at the workshop. The system studied consists of an alanine dipeptide, Figure 1, immersed in a "box" of 195 water molecules. Each of the 22 dipeptide atoms is assigned Lennard-Jones 6-12 parameters and a point charge.

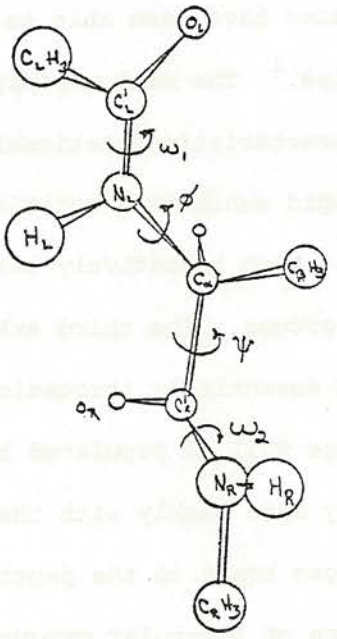


Figure 1: The alanine dipeptide-atom labels and dihedral angles are indicated. The hydrogens of the methyl groups are not shown, for clarity, but are included in the calculation.

The internal potential of the dipeptide consists of bond stretching, bond angle bending, and torsional potentials, parameterized in accord with accepted values,[4] as well as non-bonded interactions arising from the charge and Lennard-Jones terms. A "10-12" term replaces the 6-12 term between atoms O_L and H_R (see Figure 1) to provide an improved dipeptide internal

hydrogen bond.[5] A Ramachandron plot of the resulting potential energies is shown in Figure 2.

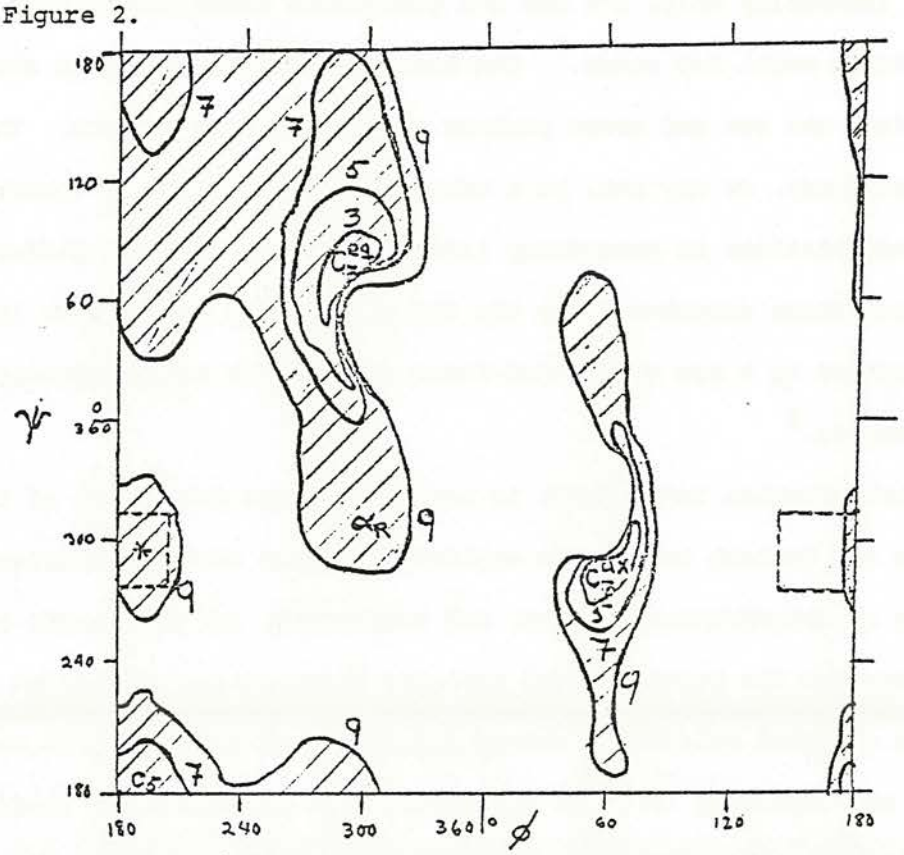


Figure 2: ϕ, ψ potential energy map for the alanine dipeptide, computed using rigid rotation about the dihedral angles. The contour energies are indicated relative to zero at the C_7^{eq} conformation. The relative energies of the other important local minima indicated are approximately: C_7^{ax}, +2kcal; C_5, +6kcal; α_R, +8kcal; extended (*), +8kcal.

The water is represented by a slightly modified version of the ST2 model of Rahman and Stillinger.[6] In the ST2 model, each water consists of a single Van der Waals sphere, within which are located four point charges directed toward the vertices of a tetrahedron and representing the hydrogens and lone pair. The change made is to allow flexibility in each water molecule by

introducing, internally only, O-H and H-H potentials taken essentially from the central force model for water.[7] The lone pair charge locations are constructed from the sum and cross product of the O-H bond vectors. This flexible version may, or may not, be a more "realistic" model of water; preliminary calculations on pure water indicate that the results differ negligibly from those computed using the ST2 model. The water-water interaction is computed as a sum of Lennard-Jones and Coulomb terms, exactly as in the ST2 studies.[6]

The water-dipeptide interaction is computed as the direct sum of the Lennard-Jones and Coulomb terms. No explicit hydrogen bonding function is added. Using an appropriately chosen, and reasonable, set of charges and 6-12 parameters for the peptide atoms produces interactions sufficient to reproduce the expected relatively strong dependence of the energy on radial distance (r) and linearity (ϕ), and relatively flat dependence on bending (θ) (see Figure 3). The resulting optimal bond energies and geometries for one ST2 water with alanine dipeptide are given in Table I.

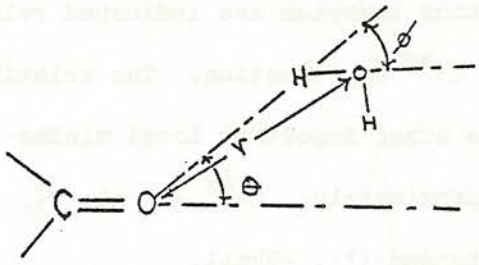


Figure 3: Geometrical parameters important to hydrogen bond description.

Table I: Optimal Hydrogen Bonds

Type	r^b(A°)	θ^b	E(kcal/mole)
O_wq--HN	2.9	10°	-6.1
NH--O=C	2.8	15°	-6.6
O_wH_w--qO_w	2.85	0°	-6.8
O_wH_w--O=C	2.6	55°	-7.4

a) O_w, H_w and q are the water oxygen, hydrogen and charge, respectively.

b) As defined in Figure 3; r is always the heavy atom distance; $\phi = 0$ (since $\phi = 0$, θ is just the acute angle formed by drawing one line through each of the two atom pairs given).

These results are consistent with the (somewhat diffuse) quantum mechanical and experimental results available.[5,8] In particular, all the values are fairly close to one another, and, in order of decreasing strength, OH - O = C > Oq - HO > NH - O = C > Oq - HN. (OH and Oq refer to the water molecule.)

The dipeptide was placed in a "box" of 216 water molecules whose configuration was the result of previous equilibrium dynamics on pure water. The water molecules with close contacts to the dipeptide were removed and the (cubic) box edge length was adjusted from 18.62 to 18.2194Å yielding a system of 195 water molecules and one alanine dipeptide at a density of 1.004 g/cc, in accord with experiment.[9]

In order to simulate an infinite system, periodic boundary conditions and the method of the minimum image[10] were applied during the calculation. The classical equations of motion for the specified system were then numerically integrated to simulate the dynamics.

As is common in such problems a cutoff distance for the intermolecular potential was introduced. The choice initially made was rather short, approximately 4.5Å, with the intention of reducing the effects of the relatively high dipeptide concentration (∿.25M) on the water behavior. Although this range excludes direct interaction with second nearest neighbor water molecules in most cases, the indirect effects transferred by the preferred orientation of nearest neighbors is present. Thus one expects that the most important forces acting on each water are still included. The introduction of this cutoff, however, causes a large, orientation dependent discontinuity in the potential and the system is found to heat up relatively rapidly during the dynamics run. A periodic readjustment of the velocities is needed to maintain a reasonable temperature. (This heating is a distinct effect from that related to the choice of too large a stepsize in integration.) Additional computations have been carried out with a more suitable potential cutoff of 8Å. This range is still short enough to exclude direct dipeptide-dipeptide interaction (resulting from the periodicity), while the water-water interaction at the cutoff distance (8Å) is, on the average, essentially zero. In computing water-dipeptide interactions, the dipolar atom pairs (C'-O, N-H) are always considered together with respect to the potential distance cutoff, so that either both atoms, or neither, in each pair interact with any given water molecule.

A Gear algorithm[12] has been used for the integrations, with a stepsize of 3.67×10^{-16} sec. One step takes 4.5 sec. on an IBM model 360/91

(approximately 5.2 sec. on the 370/168) using a potential range of 8Å, and approximately 1.3 sec. with a range of 4.5Å. This large difference in execution time with a change in range can be used to advantage in carrying out the equilibration of a "new" system, i.e., one in which there are large deviations from an equilibrium configuration. Our calculations indicate that after initial equilibration at a relatively short range, readjustments in the system, due to increasing the range from 4.5 to 8Å, takes only about 500 to 1,000 steps.

The system was initially equilibrated for approximately 6,000 steps. Two rather short additional runs have thus far been carried out; one of 0.67 psec with a 4.5Å potential range and a second of 0.85 psec with an 8Å potential range. The latter run has been used to compute all quantitative properties discussed below. The mean temperature of the system during these runs was rather high. Apparently due to the short length of these runs, various temperature estimates are not identical; the water temperature based on an average of rotational and translational motion (385°K and 354°K, respectively) was 370°K during the longer run. The dipeptide temperature was approximately 320°K.

III. Qualitative Solution Structure

The dipeptide, initially in an α helical conformation, passed during the time of equilibration through the equatorial C_7 (C_7^{eq}) region and into the region of a local minimum near $\phi = 180°$, $\psi = 300°$ (see Figure 2). The origin of this latter local minimum is a valley in the repulsive potentials: Coulombic repulsion, primarily, between the amide hydrogens and steric repulsion with the methyl side chain for the peptide hydrogens and oxygen.

In glycine this minimum is essentially missing. The bottom of this local minimum lies (in vacuum!) approximately 8 kcal above the absolute minimum, the C_7^{eq} conformation. The transition into this minimum initially was most likely an artifact of the initial water configuration. However, one can estimate the least maximum in the dipeptide (ϕ, ψ) potential energy between the C_7^{eq} minimum and this extended minimum by performing a series of conformational energy minimizations with constrained values of ϕ and ψ. Along the path observed in the dynamics, the maximum barrier occurs near (ϕ, ψ) = (210°, 30°) with a height of approximately 11 kcal above the C_7 minimum. An alternate route, approximately 1 kcal lower, corresponds to ψ passing through ψ = 180°, into the C_5 region near (180°, 180°), and having its maximum near (180°, 240°).

It is interesting to note that recent calculations of peptide atom solvent accessibility indicate that this extended conformation (ϕ = 180°, ψ = 300°) might have a relatively high (favorable) hydration energy.[11]

Figure 4 shows a history of the dihedral angles ϕ, ψ and ω_2 (see Figure 1) in solution, and also in the absence of the water for comparison. Note that they appear, qualitatively, to include similar high and low frequency components in either environment. The dashed bordered regions in Figure 2 roughly indicate the (ϕ, ψ) region traversed in solution.

There are, at any time, several water molecules bonded to the dipeptide. There is a definitional problem in this respect, since there are a wide range of possible interaction energies of a water molecule with the dipeptide, depending on the precise distance and angular geometry of association. As a convenient operational definition, a cutoff of -2.5 kcal for the total interaction of a water with the dipeptide will be used to define the difference between bonded and non-bonded molecules.

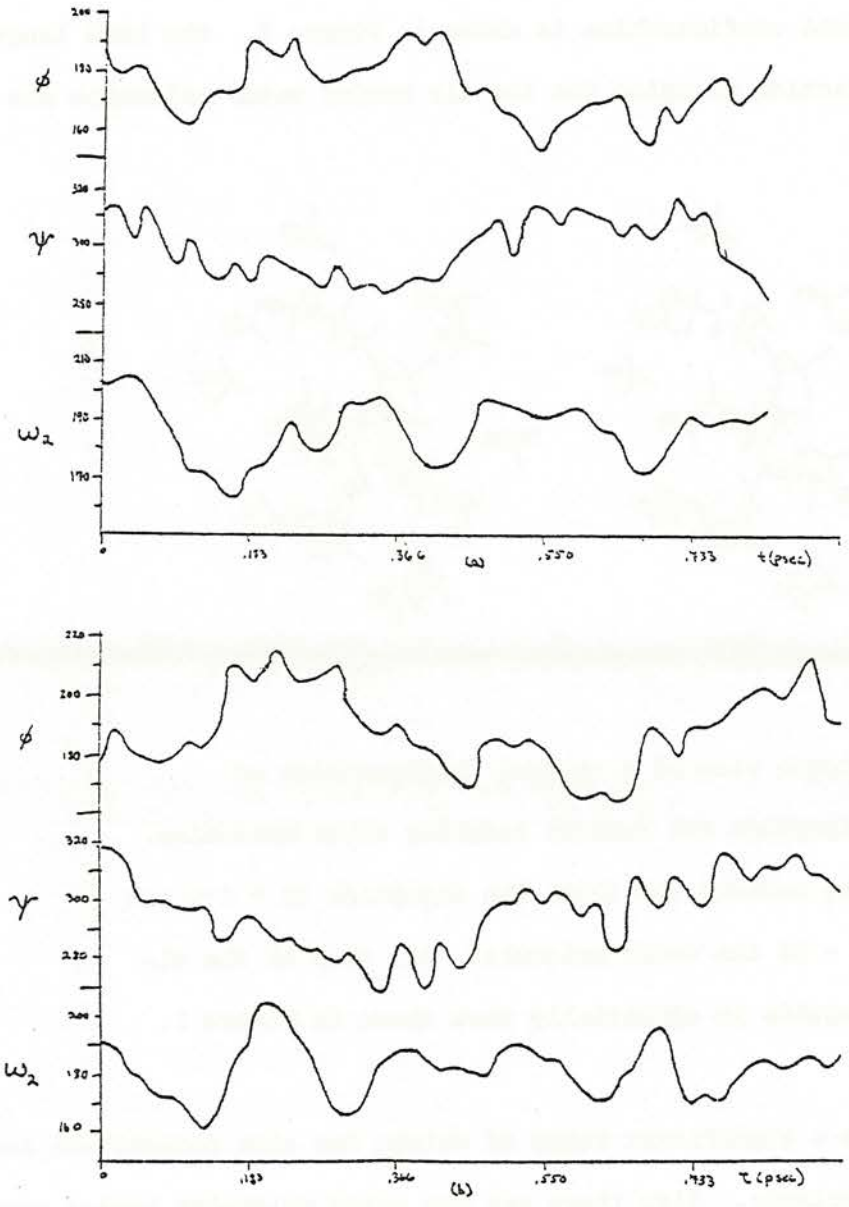


Figure 4: History of the dihedral angles ϕ, ψ and ω_2 in the dipeptide. That shown in (a) is in solution, that in (b) in vacuum.

This value effectively includes the few most strongly interacting pairs, but not the many weakly interacting molecules (see Figure 6). A stereo picture of part of a typical configuration is shown in Figure 5. The bond lengths, angles, and interaction energies for the six bonded water molecules are given in Table IIa.

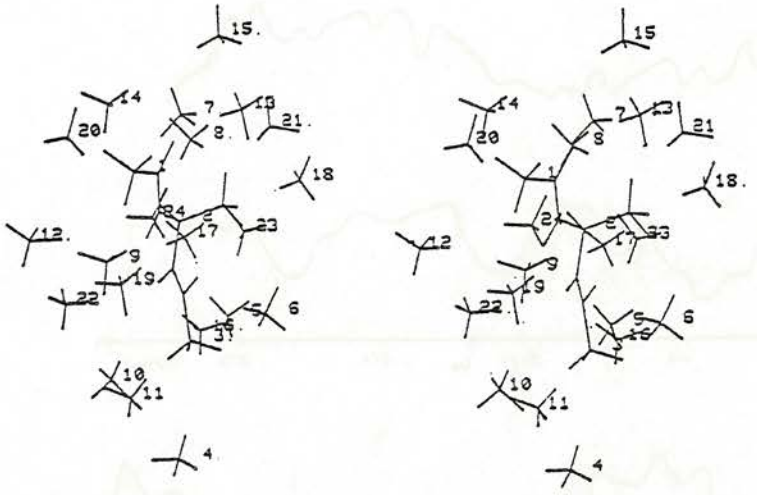


Figure 5: Stereo view of a typical configuration of dipeptide and nearest neighbor water molecules. The numbers 1-3 label the dipeptide (N → C') and 4 - 24 the water molecules. The view of the dipeptide is essentially that shown in Figure 1.

Note that there is a significant range of values for each geometrical and energetic characteristic. Also there are two water molecules bonded essentially perpendicularly to the carbonyl groups. These may be thought of as corresponding to bonding to the π orbital of the C = 0 group and are physically reasonable.

Table IIa: Water-Dipeptide Hydrogen Bonding[c]

H_2O[a]	r_{D-O_W} to	D	E_{D-H_2O}	θ[b]
6	3.19/2.60	N_R/H_R	-2.81	48°
13	3.50/2.71	C_L'/O_L	-4.05	37°
16	3.80/2.90	C_R'/O_R	-3.59	55°
19	3.03/2.03	N_L/H_L	-6.33	32°
20	3.38/3.52	C_L'/O_L	-2.89	110°
22	3.73/3.09	C_R'/O_R	-4.26	92°

a) with reference to Figure 5

b) angle between the vectors $\overrightarrow{HO}_w(\overrightarrow{qO}_w)$ and $\overrightarrow{CO}(\overrightarrow{NH})$, such that a linear bond: ($\theta = 0$) corresponds to exactly parallel bond vectors. When $\phi = 0$ (Figure 3), the angle above reduces to precisely the angle θ shown in Figure 3.

c) distances in Å, energies in kcal/mole.

Table IIb: Water-Water Hydrogen Bonding in the First Shell[b]

H_2O[a]	H_2O[a]	r_{OO}	E pair
5	6	2.96	-2.60
5	23	3.19	-4.23
8	14	2.87	-4.01
8	24	2.78	-6.51
9	17	2.67	-3.67
10	19	2.99	-5.49
13	18	3.17	-2.57

a) with reference to Figure 5

b) distance in Å, energy in kcal/mole.

In this particular configuration there are 21 water molecules with oxygen atoms within 4Å of any heavy atom of the dipeptide. This value is typical of those observed throughout the run. The bonding among these waters is given in Table IIb.

It is enlightening to compare this to another typical configuration removed in time from this one (by .64 psec). There are, here, 23 water molecules within the 4Å shell. Four are bonded to the dipeptide, while there are 16 water-water hydrogen bonds (compared to 8, above) among these 23. Only one of these 16 pairs also occurs among the 8 in the earlier configuration. The general rearrangement of bonded pairs on this time scale is consistent with experimental evidence that the "lifetime" of a hydrogen bond in water is about .5 to 1 psec.[13,14] The occurence of this larger number of intra-shell bonds does not imply that the water in the first shell is more significantly bonded in the latter configuration than in the former one. If one includes bonding to molecules in the second shell, as well as to the dipeptide, the molecules in the first shell participate on the average, in approximately the same number of hydrogen bonds in either configuration.

Figure 6 shows accumulated distributions of pair interaction potential energies among the water and dipeptide molecules. In Figure 6a, the solid line shows the distribution for water-water pairs, including all pairs in the entire sample. This result is consistent with that found previously for pure water.[6] The dotted line shows a distribution obtained by including only those distinct pairs which include at least one molecule in the 4Å shell around the dipeptide. This result would, in principle, reflect both the average strength of individual water-water interactions at, for example, a given distance apart, and, in addition, any changes in the average spatial distribution of other water molecules around each molecule in this shell.

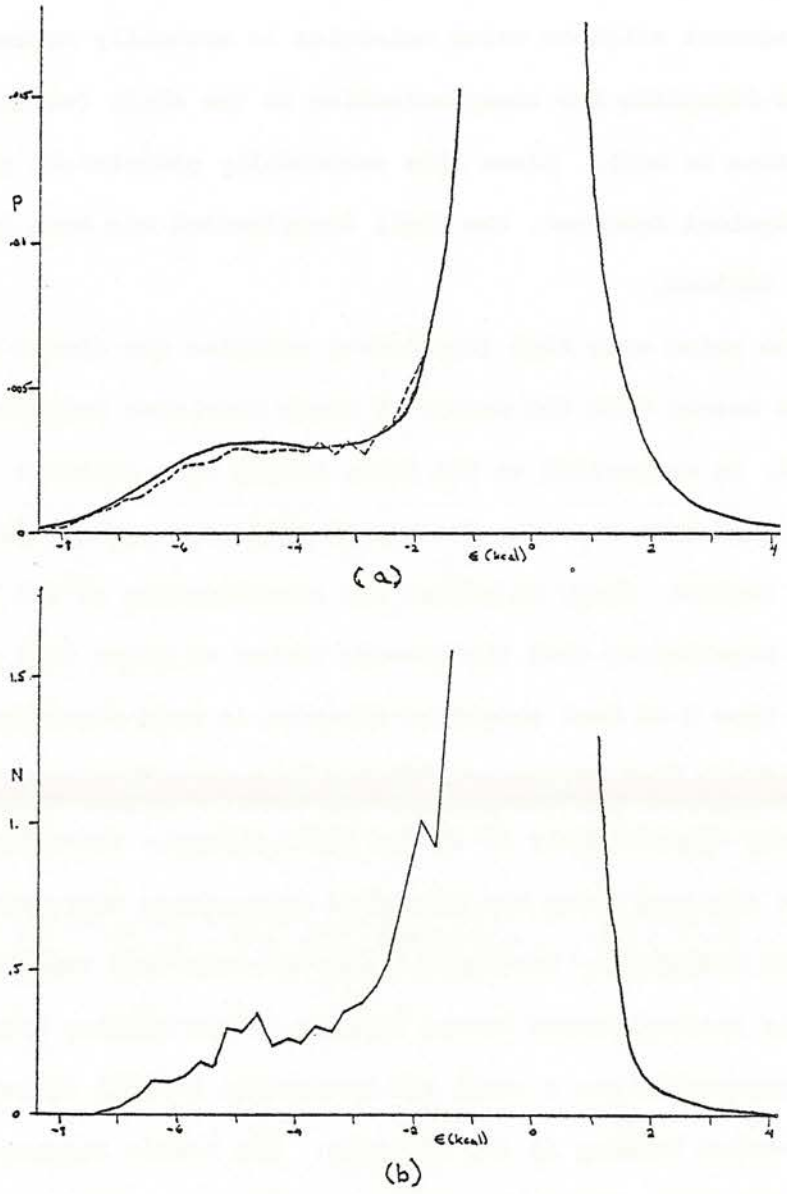


Figure 6: Pair energy distributions

a) water-water pairs; (——, all water pairs; ---, at least one of each pair in "hydration shell). The solid curve is normalized to unity.

b) water-dipeptide pairs; the normalization is to the total number of water molecules, 195.

Now, the number of nearest neighbor water molecules is naturally reduced by the proximity of the dipeptide for those molecules in the shell (while the number at long distance is not). Since this essentially geometrical effect is not of primary physical interest, the shell distribution has been approximately corrected as follows.

We know that the pairs with high (repulsive) energies are always nearest neighbors. We first assume that the number of these repulsive neighbors is reduced in the shell, in comparison to the bulk, simply by a constant fraction. Second, we assume that the same fractional reduction applies for the bonded (attractive) region. Thus, we adjust the normalization of the shell distribution by the requirement that the average number of pairs with potential energy greater than 2.25 kcal should be the same in both distributions. We find that this differs from simply normalizing both distributions to unity by an increase of only approximately 8% in the shell values. The result shown in Figure 6a is that obtained using the described approximate correction for the number of nearest neighbors. As can be seen, the resulting shell distribution is slightly shifted toward weaker bonding in the bonding region. Thus the dipeptide apparently has a small but noticeable overall disruptive effect on the water-water bonding in its vicinity. The result suggests that perhaps the process of dissolving the dipeptide occurs in part at the expense of the strength (as well as number) of water-water interactions. Figure 6b shows a corresponding distribution for water dipeptide pairs. The values (N) include all water molecules in the system, and thus the ordinate corresponds to the actual number of water molecules (out of 195) contributing for each energy. The general shape (i.e., the broad bonding region) is similar to that seen in the water-water results, although the statistical noise is necessarily much larger. By comparing results accumulated over various

portions of the run, it would appear that the "bump" at about -5 kcal is somewhat questionable but that the shoulder at about -2 kcal is more likely to be a real effect. This could result from the relatively large number of water molecules which contact the peptide groups, but not in a particularly favorable geometry. By summing the contributions to the distribution, we find that an average of 4.8 water molecules (2.5%) are bonded to the dipeptide (pair energy $\leq$ -2.5 kcal), a result consistent with the earlier observations of particular configurations. The sum of all those with pair energies less than -1 kcal or greater than +.5 kcal is 21.9 molecules. Since some of the water molecules near the non-polar groups may also be expected to contribute in the -1 to .5 kcal range, these 21.9 molecules include some second neighbors to the polar peptide groups.

The group of molecules with interaction energies between -.25 kcal and +.25 kcal constitute 70% of the 195 molecules. The fact that such a large percentage interact very weakly with the dipeptide is supportive of our implicit assumption that the effective concentration we are using is not unreasonably high for a useful study of this system.

IV. Solvent Dynamics

One is particularly interested in the water molecules which have average positions within the 4Å "shell" around the dipeptide; during the current run there are 23. Of these, 8 are separated by an average distance of more than 4Å from any polar atom, and these constitute a "non-polar" group. In addition, 10 of these 23 are either solely within the 4Å radius around the carbonyl oxygen or amide hydrogen atoms, or are significantly closer to these than to any non-polar atom. These are classified as a "polar" group. The "bulk" consists of the 172 non-"shell" water molecules. This classification scheme seems to be a reasonable approach to isolating peculiar behavior associated with particular solvent-solute interactions.

Various unequal time auto-correlation functions have been examined. We define these functions for a property A(t) as,

$$C(t) = \langle A(0)A(t)\rangle / \langle A^2(0)\rangle$$

where the bracket indicates an average over the simulation. If A is a vector quantity, the ordinary product is replaced by the vector dot product in the current applications. These functions, C(t), have been computed for each water molecule and then averaged in groups according to the defined classes. The increased statistical error inherent in this subdivision procedure, due to the reduced number of water molecules involved in each average, should be noted. The correlation functions considered are the center of mass velocity, probing translational motion, and the total angular momentum and dipole reorientation, probing rotational motion. We use a dipole defined by

$$\underset{\sim}{\mu}_i = \left(\underset{\sim}{r}_{H_1^i} - \underset{\sim}{r}_{O^i}\right) + \left(\underset{\sim}{r}_{H_2^i} - \underset{\sim}{r}_{O^i}\right)$$

where H_1^i and H_2^i are the two hydrogens of the water molecule 'i'; this vector has a mean magnitude of 1.286Å and fluctuates (due to internal vibration) relatively little.

Due to the short time interval examined, the mean square center of mass velocities of the groups varied sufficiently (a range of 80°K) that substantial difference in their velocity autocorrelation functions are to be expected from temperature effects alone.[6] Thus, the only conclusion that could be drawn from these was that the differences observed could be acccunted for solely by these temperature differences. Estimates of rotational temperatures for these groups of water molecules were similar (mean deviation from the mean of 8°K), and the resulting correlation functions (for short times) are shown in Figures 7 and 8. The differences

among the angular momentum correlation functions do not seem substantial enough to suggest large differences in rotational diffusion rates, when compared to bulk. The characteristic frequency for the oscillatory decay of the angular momentum also appears to be very similar for each class. The dipole reorientation rates, as indicated by the slopes of the curves in Figure 8 for times greater than about .1 psec, appear to be very similar. These results, however, are not inconsistent with the possibility that differences exist at lower temperature.

V. Dipeptide Internal Structure and Motion

In order to compare results in solution to those in vacuum, a parallel simulation has been done in the absence of solvent starting from a typical configuration in the solution dynamics. This yields the results referred to below as "vacuum."

A. Average Structure

Bond lengths and bond angles are affected very.little by the solvent. The mean bond lengths in solution differ from those in vacuum by less than .005Å. They are all slightly smaller in solution with the exception of the polar peptide bonds (C-O, N-H) which are slightly longer. The root mean square fluctuation in bond lengths are about .025Å and essentially the same in solution. Similar results apply to bond angles; they differ typically by only about 1° between solution and vacuum and the rms fluctuations are about 3.7° in both cases. There is somewhat more difference in dihedral angles. Tables IIIa and IIIb summarize the results. The mean values in vacuum are essentially the values at the dipeptide potential energy minimum. For the peptide angles ω, we note that they are essentially planar in vacuum, while there is a small but distinct non-planarity in solution.

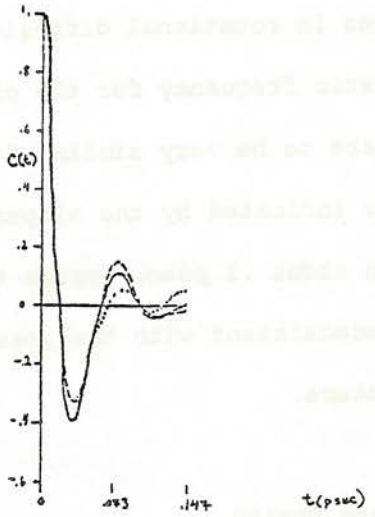


Figure 7: Autocorrelation function of water molecule total angular momentum: (——, bulk; ····, non-polar; ---, polar).

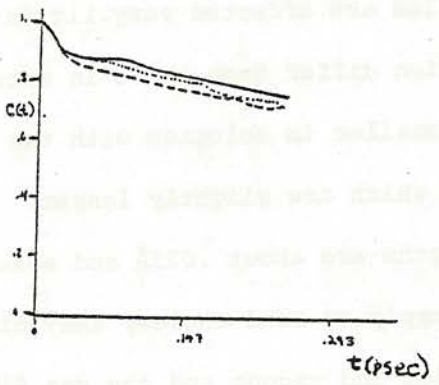


Figure 8: Autocorrelation function of water molecule dipole reorientation: (——, bulk; ····, non-polar; ---, polar).

Table IIIa: Dihedral Angle Mean Values and Fluctuations[a]

	Vacuum	Solution
$\langle\phi\rangle$	187.6	174.6
$\langle\psi\rangle$	295.8	299.0
$\langle\omega_1\rangle$	180.0	169.0
$\langle\omega_2\rangle$	181.3	186.1
$\langle\Delta\phi^2\rangle^{1/2}$	13.0	10.3
$\langle\Delta\psi^2\rangle^{1/2}$	12.3	10.0
$\langle\Delta\omega_1{}^2\rangle^{1/2}$	12.5	8.3
$\langle\Delta\omega_2{}^2\rangle^{1/2}$	9.0	10.0

(a) $\Delta\theta = (\theta - \langle\theta\rangle)$, in degrees.

Table IIIb: Dihedral Angle Equal Time Correlations[a]

vac \ sol	ϕ	ψ	ω_1	ω_2
ϕ	1.	-0.573	-0.259	-0.004
ψ	0.005	1.	-0.214	0.311
ω_1	0.082	-0.124	1.	-0.100
ω_2	0.309	0.032	-0.195	1.

a) The values given are $\dfrac{\langle\Delta\theta_i\Delta\theta_j\rangle}{\langle\Delta\theta_i{}^2\rangle^{1/2}\langle\Delta\theta_j{}^2\rangle^{1/2}}$

Since the ϕ-ψ potential well is relatively narrow, the fluctuations in ϕ and ψ are relatively small, and comparable in vacuum and solution. A distinct difference is seen in the equal time correlations among the angles, in particular for the ϕ, ψ pair. The observation of this negative correlation could be a result of the fact that the requisite motion (rotation of ϕ and ψ in opposite senses) can be carried out with relatively small changes in the overall positions of the two peptide groups. However, we cannot make a strong conclusion from the limited simulation.

B. Dynamics

Time correlation functions (computed to a maximum of .55 psec) for the internal motions of the dipeptide have also been examined. Figures 9 through 12 show typical resulting functions and their Fourier transforms. In each case, the vacuum result is shown as a solid line and the solution result as a dashed line. Due to the limited "sampling" time, there are significant sidebands in the transforms. The "broadening" function, $\sin \omega T/\omega T (T = .55$ psec), is shown in the insert in Figure 10b. These sidebands can be reduced by the use of weighting function, as is done in modern NMR work, but this has not yet been carried out. Also, relatively small differences in amplitudes between solution and vacuum results can easily be the result of unequal distribution of internal energy among the various modes.

As can be seen, the "hard," high frequency, modes associated with bonds and bond angles behave quite similarly in either environment. Each shows evidence for very long lived oscillatory correlations. The bond angle $N_L - C_\alpha - C_R$, however, might be expected to include lower frequency motions corresponding to the overall bending of the dipeptide about this central angle. In solution (Fig. 11b), the peak at about 29 $psec^{-1}$ ($154cm^{-1}$) is very much larger than in vacuum.

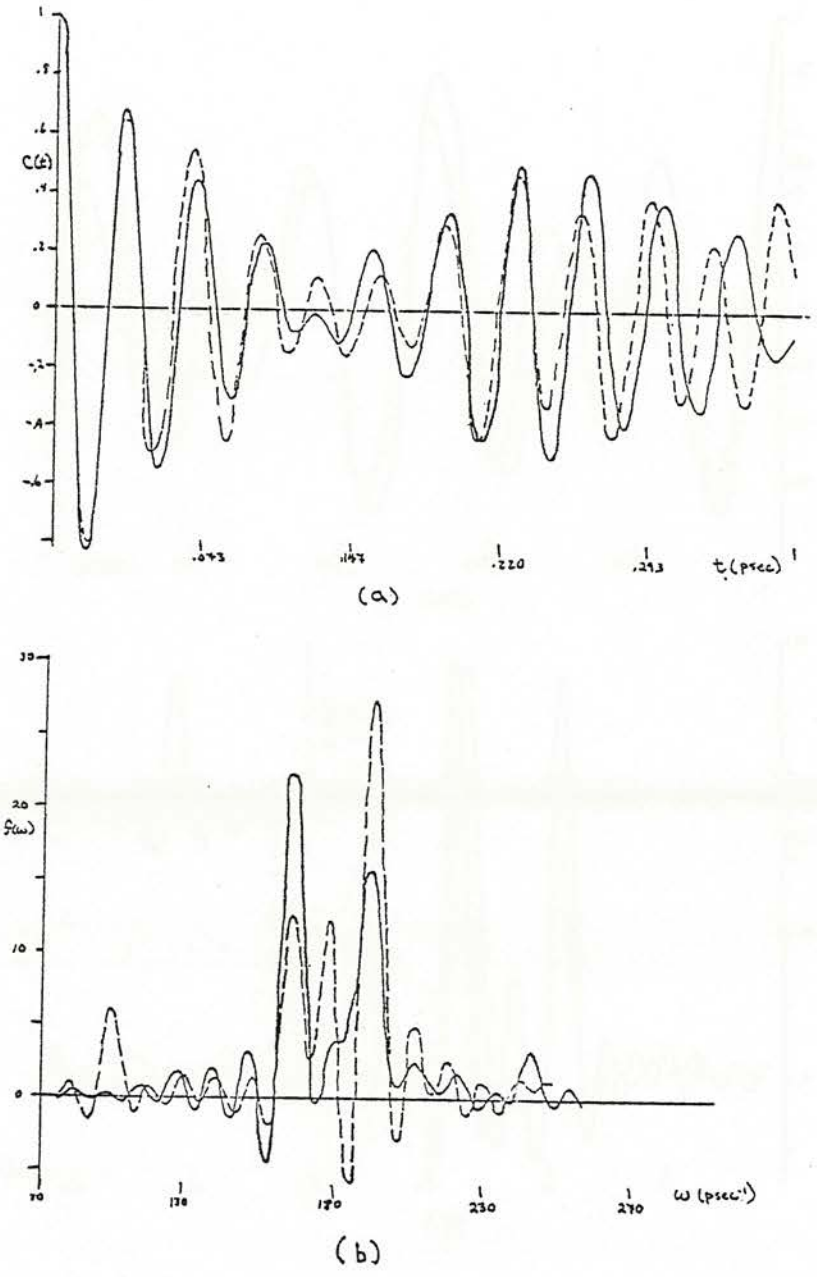


Figure 9: a) Autocorrelation function for the bond length, $C_\alpha - C_\beta$;
b) Fourier transform of C(t), arbitrary units. (——, vacuum; – – –, solution.)

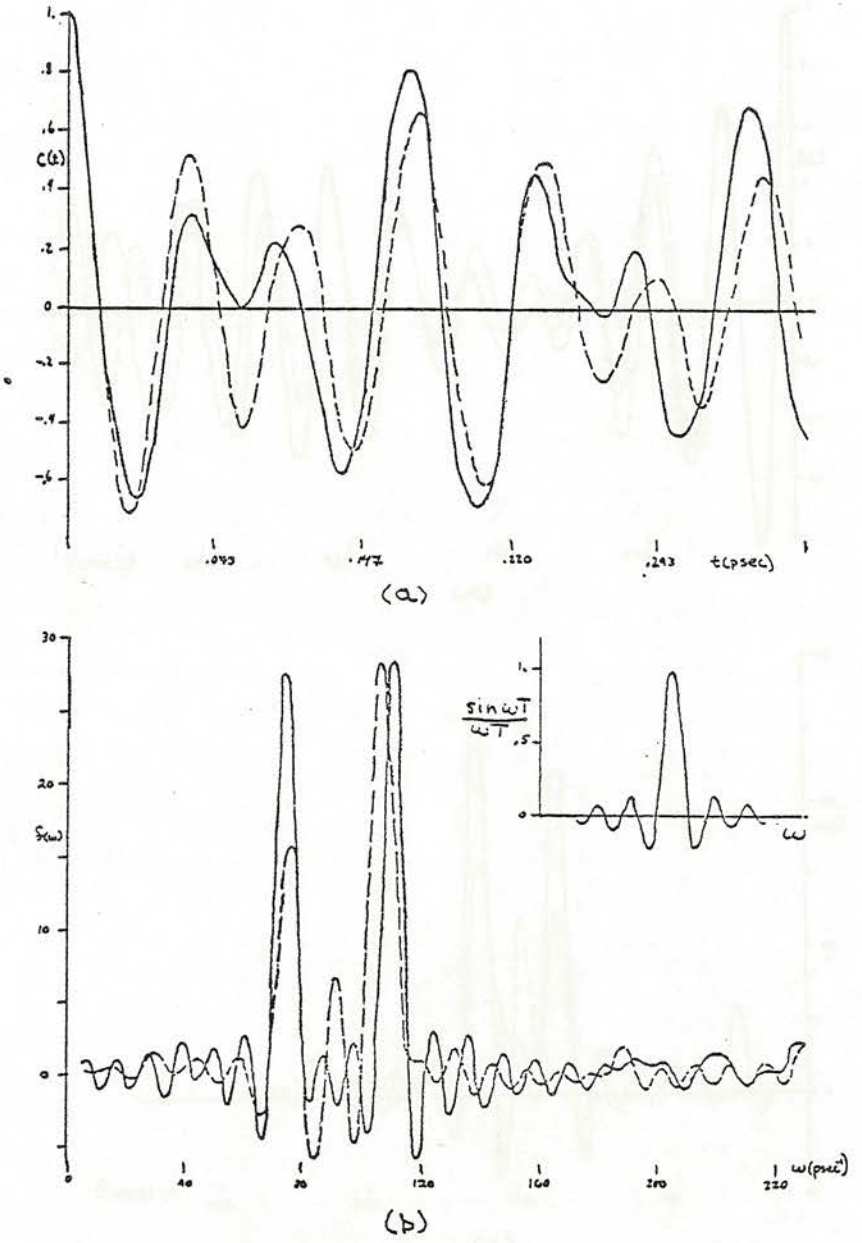
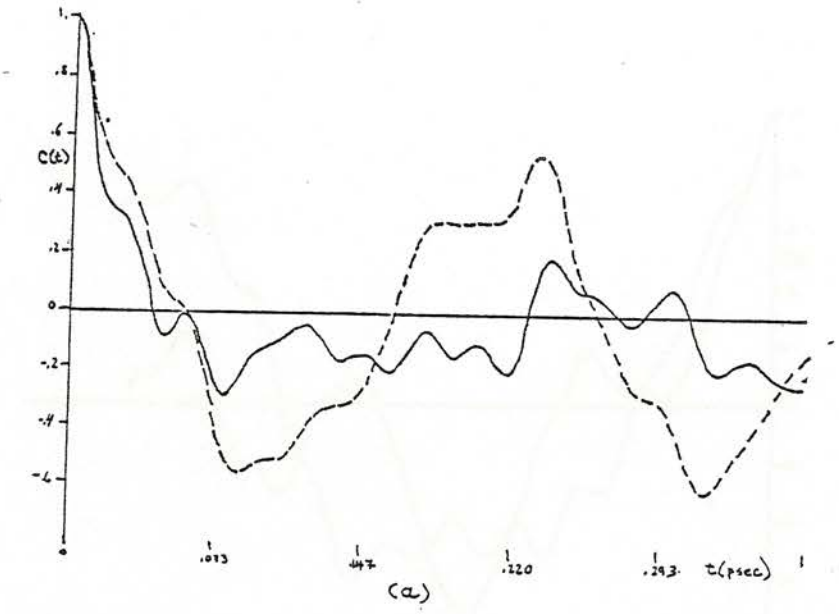


Figure 10: a) Autocorrelation function for the bond angle, $C_L - C_L' - O_L$;
b) Fourier transform of C(t), arbitrary units. (——, vacuum;
---, solution.)



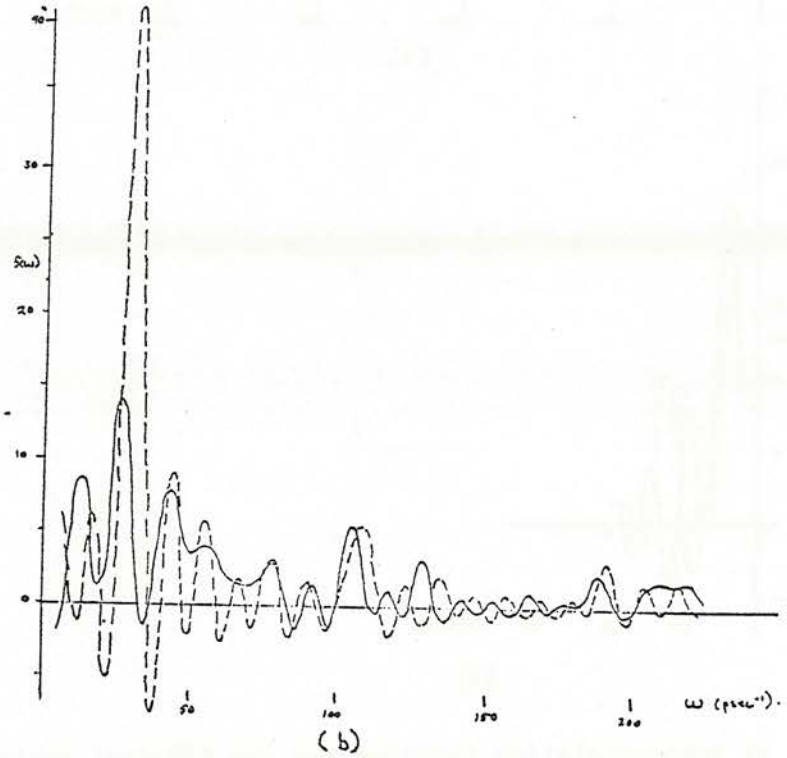


Figure 11: a) Autocorrelation function for the bond angle, N_L - C_α - C_R';

b) Fourier transform of C(t), arbitrary units. (——, vacuum; ---, solution.)

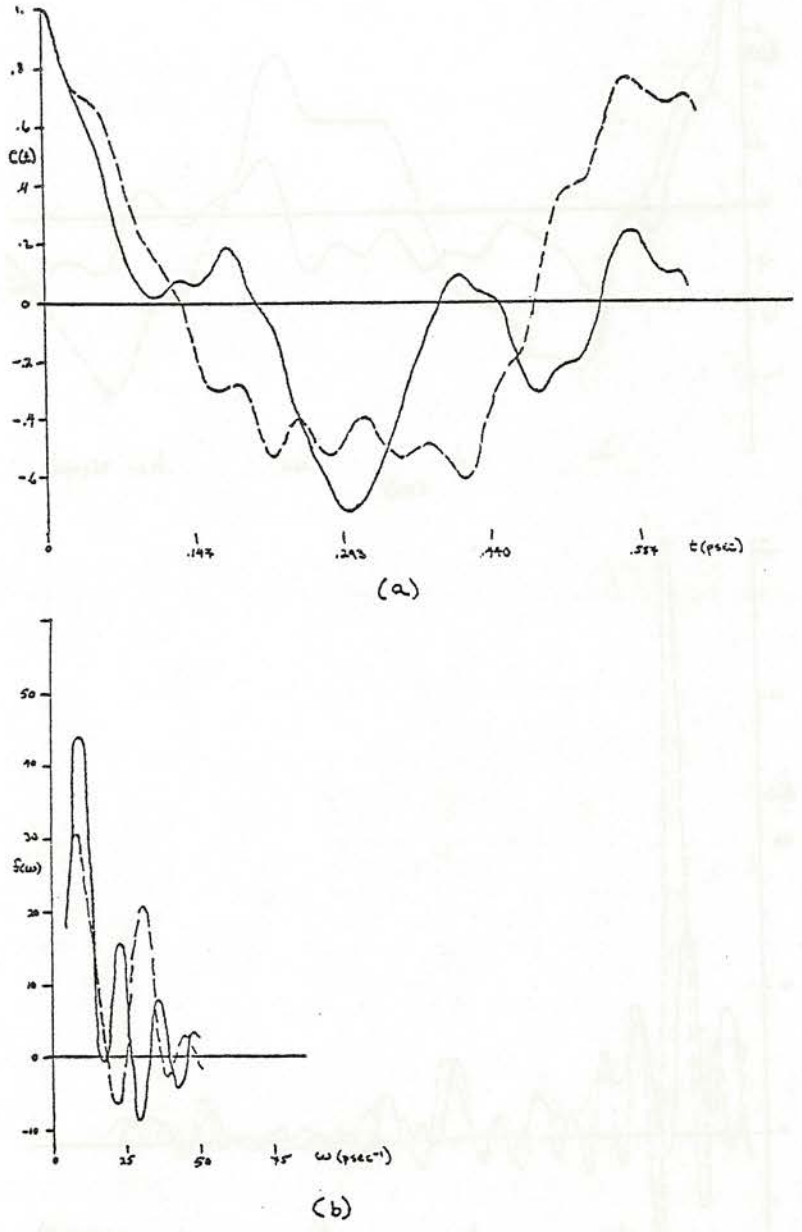


Figure 12: a) Autocorrelation function for the dihedral angle, ϕ;
b) Fourier transform cf C(t), arbitrary units. (——, vacuum; ---, solution.)

It is quite possible that this reflects coupling to low frequency collective translational motions seen in pure water studies,[6] as well as in our own simulations. For the representative dihedral angle, ϕ, (Figure 12) the principal component is at low frequency in vacuum, and therefore it is inherently difficult to see this possible effect here.

VI. Conclusion

We have found that the model proposed for simulation of this solution gives physically reasonable bonding structures and energies. Analysis indicates that the bonds break and reform fairly quickly, as expected. Analysis of the water-dipeptide pair energies suggests that the system used is not unreasonably small for an investigation of this type.

Study of the dipeptide's "hydration shell" points out the inherent statistical problems associated with a solution as opposed to a pure liquid. However, the described classification of the water molecules in this shell according to the functionality of the neighboring dipeptide atoms seems to be a useful approach for analysis, while retaining some statistical advantages.

No significant differences in the motion of these solvent molecules could be discerned from the limited analysis that was possible. The average structure of the dipeptide, at least in the neighborhood of the local minimum studied, is quite similar in solution to that in vacuum. The most significant difference found was in the peptide non-planarity in solution, perhaps allowing greater accommodation of solvent bonding. The size of fluctuations in internal coordinates are also quite similar in vacuum and solution.

Currently we are carrying out a longer simulation with the dipeptide in the equatorial C_7 conformation and at a temperature of approximately

40°C. The results of the analysis presented above can be used as a very valuable guide in this study.

Acknowledgements

Martin Karplus, and Bruce Gelin have been collaborators in this work. Shoshana Wodak kindly provided the stereoscopic picture.

References

1. a) F. Franks, D. Eagland, Critical Reviews in Biochemistry 3, 165 (1975).
   b) R. Cooke, I. D. Kuntz, Ann. Rev. Biophys., Bioeng. 3, 95 (1974).
2. See also, for example, F. Franks, in Water, A Comprehensive Treatise, F. Franks, ed., Plenum (New York, 1972).
3. M. Avignon, C. Garrigan-Lagrange, P. Bothorel, Biopolymers 12, 1651 (1973).
4. a) A. Warshel, M. Levitt, S. Lifson, J. Mol. Spec. 33, 84 (1970).
   b) B. R. Gelin, thesis, Harvard University (1976).
5. R. F. McGuire, F. A. Momany, H. A. Scheraga, J. Phys. Chem. 76, 375 (1972).
6. F. H. Stillinger, A. Rahman, J. Chem. Phys. 60, 1545 (1974).
7. A. Rahman, F. H. Stillinger, H. L. Lemberg, J. Chem. Phys. 63, 5223 (1975).
8. A Johansson, P. Kollman, S. Rothenberg, J. McKelvey, JACS 96, 3794 (1974), and references therein.
9. L. Bǿje, A. Hvidt, J. Chem. Thermo. 3, 663 (1971).
10. W. W. Wood, in Physics of Simple Liquids, H. N. V. Temperley, J. S. Rowlinson, G. S. Rushbrooke, ed.s, American Elsevier, (New York, 1968).
11. P. K. Ponnuswamy, P. Manavalan, J. Theor. Biol. 60, 481 (1976).
12. A. Rahman, F. H. Stillinger, J. Chem. Phys. 55, 3336 (1971).
13. C. J. Montrose, J. A. Bucaro, J. Marshall-Coakley, T. A. Litovitz, J. Chem. Phys. 60, 5025 (1974).
14. C. M. Davis, Jr., J. Jarzynski, in Water, A Comprehensive Treatise, op. cit.

References

1. [illegible]
2. [illegible]
3. [illegible]
4. [illegible]
5. [illegible]
6. [illegible]
7. [illegible]
8. [illegible]
9. [illegible]
10. [illegible]
11. [illegible]
12. [illegible]
13. [illegible]
14. [illegible]
15. [illegible]

II.3

MOLECULAR DYNAMICS STUDY OF THE BOVINE PANCREATIC TRYPSIN INHIBITOR

J.A.McCammon

Harvard University, Department of Chemistry, Cambridge, Mass 02138 (USA).

I. INTRODUCTION

Little is known at present about the details of motions within protein molecules. One method for exploring these motions is the solution of the classical equations of motion for the atoms comprising a protein and analysis of the resulting phase-space trajectory. Here, we report preliminary results of an application of this "molecular dynamics" method to a small (58 residues) protein molecule, the bovine pancreatic trypsin inhibitor (BPTI). Our calculation is for an isolated molecule and accordingly will not incorporate features due to the solvent environment.

II. METHODS

The essential component of any molecular dynamics study is the potential function used to describe the interactions of the atoms in the assembly of interest. Once a potential function is chosen, any of several numerical schemes may be used to determine changes in atomic positions and velocities at a succession of times. The desired trajectory may be developed from any set of initial atomic positions and velocities.

Empirical potential energy functions have been used for several years as a tool for refinement of x-ray structures [1] and, more recently, to assess the consequences of protein flexibility. [2-4] Here, we use an empirical potential function developed by Gelin and Karplus [2] as the basis for our molecular dynamics study of BPTI. This function is composed of a sum of terms representing contributions from bond stretching, bond angle bending, dihedral angle twisting, hydrogen bonds, non-bonded (van der Waals) interactions, and electrostatic interactions. Hydrogen atoms are not explicitly considered in this model, but are combined with the heavy atoms to which they are bonded by a suitable adjustment of heavy-atom parameters. This use of "extended atoms" reduces the number of interactions which must be calculated at each point in time and also permits larger steps in the trajectory calculation since the high frequency hydrogen vibrations have been eliminated.

Integration of the equations of motion was performed by using the Gear algorithm [5] and a time step of 9.78×10^{-16} sec. X-ray coordinates were used for the initial positions, and the initial velocities were zero. After 100 steps, the stresses in the initial structure had relaxed and the potential energy had developed into an internal kinetic energy corresponding to a temperature $T \approx 140°$ K. At this point, all velocities were multiplied by a factor of 1.5 and 250 more steps were taken. The added kinetic energy was allowed to partition itself between kinetic and potential terms during this interval, and the system achieved an average temperature $T \approx 295°$ K. Finally, 9,000 further steps were taken for the purpose of analysis. The analysis period corresponds to

8.8 psec. During this 9,000 step series, the total energy of the protein was well-conserved, changing by only 0.7 kcal/mole. The 9,000 step trajectory calculation required approximately 2 hours of CPU time on the IBM 370/168.

III. PRELIMINARY RESULTS

A. General Features of the Protein Motion

At 300° K, almost all the atoms of BPTI remain near their average positions, performing motions of small amplitude. In a few instances, atoms are observed to move from the region of one energy minimum to another (e.g. the sterically unhindered rotation of small groups of atoms). No tendency of the protein to unfold is observed. The root-mean-square (RMS) atomic displacement per Cartesian component was $\langle (x - \langle x \rangle)^2 \rangle^{1/2} = 0.52$ Å. Considered as a set of harmonic oscillators, the atoms move subject to apparent restoring forces with an average force constant $k = k_B T / \langle (\Delta X)^2 \rangle \approx 1.7 \times 10^3$ erg/cm^2 (2.4 kcal/mole Å^2) for displacements in the X,Y, or Z directions. This force constant is about two orders of magnitude smaller than those associated with bond stretching (e.g. $k \approx 4.2 \times 10^5$ erg/cm^2 (600 kcal/mole Å^2) for C-C stretching).

Unusually large RMS displacements (1.5 - 3.0 Å in one or more components) are observed in the side chains of Lys 26, Arg 39, Arg 42, Lys 46, Met 52 and in the C-terminal residue, Ala 58. All of these groups are at the surface of the protein. Some of these mobile regions correspond to groups of atoms which are not located or which have high temperature factors in the x-ray diffraction work on this protein (Lys 26, Arg 42, Lys 46, Ala 58), but there is not an exact one-to-one correspondence between mobile atoms and atoms with high temperature factors.

Even in regions of the protein with more typical RMS coordinate fluctuations, the atomic mobility is great enough that substantial displacements occur in the internal coordinates (bond lengths, bond angles, and dihedral angles). In Figs. 1 and 2, we show the evolution of several internal coordinates of the Phe 22 residue, beginning with step 5,000 and continuing for about 0.5 psec.[6] Phe 22 is one of the BPTI residues which is relatively well buried in the protein matrix. A statistical analysis of such internal coordinate motions is provided in the subsequent sections.

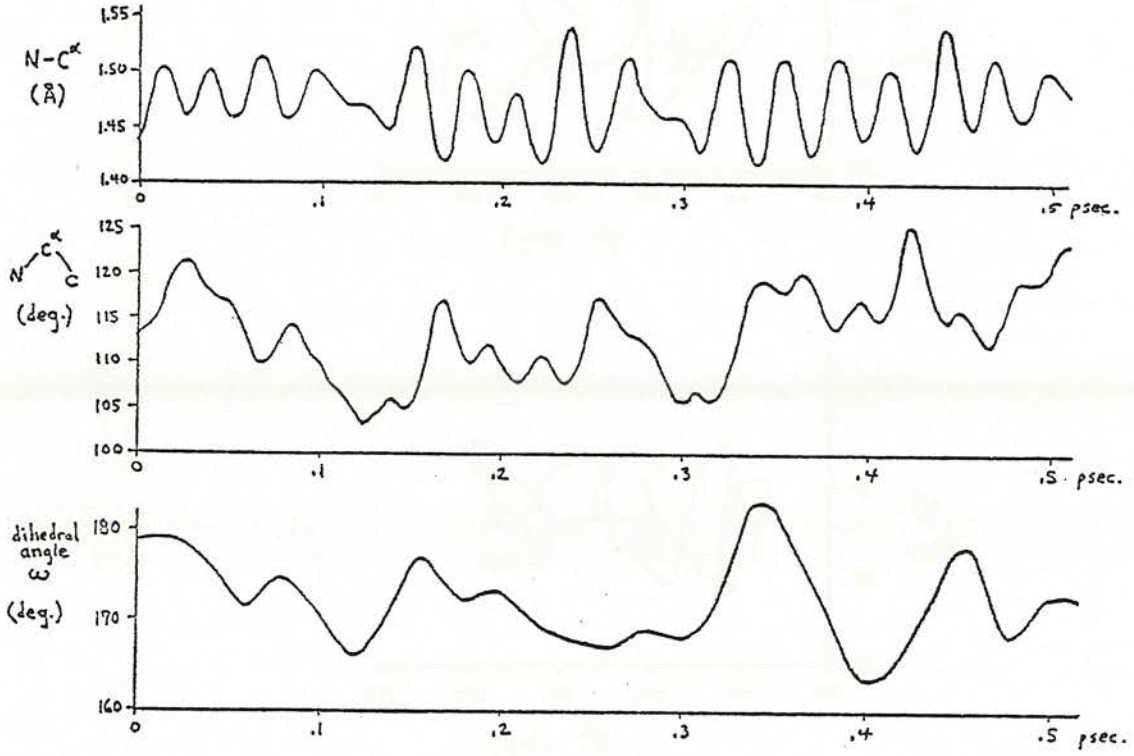


Fig. 1

Evolution of three internal coordinates of the Phe 22 backbone, starting at step 5000 and continuing for 0.5 psec. Histories of the N-C^{α} bond length, the N-C^{α}-C bond angle, and the (relatively stiff) dihedral angle ω are shown.

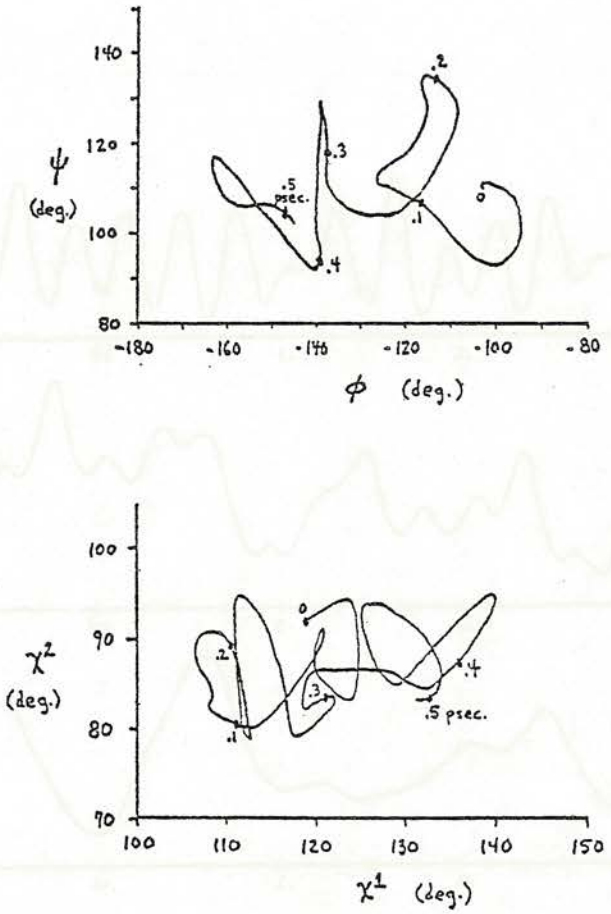


Fig. 2

Evolution of the relatively soft Phe 22 dihedral angles ϕ, ψ (in the backbone) and χ^1, χ^2 (in the side chain), starting at step 5000 and continuing for 0.5 psec.

III. PRELIMINARY RESULTS (cont.)

B. Backbone Internal Coordinates

In this section, we present the results of a preliminary analysis of the mobility of internal coordinates in the polypeptide backbone.

1) Equilibrium Correlations

Mean values and root-mean-square fluctuations for backbone bond lengths are given in Table 1. The corresponding quantities for the bond angles are given in Table 2, and those for torsional angles in Table 3.

TABLE 1. Backbone Bond Lengths: typical mean values and RMS flucuations.

Bond	Mean Length (Å)	RMS Fluctuation (Å)
$N-C^{\alpha}$	1.47 - 1.48	0.02 - 0.03
$C^{\alpha}-C$	1.53 - 1.54	0.02 - 0.03
C-O	1.23 - 1.24	0.01 - 0.02
C-N	1.32 - 1.33	0.01 - 0.02

TABLE 2. Backbone Bond Angles: typical mean values and RMS fluctuations.

Bond Angle	Mean Angle (deg.)	RMS Fluctuation (deg.)
$C-N-C^{\alpha}$	123 - 129	4 - 5
$N-C^{\alpha}-C$	111 - 118	4 - 5
$C^{\alpha}-C-N$	113 - 116	4 - 5
$C^{\alpha}-C-O$	117 - 121	4 - 5
O-C-N	122 - 123	4 - 5

TABLE 3. Backbone Torsional Angles: typical mean values and RMS flucutations.

Torsional Angle	Mean Angle (deg.)	RMS Fluctuation (deg.)
ϕ	wide variation	15 - 30
ψ	wide variation	15 - 30
ω	175 - 185	7 - 10

The RMS fluctuations of bond lengths and bond angles are fairly constant along the backbone except for slightly (~20%) larger fluctuations in the residues nearest the C-terminal end. The same behavior is seen in the dihedral angle ω. The softer dihedral angles ϕ and ψ show markedly greater variation; large RMS fluctuations in these angles sometimes appear to be associated with specific interactions connecting different parts of the polypeptide chain. Particularly large RMS fluctuations are found in ψ_{38} and ϕ_{39} (near the 14-38 disulfide bond), ψ_{44} and ϕ_{45} (near the backbone hydrogen bonds $NH_{21}\ldots O_{45}$, $O_{21}\ldots HN_{45}$), and in several ϕ,ψ angles at the C-terminal end of the chain.

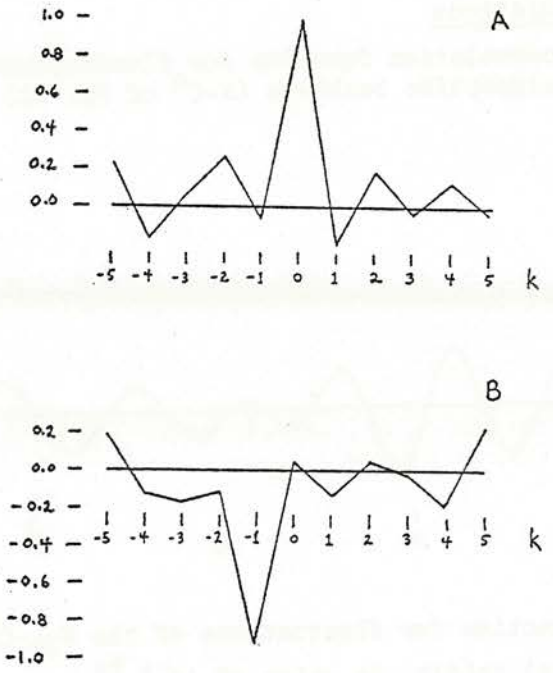
Examination of the time evolutions of ϕ ψ angles with large RMS fluctuations revealed several instances of concerted transitions in ψ_{i-1}, ϕ_i pairs. In these transitions, ψ_{i-1} and ϕ_i rotate with opposite senses by about 180°; the intervening amide group flips over, but backbone and side chain directions are undisturbed. Such transitions occurred in the pairs ψ_{38}, ϕ_{39}; ψ_{44}, ϕ_{45}; and ψ_{56}, ϕ_{57}. Concerted ψ_{i-1}, ϕ_i transitions are dramatic examples of the correlation which is generally found in the flucutations of these angles (see below).

Backbone hydrogen bonds typically have lengths (N to O distance) of 2.9 Å to 3.1 Å and RMS length fluctuations of 0.1 Å to 0.2 Å. Two backbone hydrogen bonds showed marked stretching in our dynamical simulation; they are $NH_{16}\ldots O_{36}$ (RMS fluctuation = 0.7 Å) and $NH_{27}\ldots O_{24}$ (RMS fluctuation = 0.6 Å).

The polypeptide backbone of BPTI forms two major pieces of secondary structure. Residues 18 through 24 are associated with residues 29 through 35 in an antiparallel β-sheet, and residues 47 through 56 form two complete turns of α-helix at the C-terminal end of the protein. The average RMS length fluctuation of hydrogen bonds in the β-sheet is 0.13 Å,

while the corresponding average for the α-helix is 0.22 Å. Thus, the β-sheet seems to be a more stable element of secondary structure in our model of BPTI.

In addition to the RMS fluctuations discussed above, we have calculated the equilibrium correlations of <u>different</u> dihedral angles. Correlations of fluctuations in ϕ_{33} (which is at the center of a strand of β-sheet) with fluctuations in ϕ_{33+k} and ψ_{33+k}, k = -5, -4,..., +4, +5, are presented in Fig. 3.



<u>Fig. 3</u> Normalized equal-time correlations of fluctuations in ϕ_{33} with fluctuations in ϕ_{33+k} (A) and with fluctuation in ψ_{33+k} (B). The quantity plotted in A is $\langle\Delta\phi_{33}\Delta\phi_{33+k}\rangle/\langle(\Delta\phi_{33})^2\rangle^{1/2}\langle(\Delta\phi_{33+k})^2\rangle^{1/2}$. The quantity plotted in B is $\langle\Delta\phi_{33}\Delta\psi_{33+k}\rangle/\langle(\Delta\phi_{33})^2\rangle^{1/2}\langle(\Delta\psi_{33+k})^2\rangle^{1/2}$. Fluctuations are measured with respect to mean values:

$$\Delta\phi_{33+k} = \phi_{33+k} - \langle\phi_{33+k}\rangle, \text{ etc.}$$

Over distances which are not too great (i.e. up to $k \simeq \pm 3$), these correlations follow the pattern of the dihedral angle fluctuation correlations recently found for an isolated α-helix [7]. It seems likely that this pattern of dihedral angle corrrelations will be found commonly to occur in strongly conserved regions of secondary structure in proteins. A fortiori, the highly localized correlation $\langle \Delta\phi_i \Delta\psi_{i-1} \rangle$ is found typically to be negative and of similar absolute magnitude to $\langle (\Delta\phi_i)^2 \rangle$ in BPTI, even outside regions of formal secondary structure. Such correlations of $\Delta\phi_i$ with $\Delta\psi_{i-1}$ conserve the general direction of the polypeptide backbone and may be expected to occur generally in proteins where the backbone folding is strongly conserved.

2) Time Correlations

The time correlation function for fluctuations in the length of a bond in the polypeptide backbone (N-C^{α} of Phe 22) is shown in Fig. 4.

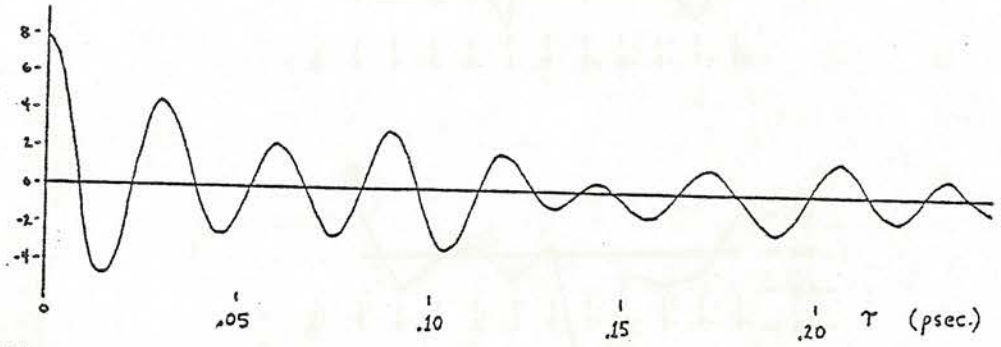


Fig. 4

Time correlation function for fluctuations of the N_{22}-C^{α}_{22} bond length, $\langle \Delta b(\tau)\, \Delta b(0) \rangle$, in units of 10^{-4} Å^2.

It is worth noting again that Phe 22 is a relatively well buried residue. There is a partial loss of correlation in the first 0.1 psec., followed by a long interval in which a residual small amplitude oscillation decays very slowly. The amplitude of the correlation function is 1.5×10^{-4} Å^2 after a time $\tau = 0.6$ psec. Such long-lived correlations appear to be a typical feature of bond length fluctuations. The Fourier transform of the the N_{22}-C^{α}_{22} time correlation function is shown is Fig. 5.

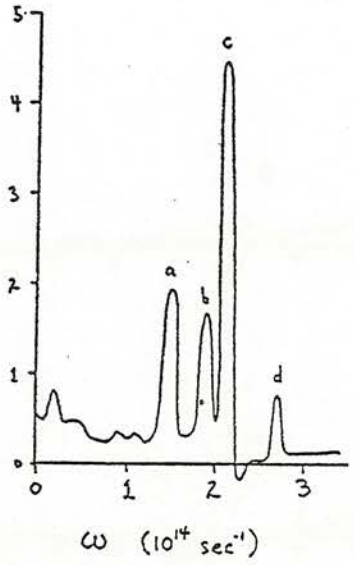


Fig. 5. Fourier transform of the time correlation function for fluctuations of the N_{22}-C^{α}_{22} bond length, in arbitrary units. Peaks labeled a,b,c,d occur at approximate wavenumbers of 800 cm^{-1}, 1010 cm^{-1}, 1110 cm^{-1}, 1430 cm^{-1}, respectively.

In addition to the four explicit peaks at wavenumbers of about 800 cm^{-1}, 1010 cm^{-1}, 1110 cm^{-1}, and 1430 cm^{-1}, there is probably a fifth peak near 1170 cm^{-1} which cancels the negative "wing" at the right of the 1110 cm^{-1} peak; these negative "wings" are expected to occur as a result of the truncation of the time correlation function at $\tau \simeq 0.6$ psec. before Fourier transforming. In addition to these peaks, the spectrum shows relatively small contributions from many low frequency motions to fluctuations in the N_{22}-C^{α}_{22} bond length.

In an attempt to identify the origins of the peaks in the N_{22}-C^{α}_{22} frequency spectrum, we performed a normal-mode analysis of the isolated fragment CH_3-CO-Phe-NH-CH_3. The vibrational motions of this fragment which involve the largest displacements in the N-C^{α} bond occur at wavenumber 1425 cm^{-1}, 1161 cm^{-1}, 1099 cm^{-1}, 952 cm^{-1}, and 787 cm^{-1}. One may conclude that the N_{22}-C^{α}_{22} bond fluctuations are simply those expected for a strongly coupled system of local oscillators which is weakly coupled to the rest of the protein molecule.

Time correlation functions for the fluctuations of the three backbone dihedral angles ϕ,ψ and ω of Phe 22 are shown in Fig. 6

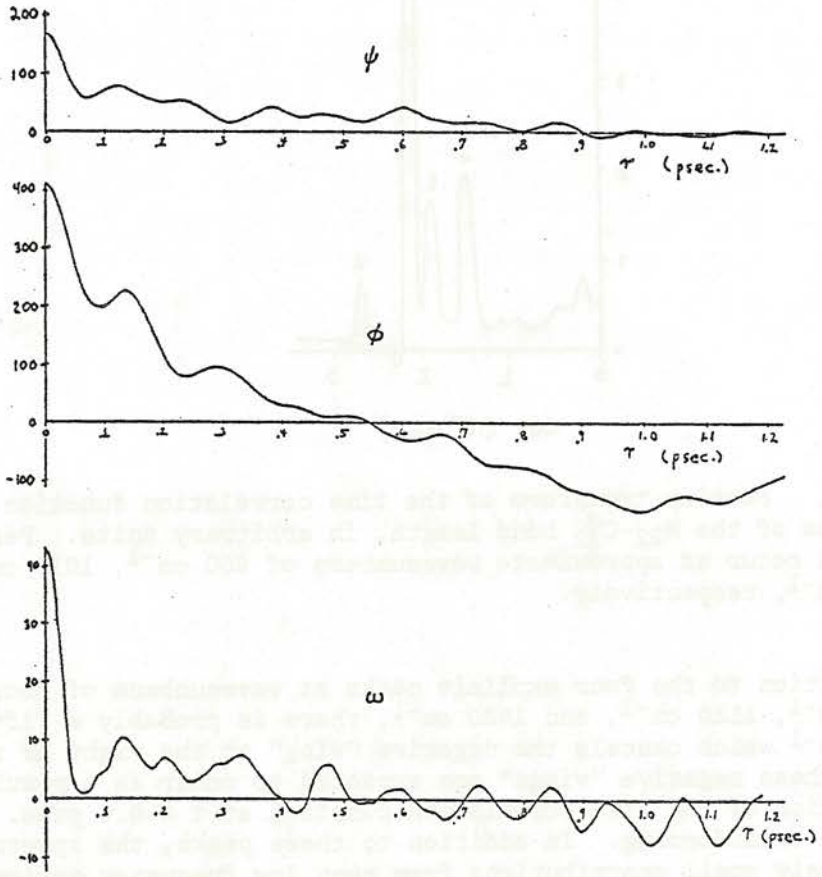


Fig. 6
Time correlation functions for fluctuations of the backbone dihedral angles ϕ_{22}, ψ_{22} and ω_{22}, in deg^2.

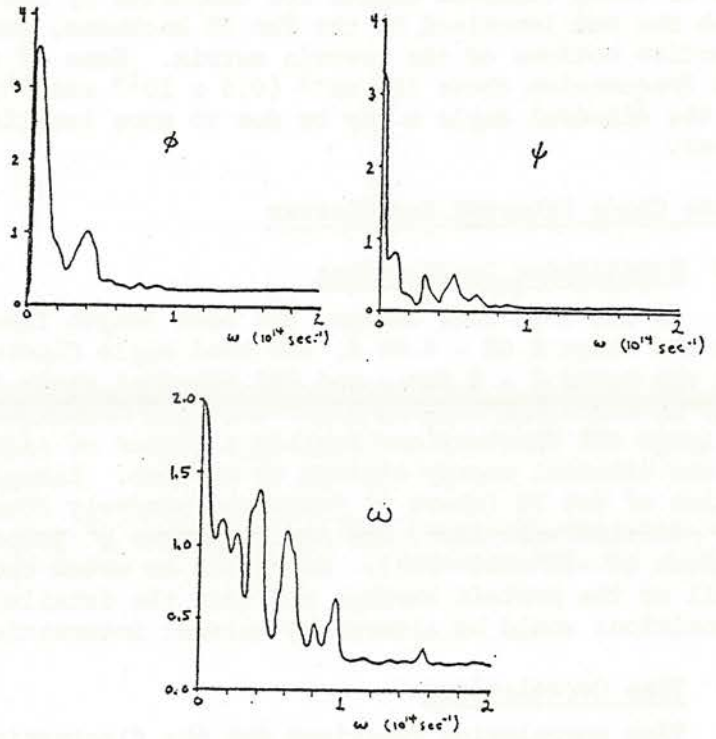


Fig. 7

Fourier transforms of the time correlation functions for fluctuations of the backbone dihedral angles ϕ_{22}, ψ_{22} and ω_{22}, in arbitrary units.

The dihedral angle ω, which has a relatively large intrinsic restoring force, exhibits behavior similar to the N_{22}-C_{22} bond. There is a partial loss of correlation in the first 0.1 psec., followed by the slow decay of a residual small amplitude oscillation. In the softer dihedral angles ϕ and ψ, initial correlations die out more slowly.

The Fourier transforms of these dihedral angle time correlation functions are shown in Fig. 7. In general, it appears that the fluctuations of these dihedral angles are dominated by low frequency motions which are not localized in the Phe 22 backbone, but which are rather collective motions of the protein matrix. Some of the structure appearing at frequencies above 300 cm^{-1} (0.6×10^{14} $sec.^{-1}$) in the spectrum of the dihedral angle ω may be due to more localized normal modes, however.

C. Side Chain Internal Coordinates

1) Equilibrium Correlations

In the BPTI side chains, RMS bond length fluctuations are typically in the range 0.02 - 0.04 Å, RMS bond angle fluctuations are typically in the range 3 - 6 deg., and RMS dihedral angle fluctuations are typically in the range 10 - 40 deg. The time evolutions of dihedral angles with large RMS fluctuations exhibit a number of side-chain transitions from one dihedral energy minimum to another. Examples include the side chains of Met 52 (where χ^2 jumps successively from the regions of $-60° \rightarrow 60° \rightarrow -60° \rightarrow 180° \rightarrow 60° \rightarrow 180°$) and Arg 39 (where χ^2 jumps successively from the regions of $-60° \rightarrow 180° \rightarrow 60°$). It should be noted that these side chains are all at the protein surface and that the detailed dynamics of their transitions would be altered by solvent interactions in solution.

2) Time Correlations

Time correlation functions for the fluctuations of side chain bond lengths are similar to those for backbone bonds: there is a partial loss of correlation in the first 0.1 psec., followed by slow decay of the resudual small amplitude oscillation. Fourier transforms of these correlation functions are dominated by a few discrete peaks, suggesting that bond length fluctuations are primarily those of a local system of coupled high frequency oscillators.

The time correlation function for fluctuations of the dihedral angle χ^2_{22} (which is the rotation angle for the phenyl ring of the relatively buried residue Phe 22) is shown in Fig. 8.

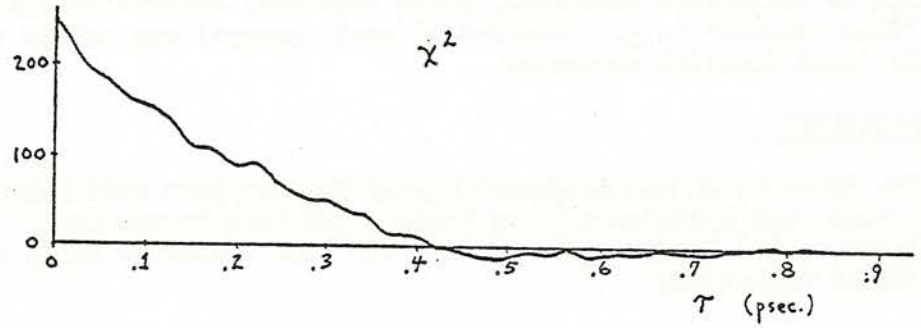


Fig. 8

Time correlation function for fluctuations of the sidechain dihedral angle χ^2_{22} in deg.^{2}

The correlation decays almost monotonically, with a time constant $\tau \simeq 0.2$ psec. The rotation of the phenyl ring is evidently strongly coupled to collective motions of the surrounding protein. Inertial motions of the ring are largely damped out, though the detailed physical mechanism of this damping has not been determined. The Fourier transform of the χ^2_{22} time correlation function, shown in Fig. 9, is dominated by a single large peak at zero frequency.

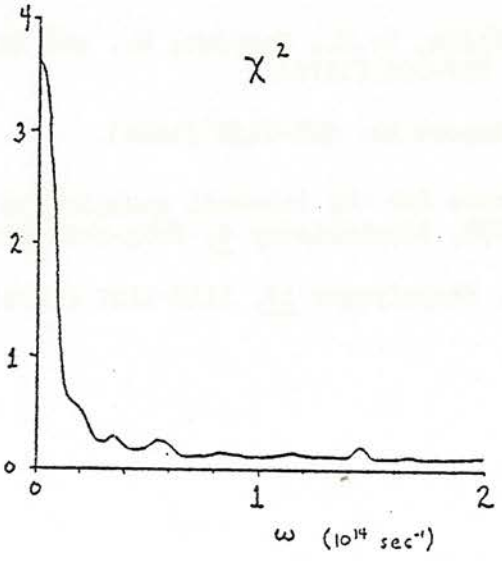


Fig. 9

Fourier Transform of the time correlation function for fluctuations of the sidechain dihedral angle χ^2_{22}, in arbitrary units.

It is reasonable to suppose that torsional motions which involve substantial displacements of large groups inside proteins will generally have this kind of collective character, while torsional motions requiring small "free volumes" (e.g., rotation of small groups) may retain some of their local inertial character.

ACKNOWLEDGEMENT

Dr. Bruce Gelin and Professor Martin Karplus have collaborated in this work, and Professors Peter Wolynes and John Deutch have contributed useful suggestions. J.A.M. has been supported by an NSF postdoctoral fellowship.

References

1. Levitt, M., J. Mol. Biol. 82, 393-420 (1974).

2. Gelin, B. R., and Karplus, M., Proc. Natl. Acad. Sci. USA 72, 2002-2006 (1975)

3. Honig, B., Ray, A., and Levinthal, C., Proc. Natl. Acad. Sci. USA 73, 1974-1978 (1976).

4. McCammon, J. A., Gelin, B. R., Karplus, M., and Wolynes, P. G., Nature 262, 325-326 (1976).

5. Gear, C. W., ANL Report No. ANL-7126 (1966).

6. Rules of nomenclature for the internal coordinates of proteins are given in: IUPAC-IUB, Biochemistry 9, 3471-3479 (1970).

7. Gō, M. and Gō, N., Biopolymers 15, 1119-1127 (1976).

II.4

STUDY OF WATER DYNAMICS IN PTI SINGLE CRYSTALS

J.Hermans[1]
A.Rahman[2]

[1]Department of Biochemistry, University of N.Carolina, Chapel Hill NC 27514 (USA)

[2]Argonne National Laboratory, Solid State Physics Division, Argonne, Illinois 60439 (USA).

Crystal Structure of PTI

Pancreatic typsin inhibitor is a 58 amino acid residue protein. Its crystal structure has been determined by a German group working under the direction of Huber.[1] Subsequent to the initial structure determination, the structure has been refined and 47 preferred locations of solvent (water) molecules have been determined. Of these, four molecules are inside the folded protein molecule, the remainder are held in certain preferred positions by hydrogen bonds to polar groups on the surface of one, or several PTI molecules. The crystallographic R-factor at a resolution of 1.5 Å is 0.23.

The crystal structure has symmetry $P2_12_12_1$ and contains four protein molecules per unit cell of dimension $43.1 \times 22.9 \times 48.6$ Å^3. (The asymmetric unit contains a single PTI molecule.) On the basis of a typical partial specific volume of 0.74 ml/g, we calculate that there are some 500 water molecules in the unit cell, i.e., about 125 per PTI molecule. There are therefore approximately 80 water molecules in the asymmetric unit which are presumably more or less free to move in the interstices between the protein molecules.

Objectives

We have begun a simulation of the water in the PTI crystal using molecular dynamics calculations. The success of these calculations is to be judged by the following criteria:

a. The time average distribution of the waters in the model should show maxima near the positions indicated to be preferred by the crystallographic work.

b. A smooth distribution should obtain in the interstices.

c. The crystallographic R-factor should be calculated on the basis of the refined protein coordinates together with the distribution of the water molecules obtained with molecular dynamics. This value should be low.

Also, the intensities of the low-order reflections should be well reproduced. These are at present left out in the calculation of Fourier sums and R-factors in protein structure work. This is so because the low order structure factors contain much more important contributions from the disordered solvent than do the high order ones.

d. The correct prediction of the dynamics (rotational and translational diffusion) of water molecules in crystals as judged by relevant spectroscopic data (e.g. nmr).

Description of the Molecular Dynamics System

The positions of all polar hydrogen atoms were estimated from the published nonhydrogen coordinates. The coordinates of all symmetry related copies of the atoms lying within the unit cell were calculated by applying the appropriate symmetry operations and lattice translations to the coordinates of the nonhydrogen atoms and these hydrogens. The apolar (i.e., carbon bound) hydrogens were not considered in this initial calculation. (The oxygen atoms of the four water molecules enclosed within the PTI molecule were also subjected to this process.) This procedure led to a total of 4×579 = 2316 atomic positions in the unit cell.

In this preliminary study the protein atoms were held in fixed positions and each atom (hydrogen included) was allotted a charge Q and Lennard-Jones parameters σ, ε; since the ST2 model for water also is a "σ, ε, Q" type model we have a very simple situation as far as the potential of interaction is concerned. For atom A of the PTI we will call the distance $(\sigma_{ST2}^{H_2O} + \sigma_{PTI}^{A})/2$ "the van der Waals distance" for H_2O and A.

Initial positions of 508 solvent water molecules were estimated by superimposing the contents of the unit cell on a volume of "thermal equilibrium" water, that is an instanteous configuration occurring in the molecular dynamics study of liquid water done by Rahman and Stillinger; their 216 molecule cell was extended by lattice translations to fill the unit cell volume. Water molecules whose oxygen atoms were farther than a certain distance from the centers of all PTI atoms were considered to be solvent molecules. The critical distance was first set at the van der Waals distance and subsequently decreased until 508 molecules were included in the set to be retained as solvent. This happened when the distance was a fraction 0.69 of the van der Waals distance.

As mentioned above, the water was treated according to the ST2 model and allowed to move. Apart from the water-water interactions according to the ST2 model, the protein molecules provided a set of nonmoving repulsive and attractive centers that constrained the motion of the water molecules. The calculation was run for 4500 cycles, each cycle corresponding to an interval in real time of 5×10^{-16} sec. During the first 2000 cycles the kinetic energy was adjusted to maintain a temperature of 300°K, and the results of these cycles were not retained for analysis. The data for the subsequent 2500 cycles covering a time span of nearly 10^{-12} sec. was analyzed.

After the workshop the work was continued in the U.S. and upto date a fairly long run of 10,000 more steps has been made. The description below includes all the pertinent results and not just those obtained during the workshop.

Analysis of Results Obtained So Far.

In Figure 1 we show maps of 1.5 Å thick layers cut out of the unit cell. The behavior of the water molecules in these cross-sections is typical of that throughout the crystal. A number of points are worth mentioning.

1. The frame at z=.26 contains a dynamic water molecule inside a protein molecule. This is one of some ten misplaced waters. These tend to remain in a very small volume and so far have not interfered with the calculation. These molecules will be omitted in subsequent calculations.

2. The behavior of the water in the four equivalent environments is not (yet) the same, i.e. the behavior in each slice (still) depends on the original placement of the water molecules vis-à-vis the protein molecules.

3. Some of the crystal water positions have a nearby dynamic water molecule in nearly all configurations in all four slices, but most crystal water positions are highly occupied by dynamic water in fewer than four of the slices, and at some crystal water positions the occupancy is zero in all four symmetry related equivalent positions.

4. The occupancy at the crystal water positions is not very high and, worse, is not increasing perceptibly as the calculation proceeds. In 1.5×1.5×1.5 $Å^3$ cubic volumes containing the crystal water position, the average dynamic water occupancy is 0.19. (In liquid water the average occupancy of this volume is 0.11.)

5. Dynamic water molecules did move into areas near the protein molecule where they were not found at the start of the calculation.

6. The distribution of the water in each frame appears to possess a good deal of structure. This structure is not of a kind a crystallographer would easily determine from an electron density map. There is therefore reason to believe that the solvent structure makes a significant contribution to the high order reflexions, and that inclusion of the complete water structure in the calculation the crystallographic R-factor will lead to considerable improvement. (For the moment, the calculated water distribution does not appear to be an equilibrium distribution, and calculation of an R-factor is not warranted.)

Further work

This project is being continued at Argonne Natl. Lab and the Univ. of North Carolina.

Acknowledgement

Work performed in part under the auspices of the U. S. Energy Research and Development Administration, and in part under a grant from the National Science Foundation.

References

Deisenhofer & Steigemann, 2nd Intl. Conference Proteinase Inhibitors, Springer Yerlag, 1974; Huber, Kukla, Epp & Formanek, Naturwiss. 57, 389, 1970; Huber, Kukla, Rühlmann & Steigemann, Cold Spring Harbor Symp. Quant. Biol. 36, 141, 1971.

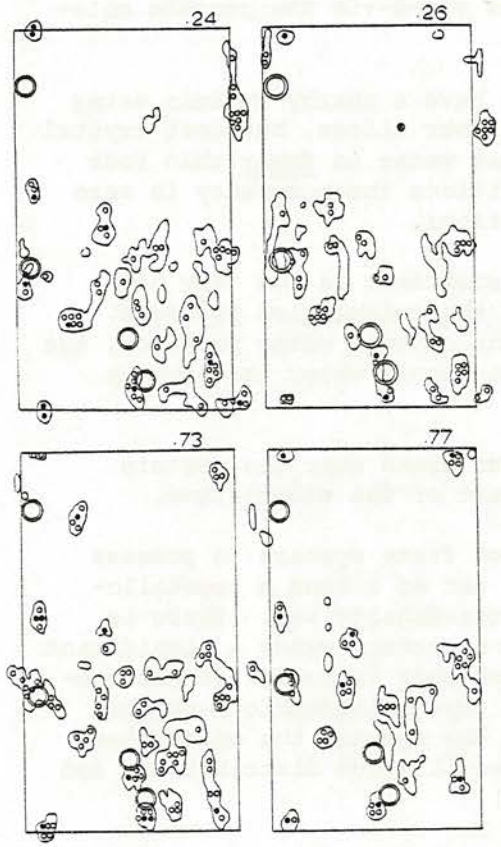


Figure 1.

Dynamic and crystal water positions in four slices of the unit cell. The z-direction is perpendicular to the plane of the slices and each slice is 1.5 Å thick; each frame spans horizontally half the unit cell in the x-direction (21.55 Å) and vertically the full unit cell in the y-direction (22.9 Å, the horizontal and vertical scales are different). The slices are equivalent, symmetry related portions of the unit cell and they have been so oriented that the protein molecules occupy the same positions relative to each frame. The area occupied by the protein is blank. Four large doubly outlined circles indicate the four crystal water position in each slice. The irregular outlines enclose all cells of 0.5×0.5×0.5 Å in which a water molecule was found in one of 214 configurations recorded after each successive 30 steps of the molecular dynamics simulation. The open circles indicate cells where a water molecule was found in 10 or more configurations, the filled circles cells where a water molecule was found more than 35 times.

III

APPROXIMATE METHODS ON STRUCTURE AND DYNAMICS OF BIO-MACROMOLECULES

III.1

THEORETICAL STUDIES OF THE DIMENSIONS OF ADRENOCORTICOTROPIC HORMONE

M.Leclerc
A.Englert

Université Libre de Bruxelles, Chimie Générale I,
Av. F. Roosevelt 50, 1050 Bruxelles (Belgique).

The long-range nonradiative electronic excitation energy transfer between luminophores attached to the two ends of a molecule depends on the molecular dimensions (Förster, 1951). From the experimental data on energy transfer, molecular dimensions and conformations can be obtained by the methods of theoretical conformational analysis of the systems considered (Leclerc et al, 1977a).

Of particular interest are biological molecules since the biological activity depends on conformation. The adrenocorticotropic hormone (ACTH), a biologically active molecule, has been extensively studied in solution by energy transfer. The efficiencies of energy transfer from Tyr^{2} and Tyr^{23} to Trp^{9} (Eisinger, 1969) in the fully active fragment ACTH (1-24)* and from Trp^{9} to N^{ε}-dansyllysine^{21} in the same fragments but dansylated (Schiller, 1972 ; Schiller and Schwyzer, 1973) have been determined in

---

*Sequence of amino-acids in ACTH (1-24):

```
          (2)                                   (9)
(H)-Ser-TYR-Ser-Met-Glu-His-Phe-Arg-TRP-Gly-Lys-Pro
          (23)      (21)                          |
(OH)-Pro-TYR-Val-Lys-Val-Pro-Arg-Arg-Lys-Lys-Gly-Val
                      |
                    (DNS)
```

various solvents, in particular in water.
These latter studies as well as the dialysis studies of Craig et al. (1965) indicate that the behavior of ACTH in solution "is best explained by some type of mobile equilibrium involving many different shapes". In these conditions, it seems appropriate to use a statistical approach to describe the various conformations in order to reproduce the experimental efficiencies of energy transfer and to obtain molecular dimensions consistent with these data.

The results of such a study are reported here. A set of chains have been generated by the Monte Carlo method of Prémilat and Maigret (1976) which is derived from that of Metropolis et al. (1953). Interactions between all the atoms are taken into account and the conformational energy is calculated from semiempirical potential functions.

The chain is constructed from standard geometrical parameters of amino-acids (Scheraga, 1968). The number of degrees of freedom has been reduced by replacing some of the side chains by a composite atom with a van der Waals radius of 1.75 Å (Pletnev et al., 1974). The important conformational features of amino-acids having a methylene group in their side chain are correctly simulated by this simplified model (Flory, 1969). The exact representation of a number of other amino-acids (Gly, Pro, Tyr, Trp, Phe, Lys-(Dns)) was required either by specific conformational features (Gly and Pro) or by the importance of mutual orientations of the aromatic rings for the transfer (Leclerc et al., 1977 a)

and for the attractive interactions between them (stacking) (Leclerc et al., 1977 b). The geometrical parameters of the dansyl luminophore (Fig. 1) have been obtained from X-ray data of related compounds, naphthalene(Cruickshank, 1957) and sulfanylamide (O'Connell and Maslen, 1967). It has been assumed that the S-N^{ε} bond has a double-bond character as in sulfanylamide and that it is trans with respect to the C^{ε}-N^{ε} and S-C bonds. The other features of the tridimensional structure of the dansyl group, in particular the tetragonal hybridization of N in $N(CH_3)_2$ and the values for the angle of rotation around the S-C bond, limited to $\pm$ 120°, are imposed by steric factors.

The direction of the transition moments of the luminophores (Leclerc et al., 1977 a and Fig. 1) has been determined by quantum mechanical calculations (Oth, 1977).

The effect of the solvent (water) has been introduced into the model by distinguishing between hydrophobic and hydrophilic amino-acids (Prémilat and Maigret, 1976). Hydrophobic amino-acids

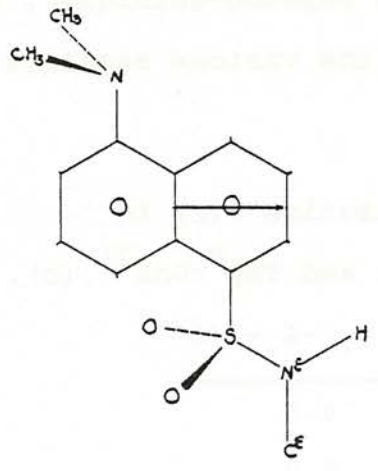


Fig. 1.

The dansyl-NH^{ε} part

of N^{ε}-dansyllysine.

have been underlined in the footnote on page 1. The attractive part of interatomic interactions between two hydrophilic or a hydrophilicand a hydrophobic amino-acid has been replaced by zero, because of the shielding effect of the surrounding water molecules. On the contrary, for two hydrophobic amino-acids the overall effect of the attractive part is predominant. Apart from this solvent effect, all the interactions have been represented by the potential functions of Scheraga (1968). Partial charges associated with the atoms contribute to the electrostatic interactions. The distribution of charges of Poland and Scheraga (1967) has been used in conjunction with a dielectric constant ε equal to 3.5, which is considered as a "mean value" of the effective dielectric constant depending on the distance between atoms. The full charges of those side chains which are ionized at neutral pH (Glu, His, Lys and Arg) have been decreased by half, because of the above-mentioned low value of ε. The charge distribution on the dansyl group has been computed by an INDO quantum-mechanical program, obtained from the "Institut de Biologie Physico-Chimique", Paris.

The calculated mean dimensions of the various segments of the chain are shown in Table I.

Table I - Various moments of the distribution f(r) in Tyr^{2}-Trp^{9},(a); Trp^{9}-Tyr^{23},(b) and Trp^{9}-Dns^{21},(c).

	$\langle r \rangle$	$\langle r^{2} \rangle^{1/2}$	$\langle r^{-6} \rangle^{-1/6}$
(a)	10.6	11.9	6.5
(b)	24.3	26.7	8.0
(c)	21.7	24.0	6.5

Only the segments 2-9, 9-23 and 9-21 (Dns) have been considered in the calculations. The distribution function f(r) of the distances r between the luminophores Trp 9 and Tyr 23 is shown in Fig. 2, together with the associated mean energies $\bar{E}(\bar{r} \pm 0.5$ Å), where $\bar{r} = i$ Å. The maxima of the distribution apparent at short distances are due to the stacking between the aromatic rings of Trp 9 and Tyr 23. It is shown that these short distances correspond to the lowest mean energies.

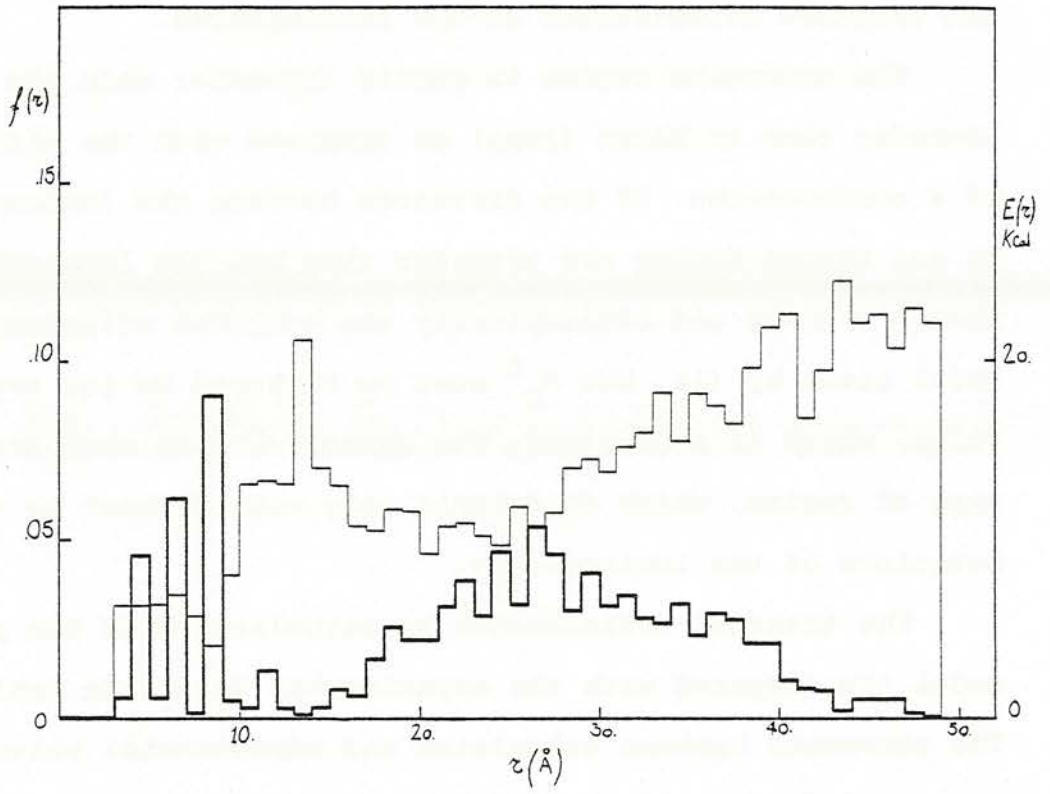


Fig. 2. The distribution function f(r) (heavy line) of the distances r between Trp^9 and Tyr^{23} and the associated mean energy E(r).

1

The transfer efficiency (Dale and Eisinger, 1974) is given for the static regime by the expression:

$$\langle T \rangle_s = \left\langle \frac{R_o^6/r^6}{1+ R_o^6/r^6} \right\rangle \qquad (1)$$

and for the dynamic limit by

$$\langle T \rangle_d = \frac{\langle R_o^6/r^6 \rangle}{1+\langle R_o^6/r^6 \rangle} \qquad (2)$$

where R_o is a parameter depending on the spectroscopic properties of the luminophores and of the solvent, and on the relative orientations of the luminophores.

The averaging regime is static (dynamic) when the transfer time is short (long) as compared with the lifetime of a conformation. If the distances between the luminophores do not change during the transfer time but the luminophores rotate rapidly and isotropically the transfer efficiency is still given by (1), but R_o^6 must be replaced by its average value, which is a constant. The symbol $\langle T \rangle_s$, is used for this type of regime, which is dynamic only with respect to the rotations of the luminophores.

The transfer efficiencies as calculated from the present model are compared with the experimental values in Table II. The agreement between calculated and experimental values is very satisfactory if the averaging regime corresponding to the experimental conditions is "static" with respect to the distances between the luminophores. Although a truly static regime is obtained only in highly viscous solvents, it can

be argued that in molecules with a considerable number of internal rotation angles the regime is always nearly static. It is indeed reasonable to assume that only a very small fraction of the total amount of molecules undergoes a conformational change which leads to a noticeable variation of interluminophore separation during the transfer time of electronic excitation energy. If, however, the regime were purely dynamic, there would be no agreement between the experimental and calculated transfer efficiencies.

Table II

Luminophores implied in the transfer*	$\langle R_o \rangle$	Experimental efficiencies $\langle T \rangle$	Calculated efficiencies $\langle T \rangle_s$	$\langle T \rangle_{s'}$	$\langle T \rangle_d$
Tyr^2-Trp^9	9.8	0.5±0.15	0.48	0.55	0.90
Tyr^{23}-Trp^9	11.0	0.15±0.10	0.18	0.20	0.84
Trp^9-Dns^{21}	19.4	0.45	0.39	0.44	0.98

The above considerations relative to the type of averaging regime raise the following question: what is the lifetime of a conformation and how does it depend on the number of monomers in an oligomer ? Of course, it could depend on the sequence as well. The answer could be given by molecular dynamics studies of the systems of interest. However, experimental

*The results are almost identical for the two transition moments of Trp (Leclerc et al, 1977a).

evidence already available supports strongly the idea that, at least in ACTH, the purely dynamic limit is never attained. As a matter of fact, the results of dialysis experiments by Craig et al (1965) imply some kind of conformational stability.

As regards the conformational calculations discussed above, it should be added that the dimensions of the segments (9-23) and (9-21) decrease considerably if all the side chains are treated as electrically neutral, because of the actual presence of six or five positive charges in these segments in neutral water. As expected there is in this case no agreement between calculated and experimental values of the efficiency in the segment (9-23) ; for the dansylated segment (9-21) the computed value of the efficiency without charges on the side chains agrees rather well with the experimental value for the protected (uncharged) peptide in water: $\langle T \rangle_{exp}=0.64$, while $\langle T \rangle_{s'\ calc.}=0.57$.

The program used in this work has been written by Prémilat, Maigret and one of us (M.L.). The calculations were performed on the IBM 370/168 computer at Orsay during the Workshop on Models for Protein Dynamics and took approximately 170 minutes for a chain consisting of 16 units.

References

Craig, L.C., Fisher, J.D. and King, T.P. (1965) Biochemistry 4, 311-318.

Cruickshank, D.W.J. (1957) Acta Cryst. 10, 504-508.

Dale, R.E. and Eisinger, J. (1974) Biopolymers 13, 1573-1605.

Eisinger, J. (1969), Biochemistry 8, 3902-3907.

Flory, P.J. (1969) Statistical Mechanics of Chain Molecules, Interscience, New York, p. 261.

Förster, Th. (1951) Fluoreszenz Organischer Verbindungen (Vandenholck and Rupprecht, Göttingen)..

Leclerc, M., Prémilat, S., Guillard, R., Renneboog-Squilbin, C., Englert, A. (1977a) Biopolymers 16, in press.

Leclerc, M., Prémilat, S. and Englert, A. (1977b), manuscript in preparation.

Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) J. Chem. Phys. 21, 1087-1092.

O'Connell, A.M. and Maslen, E.N. (1967) Acta Cryst. 22, 134-145.

Oth, J. (1977) manuscript in preparation.

Pletnev, V.Z., Popov, E.M. and Kadymova, F.A. (1974) Theoret. Chim. Acta 35, 93-96.

Poland, D. and Scheraga, H.A. (1967) Biochemistry 6, 3791-3800.

Prémilat, S. and Maigret, B. (1976) C.R. Acad. Sci. Paris, 282, série C, 225-228.

Scheraga, H.A. (1968) Adv. Phys. Org. Chem. 6, 103-184.

Schiller, P.W. (1972) Proc. Nat. Acad. Sci. USA 69, 975-979.

Schiller, P.W. and Schwyzer, R. (1973) in "Peptides", Nesvadba, H., Ed., North-Holland Publishing Company, Amsterdam.

III.2

DISTRIBUTION FUNCTIONS OF THE END-TO-END DISTANCES IN OLIGOPEPTIDES

M.Leclerc
A.Englert

Université Libre de Bruxelles, Chimie Générale I,
Av. F. Roosevelt 50, 1050 Bruxelles (Belgique).

The possibility of obtaining the distribution functions $f(r)$ of the end-to-end distances r in flexible biopolymers from experimental data on electronic excitation energy transfer by Förster mechanism (Förster, 1948) has been suggested by Cantor & Pechukas, 1971, and by Grinvald et al, 1972. The excitation energy in these experiments is transferred from the donor to the acceptor luminophore attached to the two ends of the biopolymer. The experimental data provide a set of values of a property $X(r,\alpha)$, averaged over the whole spectrum of the distances r present in solution, and depending each on a particular value of the parameter α, equal either to the spectroscopic constant R_o (Förster, 1948) either to the time t. The comparison of the experimental values of $\langle X(\alpha) \rangle$ with the values calculated from theoretical distribution functions enables one thus to select the distribution functions which are in agreement with experiment.

Haas et al, 1975, have reported experimental data of the decay of the intensity of fluorescence with time, $I(t)$, in oligopeptides with N^5-(2-Hydroxyethyl)-L-glutamine (G) as the repeating unit and the luminophores dansyl (Dns) and naphthyl (Nph) attached to the two ends of the molecules.

The number of repeating units was 4,5,6,7,8 and 9. In these experiments the lifetime of a conformation does not vary during the transfer time (about 10^{-9} sec) since the measurements were performed in a viscous solvent. The data were analysed in terms of four different analytical distribution functions f(r) assuming a fast and isotropic movement of the transition dipole moments of the luminophores.

In order to avoid the above assumption and to take the correlation between the distances r and the orientations of the transition dipole moments (Dale and Eisinger, 1976) into account, we have started a study of a theoretical model representing the oligopeptides $Dns(G)_nNph$ with the aim of reproducing the experimental data. The theoretical distribution functions and a description of the conformations of the oligopeptides are derived from the model by the methods of theoretical conformational analysis (Leclerc et al, 1977). The calculations are based on semi-empirical potential functions commonly described in the literature (Scheraga, 1968). The representative set of chains are generated by a Monte Carlo method as described by Prémilat and Maigret, 1976. Long-range interactions are taken into account in the calculations.

The amino-acid residues are constructed according to standard geometrical parameters (Scheraga, 1968) but the side chain is represented by a composite atom with a van der Waals radius of 1.75 Å (Pletnev et al, 1974). The geometrical

parameters of the dansyl have been obtained as described in the companion report (Leclerc and Englert, 1976) and those of naphthyl from the relevant X-ray data (Cruishank, 1957).

The interactions with the polar solvent have been taken into consideration in a simplified manner through the interaction potential between the atoms of the chain, following Prémilat and Maigret, 1976. It has been assumed that the presence of solvent cancels the attractive part of the potential for hydrophilic groups. The repeating unit G is considered as hydrophilic while the aromatic rings of the luminophores are hydrophobic. The attractive interactions between the aromatic rings (stacking) tend to decrease the chain dimensions. The electrostatic interactions are represented through partial charges associated with the atoms (Poland and Scheraga, 1967), those on the dansyl and naphthyl have been calculated by an INDO quantum-mechanical program obtained from the "Institut de Biologie Physico-Chimique", Paris. The dielectric constant was taken to be 3.5 (Scheraga, 1968). The transition dipole moments of the luminophores have been obtained by quantum mechanical calculations (Oth, 1977). We have performed calculations on chains with n equal to 4 and 8. All the figures show the results for the latter oligopeptide.

The dimensions of the chain $Dns(G)_4Nph$ and $Dns(G)_8Nph$ as represented by its various moments are shown in Table I.

Table I

Mean dimensions of the oligopeptides $Dns(G)_nNph$

n	$\langle r \rangle$	$\langle r^2 \rangle^{1/2}$
4	15.1	16.6
8	24.1	25.9

The distribution function f(r) is shown in Fig.1, together with the associated average values of the conformational energy of the chain.

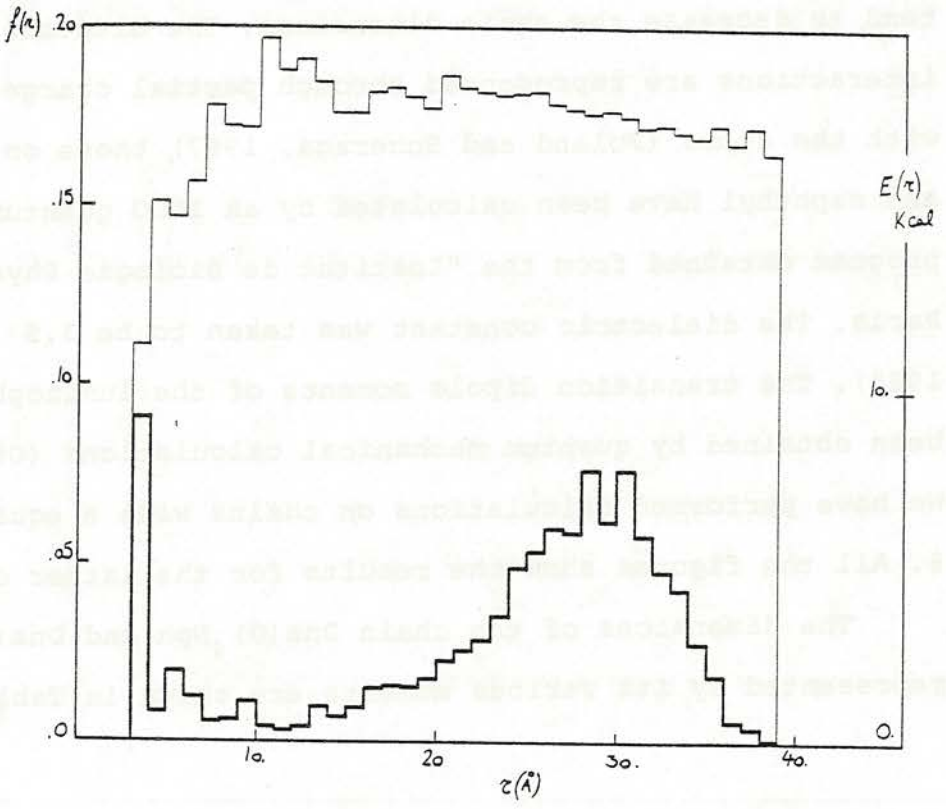


Fig. 1. The distribution function f(r) (heavy line) of the distances r between luminophores in $Dns(G)_8Nph$ and the associated mean energies E(r).

The interesting feature of these distributions is the presence of two maxima, destroying the symmetry. The maximum at the short distance arises from the stacking between the aromatic rings of the luminophores.

The orientation of the transition moments of the luminophores is represented by the orientation factor K^2, defined in the literature (see for ex. Dale and Eisinger, 1974). The distribution of K^2 shown in Fig.2 does not differ significantly from that corresponding to a hypothetical completely random orientation in space (Guillard and Englert, 1976).

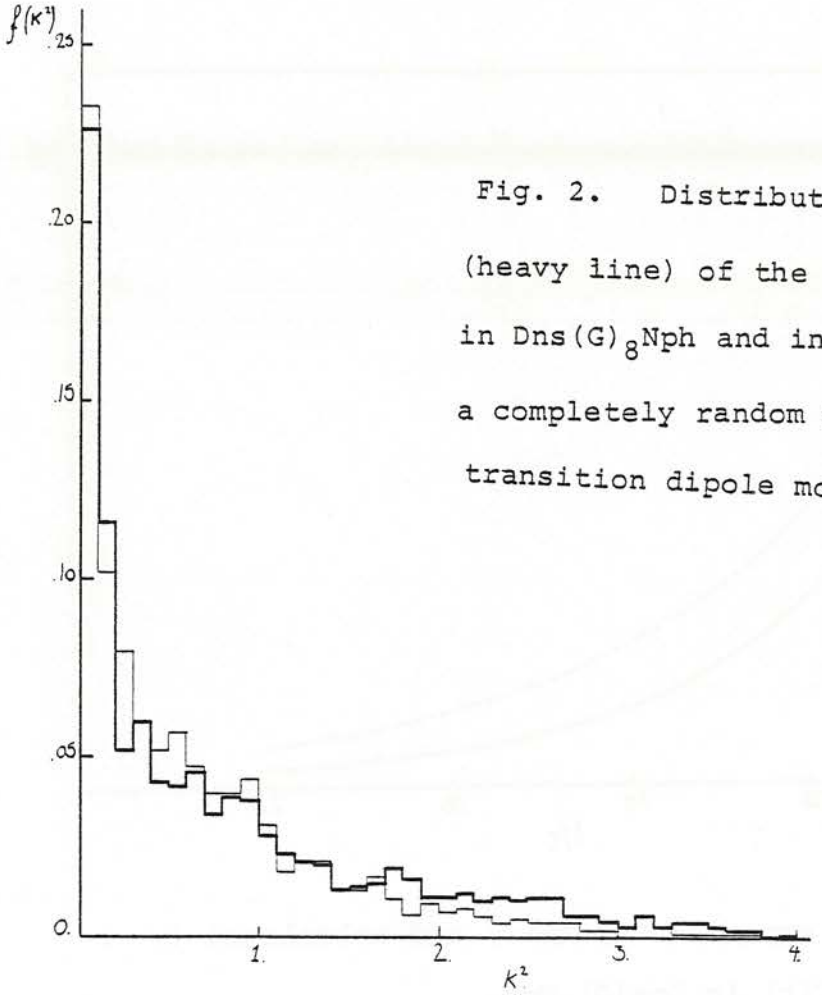


Fig. 2. Distribution function $f(K^2)$ (heavy line) of the orientation factor K^2 in $Dns(G)_8Nph$ and in a hypothetical case of a completely random orientation in space of transition dipole moments.

The experimental and theoretical decay curves, I(t), given by the expression (Grinvald et al, 1972)

$$I(t) = I(o) \int_o^\infty \int_o^4 f(K^2,r) \exp\left[-\frac{t}{\tau} - \frac{t}{\tau}\left(\frac{So}{r}\right)^6 K^2\right] dK^2\, dr \qquad (1)$$

where $S_o^6 K^2$ is equal to the spectroscopic factor R_o^6, are shown in Fig.3. The spectroscopic factor R_o was taken equal to 23 Å, a value corresponding to the experimental conditions of Haas et al, 1975, for $\langle K^2 \rangle$ equal to 0.667.

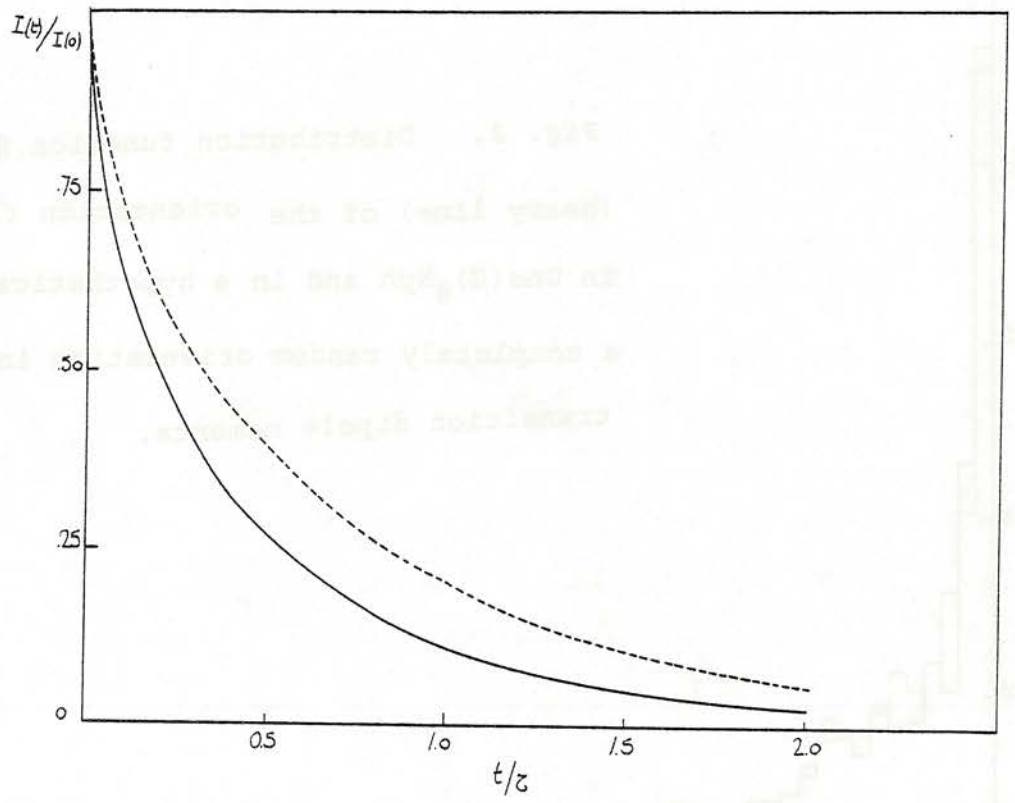


Fig. 3. The experimental (——) and calculated (---) decay curves I(t) in $Dns(G)_8Nph$.

The agreement between the experimental and calculated curves I(t) is not satisfactory. The mean dimensions obtained here are also larger than those calculated by Haas et al, 1975.

We are trying at present to modify the interatomic potential between the side chains in order to get a better agreement between the model and the experimental. It is also essential to reduce the computer time required for these calculations (85 minutes for the chain Dns$(G)_8$Nph on the IBM 370/168 at Orsay). The program was written in collaboration with Prémilat and Maigret.

References

Cantor, R.C. and Pechukas, P. (1971) Proc. Natl. Acad. Sci. USA 68, 2099-2101.

Cruishank, D.W.J. (1957) Acta Cryst. 10, 504-508.

Dale, R.E. and Eisinger, J. (1974) Biopolymers 13, 1573-1605.

Dale, R.E. and Eisinger, J. (1976) Proc. Nat. Acad. Sci. USA 73, 271-273.

Förster, Th. (1948) Ann. Phys. (Leipzig) 2, 55-75.

Grinvald, A., Haas, E. and Steinberg, I.Z. (1972) Proc. Natl. Acad. Sci. USA 69, 2273-2277.

Guillard, R. and Englert, A. (1976) Biopolymers 15, 1301-1314.

Haas, E., Wilchek, M., Katchalsky-Katzir, E. and Steinberg, I.Z. (1975) Proc. Natl. Acad. Sci. USA 72, 1807-1811.

Leclerc, M. and Englert, A. (1976) "Theoretical studies of the dimensions of the adrenocorticotropine hormone", Annual Report of CECAM.

Leclerc, M., Prémilat, S., Guillard, R., Renneboog-Squilbin,C. and Englert, A. (1977), to appear in Biopolymers.

Pletnev, V.Z., Popov, E.M., Kodymova, F.A. (1974) Theoret. Chim. Acta 35, 93-96.

Poland, D. and Scheraga, H.A. (1967) Biochemistry, 6, 3791-3800.

Prémilat, S. and Maigret, B. (1976) C.R. Acad. Sci. Paris, t 282, Série C, 225-228.

Scheraga, H.A. (1968) Adv. Phys. Org. Chem. 6, 103-184.

III.3

FLUCTUATIONS OF PARTLY HELICAL CHAINS

M.Leclerc
A.Englert

Université Libre de Bruxelles, Chimie Générale I,
Av. F. Roosevelt 50, 1050 Bruxelles (Belgique)

The helix-coil transitions in polypeptides have been studied theoretically mostly by the methods of statistical mechanics, in terms of the one-dimensional Ising model (for a review, see Poland and Scheraga, 1970). However some of the equilibrium properties related to the description of the transition on the molecular level are still being investigated. The present report is concerned with the fluctuations of the torsion angles ϕ and ψ in a chain composed of both α-helical and random-coil sections, in particular at the junction between the two sections. In previous calculations the distribution of torsion angles in helical chains (Skvortsov et al., 1971; Skvortsov et al., 1972; Birshtein et al., 1976; Gō and Gō , 1976) and in randomly coiled chains (Prémilat and Hermans, 1973 and Prémilat and Maigret, 1976) has been derived considering each of them separately. A model of a partly helical chain should also be useful for the interpretation of the frequency of occurrenceof various amino-acids at the helical ends as observed in proteins (Chou and Fasman, 1974) and for the interpretation of spectroscopic observations, such as the broadening of the infra-amide bands (Chirgadre, 1976).

Some preliminary calculations have been performed at the workshop on "Models for Protein Dynamics". Only the randomly coiled section at the junction with a rigid helical section has been considered so far. The coordinates of the atoms have been obtained from standard geometrical parameters

of polypeptides (Scheraga, 1968). The conformational energy, assumed to be equal to the sum of pairwise interactions between all the atoms, has been computed from semiempirical potential functions of Scheraga, 1968. The model has therefore the important feature of including long-range interactions. The computations are performed on a set of N chains, generated by a Monte Carlo method of Metropolis et al., 1953, adapted to problems of chain configurations by Prémilat and Maigret, 1976. The solvent (water) is not considered explicitly, but is included in the model through the potential functions. The van der Waals attractions, dominant over the repulsions, lead the chain to adopt compact structures and thus minimise the contact between the hydrophobic chain and the polar solvent (Prémilat and Maigret, 1976). The dielectric constant has been taken equal to 3.5 (Scheraga, 1968) and the hydrogen bond is described by the potential of Poland and Scheraga, 1967. The side chain of the amino-acid residue is represented by a single composite atom with a van der Waals radius of 1.75 Å. The parameters for the potential functions of this atom have been given by Pletnev et al., 1974. This atom reproduces correctly the most important conformational features of any of the amino-acids with a methylene group in the side chain, but we shall refer to the model as the poly-L-alanine model.

The torsion angles ϕ,ψ accessible to each amino-acid residue are chosen among those of an independent peptide unit of L-alanine by varying each of them in steps of 20° starting with the values of $\phi_\alpha^\circ, \psi_\alpha^\circ$ (-48°, -57°, Knaell and Scott, 1971) of a perfectly ordered α-helix and neglecting the states with energies higher than 3,5 kcal, above the absolute minimum.

The average conformational energy map of a residue in a completely random chain can be assimilated to the "matrix of success" representing the number of residues among the

total in the N generated chains having a particular value of the set of (ϕ,ψ) torsion angles. In our calculations N was equal to 10 000 and each chain was composed of 11 residues. The matrix of success for a randomly coiled poly-L-alanine is shown in the Appendix. It is of interest to compare the frequencies of occurrence near the α-helical states ($\phi_\alpha^\circ \pm 20°$, $\psi_\alpha^\circ \pm 20°$) in this map and in the dipeptide map. While in the latter this frequency is extremely small (0,8%), in the former it amounts to 12%.

The observed increase in the frequency is not the result of hydrogen-bond formation but is mainly due to hydrophobic interactions.

The partly helical chains were composed of six amino-acid residues in a fixed perfectly ordered α-helical state and of a random-coil part, adjoining the C-terminal end of the helix, composed of four or five residues.

The percentage of α-helical states is greatly enhanced in the random-coil section adjoining the α-helix (Table I). The reason for this appears from the comparison of the matrices of success for the residues 6 to 11, shown in the Appendix. In those representing the state of residues 7 and 8 adjoining the last helical residue 6, a great number of states in the left upper part of the matrix have zero frequences, apparently because of steric conflicts with atoms in the helical section.

Relaxing the constraintin the $\phi_\alpha^\circ, \psi_\alpha^\circ$ angles in the helical section may change qualitatively the values obtained. However, the observed alteration of the conformation at the random chain in the vicinity of the helix can be important for the dynamics of the helical growth and should influence

Table I

The frequencies of occurrence of α helical states in randomly coiled sections adjoining the C terminal end of the helix computed from the chain $(Ala)_6$helical-$(Ala)_n$ random.

	total	res. 7	res. 8	res. 9	res. 10	res. 11
			n=4			
% α	9.6	21.	9.	8.	0.2	
% α±20°	26.2	38.	36.	14.	5.	
			n=5			
% α	10.7	34.	5.	11.	2.	1.
% α±20°	25.1	54.	25.	26.	15.	6.

the value of the Zimm and Bragg cooperativity factor σ for helix initiation (Zimm and Bragg, 1959).

The program was written in collaboration with Prémilat and Maigret. The calculations took 36 minutes for a partly helical chain composed of 11 Ala residues, on the IBM 370/168 at Orsay.

References

Birshtein, T.M. and Skvortsov, A.M. (1976) Biopolymers 15, 1061-1080.

Chirgadze, Yü.N., Brazhnikov, E.V. and Nevskaya, N.A. (1976) J. Mol. Biol. 102, 781-792.

Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 211-222.

Gō, M. and Gō, N. (1976) Biopolymers 15, 1119-1127.

Knaell, K.K. and Scott, R.A. (1971) J. Chem. Phys. 54, 566-575.

Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) J. Chem. Phys. 21, 1087-1092.

Pletnev, V.Z., Popov, E.M., Kodymova, F.A. (1974) Theoret. Chim. Acta 35, 93-96.

Poland, D. and Scheraga, H.A. (1967) Biochemistry 6, 3791-3800.

Poland, D. and Scheraga, H.A. (1970) "Theory of Helix-Coil Transitions in Biopolymers", Academic Press, New York and London.

Prémilat, S. and Hermans, J., Jr. (1973) J. Chem. Phys. 59, 2602-2612.

Prémilat, S. and Maigret, B. (1976) C.R. Acad. Sc. Paris, t 282, Série C, 225-228.

Scheraga, H.A. (1968) Adv. Phys. Org. Chem. 6, 103-184.

Skvortsov, A.M., Birshtein, T.M. and Zolinski, A.O. (1971) Mol. Biol. 5, 69-77.

Skvortsov, A.M., Birshtein, T.M. and Aleksanyan (1972) Mol. Biol. 6, 394-399.

Zimm, B.H. and Bragg, J.K. (1959) J. Chem. Phys. 31, 526-535.

APPENDIX

Matrix of success of a randomly coiled chain

4384	3798	2763	1025	583	2448	604	0	0	0	0	0	0	0	0	0	0	0
3402	2036	1935	696	1267	1583	1085	43	0	0	0	0	0	0	0	0	0	0
613	1704	811	735	1689	4704	787	1161	0	0	0	0	0	0	0	0	0	0
652	717	692	546	3198	4919	2070	161	0	0	0	0	0	0	0	0	0	0
370	585	303	415	3315	2565	0	0	0	0	0	1644	46	0	0	0	0	0
208	1286	539	1913	906	0	0	0	0	0	0	3200	56	0	0	0	0	0
0	155	508	1403	1090	0	0	0	0	0	0	3631	1574	0	0	0	0	0
0	0	111	865	127	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	462	524	130	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	2334	3817	3982	92	0	0	0	0	0	0	0	0	0	0
4311	1834	427	801	996	2749	2523	206	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1373	1430	531	680	742	225	1	0	0	0	0	0	204	0	0	0	0	0

Origin at $\phi = -168°$, $\psi = -177°$; $\Delta\phi = \Delta\psi = 20°$.

```
Matrix of residue 7

    0     0     0     0     0     0    10     0     0     0     0     0     0     0     0     0     0     0
   58     0     0     0     0    10    12     1     0     0     0     0     0     0     0     0     0     0
  989  1128     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
  360   415    40     0     0    68    16    11     0     0     0     0     0     0     0     0     0     0
    0     0    22    23     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0   101    45     0     0     0     0     0     0    18     0     0     0     0     0     0
    0     0     0     2   205     0     0     0     0     0     0     0     5     0     0     0     0     0
    0     0     0     0     4     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0  1025     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0  1848     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0  3407   170     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     7     0     0     0     0     0

Matrix of residue 8

   49    10     0     0     3     0     0     0     0     0     0     0     0     0     0     0     0     0
   84   128    15     0    57    16     0     0     0     0     0     0     0     0     0     0     0     0
   12   404    30    37    14    17    15     0     0     0     0     0     0     0     0     0     0     0
  169    30     7     0    69    96   916     7     0     0     0     0     0     0     0     0     0     0
   17   100     0     0   105    40     0     0     0     0     0     0     0     0     0     0     0     0
    0     0    23    22  1793     0     0     0     0     0     0     0   865     0     0     0     0     0
    0     1     0   686   206     0     0     0     0     0     0   101     0     0     0     0     0     0
    0     0    12   894     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0   231     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     5  1903     0     0     0     0     0     0     0     0     0     0     0
  185     0    12     1     0    33   527     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    0     2     5     2     2     0     0     0     0     0     0     0     0     0     0     0     0     0
```

Matrix of residue 9

```
  17   25    4    0   38    7    0    0    0    0    0    0    0    0    0    0    0    0
  39    0   49   15   37   49   32   12    0    0    0    0    0    0    0    0    0    0
   4    0    2    0   28  246   40    0    0    0    0    0    0    0    0    0    0    0
   0    0   10   19  335  627   23    0    0    0    0    0    0    0    0    0    0    0
   9    0    7   20  164   10    0    0    0    0    0  718  103    0    0    0    0    0
   0   34   13   51  260    0    0    0    0    0    0   27    8    0    0    0    0    0
   0   64   16   22   35    0    0    0    0    0    0   20   80    0    0    0    0    0
   0    0    4    0   10    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0  479    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0   56  684   39    8    0    0    0    0    0    0    0    0    0    0
1554 1879    0   50   33  145 1140  148    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  47    0    3    1    0    1    0    0    0    0    0    0    0    0    0    0    0    0
```

Matrix of residue 10

```
 257  160   44    0  150   41    0    0    0    0    0    0    0    0    0    0    0    0
 118   63  134  208   55   54   12    0    0    0    0    0    0    0    0    0    0    0
  63  129   83   52   86   82   21    0    0    0    0    0    0    0    0    0    0    0
 138   21   13  330  107  204   19    0    0    0    0    0    0    0    0    0    0    0
 170  135   36   52  486  216    0    0    0    0    0    4    6    0    0    0    0    0
  80   22  169   52  482    0    0    0    0    0    0   52   25    0    0    0    0    0
   0   91    3  115  110    0    0    0    0    0    0   79    9    0    0    0    0    0
   0    0   23  106    6    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0   81   21    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0  514 1189   55    0    0    0    0    0    0    0    0    0    0    0
 927  450   57   57  337   65  185    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 481    5    5    0   12    2    0    0    0    0    0    0   10    0    0    0    0    0
```

Matrix of residue 11

```
317   72   98    0   64  153   57    0    0    0    0    0    0    0    0    0    0    0
463   65  204  158  175  280  327    0    0    0    0    0    0    0    0    0    0    0
724  460  196  225  258  357  102    0    0    0    0    0    0    0    0    0    0    0
173  112   92   87  138   67  101   17    0    0    0    0    0    0    0    0    0    0
 76  201   12  101   23   64    0    0    0    0    0   12    0    0    0    0    0    0
 21   76   16   28   37    0    0    0    0    0    0   67    0    0    0    0    0    0
  0   20    8   25    5    0    0    0    0    0    0  121    0    0    0    0    0    0
  0    0   45   53  123    0    0    0    0    0    0    0    0    0    0    0    0    0
  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  0    0    0    0  253  331   99    0    0    0    0    0    0    0    0    0    0    0
  0    0    0    0   47  239   72   20    0    0    0    0    0    0    0    0    0    0
311  165   40   79  112  134   75   64    0    0    0    0    0    0    0    0    0    0
  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
  0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
224  427  138   89  142   98    0    0    0    0    0    0   30    0    0    0    0    0
```

III.4

STATISTICS AND DYNAMICS OF PROTEIN STRUCTURES

K.Nagano

Faculty of Pharmaceutical Sciences, University of Tokyo,
Hongo, Bunkyo-ku, Tokyo (Japan).

Primary Structure Translating System

What I have been working on for recent years (Nagano, 1973, 1974, 1976a, b; Nagano & Hasegawa, 1975) is to create an artificial intelligence, which translates a primary sequence of protein spelled by a string of alphabets into a tertiary structure represented by a set of atomic co-ordinates as shown schematically in the following.

PRIMARY SEQUENCE (MKIVYW....KIANI)

⇓ translation

TERTIARY FOLD 

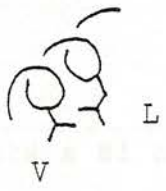
Here, an alphabet represents a type of amino acid. This is a kind of language translator. An arbitrary combination of alphabets does not necessarily correspond to a unique tertiary structure, because some peptide sequences are known to be structureless in aqueous

solution. This is quite similar to the case of human languages, in which an arbitrary string of alphabets and blanks very often does not make any sense at all. For this reason my primary structure translating system is, at present, oriented mainly for treating the sequences naturally found.

This system has been designed as a type of learning machine. It behaves itself obeying the rules of a kind of molecular dynamics, but does not assume *a priori* potential functions. I think that a good artificial intelligence must not consume a lot of computer time. Then, how can we save computer-time in estimating the interactions between atoms of amino acid residues?

Measure of Stability caused by the Short-range Interactions

In the case of the short-range interactions, the problem can be compared to a rain-umbrella relationship. When we have a thick window in a room and cannot see how much it rains outside, we just count how many umbrellas are open and directing upwards. Statistical treatment of such a number could provide us with a measure of how heavily it rains within a limited range. Similarly, the interactions between the side chains as well as the main chain of, say, V and L residues separated by m residues in a helix might be inferred by a statistical quantity representing how the region of the doublet $d \equiv (V,L;m)$ is significantly predicted as a helix, though the conformation type of the region would be perturbed by the combination of



other residues around the doublet, or under the influence of long-range interactions. Here, V and L denote valine and leucine residues, respectively, in a conventional one-letter-code notation.

When the total number of residue pairs separated by m residues, $N_{(m)}$, the total number of such pairs found in a helix, N_{m1}, the total number of a doublet (residue pair) d, N_d, the total number of d found in a helix, n_{d1}, are given, the over-all proportion, q_{m1}, and the local proportion, q_{d1}, are obtained as

$$q_{m1} = N_{m1}/N_{(m)} \tag{1}$$

and

$$q_{d1} = n_{d1}/N_d \ . \tag{2}$$

Those proportions are considered to be some kinds of probabilities. If the difference between q_{d1} and q_{m1} is very large, a region of d on a primary sequence will be predicted as a helix. The distribution of q_{d1} is binomial and is known to approach to the gausian type when $N_{(m)}$ increases. Its standard deviation can be given by

$$\sigma_{m1} = \sqrt{q_{m1}(1 - q_{m1})}. \tag{3}$$

Accordingly, a statistical quantity

$$z_{d1} = (q_{d1} - q_{m1})/\sqrt{(q_{m1} - q_{m1}^2)} \tag{4}$$

is a normalized measure of the statistical significance of the prediction when N_d and n_{d1} are large enough, and when q_{m1} is neither very close to 0 nor to 1.

If N_d is not very large, the following quantity

$$a_{d1} = (q_{d1} - q_{m1})/\sqrt{\frac{(q_{m1} - q_{m1}^2)}{N_d}} \tag{5}$$

is used to test the over-all statistical significance of a hypothesis in the χ^2-test. However, the quantity a_{d1} does not converge to a certain value when $N_{(m)}$ increases. Accordingly,

$$b_{d1} = \{N_d/N_{(m)}\}^{1/2}(q_{d1} - q_{m1})(q_{m1} - q_{m1}^2)^{-1/2}$$
$$= \{N_d/N_{(m)}\}^{1/2} z_{d1} \quad (6)$$

was used as a measure of the statistical significance of the prediction by Nagano (1973), but, recently,

$$y_{d1} = \{N_d/N_{(m)}\}^{1/4} z_{d1} \quad (7)$$

has been adopted to reduce the difference in the statistical weights $\{N_{d1}/N_{(m)}\}^{1/2}$ of b_{d1} between doublets composed of Ile, Met, Phe *etc.* and those composed of Ala, Gly, Leu, Lys, Ser and Val, because the value of $N_{(m)}$ has exceeded 6500. Besides, a similar quantity y_{t1} has been introduced due to triplet information, which represents the helical wheel effect (Schiffer & Edmundson, 1967). The doublet analyses are also applied to the prediction of loops (or β-turns) and β-structures (Nagano, 1973, 1976b).

The most important assumption is, in this context, that the statistical quantity y_{d1} is proportional to a thermodynamical quantity, though it might not be energy or free energy itself, representing how much the helical conformation of the region of d can be stabilized by the interactions of the side-chains of the residue pair as well as of the main chain. The stability of a conformation type of the region around a residue i of pth protein due to the short-range interactions can be approximately estimated as a linear combination of many conceivable interactions in the short-range, and expressed as a prediction function in the following form.

$$g_1(p,i) = -\sum_{m=0}^{4} c_{m1}\left(y_{d1} + y_{d'1} + x_{m1}\sum_{d''}^{all} y_{d''1}\right)$$

$$- \sum_{m,m'=0,2 \text{ and } 2,0} c_{mm'1} \sum_{t'' \text{ all}} y_{t1} \qquad (8)$$

The coefficients, c_{m1}, x_{m1} and c_{021} $(= c_{201})$, were adjusted so as to get the success of the helix prediction as good as possible. Similar calculations were also made for loops and β-structure. The values of the coefficients are presented in Nagano (1976b). The percentages of residues correct for the predictions of helices, loops and β-structures are 78.9%, 70.7% and 80.8%, respectively, with a suitable set of threshold values (Nagano, 1976b).

Angle Probability Calculation Applied to the Monte-Carlo Simulation

The helix, β-structure and loop prediction functions, $g_1(p,i)$, $g_1^{**}(p,i)$ and $g_1^{*}(p,i)$, are also used to calculate the dihedral angle probability in the Monte-Carlo simulation of polypeptide conformation made by Drs. S. Premilat and B. Maigret in the present workshop. The respective probabilities are obtained as follows.

$$P_\alpha = e^\alpha/(e^\alpha + e^\beta + e^\gamma) \quad \text{for helical angle} \qquad (9)$$

$$P_\beta = e^\beta/(e^\alpha + e^\beta + e^\gamma) \quad \text{for β-structural angle} \qquad (10)$$

$$P_\gamma = e^\gamma/(e^\alpha + e^\beta + e^\gamma) \quad \text{for β-turn angle} \qquad (11)$$

Here,

$$\alpha \equiv g_1(p,i) \quad \text{or} \quad \alpha \equiv A_1\{g_1(p,i) - A_2\} \qquad (12)$$

$$\beta \equiv g_1^{**}(p,i) \quad \text{or} \quad \beta \equiv B_1\{g_1^{**}(p,i) - B_2\} \qquad (13)$$

$$\gamma \equiv g_1^{*}(p,i) \quad \text{or} \quad \gamma \equiv C_1\{g_1^{*}(p,i) - C_2\}. \qquad (14)$$

The parameters A_k, B_k and C_k $(k = 1, 2)$, are at present assumed to be 1 for $k = 1$ and 0 for $k = 2$, but might be modified in future to enhance the success of the Monte-Carlo simulation.

In order to make the Monte-Carlo simulation more successful, the choice of starting structure as well as the introduction of long-range forces should be essential. It is found that the important long-range interactions occur among the regions having both helical and β-structural potentials, which are defined by the negative values of the respective prediction functions (Nagano, 1974; Nagano & Hasegawa, 1975).

Prediction of Super-secondary Structure

As the extention of the secondary structure prediction, the super-secondary structures of many proteins of known sequence seem also predictable (Nagano, 1976a). Super-secondary structures are typical building blocks, such as mono-nucleotide binding folds, β-rich structures and helix-rich structures. It is also noticed that a β-strand-α-helix-β-strand unit (βαβ-unit) of a mono-nucleotide binding fold has a preference of a right-handed sense. Fig. 1 summarizes the findings of Nagano (1976a), which is based on the statistical predictabilities. The circles represent β-strands, while the hexagons represent helices flanking the β-sheets. The numbers written in circles represent β-structural strength orders of the β-strands, while those in hexagons helical strength orders of the helices. The strength orders are defined by a couple of numbers such as $_{21}^{3}$ to show the ranking of the potential (the sum of the values of the prediction function concerned over 5 residues around a central residue) at a region by the upper number over a whole range of the polypeptide chain and by the lower number over

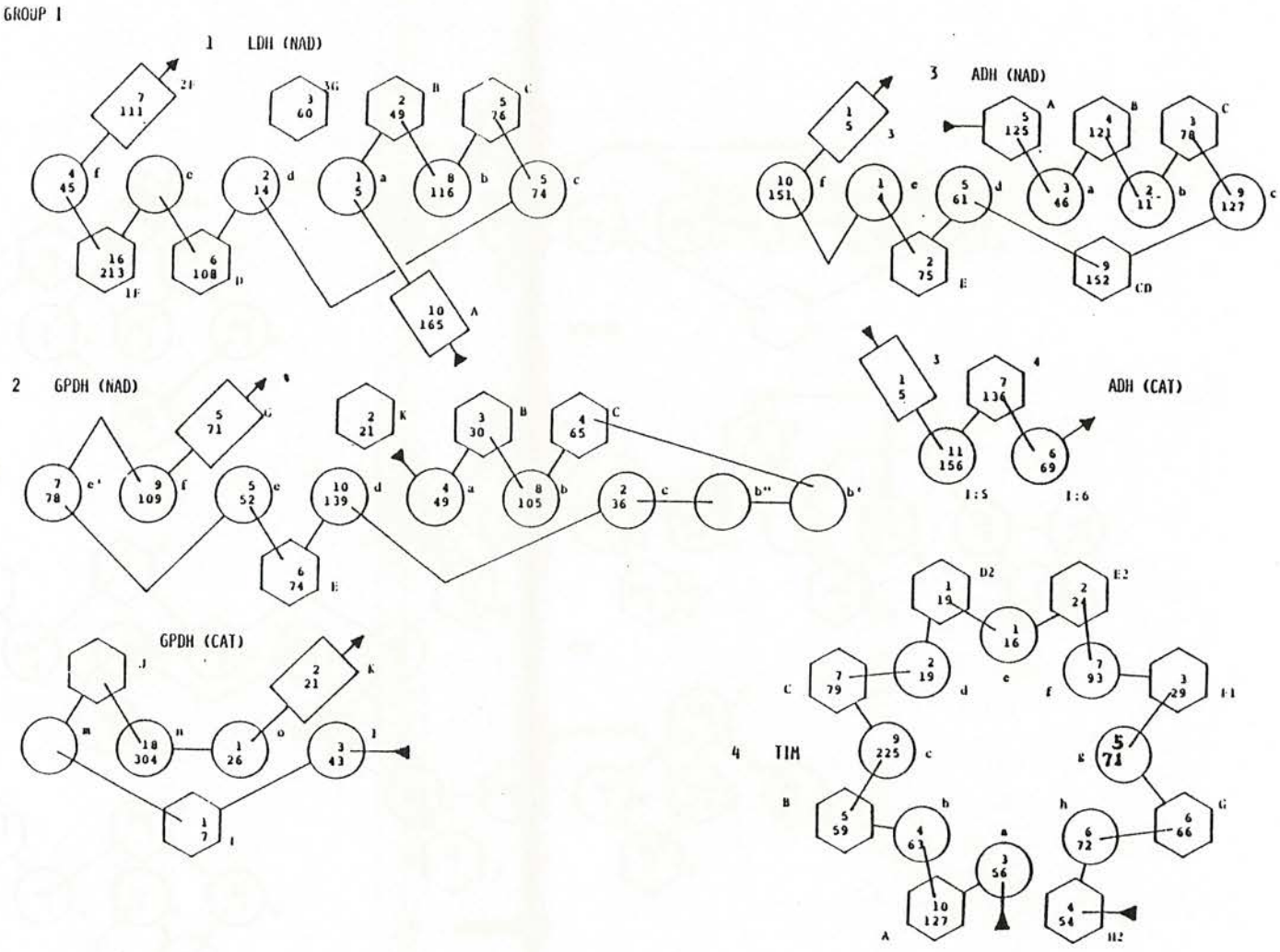


Fig. 1 -1

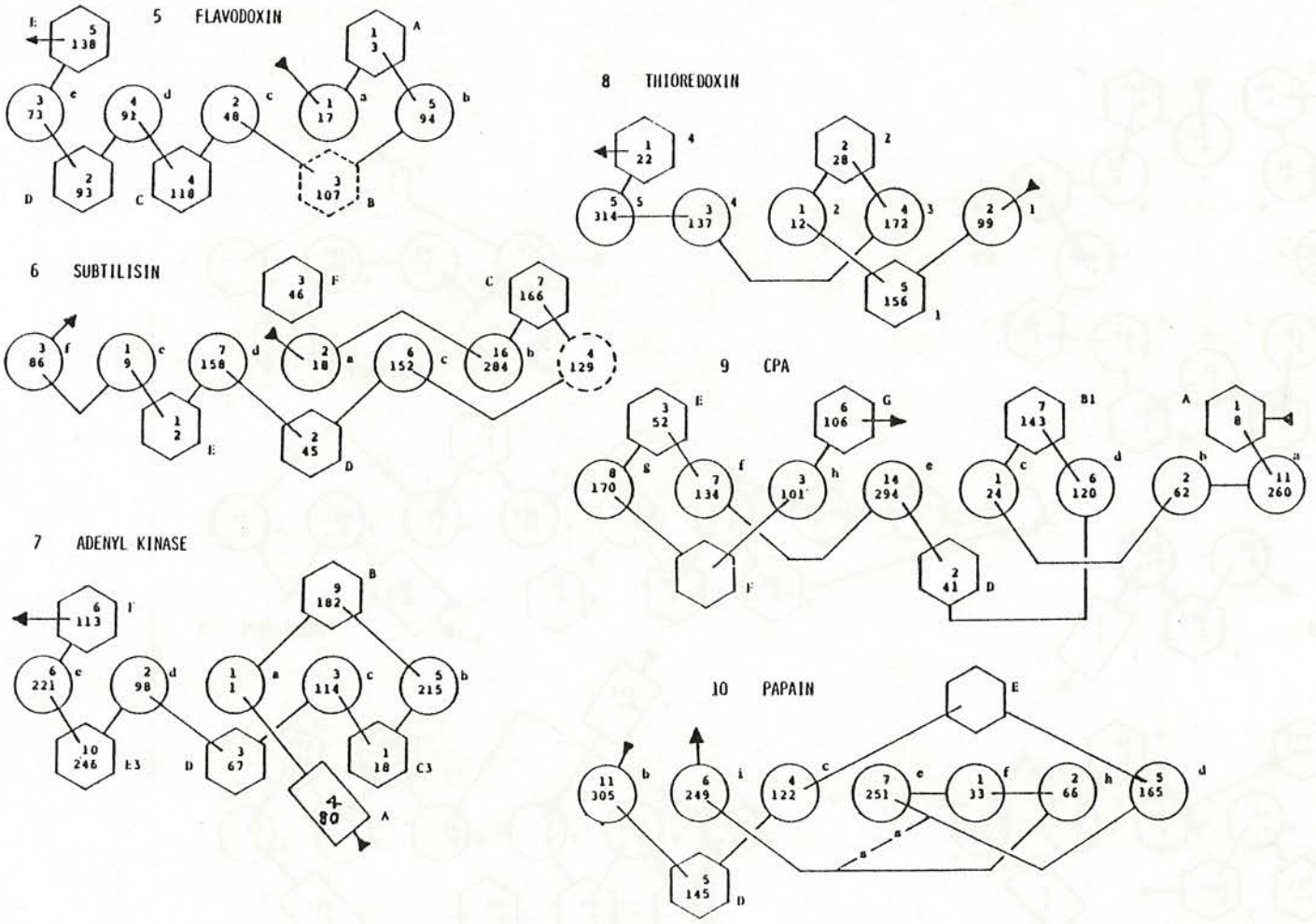

5 FLAVODOXIN
8 THIOREDOXIN
6 SUBTILISIN
9 CPA
7 ADENYL KINASE
10 PAPAIN

Fig. 1-2

the whole ranges of representative 36 proteins used in the analysis. Fig. 1 also shows that strong β-strands and strong helices come close together. Furthermore, when we concentrate our attention on the regions of such strong β-strands and helices, which are close enough, the handedness of the βαβ-unit is uniquely right-handed. This provides us with the theoretical basis for the computer simulation of possible combinations of such βαβ-units by a program called PLAYGAME.

The program defines a kind of chessboard, made of 3 X 11 boxes as shown in Fig. 2. The 11 boxes of the second row are the

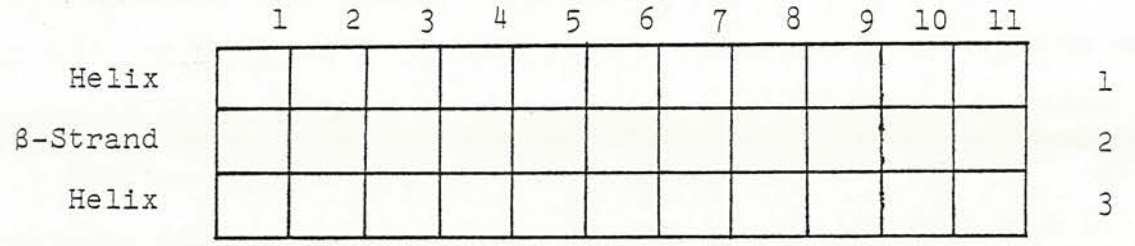


Fig. 2 Board for the sites of β-strands and helices

sites of the strong β-strands (consisting of 5 residues each) to occupy, while the upper or lower row is provided for the accompanying helices. An example of β-sheet barrel of triose phosphate

	1	2	3	4	5	6	7	8	9	10	11	
Helix		241	220	182	148	112	81	54	24	241		1
β-Strand	6	228	207	160	124	89	59	38	6	228		2
Helix												3

Fig. 3 An example of TIM barrel (See Fig. 1)

isomerase (abbreviated as TIM) (Banner *et al.*, 1975) predicted by the present method can be represented as shown in Fig. 3. The numbers in the respective boxes represent residue numbers of the central residue of individual 5 residues picked up as candidates for β-strands and helices.

When the program finds out a strong α-candidate (helix) in a given sequence data, it searches fairly strong β-candidates in its neighbourhood successively, and combines them with each other as a βαβ-unit. Besides, it also starts searching for neighbouring α- and β-candidates from a strong β-candidate. There are some occasions of having no appropriate α-candidate between two β-candidates or no appropriate β-candidate on one side of a β-candidate. The βαβ candidate can generally take three types of conformations as shown in Fig. 4, where *a* and *b* represent the central residue numbers of the two β-candidates, *B* the central residue number of the intervening α-candidate, and * a vacant site between two occupied sites of β-strands. Accordingly, it can easily be understood that the

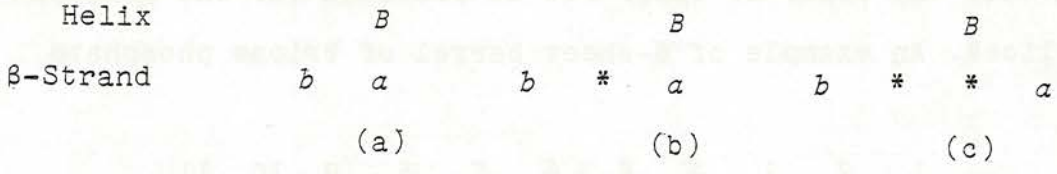


Fig. 4 The most probable 3 conformations of a βαβ-unit

total number of the possible combinations of such βαβ-units soon exceeds several hundreds. So, the program tries to calculate the score of each βαβ candidate and put the best one in the centre of the board. It searches a strong helix or a strong βαβ candidate

outside the regions already used on the board, and tries to simulate the propagation of the nuclei. This strategy seems to work quite well (Nagano, manuscript in preparation). In this way, we can still have a hope that the Monte-Carlo simulation or the energy minimization procedure would be very successful if the starting set will be taken from the one presented by the program PLAYGAME. In conclusion, we can say that the statistical aspects of protein structures reflect their dynamic characters, and that the protein sequence itself tells us how a complex protein organises its folding.

REFERENCES

Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D. & Waley, S.G. (1975). *Nature (London)*, 255, 609-614.

Nagano, K. (1973). *J. Mol. Biol.* 75, 401-420.

Nagano, K. (1974). *J. Mol. Biol.* 84, 337-372.

Nagano, K. (1976a). *J. Mol. Biol.* in the press.

Nagano, K. (1976b). *J. Mol. Biol.* in the press.

Nagano, K. & Hasegawa, K. (1975). *J. Mol. Biol.* 94, 257-281.

Schiffer, M. & Edmundson, A.B. (1967). *Biophys. J.* 7, 121-135.

III.5

A MONTE-CARLO STUDY OF THE FOLDING OF POLYPEPTIDE CHAINS

B.Maigret[1]
S.Premilat[2]

[1] Institut de Biologie Physico-Chimique 13, rue Pierre et Marie Curie, Paris V°)

[2] Laboratoire de Biophysique, Centre de 1 er Cycle, Bd. des Aiguillettes, 54 000 Nancy.

Presently about 80 protein structures are known by X-rays, but nobody is able to explain satisfactorily how a one dimensional information, the genetic code, can drive a polypeptide chain towards a unique tridimensional configuration. Hence, the state of this problem is such that we have a good sample of experimental data, but only poor theoretical tools to analyse it and make predictions. Nethertheless, from the pioneering ideas of Chantrenne (1) or Phillips (2), many works have been done. Recent papers (3) summarized well the matter. Several approaches exist to tackle theoretical solution to protein folding problems :

. Search for statistical correlations between amino-acid sequences and secondary structures. Using the known tridimentional structures as a sample, many prediction algorithms are used (4).

. Conformational energy calculations (5). A stereochemical conformational code is often used to reduce the cost of the search in the whole conformational space. This method supposes that the number of intramolecular energy minima is more and more restricted as the peptide chain growths unit by unit (6).

. Levitt (7) introduced a new and interesting method for the simulation of the folding of polypeptide chains. This promising work uses global potentials, virtual $C^{\alpha} - C^{\alpha}$ bonds and thermalisation effects.

. Tanaka and Scheraga (8) separated interactions in the polymer between short-range, medium range and long-range effects, and recently simulated the chain folding using Monte-Carlo procedure.

As a contribution to the general problem of protein folding, we proposed to develop during the workshop a method of simulation. This algorithm is based on a Monte-Carlo technique and has firstly been applied on small chains of 5 to 20 amino-acid residues. As we obtained fairly good results with these small chains, we planed to extend the method to a complete protein chain. The P.T.I. molecule (Bovine Pancreatic Trypsin Inhibitor) was choosen because of its size and of the existence of good X-ray coordinates for its native structure(9).

METHODS OF CALCULATION

The Metropolis method (10) used in the present work is outlined elsewhere (11) and can be summerized as followed :

. an arbitrary chain conformation defined by a set of (φ, ψ) angles, taken at random in a given list, is built. This step is done n_1 times and corresponds to the starting of the samples. By repetition of this stage conformations taken in the complete conformational space can be choosen with equal probabilities.

Starting from each of the n_1 conformations, one or several pairs of (φ, ψ) angles, randomly choosen in the sequence of the chain, are given other arbitrary values. The perturbation introduced can be more or less important depending on the number of (φ, ψ) angle pairs changed and on the magnitude of the variations on each angle value . This perturbation is done n_2 times and corresponds to the improvement of the samples.

The main difficulty encountered for the transposition of the Metropolis method from small oligopeptides to proteins concerns the way to obtain representative samples for these long chains.

In our previous calculations, we choosed the (φ, ψ) angles at random in the allowed area of the well-know "dipeptide" conformational maps. In this case, samples of 10.000 chains give fairly good results. But in the calculations performed on longer polymeric chains, the amount of the rotational (φ, ψ) angles is very much increased. Then the probability to obtain structures of low energies among all the possible chain conformations is poor. In order to limit in a reasonable good space the search for tertiary structure, one can drive the process using constraints coming from informations on the secondary structures of proteins. The (φ ψ) values for each residue in the chain can thus be choosen according to the predictabilities obtained for this particular residue in the field of its neighbours.

In this way we inducted the system to converge more rapidly. But the signification of the final states obtained is not clear. At least these states are of high probabilities, but do they correspond to the native structure of proteins ? To analyse this point we compared, during the calculations, the conformational energies of the computed structures and of the native one. Moreover, the root-mean-square (R.M.S.) * deviation on the C^{α} coordinates were calculated to measure "the distance" between the structures.

$$* \text{ R.M.S.} = \left[\frac{1}{N} \sum_{i,j>i}^{N} (r^{*}_{ij} - r^{c}_{ij})^2 \right]^{1/2} \quad ; \quad r^{*}_{ij} \text{ from X-rays}, \quad r^{c}_{ij} \text{ calculated}$$

As we were not presently interested to obtain accurate results in the simulation of P.T.I.folding, only the Alanyl, Glycyl and Prolyl amino-acid residues were used (the Alanyl residue is used for all amino-acids with side-chain).

Hence our work was composed of the following stages :

. fit of a chain of standard amino-acid geometry to the available X-ray data of P.T.I. This chain had to correspond to a low conformational energy and to a small R.M.S value. This structure was taken as a reference conformational state for all the other chain conformations computed in the course of this work.

. Monte-Carlo calculations : the procedure was starded using several different choices for the (φ, ψ) angles :

(1) The (φ, ψ) angles of the reference conformation.

(2) a completely random selection of the (φ, ψ) angles in the allowed conformational area of the "dipeptide" map of each kind of residue.

(3) A biased selection of the (φ, ψ) angles according to Nagano's predictabilities (12).

RESULTS

1) Determination of the (φ, ψ) angles of the reference conformation

Many algorithms exist to compute tridimentional protein chain models close to X-ray data. But our aim is not to refine the atomic coordinates obtained previously by the cristallographers. This has already been done for P.T.I. for which accurate refined coordinates exist. In our calculations this set of coordinates was used as the native conformation of the protein.

But as our building program needs as input data the (φ, ψ) angle values and uses standard amino-acid geometry we calculated the (φ, ψ) values which give a good fit between the computed and "native" C^{α} coordinates. This has been realised using several minimisation procedures (13) working on the energy or the R.M.S. The (φ, ψ) angle values corresponding to a good fit and a low conformational energy are given in table I.

Initial values (X-rays)				Final values(reference)			
PHI	PSI	0.0	144.00	PHI	PSI	27.46	153.07
PHI	PSI	-60.	152.00	PHI	PSI	-60.0	129.46
PHI	PSI	-61.00	-30.00	PHI	PSI	-64.04	-31.40
PHI	PSI	-74.00	-12.00	PHI	PSI	-82.30	-29.68
PHI	PSI	-65.00	-17.00	PHI	PSI	-37.17	-23.02
PHI	PSI	-92.00	-5.00	PHI	PSI	-100.28	19.00
PHI	PSI	-75.00	147.00	PHI	PSI	-80.25	136.69
PHI	PSI	-60.	157.00	PHI	PSI	-60.0	162.83
PHI	PSI	-60.	145.00	PHI	PSI	-60.0	142.40
PHI	PSI	-126.00	106.00	PHI	PSI	-130.90	95.97
PHI	PSI	-70.00	-41.00	PHI	PSI	-71.51	-50.73
PHI	PSI	94.00	179.00	PHI	PSI	110.53	153.71
PHI	PSI	-60.	-7.00	PHI	PSI	-60.0	-33.00
PHI	PSI	-89.00	165.00	PHI	PSI	-49.79	182.71
PHI	PSI	-127.00	32.00	PHI	PSI	-147.80	33.23
PHI	PSI	-77.00	178.00	PHI	PSI	-91.09	184.02
PHI	PSI	-134.00	88.00	PHI	PSI	-128.53	91.21
PHI	PSI	-125.00	118.00	PHI	PSI	-126.87	114.69
PHI	PSI	-84.00	118.00	PHI	PSI	-75.77	120.48
PHI	PSI	-126.00	172.00	PHI	PSI	-131.05	171.21
PHI	PSI	-110.00	146.00	PHI	PSI	-110.53	145.11
PHI	PSI	-129.00	151.00	PHI	PSI	-140.63	127.95
PHI	PSI	-80.00	130.00	PHI	PSI	-73.71	128.67
PHI	PSI	-110.00	104.00	PHI	PSI	-111.85	96.59
PHI	PSI	-65.00	-28.00	PHI	PSI	-65.09	-5.67
PHI	PSI	-67.00	-34.00	PHI	PSI	-88.73	-32.02
PHI	PSI	-93.00	-20.00	PHI	PSI	-100.67	-37.43
PHI	PSI	81.00	15.00	PHI	PSI	89.93	34.77
PHI	PSI	-159.00	175.00	PHI	PSI	-161.13	181.21
PHI	PSI	-98.00	146.00	PHI	PSI	-99.79	151.32
PHI	PSI	-133.00	161.00	PHI	PSI	-140.41	159.10
PHI	PSI	-83.00	150.00	PHI	PSI	-74.22	161.94
PHI	PSI	-146.00	166.00	PHI	PSI	-165.50	160.27
PHI	PSI	-96.00	120.00	PHI	PSI	-100.64	104.53
PHI	PSI	-100.00	134.00	PHI	PSI	-88.00	126.62
PHI	PSI	-73.00	-11.00	PHI	PSI	-52.39	-56.77
PHI	PSI	105.00	0.0	PHI	PSI	121.53	30.59
PHI	PSI	-146.00	151.00	PHI	PSI	-173.22	163.44
PHI	PSI	61.00	40.00	PHI	PSI	39.88	63.90
PHI	PSI	-61.00	152.00	PHI	PSI	-95.62	147.64
PHI	PSI	-103.00	180.00	PHI	PSI	-80.92	208.35
PHI	PSI	-85.00	-20.00	PHI	PSI	-92.37	-47.87
PHI	PSI	-77.00	74.00	PHI	PSI	-71.16	72.40
PHI	PSI	-166.00	107.00	PHI	PSI	-162.30	108.34
PHI	PSI	-129.00	158.00	PHI	PSI	-112.89	171.35
PHI	PSI	-88.00	-7.00	PHI	PSI	-109.15	-1.59
PHI	PSI	-154.00	162.00	PHI	PSI	-149.91	164.28
PHI	PSI	-67.00	-34.00	PHI	PSI	-90.33	-13.06
PHI	PSI	-69.00	-42.00	PHI	PSI	-88.10	-40.91
PHI	PSI	-65.00	-41.00	PHI	PSI	-68.48	-37.53
PHI	PSI	-59.00	-50.00	PHI	PSI	-59.69	-64.01
PHI	PSI	-71.00	-32.00	PHI	PSI	-55.65	-48.92
PHI	PSI	-63.00	-42.00	PHI	PSI	-48.40	-19.49
PHI	PSI	-76.00	-46.00	PHI	PSI	-104.45	-78.09
PHI	PSI	-106.00	-4.00	PHI	PSI	-93.50	66.11
PHI	PSI	-78.00	-14.00	PHI	PSI	-114.86	-41.42
PHI	PSI	0.0	0.0	PHI	PSI	0.0	0.0

TABLE I

X-ray and reference conformations.(φ, ψ) angle values

2) Stability of the reference conformation

Starting from the structure defined above, the energy and R.M.S. have been modified during the n_2 improvement steps (n_1 is here fixed to 1) and their variations during the perturbations of the chain are depicted in figures 1,2.

Results are unsatisfactory because of the great differences between the reference conformation and the final ones. Our reference state does not correspond to a real minimum of conformational energy. This result is obtained with different sets of potential functions (the classical ones and also potentials modified to take into account the hydrophobic or hydrophylic behaviour of the amino-acids in the sequence).

3) Convergence of random conformations

When the chain conformations choosen to initialize the Monte-Carlo computation are selected at random, the procedure drives the chain towards low energy conformations but the R.M.S. is increased drastically (fig.1,2). Nethertheless, no conformation is reached with a lower conformational energy than those obtained when starting from the reference conformation.

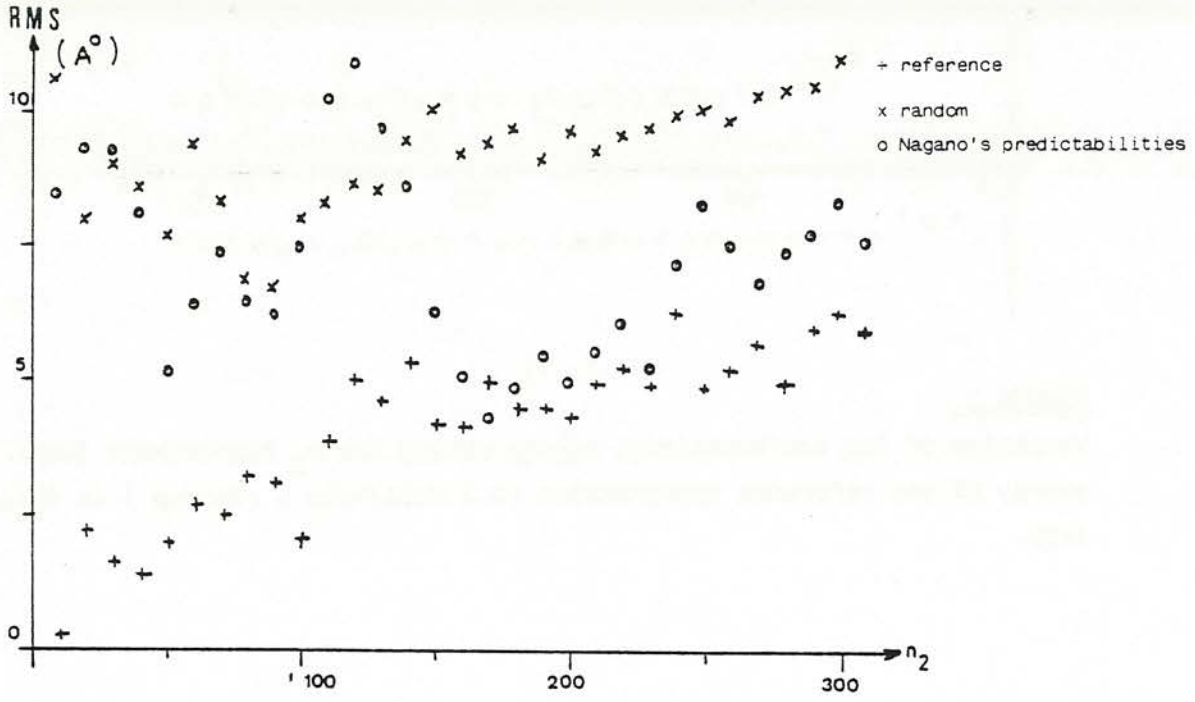


FIGURE 1

Variation of the R.M.S. during the n_2 improvement loop. The R.M.S. of the reference conformation (1.57 Å) is taken as zero.

4) Use of Nagano's predictability

When we use Nagano's predictabilities for each residue of the P.T.I. molecule (table II) conformations of low energies are more easily obtained during the n_2 improvement steps, but the R.M.S. is still increased (figures 1,2).

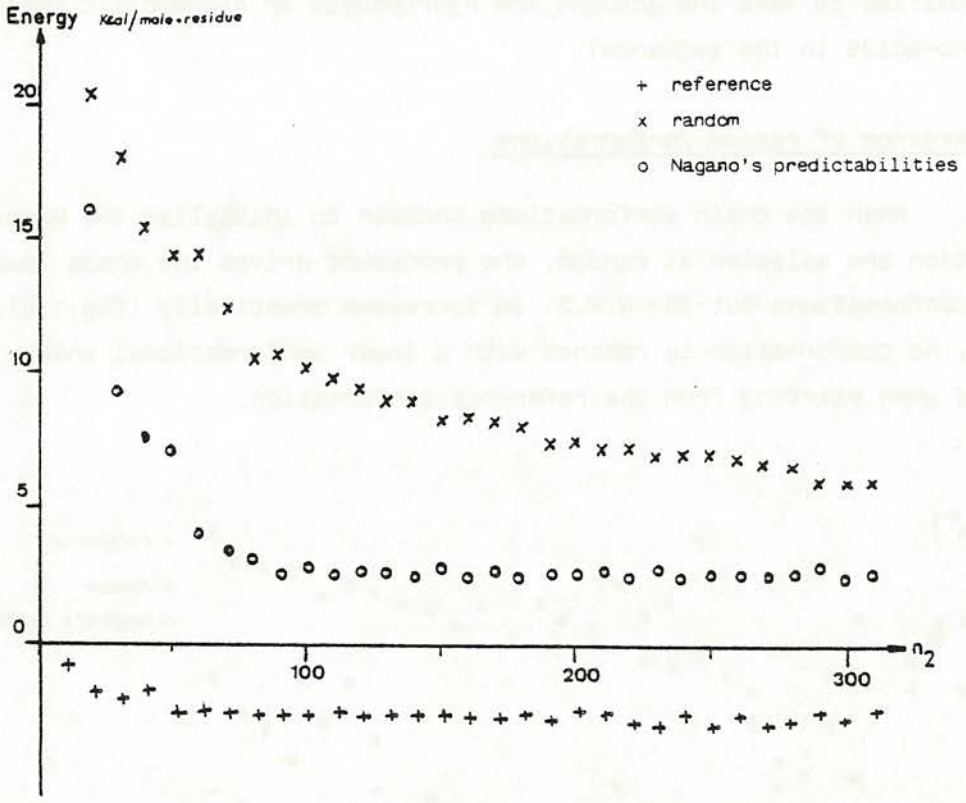


FIGURE 2

Variation of the conformational energy during the n_2 improvement loop. The energy of the reference conformation (0.2 Kcal/mole x residue) is taken as zero.

DISCUSSION

Up to now we are not able to realize the convergency of the polypeptide chain towards conformations which simultaneously correspond to low energies et small R.M.S.We tried to improve the movements of the chain towards the "native conformation" by changing the potential functions. This was done in order

Residue		α Predicta-bility	Turn Predicta-bility	β Predicta-bility
1	ARG	0.0452	-0.0894	0.0647
2	PRO	0.0987	-0.2189	0.1253
3	ASP	0.0614	-0.0681	0.0763
4	PHE	0.1703	0.1696	-0.0937
5	CYS	0.2300	0.2303	-0.1198
6	LEU	0.2205	0.1547	0.0062
7	GLU	0.2973	0.1339	0.0769
8	PRO	0.4957	-0.0556	0.1604
9	PRO	0.5124	-0.3329	0.2128
10	TYR	0.4718	-0.3032	0.1800
11	THR	0.5289	-0.2756	0.1584
12	GLY	0.4723	-0.4031	0.2394
13	PRO	0.4378	-0.3509	0.1614
14	CYS	0.3336	-0.1534	0.1310
15	LYS	0.1516	0.0625	0.0869
16	ALA	0.0246	0.2496	-0.0198
17	ARG	0.0028	0.2204	-0.3028
18	ILE	-0.1393	0.2355	-0.4539
19	ILE	-0.1742	0.2234	-0.4023
20	ARG	-0.1323	0.1393	-0.2768
21	TYR	-0.1517	0.0576	-0.2242
22	PHE	-0.1532	-0.0092	-0.1439
23	TYR	0.0480	-0.0789	-0.0753
24	ASN	-0.0007	-0.0414	0.0901
25	ALA	-0.0584	0.0270	0.2672
26	LYS	-0.0050	0.0763	0.2467
27	ALA	-0.0450	-0.0354	0.1997
28	GLY	0.1307	-0.1202	0.1104
29	LEU	-0.0421	-0.0321	0.0358
30	CYS	0.0196	-0.0906	-0.0195
31	GLN	-0.0553	-0.0407	-0.1498
32	THR	0.0930	-0.0533	-0.2881
33	PHE	0.2391	0.0520	-0.3417
34	VAL	0.4099	0.0935	-0.4450
35	TYR	0.4682	-0.1636	-0.2505
36	GLY	0.4831	-0.3705	-0.0447
37	GLY	0.3748	-0.3310	0.0739
38	CYS	0.2298	-0.0844	0.1075
39	ARG	0.0922	0.0401	0.0792
40	ALA	0.0960	0.1071	0.1168
41	LYS	0.1105	0.0951	0.0548
42	ARG	0.0085	-0.0375	0.1026
43	ASN	0.0399	-0.1483	0.0173
44	ASN	-0.0515	-0.1184	0.0561
45	PHE	-0.0908	0.0417	0.0454
46	LYS	-0.2497	0.0848	0.1529
47	SER	-0.2554	0.0026	0.1206
48	ALA	-0.3857	-0.0580	0.1819
49	GLU	-0.3571	-0.0372	0.2765
50	ASP	-0.4794	0.0039	0.2013
51	CYS	-0.4546	-0.0878	0.1680
52	MET	-0.3508	-0.0746	0.0947
53	ARG	-0.1374	0.0041	0.0675
54	THR	0.1187	-0.1157	-0.0937
55	CYS	0.1947	-0.2836	-0.0729
56	GLY	0.2174	-0.4161	-0.1153
57	GLY	0.0815	-0.1807	-0.0465
58	ALA	0.0100	0.0285	0.0358

TABLE II

Nagano's predictabilities for all the residues of P.T.I.

to take account of the different behaviours presented by the twenty amino-acids when in contact with water. It seems that the modifications introduced on the potential functions (suppression of the attractive parts of Van der Waals "6-12" potentials for instance) are to crude and more work is necessary to define on a computational level the hydrophobic effects.

Besides, it seems that when informations on secondary structures is used as constraints for choices of chain conformations, the chains adopt more easily better tertiary structures. Hopefully, this kind of results obtained with our Monte-Carlo method shall give very soon valuable indications on the problem of polypeptide chains folding.

REFERENCES

1) H. CHANTRENNE in "The biophysics of Proteins", Perganom Press, New-York (1961) p. 122

2) A.C. PHILLIPS, Proc. Nat. Acad. Sci. U.S.A., (1967), 57, p.484

3) C.B. ANFINSEN and H.A. SCHERAGA, Adv. Protein Chem., (1975)

4) T.T. WU, W.M. FITCH and E. MARGOLIASH, Ann. Rev. Biochem. (1974), 539

5) H.A. SCHERAGA, Pure and Appl. Chem. (1973), 36, 1

6) E. RALSTON and J.L. DE COHEN, J. Mol. Biol. (1974), 83, 393

7) M. LEVITT, J. Mol. Biol. (1976), 104, 59

8) S. TANAKA and H.A. SCHERAGA, Macromolecules, (1976), 9, 168

9) J. DEISENHOFER and W. STEISEMANN, Acta Cryst. (1975), B 31, 238

10) N.A. METROPOLIS, A.W. ROSENBLUTH, M.N. ROSENBLUTH, A.H. TELLER and E. TELLER, J. Chem. Phys. (1953), 21, 1087

11) S. PREMILAT and B. MAIGRET, C.R.A.S. Paris (1976), 282, 225

12) K. NAGANO and K. HASEGAWA, J. Mol. Biol. (1975), 94, 257

13) VA05A, VA09A, M.J.D. POWELL, HARWELL SUBROUTINE LIBRARY, HARWELL, England

CONMIN, P.G. HAARHOF, J.D. BOYS and H. VON MOLENDORF, PEL 190, Atomic Energy board, Pretoria, South Africa.

III.6

THE DYNAMIC BEHAVIOUR OF A SIMPLIFIED REPRESENTATION OF PANCREATIC TRYPSIN INHIBITOR.

M.Levitt

MRC Laboratory of Molecular Biology, Hills Road,
Cambridge CB2 2QH (England).

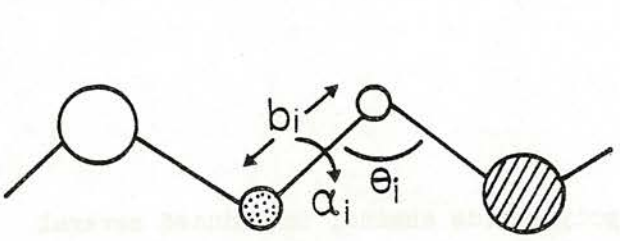
Simplified representations of polypeptide chains, introduced several years ago by Levitt and Warshel (1975), enable the energy and conformations of proteins to be computed very efficiently. In the past, these representations have been used to simulate protein folding by changing the molecular conformation so as to minimize the simplified energy of the system (Levitt & Warshel, 1975; Levitt, 1976; Warshel & Levitt, 1976; Kuntz et al., 1976). In such a minimization, the randomizing of effect of temperature has been introduced artificially either by thermalization between minimization runs or by artificial "pushing potentials" that force the conformation out of local energy minima.

A more rigorous way to simulate the conformational changes of the simplified protein chain is the method of molecular dynamics. This method, which simply involves applying Newton's third law relating force and acceleration to the system, is extremely powerful in that the way the conformation changes with time (a trajectory) is an accurate representation of real dynamic behaviour.

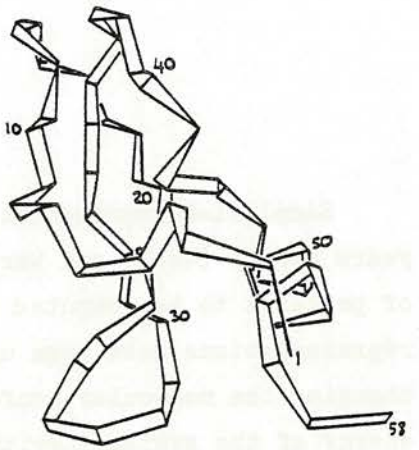
We have applied molecular dynamics to two different starting conformations of a small protein, pancreatic trypsin inhibitor (abbreviated as PTI). One conformation, the native X-ray conformation, was used to investigate the modes of vibration, deformations, and stability of a native protein. The other conformation, a fully extended chain, was used to investigate protein folding dynamics and the motion of denatured protein chains.

Methods

First, I will describe the molecular representation used here. The protein structure, which consists of a sequence of 58 amino acids each consisting of between 7 and 20 atoms, was simplified to a chain of 58 spherical groups. Each group was placed at the centroid of the atoms of the side chain it represented and assigned van der Waals and solvent interaction potentials typical of the amino acid (we used set C parameters taken from Levitt, 1976).



(a)



(b)

Fig. 1a. Showing the simplified geometry used here. The amino acid side chains represented as shaded or clear spheres to indicate their different solvent affinities are bonded together to form a simple chain.
Fig. 1b. Showing a simplified drawing of Pancreatic Trypsin inhibitor in which a single atom of each amino acid (the C^{α}) are connected by a ribbon. There are SS bonds between residues 5:55, 14:38 and 30:51.

This representation is simpler than that used previously (Warshel & Levitt, 1975), where each amino acid was represented by three groups. Another difference is that here the groups are given all degrees of freedom in a Cartesian space, whereas previously only the chain torsion angles were allowed as degrees of freedom. Additional energy functions must, therefore, be used to allow bond stretching, bond angle bending and bond twisting (torsion). The bond energy is taken as $E_{BOND}=\Sigma_{\text{all bonds}} K_b(b_i-b_o)^2$ where the force constant K_b was 40 kcal/mole-Å^2, and the i-th bond length value constrained to remain close to its initial value b_o. The bond-angle energy is taken as $E_{BOND\ ANGLE} = \Sigma_{\text{all angles}} K_\theta(\theta_i-\theta_o)^2$ where K_θ=40 kcal/mole-Å^2. The torsion energy was calculated as before (Levitt, 1976), but now the angle α is defined by residue centroids not α-carbon positions. There are bonds and bond angles between all groups adjacent along the amino acid sequence. In addition, groups representing half-cystine residues known to be connected by S-S bonds in the native protein were also sometimes connected by a bond.

With this analytical energy function, it is an easy matter to calculate the energy value, E, and its gradient, the forces F, at any point in Cartesian space. These quantities were used to calculate how the conformation evolved with time using an algorithm recommended by Beeman (1976) for its simplicity, stability, and convenience.

$$a_n = F_n / m$$

$$r_{n+1} = r_n + V_n \Delta t + (4a_n - a_{n-1})\Delta t^2/6$$

$$V_{n+1} = V_n + (2a_{n+1} + 5a_n - a_{n-1})\Delta t/6$$

Because the new position vector r_{n+1} depends on the velocity V_n it is easy to heat or cool the system by increasing or decreasing V_n. The only adjustable parameter in these equations is the time interval Δt at which the energy and forces are calculated. For computational efficiency, Δt must be as large as possible yet still allow perfect (say 0.1%/1000 cycles) energy conservation. This maximum value of Δt is related to the period of the most rapid vibration of the system, but its value is usually found by trial and error. It is very convenient to use special units in these calculations in which energy is measured in kcal/mole, distance in Angstrom units, and mass in gms/mole. When this is done, one unit of time becomes 4.86×10^{-14} seconds. Here, the optimum value of Δt was 0.5 in these natural units (about 2.5×10^{-14} seconds). In conventional molecular dynamics calculations that use atoms rather than the much heavier groups of atoms Δt is between one and two orders of magnitude smaller.

In the calculations, extensive use was made of the facility to control the kinetic energy and hence temperature of the system. When folding from an open chain, the potential energy drops considerably and energy must be removed if the molecule is ever to become compact. When studying the dynamics of the native protein, the system must first be cooled to dissipate the high initial potential energy due to bad van der Waals contacts, and then later slowly heated up to study the dynamics of denaturation.

Results

Because of the limited time available and the incomplete state of the work, these results are necessarily preliminary. First, I report on the studies of the native protein conformation. In this test, the molecular dynamics trajectory was started with each group at the centroid of the corresponding amino acid in the X-ray conformation, and the native SS bridges were bonded together. Initially, there are a few bad van der Waals contacts that give a very high starting potential energy. As a result, the kinetic energy rises very rapidly and the protein denatures within about 100 steps (2.5×10^{-12} secs). To prevent this, rapid cooling was applied for the first 1000 steps until the energy remained constant and the temperature was about 100°K. The temperature was then increased by 20° K every 1000 cycles and the trajectory recorded on magnetic tape for subsequent analysis. The analysis, which is still incomplete, shows several things.

The simplified model stays within 4Å root mean squared deviation of native PTI throughout the simulation of 10,000 steps (2.5×10^{-10} seconds), during which the temperature had reached 290°K. This deviation calculated as described before (Levitt, 1976) is comparable to the deviations of about 3Å obtained for the near-native energy minima previously, especially when one considers that the present geometry is simpler.

The amplitude of vibration of the simplified model is small at about 1Å r.m.s., and the distribution of vibration along the chain (Fig. 2) is reasonable.

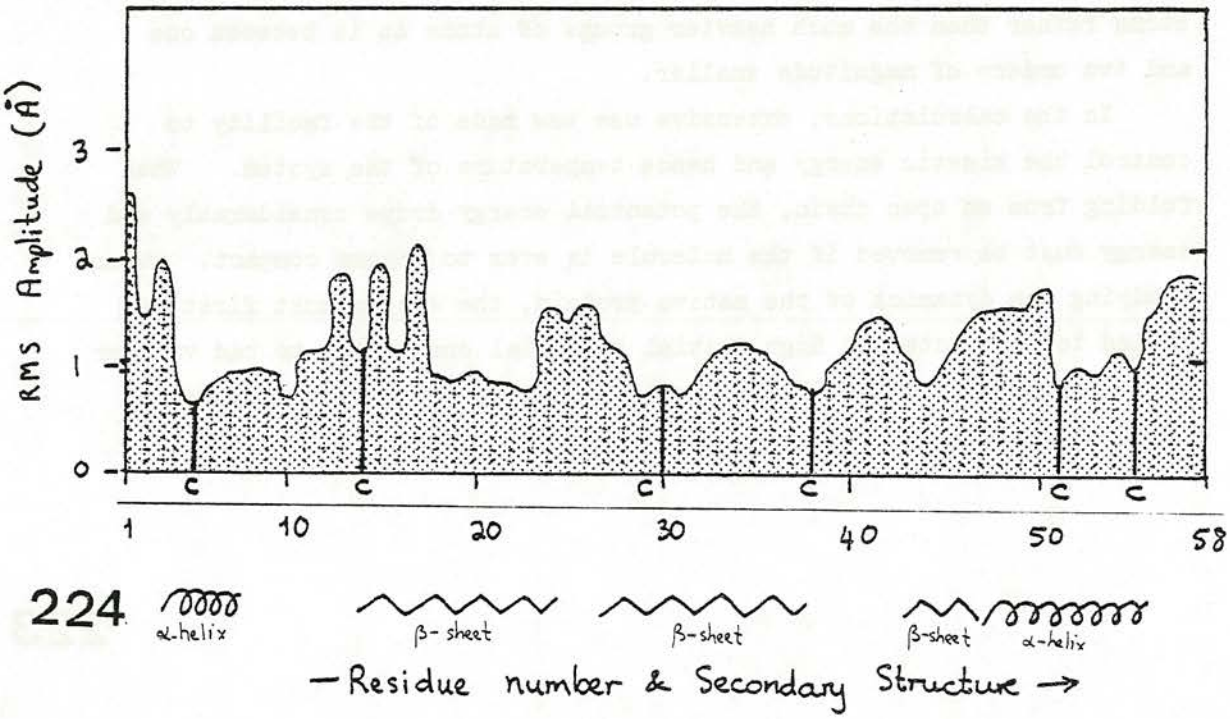


Fig. 2

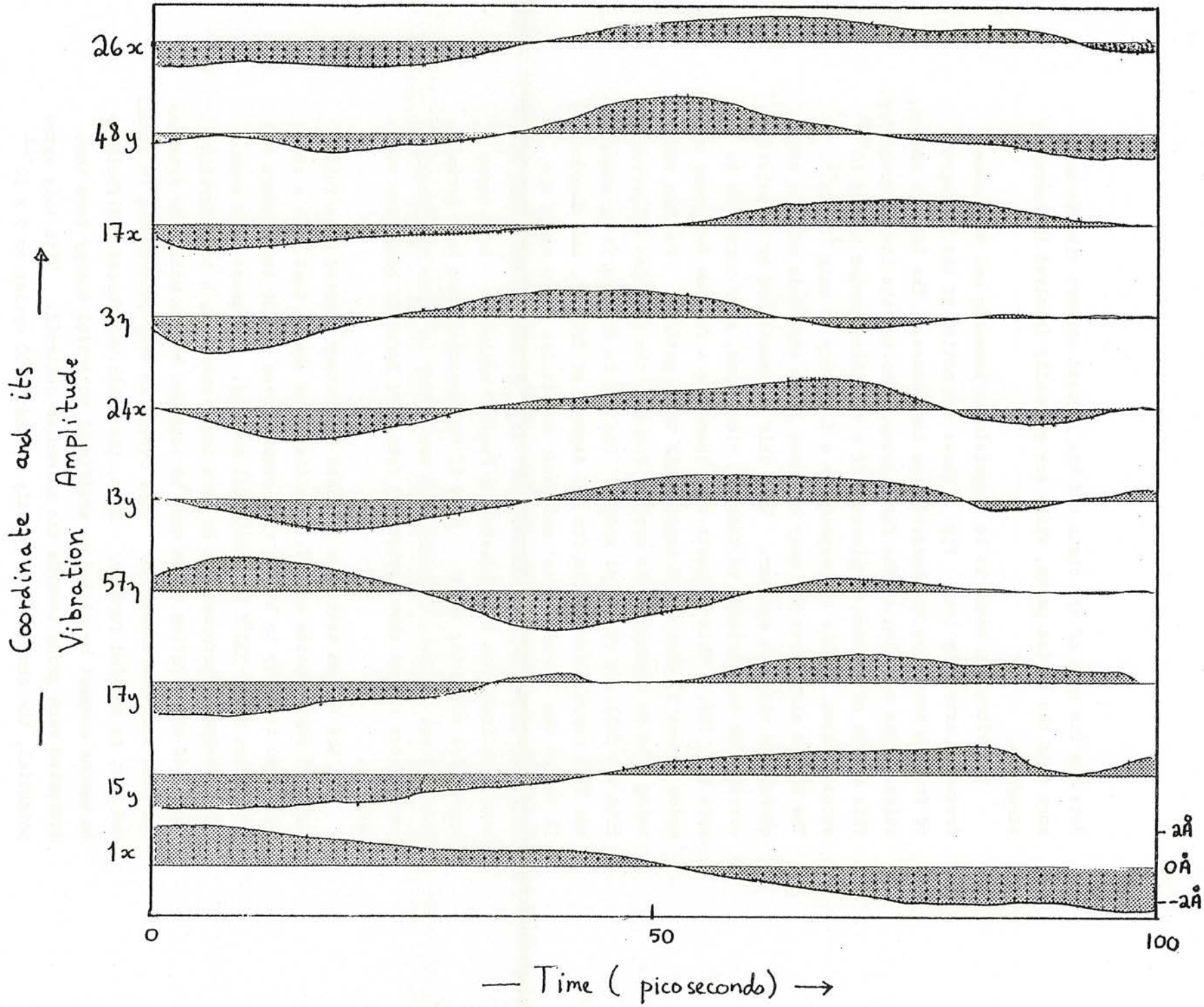

26x
48y
17x
3z
24x
13y
57z
17y
15y
1x
Coordinate and its
Vibration Amplitude
2Å
0Å
-2Å
0
50
100
Time (pico secondo)

Fig. 3.

Note how the ends of the chain and the exposed corners vibrate much more than the buried parts, which are generally involved in secondary structure.

The vibration seems to be surprisingly harmonic and the lowest frequency extremely low. Fig. 3 shows the motion of the 10 degrees of freedom that have the largest mean amplitudes. The largest motion, which is the α motion of the first group, also has the lowest frequency. This degree of freedom completes half a vibration period in the 10^{-10} seconds shown. This corresponds to a frequency of only $1/6$ cm^{-1}. The figure also shows how many of these large amplitude motions are correlated with one another. When this is quantified by calculating correlation coefficients of these 10 vibrations, many correlate to more than 70%. This suggests that there are a few low frequency normal modes (about 3) that would explain all these motions. Further work is being done to decompose the motions found by the dynamics trajectory into the full set of normal modes. One way to do this is to calculate the full correlation matrix for all degrees of freedom, and diagonalize it to get the 'normal modes' and their amplitudes (from which the frequency can easily be calculated). Alternatively, the motion of each coordinate can be subjected to Fourier analysis. Both these approaches are being tried. Even if the decomposition into normal modes is not perfect, following the trajectory in these transformed coordinates should show deviations from pure harmonic behaviour very clearly.

In the second test, the dynamics trajectory started at a fully extended conformation of PTI. Initial runs showed that such a chain showed no tendency to become very compact even if the temperature was kept below about 330°K by artificial cooling. It seems that even though compact conformations do have lower energies, a considerable volume of conformation space must be searched before molecular dynamics can find such conformations. This result is very interesting in itself and must be studied further. Here, the chain was forced artificially to become compact by adding an artificial potential energy term that attracted each group towards the molecular centroid. With this extra potential, the chain folded rapidly (within 2000 cycles or 5×10^{-11}

seconds) to within 6Å r.m.s. of the native PTI structure. This deviation is comparable to that obtained using energy minimization.

In the above trajectory, kinetic energy had to be removed to keep the temperature below 330°K. When this was not done, the temperature rose until it settled at 700°K. The chain did not become compact but assumed a labile open structure. The movements of this denatured protein chain were still slow as it took about 10^{-10} seconds for the tails of the molecule to swing through about 180°.

Conclusions

Simplified representations can be used to study the dynamics of protein chains. The main advantages of the representation is increased computational efficiency in that for the same amount of real time one can simulate about 1000 times more trajectory time. This allows one to study events on time scales of about 10^{-9} seconds. Another advantage is the smaller number of coordinates (for PTI there are 174 simplified coordinates and about 1374 detailed ones) allowing detailed analysis of correlation matrices, to normal modes, etc.

The big disadvantage of the representation is that one cannot know how trustworthy such a representation will be. I feel that the technique will be useful to studying the dynamics of open chain conformations as well as partially denatured native conformations. The detailed all-atom representation seems better suited to studies of the dynamics of the native protein, though it will be very interesting to compare the average residue amplitudes of motion in Figure 2 with those obtained by McCammon on the detailed model of PTI. Another way to get more reliable dynamics is to use effective energy terms derived as time averages from the more detailed treatment.

Acknowledgements

I am extremely grateful for the generous support that enabled me to attend this CECAM Workshop. The ideas exchanged, the new techniques learnt, and the new friends made, will have a strong influence on my future research.

References

Beeman, D. (1976). J. Computational Phys. 20, 130-139.

Levitt, M. and Warshel, A. (1975). Nature, 253, 694-698.

Levitt, M. (1976). J. Mol. Biol. 104, 59-107.

Kuntz, I.D., Crippen, G.M., Kollman, P.A. and Kimelman, D. (1976). J. Mol. Biol. 106, 983-994.

Warshel, A. and Levitt, M. (1976). J. Mol. Biol. 106, 421-437.

III.7

THE PACKING OF α-HELICES onto β-SHEETS IN PROTEINS

C.Chothia[1]
W.Ramsay[2]

[1] Institut Pasteur, Paris XIV (France)

[2] Ralph Forster and Christopher Ingold Laboratories, University College London, Gower Street, London WC 1.

INTRODUCTION

I present here a model that describes the stereochemical rules which govern how α-helices pack onto β-pleated sheets to form the three-dimensional (tertiary) structure of certain protein molecules. The model was developed empirically: a priori, models were checked and refined by a detailed analysis of the residue to residue contacts that occur between the α-helices and β-sheets in four proteins.

THE DETERMINING PRINCIPLES

Two principles have a dominating influence on the way in which secondary structures associate:

1) Residues that become buried in the interior of a protein close pack: they occupy a volume similar to that which they occupy in crystals of their amino acid[2,3].

2) Associated secondary structures retain a conformation close to the minimum free energy conformation of the isolated secondary structures[4,5,6].

These two principles imply that secondary structures interact to form a protein molecule in a manner that maximizes the van der Waals energy without inducing

appreciable steric strain. The rules described below for secondary structure associations arise from these two principles and from the intrinsic geometrical properties of polypeptides.

HELIX-SHEET PACKING

The model for packing an α-helix onto a parallel or anti-parallel β-pleated sheet is illustrated in Figure 1.

There are two general features of β-pleated sheets that are important for this model: the first is the packing between neighbouring residues within a sheet and the second, its tendency to have a right-hand twist. If we consider the C_α atoms on the same side of a sheet the distance between neighbours along the strands is $\sim 7\mathring{A}$ and between those in adjacent strands it is $\sim 5\mathring{A}$. Side chain volumes average $\sim 100\mathring{A}^3$ and vary between $25\mathring{A}^3$ (ala) and $170\mathring{A}^3$ (trp)[3]. This means that most side chains in a β-sheet will be in contact with their neighbours so the surface will not have ridges and grooves but can be considered as flat with, usually, only small irregular holes and protuberances.

Both parallel and anti-parallel β-pleated sheets in globular proteins have a right-hand twist when viewed

along the polypeptide chain[7]. The effect of this twist is that neighbouring chains wind around each other whilst remaining a constant distance apart. The same thing happens to the ropes in a rope ladder if a far rung is given a right-hand twist relative to a near rung. For a typical β-pleated sheet the right-handed twist about an axis parallel to the chain direction is observed to be about -5° per Å.

Now let us consider an α-helix. The residue pairs (i, i+1), (i+4, i+5), (i+8, i+9) ... wind round the helix with a right-hand twist (Fig. 1). For a regular α-helix this twist is about -5° per Å. Also, given suitable side chain size and conformation, these residues can form a flattened, though irregular, surface. For reasons that will become apparent just below we will call these residues the "normal contact residues".

In its simplest form our model for α-helix β-sheet packing can be stated as follows: an α-helix will pack onto a β-pleated sheet with its axis parallel to the strands of the sheet because, in this orientation, the normal contact residues form a surface complementary to that of the sheet. Such a model would predict that helix residues in contact with the sheet will be i, i+1, i+4, i+5, i+8, i+9 ..., and that the angle between the

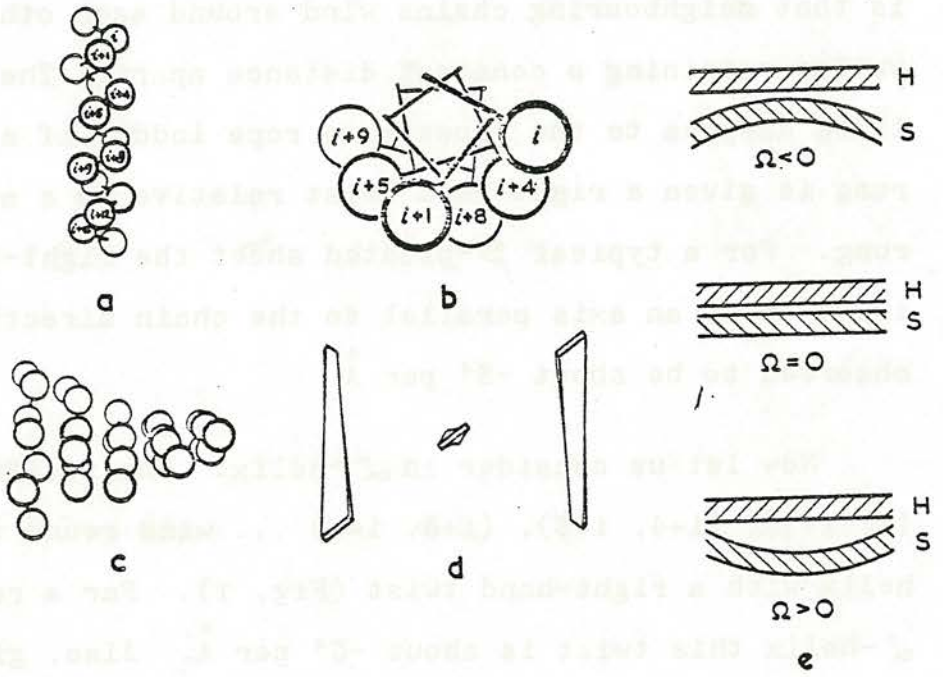


Figure 1

Figure 1: The model for helix-sheet packing: a and b show how the helical residues, i, i+1, i+4, i+5, i+8 ... form a surface with a right-hand twist. The right-hand twist of a β-pleated sheet is shown by (c), the C atoms from part of the sheet in carboxypeptidase a, and (d) a three stranded schematic sheet. (e) shows sections perpendicular to an idealised helix-sheet interface for different values of Ω (see text).

helix axis and the strands in the sheet (Ω)* will be zero.

On a more detailed level we can see that the twist of the β-sheet means that orientations of the helix away from the parallel position (Ω = 0°) are more likely to occur in the negative (clockwise) direction (the helix is assumed to be above the sheet). In this orientation (Ω < 0°) though the exposed ends of the helix move away from the sheet, its centre is still able to close pack.

Footnote:

*We use Ω to describe the relative orientation of two pieces of secondary structure in contact. Ω is defined as the angle between the β strands and/or helix axes when projected onto their plane of contact. We ignore the direction of individual helices and the strands of β-sheets so Ω is defined between -90 and +90 rather than -180 and +180. The angle is negative (0 → -90°) if the near helix or strand is rotated in a clockwise direction relative to the far strand. If this rotation is anti-clockwise the angle is positive (0° → +90°).

In the opposite orientation ($\Omega > 0$) the two ends of the helix will pack onto the sheet but they lift its centre off the sheet and so create an internal cavity in the protein. Using the same argument we can show that helices packed on sheets with a large twist will have negative Ω values.

A dominant structural feature of the proteins flavodoxin, carboxypeptidase, subtilisin and triose phosphate isomerase is a large central β-sheet flanked by α-helices. We have examined the contacts that occur between the residues in these proteins and have found that they contain 19 helices that have four or more residues in contact with the face of a β-sheet. In total, the 19 helices have 129 residues in contact with the β-sheets of which 112 (87%) are what we defined above as "normal contact residues" for sheet-helix packing. In Figure 2 we show for flavodoxin the helical residues that are contact with the central β-sheet. We also show in Figure 2 the values of the angle between the axes of these helices and the strands of their β-sheet (Ω). All these angles are in reasonable agreement with our model. Thirteen have values in the range $-10° \pm 10°$ and the distribution is skewed towards negative values.

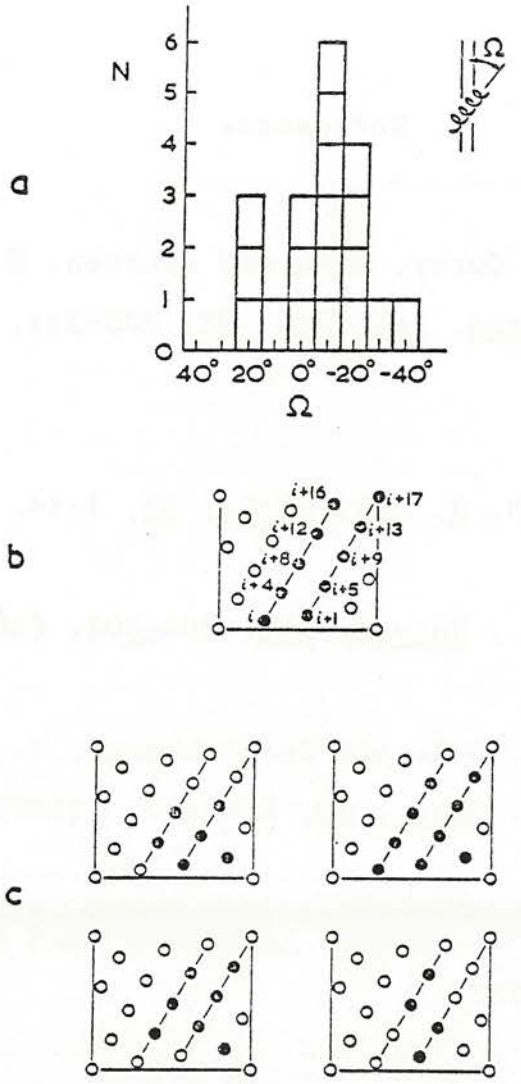


Figure 2: The observed helix-sheet packing.

(a) Histogram of the angles observed between helix axes and β -strands (Ω). The values are for the 19 helices in carboxypeptidase a, flavodoxin, triose phosphate isomerase and subtilisin which have four or more residues in contact with a β sheet. The helix is on top of its sheet.

(b) The normal contact residues (see text) are shown as filled circles in a flattened projection of a helix

(c) The helical residues in contact with the β -sheet (filled circles) in flavodoxin.

References

1. Pauling, L., Corey, R.B. and Branson, H.R. Proc. Nat. Acad. Sci. USA, 37, 205-211, 235-285, (1951).

2. Richards, F.M., J. Mol. Biol., 82, 1-14, (1974).

3. Chothia, C.H., Nature, 254, 304-308, (1975).

4. Ramachandran, G.N. and Sasisekharan, V., Advanc. Prot. Chem., 23, 284-437, (1968).

5. Richards, F.M., Advanc. in Biophy. and Bioeng., 6, in press, (1977).

6. Gelin, B. and Karplus, M., Proc. Nat. Acad. Sci. USA, 72, 2002-2006, (1975).

7. Chothia, C.H., J.Mol. Biol., 75, 295-302, (1973).

III.8

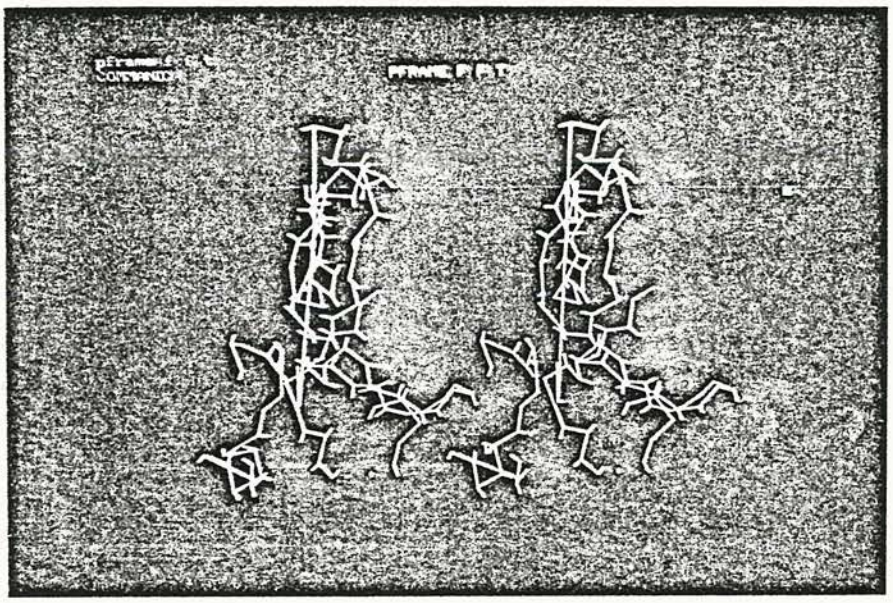
THE STUDY OF PROTEIN-PROTEIN INTERACTION: CALCULATION OF THE INTERMOLECULAR CONTACTS OF TRYPSIN WITH THE PANCREATIC TRYPSIN INHIBITOR

S.Wodak

Department of Chemical Biology, Université Libre de Bruxelles,
67 rue des Chevaux, Rhode St.Genese (Belgique).

Stereo Picture of the BPTI-Trypsin complex taken from the Tektronix CRT at CIRCE.

Onle Cα and the amino acid centroids are shown. The straight lines are part of the pseudo bond system used to simulate the rigid body motion.



BPTI

Trypsin

active site.

Introduction

A large number of studies have been directed towards elucidating the laws which govern protein structure and protein folding. Empirical energy functions have been used to "refine" protein structures starting from x-ray coordinates and to survey their conformational space. A major obstacle in these studies lies in the complexity of the molecular systems examined. The number of degrees of freedom for a given protein is large and the forces involved depend on the nature of the atoms in the protein as well as on the properties of the surrounding solvent.

There has been a developping interest in using extremely simplified models to predict the structure of small globular protein (Levitt and Warshel, 1975; Ptytsyn and Rashin, 1975; Burgess and Scheraga, 1975; Kuntz et al., 1976). In these models, the protein backbone and side-chains are replaced by one or two interactions centers per amino acid representing a time averaged local structure. Approximate empirical potential functions are used for the steric and energetic requirements of the peptide backbone and for selected long range interactions of the amino acid side-chains. As a result, the number of degrees of freedom as well as the roughness of the conformational space are greatly reduced.

The studies by Levitt and Warshel (1975), Levitt (1976) and Kuntz et al. (1976) which use such a simplified model show that a number of compact structures can be obtained starting from an extended chain, and that these compare moderately well with the known crystal structure. These results are very encouraging given the drastic approximationsused. It remains true, however, that a complete systematic survey of the conformational space is still computationally prohibitive, the reliability of the simplified models is not yet established in so far as it is not known what percentage of the minimum energy conformations generated actually correspond to physically meanigful structures or how dependent these minimum conformations are on the amino acid sequence.

The phenomenon of protein association is wide spread in biological systems. Proteins aggregate to form larger structures whose thermodynamic properties regulate many vital functions of the living cell. Several such aggregates have been thoroughly studied and their three dimensional structure determined to high resolution by x-ray crystallography. Often however, the size or geometry of the aggregate precludes studies at

high resolution, which can then be performed only on the individual protein building blocks. It would thus be very instrumental to learn the essential facts about protein association and to be able to predict the possible modes of interaction from the 3 dimensional structure of the individual proteins. In addition, the study of protein association would be extremely well suited for testing the validity of simplified models of the protein surface as well as for a systematic evaluation of the different ways used to map conformational space.

The interaction between two proteins is generally governed by the overall rotation and translation of the molecules as well as by the degrees of freedom of the individual side-chains, mostly on the surface of the protein. Protein association can also be viewed as the last step of protein folding when two or more prefolded lobes come together. In which case the rigid body movements are governed by the degrees of freedom conveyed by the flexible hinge.

From x-ray crystallography studies it is known that many of the side chains on the surface of proteins are free to rotate. Hence the notion of a time averaged side-chain for these amino acids seems as appropriate here as in the case of a flexible chain. One should therefore be able to generate all possible modes of interaction between two proteins as a function of six degrees of freedom only, by using a simplified representation for the protein surface. Moreover, as a given protein-protein interace contains only a small number of amino acids (20 or 30 at most), a representation with full atomic detail can be used to evaluate the physical chemical characteristics of the structure and a comparison with the simplified model can be made at any stage.

Studies by Cothia (1974), Cothia and Janin (1975) show that the large loss of translational and rotational entropy which occurs upon protein association is balanced by the hydrophobic free energy due to smaller accessible surface area to solvent. While the contribution of the polar interactions (h-bond, salt bridges) to the energy of association is small compared to that of the hydrophobocity, since similar interactions can be formed with the solvent in the free protein. The ability to form close packed interfaces with the polar atoms properly positioned to form h-bond plays an essential role in the complementarity and specificity of the interaction. We chose to study in this workshop the association of the enzyme trypsin with the Bovine Pancreatic Trypsin Inhibitor (BPTI). The two proteins form a very tight complex; K dissociation $\sim 1.4\ 10^{-11}$. In addition, the three dimensional structure of the complex (called here for convenience, "native" contact) is known from an x-ray crystallography study at 1.9Å resolution, Huber et al. (1974). By generating all the possible modes of interaction

of Trypsin with BPTI we hope to be able to test the protein association "rules" cited above as well as determine the effectiveness of the simplified representation in determining the known interface of the Trypsin-BPTI complex. We will also be able as a result of such a survey to learn about the properties of the protein surface (i.e. : which regions have a propencity to aggregate) and about the energy barriers involved in the search for favourables modes of association.

During the 8 weeks of the workshop, we have been able to carry out a complete survey of all the modes of interactions of the BPTI molecule with the active site of Trypsin using the simplified rigid amino acid representation as derived by Levitt(1976), where each amino acid side-chain is represented by one ball and has no freedom of movement. Over 2300 protein-protein contacts were generated.For each of these contacts the identity of the participating amino-acids, the number of interactions and the interaction energy were computed. A rough screening allows one to single out about 50 contacts for beeing as "extensive" as the observed native contact. These correspond to well defined regions on the surface of the BPTI molecule and can be correlated to the geometrical properties of the surface at these regions. Additional criteria such as a detailed "packing" test based on the averaged representation indicate that only very few of these "extensive" contacts consist of well packed interfaces. Whether this is due to the use of the simplified representation or is a physically meaningful fact related to the specificity of the association remains to be determined. As a step towards a more sophisticated evaluation of the generated interfaces in terms of their burried surface area and the detailed packing, we have also carried out energy refinement on the "extensive" contacts in the averaged representation. Some of the results of these refinements will be reported here. Further work on this and other studies using a detailed atomic representation of selected interfaces are now in progress.

The simplified model.

In the simplified protein model used in this work, each amino acid was replaced by one effective interaction center whose steric properties are as in Levitt (1976). The crystallographic CA and CB positions were used to define the direction along which the interaction center of a given amino acid was to be palced. The position of the center along this direction differed according to the amino acid considered (see Table 1).

The potential function.

The potential function representing the pairwise interactions of the locally averaged amino acids was as derived by Levitt (1976). This potential contains two contributions : a non-bonded contribution and a contribution due to solvent interactions. The non-bonded contribution is of the from :

$$V_{\text{non bonded}} = \mathcal{E}\left\{3(r^\circ/r)^8 - 4(r^\circ/r)^6\right\}$$

Where r° is the equilibrium distance of the interaction and $\mathcal{E}$ is the depth at the minimum of V non bonded. This is a 6-8 Lennard-Jones type potential which is less steeply repulsive than the usual 6-12 function. The parameters used are listed in Table 1.

The solvent contribution was again taken from Levitt (1976). It is of the form :

$$V_{\text{solvent}} = (S_i + S_j) \cdot g(r_{ij})$$

Where $$g(r_{ij}) = 1 - \tfrac{1}{2}\,(7X^2 - 9X^4 + 5X^6 - X^8)$$

$$X = r_{ij}/r_{max} \quad \text{with } r_{max} \text{ having a fixed value of } 9\text{Å} .$$

The function $g(r_{ij})$ represents the fraction of water lost from atom i due to the approach of atom j. It is a simple sigmoidal function. Its first and second derivatives are zero at X = 1 making the change to $g(rij) = 0$ continuous.

Each side chain is assigned a solvent interaction energy S (see Table 1) mostly estimated from the solubilities of amino-acids in water and ethanol (Nozaki and Tanford, 1971), and the energy due to the interaction of residue i and residue j is $(S_i+S_j).g(r_{ij})$. Thus the number of nearest neighbour contacts is used here to estimate the burried surface area and the resulting free energy gain. We are well aware that the correlation between the number of contacts and the surface area might be weak. Preliminary tests showed however, that excluding the solvent term altogether made the native contact more unstable. We intend to explore ways of improving the solvent term in the future.

TABLE 1.

THE PARAMETERS OF THE SIMPLIFIED MODEL*.

Amino acid Side-chain	CA-R (A)	r° (A)	ε Kcal/mol	S Kcal/mol
Ala	0.77	4.6	0.05	-0.5
Val	1.49	5.8	0.16	-1.5
Leu	2.08	6.3	0.21	-1.8
Ile	1.83	6.2	0.21	-1.8
Cys	1.38	5.0	0.10	-1.0
Met	2.34	6.2	0.21	-1.3
Pro	1.42	5.6	0.16	-1.4
Phe	2.97	6.8	0.39	-2.5
Tyr	3.36	6.9	0.45	-2.3
Trp	3.58	7.2	0.56	-3.4
Asp	1.99	5.6	0.21	2.5
Asn	1.98	5.7	0.21	0.2
Glu	2.63	6.1	0.27	2.5
Gln	2.58	6.1	0.27	0.2
His	2.76	6.2	0.33	-0.5
Ser	1.28	4.8	0.10	0.3
Thr	1.43	5.6	0.10	-0.4
Arg	3.72	6.8	0.39	3.0
Lys	2.94	6.3	0.27	3.0
Gly	0.0	3.8	0.025	0.0

CA-R is the distance of the side chain centroid from the Ca atom.

r°_{ij} is the equilibrium distance calculated from the expression

$$r^\circ_{ij} = (r^\circ_i \cdot r^\circ_j)^{\frac{1}{2}}$$

ij is the depth of the potential function at the minimum,

$$\varepsilon_{ij} = (\varepsilon_i \cdot \varepsilon_j)^{\frac{1}{2}}$$

S is the solvent interaction energy.

* These correspond the rigid side chain geometry and to set C of Levitt (1976).

The total potential function calculated was thus :

$$V_{tot} = \sum_{ij} \varepsilon_{ij} \left\{ 3(r^{o}_{ij}/r_{ij})^{8} - 4\,(r^{o}_{ij}/r_{ij})^{6} \right\} + \sum_{ij} (S_i + S_j).g(r_{ij}).$$

Given that the protein moieties were considered as rigid bodies, only the pairs that contributed to the intermolecular interaction were included in the energy calculations. The evaluation of the averaged potential function, (V_{tot}), and its derivatives was incorporated into the Protein Manipulation Package. This is a package of computer programs written at Columbia University (Levinthal et al., 1975), and which allows the manipulation of protein molecular models. Using these programs one can evaluate and minimize the internal energy of a protein or any part of it. The programs have interactive capabilities with an option to obtain graphic output on a CRT screen. They were modified to run on the 370/168 - IBM computer/Tektronix CRT, at the CIRCE Computer Center at the onset of the workshop.

Refinement tests using the native conformation of the complex.

In order to find out if the averaged potential functions could be used in the study of the association of trypsin and BPTI, several tests were performed starting from the native conformation of the enzyme-inhibitor complex.

The native conformation was refined by minimizing the expression of V_{tot} as a function of the six degrees of freedom which govern the rigid body motion of the molecules.

The refinement lead to structures which had an rms deviation of 2.3 to 3.1 Å from the native conformation when the full potential function described in the previous section was used. The lower rms value, 2.3 Å was obtained when a pulling potential of the form $k(r_{ij}-r^{o}_{ij})$ was used during the first 10 iterations of the refinement. i and j are atoms at the trypsin -BPTI interface and r^{o}_{ij} are the distances between these atoms in the native structure. The higher rms values : 3.0 and 3.1 Å were obtained without pulling potentials but with restrictions on the maximum angular change for each step of the energy minimization procedure. Next, the native structure was refined using only the solvent term. This led to rms values greater than 15 Å. While refinement of the native structure using only the non-bonded term of the potential function led to a conformation with an rms difference from the native complex of 6 Å and in which the two proteins were drawn apart.

The complete averaged potential function : the non-bonded term and the solvent term were therefore used throughout this study.

The coordinate system.

In order to generate all possible intermolecular interactions between the trypsine molecule and BPTI, one had to define the relative position of the molecules as a function of six independent variables, namely the rigid body rotational and translational degrees of freedom. For that purpose, one can consider the position of the trypsin molecule as fixed in space and move the BPTI molecule about it. Such an arrangement is illustrated in Figure 1. The longitude θ_1 and the latitude φ_1 in the frame of the trypsin molecule (molecule 1) determines the direction along which the center of BPTI is placed. Then the longitude θ_2 and latitude φ_2 determines the orientation of the molecular center of Trypsin in the local frame of BPTI or, in other words, the direction of line joining the molecular center in the frame of molecule 2. χ is the rotation of molecule 2 about that line, it determines the relative "spin" of the two molecules and ρ is the center to center distance.

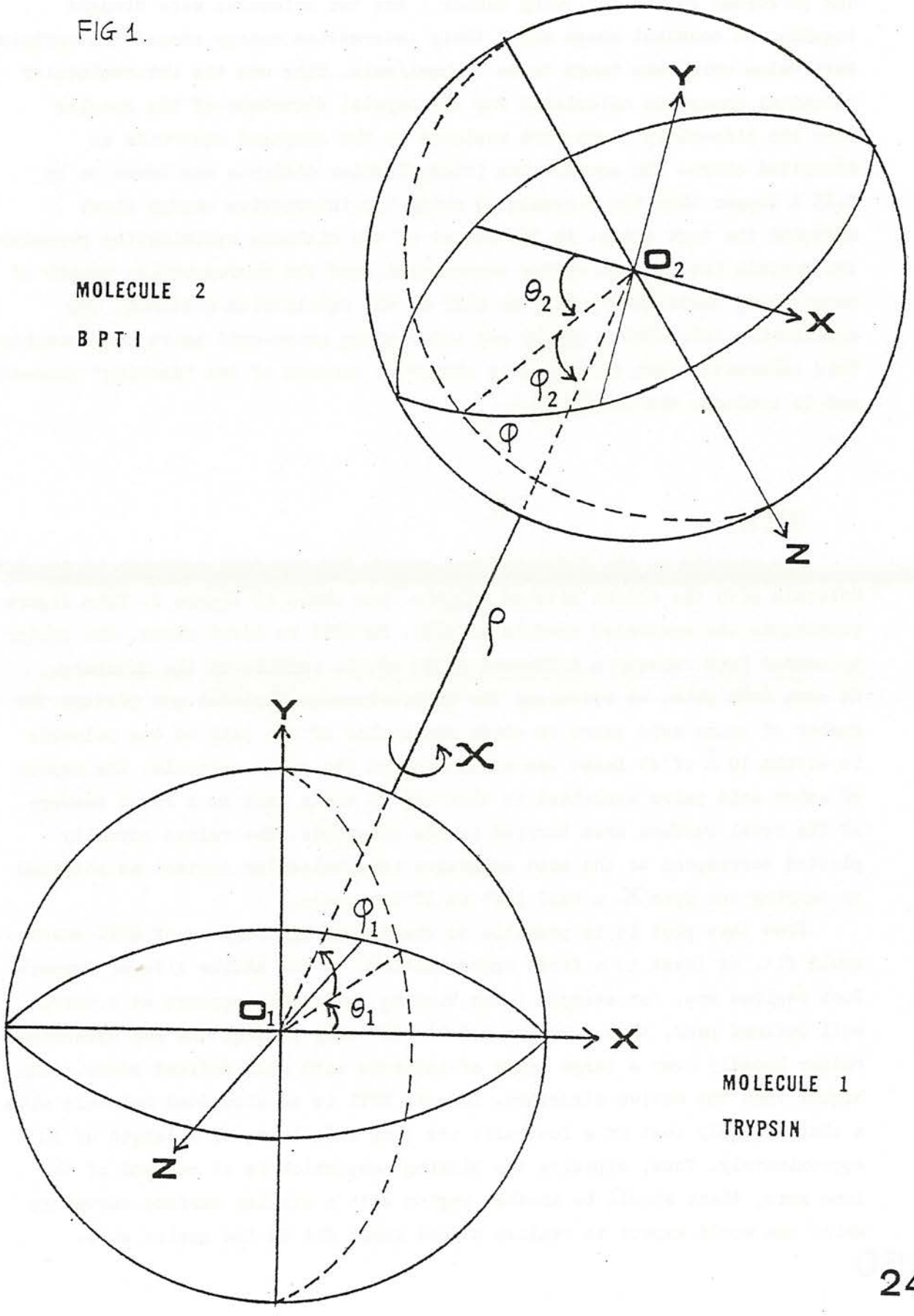
In the study we report here θ_1 and φ_1 were kept fixed at the native position. This amounts to studying all the possible associations of the active site of trypsin with the BPTI molecule. The systematic generation of intermolecular contacts was obtained as a function of θ_2, φ_2 and χ at regular intervals representing a movement on the surface of the BPTI of about 4.5 Å.

FIGURE 1

The coordinate system defining the relative positions of trypsin and BPTI. The arrangement shown corresponds to the native position, or the complex conformation found in the crystal structure. The values of the parameters whose definition is given in the text are:

$\theta_1 = 75.4°$	Corresponds to a rotation about the Y axis of molecule 1.
$\varphi_1 = 63.9°$	" " " " Z axis " " ".
$\theta_2 = 83°$	" " " " Zaxis in molecule 2.
$\varphi_2 = 25°$	" " " " Y axis in molecule 2.
$\chi = -74°$	Is the rotation about the line of centers
$\rho = 25.77$ A	Is the center to center distance.

FIG 1.


MOLECULE 2
BPTI
Y
O_2
θ_2
X
φ_2
φ
Z
ρ
Y
X
φ_1
O_1
θ_1
X
Z
MOLECULE 1
TRYPSIN

At each point a one dimensional search for the optimal "docking" distance was performed in the following manner : the two molecules were brought together in constant steps until their interaction energy exceeded a certain test value which was taken to be 70 kcal/mole. This was the intermolecular potential energy as calculated for the crystal structure of the complex when the side-chain atoms were replaced by the averaged centroids as described above. The equilibrium intermolecular distance was taken to be 0.25 A longer than the distance at which the interaction energy first exceeded the test value. At the outset of the distance optimization procedure, the maximum interaction energy encountered, and the corresponding number of interacting amino acid pairs, as well as the equilibrium distance, the equilibrium interaction energy and interacting amino-acid pairs were recorded. This information was then used to check the success of the "docking" process and to evaluate the interface.

RESULTS.

The results of the 3 dimensional search for the best contacts of the BPTI molecule with the active site of trypsin are shown in Figure 2. This figure represents the enraveled surface of BPTI. As BPTI is moved about, the center to center line crosses a different point on the surface of the molecule. At each such point we optimized the intermolecular distance and plotted the number of amino acid pairs in which one member of the pair on one molecule is within 10 Å of at least one amino acid on the other molecule. The number of amino acid pairs evaluated in this way is taken here as a rough measure of the total surface area burried in the interface. The values actually plotted correspond to the most extensive intermolecular contact as obtained by varying the spin χ a full 360° in 30°intervals.

From this plot it is possible to single out the regions of BPTI which would fit, at least to a first approximation, to the active site of trypsin . Such regions are, for example : the binding loop which appears as a narrow well defined peak, then a region a full 180° away in longitude and spreading rather broadly over a large range of latitude with well defined peaks, some higher than the native structure. Indeed, BPTI is an elongated molecule with a shape roughly that of a football, its long axis being of a length of 30 A approximately. Thus, opposite the binding loop which is at one end of the long axis, there should be another region with a similar surface curvature which one would expect to realize a good rough fit to the active site.

Further analysis of the rough fit found here and the corresponding surface features of the molecule such as curvature and roughness is now in progress.

If curvature complementarity seems to be a promising handle to matching two proteins, the specific interactions involved must be further characterized and examined in greater detail before one could hope to single out the correct mode of association.

FIGURE 2

The enraveled surface of BPTI is shown. On regular intervals on this surface are plotted the number of amino acid pairs which are responsible for the interaction with trypsin when a particular region is brought into close contact with the active site of the enzyme (see text). The molecular Z axis points upward and can be considered as the "north pole" of BPTI. The +Z -Z line corresponds to the longitude of the native interface in the molecular frame of BPTI. It is taken as the reference longitude for the plot, with the -X direction being " Greenwich".

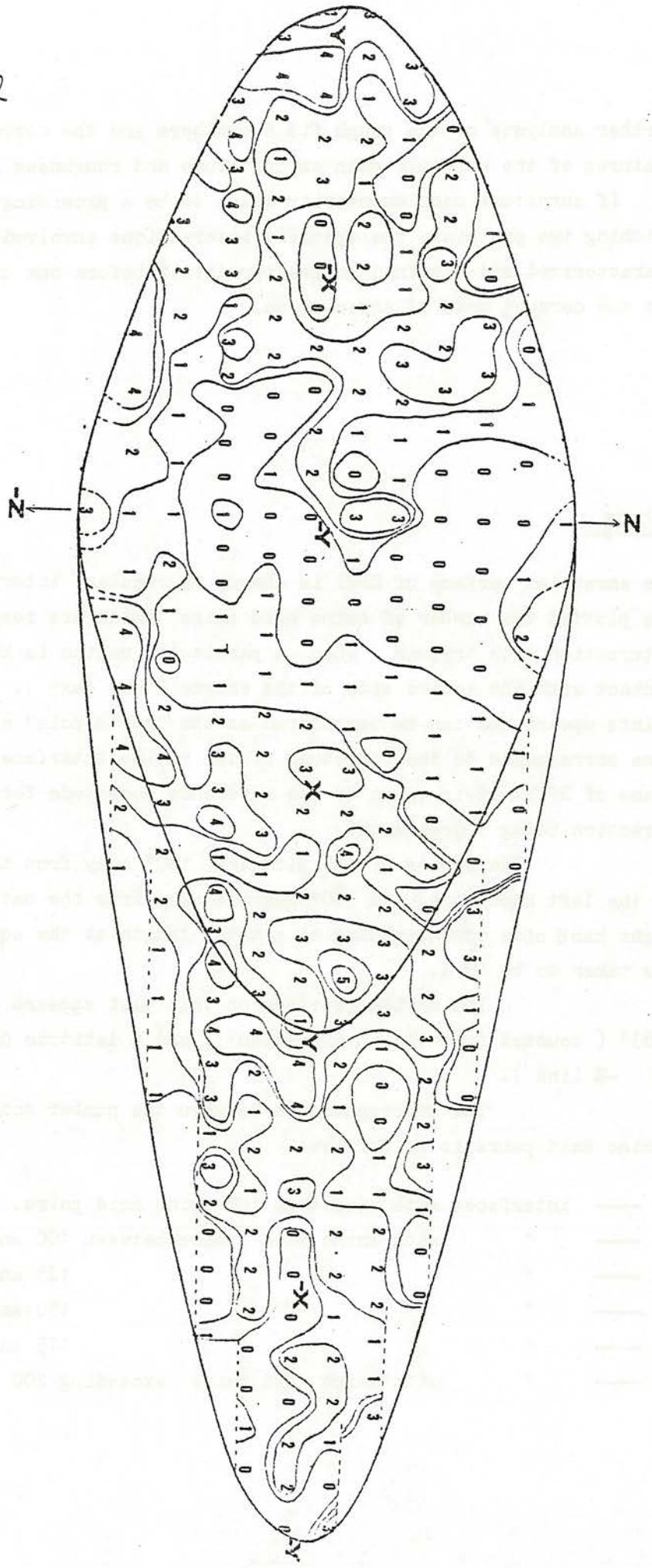
The limits of the plot are 180° away from the native longitude on the left hand side, and 360° degrees away from the native poistion on the right hand side corresponding to a total length at the equator of 3π R where R is taken to be 15 Å.

The native position on this plot appears at a longitude of + 83° (counted from the -X direction), and a latitude of +25° (along the +Z -Z line).

The correspondence between the number code and the number of amino acid pairs is as follows:

0 ——— interfaces with less than 100 amino acid pairs.
1 ——— " with amino acid pairs between 100 and 125
2 ——— " " 125 and 150
3 ——— " " 150 and 175
4 ——— " " 175 and 200
5 ——— " with amino acid pairs exceeding 200

FIG. 2



Towards a more detailed examination of the generated interfaces.

All the interfaces which had more than 150 amino acid pairs were refined using a procedure (Fletcher and Powell,1963) which minimizes the expression of V_{tot} as a function of all six degrees of freedom. Out of 50 refinements only 22 lead to structures with an rms deviation of less than 3.0 Å from the starting grid point. Most of the refined structures with rms values greater than 3 A correspond to a dissociated complex or to an interaction with trypsin further away from the active site. The interpretation of these results in terms of the stability of the interfaces has to be postponed until more and better refinement tests are performed.

Histograms of the number of contacts that the amino acids in BPTI make with trypsin in a given interface was computed for about 300 non refined intermolecular interactions. It was found that extremely few interfaces had at least one amino acid on BPTI with more than 6 contacts with trypsin . The native contact was one of these interfaces. When similar histograms were computed for the interfaces which were first energy refined, only the refined native contact had an amino acid (lys-15') that had 10 contacts with trypsin . Most other interfaces had no amino acid with more than 6 contacts with trypsine.

The correlation of these results with detailed calculations of packing and accessible surface area to solvent seems possible and is to be desired since the latter are expensive computationally and could not be performed on a great number of interfaces. Work in that direction is in progress.

REFERENCES

- Burgess, A. W.; and Scheraga, H.A., (1975), P.N.A.S., U.S.A., 72, 1221.
- Cothia, C. H., (1974), Nature (London), 248, 338.
- Cothia, C. H., and Janin, J. (1975), Nature (London), 256, 705.
- Fletcher R. and Powell, M.P.J., (1963), Comp. J., 6, 163.
- Huber, R., Kukla, D., Bode, W., Schwanger, P., Bartels, K., Deissenhofer, J. and Steigemann W., (1974), J.M.B., 89, 73.
- Kuntz I.D., Crippen, G.M.,Kollman, P.A., and Kimelman , D., (1976), J.M.B., 106, 983.
- Levinthal, C., Wodak S. J., Kahn, P., Dadivanian, A., P.N.A.S., (1975), 72, 1330.
- Levitt, M. and Warshel, A., (1975), Nature 253, 694.
- Levitt, M., (1976), J.M.B., 104, 59.
- Nazaki, Y., and Tanford, C., (1971), J.B.C., 246, 2211.
- Ptytsyn,O.B., and Rashin A. A., (1975), Biophysical Chemistry, 3, 1.

IV

THE POTENTIAL FUNCTIONS FOR PROTEINS

IV.1

A COMPARISON BETWEEN DIFFERENT POTENTIAL FUNCTIONS USED IN THE STUDY OF PROTEIN CONFORMATION

D.R.Ferro

Istituto di Chimica delle Macromolecule del CNR - Via A.Corti 12,
20133 Milano - Italy.

Introduction

As much of the work done during this workshop has shown, the recent advances of mathematical methods (function minimization, Monte Carlo techniques, molecular dynamics) and of computing facilities offer concrete possibilities to the simulation of the behaviour of proteins. This progress makes even more crucial the ability to calculate the molecular potential energy surface with good accuracy. Theoretical approaches to determine energy and structure of large flexible molecules can be divided into three types: empirical, semiempirical quantum mechanical and ab initio quantum mechanical methods. The empirical or force field method consists in the partitioning of the molecular energy into additive terms, such as dispersion, repulsion and coulombic interactions between non-bonded atoms, hydrogen-bond energy, torsional barriers, etc. This method was first used by Liquori and coworkers [1] for synthetic polymers, and it has been extensively applied to polypeptide chains, mainly by Scheraga's group [2,3]. The parameters that enter in the expression of the molecular energy are in part derived by fitting experimental data, in part based on theoretical estimates.
Semiempirical molecular orbital methods, such as EHM [4], CNDO [5], PCILO [6], have been applied to model compunds and in some instances directly to the study of the stable conformations of polymers, notably by Pullman and his group on polypeptides and other biopolymers.

Semiempirical methods have been used also to derive some parameters needed in the empirical approach (partial charges, hydrogen-bond potentials). However no author, to my knowledge, has attempted to express the results of semiempirical calculations in the form of a set of analytical potential functions, suitable for example for energy refinement or molecular dynamics of proteins.
Ab initio computations, in the SCF-LCAO-MO approximation, have been used by Clementi and coworkers [7] to derive analytical interatomic potentials for the interaction between water and aminoacids. It is likely that complete sets of ab initio interatomic potentials for the study of proteins will be available in the near future.

In this appendix we shall deal only with the empirical method, confining ourselves to the potential functions for the atoms commonly found in proteins, i.e. carbon, hydrogen, nitrogen, oxygen and sulfur. Also, this is far from being a complete review of force fields used so far in the conformational analysis of polypeptides. I have selected five sets of functions among the most recent ones that have been proposed and applied to proteins by the various groups of workers in this field. A general survey of the force-field method is presented in the recent article by Allinger [8]. Information about other force fields can be found in the reviews by Ramachandran and Sasisekharan [9], by Venkatachalam and Ramachandran [10] and by Scheraga [2, 3], and in the book by Hopfinger [11].

Molecular energy

The expression of the potential energy of a protein molecule used by most authors can be written in the general form:

$$
\begin{aligned}
E_{intra} &= E_{bonded} + E_{nonbonded} \\
E_{bonded} &= E_{str} + E_{bend} + E_{tors} \\
E_{nonbonded} &= E'_{tors} + E_{VDW} + E_{el} + E_{HB}
\end{aligned}
\qquad (1)
$$

Also the interaction between two or more molecules is usually represented by the sum of terms analogous to the last three terms of the intramolecular nonbonded energy:

$$E_{inter} = E_{VDW} + E_{el} + E_{HB} \qquad (2)$$

This expression may well include the case of interactions between macromolecule and solvent; some authors [12,13], however, have attempted to represent the average interaction with solvent by means of an effective free energy term to be added to equation (2).

The E_{bonded} contribution to the intermolecular energy arises from deformations of bond lengths (E_{str}), bond angles (E_{bend}) and torsional angles (E_{tors}) around double bonds or from similar distorsions of highly rigid groups of atoms. This term is obviously included only by those authors [14,15,16], who allow other degrees of freedom besides the torsional angles around single bonds. Recent results [17] indicate that the calculated structure of a globular protein may be significantly affected by not releasing at least some degrees of freedom, such as the peptide bond dihedral angle ω or the bond angles at the C^{α} atom.

For some homologous series of compounds, e.g. the saturated hydrocarbons, sophisticated force fields have been proposed [18,19,20], which reproduce structures and thermodynamical data with good accuracy. These force fields are similar to those (e.g. Urey-Bradley's force field) earlier proposed by spectroscopists to reproduce vibrational frequencies, but their parameters have been derived by fitting a great deal of structural and thermodynamical data as well as spectroscopic data. For systems containing heteroatoms, as the peptides, less abundant information is available and simpler potentials, of the valence-force-field type, have been used so far. Here E_{bonded} is expressed by quadratic functions corresponding to stretching, bending and torsional deformations with respect to the ideal geometry:

$$E_{str} = \Sigma \tfrac{1}{2} K_{\ell} (\ell-\ell_o)^2$$

(3) $$E_{bend} = \Sigma \tfrac{1}{2} K_{\theta} (\theta-\theta_o)^2$$

$$E_{tors} = \Sigma \tfrac{1}{2} K_{\chi} (\chi-\chi_o)^2$$

The summation for the torsional energy is carried over those dihedral angles χ showing a large barrier to rotation, say > 20 Kcal/mole, and includes also terms representing out-of-plane deformations of planar groups. Values of the force constants K_{ℓ}, K_{θ}, K_{χ} and of the ideal parameters ℓ_o, θ_o and χ_o have been selected by Levitt [15]; similar values have been used by Hermans et. al.[16]. Levitt and Lifson [14], Winkler and Dunitz [21] and Ramachandran et. al. [22] have proposed potentials for the distortion from planarity of the peptide bond.

At the present time it seems that these "bonded potentials", although they may be too crude approximations for other purposes, can be applied satisfactorily to conformational analysis of proteins.

Nonbonded energy

The torsional contribution E'_{tors} to the nonbonded energy is given by

(4) $$E'_{tors} = \Sigma \tfrac{1}{2} V_o \left\{1-\cos\left[n(\chi-\chi_o)\right]\right\}$$

where the summation is carried over torsional angles around single bonds.

This term must be included in the expression of the nonbonded energy because the van der Waals (and electrostatic) energy alone predicts much lower rotational barriers than the values observed in small molecules.

Therefore the values of V_o are chosen so that the sum of all terms best fit the spectroscopic data. Some authors[12,23], however, have omitted the van der Waals interactions between atoms separated by

three bonds, taking the values of V_o equal to the observed barriers. Selected values of V_o, n and χ_o can be found in the works of Scheraga [2,24]. In general, due to the just mentioned procedure to determine V_o, the torsional term is not the source of major discrepancies between the various sets of nonbonded potentials.

The van der Waals energy, which accounts for steric repulsions at short interatomic distances and for the attractive dispersion forces at larger separations, is commonly expressed in the form of sums of interatomic interactions, pairwise additivity being assumed. Lennard--Jones (6-12) or Buckingham (6-exp) potentials are usually chosen to represent each interaction as a function of the interatomic distance. We can combine these two forms of potentials by using the expression

$$U_{ij} = -A\ r_{ij}^{-6} + B\ r_{ij}^{-n} \exp(-\beta\ r_{ij}) \tag{5}$$

where r_{ij} is the distance between atoms i and j, and the parameters A,B,β and n depend on the chemical types of two atoms. The total van der Waals energy is then given by

$$E_{VDW} = \sum_{i<j} U_{ij} \tag{6}$$

where the summation is extended to pairs of atoms separated by at least three bonds. From the summation interactions are excluded between the hydrogen and the acceptor atoms of hydrogen bonds, which are treated with special potentials; however the interaction between the donor (i.e. the atom covalently bonded to H) and the acceptor is usually calculated with normal nonbonded potentials. As mentioned above some authors omit interactions between atoms separated by three bonds, while others use a different set of potentials for these interactions; we shall return to this point later on.

In the recent years authors have made more and more use of information available from known crystal structures of model compounds to calibrate potential functions for conformational studies of polypeptides. Here we shall examine five sets of potentials obtained by relying in various degrees on crystal structure data. Among several other potentials not taken into consideration for obvious limitations, here I shall simply quote the set of functions used by De Santis and Liquori [25], somewhat similar to set III selected by Giglio, and the functions used by Warshel and Levitt [23], related to set IV proposed by Lifson's group.

Karplus' group adopts Lennard-Jones potentials [26] (set V), obtained by the method first used by Scott and Scheraga [27]. The coefficients A_{ij} of the attractive part of equation (5) are calculated from the Slater-Kirkwood expression of the dispersion energy:

$$A_{ij} = \frac{(3/2)e(\hbar/m^{\frac{1}{2}})\ \alpha_i\alpha_j}{(\alpha_i/N_i)^{\frac{1}{2}} + (\alpha_j/N_j)^{\frac{1}{2}}} \qquad (7)$$

where e is the electronic charge, m the electronic mass, α_i and α_j the atomic polarizabilities, and N_i and N_j are effective numbers of electrons for atoms i and j. Crystal structure data are utilized indirectly in the determination of the repulsive parameters of the potentials, which are obtained by requiring that U_{ij} be a minimum at the distance r^o_{ij} equal to the sum of the van der Waals radii. This procedure presents the problem that the interatomic contacts observed in the crystals, generally used to determine the van der Waals radii, are actually significantly shorter than the distance r_o of minimum energy, due to the compressing net effect of the other interactions. For example, Scott and Scheraga's potentials lead to unacceptably low values of the calculated lattice constants [28]; these potentials, however, have been applied successfully in several intramolecular energy calculations. The above procedure has been used to calculate only the coefficients A_{ii} and B_{ii} of set V; for interactions between unlike chemical types the following averaging rules have been assumed:

$$r^o_{ij} = (r^o_{ii} + r^o_{jj})/2 \qquad (8)$$

$$\varepsilon_{ij} = (\varepsilon_{ii} \cdot \varepsilon_{jj})^{\frac{1}{2}}$$

where $-\varepsilon$ indicates the minimum energy and r^o the corresponding distance.

Giglio [29] has analyzed various nonbonded potentials proposed in the literature by calculating the minimum energy structure of several crystals in which a specific interatomic interaction is predominant. By comparison of calculated and observed lattice constants he has selected a set of "best" potentials, including set III shown here as well as functions for halogenic atoms.

Other authors [30,31,32] have utilized crystal structure data in a more systematic manner; they have obtained consistent force fields by fitting simultaneously structures and lattice energies of several crystals. These force fields include electrostatic and H-bond terms as well as the van der Waals energy; all these terms should be used consistenly. The various derivations of these force fields differ in the choice of the crystals used in the fitting, in the a priori assumption of some parameters of the force field, and in the averaging of interactions between different chemical types.

In the Lennard-Jones potentials given by Scheraga and coworkers[24,31] (set I), the attractive coefficients are calculated from Slater--Kirkwood's equation (7), while no use is made of equations relating the lattice energy to the heat of sublimation. The electrostatic term is calculated on the basis of point charges obtained by the CNDO/2(ON) method. For hydrogen-acceptor interactions of H-bonds the 6-12 potentials are replaced by (10-12)potentials, whose attractive coefficients are also derived from CNDO calculations. The following averaging rules are applied:

$$r^o_{ij} = (r^o_{ii} + r^o_{jj})/2 \tag{9}$$

$$\varepsilon_{ij} = A_{ij}/2\ (r^o_{ij})^6$$

where A_{ij} is obtained from equation (7). Therefore only the B_{ii} (including those of 10-12 potentials) are the independent best-fitted parameters.

Ferro and Hermans [30] derived A's and B's of Lennard-Jones potentials from crystal structures of non hydrogen-bonded organic molecules; available heats of sublimation were equated to the calculated lattice energies, since equilibrium conditions alone were not sufficient to determine the magnitude of the interactions. Mixed coefficients A_{ij} and B_{ij} were calculated as geometric averages, while the atomic charges for the electrostatic term were obtained with the Del Re--Pullman method previously used by Poland and Scheraga [33]. These potentials, combined with Poland and Scheraga's H-bond potential gave satisfactory results in the calculation of crystal structures of dipeptides [28], but proved to be unsuitable for intramolecular energy calculations. Nelson and Hermans [34] corrected these functions in order to account for the thermal expansion, obtaining a less repulsive set of potentials to be used with macromolecular systems (set II).

Hagler et al. [32] have derived an intermolecular force field for amides by a least-squares fitting of heats of sublimation, dipole moments and unit cell vectors of nine crystals. The van der Waals potentials are either Lennard-Jones (set IV considered here) or 6-9 functions, geometric averages being assumed for A_{ij} and B_{ij} coefficients in both cases. The atomic charges were not taken from quantum mechanical calculations, but were obtained by the least-squares fitting. These authors found that hydrogen bonding could be accounted for by using the van der Waals and electrostatic terms only and omitting the nonbonded interactions for amide hydrogens H_N.

The coefficients of the five sets of potentials, for atom pairs of the same type, are given in Table I. For convenience of the reader also the mixed type coefficients are given in Tables II, III and IV for sets I, III and V of functions respectively. Coefficients involving the sulfur atom are not listed in Tables II and IV, since the available values of α_s and N_s for equation (7) seem inconsistent with the given A_{ss}.

The parameters of the potentials obtained by fitting to experimental data are highly correlated. Therefore each force field should

TABLE I

Nonbonded potential functions [a]. Interactions between atoms of the same chemical type.

Atom type	Valence state or chemical environment	Set	A	$Bx10^{-3}$	n	β	r_o	ε_o
C_1	aliphatic carbon	I[b]	370.5	906.03	12	0	4.12	.038
C_2	carbonyl, carboxyl, carboxylate and peptide bond	I	766.6	1048.98	12	0	3.74	.141
C_3	aromatic and other unsaturated	I	370.5	475.30	12	0	3.70	.073
C1	saturated carbon	II[c]	385	915	12	0	4.10	.040
C2	aromatic and unsaturated	II	560	805	12	0	3.77	.097
C	all carbons	III[d]	327.2	301.2	12	0	3.50	.089
C1	saturated carbon	IV[e]	532	1811	12	0	4.35	.039
C2	amide carbon	IV	1340	3022	12	0	4.06	.148
C	all carbons	V[f]	391.8	426.5	12	0	3.60	.090
H1	aliphatic hydrogen	I	45.5	14.1	12	0	2.92	.037
H2	amine and amide	I	45.5	8.43	12	0	2.68	.062
H3	aromatic and sulfhydryl	I	45.5	14.39	12	0	2.93	.036
H4	hydroxyl and carboxyl	I	45.5	11.69	12	0	2.83	.044
H	all hydrogens	II	20.4	3.55	12	0	2.65	.029
H	all hydrogens	III	49.2	6.6	0	4.08	2.98	.036

./.

TABLE I (cont.)

Atom type	Valence state or chemical environment	Set	A	$Bx10^{-3}$	n	β	r_o	ε_o
H_1	aliphatic hydrogen	IV	32.9	7.15	12	0	2.75	.038
H_2	amide hydrogen	IV	0	0	0	0	0	0
H	all hydrogens	V	5.75	1.842	12	0	2.94	.0045
N_1	amide and NH_3^+ nitrogen	I	363.1	732.54	12	0	3.99	.045
N_2	amine nytrogen	I	401.0	374.94	12	0	3.51	.107
N	all nitrogens	II	580	405	12	0	3.34	.208
N	all nitrogens	III	354	387	12	0	3.60	.081
N	amide nitrogen	IV	1230	2271	12	0	3.93	.167
N	all nytrogens	V	289.5	131.0	12	0	3.11	.160
O_1	C=O oxygen	I	369.0	170.19	12	0	3.12	.200
O_2	C-O- oxygen	I	217.2	125.63	12	0	3.24	.094
O	all oxygens	II	540	177	12	0	2.95	.412
O	all oxygens	III	358	259	12	0	3.36	.124
O	amide oxygen	IV	502	275	12	0	3.21	.228
O	all oxygens	V	312.5	105.9	12	0	2.96	.230
S	all sulfur atoms	I	249	363.18	12	0	3.78	.043
S	all sulfur atoms	II	2560	4540	12	0	3.91	.361
S	all sulfur atoms	III	1430	220.8	0	3.621	3.87	.244
S	all sulfur atoms	V	286	432	12	0	3.80	.047

[a] In this table and in tables II, III and IV energies are expressed in Kcal/mole and distances in Å. [b] References 24 and 31. [c] Reference 34. [d] Reference 29. [e] Reference 32. [f] Reference 26.

TABLE II

Parameters of mixed interactions for set I of 6-12 nonbonded potential functions

Atom pair	A	$Bx10^{-3}$	Atom pair	A	$Bx10^{-3}$	Atom pair	A	$Bx10^{-3}$	Atom pair	A	$Bx10^{-3}$
$C_1..C_2$	529.1	974.7	C_1 H_1	125.7	119.5	C_1 N_1	366.2	814.0	C_1 O_1	367.9	414.0
$C_1..C_3$	370.5	661.9	C_1 H_2	125.7	97.1	C_1 N_2	385.3	593.9	C_1 O_2	278.8	346.2
$C_2..C_3$	529.1	701.1	C_1 H_3	125.7	120.6	C_2 N_1	519.3	865.5	C_2 O_1	519.1	422.7
$H_1..H_2$	45.5	10.96	C_1 H_4	125.7	110.7	C_2 N_2	547.5	621.2	C_2 O_2	389.3	351.7
H_1 H_3	45.5	14.25	C_2 H_1	185.1	126.2	C_3 N_1	366.2	591.7	C_3 O_1	367.9	289.2
H_1 H_4	45.5	12.85	C_2 H_2	185.1	101.3	C_3 N_2	385.3	422.9	C_3 O_2	278.8	243.4
H_2 H_3	45.5	11.08	C_2 H_3	185.1	127.3	H_1 N_1	122.5	104.2	H_1 O_1	121.8	46.2
H_2 H_4	45.5	9.95	C_2 H_4	185.1	116.3	H_1 N_2	129.4	71.4	H_1 O_2	90.4	38.6
H_3 H_4	45.5	12.98	C_3 H_1	125.7	82.7	H_2 N_1	122.5	84.3	H_2 O_1	121.8	36.2
N_1 O_1	365.7	369.1	C_3 H_2	125.7	66.2	H_2 N_2	129.4	56.9	H_2 O_2	90.4	30.4
N_1 O_2	278.4	310.7	C_3 H_3	125.7	83.4	H_3 N_1	122.5	105.1	H_3 O_1	121.8	46.7
N_2 O_1	383.9	254.7	C_3 H_4	125.7	76.1	H_3 N_2	129.4	72.1	H_3 O_2	90.4	39.0
N_2 O_2	291.9	215.7	N_1 N_2	381.7	530.7	H_4 N_1	122.5	96.3	H_4 O_1	121.8	42.2
			O_1 O_2	282.0	517.0	H_4 N_2	129.4	65.7	H_4 O_2	90.4	35.3

TABLE III

Parameters of mixed interactions for set III of nonbonded potential functions

Atom pair	A	$Bx10^{-3}$	n	β
C..H	125.0	44.8	6	2.04
C..N	340.0	340.0	12	0
C..O	342.3	278.7	12	0
C..S	684.0	255.4	6	1.811
H..N	132.0	52.1	6	2.04
H..O	132.7	42.0	6	2.04
H..S	265.2	40.5	0	3.851
N..O	356.0	316.2	12	0
N..S	711.5	288.6	6	1.811
O..S	715.5	239.2	6	1.811

TABLE IV

Parameters of mixed interactions for set V of 6-12 nonbonded potential functions

Atom pair	A	$Bx10^{-3}$
C..H	49.0	29.83
C..N	342.3	244.1
C..O	359.9	224.9
H..N	40.9	15.61
H..O	42.4	13.97
N..O	301.3	118.2

be considered (and utilized) as a whole. Keeping this fact in mind, it seems useful, nevertheless, to analyze common features or striking differences in Figures 1-7, where the various functions for atom pairs of the same type are plotted versus the interatomic distance.

As more data have become available and have been utilized in the parametrization of force fields, several authors have begun to distinguish interatomic potentials according to the valence state, the charge or the chemical environment of the atoms. A more rigorous, but similar, criterium is being used in deriving ab initio potentials.[7] Figures 1 and 2 show that in sets I, II and IV, where different potentials are given for saturated and unsaturated carbons, the latter present much deeper minima and shorter r^{o} distances. This fact may be due to the larger mobility of π electrons; it is significant that larger values of A for unsaturated carbons than for aliphatic carbons occur in sets II and IV, obtained without use of atomic polarizabilities. The carbon-carbon functions of sets III and V are rather similar to those for unsaturated carbons of the other sets.

Only Scheraga has distinguished different types of hydrogen atoms, finding less repulsive potentials for hydrogens of polar bonds than for the similar potentials of aliphatic and aromatic hydrocarbon hydrogens (Figure 3). The H-H functions of sets II, III and IV were essentially derived from hydrocarbons and are rather close to Scheraga's potentials. We also observe that the distance of minimum energy is in all cases significantly larger than twice Pauling's van der Waals radius of 1.2 A. Finally we note that the H-H potential of set V is unusually shallow due to the very small value assumed for α_H (Figure 4).

Large discrepancies are observed in Figure 5 among the various N-N potentials. The differences may be partially explained (and balanced, in the applications) by the different treatments of the electrostatic and H-bond terms. The very deep minimum of potential II probably depends on having been derived mainly from heterocyclic molecules having aromatic character. Further work appears necessary, possibly considering more than one type of nitrogen, as done by Scheraga.

FIGURE 1

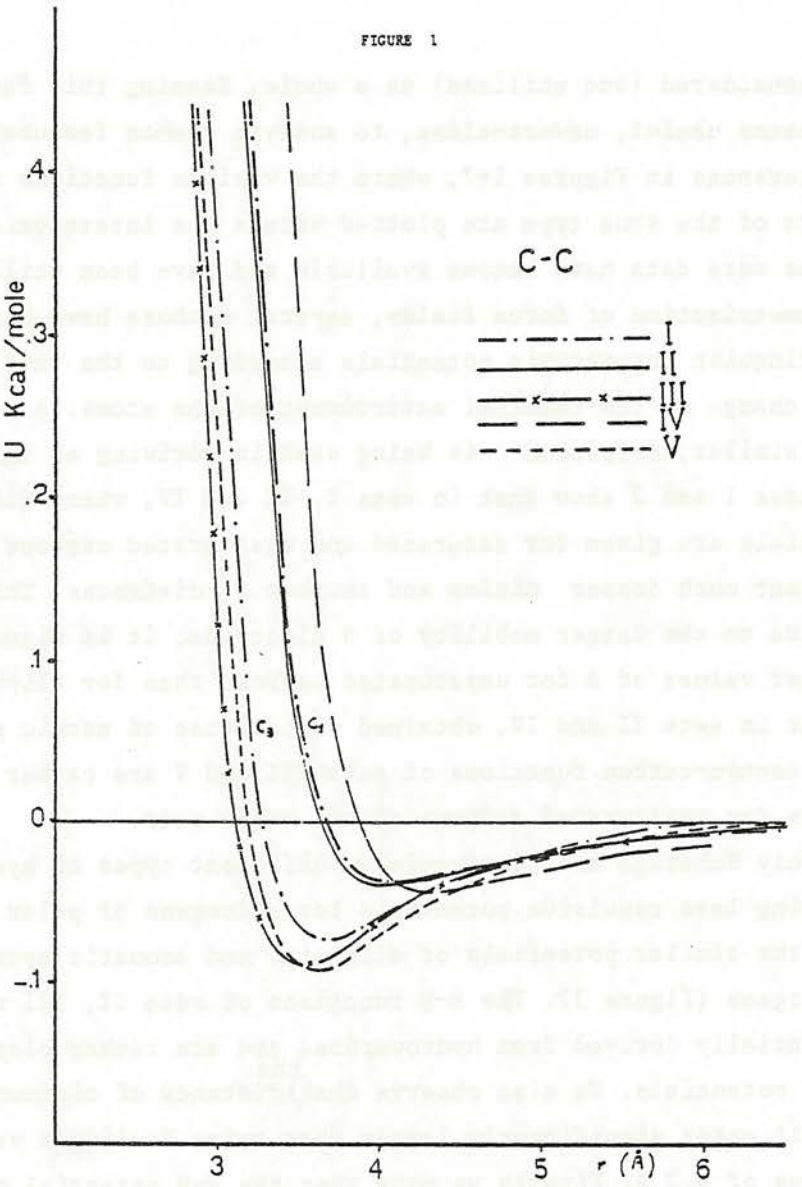


Fig. 1 Nonbonded potentials for the carbon-carbon interaction plotted versus the interatomic distance: functions for saturated carbons of sets I,II and IV, together with the only one type of carbon potential for sets III and V. The aromatic carbon potential of set I is shown for comparison.

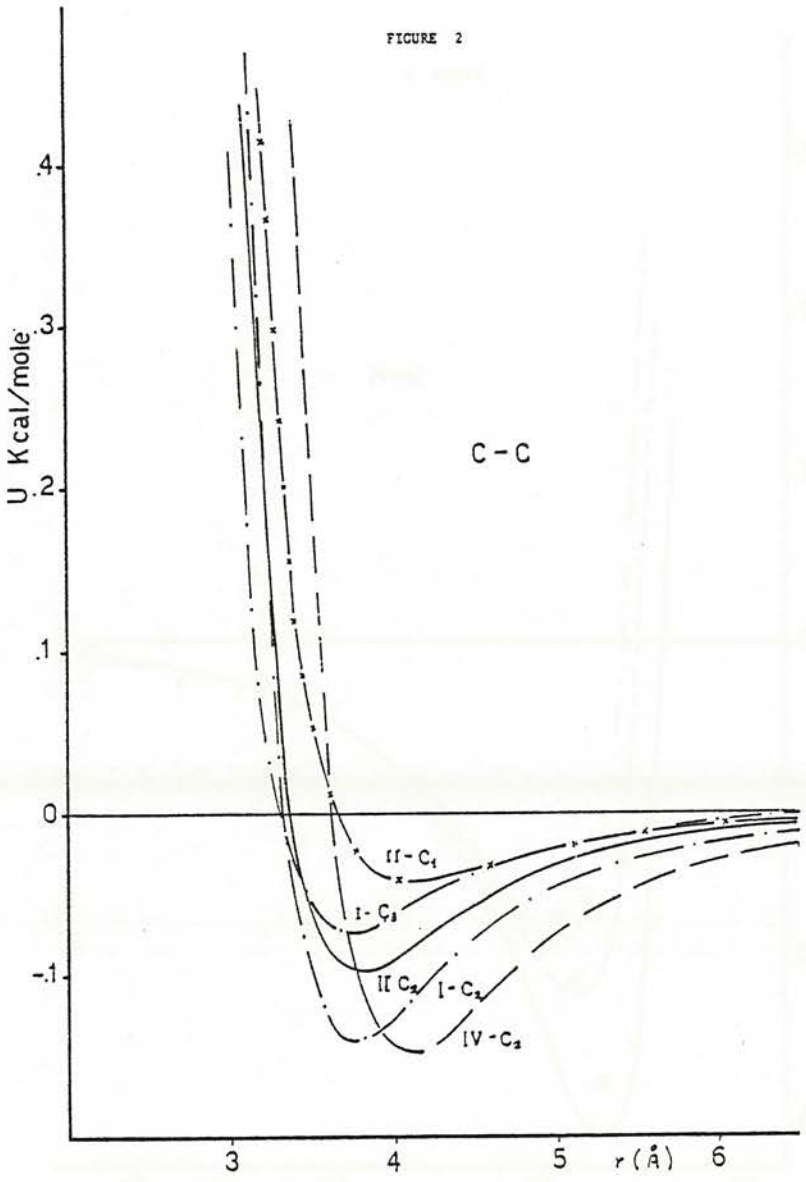


Fig. 2 Nonbonded interatomic potentials for the carbon-carbon interaction: functions for unsaturated carbons of sets I, II and IV, together with the aliphatic carbon potential of set II shown for comparison.

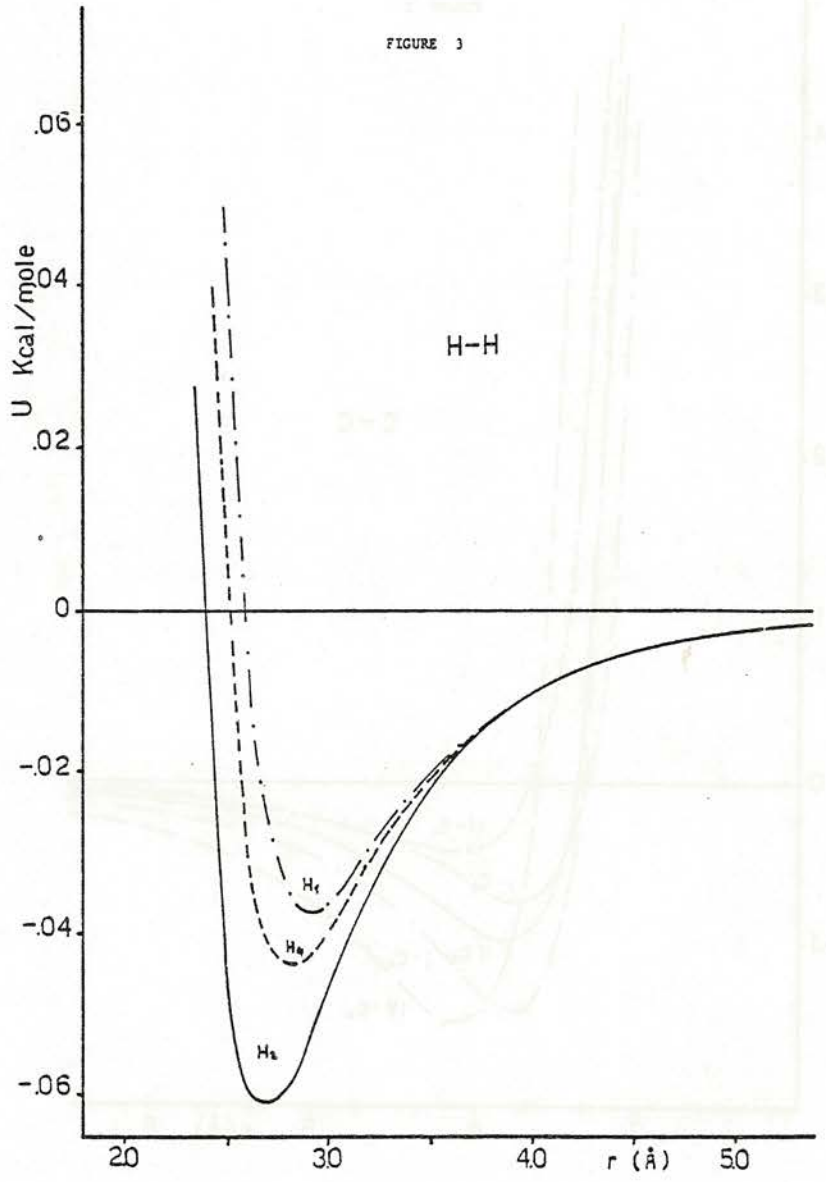


Fig. 3 Nonbonded interatomic potentials for the hydrogen-hydrogen interaction: the functions of set I. The potential for H_2 has been omitted because almost identical with H_1.

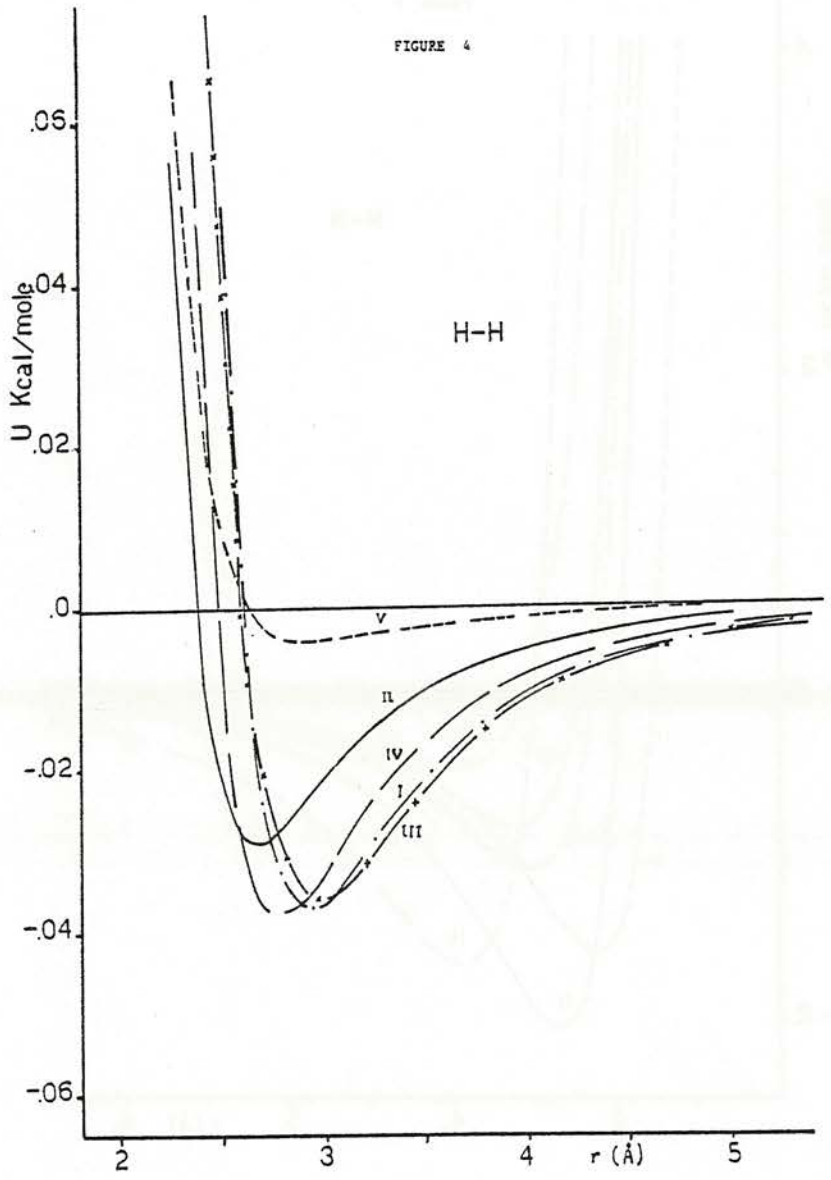


Fig. 4 Nonbonded interatomic potentials for the hydrogen-hydrogen interaction; the H_1 function is shown for set I.

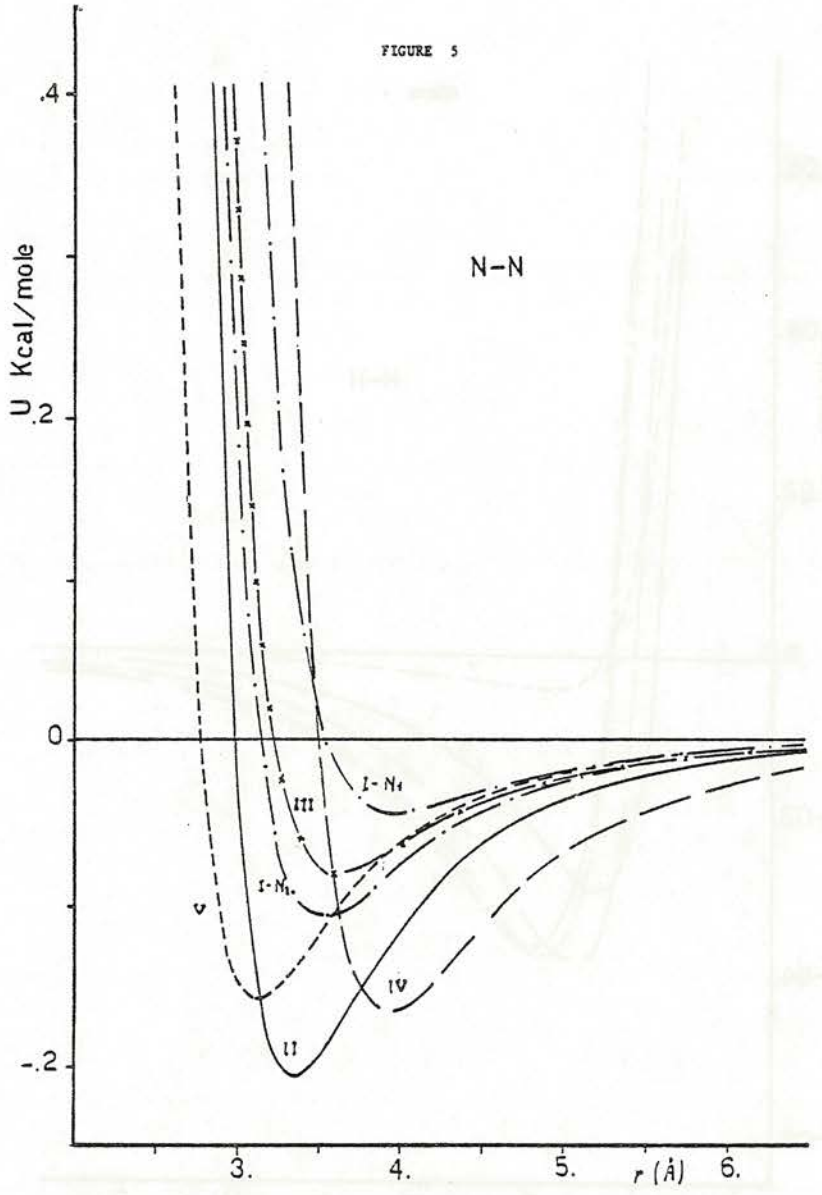


Fig. 5 Nonbonded interatomic potentials for the nitrogen-nitrogen interaction.

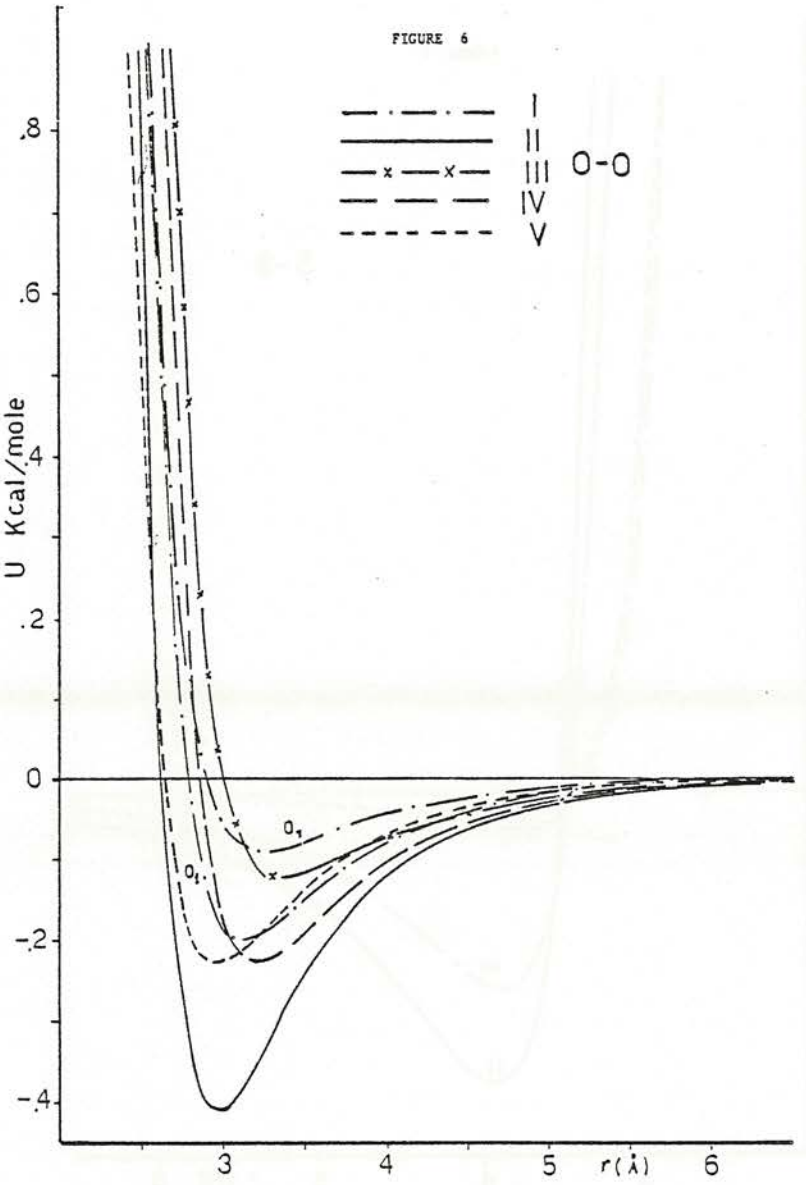


Fig. 6 Nonbonded interatomic potentials for the oxygen-oxygen interaction.

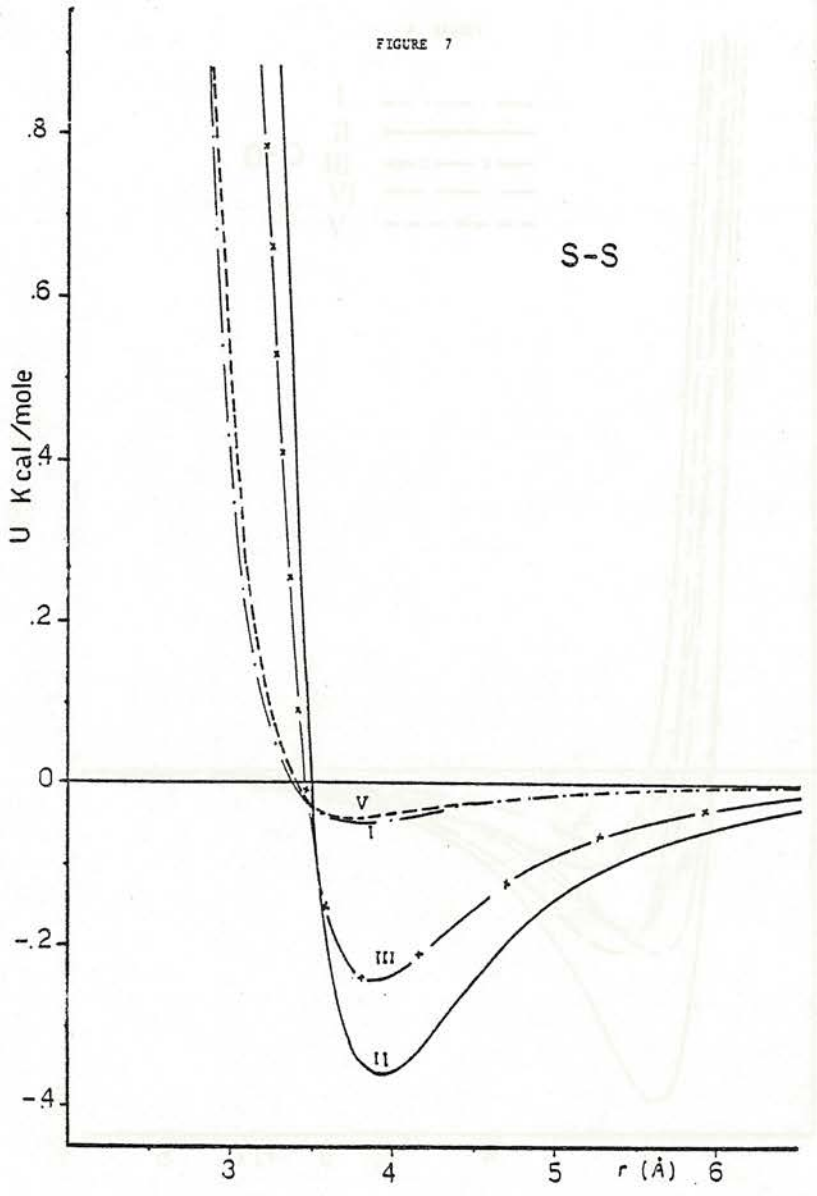


Fig. 7 Nonbonded interatomic potentials for the sulfur-sulfur interaction.

The various oxygen-oxygen potentials show a good agreement in the position of the minimum, while the values of ε vary considerably. Again potential II shows the deepest minimum, due partially to the significant temperature correction and perhaps again to the compounds used in the fitting. The function of set I for the carbonyl oxygen is less repulsive than that for -O- atoms, in agreement with the previous arguments on unsaturated systems.

Finally, the similar sulfur-sulfur potentials of sets I and V, based on quite small values of α_s, appear unable to account for the heat of sublimation of elemental sulfur. On the other hand the attractive coefficient of potential II may have been affected (i.e. increased) by the truncation of the lattice sum $\Sigma\, r^{-6}$ in the fitting of the heat of sublimation.

Electrostatic and hydrogen-bond energy

Most authors include an electrostatic term in the expression of the nonbonded energy. This term is usually calculated by using the monopole approximation:

$$E_{el} = \Sigma_{i<j} \frac{q_i\, q_j}{D\, r_{ij}} \qquad (10)$$

where the q's are point charges, normally located at the positions of the atoms, and the D is an effective dielectric constant. The atomic charges are either fitted to experimental data or obtained by means of semiempirical or ab initio quantum-mechanical methods. Here we shall consider only the charge distributions used in connection with the five sets of van der Waals potentials.

Giglio [29] considers the electrostatic term only in the presence of ionic groups and of hydrogen-bonds, in the latter case in the form of point dipole-dipole interactions. The point charges used by the other authors for a glycyl peptide unit are listed as an example in Table V. The dipole moment μ of the rigid peptide unit (from

C_i^α to C_{i+i}^α, not including the contributions of other atoms bonded to the C^α's) calculated with the respective charges is also given, together with the angle θ formed by $\vec{\mu}$ with the $C_i^\alpha \rightarrow C_{i+i}^\alpha$, direction.

TABLE V

Atomic charges for the glycyl peptide unit[a]

Set	C^α	H^α	C'	O	N	H_N	μ	θ
I[b]	-.008	.055	.450	-.384	-.344	.176	3.61	74.0°
II[c]	.000	.051	.318	-.422	-.202	.204	3.70	88.3°
IV[d]	-.200	.100	.380	-.380	-.280	.280	3.58	100.6°
V[e]	.070	.008	.433	-.428	-.274	.183	3.61	116.5°

[a]Charges in electron units, dipole moments in Debye. [b]Reference 24. [c]Reference 33. [d]Reference 32. [e]Reference 26.

We observe that, while there are large differences in some single atomic charges, the resultant dipole moments show very close absolute values. Notice that charges V were obtained as least-squares parameters together with the van der Waals coefficients.

Much confusion still exists with regard to the value of the dielectric constant to be used in equation (10). The authors of sets II, IV and V take D=1, while Scheraga [24] uses D=2. Values of 3 or 4 have been used in the past (of course, apart from the effects of polar solvents), on the basis of experimental values of the macroscopic dielectric constant ε. The parameter D should account for the fact that in the expression of the nonbonded energy we neglect polarization terms and use fixed atomic charges calculated for isolated molecules. I would like to quote two results, which argue in favour of a value of D

close to 1. In a precise work on Monte Carlo simulation of liquid water, Lie et al. [35] used a pair potential fitted to the results of ab initio computations. The potential includes a coulombic term based on D=1 and on point charges (determined by the energy best fitting) that yield a dipole moment for water even larger than the experimental value. In the application of their force field (I) to amino-acid crystal structures, Scheraga and coworkers [36], using D=2, calculate a binding energy of -26 Kcal/mole for α-glycine. This value appears to be far too low when compared to the sum of the reported heat of sublimation (31 Kcal/mole) and the estimated proton transfer energy (from 60 to 100 Kcal/mole [37,38]).

The first hydrogen-bond potential for the calculation of polypeptide conformations was proposed by De Santis et al. [39], who used the Stockmayer relation for the interaction between polar molecules. This potential, which includes a 6-12 function and one electrostatic term in the form of a dipole-dipole interaction, is still used by Giglio. Scott and Scheraga [40] adopted the potential of Lippincott and Schroeder, but since Poland and Scheraga [33], most authors have used simpler expressions, in which a properly modified van der Waals term for the interaction between hydrogen and acceptor allows the closer approaching of the two atoms and most of the H-bond energy is accounted for by the coulombic term.

Hermans et al. [16] still use Poland and Scheraga's (6-12) potential, slightly adjusted to set II of functions, while Scheraga [24] has proposed (10-12) functions to be used with set I; these functions are taken also by Karplus' group for set V of potentials. The coefficients for a few hydrogen-bond potentials of this type are listed in Table VI. As we have mentioned earlier, Lifson's group has found that no hydrogen-acceptor potential is needed in set IV.

TABLE VI

Coefficients for (m-12) potentials for the hydrogen-acceptor interaction

Set	Atom pair	m	A	B
I	$H_2..O_1$	10	12,040	4014
I	$H_4..O_1$	10	13,344	5783
I	$H_2..N$	10	32,897	8244
II	$H_2..O_1$	6	40	1600

Interaction between two molecules of formyl-tri-glycyl amide

A more meaningful comparison of the force fields examined above can be obtained by applying them to the study of some model system. I have chosen to calculate the interaction between two molecules of formyl-tri-glycyl amide (FTGA) in a few relative configurations. The purpose of these calculations is not to test the suitability of the various sets of functions for protein conformational studies, but rather to see whether it is possible to characterize the behaviour of each force field as a whole. The FTGA molecule is drawn in Figure 8 in its planar fully-extended conformation: the heavy atoms lie in the xy plane, with the two-fold screw axis (of the infinite chain) coinciding with y. Two groups of calculations were performed in which one FTGA molecule was kept fixed at the position of Figure 8 and an identical copy, respectively parallel and antiparallel to the first one in the two cases, was translated along the three cartesian axes. The configurations of minimum energy are listed in Tables VII, VIII and IX.

FIGURE 8

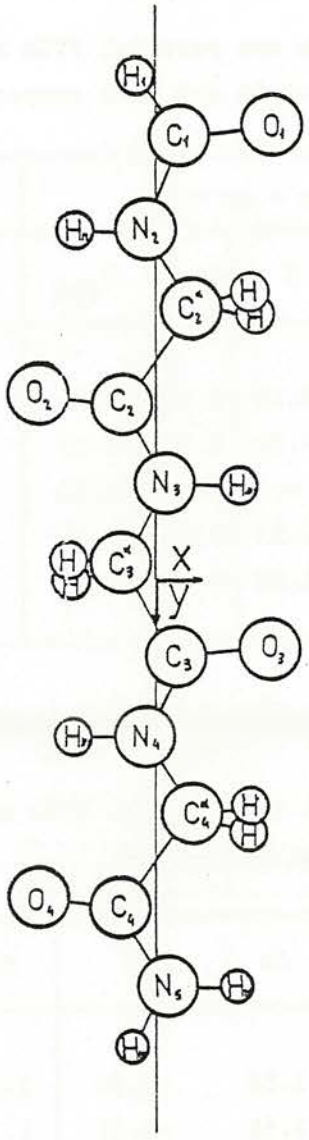


Fig. 8 Schematic representation of a molecule of formyl-tri-glycyl amide in the fully extended conformation.

TABLE VII

Interaction between two parallel FTGA molecules. Minima of E and E_{VDW} with respect to Δz and with respect to Δx

Set	$\Delta x = \Delta y = 0$				$\Delta y = \Delta z = 0$			
	Δz	E	Δz	E_{VDW}	Δx	E	Δx	E_{VDW}
I	4.12	-5.12	4.09	-6.81	5.13	-7.53	5.21	-5.03
II	3.90	-4.20	3.81	-9.15	4.89	-16.46	5.06	-6.41
III	--	--	3.94	-6.54	--	--	--	--
IV	4.05	-6.52	3.98	-10.26	5.12	-14.28	5.28	-6.08
V	3.92	-1.68	3.78	-5.79	4.79	-13.45	4.98	-4.22

TABLE VIII

Interaction between two parallel FTGA molecules. Energy minima with respect to Δy and Δz for $\Delta x = 0$

Set	Δy	Δz	E	Δy	Δz	E_{VDW}
I	3.53	3.58	-8.60	3.36	3.60	-8.88
I	1.97	3.66	-8.31	1.72	3.65	-9.50
II	3.65	3.31	-12.45	3.40	3.36	-11.51
II		not found		1.58	3.38	-12.87
III	--	--	--	3.52	3.45	-9.06
III	--	--	--	1.66	3.50	-9.79
IV	3.56	3.66	-11.74		not found	
IV		not found		1.46	3.74	-11.72
V	3.69	3.24	-7.42	3.50	3.26	-7.78
V		not found		1.58	3.28	-8.86

Figure 9 shows the intermolecular energy curves for two molecules sitting one on top of the other. If we consider only the van der Waals term (Figure 9b), we can classify the potentials according to their repulsive character, i.e. the equilibrium intermolecular separation (V $\sim$ II < III $\sim$ IV < I), and according to the magnitude of the attractive interactions at larger separations (V < III < I = II < IV). Since this molecular arrangement is largely unfavoured by coulombic interactions, the addition of E_{el} (Figure 9a) shifts the position of minima V and II close to the minimum of set III (which neglects the electrostatic term), while minimum IV is shifted toward the minimum of set I (which uses dielectric constant 2). If the upper molecule is now translated along the chain axis by about one peptide unit length (i.e. approximately the most favourable Δy value), positive and negative coulombic terms almost balance each other. The plot of E versus Δz (Figure 10) again shows the features of Figure 9b, with the exception of a more repulsive character of potentials IV with respect to set I.

The energy profiles obtained when the two parallel molecules are shifted along the chain axis at a constant separation Δz, are shown in Figure 11. The five E_{VDW} curves (part b) have a rather parallel behaviour, although the less repulsive potentials present shallower minima; the magnitude of the interactions still follows the same order as in Figure 9b. The electrostatic interactions move the minima to $\Delta y \sim 3.6$ Å, except in the case I owing to the high dielectric constant (part a). When the intermolecular energy is minimized simultaneously with respect to Δy and Δz (see Table VIII), two minima are generally found, whose relative stability is reversed by omitting the electrostatic term.

The same behaviour of E_{VDW} is still observed in Figure 12, where the interaction energies between two adjacent parallel FTGA molecules are plotted versus the distance between the axes. Sets II, IV and V show remarkably close curves of the electrostatic energy, while in case I the magnitude of E_{el} is considerably smaller than simply one half of the values given by the other potentials. Notice that in all cases the minima

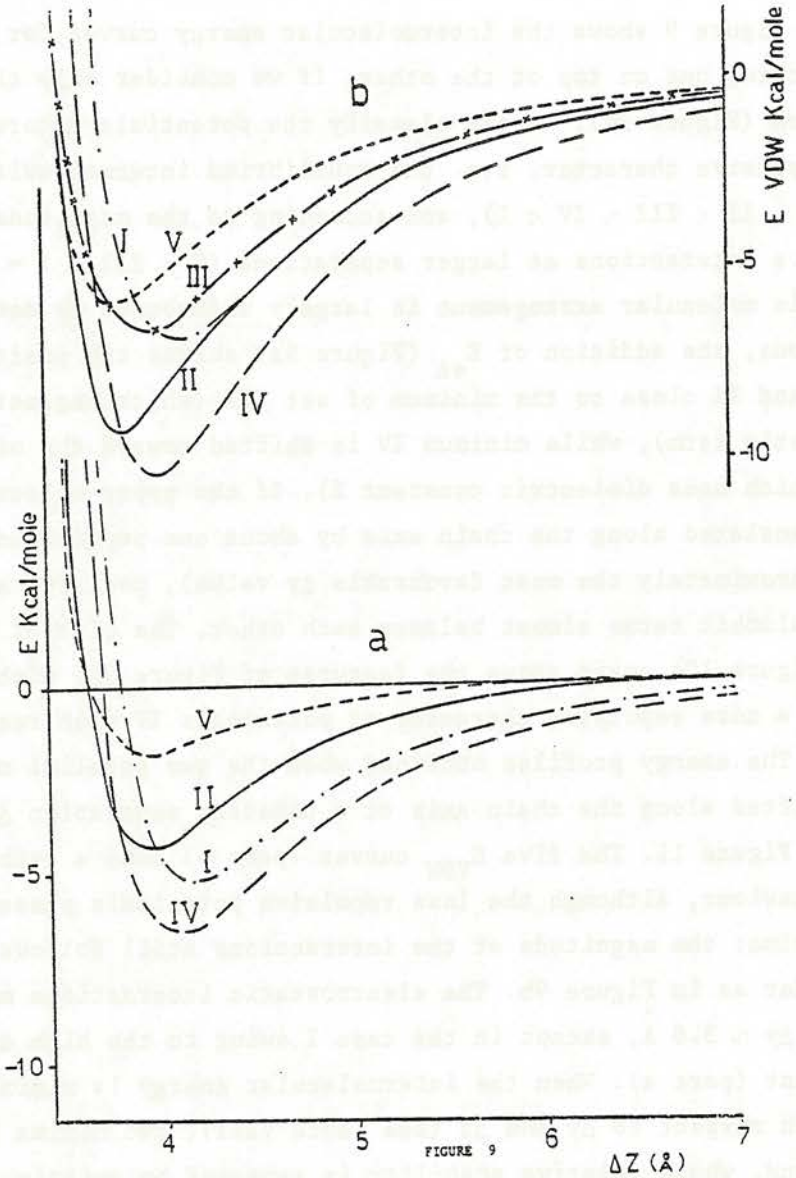


Fig. 9 Interaction between two parallel FTGA molecules plotted versus the separation between the molecular planes, for $\Delta x = \Delta y = 0$:
a) total intermolecular energy, b) van der Waals contribution.

FIGURE 10

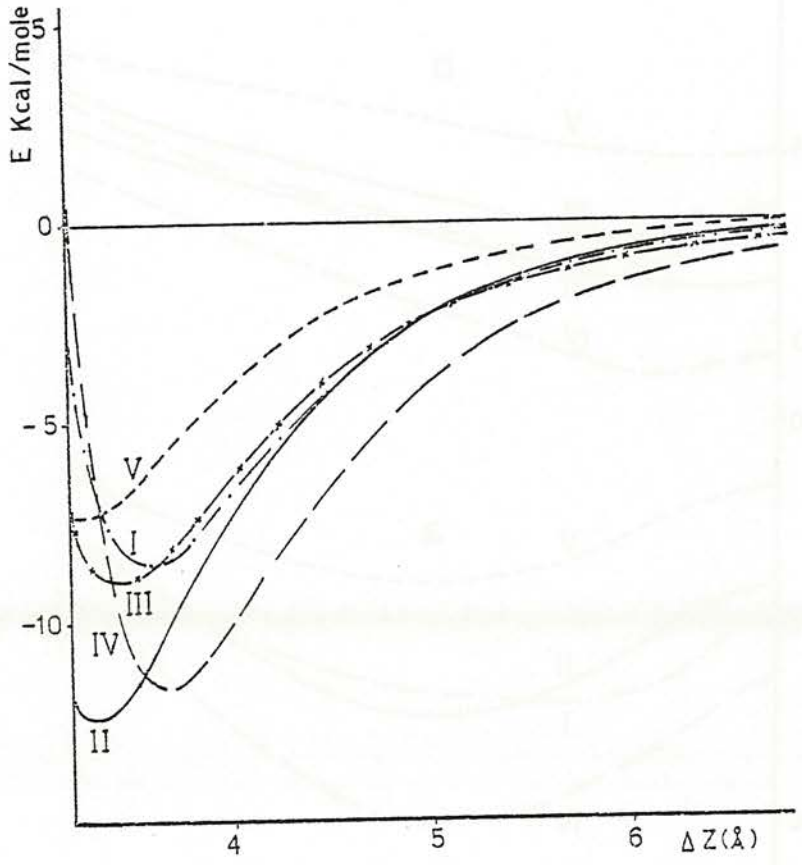


Fig.10 Interaction between two parallel FTGA molecules versus the separation between the molecular planes, for $\Delta x = 0$ and $\Delta y = 3.6$ Å.
The net electrostatic contribution is quite small for $\Delta z > 4.5$ Å

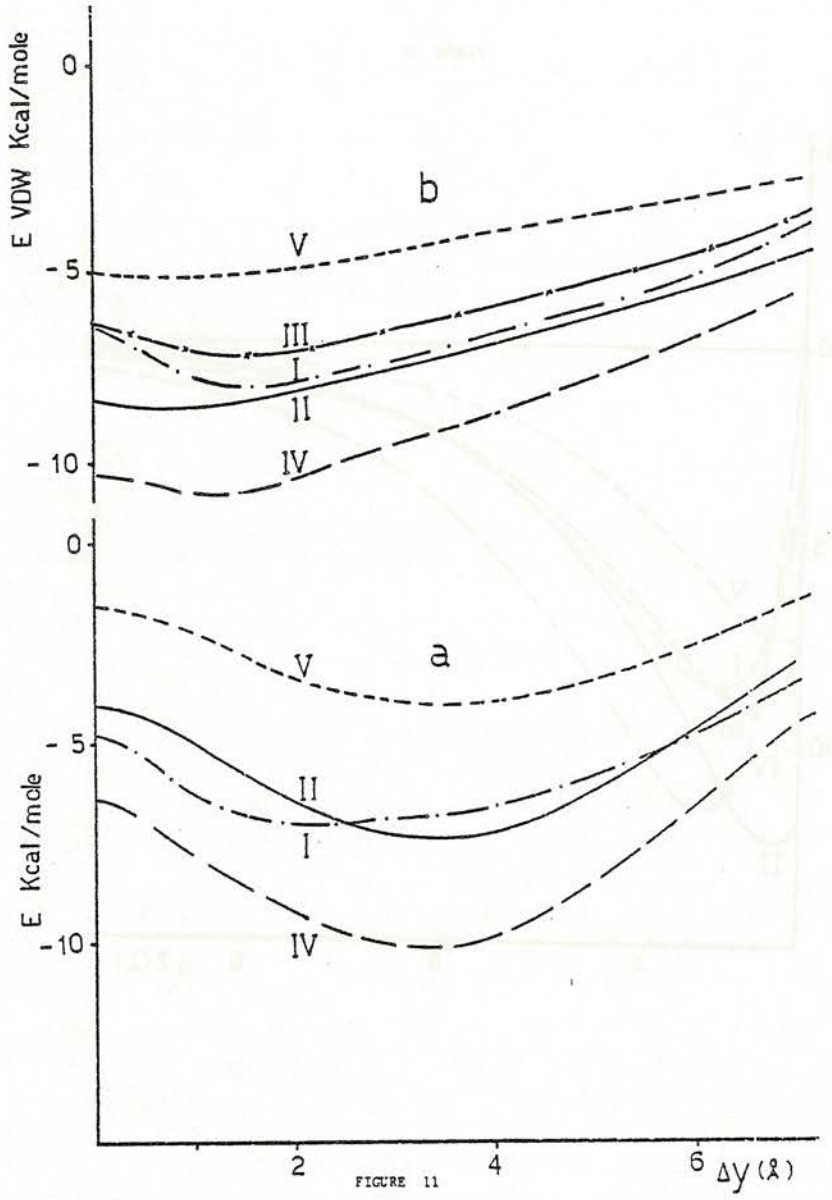


FIGURE 11

Fig.11 Interaction between two parallel FTGA molecules versus the relative shift along the chain axis, for $\Delta x = 0$ and $\Delta z = 4$ Å; a) total energy, b) van der Waals contribution.

FIGURE 12

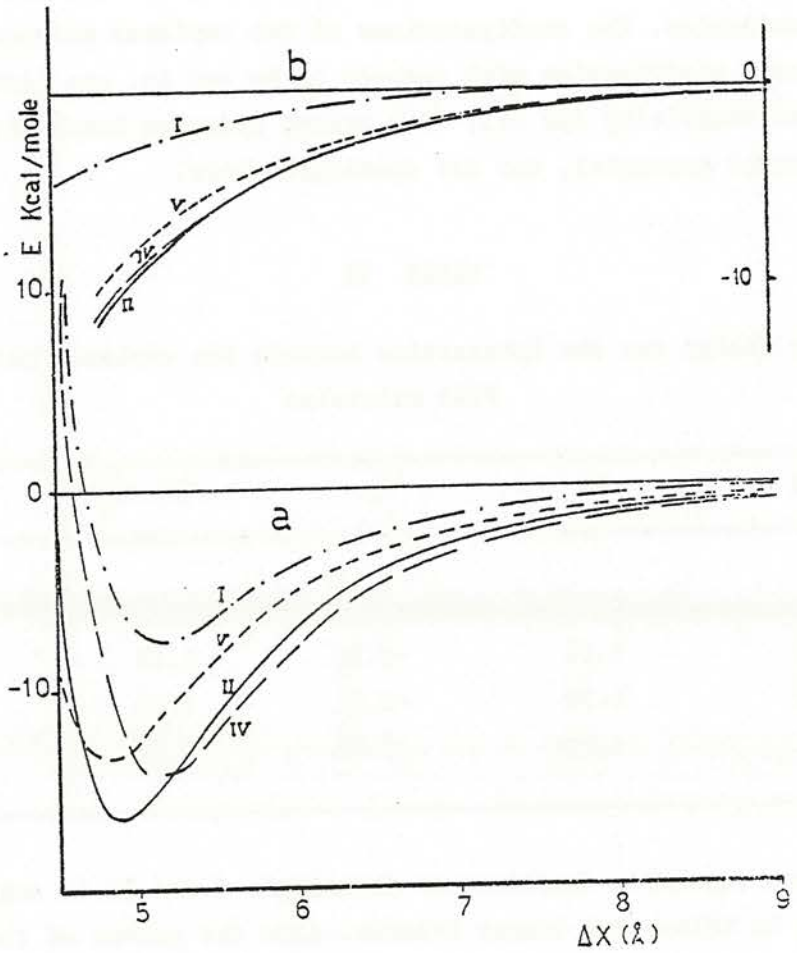


Fig.12 Interaction between two parallel FTGA molecules versus the distance Δx between the chain axes, for $\Delta y = \Delta z = 0$: a) total energy, b) electrostatic contribution.

with respect to Δx and Δy are not hydrogen-bonded structures, because of the CH_2 repulsion.

Finally we come to analyze the interaction between two antiparallel FTGA molecules. The configurations of two coplanar molecules, obtained by energy minimization with respect to Δx and Δy, are listed in Table IX. For simplicity set III, wich treats hydrogen bonds with the Stockmayer potential, was not considered here.

TABLE IX

Energy minima for the interaction between two coplanar antiparallel FTGA molecules

Set	Δx	Δy	$r(O..H_N)$	E
I	4.74	-0.41	1.97	-16.06
II	4.61	-0.54	1.87	-28.10
IV	4.78	-0.31	1.98	-24.20
V	4.57	-0.58	1.85	-22.72

The more repulsive character of potentials I and IV is expressed by larger Δx values and longer H-bonds. Also the curves of the interaction energy plotted versus the relative shift along the chain axis (Figures 13 and 14) show somewhat narrower minima in cases I and IV. The order in the magnitude of the E_{VDW} is not the usual one, the reason being that the curves are drawn at a separazion Δx where repulsion is important. The differences in the electrostatic term (Figure 14 b) between sets II, IV and V appear larger than in previous molecular arrangements, contributing significantly to the differences between the total energy minima.

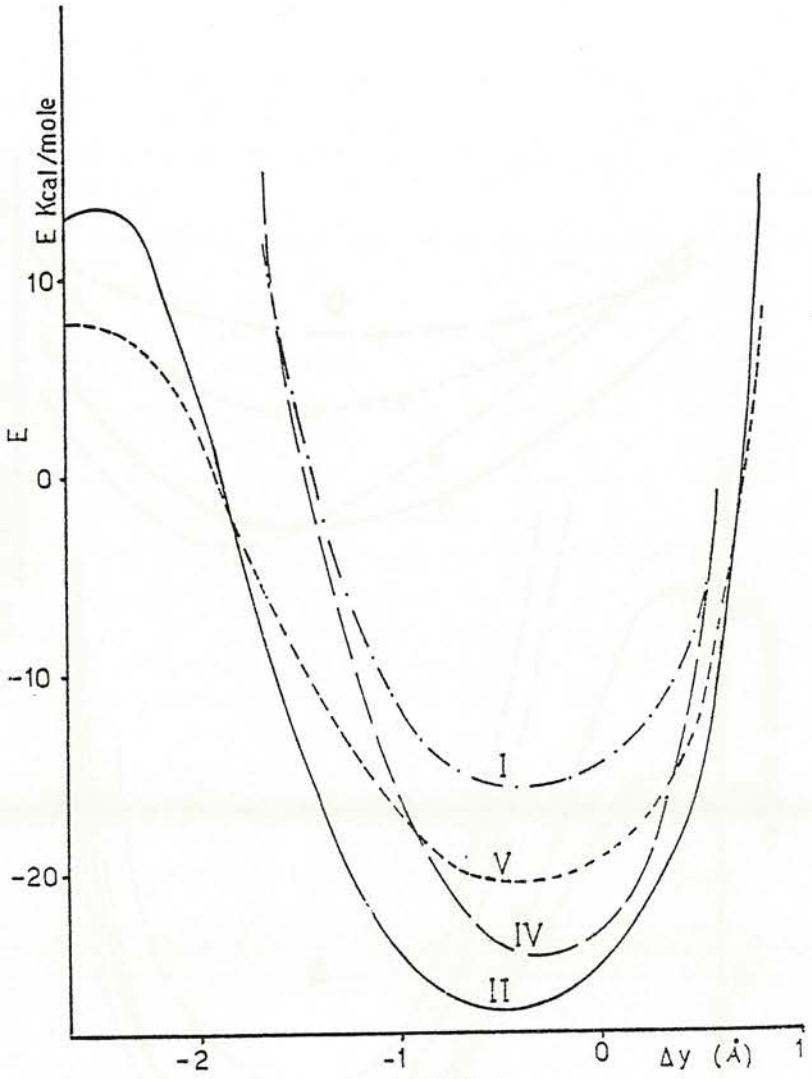


FIGURE 13

Fig.13 Interaction between two antiparallel FTGA molecules. The total intermolecular energy E is plotted versus the relative shift along the chain axis. The two molecules are coplanar, at an axial separation of 4.74 Å; $\Delta y = 0$ corresponds to hydrogen bonds $O..H_N$ normal to the y axis, while negative shifts correspond to O atoms of one molecule lying between H_N and C^{α} of the opposite molecule.

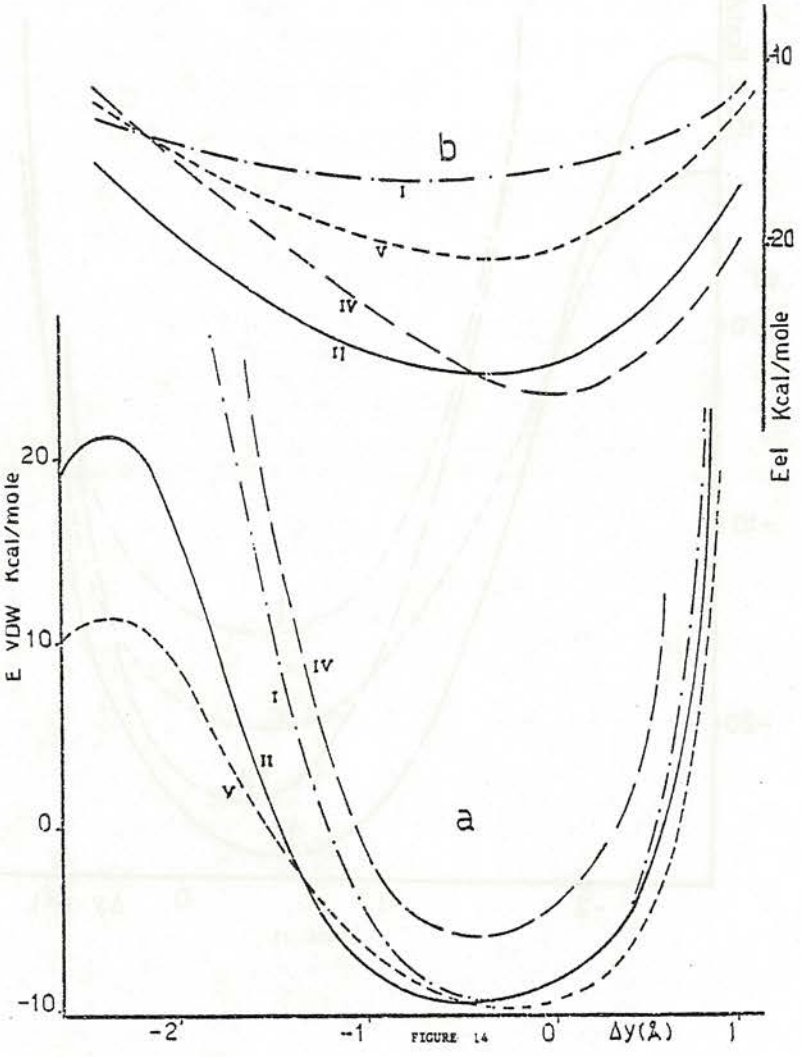


FIGURE 14

Fig.14 Interaction between two antiparallel FTGA molecules. Plot of the van der Waals (a) and electrostatic (b) components of the preceding figure. Here E_{VDW} contains also the hydrogen-bond contribution.

Conclusions

The discrepancies between the potentials examined here are in general not as large as those observed in past reviews. The potentials for N-N and S-S interactions seem the ones for which further work to clear the present disagreement is most needed.

Considering the foreseeable effects on the calculated protein structures of minimum energy, the varying repulsive character of the van der Waals potentials appears to be the most serious source of disagreement. If we look in particular at sets I and II of potentials, we observe that they give very close curves at large separations, while set I, derived directly from the analysis of crystals of small molecules, yields equilibrium distances larger by 0.1-0.2 Å than set II, and therefore also less deep minima. The difference is mainly due to the thermal correction applied by Nelson and Hermans[34] to the functions originally derived from crystal structure analysis. Scheraga and coworkers [24] account for the thermal vibrations by multiplying by a factor 0.5 the B coefficients of set II only for 1-4 interactions, leaving unchanged the functions for atoms separated by more than three bonds. Future calculations will indicate which procedure is correct. Considering that potentials derived from crystals lead to a certain shrink of the structure, it may turn out that a third intermediate set of functions is preferable for use with macromolecular systems.

Less serious effects on the calculated structures should come from differences between the various potentials as far as the absolute magnitude of the van der Waals energy is concerned. In fact these differences are rather uniform in the different molecular arrangements considered, and they may lead to different minima only if the van der Waals energy is strongly counter-balanced by other terms.

Finally we have seen that the use of a large dielectric constant in the calculation of the electrostatic energy has much more significant effects than the differences between the charge distributions currently assumed by the various authors.

References

1. P.De Santis, E.Giglio, A.M.Liquori and A.Ripamonti, J.Polymer Sci. Part A, 1, 1383 (1963)
2. H.A.Scheraga, Adv. Phys. Org.Chem. 6, 103 (1968)
3. H.A.Scheraga, Chem. Rev. 71, 195 (1971)
4. R.Hoffmann, J.Chem.Phys. 39, 1397 (1963)
5. J.A. Pople and D.L.Beveridge, "Approximate Molecular Orbital Theory" McGraw-Hill, New York (1970)
6. S.Diner, J.P.Malrieu and P.Claverie, Theoret. Chim. Acta (Berl.) 13, 1 (1969)
7. E.Clementi and coworkers, papers in press in J.Amer.Chem.Soc.
8. N.L.Allinger, Adv. Phys. Org. Chem. 13, 2 (1976)
9. G.N.Ramachandran and V.Sasisekharan, Adv. Protein Chem. 23, 284 (1968)
10. C.M.Venkatachalam and G.N.Ramachandran in "Conformation of Biopolymers"
    G.N.Ramachandran Ed., Academic Press, New York, Vol. 1, p.83 (1967)
11. A.J.Hopfinger "Conformational Properties of Macromolecules" Academic Press, New York (1973)
12. K.D.Gibson and H.A.Scheraga, Proc. Nat. Acad. Sci. U.S., 58, 420 (1967)
13. A.J.Hopfinger, Macromolecules, 4, 731 (1971)
14. M.Levitt and S.Lifson, J.Mol.Biol. 46, 269 (1969)
15. M.Levitt, J. Mol. Biol. 82, 393 (1974)
16. J.Hermans, D.R.Ferro, J.McQueen and S.C.Wei, in "Environmental Effects on Molecular Structure and Properties", B.Pullman Ed., Reidel, Dordrecht, p.459 (1976)
17. D.R.Ferro and J.Hermans, unpublished data.
18. D.H.Wertz and N.L.Allinger, Tetrahedron 30, 1579 (1974)
19. R.H.Boyd, J.Chem.Phys.49, 2574 (1968)
20. A.Warshel and S.Lifson, J.Chem.Phys. 53,582 (1970)
21. F.K.Winkler and J.D.Dunitz, J.Mol.Biol. 59, 169 (1971)
22. G.N.Ramachandran, A.V.Lakshminarayanan and A.S.Kolaskar, Biochim. Biophys. Acta 303, 8 (1973)

23. A.Warshel and M.Levitt, J.Mol.Biol. 103, 227 (1976)

24. F.A.Momany, R.F.McGuire, A.W.Burgess and H.A.Scheraga, J.Phys. Chem., 79, 2361 (1975)

25. P.De Santis and A.M.Liquori, Biopolymers, 10, 699 (1971)

26. P.Rossky, private communication.

27. R.A.Scott and H.A.Scheraga, J.Chem. Phys. 45, 2091 (1966)

28. D.R.Ferro and J.Hermans, Jr, Biopolymers, 11, 105 (1972)

29. E.Giglio, lecture given at the Summer School on "Weak Interactions in Molecules and Crystals", Bologna (Italy), Sept. 1976

30. D.R.Ferro and J.Hermans, Jr, in "Liquid Crystals and Ordered Fluids", J.F.Johnson and R.J.Porter, Eds,Plenum Press, New York, p.250 (1970)

31. F.A.Momany, L.M.Carruthers, R.F.McGuire and H.A.Scheraga, J.Phys. Chem. 78, 1595 (1974)

32. A.T.Hagler, E.Huler and S.Lifson, J.Amer.Chem.Soc. 96, 5319 (1974)

33. D.Poland and H.A.Scheraga, Biochemistry, 6, 3791 (1967)

34. D.J.Nelson and J.Hermans, Jr, Biopolymers, 12, 1269 (1973)

35. G.C.Lie, E.Clementi and M.Yoshimine, J.Chem.Phys., 64, 2314 (1976)

36. F.A.Momany, L.M.Carruthers and H.A.Scheraga, J.Phys.Chem., 78, 1621 (1974)

37. S.Takagi, H.Chihara and S.Seki, Bull. Chem. Soc./Jap., 32, 84 (1959)

38. W.R.Oegerle and J.R.Sabin, J.Mol.Struct., 15, 131 (1973)

39. P.De Santis, E.Giglio, A.M.Liquori and A.Ripamonti, Nature, 206, 456 (1965)

40. R.A.Scott and H.A.Scheraga, J.Chem.Phys., 44, 3054 (1966)

IV.2

THE ELECTROSTATIC INTERACTION

H.J.C.Berendsen

University of Groningen, Laboratory of Physical Chemistry,
Zernikelaan, Groningen (Pays Bas).

Interaction functions in proteins involve electrostatic interactions between partial charges and dipole moments on individual atoms or molecular groups. For computational reasons it is customary to include interactions up to a given cut-off radius. Atomic polarisability is not often included although a homogeneous polarisability is sometimes accounted for by the use of an effective dielectric constant. The influence of a highly polarizable solvent (water) outside the macromolecule on the internal fields is generally ignored. Thus, as most investigators will undoubtedly realize, the usual computation of electrical interactions in proteins is inaccurate. The following discussion is intended to estimate the importance of the errors inherent in the usual methods and to discuss possible improvements. We will not discuss the inaccuracies that result from the use of CNDO charges ignoring the contribution of bond and orbital dipole moments, nor shall we discuss the problem that improvement of electrical interactions calls for the formidable task of readjustment of other contributions to empirical interaction functions.

A. The use of a cut-off radius

With Coulombic interactions, the use of a cut-off radius is in general not allowed, because the potential may not converge with increasing radius. In globally neutral charge distributions to which we may limit our considerations, the convergence will depend on the characteristics of the charge distribution. An unbounded random distribution of charges in vacuum leads to divergence of the Coulombic potential; it is well-known that electrostatic interactions in ionic liquids or plasmas can only be treated by infinite summation methods.

The situation in a protein is slightly more favourable than in ionic liquids, because atomic charges are locally neutralized due to the invariant covalent structure of the molecule. This means that the Coulombic convergence is that of a collection of dipoles, the dipole being the lowest non-zero moment of the charge distribution of neutral molecular groups. For a collection of randomly oriented dipoles μ, the r.m.s. potential due to dipoles at distances between r_1 and r_2 is equal to

$$\langle \phi^2 \rangle^{\frac{1}{2}} = \frac{\mu}{h^2} \left[\frac{4\pi}{3} h \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \right]^{\frac{1}{2}},$$

where h is the separation of the dipoles (taken on a grid). This indeed converges with r, but very slowly. Thus the validity of the use of a cut-off radius remains questionable, even in the case of random distributions.

When spatial correlations exist between dipole orientations, such as frequently occurs in biopolymers (e.g. in helical structures), one has to exercise particular prudence.

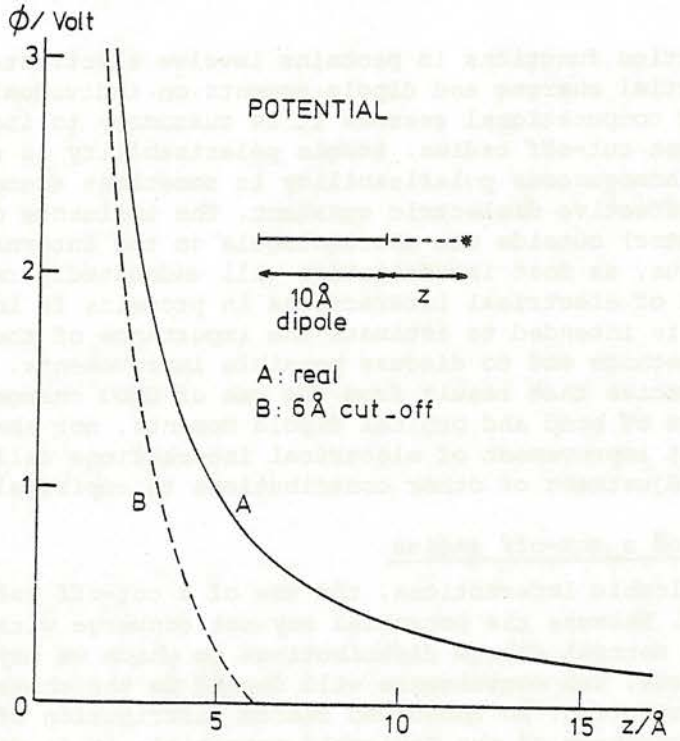


Fig. 1 The electrical potential in vacuum on the axis of a line of homogeneous dipole density of 10 Å length, at a distance z from its end. The dipole density is 8.06 x 10^{-20} Cm/m, representing a α-helix. The drawn line A gives the exact potential; the broken line B gives the potential using 6 Å cut-off radius.

The followingexample may illustrate this point. In fig.1 the potential is plotted outside a line of constant dipole density 8.06 × 10^{-20} Cm/m, representing an α-helix of 1.2 × 10^{-29} Cm per residue. The potential is calculated in vacuum on the extension of the dipole axis at a distance z from its end. The length of the line is taken as 10 Å, corresponding to about two turns of a helix with about 7 residues. It is observed that the use of a cut-off radius of 6 Å produces severe errors, also in the field used for force calculations, given in Fig.2. The actual fields are quite large: at 7.5 Å distance the field is as large as 10^{9} Vm^{-1}. In this field a water dipole of 6.13 × 10^{-30} Cm has an electrostatic energy of 1.5 kT. With a 6 Å cut-off such interactions are entirely ignored.

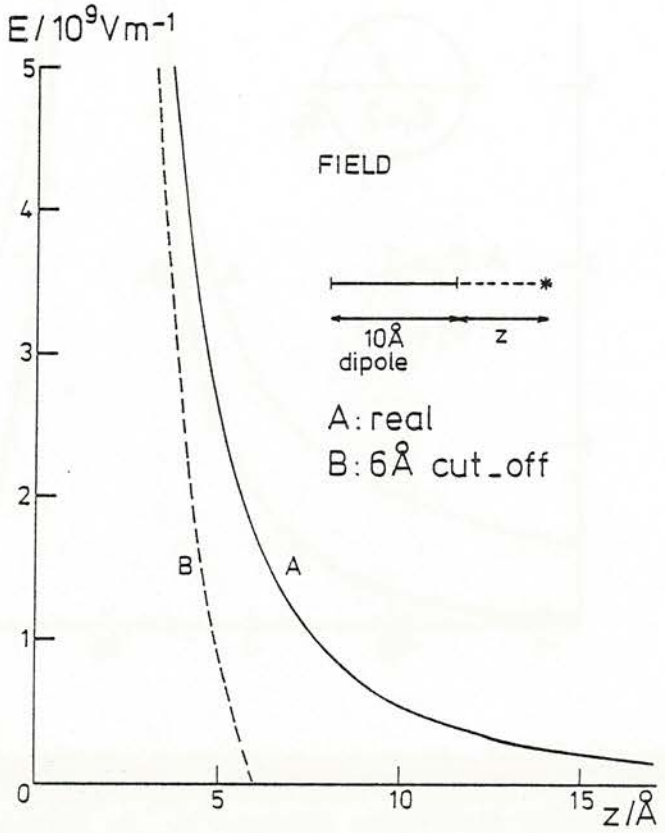


Fig. 2. The electric field calculated for the same case as in fig. 1.

B. Influence of the reaction field

A protein molecule is surrounded by a medium (water) with a high dielectric constant (80), while the dielectric constant inside the protein arises from atomic polarizability and can be taken to be about 2. The outside medium produces a reaction field inside the protein, modifying the Coulombic interaction.

For a simple case the effect of the outside medium can be calculated. Assume a charge of +q is located inside a sphere with $\varepsilon = 2$ in a medium with $\varepsilon = 80$. We take the position of the charge arbitrarily at a distance of half the radius from the centre. The reaction field can be approximated[1] with high accuracy by the field of an image charge outside the sphere at a distance of twice the radius from the centre, of magnitude $-(78/41)q$, with both charge and image charge in a medium of $\varepsilon = 2$. Fig.3 shows the simple Coulombic and the real potential of this charge along the radius through the charge (and image charge). No cut-off radius has been used. The errors obtained by neglect of the reaction field are severe. Only when the distance to the charge (or dipole) is small compared to the distance to the outer surface of the macro-molecule are the electrical interactions somewhat reliable. The incorporation of the reaction field seems to be essential for molecular dynamics and energy minimization procedures.

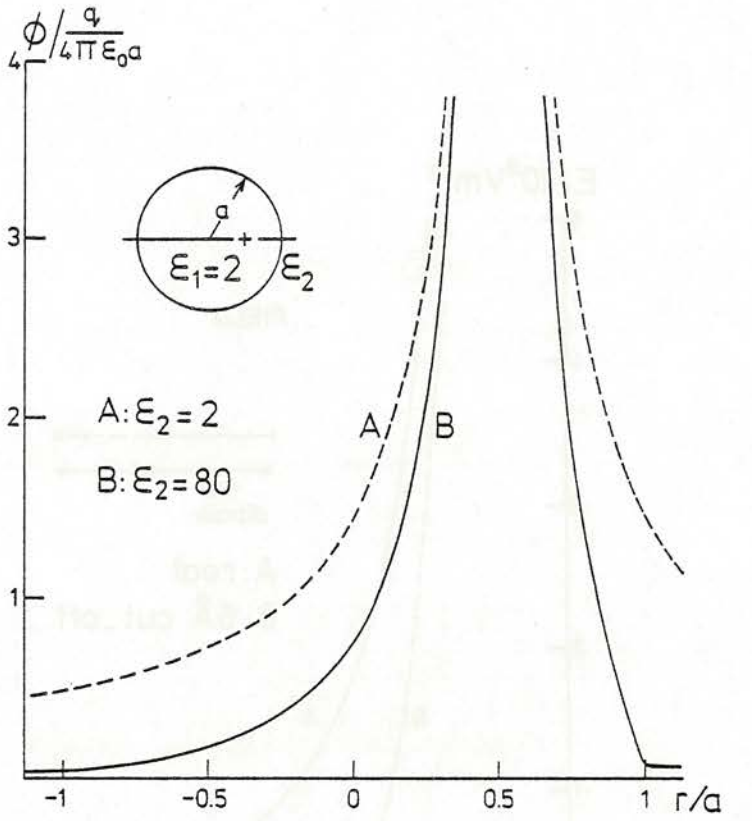


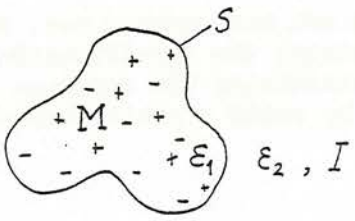
Fig. 3. The potential of a charge + q situated halfway the radius of a sphere with internal dielectric constant $\varepsilon_1=2$, which is embedded in a continuous medium with dielectric constant ε_2. A: normal Coulomb potential disregarding the reaction potential. B: Coulomb potential including the reaction potential for $\varepsilon_2=80$.

C. Methods to improve the calculation of electrostatic interactions

There are two different ways to compute electrostatic interactions without the disadvantages resulting from cut-off radius and reaction field. The first is the use of periodic boundary conditions, as is usual practice for simulation of liquids, combined with infinite summation methods, such as the often applied Ewald summation or the Fourier methods developed by Hockney and Eastwood[2], referred to as the PPPM method, for solving Poisson's equation with periodic conditions. For proteins in water these methods have the serious disadvantage that water molecules (or at least reorienting dipoles) have to be individually specified throughout a cubic or rectangular volume of sufficient size that the protein molecule does not interact significantly with its images. This requires the computation of details in the aqueous solvent that are not relevant for the macromolecular behaviour and thus can be considered as a waste of computing effort. Moreover, the periodic conditions do not guarantee a proper treatment of the reaction field unless the periodic box is very large.

The second way in which the reaction field can be taken into account is to solve the Poisson potential equation for the macromolecule embedded in a continuous medium with high dielectric constant. The problem can be formulated as follows.

Given a macromolecule M with irregular surface S, with a given internal distribution of charges $\rho\ (r_1)$ and an internal dielectric constant ε_1. The macromolecule is embedded in a homogeneous medium with dielectric constant ε_2 and (if applicable) an ionic strength $I = \frac{1}{2} \Sigma c_i z_i^2$ due to concentrations c_i of ions with charge $z_i e$.



The assumption is made that, both inside and outside M, the polarisability is homogeneous and non-saturating. If this assumption is not expected to be valid, a limited number of water molecules have to be included specifically.

The differential equations for the potential $\emptyset$ are:

inside M : $\nabla^2 \emptyset\ (\underline{r}_1) = -\rho(\underline{r}_1)/\varepsilon_1$

outside M: $\nabla^2 \emptyset(\underline{r}_2) = -\rho\ (\underline{r}_2)/\varepsilon_2$, with

$\rho(\underline{r}_2) = 0$ in a non-electrolyte, or

$\rho(\underline{r}_2) = Ne\ \Sigma c_i z_i e^{-ez_i\emptyset/kT}$ for electrolytes,

where N is Avogadro's number. The latter equation can be reduced in the Debye-Hückel limit of $e\emptyset << kT$ to:

$$\rho(r_2) = -\frac{2Ne^2 I}{kT}\,\emptyset(r_2)$$

The boundary conditions are:
at the surface S:

a) $\emptyset$ is continuous: $\lim_{r_1 \to r_s} \emptyset(r_1) = \lim_{r_2 \to r_s} \emptyset(r_2)$

b) $\varepsilon(\underline{\nabla}\emptyset.\underline{n})$ is continuous ($\underline{n}$ is the normal to the surface):

$$\varepsilon_1 \lim_{r_1 \to r_s}\left[\underline{\nabla}\emptyset(\underline{r}_1).\underline{n}\right] = \varepsilon_2 \lim_{r_2 \to r_s}\left[\underline{\nabla}\emptyset(\underline{r}_2).\underline{n}\right]$$

at $r_2 \to \infty$: $\emptyset(\underline{r}_2) \to 0$ and $\underline{\nabla}\emptyset(\underline{r}_2) \to 0$

An elegant treatment of the electrostatic interaction of a molecule with its polarisable surroundings has been given by Huron and Claverie [3] *. In their theory the potentials inside and outside a molecule of irregular shape are expanded in series of harmonic functions using a least squares criterion for the fulfilment of the boundary conditions. The outside potential is expanded on a multicentered basis, using the positions of the internal charges as centers. Unfortunately, for a macromolecule with hundreds of charges, the number of harmonic functions to be included becomes very large and the required matrix inversion becomes completely impractical.

Thus far, a satisfactory solution has not been worked out. We are at present developing such a solution, involving the specification of a limited number of water molecules and determining the reaction field in part by using image charges and in part by using a series harmonic expansion.

References

1. H.A.Friedman, Mol.Phys. 29 , 1533 (1975).
2. R.W.Hockney, S.P.Goel, J.W.Eastwood, Chem.Phys.Lett. 21, 589 (1973); J.W.Eastwood, Report RCS 34, Dept.of Computer Science, University of Reading, U.K. (1975).
3. M.J.Huron and P.Claverie, J.Phys.Chem. 7, 1853 (1974).

* We thank Drs. P.Claverie and Daudey for their discussion on molecular interaction during the Workshop.