THE ENERGETICS OF NUCLEOTIDE BINDING TO *RAS* PROTEINS

A thesis submitted for the degree of Doctor of Philosophy
in the University of Oxford

by

Graham A. Worth

St. Edmund Hall

September 1992

The Energetics of Nucleotide Binding to *Ras* Proteins

Abstract of a thesis submitted for the degree of Doctor of Philosophy in the University of Oxford.

Graham A. Worth
St. Edmund Hall
September 1992

Ras proteins are a special class of proteins that mediate cell growth signals. Their importance lies in the fact that they are products of a proto-oncogene. This means that under certain conditions the gene that determines its structure is altered and a mutant protein results that is involved in the transformation of normal cells to cancer cells.

The actual function by which the protein acts in the signal pathway is not known. However it is known that they act as a switch, undergoing a cycle involving the exchange of guaninosine nucleotides in the binding site. This thesis uses computer simulations to study the energetics of this binding, with the long term aim of developing a drug to inhibit the transforming activity of the oncogenic protein.

To begin with, a model of the protein based on a crystal structure is built. Using Molecular dynamics the motion of this model is studied. A possible mechanism by which one half of the nucleotide cycle could be induced is investigated, with the result that phosphorylation of the protein may be involved.

The main part of the thesis is then devoted to using the free energy perturbation (FEP) method to calculate the difference in Gibbs binding free energy between the nucleotides in the protein.

Using histamine as a model, a method of dealing with charged, flexible molecules is developed; namely the inclusion of a reaction field and comprehensive conformational analysis. The results from the associated calculations are seen to be very close to experimental data. The same procedures are then applied to the much more complex *ras*: nucleotide system with less successful results, the reason for which is mostly due to the restriction of limited computer resources to tackle such a problem.

The conclusion is that given the resources and by using the techniques developed in this thesis, this type of calculation is a feasible way to study such systems.

Acknowledgements

The first person to thank is my supervisor for the last four years, Dr. Graham Richards. I have enjoyed my time in his group immensely and he has always been a big encouragement. I would also like to thank my industrial supervisor, Dr. Colin Edge, firstly for the project and useful input to it, and secondly for not forcing me to spend the obligatory three months with my CASE sponsors, SmithKline Beechams, although their support has been much appreciated.

The rest of the group have also been wonderful. In particular: Jon Essex for his belief in rigour that often prevented me from doing wonderfully inventive rubbish; Susan West for her knowledge of proteins and computers; Chris Reynolds, Paul King and Ian Haworth for general help at times during the project and Maria Ramos for just being Maria. The other people who have been around at various times have all made the lab a friendly and stimulating place to work.

Thanks must go to SERC for supporting me for the last three years. However it must be said that their crass slashing of the Richards group computer budget in the final year of this thesis caused much frustration and aggravation.

Thanks must also go to my proof readers, Jon, Alison, Guy, Andy and Christiane for their time spent and minimal abuse of my poor grasp of the English language. In the end it is, however, all down to me and I must take the blame for any mistakes, inaccuracies or falsehoods that may be present; whether due to typing, laziness or sheer stupidity.

Special thanks go to Christiane for supporting me in the last throes of producing this tome.

Contents

Introduction

Theoretical chemistry is a very young subject, and computer involvement even younger. This project is thus part of a rapidly growing field of knowledge and understanding. Molecular dynamics on proteins was first realised in the mid 1970's Meanwhile, Free energy perturbation (FEP) methods have only been used in conjunction with computers since around 1980. The bibliography below gives the best all-round texts on both topics.

This thesis will show that computer chemistry is still coming of age. Reasonable simulations of macromolecules are possible, but certain systems are still outside the bounds of present resources. This is especially true if quantitative results are desired.

A major problem has been that thoughts about how these techniques should be carried out, and even what the results actually mean, are still under discussion. What is more they have changed significantly over the last three to four years - the period of time over which the material for this thesis was accumulated.

Due to this the study was not linear, starting with a detailed experimental plan, followed by the study, dutifully followed by the results which corroborate or invalidate the original hypothesis. Calculations were often started for one reason, then finished for another. Other results prompted more thought about the underlying principles and a return to earlier calculations.

This has made the thesis difficult to deliver as a cohesive argument. Instead it is presented as a journey through the project, showing the logical links between calculations and how these results go on to stimulate the next.

In October 1989 when this project began, FEP calculations had reached a zenith. They seemed to be both accurate and widely applicable. This project was thus conceived as a difficult, but possible, way to study an important biological system - the *ras* proteins. During the course of the study the simulation techniques were pushed to their limits. Much useful information was gained, some of it applicable to the system under study, but more to simulations of macromolecules in general.

The original project will now be outlined and the parts of the thesis discussed in general terms.

The first chapter outlines the theory of energetics of chemical systems. It tries to draw together often differently treated subjects, namely thermodynamics, statistical mechanics and quantum mechanics, to show that they are in fact merely different facets of the same set of concepts.

Chapter II then introduces the system that will be studied. The *ras* oncogene is one of the most commonly expressed oncogenes found in human cancers. Its product, the *ras* or p21 protein, has as a result been subject to many experimental studies and much has

been learnt about it. Various crystal structures, of both wild and oncogenically active types, have been elucidated and the protein's biochemistry studied in some depth. Despite this depth of knowledge, much is still to be learnt. For example it is not known where the protein acts in the complex cell growth pathways of which it is definitely part. The most important feature of these proteins is that they are nucleotide binding proteins. There are various proteins of this type e.g. G proteins, and the property that all have in common is that they cycle between binding guanosine 5' diphosphate (GDP) or guanosine 5' triphosphate (GTP), enabling them to act as switches in a signal pathway.

The differences between the mutant, oncogenic proteins and the naturally occurring proteins are also known. It is found that this switch mechanism is so sensitive to the protein structure that any one of three point mutations of the amino acid primary sequence is enough to interfere with it, leaving the protein permanently in the 'on' position. This interference leads to a dysfunction of the growth messenger signals and the result can be an abnormal growth - cancer. A highly desirable, but not yet attained, aim is to find some way to block these mutated forms to try and inhibit cancers.

It is this topic that this thesis will try to explore. Computer simulation techniques have become more powerful and gained much respect over the last few years. The growth in supercomputing power has increased enormously and with it the ability to study large biological systems to a good degree of accuracy has become a reality. Hopefully these theoretical techniques can be exploited to gain insight into how the point mutations affect the nucleotide binding cycle and even to propose a molecule such as a nucleotide analogue that can effectively act as a specific drug to block these changes.

The topic lends itself well to simulation studies in that crystal structures are known and various biochemical properties have been measured to calibrate calculations against. However the guanosine nucleotides pose large problems due to their size and complex electronic structure. Ions are notoriously difficult to study by these means, and ions of heavy atoms such as phosphorus even more so. These simulations will thus use well established techniques, but take them beyond the limits of previous calculations.

Chapter III sets up a model for the protein. It then studies the motion of this system, both in the gas phase and in solution. At the same time a possible mechanism is explored to explain how the exchange of the nucleotides in the binding site is induced.

Chapter IV starts the attempt to calculate the binding free energy differences desired by the scheme set out in chapter II. The difficult areas of simulations, namely the treatment of ions and flexible molecules, are explored by the use of histamine monocation as a model. As well as carefully studying the conformational and electrostatic dependence of the system, a new way of treating the long range forces in a simulation is introduced, and by the end of the chapter, values for the tautomerism ratio of histamine in solution are very close to experiment.

Chapter V carries this study on by studying the dependence of the methyl-phosphate - magnesium system on the simulation parameters used. A way is set out to calculate good atom centred charges that can be used in simulations. These methods are then applied to the guanosine nucleotides to be used in the calculations.

Finally in chapter VI a perturbation calculation is attempted on the nucleotides, both in aqueous solution and in the *ras* protein binding site. To begin with a different way of using the FEP method is set up that is much shorter than the standard windows technique. This is then used to calculate the difference in free energy of binding of GDP and GTP to the protein, and an estimate is made for the energy of hydrolysis of GTP. Both results are seen to be in some error compared to experiment, but this can be put down the limited resources available at the end of the project, which did not allow the methods to be properly utilised.

Unfortunately the problems involved with the simulation meant that no further simulations on the protein system were attempted. The conclusion therefore outlines possible further calculations that could be undertaken and tries to bring together all the ideas from the thesis to create a stepping stone to further work.

General Bibliography

General texts on protein dynamics

Brooks, C.L., Karplus, M. and Pettitt, B.M. "Proteins" (1988) John Wiley and Sons

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

General texts on FEP and other aspects of simulation of biomolecules,

"Computer Simulation of Biomolecular Systems" Eds. van Gunsteren, W.F. and Weiner, P.K. (1989) Escom

"Computer Simulation of Chemical and Biomolecular Systems" Eds. Beveridge, D.L. and Jorgensen, W.L. (1986) New York Academy of Sciences.

CHAPTER I

The Theory of Energy and Chemical Systems

Before beginning a study on anything, it is necessary to define and describe the theory that will be used. Theories are the language of science, with mathematics as the laws, or grammar. Observations are explained in their terms and if that is not possible, a new theoretical concept must evolve in an analogous manner to a new word in everyday speech.

In this thesis, the quantities being measured are the differences in energy between different chemical systems. The energies of such systems are governed by the laws of thermodynamics. However these are relationships between macroscopic quantities. In order to study chemical systems, it is desirable to look at the molecular behaviour. Thermodynamics is connected to this behaviour by statistics.

Much of the theory used in this thesis is well known and written about elsewhere. This chapter will only give an overview of the main parts of thermodynamics and how this is related to the energy of the systems constituent particles. The starting point will be to define a system in terms of its energy. This will then be connected to the arrangement of the molecules in the system and hence their properties. After examining this theory, a description will then be given of how it is implemented in practice, at the same time introducing and describing the computer programs used as a basis for this thesis.

I.1. The Theoretical Description of Chemical Systems

A system amenable to study is a portion of the universe that can be defined as a whole, separable from its surroundings. This system has properties that are measurable quantities. Any system can be viewed in two ways; the external and the internal. The former is the appearance of the system, the latter the behaviour of the individual species making up the whole. Depending on how it is has been specified, certain external properties are constrained; the size or the weight etc. Other properties can however vary and if the time scale of measurement is longer than these fluctuations, only their average value is observed.

The branch of science known as chemistry is in many ways special due to the range spanned by its observations. Systems as small as electrons, only viewable by their tracks, to huge systems of many millions of atoms are covered. Both ends of the scale appear to behave very differently and indeed the theories are often presented so.

The study of chemical systems started with viewing them as spirits. Nothing was known about the fundamental particles, although many postulates were made. The first theories thus developed about the external properties of the system; the kinetic theory of

gases is one example. It was only later that, with the reintroduction of the concept of atoms, the properties of the microscopic particles underneath could be described and related to these macroscopic observables.

The next step in chemical theory was the development of thermodynamics. The triumph of 19th century science, this was based on the realisation that heat and work are different facets of the same thing; energy. Four empirical laws, all seemingly intuitive, were deduced that combine to give all measurable properties of a system in terms of its energy.

I.2. The System in Terms of Energy

Everything is energy. Einstein showed the equivalence of mass and energy. Forces are energy gradients in space. Temperature is the average kinetic energy of a system of molecules. And so on. You cannot however measure energy directly, only by its acting on something - moving a weight or heating a beaker of water i.e. by changes in forces and co-ordinates.

In Newtonian mechanics, energy is a function of force and position, $E(\mathbf{F},\mathbf{x})$. A constant force moving a system changes the co-ordinates, and uses energy. Also, a changing force with no change in position expends energy. Hence any change in energy is given by the differential

$$\mathrm{d}E = \mathbf{F}\mathrm{d}\mathbf{x} + \mathbf{x}\mathrm{d}\mathbf{F} \qquad \text{(E1.1)}$$

Forces can be generalised to include more than its everyday terminology. Pressure ($-p$, it is negative by convention), electric field ($\mathbf{E}$), magnetic field ($\mathbf{H}$), surface tension (σ) or in fact any mechanical intensive variable can be considered a force; less obviously so can temperature (T) and chemical potential (μ). With each force there is associated a generalised co-ordinate, the property that is changed under the influence of the force. These are, listed in the same order as the above forces, the mechanical extensive variables volume (V), polarisation ($\mathbf{P}$), magnetisation ($\mathbf{M}$) and surface area (A) and also the entropy ($-S$, again it is negative by convention) and number of particles (n). Using these variables, E1.1 is expanded to be

$$\mathrm{d}E = -p\mathrm{d}V + V\mathrm{d}p + \mathbf{E}\mathrm{d}\mathbf{P} + \mathbf{P}\mathrm{d}\mathbf{E} + \mathbf{H}\mathrm{d}\mathbf{M} + \mathbf{M}\mathrm{d}\mathbf{H} - T\mathrm{d}S + S\mathrm{d}T + \mu\mathrm{d}n + n\mathrm{d}\mu + \ldots\ldots \qquad \text{(E1.2)}$$

For chemical systems of interest to thermodynamics, only p, V, T, S, μ and n are important and we shall restrict ourselves to just these variables. The exact relationship of the energy to the variables now depends on the conditions the system is under.

I.2.1. The System at Equilibrium

One possibility is to take the energy of a system with a fixed composition as a function of just the extensive variables of the system, the volume and entropy,

$$dE = -pdV + TdS$$

Any change in μ is not included as it is a constant by definition. This type of energy is called the internal energy of the system, denoted by U. Under these conditions, the system will spontaneously expand and increase its entropy until equilibrium is achieved. At this point dU, dV and dS are all zero. U can thus be thought of as a potential, driving the system towards equilibrium. At this point the system variables are constant (denoted an *NVS* system, although conventionally entropy is equated with heat and hence energy and this is called an *NVE* system) and the internal energy is the energy measured. This is in fact the first law of thermodynamics. The first term on the right hand side is the mechanical work done on the system, dw, and the second term the heat given to the system, dq.

These variables are then constrained to keep the system in this state, which means that it is isolated in a region of space; no energy or matter can flow in or out of the system. This makes it the simplest type of system with respect to its energy. It must now be remembered that the system is composed of molecules, or more accurately in view of quantum mechanics, electrons and nucleii. The simplicity of these constraints is that all the energy in the system must be possessed by these molecules.

I.2.2. The underlying nature of the system

Nothing is specified about these particles, except that they are somewhere in the system and their total energy is U. At a particular instant in time, the molecules have a particular arrangement, constrained by the physical limits and equilibrium potential of the system. They are also in motion, so an instant later they will have moved to a different arrangement.

This motion is due to their individual energies, of which there are two types. When a particle collides with something, it gives off or receives energy and its momentum changes. This is the kinetic energy, due to the heat in the system which is translated as motion. They may also have various properties, an electric or magnetic field for example, which can interact producing forces and changes of energy. This is the potential energy, and is related to the relative spatial positions of the particles, constrained by how tightly they are packed together. They also have intrinsic energy due to their mass, but under usual conditions this is fixed.

The most elegant and general way of expressing the motion of particles in terms of their energy is by the method of Lagrange. The first thing to be considered is a particle of mass m at equilibrium. It is at rest or moving in a straight line due to its inertia. If this particle is subject to a force, its velocity will be changed according to the relationship derived from Newton's second law, $\mathbf{F} = m\dot{\mathbf{v}}$. In a system with a lot of particles, all mutually interacting, at any one instant a particle will be subject to a force. The method of

Lagrange was to imagine applying a force in the opposite direction to return the system to equilibrium. This can be expressed by

$$\sum_i (\mathbf{F}_i - m_i \ddot{\mathbf{r}}_i)\delta\mathbf{r}_i = 0 \qquad \text{(E1.3)}$$

where F is the force applied to balance the acceleration, $\ddot{\mathbf{r}}$. When this expression is transformed to general co-ordinates we obtain the Lagrangian differential equation of motion

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}}\right) - \frac{\partial \mathcal{L}}{\partial \mathbf{q}} = 0 \qquad \text{(E1.4)}$$

where $\mathcal{L} = \mathcal{T} - \mathcal{V}$ i.e. the difference between the kinetic energy of the system,

$\mathcal{T} = \sum_i \frac{1}{2} m_i \dot{\mathbf{r}}_i^2$ .and the potential energy $\mathcal{V}$.

I.2.3. The System as a Probability in Phase Space

A useful way of describing a system is in terms of its phase space. This is an artificial construction of all the available degrees of freedom of the system. For example particles have six degrees of freedom; three position and three momenta co-ordinates. A particular arrangement of the molecules is specified by these variables. Each arrangement can be called a different state and each state, A, has an energy, E_A. If there are N molecules, the system has 6N degrees of freedom. Each degree is a basis vector in the phase space and thus any state of the system E_A can be described as a point in this space by these basis vectors. The region of this space available to the system under its constraints is thus a surface of all the possible points. As the particles move, according to the Lagrangian equation, the system moves between corresponding points.

The probability that the system is in any one of these states is given by.

$$P_A = \frac{1}{\text{Number of equally likely states}} \qquad \text{(E1.5)}$$

The important transition to statistical thermodynamics is the assumption that the system visits all the points in phase space. This is known as ergodic theory. It has amazing results. If a system is in a unique state which evolves by Lagrange's dynamics, it defines a definite trajectory over the phase space surface. This means that the trajectory cannot cross itself; if this happened it would not be definite, when a point is revisited it could travel in multiple directions and so all of phase space is not covered. However, if this was so, the properties of a system could vary depending on what state it starts from and this ~~is known~~ can not to be so. For example if an ideal gas is at equilibrium with a defined volume and temperature, the pressure is given by the ideal gas law and this is found to be always the case.

The important factor is that the system is not in a definite state, but can be in a range of states given by the probability above. Hence

$$P_A = \frac{1}{\text{Number of possible states with energy } E_A} \qquad \text{(E1.6)}$$

How many states can there be with this particular energy? This is related to the arrangements of the particles making up the system. This can be specified in terms of the positions and momenta of the particles, or more simply their energies. We then want to know how many particles have a particular energy.

The number of ways N particles can be arranged is given by

$$W = \frac{N!}{n_1!\,n_2!\,n_3!\ldots} \qquad \text{(E1.7)}$$

where n_i is the number of molecules with energy ε_i. With the constraints that the total energy and the number of particles are constant, this number, n_i, is then given for the most likely configuration by the quotient calculated by Boltzmann

$$\frac{n_i}{N} = \frac{\exp -\varepsilon_i \beta}{\sum_i \exp -\varepsilon_i \beta} \qquad \text{(E1.8)}$$

from which, substituting back into W in E1.7 and rearranging using Sterling's approximation, gives

$$P_A = \frac{\Pi_i \exp -n_i \varepsilon_i \beta}{\Pi_i \left(\sum_i \exp -n_i \varepsilon_i \beta \right)} = \frac{\exp -E_A \beta}{\sum_A \exp -E_A \beta} \qquad \text{(E1.9)}$$

where Π_i denotes a product over all the particles i. The denominator of the right hand side is a very special function. This sum over all the energy states of the system is called the partition function and is denoted by Q.

How does the probability change with time? Providing the energy states do not change as the system changes, the denominator in the system is a constant with time,

$$\frac{\partial P_A}{\partial t} = -\frac{\beta}{Q} \frac{\partial E_A}{\partial t} \exp -E_A \beta \qquad \text{(E1.10)}$$

Under the conditions of constant entropy, there is no heat flow in or out of the system. All the energy states have the same energy and so this becomes

$$\frac{\partial P_A}{\partial t} = -\beta P_A \qquad \text{(E1.11)}$$

The significance of this equation will become apparent later.

Now we have the situation that if all the possible energies of the particles are known, then we know the probability of the system being in a particular state and how that probability changes with time.

The measured potential is given by the average energy of the system as it travels through this phase space. i.e.

$$U = \sum_{A} \frac{n_A E_A}{\mathbf{N}} \qquad \text{(E1.12)}$$

where **N** is the total number of states.

This may not seem like a particularly interesting relationship. We have constrained all the possible states of the system to have the same energy, $E_A = U$, and then found the average. It is however more important when systems under different constraints are studied, then the energy of the states may differ but this equation still holds.

By noticing that

$$E_A \exp - E_A\beta = -\frac{\partial}{\partial\beta} \exp - E_A\beta \qquad \text{(E1.13)}$$

this expression can be rearranged to give

$$U = -\mathrm{kT}\ln Q \qquad \text{(E1.14)}$$

This summarises the development so far. If the system is ergodic i.e. its probability density covers all the possible state points, then the time evolution of the system can be replaced by a static representation. An ensemble can be set up, each member of which is one microstate. This idea will be expanded below.

We now have to calculate all the possible energy states. To do this we need to calculate all the possible energies of the particles. This, often called the Hamiltonian, $\mathcal{H}$, of a particle is simply the sum of its kinetic and potential energies. As described in I.2.2, $\mathcal{T}$ the kinetic energy is dependent on the momentum of the particles in the system and $\mathcal{V}$, the potential energy, on some function of their positions.

$$\mathcal{H}(\mathbf{p},\mathbf{q}) = \mathcal{T}(\mathbf{p}) + \mathcal{V}(\mathbf{q}) \qquad \text{(E1.15)}$$

In order to calculate the particle energies, we have to solve this equation. This is the study known as quantum mechanics.

I.3. Quantum Mechanics.

At the turn of the century, scientists discovered a property of nature that had previously not been noticed. Systems (which were initially confined to harmonic oscillators and only later extended to all matter) cannot accept or radiate energy in a continuous manner. Instead the system passes between states that are a finite energy amount apart. This is Planck's hypothesis, summarised by the equation

$$E = h\nu \qquad \text{(E1.16)}$$

This was necessary to describe various atomic phenomena such as how energy is radiated by a black body, or the existence of a work function of a metal.

This concept is difficult as this is not how nature appears to man. Starting from these observations, a new theory was developed that not only could explain many seemingly anomalous phenomena, but also predict amazing new properties of atomic systems that were then found to be present. This overcame the early scepticism and quantum mechanics is now an accepted theory.

The consequence for a system is that only certain energies are allowed. These are discrete states, a finite distance apart and each particle is free to occupy any state, bound by a Boltzmann distribution and the type of the particle. The fundamental particles have what is called spin. This is an angular momentum that arises entirely from quantum mechanics. If the spin is half integral, the particles are fermions and only one pair of opposite spin are able to occupy each level. If the spin is integral, any number of these bosons can be in a particular state.

The particle energies in the Boltzmann distribution (E1.8) can now be seen to be the energy levels allowed to the system. We now need to calculate these allowed energies of the system.

De Broglie, from work on electronic diffraction, derived the relationship.

$$p = h / \lambda \qquad \text{(E1.17)}$$

This states that a particle with momentum p has a corresponding wavelength λ and so electrons can be diffracted like light. The conceptually difficult result of this equation is wave - particle duality; the property that all matter can be described by either particle or wave properties. Again this is implicit in the idea of quanta. Light, which is commonly thought of as a wave, must be corpuscular; a small packet of energy. That it acts like a particle is known due to its impact on small systems. In the same way particles can be measured as waves.

The important link from the classical mechanics to a mechanics incorporating these new ideas was the work of Hamilton. In trying to unify the motions of particles to waves, he deduced that particles follow a path according to a principle of least action in an analogous manner to the phase length of light. The action is defined by

$$S = \int \mathcal{L} dt \qquad \text{(E1.18)}$$

and this function is minimised for the path of a particle.

At this time there was no idea that a particle had a wave nature. Hamilton apparently used this principle in a purely empirical manner. In a manner analogous to the replacement of Lorentz's apparent time for real time, which produced Einstein's Theory of Relativity, Schrödinger replaced this apparent wave nature of particles by a real wave nature. This allowed him to derive from Hamilton's equations of motion, the time independent wave equation

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}(E-U)\psi = 0 \qquad \text{(E1.19)}$$

or in a more general form,

$$\hat{H}\Psi = E\Psi \qquad \text{(E1.20)}$$

where $\hat{H}$ is the quantum mechanical Hamiltonian operator that represents the energy of the system. This acts on a function, Ψ, known as its eigenfunction, to give the measurable energy E. This, the fundamental equation for quantum chemistry, beautifully shows the symmetry between quantum and classical physics with the two sides of the equation seemingly equal except that one side is an abstract operator equation representing the measurable single value shown on the other.

What is the function Ψ? It is commonly known as a wavefunction. Its interpretation was given by Born who said that $\Psi^*\Psi d\tau$ is the probability that the particle is in the space $d\tau$. A light wave in free space can be described by the equation

$$f(x,t) = \exp\left[2\pi i\nu\left(t - \frac{x}{c}\right)\right] \qquad \text{(E1.21)}$$

which as $E = h\nu$ and $p = h/\lambda$

$$= \exp\left[\frac{2\pi i}{h}(Et - px)\right] \qquad \text{(E1.22)}$$

This turns out to be an eigenfunction of the Schrödinger equation for a particle in a vacuum. The intensity, or amplitude squared, is a measure of the number of photons at that point in the wave, or equivalently the probability for a particular photon being present. Einstein in fact made this connection, calling the wave a "phantom field", guiding the photons.

This appears to solve the problems with the concept of wave-particle duality. Matter is corpuscular in nature, being packets of energy. It also has an oscillating electromagnetic field associated with it. The wave is the containing envelope of this field and the intensity of this field is proportional to the probability that the particle is at that point in space. Classical mechanics is to quantum mechanics what geometrical optics is to the wave theory of light.

I.3.1. The Uncertainty Principle

There is one other formulation of quantum mechanics. Before Schrödinger's wave mechanics, Heisenberg had published the first complete quantum mechanical theory. He rejected position and momentum as a basis of mechanics for the reason that they are not observable quantities. Using a heuristic approach and fitting to known experimental facts, matrix mechanics was developed. The crux of this theory is that observables are not single

values, but matrices. The central result, actually produced later by Born and Jordan was the commutation relation

$$xp - px = i\hbar \qquad \text{(E1.23)}$$

From this new theory, Heisenberg also deduced his famous Uncertainty Principle

$$\Delta x \Delta p \geq h / 4\pi \qquad \text{(E1.24)}$$

This is a direct consequence of the discrete nature of energy states. In phase space the energy states are separated by h in each degree of freedom. They are now no longer surfaces but volumes. The system can be anywhere in this volume and hence the uncertainty arises. The major result of this equation is that there is a fundamental limit on the accuracy with which we can measure the position or momentum of a system. The belief that everything is knowable with the correct equipment disappears. Heisenberg also believed that causality also disappears. While it can be known where a particle is on measurement, where it was between measurements and what it does is unknown.

This equation again emphasises what was wrong with classical physics. A small quantity h was missed in the description of the motion of particles. The two separate quantum formulations of wave and matrix mechanics are in fact equivalent. This can be seen by the way momentum can be described in both formulations by the same operator

$$\hat{p} = i\hbar \frac{\partial}{\partial x} \qquad \text{(E1.25)}$$

The difference in the two approaches is shown in the way they treat time dependence. In matrix mechanics the time dependence of the system is taken up in the property, in wave mechanics it is in the wavefunction. This mirrors the situation in the system mechanics talked about above, either the particles themselves move or else the probability density evolves in time.

The importance of the uncertainty principle for statistical thermodynamics can be seen as follows. If the energy of the states is allowed to vary continuously, the state points form a continuum and the sum E1.12 becomes an integral over phase space

$$U = \iint E_A P_A \mathrm{dpdq} \qquad \text{(E1.26)}$$

Owing to the uncertainty in the energy, this is in fact not the case and a modification must be made to this integral. A further modification must first be introduced. If the particles are indistinguishable, as in a gas or solution, then many of the arrangements are identical and so should not be counted. For N particles, each arrangement has $N!$ equivalent ways. The equation E1.7 should therefore be divided by this number and so

$$U = \frac{1}{N!} \iint E_A P_A \mathrm{dpdq} \qquad \text{(E1.27)}$$

Each state, an element of phase space, is bounded by its uncertainty. The volume of each element is given by $h^{d/2}$ where d is the dimensionality of phase space. Hence,

again for indistinguishable particles, a further correction term must be added to give the partition function for the system as

$$Q = \frac{1}{N!}\frac{1}{h^{3N}}\iint \exp(-E_A\beta)\mathrm{d}\mathbf{p}\mathrm{d}\mathbf{q} \qquad \text{(E1.28)}$$

I.4. The Similarity Between Quantum Mechanics and Statistical Thermodynamics.

The idea that has been introduced by quantum mechanics is that a particle is described, not by an exact position and momentum but by a probability. This idea has already been introduced for systems via the ergodic hypothesis. It was discovered that to correctly link observed properties of systems to their underlying particles, it was necessary to allow the system to cover all the possible arrangements i.e. the probability.

The relationship between Ψ and the probability for a total system introduced above, P_A, can be seen as follows. Ψ is a function of all the co-ordinates and momenta of the particles in the system. The system has various possible states, E_i, the solutions to the Schrödinger equation. By multiplying the Schrödinger equation by Ψ^* and integrating over all the variables gives

$$\iint \Psi_i^* \hat{H}\Psi_i \mathrm{dpdq} = \iint E_i \Psi_i^* \Psi_i \mathrm{dpdq} \qquad \text{(E1.29)}$$

This equation now has the same form as E1.26 from which the left hand side can be interpreted as the expectation, or average, value of the energy.

$|\Psi|^2$ is thus the probability density in phase space and so contains all the dynamical information about the system that when operated on by the appropriate operator can give all the properties of a chemical system.

This connection can be carried one stage further by looking at the time dependence of the wavefunction. This is given by the time dependent Schrödinger equation,

$$\hat{H}\Psi = i\hbar\frac{\partial\Psi}{\partial t} \qquad \text{(E1.30)}$$

which comes directly from Hamilton's work. If this relationship is derived in terms of the square of the wavefunction, this gives

$$E\Psi^2 = i\hbar\frac{\partial\Psi^2}{\partial t} \qquad \text{(E.1.31)}$$

which is similar to E1.11, the equation for the evolution of the probability density in phase space. Interestingly this equates $-\frac{E_A}{\beta}$ to $i\hbar$.

The probability distribution for a system derived from classical theory, is therefore related to a quantum mechanical wavefunction for the system. This wavefunction is

related to the wavefunctions of the separate states, which in turn are related to the molecular wavefunctions.

I.5. Generating An Ensemble

Equation E1.28 produces an idea for a different way to calculate the thermodynamic properties of a system. If the energy is an integral over the phase space accessible to the system, then rather than evolving the system with time and calculating the average energy, if it is possible to calculate all the possible energies the system can take, the average of this static picture will give the same result.

The system is taken and replicated an infinite number of times. Each replication is one possible state of the system. This whole construct is called an ensemble. The replications can interact with each other in different ways according to the conditions the system is under. For example if the system is at constant volume, temperature and composition, heat but no matter can be exchanged between the ensemble members, which are all of the same size. This is called the canonical ensemble. Alternatively, in the isobaric-isothermal ensemble the number of particles, temperature and pressure are held constant. Again heat but no matter can be exchanged, but this time different replications can have different volumes.

These ensembles can be imagined to be of infinite size. In this, the thermodynamic limit, all possible configurations are present in the numbers appropriate to their probability. This probability is given by equation E1.9. While this change of view makes the theory possible, the generation of an infinite ensemble is clearly impossible. There are now two problems; how to generate a large enough ensemble and how to evaluate the energy of each member.

At the moment we have derived an equation for the internal energy in terms of the average energy of an ensemble member. This in itself is not a particularly useful equation. Owing to the system constraints, each state has the same energy and this energy is U. Hence whether we measure one state, or all of the states and divide by the number of them the answer will be the same. The important thing is that we have changed from working with a system that evolves in time to a static ensemble which can be averaged over.

From this it can be realised that the Boltzmann distribution for the probability of being in a particular state can now be used to give the number of state points, n_A, with energy E_A. Even though this equation was generated using the condition that the energy was constant for all possible states of the system, it is still correct if the energy is allowed to vary.

From the energy equation E1.2, it can be seem that there are five ways a system can be kept by allowing pressure or volume, temperature or entropy, or the number of particles in the system to vary. One is the internal energy used above. The other four

possibilities for potentials are written below. These are the common thermodynamic equations.

$$dH = Vdp + TdS \qquad (NPE)$$
$$dA = -pdV - SdT \qquad (NVT)$$
$$dG = Vdp - SdT \qquad (NPT) \qquad \text{(E1.32)}$$
$$d\Omega = -pdV - SdT - \Sigma n_i d\mu_i \qquad (\mu VT)$$

Possibly the most important of these for chemists is the Gibbs Free Energy, G, due to the reason that these are the usual conditions chemical experiments are conducted under.

Firstly we shall look at a system equilibrated and then kept under conditions of constant enthalpy i.e. constant pressure and entropy. The constant pressure allows the system to expand and contract. When this happens, the energy with which the system was created can be divided into two parts, that which is held by the particles in the system and that used to do work on the outside world.

$$H = \mathcal{H} + pdV \qquad \text{(E1.33)}$$

We now have the probability of being in a state A of

$$P_A = \frac{\exp -E_A\beta}{Q} = \frac{\exp -(\mathcal{H} + pdV)}{Q_{NPE}} \qquad \text{(E1.34)}$$

where Q_{NPE} is the partition function summed over the possible states at constant NPE.

As there is no heat flow to the system, all the energy states are equal and so this again reduces to 1/**N**, where **N** is the number of states. The difference from the internal energy is that now the volume of the system is also a variable.

If the system constraints are relaxed further to allow heat to flow in and out of the system, it can be kept at constant temperature as well as pressure. The probability distribution is still given by E1.9, but now the partition function is over all the states possible in an NPT, the isobaric-isothermal, ensemble.

The probability of being in a particular state is given by the same equation as that for the enthalpy above, but now the partition function is over all the states of the NPT ensemble. The energy of each state can vary and the probability can now not be reduced to 1/**N**. The Gibbs function is thus given by the analogous equation to E1.14 in the isobaric - isothermal ensemble. i.e.

$$G = -RT\ln Q_{NPT} \qquad \text{(E1.35)}$$

It can be seen that there are general equations relating a thermodynamic potential,

$$\Psi_{ens} = -RT\ln Q_{ens} \qquad \text{(E1.36)}$$

or alternatively,

$$\Psi_{ens} = \frac{1}{h^{3N}N!}\int E_A P_A d\tau_{ens} \qquad \text{(E1.37)}$$

where $d\tau_{ens}$ is the ensemble phase space. If an infinite ensemble is set up, then this integral is just the average energy calculated over all the members of the ensemble i.e.

$$\Psi_{ens} = \langle E_A \rangle \qquad \text{(E1.38)}$$

All the physical properties of a system are related to the thermodynamic potentials by standard manipulations. All we now have to do is to set up an appropriate ensemble and evaluate its average energy. These system properties have also been connected to the properties of the constituent atoms. It is thus possible to set up a model of how the system works and test it.

I.5.1. Classical Methods

The obvious way to generate the ensemble is to solve the Schrödinger equation for the system. All the energy states are then known and the ensemble is specified. At present, the best computers can calculate a reasonable semi-empirical wavefunction for a 150 atom system. Full *ab initio* methods can treat up to 50 atoms. This is possible for discrete molecules. It is also possible to calculate the partition function for a large system of non-interacting molecules, such as a perfect gas. For these systems, the possible energy states of each molecule are due just its own kinetic and internal potential. These are the same for each molecule, and so the denominator of E1.9 reduces to give

$$Q = q_N \qquad \text{(E1.39)}$$

where q is the molecular partition function which can be obtained by *ab initio* computer methods.

By the ~~of the~~ Correspondence Principle of Bohr, classical mechanics is quantum mechanics in the limit of infinitely small* h or equivalently infinitely large wavelength. At this point, the uncertainty is in effect zero and the energy levels are not noticeably discrete. From this, larger entities should behave more and more like classical systems.

This is seen to be the case. While the exact description of a system might be a probability wave for sub-atomic particles extended throughout space, it is known from experiments that atoms can be defined that seem to behave in a classical way - it is possible to attribute both a position and a momentum to an atom or molecule. For example chromophores are identifiable in numerous different molecules. In quantum terminology, these atoms can be thought of as areas of space containing a certain nucleus and its typical number of electrons - or at least such a high percentage of these particles that for practical purposes they are entirely there.

If the properties of these atoms and how they interact are known, empirical potentials using parameters for the different atoms might be set up. Classical equations of physics could then be used to describe the system. This has the advantage of simplicity - a nucleus and its electrons are treated as one unit rather than separate entities. Newton's equations of motion, or any of the known descriptions of dynamics, ~~are much simpler to~~ can then be

* in the sense of negligible

easily solved to give dynamical behaviour
~~solve and so dynamical behaviour is also possible~~. This is now the domain of molecular mechanics and simulations.

The problem with this idea is that the description of atomic properties in a potential is very difficult. Electrons and nuclei to describe; are simple, the potential and kinetic energy operators can be deduced exactly. In contrast the interactions between atoms ~~is~~ are very complex and what is worse, this varies in behaviour depending on their surroundings.

To try and circumvent these problems, the latest theory, using *ab initio* techniques in a time dependent manner on large systems is the Car-Parinello method (Car and Parinello 1989; for review see Remler and Madden 1990). This is as yet only a few years old and still undergoing teething trouble. However it does promise well. This method calculates a wavefunction and then uses Newton's equations to allow it to evolve with time. On periodic systems, such as silicates, of the order of 100 atoms have been dealt with, including all electronic effects. (Bukowinski et al, 1989, Fiocchi et al 1992)

Two methods are available to create ensembles. Molecular dynamics (MD) solves classical equations of motion in a stepwise manner. At time t the energy and energy gradients on each atom are calculated. From a knowledge of the atomic positions and current velocities, it is possible to know where the system will be at time $t+\delta t$, providing the time step is not too large. Each step generates a new member for the ensemble. At the same time dynamic properties can be monitored and the behaviour of the system checked against experimental data. These simulations can also be used as a pictorial tool, showing the spatial interactions of molecules. This is often useful for seeing how a substrate binds to a macromolecule, or how molecules can approach each other in solution before reacting.

In contrast, Monte-Carlo (MC) simulations are only able to produce equilibrium data. The energy of the system is calculated. A random step is made and a new energy calculated. If the energy has decreased, the new configuration is accepted to be the start for the next step. If the energy has increased, the new configuration is accepted depending on a Boltzmann weighting. This enables high energy configurations to be sampled and small energy barriers surmounted (for general review of simulations see van Gunsteren and Berendsen 1990).

Now that the ensemble has been generated what can we do with it? As an example we shall derive the Gibbs free energy, but the same principles apply to any other function.

I.5.2. Calculating Free Energies

For chemical studies the free energy, usually the Gibbs free energy, is the most useful property to know. This is because it is directly related to equilibrium constants and other measurable quantities.

Writing E1.37 for the Gibbs free energy, with the probability P_A, explicitly included, we have the expression

$$G = \frac{1}{h^{3N} N!} \int \frac{E_A \exp - E_A}{Q_{NPT}} \qquad \text{(E1.40)}$$

Under *NPT* conditions, the Hamiltonian can vary and high energy states, with a low probability are thus included. In a simulation, which by necessity is of finite length, these high energy states are rarely sampled, and may be important in the integral. This is especially true in MD simulations as this technique keeps the system in the low energy conformations. It can be largely overcome by using Monte Carlo methods, but the size of the calculation is still a problem. MD is also the preferred method for dealing with proteins as the acceptance probability in MC is very small in a macromolecule where the motion is largely co-operative in nature.

I.5.2.1. Free Energy Perturbation Theory

To circumvent this problem a method based on the work of Zwanzig (Zwanzig 1954) has been developed that can realistically measure relative free energies. He developed an equation of state for the difference in free energy between two states of a system by considering the effect of a perturbation on the partition functions. His method can be adapted to fit any systems that can be linked by a possible perturbative change. Examples are differences in hydration energy, tautomeric equilibria, partition coefficients and so on.

If a system has two states, A and B, the difference in free energy between these states is related to the partition functions of the separate states by

$$\Delta G_{BA} = \Delta G_B - \Delta G_A = -kT \ln \frac{Q_B}{Q_A} \qquad \text{(E1.41)}$$

If these two states are very similar, the state B can be treated as a perturbation of A and the Hamiltonians are linked by

$$\mathcal{H}_B = \mathcal{H}_A + \Delta\mathcal{H} \qquad \text{(E1.42)}$$

The system work by both states is the same and so the partition functions can be written just in terms of the Hamiltonian $\mathcal{H}_A$,

$$\Delta G_{BA} = -kT \ln \frac{\iint \exp(-\mathcal{H}_A \beta) \exp(-\Delta\mathcal{H}\beta) d\mathbf{p} d\mathbf{r}}{\iint \exp(-\mathcal{H}_A \beta) d\mathbf{p} d\mathbf{r}} \qquad \text{(E1.43)}$$

This quotient contains the probability that the system is in a state A with energy $\mathcal{H}_A$ and so this becomes an ensemble average over the perturbation Hamiltonian

$$\Delta G_{BA} = -kT \ln \langle \exp(-\Delta\mathcal{H}\beta) \rangle_A \qquad \text{(E1.44)}$$

This equation is the FEP master equation. In theory it is possible to calculate the difference in free energy between any two similar systems. How this is actually done will

be discussed in chapter IV and then a different formulation will be produced and used in chapter VI.

I.6. Practical Applications

Before continuing with the thesis it might be useful to describe the major programs used for the calculations and how they apply the theories outline above.

I.6.1. *Ab initio* Calculations

Ab initio is the term given to calculations that involve solving the Schrödinger equation. First we must construct an energy operator, $\hat{H}$. This is constructed in an analogous manner to the classical Hamiltonian E1.15 above, but now the kinetic and potential energies are operators. Hence

$$\hat{H} = \hat{T} + \hat{V} \qquad \text{(E1.45)}$$

The kinetic energy is given by

$$\hat{T} = \frac{\hat{p}^2}{2m} = -\frac{\hbar^2}{2m}\nabla^2 \qquad \text{(E1.46)}$$

where ∇^2 is the Lagrangian, or second derivative operator. The potential energy depends on the system. However, for charged particles, such as electrons and protons, the energy for one particle is given by the interaction of the charge with the potential at that point. The total energy is thus given by the Coulombic interaction between all the charged particles

$$\hat{V} = \sum_{i,j} \frac{q_i q_j e^2}{4\pi\varepsilon_o r_{ij}} \qquad \text{(E1.47)}$$

This term causes problems. It is impossible to solve the Schrödinger equation exactly for systems with more than two particles i.e. the hydrogen atom. With only two particles, it can be written with one particle relative to the other and the distance between them determined. With more, this is not possible. For a many electron and nucleii system, the Hamiltonian operator is thus

$$\hat{H} = -\frac{1}{2}\nabla_e^2 - \frac{1}{2}\nabla_n^2 + \sum_{i\neq j}\frac{1}{r_{ij}} - \sum_{iA}\frac{Z_A}{R_{iA}} + \sum_{A\neq B}\frac{Z_A Z_B}{R_{AB}} \qquad \text{(E1.48)}$$

where A, B are the nucleii with nuclear charge Z_A and Z_B, and i, j the electrons. The first two terms are the kinetic energies of the electrons and nucleii respectively.

The fist step to make is to apply the Born-Oppenheimer approximation. In essence this is the hypothesis that the electrons move much faster than the nucleii. This means that the electronic and nuclear motions can be separated out. The approximation loses its validity for highly excited molecules where the nuclear motion is very fast, but is fine for

molecules in their ground state. However, this still only enables the hydrogen molecule ion H_2^+ to be solved exactly.*

The approximation that enables larger systems to be described is to construct a wavefunction for each electron, thus taking it in the field of all the other electrons. This involves approximating that the field experienced by electron *i* due to the other *n*-1 electrons is spherically symmetrical about the nucleus. The central field approximation allows the potential energy to be separated into a radial and spatial part. Hence the exactly known hydrogenic spatial orbital shapes apply to all atoms.

I.6.1.1. Self Consistent Field Theory

The Hamiltonian can be written as nearly a summation of one electron operators

$$\hat{H} = \sum_i h(i) + \sum_{ij} \frac{1}{r_{ij}} \qquad \text{(E1.49)}$$

where *h(i)* is an operator for electron *i* moving around the fixed nucleii. This points towards the wavefunction being a product of one electron wavefunctions.

$$\Psi = |\phi_1 \phi_2 \phi_3 ..... \phi_N\rangle \qquad \text{(E1.50)}$$

If the Hamiltonian was exactly

$$\hat{H} = \sum_i h(i) = -\frac{1}{2}\nabla_i^2 - \sum_{iA} \frac{Z_A}{R_{iA}} \qquad \text{(E1.51)}$$

then this function would be an eigenfunction of the Hamiltonian with eigenvalue of

$$E = \Sigma \varepsilon_i \qquad \text{(E1.52)}$$

One modification must be made to this product. It has been found that there are two different types of particles. Fermions have spin 1/2, bosons integral spin. In general, the former are particles that form matter; electrons and protons belong to this class, while the latter form forces; photons, gravitons etc. belong to this class. The peculiar property of fermions is that their wavefunction must change sign on exchange of particle positions. This is the general result from the Pauli principle that no two electrons ~~must~~ may share the same quantum numbers. In order to form this anti-symmetrised function, the product given above becomes a Slater determinant,

$$\Psi_{SL} = (N!)^{-1} \sum_i^{N!} (-1)^i \mathcal{P}_i (\phi_1 \phi_2 ..... \phi_N) \qquad \text{(E1.53)}$$

Often just the diagonal is written in which case this wavefunction has the same form as the Hartree product E1.50.

Solving the Schrödinger equation with this wavefunction

$$\langle \Psi_{SL} | \hat{H} | \Psi_{SL} \rangle = E \qquad \text{(E1.54)}$$

* or any one electron molecule

writing the Hamiltonian as the above sum. When integrating over all space, due to the orthogonality of the individual functions, the only terms that do not go to zero are

$$E = \sum_i \langle \phi_i | \hat{h}(i) | \phi_i \rangle + \sum_{i \neq j} \langle \phi_i \phi_j | r_{ij}^{-1} | \phi_i \phi_j \rangle - \sum_{i \neq j} \langle \phi_i \phi_j | r_{ij}^{-1} | \phi_j \phi_i \rangle \qquad \text{(E1.55)}$$

the second term is called the Coulomb, the third term the exchange integral. With a little manipulation, this can be written as a set of one electron integrals

$$\hat{f}\phi_i = \varepsilon_i \phi_i \qquad \text{(E1.56)}$$

where $\hat{f}$ is called the Fock operator.

Conceptually each electron is now being treated in the field due to all the others, contained in the Fock operator. and so their mutual interactions are not dealt with directly. A wavefunction can be calculated, from which the molecular potential is known. The Schrödinger equation can then be solved for one electron in this field and a new potential calculated. This next electron can be dealt with in the same manner and a new potential calculated. Each electron can be treated in turn, going around the molecule in repeated cycles, calculating a new potential each time, until the potential does not change i.e. a self consistent field (SCF) has been obtained. However, in practice this is not how the calculation is made and matrix operations are performed. For a molecule, the molecular orbitals are represented as linear combinations of atomic orbitals (the LCAO approximation) i.e.

$$\psi_a = \sum_i c_i \chi_i \qquad \text{(E1.57)}$$

and these orbitals provide a basis set for the calculation to work in.

Various functions have been proposed to expand the orbitals. The ones usually used by computers are combinations of gaussian functions. Exponential functions give better results (e.g. Slater orbitals) but gaussian functions are computationally easier. In general, the more gaussian functions that are used to describe an orbital, the better the basis set is at representing the wavefunction for the system i.e. one of lower energy as predicted by the variation principle

When the one electron equation E1.56 is written with the molecular orbital in E1.57, this equation can now be written as a matrix equation, known as the Roothaan equations. The problem is now to find the expansion coefficients, c_i. An *ab initio* program works in the following way. The integrals over atomic orbitals, are calculated for the given basis set and geometry and then stored (if there are a very large numbers of integrals they have to be calculated when they are needed). An initial guess is made at the coefficients. This can be done by use of a simple MO theory such as extended Hückel, or a semi - empirical method. The Roothaan equations are now solved, using matrix techniques, for these coefficients and the stored integrals. Improved coefficients are then

calculated from the molecular energies obtained from this first step. These are then used for the next cycle. This iterative process is continued until convergence is achieved i.e. a self consistent field is found.

There is a different way that this problem can be treated. The Fock operator contains the field due to the electrons in terms of the wavefunction. For large numbers of electrons this can be replaced by the electron density. This method, known as density functional theory (see Lundquist and March 1983) is just gaining beginning to gain respect amongst chemists and will not be dealt with here.

I.6.1.2. Beyond Hartree - Fock

There are two major problems with Hartree-Fock theory. One is the problem that the basis set is finite. However, if the basis set is made larger, there will come point when it, for practical purposes, approaches infinity. The energy will converge and is said to be exact in the Hartree-Fock limit.

The second problem is more fundamental and difficult to treat. In an atom or molecule, electronic motion is to some degree coupled due to the Pauli principle. Because they cannot have the same quantum numbers, electrons have a tendency to stay apart and a potential hole is formed around each one. Hartree-Fock does couple electrons in this way, but only if they have the same spin. Thus there is a finite possibility that electrons with different spin can occupy the same point in space. This is a consequence of the orthogonality of spin as treated in the Slater determinant which removes all 'exchange' terms between these electrons. This lack of correlation makes the molecule more stable than Hartree - Fock methods calculate as the tendency for electrons to keep apart lowers the repulsion between them. This energy difference, the electron correlation energy is defined as

$$E_{corr} = E_{exact} - E_{HF} \qquad \text{(E1.58)}$$

where E_{exact} is the exact non - relativistic energy of the molecule and E_{HF} the energy in the Hartree-Fock limit. The solution to this problem is to allow an electron to occupy any orbital, occupied or virtual, thus mixing in excited states to the ground state wavefunction. The exact wavefunction is thus written as an expansion of all these possible Hartree-Fock wavefunctions

$$|\Phi_0\rangle = c_0|\Psi_0\rangle + \sum_{ar} c_a^r|\Psi_a^r\rangle + \sum_{\substack{ab \\ rs}} c_{ab}^{rs}|\Psi_{ab}^{rs}\rangle + \ldots. \qquad \text{(E1.59)}$$

where the electron in orbital a or b has been excited to the virtual orbital r or s. The problem is now to find not only all the wavefunctions, but their expansion coefficients. Including this configuration interaction (CI) makes any calculation much larger. Fortunately the main contribution to the energy comes from the doubly excited terms, and

the contribution from this term can be included using a perturbation treatment. This is called Møller-Plessett second order perturbation theory (MP2) (Møller and ~~Palest~~ Plesset 1934) and it can account for up to 85% of the correlation energy, providing the initial wavefunction is a reasonable one.

I.6.1.3. GAUSSIAN88 / 90

All the *ab initio* calculations in this thesis were made using GAUSSIAN88 and later GAUSSIAN90. These are SCF programs, based on the HF method but also able to do CI, MP2 etc. Starting from a set of co-ordinates and atomic numbers, the program can calculate a wavefunction. From this it can optimise the structure to find the local minima. Post SCF analysis allows various properties such as electrostatic potentials to be calculated. The only other input necessary is the level of approximation wanted; which atomic orbitals are included in the LCAO. As the name implies, these orbitals are approximated by combinations of gaussians. How many of these gaussians are used to make up the orbitals must also be specified.

These different orbital sets form the basis set for the calculation. There are known standard sets with orbital coefficients for the common atoms stored in the program. Those used in the thesis will now be briefly described. For a full account of these orbitals see Hehre *et al.* (1986).

STO-NG This is a minimal basis set. Each Slater orbital for quantum numbers l,m and n has a function $\chi_{lmn} = r^{n-1}\exp(-\xi r)Y_{lm}(\theta,\phi)$ is described by one function. This basis set is a combination of N gaussian functions fitted to a Slater orbital by a least squares fit technique. The gaussian exponents are then found by Hartree-Fock SCF optimisation. The only set of this type used was with N = 3 i.e. the STO-3G (Hehre *et al.* 1969) basis set.

M-NPG Basis sets with this notation are termed split basis sets. The core atomic orbitals are defined by M gaussians, while the valence orbitals are split into an inner and outer part. The inner part consists of N gaussians and the outer has P gaussians. 3-21G (Binkley *et al.* 1980) and 6-31G (Hehre *et al.* 1972) were used. The valence orbital exponents for 3-21G were calculated using a 6-21G basis set to prevent the valence orbitals 'falling in' to the core under optimisation due to the lack of gaussians in the core. 3-21G gives good geometries, due to these high quality valence orbitals, but poor energies.

The advantage of these split basis sets over the minimal basis set is that to a small degree the atomic orbitals can adapt to the molecular environment. This is achieved because the molecular orbital is a linear combination of the two parts of the orbital. By taking more or less of the outer part, the orbital can expand or contract.

*, ** (Binkley *et al.* 1977) A further way that orbitals can adapt to the environment is to add polarisation functions. For a hydrogen atom this involves adding a set of 2p orbitals, for first row atoms a set of 3d orbitals. This allows directionality to the molecular orbital to be added. These are added to the standard basis sets e.g. 6-31G* has polarisation functions added, but on heavy atoms only; 6-31G** has these functions also on hydrogen atoms. 3-21G* has polarisation functions added only onto second row atoms.

+,++ (Clark *et al.* 1983) In negative ions, the electron density is further from the nucleus. It is then necessary to add diffuse functions. These are p orbitals that are larger than normal. Again these are added to the standard sets e.g. 6-31+G or 3-21+G. A single + means diffuse functions are added to just heavy atoms, ++ includes the hydrogens.

I.6.2. Semi-Empirical Methods

With the help of supercomputers, it is now possible to study molecules up to around 40 atoms with a fair degree of accuracy, although by this point the approximations start to be become evident. Larger systems are impossible due to the fact that the size of a problem scales as n^2N^2 where n is the number of electrons and N the number of basis functions. The next level of approximation is the so-called semi-empirical programs. These are based on SCF methods, but use parametrization to solve some of the two electron integrals and ignore others, rather than evaluating them. Only the valence electrons, treated by a minimal basis set, are treated. These move in a field due to the nucleii and core electrons constructed from atomic parameters. The parameters are chosen so as to reproduce experimental results, such as heats of formation. By ignoring, or simplifying the hard parts of the SCF calculation, this parametrization means that the calculation is much simpler and requires less computing power. It also has the advantage that some of the electron correlation is included indirectly in the experimentally determined parameters.

I.6.2.1. MOPAC

The most commonly used program to calculate wavefunctions using these semi-empirical methods is MOPAC (Stewart). There are various different levels of approximation possible with this program. The commonly used Hamiltonians are called AM1 (Dewar *et al.* 1985) and PM3 (Stewart 1989). They are updated versions of the MNDO (moderate neglect of differential overlap) (Dewar *et al.* 1977) approximation, for which integrals involving different centres are parametrised. The various methods have different strengths and weaknesses and only trial and error can really tell which is best for a particular problem.

I.6.3 Molecular Mechanics

As mentioned above, the problem with molecular mechanics is the form of the potential function to be used. This is critical. The potential both evolves the system in time and calculates the energy. If this is bad, both the ensemble members will be a poor approximation to the real thing and the energy will be poor. An atom is described by its electronic orbitals. In a system, these orbitals can interact and overlap with those of a neighbouring atom. A bond is then said to have formed and a molecule is present. Alternatively, there can be interaction with no appreciable overlap. The potential between atoms can thus be divided into two separate sorts of interaction; bonded or non-bonded. The Born-Oppenheimer approximation is also implicitly used. The nucleii are thus able to move over a potential surface generated by the electronic interactions

The intra-molecular part of the potential is easier to describe. In their ground state, molecules are known to act like harmonic oscillators. There are 3N-6 degrees of internal freedom for a non-linear molecule where each atom can describe simple harmonic motion in any direction around its mean point. The motion can then be described by the normal modes of vibration which are orthogonal linear combinations of these motions. Unfortunately these normal modes are peculiar to each molecule and so are not a useful property for an empirical potential.

It is therefore easier to turn to internal co-ordinate space. The degrees of freedom can then be described along each bond, angle or dihedral in the system. The former two types of freedom can then be described by harmonic functions as they are known to behave this way from spectroscopy. The dihedral function is more difficult. Owing to the possibility of internal rotations the function is not harmonic but must be some sort of phase function, with a potential barrier which allows different amounts of rotation depending on the energy present. A function of this form can be parametrised with spectroscopic data and actually works quite well in reproducing geometries.

The non-bonded terms pose a much bigger problem. They are due to the mutual attraction or repulsion of the electron cloud around each nucleus. The problem is that unlike simple charges, these charge clouds are in motion. They are also soft and distorted by neighbouring charges. This makes the interaction very hard to model. It is found that there are four possible types of interaction. The total energy is thus described by

$$E_{\text{Non-Bond,Tot}} = E_{\text{Electrostatic}} + E_{\text{Induction}} + E_{\text{Dispersion}} + E_{\text{Repulsion}} \qquad \text{(E1.60)}$$

The electrostatic term is due to the static electronic distribution. This gives rise to the molecular dipole and higher moments, or in the case of ions also a charge is present. The easiest way to represent these interactions is by placing a charge on each atom centre. These charges represents the electron distribution and reproduces all the electronic moments.

Once a suitable form for a potential has been chosen, the problem is to try to create a simple system that will deal with all types of atoms in all types of molecules. The usual way this is done is to chose common chemical atom types using intuition. Each type is then given a set of parameters so that, when combined with the potential function, the correct interactions are modelled. This is often very difficult as the compromise between a small number of atom types for simplicity and exact parameters for each system must be made. A nitrogen atom may be similar in all simple amines and so one set of parameters should cover all of these. What happens if a strongly electronegative group is nearby? It will certainly be different from an amide nitrogen, but how similar is an amide nitrogen in a protein backbone to that in acetamide etc.? This problem is the weak point of all empirical methods.

Once the potential has been set up, it is a fairly simple task to solve a set of dynamics equations to describe the trajectory. The next problem is how to apply the desired constraints to measure the correct ensemble, constant temperature and pressure etc. There are various different ways this can be achieved. They will be described as used in just one program AMBER3.1 (Singh *et al.* 1988). This was used for the majority of calculations in this thesis and is representative of the typical way these problems are tackled.

I.6.3.1. AMBER

The potential function used in AMBER is (Weiner *et al.* 1984)

$$E = \sum_{\text{bonds}} K_b (r - r_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}(1 + \cos(n\phi - \gamma)) + \sum_{ij} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{K_e q_i q_j}{\varepsilon r_{ij}} + \sum_{\text{H-bonds}} \frac{C_{kl}}{r_{kl}^{12}} - \frac{D_{kl}}{r_{kl}^{10}} \qquad \text{(E1.61)}$$

This is possibly the simplest form of a potential ~~possible~~. It has terms for all of the types of energies discussed above, with the exception of inductive effects. This missing term will be further discussed in chapter V. The addition of a hydrogen bonding term was deemed necessary to keep these bonds at the correct length. This demonstrates a shortcoming of the potential. This should not be necessary as hydrogen bonds should be dealt with by the van der Waals (in this case a Lennard-Jones term) and electrostatic interactions. As a contrast, the MM2 potential (Allinger, N.L. 1977) will be given. This has proved to be extremely accurate for small molecules. In a study of 55 random alkanes, agreement of bond lengths to 0.01Å, angles to 2° and heats of formation to within 1 kcal mol^{-1} were obtained (Engler *et al.* 1973).

$$E = \sum_{\text{bonds}} K_b(r-r_0)^2(1-2(r-r_0)) + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2(1+K'_\theta(\theta-\theta_0)^4)$$
$$+ \sum_{\text{dihedrals}} \frac{V_1}{2}(1+\cos\phi) + \frac{V_2}{2}(1-\cos 2\phi) + \frac{V_3}{2}(1+\cos 3\phi)$$
$$+ \sum_{\text{stretch-bend}} K_{sb}(\theta-\theta_0)(r-r_0)(r'-r'_0)$$
$$+ \sum_{ij} K_{ij}\left(k'\exp(-12.5\left(\frac{r_i+r_j}{r}\right) - k''\left(\frac{r_i+r_j}{r}\right)^6\right)$$
$$+ \sum_{\text{bond dipoles}} \mathrm{K}_\mu\left(\frac{\mu\mu'(\cos\chi - 3\cos\alpha\cos\alpha')}{\varepsilon r^3}\right)$$

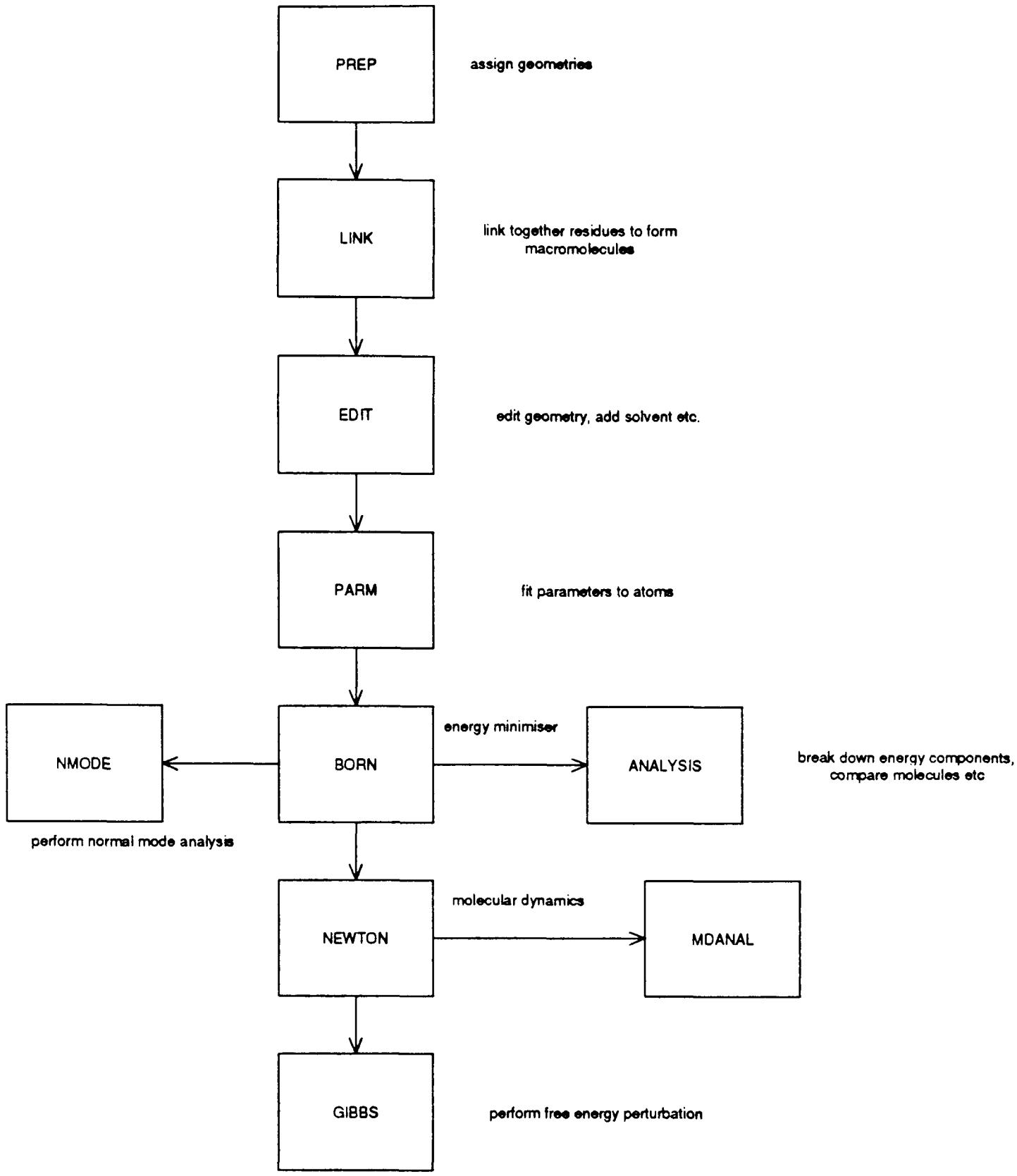
E(1.62)

The exact meaning of all these variables is not important. This equation is being used as a demonstration of how complex an empirical potential must be to model accurately even simple systems such as alkanes. As well as the anharmonicity built into the bond and angle stretching terms, the addition of a bond - angle combined stretch is interesting. The non - bonding terms are also interesting. Notice that an exponential repulsion term is used, rather than the r^{-12}. This is a more accurate, but also computationally more difficult form of this potential. Notice also that bond dipoles rather than coulombic forces are used for the electrostatic terms.

These electrostatic terms are in some ways the hardest to model. This is due to the problem of how the electron density is best modelled. In AMBER, the approach is to produce a set of charges on the atomic centres that reproduce all the electric moments of the molecule. This is done by fitting charges at these sites to reproduce the molecular electrostatic potential (MEP) calculated from an *ab initio* wavefunction. This method has become popular, but is beset by many problems, as will be discussed in chapter V.

Owing to its simplicity, AMBER does not reproduce simple molecule properties as well as MM2. However both are found to be comparable for macromolecules. When the crambin crystal structure was energy minimised by both methods, MM2 produced a structure with a radius of gyration 1.1% larger, whereas the AMBER structure was 1.1% smaller. (Lii *et al.* 1989). These large systems are found to be in general less sensitive to the potential. Possibly the rigidity of the system and its shear size dampen out any inaccuracies.

AMBER has the advantage that it has a large selection of ready tested parameters, its force field (Weiner *et al.* 1984 and 1986). It is also flexible enough that other parameters can be easily added if needed. It is modular in design and has that enormous advantage that is supplied with the source code, which enables the user to check exactly what the program is doing and to make alterations if necessary. The modules are described below in the flow chart F1.1.



F1.1. The functions and links between the various modules of the program AMBER.

I.6.3.2. Running a Simulation in AMBER

The usual way a simulation is run is as follows. A geometry for a molecule is fed into PARM. This comes from a crystal structure, standard bond data, imagination or wherever. LINK is used to join these residues together, enabling the amino acids for proteins, or the bases for DNA, to be separately specified only once and linked together in the appropriate order with any other substrates present. EDIT is then used to tailor the system to desired specifications. The macromolecule is matched to its crystal structure,

water is added in specified areas and so on. Also the whole system can be solvated. This is done by boxes of Monte - Carlo minimised waters being placed around the system and any waters that overlap the solute being discarded according to specified criterion, usually that the water atoms are more than 2Å from any solute atom. In PARM the system set up so far is matched to the parameters in the force field. The system is now ready.

I.6.3.3. Optimising the System

The first step is to run the system through the energy minimiser in BORN. This calculates all the interactive energies between the atoms and the gradient of these energies determine the direction of motion. It cycles through this procedure until the gradient of the energy, or the energy itself, is nearly converged on a specified value. Two implementations are used to speed up the process. The most time and memory consuming part of a molecular mechanics calculation is working out the non-bonded pairs i.e. how far all the atoms are from each other and their interactions. The calculation is speeded up by working out this list only once every few iterations e.g. 100 steps. This approximates that the atoms do not move far enough for these distances to radically change during this time, which is reasonable. To save memory and time however, a cutoff is used. If the atoms are more than a certain distance, usually 8Å, apart this interaction is ignored. This is not such a good approximation, especially when electrostatic forces are involved. The problem of modelling the electrostatics without increasing the memory requirements too much will be returned to in chapter IV. Due to the non-bonded interactions, a problem will increase in size as N^2, where N is the number of atoms, Hence while a 6000 atom system may be feasible with a given resource, a 6500 atom system may not.

This process removes any bad forces that may have occurred on setting up the system. It also allows any solvent to fill in the gaps left by the method of solvation. The system will now be at its energy minimum according to the parameters used. These minimised structures may be of value in themselves. A search for various minimum energy conformers may be required for example. However, in general these are just the prelude to dynamics simulations.

I.6.3.4. Simulating the System Motion.

Simulations are done in NEWTON. As the name suggests, this solves Newton's equations of motion. These cannot be evaluated exactly over a long trajectory. Instead a time step, δt, is used and these second order differential equations are integrated to predict the positions and accelerations at this new time. This is implemented in AMBER using the leap-frog algorithm (Hockney 1970). The minimised structure is taken. All the forces between the atoms are calculated, again using the non-bonded list and cutoff described

above. From the forces the accelerations on each particle are known. The velocities are then calculated at time ½δt from

$$\mathrm{v}(t+\frac{1}{2}\delta t)=\mathrm{v}(t-\frac{1}{2}\delta t)+\delta t\mathrm{a}(t) \qquad \text{(E1.63)}$$

The new positions at time t+δt are then obtained from

$$\mathrm{r}(t+\delta t)=\mathrm{r}(t)+\delta t\mathrm{v}(t+\frac{1}{2}\delta t) \qquad \text{(E1.64)}$$

and the process is repeated. This method has the advantages of simplicity and, unlike other methods has the velocities explicitly calculated, although the velocity at time t must be calculated from the average of those at the half times when this is needed in the total energy.

I.6.3.4.1. Keeping the Pressure and Temperature Constant

One of the problems with the Newtonian equations of motion is that it is the energy that is conserved. In order to keep the temperature or pressure constant, it is necessary to introduce constraints. From the virial theorem. It is possible to write an instantaneous temperature

$$\mathcal{T}=\frac{1}{3Nk}\sum_i\frac{|\mathbf{p}_i|^2}{m_i} \qquad \text{(E1.65)}$$

In order to keep the temperature constant, it is possible to add a constraint to the system, scaling the velocities by the factor

$$\chi=\left(1+\frac{\delta t}{t_T}\left(\frac{T}{\mathcal{T}}-1\right)\right)^{\frac{1}{2}} \qquad \text{(E1.66)}$$

(Berendsen *et al.* 1984) where t_T is a relaxation factor designed to perturb the system as little as possible. In water this is found to be 0.4ps. This velocity scaling is the advantage of the leap-frog algorithm which deals explicitly with velocities so that they are readily available. The temperature is found to be kept constant by this method to within 5K in a typical simulation

In a similar manner, an instantaneous pressure function can be defined as

$$\mathcal{P}=\frac{N}{V}k\mathcal{T}+\frac{\mathcal{W}}{V} \qquad \text{(E1.67)}$$

where $\mathcal{W}$ is an internal virial. Again this can be made a constant of the motion, this time scaling the box dimensions by a factor

$$\chi=1-\beta_T\frac{\delta t}{t_p}(P-\mathcal{P}) \qquad \text{(E1.68)}$$

(Berendsen *et al.* 1984) where β_T is a compressibility factor, which at STP has a value of 44.6 bar and t_p again is a relaxation constant in the range 0.1 - 0.4 ps. The co-ordinates are therefore scaled by a factor of $\chi^{1/3}$. This method, while giving an average constant pressure is not sensitive enough to instantaneous fluctuations and during the course of a simulation the pressures typically vary by ±200 atm.

I.6.3.4.2. The Description of the Solvent

One problem of simulations in solution is how to describe the solvent. The first problem is the water potential. Water has proved to be very difficult to describe empirically. AMBER uses the TIP3P (Jorgensen *et al.* 1983). TIP4P, which has a charge not on the oxygen atom, but in between the O - H bonds, is a better model, but the TIP3P water is adequate and simpler.

The nest problem is how to simulate a solvent phase. A simple sphere of solvent around the solute has three problems. Firstly it must be large enough to avoid edge effects i.e. to the ~~solvent~~ solute it must appear like bulk water, whereas the molecules at the edge of the solvent are under co-ordinated. This can only be determined by trial and error. Secondly it must be possible to keep the water in the sphere and not boil off into outer space reducing the solvation. This can be attained by adding a cap force pulling the surface waters back into the solvent. The third problem is however insoluble. Using a simple sphere it is impossible to keep the system at constant pressure.

The solution to all these problems is to use periodic boundary conditions. The solvent is in a box. When a solvent molecule reaches the edge of the box and is about to move out of it, its image appears at the other side entering the box. Interactions behave in the same way so that all the waters are correctly solvated as they "see" waters all around them from the periodic images. A smaller amount of water can now be used as there are no edge effects and constant pressure is kept by scaling the box dimensions. The only criterion necessary for this to work is that the box side is over twice as big as the non-bonded cutoff. If this is not the case, a molecule may interact with the same molecule twice; once the real version and once a periodic image.

I.6.3.4.3. Constrained Dynamics; SHAKE

One problem with the time step integration method of solving Newton's equations is that the time step must be small enough to prevent instabilities in the trajectory occurring. An example of this happens if the time step is longer than a typical vibrational frequency of a bond. If this is so, it is possible that a step will sample the bond at its full stretch. The next step may then miss the return movement and break the bond. The highest frequency modes are for bonds involving hydrogens, due to their low weight and short bond lengths. It is generally found that if the time step is longer than 0.0005ps then this

bond breaking occurs. Longer time steps can be achieved by constraining these degrees of freedom.

The problem in adding arbitrary constraints is to find out how they affect the simulation. Newton's equations of motion are calculated in Cartesian space. In order to work out the effect of constraining the bond lengths, it is necessary to transform to internal co-ordinates (van Gunsteren 1980). When this is done it is found that the ensemble average with constraints is the same as the unconstrained average, providing the parts removed from the energy are independent of the others. This is so for bonds, providing the angles do not vary far from their equilibrium distances, a reasonable approximation if the angle potential is harmonic. It is not true for other degrees of freedom.

The common way of constraining bonds is to use the SHAKE algorithm (Ryckaert *et al.* 1977). This iteratively cycles around a molecule, adjusting the bond lengths to their equilibrium distances, until a specified criterion is met. There is still a limit on the time step. If the molecule moves too far from its equilibrium distances, adjusting the first bond length will leave the next further from the required distance. This gets worse along a chain until finally the resetting is not possible. The maximum time step using this approach is 0.002ps, a four fold increase in speed.

I.7. Summary of Chapter I

In this chapter the theory to be implemented in the thesis has been presented. The first sections dealt with describing a system in terms of its energy, the theories of thermodynamics. The next section, by the use of the ergodic hypothesis, went on to explain these system properties in terms of the molecular nature of the system, statistical thermodynamics.

In order to calculate the energies of these particles it was the necessary to turn to the ideas of quantum mechanics, the study of energetics of systems small compared to the wavelength of light. From this it was discovered that the ergodic hypothesis actually had produced a system that behaved in a similar manner to a quantum system, even when no allowance for the discrete nature of energy states had been made.

From this it was also deduced that it was necessary to set up an ensemble of all the possible energy states of a system, under specified conditions, in order to calculate a thermodynamic quantity such as the Gibbs free energy.

The last sections of the chapter were then devoted to describing how this is done in practice with the use of SCF theory for quantum mechanical calculations on discrete molecules and classical simulations for calculations on many molecule systems. At the same time the workings of the programs used in this thesis, GAUSSIAN and AMBER were described.

I.8. Bibliography

The most useful general texts for the theory discussed in this chapter are

Allen, M.P. and Tildesley, D.J. "Computer Simulations of Liquids" OUP (1987)
Wonderful overview of all the techniques used in classical simulations.

Atkins, P.W. "Physical Chemistry" Second Ed. OUP (1982)
The best general physical chemistry textbook, with good introductions to both quantum mechanics and statistical thermodynamics.

Atkins, P.W. "Molecular Quantum Mechanics" Second Ed. OUP (1983)
General text on quantum mechanics.

Jammer, M. "The Conceptual Development of Quantum Mechanics" Second Ed. American Institute of Physics, Tomash Publishers (1989)
A history of quantum mechanics, enabling the meanings behind the jargon to be understood.

Reichl, L.E. "A Modern Course in Statistical Physics" Arnold (1980)
Very thorough text on statistical mechanics, with a good section on the ergodic hypothesis.

Szabo A. and Ostlund, N.S. "Modern Quantum Chemistry" Revised first Ed. McGraw - Hill (1989)
Detailed text on SCF *ab initio* methods.

CHAPTER II

An Important System To Study - The *Ras* Proteins

The last chapter discussed how a chemical system behaves and how it may be simulated using computers. Now we need a system to explore, preferably one that both tests these methods and is interesting in its own right. Such a system is the product of the RAS gene. These are relatively small proteins, known as p21 or simply *ras* proteins, for which a crystal structure has recently been determined. While a lot of data about them have been gathered, there are still many parts of their behaviour and interaction in the human body that are not understood. There are a few areas were simulations could possibly be of use. In this chapter we shall explore this system, analyse some of the experimental data that has been produced about its behaviour and try to point out some areas where theoretical study could be useful.

Firstly the general characteristics and basic biochemistry of the *ras* family of proteins will be described. Then the more difficult topic of its role in human physiology will be discussed in the light of known data. Finally a scheme will be proposed that this thesis will aim towards solving.

The nomenclature used is that the gene is denoted by capitals i.e. RAS and the protein by lower case italics, i.e. *ras*.

II.1. The *Ras* Proteins

The RAS gene is present in all types of human tissue. Similar genes are also present in many other organisms, demonstrating their importance. There are three different *ras* proteins. All are 21kD in weight, hence they are given the generic name of p21, and are composed of 189 amino acids. The types go by the names of Kirsten (K-*ras*), Normal (N-*ras*) and the most studied Harvey (Ha-*ras*).

They have many features in common, belonging to a large class of proteins that bind nucleotides, often called G - proteins. The main function of all these proteins appears to be to act as switches, transmitting and often amplifying signals. This is achieved by being able to bind the nucleotide in different states of phosphorylation. The physiological importance of this class of proteins, including *ras*, is that they are somehow involved in the signal transduction of growth factors.

By comparison with the other G-proteins, the *ras* proteins have received a huge amount of study. The reason for this is that RAS is a proto-oncogene. An oncogene is a gene whose product is involved in transforming cells to produce cancerous growth. Proto-oncogenes are common cellular genes that under certain conditions are mutated. This mutation produces a protein that is slightly different from the wild type. The mutant

protein then malfunctions in some way to produce abnormal cell growth. The majority of oncogenic products are proteins involved in the growth signal pathway. The connection is clear. If this cycle is disrupted in some way, abnormal growth can occur.

Ras proteins are among the most commonly mutated proteins found in human cancers. In over 30% of human cancers they have undergone point mutations, where single residues in the protein differ from the normal. In many more cases they are overexpressed i.e. there is a larger concentration of the normal protein in the cancer cells. Because of this much effort has been put into their study. Ha-*ras* was also the first oncogene product with a completely characterised crystal structure.

The three members of the family are involved in different cancers. Kirsten is the most active; around 95% of pancreatic cancers have a mutant form of this protein. The differences between them are small, but make the proteins act in a similar way in specific cell environments.

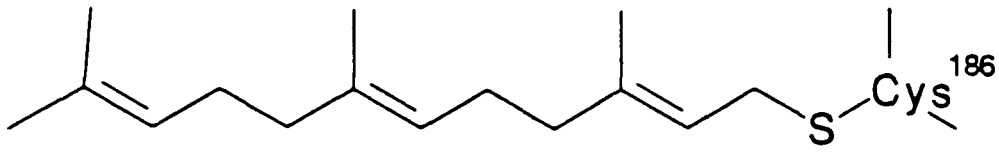
II.2. The Formation Of the Protein

The amino acid sequence of a protein is determined by the ordering of base pairs in a section of DNA. When a signal is received by the nucleus of a cell the DNA double helix unwinds in this region and via a complex signal system involving RNA causes free amino acids to be joined together in the desired sequence. The Ha-*ras* protein sequence is shown below. The other two *ras* proteins have a very high homology with this sequence, especially in the region of residues 1 to 120 which contains the active site of the protein (Barbacid 1987). In fact all the G - proteins have many similarities with this primary sequence, showing the similarity of function.

```
                5         10        15        20        25        30
  1 M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D
 31 E Y D P T I E D S Y R K Q V V I D G E T C L L D I L D T A G
 61 Q E E Y S A M R D Q Y M R T G E G F L C V F A I N N T K S F
 91 E D I H Q Y R E Q I K R V K D S D D V P M V L V G N K C D L
121 A A R T V E S R Q A Q D L A R S Y G I P Y I E T S A K T R Q
151 G V E D A F Y T L V R E I R Q H K L R K L N P P D E S G P G
181 C M S C K C V L S
```

T2.1. The primary structure of the Ha-*ras* protein.

The protein is formed in the cytosol of the cell. *Ras* proteins then undergo what is known as post translational processing. Cys186 is turned into the thioether shown in F2.1 .



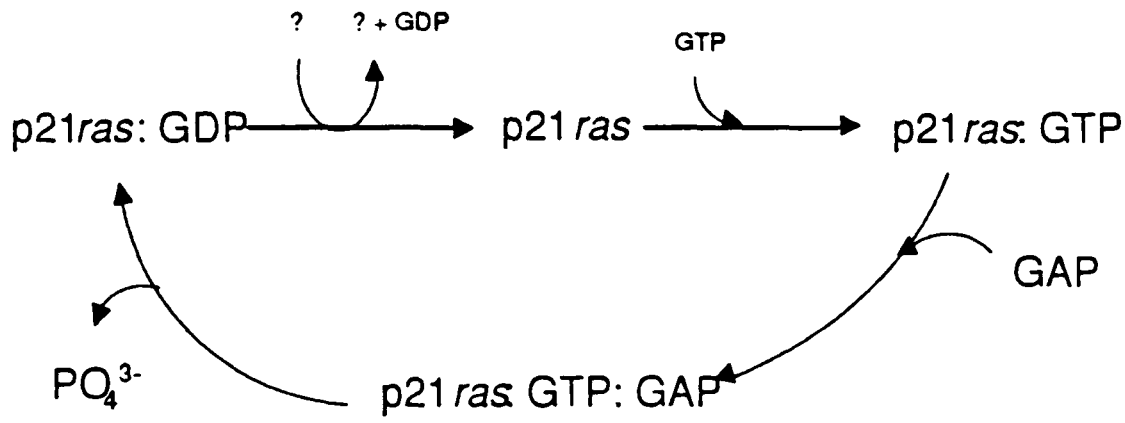
F2.1. The result of post translational processing to cysteine186 in *ras* proteins.

This long hydrocarbon chain is able to bury into the cell wall and fix the protein. In Ha- and N- *ras* Cys181 and Cys184* are also altered. These are reversibly palmitolysed (palmitic acid is $CH_3(CH_2)_{12}COOH$) and are then also able to bind the lipid membrane but allow some movement. K-*ras* does not have ~~this~~ these cystine residues and hence it is more mobile. What effect this has on its action is unknown. This is an area that simulations could help, given a model for a membrane and a very large amount of computer time.

II.3. The Nucleotide Cycle

Once held at the cell wall what happens then? During the travel through the cytosol, the protein has also bound a nucleotide, guanosine-5'-diphosphate (GDP), into its binding site. For some reason this molecule comes to be exchanged for guanosine-5'-triphosphate (GTP). During this exchange the protein has been changed in some way, it has been 'turned on'. It then acts on other species in the cell producing the effect desired by the original signal. The next stage in the cycle is the dephosphorylation of the GTP, returning the protein back to its inactive GDP bound form. The cycle can then start again (for reviews see McCormick 1989, Downward 1990 or Bourne *et al.* 1991).

The full cycle is shown below. Each half of the cycle - exchange and dephosphorylation will now be discussed separately.



F2.2. The *ras* protein nucleotide cycle.

II.3.1. Activation; Nucleotide Exchange

At the present time the cause of the nucleotide exchange is unclear. One value that has been measured is the dissociation constant for the equilibrium between the protein and the nucleotide. This is defined by the equation

* N-ras does not possess this residue.

$$p21{:}\ GXP \longrightarrow p21 + GXP \qquad K_d = \frac{[p21][GXP]}{[p21{:}\ GXP]}$$

where X can be either D or T depending on the nucleotide and p21 any of the *ras* proteins. Three sets of values for these equilibria are known.

T2. 2. The dissociation constants for the equilibrium between *ras* proteins and the bound nucleotides.

Protein	Nucleotide	Conditions	K_d / M
N-*ras*[c]	GDP	a	1.3×10^{8}
N-*ras*[c]	GTP	a	1.2×10^{8}
N-*ras*[d]	GDP	b	5.7×10^{10}
N-*ras*[d]	GTP	b	6.0×10^{10}
v-Ha-*ras*[e]	GDP	a	0.83×10^{8}
v-Ha-*ras*[e]	GTP	a	1.0×10^{8}

a equilibrium obtained with purified protein bound to GDP

b protein was initially isolated without any nucleotide present.

c Trahey *et al.* (1987)

d Feuerstein *et al.* (1987)

e Hattori *et al.* (1985).

These equilibrium constants show that the equilibrium is dominated by the bound complex. They also show that the binding of the two nucleotides is of similar strength. This means that an external impulse is needed to remove the GDP. Once the GDP is removed, given that the concentration of GTP in a cell (ca 10^{-3} M) is much higher than GDP, the GTP will end up in the binding site and the protein will be activated. The possible nature of this signal will be discussed later.

II.3.2. Deactivation; Dephosphorylation

Once bound to GTP it is known to then interact with another protein. It is called GTP-ase activating protein (GAP). The proteins themselves have an innate ability to remove the γ-phosphate. This process has a half life of the order of 40 minutes. In the presence of the GAP protein, this half life reduces to around 2 minutes (Trahey and McCormick 1987, Rey *et al.* 1989). However if the common oncogenic mutations have occurred, the half life is of the order of 10 hours and while the mutant protein is able to bind to GAP it makes no difference to this time.

Not much is known about this protein. The primary sequence has been deduced for the protein purified from bovine brain and human placenta (Trahey *et al.* 1988), but no secondary or tertiary structure has been defined. What it does after interacting with *ras* is

~~interacting with *ras* is~~ also not known. The only clue is by seeing if it is similar to any other known proteins. This matching, known as a homology study, aligns two protein sequences matching like residues with like to give a score. If two proteins are very homologous, it is likely that they have a similar structure and hence function. A database of protein sequences can be searched and a list of similar types of protein noted. When the NBRF database was matched against the GAP sequence, using the FASTP sequence alignment program (Lipman and Pearson 1985), the only noticeable similarities were with tyrosine kinases. The homology was however low, around 30%, and so nothing is conclusive. Again this is another area that simulations could be of use. If a model of the GAP protein was obtained, either by crystallography or that black art known as homology modelling, the two proteins could be docked and simulations run to see what is affected by the presence of the GAP.

The evidence also seems to point to GAP actually being the downstream target of *ras* (Calés *et al.* 1988) as well as accelerating the dephosphorylation. If it is indeed a tyrosine kinase, this would fit into the idea of the *ras* protein being involved directly in this pathway.

A mechanism has been postulated by Pai *et al.* to explain this dephosphorylation (Pai *et al.* 1990). The loop 61 -67 is very flexible. It appears that it is possible for Gln61 to move into the binding site. Here it is able, combined with Thr35, to hold a water molecule in the correct orientation for attack on the γ-phosphate. Thr58 is then involved in removing the phosphate from the binding site. This mechanism explains the long half life of the process. The flexible loop only occasionally is in the right orientation with respect to the Thr35 for the reaction to occur. The GAP may speed up the process by holding the threonine permanently in the correct orientation.

II.4. The Role of *Ras* Proteins in the Human Body

The properties described above enable *ras* proteins to play an important part in the growth of a cell. Exactly what they do is not known, however it is implicated in all parts of the cell growth cycle. This section will give an overview of this cycle and try to use known experimental data about *ras* to work out a possible role.

At a certain time the body sends signals to cells to reproduce. These signals are in the form of hormones. Hormones are usually unable to cross the cell membrane and so the first problem is how to get the signal inside the cell.

The hormone binds to a specific receptor bound across the membrane. This receptor is changed by the binding to stimulate secondary messengers inside the cell that carry the signal across to the nucleus. As soon as the signal is received by the cell the cycle begins. Ion exchange across the membrane occurs and the pH and calcium ion concentrations rise. The signal reaches the DNA in the cell nucleus. New proteins are

formed, some of which then interact with the DNA to produce more proteins. The timing of all these steps allow the cycle of cell growth, division and differentiation to occur.

Ras is known to be involved in the cell growth cycle both during the early and late parts of the cycle (Stacey *et al.* 1987). It is needed both to enable DNA replication (Mulcahy *et al.* 1985) and for the differentiation of the cell into its eventual type (Noda *et al.* 1985)

Each hormone has a specific receptor. There are however only three known signal pathways and the receptors can thus be divided into groups. One path goes via a G-protein to stimulate the production of cyclic adenosine-5'-monophosphate (c-AMP) by adenylate cyclase. A second involves a receptor that stimulates the breakdown of a small phospholipid, phosphatidylinositol (PIP_2) by phospolipase C (PLC). This breaks into two parts, inositol triphosphate travels into the cytosol and activates a cycle involving calcium. The remaining diacylglycerol stays in the cell wall and activates protein kinase C (PKC), a serine kinase. The last class of hormone activates a receptor with tyrosine kinase abilities. This appears to phosphorylate a certain protein, again a tyrosine kinase which in turn acts on another protein and to produce a cascade of phosphorylated proteins carrying the signal. These paths also interact in a co-operative manner. For example insulin, which works via a tyrosine kinase chain, is known to increase the speed of cell growth when administered to cells in addition to most other growth hormones. A sketch of the main signal pathways is shown in F2.3.

A huge amount of data about the *ras* protein and its involvement in all these pathways has been built up. One of the problems in trying to deduce a mechanism for its action is that many studies are done in different environments. Often non-mammalian cells are used and parallels drawn to human cells. This can be dangerous. For example studies on yeast cells indicate that *ras* is involved in the c-AMP cycle (Broek *et al.* 1985). However in human cells there is no evidence for this (Levitski *et al.* 1986).

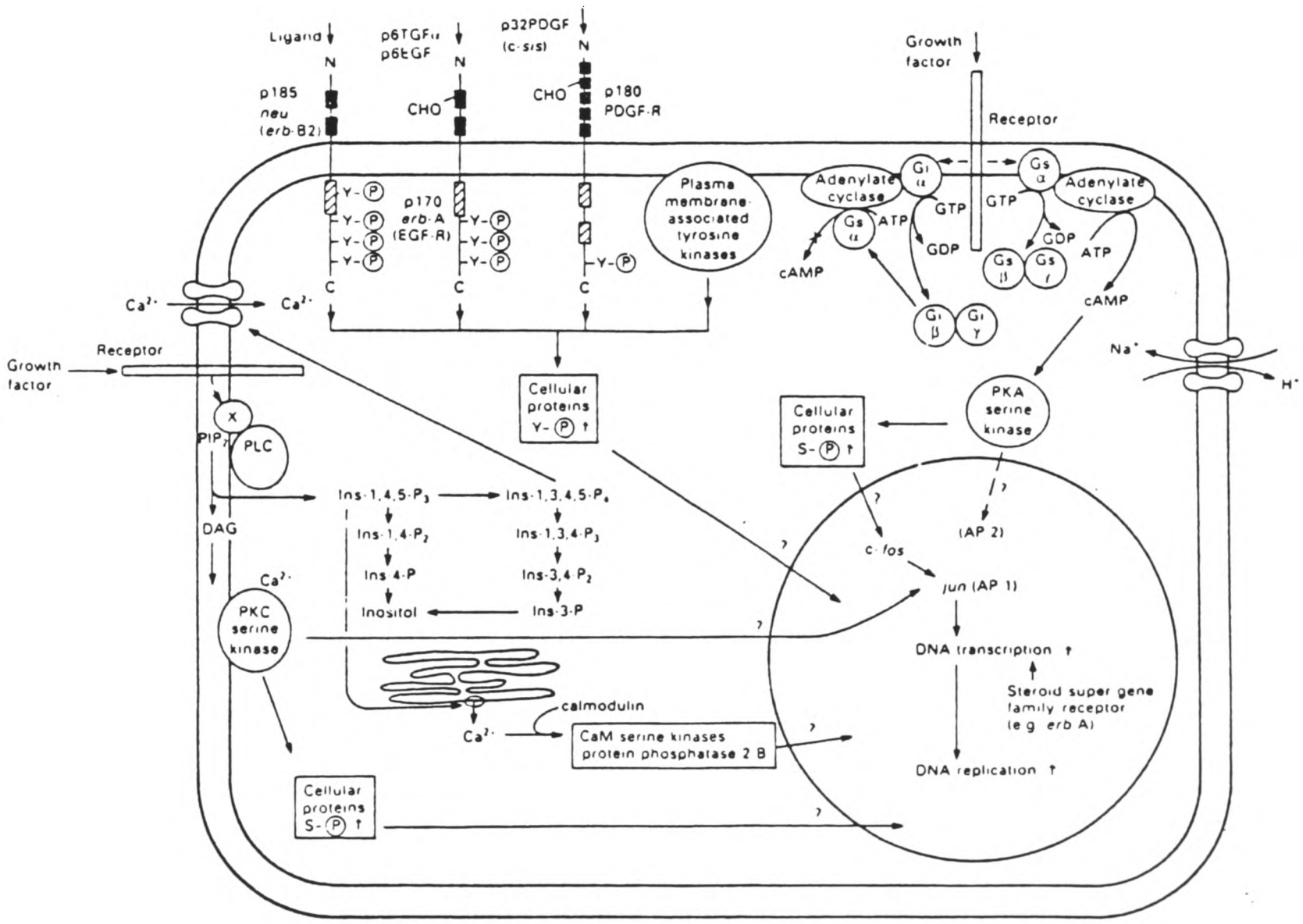
There is a link between *ras* and PKC (Lacal *et al.* 1987). One possibility is that *ras* is the link between the receptor and PLC. It increases the amount of DAG in the cell and stimulates IP_3 production. However it appears that the *ras* protein enhances the responsiveness to the growth factor without being directly involved. Also the DAG appears to come from *ras* activation of choline kinase which produces DAG by cleavage of Phosphatidylcholine (Macara *et al.* 1989).

Ras is also known to be involved in the tyrosine kinase cycle. It is essential for insulin maturation, but only accelerates progesterone (Korn *et al.* 1987). This suggests that it has two places in the cycle. Directly affected by the insulin receptor, but acting in a feedback loop for the progesterone. This feedback loop would make sense if the protein is activated by PKC and then activates choline kinase, which in turn produces DAG to activate more PKC etc.

What is also known is that both the insulin receptor (O, Brien *et al.* 1987, Kamata *et al.* 1987) and PKC (Jeng *et al.* 1987, Ballestr *et al.* 1987) phosphorylate *ras*. Phosphorylation seems to often be involved in growth factor cycles. It seems like a convenient way to regulate proteins; a large charged sphere added to a protein is bound to have some effect on its structure and hence function.

An uknown protein has been found to influence this exchange (West *et al.* 1990, Wolfman and Macara 1990, Downward *et al.* 1990) but its action is otherwise unknown. Other work seems to definitely support the idea of phosphorylation, involving *ras* with a tyrosine kinase (Bourne *et al.* 1990) and a protein cascade of MAP (mitogen activated protein) kinases (Woodgett 1992).

This is one idea that simulations might shed some light upon, but due to the speed of the process could be very difficult to monitor experimentally. In the next chapter this will be attempted, adding phosphate groups to suitable sites and recording any affects this makes to the protein.



F2.3. The major signal pathways involved in the cell growth cycle.

Reproduced from Glover and Hames (1989) "Oncogenes" ch IV p115.

II.4.1. Oncogenic Activity of the *Ras* Proteins.

Under certain stimuli, the DNA responsible for the amino acid sequence is damaged. This can happen due to chemicals, viral processes or energy such as UV or X-ray radiation. Either sections of the DNA are deleted or inactivated, in which case sections of the protein will be missing, or else bases can be changed, causing a different sequence to be produced.

In *ras* proteins it is the latter processes that are important. A single residue change at certain positions can radically alter the protein function. For example, if the residues involved in the dephosphorylation step are changed this step may be inhibited. The protein will remain in its active form for longer and so the signal time increases. The common point mutations that ~~occur~~ are found in nature are at positions 12, 13, 59 or 61. When these mutations occur, the protein has properties that cause cell transformation into a cancer cell. These proteins are henceforward denoted by the name "transforming". (for review see Barbacid 1987)

If Gly12 is changed for any other residue, with the exception of proline, the biochemical change that has occurred is that the GTP-ase activity has been lowered. This residue lies on the loop that runs alongside the phosphates. The mutation presumably forces a conformational change that affects the phosphate binding. The same is true of Gly13, although this mutation is less common.

The other region of the transforming protein binds the γ-phosphate in the GTP complex. Mutations at Gln61 are the second most common types. Any other residues, with the exception of proline and glutamic acid activate the protein. Threonine substitution for Ala59 has the same result.

There also may be some mechanism other than simple preventing of hydrolysis. Trahey *et al.* (1987) discovered that a Val12 mutant had 12% of the wild type GTP-ase activity and an Asp12 mutant 43%. Yet both had identical transforming ability.

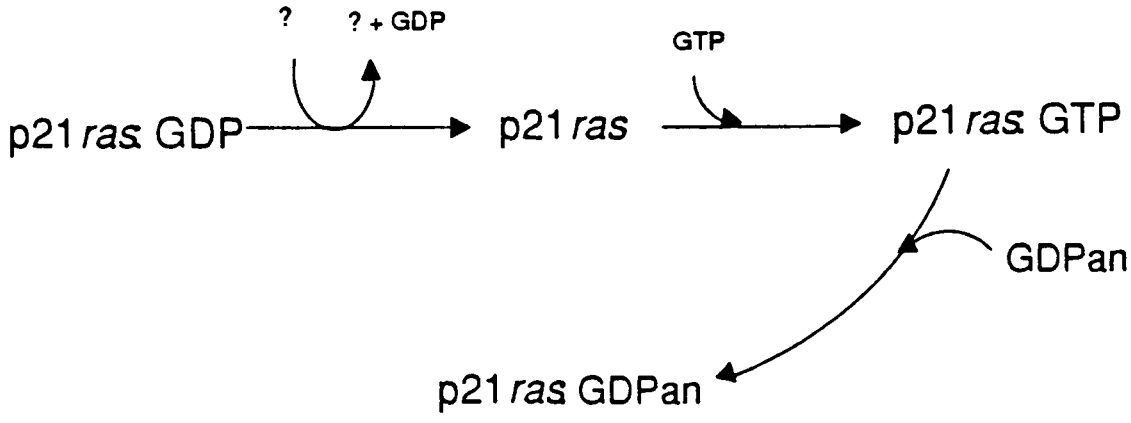
II.5. Inhibiting the Transforming Protein by Simulation

The information presented above leads to the idea that it may be possible to inhibit the transforming protein. In essence what must occur is that the GTP bound to the mutant protein must be displaced by a drug capable of changing the protein back to its inactive form, effectively the GDP bound complex. The obvious choice for such a molecule is an analogue of GDP that can bind irreversibly, preferably just to the transforming and not the normal protein.

In order to be able to make predictions about such a compound, it is necessary to know the binding free energy of the various nucleotides to the different forms of the protein. From this the equilibrium data for all the combinations are known and a molecule

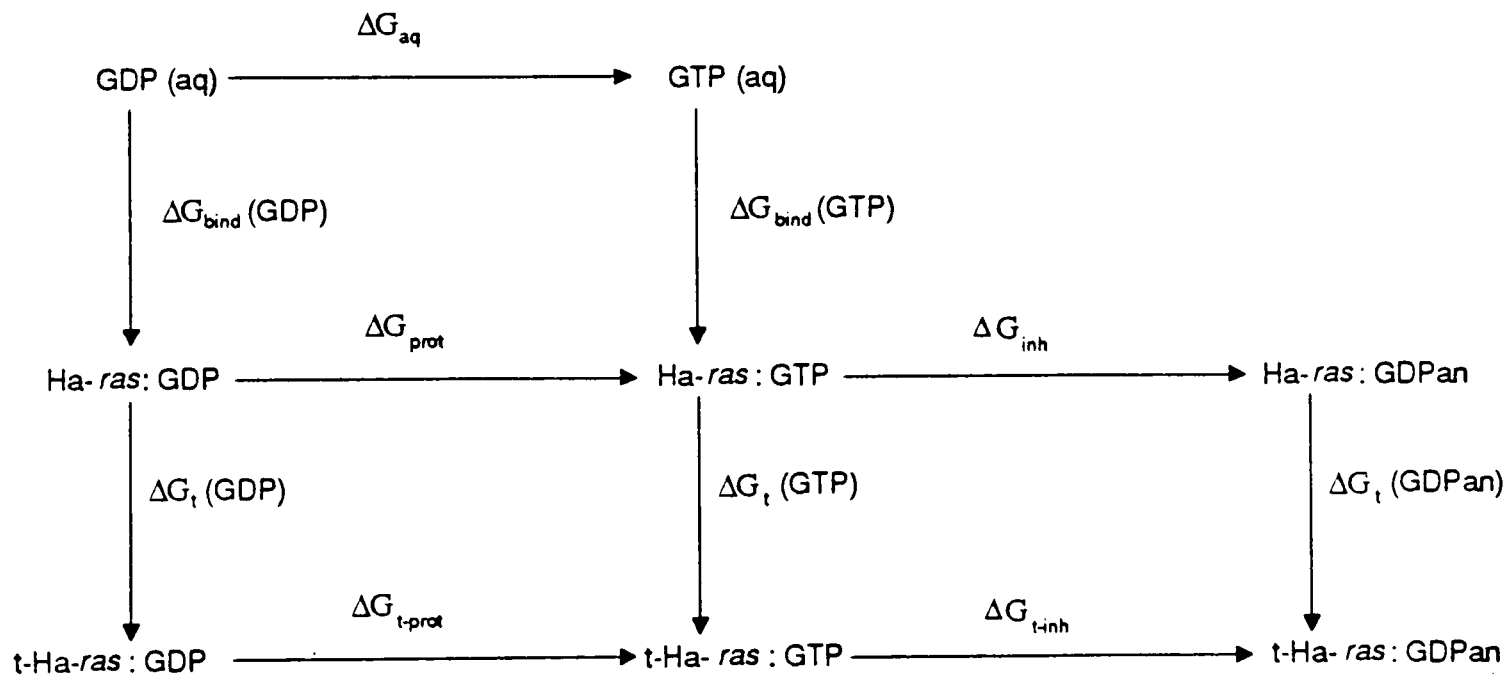
with the desired properties selected. One way this can be achieved is by the use of the FEP equation derived in chapter I.5.2.1.

For example the *ras*: GDP complex could be taken. During the course of a simulation the GDP is mutated to GTP and the energy difference produced by the change monitored. From the ensemble average of this perturbation Hamiltonian, the free energy difference between the binding of the two nucleotides is known. The GTP could then be mutated into a GDP analogue and this free energy difference calculated, and so on. This could then be repeated after mutating the protein to a transforming type.



F2.4. Inhibiting the transforming *ras* protein by the use of a GDP analogue, GDPan.

A thermodynamic cycle can be constructed, connecting all these structures



F2.5. A thermodynamic cycle to develop a transforming *ras* inhibitor.

The advantage of a cycle such as the one above is that there are comparisons with experimental results that can be made to check the model. For example the solution phase equilibrium of GTP and GDP is related to the Gibbs free energy of hydrolysis of GTP. The difference of binding energies for the nucleotides in the normal and transforming proteins can also be

compared to the binding energy data given in T2.2. For an analogue to be a good drug, it is simply necessary that $\Delta G_{t\text{-}inh}$ is large and negative while ΔG_{inh} is positive.

The only necessary requirement to begin calculations on this cycle is that information about at least one of the protein structures is known. This enables the simulations to have a good starting point from which mutations can be made. There are various ways a protein structure can be obtained which will be mentioned in the next chapter. However one of the inspirations for this project was the availability of a crystal structure. This is the best start for a simulation: an accurate three dimensional structure in terms of its co-ordinates.

The Ha-*ras*: GDP complex crystal structure has been deposited in the Brookhaven data bank. Other crystal structures have also been determined, but only qualitative descriptions were initially available. In the next section the structural data will be discussed. As well as the actual structure, a lot of information is known about the importance of different regions of the protein. This is important in order to judge whether a model, even based on a crystal structure, is reasonable; it must have the same properties as the real thing. It is also important to know the differences between possible endpoints of mutations in the cycle in F2.4 to check how well the simulations have worked.

II.6. The Protein Structure

From mutation studies and the use of antibodies, it is known which residues are important for the interaction with various other species. For example, the monoclonal antibody Y13-259 is known to inhibit the activation of *ras*. The region it binds is therefore likely to be the effector binding region (Sigal *et al.* 1986), or alternatively this could be due to the conformational change seen to take place in this region being prevented. Mutation studies have also identified the GAP binding region (Adari *et al.* 1988). The residues known to have important function are shown in T2.3.

Various crystal structures have also been produced, both of normal and oncogenic proteins. The first Ha-*ras*: GDP structure (deVos *et al.* 1988) gained some notoriety for its incorrect joining of two secondary structure regions. This demonstrates that crystal structures are not by themselves exact, they take refinement and guesswork to resolve the electron density and sometimes this can be incorrect (Brändén and Jones 1990).

The GTP complex has also been deduced. Due to the dephosphorylation of GTP, it is impossible to crystallise this complex directly. However crystals of the slowly hydrolysing complex bound to GPPNP, in which the oxygen between the β- and γ-phosphates is replaced by an -NH group, have been analysed at 1.35Å resolution (Pai *et al.* (1990). It was in fact after an initial 2.6Å resolution structure of this complex (Pai *et al.* 1989) that the revised GDP complex structure was released (Tong *et al.* 1989). The traces

of these crystal structures are shown below in figures F2.6 and F2.7. These show the nomenclature for the secondary structure elements.

T2.3. The Important Regions of the HA-*ras* protein and their targets.

residues	species bound
32 - 40	GAP
60 -76	Y13-259 monoclonal antibody
28, 116 - 119, 145 - 147	guanine
29, 117	ribose
18	α-phosphate
10 -17	β-phosphate
35, 59 -61[a]	γ-phosphate
17, 35[b]	Magnesium
58[c]	auto-phosphorylation

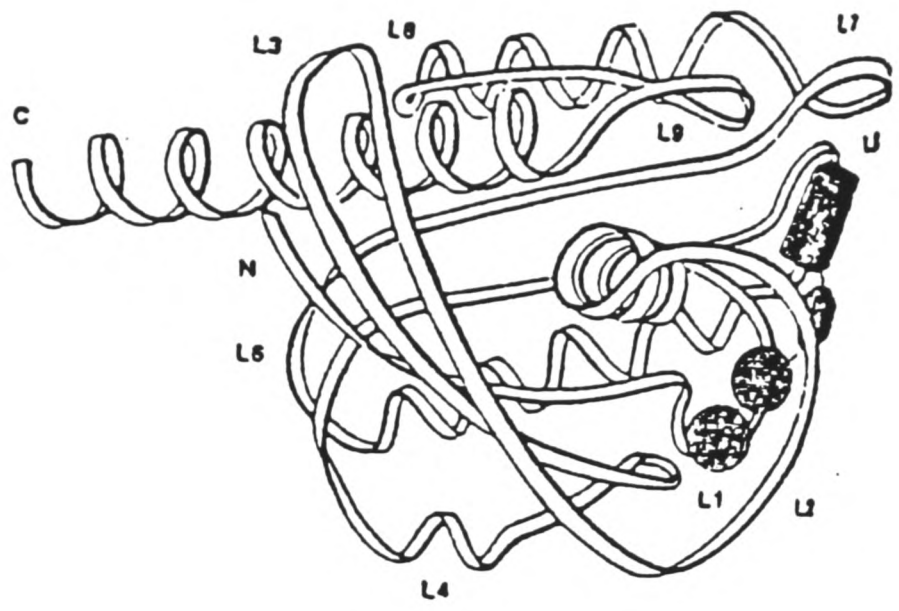
a Only applicable to GTP bound complexes

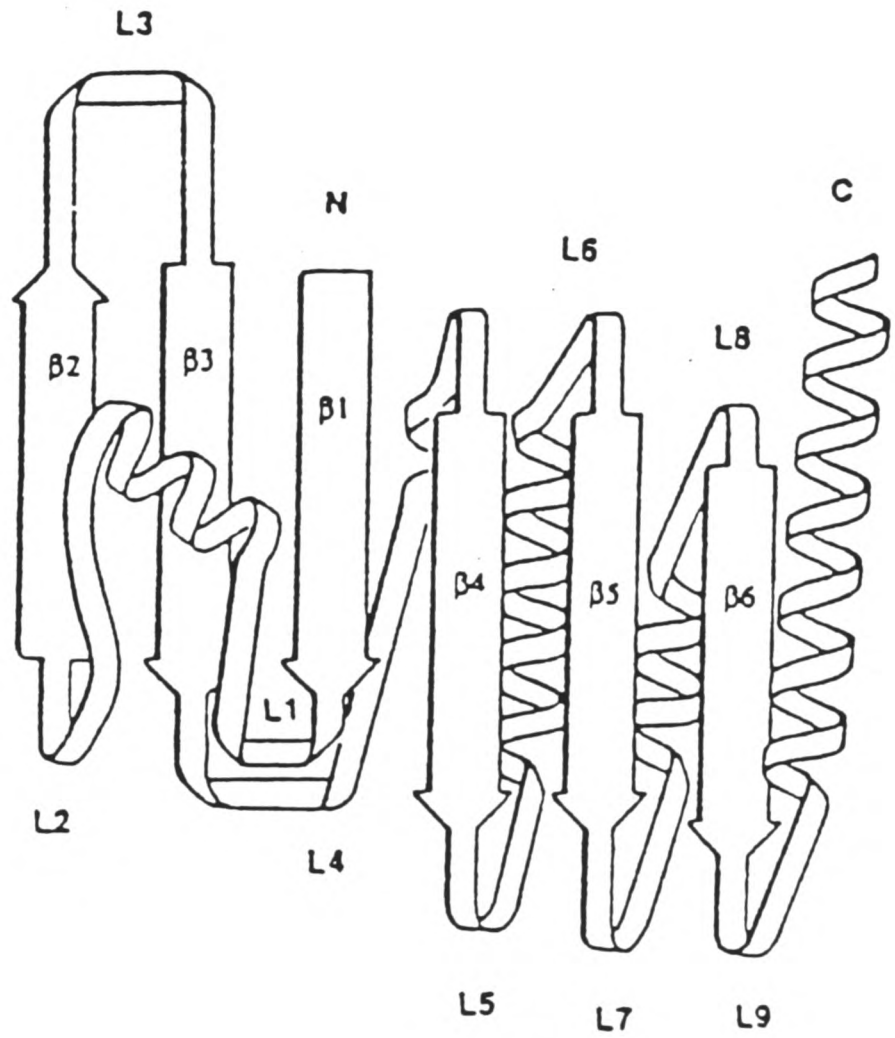
b Thr35 is not a ligand in the GDP complex.

c This residue is not bound to anything specific, but is necessary for the removal of the γ-phosphate.

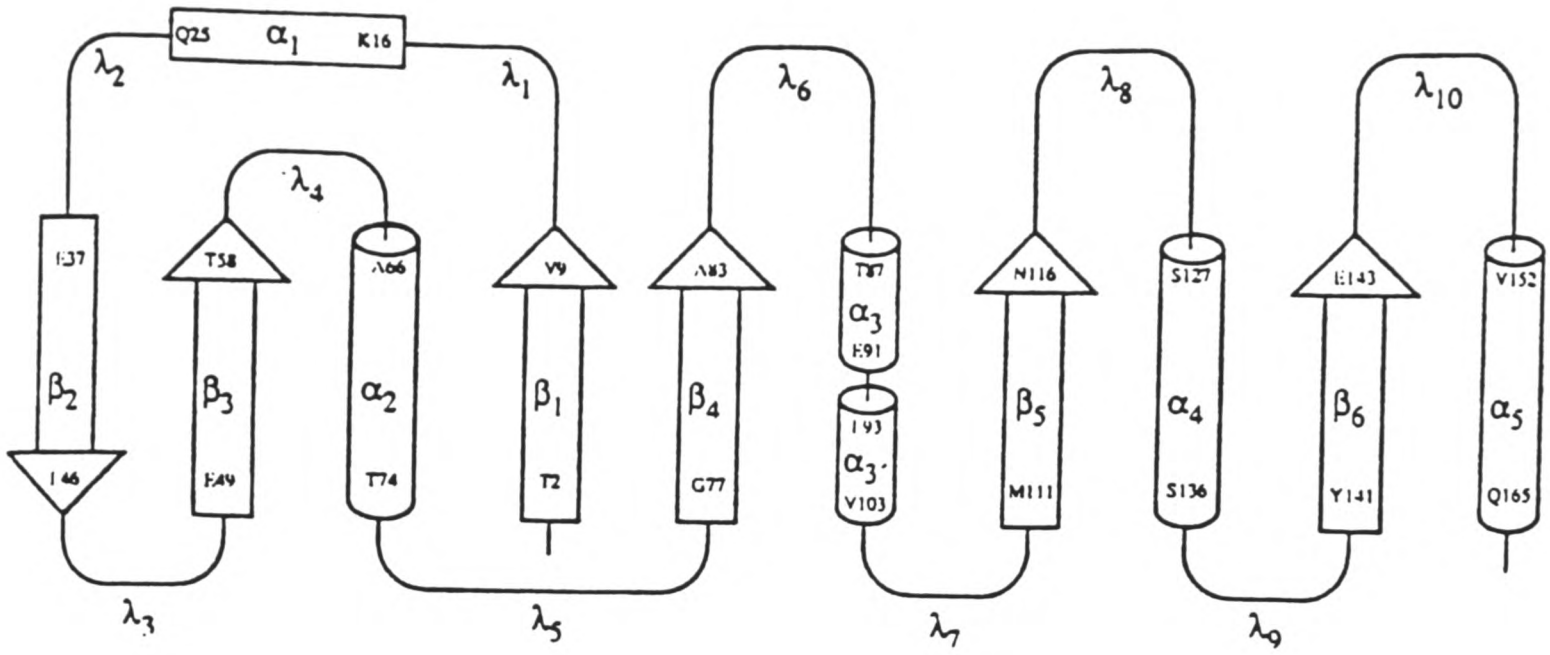
The GTP complex has also been studied by the use of caged GTP, which after photolysis undergoes hydrolysis (Sclichtling *et al.* 1990). By this technique time resolved X-ray diffraction has been used to follow the dephosphorylation. Unfortunately this technique does not resolve the structure to a degree accurate enough for a definite mechanism to be provided. However there appears to be essentially no difference between the GPPNP and GTP bound complexes.

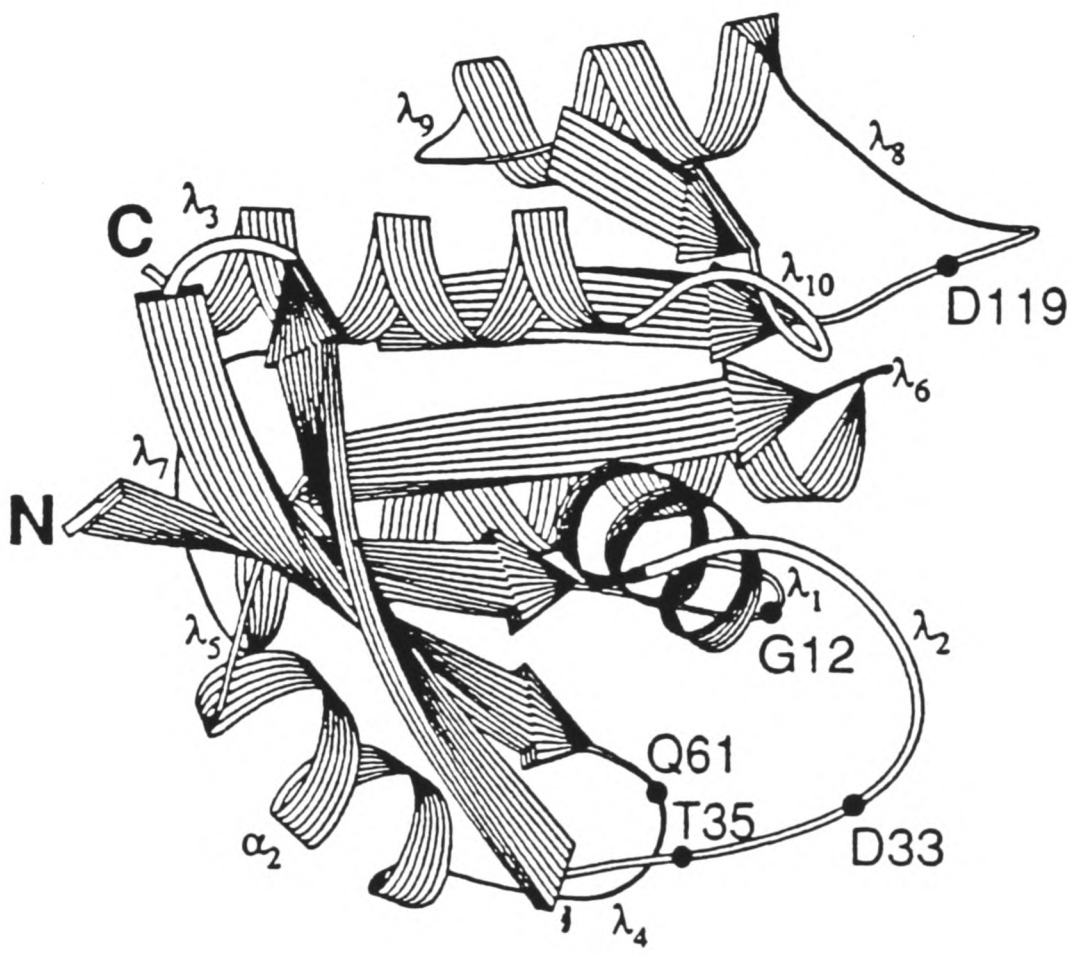
The GDP and GTP complexes are very similar in overall topology. In only one area is a change in secondary structure seen. In Ha-*Ras*: GDP complex there is a loop, λ_4, including a single helical turn between residues 59 and 76 while in the active form this loop has a three turn alpha helix from residues 66 to 74. A large conformational change is also seen between residues 32 and 40, the GAP binding region, with this loop being more extended in the GTP complex. Other differences are that Thr35 is a magnesium ligand in the GTP complex, but is separated by a water molecule in the GDP complex. The flexible loop, λ_4 is also more extended from the protein in the GDP complex. As mentioned above, the other *ras* proteins, K-*Ras* and N-*Ras*, show very high homology in these regions and around the nucleotide binding site so are likely to undergo a similar change.





F2.6. The secondary structure trace and a diagram showing the three dimensional arrangement from the Ha-*ras*: GDP crystal. Reproduced from Tong *et al.* 1989.





F2.7. The secondary structure trace and a diagram showing the three dimensional arrangement from the Ha-*ras*: GTP crystal. Reproduced from Pai *et al.* (1989).

The secondary structure elements are shown in plates 1 - 4. In all the plates the important residues 12 and 61 and the effector loop of residues 32 to 40 are marked to help viewing. The first two plates show the secondary structure from the Ha-*ras*: GDP and then the Ha-*ras*: GTP crystal with the binding site to the right of the protein. It is noticeable how similar the two structures are. The major difference that can be seen from this angle is the extra turn on the helix α_2 in the GTP complex. This noticeably tightens up the loop λ_4 and Gln61 is now into the binding site. The movement of the effector loop away from the protein, possibly due to the presence of this Gln61, can also be seen. Plates 3 and 4 show the same structures, only now looking down the binding pocket. From this angle the difference in the region around residue 61 is even more marked. A review of these differences is contained in Milburn *et al.* (1990) and Jurnak *et al.* (1990).

In all four plates it can be clearly seen how Gly12 is at the tightest point of the loop lying alongside the nucleotide. This would explain the importance of this site, a larger residue would prevent this tight turn from occurring. This was indeed found to be the case when the crystal structure of a Val12 mutant GDP bound protein was determined (Tong *et al.* 1989a).

The exposed nature of the nucleotide is also clear, without the sidechains it looks as if it is not held in the protein at all. Plate 5 shows exactly how it is bound in the GDP complex. Even here it is clear that the hydrogen bonding is sparse and there is a lot of room for water. While residues 57 to 61 do not directly interact with the nucleotide, they can be seen to be important in stabilising the loop λ_1.

Plate 1 (following page). The Ha-*ras*: GDP complex with the nucleotide binding site to the right of the picture. The GDP and magnesium ion are drawn in grey ball and stick plots, while the protein is drawn in cartoon style to emphasise the secondary structure.

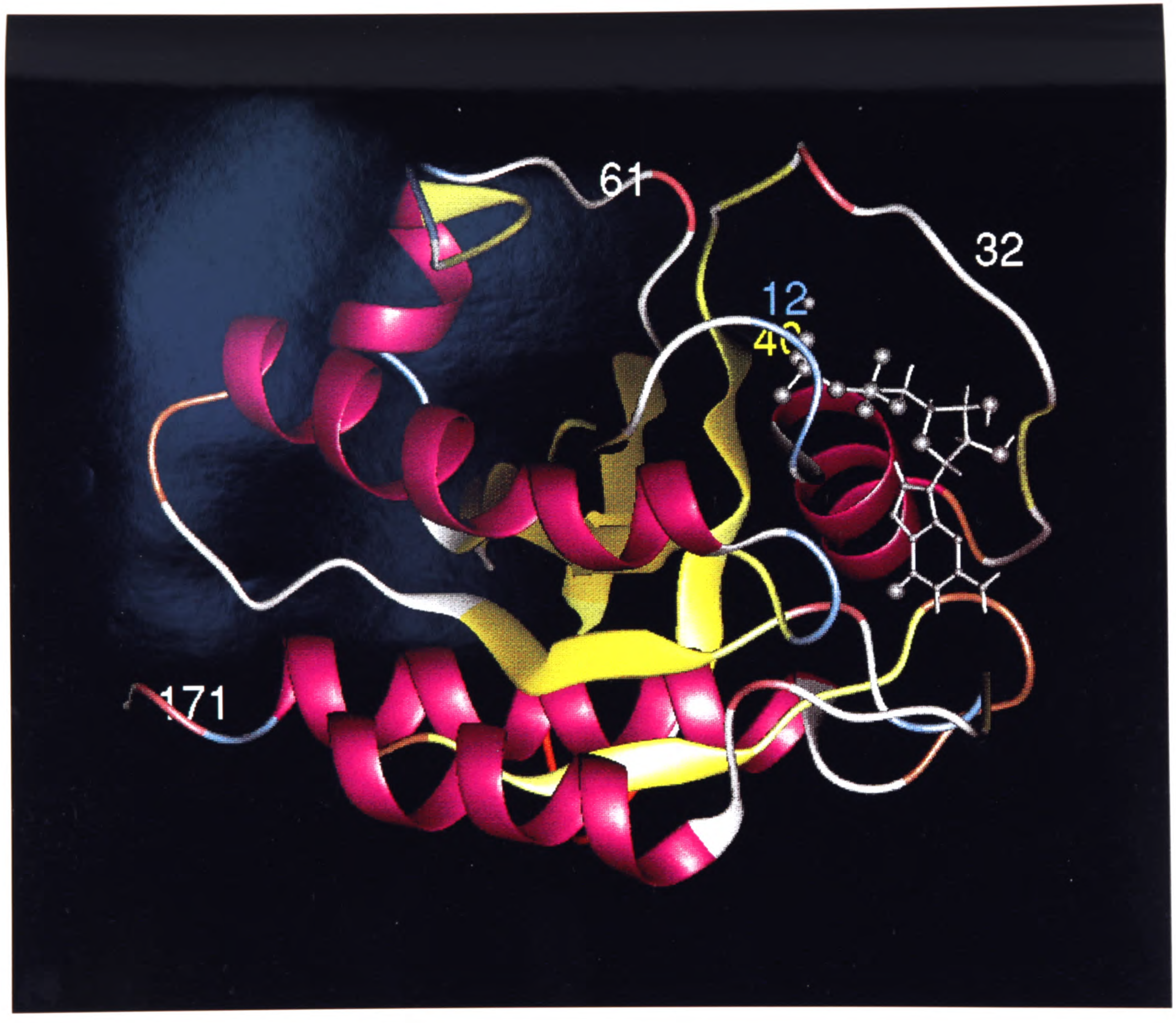

61
32
12
171

Plate 2 (following page). The Ha-*ras*: GTP complex with the nucleotide binding site to the right of the picture. The GTP and magnesium ion are drawn in grey ball and stick plots, while the protein is drawn in cartoon style to emphasise the secondary structure.

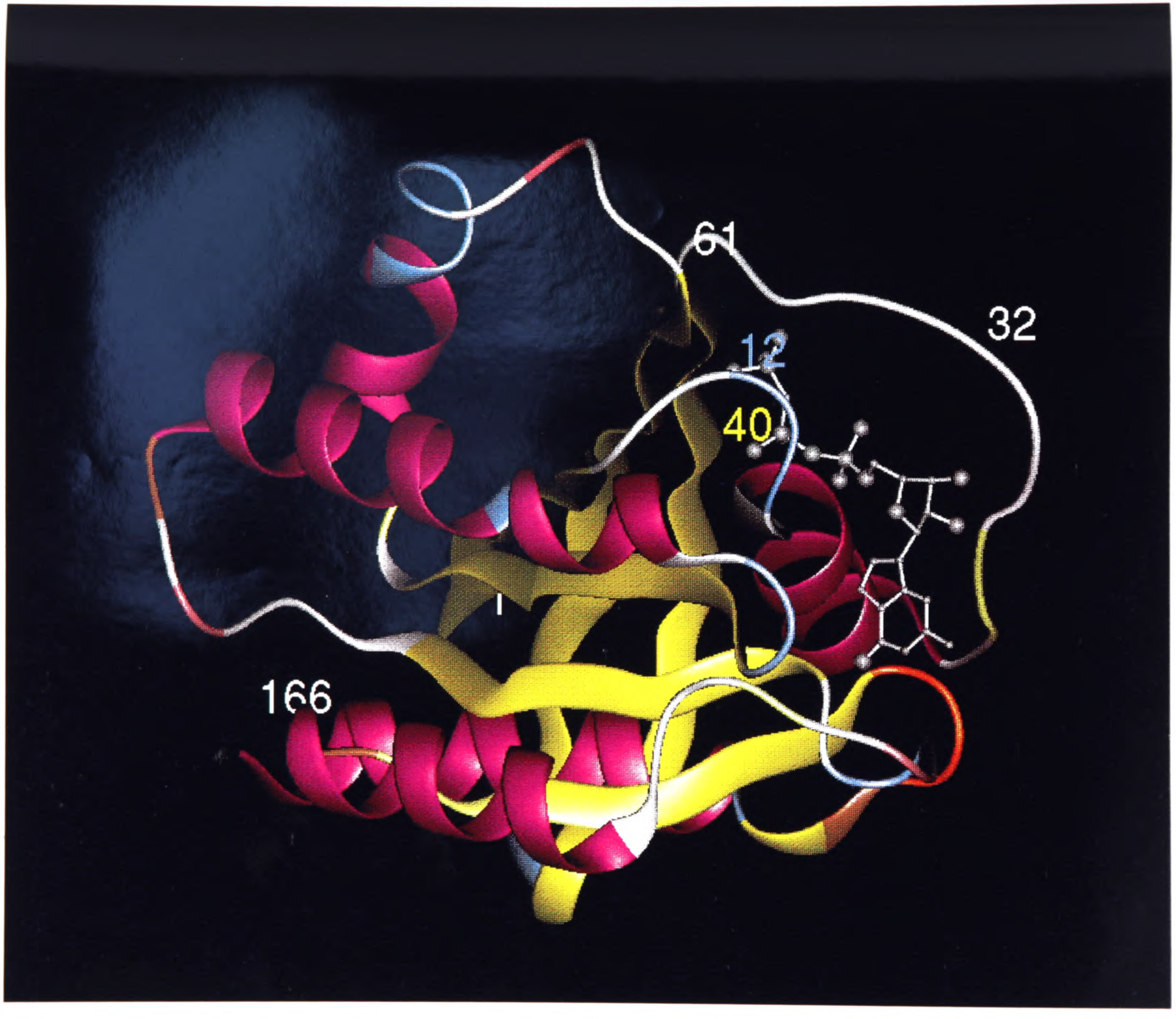

61
32
12
40
1
166

Plate 3 (following page). The Ha-*ras*: GDP complex looking down the nucleotide binding site. The GDP and magnesium ion are drawn in grey ball and stick plots, while the protein is drawn in cartoon style to emphasise the secondary structure.

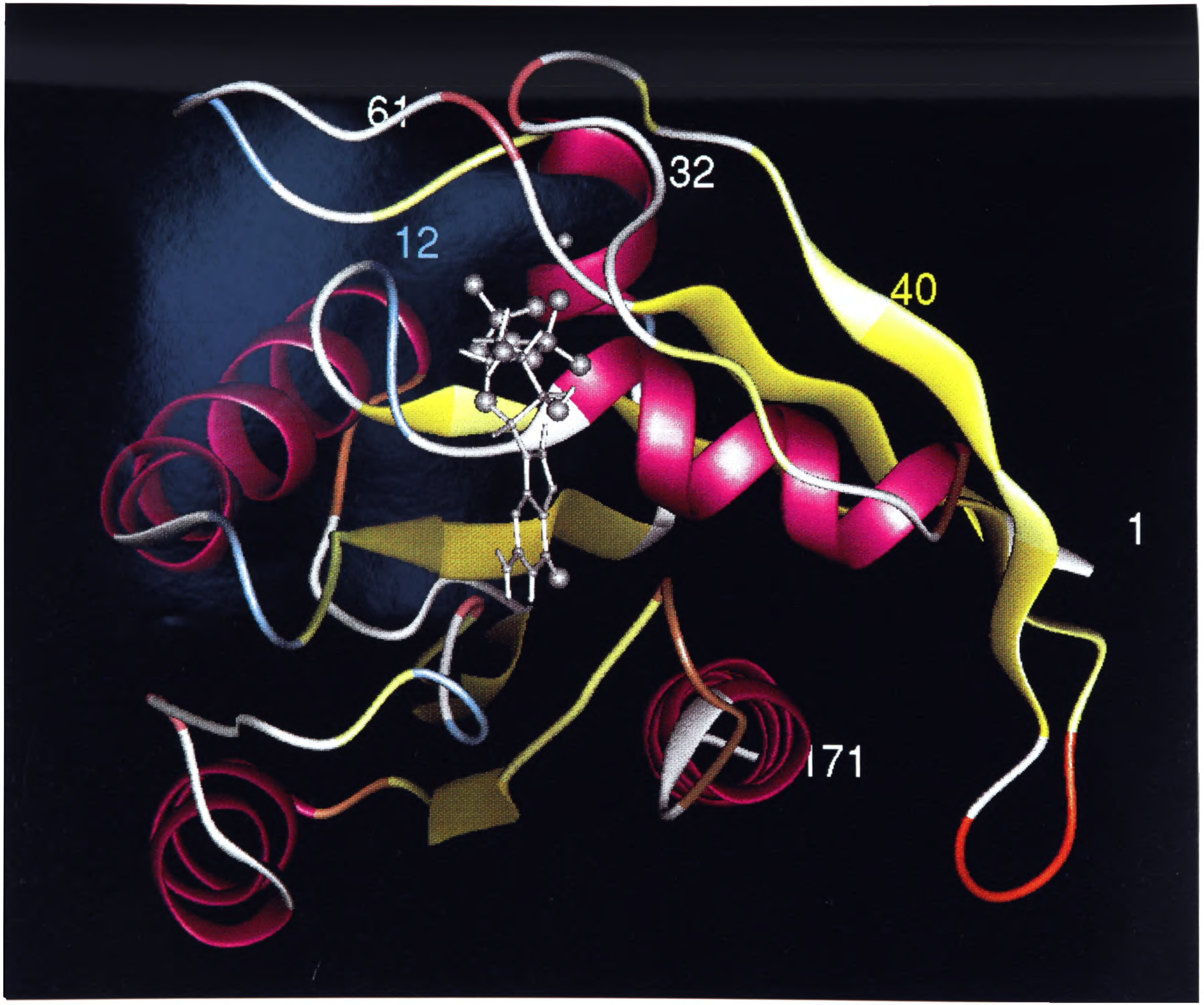
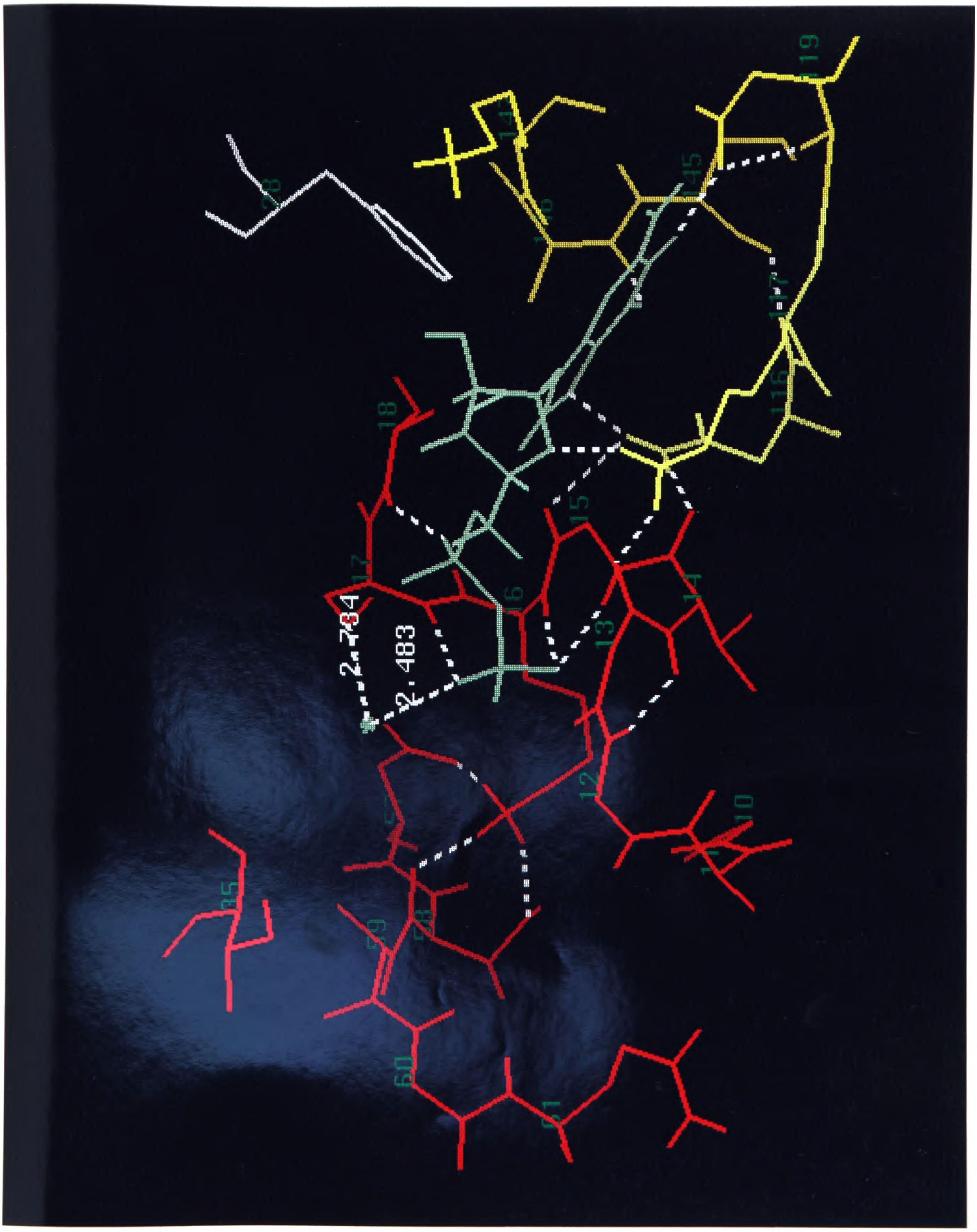

61
32
12
40
1
171

Plate 4 (following page). The Ha-*ras*: GTP complex looking down the nucleotide binding site. The GTP and magnesium ion are drawn in grey ball and stick plots, while the protein is drawn in cartoon style to emphasise the secondary structure.


61
32
12
40
1

Plate 5 (following page). The interaction of the GDP nucleotide with the Ha-*ras* protein binding site. Residues that interact with the phosphates are in red, those that interact with the nucleotide in yellow and with the ribose in white. The magnesium ion and GDP are in green.


2.483

II.7. Summary of Chapter II.

In this chapter an overview of the properties and importance of the *ras* proteins was presented. Based on the knowledge to date, a number of areas where computer simulations could help solve the outstanding problems in the *ras* cycle were suggested.

The idea of using simulations to develop a drug to inhibit the mutated protein, an active constituent of cancer cells, was also developed and a scheme produced to enable this to be possible. The necessary knowledge of the protein structure and behaviour was then discussed.

II.8. Bibliography for Chapter II

Glover, D.M. and Hames, B.D. "Oncogenes" (1989) IRL Press (OUP)

Good review on oncogenes and signal mechanisms in general. Chapter 4 has a fairly comprehensive section on *ras* biochemistry known to date.

CHAPTER III

The Activation of the *Ras* Protein

At the start of the project, the only available *ras* protein structure was for Ha-*ras*: GDP. In order to study the difference in nucleotide binding of both forms of the protein it was necessary to have the GTP bound complex also. X-ray structures of both complexes have actually been determined, Ha-*ras*: GDP to a low resolution and Ha-*ras*: GTP to high resolution as the slow hydrolysing GPPNP complex and at ~~too~~ low resolution as a pure GTP complex using time resolved methods. Both "GTP" bound structures have not been released, but qualitative descriptions of all the structures are available (see chapter II.6 for references).

The Ha-*ras*: GDP structure, which is deposited in the Brookhaven databank (code 2p21) (deVos *et al.* 1988), only gives co-ordinates for the backbone alpha carbon atoms and the GDP heavy atoms. There is also a gap between residues 60 and 66 where the loop is very flexible and the electron density was too low to be certain of atom positions at this resolution. The first problem was to build a full protein structure based on these co-ordinates. Using the known differences between the two forms, it would then be possible that the GDP complex could be transformed into the GTP complex.

In the last chapter (II.4) it was suggested that a possible mechanism by which the nucleotide exchange is induced in the cell is by phosphorylation of certain residues. This idea will also be looked at by running simulations, first in the gas phase and then aqueous solution, to see the effect of the introduction of phosphate groups on GDP binding.

III.1. Creating a Full Ha-*Ras*:GDP complex Based On An Incomplete Crystal Structure

Working out the three dimensional structure of a protein is a very difficult problem. The common way is by X-ray crystallography. Firstly the protein must be crystallised, an enormous problem in itself. Then the complex diffraction pattern caused by such an irregular system must be analysed. The problem then remains as to how similar this structure is to that seen in solution. NMR techniques can give an insight into this, but are only able to produce data on how far apart atoms are on average, rather than the full co-ordinate structure obtained by crystallography. Computer methods to generate an unknown three dimensional structure by its similarity, or homology, to a protein with a known structure are also becoming popular. However at the present time they are only able to produce a structure if the protein with known structure is nearly identical to the protein being modelled.

The Ha-*ras* protein is the only one of the family to be crystallised to date. While it has been possible to crystallise the full protein (Jancarik *et al.* 1988), it has proved easier to truncate the flexible C-terminus. In the Brookhaven GDP complex, residues 172 to 189 were deleted. Fortunately this does not seem to have any effect on the biochemistry of the system. It is still able to undergo nucleotide exchange and dephosphorylation. It is also still able to bind the GAP protein and function as normal in the cell. Presumably however as it is no longer able to bind to a cell wall the incoming signal would not be picked up, this is not important for the study of its behaviour. This is also true of the GTP complex crystals which were made using only the first 166 residues (Pai *et al.* 1989).

As well as the C_α co-ordinates, the primary sequence and the backbone tracing are known. The first task is to fit the amino acids in their correct sequence onto the known atom positions. All the early parts of the modelling - the formation of the protein model from the crystal - were done using the QUANTA graphics package incorporating CHARMm. The interface between these two programs allow simple molecular mechanics calculations to be easily viewed and monitored. Later energy and dynamics calculations on the model were made using AMBER due to its greater flexibility and better parameters.

While this fitting may sound simple, in practice it involved several stages to produce a structure. Many small adjustments were then needed to give a model that fitted the crystal co-ordinates under the influence of only the molecular mechanics force field. The major steps and problems are described below.

A 171 residue polyalanine chain was set up to become the backbone (the crystal structure has 171 residues). A polyalanine chain was used for reasons of simplicity. Larger sidechains could cause packing problems during the formation of the backbone and the alanine methyl groups can be readily mutated into the relevant sidechains after folding. Glycines were however initially placed in the chain in their relevant positions. At this point the GDP was completely ignored. It could be added later when the binding site was complete.

This chain was then energy minimised using CHARMm with a constraining potential added to fix the C_α carbons of the polyalanine/glycine chain to their relevant protein positions. The constraining potential was harmonic in form with a force constant of 50 and initially an exponent of one followed by an exponent of two as the simulation neared its completion to pull the alpha carbons harder onto their targets. The dihedrals of the chain were kept linear and the non-bond pair list was never updated. This means that because the chain is initially linear, any atom only interacts with its neighbours in the chain, so while folding different areas of the chain can pass through each other.

The sidechains were then changed so that the correct primary sequence was present. The constrained minimisation was then repeated. The whole model was then allowed to optimise freely. The RMS deviation from the crystal structure was $\cong$1.3Å. GDP

was added to the binding site and the model put into AMBER. The AMBER standard united atom parameters (Weiner *et al.* 1984) were taken for the amino acids. All atom parameters were used for the nucleoside (Weiner *et al.* 1986). The mixing of parameter sets is frowned upon by some people, but polar hydrogens are included in the AMBER united atom set and so the important GDP - protein interactions are included. It was also thought important to include the non-polar hydrogens on the ribose ring to correctly model the spatial extension of this region which might be important in the motion of the guanosine and phosphate chain.

The only part of the AMBER force field which it is normal to alter are the atom centred charges needed to model the electrostatic interactions. These are tailored to suit a small molecule by fitting charges to the molecular electrostatic potential (MEP) calculated from an *ab initio* or semi-empirical wavefunction. For the GDP the charges were fitted to an AM1 MEP, using the method of Chirlian and Francl (1987) adapted to accept semi-empirical potentials (Ferenczy *et al.* 1989).

III.1.1. Minimising the model with AMBER

The BORN module is used to optimise the geometry of a system with respect to the energy. It contains two different methods for achieving this. Steepest descent calculates the energy and its gradient then simply follows the way downhill to the minimum. The conjugate gradient method also uses the size of the gradient to enable it to adjust the step size better as the minimum is approached. The former method is more robust and so it is normal to run a few steps of steepest descent to remove any bad clashes in the starting structure. The conjugate gradient minimiser is then ~~turned~~ used to converge quickly onto the optimum energy.

On using this program on the Ha-*ras* model produced above, it was found that the final energy was dependent on the method used. Single precision minimisation (when the computer stores the number using only 16 bits) was unable to converge. This is due to the flat potential energy surface for such a system which can sometimes prevent the minimiser knowing which way to go. The double precision minimiser (using 32 bit numbers) did converge. Depending on the number of initial steepest descent steps, different energies were obtained. These are shown below in T3.1

This demonstrates that there are many different minima close together, hence a slight difference in starting conditions leads to a different end result. The crystal structure is only in one of these minima. Fortunately due to the rigid nature of a protein secondary structure these different minima contain only minor structural differences and during dynamics these regions tend to be flexible. The structure with 20 steps of steepest descent was taken due to its quick convergence. The RMS deviation from the crystal structure was now $\equiv$ 1.15Å.

T3.1. Differences in the minimum energy of Ha-*Ras*: GDP produced by BORN with different methods

Number of SD steps	Number of steps before convergence	Energy (kCal mol^{-1})*
0	7412	-9.195 x 10^3
20	5910	-9.204 x 10^3
100	> 10000	-9.298 x 10^3
1000	7214	-9.069 x 10^3

III.2. Transforming The GDP Complex Into the GTP Complex.

For the calculation outlined in chapter II.5 to be made it was necessary to have a structure for the GTP complex as well. The differences between the two structures are known, the problem was how to induce this change to occur.

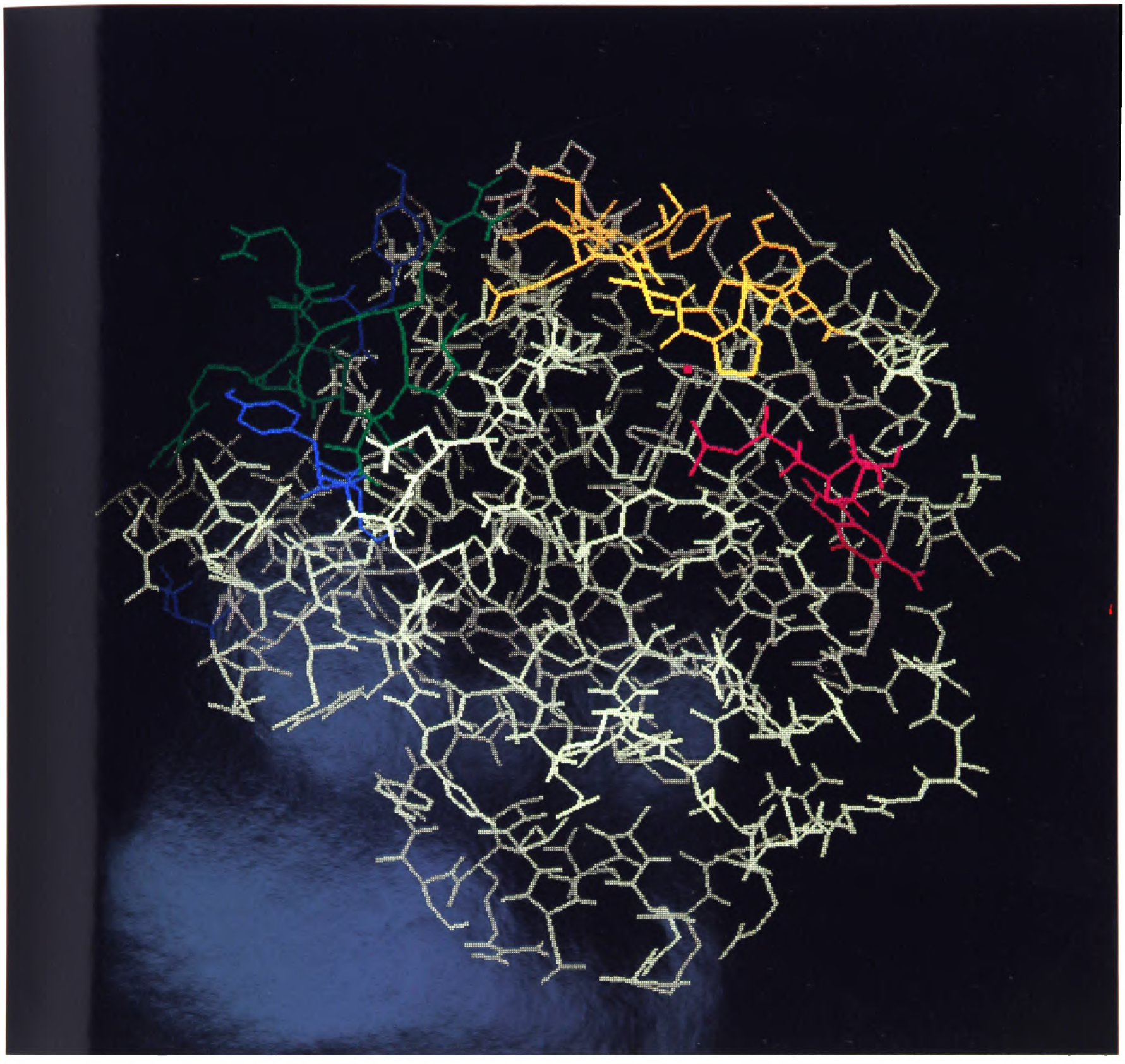
One possibility mentioned previously is that this change occurs due to the presence of phosphate groups. Even if the phosphorylation does not cause the change in secondary structure, it is possible that the nucleotide exchange is helped by the phosphate interaction with the nucleotide in the binding site. It is known from antibody studies that the probable interaction site of p21-*ras* with its upstream regulator is in the region 60 - 76 and so potential phosphorylation sites must be accessible to this area. There are three serines on this side of the protein, but Ser17 is involved in the binding of the magnesium, so this leaves Ser65 and Ser106 as the probable phosphorylation sites. Similarly, while there are many potential tyrosine phosphorylation sites, Tyr64 and Tyr71 are the nearest and most accessible to attack in this region. Plate 6 shows how close these residues are to the regulator loop. It also shows that all the regions important to the function of the protein are on the same side of the protein, that exposed to the cytosol.

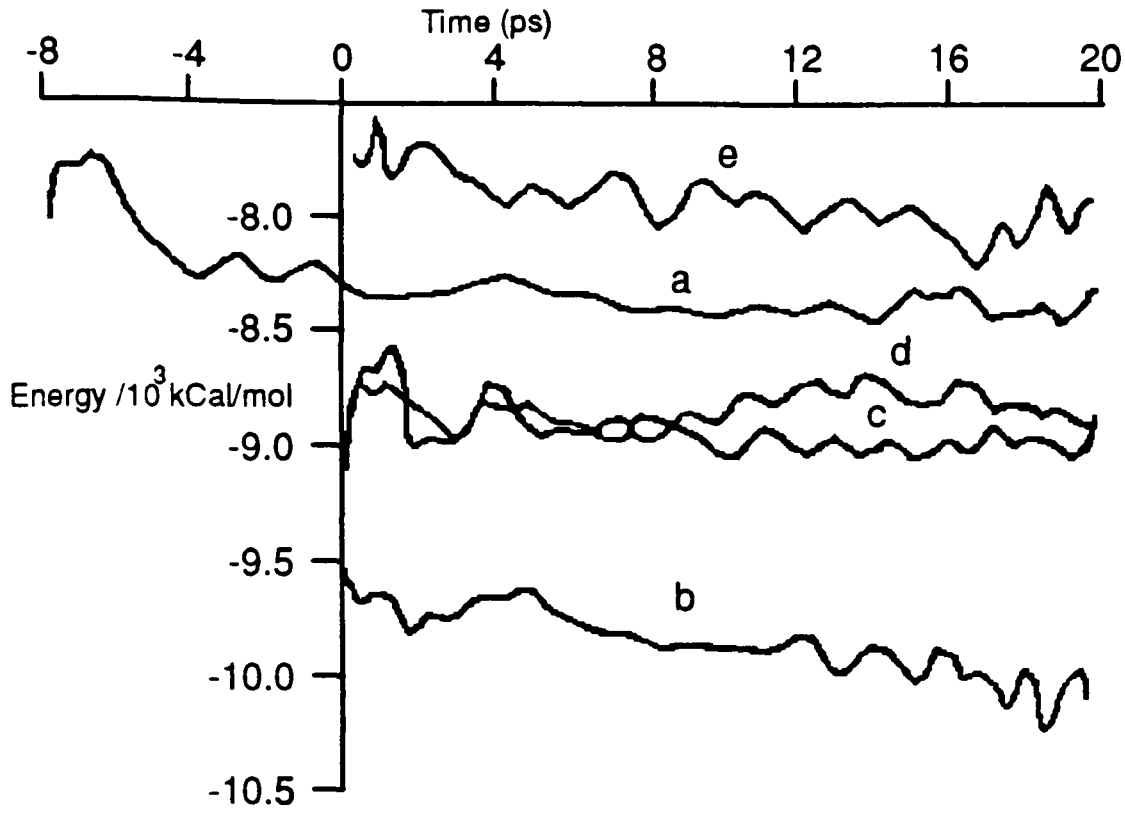
To test the hypothesis, it is necessary to simulate the protein complex with and without the phosphate groups. Any structural changes with time can be monitored and also the interaction between the nucleotide and the protein can be calculated.

The protein complex modelled above was brought up to a temperature of 298K by running 8ps of dynamics with a time step of 0.0005psIt seemed impossible to get SHAKE to work and so the step length could not be increased. At this point three more simulations were set up, each phosphorylated at one of the potential sites. 20ps more of dynamics were run on all four complexes. After some fluctuation at the start of the runs due to the phosphate groups, the potential energy of all the simulations seemed reasonably stable. This is plotted below in P3.1

* This energy and all the following energies for the protein or protein / water system, ~~are~~ is the molecular mechanics potential energy ie the sum of the bond, angle, dihedral and non-bonded interactions

Plate 6 (following page). The spatial relationship between the functional regions of the Ha-*ras*: GDP protein. The upstream regulator loop is in green. The effector region in yellow and the nucleotide in red. Putative phosphorylation sites are shown in blue.





P3.1. The potential energy of Ha-*ras* complexes during dynamics simulations in the gas phase. The region of negative time is the initial equilibration of the protein before it is transformed into the other complexes, either by exchanging the GDP for GTP or by phosphorylating different sites. The noise has been removed from the energy traces for clarity by fitting a 40th order polynomial. The RMS deviation of energies was not greater than 80 kcal mol^{-1} from these curves.

a Ha-*ras*: GDP

b Ha-*ras*: GDP with S65 and S106 phosphorylated

c Ha-*ras*: GDP with Y71 phosphorylated

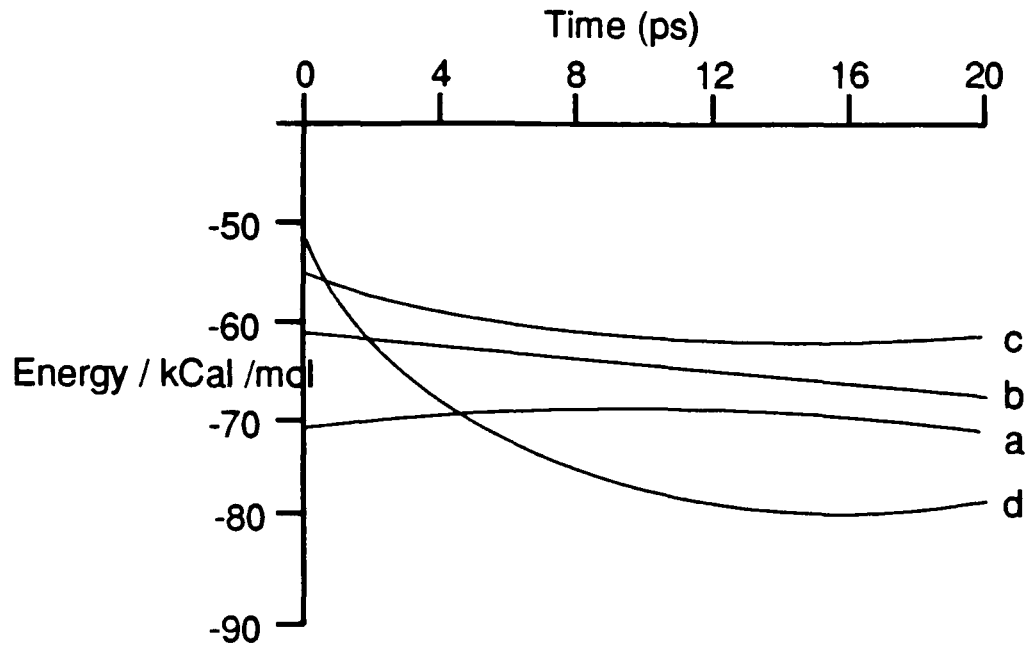
d Ha-*ras*: GDP with Y64 phosphorylated

e Ha-*ras*: GTP

III.2.1. The Effect of Phosphorylation on the Nucleotide - Protein Binding Interaction.

During the simulations described above, co-ordinates were dumped every 0.1ps. The interaction energy between the protein and the nucleotide was then calculated using the ANALYSIS module, which, amongst other things, can break the total energy down into that due to specific groups. The electrostatic term was scaled by a factor of 10 to estimate the effect of water around the solvent accessible binding site. This is plotted below in P3.2.

The plot indicates that phosphorylation at S65 and S106, or at Y71 weakens the interaction between the nucleotide and the protein. This could lead to an enhanced nucleotide exchange.



P3.2. The interaction energy between GDP and the Ha-*ras* protein phosphorylated at different sites. The noise has been removed from the curves by fitting a third order polynomial to the points calculated at 0.1ps intervals. The RMS deviation of the points from these curves was less than 7 kcal mol^{-1}. The notation of the curves is as in P3.1.

a Ha-*ras*: GDP

b Ha-*ras*: GDP with S65 and S106 phosphorylated

c Ha-*ras*: GDP with Y71 phosphorylated

d Ha-*ras*: GDP with Y64 phosphorylated

III.2.2. The Cause of the Structural Differences in the two Ha-*Ras*: nucleotide Complexes.

As well as calculating interaction energies, the co-ordinates from the simulations were analysed for secondary structure elements. The standard method for evaluating secondary structure is from Kabsch and Sander (Kabsch and Sander 1983). This evaluates hydrogen bonding along the backbone by assigning partial charges and calculating the electrostatic interaction. The pattern of this bonding is related to the various possible geometrical shapes characteristic of a protein. This method is implemented in QUANTA. Structures 4ps apart along the dynamics trajectory were energy minimised and analysed by this method. The major change looked for was the transition of residues 66 - 74 from a single helical turn to the triple turn α-helix seen in the GTP bound crystal.

It is interesting to note however that the Ha-*ras*: GDP crystal structure could not geometrically support a 3 - turn alpha helix in the position between residues 66 and 74 where it is observed in the Ha-*ras*: GTP structure. After 8ps equilibration this was now possible and in fact a second turn was present in this region. The discussed difference in secondary structure between helix α_2 (active GTP complex) and loop λ_2 (inactive GDP

complex) thus becomes much smaller when dynamic rather the static crystal structures are studied.

T3.2. Secondary structure elements of the normal Ha-*ras* complexes during dynamics.

Secondary structure name[a]	Residue numbers in element			
	Ha-*ras*: GDP[b]	Ha-*ras*: GDP[c]	Ha-*ras*: GTP[d]	Ha-*ras*: GTP[c]
β_1	2 - 7	(4) 5 - 9 (11)	2 - 9	(3) 5 - 7 (8)
α_1	16 - 25	16 - 25	16 - 25	(16) 18 - 25
β_2	38 - 45	(37) 38 - 42 (44)	37 - 46	(39) 40 - 44 (45)
β_3	49 - 57	(50) 53 - 57 (58)	49 - 58	(50) 51 - 57 (58)
α_2	68 - 72	(66) 68 - 74[e]	66 - 74	66 - 74[f]
β_4	77 - 81	(76) 79 - 81 (83)	77 - 83	(76) 77 - 81 (82)
α_3	88 - 103	89 - 99 (104)	87 - 91	89 - 104
α_3'		(101 - 104)	93 - 103	
β_5	110 - 117	(108) 109 - 113 (115)	111 - 116	109 - 114 (116)
α_4	126 - 137	128 - 136	127 - 136	(127) 128 - 137
β_6	140 - 145	139 - 142 (144)	141 - 143	(138) 139 - 144 (145)
α_5	152 - 169	(152) 154 - 169 (170)	152 - 165	152 - 167 (168)

a Secondary structure elements are named after Pai *et al.* (1989).

b Crystal structure from Tong *et al.* (1989)

c These figures are from the simulations. The first Ha-*ras*: GTP structure obtained i.e. after 4ps of MD was ignored as this structure was not equilibrated. The figures show the residues always present in each element. The figures in brackets are the maximum extent of the element.

d Crystal structure from Pai *et al.* (1989)

e Only once (after 16ps) was this region a true α-helix. The other times it was a combination of 3- and 4- turns.

f Once (after 8ps i.e. at the start) this area was not an α-helix but a series of 3- turns

On studying the secondary structure of the phosphorylated complexes during the simulation, there was no indication that the phosphates made any significant difference to the protein. In particular there were no signs of the transition mentioned above occurring.

To see if the transition is due just to the different nucleotide, in the next simulation the GDP of the minimised and 8ps equilibrated structure was swapped for GTP, putting it in the orientation seen in the Pai crystal structure. This new complex then also ran for 20ps. The potential energy dropped during this time as shown in plot P3.1. On analysing the secondary structure of this complex during the simulation, it could be seen that the

desired helix had indeed formed by 12ps. The differences in structure between the complexes is thus due entirely to the nucleotide, not to any external influence. A brief description of the secondary structure elements present during the simulations is given in T3.2.

The transformation that has occurred to the protein during this simulation with GTP in the binding site is shown in plate 7. It can be seen that after the simulation, the effector loop is further away from the protein and has two aspartate sidechains pointing into the cell ready to accept the incoming GAP protein. The change in the transformed loop, α_2, can also be seen in the ~~top right~~ bottom left hand corner.

III.2.3. Checking the model.

At the end of the simulations, the final minimised GDP complex was compared to the crystal structure. It was immediately obvious that the model was not good around the binding site. The aspartate residues 33 and 57 were much too close to the magnesium, which had distorted the loop λ_2. Also the gap between residues 60 and 66 had been filled by the fitting procedure, but no attention had been paid to whether this loop was reasonable.

Later, when the Ha-*ras*: GPPNP crystal structure had been obtained, the GDP complex was also compared against this structure to compare the unknown loop. It was seen that it did not match very well, and in fact the sidechains of Tyr64 and Glu63 were on the opposite side of the loop. This difference had remained in the simulated GTP complex.

III.3. Improving the Ha-*Ras*: GDP model.

Due to the problems mentioned above, the initial AMBER minimised protein was put back into QUANTA / CHARMm for further modelling.

III.3.1. Improving the binding site.

It was known that the GPPNP crystal contains two water molecules as ligands to the magnesium. It seems reasonable that these waters will also be present in the GDP complex. Two waters in comparable positions were thus added and the structure minimised. 1ps of dynamics, heating the complex up to 500K was then run to find reasonable minima for the waters. The structure was then minimised again using the alpha carbon constraints. The RMSD of the alpha carbons at this point was $\cong$2.6Å. At this point QUANTA version 3.1 was obtained. This contained a new option that enabled flexible fitting to occur. The alpha carbons could now be exactly overlaid onto their crystal counterparts, ignoring any potential functions. After this fitting was done, the RMSD was now $\cong 0.66 \times 10^{-2}$Å but the energy was 6.46×10^{8} kcal mol^{-1}. This structure was then

Plate 7 (following page). The difference in the HA-*ras* secondary structure before and after a simulation with GDP replaced by GTP. The GAP binding loop, residues 32 - 40 is shown with its sidechains, while the rest of the protein shows just alpha carbons. This loop at the end of the simulation (GTP complex) is highlighted in red and at the start of the simulation (GDP complex) in blue.

minimised, with the alpha carbons constrained i.e. they did not move from their positions. The energy was -9.9×10^3.

The sidechains now had to be relaxed with the new backbone structure. Firstly the complex was heated to 500K over 0.5ps. 7.5ps of dynamics at 500K, with a time step of 0.005ps, were then run and the system again energy minimised. During this time the alpha carbons were still kept fixed. The energy came down to -10.5×10^3 kcal mol^{-1}. Full conjugate gradient minimisation was now performed without constraints. During this, the magnesium ion moved to between Pα and Pβ of the GDP. One water of the magnesium ligands was expelled from the binding site by this movement and residues 31 -35 moved in close onto the magnesium.

The magnesium was moving in order to be better co-ordinated. In the GPPNP complex, the γ-phosphate is also a ligand. A water molecule was therefore added in the appropriate position to replace this phosphate. The binding site was then solvated, adding water in a sphere 5Å around the magnesium, as it was thought there might be a complex water structure that holds the magnesium and ligands in place.

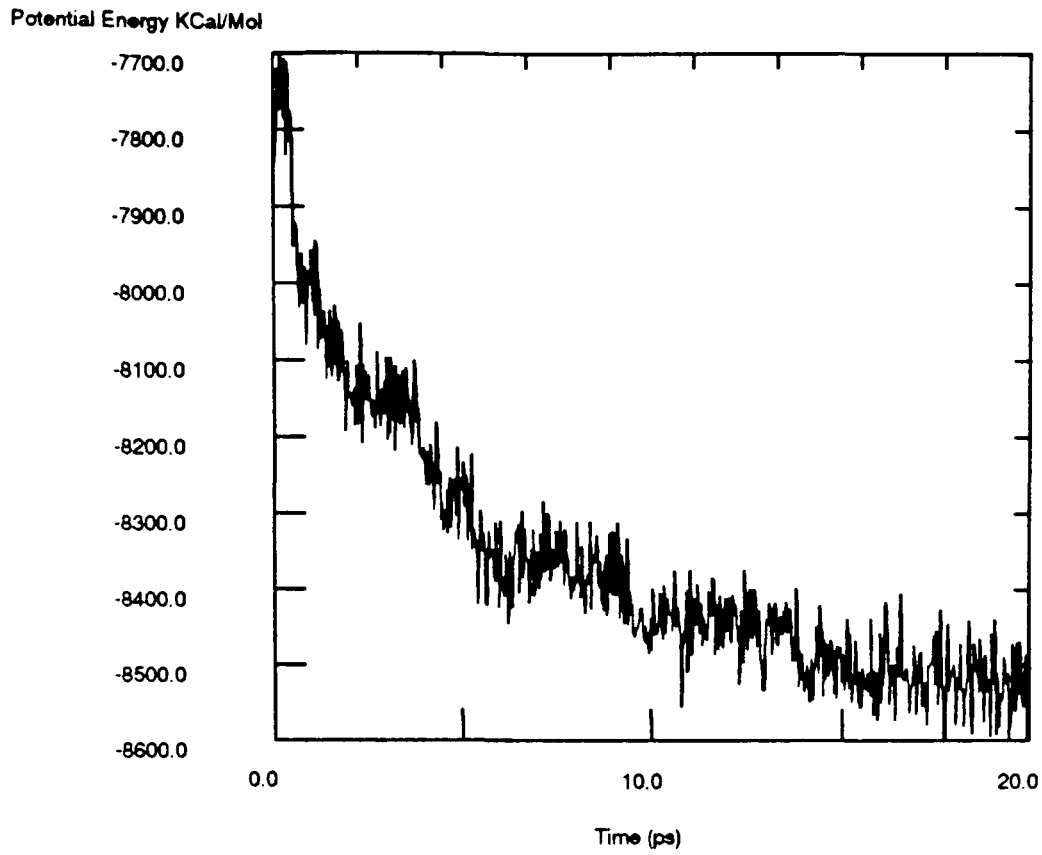
This solvated complex was minimised, firstly with alpha carbon constraints then with no constraints. It was found that the magnesium had not moved, but Asp33 had moved away from its crystal position to act as a ligand. A fourth water molecule was thus needed between the magnesium and Asp33. This complex was then minimised, first with restrained alpha carbons to allow the new water to find a good site, then without constraints. The binding site now matched the crystal structure.

Interestingly, the Max Planck group (Schlichting *et al.* 1990) have found that Ha-*ras*: GDP formed after GTP hydrolysis has Asp57 as a direct ligand to the magnesium, not mediated by a water as in the GPPNP complex. However this is not possible in the crystal structure modelled here. This would indicate that the magnesium ligands may vary, with sometimes water and sometimes residues in the first interaction shell.

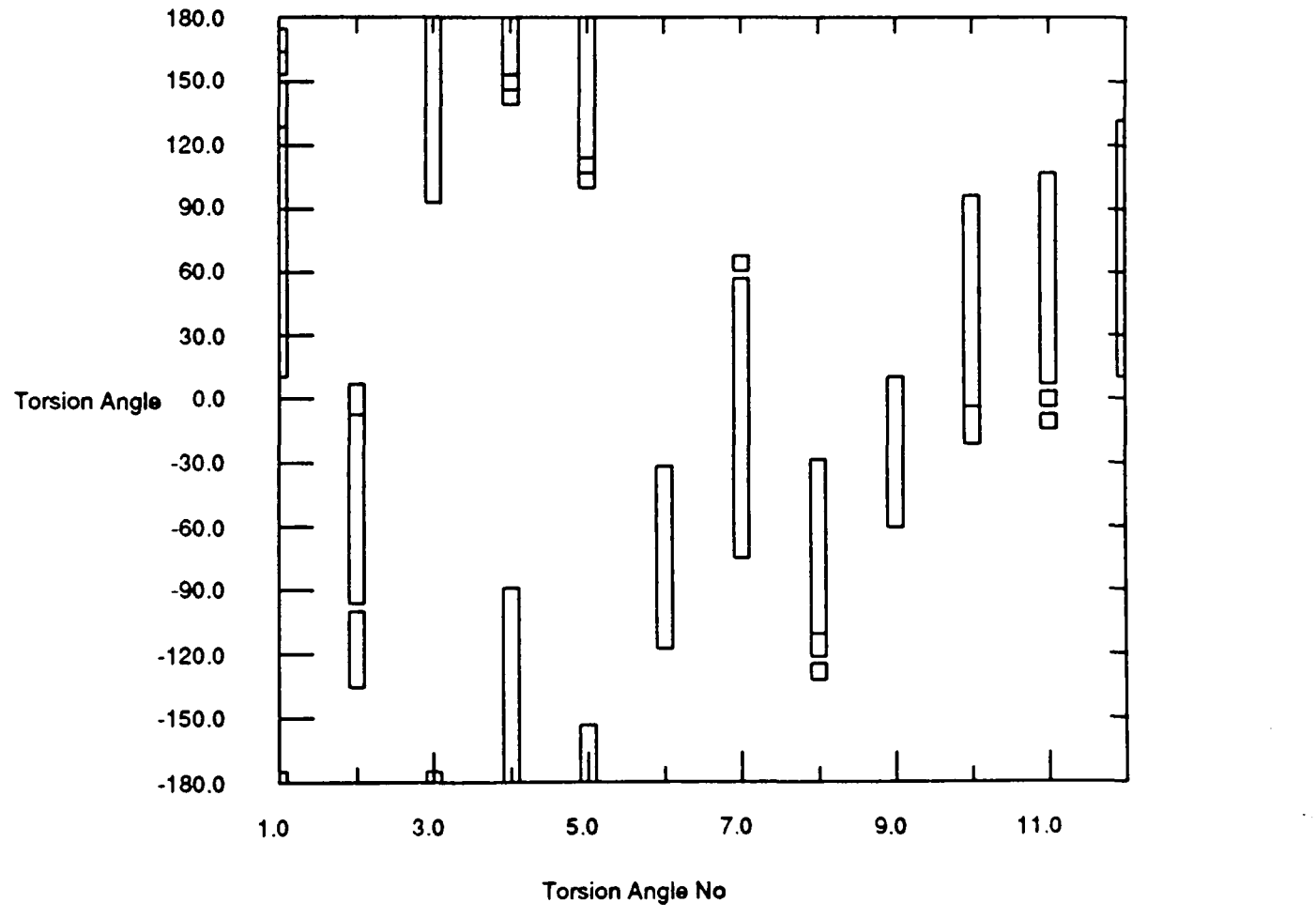
III.3.2. Finding a Good Conformation for the Undefined Loop.

Up to this point the loop between residues 61 and 65, which is missing in the crystal structures had been ignored. As it is very flexible, its minimum conformation is likely to dependant on the atoms around it. It lies sandwiched between two more rigid loops and so the conformation of these loops will affect the unknown loop. Because of this need to include the environment, a database loop search would be unlikely to produce a reasonable structure. Similarly the conformational search programs in QUANTA and other packages are unable to handle such a problem. It was thus decided to use MD to search the conformations. Solvent, while almost certainly being important for this loop - it is very exposed- would make the calculation much longer. This is due both to the larger system

required and the damping of motion by the solvent. When solvent is added it would be thought that the loop should quickly adapt to the new environment.



P3.3. The change in potential energy of an Ha-*ras*: GDP model under dynamics at 500K, constrained to the crystal structure.



P3.4. The variation in torsion angles of residues 61 - 66 during the dynamics described in P3.3. The angle numbers are in the order ϕ(C-N-C_α-C), ψ(N-C_α-C-N) for residues 61 through 66.

The flexible-fitted, minimised protein model was thus taken with no waters present. All the known crystal positions (alpha carbons, GDP and Mg) were restrained. 20ps of MD were run at 500K. How the energy dropped during this time is shown in plot P3.3. This large stabilisation is entirely due to the reordering of the sidechains and the loop 60 - 66. Co-ordinates were dumped every 20 steps. Plot P3.4 shows the variation of torsion angles for the unknown loop. The dynamics searched quite a reasonable number of different torsions. Hence the final structure is likely to be a much better model than the original. The lowest energy structure from this simulation was taken and optimised, with the alpha carbons constrained.

The last stage was to solvate the binding site. First the four waters were replaced around the magnesium and minimised to the new positions of the residues in the site. A 15Å sphere of water was added around the magnesium. Unfortunately QUANTA / CHARMm adds completely ordered waters. This order was shaken out by minimisation, followed by 4ps of MD to give better solvation. At this time the alpha carbons still had an RMSD of 0.006Å. The whole system was then freely minimised. After this the RMSD had dropped to 0.75Å. This structure looked very close to that in the crystal. Only the loop from Glu47 to ASP49 showed to have a noticeable deviation, of the order of 1Å. This loop, λ_3, is on the outside of the protein so this could be due to the different environment.

This model was then put back into AMBER for further calculations.

III.4. Using the Improved Model

Dummy atoms were placed ready to become phosphates at all the sites mentioned above. Extra solvent was added to give a sphere with an 8Å deep layer of water all round the protein. The protein was also cut down from the 171 residues of the 2p21 crystal to 166. The residues removed stick out from the side of the protein and their removal allowed a sphere to cover the whole protein, otherwise this region would be exposed. This needed 1758 water molecules (a total of 6969 atoms) which made the system cumbersome to handle. No more solvent would be possible if a reasonable length of simulation was to be run.

It had been intended to use a cap force to keep the solvent in place around the protein. However early dynamics with this force present moved the waters in between the loops of the protein, destroying its tertiary structure. The water was thus left free to wander and it was hoped that not too much would evaporate during the simulations.

The parameters were as before, except for the atom centred charges on the GDP. These were initially calculated by fitting to a 3-21G*+ MEP calculated on the nucleoside in its binding site conformation. This basis set, 3-21G* with diffuse functions added on the phosphate atoms only, is discussed further in V.1.2.1.

This system was minimised with the protocol used before in III.1.1, of 20 steps of steepest descent followed by the use of the conjugate gradient minimiser. For this system, the AMBER force field was different enough from the CHARMm force field to produce new problems in the binding site. In fact AMBER minimisation of the CHARMm minimised structure completely distorted the GDP. Fortunately the model that was CHARMm minimised, but still constrained to the crystal positions was acceptable to the AMBER minimiser.

III.4.1. Problems With the Magnesium Ion During Solvated Dynamics.

A new problem now arose. It proved impossible to keep the magnesium ion in the desired position. No matter what methods were tried - slowly minimising segments separately to shake out minor problems, or the use of restraining potentials - the magnesium always drifted to lie between the two phosphate groups, or at the centre of the β - phosphate terminal oxygens.

At first it seemed to be the fault of the solvation as the water ligands very often moved out of place around the magnesium. There were also visible holes in the water structure in the binding site due to the contraction of the water onto the phosphates under minimisation. Because of this it was postulated that the low density of water was allowing the magnesium to move. This low density was rectified by re-solvating the binding site. Waters were kept that fitted the criterion that any atom was more than 1.5Å away. Five more waters were added within a distance of 14Å from the magnesium (it was also tried keeping waters more than 1Å from anything. This added a further 21 waters, but steric clashes arose and the minimiser failed). Despite this the magnesium still refused to stay put.

Interestingly, when the magnesium moved to lie between the phosphates, the β - phosphate rotated. The alpha carbons were virtually unchanged by this with the exception of Gly12 which moved as far as 3Å from the crystal position. This residue is thus linked to this phosphate position. In the last chapter it was explained that mutations in this position cause the protein to have decreased GTP-ase activity. All residues are more bulky than glycine and so any mutation could cause the phosphates to be in a different orientation, preventing the attacking water from reaching its target. The one exception would be proline which, due to its shape, causes tight turns in the protein backbone similar to glycine.

Despite noting this fact, three months had now passed and AMBER still refused to accept the crystal structure. It was tried ~~to~~ lowering the charge on the magnesium, moving electron density from the phosphates in the vague hope that an electron transfer had occurred due to the high electron density on the terminal phosphate. A charge calculation on GDP including the magnesium ion seemed to support this as the charge on the

magnesium was 1.8 and there was appreciable density - a weak bond - between O3B and the magnesium. The magnesium still moved as before. Even with the magnesium charge lowered to 1+ it still moved between the phosphates, but now stayed much closer to the β - phosphate.

There were now two possibilities. Either the crystal structure was misleading, or the AMBER parameters were wrong. The Brookhaven file, 2p21 shows a thermal factor for magnesium of 8.88. When compared to the B values of its nearest neighbours, Ser17 (17.45), Asp33 (36.05), Asp57 (19.96) or the GDP oxygen atom O3B (24.21), the difference of this value is apparent.

This seemed strange as the magnesium is only held in place by these neighbours and water, so it would be expected to move to the same extent. Another factor was that the magnesium position was seen to be similar to that in the GTP structures. In fact after the GPPNP complex structure was obtained, the magnesium ions were seen to be in almost identical sites. It was thus a possibility that the crystal had been trapped in a metastable state and had not rearranged fully into the GDP structure. It seemed reasonable that a magnesium ion would prefer to be between two phosphates, rather than associated with just one. It was also possible that the crystal had been refined onto another structure and the magnesium wrongly assigned.

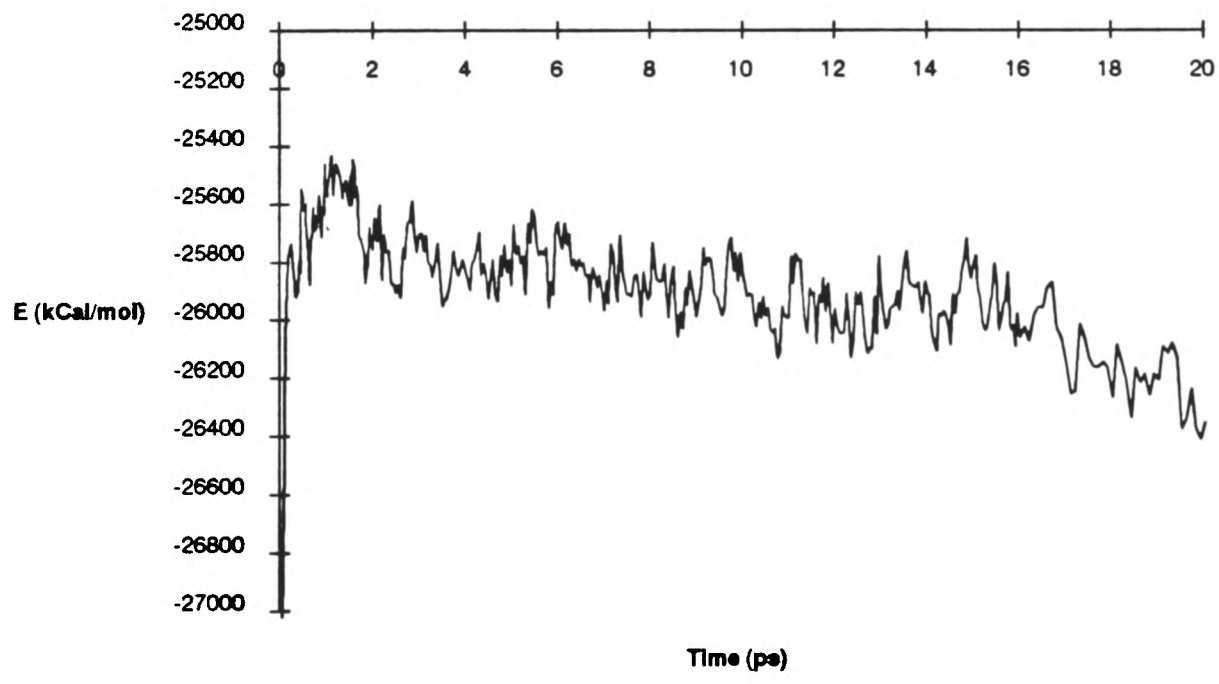
In order to check this the AMBER standard STO-3G charges for nucleosides were used. These are independent of the GDP geometry and the presence of the magnesium, although they are probably not a good model of a diphosphate chain and in fact there are major differences in these charge sets at the β - phosphate. Amazingly, with these charges the magnesium did not move. Even at the end of prolonged dynamics it did not associate with the α - phosphate. This vindicated the crystal structure and demonstrated the problems of molecular mechanics. Why did these generic DNA charges work better than the custom calculated ones which should have fitted the force field? This question is carried on in chapter V. Presumably the reason why the B values are so different is that the crystal co-ordinates are only for the alpha carbons. It is possible that the magnesium and its ligands are fairly rigid, but the backbone actually moves more.

With the STO-3G charges the minimisation had been done in two stages: waters first, then everything. After this the RMSD was 0.666Å. The comparison of this structure, taken to be an adequate model of the protein, with the crystal structure showed that now even λ_3 was now almost exactly following the crystal structure. 2ps of MD were then run to heat the system from 0 to 298K. A time step of 0.005ps without SHAKE was used in order to be gentle. The time step could then be increased to 0.001ps and SHAKE turned on. After a further 2ps of MD, the RMSD was now 1.520Å. The magnesium and GDP had moved compared to the crystal position, but were still in the correct orientation and the magnesium co-ordination was unchanged.

III.4.2. Problems with the Use of SHAKE.

The next problem was to get SHAKE to work with a time step of 0.002ps. Without this simulations would, by necessity, be very short. Unfortunately this proved to be difficult. SHAKE is known to be a problem with macromolecules. This is due to its implementation. It adjusts bond lengths down the chain in an iterative manner. This means that large deviations can add up until a bond can no longer be brought within the tolerance. At first it was thought that the problem was just due to bad forces in a large system. Dynamics were thus continued with a 0.001ps time step, every 2ps checking whether a larger time step would work. However even after 16ps this still was not to be. It was however found that SHAKE worked best with a tolerance of 0.0001; smaller than this and the bonds just could not meet this criterion of larger than this the bonds in the chain could be outside the ~~tolerance limits~~ equilibrium value and high forces arose.

The problem was eventually traced to the AMBER force field. The sulphur atoms have two "lone pairs" with charges. These have a mass of 1. Also there is a bond force constant between the sulphur and lone pair of 600 kcal mol^{-1}. This gives rise to very high frequencies of vibration. In fact no hydrogen atom in the force field has such a high force constant and these are known to cause problems at large time steps due to sampling at a point in a vibration when the bond is at one end of its extension. To overcome this, the masses were shifted. The sulphur itself was given a mass of 20 and the lone pairs 6 each. This should have the effect of just shifting the centre of gravity of the sulphur a small distance from its atomic centre (the S - LP bond length is only 0.679Å).



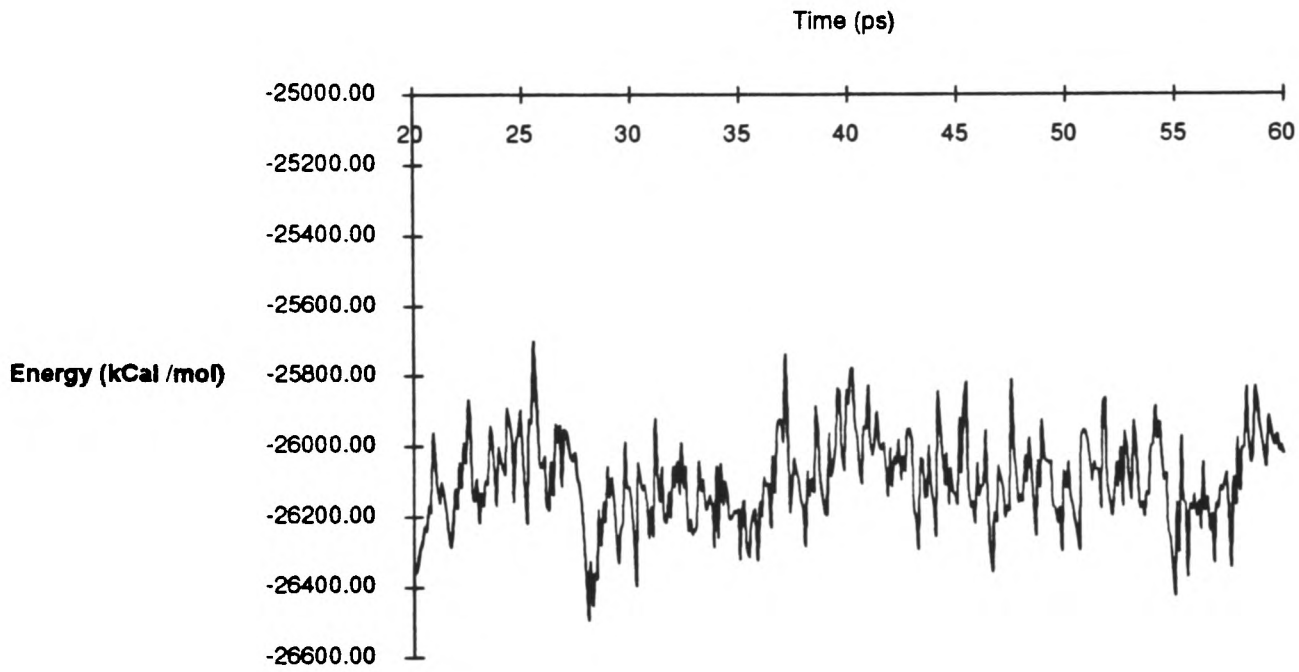
P3.5. The change in potential energy of the Ha-*Ras*: GDP complex over 20ps of dynamics, starting from the AMBER minimised model of the crystal structure.

This higher mass did indeed damp out the motion and SHAKE worked immediately with a 0.002ps time step. After a further 4ps under these conditions, the system was deemed equilibrated. This is shown by the potential energy trace below in P3.5. The temperature was stable at 298K and the potential energy had come down to around -2.63×10^{-4} kcal mol^{-1}.

III.4.3. Simulation of the Dynamic Motion of the Ha-*Ras*: GDP complex

The gas phase calculations could now be repeated in solution. Again the first simulation to be run was on the normal system dynamic behaviour. Later phosphate groups would be grown onto the dummy atoms and any changes in structure and binding energy noted.

Taking the solvated and equilibrated protein, a further 4ps of dynamics were run for good luck. 40ps were then run, broken up into 10 sets. The change in potential energy during this time is shown in P3.6.

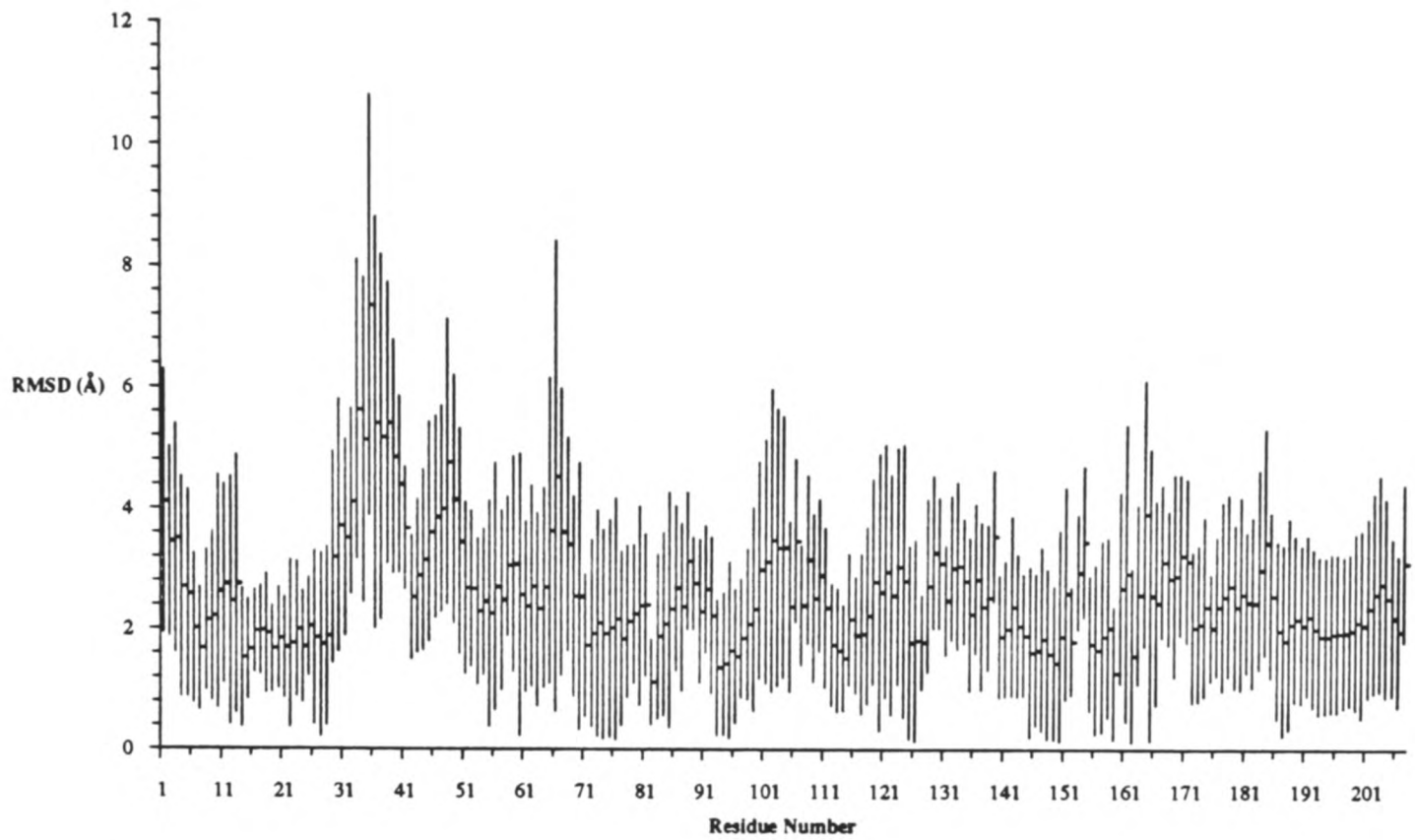


P3.6. The change in potential energy of the Ha-*Ras*: GDP complex over 40ps of dynamics, starting from the AMBER equilibrated model.

During this time the system co-ordinates were dumped every 100 steps, or 0.2ps. Each co-ordinate set was matched to the minimised AMBER "crystal structure". The deviations of the alpha-carbons, nucleoside and magnesium were then calculated. The maximum and minimum of these values are plotted in P3.7.

Several points of interest can be noted from this plot. Firstly, in general residues that are inside the protein are closer to their crystal positions, often being within 1Å, compared to the external residues that typically only get to within 1.5Å. In general the protein was seen to have expanded in the solvent.

The most obvious deviation is between residues 32 and 40. This is the GAP binding region. It is a very exposed and in bulk solution moves away from the protein bulk. It may be noted that in fact this region moves further than the loop 61 to 65 which was missing in the crystal. Residue 66, the start of the region that becomes an alpha helix in the GTP structure is also seen to move a long way.

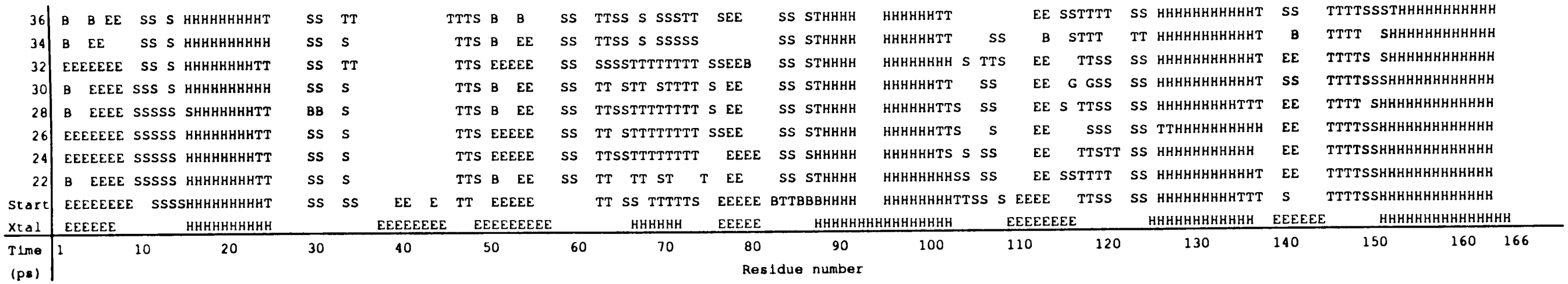


P3.7. The RMS deviation of the Ha-*ras*: GDP protein from the initial minimised model during a 40ps simulation. Only the alpha carbons, nucleotide and magnesium i.e. the atoms in the crystal were measured. Each line represents the maximum and minimum values from structures taken at 0.2ps intervals.

The secondary structure elements present during the simulation were also calculated. As before, this was done by applying the Kabsch-Sander method to structures taken at 4ps intervals. This time the original code was used for this rather than the QUANTA implementation. This secondary structure is shown in P3.8.

The ~~above~~ table shows the stability of the secondary structure of the protein. Even during the dynamics at 298K the structure elements, especially the helices, stay in the same regions of the primary sequence, although the ends of the elements do vary. On comparison with the RMS deviations, P3.7, it can be noted that the regions marked as helices, or turns move much less than bends, or in the case of the GAP binding region where no secondary structure is detected.

The region between 66-74 is interesting. The crystal structure shows this as a one turn helix. The AMBER minimised structure makes this a turn sequence, but after 30ps of dynamics this has definitely changed to a bend sequence, a sign that a transition of some



P3.8. The Secondary structure of the Ha-*ras*: GDP complex during a 40ps dynamics simulation. The notation is that from the Kabsch-Sander method.

H	4-helix (α-helix)
B	isolated β-bridge
E	extended β-strand
G	3-helix (3_{10}-helix)
I	5-helix (π-helix)
T	H-bonded turn (residues with 3- or 5-helix bonding pattern, but the element length is too short to be classed as a helix)
S	bend (a region of high curvature)

sort is occurring. In the gas phase simulation this region became a two turn helix. It is possible that this is happening, but the solvent is slowing down the transition.

III.4.4. The Effect of Phosphorylation on the Nucleotide - Protein Interaction in Solution.

The next step was to introduce the phosphate groups and see how this affects the nucleotide binding. Three simulations were again carried out, to correlate with the possible sites of phosphorylation mentioned above. The 20ps equilibrated Ha-*ras*: GDP system was put into the GIBBS module. The relevant dummies were mutated into a phosphate group at either Tyr71, Tyr64 or Ser65 and Ser106. This was done using a slow growth perturbation over 8ps. This is not very long for a perturbation, but should produce a structure which could then relax with further dynamics.

The charges on the phosphorylated residues were a combination of AMBER standard charges smoothed to give a charge of minus two, mostly on the phosphate group. It was hoped that exact electrostatics were not necessary here as the effect is a long range one - the introduction of a large charged group should be the important factor.

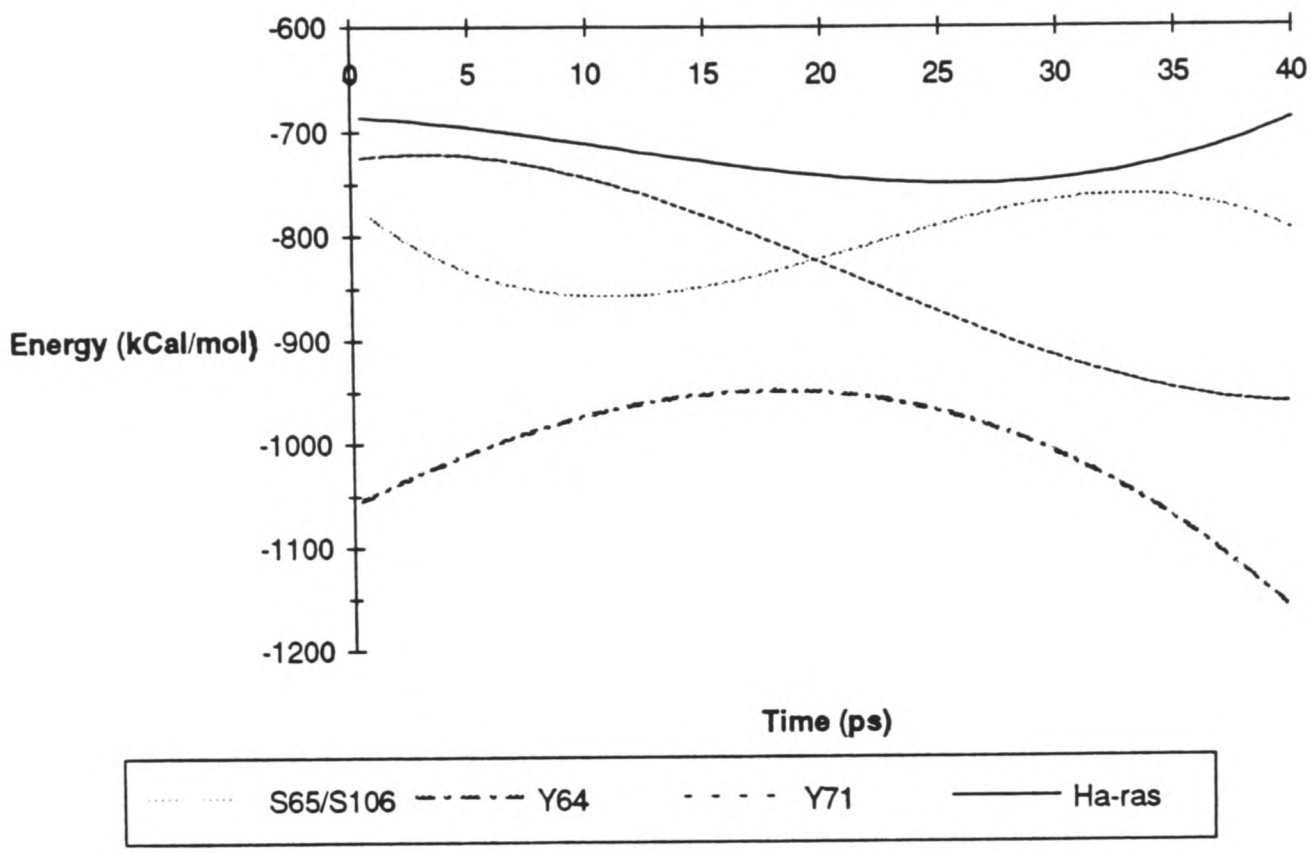
4ps of MD was then run on each phosphorylated system. As with the unphosphorylated Ha-*ras*: GDP system 40ps of dynamics were then run, collecting co-ordinates every 200 steps (collecting every 100 steps as before produced ludicrously large data sets and it was not likely to be necessary for the calculation of the average interaction energy).

The interaction energies between the nucleotide and the protein (including the binding site magnesium as part of the protein) were calculated on each co-ordinate set i.e. at 0.4ps intervals. These energies are plotted below in P3.9. The points have a curve fitted to them to reduce the noise and allow comparisons to be made. One artefact of this curve fitting is that the plots tend to swerve near the ends towards the final value. However they do give a good representation of the average energy during the time of the simulation.

In comparison to the gas phase energies, none of the phosphorylations has lowered the interaction energy. However the phosphate groups in all the positions have interfered with the binding, although this actually appears to have made the binding stronger. The curves also show that the systems were not very well equilibrated at the beginning of the data collection. Both the phosphorylated-tyrosine runs drop markedly in energy over the 40ps duration. In contrast the phosphorylated serine run is going up in energy as the time increases. Note also that the RMSD for this run is much higher than the others.

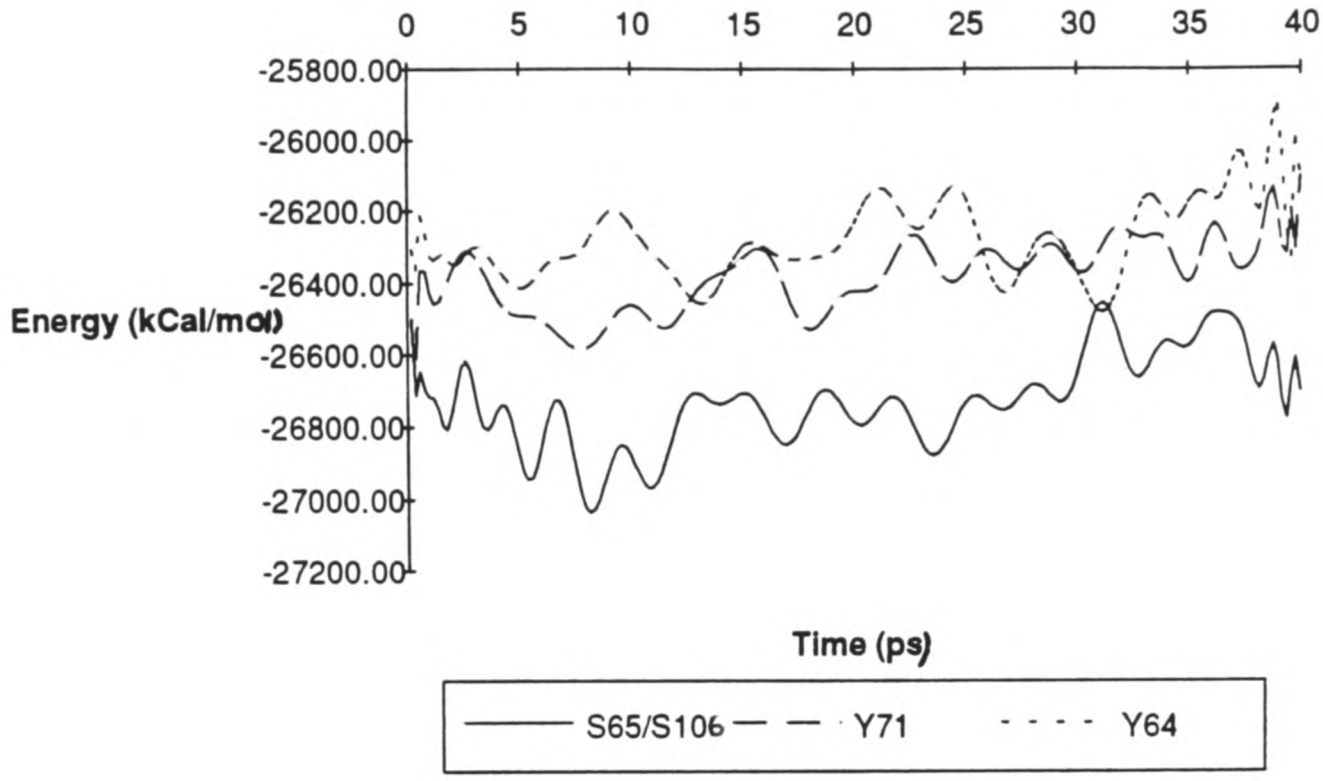
This indicates that the phosphorylated systems had not yet stabilised. The potential energies for these systems are plotted below in P3.10. This confirms that the systems have not equilibrated; the energy is still increasing at the end of the

simulation. However the indications are that the addition of a phosphate group would not enhance nucleotide exchange.



P3.9. The interaction energy of GDP with the Ha-*ras* protein after phosphorylation at various sites. The curves are third order polynomials fitted to the energies calculated at 0.4ps intervals. The RMS deviations of all the energies to the curves are as follows

Curve	RMSD
Ha-*ras*	38.74
S65/S106	52.05
Y71	35.17
Y64	40.58

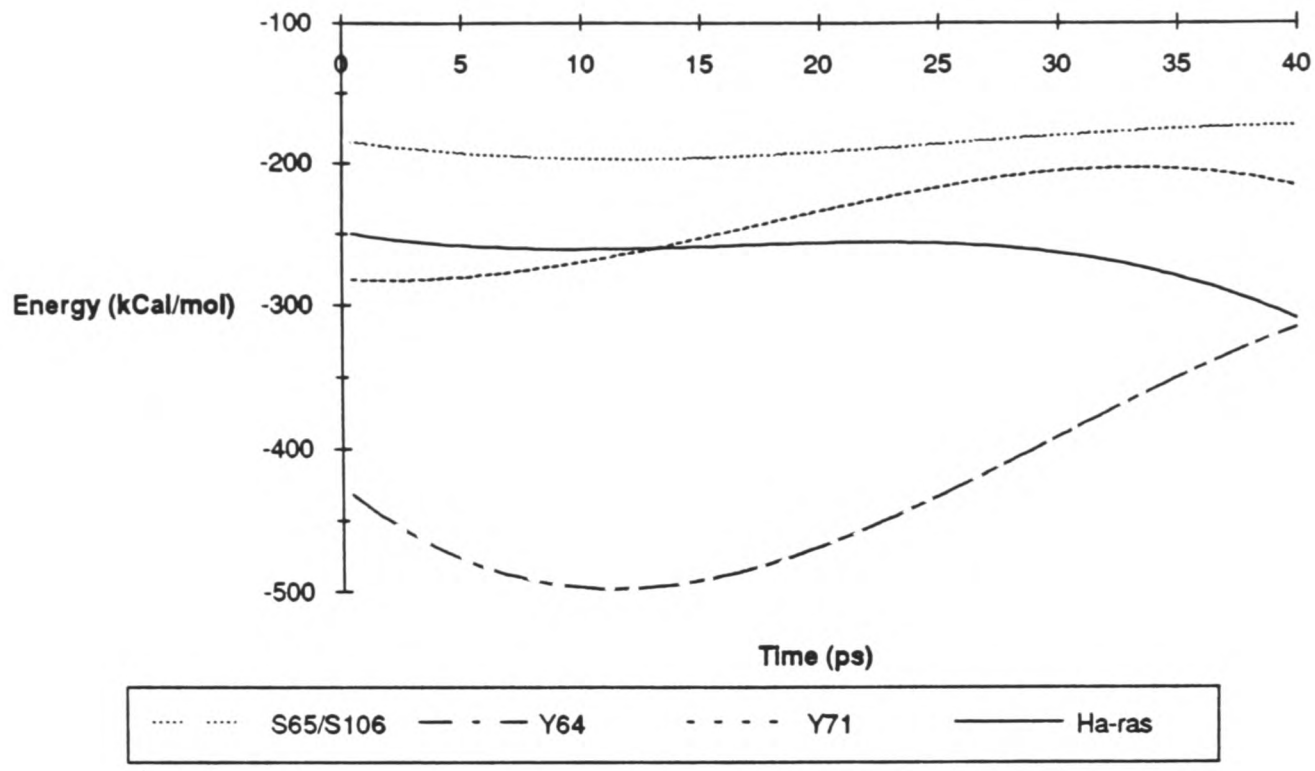


P3.10. The potential energy of Ha-*ras*: GDP complexes phosphorylated at various sites during 40ps dynamics simulations. The noise has been removed by fitting a polynomial of order 40 to the energies. The RMS deviation of these points from the plotted curves are given below

Curve	RMSD kcal mol^{-1}
S65 / S106	140.57
Y71	101.79
Y64	88.71

It is also possible that while the protein - nucleotide interaction has increased, the nucleotide - water interaction has also changed, so encouraging the nucleotide into solution. These energies are plotted below in P3.11.

These curves indicate that the nucleotide actually sees less water after the phosphorylation has occurred, perhaps indicating that the binding site has closed up. The Y64 run seems to have an increased interaction, but it is rising steeply by the end of the run. Given the high RMSD for this plot, the water might not be properly relaxed.



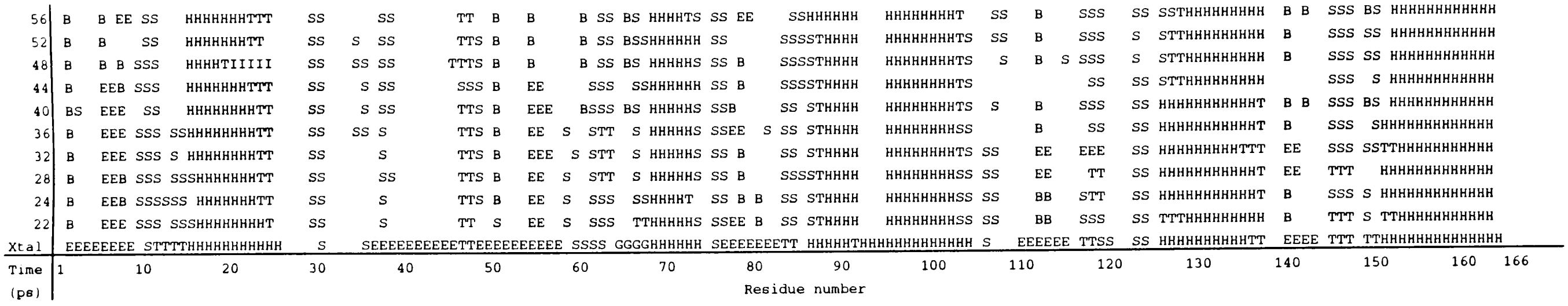
P3.11. The interaction energy of GDP with the water around the Ha-*ras* protein after phosphorylation at various sites. The curves are third order polynomials fitted to the energies calculated at 0.4ps intervals. The RMS deviations of all the energies to the curves are as follows

Curve	RMSD
Ha-*ras*	16.44
S65/S106	15.90
Y71	16.31
Y64	33.81

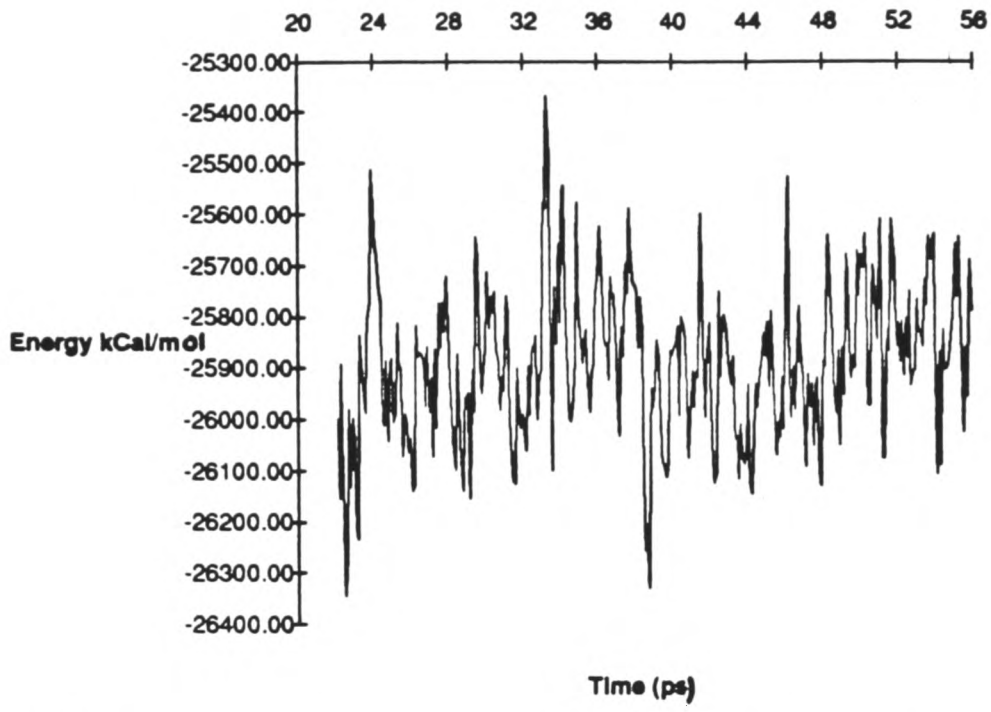
III.4.5. The Effect of the Presence of GTP in the Binding Site

To check the structural transition seen in the gas phase on exchanging the nucleotide, the GDP was mutated into GTP over 12ps. Dynamics were then run for 56ps. As before, the co-ordinates were saved every 4ps and analysed for the secondary structure elements. These are shown below in P3.12.

It can be seen that while the simulated GTP complex has not transformed fully into the structure seen in the crystal, it is halfway there. The helix between residues 66 and 74 has two turns and the area that should be the third is a bend sequence, unable to flip over into the necessary form due to the inertia of the solvent. The potential energy drop over this time is shown below in P3.13.



P3.12. The Secondary structure of the Ha-*ras*: GTP complex over the last 40ps of a 56ps dynamics simulation. The notation is that from the Kabsch - Sander method described in T3.11. Xtal denotes the secondary structure of the Ha-*ras*: GPPNP crystal structure analysed in the same manner.



P3.13. The potential energy of Ha-*ras* GTP over the last 36ps of a 56ps simulation. The complex was created from the GDP complex by the creation of the third phosphate group.

The potential energy trace can be seen to be rising slightly and the fluctuations are still decreasing at the end of the 56ps. This indicates that the complex has not yet finished its relaxation.

There was also a problem with this complex. During the formation of the third phosphate group, it had been difficult to stop the growing phosphate interacting with the magnesium in such a way as to distort the binding region. Eventually the magnesium ion had been fixed in place to prevent this happening. Even so the final orientation of the new group was not correct and the γ-phosphate was actually lying on the wrong side of loop λ_2 compared to the crystal structure. This had distorted the loop and pushed the effector loop out into solution.

III.5. Conclusions

The simulations presented in this chapter throw some light on the problems associated with the simulation of protein motion. One obvious problem is the inclusion of solvent. While this is thought necessary for a good description of the system, it slows down any transitions so that very long simulations are required. There is evidence that solvent does not affect the actual motions of the protein residues, merely the time scale (Karplus and McCammon 1983), but there is something unsatisfactory about running simulations in the gas phase. Attention was also drawn to the problem of finding a good model of the electrostatics involved in such simulations, differently calculated charges producing very different results.

The purpose of the simulations in this chapter was two fold. The principal aim was to set up a model of the *ras*: GDP and *ras*: GTP complexes to be used in the later simulations. The former was reproduced accurately from an incomplete crystal structure, but the latter was only produced in the gas phase in an unsatisfactory manner. Fortunately a crystal structure of the similar Ha-*ras*: GPPNP was received which could be used instead.

The second objective of the simulations was to try and investigate a possible mechanism for enhancing the nucleotide exchange of the *ras* proteins. It is hard to draw any conclusions about the effect of phosphorylation on the *ras* - nucleotide interaction. The gas phase simulations indicate that this is a possible mechanism for enhancing nucleotide exchange, however given the problems in the binding site structure and lack of solvation, which would obviously affect the action of a charged group like a phosphate, these results are dubious.

The solvated results are also not clear cut. While the model used should be adequate, it is not definite that the protein complex has fully relaxed. This can be clearly seen by the lack of structural transition in the case of the GTP bound complex, which was seen in the gas phase. Longer simulations are required to check this problem. The poor model of the electrostatics may also be a problem.

The only definite conclusions are that the structural differences between the two complexes are due to the presence of the different nucleotides. This is fortunate for the later calculations. Also, phosphorylation does affect the binding energy, but exactly how is not clear.

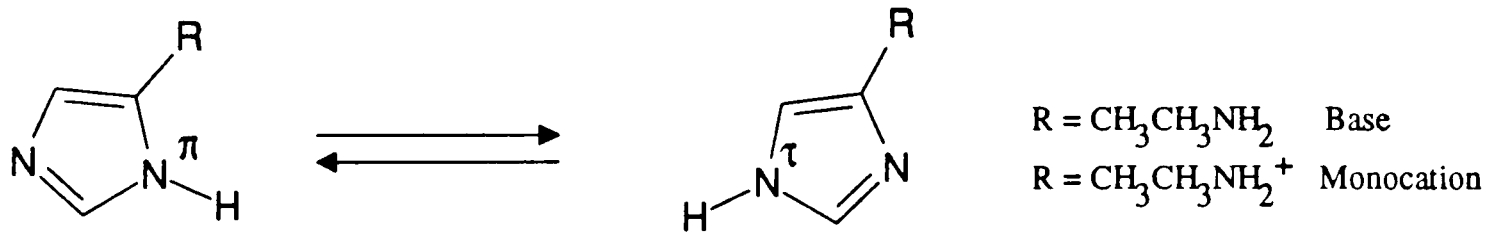
CHAPTER IV

Improving The Model

IV.1. A Model System for Study: Histamine Tautomerism

Histamine is a biologically active small molecule that is involved in various response mechanisms of the body. As everyone knows it is also responsible for many human problems such as allergies and stomach ulcers. For this reason it has attracted a large amount of research prompted by the drug companies. It has also been well studied theoretically, using both early gas phase *ab initio* and more recent solution phase molecular mechanics calculations. In fact it could be claimed that the successful H2-antihistamines Tagamet (cimetidine) and Zantac (ranitidine) were among the first drugs (and certainly the most financially successful) whose development incorporated the use of theoretical calculations.

Its main feature is three nitrogen atoms available for protonation; two in an imidazole ring and one on the sidechain. It can thus exist in three differently protonated states: the uncharged base has only one ring nitrogen protonated; the monocation where the sidechain amino group is also protonated; or a dication where all the nitrogen atoms are protonated. All three species coexist at physiological pH in a ratio of base: monocation: dication of around 1:96:3 (Ganellin 1973). Both the monocation and base are able to undergo tautomerism, having the ring proton on either nitrogen. The convention used for naming these tautomers is shown in F4.1. This gives five separate species, all able to undergo internal rotations along the sidechain to produce a whole host of different conformers, which are differently stabilised by different conditions of acidity. Hence its widespread use by the body.



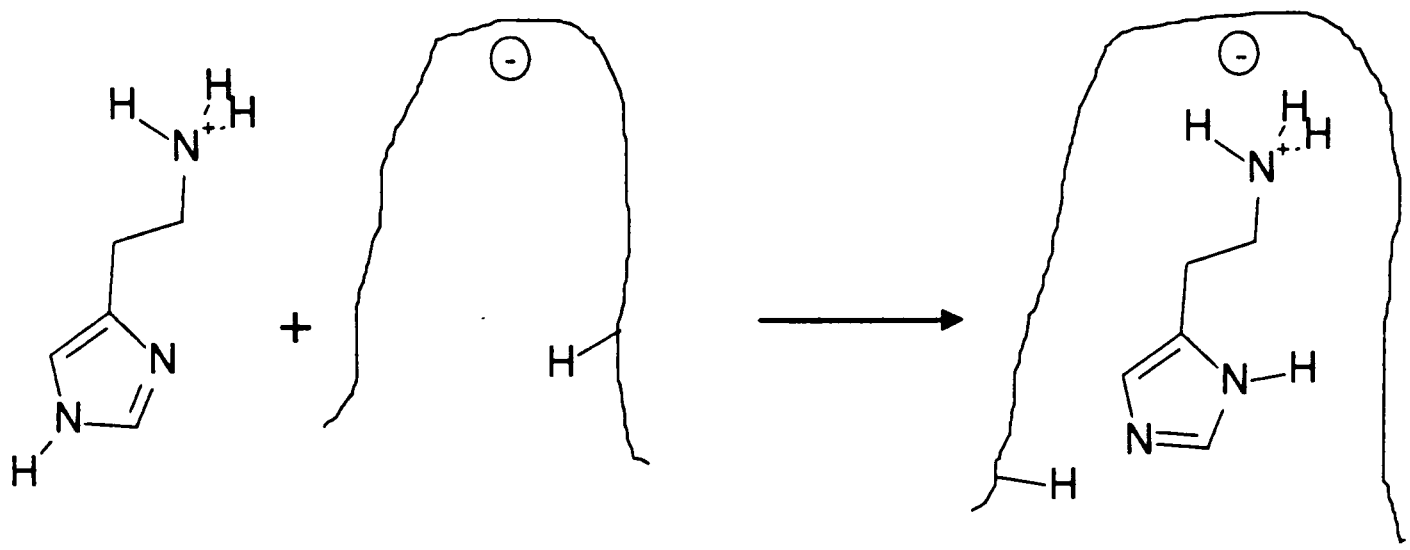
N(π)H or proximal Tautomer

N(τ)H or distal Tautomer

F4.1. The Tautomerism of Histamine

There are three known receptors. For one of these, the H2 site (responsible for gastric secretions), a mechanism has been proposed (Weinstein *et al.* 1976, Weinstein *et al.* 1986). The monocation, in the more common N(τ)H tautomer, enters the receptor and is neutralised to effectively become the uncharged base. The equilibrium constant for the tautomerism is changed significantly enough by this to promote a proton shift across the

ring. To study this mechanism, it is thus necessary to know the equilibrium data for these two species in solution and in the receptor. Experiment has so far failed to make any measurements on the receptor. Under these circumstances one must turn to theory in order to tackle this problem.



F4.2. The proposed H2 mechanism for histamine. The receptor is unknown, but is postulated to contain a negative charge, a proton and a proton acceptor site.

To calculate the ratio of tautomers present under different conditions, i.e. in solution and the receptor, it is necessary to know the equilibrium constant connecting them, defined by

$$K_T = \frac{[\text{Histamine N}(\tau)\text{H}]}{[\text{Histamine N}(\pi)\text{H}]} \qquad \text{(E4.1)}$$

This is of course related to the Gibbs free energy difference between the tautomers in the appropriate phase by

$$K_T = \exp(-\Delta G/RT) \qquad \text{(E4.2)}$$

IV.2. Making A Free Energy Perturbation Calculation

As mentioned in I.2.2, absolute free energies cannot be calculated using molecular mechanics. A non-physical calculation is thus used instead, applying the FEP equation E1.44. In this problem, during the course of a simulation, the N^{π} proton is mutated into a dummy atom and a dummy atom on N^{τ} becomes a proton (or vice versa).

In practice the calculation works as follows. The system is equilibrated using molecular dynamics so that it is at a constant temperature and pressure. Dynamics are then continued, but after each step the Hamiltonian is changed from that for the starting tautomer to the other tautomer. The difference in energy between the states over the configuration of the starting system for each ensemble member is thus measured. The exponential values of these energy differences are stored and from an average of these, the free energy is obtained.

Only the inter-molecular part of the potential is monitored. The perturbation Hamiltonian can be divided up into intra- (containing the bond, angle and dihedral energy terms) and inter-molecular (containing the non-bonded terms). When this is done equation E1.44 can be written

$$\Delta G_{BA} = -RT \ln\left\langle \exp(-\Delta\mathcal{H}_{inter}\beta)\exp(-\Delta\mathcal{H}_{intra}\beta)\right\rangle_A \qquad \text{(E4.3)}$$

The terms inside the brackets cannot be split up because they are an average value and

$$\frac{1}{N}\sum_{A,B} e^A e^B \neq \frac{1}{N}\sum_A e^A \sum_B e^B \qquad \text{(E4.4)}$$

i.e.

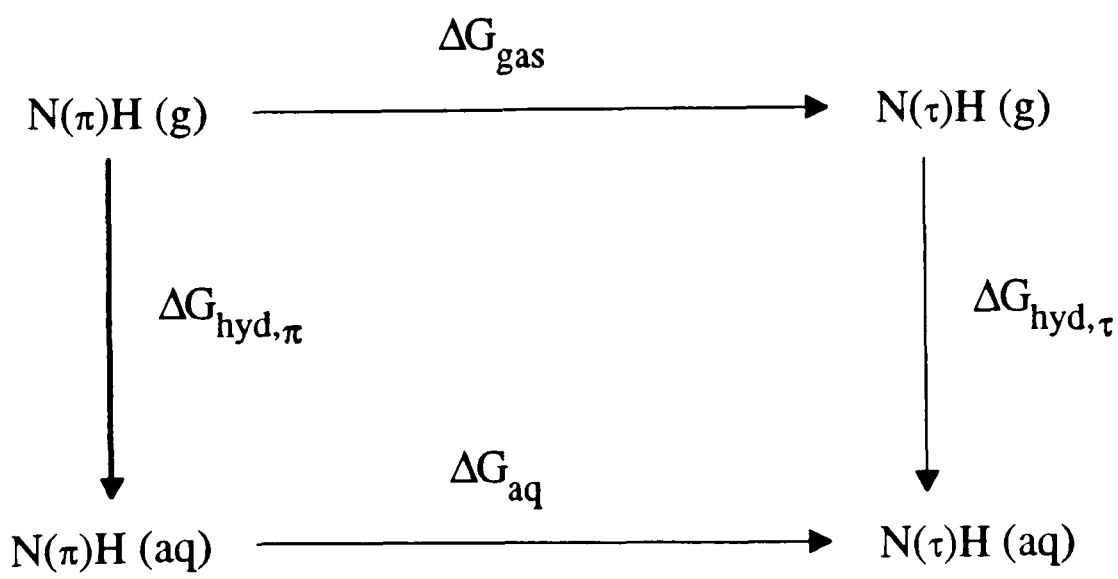
$$\left\langle e^A e^B \right\rangle \neq \left\langle e^A \right\rangle \left\langle e^B \right\rangle \qquad \text{(E4.5)}$$

To enable just the inter-molecular potential to be averaged, an approximation must be made. If one of the energy parts, most likely the intra-molecular Hamiltonian, is effectively constant, then E4.5 becomes an equality and the ensemble average can be split to give

$$\Delta G_{BA,\,inter} = -RT \ln\left\langle \exp(-\Delta\mathcal{H}_{inter}\beta)\right\rangle_A \qquad \text{(E4.6)}$$

and the intra-molecular Hamiltonian can be ignored.

The reason why want to do this is because the intra-molecular terms are large and nearly the same for each system. Thus including them would make the result the small difference between two large numbers and so reduce accuracy. Also excluding bonding terms means that SHAKE can be used to constrain the bond lengths. The inter - molecular part of the energy is incomplete, but this allows simulations of reasonable length to be run. The quantity calculated is thus the difference in free energy of hydration rather than the simple difference in free energy between the two solvated tautomers. The free energy difference in aqueous solution can then be completed by using the thermodynamic cycle in F4.3, where N(x)H stands for the appropriate tautomer



F4.3. The thermodynamic cycle used in the calculation of histamine tautomer ratios by the FEP method.

The gas phase difference in Gibbs free energy, ΔG_{gas}, can be calculated absolutely using *ab initio* calculations. As described above, FEP calculations can then be used to determine the difference in free energies of hydration of the two tautomers,

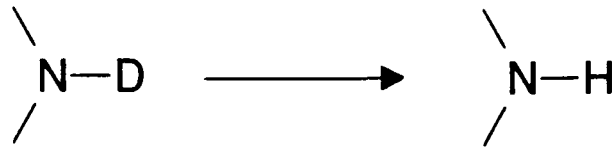
$$\Delta\Delta G_{hyd} = \Delta G_{hyd,\tau} - \Delta G_{hyd,\pi} \qquad \text{(E4.3)}$$

The combination of these two calculations then gives the desired quantity,

$$\Delta G_{aq} = \Delta G_{gas} + \Delta\Delta G_{hyd} \qquad \text{(E4.4).}$$

IV.2.1. FEP Simulation Methodology

As well as problems in parametrisation, the actual methodology that is used for the perturbation is still under discussion. For example a bad method can lead to so-called end point catastrophes. These are inaccuracies at the ends of the simulation produced when an atom "disappears" or "appears". The situation in these simulations is described by the following diagram F4.4.



F4.4. During the course of an FEP simulation, a dummy atom which has no non-bond interactions becomes a hydrogen atom.

At the end points, a dummy atom D, which has no interaction with surrounding atoms, suddenly becomes an interacting atom with parameters that are fractional values of a hydrogen atom. Two problems can arise from this. Firstly while the dummy has no extension in space, there is a distance defined between it and the nitrogen. If this distance is large e.g. the final N-H bond length, then an atom suddenly appears in space which may be occupied by a solvent molecule. This obviously produces a spurious energy. For this reason the N-D bond lengths are started at a value of 0.4 Å and are scaled with λ to finally be the N-H length (Pearlman and Kollman 1989).

While the dummy atom has no non-bonded interactions, it does have a bonded energy to keep it in position. This has lead to some debate as to whether this procedure of growing the bond introduces an extra contribution to the free energy (Pearlman 1991). The important point is that if the bonded contribution is included in a perturbation (which in these simulations it is not) then the two endpoints must be symmetrical e.g. one dummy grows into a hydrogen and one hydrogen shrinks into a dummy. This ensures that this bonding contribution to the free energy cancels.

Another factor that seems to affect the result is the time step used in the integration of the motion equations. There is a trade off between the amount of phase space covered and the coarseness of the sampling. Kollman and Pearlman have found that simulations

with a 1fs time step converge quicker than using 2fs (Pearlman and Kollman 1989). However this difference is not large and the greater sampling makes the larger time step a necessity for large systems.

A device commonly used in molecular dynamics simulations is to remove the centre of mass motion of the solute. This helps to keep the solute in the centre of the box. In fact it was found essential to use this method in the earlier calculations on histamine as there is a problem in AMBER3.1 with the solute co-ordinate resetting. It was found that often the solute would fall out of the box of waters, something that obviously should not happen with periodic boundary conditions. Removing the solute centre of mass motion cured this problem and did not seem to interfere with the result. However this does not seem to be a good idea in FEP runs. Removing this motion forces the solvent to relax around the stopped solute. That breaks the equilibrium and biases the sampling. In order to keep the solute in the box, the motion was removed at the start of each window. This worked, although the solute was often at the edge of the box at the end of a run.

IV.2.2. Results for the Calculation of Histamine Tautomer Ratios Made Previously.

Results for the calculation described above on 4-(5-)methylimidazole, histamine base and histamine monocation have been presented previously (Worth *et al.* 1989, 1990) They are summarised in Table T4.1, and the full results are in the appendix. It is easily seen that the monocation results, as previously mentioned are much less reliable than those for the other systems.

T4.1. Results for the Calculation of K_T for the Tautomeric Shift N(π)H ——> N(τ)H of 4-(5-)Methylimidazole

Molecule	ΔG(g,298) kJ mol^{-1}	K_T(g,298)	$\Delta\Delta G_{hyd}$[a] kJ mol^{-1}	K_T(aq,298)[b]	K_T(Expt)
3-(5-)methylimidazole	-1.37	1.74	0.75 ± 2.93	1.29	0.5 - 1.5
Histamine Base	-2.17	2.48	-2.09 ± 8.09	5.75	4
Histamine monocation	-49.02	3.87×10^8	33.57 ± 13.49	48.64	2.3 - 9

a The error value is discussed below in IV.3.3

b The error is not given for this result. However it is the same as the relevant $\Delta\Delta G_{hyd}$ error manipulated to give an equilibrium constant.

The first thing to be noted is that the error on the FEP result increases with the complexity of the system. What is this error?

IV.2.3. Errors in Calculated Free Energy Differences

The first point to notice is that there is no error in the gas phase *ab initio* calculations. This is not to say that these results are exact. There are errors due to the method itself, due to how far the basis set is from the Hartree-Fock limit, due to the gaussian approximation to the orbital shapes, use of atomic orbitals in a molecular environment, possibly lack of electron correlation etc. However they are exact under the level of approximation used in the calculation. There is no way of knowing this error, except for improving the model and hoping the results converge on a value. This happened with the above calculations at a level of MP2/6-31G*.

IV.2.4. The Size of a Perturbation, the Windowing Method

The error in an FEP simulation comes from two sources. The first is a measure of how close the run is to a true perturbation. Ideally a run will be reversible. Measuring the energy difference between two states A and B should be identical whether over configurations of A or B. If A and B are significantly different this will not be the case and an error results. From this comes the rule of thumb that a window should not be larger than 1 kCal mol^{-1}. This rule also ensures reasonable convergence of the ensemble average energy. This can be seen by expanding the exponential in the FEP formula to give

$$\Delta G_{AB} = -RT \ln\left[1 + \left\langle \Delta H_{AB} / RT \right\rangle_A + \frac{1}{2!}\left\langle (\Delta H_{AB} / RT)^2 \right\rangle_A + \ldots\ldots\right] \qquad \text{(E4.7)}$$

From this if $\Delta H_{AB} > RT$ this series is divergent and so by keeping below 1 kCal mol^{-1}, this divergence is avoided.

The calculations made on the histamine monocation are of the order of 10 kCal mol^{-1}. This size of perturbation therefore has to be made by using a coupling parameter. The Hamiltonian can be written as a function of λ. This function defines the path between two states and the Gibbs free energy now also becomes a function of λ. A simulation starts with a value of $\lambda = 1$. At this point it is entirely in state A. A step size is then used and the endpoint of the perturbation is taken at $\lambda = 1 - \delta\lambda$, where $\delta\lambda$ is chosen so as to keep the perturbation less than the magic 1 kCal mol^{-1}. The calculation just performed has calculated the difference in free energy over the step $\delta\lambda$. λ is now changed to $\lambda - \delta\lambda$. All the terms in the Hamiltonian are scaled accordingly and the calculation now run over these new states. This is continued until $\lambda = 0$ and the system is totally in state B. The difference in free energy between A and B is simply the sum over all these small calculations

$$\Delta G_{BA} = \sum_{\lambda=1}^{0} \Delta G_{\lambda-\delta\lambda,\lambda} \qquad \text{(E4.8)}$$

This method is known as windowing.

For the histamine system, this means starting with the system as one of the tautomers, say N(π)H. In the first perturbation the parameters are flipped between this

tautomer and one in which there is 95% of a proton on the N^{π} and 5% on the N^{τ}. The next window begins with this mutant tautomer and so on until the proton is completely on the N $^{\tau}$ nitrogen.

A value for an error in a simulation is given by a technique ~~given~~ with the grand title of double wide sampling. At each value of λ the energy difference is measured in both directions i.e. $\pm\delta\lambda$. One simulation thus gives ΔG_{AB} and ΔG_{BA}. The difference between these values, the internal error, is due to the lack of reversibility i.e. the fact that the simulation is not a true perturbation.

The second form of error is more important in many ways. For a correct result the whole of phase space must be sampled. This is not possible due to the finite nature of ensemble generated by simulations. An estimate of how close the calculated value to the value covering the whole space must therefore be made. This can be done in different ways. If all the important configurations have been sampled then the calculated value will have converged. This can be checked by seeing how the value varies in subseries taken during the simulation (Straatsma *et al.* 1986). Alternatively, if the value has converged a longer simulation should also give the same result.

Once a converged result is obtained, the statistical error can then be evaluated from a comparison of the results of different simulations (Singh *et al.* 1987). If phase space has bottlenecks that allow the system to fully sample only one small region, the value could be converged but a simulation in a different area of phase space will give a different result.

The error given in the above values is the standard deviation of four independent runs. This gives a measure of the spread of possible values. In addition to this, in this thesis the average of the internal error is added.

IV.3. Effects of Phase Space Sampling on FEP Results.

The larger errors in the histamine results in T4.1 can therefore be traced to the more complex phase space of these flexible molecules compared to the rigid methylimidazole. The extra large error for the monocation must be due to the additional complication of charge. The variance of these four runs was amazing; 24.98, 26.57, 30.54 and 52.17 kJ mol^{-1}. This error must be reduced before any improvements can be made to the method or else any changes in value will be swamped by the error.

IV.3.1. Checking the Simulations for Convergence

The first stages towards improvement were to check the FEP simulations for convergence. It was also thought that the electrostatic problem could be aggravating poor sampling. If the interaction of the solute with molecules at the boundary of the non-bond cutoff is significant, the energy of a particular unit in the ensemble could fluctuate markedly depending on the number of molecules within the cutoff distance. Thus if one

run samples configurations that have on average say 20 waters within the cutoff and the other run samples configurations with 21 waters, the energies could be very different.

In order to test these hypotheses, it was decided simply to run the FEP calculations with various numbers of data collection steps and different cutoffs. It would be hoped that as the ensemble size and cutoff increased, the result should converge. Only the N(π)H to N(τ)H perturbation was studied, as this run gave the result most different from the mean.

The system was restarted, with the histamine monocation N(π)H tautomer in a larger box of waters than previously, to allow for the larger cutoffs. To form the box, waters were included up to 13Å away from the solute, previously this had been 11Å, and there were now 858 waters solvating the solute. This box was minimised in the AMBER module BORN using periodic boundary conditions. The cutoff was 9Å and a constant dielectric of 1 was used. Initially 20 steps of steepest descent were used to remove any bad contacts formed by the solvation process. This was followed by conjugate gradient minimisation.

After minimisation, the system was put into NEWTON for 8ps of dynamics to equilibrate the system, starting at 0.2K and finishing at 298K and at a constant pressure of 1atm. In addition to the conditions applied in BORN, SHAKE was on constraining all bonds and thus a time step of 0.002ps was used. The centre of mass motion was also removed every 50 steps to keep the solute in the centre of the box.

The GIBBS module was then used to carry out the perturbation. As before, the run was divided up into 21 windows. λ started at a value of 1 and decreased in each window by 0.05 until in the last window it had a value of 0. The centre of mass motion was removed only at the beginning of each window, otherwise the conditions were the same as for the dynamics. Four runs were done using different amounts of data collection. Initially 3ps more dynamics were run to ensure equilibration in this module. Then at each value of λ, the system was equilibrated for 1ps before data collection commenced. The results for these runs are given in table T4.2

These results seem to converge nicely as data collection is increased. It can be seen that with 2000 data collection steps the $\Delta\Delta G_{hyd}(298)$ for this perturbation is now very similar to the three others calculated in the original study. These other simulations were thus repeated using these conditions i.e. a large box of waters and 4ps data collection. Firstly the above run was performed in reverse. That is starting at $\lambda = 0$ with the N(τ)H tautomer in the conformation left at the end of the run, and perturbing the ring to give the N(π)H tautomer and $\lambda = 1$ (run 3R). An independent run, starting with the N(τ)H tautomer which was minimised and equilibrated using the above conditions was also set up and run in both directions, starting at $\lambda = 1$ (run 4) and then from this final configuration at $\lambda = 0$ (run 4R). The results for these three runs are given in table T4.3

T4.2. $\Delta\Delta G_{hyd}(298)$ for histamine monocation N(π)H perturbed into N(τ)H, with varying amounts of data collection at each λ. Results are given for the change N(π)H -----> N(τ)H.

Run No.	Number of data collection steps (Time (ps))	$\Delta\Delta G_{hyd.f}(298)$ [a] kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ [a] kJ mol^{-1}	Average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
1	1000 (2)	48.150	49.685	48.918 ± 0.768
2	1500 (3)	42.131	48.698	45.415 ± 3.284
3	2000 (4)	32.218	32.632	32.425 ± 0.207

a The two results are due to the use of double wide sampling. The subscript f denotes the forward value i.e. data collected in the direction of the perturbation while the run with subscript b is collected in the backward direction.

T4.3. Further calculations of $\Delta\Delta G_{hyd}(298)$ using 2000 data collection steps. Results are given for the change N(π)H -----> N(τ)H.

Run no.	$\Delta\Delta G_{hyd.f}(298)$ kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ kJ mol^{-1}	average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
3R	29.662	29.227	29.445 ± 0.218
4	36.083	34.827	35.455 ± 0.628
4R	31.214	33.980	32.597 ± 1.383

These results are all very similar and the internal errors are small. At this point it appeared that the problem was solved, and merely more data collection would provide adequate sampling. Unfortunately, at this point run 1 was repeated using 3000 data collection steps just to check for convergence the result is given below in T4.4.

T4.4. $\Delta\Delta G_{hyd}(298)$ for histamine monocation N(π)H perturbed into N(τ)H, with 3000 steps of data collection at each λ. Result is given for the change N(π)H -----> N(τ)H.

Run No.	Number of data collection steps (Time (ps))	$\Delta\Delta G_{hyd.f}(298)$ kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ kJ mol^{-1}	Average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
5	3000	43.503	47.012	45.258 ± 1.755

Not only has the value moved back to its original high value, but the hysteresis has increased. This result meant that either the ensemble average was still a long way from convergence, even after 3000 steps, or more likely, something else was not being considered.

IV.3.2 Effects of Non-bond cutoff on Phase Space Sampling

The next step was to vary the non-bond cutoff distance and see the effect this has on the calculation. Using run 1 again, starting from the 8ps equilibrated configuration, simulations were run in the same way. Initially 3ps of equilibration were followed by 21 windows with 1ps of equilibration at each λ and 2000 steps (4ps) of data collection. 2000 steps were chosen as opposed to 3000 for reasons of speed and because more results were known at this level of accuracy to yield comparisons.

The results for these simulations are in table T4.5.

T4.5. $\Delta\Delta G_{hyd}(298)$ for histamine monocation N(π)H perturbed into N(τ)H, with varying non-bonded cutoffs. Results are given for the change N(π)H -----> N(τ)H.

Run No.	Cutoff (Å)	$\Delta\Delta G_{hyd.f}(298)$ kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ kJ mol^{-1}	average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
6	8	31.547	32.610	32.079 ± 0.532
7	10	29.066	29.292	29.179 ± 0.113
8	11	34.962	31.108	33.056 ± 1.948

While these results do not seem to be very different from the original run 1 with a 9Å cutoff, it is noticeable both that the result does not converge and also that with the largest cutoff used the internal error increases, indicating that the configurations sampled by successive λ are different. Thus while the cutoff in the region of 8 - 11Å does not seem to have much direct effect on the result, it does affect the configurations seen and this could be significant if interactions out to, say 20Å, are included.

The next calculations attempted to damp out the long range forces by the inclusion of a counter ion. It was argued that if a 1- ion is present in close proximity to the 1+ ion, the effective field will drop faster to zero. This would mean that the long range interaction of the histamine monocation would be effectively modelled by the interaction with the counterion.

IV.3.3. Effects of a Counterion on Phase Space Sampling

A chloride ion was added 2Å from the —N^+H_3 group of the histamine sidechain. The parameters for chloride were not in the standard AMBER parameter set but were taken from Cieplak *et al.* (1987). This system was then solvated in the same size box as run 1 and minimised with the same conditions. Again 8ps of dynamics were used. During this time it was noted that the counterion did not travel more than 6Å from the quaternary nitrogen atom. Thus it was deemed unnecessary to constrain the ions to be close together.

Three simulations were run using various numbers of data collection steps to see if the $\Delta\Delta G_{hyd}$ value converged quicker when the long range forces were damped. The results are below in table T4.6.

T4.6. $\Delta\Delta G_{hyd}(298)$ for histamine monocation N(π)H perturbed into N(τ)H, with a chloride ion near to the N^+H_3 and with varying amounts of data collection at each λ. Results are given for the change N(π)H -----> N(τ)H.

Run No.	Number of data collection steps (Time (ps))	$\Delta\Delta G_{hyd.f}(298)$ kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ kJ mol^{-1}	Average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
9	500 (1)	34.049	34.028	34.099 ± 0.011
10	1000 (2)	35.761	34.823	35.292 ± 0.469
11	2000 (4)	31.326	32.280	31.803 ± 0.477
11R	2000 (4)	37.012	36.355	36.684 ± 0.329

From this it can be seen that the results are much less dependent on the data collection size. This could either mean that eliminating the long range forces prevents the problems seen earlier with sampling, or it could be that the higher electrostatics in the vicinity of the solute mean that there are fewer configurations available to be sampled due to an ordering of the solvent.

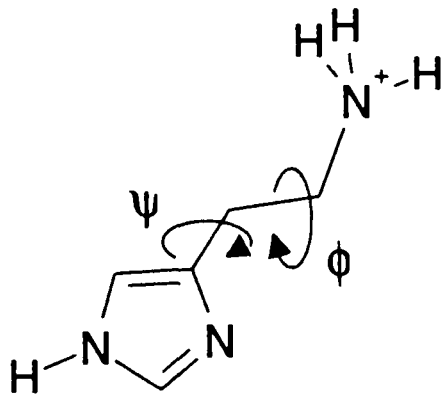
The next stage in this series of results was to add a reaction field that effectively models the long range solute - solvent interactions without the need of a counterion which could bias the available configurations. However before this topic is carried further, the possibility that the sampling may be affected by the ability of histamine to adopt various conformations was investigated. For this the equilibrium in the gas phase was returned to. It showed that in fact the possible conformers have very different energies which were not included in the original result. New parameters for the charges were then obtained to check how well the simulations sample these conformers. As well as discovering some interesting facts about equilibria of flexible molecules, during these tests another reason for the early inconsistent results was discovered almost by accident.

IV.4. Histamine Molecular Conformations and Gas Phase Energies

Previously, only the 'crystal structure' was used to calculate both the atomic charges for the simulations and the gas phase energies. It was not determined whether these were the global minima, or how the equilibrium is affected by the possibility of histamine adopting different conformations. The next step was thus to investigate the effects of the flexibility of the sidechain. If conformations other than the crystal one are important, ΔG_{gas}

would have to be modified accordingly. A method would also be needed that can better model the changing electrostatics of the molecule in each conformer.

A non-linear molecule has 3N-6 internal degrees of freedom for its nucleii. However only torsional motion can produce new conformers - bond and angle motions can only give vibrations from the mean position. In Histamine the ring torsions are fixed and the sidechain hydrogens are fixed relative to the carbon they are bonded to. Just three degrees of torsional freedom exist along the sidechain. The torsion involving the quaternary hydrogens can be ignored as long as they are free to rotate and follow the conformer into its minima. The important degrees of freedom are the torsion angles shown in figure F4.5.



F4.5. The torsion angles in histamine monocation that produce changes in conformation.

With only two torsions to worry about, a potential energy surface is easy to calculate and plot. One torsion is fixed and the other rotated through 360°. At specified degree intervals the energy is calculated. The 'fixed' torsion is then moved on the specified amount and the second torsion again varied. In this way a grid of energies depending on the torsion angles is built up which can then be interpolated to give a contour plot of the potential energy surface.

What method of energy calculation should then be used? Molecular mechanics would perform the task very quickly. However, because of the nature of the potential function, the conformational dependence of the system is assumed and so the results will be biased. *Ab initio* methods are thus preferred. Unfortunately full *ab initio* is impossible. A grid must contain at least 36 points to give any sort of reliability and the more points the better. This number of calculations, each of which must include geometry optimisation to include relaxation of the rest of the system, is clearly unfeasible. Semi - empirical *ab initio* methods are the best compromise.

Earlier studies of this subject have in fact been done in this way. One of the reasons why it was deemed necessary to look closer at this part of the cycle was the fact that these potential energy surfaces show that there are multiple minima in this torsion space.

MOPAC5 is set up to easily perform the exact calculation desired using the AM1 Hamiltonian. This calculation method was used in the original calculations of K_T(histamine,aq) and it was decided that it performed poorly. However on re-evaluating this data it was discovered that a mathematical error had occurred and in fact the AM1 Hamiltonian result was of the same order of magnitude as the full *ab initio*. This was a relief as it meant that the AM1 wavefunctions are reasonable for this system.

Two contour plots - one for each tautomer - were thus calculated. The grid points were at 36° intervals i.e. there were 122 grid points and the geometry, excluding the torsions φ and ψ, was optimised at each point. The resulting surfaces are shown in plots P4.1 and P4.2. Both the surface and a contour plot of this surface are shown.

These potential energy surfaces show the heat of formation of the tautomer conformations in the torsional space due to the angles φ and ψ. From these surfaces it must be decided what energy each tautomer has, taking into account all the different conformers. This enthalpy must then be altered to a free energy.

Each minimum is a stable conformation and can be thought of as different species which are in equilibrium. Free Energy is a state function and so

$$G_m^o = \sum_i n_i G_{m,i}^o \qquad \text{(E4.9)}$$

where n_i is the number of moles of species i with standard molar Gibbs free energy $G_{m,i}^o$. It is thus necessary to know how many of each species is present as well as their Gibbs free energy in order to know the Gibbs free energy of the tautomer.

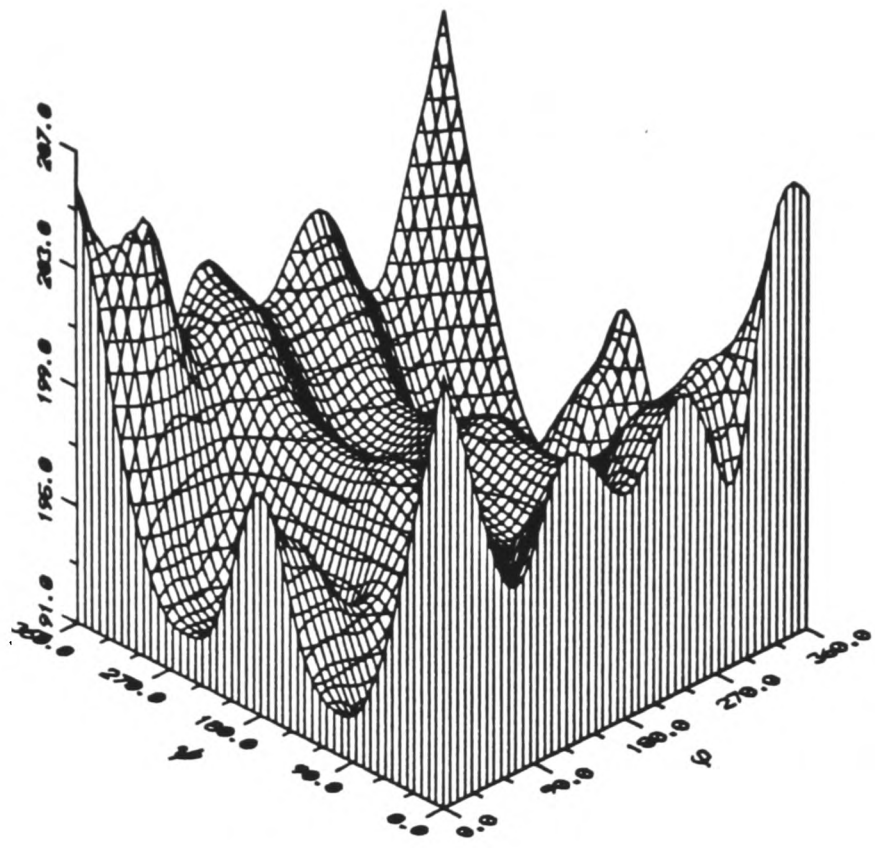
From statistical mechanics it is known that an equilibrium between two species, A —> B is governed not only by the energy difference between the species but also by the ratio of partition functions

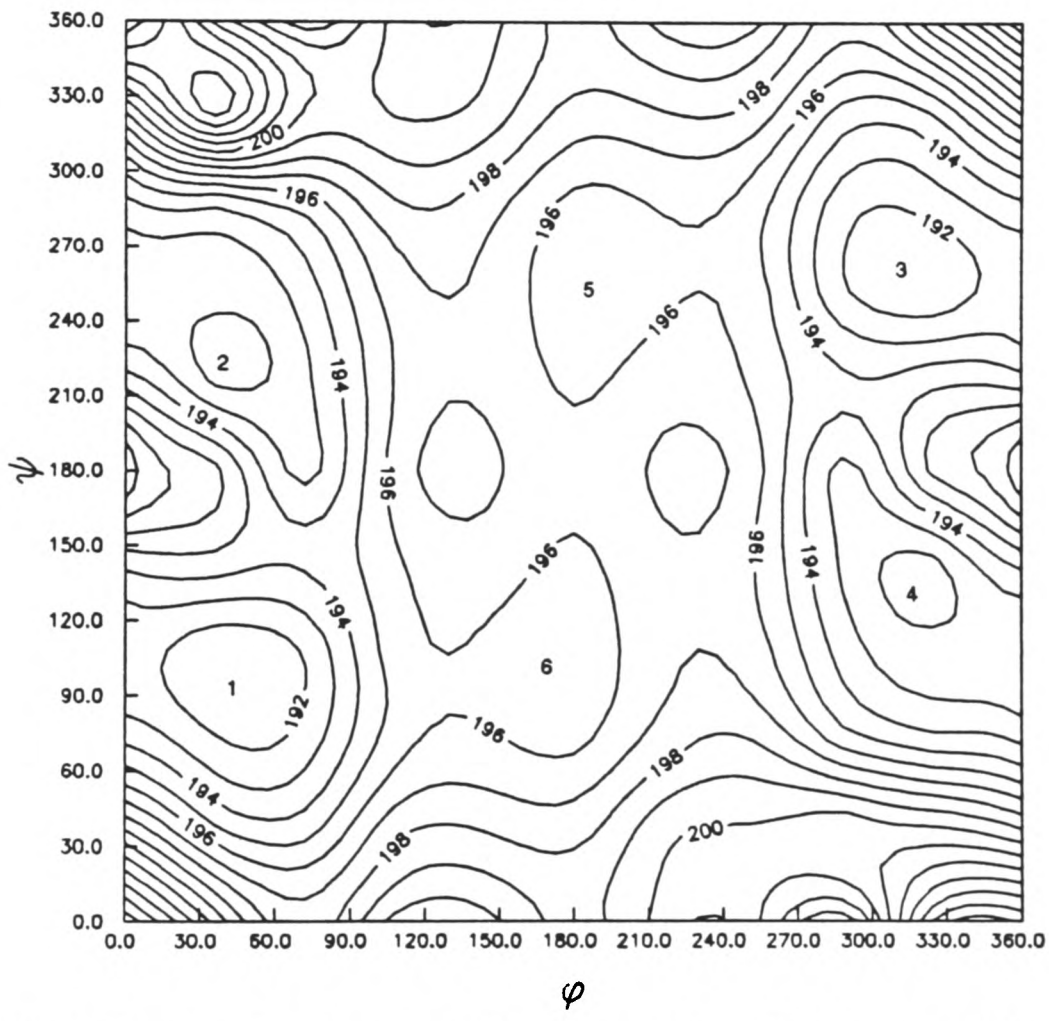
$$K = \frac{q_B}{q_A} \exp(-\beta \Delta E_0) \qquad \text{(E4.10)}$$

where ΔE_0 is the difference in ground state energies between A and B and q_A, q_B their molecular partition functions. The ground state energies are just the values shown on the potential surface. Now the partition function must be found. This is a sum over all the molecular energy levels available at the system temperature

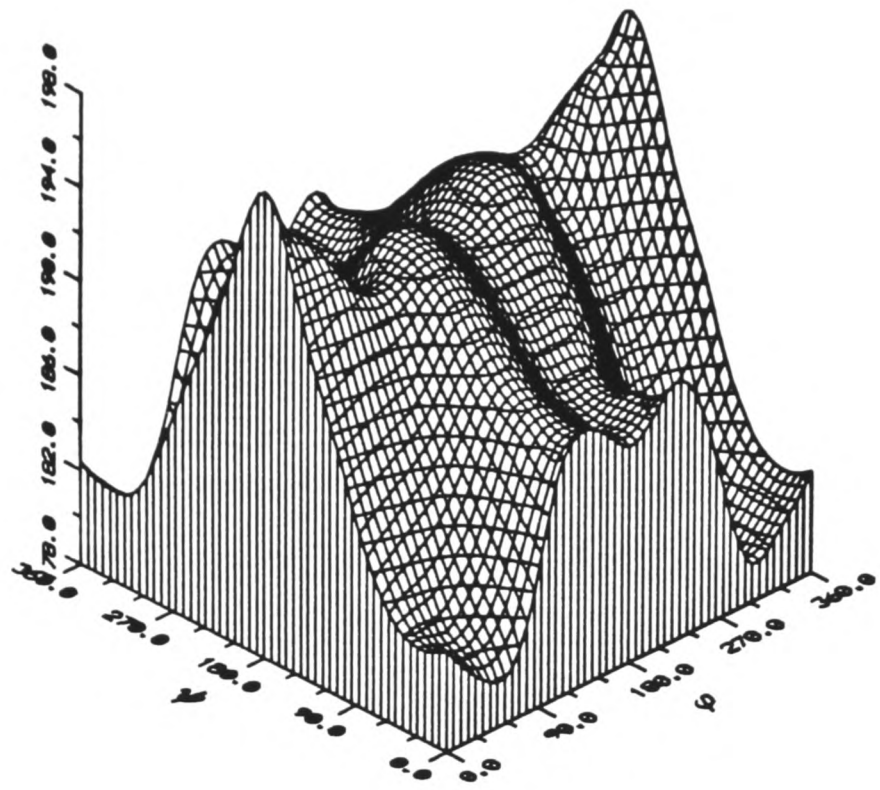
$$q_A = \sum_{i,A} \exp(-\varepsilon_i \beta) \qquad \text{(E4.11)}$$

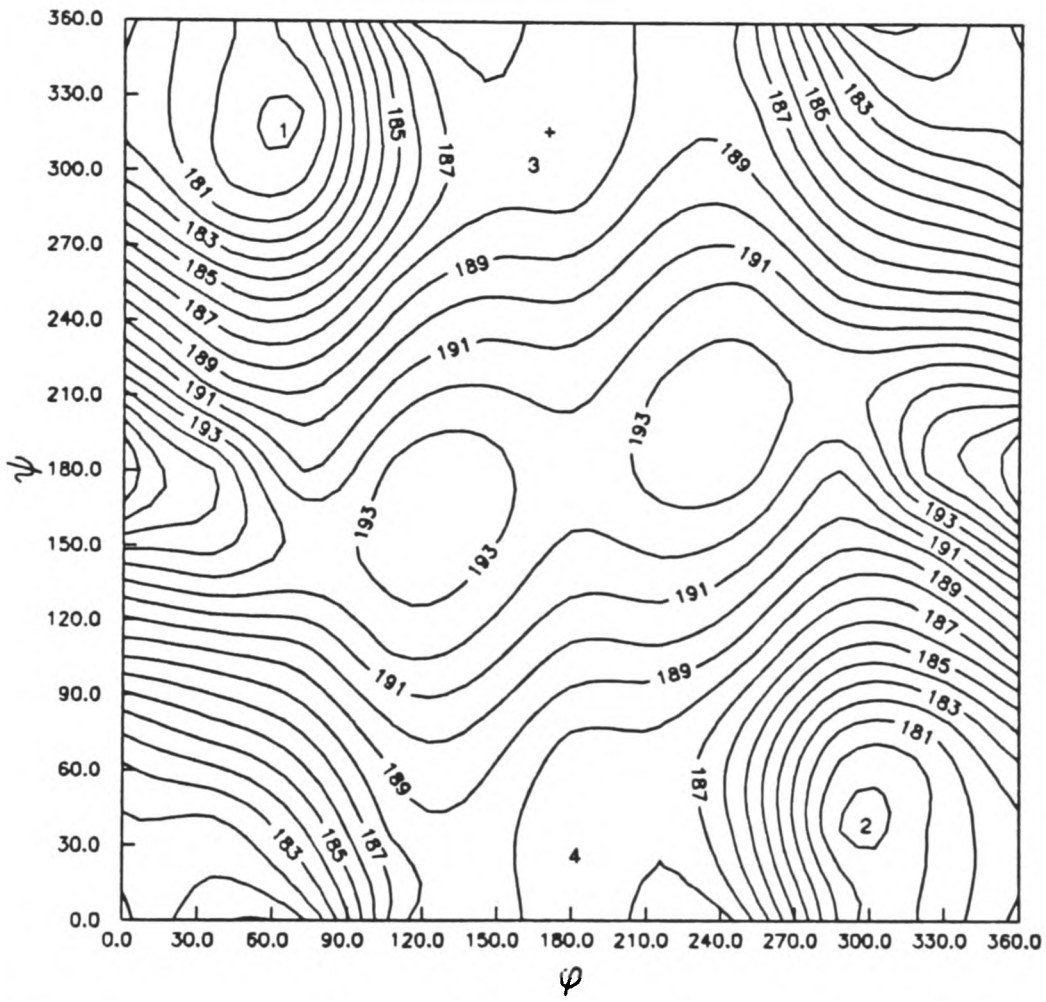
As this is related to all the thermodynamic functions once it are known, the corrections necessary to change the energies to free energies are also known.





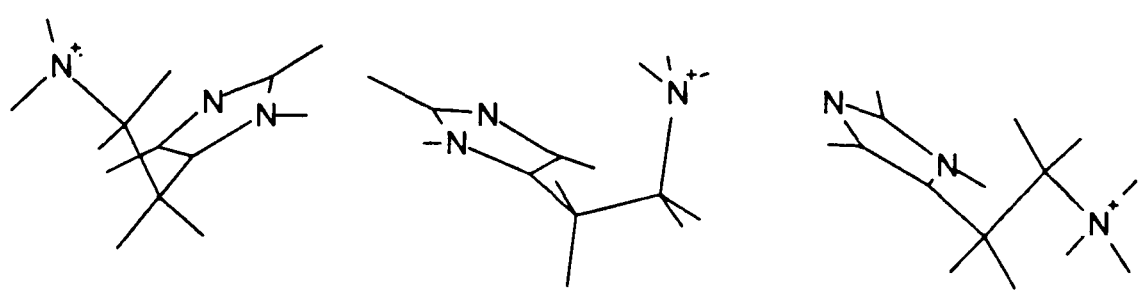
P4.1. Energy Surface and Corresponding Contour Plot for the Variation in Energy of Histamine Monocation N(π)H With the Change in Torsion Angles φ and ψ. The minima are numbered.





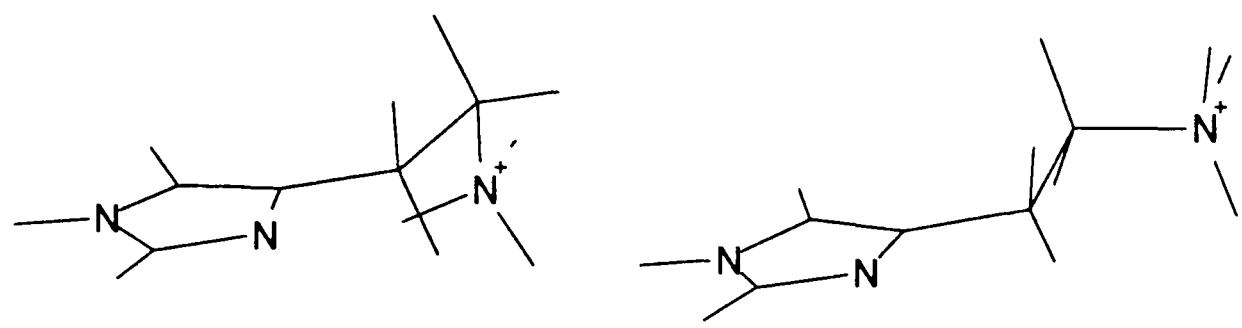
P4.2. Energy Surface and Corresponding Contour Plot for the Variation in Energy of Histamine Monocation N(τ)H With the Change in Torsion Angles φ and ψ. The minima are numbered and the crystal structure geometry is denoted by a cross.

What are the species involved? The plots show that N(π)H has four deep minima (1 - 4) and two flat, high energy minima (5 and 6). N(τ)H has two deep minima (1 and 2) and again two minima on the high plateau (3 and 4). The symmetry of the system can also be seen. The imidazole ring lies in a plane of reflection and so each minima has a mirror image. The crystal geometry lies in minima 3 of the N(τ)H plot. The corresponding geometry used in the N(π)H calculations lies in minima 5. For the N(π)H tautomer this is ≅5 kCal mol^{-1} above the lowest minima, while for the N(τ)H this height is ≅10 kCal mol^{-1}. Of the six minima for N(π)H, the molecular symmetry means that only three are unique. Similarly, for N(τ)H, only two minima are unique. Thus only five conformations need to be studied in depth. Pictures of these minima conformers are shown in F4.6 and F4.7.



(a) (b) (c)

F4.6. The various histamine N(π)H conformers in the minima found by an AM1 conformational search. (a) minimum 1 (b) minimum 2 (c) minimum 5. The conformers of minimum 3, 4 and 6 are found by inverting (a), (b) and (c) with respect to the plane of the ring.



(a) (b)

F4.7. The various histamine N(τ)H conformers in the minima found by an AM1 conformational search. (a) minimum 1 (b) minimum 3. The conformers of minimum 2 and 4 are found by inverting (a) and (b) with respect to the plane of the ring.

These conformers were taken and optimised using AM1 and the *ab initio* program GAUSSIAN88 with a basis set of 3-21G. 6-31G* single point energies were then calculated using these 3-21G geometries. The geometries and energies are given in table T4.7.

T4.7. Geometries and Energies of the Lowest Energy Conformers of Histamine Monocation Conformers From Mopac AM1 Potential Surfaces, Calculated Using Mopac and GAUSSIAN88

Histamine monocation N(π)H

minima[a]	optimised ϕ 3-21G (AM1)	optimised φ 3-21G (AM1)	AM1 $\Delta H_f(0)$ kCal mol^{-1}	RHF/3-21G//b a.u.'s	RHF/6-31G*//b a.u.'s
1	56.34 (50.67)	80.25 (96.87)	190.75	-356.28855	-358.27731
2	55.71 (49.27)	208.57 (213.64)	191.23	-356.28653	-358.27601
5[c]	181.73(186.56)	279.49(258.13)	195.265	-356.27894	-358.26993[d]

Histamine monocation N(τ)H

minima[a]	optimised ϕ 3-21G (AM1)	optimised φ 3-21G (AM1)	AM1 $\Delta H_f(0)$ kCal mol^{-1}	RHF/3-21G//b a.u.'s	RHF/6-31G*//b a.u's
1	58.86 (61.57)	319.63 (324.87)	178.54	-356.32720	-358.30586
3[c]	168.73(180.28)	318.77(326.21)	187.15	-356.29942	-358.28819[d]

a Minima numbered as contour plots P(4.1) and P(4.2).

b Energy calculated on RHF/3-21G optimised geometry.

c These structures were started from the earlier 6-31G optimised structures based on the N(τ)H crystal structure, with the ring proton appropriately moved. The crystal angles are ϕ = 170.22° and φ = 317.05°.

d These energies were actually calculated on the 6-31G optimised geometry, for which the relevant dihedrals vary by about a degree from the 3-21G angles.

A normal mode analysis was then performed using AM1 on the AM1 optimised structure. This works by calculating the second derivative matrix, called the Hessian, for displacements of the atoms. That then produces the force constants for the bonds and when diagonalised, produces the normal modes of the vibrations. In the surfaces above, the internal modes of rotation mean that these modes of freedom are not harmonic functions.

The translational and rotational energy levels are also calculated. All the levels are now summed to give the partition function at a particular temperature. From this thermodynamic data could be again calculated. These results are shown in table T4.8.

From these results it can be seen that the extended crystal conformation has many more accessible vibrational levels than the curled up absolute minima conformations. Presumably the intra molecular bonding that occurs in these structures makes them more rigid, pushing up the vibrational energy and so widening the levels. Also the slightly higher moment of inertia of the extended sidechain gives closer rotational levels.

Due to these effects the numbers of molecules present in one mole of histamine in the gas phase at 298K are calculated to be as shown in T4.9.

T4.8. Partition function and thermodynamic data from an AM1 vibrational mode analysis of histamine monocation tautomers.

Histamine monocation N(π)H

	Minima 1	Minima 2	Minima 5
$q_{vibrational}$	71.36	65.04	257.4
$q_{rotational}$	3.26×10^5	3.24×10^5	3.57×10^5
$q_{translational}$	1.15×10^{27}	1.15×10^{27}	1.15×10^{27}
q_{total}	2.68×10^{34}	2.42×10^{34}	10.57×10^{34}
Zero Point Energy	420.877	420.835	419.852
H(298,vib)	12649.475	12648.805	13644.606
H(298,rot)	3716.668	3716.668	3716.668
H(298,trans)	6194.45	6194.441	6194.45
H(298,total)	22560.597	22559.919	23555.723
S(298,vib)	77.931	77.161	91.939
S(298,rot)	118.018	117.976	118.788
S(298,trans)	167.523	167.523	167.523
S(298,total)	363.477	362.661	378.25

Histamine monocation N(τ)H

	Minima 1	Minima 3
$q_{vibrational}$	59.99	229.7
$q_{rotational}$	3.24×10^5	3.50×10^5
$q_{translational}$	1.15×10^{27}	1.15×10^{27}
q_{total}	2.24×10^{34}	9.24×10^{34}
Zero Point Energy	421.224	420.923
H(298,vib)	12475.186	13345.546
H(298,rot)	3716.668	3716.668
H(298,trans)	6194.45	16854.566
H(298,total)	22386.3016	6152.873
S(298,vib)	75.622	89.989
S(298,rot)	117.972	118.608
S(298,trans)	167.523	167.523
S(298,total)	361.121	376.121

T4.9. The number of molecules of each histamine conformer and the resulting free energy of the tautomers. Energies are in kJ mol^{-1}.

Histamine monocation N(π)H.

Minima	1 and 3	2 and 4	5
q_i/q_1	1	0.906	3.625
$\exp(-\beta\Delta E_0)$	1	0.827	0.163
n_i	0.428	0.320	0.252
$\Delta H(0 \longrightarrow 298K) - TS$	-85.81	-85.5674	-89.006
$n_i \Delta G_i(0 \longrightarrow 298K)$	-36.706	-27.359	-22.430

$\therefore \Delta G(N(\pi)H, 0 \longrightarrow 298K) = -86.538$ kJ mol^{-1}.

Histamine monocation N(τ)H.

Minima	1 and 2	3 and 4
q_i/q_1	1	3.767
$\exp(-\beta\Delta E_0)$	1	0.031
n_i	0.896	0.104
$\Delta H(0 \longrightarrow 298K) - TS$	-85.282	-88.655
$n_i \Delta G_i(0 \longrightarrow 298K)$	-76.404	-9.23

$\therefore \Delta G(N(\tau)H, 0 \longrightarrow 298K) = -85.634$ kJ mol^{-1}.

Therefore the extended conformation is present in appreciable amounts in the N(π)H system and in small amounts in the N(τ)H system, despite it lying a long way above the absolute minima. However it may be noticed that although the high energy conformations are present due to their high partition functions, this same property lowers these conformers' free energies so that in fact the free energy is almost identical to just that of the lowest energy state.

The data in table T4.9 can now be used to convert the energies in T4.7 to find the difference in free energy of the tautomers in the gas phase. For the GAUSSIAN energies E4.12 was used to find the free energy of each tautomer and the difference taken. The MOPAC results are treated differently and E4.13 was used.

$$G(\text{tautomer}) = \sum_i n_i \left(E_i + ZPE_i + \Delta G_i(0 \rightarrow 298K) \right) \qquad \text{(E4.12)}$$

$$G(\text{tautomer}) = \sum_i n_i (\Delta H_{f,i} - TS_i(298)) \qquad \text{(E4.13)}$$

The results are given in T4.10

T4.10. The difference in free energy between histamine monocation tautomers in the gas phase and their equilibrium constant.

Basis set	ΔG (g,298) kJ mol^{-1}	K_T (g,298)
AM1	-51.51	1 x 10^9
3-21G	-100.41	4 x 10^{17}
6-31G*	-74.58	1 x 10^{13}

IV.5. Multi-conformational Atom Centred Charges for the Histamine Monocation

It appeared that using a counterion to damp out long range forces the FEP result had converged. Combining the value of $\Delta\Delta G_{hyd}$ in T4.6 with the 6-31G* ΔG_{gas} in T4.10 we now have a value for ΔG_{aq} of -40.336 ± 3.854 and a K_T of 12 x 10^7. This is even further from the experimental value than before.

One error might be that the charges used in the solution simulations bias the conformations sampled so that only the extended conformers are seen. If the other species are important in solution, they could drastically change the results. Charges that model all these conformations are thus needed.

In the AMBER force field, the electrostatic interaction between two atoms is modelled with a coulomb potential where the relevant charges are atom centred partial charges calculated from fitting to *ab initio* molecular electrostatic potentials (MEP's) on random points on surfaces beyond the van der Waals surface of the molecule. In all the above simulations the atom centred charges had been fitted to the MEP from a 6-31G wavefunction calculated on a 6-31G optimised structure. This structure was, for both tautomers, based on that found in the histamine hydrochloride crystal structure (Pratt et al 1974). This is the monocation in the N(τ)H form. It was argued that this would be the most stable conformer in the gas phase for all the histamine species as the proton on the ring does not interfere with the sidechain in this position.

The charges calculated at one geometry fit the dipole moment of that geometry very well and so are thought to model the electric properties of the molecule adequately. Histamine possesses the two internal rotations, ϕ and ψ shown in F4.5. In a different conformation from that in which the charges are calculated, the charges may no longer be able to reproduce the relevant dipole. This has been found to be a problem with simulations of torsionally flexible molecules (Essex *et al.* 1992) In the simulations of propanol for example it was found that the charges calculated on just one conformation prevented the molecule from accessing all its conformations. On calculating an FEP for the change propanol to ethanol, some of the configurations that were important for the ensemble averages were not seen and the result was a poor agreement with experiment.

A method has been developed to try and take this conformational flexibility into account (Reynolds *et al.* 1992). The MEPs are calculated for all the important conformations of the molecule i.e. those that would be present under the conditions of the simulation. A weighted average of all these MEPs is then taken and the charges fitted to this. So far it has been found that these multi-configuration charges reproduce the dipole at all the included geometries better than any single set.

The minima conformations have already been optimised to 3-21G. Wavefunctions at 6-31G* have then been calculated on these structures. Charges obtained by fitting to MEP's from these wavefunctions are thought to be the best that can be practically obtained. This was therefore done. The results are given below in table T4.11.

Table T4.12 shows the dipole moments calculated with these charge sets at the various geometries. Interestingly none of the charge sets reproduce the quantum mechanical dipole for the extended conformers. This is presumably a problem due to the fact that quantum mechanics predicts that there is almost no dipole moment, hence any small deviations from the exact electronic distribution will give a large change in dipole.

It can be seen that the weighted charge sets are indeed able to reproduce the dipole moments at all the other geometries better than any one *set al.*one. Hence if the molecule is seeing all these conformers, these charges should be the ones used.

IV.5.1. Simulations Using The New Charges

Using the new conformationally adjusted charges, perturbations were then carried out as before, using 21 windows. Again runs with various amounts of data collection per window were run to see both how the new charges affect the result and if the results converge without 40ps of data collection being necessary. These results are in T4.13. Runs 12 - 15 all start with the N(π)H tautomer at λ=1, while run 16 starts with N(τ)H.

The result for shortest run, number 12, is significantly higher than the others. The value appears to have converged by the time 8ps of equilibration and 8ps of ~~dynamics~~ data collection are used at each window. However this converged result is about 8 kJ mol^{-1} higher than using the old charges (the results in T4.2 and T4.3). The final value, taken from the average of the 4000 / 4000 runs is therefore $\Delta\Delta G_{hyd} = 41.429 \pm 6.526$

When this is combined with the gas phase results obtained above, this gives values for the difference in free energy of the tautomers in solution shown in T4.14. These values are still a long way from the experimental value.

T4.11. Charges calculated for the various conformers of histamine and fitted to the boltzmann averaged MEP. The geom number shows which minima conformer the calculation was on. The ratio in the weighted MEP column shows the weighting of the various geometries. The final column in both tables is the 6-31G charges used in the original simulations. All MEP's were calculated with a RHF wavefunction.

Histamine monocation N(π)H

	6-31G*//a Geom 1	6-31G*//a Geom 2	6-31G*//b Geom 5	6-31G*//3-21G Weighted MEP 1: 2: 5 0.43: 0.32: 0.25	6-31G//b Geom 5
HN11	0.361	0.376	0.321	0.351	0.345
N	-0.593	-0.587	-0.379	-0.496	-0.459
HN12	0.361	0.376	0.321	0.354	0.345
HN13	0.361	0.376	0.321	0.337	0.345
CA	0.399	0.088	0.344	0.260	0.354
HA1	0.026	0.108	0.034	0.066	0.039
HA2	0.022	0.139	0.017	0.075	0.018
CB	-0.458	-0.497	-0.363	-0.502	-0.426
HB1	0.190	0.208	0.153	0.181	0.184
HB2	0.146	0.202	0.120	0.208	0.138
CG	0.226	0.270	0.096	0.241	0.093
ND2	-0.523	-0.583	-0.472	-0.501	-0.526
HD2	0.395	0.414	0.382	0.396	0.390
CE2	0.287	0.354	0.262	0.266	0.378
HE2	0.153	0.138	0.149	0.154	0.128
CD1	0.057	0.050	0.071	-0.056	0.194
HD1	0.129	0.151	0.146	0.162	0.116
NE1	-0.541	-0.583	-0.525	-0.496	-0.657

T4.11. cont.

Histamine monocation N(τ)H

	6-31G*//a Geom 1	6-31G*//b Geom 3	6-31G*//3-21G Weighted MEP 1: 3 0.90: 0.10	6-31G//b Geom 3
HN11	0.295	0.325	0.350	0.363
N	-0.414	-0.425	-0.551	-0.568
HN12	0.295	0.325	0.358	0.363
HN13	0.295	0.325	0.331	0.363
CA	0.302	0.551	0.494	0.567
HA1	0.057	-0.026	0.018	-0.021
HA2	0.028	-0.028	-0.007	-0.008
CB	-0.448	-0.615	-0.634	-0.632
HB1	0.168	0.178	0.206	0.199
HB2	0.166	0.159	0.191	0.162
CG	0.357	0.522	0.491	0.497
ND2	-0.389	-0.575	-0.619	-0.653
CE2	0.001	0.146	0.198	0.266
HE2	0.219	0.178	0.178	0.146
CD1	-0.388	-0.410	-0.370	-0.339
HD1	0.269	0.258	0.264	0.247
NE1	-0.167	-0.256	-0.271	-0.326
HE1	0.354	0.368	0.374	0.375

a 3-21G optimised geometry

b 6-31G optimised geometry

T4.12. The dipole moment produced along the Cartesian axes by the various charge sets in T4.8 at the various conformer geometries.

Histamine monocation N(π)H

Charges		Geom 1	Geom 2	Geom 5
QM[a]	x	9.716	-9.421	0.000
	y	-0.566	1.335	0.001
	z	-2.693	3.517	0.013
Geom 1	x	9.709	-9.357	-13.769
	y	-0.547	1.092	-2.939
	z	-2.753	3.374	0.430
Geom 2	x	9.893	-9.510	-13.771
	y	-0.379	1.308	-2.809
	z	-2.369	3.549	0.566
Geom 5	x	9.976	-9.659	-14.807
	y	0.463	0.410	-2.999
	z	-2.817	3.913	-0.084
Weighted	x	9.854	-9.459	-14.301
	y	-0.559	1.206	-3.103
	z	-2.798	3.583	0.157

Histamine monocation N(τ)H

Charges		Geom 1	Geom 3
QM[a]	x	3.740	0.000
	y	-1.455	0.067
	z	0.394	0.311
Geom 1	x	3.771	-6.849
	y	-1.448	1.286
	z	0.430	-1.210
Geom 3	x	4.708	-9.043
	y	-2.230	1.303
	z	0.393	-0.939
Weighted	x	3.813	-7.871
	y	-1.466	1.523
	z	0.444	-0.860

a QM indicates that these are the dipole moments from the quantum mechanical calculation.

T4.13. $\Delta\Delta G_{hyd}(298)$ for the Histamine monocation N(π)H perturbed into N(τ)H, using Conformationally Weighted Charges, with varying amounts of data collection and equilibration at each λ. Results given for N(π)H -----> N(τ)H.

Run No.	Number of steps equilibration /data collection	$\Delta\Delta G_{hyd.f}(298)$ kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ kJ mol^{-1}	average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
12	500 / 1000	49.823	49.380	49.601 ± 0.222
13	1000 / 1000	38.514	41.438	39.978 ± 1.464
13R	1000 / 1000	43.271	43.999	43.635 ± 0.364
14	2000 / 2000	40.338	38.100	39.219 ± 1.119
15	4000 / 4000	38.878	38.409	38.643 ± 0.235
15R	4000 /4000	48.426	38.677	43.551 ± 4.875
16	4000 /4000	38.945	45.932	42.438 ± 3.494
16R	4000 / 4000	40.940	41.225	41.083 ± 0.143

T4.14. The difference in free energy and the equilibrium constant between histamine monocation tautomers in aqueous solution.

Basis set	ΔG (g,298) kJ mol^{-1}	K_T (g,298)
AM1	-10.081 ± 6.526	58.5
3-21G	-58.981 ± 6.526	2 x 10^{10}
6-31G*	-33.151 ± 6.526	6 x 10^{6}

To check that this difference is not just an artefact of using a larger basis set (the original charges were 6-31G, these are 6-31G*), two further runs were made using the 6-31G* charges calculated on the extended geometries i.e. the charges for geometries 5 and 3 in T4.11. These results are in T4.15. Run 17 starts with the N(π)H tautomer at λ=1, while run 18 starts with N(τ)H.

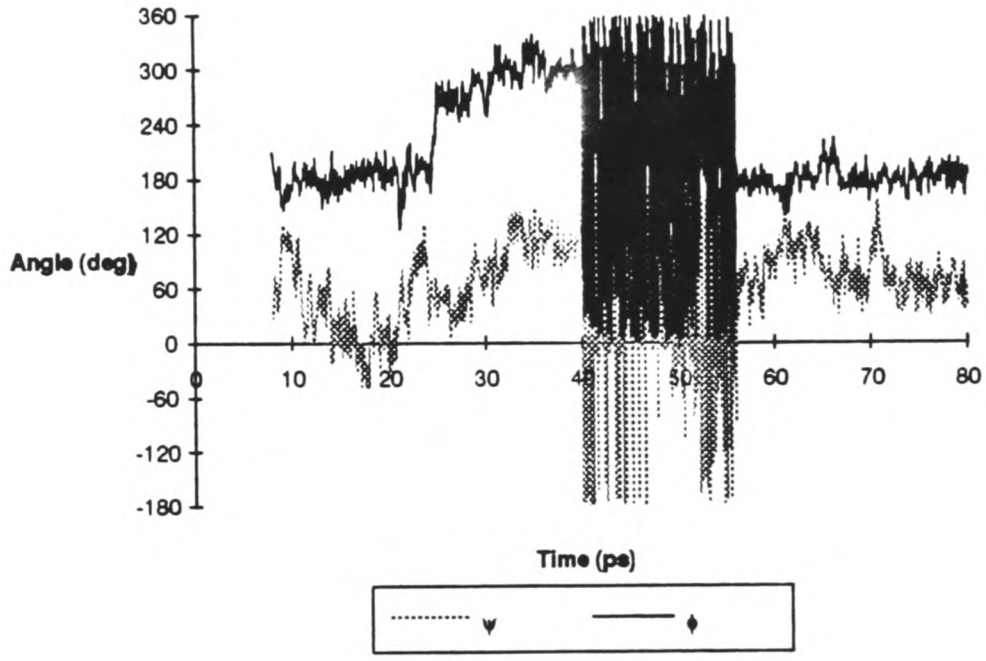
T4.15. $\Delta\Delta G_{hyd}(298)$ for the Histamine monocation N(π)H perturbed into N(τ)H, using 6-31G* charges. Results given for N(π)H -----> N(τ)H.

Run No.	Number of steps equilibration /data collection	$\Delta\Delta G_{hyd.f}(298)$ kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ kJ mol^{-1}	average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
17	4000 /4000	35.811	36.451	36.131 ± 0.320
18	4000 / 4000	37.275	33.288	35.282 ± 1.994

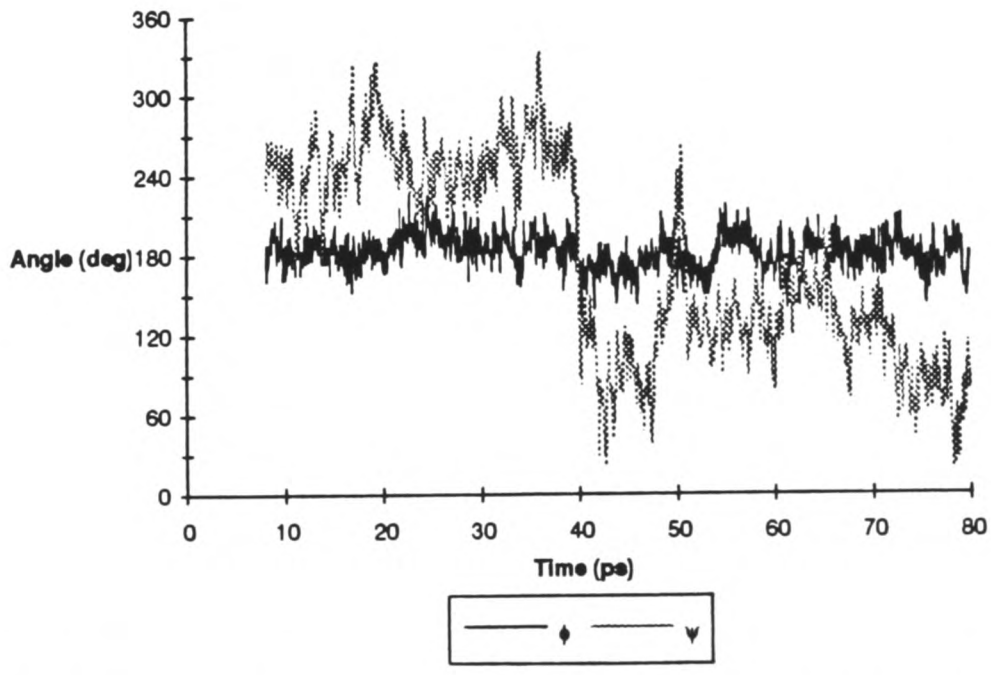
These are fortunately very similar to the runs made using the 6-31G charges. Even better, they appear to have converged on the result given by the runs in T4.2 and 4.3 using 4ps of data collection.

IV.6. Histamine Conformers in Solution Simulations

The question to be asked now was; exactly which conformers are being seen in solution? To see, 80ps of dynamics were run on each tautomer. Co-ordinates of the histamine were dumped every 16 steps (0.032ps). The variation of the torsion angles ϕ and φ over this time were then calculated. This variation is plotted in P4.3 and P4.4.



P4.3. The variation of histamine monocation N(π)H torsions with time.



P4.4. The variation of histamine monocation N(τ)H torsions with time.

These plots show that conformational changes seem to occur on roughly 20ps cycles. The N(π)H tautomer started in minima 6, then moved into minima 4 after 20ps. After 40ps a large fluctuation between the conformers of minim 2 and 4 occurs for around 15ps before the molecule returns to minima 6. In contrast, the N(τ)H tautomer only saw the extended conformation during the whole time, with ϕ staying around 180°. The only change was a move from minima 3 to minima 4 after 20ps.

This type of behaviour is a serious problem for searching phase space. Even after 80ps the N(π)H tautomer has not sampled the minima 1 and 3 at all and N(τ)H does not at any time go into minima 1 or 2.

Using proton NMR, Ganellin *et al.* (1973) deduced that only 45% of histamine in solution is in the trans form i.e. minima 3 for N(π)H and minima 5 for N(τ)H. Combined with the known ratio of the tautomers, a minimum of 20% of the N(τ)H tautomer must be in the curled up, gauche form. This indicates that in fact all the species are seen in solution and so the simulation is either incorrectly sampling phase space or, more likely, simply not running for long enough.

IV.6.1. Gas Phase Energies Revisited

In Chapter I it was explained that the calculation of ΔG_{aq} had been split up into the two parts to avoid calculating the large intra-molecular potentials. The mistake that had then been made was to take the ΔG_g in equation E4.4 to be the total free energy difference between the tautomers in the gas phase. After these results were obtained, it was realised that this in fact was not the case.

One clue is that for E4.4 to be correct all the free energies must be evaluated under the same conditions, or else extra terms are needed. Under 1atm pressure at 298K, pure histamine is unlikely to be a gas. The next clue is that the FEP result analysing only the interaction between the solvent and the solute is just the difference in solvation energies i.e. by definition $\Delta\Delta G_{hyd}$. If $\Delta G_{aq,298}$ is split up into intra and inter terms, and the latter is $\Delta\Delta G_{hyd}$, then the gas phase energies should in fact only be the intra-molecular energy missed in the aqueous phase simulations.

It might seem strange that a conformer plays no part in the calculation, even if it is predominant in the gas phase equilibrium. A physical picture is derived by imagining a box of gas of substance A, which has two conformers A1 and A2. This is an equilibrium mixture and has a Gibbs free energy of $n_{A1}G(A1)+n_{A2}G(A2)$. If this gas is then placed over water and only A2 molecules dissolve, the gas phase equilibrium restores itself by A1 converting into A2, which can then again dissolve. The ΔG_{hyd} measured is therefore entirely due to the solvation of A2 and A1 is not involved in the process.

The reason for the error in the calculation of $\Delta G_{T,aq}$ was thus due to the fact that different conformations are seen in the gas phase than in solution. Only the extended

N(τ)H conformer is seen and so the gas phase free energy required is just for this conformer. Similarly, $G_{gas}(N(\pi)H)$ should be calculated using the weighting of conformers as seen in solution. For the simulations run above, this appears to be around 50% for minima 2 and 50% for minima 5. Using these values, the corrected gas phase free energies required are given in T4.16.

T4.16. The difference in free energy and the equilibrium constant, between histamine monocation tautomers in the gas phase taken over the conformers seen in a solution simulation.

Basis set	ΔG (g,298) kJ mol^{-1}	K_T (g,298)
AM1	-24.25	2×10^5
3-21G	-42.29	3×10^8
6-31G*	-38.41	5×10^7

T4.17. The difference in free energy and the equilibrium constant, between histamine monocation tautomers in aqueous solution, using the gas phase energies in T4.16.

Basis set	ΔG (aq,298) kJ mol^{-1}	K_T (aq,298)
AM1	$+17.179 \pm 6.526$	1×10^{-3}
3-21G	-0.861 ± 6.526	1.4
6-31G*	$+3.019 \pm 6.526$	0.3

These answers are much closer to those desired. One last possibility must be looked at. Having made the calculations predominately over the extended conformations, it might be possible that the charges should have been those calculated on these geometries. Runs 17 and 18 can be taken as representative of calculations made with these charges, even though there are only two values so the error estimation will be poor. These values give $\Delta\Delta G_{hyd} = 35.707 \pm 4.187$ kJ mol^{-1}. Table T4.18 thus gives the aqueous solution equilibria data combining this result with the gas phase energies in T4.16.

These values are better still. All the methods now give a result that agrees with the experimental value within its error bounds, with the 6-31G* result being almost identical to the experimental value. These results demonstrate both the power and problems of FEP. Despite bad sampling, and despite the incorrect treatment of the long range forces a reasonable result has been achieved. Sometimes it seems as if all you have to do is run enough simulations and combinations of conditions to achieve the desired result, but in

fact this last result is now using a justifiable approach to the problems and a good value has resulted.

T4.18. The difference in free energy and the equilibrium constant, between histamine monocation tautomers in the gas phase taken over the conformers seen in a solution simulation.

Basis set	ΔG (g,298) kJ mol^{-1}	K_T (g,298)
AM1	10.457 ± 4.187	1 x 10^{-2}
3-21G	-6.583 ± 4.187	14.3
6-31G*	-2.703 ± 4.187	3.0

It would still be possible to complain that the methodology of splitting the free energy up into intra and inter components is incorrect. It could also be in error to take the gas phase correction for just the most common species seen. A more accurate approach might be to include the full potential, despite the fears aired at the beginning of this chapter about such an approach. To test this, during runs 15 and 17, the intra-molecular term was also evaluated.

T4.19. $\Delta\Delta G_{hyd}(298)$ for the Histamine monocation N(π)H perturbed into N(τ)H, evaluating both inter and intra molecular terms. Results given for N(π)H -----> N(τ)H.

Run No.	Number of steps equilibration /data collection	$\Delta\Delta G_{hyd.f}(298)$ kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ kJ mol^{-1}	average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
15	4000 /4000	-55.103	55.664	-55.383 ± 0.281
17	4000 / 4000	-98.027	100.090	-99.059 ± 1.032

These results are obviously terrible. Even the sign is incorrect. This vindicates the protocol used, evaluating only the inter-molecular free energy in solution and adding on the intra-molecular term.

IV.7. Relaxation of a solution

A second point was also raised by the last set of FEP calculations using the conformationally optimised charges. It can be seen in T4.14 that the hysteresis increases as the amount of both data collection and also initial equilibration is increased. By 8ps MD / 8ps DC, the result appears to have converged on average, but the sampling appears to be much worse than with the early runs.

This means that the system was not being allowed to relax properly at each value of λ. When collecting data for the ensemble average, it is important that the system is at equilibrium at that value of λ. Initially it was believed that water relaxes around a solvent in under 1ps, hence this was the value used in the early FEP protocols. A major problem is that the solvation energy could be affected as the relaxation energy will also contribute. This is demonstrated by the histamine system. In table T4.14 the difference in calculated free energy is 49.601 kJ mol^{-1} with 1ps of equilibration but 39.978 kJ mol^{-1} with 2ps. Assuming that the extra time does not mean significantly different conformations are sampled, this difference is due entirely to the extra relaxation.

A more subtle effect is also produced. If full relaxation does not occur between windows, the system at λ+1 will be more like that at λ than it should be. Hence running over a window from both directions will automatically be very similar. This leads to a small hysteresis on the double wide sampling and a spurious belief that the calculation is accurate. This effect can also be seen with the histamine results above. The hysteresis rises from 0.222 kJ mol^{-1} in run 12 to 1.464 kJ mol^{-1} in run 13 before dropping back to around 0.2 kJ mol^{-1} in runs 15 and 16R. When more equilibration is run between windows, allowing the system to relax properly, it becomes apparent that sampling in the smaller runs was very different from either direction. The large hysteresis on the N(τ)H runs, even with lots of equilibration and sampling must be due to the conformational problems discussed above.

The water thus seems to need in the order of 2-3ps to relax, not the 1ps as thought earlier. The reason for this mistake is due to fact that water does indeed have an average relaxation time of around 1ps. The molecular nature of solutions however means that average values are not the required data in simulations.

The problems of relaxation are well demonstrated by the work of Maroncelli and Fleming (Maroncelli and Fleming 1988 and Castner *et al.* 1988). They have used various models of solvation, ranging from standard continuum models, to inhomogeneous models through to molecular dynamics simulations. The relaxation time depends on the model chosen and also the way of defining it. However the general result obtained is that due to the molecular nature of solvation, the solvent relaxes at different rates depending on the nature solute and how far the solvent molecule is from it.

Measured solvation times are found to be reasonably similar to what is known as the solvent longitudinal relaxation time, τ_L, in continuum dynamics. For water this value is 0.42ps and the experimental solvation time 0.54ps. However on running simulations of various simple, i.e. spherical, solutes in ST3 water, they monitored the correlation functions for the decay of the solvation energy components. It was found that if the whole solvent was included, the solvation energy response decayed after around 1ps. If this was

broken down into the separate solvation shells, the first shell response had only decayed to half its initial value by this time.

These studies indicate that the much longer equilibration times of over 2ps before each window are necessary to allow the first solvation shell to be at equilibrium before data collection begins.

IV.8. Including Image Charges into GIBBS

The stage now reached is that a consistent result can be obtained from FEP calculations on the histamine monocation system. The next stage is to try and improve that result by including the long range interactions missed by using a non-bond interaction cutoff. Different ways have previously been implemented to model these forces, either based on the Ewald sum method or including a so-called reaction field.

The former method allows the system to interact with all its periodic images and sums the interaction effectively to infinity. This has the advantage of evaluating the interaction exactly. However it is really only of use in periodic systems. In random phases such as solution, this method gives a spurious order to the calculation.

The alternative is to make a cavity in a continuum. The system resides in the cavity and undergoes normal MD, or MC. The dipole inside the cavity polarises the continuum. This polarisation gives rise to a dipole, the reaction field, inside the cavity. The interaction of the two dipoles is the long range interaction energy. Various different ways of including reaction fields to simulations have been tried (see Allen and Tildesley 1987, van Gunsteren and Berendsen 1990 for reviews). They all have. the problem that the system must be spherical. This obviously gives problems for periodic boundary conditions and discontinuities at the boundary mean that a simulation will be slow to converge.

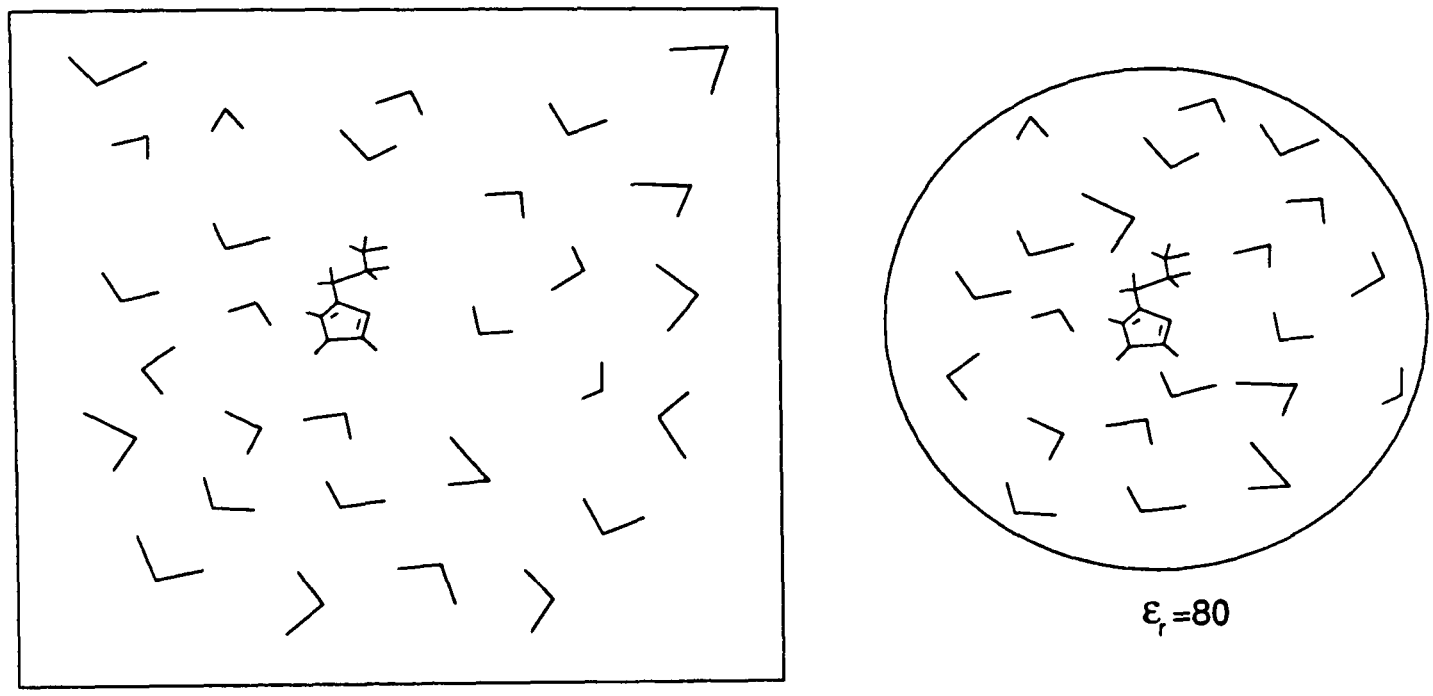
This study uses an approximation to the full reaction field method allowing periodic boundaries to be used. The approximation that is made is that in an FEP calculation only the energy difference between the endpoints for the perturbed molecule matters. The reaction field is thus only applied to this molecule, the solute. For example for the histamine system, the total energy can be broken down into

$$V = V_{\text{solute - solute}} + V_{\text{solute - solvent}} + V_{\text{solvent - solvent}}$$

The first approximation is that the solvent - solvent potential is fairly well described with an 8Å cutoff. It will also not change an FEP calculation result, except by biasing the sampling if the truncated potential is a very bad model. Only Lennard - Jones and dipolar interactions are present and so this seems unlikely. Also the solute - solute interaction is completely described with the cutoff as it reacts completely with itself inside this distance. Thus the only term that needs modifying is the solute - solvent interaction. This is broken down further into

$$V_{\text{solute - solvent}} = V_{\text{short range}} + V_{\text{long range}}$$

The idea is now to put the solute inside a sphere inside the solvent. All interactions inside this sphere are calculated explicitly. They form the short range part of the potential. A reaction field is then included for the long range part. All the solvent behaves as normal and so all boundary conditions are possible. What has changed is that the charged solute now feels a greater potential, that due to the explicit solvent and that due to the continuum model of the distant solvent. The difference in long range interaction between the endpoints of the calculation will thus be included. The following diagram F4.8 sums up this implementation



(a) (b)

F4.8. Changing a periodic system to allow the use of a reaction field. (a) shows the total system, a solute atom in a periodic box of solvent molecules. (b) represents how the system seems to the solute, a sphere of solvent molecules inside a continuum of dielectric 80.

The next problem is how to add the reaction field to the potential. The method that was chosen was to model the field as a set of point charges that are related to the solute atoms. This was taken from the work of Friedman (1975). The electrical potential, ϕ, due to a point charge, in the absence of a charge distribution is given by Laplace's equation

$$\nabla^2\phi = 0$$

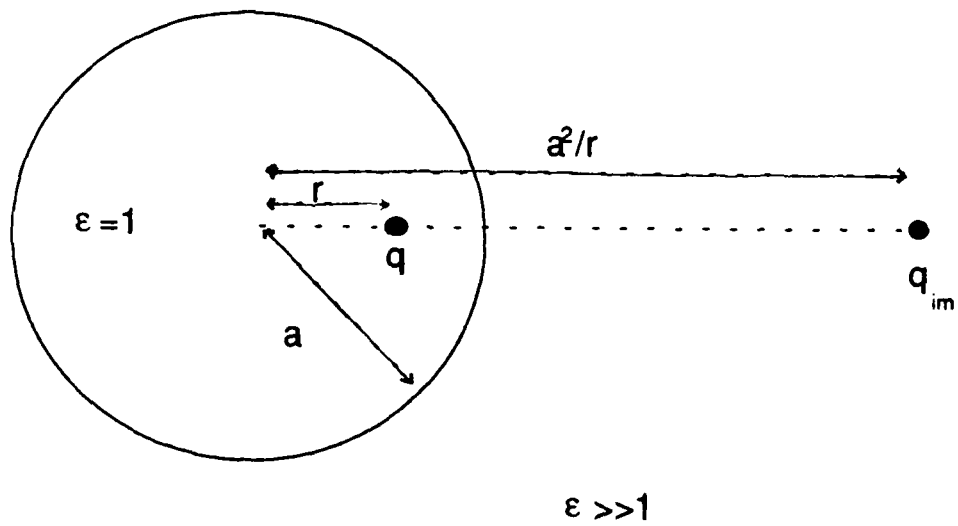
The solutions to this equation are the spherical harmonics. However if the point charge is in a spherical cavity with a dielectric constant of unity, inside a continuum, boundary conditions alter the solution. The direct coulombic potential of the charge can now be subtracted from this potential to give the potential due to the continuum i.e. due to the reaction field. This potential can then be modelled by an imaginary charge, an image charge due to the source. Under the approximation that the dielectric outside the cavity, ε, is much larger than unity, this image charge has magnitude

$$q_{im} = -\frac{\varepsilon - 1}{\varepsilon + 1}\frac{a}{r}q \qquad \text{(E4.14)}$$

and position

$$\vec{r}_{im} = \frac{xa^2}{r^2}\hat{i} + \frac{ya^2}{r^2}\hat{j} + \frac{za^2}{r^2}\hat{k} \qquad \text{(E4.15)}$$

where the source charge is at (x,y,z) a distance r from the centre of the cavity which has radius a. This is displayed below in F4.9.



F4.9 The dimensions used in the calculation of a reaction field image charge.

From this, all that must be known is the magnitude of the source charge and its position. The appropriate scaling factors then give the model of the reaction field. This was added into GIBBS3.1 as follows. The cutoff for the solute was changed from an atom based to a residue based one, with its origin at the centre of mass. Every time the non-bonded pair list was calculated (every 100 steps of dynamics) the image charge for each solute atom was calculated according to the above prescription. These were then added into the pair list for the solute coulombic potential energy. This meant that an extra n^2, where n is the number of solute atoms, interactions were added to the calculation, an insignificant number for a 40 atom solute in a box of solvent. The forces due to the long range solute - solvent interaction were thus also added onto the solute as these are calculated from the gradient of the potential.

This method was then used to recalculate the FEP energies calculated above for the histamine tautomerism. These results are given in T4.15. Runs 20 and 22 start with the N(π)H tautomer, while runs 21 and 23 with the N(τ)H.

It can be seen that the reaction field has changed the FEP result by ca. 1Kcalmol^{-1}. Interestingly with the conformationally averaged charges this change favours the N(π)H tautomer, whereas with the extended charges this is reversed.

The equilibria data calculated, combining these results with the gas phase energies in T4.16, is given in T4.21.

T4.20. $\Delta\Delta G_{hyd}(298)$ for the Histamine monocation N(π)H perturbed into N(τ)H, using different charge sets, with the inclusion of the image charge reaction field.

Run No.	Charge set[a].	$\Delta\Delta G_{hyd.f}(298)$ kJ mol^{-1}	$\Delta\Delta G_{hyd.b}(298)$ kJ mol^{-1}	average $\Delta\Delta G_{hyd}(298)$ kJ mol^{-1}
20	conf	40.033	45.572	42.802 ± 2.770
20R	conf	42.078	48.806	45.442 ± 3.364
21	conf	41.685	48.940	45.313 ± 3.628
21R	conf	42.702	42.953	42.827 ± 0.126
22	6-31G*	37.514	36.669	37.091 ± 0.423
23	6-31G*	30.815	29.769	30.292 ± 0.523

a conf denotes the conformationally adjusted charges, while 6-31G* denotes the 6-31G* charges calculated on the extended geometries.

T4.21. The difference in free energy and the equilibrium constant, between histamine monocation tautomers in solution, including a reaction field correction for the long range energy.

	Conformational charges		Extended charges	
Basis set	ΔG (g,298) kJ mol^{-1}	K_T (g,298)	ΔG (g,298) kJ mol^{-1}	K_T (g,298)
AM1	19.846 ± 7.948	3 x 10^{-4}	9.442 ± 5.281	2 x 10^{-2}
3-21G	1.806 ± 7.948	0.5	-8.598 ± 5.281	32.2
6-31G*	5.686 ± 7.948	0.1	-4.718 ± 5.281	6.72

The inclusion of the image field has changed the result slightly. It may even be considered more accurate (the experimental value of 4 seems to be a lower bound to the exact constant) but with the error involved, this sort of statement is not a reasonable one. However it appears that the reaction field correction is giving the correct result and should be theoretically more exact than the earlier calculations ignoring the long range part of the potential. When making calculations on the high charges in GDP and GTP, this correction should be even more important.

IV.9. Conclusions about Simulations of Charged, Flexible Species

The series of calculations described in this chapter show how the calculation of free energy differences between flexible, charged species should be performed in order to obtain a reasonable result. The methods that should be adopted can be summarised as follows.

Firstly it must be ascertained which conformational species are present during the simulation. Charges that represent the electronic distribution for all these species should then be used. Also the intra part of the free energy term can be added by gas phase calculations on the appropriate conformations, thus avoiding the need to include these large terms in the molecular mechanics simulation.

The treatment of the long range forces should also be included. An image charge approximation for the perturbed species seems to be a relatively cheap and simple way of doing this.

When these points are noted, the results are seen to be in remarkably good agreement with experiment.

CHAPTER V

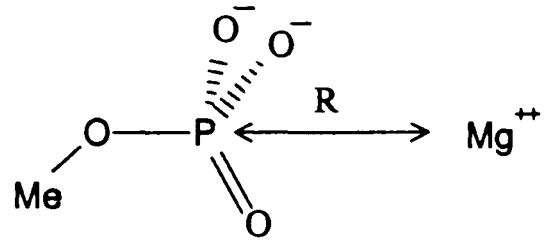
Calculating Suitable Parameters for the Guanosine Phosphate Systems

The next stage is to go on to calculate the difference in binding free energies of GDP and GTP to *Ha-ras*. Before this can be done, the parameters that will be used in the molecular mechanics simulations must be carefully chosen. The standard AMBER parameters are known to simulate proteins in a reasonable manner, and can be taken as a complete set. Parameters are also available for RNA nucleotides, although these do not seem to be as well optimised as those for the amino acids. These parameters should reproduce the bonding and non-bonding characteristics of GDP and GTP well enough with one exception: the electrostatic terms. The AMBER force field phosphate charges were calculated on dimethylphosphate, and so are unlikely to reproduce a di- or tri- phosphate electron distribution accurately. This is the area in the system that will be perturbed in the calculations, so it was thought worthwhile to spend time on getting these "atom charge" parameters as reasonable as possible.

V.1. Checking the AMBER Parameters for the Phosphate - Magnesium Interaction.

A large term in the nucleoside - *ras* protein interaction potential will be due to the proximity of the negative phosphates and the positive magnesium ion. The way AMBER models this interaction was thus checked against high level *ab initio* calculations. In principle this is simple. Using both methods, potential energy curves are calculated from the interaction energy between the two ions at various distances. These are then compared for shape and depth etc.

It was decided to use the system $MePO_4^{2-}$ with the magnesium ion approaching along the P—O bond towards the methyl group as shown below in figure F5.1.



F5.1. The magnesium - phosphate system used to calculate interaction potentials with various parameter sets

This would be expected to be a favourable approach with the magnesium interacting with all three charged oxygen atoms equally.

V.1.1. Calculating an *Ab Initio* Potential Curve for the Magnesium - Phosphate Interaction.

The *ab initio* calculations were performed first. The $MePO_4^{2-}$ ion structure was initially optimised to STO-3G and then 3-21G* level. This two stage calculation was made to speed up the optimisation, as the initial structure was fairly random. In retrospect it is surprising that the STO-3G basis set managed to converge. The 3-21G* basis set was not perfect; in fact even when optimised there were only 52 negative eigenvalues for 58 electrons. This means that six electrons are in effect unbound. In general it is thought that this level of optimisation is sufficient for negative ions, although any electronic properties calculated from this wavefunction would obviously be dubious.

The magnesium ion was then added and 6-31++G* single point energies were calculated with various values of R. While this is not the highest level of basis set available with GAUSSIAN88, there are diffuse functions on all the ~~heavy~~ atoms and polarisation functions on ~~all~~ heavy atoms. With this basis set, methylphosphate had 150 basis functions and gave 58 negative eigenvalues. All the electrons are now bound and the wavefunction should be adequate.

In order to calculate the interaction energy, it is then necessary to remove the energy of the separate ions from this value.

$$E(\text{interaction}) = E(\text{total}) - E(MePO_4^{2-}) - E(Mg^{2+}) \qquad \text{(E5.1)}$$

At the same time it is possible to correct for the so-called basis set superposition error (BSSE). In early calculations on helium dimers, it was found that using a basis set of 1s, 2s, 2s' and 2p orbitals, the interaction energy agreed well with experiment. The only problem was that as the atoms came closer together than 2Å, there was a large increase in the mixing of the 2p orbitals. This extra mixing gives rise to a spurious stabilisation energy, due in effect to the migration of the 2p orbital on one helium atom onto the other to form a beryllium atomic orbital (which is what happens in the limit of 0Å between the helium atoms) (Kestner 1968).

This error in the superposition of the bases, seen when a long way below the Hartree-Fock limit, is especially marked when an electronegative atom is next to an electropositive one. In this situation, the electron rich species tends to steal functions from the function rich species to help stabilise its high energy electrons. The system under study is such a "super molecule", as isolated combinations of molecules are called. The phosphate group has all its valence orbitals occupied, and two fairly high energy electrons, while the magnesium has an empty valence shell.

Because of this, rather than simply calculating the energy of the free ion, the following procedure was carried out. For each value of R used in the total energy calculations, two more calculations were made. One had an explicit methylphosphate ion

and a ghost magnesium ion, while the other had the reverse situation. Ghost atoms have no nuclear charge or electrons, but have all the basis functions of the normal atom. This is known as the Counterpoise Correction (Boys and Bernardi 1970). Calculating in this way thus allows the same basis function migration as in the total energy calculations, and so corrects for the basis set superposition error. In fact it allows a greater migration, as in the counterpoise calculation all the ghost orbitals are empty and available rather than just the virtual and empty valence orbitals in the supermolecule calculation (Yang and Kestner 1990). This overestimates the BSSE, but with what is hoped to be a negligible error, as more accurate methods are horrendously complicated.

The interaction energies obtained in this manner at various values of interaction distance, R, are given below in table T5.1. E(Total) is the supermolecule energy, while the CP corrected energies are for the atom - ghost atom complexes described above. E(Interaction) is the potential calculated according to E5.1. The counterpoise corrected ionic energies show that the BSSE is negligible for magnesium, but for the methylphosphate it is of the order of 0.005 a.u's (13 kJ mol^{-1}).

T5.1. The *ab initio* 6-31++G* interaction energies, corrected for BSSE, for magnesium - methylphosphate calculated as described in the text.

R (Å)	E (Total) (a.u's)	E ($MePO_4^{2-}$) CP corrected (a.u.'s)	E (Mg^{2+}) CP corrected (a.u.'s)	E (Interaction) kJ mol^{-1}
1	-877.45832	-679.78068	-198.81263	2867.39
1.5	-879.10126	-679.77762	-198.81246	-134.02
2	-879.39742	-679.77663	-198.81228	-2122.62
2.5	-879.37400	-679.77633	-198.81219	-2062.15
3	-879.28933	-679.77605	-198.81217	-1840.66
3.5	-879.21076	-679.77556	-198.81215	-1635.72
4	-879.15449	-679.77521	-198.81215	-1488.91
4.5	-879.18218	-679.77508	-198.81214	-1561.97
5	-879.15893	-679.77503	-198.81214	-1501.06
6	-879.15921	-679.77496	-198.81214	-1501.99

V.1.2. Calculating AMBER Interaction Energies to Compare With the *Ab Initio* Data.

The calculations above were then repeated using AMBER. Before using AMBER however it was necessary as always, to choose the parameters. Initially for the methyl phosphate, all the parameters were taken from the standard force field. The charges for the phosphate were taken from the RNA standard set, and the methyl group charges from the

serine β-carbon because it is next to an electronegative oxygen. These charges were then adjusted slightly to make the total charge 2-.

The Lennard-Jones parameters in AMBER are specified by two values, ε* and R*. The former is the well depth and the latter the equilibrium distance associated with the atom. These combine to give the A_{ij} and B_{ij} parameters in E1.61 by using the combination rules

$$A_{ij} = \sqrt{\varepsilon_i^* \varepsilon_j^*}\left(R_i^* + R_j^*\right)^{12} \qquad \text{and} \qquad B_{ij} = 2\sqrt{\varepsilon_i^* \varepsilon_j^*}\,(R_i^* + R_j^*)^6 \qquad \text{(E5.2)}$$

For the magnesium ion, two sets of non-bonded parameters were found, and both tried. The first set were from the standard AMBER3.1 force field and the second from Cieplak *et al.* (1987). For comparison the values are given in T5.2.

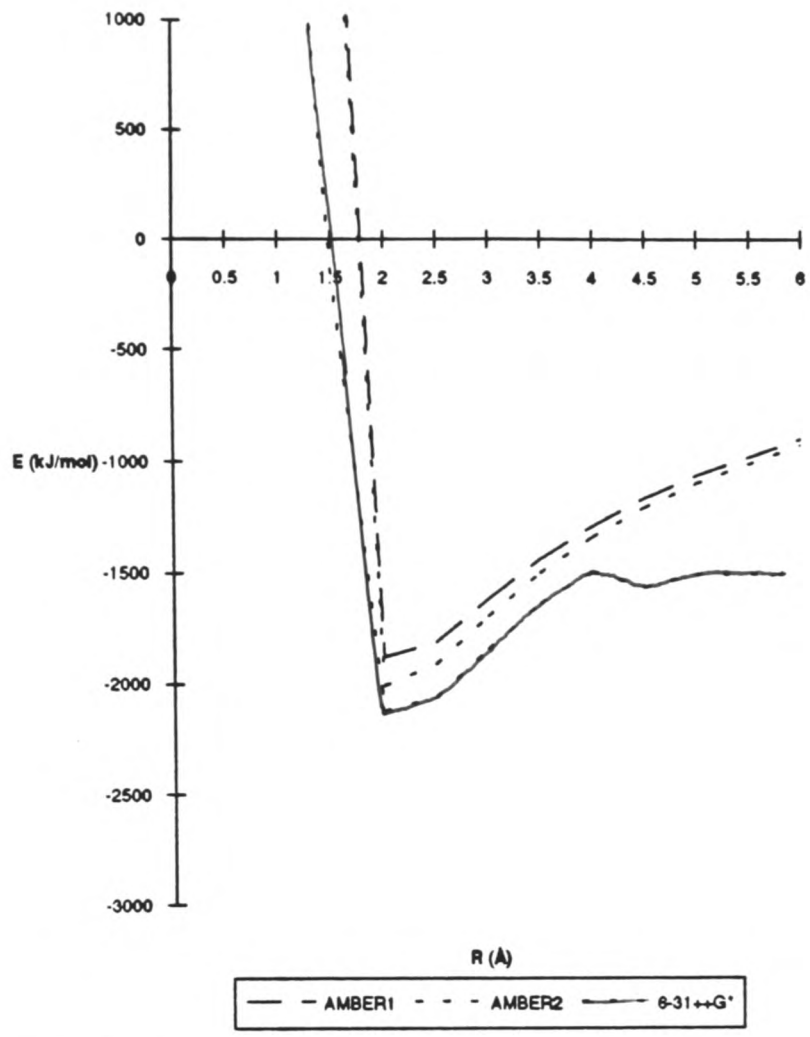
T5.2. Different Lennard-Jones parameters for the magnesium ion

	R* (Å)	ε (Kcal)
set 1	1.17	0.10
set 2	1.16	0.05

The set 2 data also included a polarisability parameter, but AMBER3.1 does not have the facility to use this data so it was ignored. A charge of 2+ was also obviously used.

Using AMBER the methylphosphate was geometry optimised. The EDIT module was used to add the magnesium ion with respect to the phosphate in positions identical to those used in the *ab initio* calculations. The ANALYSIS module of AMBER was then used to calculate single point energies. This program calculates the total energy and breaks it down into parts to give directly the interaction energy between two groups in a system. For an accurate interaction energy, the system should be optimised at each point to allow it to relax. This was not done to make the calculations equivalent to the *ab initio* data.

These interaction potentials from the GAUSSIAN and AMBER calculations are plotted below in P5.1. The potential energy curves are very encouraging. The AMBER potentials ~~has~~ have a similar shape to the *ab initio* one and the minima are also in the same place. The differences are that at short range the AMBER potentials ~~is~~ are much steeper, and at long range it does not flatten off as quickly. Fortunately the short range part of the potential is not important since simulations will not visit this unfavoured area. The inclusion of a surrounding would also moderate the long range part by reducing the electrostatics and in the protein the magnesium cannot get this far from the phosphate. The more recent set of Lennard-Jones parameters are also seen to give a potential closer to the *ab initio*. This parameter set was thus taken for all other calculations. The difference between the two sets is striking. Nearly 200 kJ mol^{-1} difference in minima energy is obtained from what appears to be a small change in the ε* and R* values. The AMBER potential is thus adequate, although a little high in energy.



P5.1. The potential energy curves for the interaction between magnesium and methylphosphate ions calculated with a 6-31++G* wavefunction and with AMBER using the two parameter sets in T5.2. AMBER1 is with set 1 and AMBER2 with set 2.

V.1.2.1. The Effect of Different Charge Sets on AMBER Interaction Energies.

The only problem left now is how to calculate charges not available from the AMBER databases. At this point a very complex problem is encountered. When the AMBER parameters were calculated (Weiner *et al.* 1984, 1986) they were designed to be as transferable as possible. The database contained suitable Lennard - Jones and bonding parameters for all common chemical atom types. All one had to do was calculate the atom - centred charges and the simulation could be set up. However there are different ways to calculate these charges. In AMBER these were done by fitting to an *ab initio* MEP calculated with a STO-3G basis set. Later, people thought that the use of a higher basis set would immediately improve the quality of the electrostatic interactions.

This was the reason why the charges in the histamine calculations in chapter IV were fitted to MEPs from 6-31G or 6-31G* wavefunctions. However it does not seem to be as simple as this. For example in DNA simulations, the standard charges work passably well. If 6-31G* fitted charges are used instead (Hausheer *et al.* 1990), the DNA undergoes a transition from the common B form to the A form (I. Haworth and A. Elcock, personal

communication). In solution this should occur in conditions of high ion concentrations, but not in pure water, or worse still in the gas phase, as these simulations have predicted.

The reasons for this transition are not obvious. Higher basis sets, however, are seen to give larger atomic charges, sometimes greater than unity. Electronegative atoms and carbons in their vicinity are especially effected. It is argued that the atom centred charges are non-physical and so this is possible. However, this is disconcerting to a chemist to whom the idea of ionisation occurring in a neutral molecule does not appeal. This greater polarity is one possible explanation for the strange DNA transition; the system with 6-31G* charges are simulating a DNA molecule polarised by an ionic atmosphere.

The reason for the greater polarity is also unclear. The dipole moments of a system are much better represented by a 6-31G* wavefunction and its corresponding atom charges than those calculated using STO-3G (Hehre *et al.* 1986). However these values can only be tested against molecules that have well defined rotational spectra, which excludes complex molecules and ions. It might be possible that in a complex electronic system where neighbouring molecules have very different eletronegativities and electron densities, a sort of intra - molecular superposition problem occurs. As the number of basis functions rises, whilst remaining well below the Hartree - Fock limit, function borrowing between atoms may occur in an analogous manner to that in the magnesium - phosphate system. This could then lead to a shifting of electron density.

At a coarser level, the reason for these problems is the incompatibility of the high level basis set charges with the other AMBER parameters. Force field parameters are "fudged", or to put it technically optimised, starting from experimental or *ab initio* data, to give experimental results. In chapter I it was noted that the non-bonded interaction between two atoms has four terms:

$$U_{\text{Non-bond}} = U_{\text{Coulombic}} + U_{\text{Induction}} + U_{\text{Dispersion}} + U_{\text{Repulsion}} \qquad \text{(E5.3)}$$

In contrast the AMBER non-bond function has just three terms. The induction term has been combined into the Lennard-Jones and Coulombic terms. The induction energy is given by

$$U_{\text{ind}} = -\frac{1}{2}\alpha\mathcal{E}^2 \qquad \text{(E5.4)}$$

where $\mathcal{E}$ is the electric field experienced by the atom, and α its polarisability. This however is related to its electron distribution by

$$\alpha = \frac{2}{3}\sum_{n}{}' \frac{\langle 0|\mu|n\rangle\langle n|\mu|0\rangle}{\Delta_{n0}} \approx \frac{e^2\hbar^2 N_e}{m_e\Delta} \qquad \text{(E5.5)}$$

where N_e is the number of electrons and Δ an average value for the energy level separations. From this, changing the electronic distribution requires a modification of the induction term. In AMBER, the Lennard - Jones terms were altered so that when combined with a STO-3G electron distribution the results matched experiment. The more polar

6-31G* charges, whether or not more "correct" than STO-3G charges, are therefore inapplicable with the AMBER Lennard-Jones terms.

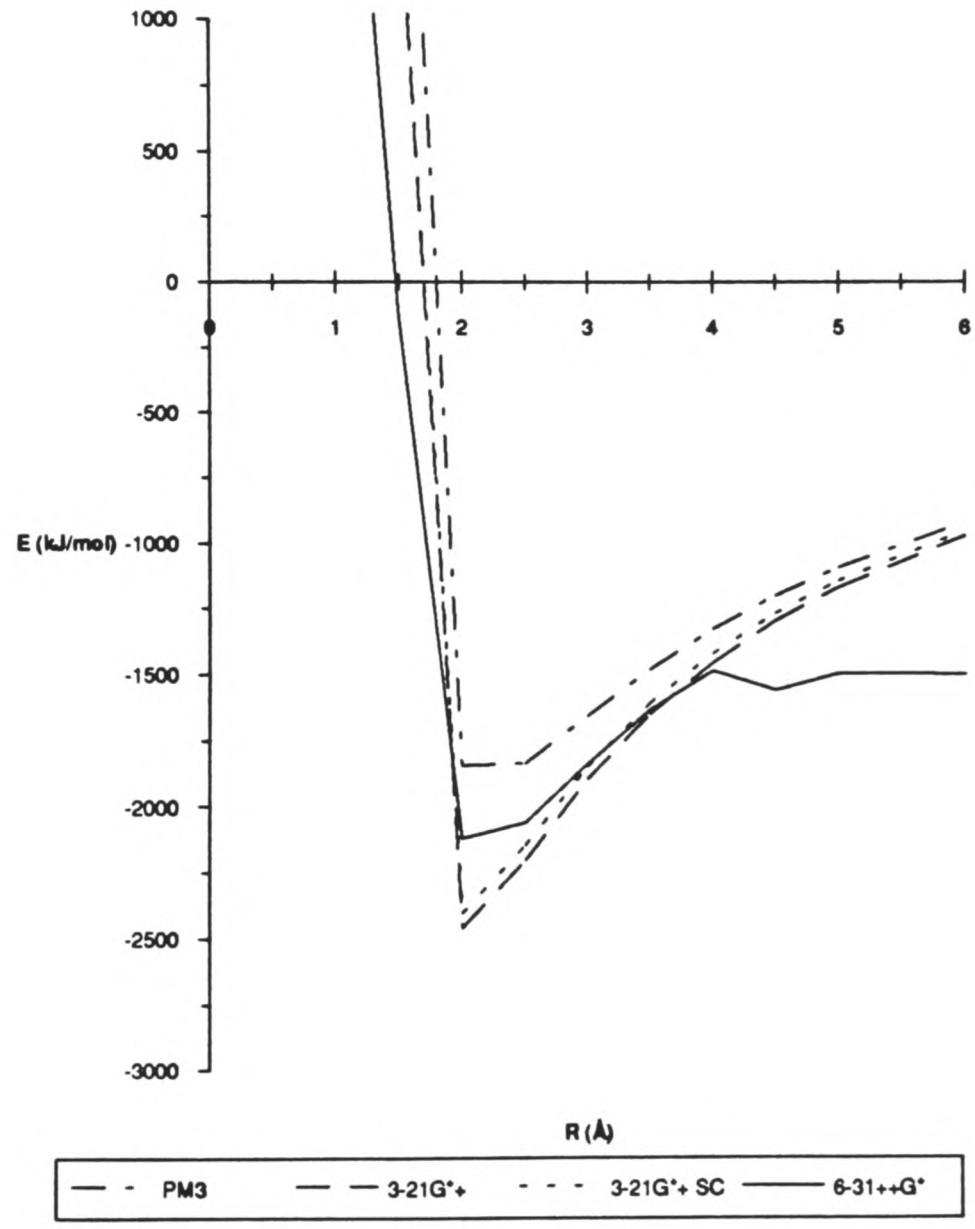
An example of the importance of force field integrity is shown by the OPLS method (Jorgensen and Tirado-Reves 1988). Here the Lennard-Jones terms are taken as fixed, and the "charges" are simply numbers optimised to give the desired results. Despite this complete lack of physical meaning in the electrostatic terms, these parameters have been shown to perform very well indeed. It is interesting to note that these optimised "charges" are in general closer to calculated 6-31G* mulliken charges than to MEP fitted charges.

A second problem with charges fitted to MEPs is how this fitting actually takes place. Charges are seen to be dependant on how the MEP surface is set up. CHELP (Chirlian and Francl 1987) used spheres of random points 1Å apart from the van der Waals surface of the molecule, out to 3Å. QUEST, a part of AMBER, again uses surfaces of points, but this program allows the number of surfaces and the point density to be varied. The default however is a single surface at 1.4 times the van der Waals surface of a molecule. The program used in this project, QGEN (C.A. Reynolds in house code), uses the MEP calculated at a random set of points in a band beyond the van der Waals surface of the molecule. This means that the charges will fit the MEP as well as possible over a whole volume of space. Even with this method the charges depend on the density of points. The differences between methods can give charges varying by 0.2 units (J.W. Essex, personal communication). This may not sound much, but for FEP calculations it can lead to a significant error. For this reason alone, the OPLS approach is more prudent than the AMBER.

The result of all this is that STO-3G charges, all produced by the same method, should be used with the standard AMBER parameters. The problem then arises that for di- and tri- phosphates STO-3G cannot produce a stable wavefunction ; for example it was unable to optimise GDP. The smallest basis set that can is 3-21G*. Even this gives some unbound electrons with GDP, and so diffuse functions must be added to both oxygen and phosphorus atoms of the phosphates. It might be thought that adding diffuse functions to just one part of a molecule might unbalance the basis set, but it is hoped that the charge is predominantly in this region and so the extra functions merely take care of these electrons without drawing any from other parts of the molecule. Adding diffuse functions to all atoms in GDP or GTP makes the basis set too large to be workable.

It has been found that there is a fairly linear correlation between charges calculated between various basis sets. 3-21G charges are a factor of 1.34 larger than STO-3G charges (Lister *et al.*). This scaling factor can therefore be used to create charges from a 3-21G MEP compatible to AMBER.

Using the 3-21G* structure, a 3-21G*+ MEP was calculated for methylphosphate and charges were fitted to this. 3-21G*+ is used to denote a 3-21G* basis set with diffuse functions on just the phosphates of the molecule. It should not ~~to~~ be confused with the GAUSSIAN standard basis set 3-21+G* which would have diffuse functions on all atoms. The charges were then scaled by the factor given above. Unfortunately when a system is charged it cannot be simply scaled as this also scales the overall charge. Assuming once more that the ionic charge is all contained on the phosphate, the charge lost by scaling was divided by five and added back to the phosphate atoms. The values on the terminal oxygens and hydrogens were then averaged. The calculation of the AMBER magnesium - phosphate interaction potential was then repeated using both the scaled and unscaled charges. As a comparison, charges were also obtained from a semi-empirical PM3 MEP. These are shown below in plot P5.2.

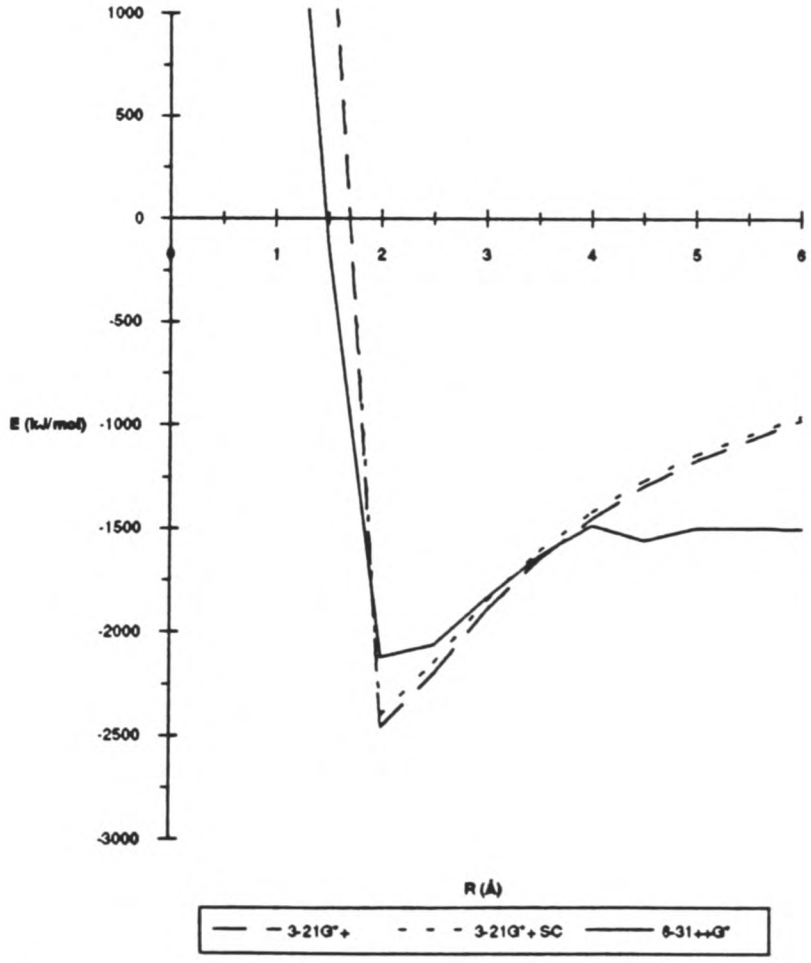


P5.2. The potential energy curves for the interaction between magnesium and methylphosphate ions calculated with a 6-31++G* wavefunction and with AMBER using charges fitted to a PM3 MEP (PM3) and a 3-21G*+ MEP either averaged over terminal atoms (3-21G*+), or averaged and scaled by 1.34 to mimic STO-3G charges (3-21G*+SC)

Again these empirical curves follow the *ab initio* one well, especially in the region from 2.5 - 4 Å which is important for simulations. The minimum is much lower than the standard AMBER charge curves above and indeed lower than the 6-31++G* potential. This may prove advantageous. The *ab initio* curve was calculated without the inclusion of electron correlation. This would be especially important around the minima and would lower the *ab initio* energy.

The major difference in the STO-3G and 3-21G*+ charges is on the phosphorus atom. The STO-3G charge is 1.337, the 3-21G*+ charge 0.182. This is due to the ability of the higher basis set to use pπ - dπ (d-orbitals are not used with STO-3G) back donation from the terminal oxygen atoms onto the phosphorus. The oxygen charge is also lowered by this from -0.943 to -0.738 in the averaged charge set.

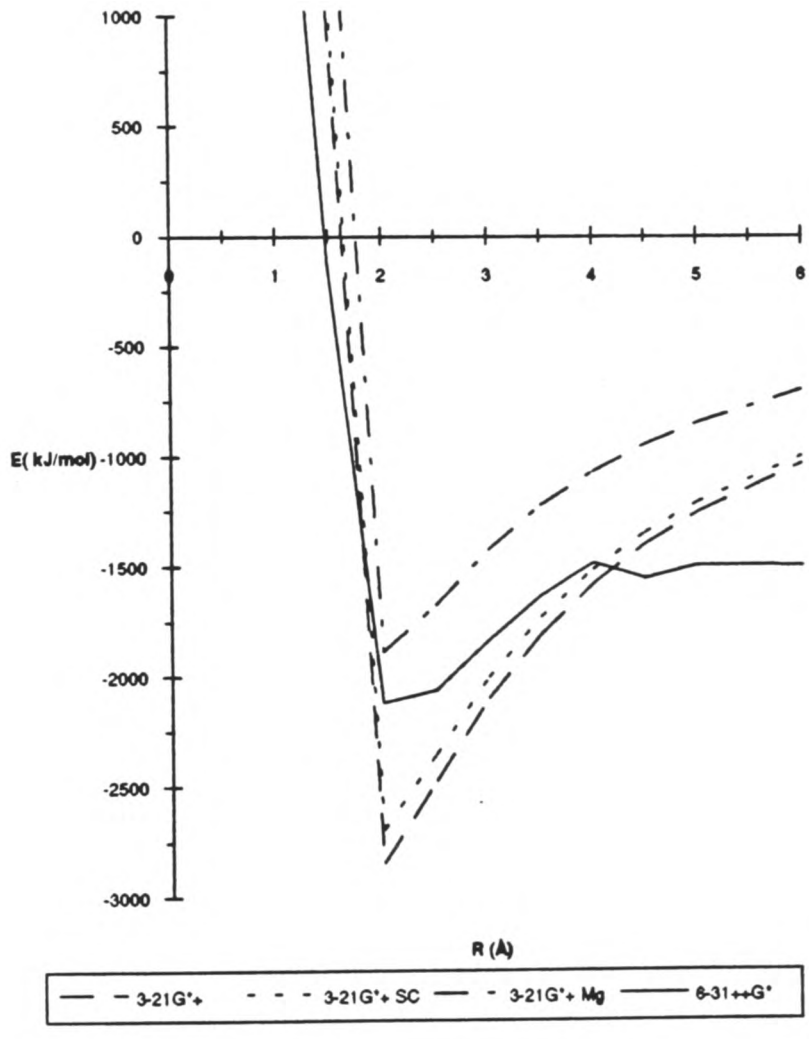
The calculated oxygen charges are in fact very asymmetric due to the geometry taken by the *ab initio* optimised molecule. Using the raw, non-averaged charges methylphosphate was minimised and another interaction potential was calculated. In solution or gas phase simulations the oxygens must be equivalent, however in a hindered environment such as a protein binding site this is not the case and the charge asymmetry may be important. It may also be important in these comparisons between methods.



P5.3. The potential energy curves for the interaction between magnesium and methylphosphate ions calculated with a 6-31++G* wavefunction and with AMBER using charges fitted to a 3-21G*+ MEP (3-21G*+) or scaled by 1.34 to mimic STO-3G charges (3-21G*+SC)

From P5.3, it can be seen that this makes very little difference, although this could be an artefact of the interaction geometry chosen. As the magnesium is equidistant from the three oxygens at all times, the field will be equivalent either way.

The next question was how the inductive effect of the magnesium affects the MEP fitted charges. A 3-21G*+ MEP was thus calculated on methylphosphate with the magnesium explicitly present in the minimum position indicated by the 6-31++G* curve. The charges obtained are quite different, and in fact give a charge of 1.8 on the magnesium ion! Interaction potentials were calculated both with the magnesium charge at this value, and scaled up to 2+, the excess charge being as usual spread over the phosphate group. These potentials are shown in P5.4.

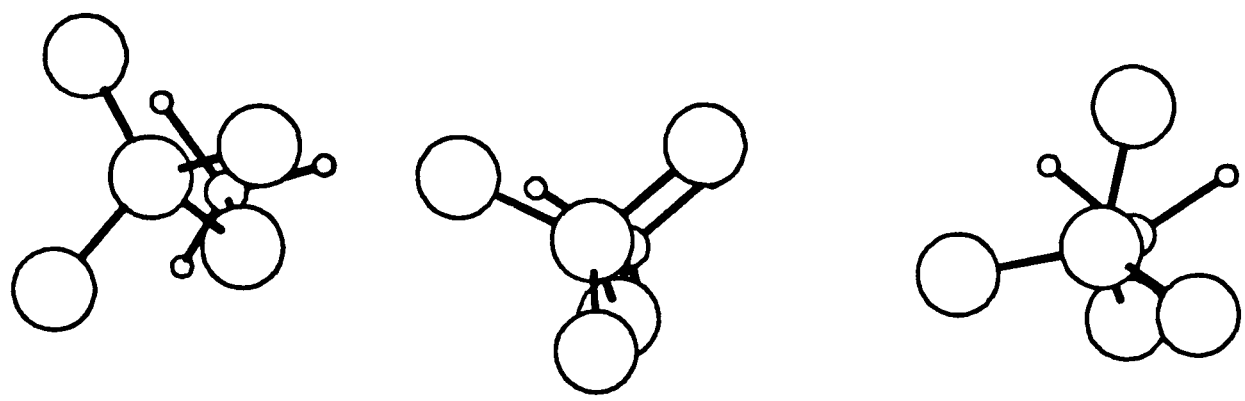


P5.4. The potential energy curves for the interaction between magnesium and methylphosphate ions calculated with a 6-31++G* wavefunction and with AMBER using charges fitted to a 3-21G*+ MEP calculated over both the methylphosphate and magnesium, either pure (3-21G*+Mg) or scaled to give a charge of 2+ on the magnesium (3-21G*+) or scaled again by 1.34 to mimic STO-3G charges (3-21G*+SC)

These potentials are not as good using the lower magnesium charge. This is a bit of a surprise, as it might be thought that the exact charges should give a very good minimum energy. The reason must again be a manifestation of the problem of including induction in the non-bond terms. In these calculations it is included twice; once in the Lennard-Jones and once in the coulombic term. This is why charges should be calculated on isolated gas phase molecules in the same way as the original parameter set. Once the magnesium is returned to a value of 2+, the original electron density is achieved and the parameters perform as well as before.

V.1.2.2. The Effect of Minima Geometry on AMBER Interaction Potentials

One interesting factor that was noticed is that different minima geometries are obtained with the different charge sets. These are represented in figure F5.2.



(a) (b) (c)

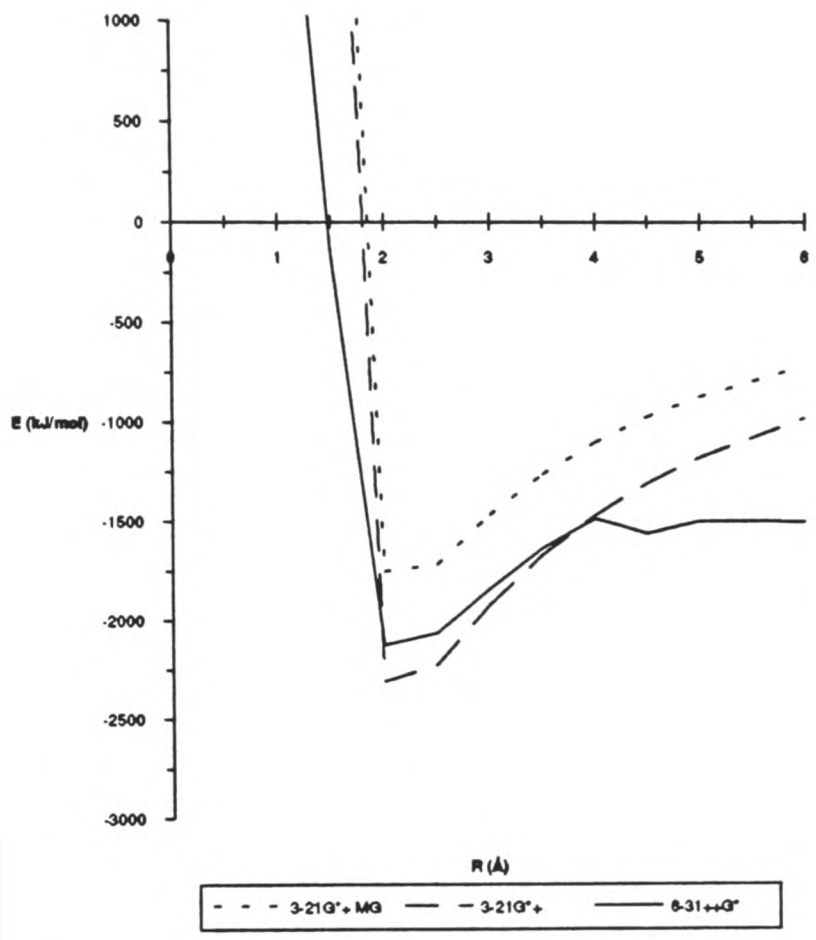
F5.2. Various minimum geometries for methylphosphate. (a) is from G88 with a 3-21G basis set. (b) is from AMBER using either STO-3G or 3-21G*+ charges which have been averaged over the terminal atoms. (c) is from AMBER using the untreated charges fitted to a 3-21G*+ MEP.

As a check on the affect the geometry has on the energy, the potential calculations were repeated with the 3-21G*+ unaveraged and unscaled charges, calculated with and without the magnesium present, but using the G88 geometry for the methylphosphate. These curves are given below in P5.5

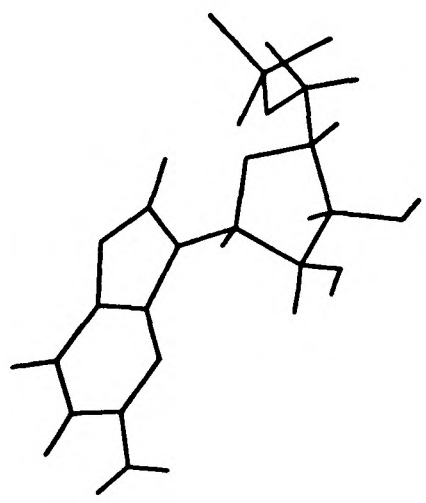
The charges calculated without the magnesium now give a curve very close to the *ab initio one*. There is now even the same shape at the minima. The problem with the geometry may be caused by the flexibility of the system. It is a multi minima molecule which could make it very sensitive to small changes in parameters. Alternatively the AMBER torsional potential for methylphosphate may be incorrect.

The answer to this problem can be found by looking at the crystal structure for guanosine 5' monophosphate (GMP). This is the only guanosine 5' phosphate structure in the Cambridge Crystallographic database (code GUANPH01, re-evaluated, 1980 Emerson and Sundaralingam). From the orientation of the crystal structure shown in figure F5.3, it

can be seen that the phosphate in fact adopts the same orientation as the AMBER minimised methylphosphate using the charges averaged over the terminal atoms (F5.2 (b)). The difference in the GAUSSIAN88 structure (F5.2(a)) must be an artefact of the starting geometry. This is good news as it means that the AMBER force field torsional parameters are quite adequate for this system.



P5.5. The potential energy curves for the interaction between magnesium and methylphosphate ions calculated with a 6-31++G* wavefunction and with AMBER using charges fitted to a 3-21G*+ MEP, either calculated over both the methylphosphate and magnesium (3-21G*+Mg), or just the methylphosphate (3-21G*+), but using the G88 optimised geometry for the methylphosphate

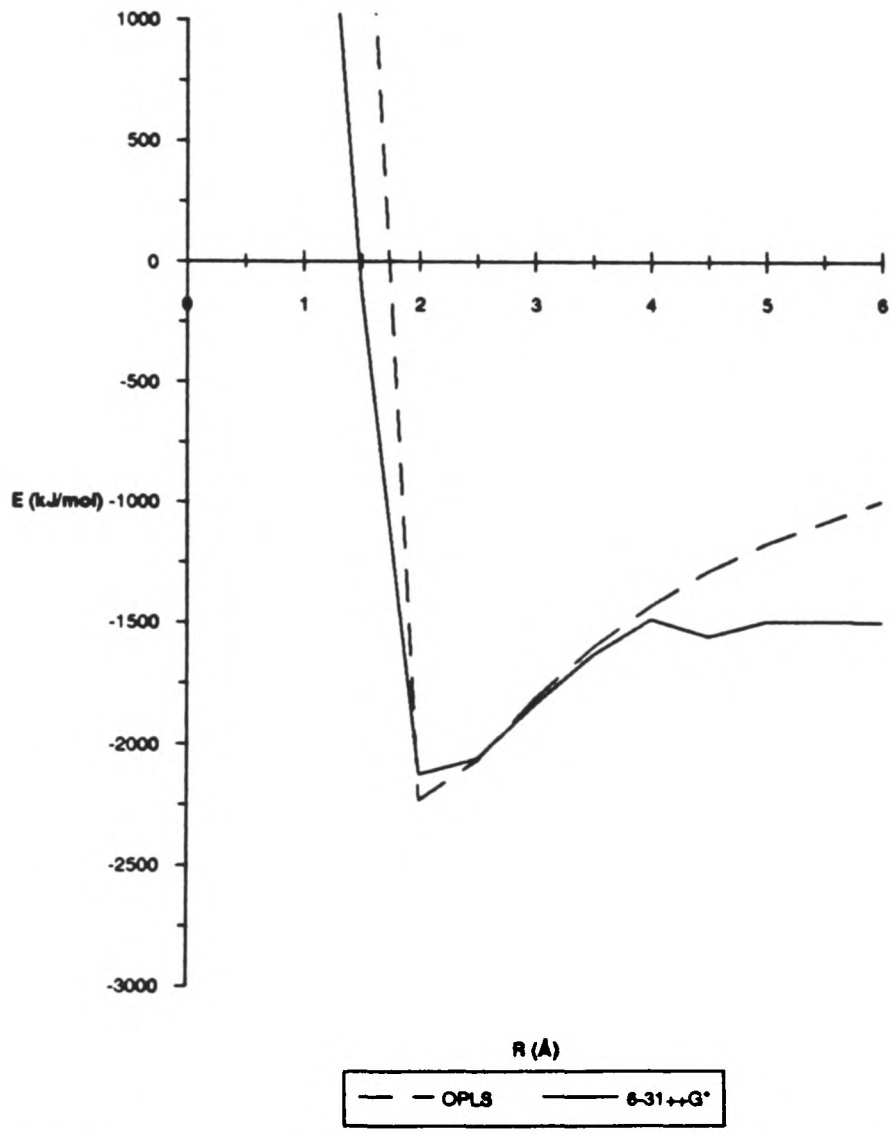


F5.3. The GMP crystal structure.

V.1.2.3. The OPLS Force Field Interaction Potential for Magnesium - Phosphate

As a final comparison of parameter sets the methylphosphate - magnesium potential was calculated using parameters from the OPLS parameter set (from the program BOSS3.1). This is shown in P5.6.

The success of these parameters is obvious. This is not altogether surprising as OPLS are optimised to 6-31G* *ab initio* calculations. However, the phosphate parameters again come from dimethylphosphate and have transferred very well to methylphosphate with just an adjustment of charge. Notwithstanding this, the major problem with the OPLS is their transferability. Due to the lack of physical meaning for the charges, if the parameters are not optimised for the particular system they are unlikely to fit, and while methylphosphate is similar to dimethylphosphate, diphosphate is less similar.



P5.6. The potential energy curves for the interaction between magnesium and methylphosphate ions calculated with a 6-31++G* wavefunction and with OPLS parameters.

V.1.3. Conclusions About Non-bond Parameters to Use With AMBER.

It appears that the AMBER parameters, combined with *ab initio* potential derived charges do reproduce to a reasonable degree of accuracy full *ab initio* interaction energies for ~~foe~~ magnesium and methyl phosphate. It may also be noted that because electron correlation has been omitted in the *ab initio* calculations that the AMBER values with the 3-21G*+ charges may in fact be the best representation of the minimum by any of the methods.

For comparison the charges and Lennard-Jones parameters used above are given in T5.3 and T5.4 respectively.

T5.3. Various charge sets calculated for methylphosphate. The heading indicates the MEP the charges were fitted to. Scaled indicates the charges were scaled by a factor of 1.34. Inc. Mg means that the MEP was calculated with the magnesium explicitly in the molecule.

	AMBER Standard (STO-3G)	PM3	3-21G*+	3-21G*+ Scaled	3-21G*+ inc Mg	3-21G*+ inc Mg Scaled	OPLS
Mg	2.000	2.000	2.000	2.000	1.844	2.000	2.000
O3A	-0.943	-1.255	-1.473	-1.201	-1.500	-0.659	-0.774
O2A	-0.943	-1.255	-0.388	-0.391	-0.340	-0.659	-0.774
O1A	-0.943	-1.255	-0.352	-0.364	-0.244	-0.659	-0.774
PA	1.337	2.300	0.182	0.034	-0.364	-0.373	0.666
O	-0.748	-0.675	-0.469	-0.452	0.038	-0.073	-0.544
C	-0.015	0.170	0.276	0.206	-0.236	-0.176	0.200
H1	0.085	-0.010	0.060	0.045	0.314	0.199	
H2	0.085	-0.010	0.067	0.050	0.214	0.199	
H3	0.085	-0.010	0.098	0.073	0.274	0.199	

T5.4. The Lennard-Jones parameters from the AMBER and OPLS force fields

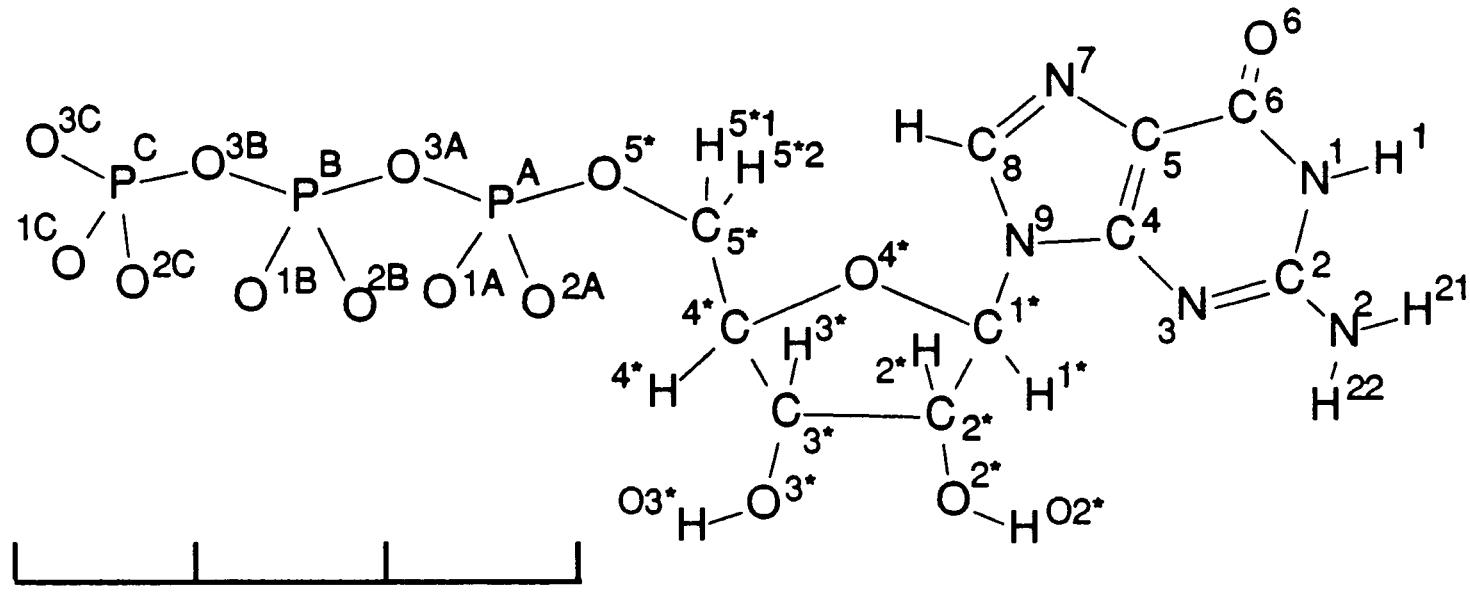
	AMBER		OPLS	
Atom Description	ε* (Kcal Mol^{-1})	R* (Å)	ε* (Kcal Mol^{-1})	R* (Å)
Terminal Oxygen	0.20	1.60	0.21	1.661
Phosphorus	0.20	2.10	0.20	2.099
Ester Oxygen	0.15	1.65	0.17	1.740
Carbon	0.06	1.80	0.17	2.133
Hydrogen	0.02	1.00		

The above set of calculations indicates that scaled 3-21G*+ charges should be both compatible with AMBER parameters, and give a reasonable description of the complex phosphate electron distribution. In all calculations later in this thesis, the charges will be calculated this way.

V.2. The Nucleotide Flexibility Problem

Both nucleotides in this study are very flexible molecules: GDP has seven possible internal rotations and GTP nine! From chapter IV, it is obvious that this will produce problems in deciding what conformations should be used in any calculations. The first thing to do was to find all the possible stable gas phase conformers. Then the important ones both for the gas phase energy difference relative to the solution phase, and the description of the electronic distribution had to be chosen.

The naming convention used for the atoms of GTP and GDP is shown in figure F5.4.



F5.4 Naming protocol for the atoms in GTP and GDP. In the text occasionally ' is used in place of * for the sugar atoms e.g. O5' for O5*.

V.2.1. Gas Phase Conformational Analysis

As well as the problem of so many degrees of freedom, there is also the complication of the high negative charge involved. In the protein, the nucleotide is associated with a magnesium 2+ ion. It is also stabilised by the amino acids in the binding site. Under these conditions GDP has a formal 3- charge and GTP 4-. Simple gas phase *ab initio* calculations, such as semi-empirical methods or minimal basis sets are unreliable, with this situation . The problem is that with negative ions the electronic distribution extends further from the nucleii than uncharged, or positively charged, molecules. After all, under usual conditions a GDP^{3-} ion in the gas phase is not very stable. With low level calculations, this leads to electrons being unbound and an unreliable wavefunction results. The only way to deal with such molecules is to include "diffuse functions". These are basis functions that enable the orbitals to be more spatially extended than normal.

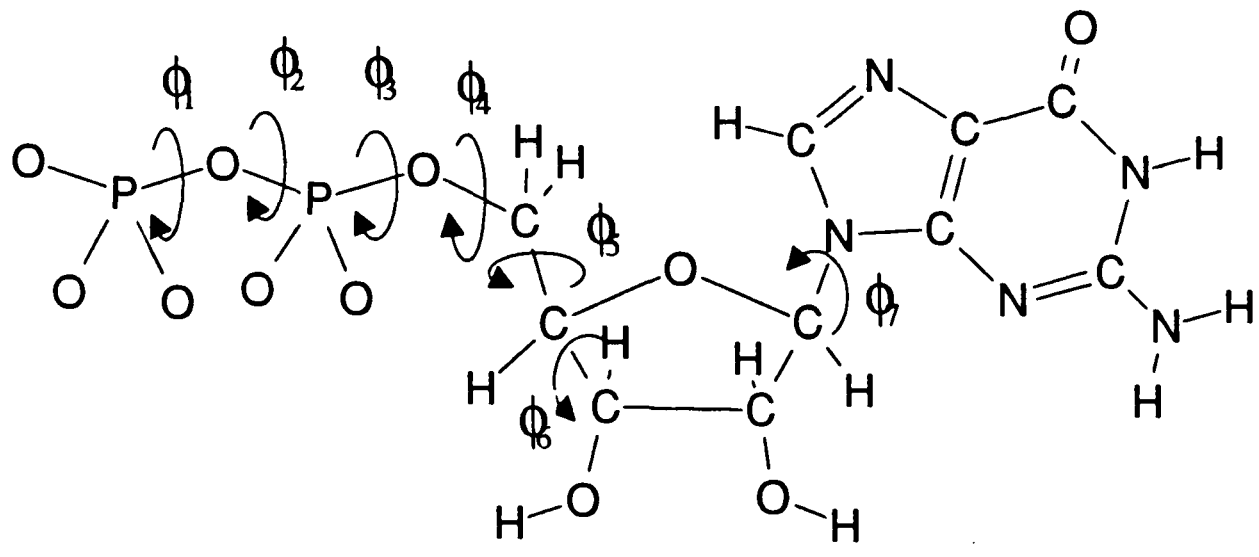
This means that gas phase energies can be calculated, but they are computationally expensive - especially when GDP has 40 atoms and GTP 44. Thus *ab initio*

conformational contour plots, like those derived for histamine are impossible. For this problem, molecular mechanics must therefore be used. Fortunately it is to be hoped that due to the high charges involved, the conformations of GDP / GTP are likely to be mainly driven by the electrostatic terms in the empirical potential, and so the problems with biasing the torsional parameters should be negligible.

The problem now is the sheer number of possible conformations. If a systematic search was performed in the way the histamine study was done, and the torsions were again varied by 36° each calculation, this would mean that 11^7 or 19,487,171 calculations must be made for GDP and 11^9 or 2,357,947,691 for GTP. Even molecular mechanics would find this a lengthy process.

It was decided to use the random Boltzmann procedure implemented in the QUANTA package. This utilises the method used in Monte-Carlo simulations. Given an initial minimised structure, it calculates the energy of the system. A random jump by perturbing the variable parameters, in this case the torsion angles shown in F5.5. This produces a different conformation and the energy is recalculated. The new structure is then accepted and saved for later analysis if the energy is lower than the starting structure. If the energy increases, the probability that the new structure is saved is according to the boltzmann factor, exp(-E/RT). The temperature was set at 350K to allow a greater acceptance of high energy conformations and so enable more phase space to be sampled.

Once the structure has moved more than a set RMS deviation from the starting structure (in these runs it was 30°) the saved structure is minimised, and the jump sequence started again. Five hundred such jump sequences were made. It would be hoped that if enough structures are generated, all the common low energy conformations would be seen. These structures could then be analysed for their torsional dependence, and so a picture of the stable species available can be built up.



F5.5. The torsions searched over for a GDP gas phase conformational analysis.

It may be noted that ϕ_6 is not a torsional degree of freedom like the other torsions, it is an angle that defines the pseudo-rotation state of the sugar ring. The Boltzmann

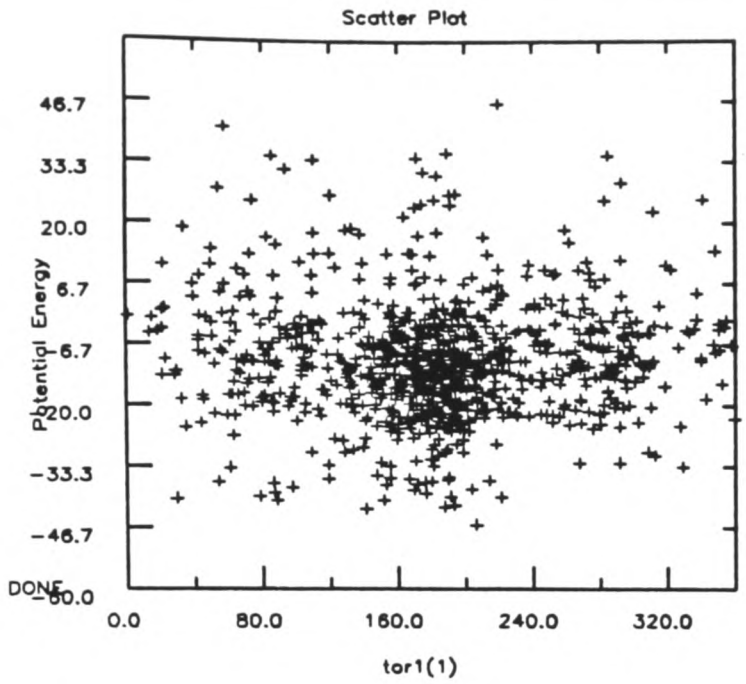
method used was able to induce changes of conformation in the ring and so this angle was sampled.

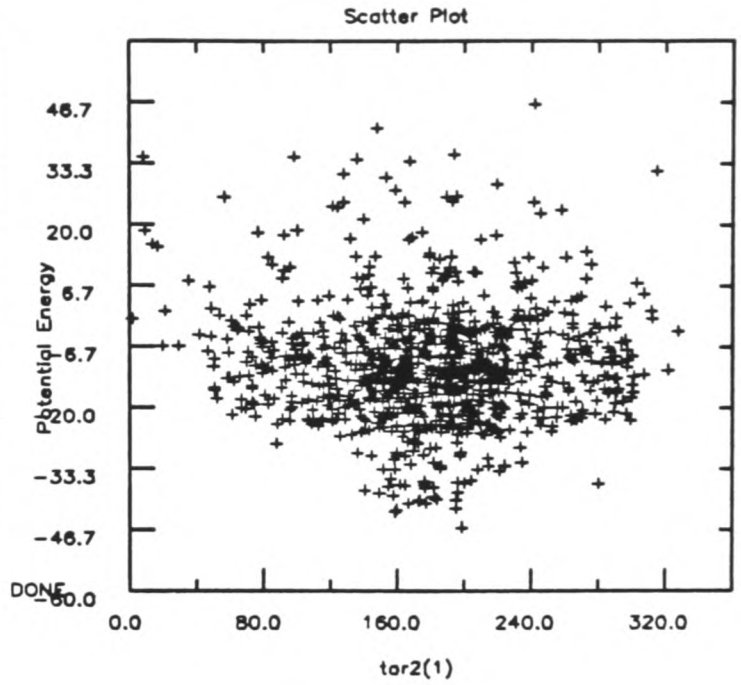
In fact conformational searches on DNA nucleosides have been made earlier, but only on the 5'-monophosphates. Using empirical and semi-empirical potentials effectively the same conclusions were obtained (Yathindra and Sundaralingam 1973, 1973a). Working on AMP, using only non-bonded terms to derive contour plots, they decided that the favoured conformation was what they term the anti-gg form. This has $\phi_7 \cong 360°$, $\phi_4 \cong 180°$, ϕ_5 has a threefold dependence with 60°, 180° and 300° seen above, and the sugar pucker of the ring is seen not to affect the rest of the system. The same conformations were found to be the most stable for the other nucleotides UMP, TMP and CMP. The low energy form of GMP was however discovered to be different, with the guanine bent back across the ribose ring to interact with the phosphate group. These studies correlated well with crystal structures and so the molecular mechanics potential is able to treat the conformational dependence of these systems.

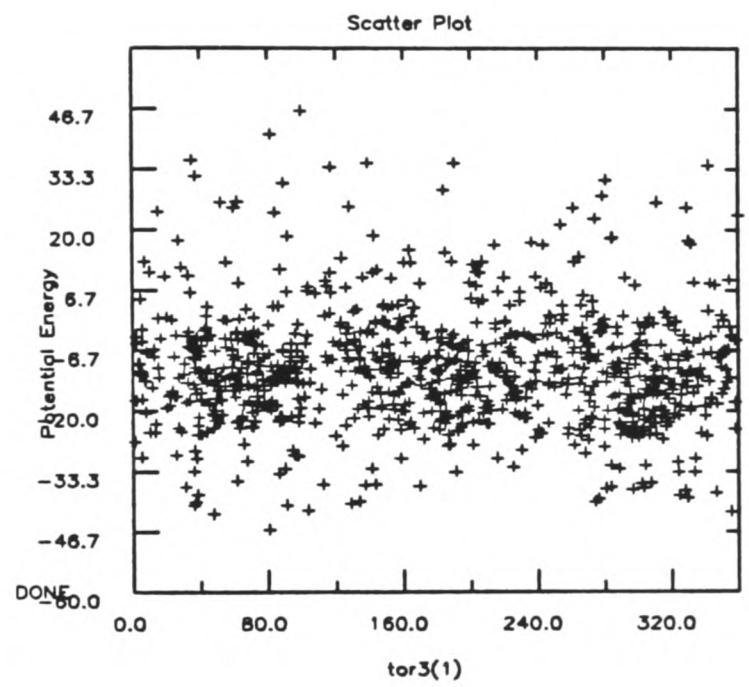
Scatter plots of the potential energy and its variation with the seven torsion angles are shown below in plot P5.7. These plots do show some clustering indicative of minima. As a check, the same calculation was run on histamine monocation N(τ)H over the same torsions in the AM1 conformational analysis in IV.5 (shown in F4.1). A scatter plot from this calculation is shown below in P5.8. Comparing this to the AM1 plot, P4.2, it can be seen that the area around the crystal minimum, marked with the number 3, has been extensively searched, but the calculation has been unable to move into the other minima.

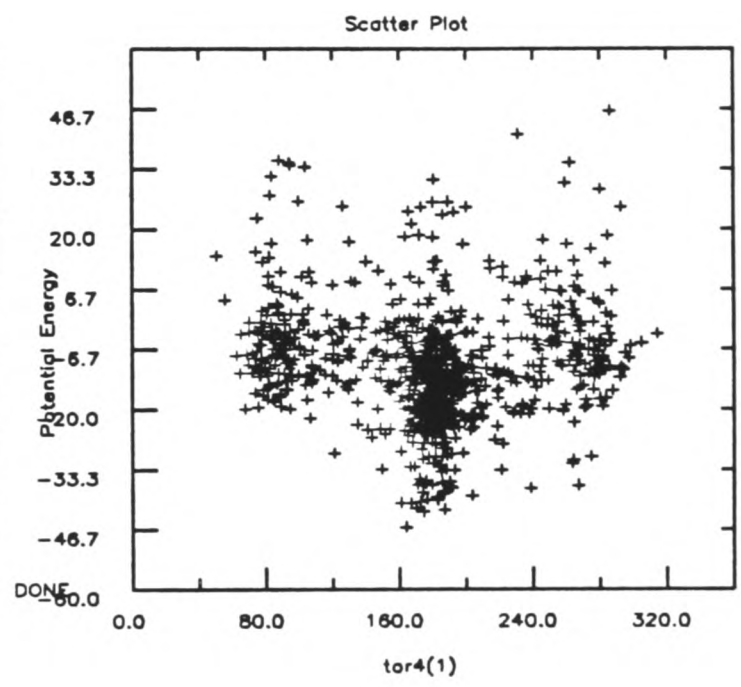
Returning to the GDP scatter plots, ϕ_7 shows two definite minima, one with a value for ϕ_7 around 72° and the other around 288°. The former is the global minimum, lying some 20 kcal mol^{-1} lower than the 288° minimum. From this it might be thought to be the dominating region for this torsion. However, when the lowest energy structure is viewed it can be seen that there is an interaction between the guanosine N2 amino group and the β-phosphate, as shown in F5.6.

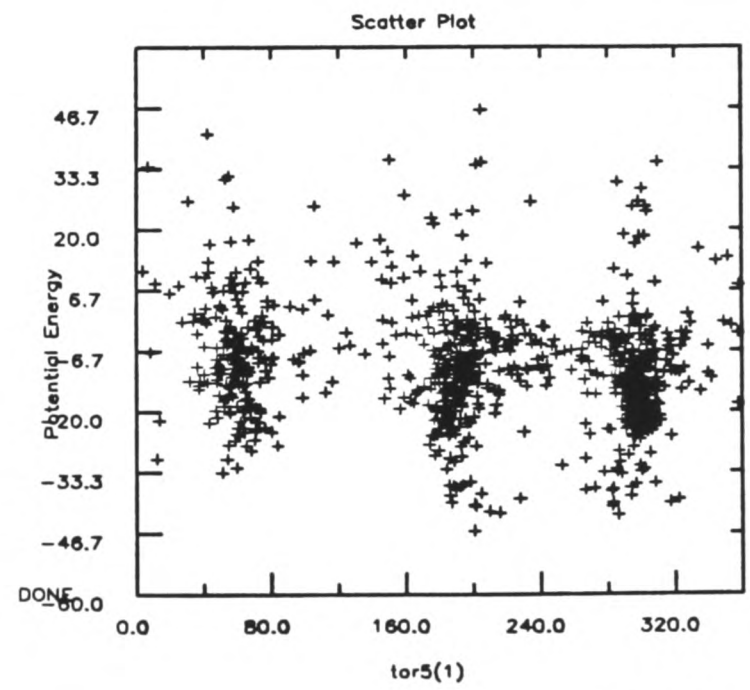
This is exactly the same minimum found by Sundaralingham *et al.* However, while it is still not exactly sure what conformations are seen in solution, data seems to indicate that there is no base - phosphate interaction of this kind. The strongest support for this comes from comparing pK_a values for all the 5' nucleosides. There is effectively no difference in pK_a for the phosphate oxygens at around pH7 between all the bases. Only guanine has an amino group in this position, and if the above interaction was seen then it would affect the electron density, and hence the protonation. Therefore as in the histamine case, the global minimum conformers can be excluded from the calculations. The 288° minimum is also more populated, which may reflect the fact, that there are more ways to arrange the molecule without this inter - molecular interaction. This would again make these conformers favoured in solution.

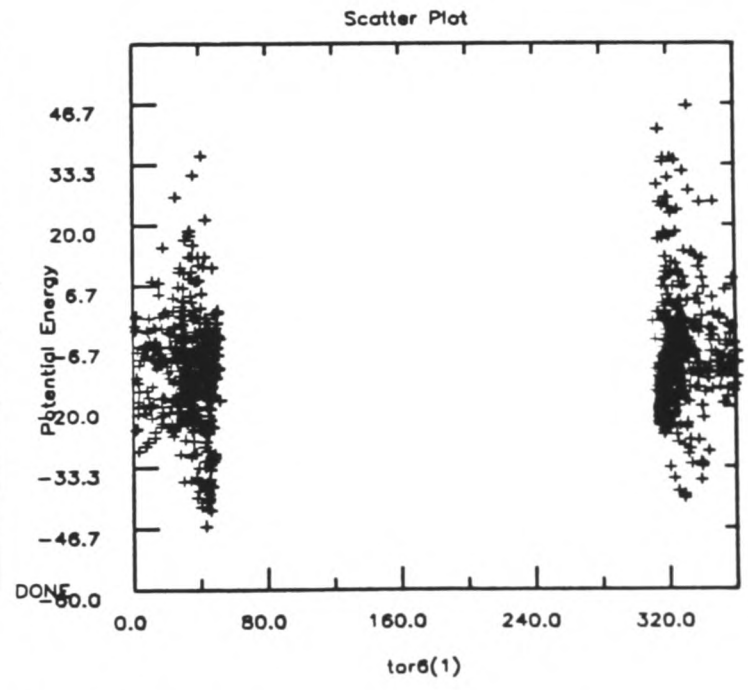




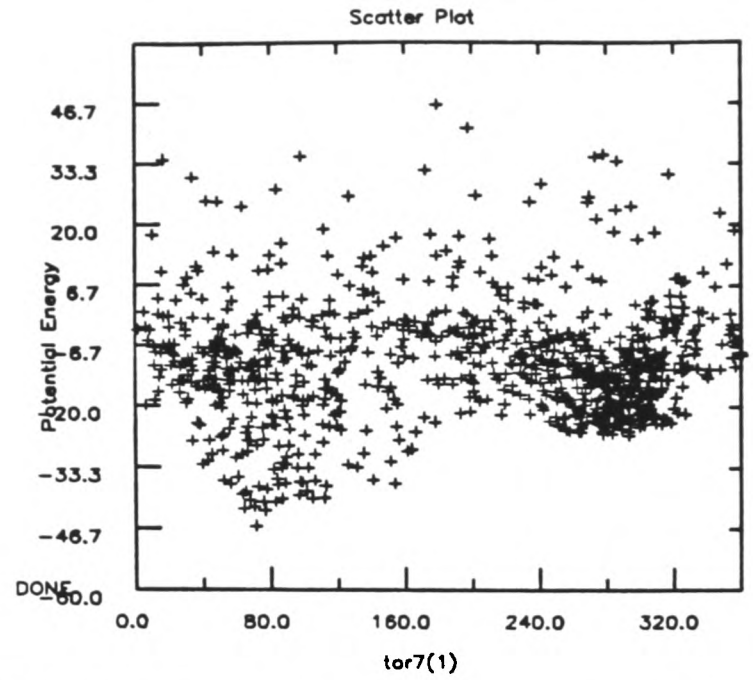




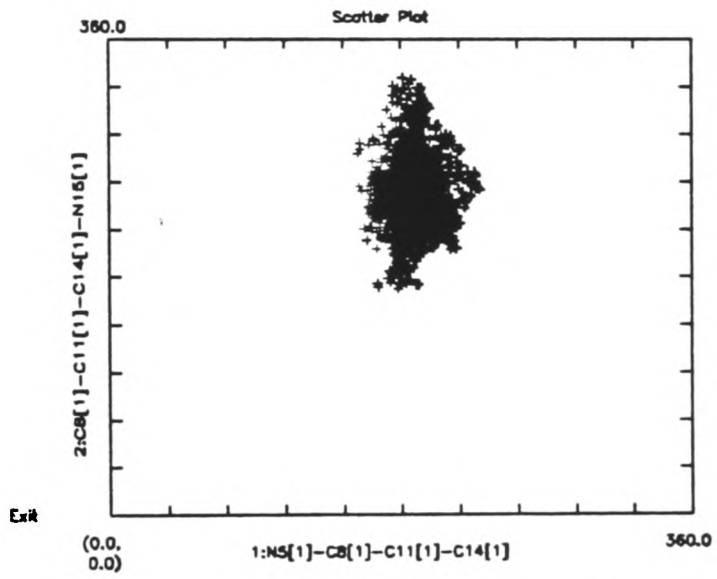




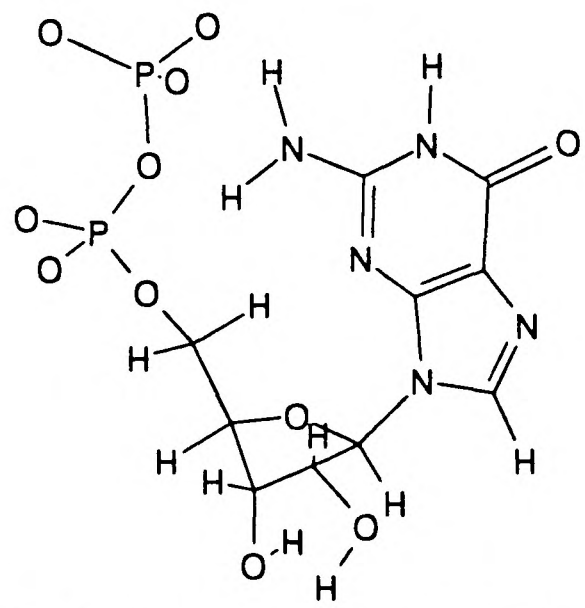
Continued overleaf.



P5.7. Scatter plots of potential energy variation with torsion angle for GDP, calculated by a random jump method. The notation for the torsions is that in F5.5.



P5.8. A scatter plot showing the correlation between the torsion angles in histamine monocation N(τ)H. The x-axis is ϕ and the y-axis ψ as defined in F4.5.



F5.6. The lowest energy structure of GDP found by a conformational search.

ϕ_6 is seen to adopt two basic angles and all values in between. This is in accord with earlier calculations (Levitt and Warshel 1978). These can exist in two minima termed 3'- and 2'- endo, shown in F5.7, with angles of 0-50 and 310-360 respectively as defined in this calculation. If the low energy groups are ignored because of their association with ϕ_7 = 72°, both minima have near identical energies. This agrees with the earlier conformation studies which found that there is a low barrier (around 0.5 kcal mol^{-1}) between these two sugar conformations but otherwise they are near identical.

(a) C3' endo (b) C2' endo

F5.7. The different Conformations of a nucleotide sugar ring. The ring oxygen is behind the c2' - c3' bond.

The remaining torsions are all seen to be very flexible and able to adopt all angles. This is especially true of ϕ_1, ϕ_2 and ϕ_3 which are at the end of the chain. However even for these angles, again ignoring the low lying energy points, there appears to be a three fold dependence. All these minima seem to be of very similar energy.

This agrees with chemical intuition. If the base and the phosphate chain can be taken as separate entities, the base has only one orientation and the chain has multiple minima all of which are staggered forms and therefore identical in energy. Only the angles taken by the torsions nearest the sugar deviate from this. They are hindered by the ring from taking certain angles. The possible conformations are shown below in a tree diagram F5.7.

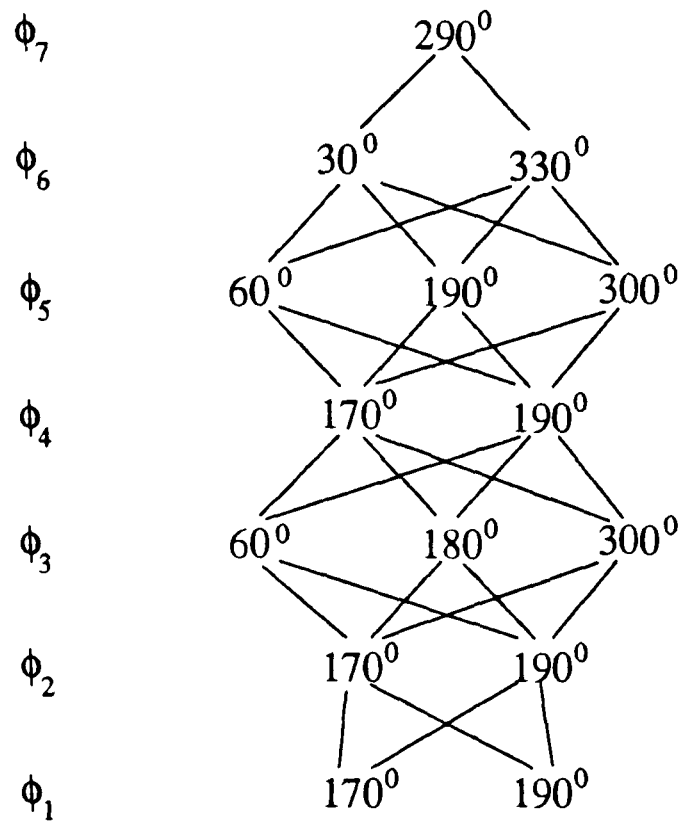
From this analysis, in order to calculate the gas phase minima energies, the fully extended conformation will be taken, with all the phosphate chain torsion angles near to 180°. This may be thought to be especially preferred in solution as there is a greater surface area for the water to solvate. Again this would also be true for GTP.

V.2.2. Charges for GDP and GTP

The conformations relevant to the calculation in solution and also in the protein are now known. The next step is to calculate the *ab initio* fitted charges. At the same time it is possible to calculate the energies to be used in the FEP cycle. These will be discussed in the next chapter; this section will be devoted to the problem of charges.

From the last section the favoured conformations of GDP in solution have hopefully been obtained. As the extended conformations are likely to be equivalent in

energy, it might be expected that the electronic distribution would also be similar for all the possible states. Optimising one of these structures using *ab initio* methods is unfortunately too lengthy a process. A 3-21G*+ basis set for GTP has 178 electrons and 413 basis functions. The calculation of a single wavefunction takes around six hours of IBM3090 cpu time and this is using GAUSSIAN90 which is around a factor of 5 faster than GAUSSIAN88.



F5.7. A tree diagram showing all the possible torsion angles of GDP.

V.2.2.1. Fragmenting GDP to Calculate Force Field Charges.

The original AMBER charges were calculated by splitting up the nucleoside into three parts: the guanine with a methyl group in place of the ribose, the ribose with an amino group in place of the base and a proton in place of the phosphate, and dimethylphosphate (the phosphate in DNA joins two ribose rings together). On cytosine 3'-phosphate with STO-3G, this fragmentation approach gave charges within 10% of those calculated on the full molecule. Fortunately also there was little difference between structures with the 3' or 2' endo ribose.

In the light of the lack of interaction between the base and phosphate chain in solution, this approach seems reasonable. However how well does it work for the boundaries of the fragments? And how well will it cope with the conformation in the protein? These two questions were checked simultaneously using the known structure of GDP in the *ras* protein binding site. Seven calculations were made, one on the full GDP and two each on the fragments. The first set of fragments were split up in an identical way to the AMBER approach, except that ethyl diphosphate was used in place of the dimethylphosphate. The geometries of the fragments were taken directly from the crystal

structure with terminating hydrogens overlaying crystal atoms where possible. These added hydrogens were then CHARMm minimised into place

The second set of calculations on the fragments took the missing group approximation one step further down the chain. The ribose attached to the guanosine was modelled with the C1*, C2* (with attached hydroxyl group), and O4* in the relevant places. These were then terminated with protons representing C3* and C4*. The ribose had the guanine modelled by a dimethylamino group with the methyl groups centred on the C8 and C5 and the α-phosphate present. For the phosphates the ribose was approximated as far as O4* and C3* with its hydroxyl group i.e. 1,2-dihydroxy-3-diphosphatepropane. By going further down the chain it should be possible to see how much affect polarisation through bonds makes.

The charges from these calculations are given in table T5.5. From these numbers it is clear that the fragmentation approach does not produce very good results. Looking at the atoms on the boundary of the guanine and ribose fragments, N9 and C1*, it can be seen that they are very different in the AMBER style fragments but similar in the second set. This shows the importance of including the electronegative groups bonded to these atoms. Away from the boundary guanine shows no significant differences, but ribose does, presumably due to the phosphate which has polarised O4*. There is also little correlation between the phosphate sidechain fragments.

All fragment charges are different from their respective charges on the full molecule. This is probably due to the effect of through space polarisation by the phosphates. For example C8 on the guanine and O4* on the ribose are much more positive with the phosphates present.

While this interaction is not present in solution, it does mean that the nucleoside in the protein cannot be broken up. Combined with the boundary problem and the fact the full calculation has to be done in any case to find the energy, it was decided to only work on the full nucleoside. The solution charges were looked at first.

V.2.2.2. Solution Phase Charges for GDP and GTP

Two conformations were taken from the areas of the CHARMm minima and AMBER minimised. In fact more conformations were tried, but on minimisation the β-phosphate always twisted round to bond to the C3* hydroxyl hydrogen. The CHARMm minimiser produced the same result, demonstrating how sensitive charged species are; a slightly different geometry and the system can move a long way. The torsions for these are given below in T5.6

These two structures were then put into GAUSSIAN and charges calculated from the 3-21G*+ MEP. The results are given in T5.7. After scaling the charge lost was redistributed back with 2/3 onto the β-phosphate and 1/3 onto the α-phosphate. The

T5.5. Charges fitted to a 3-21G*+ MEP calculated on GDP from the *Ha-ras*: GDP crystal structure and fragments of that structure. An H in front of a charge shows that it is a terminating hydrogen in place of the atom.

	Full GDP	guanine fragment 1	guanine fragment 2	ribose fragment 1	ribose fragment 2	diphosphate fragment 1	diphosphate fragment
O3B	-0.162					-1.815	-0.129
O2B	-1.659					-1.543	-1.353
O1B	-1.526					-0.124	-0.715
PB	-0.771				H0.639	-0.643	-0.523
O3A	-0.556				-0.352	-0.354	-0.456
O2A	-0.956				-0.074	-0.042	-0.736
O1A	-0.944				-0.773	-0.110	-0.767
PA	1.977			H0.420	-0.955	1.386	1.815
O5*	-0.690			-0.681	0.053	-0.406	-0.556
C5*	0.019			0.196	-0.320	0.554	-0.169
H5*1	0.121			0.072	0.294	-0.028	0.222
H5*2	0.097			0.013	0.193	0.045	0.116
C4*	0.137		H0.356	0.010	0.137	-0.437	0.283
H4*	0.140			0.104	0.225	H0.177	0.115
O4*	-0.504	H0.131	-0.674	-0.523	-0.056	H0.155	-0.718
C3*	0.272		H0.097	0.336	0.200	H0.183	0.172
O3*	-0.676			-0.755	-0.733		-0.740
HO3*	0.420			0.454	0.513		0.479
H3*	0.196			0.014	0.017		0.086
C2*	-0.010	H0.112	-0.629	0.163	0.297		H0.069
O2*	-0.658		-0.678	-0.751	-0.742		
HO2*	0.455		0.157	0.426	0.447		
H2*	0.222			0.118	0.038		
C1*	0.431	-0.265	0.589	0.680	0.519		H0.485
H1*	0.099	0.165	0.045	0.020	0.086		
N9	-0.716	-0.104	-0.332	-1.167	-0.239		
C8	1.027	0.160	0.267	H0.403	-0.352		
H8	0.072	0.155	0.123		0.080		
N7	-0.537	-0.532	-0.553		H8 0.143		
C5	0.128	-0.116	-0.045		H8 0.143		
C6	0.850	0.866	0.831				
O6	-0.323	-0.650	-0.646				
N1	-0.943	-0.899	-0.882				
H1	0.519	0.440	0.440				
C2	1.141	1.055	1.047				
N2	-0.918	-1.077	-1.077				
H21	0.499	0.456	0.456		H4 0.102		
H22	0.511	0.457	0.459		H4 0.168		
N3	-0.683	-0.776	-0.775		H4 0.124		
C4	0.896	0.423	0.430	H0.448	-0.314		

figures given at the bottom of the table are the total charges on the different parts of the molecule to demonstrate how they change with basis set.

For ease of viewing these charges are plotted in P5.7. It shows clearly that while the four sets are very similar, the major differences are in the boundary regions between the fragments. In general the 6-31G* charges are the outside limits, showing clearly their greater polarity.

Comparing the charges on the two different conformers, it can be seen that they are almost identical on the phosphates, indicating a lack of conformational dependence in this region. The major differences are at the fragment junctions; N9, C8, C5* and O5* are quite different depending on the geometry.

T5.6. The torsion angles of the AMBER minimised GDP in two different extended, and the protein, conformations. Torsion numbers are as in F5.5

torsion no.	Structure 1	Structure 2	Protein
1	168.2	197.3	298.4
2	166.5	201.5	260.0
3	185.8	67.8	327.9
4	166.2	180.4	219.1
5	182.5	308.8	65.38
6	35.03	332.0	29.23
7	287.8	268.1	297.0

Charges were then calculated in the same way for GTP using the fully extended conformation shown in T5.8. These are shown in T5.9. Again the scaled 3-21G*+ scaled charges are very similar to the standard AMBER charges.

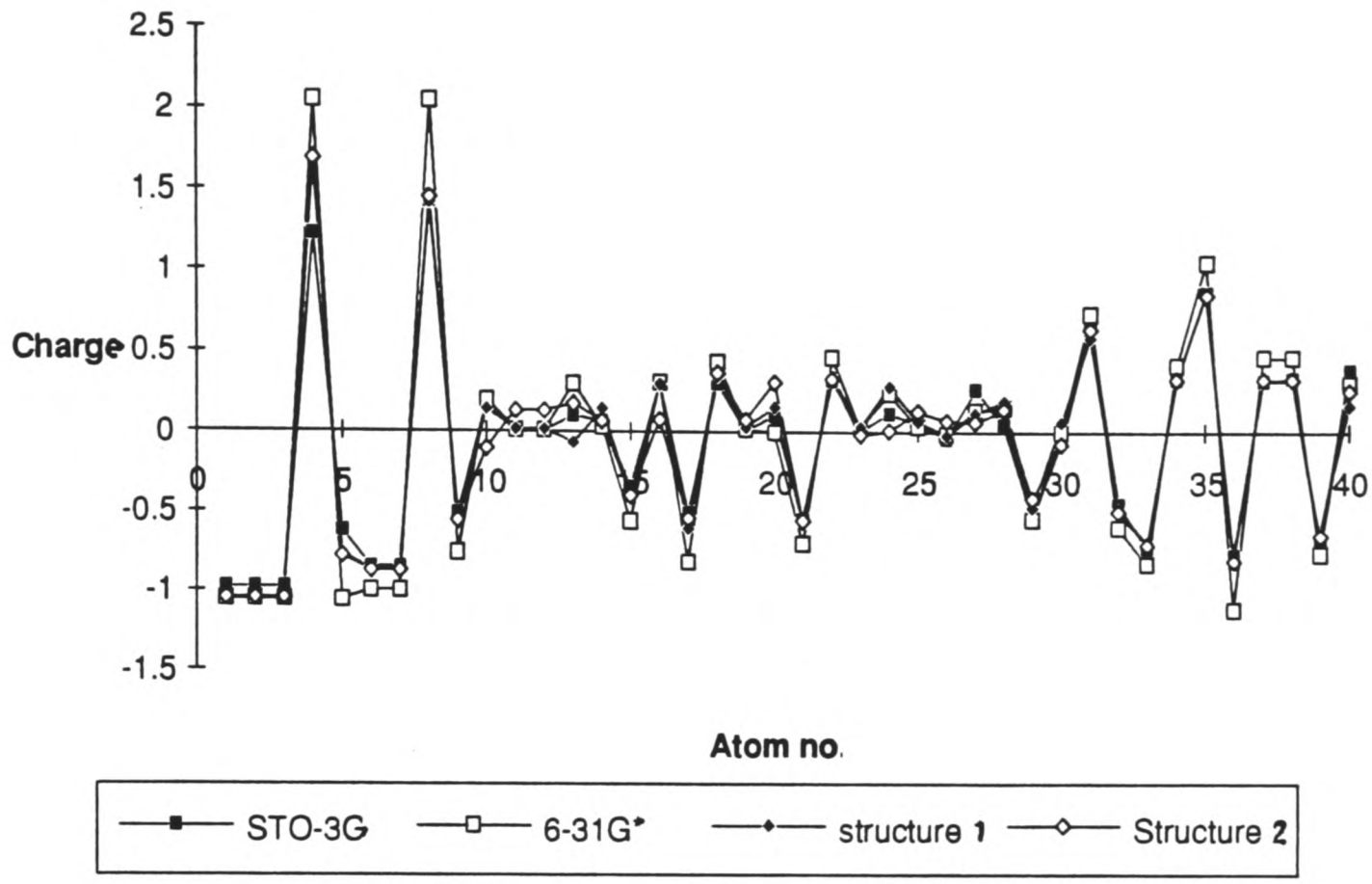
How similar will these electron distributions be to those in solution? The problem is that the higher dielectric constant in water could mean that there is a higher charge separation on the phosphate. This would move electron density from the phosphorus back out onto the oxygen atoms. A major problem is that these are unrealistic simulations. In solution GDP and GTP are protonated. These protons lower the overall charge and possibly allow the "standard" phosphate electron distribution to be resumed. However, without these protons the charge distribution will be different. The question of nucleotide protonation will be returned to in the next chapter.

As well as the charges, the *ab initio* energies of the nucleotides were also obtained. These will be required for the intra-molecular Hamiltonian when perturbation calculations are performed on these molecules in the next chapter. They are given in T5.12. The energies of the GDP structures are not very similar, differing by 0.04846 a.u.'s (141.56 kJ mol^{-1}). Structure 2 is seen to get some stabilisation energy due to its interaction with the

T5.7. Atom centred charges for GDP calculated from a 3-21G*+ MEP in the extended conformations given in T5.6. Averaged means that the charges have been scaled by a factor of 1.34 and averaged over the equivalent atoms.

	AMBER STO-3G	AMBER 6-31G*	3-21G*+ structure 1	3-21G*+ structure 1 averaged	3-21G*+ Structure 2	3-21G*+ Structure 2 averaged
1 O3B	-0.981	-1.054	-1.289	-1.049	-1.285	-1.046
2 O2B	-0.981	-1.054	-1.252	-1.049	-1.266	-1.046
3 O1B	-0.981	-1.054	-1.267	-1.049	-1.246	-1.046
4 PB	1.222	2.049	2.409	1.697	2.404	1.693
5 O3A	-0.623	-1.055	-0.897	-0.771	-0.909	-0.780
6 O2A	-0.850	-0.997	-1.073	-0.865	-1.076	-0.874
7 O1A	-0.850	-0.997	-1.077	-0.865	-1.096	-0.874
8 PA	1.429	2.049	1.991	1.422	2.031	1.452
9 O5*	-0.509	-0.763	-0.670	-0.563	-0.664	-0.559
10 C5*	0.180	0.197	0.187	0.140	-0.141	-0.105
11 H5*1	0.008	0.012	0.066	0.017	0.191	0.132
12 H5*2	0.008	0.012	-0.022	0.017	0.161	0.132
13 C4*	0.100	0.300	-0.091	-0.068	0.232	0.173
14 H4*	0.061	0.028	0.192	0.143	0.092	0.069
15 O4*	-0.343	-0.564	-0.548	-0.409	-0.542	-0.405
16 C3*	0.303	0.310	0.404	0.302	0.106	0.079
17 O3*	-0.509	-0.822	-0.811	-0.606	-0.726	-0.542
18 HO3*	0.306	0.438	0.471	0.351	0.494	0.368
19 H3*	0.007	0.018	0.040	0.030	0.101	0.076
20 C2*	0.101	-0.003	0.209	0.156	0.418	0.312
21 O2*	-0.546	-0.704	-0.771	-0.576	-0.748	-0.558
22 HO2*	0.324	0.466	0.463	0.345	0.449	0.335
23 H2*	0.008	0.026	0.044	0.033	-0.023	-0.017
24 C1*	0.117	0.240	0.381	0.284	0.011	0.008
25 H1*	0.054	0.033	0.092	0.068	0.173	0.129
26 N9	-0.042	-0.010	-0.024	-0.018	0.099	0.074
27 C8	0.266	0.134	0.159	0.119	0.085	0.063
28 H8	0.046	0.153	0.260	0.194	0.192	0.144
29 N7	-0.543	-0.547	-0.635	-0.474	-0.562	-0.419
30 C5	-0.060	0.000	0.082	0.061	-0.095	-0.071
31 C6	0.690	0.741	0.792	0.591	0.867	0.647
32 O6	-0.458	-0.602	-0.649	-0.484	-0.673	-0.502
33 N1	-0.729	-0.831	-0.939	-0.701	-0.948	-0.708
34 H1	0.336	0.422	0.451	0.336	0.449	0.335
35 C2	0.871	1.062	1.163	0.868	1.153	0.860
36 N2	-0.778	-1.120	-1.080	-0.806	-1.087	-0.811
37 H21	0.325	0.471	0.426	0.326	0.433	0.333
38 H22	0.339	0.471	0.448	0.326	0.460	0.333
39 N3	-0.709	-0.766	-0.860	-0.641	-0.868	-0.648
40 C4	0.391	0.312	0.224	0.167	0.352	0.263
PO4	-2.615	-2.113	-2.454	-2.529	-2.443	-2.521
RIB	-0.330	-0.776	-0.363	-0.335	-0.415	-0.373
GUAN	-0.055	-0.110	-0.182	-0.136	-0.142	-0.106

base. This could cause problems when trying to evaluate which conformers to include for the intra-molecular energy in the FEP calculations on these nucleotides. There is also a huge difference in energy between GTP and GDP due to the different number of atoms.



P5.7. The different charge sets for extended GDP. Notation corresponds to that in T5.7.

T5.8. The torsion angles of the AMBER minimised GTP in the extended and protein conformations. Torsion numbers are as in F5.5 with 0 and 01 being the torsions for the γ-phosphate.

Torsion no.	Extended	Protein
0	168.7	200.1
01	170.5	145.3
1	174.9	243.8
2	170.0	185.7
3	188.1	295.1
4	168.4	191.4
5	187.0	59.21
6	32.7	7.382
7	295.9	243.7

T5.9. Atom centred charges for GTP calculated from a 3-21G*+ MEP in the extended conformation given in T5.8. Averaged means that the charges have been scaled by a factor of 1.34 and averaged over the equivalent atoms.

	AMBER STO-3G	3-21G*+	3-21G*+ averaged
O3C	-0.921	-1.280	-1.064
O2C	-0.921	-1.274	-1.064
O1C	-0.921	-1.314	-1.064
PC	1.320	2.444	1.722
O3B	-0.538	-0.952	-0.812
O2B	-0.981	-1.093	-0.879
O1B	-0.981	-1.092	-0.879
PB	1.222	2.108	1.510
O3A	-0.623	-0.847	-0.695
O2A	-0.850	-1.057	-0.854
O1A	-0.850	-1.062	-0.854
PA	1.429	1.933	1.379
O5*	-0.509	-0.615	-0.522
C5*	0.180	0.136	0.101
H5*1	0.008	0.059	0.025
H5*2	0.008	0.007	0.025
C4*	0.100	-0.018	-0.014
H4*	0.061	0.157	0.117
O4*	-0.343	-0.572	-0.427
C3*	0.303	0.443	0.331
O3*	-0.509	-0.827	-0.617
HO3*	0.306	0.478	0.357
H3*	0.007	0.035	0.026
C2*	0.101	0.140	0.104
O2*	-0.546	-0.764	-0.570
HO2*	0.324	0.461	0.344
H2*	0.008	0.036	0.027
C1*	0.117	0.489	0.365
H1*	0.054	0.049	0.037
N9	-0.042	-0.045	-0.034
C8	0.266	0.175	0.131
H8	0.046	0.236	0.176
N7	-0.543	-0.639	-0.477
C5	-0.060	0.084	0.062
C6	0.690	0.770	0.575
O6	-0.458	-0.647	-0.483
N1	-0.729	-0.911	-0.680
H1	0.336	0.444	0.331
C2	0.871	1.157	0.864
N2	-0.778	-1.124	-0.839
H21	0.325	0.440	0.336
H22	0.339	0.462	0.336
N3	-0.709	-0.833	-0.621
C4	0.391	0.224	0.167
PO4	-3.615	-3.485	-3.552
RIB	-0.509	-0.881	-0.721
GUAN	0.103	0.366	0.273

V.2.2.3. Charges for GDP and GTP in the *Ha-ras* Protein Complex

In chapter III, the initial AMBER simulations were first tried with the 3-21G*+ charges in table T5.5 . These were found to give unsatisfactory results, in contradiction to general expectations which would presume that these would give a better description of the electronic distribution than the standard STO-3G AMBER charges. This was found to be true in the phosphate - magnesium interaction potential calculated above. The problem is to find out what went wrong in the protein.

It was mentioned above that the charges should be scaled to mimic the STO-3G level. This had not been done, yet when done this made little difference. A minimisation with these charges showed no differences to the unscaled charges. The problem is that the actual distribution must be incorrect.

It can be seen that the major difference between the AMBER charges set and the 3-21G*+ charge set in T5.5 is on the terminal phosphate group. As in the methylphosphate case, the excess electron density has ended up on the phosphorus atom. From the failure of the simulations it would seem that is not a correct description. Presumably the negative charge on the phosphorus was the reason why the magnesium always migrated between the phosphates where it would see this charge as well as three oxygens.

The error that was made in calculating these charges was not to consider what the charges were meant to represent. They are not meant to reproduce the gas phase electrostatics, but those in the simulation in the protein. In a solution the surroundings have no net permanent dipole. The electronic distribution will thus be similar to that in the gas phase. Changes will be fluctuating due to momentary interactions with solvent molecules, but these will be of the nature of induced charges. In contrast, in the protein the nucleoside is held fairly rigidly in an effectively static field. This could lead to a permanent induction compared to the gas phase. Hence the MEP should be calculated in the surrounding field.

In chapter III it was mentioned that charges with GDP and the binding site magnesium were calculated but gave no difference to the simulation. More of the field was thus needed.

The nucleoside was taken from the AMBER crystal structure with all atoms less than 5Å from it. The MEP was then evaluated as normal on the nucleoside, including the magnesium ion explicitly, with the surrounding atoms as a point charge field. The magnitude of the point charges were taken from the CHARMm force field for reasons of ease.

The atom centred charges fitted to this MEP were very different to those without the surrounding field. This can be easily seen by comparing the total charge on the fragments of the nucleotide. There has effectively been a shift of two electrons from the guanine ring system onto the phosphates. They are shown below both as calculated and

scaled to STO-3G in T5.10. The standard AMBER charges are repeated in this table for ease of comparison.

Interestingly, these charges calculated including the surrounding field are very similar indeed to the standard charges. This is a good victory for the concept of chemical types. It would appear that a phosphate group wants to have a certain electron distribution and the surrounding field stabilises the GDP to allow this to happen.

In minimisations of the *Ha-ras:* GDP complex, these charges keep the structure very close to that of the crystal structure. A short dynamics run also showed no tendency by the magnesium ion to migrate along the phosphates. They were thus taken for use in the simulations.

By this time the *Ha-ras*: GPPNP crystal structure had been obtained from Pai *et al.* (1990). To produce a reasonable *Ha-ras*: GTP structure the NH_2 group between the β and γ-phosphates was simply changed into an oxygen and the structure minimised in AMBER using standard charges.

This was not quite as simple as it sounds. First the system had to be solvated to keep the binding site correct. The easiest way would have been to include the crystal waters. Unfortunately they did not have any hydrogen atoms, and adding protons in a sensible orientation proved to be impossible. The system was thus solvated in the same way as the Ha-*ras*: GDP model in chapter III. First a solvent box was added around the protein. This was then minimised and any gaps in the binding site filled in with a second solvation. This kept the protein almost identical to the crystal structure, with the exception of the GTP and some residues local to the γ-phosphate which moved slightly due to the new oxygen.

The same procedure as above was then used to calculate the MEP fitted charges for the GTP in the protein, based on this model. As with the GDP, these scaled charges are very similar to the standard AMBER charges.

The *ab initio* energies of the nucleotides in the binding site were also calculated, using exactly the same method and system, only with the magnesium ion modelled as a point charge. These energies are included in T5.12. The GTP has been destabilised less by the transition to the protein bound conformation than the GDP. This is because it is closer to the extended conformation, only ϕ_1 is significantly different from the extended conformation, whereas in GDP four torsions, ϕ_1 - ϕ_4 are around 30° from the postualted solution phase minima. This supports a purely ~~kinteic~~ kinetic argument for the nucleotide exchange.

T5.10. Atom centred charges for GDP calculated from a 3-21G*+ MEP on the *Ha-ras* crystal structure conformation . Scaled means that the charges have been scaled by a factor of 1.34. inc field denotes that the protein was modelled as a point charge field around the nucleotide. inc Mg means that the magnesium ion from the crystal was explicitly calculated.

	AMBER STO-3G	3-21G*+ inc field 5Å + Mg	3-21G*+ inc field 5Å+ Mg Scaled	3-21G*+ inc Mg
Mg	2.000	1.938	2.000	1.821
O3B	-0.981	-1.406	-1.194	-1.435
O2B	-0.981	-1.161	-1.011	-1.428
O1B	-0.981	-1.049	-0.927	-0.242
PB	1.222	2.124	1.440	-1.021
O3A	-0.623	-0.809	-0.748	-0.314
O2A	-0.850	-0.969	-0.744	-0.760
O1A	-0.850	-1.013	-0.777	-0.919
PA	1.429	1.652	1.211	1.455
O5*	-0.509	-0.505	-0.398	-0.363
C5*	0.180	-0.280	-0.209	-0.150
H5*1	0.008	0.248	0.185	0.108
H5*2	0.008	0.190	0.142	0.161
C4*	0.100	0.100	0.075	0.252
H4*	0.061	0.136	0.101	0.160
O4*	-0.343	-0.735	-0.549	-0.555
C3*	0.303	0.199	0.149	-0.180
O3*	-0.509	-0.721	-0.538	-0.627
HO3*	0.306	0.457	0.341	0.468
H3*	0.007	0.160	0.119	0.283
C2*	0.101	0.088	0.066	0.321
O2*	-0.546	-0.699	-0.522	-0.744
HO2*	0.324	0.474	0.353	0.513
H2*	0.008	0.126	0.094	0.082
C1*	0.117	0.826	0.616	0.555
H1*	0.054	-0.022	-0.017	0.071
N9	-0.042	-0.234	-0.175	-0.784
C8	0.266	0.181	0.135	1.062
H8	0.046	0.152	0.114	0.031
N7	-0.543	-0.591	-0.441	-0.648
C5	-0.060	0.038	0.028	0.314
C6	0.690	0.865	0.646	0.764
O6	-0.458	-0.719	-0.537	-0.272
N1	-0.729	-1.033	-0.771	-0.946
H1	0.336	0.570	0.426	0.540
C2	0.871	1.189	0.888	1.024
N2	-0.778	-1.156	-0.863	-0.838
H21	0.325	0.555	0.414	0.517
H22	0.339	0.441	0.329	0.491
N3	-0.709	-0.841	-0.628	-0.521
C4	0.391	0.236	0.176	0.752
PO4	-2.615	-2.631	-2.750	-4.664
RIB	-0.330	0.040	0.009	0.356
GUAN	-0.055	-0.347	-0.259	1.486

T5.11. Atom centred charges for GTP calculated from a 3-21G*+ MEP on the *Ha-ras*: GTP crystal structure conformation. Scaled means that the charges have been scaled by a factor of 1.34. inc field denotes that the protein was modelled as a point charge field around the nucleotide. inc Mg means that the magnesium ion from the crystal was explicitly calculated.

	AMBER STO-3G	3-21G*+ inc field 5Å	3-21G*+ inc field 5Å scaled
Mg	2.000	1.859	2.000
O3C	-0.921	-1.081	-1.095
O2C	-0.921	-1.148	-1.095
O1C	-0.921	-1.483	-1.095
PC	1.320	2.107	1.401
O2B	-0.538	-0.746	-0.727
O3B	-0.981	-0.954	-0.729
O1B	-0.981	-0.918	-0.729
PB	1.222	2.004	1.465
O3A	-0.623	-1.076	-0.833
O2A	-0.850	-1.040	-0.917
O1A	-0.850	-1.337	-0.917
PA	1.429	2.021	1.478
O5*	-0.509	-0.198	-0.178
C5*	0.180	-0.065	-0.049
H5*1	0.008	0.162	0.063
H5*2	0.008	0.007	0.063
C4*	0.100	-0.142	-0.106
H4*	0.061	0.014	0.010
O4*	-0.343	-0.253	-0.188
C3*	0.303	0.784	0.585
O3*	-0.509	-0.843	-0.629
HO3*	0.306	0.384	0.287
H3*	0.007	-0.149	-0.111
C2*	0.101	0.026	0.019
O2*	-0.546	-0.568	-0.424
HO2*	0.324	0.384	0.287
H2*	0.008	0.187	0.140
C1*	0.117	-0.084	-0.063
H1*	0.054	0.000	0.000
N9	-0.042	-0.014	-0.010
C8	0.266	0.208	0.155
H8	0.046	0.170	0.127
N7	-0.543	-0.539	-0.403
C5	-0.060	0.027	0.020
C6	0.690	0.956	0.713
O6	-0.458	-0.639	-0.477
N1	-0.729	-0.985	-0.735
H1	0.336	0.559	0.417
C2	0.871	1.138	0.849
N2	-0.778	-1.011	-0.754
H21	0.325	0.319	0.302
H22	0.339	0.490	0.302
N3	-0.709	-0.990	-0.739
C4	0.391	0.426	0.318
PO4	-3.615	-3.792	-3.792
RIB	-0.509	-0.455	-0.370
GUAN	0.103	0.217	0.162

T5.12. The *ab initio* energies of GDP and GTP, both in extended conformations and in the binding site of the Ha-*ras* protein.

Nucleotide	Structure	3-21G*+ energy a.u.'s
GDP	extended structure 1[a]	-2151.57828
GDP	extended structure 2[a]	-2151.62674
GTP	extended[b]	-2714.04157
GDP	protein[c]	-2151.51226
GTP	protein[c]	-2713.98207

a geometry as defined in T5.6

b geometry as defined in T5.8

c nucleotide taken in Ha-*ras* crystal structure geometry with the surrounding binding site modelled as point charges.

V.3. A Summary of How to Calculate Atom Centred Charges for use in AMBER Simulations.

In this chapter the problems of using atom centred point charges to model the electrostatic part of the empirical Hamiltonian in AMBER. This term does not include inductive effects and has been parametrised to overcome this. Therefore it was decided that charges should be calculated by exactly the same method as the original AMBER force field, namely by fitting the atom centred charges to an *ab initio* MEP calculated using the STO-3G basis set.

By the use of comparison with an *ab initio* interaction potential, it was decided that the best way to calculate charges for molecules when this is not possible, e.g. phosphate groups where STO-3G is too simple to manage the high electron density, is to scale charges calculated with a larger basis set down to the level of STO-3G.

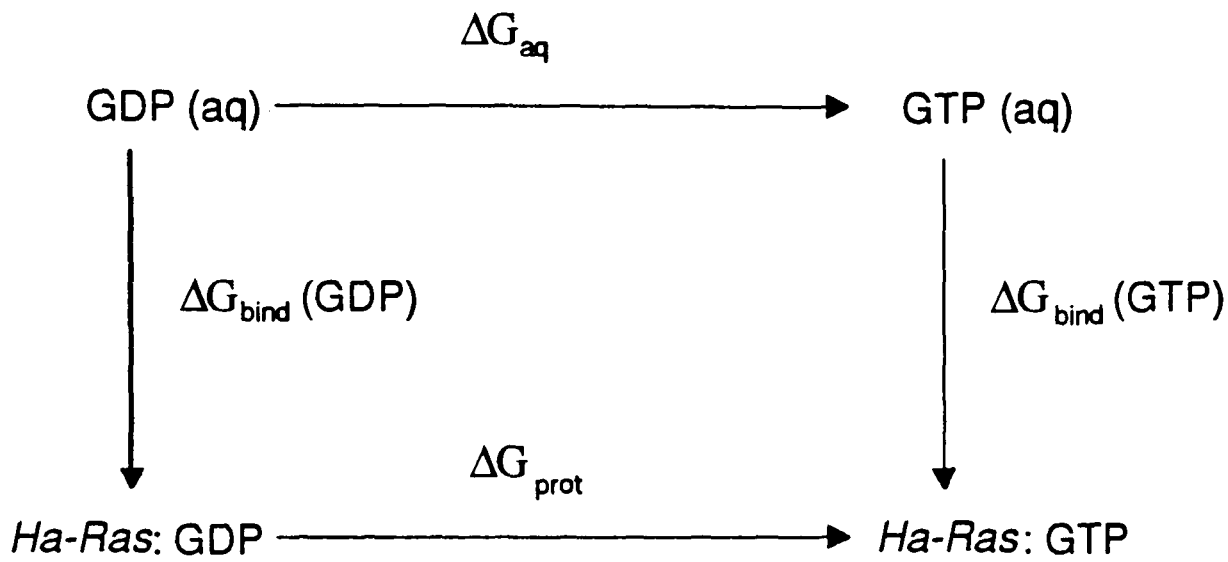
The conformations of GDP and GTP were then analysed to try and evaluate which would be the most common forms in solution. The prescription outlined above was hence used to calculate the charges necessary for solution phase simulations with these molecules.

Finally the nucleotides in the binding site of the Ha-*ras* protein were taken and charges calculated for them. It was discovered that the protein had to be modelled as a point charge field around the nucleotide to get reasonable answers, i.e. ones that work in simulations. When this was done the charges were noted to be very similar to the standard AMBER charges.

CHAPTER VI

The Final Simulations

Due to the limitations of computer time (in 1991 -2, the last year of this project, the Richards group had 90% less supercomputer time allocated than in 1990-1) and the problems encountered, the grandiose schemes laid out in chapter II were modified. Only one perturbation calculation would be attempted, that to calculate the difference in binding constants between the two nucleotides with the Ha-ras protein. This involved evaluating the cycle



F6.1. The thermodynamic cycle needed to calculate the difference in binding free energy of GDP and GTP ~~the~~ the Ha-*ras* protein.

The two values of ΔG_{bind} correspond to the experimentally determined values of K_d given in chapter II. Two simulations are required, mutating GDP to GTP in aqueous solution and then in the protein to calculate

$$\Delta G_{bind}(GTP) - \Delta G_{bind}(GDP) = \Delta G_{prot} - \Delta G_{aq} \qquad (E6.1)$$

or

$$\Delta\Delta G_{bind} = \Delta G_{prot} - \Delta G_{aq} \qquad (E6.2)$$

From the data in T2.2 the experimental value is

$\Delta\Delta G_{bind} = -3.9$ kJ mol^{-1}

This value uses the data from the protein isolated without any bound nucleotide i.e. the values marked "b" in the table. These are considered to be the most realistic.

VI.1. Speeding up the Calculation;Thermodynamic Integration

Even this one calculation using the previous protocols would be huge. The *ras* protein complex in solvent comprises at least 10,000 atoms. The perturbation protocol

used so far needs at least 300ps, this would require 150 hours of CRAY XMP time for one run and a minimum of two runs are needed for an estimation of errors. This size of calculation was unfeasible. Hence a shorter method had to be found.

In chapter IV.2.2.1 the free energy of a system was made a function of a coupling parameter, λ, which linked the state A to B. From E1.36, under isobaric-isothermal conditions, this gives the Gibbs Free energy as

$$G(\lambda) = -kT \ln Q(\lambda) \tag{E6.3}$$

An ensemble average of this expression leads to the FEP master equation (E1.44). Another formula can be obtained by differentiating with respect to λ. This gives

$$\frac{\partial G(\lambda)}{\partial \lambda} = -\frac{kT}{Q(\lambda)} \frac{\partial Q(\lambda)}{\partial \lambda} \tag{E6.4}$$

$$= \frac{\iint \frac{\partial H(\mathbf{p},\mathbf{r},\lambda)}{\partial \lambda} \exp[-H(\mathbf{p},\mathbf{r},\lambda)\beta] d\mathbf{p} d\mathbf{r}}{\iint \exp[-H(\mathbf{p},\mathbf{r},\lambda)\beta] d\mathbf{p} d\mathbf{r}} \tag{E6.5}$$

writting the expression for Q(λ) from E1.28 explicitly. This is again an ensemble average

$$= \left\langle \frac{\partial H(\mathbf{p},\mathbf{r},\lambda)}{\partial \lambda} \right\rangle_{\lambda} \tag{E6.6}$$

Which on integration from λ in state A, λ_A, to state B, λ_B, gives the expression

$$\Delta G_{BA} = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H(\mathbf{p},\mathbf{r},\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda \tag{E6.7}$$

This equation has been applied previously to problems. It can be used in two different ways. The first is known as slow growth (Berendsen *et al.* 1985). A is grown into B by a continuous method. Each point in time is a value λ and the next λ + δλ. If the gradient of the energy, H, between time steps is calculated and summed, this produces the free energy difference. The approximation made is that if δλ is small enough, the ensemble average can be represented by a single value.

This is not a good approximation when flexible systems are studied. Another way is to evaluate the ensemble average of the gradient and sum these. This is the so-called multi-configuration thermodynamic integration (MCTI) method (Straatsma and McCammon 1991). For some reason this method has not become popular, the suprise being that on examination it has many advantages over both the other methods.

Isolated λ can be taken, the system equilibrated at this value and an ensemble average of the perturbation gradient calculated. These points then need to be integrated over λ to give ΔG. The isolation of the points removes the reliance on the previous steps for the ensemble. This means the points can be fairly widely spread and the gaps filled in

by interpolation so saving on the run time needed. Later if a particular area of the plot seems to be badly defined more sampling in this region can be made. Also if one point has not converged well, only this point needs to be redone; in windows the whole run must be made again. The size of the perturbation can also be made very small, in fact it should be very small to give a good tangent. The final advantage of this method is that splitting the free energy into inter and intra components is now exact.

The only disadvantage of the method is the incomplete nature of the data. Given the errors involved in the potential and the sampling, this should not be a problem. In fact fitting a curve to a number of points may give a more accurate result as errors may be smoothed out. Also given that more separate windows can be performed in the same computer time, a better idea of the errors involved can be obtained.

This method is implemented in AMBER4.0. However it does so in a rather strange manner. Using the windowing method, instead of accumulating the exponential of the energy change it accumulates the gradient. These are then numerically integrated to give the final result. This, to my mind, gives the worst of both methods. A full simulation is still needed and if a normal size window is used (typical $\delta\lambda = 0.05$) the gradient is a poor approximation to the true one at that λ. Hence an even longer simulation must be run. Also the double wide sampling is dropped because the integral requires only information in the forward direction.

Because of this AMBER3.1, the version used for all the calculations in this project to date, was modified very simply to collect and average the gradient of the energy difference, rather than the exponential, at a particular value of λ. The gradient was calculated in both directions at each point. If the perturbation is too large, so that the gradient is not the tangent to the free energy curve, these values will vary. Hence this will be a good check on the accuracy. An average of these two gradients will then be a better approximation to the real tangent. Of course poor relaxation may mask this in the same way as it did the windowing problems seen in chapter IV.

VI.1.1. A Test Case; Histamine Monocation Tautomerism

As a test case a run was made on the histamine monocation system studied in chapter IV. The minimised and equilibrated N(π)H tautomer with the conformationally adapted charges was taken and the image charge reaction field was included in the calculation. Nine different points on the path were studied, with λ taking values of 1.0, 0.9, 0.7, 0.6, 0.5, 0.4, 0.3, 0.1, 0.0. This should give a good enough spread.

At each point, 8ps of equilibration was run as before. The value of $\delta\lambda$ was chosen to be 0.01. As the perturbation was so small (again one of the advantages of this method is that the perturbation can be made arbitrarily small), it would be hoped that the value would

converge quicker. 4ps of data collection were then made. This guess turned out to be correct and the value at each point had converged reasonably after this time.

Even though the nine points are independent simulations, at each value of λ the final configuration from the last simulation was taken. This was to allow the system to start as relaxed as possible. The equilibration at the start of each point should be enough to ensure independence. The values are given in T6.1.

T6.1. The accumulated perturbation gradients at various values of λ for the mutation between histamine monocation tautomers. The energies are given for N(π)H ———> N(τ)H

λ	$\langle \partial H/\partial\lambda \rangle_{\lambda f}$ [a] (kJ mol^{-1})	$\langle \partial H/\partial\lambda \rangle_{\lambda b}$ [a] (kJ mol^{-1})	Average (kJ mol^{-1})
1.0	129.018		129.018
0.9	125.675	125.332	125.504 ± 0.172
0.7	88.563	88.482	88.523 ± 0.041
0.6	68.944	68.948	68.946 ± 0.002
0.5	31.916	31.962	31.939 ± 0.023
0.4	19.623	19.644	19.634 ± 0.011
0.3	16.238	16.200	16.219 ± 0.019
0.1	-36.292	-36.631	-36.462 ± 0.170
0.0		-51.137	-51.137

a f and b denote the data collected in the forward i.e λ = 1 —>0, and backward i.e. λ = 0 —>1, directions respectively.

The differences between the forwards and backwards gradients are very small. Especially so in the middle of the simulation when the perturbation is not changing as fast. These low errors indicate that the perturbation is small enough to be considered a true gradient.

The next problem was how to integrate the points. The best method would possibly be to fit the functional form of $\partial H(\lambda)/\partial\lambda$. The constants in the AMBER energy equation change linearly with λ. For example at any point between the end states A and B, the bonding constant, K_b, is given by

$$K_{b,\lambda} = (K_{b,A}\lambda + (1 - \lambda)K_{b,B}) \qquad \text{(E6.8)}$$

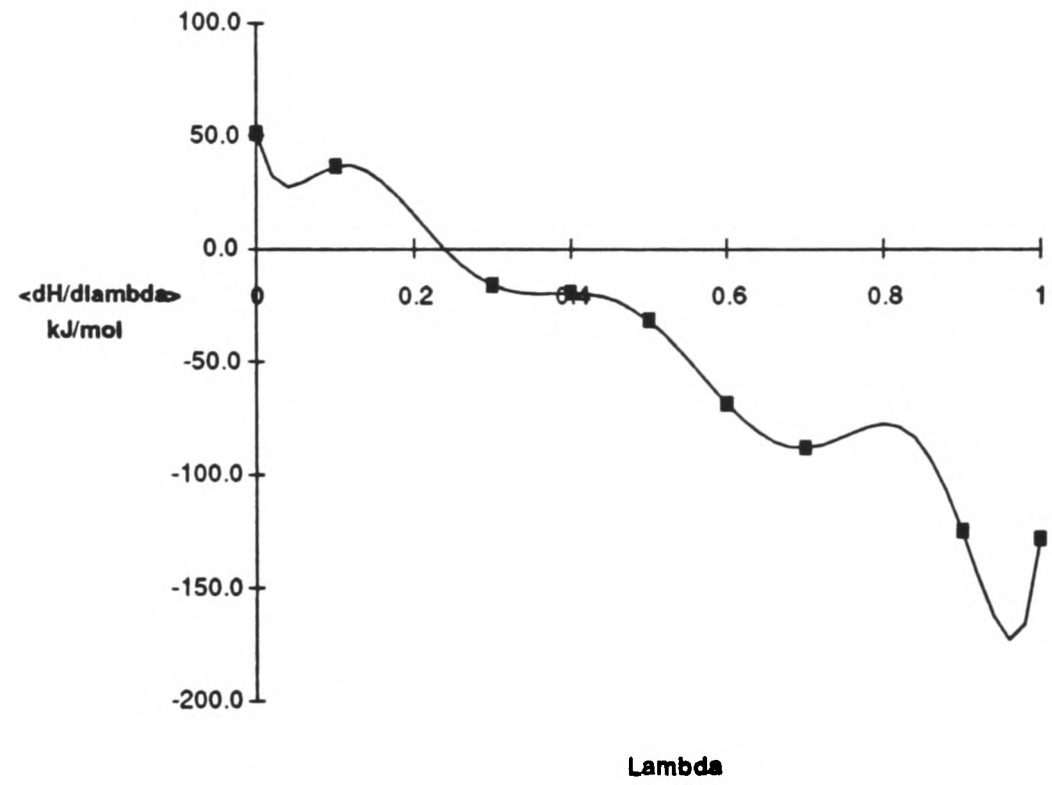
Considering the form of the energy terms in AMBER and absorbing all the force field parameters into constants, this turns out to be

$$H(\lambda)=\sum_{i=0}^{3}c_i\lambda^i+\left(c_4+c_5\lambda+c_6\lambda^2\right)^{\frac{1}{2}}\left(c_7+c_8\lambda\right)^{12}+\left(c_9+c_{10}\lambda+c_{11}\lambda^2\right)^{\frac{1}{2}}\left(c_{12}+c_{13}\lambda\right)^6$$
$$+\left(c_{13}+c_{14}\lambda\right)\left(1+\cos\left(c_{15}+c_{16}\lambda\right)\right) \qquad \text{(E6.9)}$$

The simplest way to express this is as a polynomial and so differentiating with respect to λ gives

$$\frac{\partial H(\lambda)}{\partial\lambda}=\sum_{i=0}^{\infty}c_i\lambda^i \qquad \text{(E6.10)}$$

As only nine points have been calculated, only an eighth order polynomial could be fitted to these points as a higher order polynomial would have more than one solution. As all the values of λ are between zero and one it would be hoped that the higher order terms quickly become negligible. This curve is shown below in P6.1.



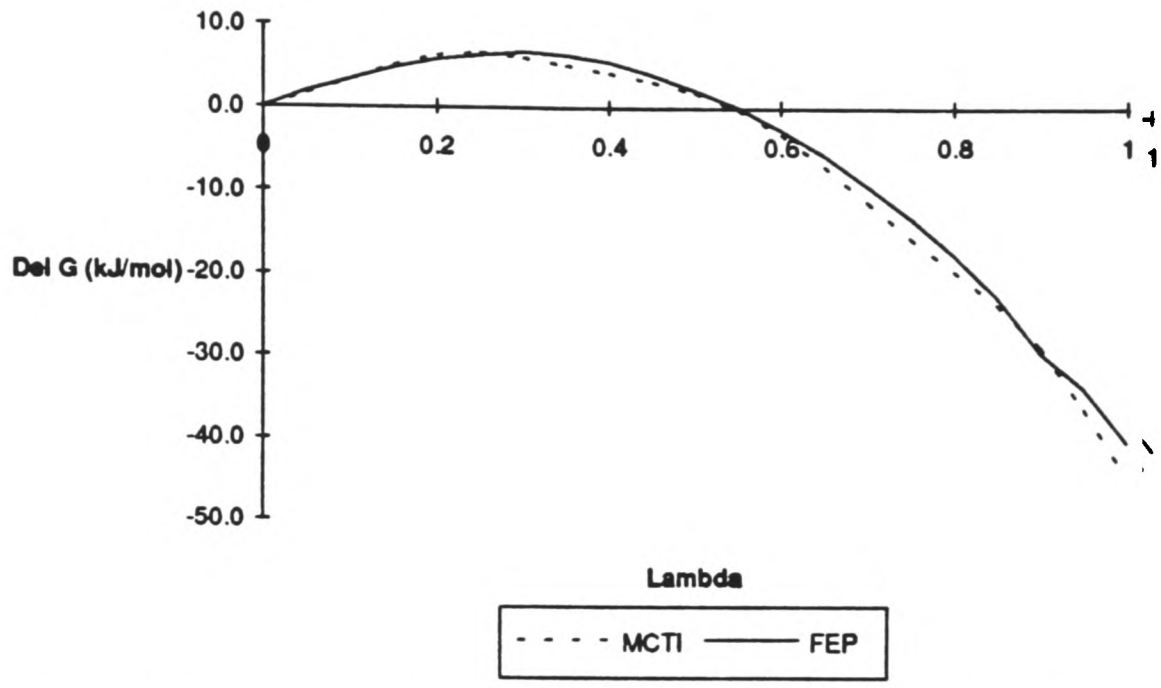
P6.1. A polynomial curve fitted through the points calculated at nine values of λ during the mutation of histamine monocation from N(π)H at 1 to N(τ)H at 0. Values are for $N(\tau)H \longrightarrow N(\pi)H$

When this curve is integrated from 0 to 1, this produces a value for $\Delta\Delta G_{hyd}$ of 44.677 kJ mol^{-1}.This compares very well to the value in T4.20 for the equivalent run using windowed FEP, run 20, of 42.802 ± 2.770 kJ mol^{-1}.

How reasonable is this curve? It looks smooth enough, but the shape of the curve at either end of the run looks a little suspicious. Another way that curves can be fitted to a set of random points is by using what are called spline curves. These are geometrical constructs, smoothly fitting all the points in the way a draughtsman uses a spline tool. This

curve can then be integrated using numerical methods such as Simpson's rule. When this was done, an answer of $\Delta\Delta G_{hyd} = 45.522$ kJ mol^{-1} was obtained. This is not in quite such good agreement with the windowed result. However given the errors involved in these calculations, there is little to choose between the two answers.

One advantage of the analytical curve is that it can be integrated up to various values of λ and directly compared to the windowed result. This is plotted in P6.2.



P6.2. The change in $\Delta\Delta G_{hyd}$ with changing λ during the mutation of histamine monocation from N(π)H at *i* to N(τ)H at *o*. MCTI is the result calculated using the multi configuration thermodynamic integration method at nine values of λ. FEP is the analogous calculation made using the windowing method with 21 windows.

The similarity between the two curves is excellent. The MCTI curve is not quite as smooth as the windowed FEP change in ΔG. However considering that this calculation had taken a third of the time, the accuracy of the result is excellent. The following perturbations on the larger systems all use this method to save time. The integration will also be done analytically on the fitted polynomial.

VI.2. The Calculation of the Difference of Free Energy in Solution Between GDP and GTP.

At this stage everything appeared to be ready for the final calculation. The nucleotide systems had all the necessary parameters. Long range forces should be adequately dealt with and a faster method to make the whole calculation was possible. The only big problems remaining were whether the simulation would give any sort of meaningful results.

The major problem is that the difference between GDP and GTP is enormous. The problem with the size of the perturbation has been eliminated with the MCTI method. However the calculated energies for both sides of the cycle will be large and the difference between the binding constants is small. This automatically produces problems as any errors will be magnified. Given the problems with sampling that are going to be present in solution, the chances are that the result will be completely inaccurate. In fact it could probably be stated that if the result is close to the experimental value it will be more by luck than judgement.

The calculation is in two parts, a simulation in each phase. From chapter IV, it was determined that the best way to calculate a value for ΔG is to break it down into two parts, the inter and intra molecular terms, treating the former by molecular mechanics and the latter by *ab initio*.

VI.2.1. Calculating the Common Conformers of the Nucleotides seen in Solution During a Simulation.

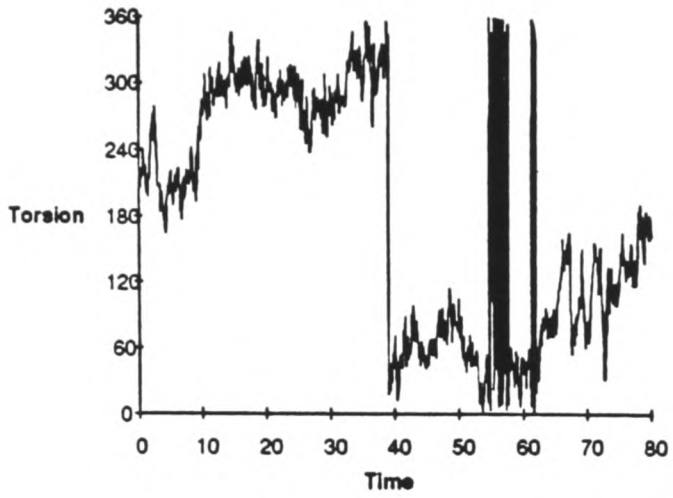
From the calculations made in chapter IV, the first thing to calculate was which conformations are seen most often in solution. This information would then be used to calculate the internal energy and the correct charges for the simulation.

GDP^{3-} and GTP^{4-} were set up in AMBER, using the charges calculated in chapter V and initially in the same extended conformation. These molecules were then solvated. In order to be as efficient as possible, in place of the standard rectangular box of waters, a truncated octahedron was used. This is a shape that approximates closely to a sphere, but is still able to be packed together and periodic boundary conditions applied.

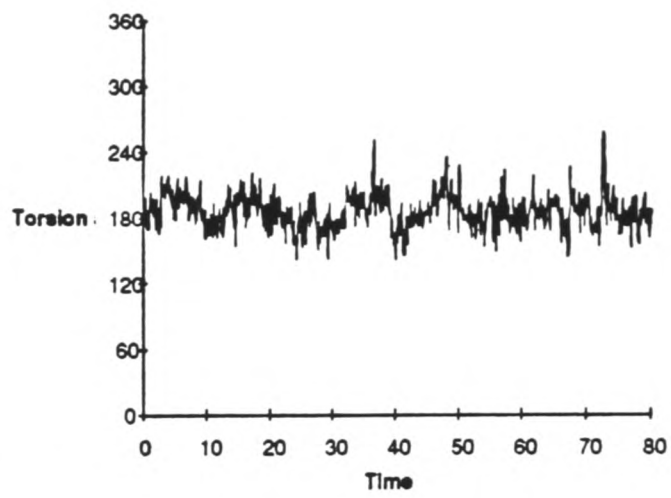
The solvated systems were energy minimised and then heated up to 298K over 4ps, under conditions of constant pressure. Molecular dynamics were then run for 80ps, using a time step of 0.002ps and SHAKE. After each 0.08ps (40 steps), the coordinates of the nucleotide were saved.

The torsion angles of all the free torsions, defined in chapter V, were then calculated at each point in time. The plots of these data are below in P6.3 and P6.4, although torsion ϕ_6 has been ignored in both cases at it remained constant at around 330°. It can be seen that several of the torsions change conformation frequently. Even so, over the 80ps only a fraction of the conformers seen in the gas phase analysis are seen. As predicted, only structures with a ϕ_7 of 280° were seen. In the GDP simulations, it can be seen that the fully extended conformations are not the most common. In fact the torsion angle ϕ_5 takes an angle between 75 and 80 degrees. This corresponds to the phosphate chain stacking over the base ring system. The most common conformers for both GDP and GTP are shown in figures F6.3 and F6.4 respectively.

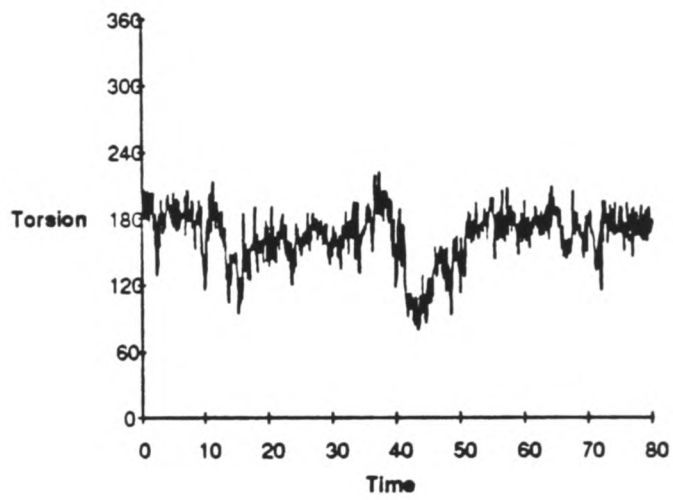
This conformation has in fact been implicated as important in this system by NMR experiments (André *et al.* 1990). In contrast the GTP does not exhibit this conformer. Whether this is a real effect due to the different charge involved, or is just due to insufficient sampling is not clear.



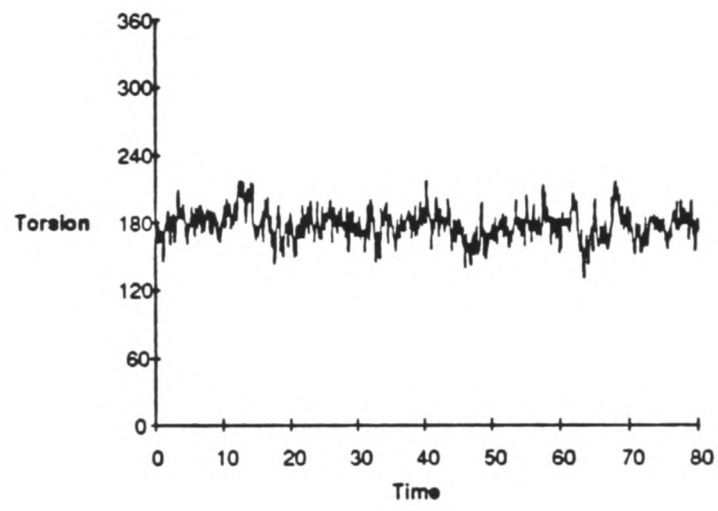
(a) ϕ_1



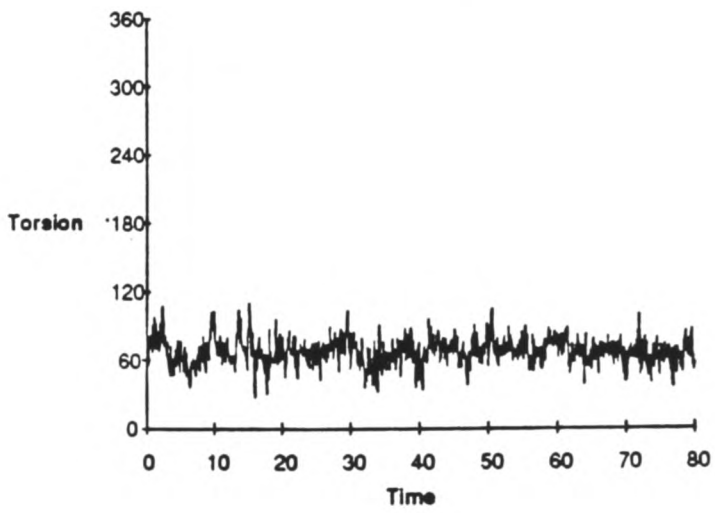
(b) ϕ_2



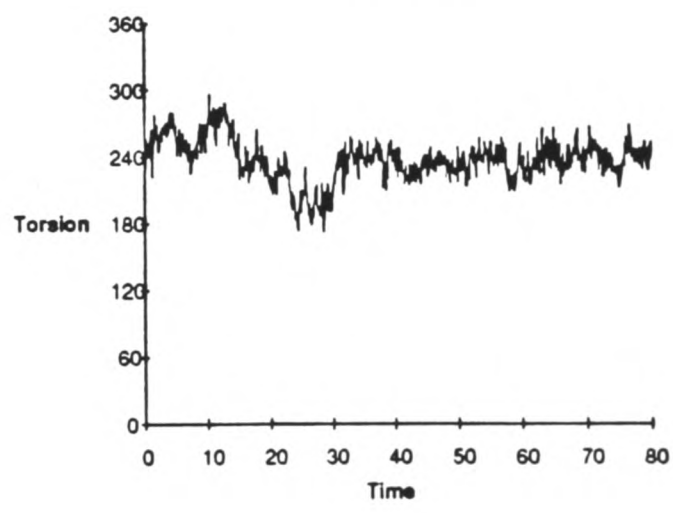
(c) ϕ_3



(d) ϕ_4

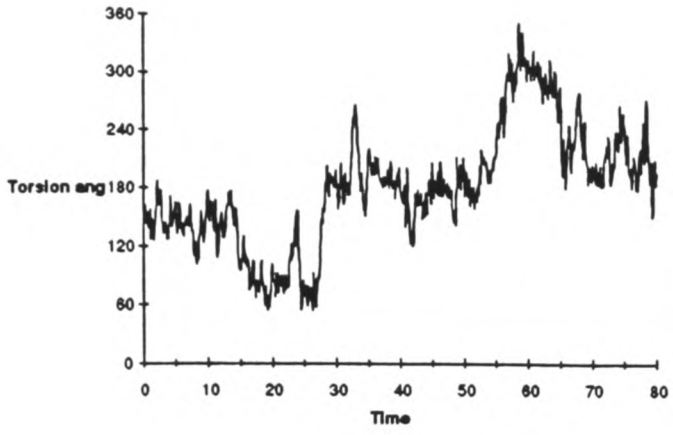


(e) ϕ_5

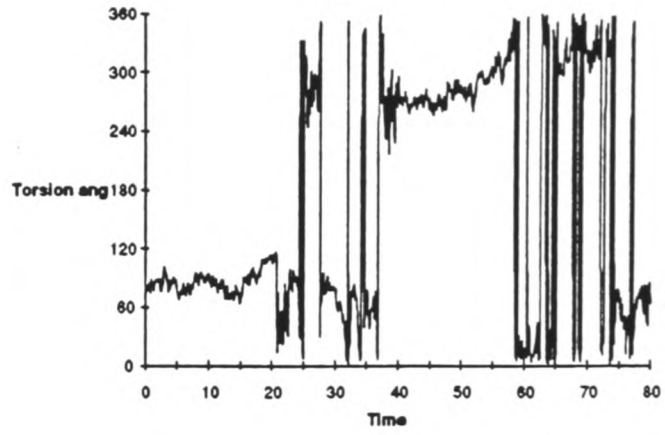


(f) ϕ_7

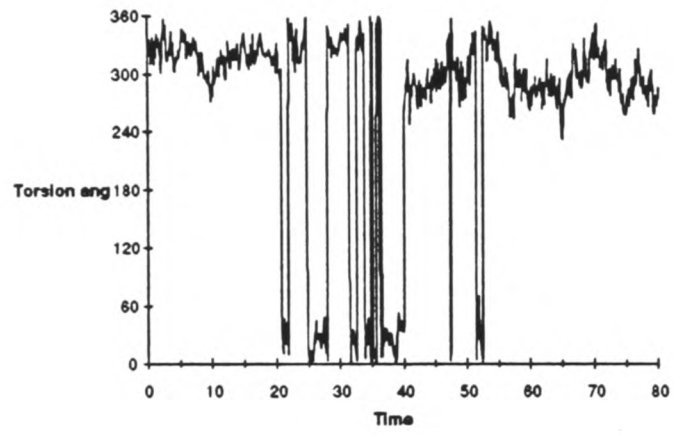
P6.3. The change in conformation of GDP during an 80ps dynamics simulation. The torsions are labelled as F5.6.



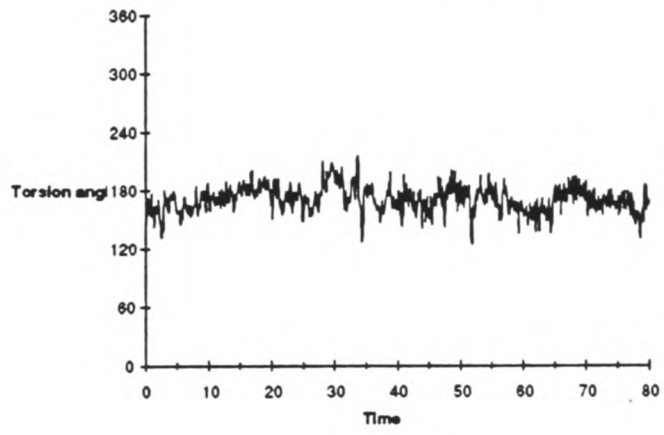
(a) ϕ_0



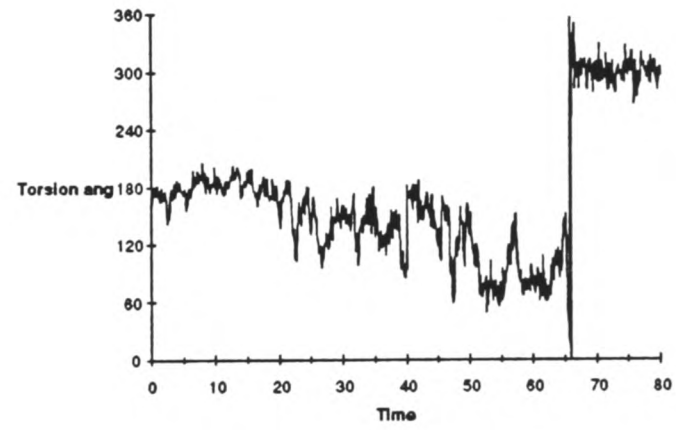
(b) ϕ_{01}



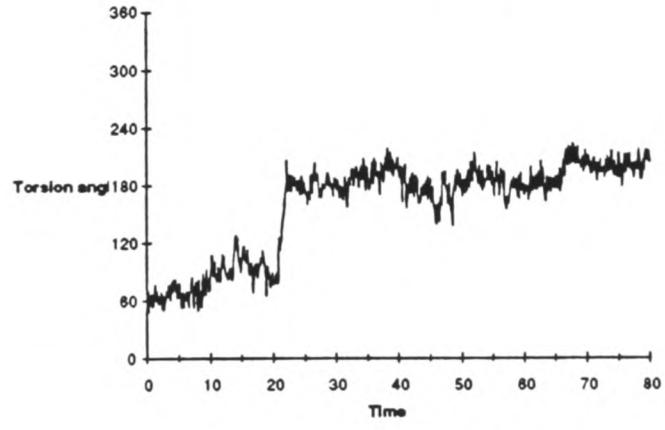
(c) ϕ_1



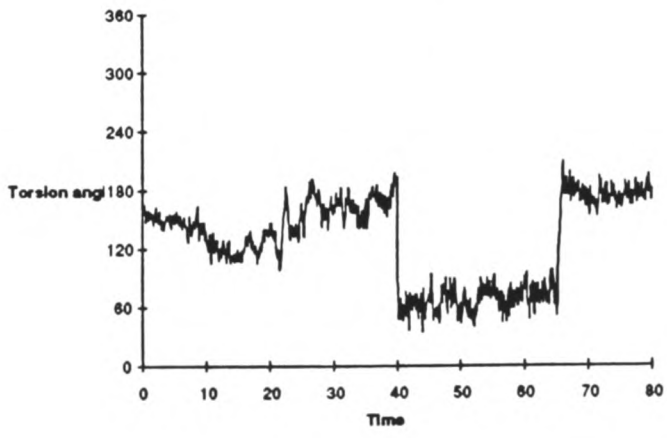
(d) ϕ_2



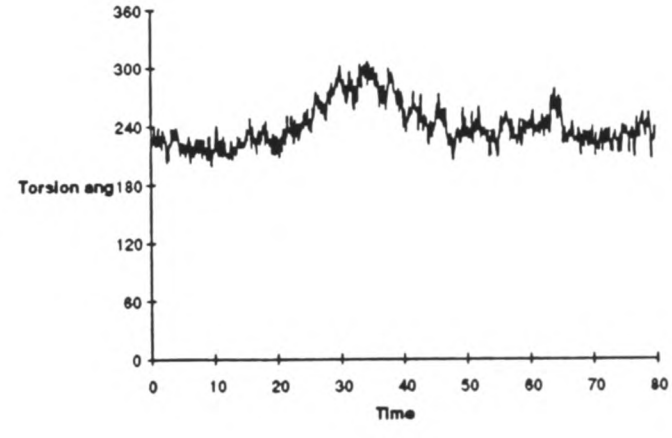
(e) ϕ_3



(f) ϕ_4

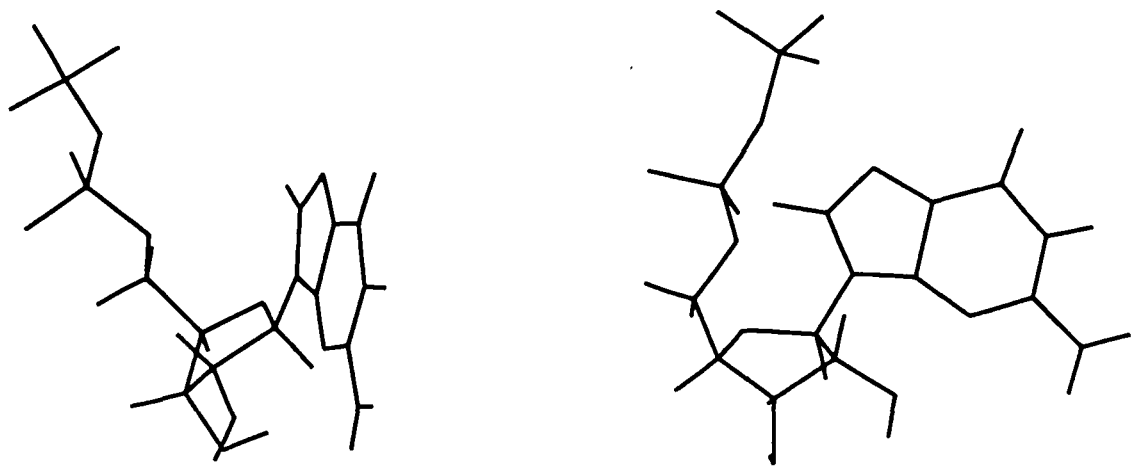


(g) ϕ_5

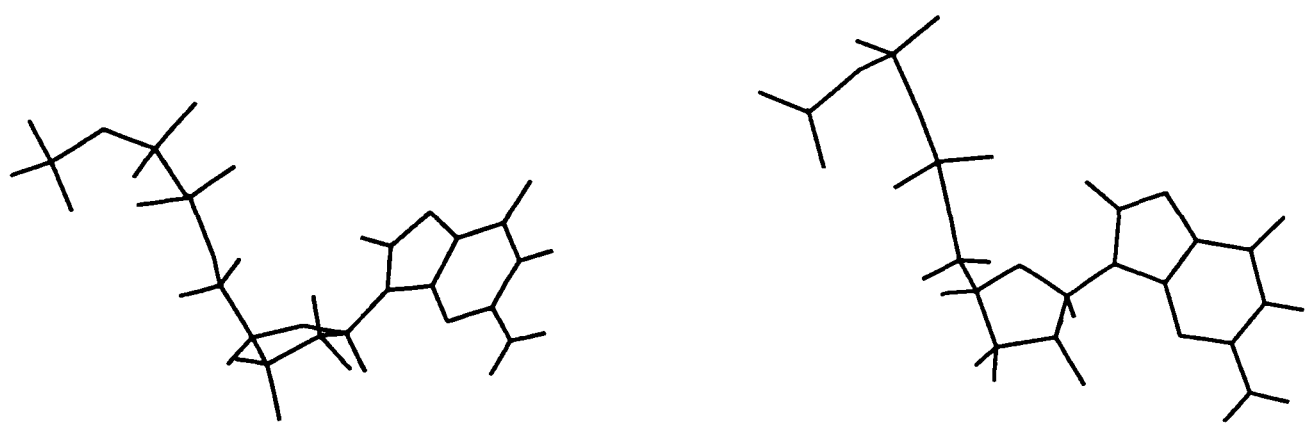


(h) ϕ_7

P6.4. The change in conformation of GTP during an 80ps dynamics simulation. The torsions are labelled as F5.6 for GDP, with ϕ_0 and ϕ_{01}.as the torsion angles along the γ-phosphate.



F6.3. The most common GDP conformer seen during an 80ps MD simulation shown from two angles to show the base - phosphate stacking.



F6.4. The most common GTP conformer seen during an 80ps MD simulation shown from two angles.

Unfortunately it was not possible to find the minima structures for these conformers using molecular mechanics. Due to the high charges, the same problems were encountered as in V.2.2.2. On minimisation the system distorts to find intra-molecular bonds. For example the phosphates in GDP moved across the ring and the amino group on the guanine bent to allow an interaction between the amino hydrogens and phosphate oxygens. This would obviously not occur in solution.

The charges and energies calculated on the extended conformers in the last chapter are thus the best approximation possible to the internal energy and charges for these systems.

VI.2.2. Carrying Out the Perturbations on GDP and GTP in Aqueous Solution.

Perturbations in solution were then carried out using the protocol developed in VI.1.1. Including the reaction field, nine separate calculations were carried out and the ensemble average of the perturbation energy gradient collected. A second set of results was then obtained. Starting with the final configuration at each value of λ, a further 2ps of dynamics was run to ensure a separate part of phase space was sampled. Data collection was then performed for 4ps. All the calculations started with GTP at λ=1 and GDP at λ=0. This direction was decided upon for several reasons. It was thought to be easier to shrink a charge away to nothing, allowing the solvent to relax into the space,rather than suddenly creating a charge which could cause terrible problems. This direction is also preferable when thinking about the perturbation in the protein. Growing a phosphate group inside the protein was shown to be very difficult in III.4.5 Depending on the initial direction of growth, the phosphate can end up in totally the wrong position. This cannot happen if the phosphate group dissapears.

Two runs were performed. The first was using 8ps of equilibration, followed by 4ps of data collection at each value of λ. The second started from the final configuration at that value of λ reached by the first run. Only 2ps of equilibration were used before data collection commenced. The system had run for 12ps at this point and so relaxation is not necessary, this should merely ensure that different regions of phase space are covered. The results are given in T6.2. and T6.3.

T6.2. The accumulated perturbation gradients at various values of λ for the mutation from GTP at λ=1 to GDP at λ=0. Results are for the direction GDP ———> GTP.

λ	$\langle \partial H/\partial \lambda \rangle_{\lambda.f}$ (kJ mol^{-1})	$\langle \partial H/\partial \lambda \rangle_{\lambda.b}$ (kJ mol^{-1})	average (kJ mol^{-1})
1.0	-743.643		-709.590 ± 34.054
0.9	-674.703	-597.295	-635.993 ± 38.698
0.7	-801.136	-736.915	-769.028 ± 32.112
0.6	-767.459	-696.237	-731.861 ± 35.597
0.5	-772.015	-704.987	-738.501 ± 33.514
0.4	-769.659	-705.360	-737.509 ± 32.150
0.3	-794.504	-741.945	-768.228 ± 26.280
0.1	-685.695	-667.063	-676.381 ± 9.318
0.0		-517.034	-519.506 ± 2.473

T6.3. The accumulated perturbation gradients at various values of λ for a second simulation of the mutation from GTP at λ=1 to GDP at λ=0. Results are for the direction GDP ———> GTP.

λ	$\langle \partial H/\partial \lambda \rangle_{\lambda.f}$ (kJ mol^{-1})	$\langle \partial H/\partial \lambda \rangle_{\lambda.b}$ (kJ mol^{-1})	average (kJ mol^{-1})
1.0	-727.790		-694.628 ± 33.162
0.9	-702.573	-629.516	-666.047 ± 36.531
0.7	-786.031	-719.690	-752.861 ± 33.171
0.6	-768.308	-696.536	-732.422 ± 35.886
0.5	-762.906	-697.427	-730.167 ± 32.740
0.4	-781.617	-721.485	-751.551 ± 30.066
0.3	-807.391	-756.907	-782.149 ± 25.242
0.1	-730.560	-708.364	-719.764 ± 11.100
0.0		-482.884	-486.666 ± 3.782

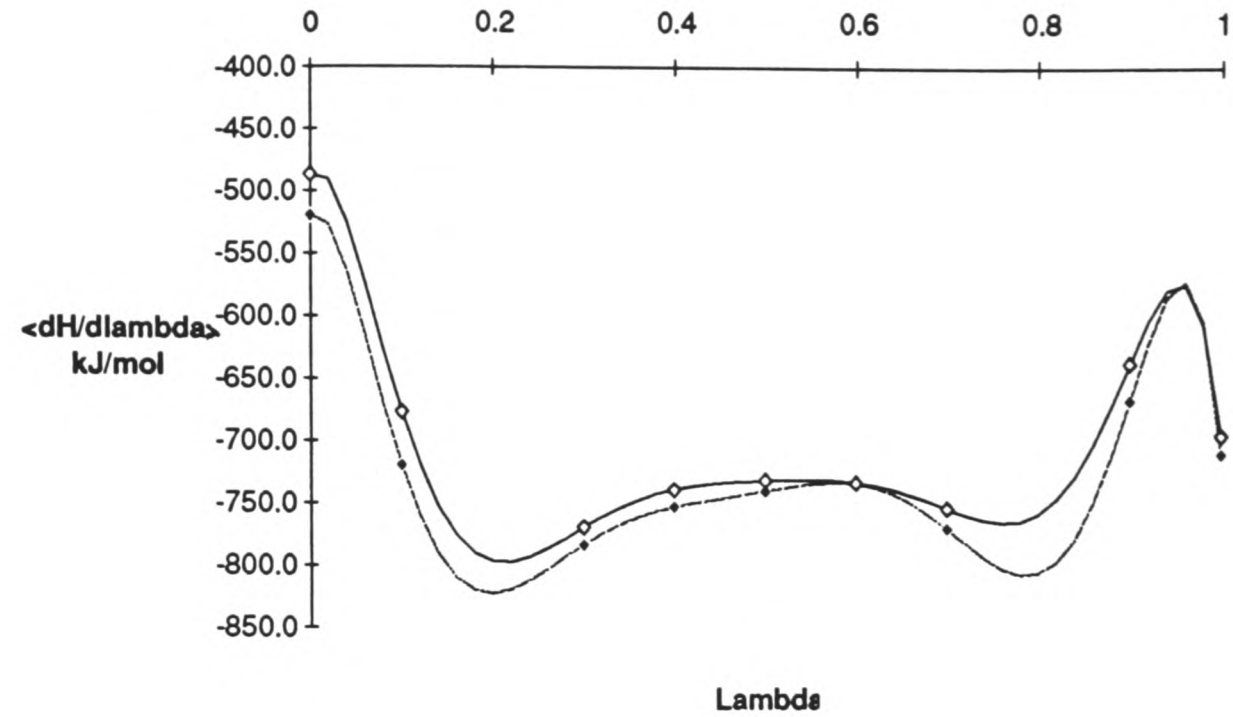
The error on the values can be seen to be much higher than in the histamine monocation case. Because of this it was decided that an estimation of the error at the endpoints would be necessary. This was achieved by fitting a cubic expression to the simulation errors and from this expression working out the likely error when λ is zero or one. A cubic equation was chosen as this seemed like the form of the known points. These errors have been added to the average values in the table.

These two sets of values are independent. It is therefore possible to take the high and low value of $\langle \partial H/\partial \lambda \rangle$ at each λ. A curve can then be fitted to the high and low sets of values and integrated to upper and lower limits of the free energy change. These two sets are plotted below in P6.3[A], with an eighth order polynomial fitted through the points. The difference between the high and low values is very small. The value given is therefore Δ G(aq,inter) = -723.334 ± 10.745 kJ mol^{-1}.

From P6.4[A] it appears that the change in ΔG_{aq} as λ changes is almost linear. A simple model of the solvation energy of an ion is given by the Born equation

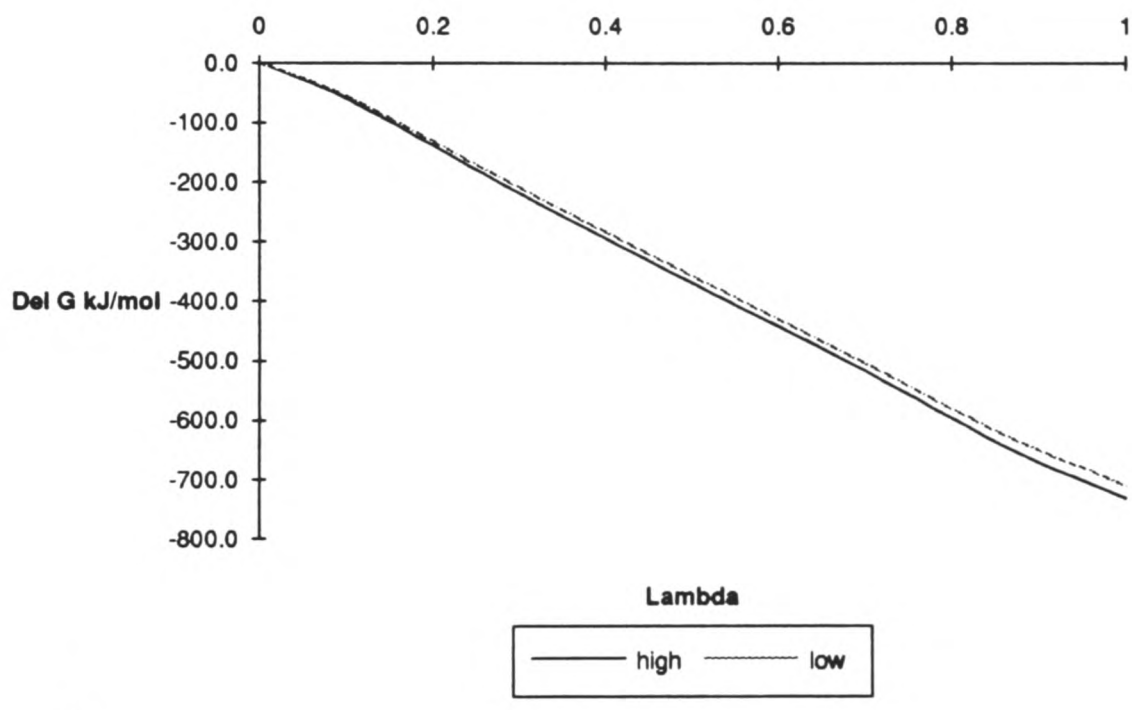
$$\Delta G^{0}_{m,solvation} = -\frac{1}{2}\left(\frac{z_i^2 e^2 L}{4\pi\varepsilon_o r_i}\right)\left(1-\frac{1}{\varepsilon_r}\right) \qquad \text{(E6.11)}$$

where z_i is the charge on an ion of radius r_i. As the radius and charge are both linear with respect to λ, this would predict a linear change of free energy for growing a purely electrostatic system. This indicates that, not surprisingly, the major term in the perturbation is electrostatic.



P6.3[A] A polynomial curve fitted through the points calculated at nine values of λ during the mutation of GTP to GDP. The upper curve are the low and the lower curve the high values taken from two independent runs.

When this is integrated, the following plot is produced



P6.4[A] The change in $\Delta\Delta G_{hyd}$ with changing λ during the mutation of GTP to GDP. High is the curve for all the highest values at that λ from two independent runs. Low is the same, only for the low energy values.

The Born equation calculates that growing an ion of charge 1+ and radius ca 3.5Å i.e. the terminal phosphate group, the gain in free energy would be -195.9 kJ mol^{-1}. This is clearly much less than the value calculated in the simulation. The model is however very simple. In the nucleotide perturbation performed above, there is also a change in charge

distribution, it is not the simple case of growing a 1+ charge. Also the growing ion is not symmetrical and cannot be solvated equally from all sides due to the presence of the base. The surface area available to solvation will be greater for GTP than GDP which might explain the more negative value.

To check on the error introduced by the curve fitting, this value was calculated again using the spline curve interpolation and numerical integration. This lead to a value of $\Delta G(aq,inter) = -719.328 \pm 10.602$ kJ mol^{-1}, which, like in the histamine case, is very close to the fitted curve result. The error on using this method still seems to be fine.

VI.2.3. Calculating the Free Energy of Hydrolysis of GTP.

A useful check on the accuracy of this calculation is to compare it to the energy of hydrolysis of GTP. An experimental value for this exact reaction could not be found, but the value for the hydrolysis of ATP is known. The similarity between the two molecules means that this value could be taken.

ATP ———> ADP + P $\qquad\qquad \Delta G = -30.5$ kJ mol^{-1}

The free energy change calculated above is not the same as this reaction for two reasons. The first is that in solution the nucleotides will be protonated, not in a purely charged state. If the triphosphate and diphosphate are on average bound to different numbers of protons, or if the binding energy of these protons was very different then this energy must be included. The second correction is to include the free energy of formation of the phosphate group that has been removed.

VI.2.3.1. Protonation States of GDP and GTP.

The acidity of a site can be defined by the acid dissociation constant, the equilibrium constant for, ignoring charges,

AH ———> A + H $\qquad\qquad K_a = \dfrac{[A^-][H^+]}{[AH]}$ $\qquad\qquad$ (E6.12)

In GDP and GTP there are various sites that can exchange protons. As well as the phosphate chain, the ribose hydroxyl groups and the guanine heteroatoms can also act as acids or bases. Whether these sites are protonated depends on the pH of the solution. Such a system, called a polyprotic acid, has a complex behaviour. The constants for the important proton equilibria for these molecules around physiological pH are shown in T6.4.

One of the problems in measuring such equilibria is that the method used affects the result. Other values are available (for example pK_4 (GDP) = 7.19 and pK_5 (GTP) = 7.65 are given in Randerath *et al.* (1972)), but the main point is the comparison of series obtained under the same conditions, provided the method is valid at the pH of interest.

T6.4. The acid dissociation constants for GDP and GTP

GDP	GTP
$pK_3 = 2.9$	$pK_4 = 3.3$
$pK_4 = 6.3$	$pK_5 = 6.3$
$pK_5 = 9.6$	$pK_6 = 9.3$

taken from "Dissociation Constants for Organic Bases in Aqueous Solution" Butterworth (1965).

A second problem is to be exactly sure where each proton is associated. As a guide, the equilibria for the simpler systems $H_4P_2O_7$ and $H_5P_3O_{10}$ can be examined, of which GDP and GTP can be thought of as substituted versions. The standard pK_a's for these acids are given in T6.5.

T6.5. The acid dissociation constants for $H_4P_2O_7$ and $H_5P_3O_{10}$

Species	pK_a	Species	pK_a
$H_4P_2O_7$	2.5	$H_5P_3O_{10}$	<1
$H_3P_2O_7^-$	2.7	$H_4P_3O_{10}$	2.2
$H_2P_2O_7^{2-}$	6.0	$H_3P_3O_{10}^-$	2.6
$HP_2O_7^{3-}$	8.3	$H_2P_3O_{10}^{2-}$	5.6
$HP_2O_7^{3-}$	8.3	$HP_3O_{10}^{3-}$	7.9

Comparing the nucleotide values with the phosphoric acid data above, it seems plausible that the values above 9 are not due to the phosphates but the base or ribose. This would then mean that of pK_1 and pK_2 for GDP, one dissociation constant is for the phosphates and one for the protonated amino group on the base. This may seem a very low bascicity for an amine, but it should be remembered that it is conjugated to the guanine ring system, so it is a stabilised form of aniline for which the pK_a is 4.6. Likewise of pK_1 - pK_3 for GTP, one dissociation is for the protonated amino group and the other two for the phosphates. It is known empirically that on the same centre, pK_a's differ by around 5 units. Hence the first two phosphate protons in T6.4. come from different phosphates.

We now have an idea about the equilibria responsible for the equilibrium constants in table T6.4. We want to know the fraction of the molecule present as a particular species. This can be thought of as being formed by the stepwise addition of protons, the opposite equilibria to the acid dissociations above.

$$H^+ + H_{j-1}A \longrightarrow H_jA \qquad K_j = \frac{1}{K_{a,n-j}} \qquad \text{(E6.13)}$$

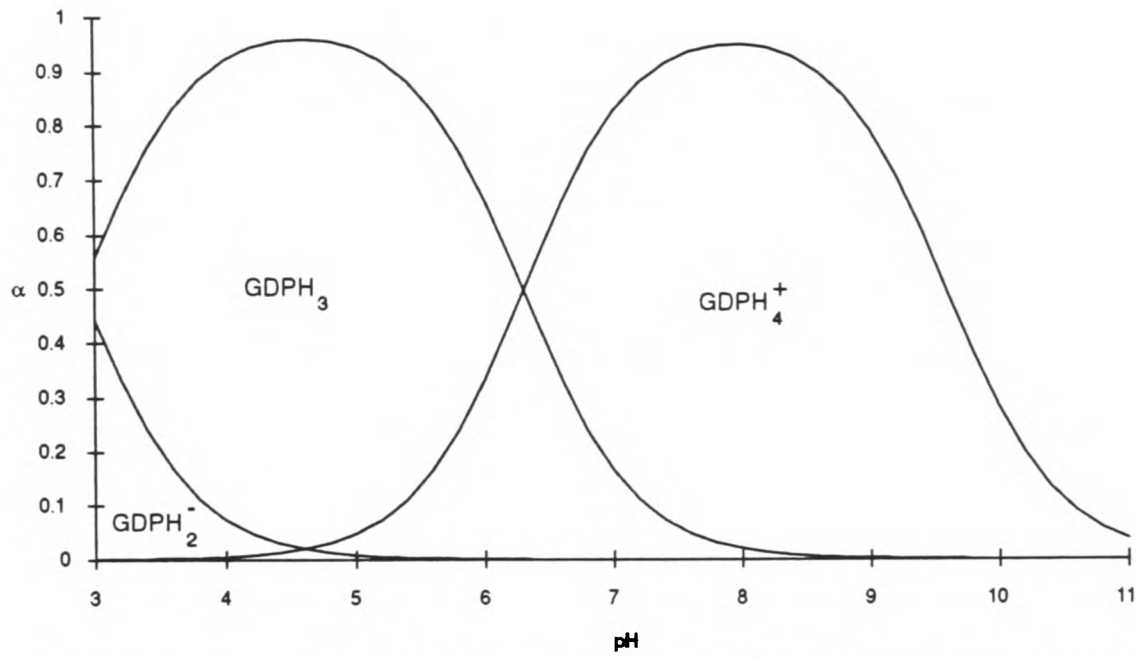
or alternatively as a one step process.

$$jH^+ + A \longrightarrow H_jA \qquad \beta_j = K_1K_2....K_j \qquad \text{(E6.14)}$$

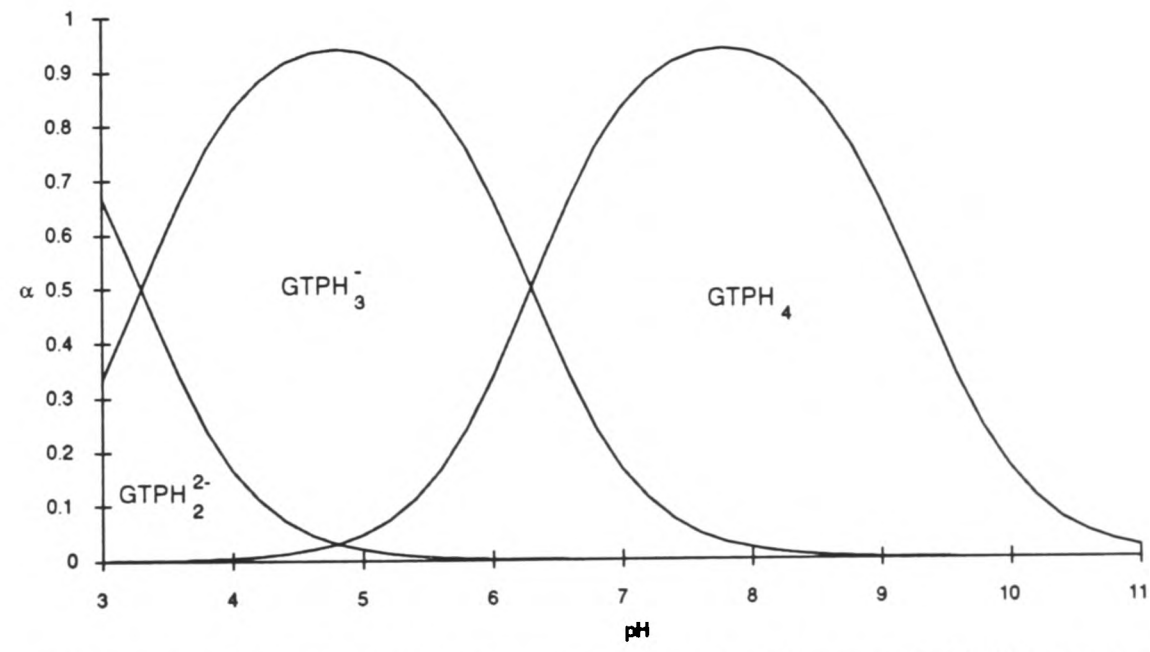
It is now possible to calculate the fraction of a particular species present in terms of these equilibria,

$$\alpha_j = \frac{[H_jA]}{[A_{Tot}]} = \frac{\beta_j[H^+]^j}{\sum_{i=0}^{N} \beta_i[H^+]^i} \qquad \text{(E6.15)}$$

These are plotted over a range of hydrogen ion concentrations for the proton equilibria for GDP and GTP in P6.5 and P6.6.

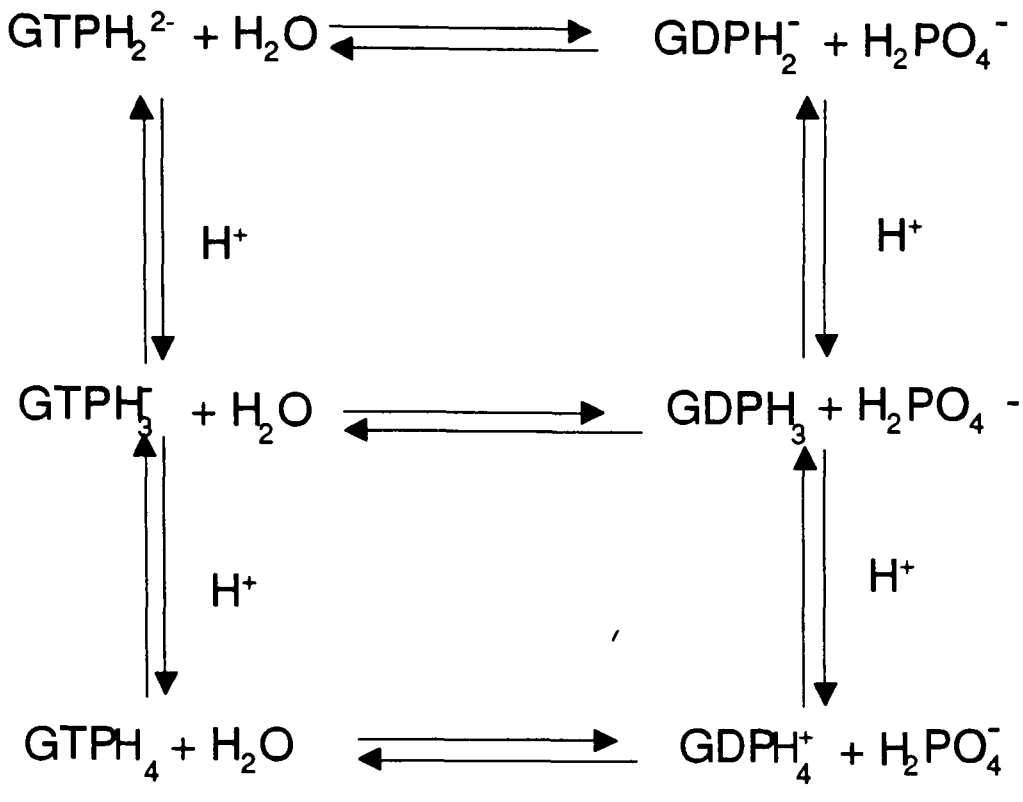


P6.5. The fraction of various protonated forms of GDP present in solution around pH7. The numbers of hydrogens shown are the number of protons gained relative to the 3- ion shown in F5.4.



P6.6. The fraction of various protonated forms of GTP present in solution around pH7. The numbers of hydrogens shown are the number of protons gained relative to the 4- ion.

Including all three possible protonation states, the complete cycle for the hydrolysis reaction is therefore as shown in F6.5



F6.5. The Complete reaction cycle for the hydrolysis of GTP including all the protonation states important around physiological pH.

Fortunately, because the protonation equilibria for GDP and GTP are virtually identical, the energies of the vertical sides in the above cycle can be ignored and only the reaction for the fully charged species needs to be considered. This is very fortuitous. If the proton equilibria were be important all the three equilibria must be studied and the appropriate average taken.

VI.2.4. Including the Intra - Molecular Part of the Hamiltonian to the Perturbation Energy

The *ab initio* energies must now be added to the calculated solvation energy difference in order to regain the full free energy. These energies are given in T5.12. In the case of the histamine tautomerism, the difference in intra - molecular free energy was simply the *ab initio* energy difference corrected to a free energy at 298K. This is not so in this case. The difference between the values for GTP and GDP is large; of the order of 550 a.u.'s (around 1.6 x 10^7 kJ mol^{-1}). The reason for this is that the two molecules have different numbers of atoms and so the dissociation limits, the energy zero for these calculations, is different for the two molecules. It is thus necessary to normalise these results.

The simplest way to do this is to add the energy of the missing particles to the equation i.e.

$GDP^{3-} + P + 3O + e^- \longrightarrow GTP^{4-}$

The problem is how to deal with the electron. The Hartree - Fock energy of atoms is a simple problem. The energy states of a free electron are the solutions to the particle in a box problem. The energy it would have on the dissociation of the GTP molecule will be its orbital ionisation energy. This orbital will be high in energy and so this should be hopefully very small. We are after an order of magnitude calculation to see how reasonable the simulation result is, and for this reason it will be ignored. From this, as these are internal energies,

$$\Delta U(g,intra,0K) = E(GTP^{4-}) - E(GDP^{3-}) - E(P) - 3E(O)$$

The Hartree - Fock energies of phosphorus and oxygen had to now be calculated. To ensure maximum cancellation of systematic errors, this was performed using the same basis set, 3-21G* with diffuse functions. These turned out to be

E(UHF, P) = -339.0632939 a.u.'s

E(UHF,O) = -74.41319190 a.u.'s

Hence

$\Delta U(g,intra,0K) = -0.1604184$ a.u.'s = -468.6 kJ mol^{-1}.

This seems reasonable. The next problem arises when the correction from this internal energy to a free energy has to be made. It is impossible to perform a normal mode analysis on these systems, it is necessary to first optimise the structure to work around the minimum. Semi - empirical methods are not reliable enough for these molecules and *ab initio*, as already explained, is just too big. This correction term could be quite large as there are different numbers of bonds and degrees of freedom. Hopefully the fact that both systems are large will mean that this difference is again small and ignoring this calculation will not be crucial.

Ignoring the correction, we have

$\Delta G(aq) = -1191.934$ kJ mol^{-1}

This energy is however not the energy of GTP hydrolysis. The full energy is

$$\Delta G(hydrolysis) = -\Delta G(aq) - \Delta G_f(H_2O) + \Delta G_f(H_2PO_4^-)$$

The free energy of formation of water is -237.2 kJ mol^{-1} and that for the phosphate ion -922.4 kJ mol^{-1}. Both these values are from standard tables (Kaye and Laby, Longmans). Combining these values, the result is that

$\Delta G(hydrolysis) = +506.7$ kJ mol^{-1}

This result is obviously the wrong sign and much too large. Interestingly if the value for the free energy difference between GDP and GTP estimated from the BORN equation is used in place of the simulation result, the value is -20 kJ mol^{-1}. Assuming that all the approximations made in the estimation of the hydrolysis energy are not too bad, this indicates that the simulation energy is much too high, overestimating the stability of the GTP ion.

VI.3. The Effect on the Free Energy of Different Charges

What could be causing this high energy? Reasons are, as always, sampling and relaxation problems and doubts about the parameters. Ignoring the sampling as one problem that cannot possibly be dealt with properly, a further calculation was made to check the parameters.

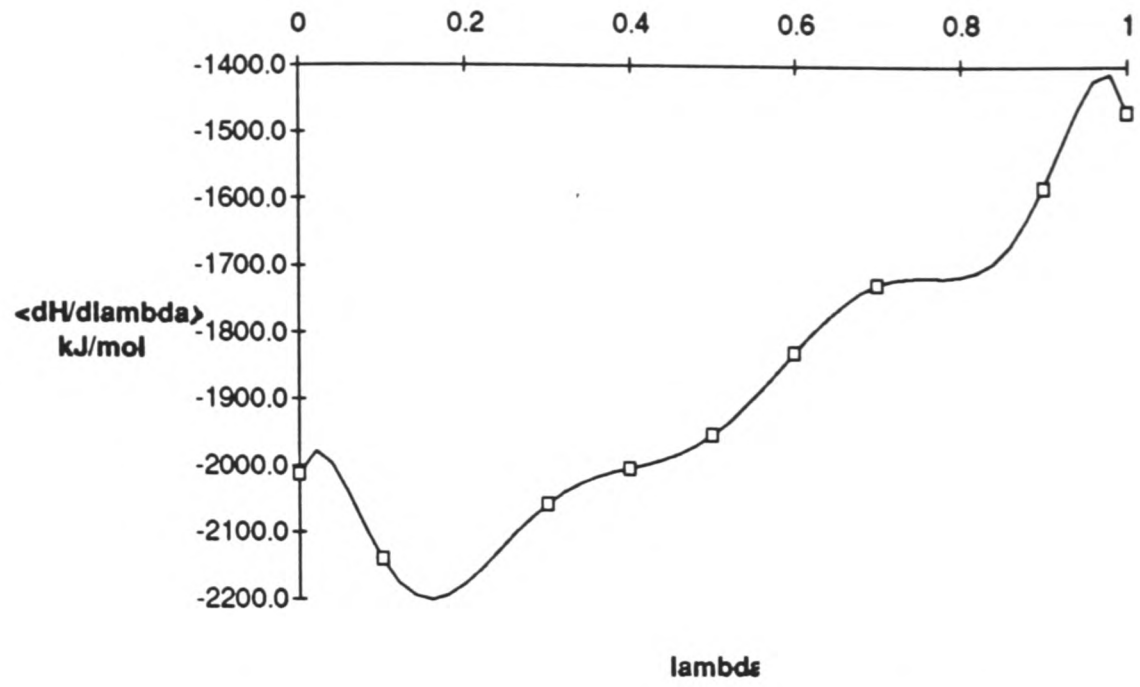
The major error in the parameters must be considered to be the atom centred charges. Rather than calculating new charges with higher basis sets, which would be rather time consuming and possibly less accurate, it was decided to repeat the calculation, only running one simulation, using the standard AMBER charges. This would at least show how sensitive the calculation is to different charges. This was duly performed and the results are given in T6.6.

T6.6. The accumulated perturbation gradients at various values of λ for the mutation from GTP at λ=1 to GDP at λ=0 using standard AMBER charges. Results are for the direction GDP ———> GTP.

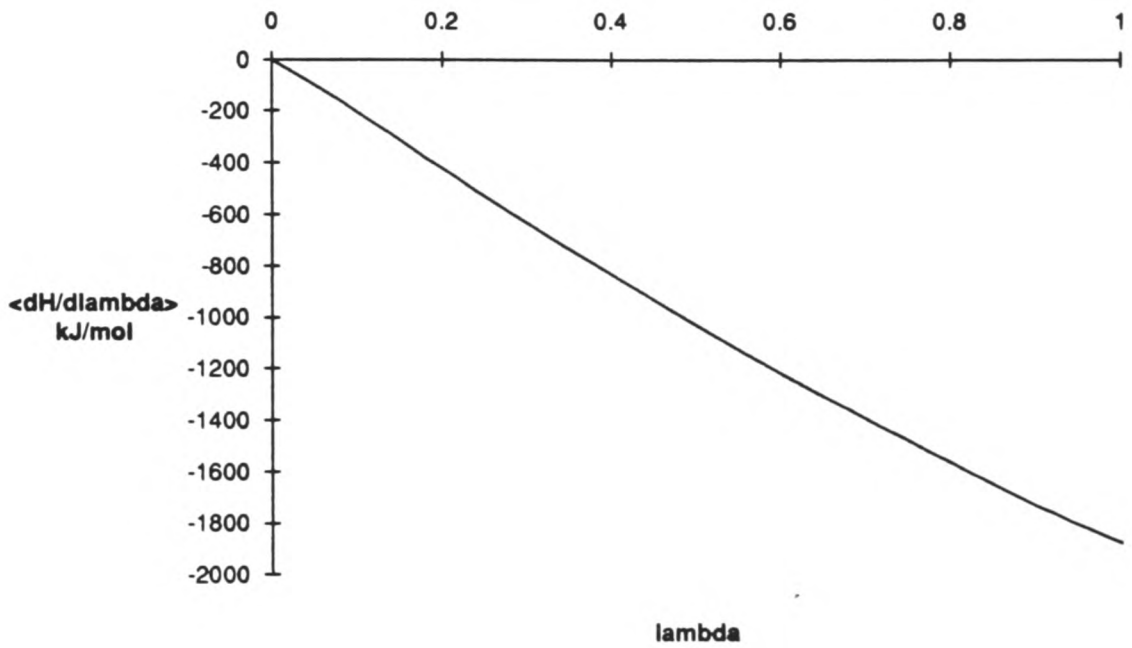
λ	$\langle \partial H/\partial \lambda \rangle_{\lambda.f}$ (kJ mol^{-1})	$\langle \partial H/\partial \lambda \rangle_{\lambda.b}$ (kJ mol^{-1})	average (kJ mol^{-1})
1.0	-1511.211		-1468.785 ± 42.426
0.9	-1619.572	-1541.519	-1580.549 ± 39.028
0.7	-1763.025	-1686.980	-1725.005 ± 38.024
0.6	-1864.420	-1789.790	-1827.107 ± 37.317
0.5	-1984.609	-1916.736	-1950.673 ± 33.936
0.4	-2034.114	-1968.472	-2001.295 ± 32.823
0.3	-2086.138	-2025.307	-2055.725 ± 31.422
0.1	-2172.856	-2150.831	-2161.844 ± 11.012
0.0		-2009.994	-2013.345 ± 3.351

From these figures it can be seen immediately that the different charges have caused a huge effect. The numbers are around 1000 kJ mol^{-1} larger than those in T6.2 and T6.3. As usual a curve was fitted to these points and then integrated. These are plotted in P6.7 and P6.8. The final value using these charges is $\Delta G(aq,inter) = -1885.321$ kJ mol^{-1}.

When this is compared to the value calculated for the hydrolysis of GTP, it can be immediately seen that this result takes the simulated value even further from the experimental. The difference produced by these charges, even though they do not seem to be that different from the ones used in the first simulations is astounding.



P6.7. A polynomial curve fitted through the points calculated at nine values of λ during the mutation of GTP to GDP using standard AMBER charges.



P6.8. The change in $\Delta\Delta G_{hyd}$ with changing λ during the mutation of GTP to GDP using standard AMBER charges.

VI.4. Calculating the Difference of Binding Free Energy of GDP and GTP to the Ha-*ras* Protein

Next the free energy difference between the nucleotides in the protein was to be calculated. A model for the protein with GTP in the binding site, based on the GPPNP complex crystal structure of Pai *et al.* had already been set up to calculate the charges. This has been described in chapter V.1.1.3. In theory all that had to be done was to equilibrate this model and then run a perturbation calculation using the protocol developed above.

The minimised model was heated to 298K and equilibrated for 20ps. The protocol developed in chapter III was used whereby the initial 1ps was with a 0.0005ps time step.

SHAKE was then turned on and the time step increased to 0.001ps for 3ps before the time step was increased to the maximum 0.002ps for the remainder of the run.

At this point the first problem was encountered. One modification had been made to the reaction field method. In the previous calculations in solution, the sphere containing the solute was based on its centre of mass. Outside this sphere it was approximated to be like bulk water. When the nucleotide is inside the protein, this obviously means that all of the protein has to be included with enough water inside the explicit solvent sphere so that this approximation holds.

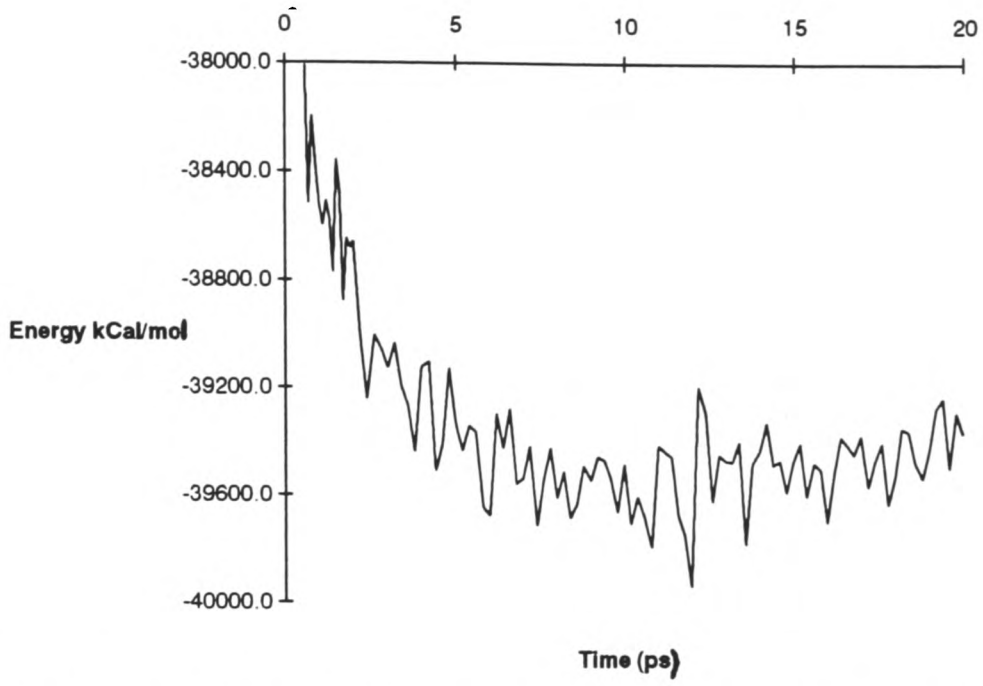
The problem is that the solute is to one side of the protein. If the sphere is based on its centre of mass, a very large sphere is needed to include all of the protein and more solvent would be required to prevent violation of the minimum image convention. The answer was to base the sphere at the centre of mass of the protein, with a radius big enough to include most of the truncated octahedron. This way the nucleotide would have 8Å of water in the direction straight out into the solvent as well as interacting with all the protein.

During the equilibration with these conditions, it was discovered that the nucleotide and magnesium left the binding site and migrated to the edge of the solvent box, lead by the phosphate chain. The reason for this is presumably an artefact of the image charge approximation. The magnitude of the image charge and its distance from the sphere containing the source charge is proportional to a/r where a is the radius of the sphere and r the magnitude of the distance of the charge from the centre of the sphere. Thus when the source charge is at the edge of the sphere, the image charge is both at a maximum and as near as possible to the solute.

As the reaction field was set up, the Coulombic forces due to the nucleotide interacting with its image field were presumably high enough to force the nucleotide out to the edge of the box. This could not happen before as the sphere followed the solute as it moved. The answer to this was to remove the interaction between the solute and the image field from the forces i.e. it was included in the energy calculation but not in the energy for purposes of calculating the gradient. This seemed to work. The fact that the nucleotide did not see as much solvent as it should for calculating its trajectory was outweighed by the need to keep it in the binding site.

The equilibration was repeated and during this time the potential energy dropped as shown in P6.9. During this equilibration, the energy was seen to drop nicely and even out. The perturbation was then run. A second problem was now encountered. With a terrible sense of déjà vu, the magnesium seemed determined to once again move around the binding site. After 3 calculations i.e. when λ was 0.7, this movement had resulted in the binding site being opened up and the magnesium moving very close to the β-phosphate. Again the nucleotide was expelled from the binding site and the γ-phosphate bonds broke.

The reason for this was that as the terminal phosphate shrank, there would appear to the magnesium to be a large movement of charge towards the middle phosphate group. It would then follow this moving charge, pulling the protein ligands with it and so disrupting the tertiary structure.



P6.9. The change in potential energy for the Ha-*ras*: GTP complex during 20ps of dynamics.

In order to overcome this problem it was decided to fix the magnesium in place. It was included in the interactions of everything around it but would not undergo dynamics itself. This should allow the system to move around the magnesium and so the trajectory should not be unduly affected.

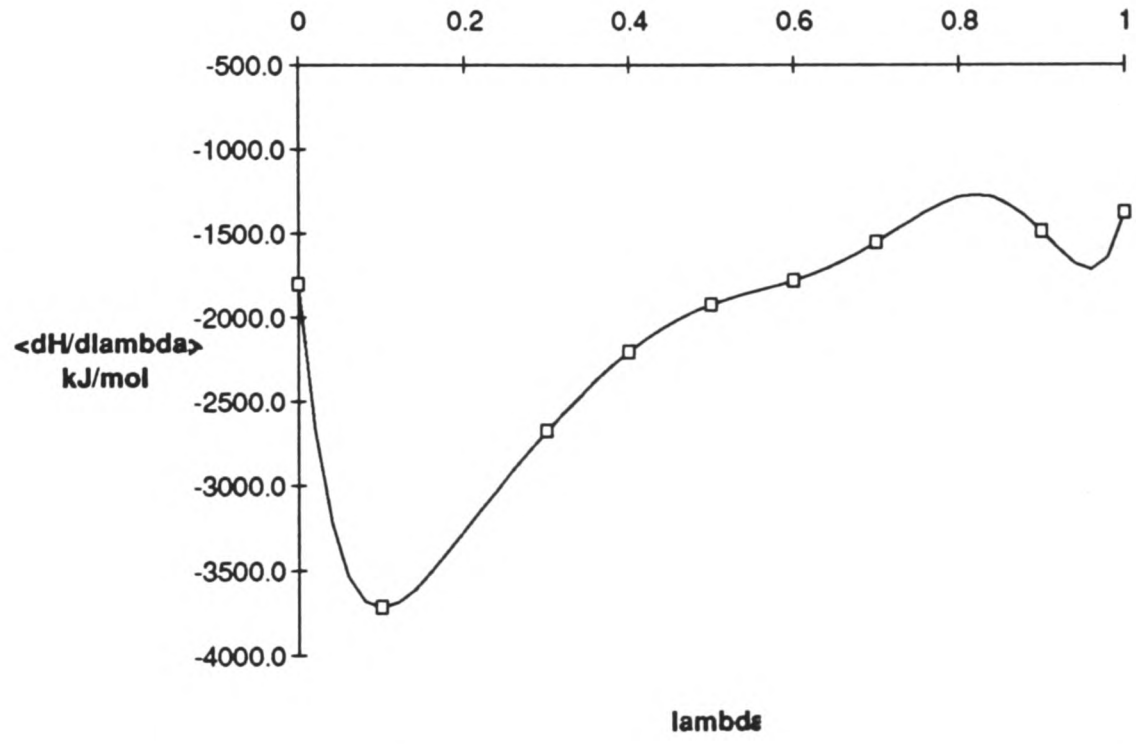
When this was done, all nine perturbation steps successfully finished. The results are shown in T6.7. It can be seen that now the errors have increased markedly compared to the solution simulations. The fitted polynomial is plotted in P6.10 and integrated in P6.11. The result from this integration is $\Delta G(prot) = -2210.179$ kJ mol^{-1}.

VI.4.1. Checking the Secondary Structure of the Ha-*ras* protein after a Mutation of the Bound Nucleotide.

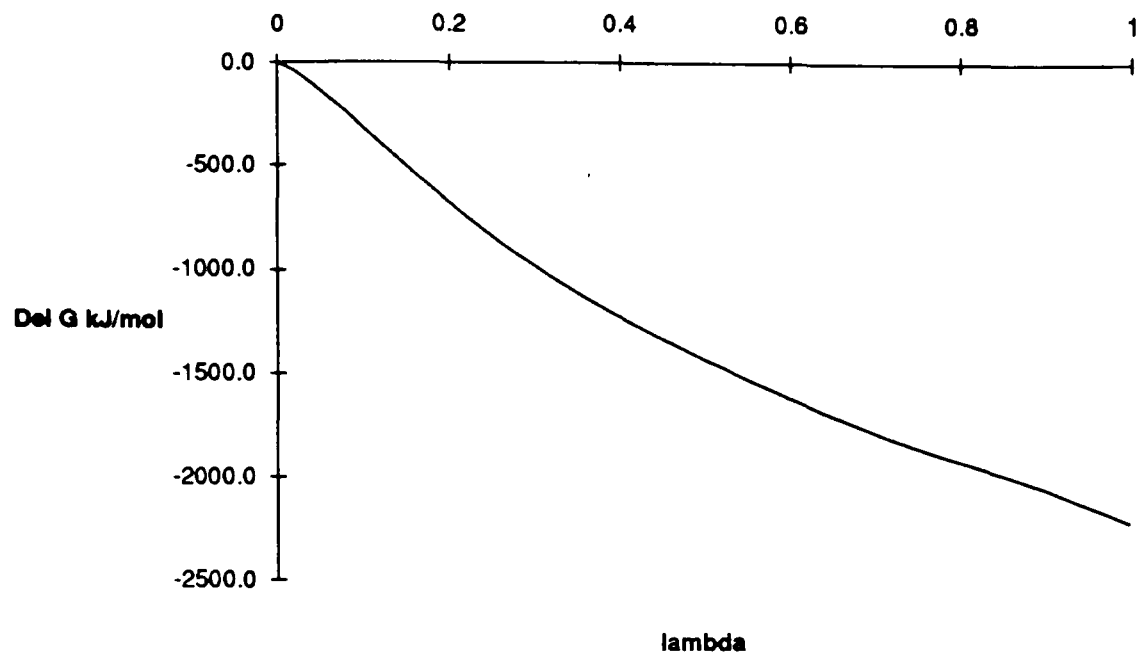
At the end of this run, it was checked to see if the correct transformation between the two protein states had occurred i.e. the change in secondary structure between residues 66 and 74 on moving from the GDP to GTP bound complex. To do this the secondary structure of the final protein conformation was calculated, using the Kabsch-Sander method in the same way as in chapter III. This is shown in T6.8.

T6.7. The accumulated perturbation gradients at various values of λ for the mutation from GTP at λ=1 to GDP at λ=0 in the Ha-*ras* binding site. Results are for the direction GDP ———> GTP.

λ	$\langle \partial H/\partial \lambda \rangle_{\lambda,f}$ (kJ mol^{-1})	$\langle \partial H/\partial \lambda \rangle_{\lambda,b}$ (kJ mol^{-1})	Average (kJ mol^{-1})
1.0	1423.744		1381.222 ± 42.522
0.9	1549.252	1439.217	1494.236 ± 55.020
0.7	1619.442	1497.223	1558.335 ± 60.693
0.6	1847.265	1723.130	1785.200 ± 62.070
0.5	1991.751	1868.633	1930.192 ± 61.559
0.4	2272.192	2151.066	2211.629 ± 60.563
0.3	2758.327	2610.724	2684.526 ± 73.802
0.1	3754.843	3683.271	3719.057 ± 35.786
0.0		1771.547	1801.630 ± 30.083



P6.10. A polynomial curve fitted through the points calculated at nine values of λ during the mutation of GTP to GDP in the Ha-*ras* binding site.



P6.11. The change in $\Delta\Delta G_{hyd}$ with changing λ during the mutation of GTP to GDP in the Ha-*ras* binding site.

```
   EE   B     SSS HHHHHHHHHT      S     TT   S EEEEETTEEEEE          TT HHHHHH
1         10         20         30         40         50         60         70

TTSS SSBEEE   TT THHHHH    HHHHHHHHTT SS    EEEEE  B     SS HHHHHHHHHHHT
         80         90        100        110        120        130        140

EEEEBTTTTBSSHHHHHHHHHHTTT
        150        160    166
```

T6.8. The secondary structure elements of Ha-*ras* after the nucleotide in the binding site has been mutated from GTP to GDP over 108ps. The notation is the Kabsch-Sander standard notation, given in T3.11 and the numbers under the line are the residue numbers.

Comparing this structure with those in chapter III, it is clear that the protein is still in the GTP form, having a helix-turn sequence from residues 66 to 74. Also despite being fixed, the magnesium has moved away from its binding site position. This means that the forces are high enough to make the entire protein move around it. The loop λ_2 along side the binding site had also moved a long way during the simulation; the middle of this loop was around 8Å from the crystal position.

This analysis means that something is wrong about the model or the method of calculation. The forces should not be so high and something must be missing. This apart, the lack of structural transformation over this simulation of 108ps means that a much longer run is needed for a reasonable result to be obtained.

VI.4.2. The Difference in Binding Free Energy.

Despite the errors accumulated in the runs so far, it would be interesting to calculate the final result to see just how inaccurate it is. The final calculation needed is

again the intra - molecular part of the potential. This involves calculating the *ab initio* energy of the nucleotides in the protein binding site. The bound nucleotide is fairly rigid and so there is little problem in deciding which conformers to choose for these calculations. Again these energies have been given in chapter V.

As noted above, it is impossible to directly subtract the *ab initio* energies of the GTP and GDP in the protein conformer to give the intra-molecular energy, due to the different number of atoms present. Instead it is possible to take the difference between the GTP in solution and protein and then subtract that from the difference between the GDP in solution and protein. This is in effect working around the cycle in figure F6.1 along the vertical sides and adding it to the difference in inter-molecular free energy along the horizontal sides. From E1.2,

$$\Delta\Delta G_{bind} = \Delta G_{prot,inter} - \Delta G_{aq,inter} + \Delta G_{prot,intra} - \Delta G_{aq,intra}$$

$$\Delta\Delta G_{bind} = \Delta G_{prot,inter} - \Delta G_{aq,inter} + \Delta G_{GTP,intra} - \Delta G_{GDP,intra}$$

This also has the advantage that the lack of correction to the *ab initio* internal energies is no longer a big problem as the difference is between two similar species which would have near identical partition functions. Hence from T5.12,

$$\Delta G_{GTP,intra} = 0.059499 \text{ a.u.} = 173.806 \text{ kJ mol}^{-1}$$

$$\Delta G_{GDP,intra} = 0.066018 \text{ a.u.} = 192.849 \text{ kJ mol}^{-1}$$

Notice again that the GDP is destabilised more than the GTP on entering the binding site. Combining all the results, we therefore have the answer

$$\Delta\Delta G_{bind} = -1506 \text{ kJ mol}^{-1}$$

Compared to the experimental result at the beginning of this chapter of around -4 kJ mol^{-1}, the kindest comment can be that at least it has the same sign.

VI.5. Conclusions

In this chapter the techniques developed during chapters IV and V were put to the test on the real problem of calculating the difference in binding free energy of GDP and GTP to the Ha-*ras* protein.

Initially a different method of calculating the difference in free energy was implemented in a novel manner using the MCTI methodology. On the histamine monocation problem this gave the same accuracy as windowing calculations taking three times as much computer time.

In contrast to the high accuracy obtained with the histamine calculations, the nucleotide perturbations when compared to experimental values were seen to overestimate the difference in free energy between the two molecules. The results were also seen to be very dependent on the charges used.

The reasons for this error were, despite the faster method, the computer time available was simply not enough to investigate the very large numbers of conformers

available. In the histamine case, good results were obtained when the intra- and inter-molecular Hamiltonians were evaluated over the same species, despite the fact they are not in the solution distribution. This is possibly only a reasonable assumption when the intra-molecular energy is the same for all the conformers. If the conformer energies are very different, as they were found to be for GDP, this could introduce some serious errors.

The problem with the protein results is for the same reason but with a different symptom. The problem here was that the simulation was not long enough for the protein to undergo the transition known to be present between the two endpoints of the perturbation. In both cases it is also likely that a poor electrostatic description had also resulted from the use of non-optimised geometries in the *ab initio* calculations.

There are also uncertainties about the protonation state of the nucleotide in the protein site. A recent simulation studying the dynamic motion of the Ha-*ras*: GTP complex (Foley *et al.* 1992) used a singly protonated GTP. They give no evidence as to why this was used, or where the proton is, but this is obviously one area that should be investigated. This simulation also found that the protein complex had in fact not equilibrated from the crystal structure until 160ps had passed.

Chapter VII

Discussion

Over the last few years the fortunes of FEP calculations have waxed and waned. When first introduced, only around 10 years ago, they seemed like the answer to many problems. The reason for this is that it is the only method of calculating a free energy difference between two similar systems including the solvent explicitly and taking into account entropy.

Early calculations on simple systems showed great promise, giving results with experimental accuracy in many cases. As time passed, people started to apply the method to all types of problems, sometimes inaccurately and often with disappointing results. At the present time FEP has lost favour with many people.

This thesis demonstrates both sides of this method. The calculations in chapter IV on histamine reproduce an experimental result to within the experimental error. In contrast the calculations on the nucleotide binding to the Ha-*ras* protein in chapter VI give results that err markedly from the expected. It must however be noted that this was also the case at the beginning of the histamine monocation calculations and it took several improvements in the method in which the calculations were made before the desired result was achieved.

This thesis has developed a method by which complex calculations should be performed. Two novel implementations of theory have been used, namely the reaction field method in chapter IV and the multi configurational thermodynamic integration in chapter VI. The first improves the accuracy of the calculation by including the long range forces in a simple manner, yet retains the boundary conditions necessary for a good simulation. The second allows a much shorter simulation to be run by taking separate simulations and interpolating between them.

The problems still remaining at the end of thesis are more due to restrictions of hardware than theory. The problems with the *ras* protein calculations are those of size. The nucleotides were too big to treat correctly by *ab initio* methods and this means that the conformational problem could not be sufficiently dealt with. The failings in the electrostatic description of these molecules, demonstrated by the poor free energy of hydrolysis calculated for GTP, also need large scale *ab initio* calculations to be thoroughly investigated. The final problem is just the size of the problem. For example recent dynamics simulation of the *ras*: GTP complex in solvent needed 300 hours of CRAY YMP time to produce reasonable results.

If the present increase in computer performance continues, it should not be long before we have the ability to re evaluate this project and, using the methods developed in this thesis accurate results for the binding energies should be obtainable.

References

Adari, H., Lowry, D.R., Willumsen, B.M., Der, C.J. and McCormick, F. (1988) *Science* **240**: 518.

Allinger, N.L. (1977) *J. Am. Chem. Soc.* **99**: 8127.

André, F., Demassier, V., Bloch, G. and Neumann, J.-M.(1990) *J. Am. Chem. Soc.* **112**: 6784

Ballester, R., Furth, M.E. and Rosen, O.M. (1987) *J. Biol. Chem.* **262**: 2688.

Barbacid, M. (1987) *Ann. Rev. Biochem.* **56**: 779.

Berendsen, H.J.C., Postma, J.P.M., Van Gunsteren, W.F., Di Nola, A. and Haak, J.R. (1984) *J. Chem. Phys.* **81**: 3684.

Berendsen, H.J.C., Postma, J.P.M. and Van Gunsteren, W.F. (1985) in "Molecular Dynamics and Protein Structure." Ed. J.Hermans, Polycrystal Book Service.

Binkley, J.S., Pople, J.A. and Hehre, W.J. (1980) *J. Am. Chem. Soc.* **102**: 939.

Binkley J.S. and Pople, J.A. (1977) *J. Chem. Phys.* **66**: 879.

BOSS3.1, Jorgensen, W.L., Yale University (1991).

Bourne, H,R., Wrischnik, L. and Kenyon, C. (1990) *Nature* **348**: 678.

Bourne, H., Sanders D.A. and McCormick, F. (1991) *Nature* **349**: 117.

Boys, S.F. and Bernardi, F. (1970) *Mol. Phys.* **19**: 553.

Brändén, C.-I. and Jones, T.I. (1990) *Nature* **343**: 687.

Broek, D., Samiy, N., Fasano, O., Fujiyama, A., Tamanoi, F., Northrup, J. and Wigler, M. (1985) *Cell* **41**: 763.

Calés, C., Hancock, J.F., Marshall, C.J. and Hall, A. (1988) *Nature* **332**: 7.

Castner, E.W., Fleming, G.R., Baghi, B. and Maroncelli, M. (1988) *J. Chem. Phys.* **89**: 3519.

Chirlian, L.E., Francl, M.M. (1987) *J. Comp. Chem.* **8**: 894.

Cieplak, Lybrand and Kollman, P.A. (1987) *J. Chem. Phys.* **86**: 6393.

Clark, T., Chandrasekhar, J., Spitznagel, G.W. and von R. Schleyer, P. (1983) *J. Comp. Chem.* **4**: 294.

DeVos, A.M., Tong, L., Milburn, M.V., Matias, P.M., Jancarik, J., Noguchi, S., Nishimura, S., Miura, K., Ohtsuka, E. and Kim, S.-H. (1988) *Science* **239**: 888.

Dewar, M.J.S. and Thiel, W.J. (1977) *J. Am. Chem. Soc.* **99**: 4899.

Dewar, M.J.S., Zoebisch, E.G., Healey, E.F. and Stewart, J.J.P. (1985) *J. Am. Chem. Soc.*. **107**: 3902.

Downward, J. (1990) *TIBS*, **15**: 469.

Downward, J., Riehl, R., Wu, L. and Weinberg, (1990) R.A. *Proc. Natl. Acad. Sci. USA.* **87**: 5998.

Emerson, J. and Sundaralingam, M. (1980) *Acta Cryst. Sect. B* **36**: 1510.

Engler, E.M., Andose, J.D. and von R. Schleyer, P. *J. Am. Chem. Soc.* (1973) **95**: 8005.

Essex. J.W., Reynolds, C.A. and Richards, W.G. (1992) *J. Am. Chem. Soc.* **114**: 3634.

Ferenczy, G.C., Reynolds, C.A. and Richards, W.G. (1989) *J. Comp Chem.* **11**: 159.

Feuerstein, J., Goody, R.S. and Wittinghofer, A. (1987) *J. Biol. Chem.* **262**: 8455.

Foley, C.K., Pedersen, L.G., Charifson, P.S. Darden, T.A., Wittinghofer, A., Pai, E.F. and Anderson, M.W. (1992) *Biochemistry* **31**: 4951.

Friedman, H.L. (1975) *Mol. Phys.* **29**: 1533.

Ganellin, C.R. (1973) *J. Pharm. Pharmocol.* **25**: 787.

Ganellin, C.R., Pepper, E.S., Port G.N.J. and Richards, W.G. (1973) *J. Med. Chem.* **16**: 610.

GAUSSIAN88, Frisch, M.J., Head-Gorden, M., Schlegel, H.B., Raghavachari, K., Binkley, J.S., Gonzalez, C., Defrees, D.F., Fox, D.J., Whiteside, R.A., Seeger, R., Melius, C.F., Baker, J., Martin, R., Kahn, L.R., Stewart, J.P.P., Fluder, E.M., Topiol, S. and Pople, J.A. Gaussian Inc., Pittsburgh, P.A. 1988.

GAUSSIAN90, Frisch, M.J., Head-Gorden, M., Trucks, G.W., Foresman, J.B., Schlegel, H.B., Raghavachari, K., Robb, M.A., Binkley, J.S., Gonzalez, C., Defrees, D.F., Fox, D.J., Whiteside, R.A., Seeger, R., Melius, C.F., Baker, J., Martin, R., Kahn, L.R., Stewart, J.P.P., Fluder, E.M., Topiol, S. and Pople, J.A. Gaussian Inc., Pittsburgh, P.A. 1990.

Hattori, S., Ulsh, L.S., Haliday, K. and Shih, T.Y. (1985) *Mol. Cell Biol.* **5**: 1449.

Hausheer, F.H., Singh, U.C., Palmer, T.C. and Saxe, J.D. (1990) *J. Am. Chem. Soc.* **112**: 9468.

Hehre, W.J., Stewart, R.F. and Pople, J.A. (1969) *J. Chem. Phys.* **51**: 2657.

Hehre, W.J., Ditchfield, R. and Pople, J.A. (1972) *J. Chem. Phys.* **56**: 2257.

Hehre, W.J., Radom, L., von R. Schleyer, P. and Pople, J.A. (1986) "*Ab Initio* Molecular Orbital Theory" John Wiley and sons Ltd.

Hockney, R.W. (1970) *Methods Comput. Phys.* **9**: 136.

Jancarik, J., deVos, A., Kim, S.-H., Miura, K., Ohtsuka, E., Noguchi, S. and Nishimura, S. (1988) *J. Mol. Biol.* **200**: 205.

Jeng, A.Y., Srivastava, S.K., Lacal, J.C. and Blumberg, P.M.(1987) *Biochem. and Biophys. Res. Comm.* **145**, 782.

Jorgensen, W., Chandrasekhar, J., Madura, J. Impey, R. and Klein, M. (1983) *J. Chem. Phys.* **79**: 926.

Jorgensen, W.L. and Tirado-Rives, J. (1988) *J. Am. Chem. Soc.* **110**: 3469.

Jurnak, F., Heffron, S. and Bergmann, E. (1990) *Cell* **60**, 525.

Kabsch, W. and Sander, C. (1983) *Biopolymers* **22**: 2577.

Kamata, T., Kathuria, and S.Fujita-Yamaguchi, Y. (1987) *Biochem. and Biophys. Res. Comm.* **144**: 19 - 25.

Karplus, M. and McCammon, J.A. (1983) *Ann. Rev. Biochem.* **52**: 263.

Kestner, N.R. (1968) *J. Chem. Phys.* **48**: 252.

Korn, J.K., Siebel, C.W, McCormick, F. and Roth, R.A. (1987) *Science* **236**: 840.

Lacal, J.C., Fleming, T.P., Warren, B.S., Blumberg, P.M. and Aaronson, S.A. (1987) *Mol. Cell Biol.* (1987) **7**: 4146.

Levitsky, A., Rudick, J., Pastan, I., Vass, W.C. and Lowry, D.R. (1986) *FEBS Letts* **197**: 134.

Levitt, M. and Warshel, A. (1978) *J. Am. Chem. Soc.* **100**: 2607.

Lii, J., Gallion, S., Bender, C., Wilkstrom, H., Allinger, N.L., Flurchick, K. and Teeter, M. (1989) *J. Comp. Chem.* 10 : 503

Lipman, D.J. and Pearson, W.R. (1985) *Science* **227**: 1435.

Lister, S.G., Reynolds, C.A. and Richards, W.G. (1992) *Int. J. Quant. Chem.* **41**: 293.

McCormick, F. (1989) *Cell* **56**: 5.

Macara, I.G. (1989) *Mol. Cell Biol.* **9**: 325.

Maroncelli, M. and Fleming, G.R. (1988) *J. Chem. Phys.* **89**: 5044.

Milburn, M.V., Tong, L., DeVos, A.M., Brünger, A., Yamaizumi, Z., Nishimura, S. and Kim, S.-H. (1990) *Science* **247**: 939.

Møller, C. and Plesset, M.S. (1934) *Phys. Rev.* **46**: 618.

Mulcahy, L.S., Smith, M.R. and Stacey, D.W. (1985) *Nature* **313**: 241.

Noda, M., Ko, M., Ogura, A., Liu, D-G., Amano, T., Takano, T. and Ikaura, Y. (1985) *Nature* **318**: 73.

O'Brien, R.M., Hall, A., Housley, M.D. and Siddle, K. (1987) *FEBS Letts.* **217**: 253.

Pai, E.F., Kabsch, W. ,Krengel, U., Holmes, K.C., Jacob, J. and Wittinghofer, A. (1989) *Nature* **341**: 209.

Pai, E.F., Krengel, U., Petsko, G.A., Goody, R.S., Kabsch, W. and Wittinghofer, A. (1990) *EMBO. J.* **9**: 2351.

Pearlman, D.A. and Kollmann, P.A. (1989) in *"Computer Simulation of Biomolecular Systems"*. Ed. van Gunsteren, W.F. and Weiner, P.K. Escom Publishers.

Pearlman, D.A. and Kollmann, P.A. (1991) *J. Chem. Phys.* **94**: 4532.

QUANTA / CHARMm version 3.2. Polygen corporation.

Randerath, Yu and Randerath (1972) *Anal. Biochem.* **48**: 172.

Rey, I., Soubigou, P., Debusche, L., David, C., Morgat, A., Bost, B.E., Mayaux, J.F. and Tocque, B. (1989) *Mol. Cell Biol.* **9**: 3904.

Reynolds, C.A., Essex, J.W. and Richards, W.G. (1992) *J. Am. Chem. Soc.* (in press).

Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C. (1977) *J. Comp. Phys.* **23**: 327.

Schlichtling, I., Almo, S.C., Rapp, G., Wilson, K., Petratos, K., Lentfer, A., Wittinghofer, A., Kabsch, E., Pai, E., Petsko, G.A. and Goody, R.S. (1990) *Nature* **345**: 309.

Sigal, I.S., Gibbs, J.B., D'Alonzo, J.S. and Scolnick, (1986) E.M. *Proc. Natl. Acad. Sci. USA* **83**: 4726.

Singh, U.C., Brown, F.K., Bash, P.A. and Kollmann, P.A. (1987) *J. Am. Chem. Soc.* **6**: 1607.

Singh, U.C., Weiner, P.K., Caldwell, J.W. and Kollmann, P.A. (1988) AMBER (UCSF), version 3.1, Department of Pharmaceutical Chemistry, University of San Francisco.

Stacey, D.W., Watson, T., Kung, H-F. and Curran, T. (1987) *Mol. Cell Biol.* **7**: 523.

Stewart. J.P.P. MOPAC 5.0, Frank J. Seiler Research Lab., U.S. Air Force Academy, Colorado 80840.

Stewart, J.P.P. (1989) *J. Comp. Chem.* **10**: 209.

Straatsma, T.P., Berendsen, H.J.C. and Stam, A.J. (1986) *J. Mol. Phys.* **57**: 89.

Straatsma, T.P. and McCammon, J.A. (1991) *J. Chem. Phys.* **95**: 1175.

Tong, L., Milburn, M.V., de Vos, A.M. and Kim, S.- H. (1989) *Science* **245**: 244 .

Tong, L., de Vos, A.M. Milburn, M.V., Jancarik, J., Noguchi, S., Nishimura, S., Miura, K., Ohtsuka, E., and Kim, S.- H. (1989a) *Nature* **337**: 90.

Trahey, M. and McCormick, F. (1987) *Science* **238**: 542.

Trahey, M, Milley, R.J., Cole, G.E., Innis,M., Paterson, H., Marshall, C.J., Hall, A. and McCormick, F. (1987a) *Mol. Cell. Biol.* **7**: 541.

Trahey, M., Wong, G., Halenback, R., Rubenfield, B., Martin, G.A., Ladner, M., Long, C.M., Crosier, W.J., Watt, K., Koths, K. and McCormick, F. (1988) *Science* **242**: 1697.

van Gunsteren, W.F. (1980) *Mol. Phys.* **40**: 1015.

van Gunsteren, W.F. and Berendsen, H.J.C. (1990) *Agnew. Chem. Int. Ed. Engl.* **992**: 1023.

Weiner, S.J., Kollmann, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S. amd Weiner, P. (1984) *J. Am. Chem. Soc.* **106**: 765.

Weiner, S.J., Kollmann, P.A., Nguyen, D.T. and Case, D.A. (1986) *J. Comp. Chem.* **7**: 230.

Weinstein, H., Chou, D., Johnson, C.L., Kang, S. and Green, J.P. (1976) *Mol. Pharm.* **12**: 738.

Weinstein, H., Muzarek, A.P., Osman, R. and Topiol, S. (1986) *Mol. Pharm.* **29**: 28.

West, M., Kung, H.F. and Kamata, T. (1990) *FEBS Letts.* **259**: 245.

Wolfman, A. and Macara, I. (1990) *Scienc*, **248**: 67.

Woodgett, J.R. (1992) *J. Current Biol.* **2**: 357.

Worth, G.A., King, P.M. and Richards, W.G. (1989) *BBA* **993**: 134.

Worth, G.A., King, P.M. and Richards, W.G. (1990) *BBA* **1036**: 158.

Yang, J.-A. and Kestner, N.R. (1991) *J. Chem. Phys.* **95**: 9214.

Yathindra, N. and Sundaralingam, M. (1973) *Biopolymers* **12**: 297.

Yathindra, N. and Sundaralingam, M. (1973a) *Biopolymers* **12**: 2261.

Zwanzig, R.W. (1954) *J. Chem. Phys.* 22: 1420.

Buda, F., Chiarotti, G.L., Car, R. and Parinello, M. (1989) Phys. Rev. Lett. 63: 294

Car, R. and Parriello, M. (1985) Phys. Rev. Lett. 55: 2471

Finocchi, F., Galli, G., Parinello, M. and Bertoni, C.M. (1992) Phys. Rev. Lett. 68: 3044

Jorgensen, W.L. and Maclwa, J.D. (1985) Mol. Phys. 56: 1381

Lundqiust, S. and March, N.M. (1983) "Theory of the Inhomogeneous Electron Gas" Plenum N.Y.

Postma, J.P.M., Berendsen, H.J.C. and Haak J.R. (1982) Faraday Symp. Chem. Soc. 17: 55

Pront, K., Critchley, S. and Carellin, C.R. (1974) Acta Cryst. Sect. B 30: 2884

Rennler, D.K. and Madden, P.A. (1990) Mol. Phys. 70: 921

Tembe, B.L. and McCammon, D.A. (1984) Comput. Chem. 8: 281

APPENDIX

Previously Published Results for the Tautomer Ratio of Histamine and 4-(5-)methylimidazole in Solution

The following results have previously been released in Worth *et al.* (1989, 1990) and Worth, G.A. (1989) Part II Thesis, University of Oxford. They are included here as they are a useful comparison to the results calculated in chapter IV on the histamine monocation system. One result that can be seen from these tables is that the calculated value for the tautomeric equilibrium constant converges in each case as the basis set used for the gas phase calculations increases in size, up to 6-31G.

The major factor to be noticed for the purposes of this thesis is that the results for the histamine base and for 4-(5-)methylimidazole are close to the experimentally determined equilibrium constant, whereas that for histamine monocation is very different. The difference between the four FEP simulations made on each system are also interesting. The 4-(5-)methylimidazole results are all close to one another, the histamine base FEP values are slightly further apart, but the histamine monocation simulations have noticeable differences - especially viewing run1 against the other three.

Table TA.1 Results for the Calculation of K_T for the Tautomeric Shift N(π)H ——> N(τ)H of 4-(5-)Methylimidazole

TA1.1. The difference in free energy, ΔG_{gas},in kJ mol^{-1}, and the equilibrium constant, K_T, in the gas phase[a].

Method/Basis Set[b]	ΔG(g,298)	K_T(g,298)
AM1	-2.73	3.21
RHF/STO-3G//RHF/STO-3G	4.22	0.18
RHF/3-21G//RHF/3-21G	1.11	0.64
RHF/6-31G//RHF/6-31G	-1.56	1.88
RHF/6-31G*//RHF/6-31G	-0.76	1.36
RHF/6-31G**//RHF/6-31g	-0.92	1.45
MP2/6-31G*//RHF/6-31G	-1.33	1.71
MP2/6-31G**//RHF/6-31G	-1.37	1.74

TA1.2. FEP results i.e. $\Delta\Delta G_{hyd}$, in kJ mol^{-1} [c].

	Run 1	Run 2
Forward	1.86 ± 1.61	0.30 ± 2.05
Reverse	-1.09 ± 0.60	1.90 ± 2.49
Average	0.39 ± 1.72	1.10 ± 3.22

$\Delta\Delta G_{hyd} = 0.75 \pm 3.65$ kJ mol^{-1}

TA1.3. The difference in free energy, ΔG_{aq},in kJ mol^{-1}, and the equilibrium constant, K_T(aq), in aqueous solution[d].

Method/Basis Set[b]	ΔG(aq,298)	K_T(aq,298)
AM1	-2.15	2.38
RHF/STO-3G//RHF/STO-3G	4.97	0.14
RHF/3-21G//RHF/3-21G	1.85	0.47
RHF/6-31G//RHF/6-31G	-0.81	1.39
RHF/6-31G*//RHF/6-31G	-0.01	1.10
RHF/6-31G**//RHF/6-31g	-0.18	1.07
MP2/6-31G*//RHF/6-31G	-0.59	1.27
MP2/6-31G**//RHF/6-31G	-0.63	1.29
Experimental		0.5 - 1.5

Table TA.2 Results for the Calculation of K_T for the Tautomeric Shift N(π)H ——> N(τ)H of Histamine Base

TA.2.1. The difference in free energy, ΔG_{gas},in kJ mol^{-1}, and the equilibrium constant, K_T, in the gas phase[a].

Method/Basis Set[b]	ΔG(g,298)	K_T(g,298)
AM1	-2.17	2.40
RHF/STO-3G//RHF/STO-3G	3.78	0.22
RHF/3-21G//RHF/3-21G	-0.83	1.40
RHF/6-31G//RHF/6-31G	-3.18	3.61
RHF/6-31G*//RHF/6-31G	-1.57	1.88
RHF/6-31G**//RHF/6-31g	-1.72	2.00
MP2/6-31G*//RHF/6-31G	-2.25	2.46
MP2/6-31G**//RHF/6-31G	-2.17	2.48

TA.2.2. FEP results i.e. $\Delta\Delta G_{hyd}$, in kJ mol^{-1} [c].

	Run 1	Run 2
Forward	0.54 ± 2.65	4.81 ± 2.80
Reverse	-3.84 ± 3.05	-9.79 ± 2.22
Average	-1.65 ± 4.04	-2.49 ± 3.57

$\Delta\Delta G_{hyd}$ = -2.09 ± 5.39 $kJmol^{-1}$

TA.2.3. The difference in free energy, ΔG_{aq},in kJ mol^{-1}, and the equilibrium constant, K_T(aq), in aqueous solution[d].

Method/Basis Set[b]	ΔG(aq,298)	K_T(aq,298)
AM1	-4.26	5.58
RHF/STO-3G//RHF/STO-3G	1.69	0.51
RHF/3-21G//RHF/3-21G	-2.92	3.24
RHF/6-31G//RHF/6-31G	-5.27	8.37
RHF/6-31G*//RHF/6-31G	-3.66	4.38
RHF/6-31G**//RHF/6-31g	-3.81	4.64
MP2/6-31G*//RHF/6-31G	-4.31	5.70
MP2/6-31G**//RHF/6-31G	-4.34	5.75
Experimental		4

Table TA.3 Results for the Calculation of K_T for the Tautomeric Shift N(π)H ——> N(τ)H of Histamine Monocation

TA.3.1. The difference in free energy, ΔG_{gas},in kJ mol^{-1}, and the equilibrium constant, K_T, in the gas phase[a].

Method/Basis Set[b]	ΔG(g,298)	K_T(g,298)
AM1	34.56	8.81×10^{-6}
AM1 correct	-34.56	
RHF/STO-3G//RHF/STO-3G	-42.15	2.42×10^{7}
RHF/3-21G//RHF/3-21G	-52.35	1.48×10^{9}
RHF/6-31G//RHF/6-31G	-53.89	2.76×10^{9}
RHF/6-31G*//RHF/6-31G	-46.53	1.42×10^{8}
RHF/6-31G**//RHF/6-31g	-46.97	1.69×10^{8}
MP2/6-31G*//RHF/6-31G	-49.15	4.08×10^{8}
MP2/6-31G**//RHF/6-31G	-49.02	3.87×10^{8}

TA3.2. FEP results i.e. $\Delta\Delta G_{hyd}$, in kJ mol^{-1} [c].

	Run 1	Run 2
Forward	52.17 ± 1.67	24.98 ± 2.93
Reverse	30.54 ± 3.31	26.57 ± 2.34
Average	41.36 ± 3.71	25.78 ± 3.75

$\Delta\Delta G_{hyd} = 33.57 \pm 5.28$ kJmol^{-1}

TA3.3. The difference in free energy, ΔG_{aq},in kJ mol^{-1}, and the equilibrium constant, K_T(aq), in aqueous solution[d].

Method/Basis Set[b]	ΔG(aq,298)	K_T(aq,298)
AM1	73.95	1.11×10^{-11}
RHF/STO-3G//RHF/STO-3G	-2.76	3.04
RHF/3-21G//RHF/3-21G	-12.96	1.87×10^{2}
RHF/6-31G//RHF/6-31G	-14.51	3.84×10^{2}
RHF/6-31G*//RHF/6-31G	-7.14	17.87
RHF/6-31G**//RHF/6-31g	-7.55	21.07
MP2/6-31G*//RHF/6-31G	-9.76	51.37
MP2/6-31G**//RHF/6-31G	-9.62	48.64
Experimental		2.3 - 9