UNIVERSITY OF SOUTHAMPTON

Methods for Investigating Conformational Change in Biomolecular Simulations

Robert John Gledhill

A dissertation submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy at the University of Southampton

Supervisor: Dr J. W. Essex

Department of Chemistry

September 2003

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SCIENCE

CHEMISTRY

Doctor of Philosophy

Methods for Investigating Conformational Change in Biomolecular Simulations

by Robert John Gledhill

A series of simulations of the Syrian hamster prion protein and three point mutated variants thereof are presented. These simulations are analysed using root mean square deviations, dynamical cross correlation maps, secondary structure versus time plots and hydrogen bonding profiles. The results of these analyses are discussed. A problem identified during the interpretation of the prion protein trajectories was the lack of sufficiently subtle analysis tools. To remedy this, the recently proposed Hilbert Huang Transform (HHT) method was considered as a basis for the development of new techniques. The HHT is examined in depth, some improvements to it are proposed, and a number of issues with its use and interpretation are identified. A set of new analysis tools based on the HHT and wavelet methods are then proposed, and these are tested on a set of dihydrofolate reductase simulations performed by other workers. The results are encouraging and a series of proposals for future research work based on the findings are made.

Acknowledgements

I would like to offer my warm thanks to Jonathan Essex for his guidance, enthusiasm and good humour throughout this project. Also to *everyone* I have known in his group, past and present. Particular gratitude goes to:

Stephen Phillips, for many pleasant and helpful conversations, along with much else.

Oz Parchment, for the prion simulation system on which part of this work is based.

Martin Swain, for the DHFR simulations used in Chapter 6, and also for some constructive criticism.

Adrian Wiley, for helpful discussions. Also for sending my programs through purgatory.

Donna Goreham, for all sorts of things.

Chris Woods, for disaster-averting logistical support on at least two occasions.

My family, for their love, support and confidence through the years.

I would also like to acknowledge the BBSRC for funding this work for three years

Contents

List of Figures

List of Tables

Chapter 1

Introduction

All living things depend on the correct functioning of many different types of protein molecules. These proteins may be structural, such as keratin, or catalytic, such as protease enzymes. Biochemists have discovered much about the properties and actions of these molecules, but one of the principal unsolved problems in the field is the relationship between their functions and their dynamical properties.[1] In the few cases where detailed information is available, the link between the two is clear though. For example, the active sites of many enzymes are obscured from the solvent by the fold of the protein for much of the time, and substrates can gain access only after a large scale conformational rearrangement has taken place.[2] Also, many proteins form *amyloid* filaments under certain conditions, and this disease-associated process is known to be accompanied by a significant change in secondary structural content.[3]

The principal reason that the relationship between function and dynamics is very difficult to resolve for specific cases, and more general conclusions have proven difficult to draw, is that the experimental tools for gathering dynamical data are currently in a primitive state of development. An alternative means of gathering data on these phenomena is available though: computer simulation. Wisely conceived computer simulations of the dynamics of protein molecules, carefully validated with experimental data, can give useful results that are currently difficult or impossible to obtain by other means. An example of where this has proven to be the case is the enzyme *acetylcholinesterase* (AChE), entry and exit to and from the active site of which is controlled by a pair of protein 'gates'.[4] The catalytic process could only be fully appreciated once the molecular dynamics were understood, and the gating mechanism was finally resolved using computer simulation studies.[5] A second example is the cytochrome P450 family of enzymes, whose active sites appear to be occluded in experimentally determined structures. The dynamical process by which substrates can gain entry was again resolved by carefully validated molecular dynamics computer simulations.[6]

Computer simulations have then, on many occasions, proved invaluable to an understanding of protein mechanism. However, there is still much work to be done in this area to improve the methods used. One area of particular concern is the analysis and interpretation of the raw data obtained from simulations.

In this thesis simulations of the *prion* protein molecule have been run in an attempt to gain understanding of the dynamics of the molecule in solution and the effects on these dynamics of a pair of point mutations associated with specific disease states. While the simulated molecules clearly exhibit significant differences in their dynamics, the exact nature of these changes was found to be very difficult to analyse constructively with the tools currently available. Although many analysis techniques were tried, forming a clear picture of the differences in behaviour proved highly problematic. A key problem identified was that the techniques were all dependent on simultaneous visualisation of the protein trajectory for correct interpretation. This proved particularly true for *dynamical cross correlation maps* (DCCMs), a further problem with which was their sensitive dependence on the fitting strategy used. In interpreting these data it proved difficult to exclude an element of subjectivity.

A very new analysis method was therefore identified, the *Hilbert Huang Transform* (HHT), which is related to the Fourier and wavelet methods, as an avenue of research. The properties of this method were investigated, along with whether it might be improved, and how it might be employed in the analysis of protein molecular dynamics simulation data. A number of new techniques using the HHT were developed, and tested on a set of simulation data obtained by other workers on the enzyme *dihydrofolate reductase* (DHFR).

In Chapter 2 the molecular dynamics simulation method is discussed, along with the techniques currently used to analyse the simulation data obtained. In Chapter 3 the setting up, execution and analysis of a series of simulations of the Syrian hamster prion protein are described. In Chapter 4 the mathematical background to the HHT method, the claims made about it, and some of the work that has been done using it to date are outlined. In Chapter 5 a number of tests of the method and comparisons of it to the related Fourier and wavelet transforms are described. Several means by which the method might be improved are also investigated. In Chapter 6 a number of new methods that make use of the HHT and wavelet transforms are presented, and the results of applying these to DHFR are reported. Finally, in Chapter 7 the results of the earlier chapters are drawn together, the advantages and disadvantages of the various methods investigated are discussed, and suggestions for future work are made.

Chapter 2

Chemistry and Computers

Chemists have been simulating the behaviour of molecular systems on electronic computers since the early 1950s.[7] The methods of using automatic calculating machinery to integrate the equations of motion for a large number of particles and study the properties of the resultant ensemble of system states have been in development since the late 1950s. As computer speed and memory capacities have increased and their prices fallen, the size of the systems that chemists have attempted to tackle, and the complexity of the theoretical models used, have risen. In the late 1970s it became practical to simulate protein systems,[8] and among the insights obtained from doing so was the *critical* importance of a protein's flexibility and dynamics to its biological function. Detailed information on this is difficult to obtain directly from experiments, and this reinforces the practical significance of computer simulation as a scientific tool in this area.

In addition to molecular dynamics simulations, computers can be used as a tool to process and graphically visualise the three dimensional structures of biomolecules. Protein 'secondary structures', the patterns of electric charge, hydrophilicity and hydrophobicity on their surfaces and in their active sites can all be examined easily on a computer display unit. The plastic and wire models of the late Max Perutz and other pioneers of crystallography have long since given way to databases containing vast numbers of structures[9] and a plethora of computer programs[10, 11] that enable the researcher to create and manipulate representations of these structures in a fashion limited by imagination and video display technology (though both these *are* serious limitations).

In the last thirty years, and particularly in the last ten, a third series of computer applications have been developed based on the global communication network known as the 'internet'. More or less every published chemical structure obtained by X-ray crystallography or nuclear magnetic resonance (NMR) is now routinely deposited in one or more online databases.[9] These databases can be rapidly and automatically searched in a very sophisticated fashion by researchers around the world, and this direct access to a vast body of experimental results has itself proven valuable as a new way to conduct research. A further use of the internet that arose within the first few years of the twenty first century is the large scale publication of scientific papers in electronic formats, and the indices, databases and search tools that can rapidly and trivially harvest them.[12]

The current work presented here is concerned with the analysis of molecular dynamics (MD) simulations using electronic computers. This chapter gives an overview of the simulation method itself, and also the methods used to extract information from the trajectories generated.

2.1 Molecular Dynamics Simulations

The computer simulation of molecular dynamics is a large and growing field and a comprehensive review of it is beyond the scope of the current work; more detailed information can be found elsewhere. Only the basics will be covered here.

In a molecular dynamics simulation the physics of a chemical system is approximated by a Newtonian model.[7] The motions of the model are integrated in a stepwise manner, creating a series of frames whose statistics correspond to one or another thermodynamic ensemble. At each step of the simulation accelerations are calculated from a force equation whose terms are parameterised to give good agreement with particular experimental quantities.[7] To speed the calculation of nonbonded interactions, various techniques are often used, such as distance based cutoffs or an inverse space (Ewald) summation;[13] to maximise the timestep that may be employed, the lengths of bonds to light (hydrogen) atoms are usually constrained, often using the SHAKE[14] algorithm or similar. The ensemble from which the frames are sampled depends is determined by whether a thermostat and barostat are applied.

A common method with which the equations of motion are integrated is the *leapfrog Verlet* algorithm;[7] this is the method in the *sander* program[15] that was used for the calculation of the prion simulation. In this approach, accelerations and positions are calculated for each timestep t, but velocity values are calculated for points halfway between timesteps. The equations solved are

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \Delta t\,\mathbf{v}(t+\frac{1}{2}\Delta t) \tag{2.1}$$

$$\mathbf{v}(t+\frac{1}{2}\Delta t) = \mathbf{v}(t-\frac{1}{2}\Delta t) + \Delta t\,\mathbf{a}(t) \tag{2.2}$$

where $\mathbf{x}$ is the vector of atomic cartesian coordinates, $\mathbf{v}$ the vector of velocities and $\mathbf{a}$ the vector of accelerations.

The timestep used in a simulation is limited by the requirement for energy to be conserved and accurate dynamics recovered. An overlong timestep results in the approximation that force is constant over the timestep to fail, causing numerical instability. Numerous methods have been devised to solve this problem by constraining the lengths of certain bonds (usually the bonds between hydrogen and heavy atoms). The method used in both the prion and DHFR simulations is called SHAKE.[14] In it, extra *constraint* forces directed along the bonds of interest are calculated such as to ensure the bond length remains constant at each integration step. To ensure all constraints are simultaneously satisfied the procedure is iterated until the errors are within a specified tolerance.

The forces at each step of an MD simulation are calculated from a potential energy function specific to the type of force field in use. Most force fields have a fairly similar functional form resembling the following

$$
\begin{aligned}
V(\mathbf{x}) = & \sum_{bonds} \frac{k_i}{2}(l_i - l_{i(0)})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i(0)})^2 \\
& + \sum_{torsions} \sum_{m=1}^{M_i} \frac{k_i}{2}(1 + \cos(m\,\omega_i - \phi)) \\
& + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right)
\end{aligned}
\tag{2.3}
$$

where $V(\mathbf{x})$ denotes the potential at coordinates $\mathbf{x}$; the k_i are force constants of the force field; l_i, θ_i and ω_i are bond lengths, angles and torsions respectively; $l_{i(0)}$ and $\theta_{i(0)}$ are equilibrium points from which the bond lengths and angles are measured; M_i is the multiplicity of the torsional term and ϕ a phase offset; ε_{ij} and σ_{ij} are parameters of the Lennard-Jones potential; r_{ij} denotes the distance between the pair of atoms i and j, and q_i and q_j the charges on those atoms; N is the number of atoms in the system and ε_0 is the permittivity of free space.

The force field parameters used in the above equation are developed to obtain empirically correct results for some particular class of systems. There is no universally applicable 'correct' force field, only force fields parameterised correctly for one type of simulation or another. For example, the AMBER[16] force field used in the prion simulations was developed specifically for modelling proteins. Other force fields have been developed for modelling hydrocarbons,[17] for example. Owing to the sparsity of experimental data on many of the parameters required for a force field, many are parameterised using *ab initio* quantum mechanical simulations.

2.2 Trajectory Analysis Methods

Once a simulation has been performed, the data must be analysed. This may serve a number of purposes. First, the simulation must be checked for stability and correctness and second, thermodynamic and structural properties may be required from the results. If the simulation is of a protein system, a variety of techniques may be brought into play. Root mean square deviations (RMSDs) yield information on the flexibility of regions of the protein, and the rate of divergence of the structure from its initial conformation (usually obtained from experimental results). The trajectory may be directly visualised as an animation on a computer screen in a number of ways[10, 11] and secondary structure determining algorithms[18] can be used on a frame-by-frame basis to give an indication of structural rearrangements.[19] More detailed examination of the dynamics can be obtained by covariance analysis techniques that indicate which parts of the protein have correlated motions.[20] For some systems it also makes sense to use one or other of a pair of mathematically related techniques that seek to obtain geometric descriptions of the key motions present in a simulation[21] (or even from a static minimised structure, given a force field).

This section will describe some of the methods outlined above.

2.2.1 Simulation Stability

Simulation Energies and Temperatures

The first and most obvious method for determining whether a simulation in the NVE ensemble is showing correct dynamics, and verifying the implementation and setup of the program and force field, is to verify that the energy is conserved as it should be. A simple plot of total energy against time should show little, if any, increase in energy over the course of the simulation. If there *is* an increase in energy, this is an indicator that something is going wrong (commonly the stepsize is too large). When simulating in ensembles where temperature is being controlled, it is sensible to check that the solvent and solute equilibrate to the same temperature; if this is not the case, the use of a dual-scaling thermostat is indicated.[22] The calculation of temperature will be described in Subsection 2.2.2.

Box Sizes

In systems where a barostat is applied, it is usual to verify that the dimensions of the simulation box remain approximately constant over the course of the simulation. This can be achieved by a simple plot of the lengths of the simulation cell sizes versus time.

2.2.2 Thermodynamic Properties

A variety of thermodynamic properties may be obtained from the ensemble of configurations generated during a molecular dynamics simulation. Some of these are described below.

Temperature

Each degree of freedom of a molecular system has, on average, $k_B T$ J of energy, where k_B is the Boltzmann constant and T the temperature. Half of this energy is kinetic and half potential, and so we may define temperature as

$$T = \left\langle \frac{1}{D-C} \sum_{i=1}^{D} \frac{m v_i^2}{k_B} \right\rangle \tag{2.4}$$

where the average is taken over the ensemble of configurations in the simulation, D is the number of cartesian degrees of freedom in the system (generally 3 times the number of atoms) and C the number of constraints in the system.

Pressure

Pressure can be defined and calculated in a number of ways,[7] one of which is as follows. If we define an intermolecular virial function as

$$W = -\frac{1}{3} \sum_{i=1}^{N} \sum_{j<i}^{N} r \frac{dv(r)}{dr} \tag{2.5}$$

where i and j are indices over the N atoms in each molecule, $v(r)$ is the potential energy of the atom pair and r the distance between them, then we may define an 'instantaneous pressure' as

$$P = \frac{N}{V} k_B T + \frac{W}{V} \tag{2.6}$$

where T is the instantaneous temperature and V is the volume of the system. The average $\langle P \rangle$ should then converge on the correct pressure of the system.

Autocorrelation

A normalised time autocorrelation function[12] of some property obtained from a simulation describes how the system's 'memory' of that property decays over time. A common property studied with this technique is velocity; velocity autocorrelation functions describe how rapidly the direction of travel of the particles in a system becomes randomised. They can be calculated as follows

$$C(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{v}_i(t)\cdot\mathbf{v}_i(0)}{\mathbf{v}_i(0)\cdot\mathbf{v}_i(0)} \tag{2.7}$$

where N is the number of atoms; $\mathbf{v}_i(t)$ is the velocity of atom i at time t and $\cdot$ is the inner product. The rate at which $C(t)$ decays to 0 indicates how long the 'memory' of the system for velocity is. In some cases, this correlation coefficient may be negative at certain times, showing that the velocities of the system's particles can become anticorrelated over certain time periods. This has been explained[12] as resulting from a 'caging' effect: the particles of a simulated liquid rebound off nearby particles after moving only a short and roughly constant distance.

Diffusion coefficients may be calculated[7] for a simulation from the time integral of the velocity autocorrelation function

$$D = \frac{1}{3}\int \langle C(t)\rangle dt \tag{2.8}$$

where $\langle C(t)\rangle$ is the ensemble average of the velocity autocorrelation function. Systems which lose their velocity 'memory' very quickly—whose motions rapidly become randomised—will have small values of D, indicating a slow rate of diffusion.

Radial Distribution

To obtain a portrait of the structure of a simulated liquid, a *radial distribution function* $g(r)$ may be calculated. This is usually obtained by taking a histogram of the distances between every pair of molecules in the system within some cutoff range of one another. In principle

$$g(r) = \frac{V}{N^2} \left\langle \sum_{i=1}^{N} \sum_{j=1;j \neq i}^{N} \delta(r - r_{ij}) \right\rangle \tag{2.9}$$

where V is the system volume, N the number of molecules and r_{ij} the distance between molecules i and j. The term $\delta(r - r_{ij})$ is a delta function equal to 1 where the difference between r and r_{ij} is zero—this counts up the number of molecule pairs which are at the given distance r. In practice, $g(r)$ is calculated by binning molecular distances to form a histogram. Essentially, this function gives the probability of finding a pair of molecules at a particular distance from one another. For an atomic liquid (where orientational effects need not be taken into account), $g(r)$ will typically contain a series of peaks and troughs at short distances that decay to a constant value of unity at long range (showing that no longer range order is present).

2.2.3 Root Mean Square Deviation

To verify the stability of a protein molecular dynamics simulation, and verify the attainment of thermodynamic equilibrium, RMSD values are commonly calculated. To do this, all or part of the protein molecule is *fitted* to a reference structure—that is, the structure under consideration is translated and rotated such that the RMSD value calculated between it and the reference is minimised. A good value for a fairly rigid system is of the order $\frac{1}{2}$ Å on the backbone atoms.

Fitting Methods

The fitting strategy employed depends on the type of property to be calculated and also on the nature of the protein molecule. In particular, very flexible regions of proteins are sometimes excluded from the fitting as their large movements over the course of a simulation would harm the fit achieved and so bias the results obtained.[19] If the divergence of the protein from its initial conformation during the course of the simulation is to be studied, each frame of the molecule would be individually fitted to the first frame, and a single RMSD value obtained for the entire structure taking into account either all the atoms, or possibly just the backbone. Where information on the relative mobility of the individual parts of the molecule is of interest, one strategy that may be used is to again fit each frame of the simulation to either the first structure or a mean structure, but then calculate the RMSD of the backbone atoms of each residue individually; the average RMSD of each residue over the course of the simulation is then taken as a measure of the amount of motion in that region of the molecule. In any case, the fitting strategy used should always be stated.[10]

Trajectory Divergence

Divergence of a simulation trajectory from its initial state is commonly used as a measure of both the stability of the simulation and the achievement of an equilibrium state. If the RMSD of simulation frames grows strongly throughout the entire simulation, the structure may be unravelling, perhaps due to incorrect simulation set up (incorrect treatment of electrostatics, for example). A significant leap at the start of a simulation, converging to a relatively static state is characteristic of a correct simulation, though: the system relaxing from the initial state (usually obtained from experimental data) to converge on a more or less stable equilibrium structure.[19]

Chain Flexibility

RMSD values calculated on a residue by residue basis are often used to characterise the relative mobility of the different parts of a polypeptide chain. Highly structured parts of a protein, particularly sheets and helices, tend to be quite rigid; unstructured loop regions tend to move much more. Although one might intuitively expect loop regions with unstable structures to have relatively little evolutionary purpose in enzyme dynamics and be quite uninteresting, this is not the case. It has been observed for many enzymes including dihydrofolate reductase that such loops are often present near to active sites and their motions are vital to the catalytic process.[23] RMSD values can help pinpoint which parts of a protein are mobile in solution.

2.2.4 Direct Visualisation

Since the emergence of graphical workstation computers, direct visualisation of three dimensional protein structure has come to widespread use in both industry and academia. Principal applications are the inspection of static structures determined by X-ray crystallography or NMR, visualisation of properties derived from these structures, such as the electric field contours in the nearby space of the protein, and of course viewing of calculated dynamics trajectories.

There are a number of standard representations for proteins, many of which reflect the plastic 'kits' used in teaching basic chemistry (CPK, etc). A small number of more sophisticated representations have also been devised to represent secondary structure in abstract cartoon forms.[11, 10, 12] Beyond this, there has been relatively little innovation, particularly in representations of dynamical properties (although the Oak Ridge thermal ellipsoid program[24] (ORTEP) plots common in X-ray crystallography are noteworthy). This scarcity of useful representations would seem to be more a failure of imagination than a failure of technical capability.

2.2.5 Secondary Structure

When crystallographers first began to obtain the structures of protein molecules, one of the first things they found was that a number of structural elements were ubiquitous. Foremost among these are the α-helix and β-sheet structures, but there are others. These elements are defined in terms of repetitive features of the protein backbone. In the late 1970s an algorithm was developed for categorising secondary structure automatically, similar to the way an experienced crystallographer would do so. This algorithm, commonly referred to as DSSP[18] has come into very common use, becoming something of a standard.

There are two common uses for this algorithm. First, the determination of secondary structure for static protein structures, particularly helpful for computer visualisation approaches such as those described in Subsection 2.2.4. Second, the automatic processing of an MD simulation to uncover the stability of these structural elements over the course of the trajectory.

In this approach the algorithm is fed each of the simulation frames in turn and the element type of each residue's backbone is determined. These are then plotted on a graph with time along one axis, residue number along the other, and secondary structure indicated by a colour code. An example of this can be found in Figure 3.4. Further discussion of the interpretation of such graphs can also be found in Chapter 3.

2.2.6 Dynamical Correlation

The motions of different parts of a protein structure may be correlated with each other. A pair of related techniques have been devised for observing this.

Covariance Maps

Consider a pair of atoms in an MD simulation. The motions of the atoms about their mean coordinates can be compared. If both atoms have a tendency to move in a given direction at the same time they are said to be positively correlated; if they tend to move in opposite directions, negatively correlated. A matrix of covariances can be obtained, each element of which is derived by taking the three coordinates of each atom in the pair with respect to their mean values, and summing the product of each coordinate with that of the other member of the pair.

$$C_{ij} = \sum_{n=0}^{N} d_i(n) d_j(n) \tag{2.10}$$

where C_{ij} is the matrix element for cartesian degrees of freedom d_i and d_j, and N is the number of simulation frames. If there are A atoms in the simulation, there will be $9A^2$ elements in the matrix.

The drawback with such covariance maps is that for systems with more than a handful of atoms they are very difficult to read. It is hard to interpret the nine elements obtained for each pair of atoms: it is difficult to develop intuition for the separate correlations of the nine direction pairs for each atom pair, *xx*, *xy*, *xz*, etc.

DCCM

A modified version of the covariance map is the dynamical cross correlation map,[25,20] which obtains for each pair of atoms a single element describing the overall correlation, but loses most of the geometric information contained in the basic covariance matrix. Very simply, for each frame of the simulation, a vector is obtained for each atom describing its position relative to its mean coordinates. The matrix element for atoms i and j is defined by the expression

$$C_{ij} = \frac{\langle \Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j \rangle}{\sqrt{\langle \Delta\mathbf{r}_i^2 \rangle \langle \Delta\mathbf{r}_j^2 \rangle}} \tag{2.11}$$

where $\Delta\mathbf{r}_i$ is the displacement from the average position of the ith atom, and $\langle\ \rangle$ denotes the ensemble average. The matrix elements take values from $+1$ indicating that the motion of the atom pair is perfectly correlated, to -1 indicating perfect anticorrelation; zero represents no correlation between their movements.

A drawback of the DCCM method is that it is highly sensitive to the fitting strategy employed.[20] Small changes to the fitting produce quite different results, and this can cloud interpretation and allow a degree of subjectivity to enter the analysis.

2.2.7 Eigenvector Methods

A number of related methods have been developed that seek to define and visualise the *globally correlated* motions of molecules. These come under the headings *essential dynamics* (ED)[21] and *normal mode analysis* (NMA).[26] ED is the application of the statistical technique called *principal component analysis* (PCA) to molecular dynamics coordinate data.[27] The application of PCA to molecular simulations was first proposed as a means of calculating the effect of anharmonicity on vibrational spectra,[28] and termed *quasiharmonic analysis*. A detailed review of the various uses of these methods can be found elsewhere.[29]

Theory

A frame of a trajectory from a molecular dynamics simulation is usually described in terms of the cartesian coordinates of the atoms, and this can be written as a single vector

$$[x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_n, y_n, z_n] \tag{2.12}$$

This vector can then be considered as a linear combination of a set of orthonormal basis vectors: $[1,0,0,\ldots,0]$, $[0,1,0,\ldots,0]$, etc. The coordinate system used for the representation of each frame can have an arbitrary origin, though. The simplest representation is to use the origin of cartesian space; but another possibility is to use the *mean* coordinates of the trajectory. This change of representation can be considered as a vector subtraction, $\mathbf{E} = \mathbf{F} - \mathbf{M}$, where $\mathbf{E}$ is the vector with the mean origin, $\mathbf{F}$ represents the frame and $\mathbf{M}$ represents the simulation mean structure. If the set of $3N$ dimensional $\mathbf{F}$ vectors for the entire simulation could be visualised, they would approximate a $3N$ dimensional *hyperellipsoid* centred around a point described by the mean vector, $\mathbf{M}$ (where N is the number of atoms in the simulation). If the same were done for the new vectors, $\mathbf{E}$, we would simply get the same ellipsoid, but centred at the origin of the space. In a mean-centred coordinate system, the $\mathbf{E}$ vectors represent deviations from the mean structure.

In practice the point distribution often deviates from a perfect hyperellipsoid, particularly when the molecular system exhibits multistate behaviour: the distribution along some axes of the ellipsoid no longer approximates a gaussian, but contains multiple maxima corresponding to the different system states. Such behaviour does not prevent the method from working, though, and can present interesting possibilities for visualisation, as described below.

The motion along each coordinate of each atom accounts for a small fraction of the total variance present in the whole trajectory, and it is clearly possible to calculate this—calculate a variance number associated with each of the cartesian basis vectors. This idea can be taken further, though. It is possible to determine an *alternative* set of orthonormal basis vectors such that the first member of the set accounts for the largest proportion of the variance. Considering the hyperellipsoid again, the vector describing the *longest* axis of the hyperellipsoid is the vector along which the variance is maximised. The next longest perpendicular vector (orthogonal to the first) describes the second largest part of the variance; in all, a set of $3N$ orthogonal basis vectors may be obtained in terms of which the simulation may be described. In other words, a given frame of the simulation can be described as a linear combination of these new vectors. The 'essential dynamics' hypothesis is that nearly all of the variance in a simulation occurs along a *very* small subset of the vectors, and where a protein undergoes some large scale coordinated motion, this subset is sufficient to describe it. If this vector set is obtained, the trajectory of the protein during a simulation can be projected onto it to obtain a highly reduced and simplified description of the motions present.

In practical terms, a covariance matrix is obtained for the simulation, as described in Subsubsection 2.2.6, commonly of the backbone C_α atoms, and this is diagonalised to produce a matrix of normalised eigenvectors along with a vector of the associated eigenvalues. These eigenvalues are the variances associated with the axes of the hyperellipsoid describing the distribution (given by the eigenvectors).

The data obtained from this can take various forms. In one representation the inner (scalar) product of each simulation frame (considered as a vector) and a chosen eigenvector is calculated, and a projection is then generated by $\mathbf{C} = \mathbf{M} + \mu\mathbf{I}$, where $\mathbf{C}$ is the coordinates of the projection (the straightforward cartesian vectors of the atoms, suitable for drawing in a molecular graphics package); $\mathbf{M}$ is the mean coordinate vector; μ is the inner product for that frame and $\mathbf{I}$ is the eigenvector. A set of such coordinate projections can be compiled together to give an animation of the motions of the protein along the chosen eigenvector. An alternative representation can be obtained by calculating the projection of the trajectory onto a set of two or three of the most important eigenvectors (as judged by their eigenvalues), and plotting these projections directly as a two or three dimensional graph. Doing this can allow straightforward visualisation of multistate behaviour: where such behaviour is present, multiple well-sampled regions of the low dimensional subspace separated by poorly sampled bottlenecks may be observed (this has been seen in DHFR, for example[30]). A further set of data that is sometimes of interest is the amount of fluctuation of each atom under consideration along a given 'essential' eigenvector. A graph of fluctuation versus residue number indicates which parts of the molecule move most along the 'important' modes.

It has been argued that in certain cases, an eigenvector may serve as a useful linear approximation to a structurally significant motion of a protein system (e.g. an active site loop movement[31]). The vector of inner products obtained for the trajectory may also provide useful information about the character of the motion—whether it is a smooth movement, or whether a discrete transition event takes place.[30]

An interesting extension to the basic ED idea has been proposed,[27] where the vector curl of atomic displacements is used instead of the straightforward displacement vectors. This has allowed the derivation of 'hinge' and 'screw' axes for multidomain proteins, which are less well approximated by modes obtained from linear displacements.

Normal Modes

Normal mode analysis(NMA) is a related technique where a Hessian matrix of force constants (obtained from the force field) is diagonalised to describe the global vibrations of the molecule about the conformation from which the force constants are obtained.[26] This is generally performed on an energy minimised structure. While normal mode analysis is certainly of considerable use in calculations of the spectroscopic properties of small, rigid molecules, its application to very large, flexible systems such as proteins raises questions.[32] The possibility of conformational transitions within the protein is not taken into account, but rather the molecule is considered as if it were a rigid body: the normal modes represent the resonant modes of the structure as defined by the second derivative of the energy function with respect to position around a single, minimised structure. We know that the conformational plasticity of protein molecules is critical to their function, though, and so important aspects of protein dynamics will be left out. However, a close correspondence between the collective modes obtained by NMA and ED has been shown to exist for some protein systems and NMA has certainly seen considerable use.[27,29]

Chapter 3

Prion Protein Simulations

3.1 Background

The transmissible spongiform encephalopathies (TSEs) are a subfamily of the protein folding diseases that are linked to problems with the secondary and higher structural-stabilities of protein molecules.[3,33] These diseases commonly progress by an aggregation process where correctly folded monomeric protein molecules polymerise to form high molecular weight partially-structured aggregates. This is generally accompanied by a conformational change forming a structure rich in β-sheet.[34,35] Current methods for determining atomically detailed structural information are unsuited to analysing such insoluble aggregates, although data on the monomers can often be obtained.

TSEs include scrapie in sheep, bovine spongiform encephalopathy (BSE) in cattle, and Creutzfeld-Jakob disease (CJD) in humans.[36,37] These diseases can be transmitted between organisms, inherited, and arise sporadically; this has been explained in terms of a 'protein only' hypothesis in which the naturally occuring prion molecule (PrP^c) is converted to an aggregated amyloid form (PrP^{Sc}) through an autocatalytic templating process.[36]

The function of PrP^c is currently unknown, though a role in copper transport has been proposed.[38] PrP^c is bound to the exterior membrane surface of cells by a glycosylphosphatidylinositol (GPI) anchor and contains a pair of large N-linked sugars. The protein is known to undergo rapid endocytosis (the process by which surface proteins are cycled into and out of the cell in vesicles), and there is evidence that the *in vivo* process of conversion to PrP^{Sc} takes place in acidic vesicles along the endocytic pathway.[39] Genetic studies[40] have suggested that conversion requires the assistance of a currently unknown chaperone, 'protein X', which binds to a four residue discontinuous epitope comprised of glutamine 168, glutamine 172, threonine 215 and glutamine 219.

The primary structure of PrP^{c} has an influence on the nature of TSE diseases.[41,42,43] A number of point mutations have been linked to various conditions in humans, including CJD, fatal familial insomnia (FFI) and Gerstmann-Straussler-Scheinker disease (GSS). A natural methionine/valine polymorphism occurs in humans at position 129; homozygosity for methionine at this position leads to a slight predisposition towards sporadic CJD.[44] Mutation of aspartate residue 178 to asparagine has been linked to inherited disease; the disease phenotype depending on the residue at position 129, valine being associated with CJD and methionine being associated with FFI.[42]

Work by Liemann and Glockshuber[43] on the thermodynamic stability of amino acid substitutions related to inherited prion diseases has concluded that thermodynamic destabilisation of PrP^{c} is not likely to be a *general* mechanism underlying the formation of PrP^{Sc}. However, their data indicate that the asparagine 178 mutation reduces the thermodynamic stability of PrP^{c}, although the residue 129 polymorphism does not cause a significant change. This suggests that more subtle structural changes differentiate disease phenotypes.

In this Chapter are reported a series of molecular dynamics simulations of four point variants of the prion protein: the wild type protein structure obtained from the protein databank with methionine at position 129 and aspartate at position 178, ***MD***; the polymorphic variation in which residue 129 is replaced with valine, ***VD***; the disease causing variant associated with inherited FFI where aspartate 178 is replaced by asparagine, ***MN***; and the variant associated with inherited CJD, ***VN***.

The purpose of these simulations was to shed light on the dynamics of the protein, and to discover any changes brought about to the structure or dynamics by a pair of polymorphs. The first was the disease causing 178 mutation and the second the polymorphism associated with it in determining disease phenotype in the inherited and sporadic disease forms. Of interest were both global changes, such as a breakdown in, or modification to, the secondary or tertiary protein structure, and local changes such as the pattern of hydrogen bonding in the immediate vicinity of the modified residues. Further, might any altered states discovered represent an intermediate step in the mechanism of the global conformational change pathway that leads to the pathogenic form of the protein?

At the time this work was started, only the solution NMR structures of the mouse[45] and Syrian hamster[46] forms of PrP^c were known. Since then, the structures of the bovine[47] and human[48] forms have been solved. The Syrian hamster prion protein used in this work is shown in Figure 3.1.

Several previous simulations of the prion protein have been done in Southampton and by others,[49, 50, 51, 52, 53] and these will be summarised later. The previous work in Southampton by Parchment and Essex[19] on mouse and Syrian hamster PrP^c concluded that extra residues present at the C- and N- termini of the hamster form stabilised regions of the protein containing parts of the binding epitope for the proposed chaperone protein, 'X'. Also, the region of the protein between and including the first β-strand and helix **A** was found to be more flexible in the hamster form, and the first strand was observed to form a stable β-bridge with N-terminal residues.

3.2 Methods

3.2.1 Simulation

The simulations were initially set up by Parchment as follows. The Syrian hamster prion protein (SHa PrP^c), was obtained from the Brookhaven protein databank, entry 2PRP. The point mutations were made by hand; the asparagine 178 residue was oriented such that the amide oxygen is closest to the hydroxyl group of tyrosine 128. It was necessary to remove residues glycine 90 through methionine 112 from the highly flexible N-terminus to reduce the necessary simulation box to a tractable size. The protein molecules were solvated in a 6 Å

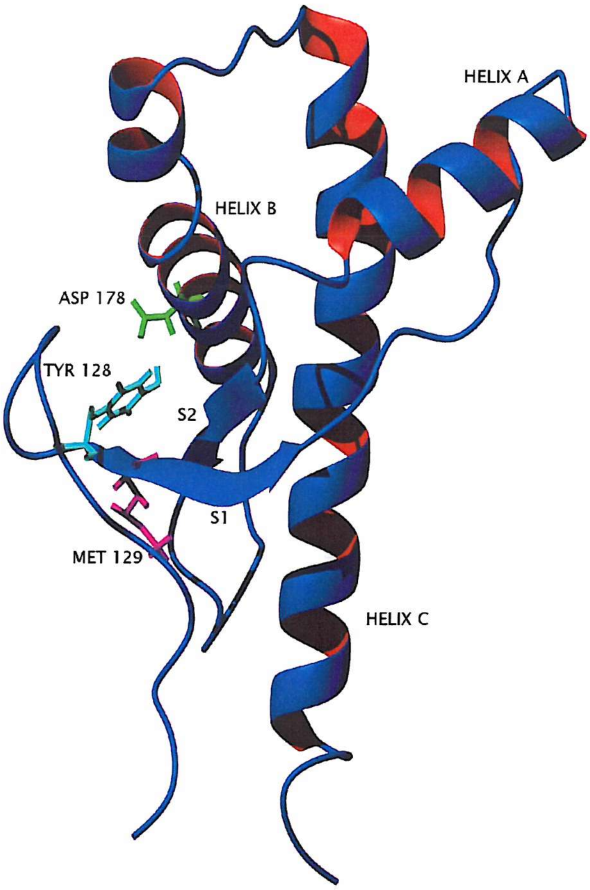

Figure 3.1: The NMR solution structure of the Syrian hamster prion protein used in these simulations. Helices A, B and C and β-strands S1 and S2 are labelled. Aspartate 178 is drawn in green, tyrosine 128 in cyan and methionine 129 in magenta. The directions of the β-strands are indicated by arrowheads.

layer of TIP3P water molecules[54] and neutralised by placing counterions into the simulation system after solvation using the AMBER 5.0 LEAP program.[15] Approximately 4800 water molecules were included in each simulation. Periodic boundary conditions were used.

Errors were found in the molecular dynamics protocol used in the original work on these systems by Parchment. Once this was discovered a corrected series of simulations were performed by this author, and these are reported here.

The four simulations were carried out with identical protocols using the AMBER 5.0 software[15] with the all-atom parm96 parameter set.[16] The solvated protein structures were subjected to 500 steps of steepest descent, followed by conjugate gradient minimisation until the RMS of the cartesian gradient was less than 0.01 kcal mol^{-1} Å^{-1}. During the first stage of molecular dynamics, the minimised structures were heated from 10 to 298 K in a series of steps lasting 60 ps in total; stability was checked after each of these 60 ps heating steps. Temperature and pressure were controlled using the Berendsen algorithm,[22] scaling solute and solvent atoms separately. All bond lengths were constrained using the SHAKE algorithm[14] and the simulation timestep was 2 fs. The particle mesh Ewald method[13] for long range electrostatic interactions was used; this was shown to produce stable simulations of the prion protein by Zuegg and Gready.[50] A 12 Å based cutoff was used for Lennard-Jones interactions.

The second and third stages of molecular dynamics simulation consisted of 1 ns of equilibration and 5.4 ns of data collection. These were also done in the isothermal-isobaric (NPT) ensemble and followed the same protocol as the heating stages. During these stages data was written to file every 100 fs.

3.2.2 Analysis

The protein trajectory was analysed using a combination of both residue- and time- based root mean square deviation (RMSD) calculations, secondary structure versus time plots, examination of the occupancy of key hydrogen bonds, dynamical cross correlation (DCCM) maps[25,20] and trajectory visualisation. Also, simulation box sizes and energies were determined as functions of simulation time to verify stability.

Standard error plots were produced for the DCCM maps.

The RMSDs were calculated with respect to the first structure in the simulation (after the energy minimisation, but before the heating period). Residues 113–123 were highly mobile during the simulation and consequently excluded from the fitting procedure. Residue based displacements were determined by fitting each molecule in the trajectory to the first molecule and calculating the RMSD for the backbone atoms on a residue by residue basis. Time based RMSDs were determined for all backbone atoms fitted as described above.

The first set of simulations of this system by Parchment were also analysed using the essential dynamics analysis method.[21] Unfortunately, the results were found to be difficult to interpret, and this method was not used for the current set of simulations.

3.3 Results

3.3.1 Simulation Stability

Box sizes and energies were plotted as functions of simulation time (data not shown). These were all found to reach reasonably stable equilibrium values and maintain these over the course of the data collection stage of each trajectory.

Time based RMSD values for the backbone atoms also give an indication of protein stability; these are shown in Figure 3.2. The ***VD*** structure proved to have the lowest deviation from the initial minimised structure, remaining within 2.5 Å throughout the trajectory. The ***MD*** structure remained within about 3 Å for most of the simulation, diverging to around 3.5 Å for a period in the middle. The ***MN*** simulation remains between 2.5 and 3 Å for most of the simulation, diverging towards 3.5 Å near the end. The trajectory with the highest RMSD values was the ***VN*** mutant, which diverged rapidly to around 3 Å within the first 1 ns, edged up slowly to around 4.5 Å by the 5.5 ns mark, and then increased abruptly to around 5.0 Å during the last nanosecond of simulation time. In interpreting this result for ***VN***, one must be mindful of the fitting strategy used; residue based RMSD plots and visual comparisons of the mean

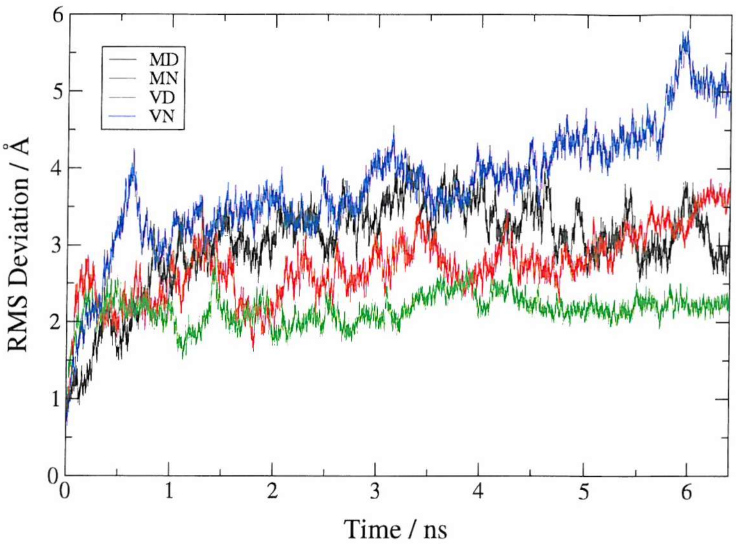

Figure 3.2: RMSD versus time over the equilibration and data collection stages of the simulation; the fitting strategy excluded residues 113–123.

simulation structures indicate that a region comprising the N-terminal end of helix **A** and the coil region immediately prior to it has a very high mobility. This will cause an overall increase in the RMSD with respect to the first structure in the equilibration stage of the simulation.

3.3.2 Residue Based RMSD

Examining the graph of residue based RMSD (Figure 3.3) it is clear that an 'arm' region consisting of residues at the beginning of helix **A** and the coil just before it has a very high RMSD with respect to the initial structure in the ***VN*** trajectory, as noted above, and a lower but still significant RMSD in ***MN***. Other differences between trajectories are less clear. The ***VN*** and ***MD*** trajectories have significantly higher values from around residue 190 to residue 200 in helix **B**. It can be seen from the secondary structure graph—Figure 3.4—of the ***MD*** simulation that the 5 C-terminal residues of helix **B** suffer significant breakage, which explains their high RMSD values and those of the coil region prior to helix **C**. Secondary structure data do not indicate a breakdown in the ***VN*** trajectory, though.

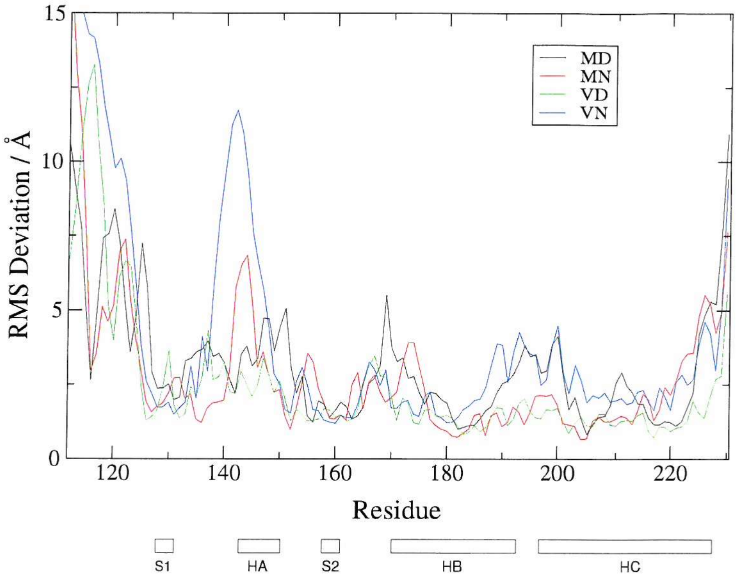

Figure 3.3: RMSD as a function of residue number. The boxes indicate the positions of secondary structure elements: HA, HB and HC are the α-helices and S1 and S2 the β-strands.

3.3.3 Secondary Structure

In an effort to gain greater insight into the structural and dynamic differences between the simulations, secondary structure data was calculated using the DSSP algorithm[18] as implemented in MOLMOL.[10] The results for each structure are plotted as a function of simulation time in Figure 3.4. These data confirm the basic stability of the simulations: overall, the secondary structure in each trajectory was very stable. The α-helices remained generally intact with minor exceptions detailed below, whereas the β-sheet showed a degree of breakup and size change dependent on the amino acid substitutions.

The DSSP method for detecting helices and sheets is based on analysis of the pattern of hydrogen bonding. The presence or absence of a hydrogen bond is defined precisely in the algorithm to obtain this, although to quote Kabsch and Sander,[18] "There is no generally correct H-bond definition, as there is no sharp border between the quantum mechanical (wave-function overlap dominates at short distances) and electrostatic (electrostatic interaction dominates at larger distances) regimes and no discontinuity of the interaction energy as a function of distance or alignment." This is clearly not a defect in the algorithm itself, which provides definitions for structural properties that are intrinsically subjective; nevertheless, care must be

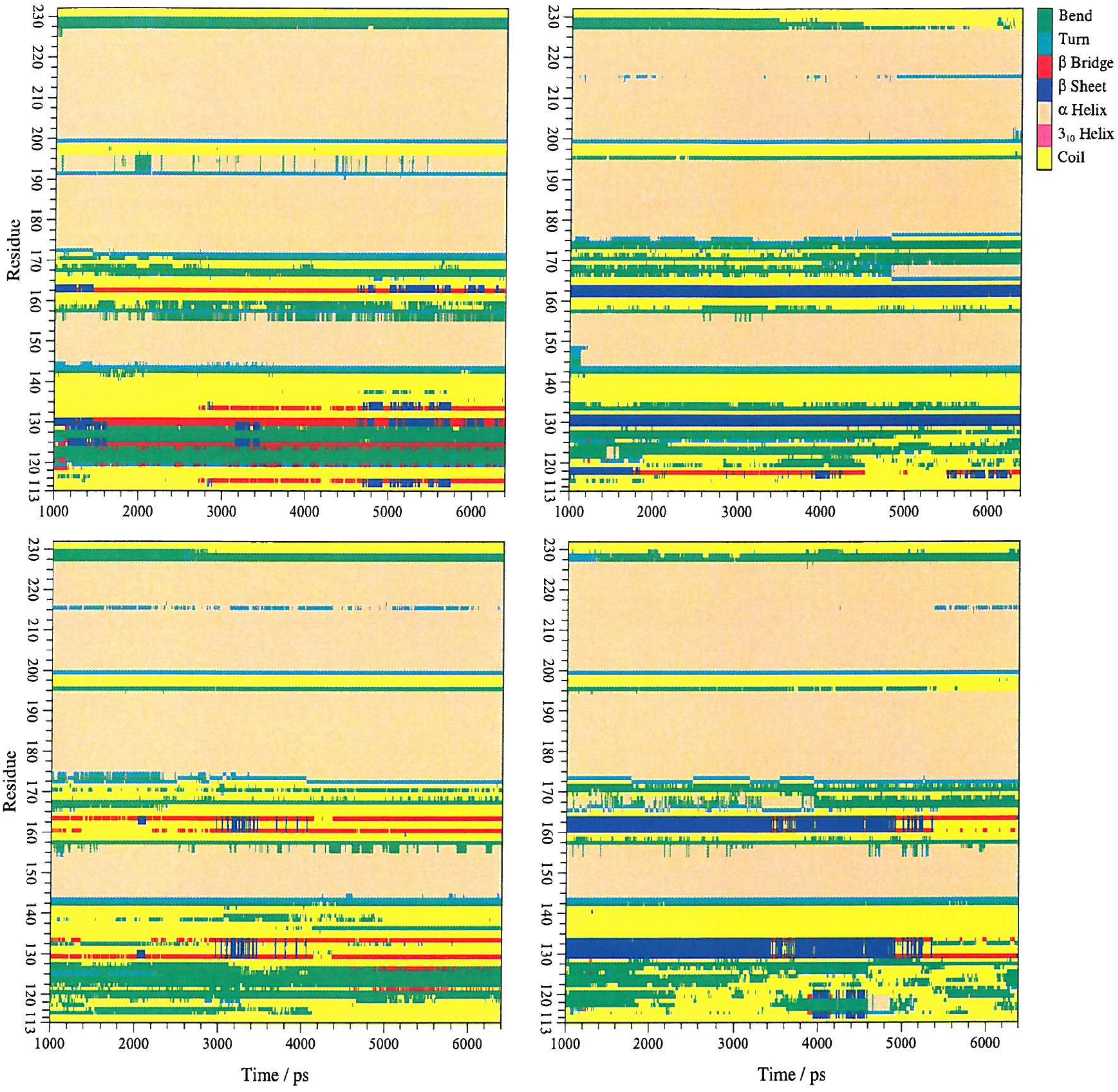

Figure 3.4: Secondary structure as a function of time, as determined by the DSSP algorithm. Clockwise from the top left: ***MD***, ***MN***, ***VN***, ***VD***.

taken not to overinterpret the results of hydrogen bond or secondary structure analysis.

Helix **A** extended from residue 144 to 156 in all simulations, with a breakdown to coil of the last two residues occasionally in ***VD*** and more commonly in ***MD***; however, visual inspection of the corresponding trajectories showed no clear conformational changes taking place.

As noted in the previous section, the last five residues of helix **B** undergo significant breakdown throughout the whole 5.4 ns of the data collection stage of the ***MD*** simulation. In the other simulations helix **B** is one residue shorter (extending up to residue 194) and consistently stable at the C-terminal end. Visual comparison of structures from the simulation showed a distinct bend at the C-terminal of helix **B** in the ***MD*** structure. The N-terminal end of this helix is poorly defined by DSSP in all trajectories other than ***MD***. The first helical residue varied between 172 and 173 in ***MD***, 173 and 175 in ***VD***, 175 and 177 in ***MN***, and 173 and 174 in ***VN***; this indicates a degree of conformational plasticity.

The length of helix **C** was generally constant throughout each simulation, with a rare and insignificant single residue gain at the C-terminal end in the ***MD*** and ***MN*** structures. A sporadic breakage occurred at residue 215 in all simulations other than ***MD***. Although visual inspection of the trajectories indicated a slight bend at this position in the ***VN*** simulation, no clear bend or break was observed in ***MN*** or ***VD***.

A short α-helix was observed to form at times between residues 166 and 168 in the ***MN*** and ***VN*** trajectories.

Examination of the positions at which β-sheet would be expected from the NMR structures shows clear differences between the simulations. The disease-associated ***MN*** and ***VN*** mutants generally have very stable β-sheet structures compared to the innocuous strains, ***MD*** and ***VD***. DSSP detects a significant breakdown in the β-sheet towards the end of the ***VN*** simulation, though, and this will be returned to below.

In the valine 129 strains, ***VD*** and ***VN*** the region pinned together by β-bonding is constantly (4,5)-sized, even though the sheet structure itself is mostly reported as broken by DSSP. The ***MN*** structure has a consistent (3,3)-sized sheet. The ***MD*** data indicate a fairly consistent

(2,1)-sized β-sheet at these same positions.

The formation of sheet- and helix- type structures of varying degrees of stability can be seen at multiple positions below helix **A** in all simulations. The ***MD*** trajectory shows sheet formations between 133–134 and 114–115, 119 and 123, 124–125 and 128–129, and 129–130 and 162–163, only the last of which is present in the NMR structure. The ***MN*** trajectory shows a β-sheet arrangement between the first β-strand and residues 117–118 and 116–117 at various points. The ***VD*** trajectory contained a degree of β-bridging between residues 121 and 126, and the ***VN*** simulation showed both a short sheet forming from 114–115 and 119–120, and a brief four residue α-helix formation from 116 to 119.

These results suggest that the N-terminal region of the protein, up to helix **A** has a high degree of conformational plasticity. The structures derived from NMR spectra indicate the presence of one sheet region, but the current simulations indicate that other, more transient structures may form in the $\mathrm{PrP^c}$ monomer.

It is known from statistical examination[55] of protein structures that valine has a greater propensity to take part in β-sheets than methionine. This is not inconsistent with the greater sheet size in the valine polymorphs, although the increase in sheet length takes place at the other end of the sheet from residue 129.

3.3.4 Dynamical Cross Correlation Maps

Dynamical cross correlation maps are a means for determining the large scale concerted motions of groups of atoms in a molecular dynamics simulation. The method, which has been described in detail elsewhere,[25,20,56] consists of calculating a matrix of normalised cross correlation coefficients of the motions of atoms with respect to some average structure. This is been described in Subsection 2.2.6.

Standard error values were obtained for all matrix elements. This was done by dividing the 5.4 ns of simulation trajectory into nine 600 ps blocks, calculating the DCCM matrix for each block separately, and then determining the mean and standard error values for each matrix element. All error values for all matrix elements of the four simulations were found to be of very small magnitude, below 0.05.

DCCM plots where all backbone atoms (excluding residues 113–123) were fitted to the mean structure calculated from the same set of atoms showed a strong correlation between helix **A** and the coil region immediately before it in the ***VN*** simulation, but none of the others. This suggests that the 'arm' region is made more rigid by the valine polymorphism, and is shown in Figure 3.5.

DCCM plots fitted to just helix **B** were produced (Figure 3.6). This has the effect of removing correlations between this helix and other structural elements, and what little remains of such correlations cannot be interpreted meaningfully.[56]

The resulting plots indicate a significant breakage in helix **C** at around residue 214 in the aspartate (***MD***and ***VD***) structures which is not present in the disease associated mutants. It appears that the helix is broken into two independent parts at this residue position, each possessing a high degree of internal correlation, but each relatively uncorrelated with the other. The detection of this *seems* to indicate that the secondary structure information—where a break in the helix was registered at residue 215 in all structures *other than* ***MD***—is limited in describing a simulation's dynamics; no clear picture of what was happening could be discerned from visualisation of the trajectories, though.

A much higher degree of correlation is seen between the unbroken helix **C**, and helix **A** and the coil region immediately before it in the ***MN*** and ***VN*** trajectories; this shows a greater degree of correlation between the rigid core of the protein consisting of helices **B** and **C**, and the flexible regions in these disease associated structures. These proteins also show highly stable β-sheet structures, which suggests that the β-sheet plays a role in pinning the more flexible regions of the protein to the rigid core.

A plot fitted to helix **C** was also created (data not shown), which provided a third view of the dynamics. Unfortunately, it was difficult to draw any further consistent conclusions from the various DCCM plots. The delicate dependence on the fitting strategy used and the contradiction between DCCM and secondary structure data and trajectory visualisation suggested that a degree of caution should be excercised with this method: it seemed difficult to prevent

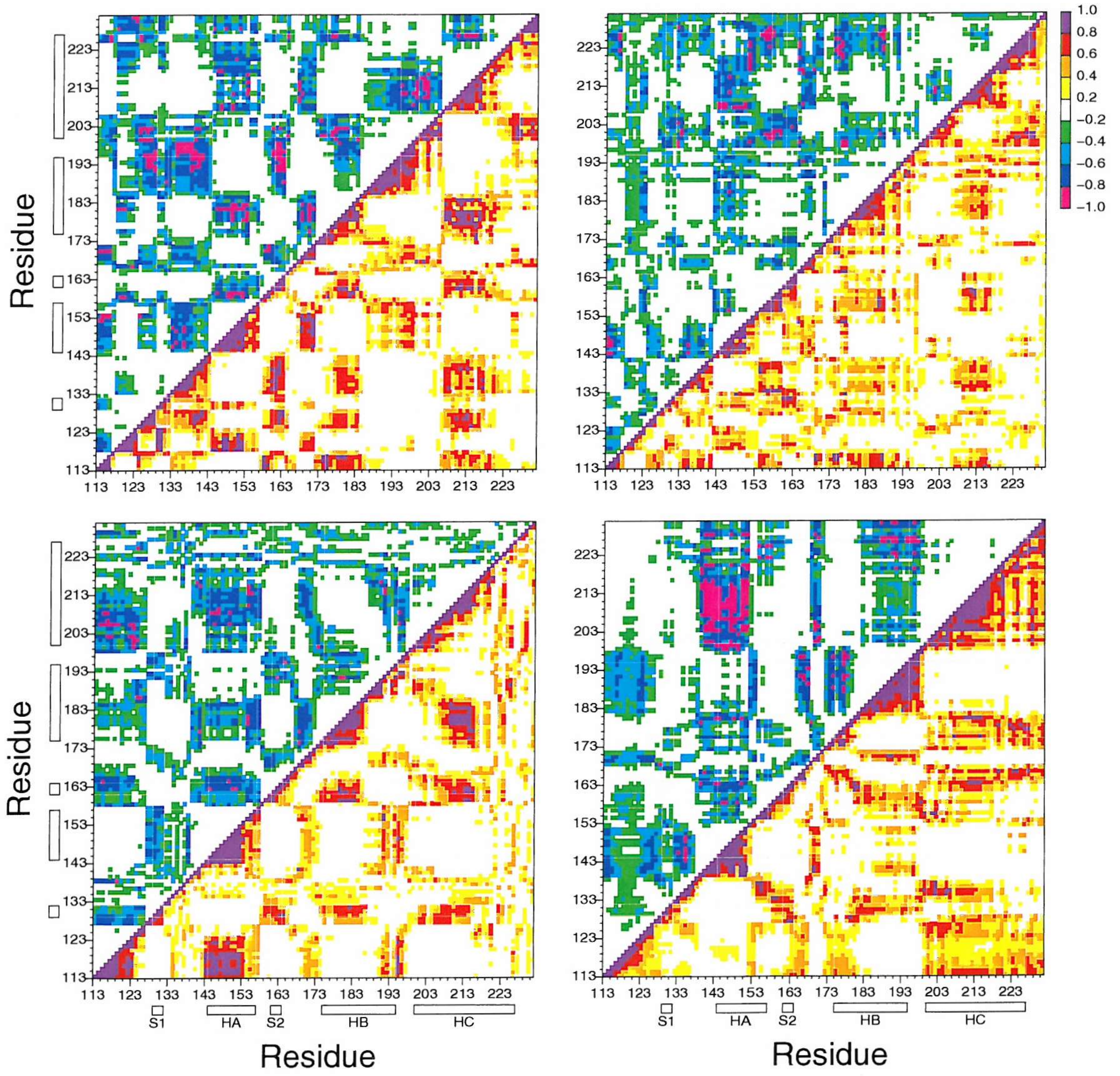

Figure 3.5: DCCM plots with backbone atoms fitted to the mean structure of residues 124–231. Clockwise from the top left: ***MD***, ***MN***, ***VN***, ***VD***.

an element of subjectivity entering the interpretation of the results.

3.3.5 Hydrogen Bond Profile

Data for the existence or nonexistence of the 128–178 bond as a function of simulation time were obtained from the DSSP algorithm as implemented in MOLMOL (Figure 3.7). This bond is of interest because residue 178 is the site of the disease causing D178N mutation, and residue 129 is the site of the methionine/valine polymorphism which determines the phenotype of the corresponding disease, as described in Section 3.1. Although, as stated above, care must be taken in interpreting such discrete bonding data, they do allow a useful general comparison to be made between the trajectories. The aspartate 178 trajectories, ***MD*** and ***VD***, showed 56% and 99% occupancy respectively. The asparagine trajectories showed a lower degree of bonding, as might be expected from replacing a charged group with an uncharged one: 37% for the ***MN*** and 2% for the ***VN***.

During the first part of the ***VN*** simulation, the bond is reported very infrequently; however from around the 4.8 ns point, a 100 ps long band of quite high occupancy is observed; after this the bond occurs much more commonly. This high occupancy band is coincident with the breakdown of the β-sheet seen in the secondary structure plots. The events appear to be linked.

Lengths of the five backbone-backbone hydrogen bonds that make up the β-sheet structure were determined from the simulations as a function of time to supplement the data obtained from DSSP (Figure 3.8). The specific bonds examined were obtained from an inspection of the ***VN*** trajectory, which possessed the largest and most stable sheet. The bonds were:

1. valine 120 NH → tyrosine 164 O
2. valine 120 O ← tyrosine 164 NH
3. glycine 122 NH → valine 162 O
4. serine 123 O ← valine 162 NH

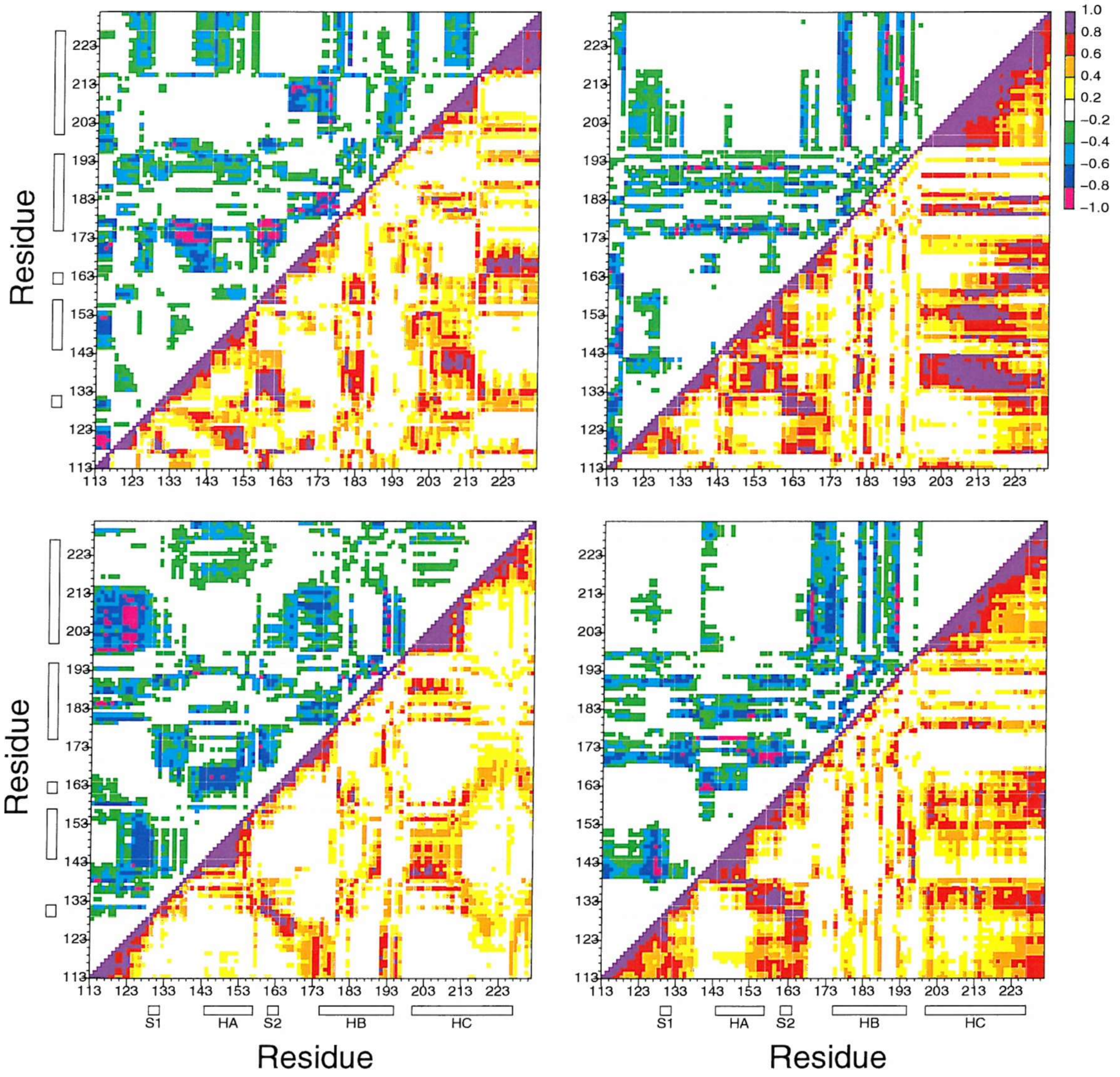

Figure 3.6: DCCM plots for the four trajectories. Only the backbone atoms of helix **B** were included in the fitting to remove correlations between this helix and other elements. Clockwise from the top left: ***MD***, ***MN***, ***VN***, ***VD***.

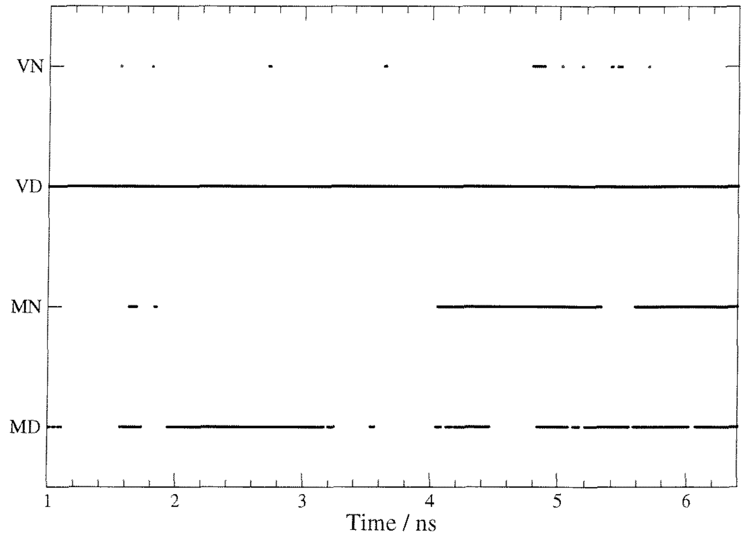

Figure 3.7: Profile of the hydrogen bond between residues 128 and 178, as determined by the DSSP algorithm. Points indicate where the bond exists.

5. methionine 125 NH → asparagine 160 O

The ***MN*** trajectory showed the first three of the five bonds to be formed constantly throughout the trajectory (staying at around 2 Å long). The last two 'bonding pairs' maintained fixed distances of over 5 Å, again over the whole simulation. This result is completely consistent with the corresponding secondary structure data.

Hydrogen bond data was obtained for the data collection stage of the simulation (1–6.4 ns). In the ***MD*** trajectory, the second and third hydrogen bond pairs listed above were formed consistently throughout the trajectory. The first pair was within plausible bonding distance for the first 500 ps of the data collection phase, jumping abruptly to a distance of around 4–5 Å for the next 3 ns, before returning to a distance of around 3 Å for the rest of the simulation. The fourth and fifth pairs remained at around 4–5 Å over the whole simulation.

Distances of all five studied bonding pairs in the ***VN*** simulation remained at around 2 Å for the first 3.4 ns of the trajectory. At this point the third bonding pair jumped very abruptly to a distance of around 4 Å. This bond length then decreased a little, to around 2.5–3.5 Å up to the 4.8 ns point, where the distance increased again to over 4 Å where it remained to the end of the simulation. The fourth bond length drifts to 4.0 Å at around the 5.4 ns point, and remains at around this separation to the end of the simulation.

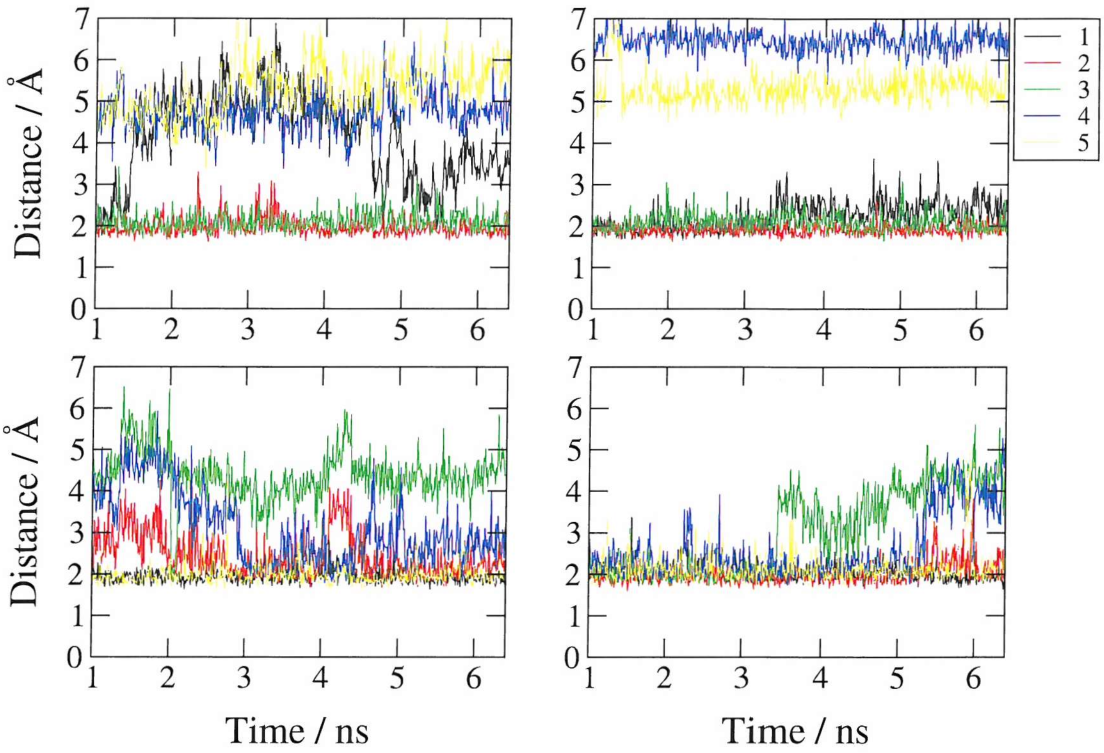

Figure 3.8: Lengths of the five backbone-backbone hydrogen bonds that make up the β-sheet structure. The numbered bonds are described in the text. Clockwise from the top left: ***MD***, ***MN***, ***VN***, ***VD***.

These results for the ***VN*** simulation are entirely consistent with the secondary structure data. The β-sheet shows initial signs of breakdown at the 3.4 ns point when the third bond distance first jumps, then reforms as the bond distance decreases again, temporarily. The second upward jump in the pair separation at 4.8 ns is coincident with the more significant breakdown of sheet structure occuring at this time, and also with the period of 128–178 bond existence described above. The sheet is still detected by DSSP to some extent up to the 5.4 ns point when the fourth bond breaks; after this, the structure is no longer recognised as sheet, even though the first, second and fifth bonds are still present, pinning the sheet region together at both ends.

The ***VD*** trajectory shows continuous bonding between the first and fifth pairs, holding the sheet structure together at both ends; the second pair remains at around 3 Å distance until around 2 ns, at which point it drifts to around 2 Å and remains there for the rest of the calculated trajectory. This is similar to the behaviour observed in the ***VN*** trajectory: the residues in the centre of the sheet wander apart a short distance, while the end residues remain bonded and prevent the structure from breaking up permanently. The appearance of this templating behaviour underlines the care that must be taken in interpreting data from secondary structure finding algorithms such as DSSP.

3.3.6 Visualisation

Various software tools were used to make a visual inspection of the trajectories generated, with comparisons being made to data obtained by other means.

Mean structures from the data collection phase of the simulation were generated and compared. These all appeared very similar, with the interesting exception being the 'arm' region in the ***VN*** trajectory; this had visibly come away from the rigid core of the protein, in contrast to the other three structures.

Other comparisons are referred to where appropriate in the rest of the text.

3.4 Discussion

The very high residue RMSD values and visualisation of the structures seen in the simulation suggest the 'arm' region acts as a rigid mobile rigid mobile unit, coming away from the protein in the ***VN*** trajectory; a high RMSD is also seen in the ***MN*** simulation. DCCM plots indicate a higher degree of positively correlated motion with the rigid core region of the protein consisting of α-helices **B** and **C** in both the ***MN*** and ***VN*** simulations. When locked to the core the 'arm' region cannot move with respect to it in a coherent way, and any relative motions present are small. Also, helix **C**'s dynamics become split at residue 214. When the 'arm' breaks away coherent motion takes place, positively correlated with helix **C**.

The β-sheet seen in the NMR structure appears to be destabilised by bonding between residues 178 and 128. Bonding length data indicate that sheet residues remained in a fairly constant arrangement throughout the ***VN*** and ***VD*** simulations, and that residues at either end of the region hold the backbone strands in close proximity, effectively forming a template for the structure. These observations suggest that the β-sheet can form and melt repeatedly over time, and the presence of a stable sheet is facilitated by the disease associated asparagine mutation. The polymorphism at residue 129 has an effect on the length of the sheet-like region, increasing it to a (4,5) pattern.

The secondary structure data for the region of the protein N-terminal to helix **A** shows a high degree of conformational plasticity. This ability to form multiple β-bonding arrangements in a flexible part of the protein is interesting in the light of the experimental conclusions of Chiti *et al.*[57] that amyloid formation takes place from a partially denatured state where backbone hydrogen bond donors and acceptors are to some extent exposed, and that understanding of the process by which a particular protein forms amyloid requires knowledge of the balance of equilibria between its various forms. This experimental data provides a degree of corroboration of these simulation results.

Zuegg and Gready have published two papers detailing simulations of the prion protein. The first[50] stressed the importance of correct treatment of electrostatic interactions: truncation methods created stability problems, while the Ewald method was found to prevent this. The second paper[51] detailed separate simulations of the protein with bound N-linked sugars, and of the GPI anchor bound to a membrane. It was found that the sugars slightly increased the rigidity of helices **B** and **C**, though other than this had little effect on the conformational structure of the protein. Alonso *et al.*[49] performed 10 ns simulations of Syrian hamster $\mathrm{PrP^C}$ at neutral and acidic pH. Their work showed the importance of simulating protein molecular dynamics for a sufficiently long period of time (multiple nanoseconds) to observe changes in structure. They found that acidic pH increased global flexibility and increased β-sheet content. El-Bastawissy *et al.*[52] performed simulations of a mouse homology model of the human protein, along with the human protein NMR structure. Both the wild type and the asparagine 178 mutant structures were simulated. They found stability problems with the cutoff and counterion methods for treating electrostatics, as also reported by Zuegg and Gready.[50] In their work the wild type structure appeared extremely stable even at high temperatures, and an increase in β-sheet content was detected in the mutant structure. A paper by Gsponer *et al.*[53] investigating the mouse protein and its asparagine variant showed a high flexibility at the N-terminal end of helix **A** and the residues immediately before it. This was in contrast to the previous results of Parchment and Essex[19] which found these regions to be quite rigid in the mouse structure. This might be due to differences in simulation protocol; for example, their simulation took place in a sphere of water treated partly by Newtonian and partly by Langevin dynamics, and they treated electrostatic interactions with a shift function, whereas the earlier simulations of Parchment and Essex[19] followed a protocol similar to the current work.

The simulations reported here were performed using the Ewald electrostatic method for a period of 6.4 ns, and found to be stable. Earlier, shorter runs with a different electrostatic treatment did not reach equilibrium during a period of 2 ns. The motion of the 'arm' region as a single rigid unit found in the ***VN*** and ***MN*** simulations also occurs in all of the simulations of the mouse prion protein reported by Gsponer *et al.*[53] suggesting that the differences between the mouse and Syrian hamster protein structures change the dynamics. The stabilisation of the β-sheet by the asparagine mutation seen in the simulations appears at least partly consistent with the conclusions of El-Bastawissy *et al.* although in their simulations an increase in sheet size to a (5,5) structure—also determined by the DSSP implementation within MOLMOL—was found, which is inconsistent with the ***MN*** results. The formation of secondary structural elements within the coil region N-terminal to sheet strand 1 is consistent with the earlier simulation of wild type Syrian hamster prion protein by Parchment and Essex,[19] the simulations reported in the paper on electrostatic treatment by Zuegg and Gready, and the simulations under acidic conditions by Alonso *et al.*

Some of the earlier work in Southampton on this protein, along with some other groups' simulations of various proteins, suffers from a degree of overinterpretation of secondary structure and hydrogen bond data. To avoid this here, effort was made to visually relate interpretations of such data to the structures present in the trajectory, and also to seek other types of observation to back up conclusions.

3.5 Conclusions

3.5.1 Prion Protein Simulations

The molecular dynamics trajectories of the various mutant structures of the protein showed a number of subtle structural changes with respect to the wild type simulations. The 'arm' region consisting of the N-terminal residues of helix **A** and the coil region before it can sometimes break away from the rest of the protein structure and move independently as a rigid body. The sheet is destabilised by the hydrogen bond between residues 128 and 178—this hydrogen bond is less stable in the disease associated (D178N) mutants—and the valine 129 polymorphism increases the length of the sheet to a (4,5) arrangement. Methionine homozygosity at residue 129 is weakly associated with sporadic CJD. Also, residues at either end of the sheet have a higher tendency to remain connected in the valine 129 polymorphs; this may form a template that aids the formation and melting of the rest of the sheet. In general, the residues N-terminal to helix **A** have a high degree of conformational plasticity.

The results regarding the stability and size of the β-sheet seem to point in opposite directions. The mutations associated with disease, ***MN*** and ***VN***, both possess more stable sheet structures, but the weakly predisposing methionine polymorphism, ***MD***, reduces the number of residues in the sheet—does the presence of β-sheet structure in the correctly folded state increase or decrease the predisposition to disease? It is possible, though, that the small β-strand which appeared to be more common in the N-terminal region of ***MD*** than ***VD*** is significant.

The general picture that emerges from these simulations is one of a protein under internal tension, whose quantity of β-sheet structure is delicately balanced. The effects of the modified residues studied are to change the size and stability of the β-sheet region, and also the internal tension. The disease associated D178N mutation tends to increase the stability of sheet and to loosen the ties between the rigid helical core and the more flexible parts of the protein, in particular the 'arm' region (suggesting that the sheet is not primarily responsible for connecting the flexible regions of the protein to the core). It is possible that this might lower the steric hindrance to other protein molecules approaching the sheet, and hence reduce the resistance to amyloid formation; similar protective structural motifs have been observed in many other proteins.[58] The codon 129 polymorphism that determines inherited disease phenotype increases the size of the sheet-like region, and creates a templating effect, helping the sheet reform after breakage events.

3.5.2 Analysis Methods

The broader conclusion to be drawn from this work is that the extraction of meaningful information about the structures and dynamics of protein simulation trajectories is a very difficult undertaking. Even in cases such as this where there clearly exist major differences between the behaviour of related structures, analysing and describing them with numbers, graphs or words is hindered by the techniques and concepts available.

The information from essential dynamics[21] was found to be difficult to interpret for this protein, and so was not included in this analysis. Both the earlier ED work by Parchment on the first simulations of this system, and a further attempt by this author, resulted in sets of data that seemed to contain little clear information. The differences in dynamics between the structures seemed to be too subtle to allow straightforward interpretation of the ED information. Statistical measures of RMS deviation were useful in determining a general picture of the relative stability of different parts of the molecules, but revealed little about the details of the dynamics. Also, RMSD information depends sensitively on the fitting strategy used, and this requires a subjective judgment on the part of the analyst. Secondary structure information seemed more useful for making sense of the differences between the proteins' structures, but it is unclear how sensitive this technique is. What sort of an effect does the discrete nature of the hydrogen bonding definition have on the data obtained, and how much emphasis should be placed on the fine details of changes between different structural element types? The 'templated' region was not picked up by the DSSP algorithm: a direct examination of the hydrogen bond lengths was required. Dynamical cross correlation maps revealed the effect of the valine mutation on the rigidity of the 'arm' region, but also presented a picture of the dynamics of helix **C** that was inconsistent with the secondary structural information. No clear explanation of this could be found from visual inspection of the trajectories: this was a purely dynamical, rather than structural effect. As with RMSD calculations, DCCM requires a subjective judgement by the analyst about the fitting strategy to be employed.

The analysis of protein molecular dynamics would be greatly facilitated by having a broader palette of tools and structural concepts to work with. Part of this problem seems to be that the forms of motion and dynamical change undergone in a protein system are as yet poorly characterised. It might be useful in developing such approaches if the motions of a protein system were characterised in a fashion more in tune with the thinking of physicists than biologists. If protein motions are divided into two categories, harmonic and discrete, then methods might be developed for comparing and contrasting the trajectories of similar proteins, and also spotting the time and pattern of any conformational changes taking place.

Chapter 4

Hilbert Huang Transform Background

As shown in the previous chapter, the analysis of biomolecular simulation data by automated procedures presents significant difficulties. It is much easier to state that slightly different structures have subtle changes in their dynamics than to describe what those changes are, or what effects they are likely to have. One avenue of attack on this problem is to separate the motions of the system on the basis of frequency and time.[59] How do the vibrational frequencies change as the conformation of the system moves across its potential energy surface? Related to this, some recent approaches[60,61] to improving the sampling observed in simulations have used digital filters to modify their energy profile in a frequency specific way. Such techniques require characterising and parameterising, and this would be greatly facilitated by having accurate time-frequency-energy profiles of their trajectories. A number of mathematical approaches to these problems have existed for many years, to which very recently has been added a new one: the Hilbert Huang transform.[62]

4.1 Fourier Transforms

The Fourier transform was originally developed as a method for solving differential equations in thermodynamics. It works on the principle that an arbitrarily complicated function may be constructed by summing a series of simple trigonometric 'basis' functions. This powerful concept has found extensive application throughout mathematics and the sciences. Indeed, the Fourier transform has become essentially synonymous with the idea of the spectral representation of a signal.

To determine the size of the contribution to a target function $f(t)$ extending over $-\infty < t < \infty$ of a trigonometric wave with angular frequency ω, one integrates the product of the functions

$$g(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt \tag{4.1}$$

where $g(\omega)$ is a complex number whose modulus represents the amplitude of the wave, and whose argument represents the corresponding phase angle; i represents $\sqrt{-1}$. If this is repeated at many separate frequencies, ω_j, then a 'spectrum' of the target function is derived. If a sufficient set of orthogonal frequency components is obtained then the spectrum will be *complete* in the sense that the original target signal may be regenerated exactly from its spectrum.

$$f(t) = \sum_j g(\omega_j)e^{i\omega_j t} \tag{4.2}$$

In the case where the target function repeats itself exactly after a finite interval Δt such that

$$f(t + n\Delta t) = f(t) \tag{4.3}$$

for any positive or negative integral value of n, the signal is described as *stationary*; an example of this is the vibration of a simple single bond. To construct the complete spectrum of a stationary signal, the longest required frequency component has a wavelength equal to the repetition length, Δt.

In the case where the target, rather than being a continuous mathematical function, consists instead of a stationary, finite set of discretely sampled values, a finite range of frequencies is sufficient for a complete spectral representation. If the representation of the function consists of n points separated by δt, then the longest required wavelength is $n\delta t$. The shortest required wavelength is $2\delta t$—in other words, a wave that can be completely represented by two elements of the function. This corresponds to the *Nyquist* critical frequency, N_f, and is clearly the highest frequency that can be extracted from the target. The set of values required for a complete spectral representation of a dataset of n points therefore spans the range $-N_f < \omega < N_f$ in frequency steps corresponding to $n\delta t$. The *discrete Fourier transform* of a target dataset $f(t)$ is the set of spectral points required for a complete reconstruction of the dataset. Practically, the discrete Fourier transform may be rapidly calculated on a computer using the *fast Fourier transform* (FFT) algorithm developed by Cooley and Tukey.[63]

Strictly speaking, stationarity is a requirement for the Fourier transform to work as expected when applied to an arbitrary dataset, although this is often neglected. To understand why, it is necessary to examine its behaviour where discontinuities exist in the target. Consider a function consisting of an abrupt pulse such that $f(t) = 1$ where $0.5 < t < 0.6$ and 0 elsewhere. The Fourier transform of this signal requires a large number of *harmonic* components to construct the shape of the pulse. These sinusoidal harmonics interfere with each other constructively and destructively in such a way as to cancel out everywhere other than the region of the pulse. This behaviour can be seen in Figure 4.1. Although these harmonic components are mathematically required for a complete reconstruction of the target function, it is not always clear what meaning to attach to them when interpreting physical data. In the situation where the target function is nonstationary, the infinite replication implicit in the Fourier transform creates a procession of step-like artefacts separated by Δt, the length of the signal. This is illustrated in Figure 4.2. This obviously complicates the interpretation of such spectra, tending to obscure the frequency components that do have clear meanings in physical situations. The name for this problem is the *Gibbs phenomenon*.[64] It is unfortunate that condensed phase molecular dynamics simulations inevitably produce nonstationary data.

A common tactic for somewhat alleviating the Gibbs phenomenon at the ends of a signal is to multiply the signal by a function which tails its amplitude smoothly down to a small value at each end before taking the Fourier transform. Many functions such as triangular waves, Gaussian curves and raised sections of cosine waves have been proposed for doing this, although most such simple windows produce unwanted artefacts in the resulting spectra and are far from optimal; see Harris[64] for a comprehensive review. A common choice with a simple functional form optimised for frequency discrimination is the ‘minimum’ 3-term

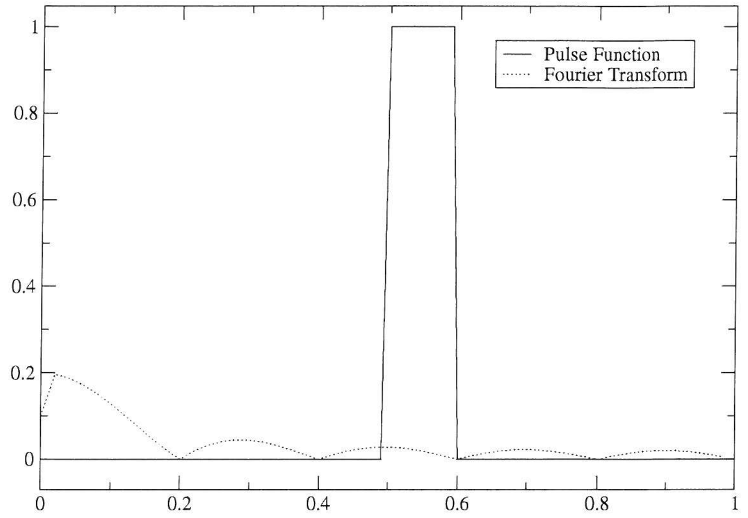

Figure 4.1: Pulse function and the positive frequency components of its Fourier transform. The magnitude of the Fourier transform is shown. The ripples in the transform are the harmonics required to recreate the sharp-edged pulse; the presence of such artefacts clouds interpretation of signals with pulse- or step-like properties.

Blackman (or 'Blackman-Harris') function[64]

$$w[x] = 0.42323 + 0.49775\cos\left(2\pi\frac{x-(l/2)}{l}\right) + 0.07922\cos\left(\pi\frac{x-(l/2)}{l}\right) \tag{4.4}$$

where l is the length of the window and x the coordinate relative to its origin.

4.2 The Spectrogram

Despite the difficulties described above, the Fourier transform has proved very useful for the interpretation of physical phenomena. The spectra obtained have a further drawback though: they are completely time independent. A nonzero spectral component of frequency ω indicates that such a wave is required for the reconstruction of the totality of the original signal. No information about the period through which a frequency component persisted is available in the spectrum.

A common means of getting around this limitation is the *spectrogram* or *windowed Fourier transform*. The idea behind this method is simply to slice the target function into a series of sections, and then obtain the Fourier transform of each section separately. This is effectively

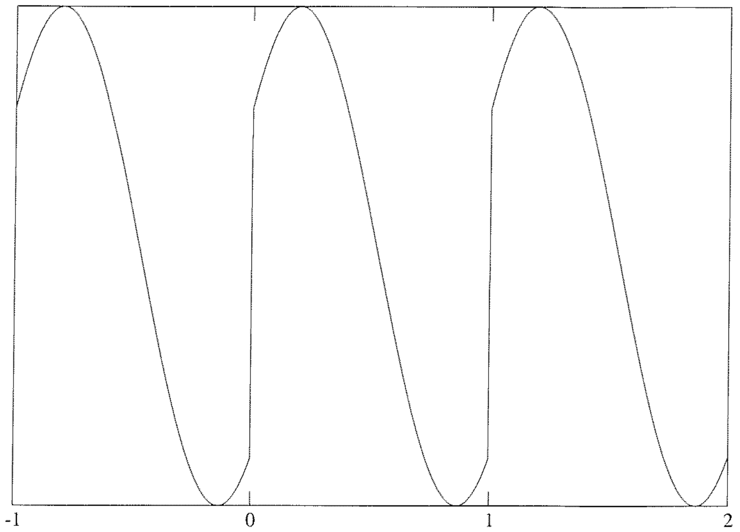

Figure 4.2: A short, nonstationary sinusoidal function on the interval 0–1, shown as the Fourier transform will process it. The discontinuity at each end of the signal will require a large number of harmonics to represent it, in a similar manner to the edges of the step function in Figure 4.1.

like sliding a window along the signal, repeatedly taking the Fourier transform of the visible data at constant intervals. This gives a measure of how the spectrum of the signal has varied with respect to time—a time-frequency distribution.

The question that must be answered before computing a spectrogram is how large the spectral windows must be. The longest wavelength that may be obtained is equal to the length of the window, and the frequency resolution of the technique is inversely proportional to the length of the window used, i.e. a longer window gives a higher frequency resolution and a lower minimum frequency bound. Conversely, the shorter the window, the higher the time resolution of the spectrogram obtained, as fewer waves of a given wavelength are packed into each.

An unfortunate drawback of the spectrogram method is its reliance on the calculation of a sequence of Fourier transforms, each of which should ideally be stationary with respect to the length of window employed. For any real-world dataset this is clearly very unlikely to be the case and this may make the interpretation of the spectrogram awkward. Windowing functions are generally used to alleviate this problem, as described in Section 4.1.

An example of a spectrogram is shown in Figure 4.3.

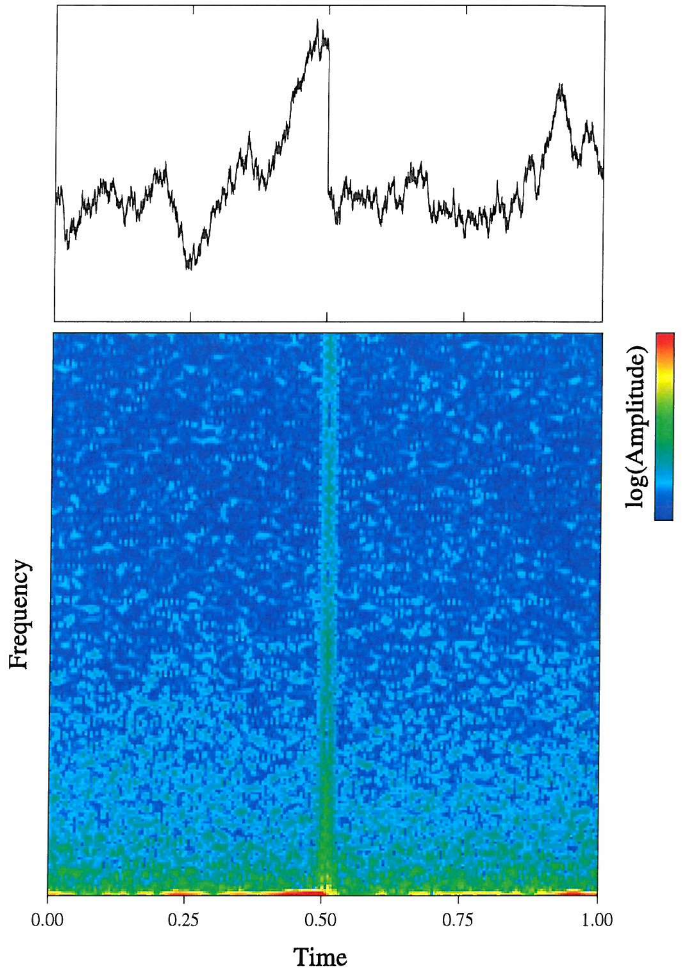

Figure 4.3: Top: a Brownian wave with a disjunction at time 0.5. Bottom: the spectrogram of the wave; the window length was here set at 0.05 time units, and 200 overlapping windows were calculated in all. The Blackman function in Equation 4.4 was used to attenuate the edges of each window.

4.3 Wavelet Analysis

If the nature of the resolution trade off in the spectrogram method is considered carefully, it becomes apparent that the method has a further drawback. For a given window length, Δt, the product integral of any wave shorter than this maximum will contain the information from more than one complete wave period. This is inefficient: a better trade off would be for the window length to be adjusted to the length of a single wave for each wavelength required. Unfortunately, there is no direct way of doing this with the conventional Fourier transform.

A solution to this problem is the *wavelet transform*.[65] This is a generalised spectrogram method where a chosen 'mother' wavelet function—not necessarily sinusoidal—is fitted to the dataset at multiple translations after being 'dilated' to a given size. To obtain the whole spectrum, a series of dilation parameters are used in succession. For a particular spectral point, the dilation parameter governs the frequency being investigated and the translation parameter the time region in the signal.

The wavelet transform has been used extensively in the fields of image compression, sound processing, fault detection and radar and sonar imaging. So far, though, it has seen little use within the field of chemistry. A review by Leung *et al.*[66] observes that common chemical applications have included smoothing, denoising and compressing experimental data, particularly from spectrometers; other physical- and quantum-chemical uses they note include modelling the Duffing and Morse oscillators, visualising atomic orbitals and providing alternative basis sets for electronic structure calculations. Li *et al.*[67] used the wavelet transform to characterise a condensed-phase simulation of argon. They decomposed the cartesian coordinates of a simulation of argon atoms into a set of 'scale' components. These were then used to calculate the 'scale spectrum' of the simulation, giving a plot of how the temperature of the system varies with respect to 'scale' of motion. Such spectra are computed for a number of temperatures, and this shows a decrease in the timescale over which particle velocities change as thermal energy increases, as would be expected. Askar *et al.*[68] have also investigated the use of the wavelet transform in analysing molecular dynamics simulations. They derive a parameter to characterise the average amplitude of the highest frequency 'scale' of motion in their system from the coefficients of the wavelet transform, and attempt to use this as a diagnostic tool to determine which degrees of freedom have made conformational transitions. Also, they compare the low-pass filtering of data using wavelets and Fourier transforms from the perspective of data compression, concluding that wavelets are an improvement.

4.3.1 Theory

The wavelet transform of a target function $f(t)$ is defined as

$$w(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) g^*((t-b)/a)\, dt \tag{4.5}$$

where a is the dilation parameter, b is the translation parameter and the function $g(t)$ is the original, 'mother' wavelet function; * represents the complex conjugate. In words, the mother wavelet function is dilated with respect to the signal to obtain the required wavelength, and then translated to the required position relative to the origin of the signal. The product integral of the two is then calculated. A wavelet spectrum is obtained by repeating this procedure for many wavelengths and positions.[65]

For a wavelet consisting of a single trigonometric wave, the dilation of the scaled function will directly correspond with the obtained wavelength. This type of wave can be considered as an optimised spectrogram transform with no attenuating function. The time-frequency resolution trade off will be optimised for any particular frequency, as only a single wave will be matched to the data at any given translation. This does not, however, imply that this specific wavelet function is in any way optimal. Such a wavelet would suffer from similar nonstationarity problems to those that affect a spectrogram with no attenuating function. The solution to this is to construct a wavelet function with an attenuation built into it. Examples of such functions will be given in the next section.

Certain classes of wavelet function, described as 'orthogonal', have been constructed with the objective of creating a reversible transform. In order for these to be used, the translation and dilation parameters must be carefully selected, generally as a binary series. In the first transform the wavelet is thus scaled to the full length of the dataset, and only a single translation (centrally located) is required. In the second the mother wavelet is scaled by $1/2$ and integrated at two translations that between them exactly cover the dataset without overlapping; in the third the wavelet is scaled to $1/4$ the size of the mother wavelet and integrated at four translations. A transform of this nature with an orthogonal wavelet basis function is referred to as a *discrete wavelet transform* (DWT); it can be considered mathematically as a linear operation—i.e. a matrix operator—and fast algorithms have been designed for computing it. The alternative to the DWT is the *continuous wavelet transform* (CWT), which will be employed throughout the current work. The CWT obtains much more information from the signal and is consequently much easier to interpret. The dilation parameter is varied smoothly over a continuous range, and for each value the transform is calculated for a large and frequency-independent set of translations through the data.

4.3.2 Wavelet Functions

A function commonly used for work where direct comparison with spectrogram and Fourier transform methods is required is the *Morlet* wavelet.[65] This is the product of a complex trigonometric wave and an attenuating Gaussian envelope

$$g(t) = ke^{i\omega t}e^{-\frac{t^2}{2}} \tag{4.6}$$

where ω and k are constants controlling frequency definition and normalisation, respectively. The larger the value of ω, the higher the frequency resolution of the wavelet, and the lower the resulting time resolution. A value of 6 has been used throughout this work to give a reasonable trade off. This function contains wavelike elements very similar to conventional trigonometric waves in shape (though not identical, due to the attenuating function), and this aids comparison with the Fourier transform. Another species of wavelet sometimes encountered is the *Marr* wavelet,[65] also known as the *Mexican hat*. This function is very sharply localised, giving it a higher time resolution than the Morlet at the expense of a lower frequency resolution.

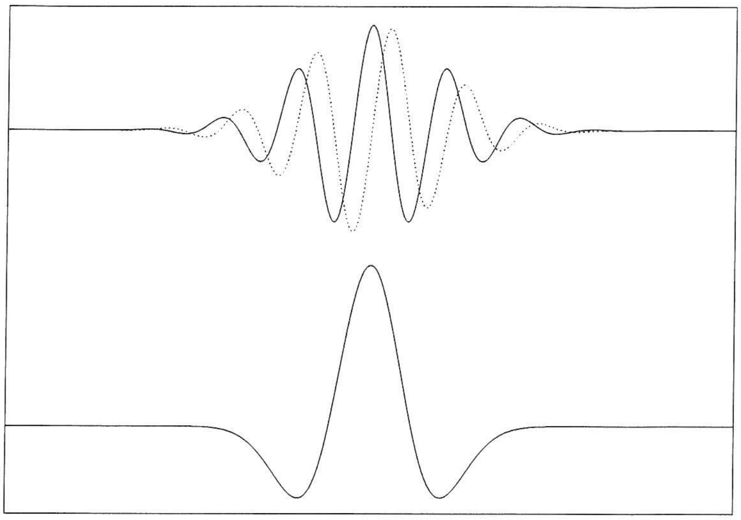

Figure 4.4: Two common types of wavelet functions. Top: the *Morlet* wavelet with $\omega = 6$. The imaginary part of the function is plotted with a dashed line. Bottom: the *Marr* wavelet, also known as the *Mexican hat.*

It is the second derivative of a Gaussian

$$g(t) = k(1-t^2)e^{-\frac{t^2}{2}} \tag{4.7}$$

where k is a normalisation constant. The resulting wavelet clearly has less in common with trigonometric functions, and so is more difficult to compare with the results of the Fourier transform than the Morlet. Both the Morlet and Marr wavelets are shown in Figure 4.4. Many other wavelet functions have been described, often with a particular application in mind, e.g. sound signal noise suppression or video compression; these are beyond the scope of the current work.

4.3.3 Practical Details

To calculate the continuous wavelet transform of a function in practise it is necessary for efficiency reasons to use Fourier convolutions. The target function is Fourier transformed and then multiplied by the transform of the correctly dilated wavelet function. The resulting product is reverse transformed to obtain the integrated product given by equation 4.5 for all values of the translation parameter, b. This is repeated for each required value of the dilation parameter, a, to yield the complete spectrum. Two example wavelet spectra are shown in Figure 4.5. Such time-frequency distributions may also be integrated in either the time or frequency domain to obtain *marginal spectra*. An example of such marginal spectra is shown in Figure 4.6.

A source of difficulty when interpreting the spectra of physical phenomena occurs where the target signal exhibits *intrawave modulation*. Consider a sinusoidal wave whose frequency is varied in a regular manner over the course of each period. For this wave to be represented by a linear combination of simple near-sinusoidal basis functions will require a large number of harmonic components if the basis does not closely match its modulated shape. This phenomenon is illustrated in Figure 4.7, where the function

$$f(x) = \sin(x + 0.6\cos(x)) \tag{4.8}$$

is plotted along with its Morlet wavelet spectrum. The problem is obviously very common: without foreknowledge of the structure of the target wave, no basis function can be constructed to closely model it, and the spectrum will generally contain complicated harmonic components. Such harmonics are generally difficult or impossible to interpret. This problem is common to all Fourier-based methods of analysis, as they rely on calculating the degree of similarity of a series of marginally different basis set functions to the target; the leakage results from the many partial matches to the basis elements.

4.4 The Hilbert Transform

There is an alternative approach to time-frequency analysis that does not rely on matching the shape of some basis function to the target signal. It relies on the properties of an *analytic signal* that may be defined in terms of the *Hilbert transform* of the target function. From this analytic signal may be derived a measure of *instantaneous frequency*.[69,70] The difficulty with the method comes in understanding and creating the conditions under which a meaningful 'instantaneous frequency' may plausibly be defined.

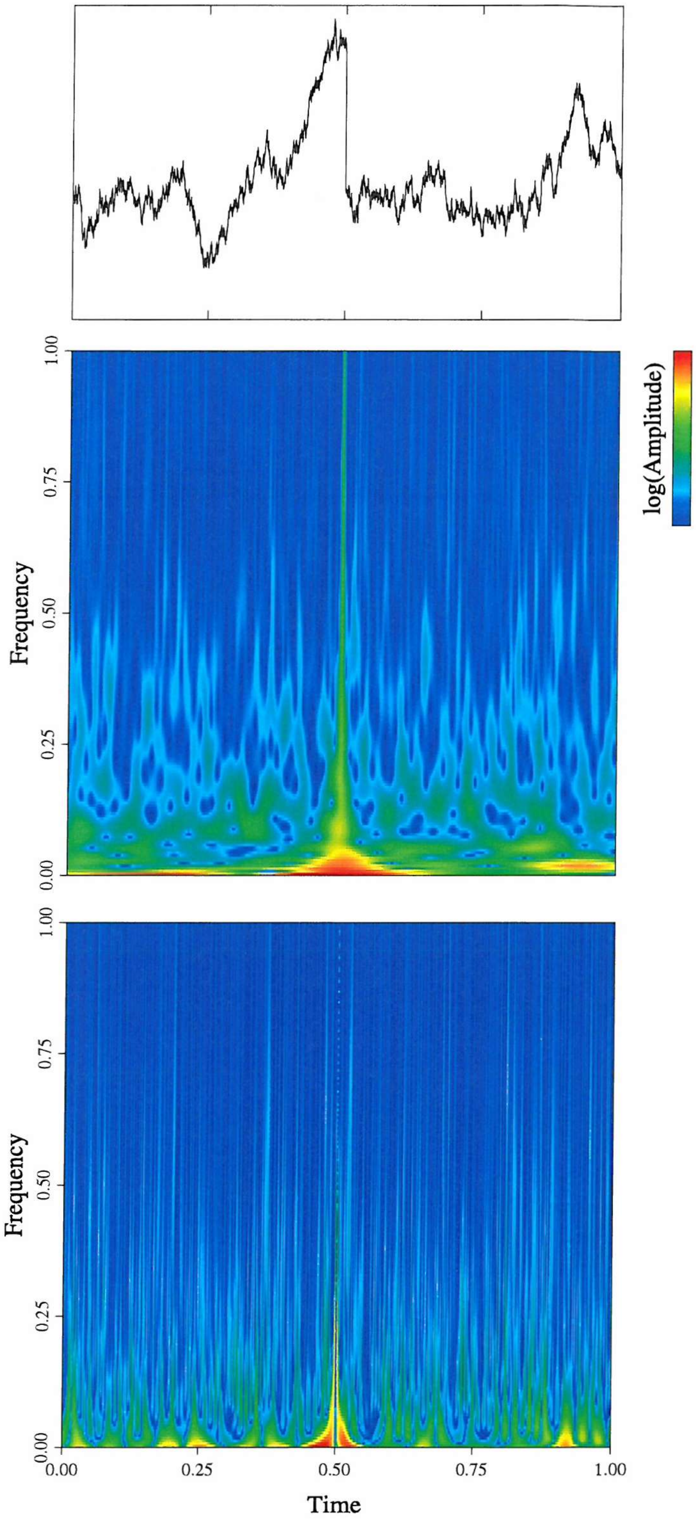

Figure 4.5: Examples of wavelet transforms. Top: a sample of Brownian noise with a disjunction at its midpoint is shown. Middle and bottom are its Morlet and Marr wavelet transforms, respectively. Notice how the peaks in the Marr wavelet are narrower along the time axis and broader along the frequency axis. The target signal used is the same as that in Figure 4.3.

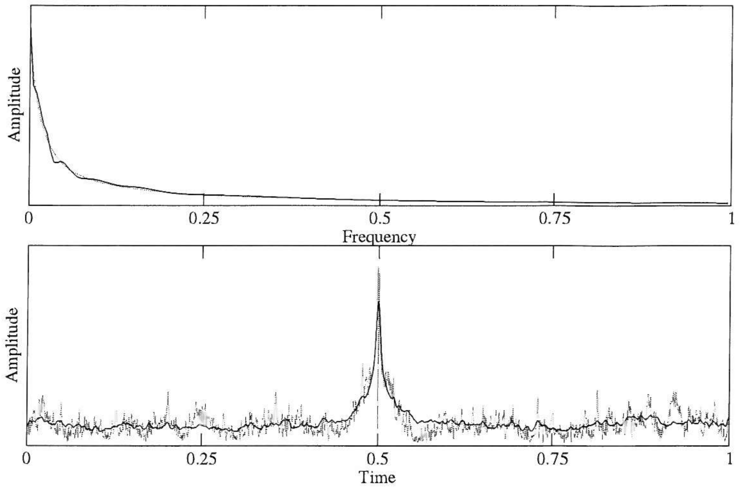

Figure 4.6: Examples of wavelet marginal spectra for the signal used in Figures 4.3 and 4.5. The upper graph is the frequency marginal spectrum; the lower is the time marginal spectrum. The disjunction is indicated by the peak. The data obtained using the Morlet and Marr wavelets are drawn with solid and dotted lines, respectively.

The concept of 'instantaneous frequency' has been controversial ever since it was first proposed in the 1930s during early work on FM radio.[71] The idea of a wave motion having properties of frequency or amplitude localised to a single instant contradicts basic intuition: wave motion is an intrinsically nonlocal phenomenon. One way of thinking about it is as follows: in the special case where the wave motion being investigated is essentially a sinusoidal wave with its frequency and amplitude being modulated 'slowly'—that is any wave motions in the modulating influences have much lower frequency than that of the oscillation being investigated—then a single varying frequency and amplitude can make sense as properties of the wave. To understand the problem of defining these properties exactly consider the sum of two sine waves, one of large amplitude and low frequency and a second 'riding wave' of small amplitude and high frequency. It is obvious that no meaningful *single* 'instantaneous frequency' can be defined for any point: two separate oscillations are present throughout each with its own amplitude and frequency. For the purposes of the next section, only 'single component' waves as described above will be considered.

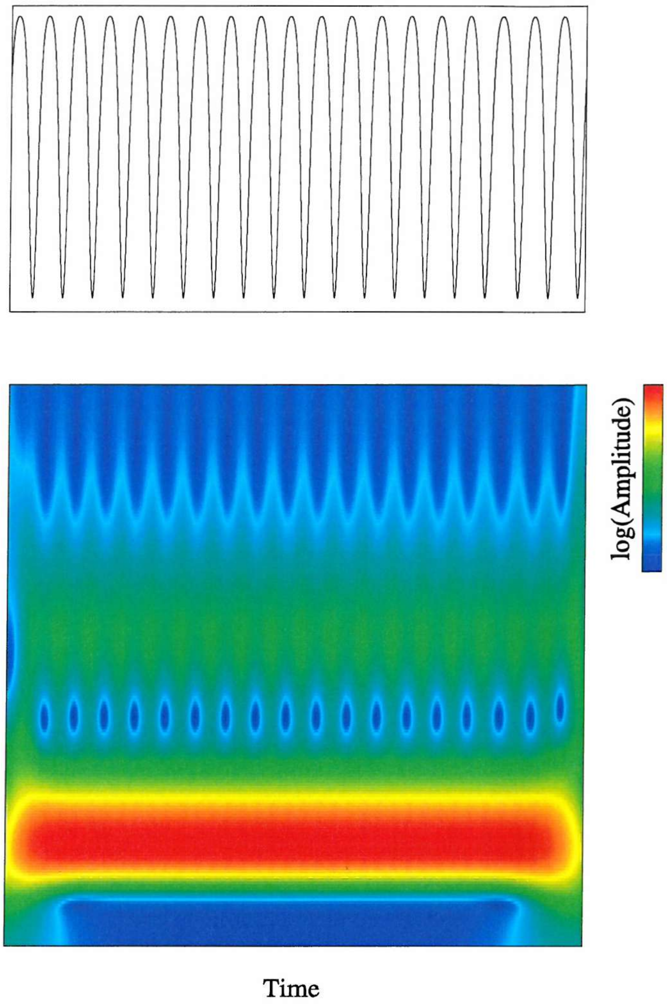

Figure 4.7: In the upper part of this figure can be seen a waveform with intrawave modulation. In the lower part is the Morlet wavelet transform of this signal, illustrating the difficulty of interpreting such phenomena. The red band corresponds to the unmodulated wave, $\sin(x)$, and the blue-green ripple artefacts at around double this frequency are produced by the modulating factor $0.6\cos(x)$.

4.4.1 Theory

The Hilbert transform $h(t)$ of a real-valued function $f(t)$ extending over $-\infty < t < \infty$ is obtained by convoluting it with $1/\pi t$

$$h(t) = \text{pr.v.} \int_{-\infty}^{\infty} \frac{f(u)}{\pi(t-u)} du \tag{4.9}$$

Where pr.v. indicates the Cauchy principal value of the integral (explained in Section A). Unlike the Fourier transform, this results in another real-valued time domain function. The process is illustrated in Figure 4.8. As shown, the Hilbert transform of a sine wave is the negative of a cosine wave of the same amplitude and frequency. More generally, it may be shown[72] that the Hilbert transform of a signal has the same magnitude, but a phase shifted by $\pi/2$. The calculation is highly localised by the rapid attenuation of the convoluting function: a point in the calculated transform is strongly affected only by a relatively narrow band of the target function. The Hilbert transform is thus a $\pi/2$ phase shift calculated from time localised data.

The analytic signal $z(t)$ is a complex function which may be defined as follows

$$z(t) = f(t) + ih(t) \tag{4.10}$$

where $f(t)$ is the target function and $h(t)$ is its Hilbert transform. This can also be written using the notation

$$z(t) = A(t)e^{i\phi(t)} \tag{4.11}$$

where A is the amplitude of the signal and ϕ is the phase angle. These are given by

$$\begin{aligned} A(t) &= \sqrt{f(t)^2 + H(t)^2} \\ \phi(t) &= \arctan\left(\frac{H(t)}{f(t)}\right) \end{aligned} \tag{4.12}$$

If the target signal is considered to be the real part of a complex sinusoidal motion, then the problem of recovering the phase angle is changed to that of recovering the imaginary part

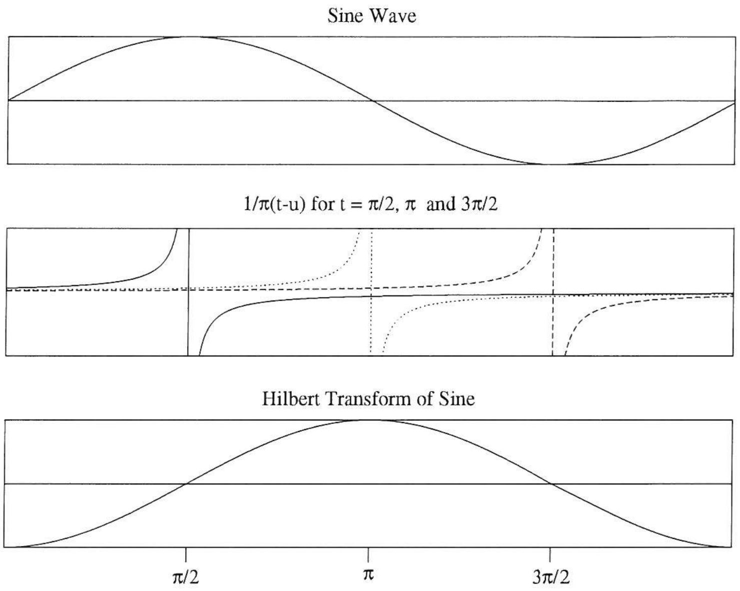

Figure 4.8: Illustration of the Hilbert transform of a sine wave. The function $1/\pi(t-u)$ is shown for $t = \pi/2$, π and $3\pi/2$ in the middle graph (drawn in solid, dotted and dashed lines, respectively). It can be seen that the signs of the functions are similar immediately before $\pi/2$ and opposed immediately after it. Owing to the symmetry of both functions about this point, the magnitudes of the integrals before and after it are identical, and so the Hilbert transform of the point is 0. The other extreme and crossing points may be evaluated by eye in a similar manner: zero-crossing points in the target correspond to extrema in the transform, and extrema in the target correspond to zero-crossing points in the transform. Between these points the transform is continuous and smooth. The Hilbert transform of $\sin(x)$ is $-\cos(x)$, as shown.

of this sinusoid. The Hilbert transform is a means of doing this and the analytic signal defined above provides a way of recovering the phase angle of the target as a function of time. Under the restriction that the target is comprised of a single wavelike component, the rate of change of this phase is a measure of frequency: the more rapidly it varies, the higher the frequency. The relationship is linear and so we may define the instantaneous frequency, $\nu(t)$ of the target as

$$\nu(t) = \left(\frac{1}{2\pi}\right)\frac{d}{dt}\phi(t) \tag{4.13}$$

Putting this together, given a real time-domain signal, $f(t)$, we can calculate the signal's Hilbert transform, $h(t)$. If we then combine the two to obtain an analytic signal, $z(t)$, we may calculate the instantaneous amplitude, $A(t)$, and phase angle, $\phi(t)$. Differentiating the latter yields the instantaneous frequency, $\nu(t)$. The instantaneous frequency of a sine wave is constant, but the method can deal with more complicated waves whose frequency and amplitude vary as a function of time. An example of the application of this process to a function exhibiting intrawave modulation is shown in Figure 4.9.

4.4.2 Applicability

In Section 4.4 the assumption was made that the signals to be examined were 'single component'. The question of what kind of waves this analytic signal approach can be applied to has been a source of contention, and a number of authors have offered different definitions.

According to Vakman *et al.*,[70] the definitions of instantaneous properties given in Subsection 4.4.1 yield data in agreement with the intuitive meanings of amplitude, phase and frequency for "narrow-band" oscillations, and for broad-band oscillations that arise under "slow" frequency modulation. If the signal does not meet these criteria directly, then, they suggest, a filtering or smoothing method must be applied. They note that under such conditions, the analytic signal derived can make sense only within the framework of the methods used to modify the target signal, which clearly creates difficulties in interpretation. Although examples are presented, no mathematical definition of "slowness" or "narrow-bandedness" is given.

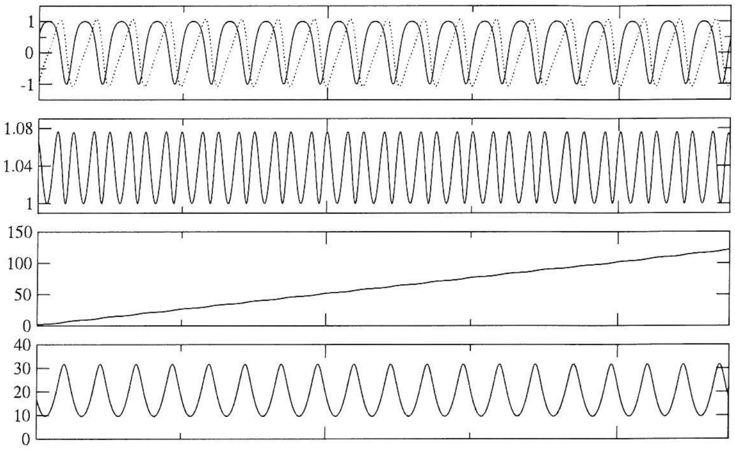

Figure 4.9: Calculation of the analytic signal and its 'instantaneous' properties. The top graph shows the target signal along with its Hilbert transform as a dashed line (the real and imaginary parts of the analytic signal, respectively). It can be seen that the target contains 19 complete wave motions. The second graph shows the instantaneous amplitude of the analytic signal (see Equation 4.10, which remains close to unity in this case. The third graph contains the instantaneous phase calculated from the signal, unrolled from the $-\pi < \phi < \pi$ range to create a continuous function. The bottom graph shows the instantaneous frequency obtained by differentiating the phase graph shown. In the units used here the frequency reflects the number of complete wave motions in the original dataset, staying close to around 19 on average, but rising coincidentally with the sharper (higher frequency) regions of the target signal. The target signal is the same as that used in the wavelet intrawave modulation example, Figure 4.7. The greater ability of the Hilbert transform to make sense of this phenomenon is clear.

Boashash[73] believes that the concept of instantaneous frequency has meaning only for signals that are "monocomponent"—where there is only a single frequency or a "narrow band" of frequencies varying as a function of time—and has no meaning for multicomponent signals. A condition for the meaningfulness of the analytic signal is presented in terms of the spectra of the amplitude modulation and the frequency modulation: the Fourier spectra $F[A(t)]$ and $F[\cos\phi(t)]$ must not overlap. If they do, it is impossible to separate the amplitude modulation from the frequency modulation and so neither is defined. The analytic signal as defined from the Hilbert transform will then be uninterpretable.

Further problems occur if the peaks of the target function are not arranged roughly symmetrically around the axis. In the extreme case where a minimum occurs *above* the axis, or a maximum below, the Hilbert transform will obtain a singularity. As the target becomes closer to symmetrical, the analytic function tends towards a circle of constant amplitude in the complex plane, and the amount of frequency modulation *caused by the asymmetry* is reduced to a minimum. For the practical analysis of data then, some sort of symmetry criterion is needed to exclude the cases where peaks occur on the wrong side of the mean, and to reduce the frequency and amplitude modulation to their minimum values.

The instantaneous frequency can be thought of as the frequency of the sinusoidal wave that can best be fitted to the *local* region of the signal. If a signal comprising a simple sinusoid varying slowly in frequency over time were to be filtered to remove everything but an infinitely narrow frequency window centred on the wave, the result from the Hilbert transform would be identical to that from the Fourier transform: a single frequency. If the window were to be broadened, a set of sinusoidal components would be obtained by the Fourier transform, each persisting throughout the whole target signal. In other words, multiple sinusoidal waves would exist at every point. The Hilbert transform, on the other hand, could be used to obtain a single-valued *function* of frequency for each point in the signal. The Hilbert picture of amplitude and frequency thus make sense locally, whereas the Fourier picture makes sense globally.

4.4.3 Practical Details

In practice, the Hilbert transform may be computed by taking the Fourier transform of the data, setting all negative frequency components to zero, doubling all positive frequency components and back-transforming;[72] this approach has been used throughout the current work. While other methods of computing the transform exist,[74] they are significantly less efficient than the Fourier approach.

Although the Hilbert transform would seem to be a useful method, until recently the restrictions on its applicability described above have tended to limit its use on real-world data. It has been used in geophysics to detect phase shifts in probe signals, for example to detect discontinuities in geological formations; it has also been used in signal processing to determine the time delays in energy propagation from correlation functions. More generally, however, the requirement for either a well behaved target signal or a case by case judgement about filtering has tended to act as a deterrent to its application.

4.5 Empirical Mode Decomposition

Empirical mode decomposition (EMD) is a new method of signal analysis developed by Huang *et al.*[62] The objective of EMD is to decompose an arbitrary signal into a small set of components to which the Hilbert transform procedure may be applied to yield meaningful instantaneous frequencies. The method obtains these components, called *intrinsic mode functions* (IMFs), by an iterative procedure. Each IMF component may then be Hilbert transformed individually, and its instantaneous frequencies and amplitudes obtained in the way explained.

The core of the method is the definition of the intrinsic mode function in such a way that its analytic signal will have meaning. An IMF is defined as a function where:

1. The number of extrema and the number of zero-crossings differ by at most one.
2. At any point, the mean of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

This definition satisfies the conditions required for the analytic signal to have a single frequency and a single amplitude value at each time point. Point 1 of the definition guarantees that no 'riding waves' are present in the function: it ensures that there can be only one maximum between a pair of zero-crossings. This removal of riding waves forces the instantaneous properties to be single valued. Point 2 is there to ensure the analytic signal remains centred at the origin of the complex plane to avoid problems with singular or excessively modulated frequencies.

The algorithm suggested by Huang *et al.*[62] for decomposing a dataset into a set of IMF components meeting the definition above is described below, with the modifications to the stopping criteria given in a later paper.[75] The algorithm has an outer *sifting* procedure and an inner *seeking* procedure. The outer part is as follows:

1. Determine the highest frequency IMF component present in the dataset using the *seeking* procedure.
2. Subtract this component from the dataset.
3. A pair of stopping criteria are evaluated. If neither is true then the process repeats at step 1; otherwise the algorithm stops. The criteria are: (i) is there exactly one maximum and one minimum in the new dataset, and (ii) is the integral of the absolute magnitude of the dataset less than a predetermined cutoff level.

The first stopping criterion determines whether the dataset has had all wavelike motion subtracted from it and so been reduced to a *trend*. The second criterion will stop the algorithm if the remainder of the signal has become negligible in magnitude (as defined by the analyst). The subtractive nature of the *sifting* process guarantees that the IMF components and the trend sum exactly to reproduce the original dataset. The second stopping criterion rarely occurs in practice.

An IMF component is determined by a *seeking* algorithm applied to a copy of the dataset as it currently stands in the outer *sifting* process:

1. A spline curve is fitted to the maxima of the dataset. A second spline curve is fitted to the minima.
2. The mean curve between the two splines is evaluated, and this is then subtracted from the dataset. This increases the local symmetry of the signal: the mean envelope of the local minima and maxima is shifted towards zero.
3. A pair of stopping criteria are evaluated. If neither is true then the process repeats at step 1; otherwise this seeking cycle stops. The criteria are: (i) does the number of extrema and zero-crossings differ by 1 or 0, and (ii) is the integral of the absolute magnitude of the dataset less than a preset cutoff level.

The first stopping criterion corresponds directly to point 1 of the definition of the IMF given above. The second criterion will stop the algorithm if the current IMF being evaluated has become negligible in magnitude. In general usage, the second criterion almost never occurs, and often indicates a problem with the seeking process when it does. It is important to note that the algorithm described above is an update to the original: the stopping criteria were changed[75] to improve the preservation of the amplitude modulation in the signal. The resulting procedure does not strictly guarantee point 2 of the IMF definition, although this point is not mentioned by Huang.[75] In practise, though, the IMFs obtained seem sufficiently symmetrical to reduce unnecessary frequency or amplitude modulation to a very low level *most* of the time. Numerical problems can occur under certain circumstances, though, and these will be returned to in Section 5.2.

Although the seeking algorithm might optimally recover a function satisfying the definition of an IMF after a single cycle, in practice this is very rare. A bump on a slope often becomes a new extremum after being translated by the mean curve, and will then affect subsequent iterations of the algorithm. Consequently, many cycles must usually take place as minima and maxima that were previously hidden by longer-scale curves in the data are recovered by the sifting algorithm. This process has been described[62] as having the effect of recovering the original "intrinsic scales" of the data.

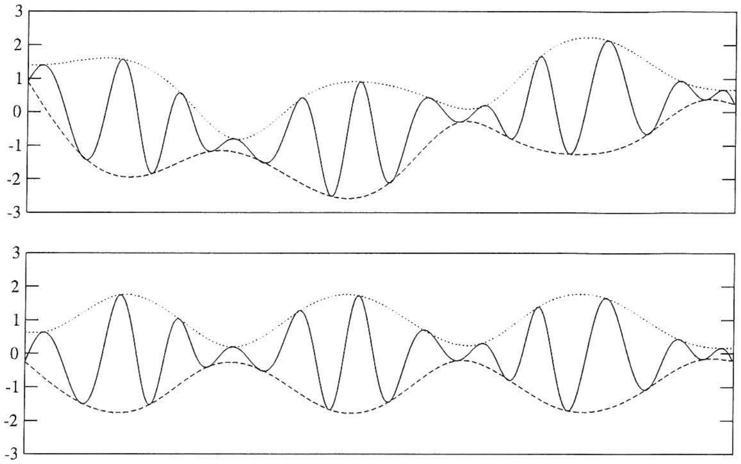

Figure 4.10: Example of the Empirical Mode Decomposition (EMD) method. The upper graph shows the original signal, with the maxima and minima fitted by spline curves. The lower graph shows the result of subtracting the mean curve between the two splines from the data, along with the new maxima and minima curves.

Although there is no formal reason why the IMF components obtained should be orthogonal, in practice they are numerically very close to being so.

The IMF procedure is illustrated in Figure 4.10 for a simple dataset. The complete set of IMFs extracted from this dataset are shown, along with the trend component, in Figure 4.11.

The combination of the empirical mode decomposition and the Hilbert transform has been called the *Hilbert Huang Transform* (HHT) by Pan *et al.*[76]

4.6 Spectra Obtainable

Once the target signal has been decomposed into a set of IMFs and the instantaneous properties of the IMFs determined as functions of time, the resulting data must be graphed in a meaningful way. There are three conventional approaches to this: frequency-time, frequency margin and time margin. All three types of plot are histograms. The marginal plots are the result of integrating the frequency-time distribution along one or other of its axes. Examples of all three graphs can be seen in Figure 4.12. Huang *et al.*[62] suggest that the frequency marginal spectrum be calculated by

$$h(\omega) = \int_0^T H(\omega,t)dt \tag{4.14}$$

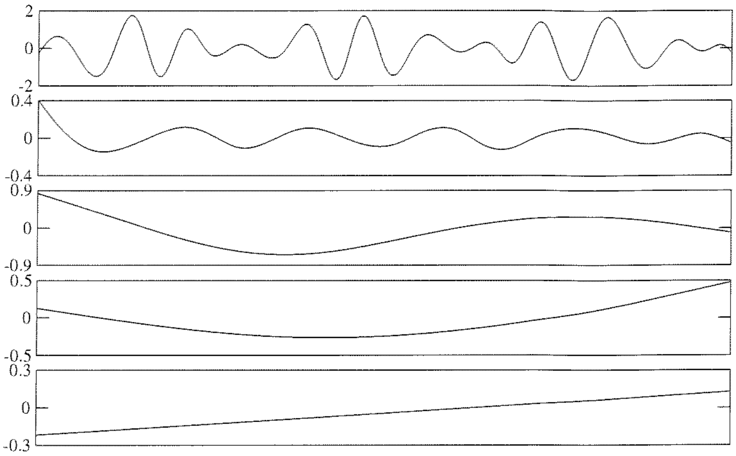

Figure 4.11: The Intrinsic Mode Functions (IMFs) obtained from the data shown in Figure 4.10. These components sum exactly to give the original target signal.

where $H(\omega,t)$ is the amplitude or signal energy at time t, and the integration is in practice performed as a sum over the HHT histogram. It should be noted that the value calculated as described above would *appear* to have units of amplitude *multiplied by time*, although the authors state that, "The marginal spectrum offers a measure of total amplitude (or energy) contribution from each frequency value. It represents the cumulated amplitude over the entire data span in a probabilistic sense". The meaning of this will be returned to in Subsection 5.7.2. Unless otherwise indicated, though, the definition of the marginal spectrum given by Huang *et al.* should be assumed, and where 'amplitude' or 'signal energy' (as defined below) are stated these should be understood in terms of the above integral.

An important issue to consider when producing such plots is which instantaneous property should be used. The most obvious is the signal amplitude, and this usually provides the most straightforward interpretation of the data. Another alternative suggested by Huang *et al.*[75] is an *energy density* defined to be the square of the amplitude, a definition commonly used in the field of signal processing. Plotting such a density would tend to emphasise more intense parts of signals. A third possibility is to plot a *signal energy* defined from the properties of the simple harmonic oscillator. This signal energy is defined as

$$E \propto A^2 \nu^2 \tag{4.15}$$

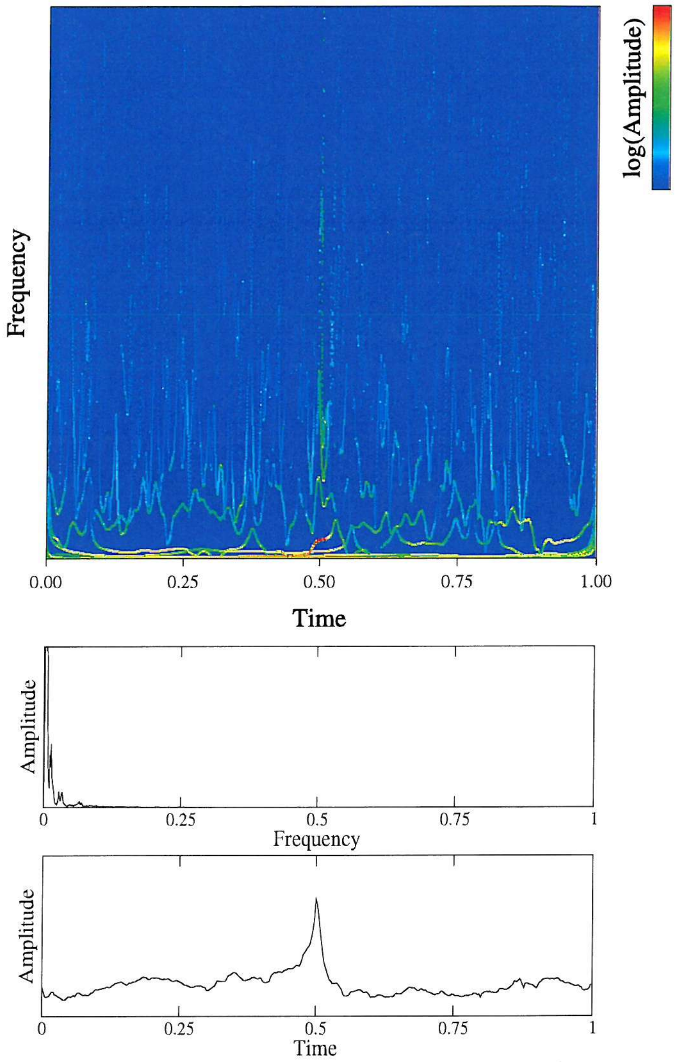

Figure 4.12: Examples of the spectra obtainable from the HHT. Top: frequency-time distribution. Middle: frequency margin spectrum. Bottom: time margin spectrum. The target signal is that used in Figure 4.3.

where A is amplitude and ν is frequency. A derivation of this property is given in Subsection B. Plotting a graph of signal energy will not only highlight larger amplitude features at the expense of smaller ones, but also emphasise the higher frequency parts of signals. This latter effect is helpful under certain circumstances, but because of the distortion of the spectrum involved, signal energy requires an explanation when used.

All of these properties may be calculated for both HHT and wavelet spectra. Signal amplitude has been used throughout the current work unless otherwise stated.

4.7 Previous Work

4.7.1 Development of the Technique

The development of the HHT method may be charted through a series of papers by Norden Huang and coworkers published since 1996.

In 1996, Huang *et al.* published a paper[77] in which they investigated the properties of mechanical wave fusion in water tanks and on the ocean surface. The authors describe the Hilbert transform and give a summary of its use in their field and the problems found with interpretation of the instantaneous frequencies calculated from it. They provide a definition for a "simple oscillatory function" to which the Hilbert transform may be applied, requiring it to have no riding waves—point 1 of the IMF definition given in Section 4.5—and a zero local mean—point 2 of the definition. They provide the skeleton outline of a "Characteristic Scale Decomposition Method" (CSDM), similar to the EMD algorithm, but with only a single iteration of the inner, sifting, procedure and hence no definition of the associated stopping criteria. They state that the outer, sifting, procedure should continue until the remainder of the signal is "less than a preassigned criterion for stopping the procedure", perhaps indicating the minimum magnitude integral (stopping criterion (ii) of the algorithm given in Section 4.5). The CSDM/Hilbert transform method is used to interpret a set of experimental data, with particular attention being paid to abrupt phase and frequency jumps in physical waves which the authors state cannot be analysed adequately with Fourier methods.

A paper referred to repeatedly in this work is listed in the bibliography as "Huang, N. E., Long, S. R., Lin, R. Q. and Shen, Z. (1995) Wave fusion as a mechanism for nonlinear evolution of water waves. *J. Fluid Mech.* (in press)". A thorough check reveals that this paper has never been published. This paper is referred to as, among other things, containing a detailed description of the CSDM method.

In 1998, Huang *et al.* published a second paper[62] describing the HHT method at length for the first time, along with applications to a series of test systems and experimental data.

The paper begins with a review of previous time-frequency analysis methods, followed by a discussion of the concept of instantaneous frequency. The authors define a class of functions they name 'intrinsic mode functions', as described in Section 4.5. An algorithm by which the IMF components may be derived is then presented under the title 'empirical mode decomposition'. This algorithm is different to that described in Section 4.5: the inner sifting process stops when the standard deviation between the signal in two successive sift cycles is less than some predefined value. The completeness and orthogonality of the expansion in terms of IMFs are then addressed, and this is followed by a section describing the use of the Hilbert transform to derive instantaneous frequency and amplitude functions from them. A short discussion of the calibration and validation of the method with simple test signals follows. A number of time series, both synthetic and experimental are then analysed, and the results examined in detail.

The principal argument of the paper is that wavelike motion that is not *intrinsically* composed from a set of stationary, perfectly sinusoidal components will produce misleading Fourier and wavelet spectra. Members of this class include nonstationary signals, for example seismic readings of an earthquake, and nonlinear signals, such as those obtained from physical systems governed by nonlinear equations. The spectra obtained suffer severe 'leakage' of signal energy into harmonic components which, although mathematically required for an exact, linear decomposition, are nonphysical and cloud interpretation. The authors argue that there is a better alternative to understanding nonlinear and nonstationary signals than harmonic components: the concept of instantaneous frequency and the intrawave modulation that may be described in terms of this. The combination of the Hilbert transform and the empirical mode decomposition method (the HHT) allow such an instantaneous frequency to be defined.

The paper's discussion section describes a number of the drawbacks and limitations of the method as its designers see them, along with a list of possible improvements and avenues for further work. The drawbacks highlighted are:

1. The cubic spline fitting algorithm tends to over- and under-shoot the envelope of the signal near the extreme points. This creates numerical instabilities with the algorithm and causes errors to build up if large numbers of seeking cycles are required. An example of this is given on page 943 of the paper. A signal comprising of two cosine waves of very similar frequencies and identical amplitudes, interfering to produce what appears as a single wave of constant frequency and modulated amplitude is analysed with EMD. As the signal is very nearly an IMF to begin with, only a single IMF component would ordinarily be extracted. To attempt to rectify this, a very large number of seeking cycles is imposed by lowering the maximum standard deviation used in the stopping criterion. By doing this a set of IMF components were recovered from the data, but unfortunately their amplitudes and frequencies were incorrect and unstable, a flaw attributed to the buildup of numerical errors from large numbers of spline fittings.

2. There are two types of end-effect problem in the HHT method. The first comes from the spline fitting in the EMD seeking procedure and the second comes from the Gibbs effect caused by using Fourier transforms to compute the analytic signal. To understand the spline fitting errors, consider the case where the last point of the dataset is a minimum (or on the downward edge of a wave). A gap will then exist between the last maximum and the edge of the dataset; to bridge this gap, the spline interpolation algorithm must be relied on to *extrapolate* points. Unfortunately, spline curves are well known to suffer from large swings near and beyond their endpoints (see for example Press *et al.*[78]). This creates errors at the edges of the dataset that have a tendency to propagate inwards; the EMD algorithm then becomes unstable, often running for too many cycles and losing all sense of the original data.

The author's solution to both of these problems is to synthesise two "characteristic waves" beyond each of the true endpoints of the signal, though the precise nature of these waves is not defined. In the seeking procedure these waves 'anchor' the ends of the spline curves at plausible values. In the Hilbert transform procedure, the extra waves are attenuated toward zero, reducing the Gibbs effect.

3. In the case where a weak signal is embedded within a stronger one, and the two are phase locked in such a way that the extrema of the weaker signal occur at the positions of maximum slope of the stronger, the EMD algorithm will not pick them up. The weaker signal will then appear as intrawave frequency modulations. The solution suggested by the authors in cases where this is suspected is to differentiate the target signal before processing it.

4. The highest frequency that may be defined for a given dataset by the Hilbert transform is commonly lower than that in the Fourier transform (the Nyquist limit). The maximum frequency obtainable from EMD is constrained by the number of points required for a stable numerical derivative to be obtained. The authors suggestion, based on their experience of using the method, is that five points is the minimum needed. A further suggestion given is to interpolate the data using spline functions.

5. The EMD method depends on the dataset containing meaningful scales: in the case where two signals of very similar scale are combined, it will not be able to separate them. In other words, the algorithm cannot separate signals when their frequencies are too close together. The authors consider that this may be the biggest drawback of the method.

   The exact limitations of the resolving capabilities of the EMD method are not explained, "Details of the restrictions will be studied later".

6. Although the EMD procedure can reduce an arbitrary signal to a set of IMF components, an individual component may not have any direct physical meaning. Individual interpretation is possible in cases where the data contains clearly separable scales, and

the IMF in question represents one such scale component. Otherwise, the HHT spectrum must be considered as a unified whole.

A further paper[75] by Huang and coworkers was published in 1999. It begins with a description of the Hilbert transform and the EMD algorithm and a section on the validation and calibration of the Hilbert spectrum with some of the same simple test systems as the 1998 paper. A series of more complicated nonlinear and nonstationary data are then examined. As in the 1998 paper, the Duffing and Rössler nonlinear systems are analysed and their Hilbert and wavelet spectra compared. A series of experimental results are then examined, including the laboratory wave data reported in both the 1996 and 1998 papers.

Two modifications to the EMD algorithm are described. The first is a modification to the original stopping criterion for the seeking process, and the second is to address a problem not noted in the 1998 paper's discussion section: intermittent data.

The original seeking procedure had a tendency to continue for too many cycles, removing all amplitude modulation from the signal. In data possessing a clear separation of scales and hence physical meaning for the IMF components, this would remove that physical meaning. To prevent this, the simplified stopping criterion based on counting the number of zero-crossings in the signal as described in Section 4.5 was proposed.

In the period between the writing of the 1998 and 1999 papers, a problem was discovered concerning data that possess oscillatory scales that appear only intermittently. The IMFs obtained by applying standard EMD to such signals exhibit mode mixing—that is, two very different scales of motion are contained within a single component. For example, during the parts of a signal where an intermittent high frequency oscillation exists, the shortest scale IMF will contain this waveform; but in the parts where it is not present, the shortest scale IMF will describe a lower frequency mode of motion instead. Such mode mixing in cases where a clear separation of scales exists will prevent physically meaningful components from being obtained. The solution the authors propose is to add a further criterion to the seeking process limiting the longest period wave that may be included in an IMF component in cases where a specific error is clearly occuring. The algorithmic details of how this should be implemented are not given in the paper; nor are any automatic means described for deciding whether the problem is occuring: the only solution seems to be judgement of each case individually.

4.7.2 Applications of the Technique

Since the introduction of the HHT method in 1998, a number of papers applying it to synthetic and experimental data have been written by authors working in various fields. Some remarks on a selection of these papers follow.

Wei Huang *et al.*[79] use the HHT method to study the variation in blood pressure within the pulmonary arteries of rats, known to be a nonstationary phenomenon. Data were recorded at a rate of 100 samples per second over a 24 hour period and both time-frequency and marginal spectra were computed. The authors conclude that the HHT method offered a more comprehensive description of blood pressure fluctuation than the conventional Fourier transform.

A paper by Lai[80] explores the transition from regular to chaotic motion in a nonlinear system. This has previously been considered in terms of a discrete change from a narrow- to a broad-band Fourier spectrum, i.e. the spectrum switches from having a finite number of peaks to having an infinite number on an arbitrarily small change of the controlling parameter. The finding of the paper is that the Hilbert Huang transform gives a *very* different picture of this transition. The number of IMFs obtained for the system studied is identical under both regular and chaotic regimes and the frequency histograms obtained from their Hilbert spectra remain clear and uncluttered, showing no discrete change as the system switched state. The conclusion is that characterising chaotic motion with analytic signals derived from EMD results in a simpler picture than with Fourier transforms.

Echeverría *et al.*[81] examine the variability in the heart rates of human test subjects in both controlled breathing and transient activity experiments, along with a number of chirp sequences and the results of a synthetic heart rate variability model. They find that the application of EMD to their experimental data consistently produces 4 components with similar amplitudes and with frequencies localised in currently recognised spectral bands of autonomic regulation, i.e. their data contained an intrinsic separation of scales that was reflected in the

IMF components obtained. The authors speculate that the scales identified may have a direct physiological meaning. They also note that the HHT appears to be good at tracking changes in the frequencies of signals in short time windows.

Komm *et al.*[82] apply the HHT method to helioseismic data obtained by the SOHO satellite in order to study rotation and oscillation patterns in the solar atmosphere. The authors use Monte Carlo simulations to model how the error information obtained from the satellite propagates through the EMD algorithm (though they comment that their approach is only likely to be practical for very short time series, such as their 49 point datasets). From a measured signal comprising 49 datapoints with known measurement errors, a large set of synthetic signals is constructed. The original data form the mean value of these signals, to which is added a random perturbation with a standard deviation equal to the known measurement error. Both the original signal and the synthetic waveforms are then processed with EMD, the original yielding 3 IMFs. Of the synthetic waves, only those which also give 3 IMFs are kept; these are then compared with the IMFs obtained from the measured signal (this does not bias the results because the variance in the random number distribution of the retained subset of synthetic signals is the same as in the complete set). The variation in each IMF component is then considered by the authors to be a measure of how the error values propagate through the algorithm into different IMFs. They conclude from the good statistical agreement shown between IMF components of the synthetic and original data that the EMD method is robust, although it is noted that the errors at the edges are significantly worse, affecting 5% or so at each end of the signals. A series of completely random datasets with a variance equal to the experimental error were also decomposed with EMD and their Hilbert spectra analysed in order to measure the size of the errors in the resulting spectra. The conclusions of the paper focus on the measurements taken and what was learned from them about solar physics.

Niethammer *et al.*[83] compare four methods for obtaining time-frequency representations of signals to characterise the dispersion properties of multimode Lamb waves. Lamb waves are guided ultrasonic waves that propagate in plates and are commonly used for detecting faults in engineered structures; their analysis is difficult and a highly specialised field. The authors examine the spectrogram and wavelet transforms, the Wigner-Ville distribution and the HHT as methods for doing this, concluding that each method has certain strengths and weaknesses with their data. The strength they see in the HHT method is that the IMF components are based on local properties of the signal itself (rather than some arbitrary basis function, for example). In the analysis of their experimental data, EMD is seen to be failing to separate the Lamb waves into meaningful components in the spectral regions where many modes coexist within a narrow frequency band, exhibiting severe mode mixing problems. Although not mentioned in the paper, it would seem that the remedy proposed by Huang[75] for this problem would be unlikely to work, as the Lamb wave being examined is simply too complicated within certain frequency ranges. There would appear to be no way of separating the many vibrational modes, which cross and recross in time and frequency within narrow frequency bands, into non-mixed IMF components. The scale separation problem identified by Huang[62] would seem to make this impossible. The authors also assert that since the EMD is an empirical method, it is impossible to analyse the HHT's time-frequency resolution in as rigorous a fashion as for the other methods they examine. They conclude that the spectrogram is the best for their needs, although it is clear that all the methods tried have substantial difficulties with these extremely complicated signals.

Chunming *et al.*[84] use the EMD method as a means of smoothing data obtained from synthetic aperture radar(SAR). Their objective is to reduce the speckle in their data without overly degrading its information content, but at the same time preserving its "visual-natural" appearance. Their interest in the EMD method is due to its ability to separate a signal on the basis of its intrinsic scales; the method they develop is based on rejecting the two shortest scale IMF components to remove short period oscillations from their data. A test image is processed using a number of standard filtering algorithms along with their proposed method, and the results compared. They show their method to have better mean and edge preserving properties than any of the others tried, although the reduction in standard deviation was the lowest. They conclude that their method is an improvement on the others.

Montesinos *et al.*[85] analyse a set of noisy, nonlinear and nonstationary neutron flux transients measured in a nuclear reactor. The authors mention that the analysis of such data with wavelet transforms is difficult due to the excessive spreading of signal energy across the expansion. They obtain IMF components from their signals with EMD, and one component is observed to contain the physical phenomenon of interest. This is then processed with an autoregressive algorithm to determine a decay parameter and an oscillation frequency for the transient. The values obtained are compared with both known-correct values and the results of applying the autoregressive model to the unprocessed original dataset according to the authors' standard practice. The numbers from the EMD-filtered signal are found to be closer to the correct ones, and the authors conclude that their EMD-based method is superior.

4.7.3 Other Work Done in Southampton

Other workers in Southampton have used the HHT algorithm and the codes described in Chapter 5. The paper published by Phillips *et al.*[59] describes the application of the HHT and wavelet methods to the analysis of chemical simulations, with particular emphasis on the reversible digitally filtered molecular dynamics (RDFMD) method of Phillips *et al.*[61] In this paper were reported some of the Brownian dynamics results given here, along with the application of the method to simulations of gas-phase pentane and alanine dipeptide, chloroform solvated alanine dipeptide, and also the pentapeptide YPGDV solvated in water. The principal aims of this paper were to draw attention to the HHT and RDFMD methods, and to provide supporting evidence for the low frequency amplification strategy taken in RDFMD.

Chapter 5

Hilbert Huang Transform Work

5.1 Implementation Details

The Hilbert Huang Transform method was implemented according to the description given by Huang *et al.*,[62] with the modifications to the seeking process's stopping criteria reported by Huang *et al.*[75]

An empirical mode decomposition code was created first, and the stability of the algorithm was verified. To prevent errors from building up at each end of the signal and propagating inwards, one extra extremum is created beyond the true endpoint whose coordinates are determined as follows. Consider a signal terminating partway through the rising edge of a wave (effectively a *false* maximum). The y coordinate of the synthesised point is that of the last occuring *true* maximum—the last maximal turning point in the signal—and the x coordinate is chosen such that the gap between the new point and the previous true maximum is equal to that between the last pair of true maxima. Without this the buildup of errors tended to cause the algorithm to 'run away', iterating forever.

The Hilbert transform was implemented according to the method of Bendat[72] using a fast Fourier transform.[63] The phase of the analytic signal is obtained using a sign-aware arctangent function, and the result is 'unwrapped' to remove the discontinuities where the angle crosses from π to $-\pi$, as shown in Figure 4.9. This phase function is differentiated with the Savitzky-Golay algorithm,[86,78] using a five element vector by default. A side effect of this is that the shortest wavelength obtainable from the transform corresponds to five datapoints, as compared to two datapoints for the Fourier transform (the Nyquist limit). A simplistic differentiation algorithm requiring just two datapoints was tried during testing, but this sometimes exhibited artefacts and was subsequently dropped.

A number of smaller programs were written to obtain the various spectra, statistics and plots from the energy-frequency-time distribution obtained using the Hilbert transform code.

5.2 Preliminary Tests

The implementation of the HHT method was checked with a number of simple test datasets, both to verify the programs, and to build confidence in the use of the method and interpretation of results.

The first test was a sine wave of constant amplitude, but with an abrupt change in frequency at its centrepoint; this test was described by Huang *et al.*[62] The signal was symmetrical and possessed no riding waves, and so already passed the criteria for being an IMF, therefore the Hilbert transform was taken directly without recourse to EMD. The amplitude-frequency-time portrait (HHT spectrum) of this system showed accurate values for the two frequencies, and a very sharp transition between them at the correct point. A frequency marginal spectrum of this signal was then taken and compared with the corresponding Fourier transform. There was little difference between the two plots, but the HHT marginal spectrum had sharper peaks, and lacked the ripple artefacts observed in the Fourier.

The second test was a sine wave with linearly increasing amplitude. This signal again matched the IMF criteria and required no EMD processing. Its time marginal spectrum showed the expected increase in amplitude.

The third test was a pair of sine waves of different frequencies, one with constant amplitude, and one with linearly increasing amplitude. This dataset did not meet the IMF criteria, and so first required processing with EMD. The IMFs obtained were *not* identical to the component sine waves, but were of the correct frequencies and did show the expected variations in amplitude over time.

The fourth test was a pair of constant amplitude sine waves, one of which increase linearly in frequency while the other remained constant. The EMD algorithm again failed to obtain the component waves precisely, but the IMFs did exhibit the correct trends in frequency.

A fifth test was to synthesise two random walk signals such that the first 50% of the data were common to both, but the second 50% were allowed to diverge. Sifting these signals with EMD obtained two completely different sets of IMFs. The IMFs at both high and low frequency were observed to be markedly different in the data section common to both signals. The stability of EMD to remote changes in the dataset will be returned to in Section 5.3.

A number of signals of more complicated phenomena, such as the Lorenz attractor[87] were constructed and tested to gain more general experience of the results obtained from the HHT. It was noticed during this testing that where an IMF possesses a low amplitude extremum immediately neighboured by high amplitude extrema, the waveform is highly asymmetric. This is caused by the EMD algorithm stopping *immediately* the zero-crossing criterion is achieved: the small extremum only just barely crosses the zero axis. Consequently the numerical errors from the phase detection stage become large in this vicinity, resulting in wildly incorrect frequency values for a few points. This problem might be avoided with some further refinement to the EMD seek-cycle stopping criteria to enforce the second element of the IMF definition.

This error behaviour is generally quite rare, though, and affects only a small region when it does occur. A simple treatment for the effects, used throughout the current work, is to *clip* the data: points whose amplitudes are less than 1% of their IMF's mean are excluded from the spectrum—they are simply not drawn. This has a negligible effect on marginal spectra, and removes visually confusing artefacts from time-frequency portraits.

5.3 Stability of the EMD algorithm

If the EMD procedure is to be used, it would seem prudent to develop some intuition as to the stability of its output against numerical errors, both in the input data, and accumulated during the seeking procedure itself. Do small changes to the input signal produce large changes in the output obtained? If so, might they be reduced through some straightforward modification to the algorithm?

5.3.1 Small Random Perturbations

The stability of the EMD method to small random perturbations of the data was tested. A random walk signal with a broadband spectrum, somewhat similar to what might be obtained from a simulation of a chemical system, was constructed. Each timestep of the signal is given by

$$y_{t+1} = y_t + \zeta \tag{5.1}$$

where ζ was chosen at random from a flat distribution in the range $-0.5 < \zeta < 0.5$. To examine the stability of the sifting algorithm, two further signals were constructed. The first of these was a copy of the original signal with a sample of white noise of amplitude 0.01 added over the complete time period. The second was similar, but with white noise added over only the first 10% of the time period. The results of applying the EMD sifting procedure to each of these are shown in Figure 5.1. From this it is obvious that the empirical mode decomposition algorithm is unstable under the addition of small amounts of noise. The effects of minor changes in the data propagate along the dataset, with large differences noticeable in the *last* 10% of the second IMF where noise was added to only the *first* 10% of the dataset. This indicates that localised changes to input have nonlocal effects; such effects are clearly meaningless artefacts of the seeking procedure. There are two possible mechanisms by which these artefacts may occur:

- There is a direct nonlocal influence between widely separated points of the signal, and localised perturbations or errors directly have a global effect.
- Perturbations or errors have a tendency to propagate laterally through the dataset in subsequent seeking cycles.

The relative sizes of these effects are difficult to quantify analytically due to the empirical nature of the EMD process.

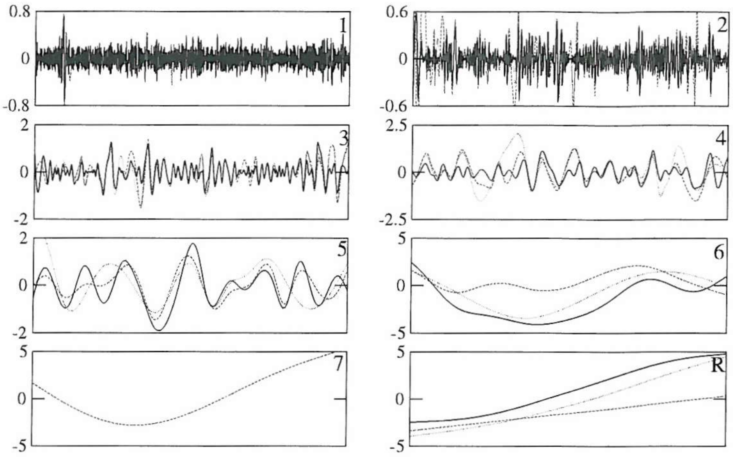

Figure 5.1: Random walk sifted by EMD; a random value between -0.5 and 0.5 was added to the signal at each timestep. The solid line shows the IMFs derived from the original signal. The dotted line shows the IMFs obtained after the addition of white noise with a maximum amplitude of 0.01 to the complete dataset. The dashed line shows the results after similar white noise was added to the first 10% of the dataset. Graphs are labelled with their IMF numbers; the graph on the bottom right is the residue component.

5.3.2 Spline Fitting Algorithms

The conventional spline algorithm used in the EMD procedure obtains a cubic polynomial for the domain between each pair of specified datapoints, i.e. each pair of maxima. The first and second derivatives are required to be continuous at the boundaries between the specified datapoints, which requires the solution of a system of simultaneous equations. An equation involving a tridiagonal matrix is set up and solved to obtain the coefficients for the cubic polynomials,[78] and a consequence of this approach is that each coefficient depends on *all* the datapoints in the signal. It follows that every point in every IMF depends on *all* of the points in the original dataset, irrespective of position, thus the first of the mechanisms described above by which nonlocal artefacts may be created is definitely present. If a spline algorithm were used in which only a local set of datapoints was needed for interpolation, such artefacts might be reduced.

One of the principal problems with EMD noted by Huang *et al.*[62] was that the cubic spline algorithm used in the seeking procedure had a tendency to over- and under-shoot the data, and this created problems where large numbers of seeking cycles were required to yield meaningful IMF components. These errors have a tendency to exaggerate abrupt changes in the amplitude of the signal, causing artefactual bumps that require further seeking cycles to flatten, and presumably create errors that could propagate through the data via either of the two mechanisms outlined above. If a spline algorithm could be found that was less prone to the under- and over-shooting problems, it is possible that both the nonlocal artefact behaviour and the overall stability might be improved.

The Akima spline algorithm[88, 89] is a method for fitting cubic polynomial functions to a dataset, the resulting curve passing exactly through all specified points. A value is obtained for the first differential (slope) of the curve at each datapoint, using the coordinates of only that point and the two adjacent points in either direction. The two required points beyond each end of the dataset are estimated using the endmost point and two points adjacent to this. Interpolation of a value between a pair of specified datapoints is achieved by deriving the coefficients of a cubic polynomial from the slopes at the points and their coordinates. The coordinates of any point along the spline curve therefore depend only on a highly localised set of datapoints: three points on each side of the required abscissa. A key consideration in the design of the algorithm was the removal of the overshooting behaviour of the conventional cubic spline. Removing the requirement for a global functional fit helps achieve this, with the side benefit of a significant speed increase. Akima reports that the procedure exhibits no problems with convergence or numerical stability.[89]

The EMD algorithm was reimplemented using the Akima spline fitting procedure. An intermediate step in the seeking procedure showing the difference between the conventional and Akima splines is shown in Figure 5.2. The reduction in the under- and over-shooting behaviour is immediately evident in this case: the Akima curve hugs the dataset much more tightly. Numerous trials of the modified EMD algorithm with different data uncovered no problems with the numerical stability of the new interpolation function.

5.3.3 Quantitative Comparison

From the reasoning above it would seem plausible that the Akima spline might improve the stability of the EMD seeking procedure, both to small random data perturbations and to the

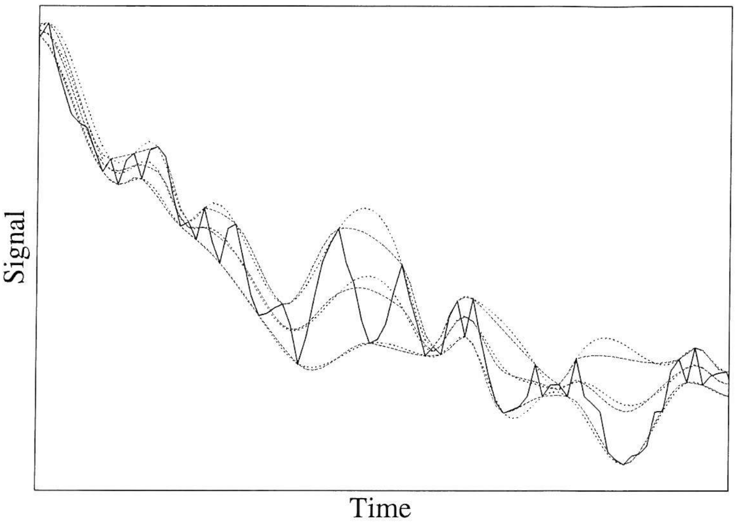

Figure 5.2: Examples of spline fitting in EMD. The solid line shows the signal whose maxima and minima are being fitted. The upper and lower dotted lines show the conventional cubic spline curves and the dashed lines the Akima splines. The curves through the centre of the dataset are the mean values between the fitted splines. The Akima curves are clearly far less prone to overshooting the data than the conventional cubics.

build up of numerical errors. It might affect both the global stability to local perturbations, due to the localised nature of the fitting, and also the global stability to global perturbations.

To quantify any improvement, a large number of test signals were processed with both the conventional and Akima versions of EMD. These were created using a similar procedure to that described in Subsection 5.3.1. A random walk reference signal was created and processed to obtain a reference set of IMFs. A small modification was then made to the reference signal (as explained below) and a new set of IMFs derived from it. The total RMS deviation between each IMF common to both datasets was calculated, and the mean of these RMSDs was taken as an indicator of the discrepancy induced by the perturbation

$$\text{discrepancy} = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\frac{1}{N}\sum_{x=1}^{N}\left(I_j^{\text{modified}}(x) - I_j^{\text{reference}}(x)\right)^2} \tag{5.2}$$

where N is the number of points in the dataset and IMF components, M is the number of non-trend IMF components common to both and $I_j^{\text{modified}}(x)$ refers to element x of the jth IMF in the set obtained from the modified trajectory.

The modifications to the reference again comprised low amplitude white noise signals

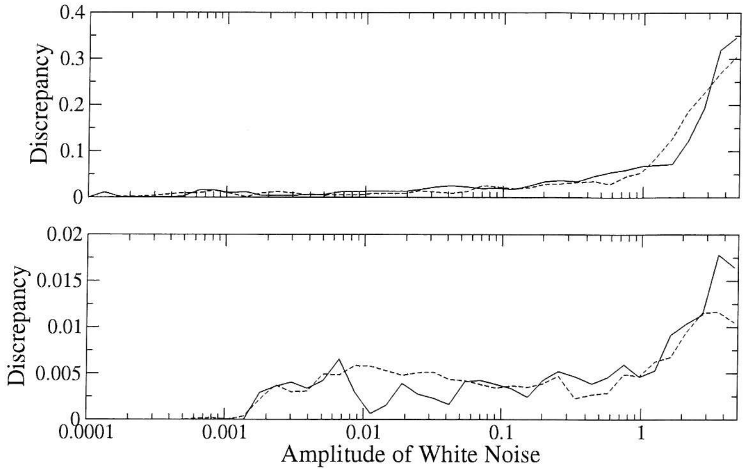

Figure 5.3: Discrepancy tests for the Akima and cubic splines. Top: wholly modified signal. Bottom: signal with first 10% modified. The Akima is drawn with a solid and the conventional spline with a dashed line.

added to either the whole or the first 10% of the dataset (*whole* or *section* modification, respectively). Noise amplitude was varied over a wide range of values, and 100 tests were performed for each. The reference signal was constant across all tests and all amplitude values. Conventional and Akima splines were compared, and the results are shown in Figure 5.3.

A number of observations may be made from this Figure. First, the wholly modified systems generally have a higher discrepancy than the sectionally modified ones, as would be expected. Also, in both cases the discrepancy increases sharply as the amplitude of the noise added approaches the amplitude of the wave-like components of the random walk signal (close to 1.0). Again this would be expected. In the lower graph, where only a short section of the trajectory is modified, there is an abrupt attenuation of the discrepancy below noise amplitudes of 0.002. The size of the modification in these cases is insufficient to perturb the algorithm from the reference solution. Overall, though, the curves are very similar for the two splines, and the broad conclusion to draw is that adopting the Akima spline gives no measurable improvement to the stability of the EMD algorithm against small discrepancies perturbing the solution set.

Two means of error propagation through the sifting procedure were described above. To recap, the first is where each point in the signal depends directly on *every* other point, and small perturbations can have nonlocal effects; in the second, a small perturbation in one part of the signal propagates laterally in successive sifting steps, multiple steps being needed for the perturbation to affect all points of the dataset. The localised nature of the Akima spline does not localise the effects of small perturbations to the dataset. As the first, nonlocal, mechanism of error propagation is no longer in operation this implies that the second, lateral, mechanism *must* be in effect and causing the global changes observed.

Using the Akima spline fitting procedure also has no clear effect on the global stability of the fitting against global changes. Reduction of the under- and over-shooting behaviour, coupled with the localised nature of the fitting appears to have no effect on the stability under small changes in the signal.

Further examination of the variability of the discrepancy reveals that the changes wrought by the addition of noise do not have a completely random distribution, as suggested by the abruptness of the attenuation of the discrepancy in the lower graph of Figure 5.3. Figure 5.4 shows a histogram where the log of the perturbation size is plotted along the horizontal axis, the log of the discrepancy along the vertical axis, and the number of occurrences is indicated by the colour; 100 trials were performed for each noise amplitude value here. The presence of distinct, discrete bands of very similar discrepancy indicates that the addition of noise tends to cause the EMD algorithm to 'flip' among a discrete set of possible solution states. That these bands have significant breadth indicates that this 'quantisation' is not perfect: rather, the bands consist of a group of closely related states. Although the data shown is for the Akima spline with a 10% modified signal similar behaviour was observed in the tests of the cubic spline and the cases where the entire trajectory was modified.

To examine the hypothesis that a tighter spline fitting procedure would improve the algorithm's stability against the build up of errors over large numbers of seeking cycles, the test in which the EMD algorithm failed to separate a pair of cosine components of very similar frequency described by Huang *et al.*[62] was performed as described. A wave was generated from the equation

$$f(x) = \cos\frac{2}{30}\pi x + \cos\frac{2}{34}\pi x \tag{5.3}$$

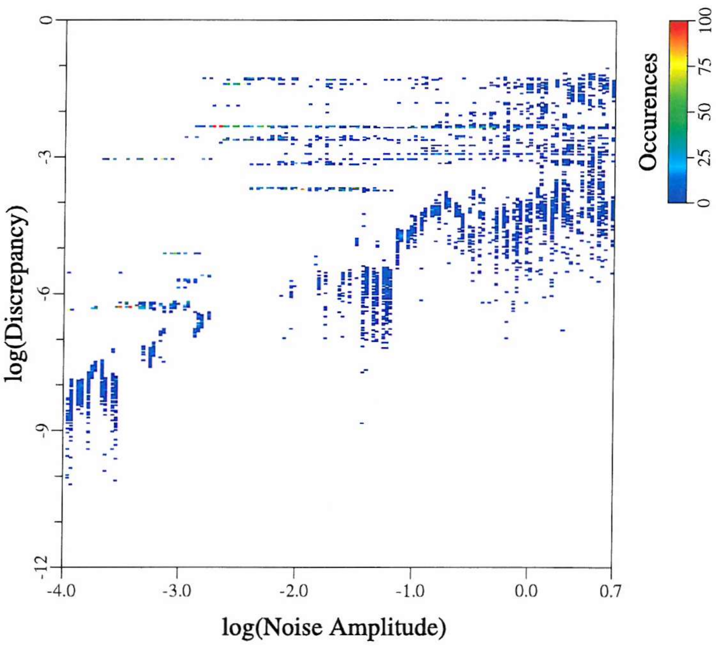

Figure 5.4: Histogram showing the occurence of discrepancy values as a function of the amplitude of noise added to the first 10% of the trajectory. The data shown are for the Akima spline fitting algorithm. Note that the axes are logarithmic to the base of 10. Note the banding, particularly evident where the discrepancy is greater than 10^{-4}.

where $1 < x < 512$. The cosines of similar frequency create 'beat' interference over the course of the dataset. As this signal is very nearly an IMF without any sifting, very stringent stopping criteria were used to force its separation into multiple components. For the original cubic spline, around 300 seeking cycles were used for each IMF, and for the Akima around 100, resulting in 11 and 14 IMF components respectively. Most of the IMF components were of very low energy, and only the first pair were of interest; these are shown in Figure 5.5, along with the test waveform. It can be seen that the cubic spline has obtained a pair of IMFs with 15 and 17 waves, while the Akima spline IMFs contained roughly 16 and 17 waves, although these latter were badly defined. The theoretically correct values are $512/34 = 15$ and $512/30 = 17$ ($\cos 2\pi x$ in the range 0–512 contains 512 waves; $\cos \frac{2\pi x}{34}$ has $512/34 = 15$ waves in the same range). Although the cubic spline obtained the correct numbers of peaks, significant contamination from the large numbers of seeking steps can be seen, confirming the result of Huang *et al.* for this case. The Akima algorithm is clearly much the worse of the pair; the signals obtained are not clear enough to allow accurate peak counting, and suffer

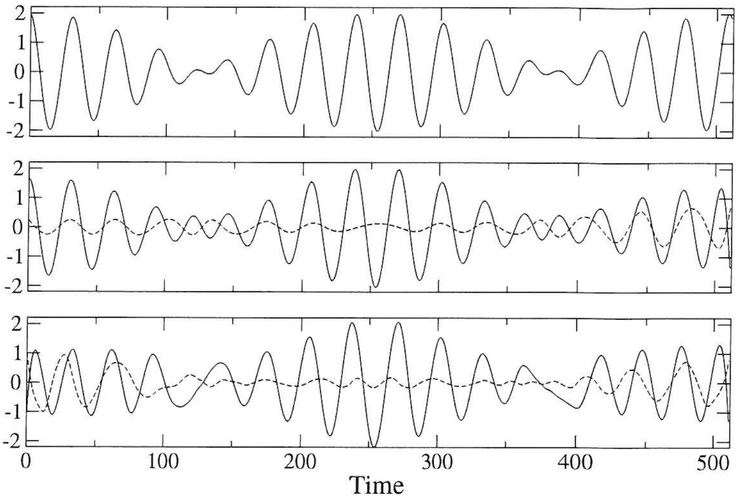

Figure 5.5: Top: the signal being processed. Middle: the first two IMFs obtained using conventional cubic splines, the first drawn with a solid and the second with a dashed line. Bottom: the first two IMFs obtained using the Akima spline, again the first is drawn with a solid and the second a dashed line. The distortion in the waveforms resulting from large numbers of seeking cycles is considerably worse for the Akima algorithm: the periods of the components are more variable with respect to time, and some wave peaks are missed.

from severe contamination.

The hypothesis given by Huang *et al.*,[62] that using a spline fitting algorithm less prone to under- and over-shooting the dataset would lead to more stable behaviour where large numbers of seeking cycles are required, thus seems to be false. The accumulation of errors in the seeking procedure cannot be judged by the intuitive tightness of the spline fitting algorithm. In summary then, although the Akima spline gives a localised and intuitively better fit to the maxima and minima, it yields no measurable improvement to the stability of the EMD algorithm. Moreover, in the event that large numbers of seeking steps are required to obtain usefully separated IMF components, the conventional cubic spline outperforms it.

In the above investigation, the empirical mode decomposition algorithm was shown to be highly sensitive to small changes in input data. Any small, local perturbation has a global effect, commonly leading to an entirely different set of IMF components being obtained. The output of the algorithm appears to switch among a discrete set of possible solutions under the action of such modifications, though there exists a degree of variability in the neighbourhood of these states. Adopting the Akima spline fitting procedure does not yield the improvements that might be hoped for from the conjecture of Huang *et al.* about the stability of the algorithm.

The effects of the instability of the EMD procedure to small changes on the resulting HHT spectra are difficult to predict, though. This will be returned to in Section 5.6.

5.4 Hilbert Boundary-Effect Modelling

For the Hilbert transform to be calculated rapidly, the Fast Fourier Transform[63] is employed. This has the unfortunate side effect of creating ripples (Gibbs effects) in the transform due to the inevitable disjunctures at the signal boundaries. Such ripples may be greatly reduced where a conventional Fourier spectrum is required by employing a windowing function, such as the Blackman window described in Section 4.1. As the Fourier transform is being used to calculate a convolution, the use of a windowing function would not be possible here. These errors are significant at the edges of the signal, resulting in incorrect frequency and amplitude values, and this fact is recognised by Huang *et al.*[62] Their prescription for solving this problem is to add extra "characteristic waves" at the edges of the dataset, which are attenuated to nothing in some unspecified manner. Their explanation of this procedure is both very short and difficult to interpret. Komm *et al.*[82] have not used Huang's approach to these edge effects; they say that the results within "about 5%" of the ends are "less reliable", without attempting to quantify how much so.

The approaches to dealing with these end effects can thus be summarised as follows

1. Do nothing, and try to ignore points at the edges of the data on a subjective, visual, case-by-case basis.
2. Clip points from the edges of the data automatically.
3. Invent extra data at the edges of the signal and reduce its amplitude gradually to zero with some form of attenuating function.

To quantify the amount of data that must be removed in the second approach described above, some criterion for the maximum allowable error is required. The simplest criterion is the frequency resolution of the HHT technique, stated by Huang *et al.*[62] to be $1/T$ where T

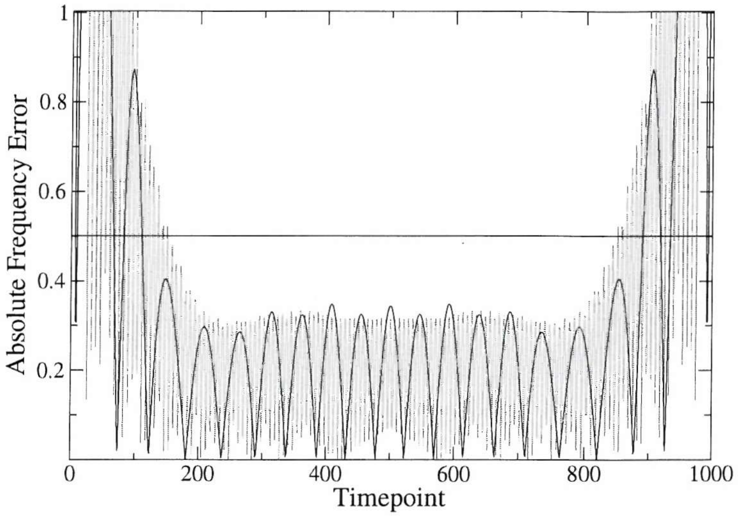

Figure 5.6: Absolute error in frequency returned by the Hilbert transform as a function of the timepoint within the signal. The errors for a signal containing 10.5 wavelengths is drawn with a solid line, and those for a signal containing 70.5 waves with a dotted line. The absolute frequency error is shown as a fraction of the resolution and a solid line is drawn at the maximum permissible error value.

is the time period of the dataset (in other words, the wavelength resolution of the technique is equal to the length of the dataset). The exclusion criterion is then, 'any point with a frequency error of more than half the resolution is unreliable'. If a datapoint could be resolved to have a frequency of f, $2f$, $3f$, etc., then this error condition restricts us to points within $f \pm \frac{f}{2}$.

A large set of signals comprised of constant amplitude sine waves at different frequencies was created, and their Hilbert transforms obtained. The frequencies were chosen to create a phase difference of π at the ends of the signal to maximise the Gibbs effects. Absolute errors in frequency from the known correct value were quantified as a function of timepoint, and the position where these errors fell below half the resolution was noted. A pair of such error functions for different wave frequencies is shown in Figure 5.6. Several tests were carried out with different sizes of dataset, and the following results were found not to vary with this.

A curve of the form

$$E = bt^c - a \tag{5.4}$$

was fitted to the data, where E is the position at which data should be clipped as a fraction of the set length and t is the frequency on a scale normalised to unity. The fit can be seen

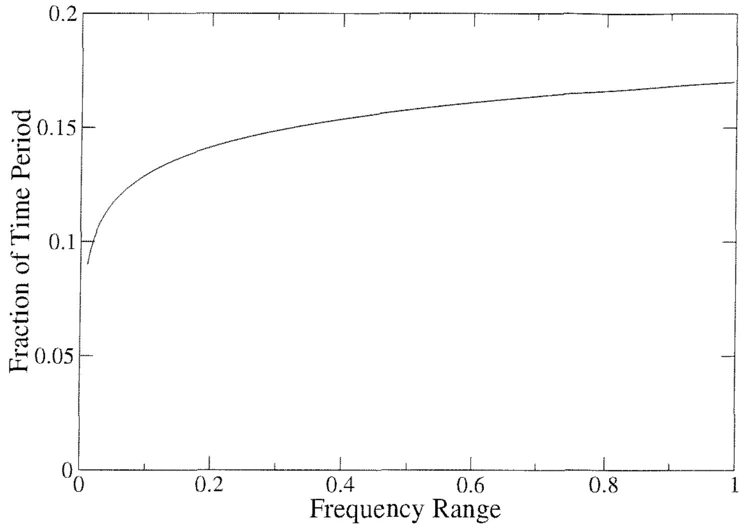

Figure 5.7: Position in a dataset at which the absolute error in frequency falls below the acceptable value of half the resolution of the transform. For low frequency values 9% (or less) of the data must be clipped from each end, rising to 15% at the maximum frequency.

in Figure 5.7. The values obtained were $a = 1.30034$, $b = 1.47041$ and $c = 1.21602 \times 10^{-2}$. This function may be used to calculate the position in an HHT spectrum at which data should be excluded for the accuracy to be within the resolution tolerance of the method. For higher frequency values to be reliable, as defined by the exclusion criterion given above, quite a large fraction of the data must be removed: up to about 17% at *each end* of the dataset. This could clearly be a significant handicap where exact figures are required from the technique.

If the third approach to the end effects described above is to be taken, the attenuating function which produces optimum results should obviously be used. As described by Harris,[64] many types of windowing function have been adopted for the conventional Fourier transform, mostly based on arbitrary mathematical functions, many of which obtain far from optimal results. His paper quantified the results of many such functions in a straightforward way, and attempted to discover the best function for frequency discrimination. In the following, a number of possible attenuating functions for the HHT transform are evaluated.

The signal examined in each case comprised a single sine wave with exactly 20.5 wavelengths, to which was added one further wavelength at each end of the set; 1000 datapoints were used for the main signal. The half wavelength discontinuity at the end of the set again had the effect of maximising the size of the Gibbs errors that are of interest. A series of possible attenuating functions were applied to the waves at each end of the set, Hilbert transforms were taken, and the absolute frequency error as a function of set position was obtained. Each function was scaled to return unity at the edge of the main signal. The functions were as follows, s being a parameter varying between 1 at the edge of the dataset proper to 0 at the point of maximum decay:

1. s (Linear decay)
2. $\sin(\pi s/2)$ (Sinusoidal decay)
3. s^2 (Square polynomial decay)
4. $s^{0.5}$ (Root decay)
5. $\exp(-k_e s)$ (Exponential decay with constant k_e)
6. s^{k_n} (Power decay with constant k_n)

The values for the constants which yielded optimum values of the total integral of the absolute error over the dataset were obtained; these were found to be $k_e = 2.3$ and $k_n = 1.1$. There obviously exists a lot of redundancy in the functional forms tried, but certain key values corresponding to the square root, etc. were tested early in the process as obvious, intuitive candidates. Their inclusion serves to highlight the problems encountered in attempting to guess attenuation functions and their parameters. The results are shown in Figure 5.8.

A number of observations can be made from this Figure. First, the effects of using an attenuation function on made up data added to the edges of the signal are dramatic. None of the functions exhibit an error greater than the resolution for points within the dataset proper. The linear window is very nearly the best of the set, with a slight improvement to be gained from a power decay with a constant of 1.1 (almost identical to the linear function). Other windows with obvious intuitive forms, and/or guessed parameters, produce much poorer results, as also observed for Fourier windowing functions.[64]

The conclusion to be drawn from these results is clear: if the analyst is content with making up fake data at the edges of his measured signal, an excellent reduction in the level

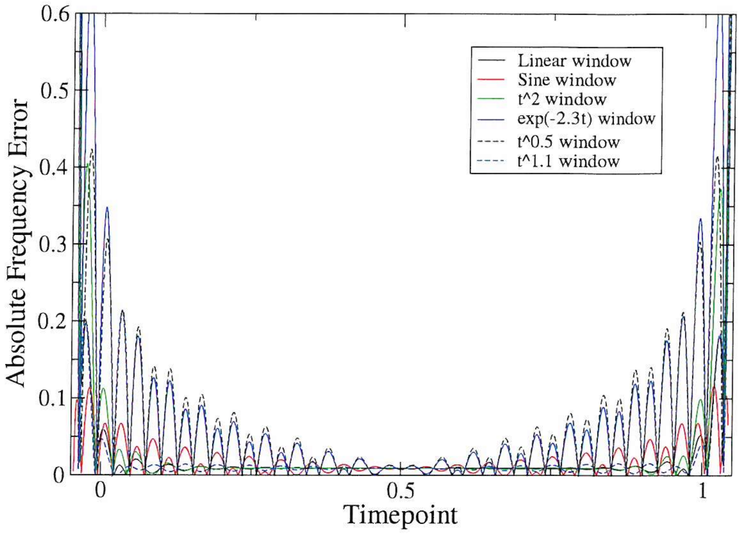

Figure 5.8: Absolute frequency errors on a sine wave with 20.5 periods as a function of timepoint for a set of attenuation functions.

of noise created by the Gibbs phenomenon can be achieved. The optimal attenuation function to use from the set described above has the form $s^{1.1}$, where s varies as explained. Using intuitively obvious, non-parameterised attenuation functions will generally produce results that are suboptimal, sometimes drastically so. If the analyst is *not* prepared to invent extra data, but still requires accurate frequency information, clipping the spectra with Equation 5.4 using the obtained parameters will remove datapoints with frequencies that have errors larger than the resolution of the transform, as stated by its designers. The drawback with doing this is that a large proportion of the dataset has to be excluded. A third intermediate possibility would be to use an attenuation function without inventing new points—less points in total would be affected than if no attenuation were used. In the current work, either the clipping function is used, or the dataset is left as-is. Where clipping is used, this will be indicated.

5.5 Brownian Dynamics

A test system was required that was both simple to implement and study, but at the same time displayed behaviour closer to that observed in condensed phase molecular dynamics, including transition events between minima on an energy landscape. To this end, a series of Brownian dynamics (BD) simulations were performed. The system studied comprised a

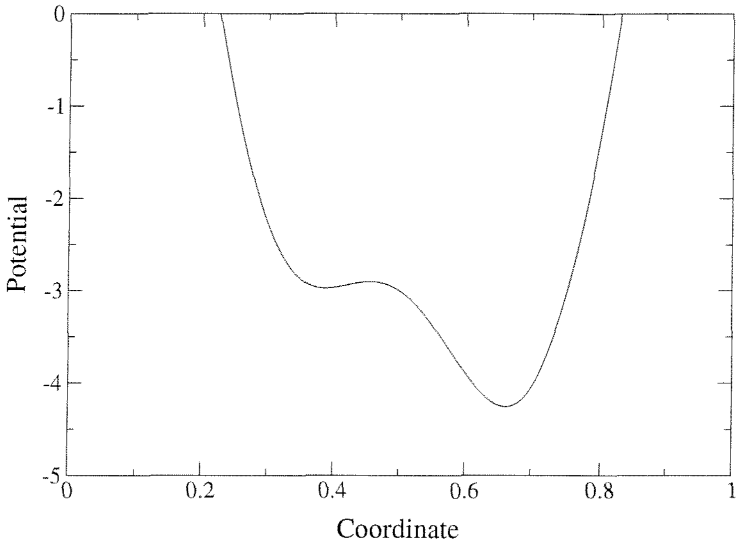

Figure 5.9: The potential well used in the Brownian dynamics simulations. The well is constructed from a short Fourier series described in the text.

single particle moving freely on a one dimensional potential energy surface driven by random thermal excitation and hindered by friction.

5.5.1 Implementation Details

The potential surface was constructed from a short Fourier series and contained two minima of different energies separated by a low barrier. The characteristic frequencies of the minima are different and the surface rises steeply on either side of the pair. The functional form of the series was as follows

$$V(x) = 5\cos 2\pi x + 2\cos 4\pi x + \frac{3}{2}\cos \pi x \tag{5.5}$$

The surface described by this is shown in Figure 5.9.

The BD algorithm used is based on that of Ermak and Buckholtz,[90] modified[7] to reduce to the velocity Verlet algorithm upon setting the frictional coefficient, ξ, to zero. The following equations are integrated in the standard velocity Verlet fashion

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + c_1 \delta t \mathbf{v}(t) + c_2 \delta t^2 \mathbf{a}(t) + \delta \mathbf{r}^G \tag{5.6}$$

$$\mathbf{v}(t+\delta t) = c_0\mathbf{v}(t) + (c_1 - c_2)\delta t\mathbf{a}(t) + c_2\delta t\mathbf{a}(t+\delta t) + \delta\mathbf{v}^G \tag{5.7}$$

Where the constants, c_0, c_1 and c_2 were given by

$$c_0 = e^{-\xi\delta t} \tag{5.8}$$

$$c_1 = (\xi\delta t)^{-1}(1 - c_0) \tag{5.9}$$

$$c_2 = (\xi\delta t)^{-1}(1 - c_1) \tag{5.10}$$

The correlated variables $\delta\mathbf{r}^G$ and $\delta\mathbf{v}^G$ are derived from a bivariate Gaussian distribution as follows

$$\delta\mathbf{r}^G = \sigma_r\zeta_1 \tag{5.11}$$

$$\delta\mathbf{v}^G = \sigma_v(c_{rv}\zeta_1 + \sqrt{1 - c_{rv}^2}\zeta_2) \tag{5.12}$$

where ζ_1 and ζ_2 are random numbers independently sampled from Gaussian distributions with means of 0 and variances of 1. The variances, σ_r^2 and σ_v^2, and correlation coefficient, c_{rv}, are given by

$$\sigma_r^2 = \delta t^2\frac{k_BT}{m}(\xi\delta t)^{-1}(2 - (\xi\delta t)^{-1}(3 - 4e^{-\xi\delta t} + e^{-2\xi\delta t})) \tag{5.13}$$

$$\sigma_v^2 = \frac{k_BT}{m}(1 - e^{-2\xi\delta t}) \tag{5.14}$$

$$c_{rv} = \frac{1}{\sigma_r\sigma_v}\delta t\frac{k_BT}{m}(\xi\delta t)^{-1}(1 - e^{-\xi\delta t})^2 \tag{5.15}$$

The parameters chosen for the friction, timestep and degree of thermal excitation in all of the

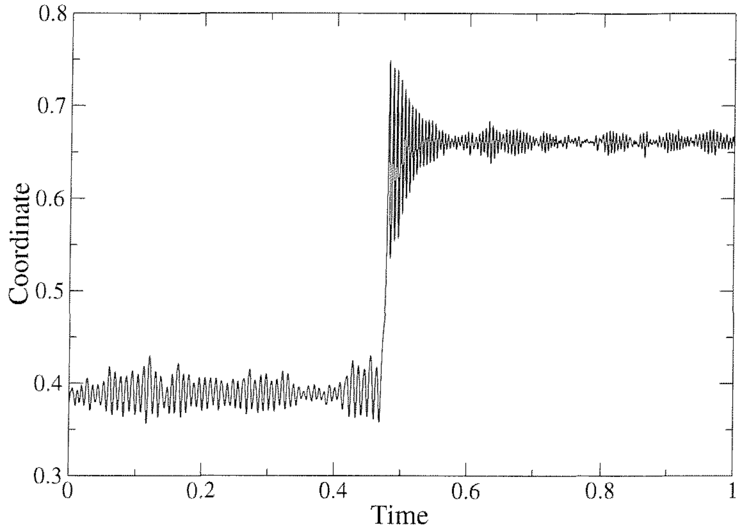

Figure 5.10: The Brownian dynamics trajectory chosen for analysis. The simulation starts with the particle in the higher potential well, a transition to the lower energy well occurs around the midpoint, and the particle then loses its excess kinetic energy to friction.

simulations were

$$\xi = 0.0001;\ \delta t = 0.02;\ \frac{k_B T}{m} = 0.0025 \tag{5.16}$$

5.5.2 Results

A series of simulations with different random number seeds were performed until a transition between the potential energy wells was observed. The simulation picked is shown in Figure 5.10. The simulation timescale is normalised to unit length, and the potential well shown in Figure 5.9.

Empirical mode decomposition of the trajectory yielded nine IMFs and a trend. Of these components, there existed a clear but unusual ‘separation of scales’: the three highest frequency IMFs contained the harmonic content of the signal, while the six lower frequency IMFs and the trend summed to give the mean values corresponding to the two minima and the transition between them. These lower IMF components were of a substantial amplitude and might better be thought of as being constituents of the true trend in the data. The sums of IMFs 1–3, and also 4–9 and the trend component are shown in Figure 5.11.

The HHT spectrum of all IMFs, along with the Morlet wavelet spectrum of the trajectory is shown in Figure 5.12. Signal energy has been plotted to de-emphasise the low frequency

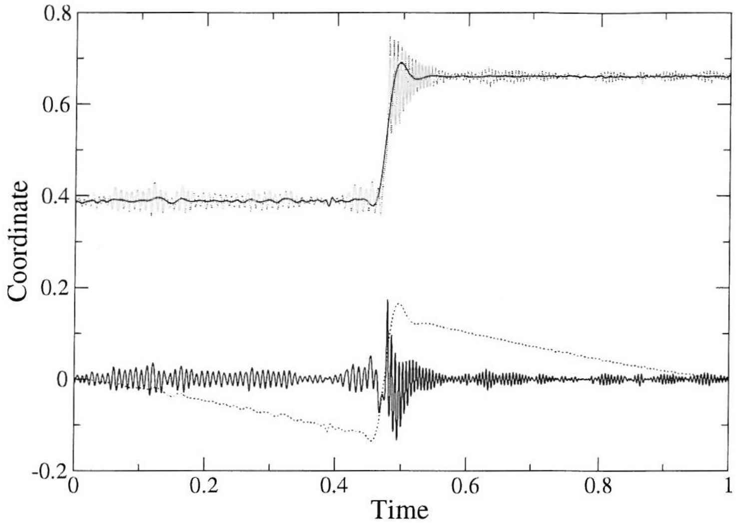

Figure 5.11: Separability of scales in the empirical mode decomposition of a trajectory containing a transition event. Top: IMFs 4–9 and the trend component are summed to give the signal drawn with a solid line. The original trajectory is drawn with a dashed line. Bottom: IMFs 1–3 together account for the harmonic content of the signal, drawn with a solid line. The sum of IMFs 4–9 *without* the trend component is drawn with a dashed line.

trend-constituent IMFs with respect to the higher frequency harmonic ones; without this, the amplitude of the low frequency constituents would distort the scale of the graphs, making them difficult to read. Log energy scales were used in both cases for ease of comparison. The edges of the Hilbert spectra were not clipped.

The first thing apparent on examining the Figure is that both spectra, as drawn, are remarkably similar. The frequencies of motion associated with the two minima are clearly visible in both wavelet and HHT plots. The transition event appears more sharply defined in frequency in the HHT spectrum, not suffering the frequency leakage evident with wavelets. Both plots also exhibit a broad feature at low frequencies close to the transition point, though the structure is visually clearer with wavelets—the sparseness of the lines in the HHT spectrum tends to confuse the eye and cloud interpretation.

What meaning can be attached to the low frequency values reported for IMF components 4–9? These are plotted in Figure 5.13. To answer this question we must remember that the EMD method decomposes a signal in terms of a basis set derived from the signal itself, rather than some function chosen arbitrarily by the analyst. Although the components recovered are not necessarily orthogonal, in practice they are numerically close to being so and the com-

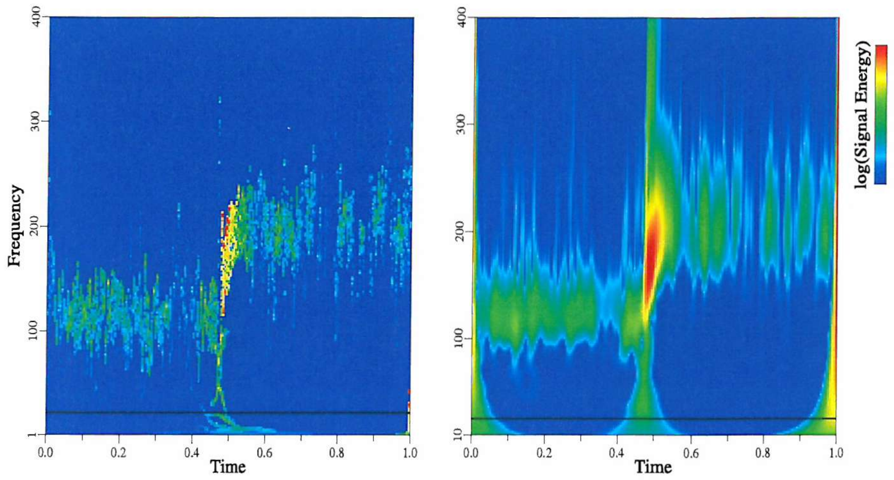

Figure 5.12: Left: HHT spectrum of the Brownian dynamics trajectory shown in Figure 5.10. Signal energy is plotted for clarity. Right: Morlet wavelet spectrum of the same trajectory, plotting log signal energy for ease of comparison. The horizontal line shown in both graphs indicates the upper limit of the low frequency region integrated over in Figure 5.18.

ponents recovered may interfere with one another in both a 'constructive' and a 'destructive' fashion, even though the decomposition is strictly subtractive. It is obvious from comparing the low frequency IMF components, shown in Figure 5.13, with their sum, shown in Figure 5.11, that the components must interfere in a destructive manner in places to equal their sum. Their wavelike character thus possesses an artificial, synthetic nature similar to that seen in Fourier analysis. Further consideration will be given to the interpretation of this in Section 5.6.

Marginal spectra from both the HHT and wavelet spectra, along with Fourier transforms were derived for the simulation shown in Figure 5.10. The Fourier transforms taken without a windowing function and also with the Blackman window described in Section 4.1 are shown in the upper part of Figure 5.14. The latter is clearly a substantial improvement, with the Gibbs artefacts resulting from the dual transitions—in the centre and at the end of the signal—much reduced. This filtered spectrum was used for the frequency calculations below; the reduction in spectral leakage will allow more accurate determination of peak frequencies.[64] Frequency marginal spectra are shown in the lower part of the Figure. The wavelet transform exhibits a smooth variation of amplitude with frequency and little detail can be resolved due to spectral

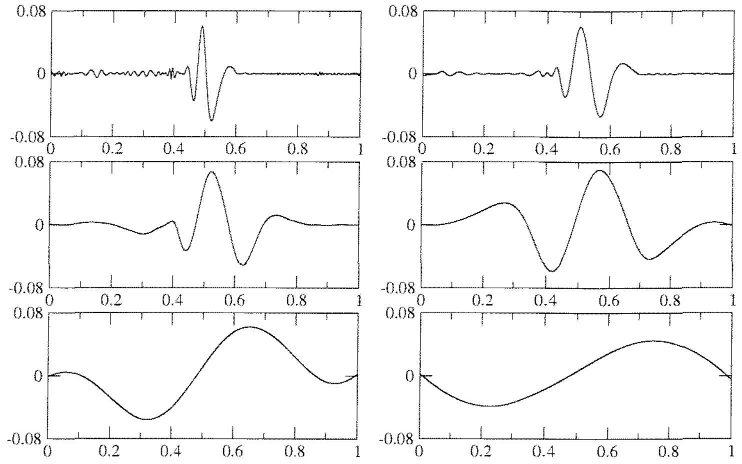

Figure 5.13: The six low frequency IMF components obtained from the trajectory shown in Figure 5.10.

leakage. Two peaks corresponding to the frequencies of harmonic vibration in the signal can be discerned, though. The Hilbert-Huang transform also shows two peaks, but separated by a much clearer amplitude minimum. The slight shift in peak frequencies in the wavelet spectrum may be caused by overlap between the peaks.

At the two minimum points, x_1 and x_2, in the potential well described in Subsection 5.5.1

$$\frac{d^2E}{dx_1^2} = 2.420094 \tag{5.17}$$

$$\frac{d^2E}{dx_2^2} = 6.356041 \tag{5.18}$$

and given that

$$\frac{dE}{dx} = -F \tag{5.19}$$

$$F = -kx \tag{5.20}$$

$$\frac{dF}{dx} = -k \tag{5.21}$$

$$k = \frac{d^2E}{dx^2} \tag{5.22}$$

and

$$\omega = \sqrt{\frac{k}{m}} \tag{5.23}$$

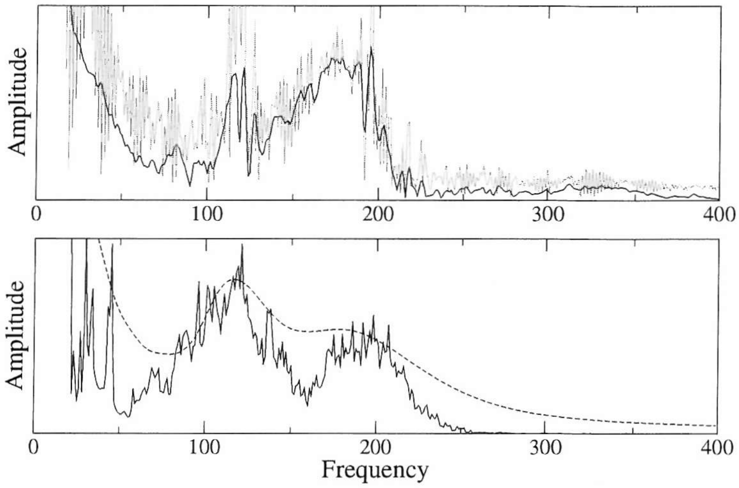

Figure 5.14: Top: Fourier transforms of the Brownian dynamics trajectory shown in Figure 5.10. The fine dotted line shows the transform of the unmodified dataset; the solid line that of the dataset multiplied by the Blackman windowing function. Bottom: frequency marginal spectra. The dashed line shows the result from the Morlet wavelet; the solid line that of the HHT.

if we assume the mass $m = 1$, then have for the minima

$$\omega_1 = \sqrt{2.420094} = 1.555665 \tag{5.24}$$

$$\omega_2 = \sqrt{6.356041} = 2.521119 \tag{5.25}$$

$$\omega_1/\omega_2 = 1.620605 \tag{5.26}$$

The frequencies of the peaks obtained from the Fourier transform were 172 and 118, the ratio between them being 1.46. Peak frequencies were obtained by fitting Gaussian distributions to the data. To obtain a stable fit to the higher frequency peak, values with frequencies below 130 were excluded from the fit; this value was judged to exclude as much as possible of the lower peak. The values measured from the wavelet marginal spectra were 178 and 116, with a ratio of 1.52; the peak values being read directly from the graph. The figures from the Hilbert spectrum were 190 and 114, with a ratio of 1.67.

These results indicate that all three spectral methods are able to obtain comparable results from the same data; the Hilbert transform, working in a completely different way to the two Fourier based methods, obtains a similar spectrum. The ratio obtained from the HHT marginal spectrum is the closest to the theoretical value, with the wavelet value coming second and the

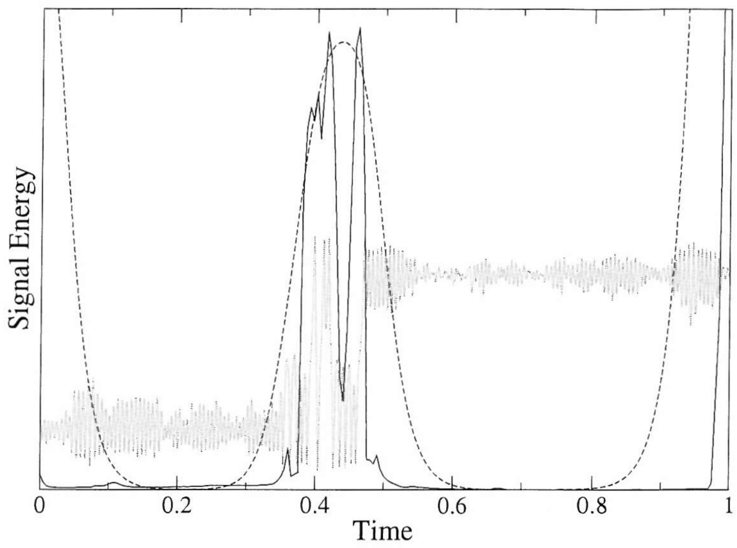

Figure 5.15: Brownian dynamics trajectory exhibiting multiple transition events. The signal energy over the low frequency region is shown as a function of time. The HHT margin is drawn with a solid line and the Morlet wavelet margin with a dashed line. The original signal is superimposed as a dotted line.

Fourier transform comes last. The broad, smooth features of the wavelet transform are probably, considering other wavelet transforms seen, the result of severe leakage in the frequency domain.

One final test using Brownian dynamics was performed on a potential energy surface modified to allow multiple transitions to take place back and forth between the two wells. The new, symmetrical potential function was

$$V(x) = 5\cos 2\pi x + 2\cos 4\pi x \tag{5.27}$$

A simulation performed on the new surface exhibiting a series of rapid back and forth transitions is shown in Figure 5.15. The time margin for both Morlet wavelets and the HHT are plotted alongside the trajectory. The signal contains multiple transitions between states between $t = 0.39$ and $t = 0.42$. There then follows a short period of 0.05 time units where the system remains trapped in a single state. This ends at $t = 0.47$ with a single, final transition event. Signal energy yielded the clearest plots for this data with its emphasis on larger amplitude features.

As can be seen, the HHT time margin registers two signal energy peaks: the first, broad and steep shouldered, follows the period of multiple transitions, starting at $t = 0.38$ and ending at $t = 0.43$. The second peak registers the final transition event at 0.47, rising at around $t = 0.45$ and falling at around $t = 0.48$. There is a clear interregnum of lower signal energy between the two peaks detected by the HHT, reflecting the period where the system is trapped in one state. This should be compared with the wavelet spectrum, which has only succeeded in obtaining a single peak for the entire sequence of events. The peak has a smooth, bell-like shape, centred on the period of temporary entrapment. None of the fine structure of the dynamics can be obtained from the wavelet margin.

The results from the Brownian dynamics simulations have shown that for the types of dataset of interest to chemical simulators, the HHT is capable of producing results directly comparable with the Fourier transform, the wavelet time-frequency spectrum, the wavelet marginal frequency spectrum and the wavelet time margin. For a signal whose average trend line exhibits a sharp transition event, EMD will derive one set of components describing the noise and wave-like behaviour of the system, and a separate set of components which, although individually meaningless, sum to yield a remarkably clear portrait of the transition. Such systems could be described as having a *meaningful* separation of scales. Taking an HHT time-frequency portrait for this system yields visually similar results to the wavelet transform. There is significantly less leakage in the frequency domain with the HHT, but difficulties emerge in interpreting regions sparsely represented by IMF components. Calculations of the peak frequency centres in this system showed an advantage for the HHT over both Fourier and wavelet transforms, the latter again significantly hampered by its frequency leakage. Also, in the time domain, the HHT performed better at discerning rare events in a complicated signal than the wavelet time marginal spectrum.

5.6 Composite Spectra

In cases where the frequencies of individual IMF components are clearly not meaningful, we might ask whether the amplitude of a point in time-frequency space defined by a set of

IMFs would appear similar to the amplitude of the same point obtained through the continuous wavelet transform (CWT). The nonzero points in the HHT spectrum—the points of the histogram through which one or more IMF components pass—are sparse, but, as shown in Subsection 5.3.1, the paths in time-frequency space taken by a set of IMFs are sensitive to small changes in the input data. Taken together, this would seem to imply that the set of paths has a somewhat arbitrary sampling of regions of the space where no meaningful scale can be defined. The IMF would then best be considered a curvilinear 'slice' through time-frequency space, whose amplitude at a given point would to some extent reflect what would be obtained through wavelet analysis. The absence of harmonic leakage should make a significant difference to such a spectrum though: regions of leakage should presumably have zero amplitude in the HHT time-frequency portrait.

To discover if the above statements are true, a test was performed, again using the signal shown in Figure 5.10. A collection of 100 copies of the signal was created, to each of which was added white-noise with an amplitude of 0.001, around 5% that of the signal's wave-like content. The HHT spectrum was derived independently for each sample, and the resulting histograms were composited into a master spectrum as follows. Where a histogram bin in the master spectrum was previously empty, it acquired the value obtained in the sample spectrum. Where the bin was nonempty, it was added to the original contents and a counter for that cell was incremented. After all data was collected, the cell contents are divided by the counter values to calculate averages. Obviously, the counter values should initially be set to unity to avoid a division by zero error.

The white noise made little difference to the overall signal energy, but was sufficient to perturb the EMD algorithm into deriving IMF components that took different paths through time-frequency space. The spectrum of the noise signal used for each sample was also taken, composited in the same way, before finally the composite noise spectrum was subtracted from the composite noise + signal spectrum (and rounded up to zero, if required, to obtain just the signal data) to produce an overall picture of time-frequency space with an information density closer to that obtained with the CWT. The result is shown in Figure 5.16. This spectrum took

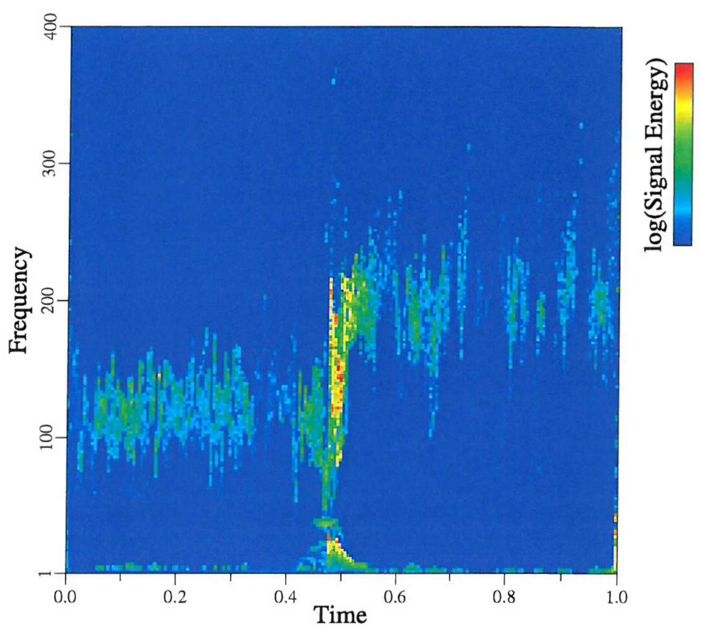

Figure 5.16: The composite spectrum obtained from 100 slightly modified copies of the Brownian dynamics trajectory shown in Figure 5.10. The log of signal energy has been plotted for clarity and ease of comparison.

a few tens of seconds to obtain on an 800 MHz PC workstation, as compared to less than one second for a conventional HHT spectrum on the same machine. This is not a drawback when producing individual plots for human inspection, but might become one if large numbers of datasets needed to be processed automatically.

This spectrum is more closely comparable to the CWT than the non-composited HHT plot shown in Figure 5.12, owing to the reduction in sparseness. The large number of IMF traces through time-frequency space build up a more complete picture of the behaviour of the system: the tiny perturbations change the IMF paths, and this results in an improved sampling of the underlying space. The spectrum still exhibits a narrower distribution of signal energy than that seen in the wavelet transform, with less leakage in the frequency domain as expected; also, the boundary effects obvious in the wavelet spectrum still remain absent from the HHT. The removal of the visually confusing artefacts in the low frequency regions of the spectrum constitutes a useful advantage.

Complicated HHT spectra, or spectra where regions of interest are under-represented by

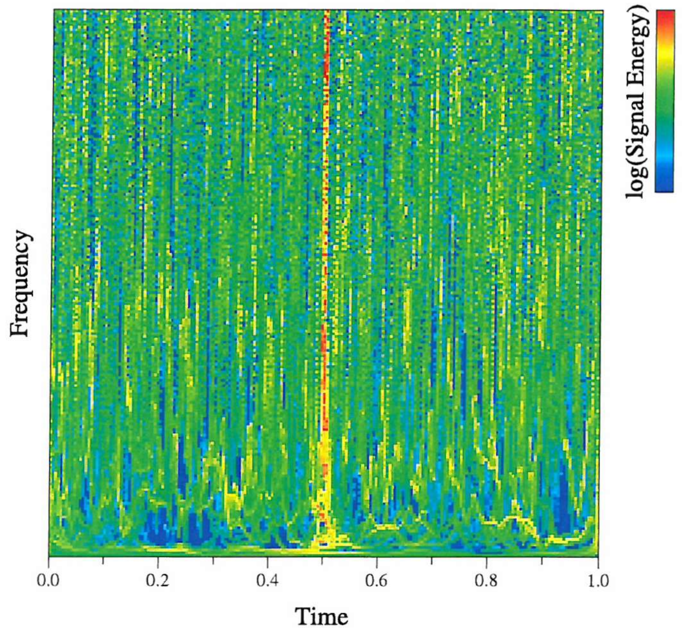

Figure 5.17: The composite spectrum obtained from the signal used in Figure 4.3. The transition event is sharply defined in time across all frequencies.

the time-frequency distribution of IMFs, may be shown clearly and completely. This method of obtaining significantly improved spectra from the HHT method can be applied to any signal. A second example, the test signal used for demonstrating the spectrogram, wavelet and HHT spectra from Figures 4.3, 4.5 and 4.12 is plotted in Figure 5.17. Note that signal energy is shown, rather than amplitude. The greater coverage of frequency-time space again generates a much clearer spectrum.

One thing important to note when constructing such spectra is that the amplitude of the perturbing noise signals must be judged correctly. A value of around 5% of the signal's harmonic content appears to be generally appropriate. If the noise amplitude is too low, the IMFs obtained will not cover the entire space. This is due to the somewhat quantised nature of the EMD solution space observed in Subsection 5.3.3. In such cases, the IMFs obtained will tend to cluster into a larger, but still discrete set of bands, particularly evident at low frequencies. But, on general grounds of computational precision and numerical error minimisation, the noise signal should also be kept as small as possible: just large enough to perturb the sifting process sufficiently, without swamping the target signal.

An obvious question to ask about this compositing procedure is how it changes the marginal signals obtained from a dataset. This was investigated for the Brownian dynamics trajectory, and the results are shown in Figure 5.18. The low frequency region of the spectrum indicated in Figure 5.12 is also examined on its own: this is where the worst problems due to under-representation occur in the conventional spectra. Time marginal plots use signal energy to emphasise high amplitude features.

The first thing to note about these graphs is that they are plotted on identical scales, except where stated for the low frequency time margin. The overall amplitudes and signal energies of the frequency and time margins are more or less unaffected by the compositing process over most of the frequency range, as are the shapes of the peaks seen. This is unsurprising, as the frequency range above about 50 units is already well sampled by the conventional HHT. It is over the low frequency bands that the differences between the results becomes significant. As can readily be seen in Figure 5.12, this area of the spectrum is poorly sampled, exhibiting a skeletal structure that is hard to make visual sense of. The integrated time margin derived from this area has a much lower overall amplitude than that obtained from the composite signal, in contrast to the situation where the entire frequency spectrum is examined, precisely because of this poor sampling. The composited spectrum reveals a central peak centred on the transition event with sharper time definition due to the larger number of points used.

In summary, composited HHT spectra yield frequency-time portraits that are similar to the more intuitive wavelet spectra, but without the problems related to spectral leakage. They are much less prone than conventional HHT to the unwanted artefacts caused by sparse sampling of the space. Application of the procedure requires a little judgement, but can be applied to any signal. The changes to the time and frequency margins are negligible where the HHT is sampling sufficient points, but obtain a distinct improvement in regions where undersampling is taking place. It clearly *is* an improvement, because the marginal energy/amplitude *increases*: This comes about by a greater sampling of the underlying frequency-time space, resulting in a better description. A drawback of the method is that it does require substantially more computing time: two orders of magnitude for the spectra shown here.

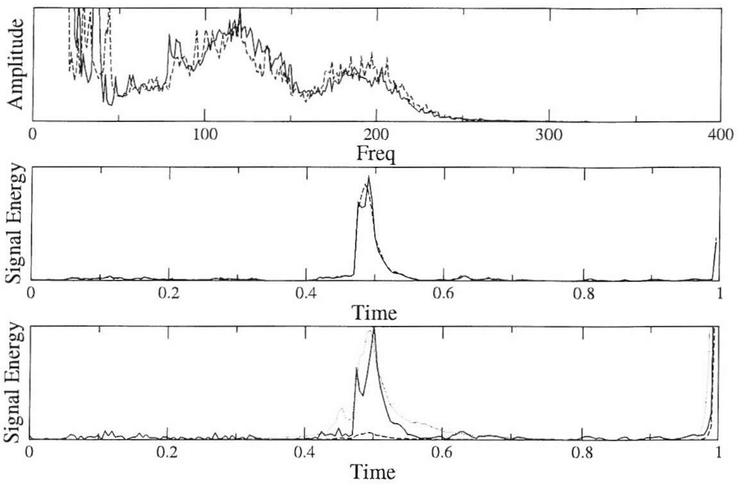

Figure 5.18: Composite and non-composite marginal spectra. Top: frequency versus amplitude. Middle: signal energy versus time integrated over entire frequency range. Bottom: signal energy versus time integrated over the low frequency range shown in Figure 5.12. The composite spectra are drawn with solid lines and the non-composite with dashed lines. In the bottom graph, the non-composite margin multiplied by a factor of 15 is drawn with a dotted line.

5.7 White Noise

Examination of HHT spectra reveals that most signals extracted from computer simulations tend to have high amplitude features at very low frequencies. Obtaining intuitively reasonable graphs of such spectra often necessitates the use of adjustment strategies, such as the 'signal energy' defined in Section 4.6, to de-emphasise these regions. To further investigate this preponderance of low frequency values, the HHT frequency marginal spectrum of a white noise signal was obtained, along with the Morlet wavelet marginal spectrum and the Fourier transform for comparison. These are shown in Figure 5.19.

As should be expected, the graph of the Fourier spectrum exhibits no trend in its amplitude profile. The wavelet frequency marginal spectrum, though not perfectly flat, again shows no systematic trend: the ripples would presumably be erased with increased data sampling, resulting in a flat spectrum. In the HHT marginal spectrum, though, a clear trend emerges. Most of the spectrum appears occupied with a fairly straight downward sloping line, while the lowest frequency 15% shows a much steeper curve. This suggests why the amplitude spectra of simulation data overemphasise lower frequencies. The cause of the trend will be discussed

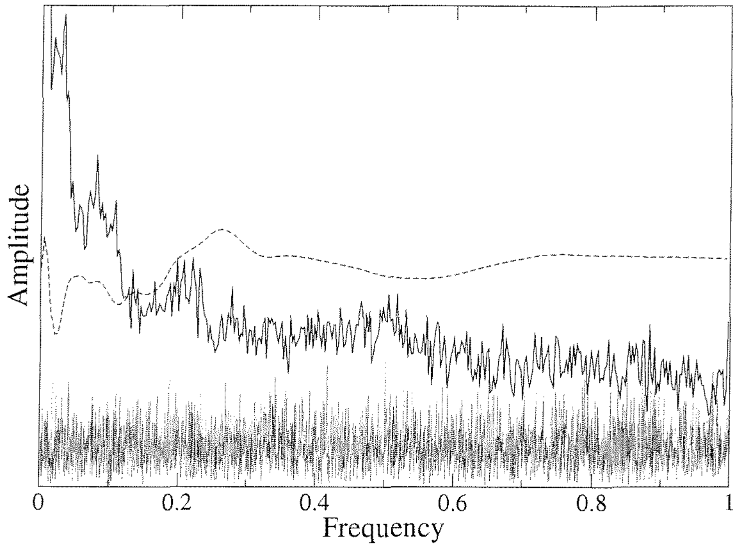

Figure 5.19: Amplitude versus frequency graphs for white noise. The HHT is drawn with a solid line, the Morlet wavelet with a dashed line and the Fourier transform with a dotted line. A distinct trend is evident in the HHT spectrum.

below.

5.7.1 Removal by Direct Fitting

In many cases it would be useful to remove this trend in such a way as to match the spectral properties of the Fourier transform of white noise. This would aid direct comparison with Fourier transforms, particularly in the low frequency regions of spectra where the distortion is most significant. To this end, a set of 1000 white noise HHT marginal spectra were obtained and a number of functions were fitted to them. The function with the best fit was of the form

$$R(f) = Ae^{Bf} + Cf + D \tag{5.28}$$

and the optimum parameters are as shown in Table 5.1. The mean of the noise spectra, along with the line of best fit and the mean residual is shown in Figure 5.20. Although the mean values for the curve parameters are accurate, as defined by their standard errors, the standard deviations observed are very broad: any particular spectrum is likely to have a mean curve significantly different from that drawn in Figure 5.20. Also, the residual data contains significant bumpy artefacts. Nevertheless, dividing the spectrum by the fitted curve obtained will

Parameter	Mean	Standard Deviation	Standard Error on Mean
A	8.33	3.65	0.115
B	-128	125	3.96
C	-1.08	0.179	0.00565
D	1.41	0.118	0.00372

Table 5.1: Parameters obtained for Equation 5.28 by fitting to white noise HHT marginal spectra. The spectra were fitted independently, and the above statistics apply to the set of parameters obtained. 1000 spectra were used.

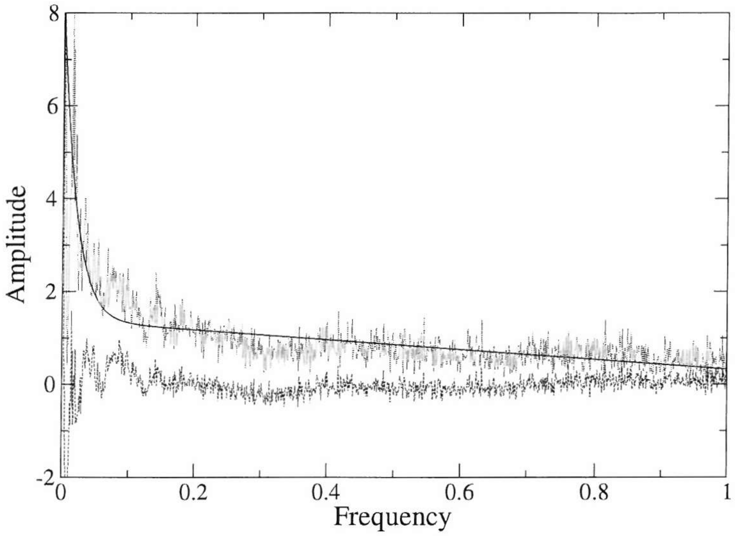

Figure 5.20: White noise model fitting. The mean spectrum of white noise is drawn with a dotted line, the mean best fit of model A is drawn with a solid line, and the mean residual is drawn with a dashed line.

produce a result somewhat closer to a flat spectrum for white noise, more similar to the Fourier result. An example of this for a single HHT marginal spectrum of noise is shown in Figure 5.21. While the trend in this spectrum has been reduced, it is still present: higher frequency elements appear to have larger amplitudes than they should. In view of the poor statistical match between this model and the data, and the consequent performance problems observed, the use of this model cannot be recommended. If the trend problem is to be ameliorated, a more detailed analysis of its causes must be undertaken.

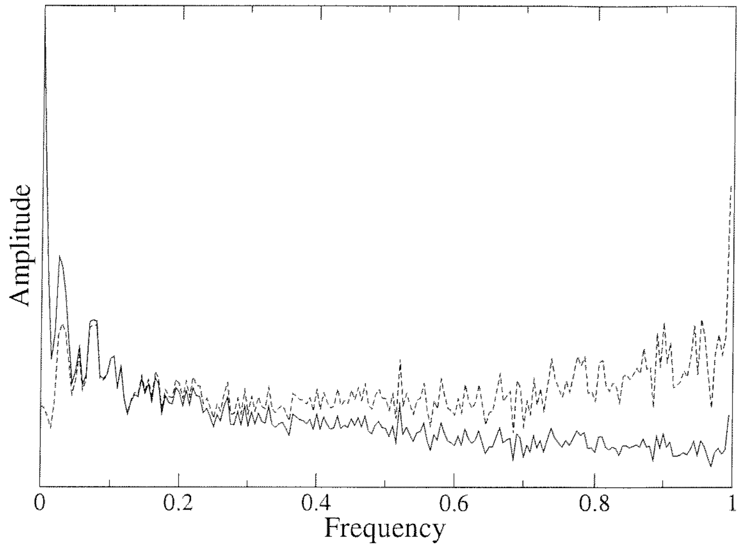

Figure 5.21: Amplitude versus frequency graphs for white noise. The solid line shows the unmodified spectrum and the dashed line shows the spectrum divided by the function from Equation 5.28. The latter exhibits somewhat less of a trend, but still leaves room for improvement.

5.7.2 Averaged HHT Marginal Spectra

Imagine that the IMF were a *continuous* function of time extending from T_{min} to T_{max}, whose amplitude was given by $a(t)$ and frequency $F(t)$. Consider some small region of frequency space, Δf, that the IMF passes through N times, entering at T_i and exiting at T_o for each of the N (if $F(T_{min})$ or $F(T_{max})$ are within Δf, then they count as entry or exit points, respectively). If we wished to calculate a mean amplitude for the region Δf, we would first evaluate

$$W = \sum_N \int_{T_i}^{T_o} a(t)dt \tag{5.29}$$

where the sum is over the N periods where the IMF passed through Δf, and W is a product with the units of amplitude *multiplied by time*. To obtain the amplitude mean, we must now divide by the sum of the time periods where $F(t)$ passed within the boundaries of Δf

$$M = \frac{W}{\sum_N (T_o - T_i)} \tag{5.30}$$

If we wished to apply this to discontinuous data, where $a(t)$ and $F(t)$ are vectors of discrete points (obtained from the data), we should use

$$W = \sum_{N} \sum_{t=T_i}^{T_o} a(t) \tag{5.31}$$

where W is now an amplitude multiplied by the *number* of discrete points occuring within Δf. Equation 5.30 should then be applied, this time considering T_i and T_o to be the integral indices of the initial and final elements of the vector occuring within segment N. This will obtain a corrected mean amplitude value, M. Put simply, we are dividing the sum of amplitudes within a frequency bin by the number of points in that bin.

The conventional Hilbert marginal amplitude spectra as described by Huang *et al.*[62] is equivalent to Equation 5.31 applied on its own, and the results obtained from it are therefore weighted by the number of points occuring within each segment Δf. Their statement, "The marginal spectrum offers a measure of total amplitude (or energy) contribution from each frequency value. It represents the cumulated amplitude in a probabilistic sense", is correct as far as it goes, but failing to take into account the number of points at each frequency can sometimes produce confusing results, as will now be shown.

Conventional HHT frequency marginal spectra are affected by the sampling of frequency space. Two factors determine this sampling in the white noise spectrum. First, the mean frequencies of successive IMF components decrease *geometrically* because the *locally* orthogonal nature of consecutive IMFs imposes this: successive mean frequencies approximately halve. Second, the standard deviations of the IMF frequency distributions also decrease geometrically, and the interaction of these effects on the distribution creates the nonuniform point density shown in Figure 5.22. The combination of these two factors leads to the observed trend and lumpiness in the sampling of frequency space and, if we make no adjustment, this nonuniformity will be carried over into the frequency marginal spectrum.

If we correct the marginal spectrum by dividing through by the total number of points falling within each histogram bin according to Equation 5.30, we obtain the normalised amplitude distribution of the IMFs. Such a graph will be referred to as the *averaged HHT marginal*

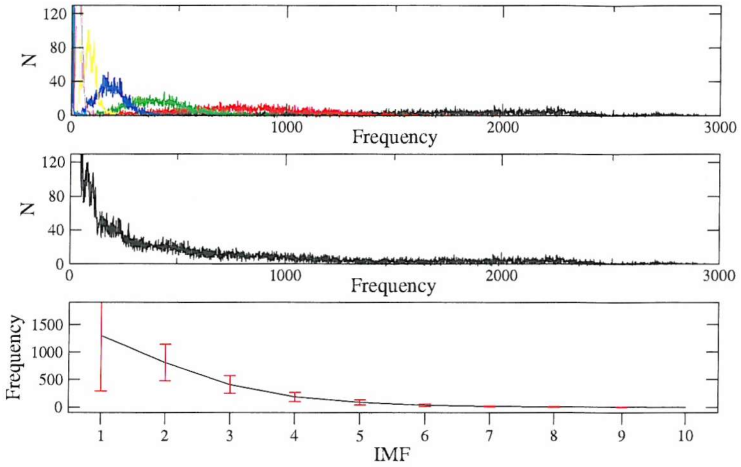

Figure 5.22: Top: histogram of occurence number versus frequency for the IMF components obtained from white noise. Each unit on the ordinate axis correspond to a point from an IMF having that frequency, making this a graph of point density versus frequency. The IMFs are shown separately in different colours, IMF 1 in black, IMF 2 in red, IMF 3 in green, etc. Middle: histogram of occurence summed over all IMF components. The lumps produced by individual IMF components are clearly visible. Bottom: the mean frequencies of the IMF components. The red bars show the standard deviations in frequency.

spectrum (AHM). The amplitude of the AHM spectrum of white noise falls off geometrically towards lower frequencies as shown in Figure 5.24. If we contrast this with the behaviour of conventional HHT marginal spectra, we see that the increase in the density of points towards lower frequencies at first nearly balances, but then wins out over, the attenuation in their amplitudes to create the observed overemphasis. The behaviour of the conventional marginal spectrum is thus rather difficult to interpret for broadband phenomena such as this, whereas the AHM spectrum removes the complicating factor of the underlying point density.

Does this amplitude attenuation make sense? Yes. A cursory examination of a white noise spectrum reveals that it is nearly flat. There exists no intuitive separation of scales: most of the amplitude appears to be present at high frequencies, and no lower frequency waves are evident from inspection. The IMF components extracted from white noise reflect this; they are plotted at constant scale in Figure 5.23. The trend component of white noise is by definition zero, and the observed reduction in amplitude with increasing IMF component number is reflected in the AHM by the attenuation toward zero at low frequencies. This reduction is *not* reflected in the conventional HHT marginal spectrum due to the influence of the point density.

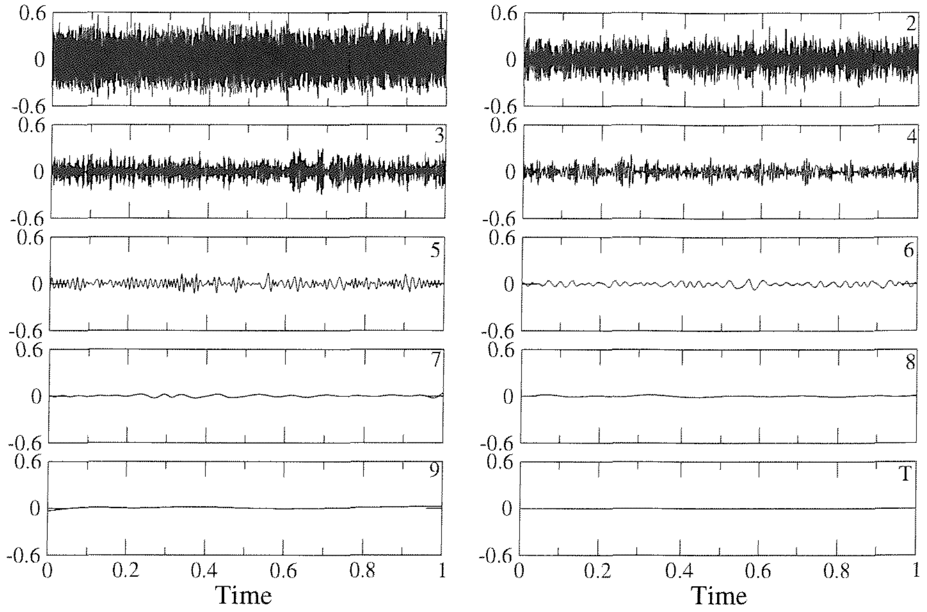

Figure 5.23: The IMF components obtained from white noise, plotted on identical scales. The trend component is labelled with a 'T'. Note the reduction in amplitude with increasing component number.

A curve was fitted to the white noise AHM spectrum of the form

$$y = Ax^B \tag{5.32}$$

where $A = 0.18884$ and $B = 0.468182$. The residual data shown are small and have no trend or pattern, save their obviously broader spread at higher frequencies caused by the relative reduction in point density; the signal consisted of 1000 points in all. This function describes the AHM spectrum of unit amplitude white noise where the frequency range is normalised to unity.

If a spectrum directly comparable with Fourier transforms is required for some arbitrary signal, then the observed AHM should be divided by the curve fitted to unit amplitude white noise given above. This will result in a spectrum whose amplitude values can be interpreted as multiples of the unit white noise spectrum. When this is done to a noise signal, a flat spectrum is obtained, as shown in Figure 5.24. The abrupt attenuation at very low frequency is a result of the large numerical errors attendant on dividing one very small number by another, and the increasing noise at higher frequencies is due to the sparser sampling.

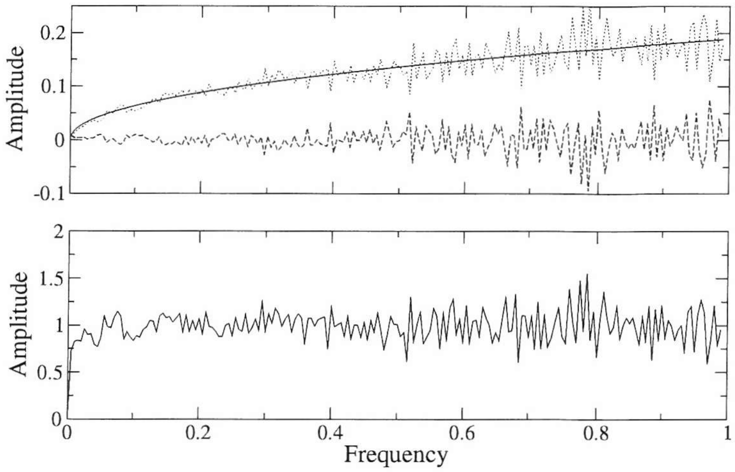

Figure 5.24: Top: the averaged HHT marginal spectrum for white noise is drawn with a dotted line. The best fit to Equation 5.32 is drawn with a solid line, and the residual from this fit is drawn with a dashed line. Bottom: spectrum of white noise flattened as described in the text to obtain direct comparability with the Fourier spectrum.

In summary, the conventional HHT frequency marginal spectrum of white noise was compared with the Fourier spectrum and found to be systematically different. Removing the observed trend component to obtain a more directly comparable spectrum would clearly be helpful in certain cases, and an initial attempt to do this by functional fitting was made. The results from this simple approach were found to be poor, necessitating a more sophisticated analysis. An averaged HHT marginal (AHM) spectrum, was devised to remove the artefacts caused by the HHT's strongly nonuniform sampling of frequency space. For an arbitrary signal, a spectrum whose amplitudes will be more directly comparable with the Fourier approach can be obtained by dividing its AHM by the mean AHM of white noise. This should be considered as a useful empirical tool for obtaining results more comparable with Fourier spectra than the conventional Hilbert marginal spectrum. The full implications of adopting this approach are still not fully understood, though. For this reason, conventional amplitude or signal energy marginal spectra will be used throughout the rest of this work. Although further analysis is required in this area, the initial results seem promising. Further work would include integrating the method with the other programs in the suite developed, testing it on a number of interesting signals to build confidence in its use, and doing detailed comparisons with the results of Fourier analysis.

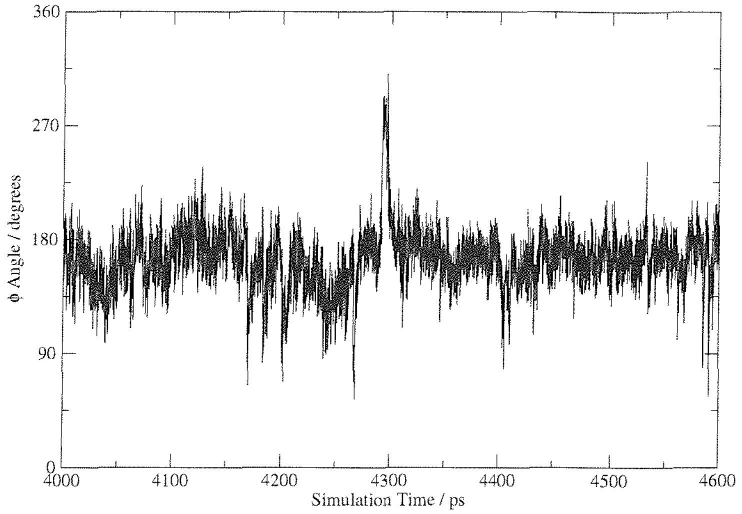

Figure 5.25: φ torsional angle of residue 119 of the **VN** prion protein simulation between 4.0 and 4.6 ns. The unstable conformational change occurs between 4291 and 4299 ps.

5.8 Simulation Torsions

Data from a protein MD trajectory was examined as a final test of the basic HHT and wavelet methods. The prion protein simulations were visualised and a degree of freedom exhibiting interesting behaviour was obtained for use as a test signal.

The φ torsional angle of residue 119 of the asparagine/valine prion protein simulation displays a brief change event close to the 4.3 ns point. The angle, elsewhere stable at around 160° abruptly changes to 280° for 8 ps. A portion of the trajectory between 4.0 and 4.6 ns in which this event occurs is displayed in Figure 5.25; the instability can be seen near the centre of the signal.

Time margins of this part of the trajectory were obtained with both the wavelet and HHT methods. The maximum frequency limit was constrained by the Morlet wavelet, and the minimum frequency by the HHT. According to these constraints, the range between 0.6 and 17.4 cm^{-1} was investigated, along with a slightly lower frequency band between 0.6 and 10.0 cm^{-1}. Both amplitude and signal energy data were obtained. The results for this are shown in Figure 5.26.

The first observation to make from this data is that the wavelet and HHT results appear remarkably similar. The amplitude margins produce much broader peaks than the signal en-

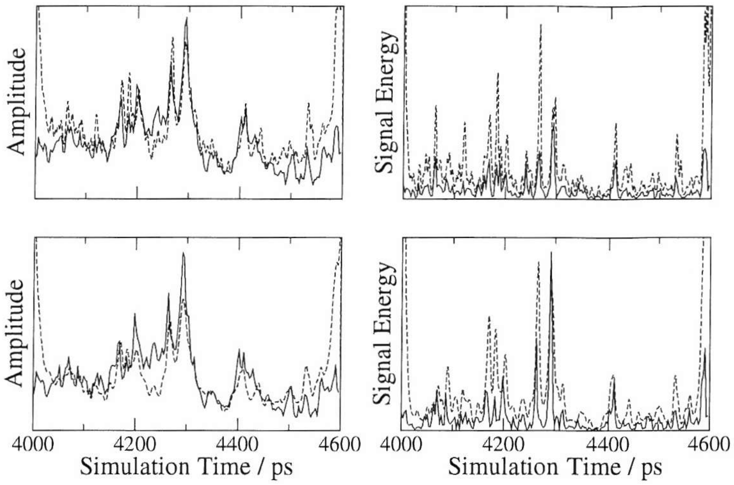

Figure 5.26: Time margins for the HHT and wavelet transforms of the prion torsional angle shown in Figure 5.25. The top pair of graphs cover the complete frequency range, 0.6–17.4 cm^{-1}, and the bottom pair the reduced range, 0.6–10.0 cm^{-1}. The left pair show the amplitude versus time and the right pair signal energy versus time. In all graphs, the HHT data are drawn with solid lines and the wavelet data with dashed lines.

ergy as expected, the latter emphasising the higher frequency parts of the region analysed. The lower frequency margins are generally very similar to those covering the complete range, although the difference in frequency bands in this case is not particularly large.

A principal question to be answered is: which of the margins is best able to discriminate the event of interest from other, less interesting noise-like elements in the signal. This test case is a useful one because a second, very short, event is registered in all spectra at around the 4.26 ns point, shortly before the unstable state transition. The differences in peak amplitudes measured for these two events indicates the relative success of the different techniques and strategies.

Where amplitude is being measured, the wavelet margins register almost identical peak heights. Where signal energy is measured, though, the first peak records a *greater* value than the second, which might be expected if its amplitude is similar and its frequency is higher (shorter time period) as would appear to be the case from the signal. For the Hilbert margins, the second peak is noticeably higher than the first in all cases, and generally much higher than the other, less interesting features of the data. Looking more carefully, the pair of amplitude margins have a similar, smaller height ratio between the peaks than the signal energy margins.

Considering just the latter, the low frequency margin shows the most dramatic difference, both between the peak heights, and also between the interesting peak and the other features of the time margin.

To summarise, the HHT and wavelet time margins have been compared and found to show a similar ability to spot the interesting events in a torsional angle signal from a real condensed phase protein simulation. HHT time margins showed the greatest ability to discriminate the interesting state change exhibited by this signal from the other, less interesting features. The best approach to applying the HHT was to plot the signal energy over the very low frequency, 0.6–10 cm^{-1} range. The use of signal energy has the effect of highlighting higher frequency and higher amplitude features.

5.9 Discussion

Biomolecules in solution exhibit many complex modes of motion; some of these can usefully be thought of in terms of vibrations, which may or may not be well described by the simple harmonic approximation. Other motions are more difficult to characterise, typically thought of as 'snap' events where the system crosses between wells on the energy surface; excess energy present in the relevant degrees of freedom immediately afterwards is then dissipated throughout the rest of the system. These 'snap' motions may be complicated and involve many residues. On crossing from one potential minimum to another, the frequencies of the vibrational modes of motion will change; furthermore, if the motions are not simple harmonic, their patterns of motion will be modified. For the chemist interested in characterising and predicting the behaviour of biomolecules, the vibrational motions present in the system encode a wealth of information. Although it may be difficult to probe the individual motions within such complicated systems experimentally, atomic level simulations can provide a means of extracting useful vibrational information. The problem then becomes one of interpretation: how is useful knowledge to be extracted from the motional data?

Recent studies with the DFMD and RDFMD methods have shown that pumping energy into low frequency motions increases the probability of observing conformational change events. To achieve this while affecting the rest of the simulation dynamics as little as possible, the smallest necessary amount of energy should be added. For this to be effective, the frequency band over which motion is enhanced must be as narrow as possible, and this necessitates accurate targetting: we must discover which frequencies of motion are involved with conformational changes. Furthermore, the degrees of freedom to which this energy should be applied should be quantified in some way. Ideally, we should examine a number of realistic simulations of typical protein systems in which conformational change events take place without any artificial stimulation. The types of change event observed, the frequency bands that show enhanced amplitude motions, and the degrees of freedom, secondary structure elements, etc. affected should be correlated. This would enable the more rational targetting of conformational change enhancing methods, such as DFMD and RDFMD.

The process most commonly used for obtaining an amplitude-frequency profile for a system is the Fourier transform, and where time information is also required, the wavelet transform. Where the signal being examined is a poor match for the basis function used, these transforms will still describe the system exactly, building up a description of the data from large numbers of 'harmonic' components. Although these harmonics constitute a complete description of the information content of the signal, they can be difficult to interpret. Huang *et al.*[62] have proposed that a better alternative to the harmonic description of complicated signals is the *intrawave modulation* concept. They also proposed a method to make this a practical possibility for arbitrary signal waveforms, the Hilbert Huang transform (HHT). This chapter has examined the properties of the HHT.

An important thing to remember about the HHT is that it is still *very* new. Publications on its methodology and applications, although coming from many different disciplines, are still few in number. There is not yet the large body of experience that exists for the Fourier and wavelet methods. A second point to keep in mind is that the HHT depends upon the *empirical* mode decomposition procedure to work. This empirical nature makes analytical approaches to its characterisation very difficult, as others have remarked.[83] To determine the properties of the HHT for particular classes of application, the most straightforward approach would seem to be just to try it.

The implementation of the method was more or less straightforward, with issues arising with the treatment of the end effects: the description in Huang *et al.*[62] is slightly vague. Creating an implementation of the codes that worked at a baseline level was rapid, although the subsequent detailed testing and adjustment of the codes required considerable time and patience. Performance is not a major issue—the mathematics used are simple and can make use of well established, highly optimised library functions for spline fitting, Fourier transforms and differentiation. The numerical stability of EMD was found to be reasonable: once the treatment of the end-points of the signal is correctly adjusted, the propensity of the algorithm to 'run away', iterating forever, appears negligible. Errors can occur where a low amplitude wave occurs between a pair of high amplitude ones, but the consequences of this seem unlikely to be of concern in MD simulation analysis, and the resulting artefacts are easily removed from spectra. If these errors were to become a concern, they could probably be ameliorated or removed by improving the stopping criteria for the EMD seeking algorithm.

Stability of the solution set of IMFs obtained by EMD from a given signal was found to be low. High sensitivity to minor changes in input data was observed; tiny changes to a small part of the signal can change the solution obtained by EMD globally. Solutions for a particular dataset tend to form clusters: many very similar solutions are located near one another with empty gaps between. The conjecture of Huang *et al.*[62] conjecture that a spline fitting procedure which obtains a tighter fit to the data will improve the stability of the algorithm was checked with the Akima spline. No improvement was observed, either to the global stability against local changes, or to the overall numerical stability where large numbers of seeking cycles are required to achieve reasonable separation of scales.

Problems occur with HHT spectra towards the edges of the signal in the time domain. A solution suggested by Huang *et al.*[62] was to extend the dataset by synthesising extra points at the edges of the dataset and attenuating them towards zero. Unfortunately, the details of how to do this were unclear. As the situation with Fourier windowing is complicated, with intuitively obvious functions producing far from optimal results, it was decided that a number of functions should be tested and the results compared. A linear attenuation was very nearly optimal, with a function very slightly modified from the linear outperforming it by a narrow margin in the tests performed. If the analyst is unwilling to make up extra data at the edges of the signal to be examined, but wishes to know the precision of the points in the spectra obtained, a clipping strategy must be used. The points at which clipping must take place for a given level of precision can be described by a simple function for single frequency, constant amplitude sinusoidal signals. Such a function was parameterised. For the points in a spectrum to have a precision equal to the resolution stated by Huang *et al.*[62] typically requires a substantial fraction of the data at the edges of the signal to be removed. A middle way between the two processing strategies would be to attenuate the edges of the data with a well chosen function, without synthesising extra datapoints. The amount of data that must be excised from the spectrum's edges would then be much reduced.

For simple signals, such as the preliminary sinusoidal test cases employed to check the implementation of the method, the interpretation of results is trivial. In more complicated cases some work is required on the part of the analyst, though. EMD will separate any signal into a series of scale-based components. All, some, or none of these may have a clear physical meaning, and a judgement on this must be made on a signal by signal basis. Where a direct physical interpretation is possible—where a clear, well-separated scale property exists—the instantaneous frequency values obtained from the Hilbert transform should have a straightforward interpretation. Where no such meaningful scale separation exists, difficulties with interpretation begin.

Simulations of a simple Brownian dynamics test system show that the method is able to produce results comparable with the wavelet and Fourier methods of signal analysis for complicated datasets. EMD can obtain two *sets* of scale components for a case where a nonstationary transition event occurs; one set sums to give the noise-like and harmonic-like behaviours of the system, and the other set sums to give the underlying path the trajectory has taken. In this case EMD has acted as an empirical scale filter. It should be borne in mind that there exists no useful separation of scales within the low-frequency set: the individual components are physically meaningless, only obtaining the correct behaviour on summing. Naive use of EMD as a filtering algorithm should therefore be avoided—although this may be possible for some applications, significant checking and monitoring of IMF properties would be required. Peak centre frequencies obtained from the HHT frequency marginal spectra show a small improvement over the Fourier and marginal wavelet spectra, the harmonic leakage of energy across the frequency domain causing problems with the latter. The HHT time margin showed better ability to discriminate transition events in the Brownian dynamics simulations than the Morlet wavelet.

The question of interpreting the low frequency components obtained where transition events take place is important. It seems that although such intrinsically unwavelike phenomena can have no meaningful *frequency* value assigned to them, their low frequency signatures in both wavelet and HHT spectra provides a useful indicator of their appearence.

Regions of the time-frequency portrait that are poorly represented by the IMFs obtained for a signal are difficult to interpret using HHT. Wavelet analysis presents an advantage in such cases, due to the more intuitively complete spectra obtainable. A modification to the HHT algorithm whereby large numbers of slightly modified signals are transformed and the results composited to yield a more complete portrait of time-frequency space is proposed. The resulting spectra seem better in line with intuition and visually closer to those obtained by the wavelet approach; the lack of harmonics resulting in clearer results in the cases studied. The change to regions of the spectrum well sampled by IMFs is negligible, while the poorly sampled regions are significantly better: clearer to view and with improved sampling and hence better marginal spectra. The drawbacks of this technique are the judgement needed about how much noise to add to each signal, and the increased computing time required.

In view of the intuitive similarity between HHT, Fourier and wavelet results, the differences between their white noise spectra are at first sight surprising. In certain cases, directly comparable spectra would be useful, and to obtain these the simplest approach would be to fit a function to the frequency marginal spectrum of white noise and divide the spectrum of the phenomenon to be analysed by this, reweighting it. Attempts to fit various functions in this manner met with little success: the spectra from the noise were simply too variable to be reliably parameterised. A more sophisticated analytic approach was then taken, and a new type of spectrum, the average HHT marginal, was created as a means of removing the effects of the nonuniform sampling of frequency space in the conventional HHT. Fitting a function to this AHM spectrum is much easier, and the results obtained look very promising as a means of adjusting HHT results to be better comparable with the Fourier transform. Further work is still required to appreciate the broader effects of doing this, though.

A set of torsional data was obtained from one of the prion protein simulations that exhibited a short, unstable transition event. This was analysed with the wavelet and HHT methods, and similar results obtained from both. The HHT time marginal spectrum showed a very slight advantage above the wavelet spectrum in this case. The best results were obtained where the HHT time marginal signal energy spectrum was obtained for a very low frequency band.

To conclude, the HHT seems to be a useful method for the time-frequency-amplitude decomposition of signals. The spectra appear broadly comparable with those obtained by conventional Morlet wavelet and Fourier transform techniques, although closer examination reveals that differences do exist. It may be possible to lessen these with further work. The method was found to be easy to implement and work with and rapid enough for large numbers of spectra to be processed in a few minutes on modern hardware. The results obtained for signals of the sort that are of interest to chemical simulators are mostly straightforward to interpret, and where difficulties exist methods such as the spectrum compilation technique may help. For the processing of simulation data to extract discrete change events the HHT and wavelet methods appear about evenly matched, in spite of their very different modes of operation. The lack of harmonic leakage present in the former may confer significant advantages if large numbers of signals need to be processed statistically by an automated procedure, though. This will be the subject of the following chapter.

Chapter 6

Simulation Statistics

6.1 Background

6.1.1 Simulation Statistics

Many of the methods of analysing molecular dynamics simulations are concerned with rendering visible to the analyst aspects of the simulated behaviour that would otherwise escape notice. There are two parts to this problem. First, developing means of abstracting well defined facets of the dataset, and second, reducing the volume of abstracted data thus obtained to manageable proportions and representing it in a form intelligible to a human being.

The wavelet and HHT transforms are both solutions to the first part of the problem as outlined above; signals obtained from the trajectory can be processed with them to obtain an interpretable abstract representation of the properties of the signal in time and frequency space. Examples of the wavelet transform doing this are shown for a pair of hydrogen bonds obtained from the ***F1*** trajectory (the nature of the different trajectories is described in Subsection 6.2.1) in Figures 6.1 and 6.2. The first hydrogen bond is between the backbone amide hydrogen of serine 3 and the backbone carbonyl of isoleucine 91; the second bond is between the backbone amide of leucine 4 and the backbone carbonyl of lysine 109. More energy is seen at the higher end of the frequency spectrum in the second bond. Unfortunately, to compare and contrast *all* of the hundreds of hydrogen bonds present at some stage in the trajectories with this method

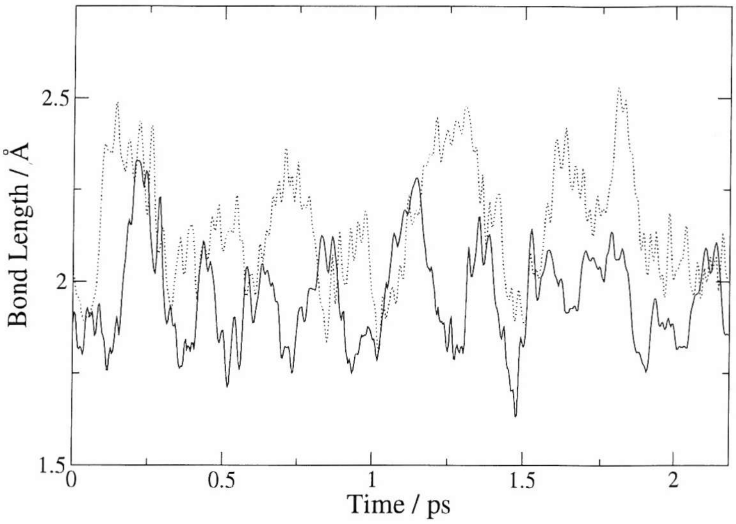

Figure 6.1: Lengths of the first and second hydrogen bonds in the ***F1*** trajectory; bond 1 is drawn with a solid line and bond 2 with a dotted line.

would be impossible, as the HHT and wavelet methods do not solve the second part of the problem described above: the hundreds of time-frequency-amplitude graphs simply contain far too much information for a human analyst to make sense of. For these methods to be more useful in analysing and comparing protein trajectories, some means must be found to reduce the data obtained to manageable quantities and represent it in an interpretable fashion. This chapter describes methods by which this can be achieved, along with some results obtained with them.

Of particular interest is the development of methods capable of spotting conformational transition events, and making sense of them; a secondary aim is to discover ways of checking and parameterising the recently described digital filtering methods for promoting such change events.

6.1.2 Dihydrofolate Reductase Simulations

To test these methods on a realistic scale a protein system was required that undergoes clear conformational change events on timescales accessible to molecular dynamics. A set of previously existing simulations of the dihydrofolate reductase (DHFR) enzyme were chosen to meet this requirement. To stimulate transitions between conformers, the 'reversible digitally

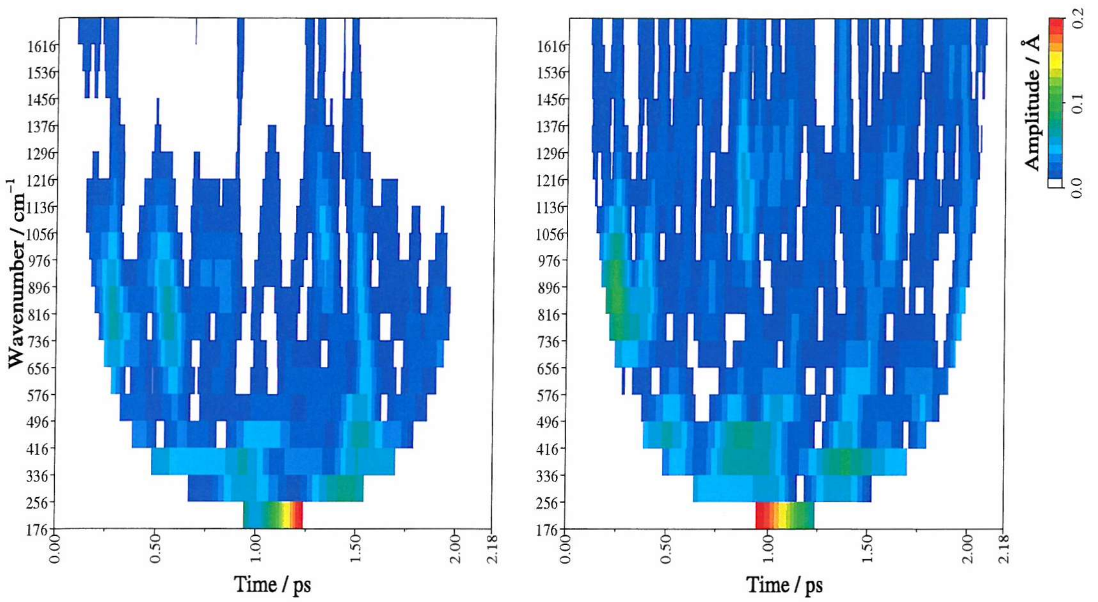

Figure 6.2: Wavelet transforms of the first and second hydrogen bond lengths in the ***F1*** trajectory. Note the difference in energy at higher frequencies. Higher amplitude elements can be seen in the lowest frequency band in both plots, although the clipping makes it difficult to see them clearly.

filtered molecular dynamics' procedure described in Subsection 6.1.3 was employed. More than one type of conformational change event was observed in the simulations, consistent with experimental evidence.

Many enzymes require additional factors to perform chemistry beyond the range of the 20 amino acid sidechains.[1] These are called *coenzymes*, and are often structurally related to vitamins—molecules required by an organism that it cannot itself manufacture and so must obtain from its environment. An example of this is the coenzyme tetrahydrofolate (THF), which is synthesised from the folic acid that human beings must obtain from their diet. THF acts as an intermediate carrier of single carbons, and is involved in the biosynthesis of purines and a number of amino acids.

Folic acid is converted to THF in a two step reduction via the intermediate dihydrofolate (DHF), both steps catalysed by the enzyme DHFR which is present in all known organisms. The second step of this reduction has a higher rate constant, and is biologically important for recycling DHF to THF and maintaining a cellular reservoir of the latter. One of the uses of THF is in the synthesis of thymidylate, a DNA precursor.[91] This role of THF in DNA synthesis makes DHFR a target for the treatment of cancer, among other things, and large amounts of work have been done in this area. Mechanistic studies of the enzyme began in the 1950s and are still proceeding; consequently there is a large and growing body of relevant experimental knowledge.[91]

The *Escherichia Coli* form of DHFR is a globular protein of 159 residues comprised of an eight stranded β-sheet with four α-helices; this is shown in Figure 6.3 (the figure was produced using the VMD program[11]). There are two subdomains, the relative rotation of which has been implicated in catalytic activity. The first subdomain comprises strands **D**, **C**, **B** and **E**, along with helices **C**, **E** and **F**. The adenosine part of NADPH is known to bind to this subdomain. The rest of the molecule, comprising strands **A**, **F**, **H** and **G** and helix **B**, is termed the loop subdomain, as it contains the three loop sections implicated in the catalytic activity of the enzyme, **M20**, **F-G** and **G-H**. Movement between subdomains can occur by rearrangement of the hydrogen bonds between helices **A** and **E**. The nicotinamide part of NADPH ordinarily resides in this position, while DHF sits between helix **C** and helix **B**.

X-ray crystallographic studies[23] have shown that the three loop regions, **M20**, **F-G** and **G-H**, exhibit a degree of conformational plasticity. A number of hydrogen bonds are formed between them and also between the **M20** loop and residues in helix **C**, on the opposite side of the active site. The physical arrangement and hydrogen bonding pattern of the three loops changes over the catalytic cycle of the enzyme. In the first three states along the reaction pathway, the enzyme adopts a *closed* conformation, where hydrogen bonds exist between the **M20** loop and helix **C**. In the last three states an *occluded* conformation is favoured, in which the loop has rotated away from the helix, breaking the hydrogen bonds between them. A third *open* state exists through which it is believed the enzyme may pass on changing between the *closed* and *occluded* states. The closed conformation, along with a structure observed towards the end of one of the DHFR simulations similar to the occluded conformation, are shown in Figure 6.4 (this figure was produced using the VMD program[11]). Also shown is one of the key hydrogen bonds that connects the **M20** loop to helix **C** in the closed conformation; this

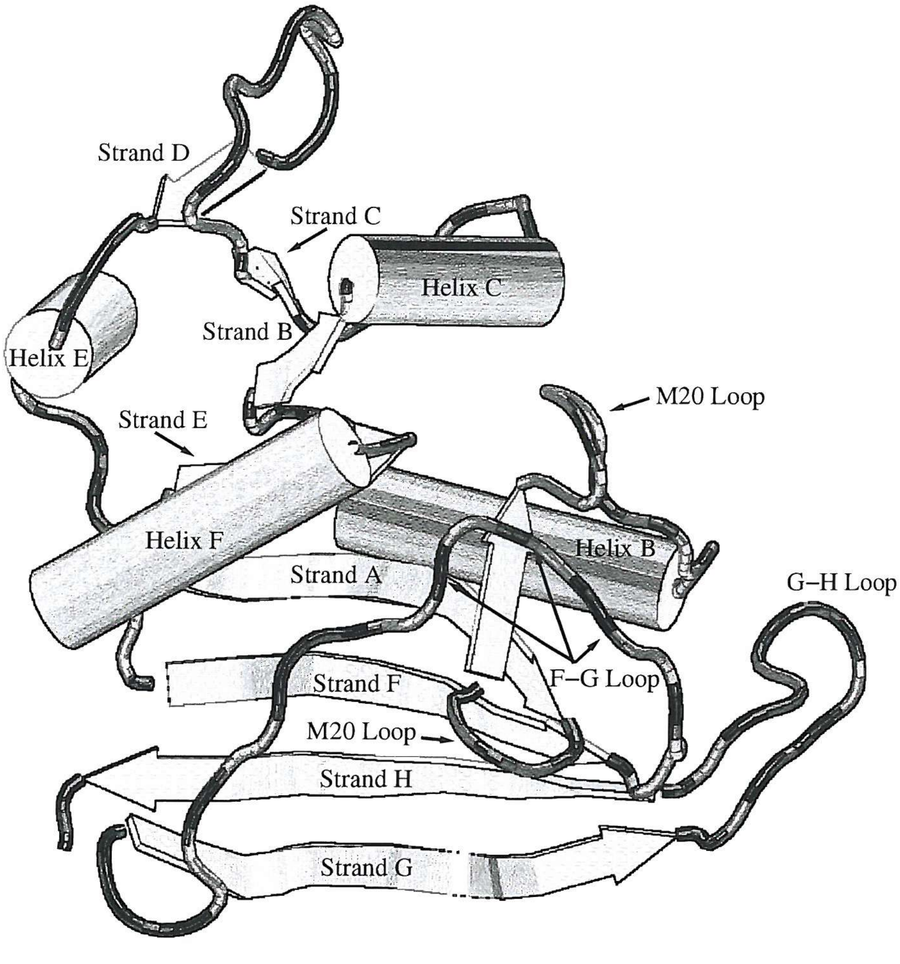

Figure 6.3: Dihydrofolate reductase (DHFR) with parts labelled.

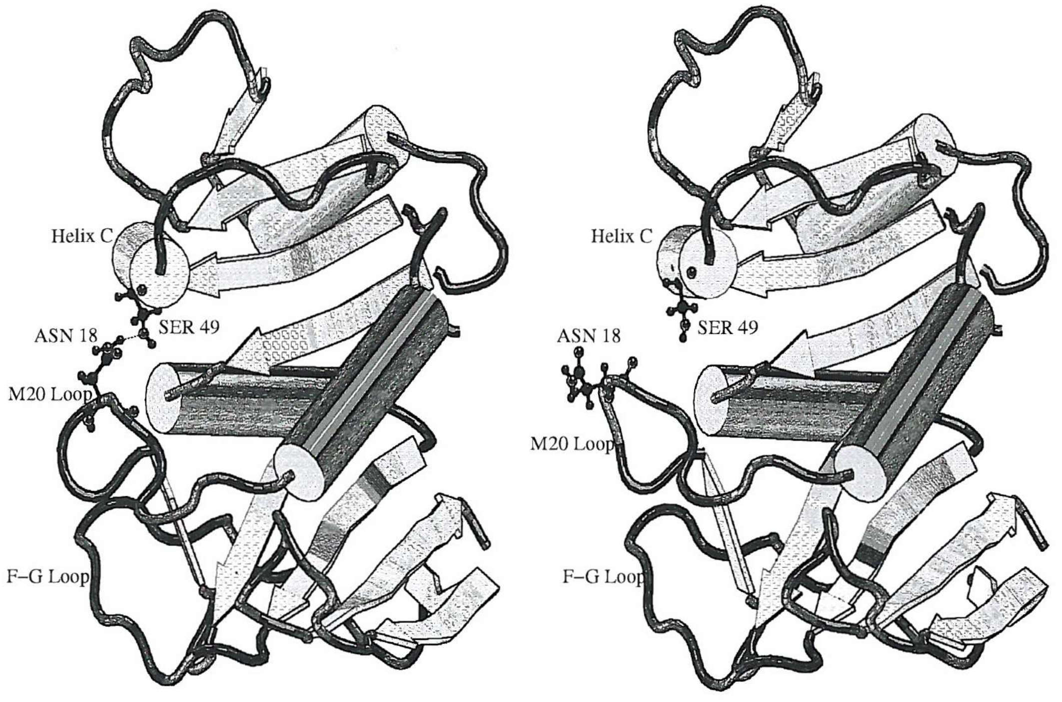

Figure 6.4: Structures from the beginning and end of the ***F2*** DHFR simulation (as defined in Subsection 6.2.1). The Asn 18 - Ser 49 hydrogen bond is present in the closed conformation of the protein seen in the first half of the trajectory (left) and absent in the structure similar to the occluded conformation seen in the second half (right).

is formed between asparagine 18 and serine 49, and is broken on transition to the occluded conformation.

6.1.3 RDFMD

The constraints imposed on the lengths of protein molecular dynamics simulations by the equipment currently affordable within an academic budget are such that the chances of observing conformational change events within such simulations is quite low. A recent approach to this problem is to use digital filters to impart energy to the molecule in a frequency specific manner. These methods have been shown[60, 61] to be effective in stimulating transition events. The *reversible digitally filtered molecular dynamics* (RDFMD)[61] method was used in the preparation of three of the four DHFR simulations used in this chapter.

RDFMD works in a stepwise manner as follows. First, a short section of molecular dy-

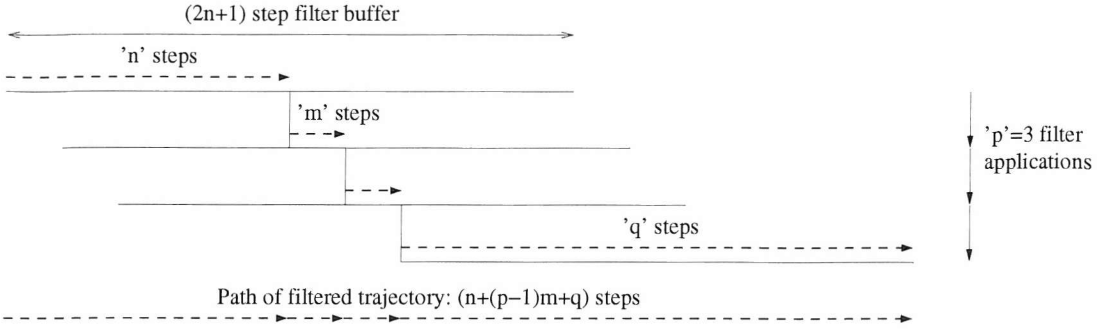

Figure 6.5: Illustration of reversible digitally filtered molecular dynamics (RDFMD). The path through the series of filter buffers from which the final simulation trajectory is constructed is indicated by the dashed arrow.

namics is run in the normal fashion, with coordinate and velocity information stored at each step; this data is referred to as the *buffer*, and consists of $2n+1$ steps, where n is a parameter of the process. The cartesian velocities of a predetermined subset of the atoms are then processed with a symmetrical, centre tapped *finite impulse response* (FIR) digital filter to obtain a new set of velocities for the MD timestep at the centrepoint of the buffer. The simulation is then restarted from this point, again with the velocity information being recorded into a buffer. A simulation section of $n+m$ steps is then collected, starting from the restart point, along with a section of $n-m$ steps integrated *in reverse* from the restart point; m is a second parameter. The filtering step is then reapplied to the new buffer, and the process is repeated as shown in Figure 6.5. Once the notional temperature of the system reaches a predetermined level, the system is allowed to evolve under normal molecular dynamics for a further q steps. This temperature value is not a meaningful temperature, as the energy is introduced in a frequency specific manner, but does serve as an indication of how much energy has been added to the system. The resultant simulation path thus contains $(n+(p-1)m+q)$ steps, where p is the number of filter pulses applied.

Typically, the digital filter is designed to amplify low frequency components of the trajectory, leaving higher frequency motions unaffected. The RDFMD procedure is designed to amplify the velocities in a frequency specific fashion as gently as possible; the absolute minimum of energy is introduced to the system to achieve the required effect.

6.2 Methods

6.2.1 Simulation

A series of simulations of DHFR performed previously were analysed with the new methods. Four simulations were chosen. A 4 ps control simulation where no digital filtering was employed (***UC***), along with a 1.904 ps simulation where RDFMD was used but no clear conformational transition took place (***FC***), a 2.184 ps simulation with RDFMD where a conformational disturbance took place, with broad activity in the **M20**, **F-G** and **G-H** loops (***F1***), and finally a 1.984 ps RDFMD simulation in which a transition took place to a structure similar to the physically observed *occluded* conformation (***F2***).

All of these simulations started out from an X-ray crystal structure obtained by Sawaya and Kraut[23] of the *E. Coli* form of the enzyme in its closed conformation. This structure has the PDB reference code 1RX2, and was crystallised with both NADP$^+$ and folate bound, acting as analogues for NADPH and DHF respectively. The analogues were removed during system preparation and replaced with water molecules. The system was solvated to a distance of approximately 10 Å using 8483 TIP3P water molecules.

All-atom simulations were performed with version 2.4 of the NAMD package modified for RDFMD. The CHARMM27 force field was used;[92,93] SHAKE[14] constraints were placed on hydrogen atoms; electrostatic interactions were treated with the particle mesh Ewald method and a switching function was employed for Lennard-Jones interactions, switching at 9 Å with a cutoff at 10.5 Å; dynamics were integrated with a timestep of 2 fs. The simulations were equilibrated in the constant volume, constant temperature ensemble at 298 K, using a Langevin dynamics thermostat.

The FIR filter was designed to amplify frequencies below 50 cm^{-1} by a factor of 4.0 on each amplification cycle, and consisted of 751 coefficients—$n = 375$ in terms of the description in Subsection 6.1.3. The period between successive amplification pulses, m, was 20 timesteps (40 fs). Amplification was repeated until the temperature of the filtered atoms reached 1500 K, although this figure should be treated with caution as the excess energy was present mainly at low frequencies. Typically around 10 to 15 pulses is expected for this condition to be satisfied. After the filter applications, the system was allowed to evolve for 396 timesteps. In summary, the trajectory data collected from the filtered simulations covered $376 + ((p - 1) * 20) + 396$ timesteps, where p is the number of filter applications required for the temperature of the filtered atoms to reach 1500 K. All atoms of residues 14–19 were selected for filtering (the **M20** loop region).

6.2.2 Analysis

The analysis procedure is best considered in two stages. In the first stage, a large volume of data was derived from the MD trajectories and stored in files. In the second stage, a variety of graph types were extracted from this data. To recap the objectives of this statistical work, a simplified representation of the information present in protein trajectories is desired, and to achieve this the very large quantity of information present in such trajectories must be 'boiled down' to a human-intelligible volume.

Data Preparation

Four signal sets were extracted from each simulation trajectory: the cartesian coordinates of all heavy (non-hydrogen) atoms, the cartesian coordinates of just the backbone atoms, the ϕ and ψ torsional angles of the protein backbone, and the lengths of a subset of the observed hydrogen bonds. The set of hydrogen bonds to examine was obtained by processing each of the four trajectories with DSSP (as implemented in Molmol[10]), selecting from each the set of hydrogen bonds existing in at least 10% of the simulation frames, and merging the lists.

For each simulation, the frequency ranges reachable by the HHT and the wavelet transform were then calculated, and a frequency resolution for the wavelet transform was decided. Just to recap, the wavelet transform obtains the amplitudes of single, specific frequencies, whereas the HHT works by a binning procedure—the sizes of those bins must be specified. To limit the volume of data storage required by the wavelets, a relatively small number of individual frequencies was desired. The maximum frequency values were limited by the wavelet transform to around 1600 cm^{-1}, and if twenty frequency values were used the frequencies would be spaced by 80 cm^{-1}. The minimum frequencies for both transforms are determined by the length of the shortest simulation, ***FC***. For the HHT this was 16 cm^{-1}, and this value was indeed used as a base frequency for the corresponding plots. Owing to the masking required to remove regions of the wavelet transforms affected by the dataset endpoints, only a single datapoint is accessible at the theoretical minimum frequency of around 80 cm^{-1}, and using this as a base would yield unhelpful statistics. To obtain better sampling, the minimum frequency used was shifted up to around 170 cm^{-1}, and to aid comparison of HHT and wavelet graphs, a figure of 176 $\mathrm{cm}^{-1}$was chosen; the maximum frequency was shifted up by a similar amount to maintain the round-number spacing of 80 cm^{-1}. In summary then, the minimum frequency chosen for the wavelet transform was 176 cm^{-1}, the maximum frequency was 1616 cm^{-1}, and a spacing of 80 cm^{-1}, lead to 19 frequencies being sampled: 176, 256, 336, . . . , 1616 cm^{-1}. The minimum frequency in the HHT plots was 16 cm^{-1}, and for the statistical data where a discrete number of frequency bands were sampled, the spacing was kept the same as the wavelet plots at 80 cm^{-1}. This resulted in 21 bins: 16–96, 96–176, 176–256, . . . , 1616–1696 cm^{-1}.

In the first step of the processing, EMD was performed on each signal in each signal set, the Hilbert transform was computed on each IMF obtained, and the frequencies and amplitudes of each IMF at each timepoint were written to file. Thus was obtained the exact Hilbert Huang transform of every signal (i.e. each hydrogen bond length) for the four signal sets (heavy atoms, backbone atoms, torsional angles and hydrogen bond lengths). Typically 5–7 IMFs were obtained for each signal, each IMF requiring two vectors of data be written to file (amplitude and frequency), each vector containing a datapoint for every timestep in the signal. Points where the boundary-effect errors become significant were masked to zero. Similarly, the Morlet wavelet transform of each signal was calculated, and the amplitude as a function of time and frequency was written to file. Each signal required 19 vectors to be written to file, one for each frequency examined; each vector containing a point for every timestep in the signal. Points where the signal boundaries would cause errors were masked to zero as described in Chapter 5.

In the second step of the processing, a set of base statistics were calculated from the data obtained from the first step. Considering first the wavelet transforms, the mean and standard deviation for each frequency of each signal were determined. Only points not subject to boundary error masking were included in the calculation. Each signal therefore required two vectors of 19 elements be stored for it: one vector of mean amplitudes, and one vector of the standard deviations of those amplitudes, one datapoint for each frequency. The HHT data were sorted into frequency bins according to the scheme given above, before the means and standard deviations of the amplitudes of the points falling into each bin were calculated. Again, points subject to boundary error masking were excluded from the calculation.

In the third step of the processing, the deviations of all datapoints from a set of reference statistics were determined, along with a record of the highest deviation obtained at any frequency for each timepoint. For the ***UC*** control trajectory, the only reference used was itself; for the filtered simulations two statistical references were used: the same filtered simulation and the unfiltered control, ***UC***. Considering first the wavelet datasets, the difference between each point and the reference mean was divided by the reference standard deviation, and the resulting signed number stored to file. Where the amplitude was higher than the reference simulation, a positive number would be stored; where lower, negative. For each timepoint, a note would also be taken of the deviation from the reference with the highest modulus, occuring at any frequency, along with the frequency at which this occured; this would be written to a second file of *maximum* deviations. The first file thus contained for each signal a vector for each frequency (19 in all), each vector having an element for each timepoint. The element was the signed number of reference standard deviations of the signal amplitude from the reference mean amplitude. The second file contained for each signal two vectors, each vector having an element for each timepoint. The elements described the number of reference deviations from the reference mean of the frequency element at that timepoint that was furthest from the reference mean (either higher or lower amplitude), along with the frequency of that element. The HHT datasets were similar to this, but the frequencies obtained were binned according to the scheme described above and the frequencies written to the maximum deviation file were the exact frequencies obtained from the Hilbert transform.

Graph Production

Many types of graph could be generated from the datasets obtained as described, and many variations were tried. It was found with the DHFR data that many of these possibilities were of little value for visualisation and interpretation. Two graph types were found to be particularly interesting, though. First, amplitude plotted against signal number and frequency, which showed at which frequency bands the different signals contained energy. Second, maximum amplitude deviation plotted against time and signal number, which showed *when* concentrations of energy occurred and *which* signals they occurred in.

It was found that the cartesian coordinate signal sets were difficult to make sense of, and the torsional angles were only slightly better. The most useful of the signal types was the hydrogen bond lengths, these providing insight into the changes taking place in the simulation.

6.3 Results

6.3.1 Wavelet Mean Amplitudes

A graph of mean Morlet wavelet amplitude versus frequency for each of the hydrogen bonds in each of the four trajectories is shown in Figure 6.6. The hydrogen bonds are ordered roughly according to the position of the donor atom in the protein's primary sequence, starting at the N-terminus. Only a small number of the bonds will be discussed, and in each case the nature and position of the bond will be described.

The first thing to note is the lower mean amplitude seen at the lowest frequencies of the three filtered trajectories. As the average is taken only over the time range over which the wavelet amplitude is valid, this is puzzling. Inspection of the plots of hydrogen bond 101 in Figure 6.12 (discussed later in this section) suggest the answer. Motion at low frequencies is heterogeneous; where high amplitude events occur, they tend to have *so* high an amplitude

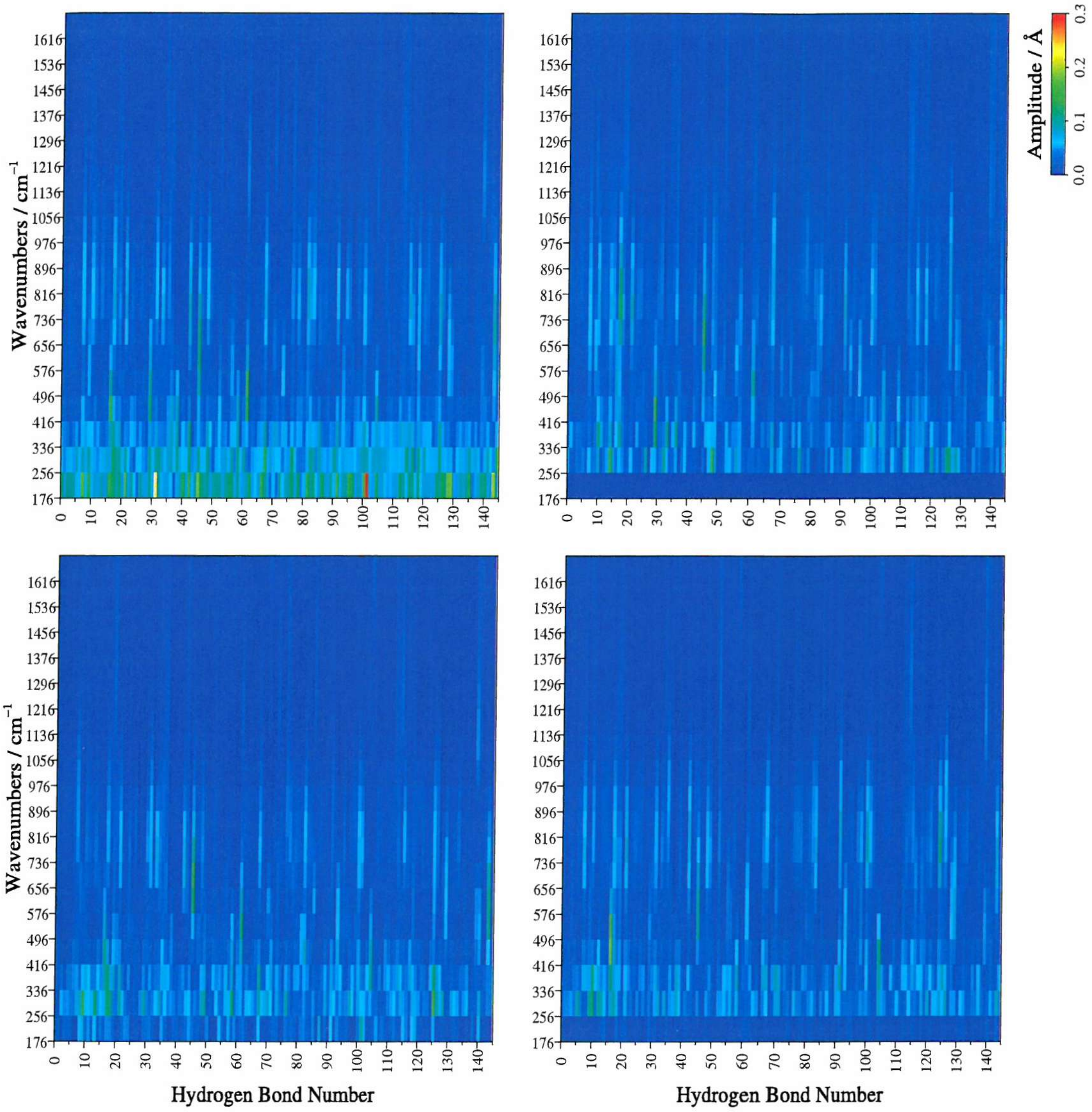

Figure 6.6: Mean amplitudes of wavelet transforms of each hydrogen bond in the four DHFR simulations as a function of frequency. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

that they dominate the computed mean, but such events may be quite localised in time. Where very few datapoints are being computed at these frequencies, the probability of seeing a high amplitude event is low, and this tends to result in lower averages for the shorter simulations: there is a lower likelihood of a high amplitude, mean-increasing event being observed in the shorter window.

A number of distinct patterns of spectral distribution can be discerned from the graph. While all elements show significant amplitude at low frequencies of 336 cm^{-1} and below, the distribution at higher frequencies can be classified as follows; an example is given for each in brackets

1. energy *only* at low frequencies, (bond 0, between the the backbone amide hydrogen of serine 3 and the backbone carbonyl of isoleucine 91)
2. energy present between about 656 and 1056 cm^{-1}, usually with a peak at around 736 or 816 cm^{-1} (bond 42, between the backbone amide of glycine 43 and the backbone carbonyl of glycine 95)
3. energy present between 976 and 1376 cm^{-1} (bond 139, between one of the sidechain amine hydrogens of arginine 158 and the backbone carbonyl of leucine 104)
4. an energy spike between 416 and 576 cm^{-1}, with a second spike between 976 and 1376 cm^{-1} (bond 61, between one of the sidechain amide hydrogens of asparagine 59 and the backbone carbonyl of arginine 57).

An obvious first question to ask is whether these behaviours are correlated to the secondary structure of the enzyme. To examine this, an indicator showing the secondary structure obtained through the DSSP algorithm as implemented in Molmol[10] has been superposed on the ***UC*** amplitude-frequency-bond spectrum; this is shown in Figure 6.7. Where a hydrogen bond occurs between the backbone atoms of a pair of residues where both members are determined to take part in β-strands, a bar is drawn in red; where both members are determined to be α-helix, a bar is drawn in green.

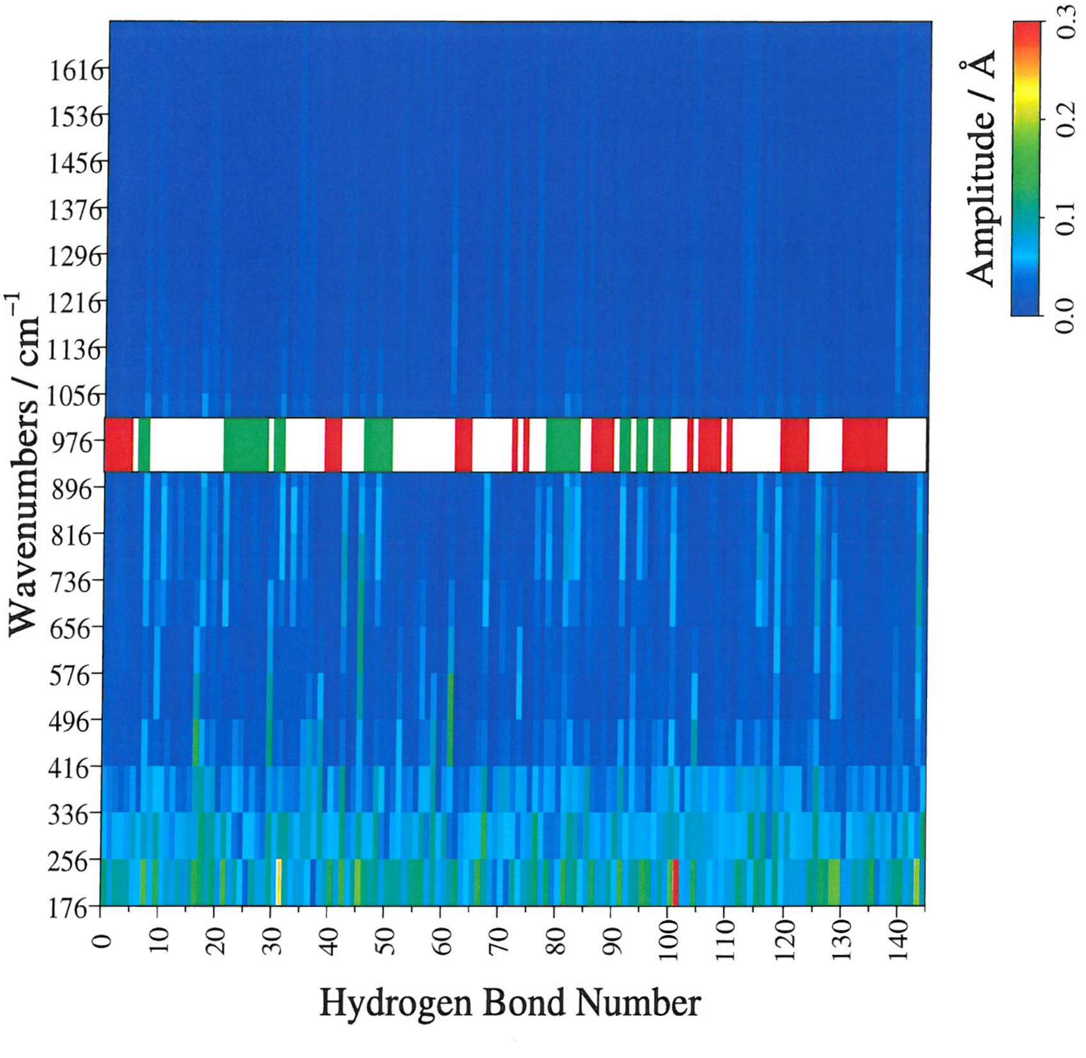

Figure 6.7: Mean wavelet amplitudes of each hydrogen bond in the UC simulation as a function of frequency. A bar indicating secondary structure has been added. Hydrogen bonds between pairs of backbone atoms where both residues are described as helix by DSSP are shown in green; pairs described as sheet are in red.

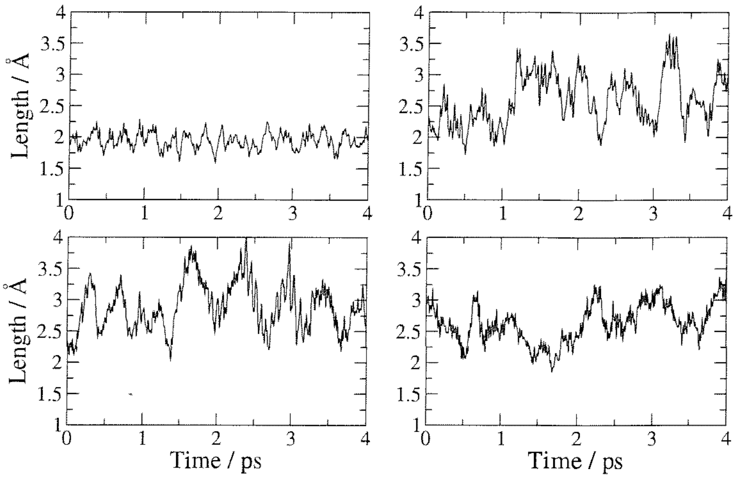

Figure 6.8: Lengths of hydrogen bonds (clockwise from top left) 0, 42, 139 and 61 in the ***UC*** simulations, illustrating the different behaviours seen in Figure 6.6

From this graph two things are apparent. First, β-sheet hydrogen bonds are all of class 1, as defined above: there is little vibration at frequencies of 416 cm^{-1} or above. Second, none of the class 4 hydrogen bonds occur as part of secondary structural backbone bonding. The first observation shows that the method has picked up on the extreme rigidity of β-sheet structures, and also suggests that all of the hydrogen bonds considered by DSSP to be part of such structures are highly stable in this trajectory. If the bonds were less stable, they might conceivably contain energy at higher frequencies.

Plots illustrating the behaviour of bonds of each class described above using the examples given in brackets are shown in Figure 6.8, along with their wavelet transforms in 6.10, and their HHT transforms in 6.9. All figures were obtained from the ***UC*** simulation trajectory, and are plotted on the same linear amplitude scale.

Hydrogen bond 0 (of class 1) remains firmly bonded at a distance of approximately 1.9 Å throughout the entire trajectory, and exhibits vibrations of only about 0.25 Å about this position; this is reflected in the wavelet and HHT spectra. These show similar behaviour, although some slightly higher amplitude activity in HHT is seen below about 176 cm^{-1} at the 1.9 ps point, seemingly reflecting an abrupt shortening and lengthening of the bond with no riding waves in the trajectory. Visualisation of the simulation shows a short, unvarying β-sheet hydrogen bond throughout the whole of the trajectory.

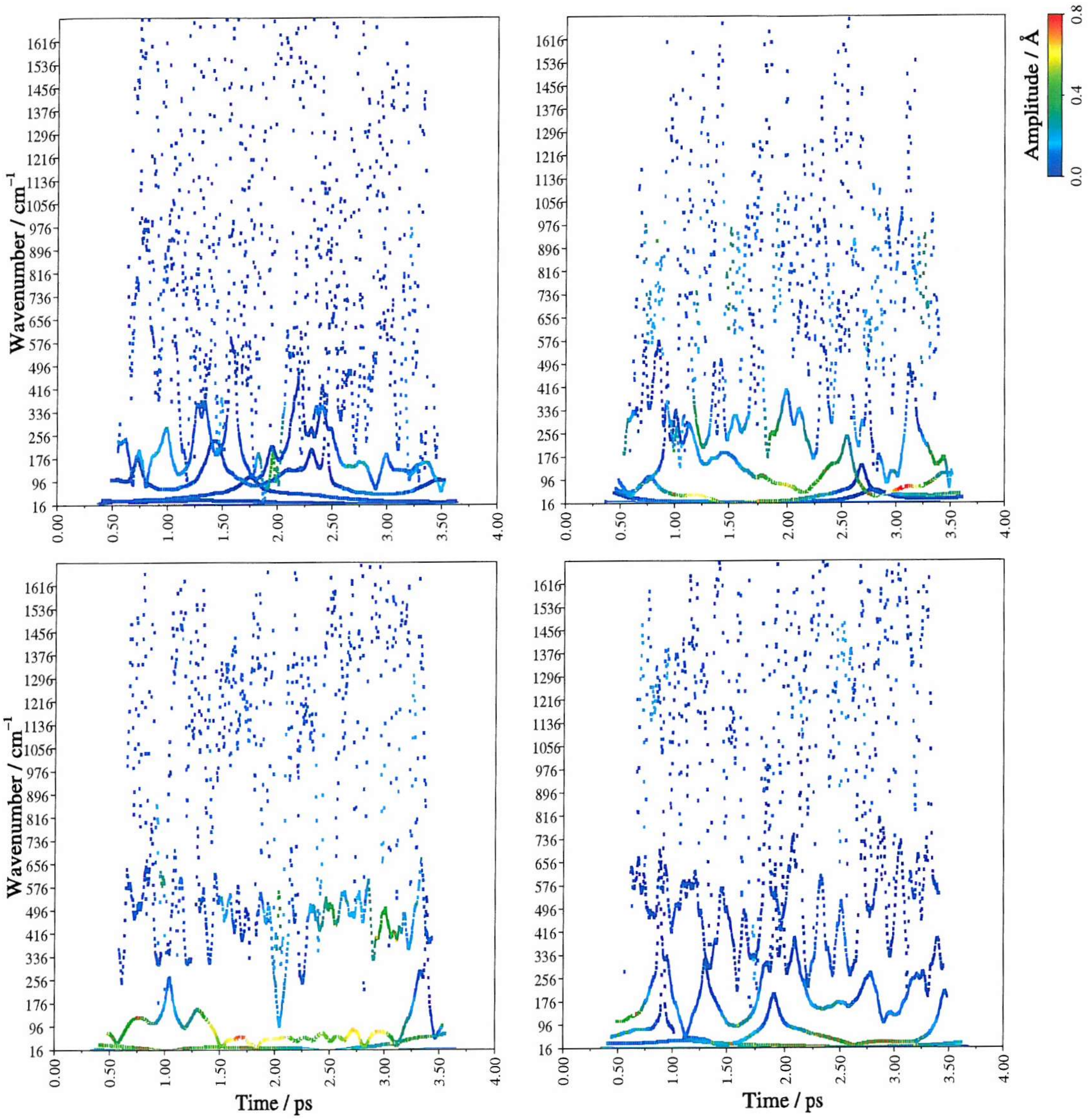

Figure 6.9: HHT transform of hydrogen bonds (clockwise from top left) 0, 42, 139 and 61 in the ***UC*** simulations.

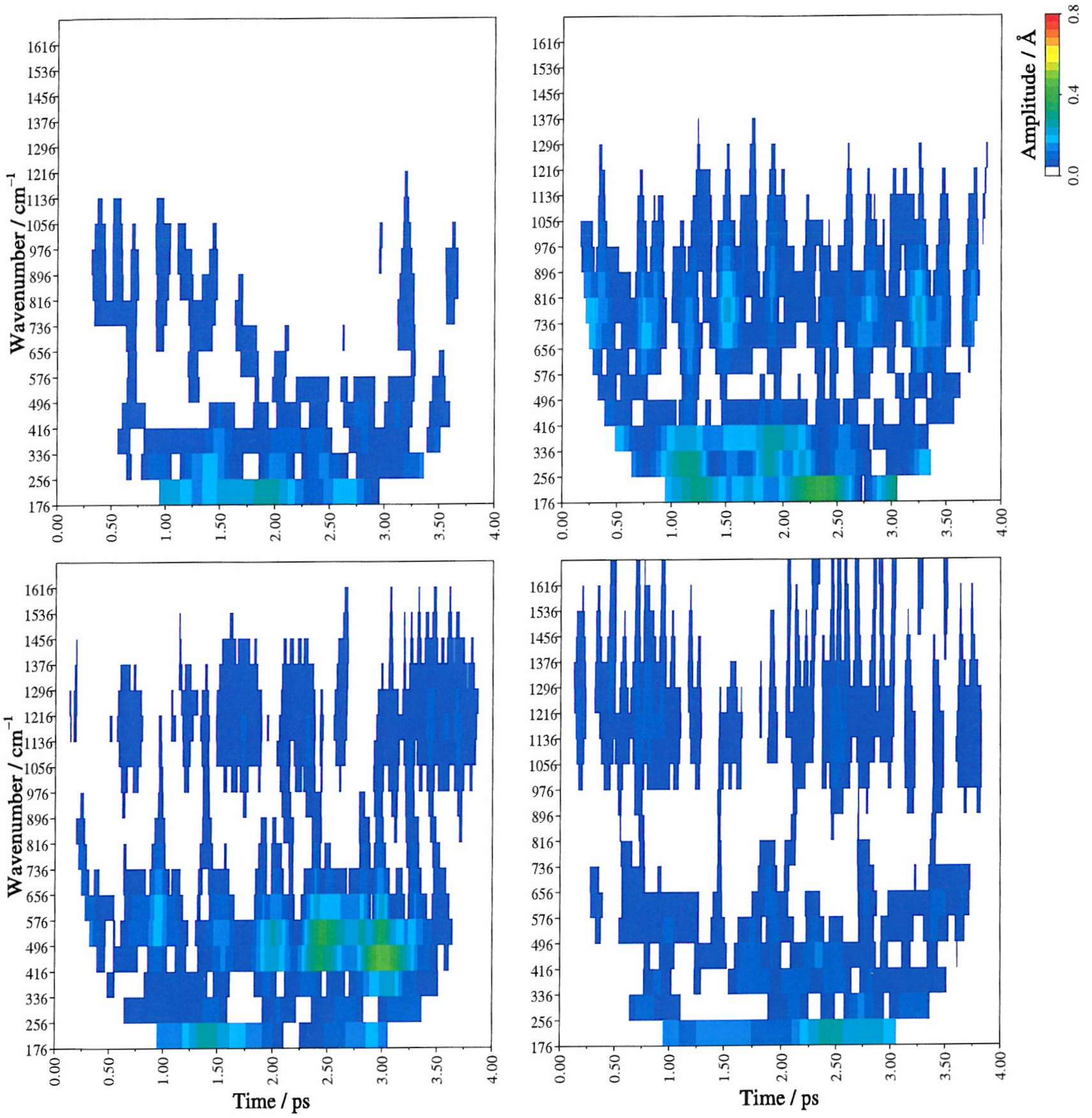

Figure 6.10: Wavelet transform of hydrogen bonds (clockwise from top left) 0, 42, 139 and 61 in the ***UC*** simulations.

Hydrogen bond 42 (of class 2) appears to switch between two states: one where the bond is present, at a distance of around 2.2 Å, and another where the bond is broken and the interatomic distance increases to around 3 Å. Direct visualisation of the trajectory reveals that the distance between amide nitrogen and carbonyl oxygen remains approximately constant throughout the simulation, and the only high frequency motion present corresponds to the C_α-N-H angle vibration of glycine 43. This suggests that vibration in the angle term of the hydrogen bond is the cause of the energy band at approximately 656–736 cm^{-1}.

Hydrogen bond 139 (of class 3) varies between 2–3 Å in length, with little obvious pattern. A low amplitude vibration at around 1136 cm^{-1} can be seen in the wavelet graph. Trajectory visualisation again shows that the distance between the heavy atoms of the bond remains approximately constant, and the high frequency motion appears to be caused by the in-plane C-N-H angle vibration of arginine 158.

Examination of the length of hydrogen bond 61 (of class 4) in Figure 6.8 shows two obvious modes of motion: a low frequency wave with a period of around 0.5 ps (67 cm^{-1}) and an amplitude of around 0.7 Å, and a higher frequency motion whose amplitude reaches a maximum of around 0.5 Å between 2.0 and 3.5 ps. The higher frequency vibration is clear from the wavelet graph, showing up in the 416–496 cm^{-1} bands, but the low frequency vibration is absent. However, the HHT graph (Figure 6.9) shows it occuring at frequencies between 16 and 96 cm^{-1}, as expected. Examination of the wavelet graph also shows a third motion with a frequency of around 1136 cm^{-1}, and an amplitude of less than 0.1 Å. Visualisation of the simulation suggests that the 0.5 Å motion (seen at around 416–496 cm^{-1}) is the H-N-C-O torsional vibration within the amide sidechain of asparagine 59. The origin of the 1136 cm^{-1} vibration is unclear, although comparison with hydrogen bond 139, above, suggests that it may be the in-plane C-N-H vibration.

Returning to Figure 6.6, one final observation is the prominent high mean amplitude feature at 176 cm^{-1} of hydrogen bond 101 in the ***UC*** trajectory. The length of this hydrogen bond in the four trajectories is shown in Figure 6.11, along with the corresponding wavelet and HHT transforms in Figures 6.12 and 6.13, respectively.

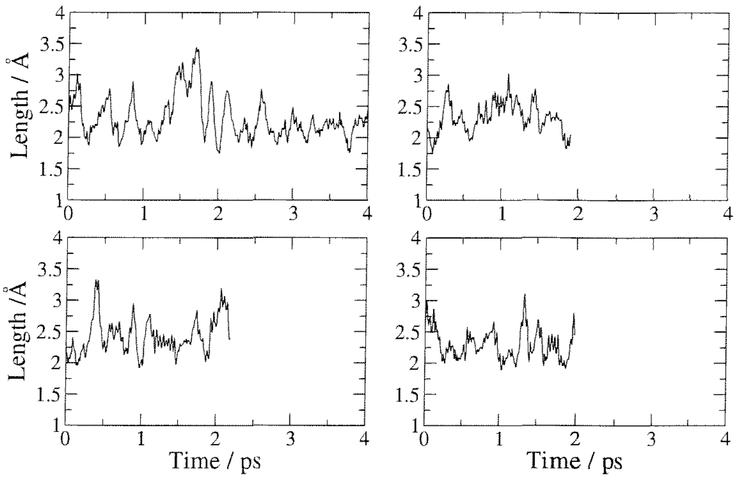

Figure 6.11: Length of hydrogen bond 101 in the four simulations. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

Hydrogen bond 101 is formed between the backbone carbonyl of leucine 104, at the C-terminal end of helix **F**, and the backbone amide nitrogen of alanine 107, just N-terminal to the beginning of strand **F**. Residues 104–107 form a type 1 β-bend at the surface of the protein (residue 105 is a proline, as is commonly seen in such structures), and bond 101 is the key 1–4 bond pinning the sides of the loop together. Visualisation of the simulation shows the random twistings of the structure pull the bond apart in the ***UC*** simulation for a period between 1.4 and 1.8 ps, followed by a pair of abrupt 'snap' events where the bond is repeatedly made and broken. Although it is interesting that the method has picked out a disturbance in a bond involved in the secondary structure, some caution is required. The high amplitude detected seems quite anomalous considering the nature of the event; the most plausible explanation seems to be that the waveform created by the repeated making and breaking of the bond has, by chance, an improbably high similarity to the Morlet wavelet basis function (at around the 1.8 ps point). It is hard to draw any other conclusions from this data; the HHT plot shows no similar high amplitude feature at this time point or anywhere else, in any of the simulations, for this hydrogen bond.

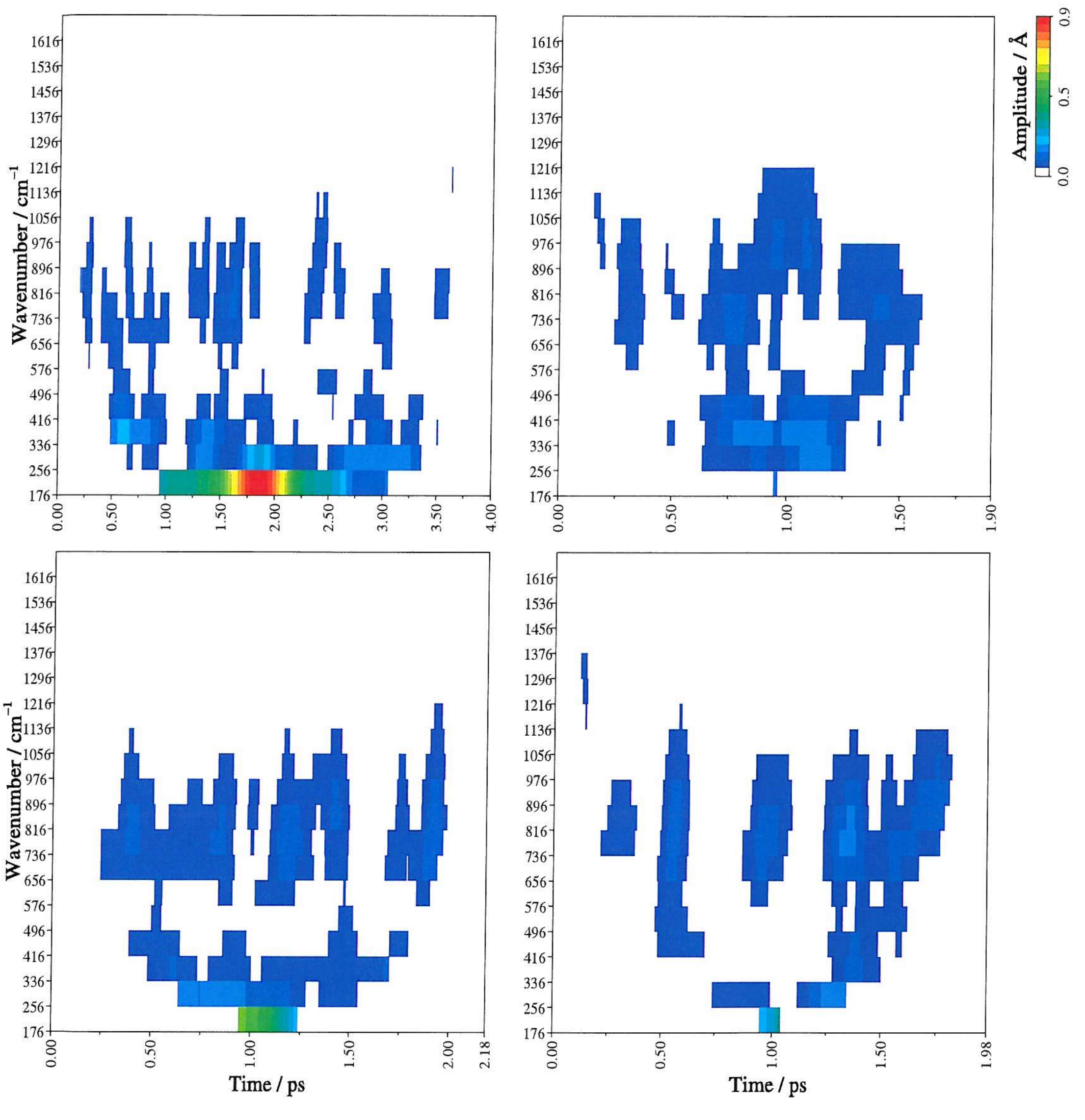

Figure 6.12: Wavelet transform of the length of hydrogen bond 101. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

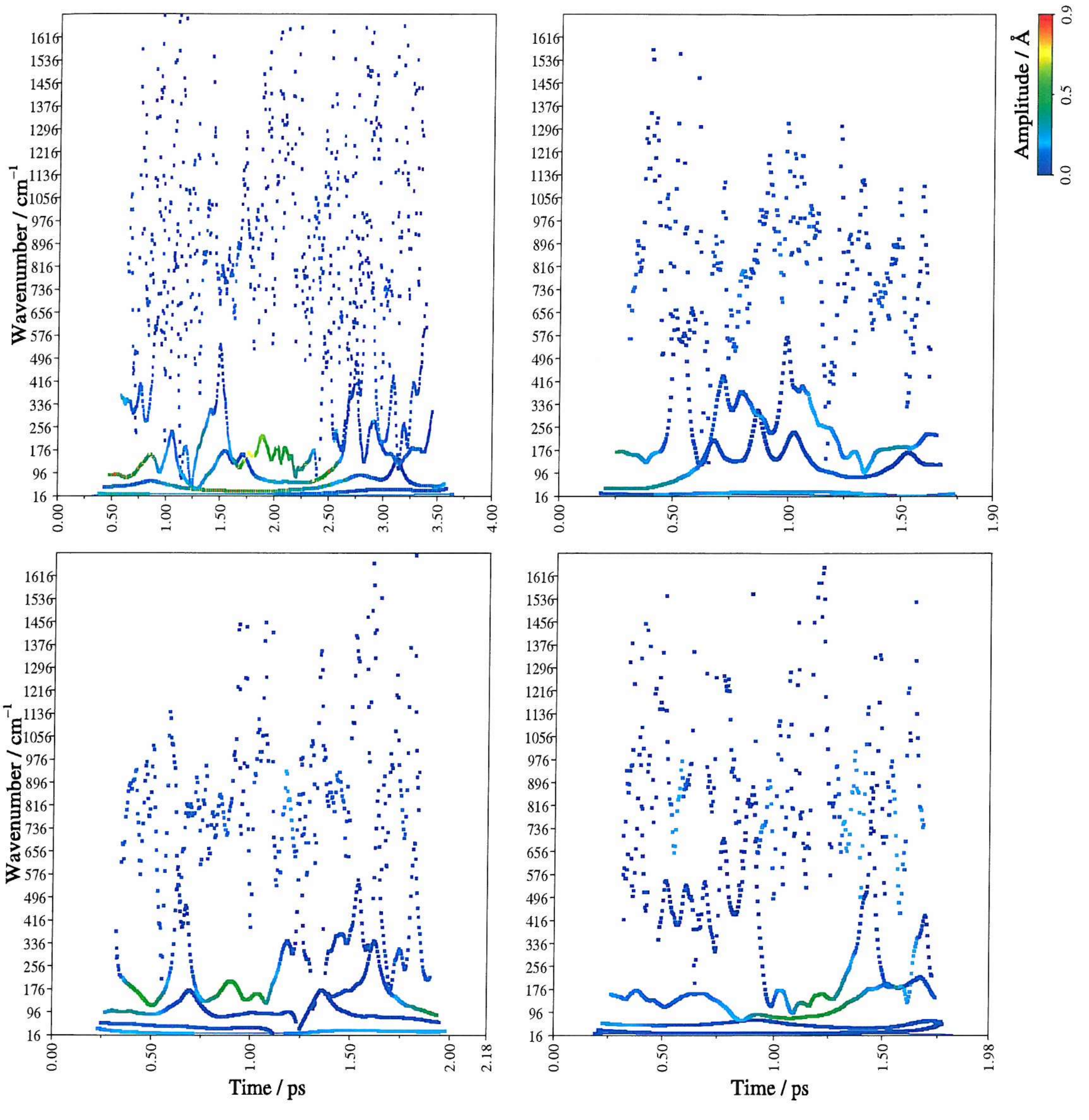

Figure 6.13: HHT transform of the length of hydrogen bond 101. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

6.3.2 HHT Mean Amplitudes

A plot of mean HHT amplitudes for each of the DHFR trajectories is shown in Figure 6.14. It is important to note that the same logarithmic amplitude scale has been used in each of the graphs. When a linear scale was tried initially, the difference between the minimum and maximum amplitude values was so large that almost no detail could be seen in the graph: all but the dozen or so highest amplitude points were represented by a handful of colours at the bottom of the scale. An alternative solution to this would be to simply clip the upper end of the amplitude scale, assigning the same colour to all points higher than a given cutoff value, but this would result in a loss of information from the plot. The logarithmic scale was chosen as a compromise.

The difference between the HHT (Figure 6.14) and Morlet (Figure 6.6) wavelet graphs is quite striking. All of the structure at frequencies higher than $576\,\mathrm{cm}^{-1}$ has been lost in the figure, and the graph is dominated entirely by the activity at low frequencies.

There are large disturbances in the ***FC***, ***F1*** and ***F2*** simulations in a number of hydrogen bonds; in contrast, there is less to see in the unfiltered ***UC*** trajectory. The bonds of special interest were judged by inspection of the graph. These hydrogen bonds are listed in Table 6.1. Each of these bonds appears significant in one or more of the filtered trajectories.

The bonds listed have been grouped roughly according to structure as follows.

Group A involve residues in the set [9, 10, 12–17, 19, 20, 22, 24, 27], which are part of the **M20** loop. These are shown in the upper left picture of Figure 6.15. Group B involve residues in the set [117, 119, 122, 123], which are part of the **F-G** loop and are shown in the upper right picture of 6.15. Group C contains just two bonds, but these specific bonds (14 and 16), are known to be important to the correct functioning of the enzyme. Bond 16 connects helix **C** to the the **M20** loop, keeping the enzyme's active site covered during the first half of the catalytic cyle, but breaking during the second half when the enzyme changes conformation, as described in Subsection 6.1.2. Bond 14 is the cross-connecting hydrogen bond of a β-hairpin formed of residues 16–19 in the **M20** loop, short-timescale rearrangements of which are known to take place in the working enzyme. Group D contains residues in the set [32, 35,

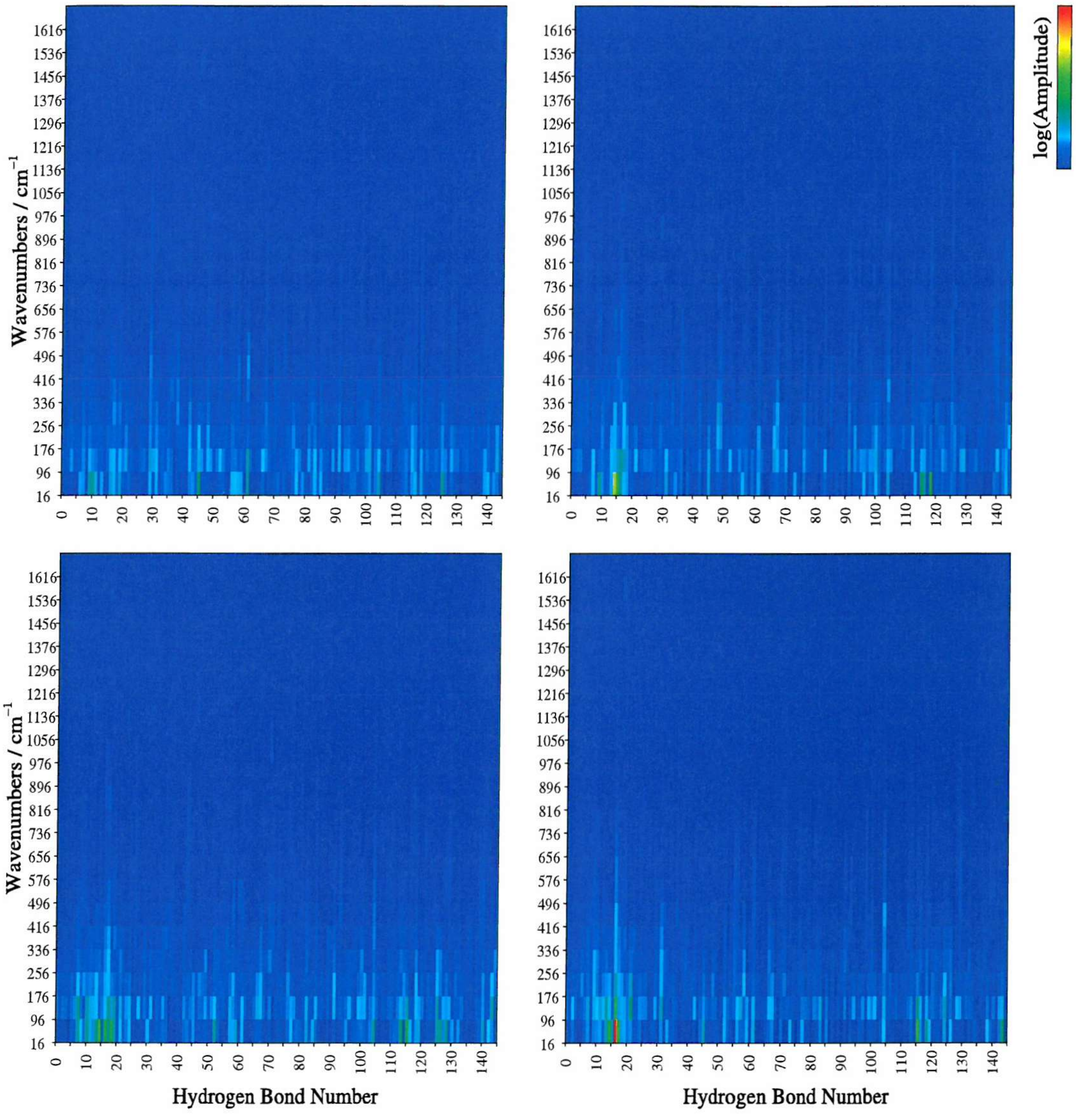

Figure 6.14: Mean HHT amplitudes of each hydrogen bond in all four simulations plotted on a log scale. Note that the frequency scale extends downwards to 16 cm^{-1}. The conformational disturbance is clearly visible at hydrogen bond 16 and in its vicinity in the filtered simulations. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

Bond	Donor	Acceptor	Group
7	166 HN ARG12	136 O ALA9	A
10	190 HN VAL13	136 O ALA9	A
11	190 HN VAL13	152 O VAL10	A
12	206 HN ILE14	1921 O THR123	B
13	225 HN GLY15	1921 O THR123	B
14	232 HN MET16	286 O ALA19	C
15	249 HN GLU17	1904 OD1 ASP122	B
16	273 HD21 ASN18	785 OG SER49	C
17	278 HN ALA19	247 O MET16	A
18	278 HN ALA19	262 O GLU17	A
19	319 HN TRP22	303 O MET20	A
21	400 HN ASP27	374 O LEU24	A
31	545 HN THR35	505 O LYS32	D
91	1519 HN VAL99	1486 O GLY96	D
104	1708 HZ3 LYS109	1743 OH TYR111	E
111	1859 HN VAL119	1842 O ALA117	B
113	1909 HN THR123	230 O GLY15	B
114	1909 HN THR123	1904 OD1 ASP122	B
116	1987 HN TYR128	1982 OD1 ASP127	E
117	2061 HN TRP133	2021 O GLU129	E
124	2233 HN ASP144	2284 O ASN147	F
125	2245 HN ALA145	2240 OD1 ASP144	F
126	2255 HN GLN146	2240 OD1 ASP144	F
127	2272 HN ASN147	2240 OD1 ASP144	F
128	2282 HD22 ASN147	2240 OD1 ASP144	F

Table 6.1: Table of hydrogen bonds that contain significant energy in one or more of the RDFMD simulation trajectories as defined from the HHT plots (Figure 6.14). The bond numbers are counted from zero: the same numbering convention is used as in the previous figures. Donor and acceptor atoms are listed according to their atom number in the PDB file (code 1RX2), and described according to the PDB convention. '166 HN ARG12' refers to atom number 166: the hydrogen attached to the backbone amide nitrogen of residue number 12, which is an arginine. The labels are referred to in the body of the text.

96, 99], which appear to be distant from the active regions of the enzyme; visualisation of the simulation reveals no clear large scale motions in these structures, or obvious disturbances in the lengths of the two hydrogen bonds associated with them. Group E contains residues in the set [109, 111, 127–129, 133], which form a physically contiguous group shown in the bottom right corner of Figure 6.15. These residues do seem to change slightly in relative position on comparing the first, middle and last frames of all four trajectories, although it is difficult to determine what meaning, if any, to ascribe to this; one possibility is that a 'breathing' motion in the structure is causing the amplitude peaks in these bonds. Group F contains residues in the set [144–147], which are situated in the **G-H** loop region, a flexible domain that is associated with conformational change events over the catalytic cycle as described in Subsection 6.1.2. The residues in this group are shown in the bottom left corner of Figure 6.15.

It should be noted that hydrogen bond 101 is not registered as having the same high amplitude motion observed in the wavelet mean amplitude figure, either in the ***UC*** trajectory or in the others. This adds weight to the interpretation that this behaviour is an artefact of the wavelet transform.

One bond, though, does seem to exhibit an extremely large amplitude motion in the ***F2*** trajectory: hydrogen bond number 16. This is extremely interesting, as the making and breaking of this bond is critical to the function of the enzyme, and the ***F2*** trajectory is the one trajectory where a significant and clear breaking event is seen. The behaviour of this bond, and also bond number 14, are examined in more detail in Subsection 6.3.3.

To summarise, the HHT mean amplitude spectrum lacks the detailed separation of different forms of behaviour seen in the wavelet mean amplitude spectrum, but has correctly identified not only the **M20** loop region of the structure where amplitudes were enhanced by the RDFMD method, but also the motions of residues to which no filtering was applied. These latter included group B in the **F-G** loop which are in physical contact with the filtered residues, and also group F, in the **G-H** loop, which are not, but which have been shown by experiment to rearrange during the conformational change events of the protein's catalytic cycle. Of concern, though, are the apparent false positive motions seen in group D. It was not

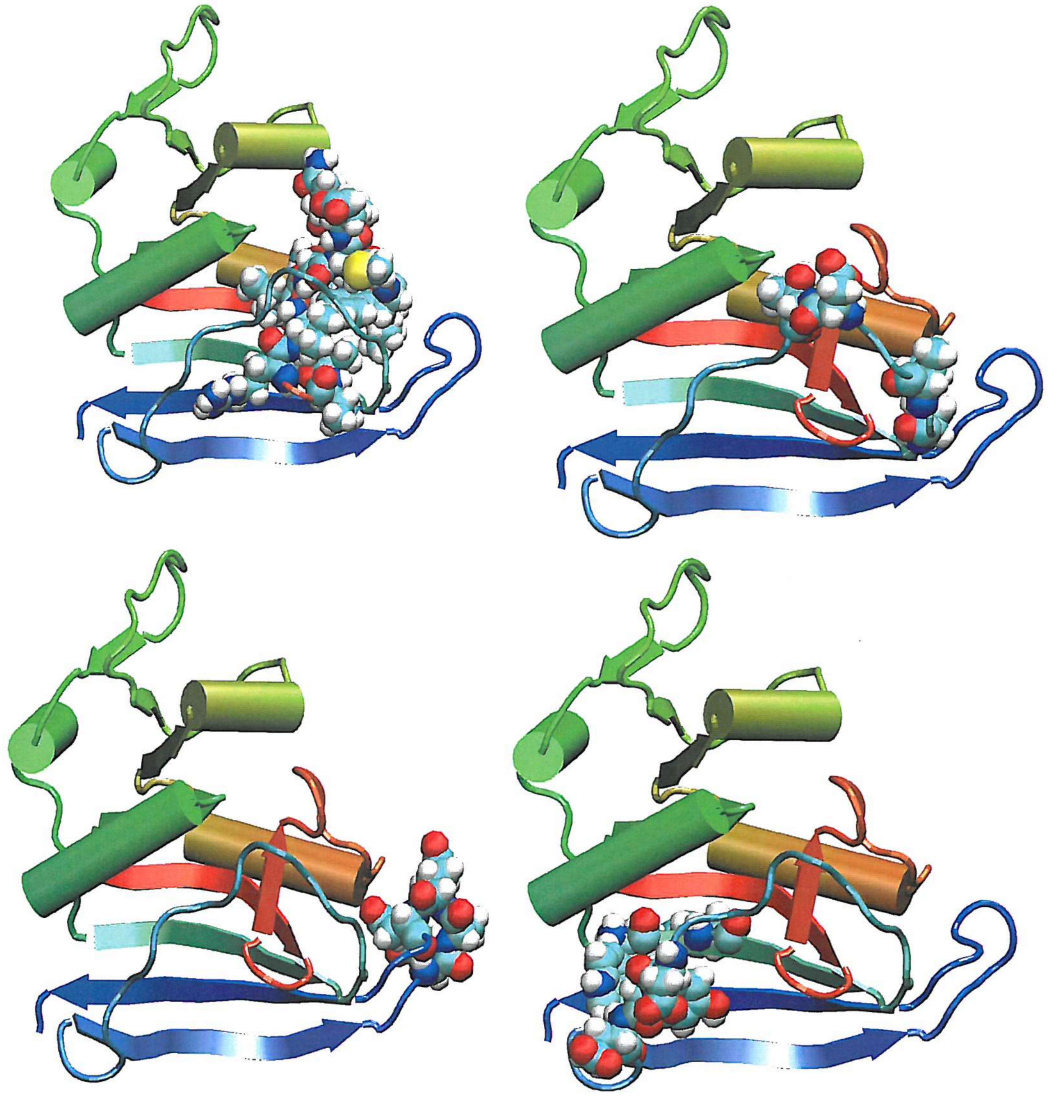

Figure 6.15: Parts of the DHFR structure shown by the HHT analysis to move during the conformational change. Refering to Table 6.1: top left are the residues in group A; top right is group B; bottom left is group F; bottom right is group E.

clear from visualising the trajectory why these bonds registered as having large scale motions associated with them. Also, why the motion of residues in group E registered as having significant amplitude was unclear from examination of the trajectories. A possible explanation of these two seemingly disappointing results is that the motions which registered are simply too slow to easily be seen from direct visualisation of the trajectory, although confirming this would require further work, possibly using a technique such as essential dynamics. It should be noted that much of the activity seen in this spectrum occurs at very low frequencies, beyond the range of the Morlet wavelet spectra.

6.3.3 Maximum Deviations

A plot of the maximum wavelet amplitude occuring at each timepoint in each hydrogen bond is shown in Figure 6.16. The ***UC*** trajectory has its deviations compared with its own statistics, while each of the filtered simulations are also each compared with the ***UC*** trajectory statistics. Examining the graphs it is clear that apart from the obvious high deviation event in the ***FC*** trajectory, there is not really very much to see.

Hydrogen bond number 14 clearly shows a well defined period of high deviation from the control simulation behaviour between 1.5 and 1.6 ps. This is obviously interesting, as this is one of the critical hydrogen bonds associated with the conformational changes the protein undergoes during its catalytic cycle. This and hydrogen bond 16 were defined as 'group C' in Subsection 6.3.2. The waveform of the length of the bond and its associated wavelet and HHT transforms will be discussed below. For the present it is sufficient to note that substantial activity only appears in the wavelet maximum deviation plot for the ***FC*** trajectory, and not for ***F1*** or ***F2***, as would be expected. These latter two filtered trajectories undergo conformational change events, and all trajectories other than ***UC*** have had energy added at frequencies 0–50 cm^{-1} to residues 14–19. The obvious conclusion is that neither the energy added by the filter, nor the conformational changes triggered by it in ***F1*** and ***F2*** can reliably be observed by the Morlet wavelet transform, due to the constraints on the measurement of low frequencies. Such constraints would apply to any simulation whose length is similar to these.

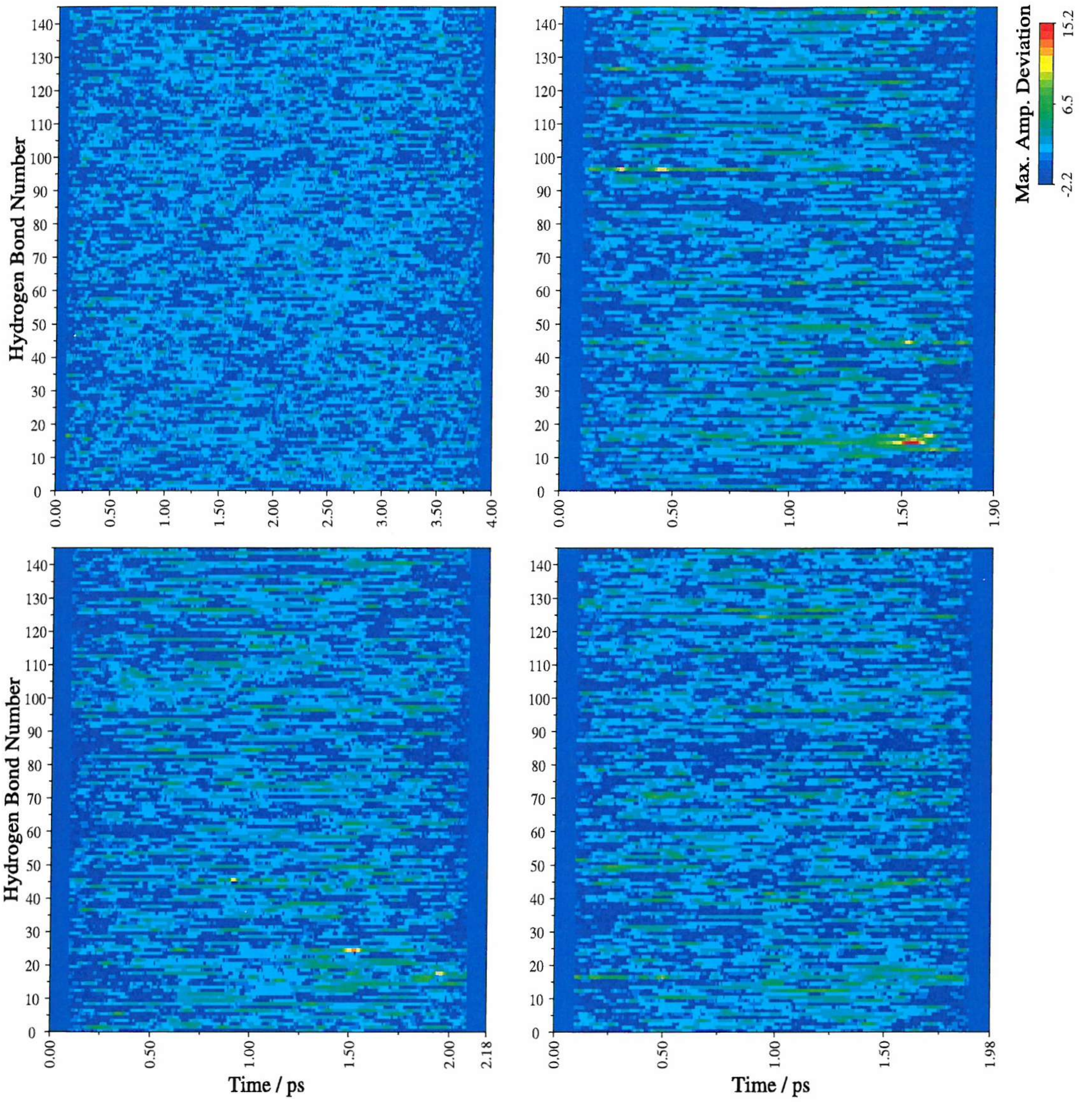

Figure 6.16: Maximum deviation in wavelet amplitude relative to ***UC*** trajectory mean at any frequency as a function of time for each hydrogen bond in the four simulations. The increase in energy in the asparagine 18—serine 49 bond is evident in the ***FC*** but not in the ***F1*** or ***F2*** plots. This misleading result shows the limitations imposed by the low frequency constraints of the Morlet wavelet. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

Maximum deviation plots were prepared for the four simulations using the HHT method in a similar way to the wavelet plots, and these are shown in Figure 6.17. The first thing it is important to note about these graphs is that the linear scale has been clipped at the top and bottom ends of its range: amplitudes more than 50 (***UC***) standard deviations higher than the ***UC*** mean value are shown in black, and amplitudes more than 5 deviations below are shown in white. Only a small number of points are affected by this clipping and so little information is lost from the graphs.

It is strikingly clear at first glance that this graphing method captures some sort of event taking place. High amplitude features are clearly detected in hydrogen bond 14 in trajectory ***FC***, ***F1*** and ***F2***, and in bond 16 in ***F1*** and ***F2***. Other bonds with significant motion are 18 in ***FC***; 7, 8, 18, 19 and 20 in ***F1***; 13 in ***F2***. These events are mainly observed during the second half of the trajectories.

The lengths of hydrogen bonds 14 and 16 are shown in Figure 6.18. The correspondence between the motions in the bonds and the HHT maximum deviation graphs is reasonably good, although not perfect. It would be expected intuitively that the change event in bond 16 in the ***FC*** trajectory should result in a higher amplitude deviation. Also, the events detected in bonds 14 and 16 should have more similar deviations in the ***F1***, and more differing ones in the ***F2*** trajectories. It is not clear why these deviations from the intuitively expected behaviour occur.

The HHT and wavelet plots for both of these datasets are shown in Figures 6.20, 6.19, 6.22 and 6.21. The key point to note from these plots is how much of the energy is present at low frequencies beyond the reach of the wavelet transform. Although the maximum *statistical deviation* event seen in the wavelet transform of the ***FC*** trajectory (Figure 6.16) is clearly a significant peak, the actual *amplitude* of the observed peak is very low, as seen from Figure 6.22. The wavelet transform has spotted the leading edge of the transition event at a low amplitude value, but its failure to detect the full scale of the event suggests that the wavelet maximum amplitude method is less reliable than the related HHT method, due to its frequency limitations.

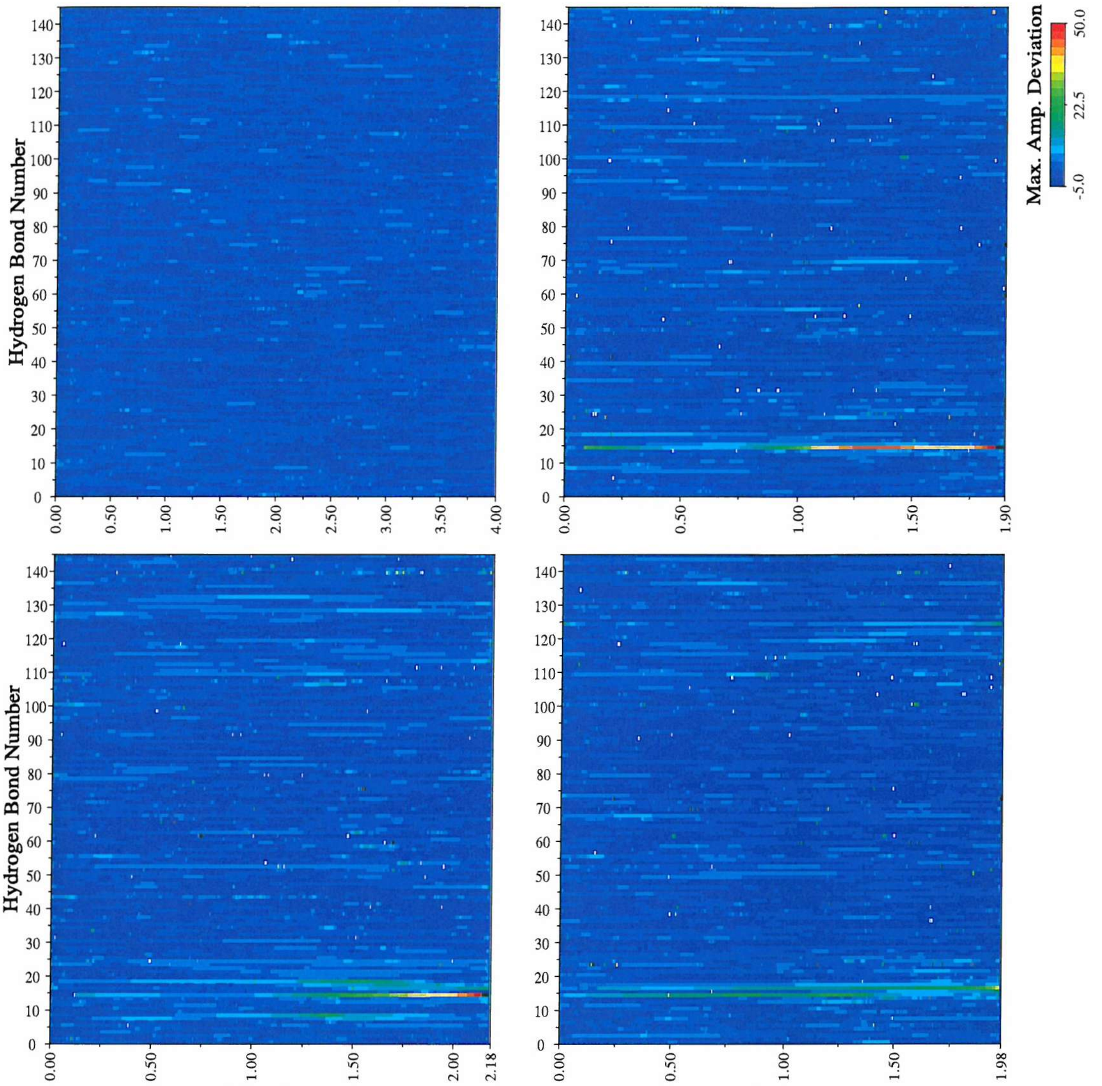

Figure 6.17: Maximum deviation in HHT amplitude at any frequency as a function of time for each hydrogen bond. The increase of energy in the region undergoing conformational change is clearly visible in the three filtered simulations, contrasting favourably with the misleading wavelet plot. The analysis of frequencies below 176 cm^{-1}is clearly vital for the correct detection of change events in these simulations. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

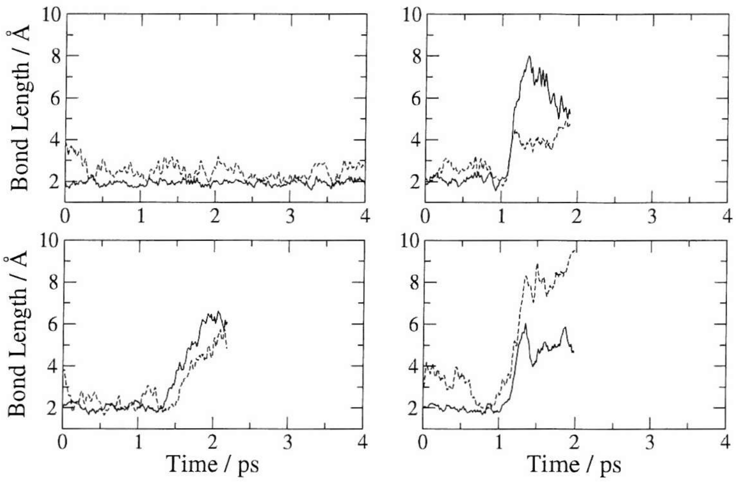

Figure 6.18: Length of hydrogen bonds 14 (methionine 16—alanine 19) and 16 (asparagine 18—serine 49) in the four simulations. Bond 14 is drawn with a solid line and bond 16 with a dashed line. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

Finally, it was noticed that the HHT torsional plots also show the presence and time locations of conformational transition events in the three filtered trajectories. These plots are shown in Figure 6.23. The linear deviation scale has been clipped in a similar manner to the HHT hydrogen bond plots, but with a maximum standard deviation value of 20, and a minimum of -5. The ψ and ϕ torsional angles are both plotted for each residue, starting with ψ. The very large number of horizontal lines shown makes the graph somewhat difficult to read on paper, but this is not a problem when the graph is shown on a computer display: the analyst can simply ‘zoom’ into the areas of interest on the plot, and reading the scale becomes easy. For the purposes of this writing, the degrees of freedom of interest will be stated.

It is immediately clear from this graph that residues 9–14, which are targetted by RDFMD and would be expected to show increased motion, do so. The angles which show motions with an amplitude deviations greater than 15 standard deviations are: in ***FC***, ϕ of residues 15 and 16 at around 1.2 and 1.1 ps respectively, followed by a larger motion in ϕ16 toward the end of the trajectory (also registered are points at the very end in both angles for their ψ components); in ***F1***, ψ20 at 1.2 ps, ϕ18 at 1.7 ps, and ψ17 at 2.0 ps; in ***F2***, ϕ16 in a broad band centred at around 1.5 ps, ψ16 at the end of the trajectory, and ψ15 at 1.1 ps, and again at 1.6 ps (the second motion seemingly weaker). The three filtered trajectories were visualised to determine the nature of any transition events taking place.

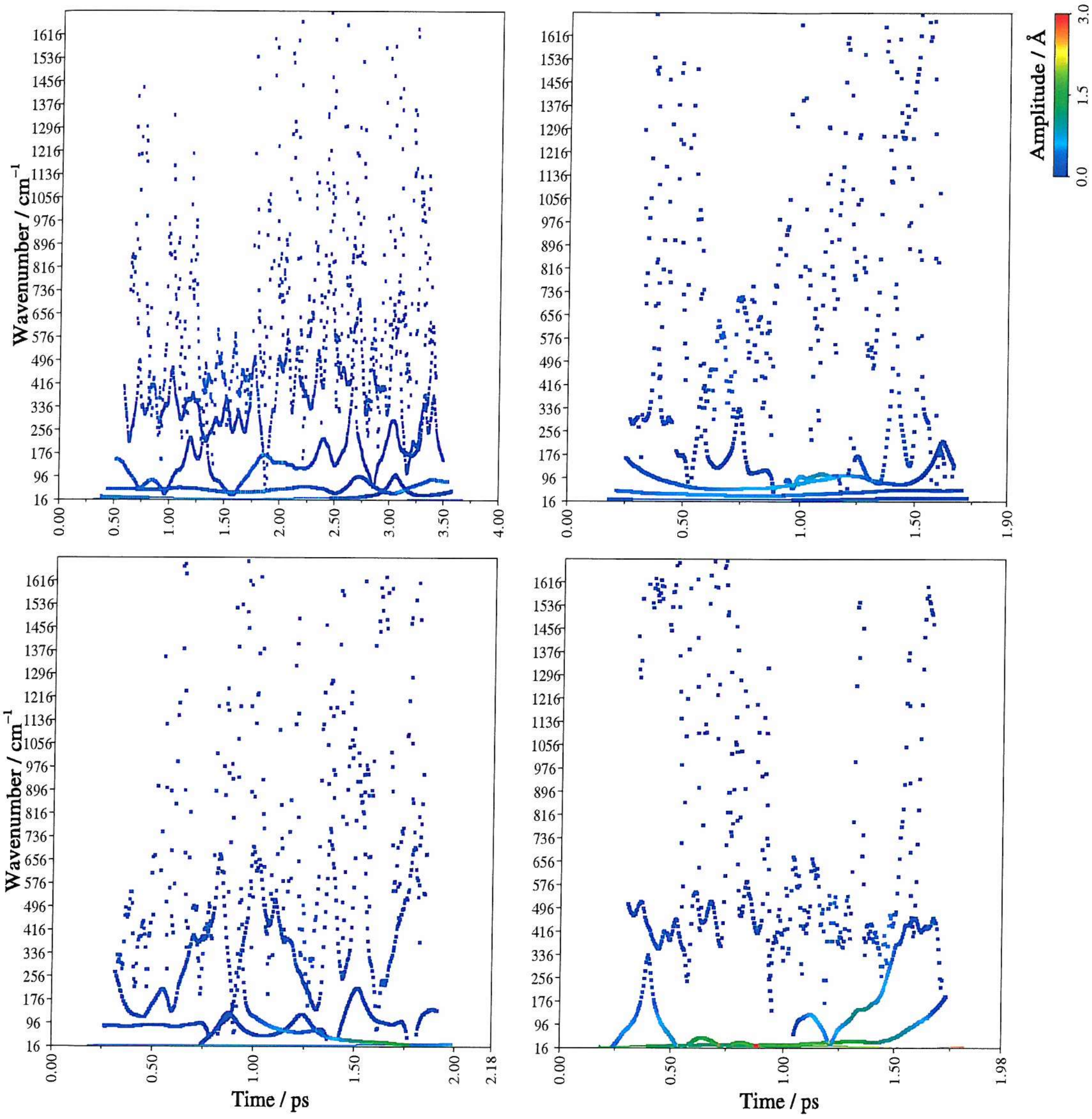

Figure 6.19: HHT transform of the length of hydrogen bond 16. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

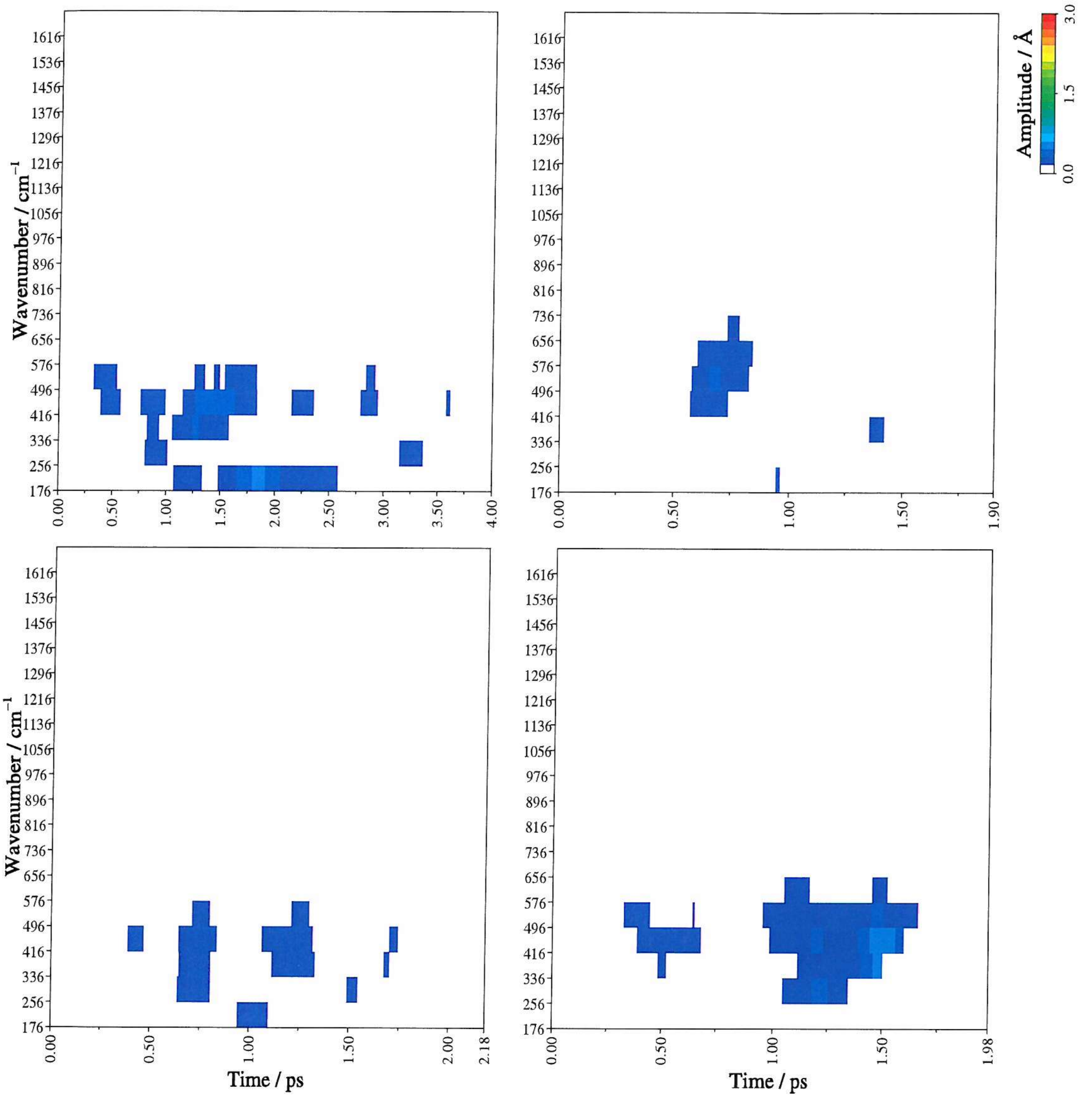

Figure 6.20: Wavelet transform of the length of hydrogen bond 16. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***

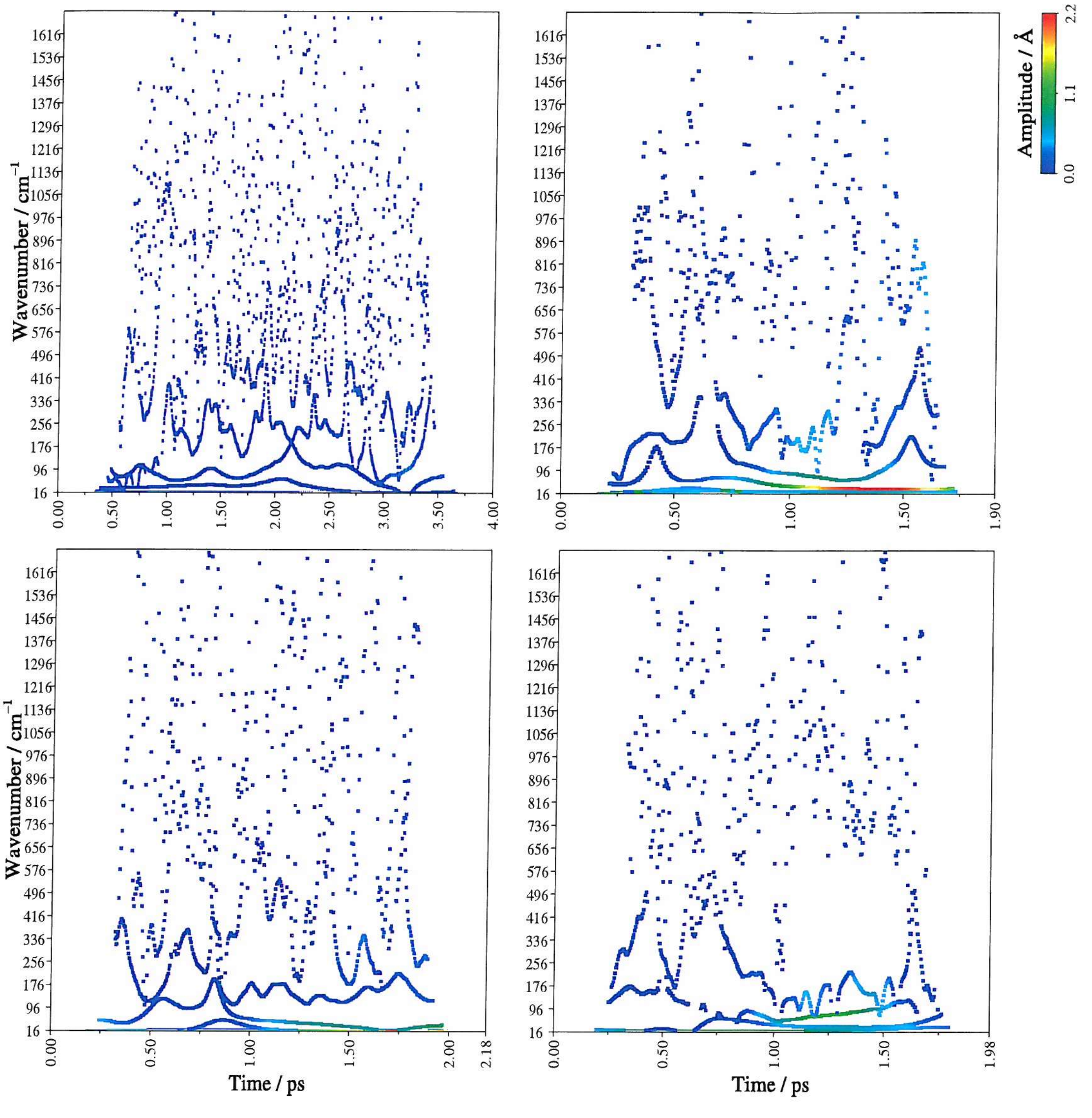

Figure 6.21: HHT transform of the length of hydrogen bond 14. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

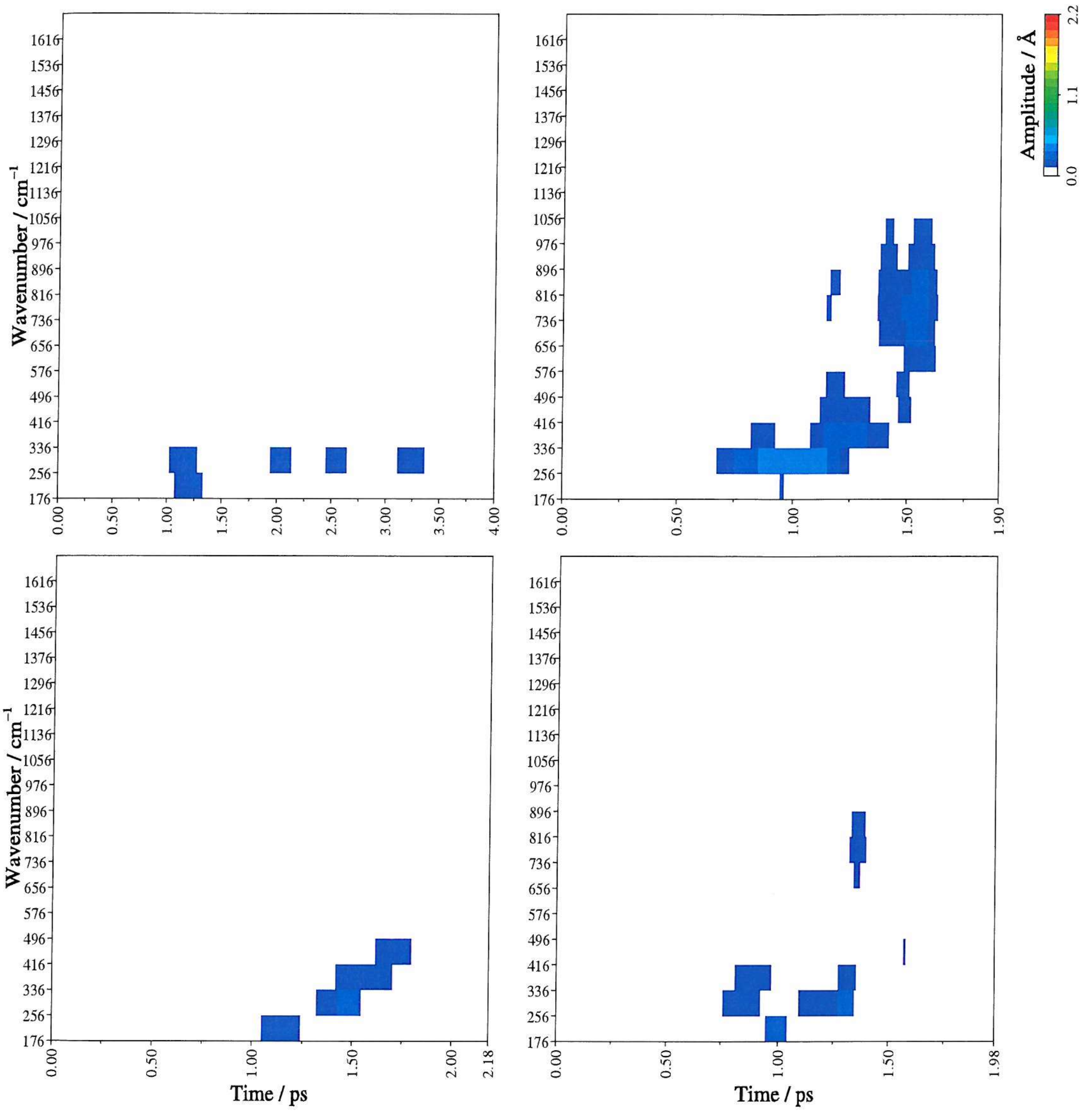

Figure 6.22: Wavelet transform of the length of hydrogen bond 14. Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***.

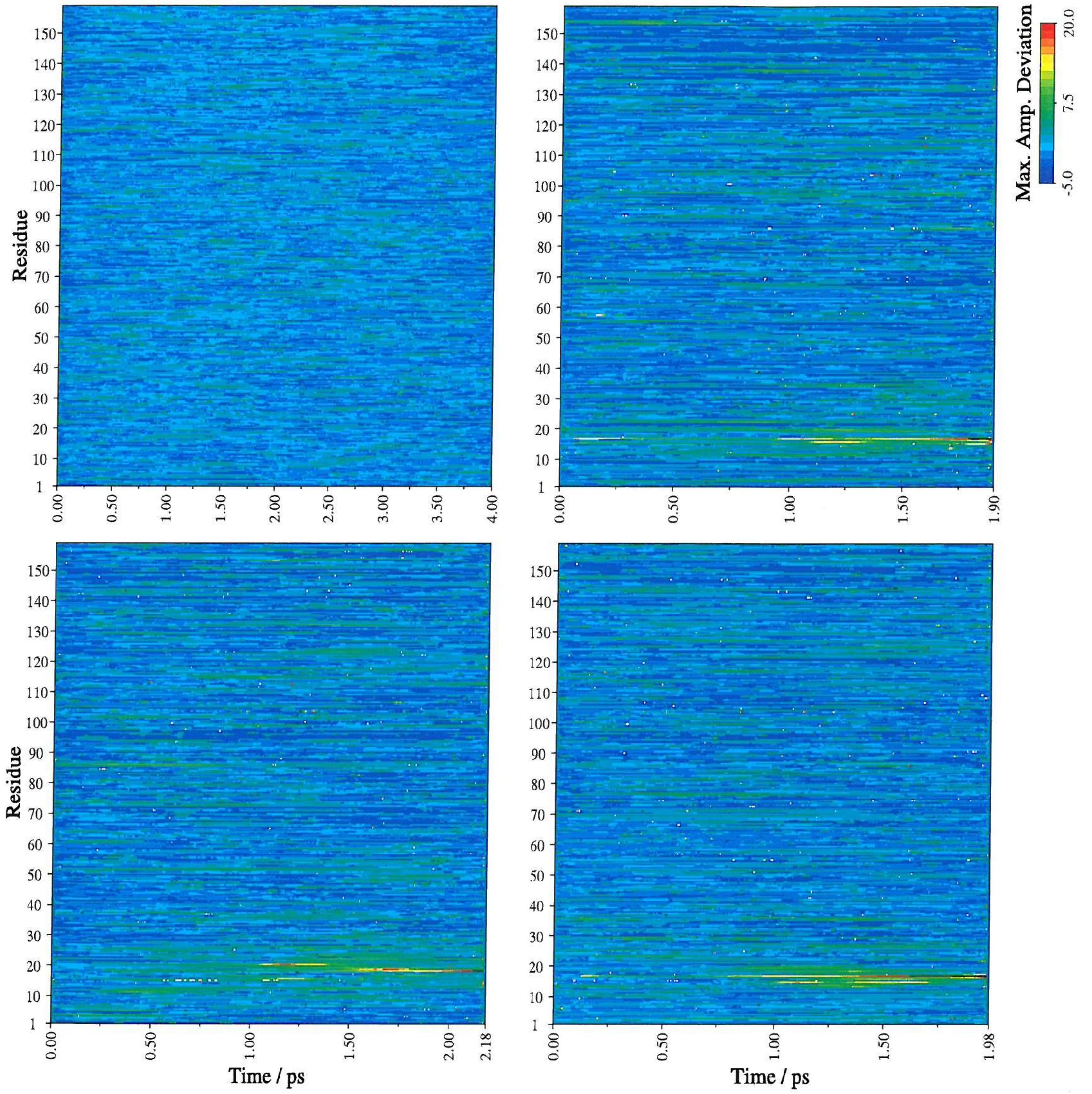

Figure 6.23: Maximum deviation in HHT amplitude at any frequency as a function of time for the ϕ and ψ backbone torsional angles (two horizontal graph lines are drawn for each residue). Clockwise from top left are ***UC***, ***FC***, ***F2*** and ***F1***. Very high amplitude points are plotted in black.

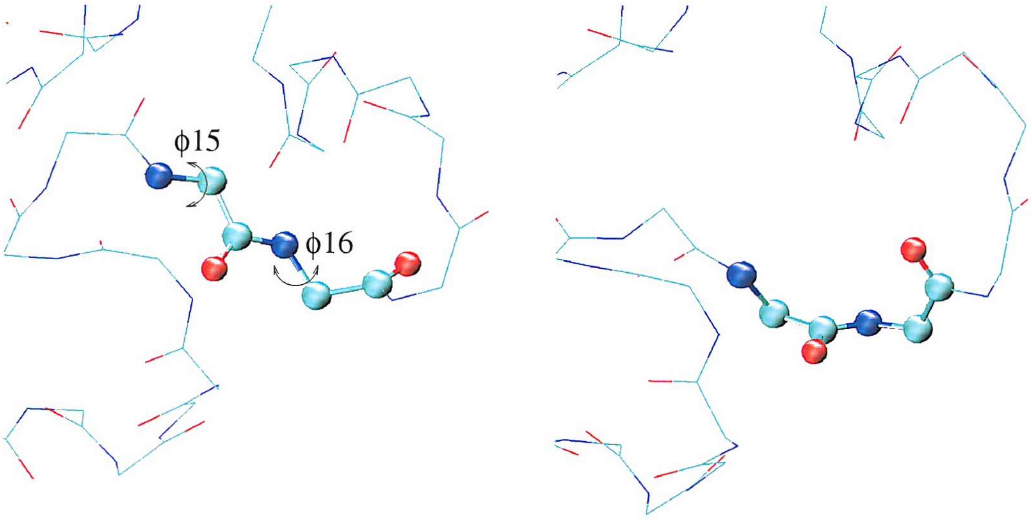

Figure 6.24: Backbone of residues 15 and 16 in the **M20** loop region of DHFR. Left: conformation seen at frame 0 of the ***FC*** trajectory. Right: conformation at final frame.

In the ***FC*** trajectory, there is a subtle disturbance in the backbone conformation of the **M20** loop at residues 14 and 16, as suggested by the torsion deviation graph. The motion, while interesting, does not cause a significantly large change in the rest of the protein. Images taken from the beginning and end of the trajectory are shown in Figure 6.24. Without knowing which residues to look at, this disturbance would be difficult to spot from a straightforward examination of the simulation trajectory.

The ***F1*** trajectory shows a clear flip event in the ψ angle of residue 20 from about 60 to about 170 degrees at the time indicated in the graph. The ϕ angle of residue 18 flips abruptly from close to 180 degrees to around 30 degrees, at the later time indicated in the graph. The ψ angle of residue 17 does not actually flip between states in the simulation, but does exhibit a lot of abrupt twisting toward the end of the trajectory. Why these should be detected as a single broad smear of energy is unclear, though.

The backbone of residues 15 and 16 in the ***F2*** trajectory exhibits a number of very abrupt movements that are difficult to put into words. The most important thing to note about these motions is that some of them occur at the very end of the trajectory: the large 'events' seen in the last 0.25 ps of the deviation graph (Figure 6.23) are *not* artefacts, but real events in this case.

6.4 Discussion

The ideas developed in this chapter to facilitate the rapid inspection of many dynamical properties of molecular simulations have been shown to have merit. Some specific refinements are still necessary, and these will be discussed below. An important general issue identified is the question of what type of *degree of freedom* within the simulation to inspect. For the DHFR simulations, hydrogen bonds gave the clearest and most interesting picture of the dynamics; torsions were useful for the HHT maximum deviation approach, but less so elsewhere; cartesian coordinates were of little use in any of the methods, as the diversity of behaviour present in each graph was too great for the eye to easily interpret, and the level of noise also seemed relatively higher. The graphing of very large numbers of degrees of freedom is another general issue identified, which, up to a point, may be helped by the use of computers for graph inspection: the ability to magnify particular regions of a graph on a VDU was very helpful on many occasions. Although perhaps less useful for the publication of results, this idea is worth noting as another means by which computers can aid scientific visualisation. An alternative possibility for these statistical techniques is to develop some form of filtering algorithm to determine the parts of a system where something significant happens during a simulation, displaying a broad variety of data for a small number of degrees of freedom.

The wavelet mean amplitude spectrum (WMA) shows interesting fine details of the motions in the simulations. The lack of high frequency vibrations in β-sheet is seen as expected, and goes to confirm the extreme stability of all such structure present in all of the DHFR simulations. It is interesting to be able to show the high frequency vibrations in the backbone, seemingly attributable to amide bending and torsions, although perhaps not very useful in itself. Unfortunately, the lack of information about low frequencies limits the usefulness of this method for detecting conformational change events.

HHT mean amplitude spectra (HMA) show no detail of the high frequency motions when drawn as they have been here. The principal problem with these plots is scaling. It should be possible to develop an improved method of scaling the data obtained such that the low frequency components do not overwhelm higher frequency ones, allowing similar high-frequency details to be seen in the HHT spectra as in the WMA. Even without this, though, the HMA is capable of detecting the very low frequency motions associated with conformational change. For the DHFR simulations, the method was able to correctly pinpoint several regions of the protein disturbed by the RDFMD technique, including points not targeted directly by the filter used. There were a number of what seemed to be false positives, though it is possible that further work using other methods, perhaps ED, might, at the very least, reveal why these regions were picked out.

Maximum deviation versus time seems the most useful of the methods tested, and also the method which most clearly showed the limitations of the Morlet wavelet to probe the low frequency motions associated with conformational change events. The graphs produced using the HHT form of this method were able to highlight regions of the protein subject to conformational changes in a very clear manner, and also gave a reasonable definition of when those changes took place. It seems plausible that improvements in the amplitude scaling method used might allow a better time definition of specific events. One avenue for further investigation here is the averaged HHT marginal spectrum; another is to develop some means by which amplitudes are accumulated across all frequencies for a time point, rather than just picking out the highest value (and losing information in the process).

These methods, as presented, enable the analyst to detect a conformational change event in a simulation with little difficulty, and a reasonable degree of reliability. Given a control simulation for which it is known that no interesting behaviour takes place, large numbers of other simulations can be rapidly scanned; interesting events are highlighted quite clearly. Fully measuring and characterising the method's ability to detect subtle changes while rejecting uninteresting motions will require further work and the examination of many more simulation trajectories. The discriminating ability seen for these DHFR simulations suggests that it may be possible to improve the methods to the point where they could be reliably automated: a simulation could be examined automatically for interesting behaviour, and classified on the basis of what type of behaviour was detected. This could be useful for projects where very large numbers of simulations must be examined with little or no human intervention, such as the ‘folding at home’ initiative and similar large scale distributed computing projects, and also parallel tempering MD methods.

Chapter 7

Conclusions

In Chapter 3 we described a series of simulations of the Syrian hamster prion protein and a set of point mutated structures associated with disease. The picture which emerged was of a protein under internal tension, whose β-sheet content was finely balanced. The mutations had significant effects on the dynamics of the structures obtained from the simulations, particularly on the sheet region. An *arm* region was also identified from the analysis, the dynamics of which were influenced by the residues whose mutations were studied. The effect of the mutation associated with disease (D178N) was to increase the *stability* of the β-sheet region, and to loosen ties between the rigid helical core of the protein and the more flexible parts, particularly the arm region. The (M129V) polymorphism that controls the phenotype of inherited prion disease increased the *size* of the β-sheet and also created a templating effect, assisting the reformation of the sheet structure after breakage events. It was also conjectured that where the arm region became more mobile, the approach of other prion molecules might be facilitated.

The deeper lesson drawn from the prion simulations, though, was that the analysis tools used in interpreting protein simulation dynamics were inadequate for the task. Even though there were clearly significant differences between the behaviours of the structures, it was extremely difficult to interpret those differences with the methods available. Problems arose from the fitting strategies required by some of the methods, which required many subjective judgements to be made. Furthermore, all of the analysis methods were dependent for their interpretation on subjective comparison with the visualised trajectory. Further still, some of the information obtained by these methods was contradictory, and these contradictions proved impossible to resolve. Improved analysis methods would be required for a full and clear understanding of simulations with similarly subtle variations in behaviour.

In Chapter 4 we described an analysis method developed very recently by NASA oceanographers for the decomposition and interpretation of wavelike phenomena: the Hilbert Huang Transform (HHT). Many workers in various fields have made use of this new method, and we decided that this would be an interesting avenue of investigation.

In Chapter 5 we explained that the method was fairly straightforward to implement, although some confusion existed over a few of the details. Particularly, problems occurred with so called *end effects*—errors arising at the edges of the dataset. Also, in certain cases where very low amplitude waves occured in the middle of a higher amplitude wavetrain the method tended to exhibit artefacts, although these were deemed of little significance for MD trajectory analysis, and a strategy for removing them from spectra was decided upon.

The stability of the algorithm to small perturbations in the input signal was investigated and found to be low. The conjecture of Huang *et al.* that a spline fitting algorithm which obtains a tighter fit to the data would improve the algorithm's stability was tested with the Akima spline. No improvement was observed.

Methods of dealing with the end effects resulting from the Hilbert transform were investigated; several windowing functions were evaluated and the optimum choice decided upon. Strategies for using this function were discussed.

A simple Brownian dynamics test system was implemented and its trajectories studied with both the HHT and the Morlet wavelet transform. The results were broadly comparable, and peak centre frequencies obtained from the methods were compared. It was found that the HHT offered a small improvement on the wavelet and Fourier transforms. The HHT marginal spectrum also seemed to have a better ability to discriminate transition events than the equivalent Morlet wavelet spectrum.

A method for improving the coverage of time-frequency space by compositing large numbers of HHT spectra was developed. This resulted in plots that are easier to interpret than the conventional HHT, and more comparable to wavelet spectra. The drawbacks of this technique are the subjective judgement of the quantity of noise to add to the spectra, and the increased processing time required relative to the original method.

HHT spectra of white noise were examined and found to differ qualitatively from the results obtained from Fourier transforms. The reasons for this were deduced and a method was proposed to modify HHT spectra to make them more comparable with Fourier spectra. Although the results obtained look very promising, further work on the method is needed before it can be applied more generally.

Chapter 5 concluded that the HHT transform is an interesting and useful development, and some of the results obtained are an improvement on other methods. The method is not without problems, though. The principal difficulty lies in the interpretation of signals which do not possess a clear separation of scales of motion, or where discrete transition events occur. Although the exact interpretation of the frequencies of these latter events is problematic, it seems fairly clear that the method is successful at *detecting* their existence, although for the data studied thus far HHT seemed about evenly matched with the Morlet wavelet transform in this respect.

In Chapter 6 a set of methods were developed that made use of both the HHT and wavelet methods to generate a statistical description of the time-frequency-amplitude space sampled by various degrees of freedom in protein simulations. These methods were applied to a series of simulations of the DHFR molecule prepared by other workers.

The results showed that the methods developed are useful, and allow a large amount of information about the dynamical properties of the enzyme to be represented in a small number of plots. Specific issues identified were the selection of useful degrees of freedom to process, the difficulty of representing very dense datasets in two dimensional plots, particularly on paper, and the problems of scaling datasets obtained by the HHT method to obtain meaningful representaions.

Mean amplitude versus frequency and maximum amplitude versus time plots were created using both the HHT and wavelet methods. The wavelet transform version of the mean amplitude versus frequency method showed interesting details of the fine vibrational structure of the protein, but these seemed of limited practical value for the DHFR data, although it is possible that they might be of more use for discriminating the behaviour observed in other protein simulations. The HHT version of the mean amplitude versus frequency plots revealed none of the fine details seen with the wavelet transform, but the ability of the HHT to probe lower frequencies was seen to be useful: hydrogen bonds implicated in the conformational change events were highlighted clearly. The utility of the low frequency probing abilities of the HHT transform were even more starkly highlighted in the maximum amplitude versus time plots: although the sensitivity of the wavelet version of this method was fairly poor, the HHT version was able to identify key hydrogen bonds that broke during the conformational transition events very clearly. The HHT maximum amplitude versus time technique seems to be the most promising of the new methods.

In summary, some interesting observations were made from the simulations of the prion protein; but this was in spite of, rather than because of, the analysis methods available. Problems identified with these tools lead us to consider alternative frequency-amplitude-time methods for the analysis and interpretation of simulation data, and the recent Hilbert Huang transform method was seen as an interesting possibility for development. This method was duly investigated, some improvements were proposed, and although a number of issues were raised over the interpretation of the data obtained, it was considered to be broadly useful. A set of new simulation analysis techniques based on the HHT and wavelet transforms was developed, and tested on RDFMD simulations of the DHFR enzyme. The methods revealed some interesting structure within the simulations, and also succeeded in highlighting where conformational change events took place in the trajectories. Although these methods appear to be useful, further work is required.

This thesis has identified a number of areas where further work would seem indicated. First, and very generally, far more work is needed in the whole field of protein simulation analysis. More varied and subtle tools would greatly facilitate our understanding of the dynamical effects critical to their biological functions.

A possible alternative to the HHT and wavelet approach to analysis described in Chapter 6 would be to process the trajectories directly with simple RMSD statistics. A significant difference between the frequency-time statistical treatment and a conventional RMSD route is that the model presented here works by describing changes in derivatives of the signal, and so highlights change *events* directly. Also, as has been shown (in Figure 6.17), conformational change events are detected by the HHT as occuring at up to *fifty* standard deviations from the mean, and this provides clearer results than might be expected from RMSD analysis. In particular, hydrogen bonds are either present, and the bonding distance almost constant, or absent, in which case the distance between the atoms being considered will often wander over quite a wide range, as seen in the prion β-sheet. Also, dihedral angles, which may rotate continuously throughout the 0–2π range, might be expected to have an almost arbitrary mean value and standard deviation, and this would make it difficult to detect events of interest using conventional statistics. For completeness then, it might perhaps be interesting to evaluate such a statistical approach, although it would seem unlikely that the results would be as clean, and obviously the extra information that could be obtained from frequency specific methods would be lost.

It would also seem that our simulations of the prion protein could usefully be reanalysed using some of the techniques devised in the later chapters of this thesis. For the HHT method, further application and refinement of the compositing technique may be of value; also, the normalisation approach developed using the averaged HHT marginal spectrum might make HHT results more directly comparable with Fourier transforms, simplifying their interpretation and perhaps alleviating the problems identified with scaling the plots obtained. Improvements to the statistical methods, besides resolving the scaling issue, might come from processing alternative degrees of freedom. One possibility for this is to obtain a covariance eigenvector basis set, and process the projections of the simulation onto the 'essential' eigenvectors that represent the bulk of the trajectory's variance. This would result in a dramatic reduction in the amount of data to be processed and represented. Another interesting avenue is the possibility of refining the statistical techniques to process large quantities of simulation data automatically, with the minimum of operator intervention. Automated detection of unusual or interesting conformational change events in large volumes of simulation data would be beneficial for highly parallel simulations, such as network distributed projects or parallel tempering methods. Besides these things, though, the most important avenue for further work is simply to apply the methods developed to many more protein simulations and so gain experience of the results that they are capable of obtaining.

Appendix A

Cauchy Principal Value

An integral of the form

$$\int_P^Q f(x)\,dx \tag{A.1}$$

where $f(x)$ tends to infinity at some point s such that $P < s < Q$ must be determined using limits that avoid the asymptote. One method of doing this is as follows

$$\int_P^Q f(x)\,dt = \lim_{\alpha\to 0}\left[\int_P^{s-\alpha} f(x)\,dx + \int_{s+\alpha}^Q f(x)\,dx\right] \tag{A.2}$$

where α tends to zero through positive values. This is called the *Cauchy principal value* of the integral

$$\text{pr.v.}\int_P^Q f(x)\,dx \tag{A.3}$$

The principal value sometimes exists and can be calculated even though the original integral is meaningless and cannot.[94]

Appendix B

Signal Energy

A simple harmonic oscillator with force constant k and amplitude A has energy

$$E = \frac{1}{2}kA^2 \tag{B.1}$$

If ν is the frequency of vibration and m is the mass of the moving particle then

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{B.2}$$

and

$$k = (2\pi\nu)^2 m \tag{B.3}$$

But a single harmonic oscillator has only one frequency. In the more general case where the frequency of a moving particle may change, or more than one frequency may exist, there is a continuous distribution of frequencies, each with a corresponding force constant, k. Substituting B.3 into B.1:

$$\begin{aligned} E(\nu) &= \frac{1}{2}k(\nu)A^2(\nu) \\ &= \frac{1}{2}(2\pi\nu)^2 mA^2 \\ &= 2\pi^2 m\nu^2 A^2 \end{aligned} \tag{B.4}$$

Applying this expression as it stands must be done with care, as the mass value is not always clear. The absolute values obtained from separate spectra of degrees of freedom with different mass values are not strictly comparable *as energies*.

Bibliography

[1] Mathews, C. K. and van Holde, K. E. Biochemistry. Benjamin Cummings, 1990.

[2] Rod, T. H., Radkiewicz, J. L., and Brooks, C. L. Correlated motion and the effect of distal mutations in dihydrofolate reductase. Proc. Natl. Acad. Sci. USA 100:6980–6985, 2003.

[3] Kelly, J. W. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. Curr. Op. Struct. Bio. 8:101–106, 1998.

[4] Wlodek, S. T., Shen, T., and McCammon, J. A. Electrostatic steering of substrate to acetylcholinesterase: Analysis of field fluctuations. Biopolymers 53:265–271, 2000.

[5] Tai, K., Shen, T., Borjession, U., Phillipopoulos, M., and McCammon, J. A. Analysis of 10 ns molecular dynamics simulation of acetylcholinesterase. Biophys. J. 81:715–724, 2001.

[6] Luedemann, S., Gabdoulline, R., Lounnas, V., and Wade, R. Substrate access to cytochrome p450cam investigated by molecular dynamics simulations: An interactive look at the underlying mechanisms. Internet J. Chem 4:6, 2001.

[7] Allen, M. P. and Tildesley, D. J. Computer Simulation of Liquids. Oxford University Press, 1987.

[8] McCammon, J. A., Gelin, B. R., and Karplus, M. Dynamics of folded proteins. Nature 267:585–590, 1977.

[9] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalor, I. N., and Bourne, P. E. The protein data bank. Nucleic Acids Research 28:235–242, 2000.

[10] Koradi, R., Billeter, M., and Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. J. Mol. Graphics 14:51–55, 1996.

[11] Humphrey, W., Dalke, A., and Schulten, K. VMD—visual molecular dynamics. J. Molec. Graphics 14:33–38, 1996.

[12] Leach, A. R. Molecular Modelling. Prentice Hall, 2001.

[13] Darden, T., York, D., and Pedersen, L. Particle mesh Ewald: a N log(N) method for Ewald sums in large systems. J. Chem. Phys. 98:10089–10092, 1993.

[14] Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Computer Phys. 23:327–341, 1977.

[15] Case, D. A., Pearlman, D. A., Caldwell, J. W., Chetham III, T. E., Ross, W. S., Simmerling, C. L., Darden, K. M., Merz, R. V., Stanton, R. V., Cheng, A. L., Vincent, J. J., Crowley, M., Ferguson, D. M., Radmer, R. J., Seibel, G. L., Singh, U. C., Weiner, P. K., and Kollman, P. A. AMBER 5.0. University of California, San Francisco, 1997.

[16] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz Jr, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 117:5179–5197, 1995.

[17] Allinger, N. L., Chen, K., and Lii, J.-H. An improved force field (MM4) for saturated hydrocarbons. J. Comp. Chem. 17:642–668, 1996.

[18] Kabsch, W. and Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.

[19] Parchment, O. G. and Essex, J. W. Molecular dynamics of mouse and syrian hamster prion PrP: Implications for activity. Proteins, Struct. Func. Genet. 38:327–340, 2000.

[20] Ichiye, T. and Karplus, M. Collective motions in proteins: a covariance analysis of the atomic fluctuations in molecular dynamics and normal mode simulations. Proteins, Struct. Func. Genet. 11:205–217, 1991.

[21] Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C. Essential dynamics of proteins. Proteins, Struct. Func. Genet. 17:412–425, 1993.

[22] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 81:3684–3690, 1984.

[23] Sawaya, M. R. and Kraut, J. Loop and subdomain movements in the mechanism of escherichia coli dihydrofolate reductase: Crystallographic evidence. Biochemistry 36:586–603, 1997.

[24] Burnett, M. N. and Johnson, C. K. ORTEP-III: Oak ridge thermal ellipsoid plot program for crystal structure illustrations. Oak Ridge National Laboratory Report ORNL-6895, 1996.

[25] Swaminathan, S., Harte, W. E., and Beveridge, D. L. Investigation of domain structure in proteins via molecular dynamics simulation: application to HIV-1 protease dimer. J. Am. Chem. Soc. 113:2717–2721, 1991.

[26] Atkins, P. W. and Friedman, R. S. Molecular Quantum Mechanics. Oxford University Press, 1997.

[27] Hayward, S., Kitao, A., and Berendsen, H. J. C. Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. Proteins, Struct. Func. Genet. 27:425–437, 1997.

[28] Levy, R. M., Rojas, O. L., and A., F. R. Quasi-harmonic method for calculating vibrational spectra from classical simulations on multidimensional anharmonic potential surfaces. J. Phys. Chem. 88:4233–4238, 1984.

[29] Kitao, A. and Go, N. Investigating protein dynamics in collective coordinate space. Curr. Op. Struct. Biol. 9:164–169, 1999.

[30] Verma, C. S., Caves, L. S. D., Hubbard, R. E., and Roberts, G. C. K. Domain motions in dihydrofolate reductase: A molecular dynamics study. J. Mol. Biol. 266:776–796, 1997.

[31] Peters, G. H., Frimurer, T. M., Andersen, J. N., and Olsen, O. H. Molecular dynamics simulations of protein-tyrosine phosphatase 1B. I. ligand-induced changes in protein motions. Biophys. J. 77:505–515, 1999.

[32] Berendsen, H. J. C. and Hayward, S. Collective protein dynamics in relation to function. Curr. Op. Struct. Biol. 10:165–169, 2000.

[33] Gazit, E. The “Correctly Folded” state of proteins: Is it a metastable state? Angew. Chem. Int. Ed. 41(2):257–259, 2002.

[34] Pan, K. M., Baldwin, M. A., Nguyen, J., Gasset, M., Serban, A., Groth, D., Mehlhorn, I., Huang, Z., Fletterick, R. J., Cohen, F. E., and Prusiner, S. B. Conversion of α-helices into β-sheets features in the formation of scrapie prion proteins. Proc. Natl. Acad. Sci. USA 90:10962–10966, 1993.

[35] Safar, J., Roller, P. P., Gajdusek, D. C., and Gibbs, C. J. Thermal stability and conformational transitions of scrapie amyloid (prion) protein correlate with infectivity. Protein Sci. 2:2206–2216, 1993.

[36] Prusiner, S. B. Prions. Proc. Natl. Acad. Sci. USA 95:13363–13383, 1998.

[37] Griffith, J. S. Self-replication and scrapie. Nature 215:1043–1044, 1967.

[38] Viles, J. H., Cohen, F. E., Prusiner, S. B., Goodin, D. B., Wright, P. E., and Dyson, H. J. Copper binding to the prion protein: Structural implications of four identical cooperative binding sites. Proc. Natl. Acad. Sci. USA 96:2042–2047, 1999.

[39] Vey, M., Pilkuhn, S., Wille, H., Nixon, R., DeArmond, S. J., Smart, E. J., Anderson, R. G. W., Taraboulos, A., and Prusiner, S. B. Subcellular colocalization of the cellular and scrapie prion proteins in caveolae-like membranous domains. Proc. Natl. Acad. Sci. USA 93:14945–14949, 1996.

[40] Kaneko, K., Zulanello, L., Scott, M., Cooper, C. M., Wallace, A. C., James, T. L., Cohen, F. E., and Prusiner, S. B. Evidence for protein X binding to discontinuous epitope on the cellular prion protein during scrapie prion propagation. Proc. Natl. Acad. Sci. USA 94:10069–10074, 1997.

[41] Baker, H. F., Poulter, M., Crow, T. J., Frith, C. D., Lofthouse, R., Ridley, R. M., and Collinge, J. Aminoacid polymorphism in human prion protein and age at death in inherited prion disease. Lancet 337:1286, 1991.

[42] Goldfarb, L. G., Petersen, R. B., Tabaton, M., Brown, P., LeBlanc, A. C., Montagna, P., Cortelli, P., Julien, J., Vital, C., Pendlebury, W. W., Haltia, L., Wills, P. R., Hauw, J. J., McKeever, P. E., Monari, L., Schrank, B., Swergold, G. D., Autilio-Gambetti, L., Gajdusek, D. C., Lugaresi, E., and Gambetti, P. Fatal familial insomnia and familial Creutzfeld-Jakob disease: Disease phenotype determined by a DNA polymorphism. Science 258:806–808, 1992.

[43] Liemann, S. and Glockshuber, R. Influence of amino acid substitutions related to inherited human prion diseases on the thermodynamic stability of the cellular prion protein. Biochemistry 38:3258–3267, 1999.

[44] Alperovitch, A., Zerr, I., Pocchiari, M., Mitrova, E., de Pedro Cuesta, J., Hegyi, I., Collins, S., Kretzschmar, H., van Duijn, C., and Will, R. G. Codon 129 prion protein genotype and sporadic Creutzfeld-Jakob disease. Lancet 353:1673–1674, 1999.

[45] Riek, R., Hornemann, S., Wider, G., Billeter, M., Glockshuber, R., and Wüthrich, K. NMR structure of the mouse prion protein binding domain PrP(121–231). Nature 382:180–182, 1996.

[46] James, T. L., Liu, H., Ulyanov, N. B., Farr-Jones, S., Zhang, H., Donne, D. G., Kaneko, K., Groth, D., Mehlhorn, I., Prusiner, S. B., and Cohen, F. E. Solution structure of a 142 residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. Proc. Natl. Acad. Sci. USA 94:10086–10091, 1997.

[47] Garcia, F. L., Zahn, R., Riek, R., and Wüthrich, K. NMR structure of the bovine prion protein. Proc. Natl. Acad. Sci. USA 97:5422–5427, 2000.

[48] Zahn, R., Liu, A., Lührs, T., Riek, R., von Schroetter, C., Garcia, F. L., Billeter, M., Calzolai, L., Wider, G., and Wüthrich, K. NMR structure of the human prion protein. Proc. Natl. Acad. Sci. USA 97:145–150, 2000.

[49] Alonso, D. O. V., DeArmond, S. J., Cohen, F. E., and Dagget, V. Mapping the early steps in the pH-induced conformational conversion of the prion protein. Proc. Natl. Acad. Sci. USA 98(6):2985–2989, 2001.

[50] Zuegg, J. and Gready, J. E. Molecular dynamics simulations of human prion protein: Importance of correct treatment of electrostatic interactions. Biochemistry 38(42):13862–13876, 1999.

[51] Zuegg, J. and Gready, J. E. Molecular dynamics simulation of human prion protein including both N-linked oligosaccharides and the GPI anchor. Glycobiology 10(10):959–974, 2000.

[52] El-Bastawissy, E., Knaggs, M. H., and Gilbert, I. H. Molecular dynamics of wild-type and point mutation human prion protein at normal and elevated temperature. J. Mol. Graph. Mod. 20:145–154, 2001.

[53] Gsponer, J., Ferrara, P., and Caflisch, A. Flexibility of the murine prion protein and its asp178asn mutant investigated by molecular dynamics simulations. J. Mol. Graphics. 20:145–154, 2001.

[54] Jorgensen, W. L., Chandrasekar, J., Madura, J. D., Imprey, R. W., and Klein, M. L. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79:926–935, 1983.

[55] Zhu, H. and Braun, W. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β-sheet formation in proteins. Protein Science 8:326–342, 1999.

[56] Hunenberger, P. H., Mark, A. E., and van Gunsteren, W. F. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. J. Mol. Biol. 252:492–503, 1995.

[57] Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G., and Dobson, C. M. Designing conditions for *in vitro* formation of amyloid protofilaments and fibrils. Proc. Natl. Acad. Sci. USA 96:3590–3594, 1999.

[58] Richardson, J. S. and Richardson, D. C. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc. Natl. Acad. Sci. USA. 99:2754–2759, 2002.

[59] Phillips, S. C., Gledhill, R. J., Essex, J. W., and Edge, C. M. Application of the Hilbert-Huang transform to the analysis of molecular dynamics simulations. J. Phys. Chem. A 107(24):4869–4876, 2003.

[60] Phillips, S. C., Essex, J. W., and Edge, C. M. Digitally filtered molecular dynamics: The frequency specific control of molecular dynamics simulations. J. Chem. Phys. 112(6):2586–2597, 2000.

[61] Phillips, S. C., Swain, M. T., Wiley, A. P., Essex, J. W., and Edge, C. M. Reversible digitally filtered molecular dynamics. J. Phys. Chem. B 107:2098–2110, 2003.

[62] Huang, N. E., Shen, Z., Long, S. R., Wu, M. L. C., Shih, H. H., Zheng, Q. N., Yen, N. C., Tung, C. C., and Liu, H. H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. London Ser. A—Math. Phys. Eng. Sci. 454(1971):903–995, 1998.

[63] Cooley, J. W. and Tukey, J. W. An algorithm for the machine calculation of complex Fourier series. Math. Comp. 19:297–301, 1965.

[64] Harris, F. J. On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE 66(1):51–83, 1978.

[65] Torrence, C. and Compo, G. P. A practical guide to wavelet analysis. Bull. Am. Met. Soc. 79:61–78, 1998.

[66] Leung, A. K., Chau, F., and Gao, J. A review on applications of wavelet transform techniques in chemical analysis: 1989-1997. Chemometrics Int. Lab. Sys. 43:165–184, 1998.

[67] Li, Z., Borrmann, A., and Martens, C. C. Wavelet analysis of condensed phase molecular dynamics. Chem. Phys. Lett. 214:362–366, 1993.

[68] Askar, A., Cetin, A. E., and Rabitz, H. Wavelet transform for analysis of molecular dynamics. J. Phys. Chem. 100:19165–19173, 1996.

[69] Cohen, L. Time Frequency Analysis. Prentice Hall, 1995.

[70] Vakman, D. E. and Vaǐnsshteǐn, L. A. Amplitude, phase, frequency—fundamental concepts of oscillation theory. Sov. Phys. Usp. 20:1002–1016, 1977.

[71] Carson, J. R. and Fry, T. C. Variable frequency electric circuit theory with application to the theory of frequency-modulation. Bell System Tech. J. 16:513–540, 1937.

[72] Bendat, J. S. The Hilbert Transform and Applications to Correlation Measurements. Brüel and Kjær, 1985.

[73] Boashash, B. Estimating and interpreting the instantaneous frequency of a signal—part 1: Fundamentals. Proc. IEEE 80:519–538, 1992.

[74] Hasegawa, T. and Torii, T. An automatic quadrature for Cauchy principal value integrals. Mathematics of Computation 56:741–754, 1991.

[75] Huang, N. E., Shen, Z., and Long, S. R. A new view of nonlinear water waves: the Hilbert spectrum. Annu. Rev. Fluid Mech. 31(1):417–457, 1999.

[76] Pan, J., Yan, X.-H., Zheng, Q., Liu, W. T., and Klemas, V. V. Interpretation of scatterometer ocean surface wind vector EOFs over the northwestern pacific. Remote Sensing of Environment 84:53–68, 2002.

[77] Huang, N. E., Long, S. R., and Shen, Z. The mechanism for frequency downshift in nonlinear wave evolution. Adv. Appl. Mech. 32:59–117, 1996.

[78] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. Numerical Recipes in Fortran 77, volume 1. Cambridge University Press, 1992.

[79] Huang, W., Shen, Z., Huang, N. E., and Fung, Y. C. Engineering analysis of biological variables: An example of blood pressure over 1 day. Proc. Natl. Acad. Sci. USA 95:4816–4821, 1998.

[80] Lai, Y.-C. Analytic signals and the transition to chaos in deterministic flows. Phys. Rev. E 58:R6911–R6914, 1998.

[81] Echeverría, J. C., Crowe, J. A., Woolfson, M. S., and Hayes-Gill, B. R. Application of empirical mode decomposition to heart rate variability analysis. Med. Biol. Eng. Comput. 39:471–479, 2001.

[82] Komm, R. W., Hill, F., and Howe, R. Empirical mode decomposition and Hilbert analysis applied to rotation residuals of the solar convection zone. Astrophys. J. 558:428–441, 2001.

[83] Niethammer, M., Jacobs, L. J., Qu, J., and Jarzynski, J. Time-frequency representations of Lamb waves. J. Acoust. Soc. Am. 109:1841–1847, 2001.

[84] Chunming, H., Huadong, G., Changlin, W., and Dian, F. A novel method to reduce speckle in SAR images. Int. J. Remote Sensing 23:5095–5101, 2002.

[85] Montesinos, M. E., Muñoz Cobo, J. L., and Pérez, C. Hilbert-Huang analysis of BWR neutron detector signals: application to DR calculation and to corrupted signal analysis. Annals of Nuclear Energy 30:715–727, 2003.

[86] Savitsky, A. and Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36(8):1627–1639, 1964.

[87] Lorenz, E. N. Deterministic nonperiodic flow. J. Atmos. Sci. 20:130–141, 1963.

[88] Akima, H. A new method of interpolation and smooth curve fitting based on local procedures. Communications of the ACM 17(4):589–602, 1970.

[89] Akima, H. Interpolation and smooth curve fitting based on local procedures. Communications of the ACM 15(10):914–918, 1972.

[90] Ermak, D. L. and Buckholtz, H. Numerical integration of the Langevin equation: Monte carlo simulation. J. Comput. Phys. 35:169–182, 1980.

[91] Huennekens, F. M. In search of dihydrofolate reductase. Protein Science 5:1201–1208, 1996.

[92] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4:187–217, 1983.

[93] MacKerell, Jr., A. D., Bashford, D., Bellott, M., Dunbrack Jr., R., Evanseck, J., Field, M., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F., Mattos, C., Michnick, S., Ngo, T., Nguyen, D., Prodhom, B., Reiher, III, W.,

Roux, B., Schlenkrich, M., Smith, J., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B 102:3586–3616, 1998.

[94] Kreyszig, E. Advanced Engineering Mathematics. Wiley, 1988.