Theoretical studies of time-resolved spectroscopy of protein folding

Jonathan D. Hirst,*[a] Samita Bhattacharjee[a] and Alexey V. Onufriev[b]

[a] *School of Chemistry, University of Nottingham, University Park, Nottingham, UK NG7 2RD*

[b] *Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla CA 92037, USA*

Received 24th January 2002, Accepted 7th March 2002
First published as an Advance Article on the web 16th July 2002

Recently, we have made significant improvements in the accuracy of calculations of the circular dichroism of proteins from first principles. The quality of these calculations (especially at 220 nm, a key wavelength, where the intensity of the band correlates well with the helical content of polypeptides) has given us confidence to use such calculations to analyse nanosecond molecular dynamics simulations of the folding of polypeptides. We use this combined approach to explore the influence of dynamics on the circular dichroism spectroscopy of polypeptides. We apply it to equilibrium molecular dynamics simulations of two β-sheet proteins with similar structures, but differing circular dichroism spectra. We analyse a molecular dynamics simulation of the acid-unfolding of myoglobin. For both α-helical and β-sheet conformations, we find that changes in dihedral angles of 30° can change intensities of bands in circular dichroism spectra by up to 5000 degree cm^2 $dmol^{-1}$. Thus, in isolation, moderate differences in circular dichroism spectra cannot be interpreted uniquely in terms of conformational changes. Examination of individual structures allows us to dissect the influence of conformation on the calculated circular dichroism spectra. Our results are aimed at providing a deeper understanding of the optical properties of proteins. An atomic level connection between molecular dynamics simulations and optical spectroscopy is increasingly desirable as theoretical and experimental studies begin to probe protein folding events reliably on the nanosecond timescale.

Introduction

Protein folding is generally rapid and strongly co-operative.[1,2] Knowledge of protein folding pathways and structural characterisation of the states that occur along them are necessary for a thorough understanding of folding. Such an understanding would have an immediate practical impact, as folding and unfolding participate in the control of a variety of cellular processes, such as cross-membrane transport of proteins and cell cycle regulation.[3] The transient nature of intermediates has limited the understanding of the folding process. Intermediate states undergo much larger structural fluctuations than native states, which makes it difficult to resolve their structures fully using techniques such as X-ray crystallography or NMR. In this regard, a quantitative understanding of the relationship between protein conformation and circular dichroism (CD) spectra would be a valuable tool.

CD is the differential absorption of left and right circularly polarised light. It arises from the asymmetry of a chromophore or its environment, and thus provides a tool for measuring both conformation and changes in conformation for proteins and peptides. For decades CD spectroscopy has been an important method used by biochemists to analyse structures in globular proteins[4,5] and it can reveal greater detail than techniques such as routine UV absorption and fluorescence spectroscopies. Most conventional UV absorption and fluorescence studies of proteins provide only qualitative information about tertiary structural changes in the micro-environments of aromatic residues. In contrast, CD spectra provide distinct information about both tertiary and secondary structure and, unlike UV absorption, particular CD spectral features characterise the specific types of secondary structures present in proteins. In particular, CD measurements in the far-UV detect transitions involving primarily the peptide chromophore and are sensitive to secondary structure conformation.[4,5] For example, an α-helical CD spectrum consists of a positive band at 190 nm and two negative bands at 208 nm and 220 nm,[6] whereas the β-sheet proteins have a maximum at 195 nm and a minimum in the region 210–220 nm.

In addition to illuminating equilibrium studies, a detailed connection between the atomic structure of polypeptides and their CD spectra would also have an impact on time-resolved CD studies of protein folding. Several experimental techniques[7] now have the time resolution to follow early events in protein folding on nanosecond time scales.[8] Schemes that rapidly photoinitiate folding through laser temperature jump methods[9] or electron transfer[10] in real-time CD studies are now being pursued by several groups. Meanwhile, the nanosecond time regime is increasingly accessible to molecular dynamics simulations.[11,12] This convergence of experiment and theory makes the development of quantitative protein CD calculations particularly timely. Time-resolved CD is usually limited to a single wavelength, often 220 nm, to monitor helix content. The technique is widely applied, and it is therefore important to understand what factors affect the ellipticity at 220 nm. Thus a detailed understanding of the relationship between conformation, fluctuations and the measured CD spectra will be necessary for the fullest interpretation of these experiments. Clearly the experimental CD intensity of a single band provides only limited information, and we use the combination of CD calculations and molecular dynamics simulations to investigate which mechanisms of protein unfolding would be consistent with the experimentally measured changes in CD.

Although CD is sensitive to protein conformation, its interpretation has been largely empirical,[13] *i.e.*, based on comparison with the CD spectra of proteins of known three-dimensional structure.[14] The validity and success of this approach rests on the structural similarity of different proteins in their native conformations. Whether non-native folding intermediates are sufficiently similar in conformation to fully folded proteins that their CD spectra can be analysed in the same way as those of native proteins is more of an open question. If we could accurately calculate the CD spectra arising from different conformations, we would be much better placed to interpret CD experiments on non-native conformations of peptides and proteins. Recently, we have made encouraging advances in the accuracy of calculations of the CD of proteins from first principles,[15,16] based on modern quantum chemical calculations on the amide group combined with a continuum model of solvent effects. Calculations of similar quality have also been realized by others.[17,18] We begin this paper with some additional benchmarking of our first principles approach against empirical approaches for estimating helicity, using the X-ray crystal structures of 29 proteins. Having established the quality of the first principles calculations on static structures, we investigate the possibility that equilibrium dynamics in solution of some β-sheet-containing proteins may influence their CD spectra.

The CD spectra of β-sheet proteins fall into two classes.[19,20] Class I β-sheet protein (β-I) exhibit a negative band at 216–218 nm and a positive band at 195 nm. The CD spectra of class II β-sheet proteins (β-II) resemble those of random coil models, dominated by an intense negative band near 198 nm. The β-II proteins tend to contain some disulfide bridges (which may contribute to the CD in the far-UV)[21] and have β-sheets that are more irregular than the β-I proteins, reflected by larger content of β-bulge structures. However, the structural origins of the CD spectra of β-II proteins are not fully understood. Calculations of CD based on static structures are unable to distinguish between β-I and β-II proteins, predicting β-I spectra for both classes. To investigate the CD calculations on β-I and β-II proteins further, we analyse the equilibrium dynamics of concanavalin A (a β-I protein) and elastase (a β-II protein). Schematic representations of the structures of these two proteins along with myoglobin are shown in Fig. 1.

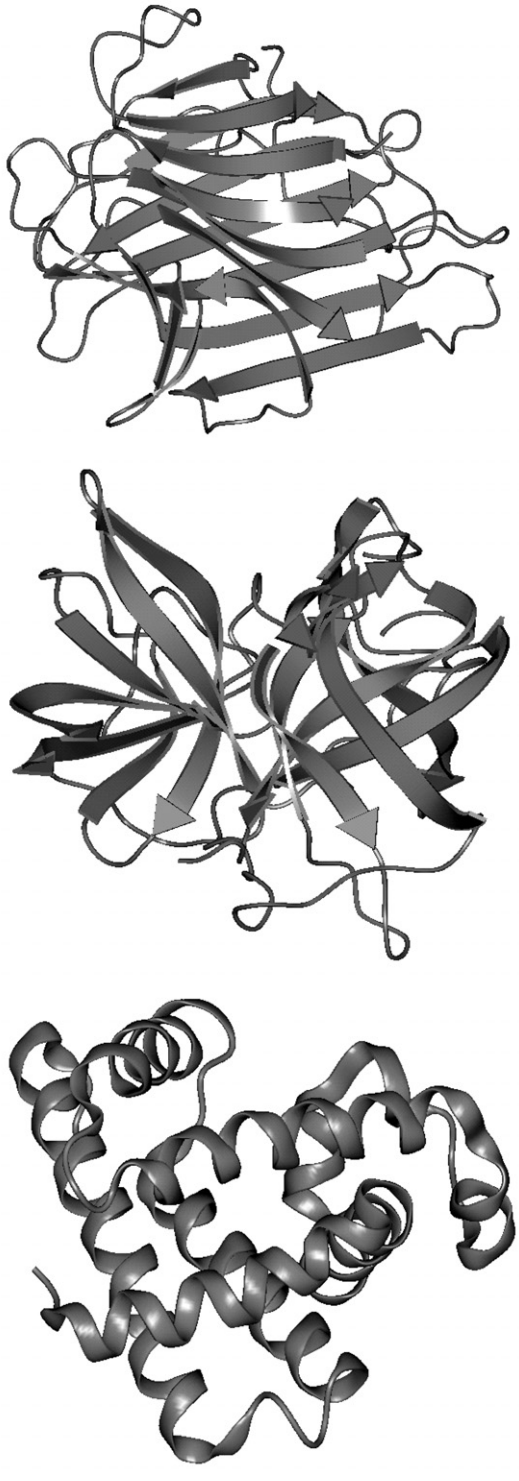

Fig. 1 Ribbon representations, rendered using PREPI (Dr Suhail Islam, Imperial College, London), of the structures of the proteins (PDB accessions codes in parentheses) whose dynamics have been studied: upper – concanavalin A (3cna); centre – elastase (3est); lower – myoglobin (2mb5).

Finally, we explore a non-equilibrium process by calculations of CD based on structures from molecular dynamics simulations of the unfolding of apomyoglobin. Myoglobin has at least two non-native apo states whose structures have been characterised using NMR and CD. These partially unfolded stable states may be good models of folding intermediates, and have attracted much attention from experimentalists and theorists. One difficulty in connecting simulation data to CD experiments is the issue of which conformations contribute to the CD spectrum, especially at 220 nm where the intensity is taken to reflect the helical content of the folding intermediates. Typically, residues are designated 'helical' based on local main-chain dihedral angles and the conformational states of neighbouring residues in any empirical calculations, whereas our calculations of CD from first principles avoid *ad hoc* definitions.

Nanosecond and increasingly microsecond time scales are accessible to molecular dynamics simulations of protein folding and unfolding, providing exquisite detail of these microscopic events.[22] The credibility of such simulations is critical, resting partly upon the quality of force field development, partly on technical issues of sampling and protocol and, perhaps most directly, upon the reproduction of key experimental data. Clear connections to experimental data obviously enhance confidence in molecular dynamics simulations and the atomic level picture that they paint.

Methods

CD calculations

Calculations of the electronic structure of polypeptides are well beyond the scope of fully *ab initio* treatments. The most common method of computing the CD spectra of polypeptides is the matrix method,[23,24] where the excited states of individual chromophores are described quantum mechanically and interactions between these states are computed classically, based on parametrizations of the *ab initio* calculations on the individual chromophores. The intensities of the bands are derived directly from the electronic and magnetic transition dipole moments through the rotational strengths corresponding to each excited state of the polypeptide. The rotational strength can be expressed as the imaginary part of the product of the electronic and magnetic transition dipole moments using the Rosenfeld equation.[25] For an electronic transition $i \leftarrow 0$, the rotational strength R_{0i} is given by

$$R_{0i} = \mathrm{Im}(\langle\psi_0|\boldsymbol{\mu}_\mathrm{e}|\psi_i\rangle)\cdot(\langle\psi_i|\boldsymbol{\mu}_\mathrm{m}|\psi_0\rangle) \quad (1)$$

where Im denotes "the imaginary part of", ψ_0 is the ground state wave function, ψ_i is the excited state wave function and $\boldsymbol{\mu}_\mathrm{e}$ and $\boldsymbol{\mu}_\mathrm{m}$ are the electronic and magnetic transition dipole moments, respectively.

In the matrix method, a polypeptide is treated as a collection of M non-interacting chromophoric groups. Electronic excitations may occur only within a group but not between the groups. The excited-state wave function of the whole molecule is expressed as a linear combination of basis functions Φ_{ia} involving the n_i excitations within each chromophoric group:

$$\Psi_\mathrm{T} = \sum_i^M \sum_a^{n_i} c_{ia}\Phi_{ia} \quad (2)$$

Each basis function is a product of M monomer wave functions. The basis set is further restricted to allow only one group to be excited. Thus:

$$\Phi_{ia} = \varphi_{10}\cdots\varphi_{ia}\cdots\varphi_{j0}\cdots\varphi_{M0} \quad (3)$$

where φ_{ia} represents the wave function of chromophore i, which has undergone an electronic excitation $a \leftarrow 0$. In general, each transition from the ground state to one of the excited states may have a nonzero rotational strength at its particular transition energy, and the CD spectrum is the sum of all these rotational strengths.

A Hamiltonian matrix is then constructed. The diagonal elements of this matrix are the excitation energies of the single chromophores, and the off-diagonal elements describe the interactions between different chromophoric groups. If the latter are assumed to be purely Coulombic in nature,

then the off-diagonal elements are computed from the electrostatic interaction between charge densities and have the form:

$$V_{i0a;j0b} = \int\int \frac{\rho_{i0a}(\boldsymbol{r}_{i1})\rho_{j0b}(\boldsymbol{r}_{j1})}{4\pi\varepsilon_0 r_{i1,j1}} \mathrm{d}\boldsymbol{r}_{i1}\mathrm{d}\boldsymbol{r}_{j1} \tag{4}$$

where ρ_{i0a} and ρ_{j0b} represent the permanent (when $a = 0$ or $b = 0$) and transition electron densities on chromophores i and j, respectively. Thus, the matrix method requires parameters that describe the above charge distributions associated with the different electronic states of the chromophoric groups of the protein. In the present study, these parameters were taken from calculations on *N*-methylacetamide (NMA) in solution using the complete-active-space self-consistent field method implemented within a self-consistent reaction field (CASSCF/SCRF).[26–29] The interaction potentials were evaluated by representing the charge densities with a set of point charges (or monopoles). The point charges were fitted to reproduce the *ab initio* electrostatic potential arising from the various states,[30] thereby improving the representation of the monomer.

In the calculations reported here, only the amide electronic transitions $n\pi^*$ (at 220 nm) and $\pi\pi^*$ (at 193 nm) are included. In such a case, for a diamide (considered here solely for illustrative purposes) the Hamiltonian matrix takes the form

$$\boldsymbol{H} = \begin{pmatrix} E^1_{n\pi^*} & V^{11}_{n\pi^*\pi\pi^*} & V^{12}_{n\pi^*n\pi^*} & V^{12}_{n\pi*\pi\pi^*} \\ V^{11}_{n\pi^*\pi\pi^*} & E^1_{\pi\pi*} & V^{21}_{n\pi^*\pi\pi^*} & V^{12}_{\pi\pi^*\pi\pi^*} \\ V^{12}_{n\pi^*n\pi^*} & V^{21}_{n\pi^*\pi\pi^*} & E^2_{n\pi^*} & V^{22}_{n\pi^*\pi\pi^*} \\ V^{12}_{n\pi^*\pi\pi^*} & V^{12}_{\pi\pi^*\pi\pi^*} & V^{22}_{n\pi^*\pi\pi^*} & E^2_{\pi\pi^*} \end{pmatrix} \tag{5}$$

Diagonalization of the above matrix by a unitary transformation yields the eigenvalues and eigenvectors of the composite transitions of the protein. The eigenvalues are the energies of the transitions of the polypeptides and the eigenvectors give the mixing coefficients describing contributions of the excited states of the individual groups to the delocalised excited states of the polypeptides. The eigenvectors are then used to calculate the rotational strengths corresponding to each excited state of the peptide, as described in eqn. (1) and subsequently, the CD can be calculated.

Empirical estimates of CD were made using two definitions of helicity, one based on main chain dihedral angles, the other on hydrogen bonds. In the former, a residue was deemed to be helical if its dihedral angles fell within a square region of the Ramanchandran plot[31] centred on the values for an ideal helix, as defined by $\phi = -57° \pm 40°$ and $\psi = -47° \pm 40°$. Occurrences of a single residue or two contiguous residues satisfying the previous constraint were designated non-helical; in other words three (or more) contiguous residues with the necessary dihedral angles were required for these residues to be classified as helical. An alternative empirical definition of helicity was explored based on patterns of hydrogen-bonding and geometrical features calculated by the popular DSSP program.[32] In this program, an ideal hydrogen bond is one where the amide N–H bond and the carbonyl CO bond are collinear and the ideal hydrogen bond length is 2.9 Å. The DSSP definition tolerates deviations in the orientation of up to 63° and a maximum hydrogen bond length of 5.2 Å. Specifically, main-chain $i, i+3$ hydrogen bonds, $i, i+4$ hydrogen bonds and $i, i+5$ hydrogen bonds were counted. Two such hydrogen bonds that were contiguous defined a helical region of the types 3_{10}, α and π, respectively. Helicity was based on the presence of these helical structures; an isolated hydrogen bond did not contribute to the helicity, as it would more likely correspond to a turn-like structure.

CD spectra were calculated for a set of 29 protein structures taken from the Protein Data Bank (PDB)[33] and compared with experimental CD data from the literature.[34–37] This set ranges from highly helical proteins to those that are largely β-sheet. It includes all-α proteins: cytochrome c (PDB accession number: 3cyt), hemoglobin (1hco), myoglobin (1mbn) and bacteriorhodopsin (2brd); some mixed α,β (mainly α) proteins: alcohol dehydrogenase (5adh), glutathione reductase (3grs), lactate dehydrogenase (6ldh), lysozyme (7lyz), papain (9pap), rhodanese (1rhd), subtilisin (1sbt), thermolysin (4tln), triose phosphate isomerase (1tim), flavodoxin (2fx2); some β-I proteins: carbonic anhydrase (1ca2), concanavalin A (3cna), λ-immunoglobulin (1rei), ribonuclease A (3rn3), ribonuclease S (2rns), erabutoxin (3ebx), plastocyanin (1plc), porin (3por), prealbumin (2pab); and some β-II proteins α-chymotrypsinogen A (2cga), α-chymotrypsin II (5cha), elastase (3est), superoxide dismutase (2sod), trypsin inhibitor (4pti) and trypsin (3ptn). The rotational

strengths computed *via* the matrix method were used to generate continuous spectra (rather than line spectra) through Gaussian functions centred on each of the transition energies with a bandwidth of 15.5 nm and an area proportional to the rotational strength of the transition.

Effect of conformational dynamics on β-sheet CD

Matrix method calculations of CD using the static X-ray structures are unable to distinguish between β-I and β-II proteins, predicting β-I spectra for both classes. Whilst there may be deficiencies in the CD calculations, we have suggested[15] that relatively minor fluctuations of 30° in backbone dihedral angles would lead to structures whose calculated CD spectra would be much closer to those observed experimentally. We have performed molecular dynamics simulations on concanavalin A (a β-I protein) and elastase (a β-II protein).

Concanavalin A and elastase both comprise approximately 240 residues. Simulations with explicit solvent would require a large number of water molecules. Therefore, to facilitate the sampling of conformational space under equilibrium conditions, a continuum model of solvent was employed. In a continuum model, one eliminates the solvent nuclear degrees of freedom and the interatomic interactions of the biomolecule are reparametrized to give structural, energetic and dynamic properties that are similar to those seen in explicit solvent. The model used here is the generalised Born model.[38] The accuracy of this model has been established in a number of applications to biopolymers. For example, the model has been compared to explicit solvent models in equilibrium simulations of interleukin-8[39] and in folding simulations of a small β-sheet protein.[40]

Simulations under equilibrium conditions were performed at 298 K for 500 ps using the generalised Born model, as implemented[41] within the CHARMM biomolecular simulation code.[42] The PARAM19 parameter set was used. The stability of the simulations was monitored by following the root mean square deviation (rmsd) of backbone atoms from the initial structure. The distributions of main dihedral angles were computed. The dynamics and mobility of the protein backbones were characterised using the generalised order parameter, which measures the angular correlation for the dynamics of the N–H bond. It is calculated from the trajectory as the plateau value of the correlation function $\langle P_2[\mu(t)\cdot\mu(t+\tau)]\rangle$, where μ is a unit vector oriented along the N–H bond and $P_2(x)$ is the second Legendre polynomial.[43] A generalised order parameter with a value close to unity indicates little motion on the picosecond timescale and greater motion for lower values. The CD was calculated using ensembles of 80 structures (one structure sampled every 5 ps from final 400 ps the trajectory) of each of the two proteins.

Molecular dynamic simulations on myoglobin

Holo myoglobin is a globular, *b* heme protein of 153 residues, comprising eight α-helices. Native apomyoglobin is produced by removal of the proto-porphyrin heme prosthetic group from the holomyoglobin, which partially destabilises the tertiary fold of the globin. Apomyoglobin (myoglobin without the heme group) is well suited for both theoretical and experimental studies of folding, as it folds through a set of well-defined intermediate states.[44] Experimental studies[45,46] reveal that the protein is structurally very similar to the native (holo) myoglobin, retaining most of its secondary, and most likely, tertiary structure. With the gradual addition of acid, apomyoglobin undergoes a two-phase unfolding, first to a molten globule intermediate (I-state) at about pH 4, and finally to an unfolded state at pH 2. In re-folding experiments, the I-state is shown to be an obligatory folding intermediate[47,48] suggesting strong similarity between the acid-unfolding and re-folding pathways of apomyoglobin.[44] This observation makes apomyoglobin a particularly interesting system for theoretical studies of folding: one can hope to gain insight into the apomyoglobin folding process by simulating its acid-unfolding, which, unlike direct, fully atomistic simulation of folding, is quite feasible computationally.

To provide a reference point, an equilibrium simulation of native holo-myoglobin was performed. We used version 5.0 of the AMBER suite of programs.[49] An all-atom force field[50] was employed and the SHAKE algorithm[51] was used to restrain hydrogen–heavy atom bond distances. The integration time-step was 2 fs, with a 12 Å cut-off for long-range interactions. The starting structure for the native holo-myoglobin simulation was the structure obtained by neutron scattering[52] (PDB accession number 2mb5). We kept all the hydrogen atoms found in the PDB set. The protonation state of this structure corresponds to neutral pH; its total charge is +5. The protein

was solvated by approximately 4000 TIP3P water molecules,[53] forming a spherical droplet of radius 37 Å around the centre of mass of the molecule. This is sufficient for the native structure under neutral pH conditions, as the protein is expected to remain in a compact conformation. A simulation cycle began with a 100 steps of steepest-descent minimization followed by 100 ps equilibration during which the temperature was gradually raised to 305 K, while the protein atom coordinates remain fixed by harmonic restraints (force constant 5 kcal mol^{-1} $Å^{-2}$) at their crystallographic positions. The Berendsen temperature coupling algorithm[54] was used, with a coupling constant for both solute and solvent of 1.0 ps. After the equilibration was completed the constraints were removed, and the simulation continued for another 200 ps at an average temperature of 305 K. Separate temperature coupling constants of 20 ps and 1 ps, for the solute and solvent respectively, were used. To prevent evaporation of water molecules, a weak (force constant 0.05 kcal mol^{-1} $Å^{-2}$) spherical cap harmonic restoring potential was applied to atoms further than 37 Å from the centre of mass of the system. Protein coordinates were saved every 0.5 ps. After 200 ps the backbone rmsd from the crystal structure was about 1.3 Å. We used the last 100 ps of the simulation for the analysis.

The starting structure for the unfolding simulation was prepared by removing the heme group from the holo-myoglobin coordinate set used above. To model the conditions of extremely low pH, all aspartate, glutamate, histidine, and the C-terminus side chain groups in the protein were protonated,[55] making the overall charge of the globule +36, in agreement with the experimentally observed value under these conditions. Simulations were carried out using version 5.1 of the AMBER suite of programs;[49] other details were as for the native state simulation, except as noted below. The protein was solvated by 10 000 water molecules, forming a spherical droplet of radius 47 Å around the centre of mass of the molecule. After the equilibration was completed the constraints were removed, and the simulation continued for another 100 ps at 300 K. Protein and solvent coordinates were saved every 10 ps. At the end of this stage, the water molecules were removed, and the protein was re-solvated by 10 000 water molecules forming a spherical droplet of radius 47 Å around the centre of mass of the protein. The above three-step cycle was then repeated 16 times, yielding a total of 1.6 ns of unconstrained trajectory and 160 molecular dynamics trajectory structures. The re-solvation procedure ensured that the protein always stayed well within the droplet during the simulation. The protonation of acidic side-chains to simulate low pH is sufficient to induce unfolding with no added biasing forces or other unusual conditions. By the end of the simulation the protein appears to be completely unfolded; its radius of gyration increased from 15.3 Å of the native holo-myoglobin to 29.6 Å (calculated as an average over the last 200 ps of the trajectory), and was close to that observed experimentally[46] for the acid-unfolded state of apomyoglobin. Other evidence that the simulation yields reasonable structures is provided by the average separations between helices A and G, and A and H. These were calculated as the arithmetic mean of distances between C_α atoms of residues 7 and 103, and 14 and 103; $|AG| = 48$ Å, $|AH| = 57$ Å, consistent with experimental[56] values of $|AG| > 50$ Å, $|AH| > 50$ Å for the acid-unfolded state.

As a further test of the above protocol, a separate simulation of apomyoglobin that models neutral pH conditions was performed. After 20 simulation cycles corresponding to 2.0 ns of unconstrained dynamics, the native fold was preserved, the radius of gyration was only 5% larger, and helical content was 25% lower than that calculated for the native holo-myoglobin. These characteristics were similar to those experimentally observed in apomyoglobin at pH 6.5.[46]

Close quantitative agreement with experiment is not required for the purposes of the current study, and a set of structures reasonably approximating the transition between the native and acid-unfolding states of apomyoglobin is sufficient. The 160 structures taken at 10 ps intervals from the trajectory were used to analyse fluctuations in main chain dihedral angles and hydrogen bonding. The influence of these properties on the predicted helicity has been explored using empirical CD calculations and the matrix method.

Results

Benchmark on static structures

Our benchmark calculations on the set of 29 proteins are summarised in Table 1. The calculation of helicity was based on the mean residue ellipticity at 220 nm, $[\theta]_{220}$. The measured helicity was

Table 1 Comparison of approaches for CD calculations

Method	Spearman rank correlation between calculated and measured helicity	Root mean square error
Matrix method	0.896	0.066
Main chain dihedral angles	0.876	0.089
α-Helical (only) H-bonds	0.874	0.093
All helical H-bonds	0.865	0.107

computed as $[\theta]_{220}/(-37\,000)$, where $[\theta]_{220}$ was taken from experimental spectra reported in the literature. An analogous expression was used to give the helicity calculated using the matrix method, with $[\theta]_{220}$ taken from the calculations. The value of $-37\,000$ degree cm^2 dmol^{-1} is based on experimental[57] and theoretical[15] estimates of the value of $[\theta]_{220}$ for an polypeptide of 100% helicity. Thus, helicity usually ranges from zero to one. Two values of helicity derived from the hydrogen bonding were computed as the fraction of helical residues in the protein, where helical was either restricted to α-helical or included all types of helices. In the former case, the fraction was calculated using the total number of residues less four (as four terminal residues would not be able to form i, $i+4$ hydrogen bonds); in the latter case, the total number of residues less three was used (as three terminal residues would not be able to form i, $i+3$ hydrogen bonds). Table 1 shows two statistical measures of accuracy, the Spearman rank correlation coefficient between the calculated (by various methods) and experimental helicities and the root mean square error in the calculated helicity. Both measures show that the matrix method calculations are marginally better than the empirical approaches, although this difference is probably not statistically significant.

β-Sheet proteins: elastase and concanavalin A

Table 1 presents data on helicity, which are based solely on the intensity at 220 nm. The accuracy of the matrix method at other wavelengths, where there is no comparison with empirical estimates of helicity, has been assessed elsewhere.[15] In contrast to the region at 220 nm, which arises predominantly from the amide $n\pi^*$ transition, the region 190–200 nm is due primarily to the amide $\pi\pi^*$ transition. This part of the spectrum is less well modelled by the matrix method calculations, possibly because the $\pi\pi^*$ transition has a large electric transition dipole moment which can couple with higher energy transitions. Nevertheless, the accuracy of the matrix method appears to be reasonable, based on the static X-ray structures. However, there are some β-sheet proteins (class β-II) for which the matrix method calculations perform poorly. Yet on other β-sheet (class β-I) proteins, with similar structures, the calculations perform well.

To investigate whether conformational dynamics in solution might explain the discrepancy between class β-I and class β-II proteins, simulations on representatives of these two classes were carried out. The chosen proteins were concanavalin A from class β-I and elastase from class β-II. The rmsd of the backbone atoms is shown in Fig. 2 over a 500 ps period. Both proteins undergo some initial rearrangements, but these appear to be complete after 100 ps. In the case of concanavalin A, the ~4.5 Å change in the backbone rmsd corresponds to one of the β-sheets (the upper one in Fig. 1) unfurling and becoming flatter.

The CD spectra, computed using the matrix method, of concanavalin A and elastase, based on the final 400 ps of the simulations are shown in Figs. 3 and 4, respectively. The thick solid line is the mean spectrum computed from 80 snapshots along the trajectory. Spectra from individual structures were qualitatively similar, but varied by up to ±5000 degree cm^2 dmol^{-1}. The calculated CD for concanavalin A remains close to that calculated for the X-ray crystal structure (Fig. 3), whereas the ensemble of elastase gives a calculated CD spectrum that is markedly different from that of the crystal structure. For the elastase ensemble, the computed intensity at 195 nm is 2400 degree cm^2 dmol^{-1} compared to 10 000 degree cm^2 dmol^{-1} calculated for the crystal structure and -7500 degree cm^2 dmol^{-1} observed experimentally (Fig. 4). Whilst a qualitative problem still remains (the sign of the peak is wrong), the results nevertheless represent a significant improvement. Quantitatively, the error in the computed CD spectrum of elastase is much lower. Qualitatively, it appears

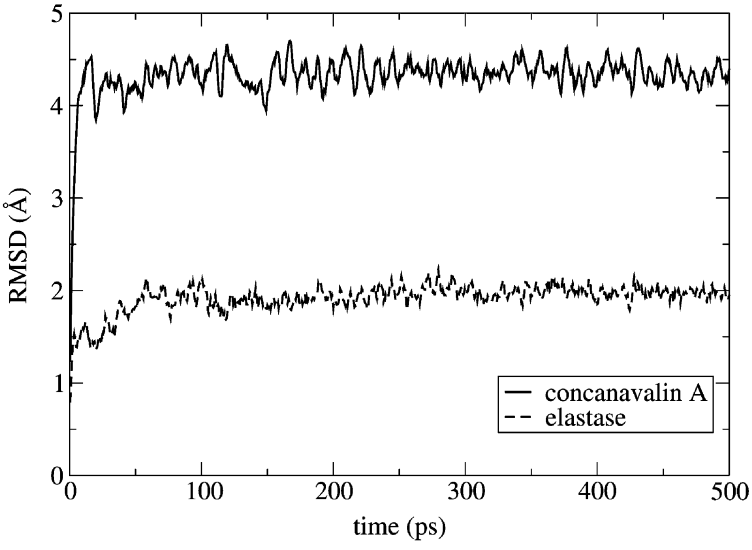

Fig. 2 Root mean square deviation of the backbone atoms during the simulations of elastase (lower curve) and concanavalin A (upper curve).

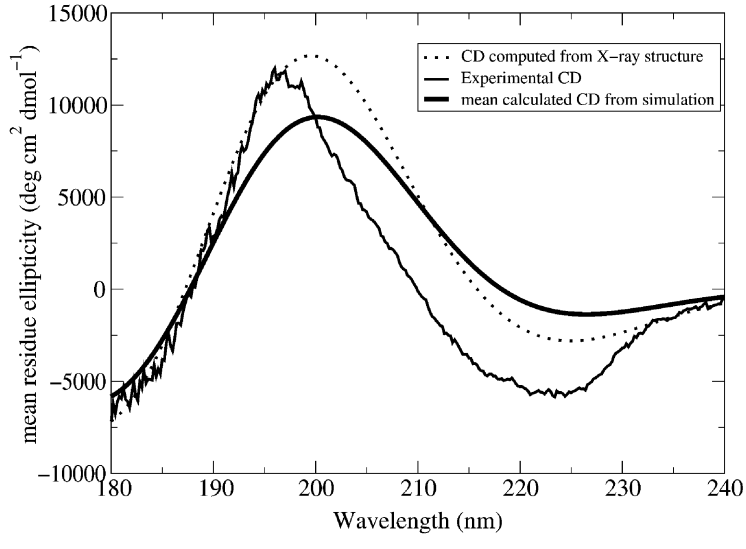

Fig. 3 Circular dichroism spectra of concanavalin A: medium solid line, experimental; dotted line, computed from the X-ray crystal structure; thick solid line, mean computed spectrum from an ensemble of structures from molecular dynamics simulation.

that relaxation of the elastase structure in solution may account for some of the previously unexplained differences between β-I and β-II proteins.

The differences between the ensembles of concanavalin A and elastase have been investigated with respect to both dynamics and structure. To probe whether elastase underwent larger fluctuations than concanavalin A, the generalised order parameters were computed for the last 400 ps of the simulations. From Fig. 5, the two proteins exhibit quite similar backbone flexibility and so this appears unlikely to be the origin of the differing CD spectra. For the last 400 ps of each simulation, the main-chain dihedral angles were computed for structures every 0.5 ps along the

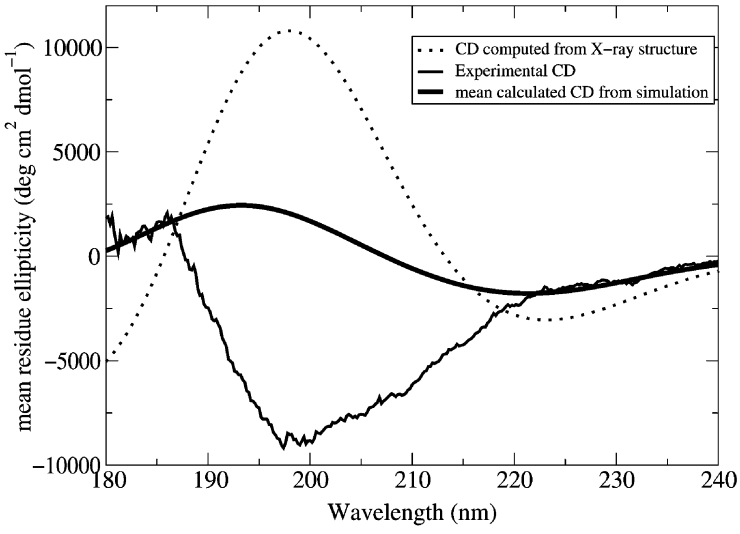

Fig. 4 Circular dichroism spectra of elastase: medium solid line, experimental; dotted line, computed from the X-ray crystal structure; thick solid line, mean computed spectrum from an ensemble of structures from molecular dynamics simulation.

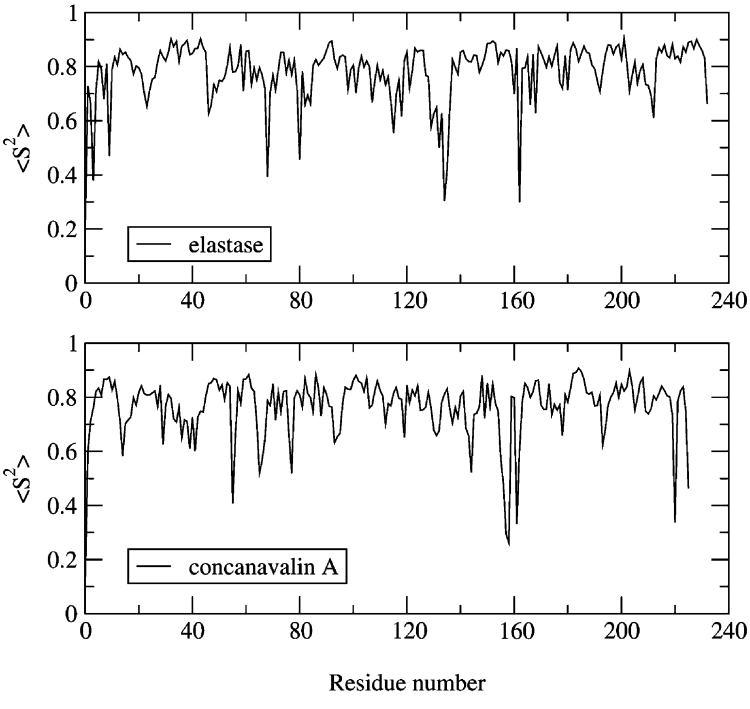

Fig. 5 Generalised order parameters, $\langle S^2 \rangle$, for N–H relaxation.

trajectories. These were used to construct histograms representing the distributions of dihedral angles populated by the ensembles. The difference between the two resulting histograms is plotted in Fig. 6. The darker areas on the figure indicate regions, which are more populated by elastase; the lighter areas indicate a greater population by concanavalin A. A shift in the population in

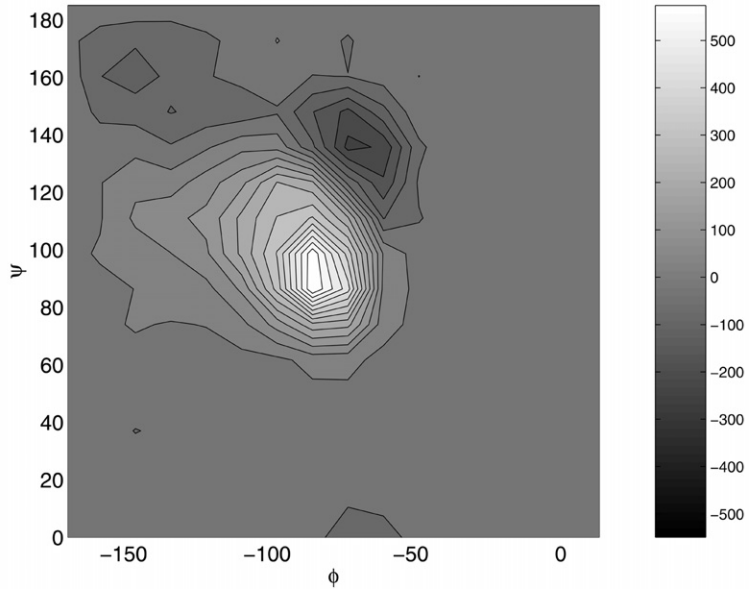

Fig. 6 Difference plot between the distributions of main chain dihedral angles sampled in the simulations (concanavalin A–elastase).

β-conformation portion of the Ramanchandran plot is evident, from (−85°, 95°) for concanavalin A to (−70°, 140°) for elastase. We discuss this further in the next section. The sensitivity of the circular dichroism to changes of this magnitude in the main chain dihedral angles also emerges in the calculations on myoglobin.

Myoglobin

In the benchmark calculations (Table 1) on 29 proteins, the structure of myoglobin 1mbn[58] was considered. In the simulation studies, the structure 2mb5 was used as a starting point.[52] These two structures appear to be very similar, with a backbone rmsd of ~0.5 Å. The computed CD spectra are compared with experiment in Fig. 7. The calculated CD spectrum is also shown from a third structure, 1bzr, which was solved most recently and to the highest resolution, namely 1.15 Å.[59] Whilst the computed CD spectra agree only qualitatively with the experimental spectrum across the entire wavelength range, the quantitative agreement at 220 nm is evident for the 1mbn structure. The experimental value of $[\theta]_{220}$ is −24 090 degree cm^2 dmol^{-1}; the value computed from structure 1mbn is −23 597 degree cm^2 dmol^{-1} and the value from structure 1bzr is −22 175 degree cm^2 dmol^{-1}. However, the intensity at 220 nm computed from the 2mb5 structure, −19 048 degree cm^2 dmol^{-1}, is less negative than anticipated, which at first glance seemed curious given the similarity of the experimental structures. Closer inspection of the 1mbn and 2mb5 revealed that there were subtle, but significant differences in the main-chain dihedral angles, as shown in Fig. 8. Although in both cases the main chain dihedral angles fall within the helical region and both structures have the same helical assignments the respective centres of the (phi,psi) distribution in the helical region are shifted relative to each other. Analysis of the last 100 ps of the 200 ps simulation of the native state of holo-myoglobin, which used the 2mb5 structure as initial coordinates, showed that over time the dihedral angle distribution moved towards that of the 1mbn structure and so did the value of $[\theta]_{220}$. The value of $[\theta]_{220}$ computed for structures taken every 0.5 ps from the 100 ps portion of the trajectory was −20 110 degree cm^2 dmol^{-1} with a standard deviation of 30 degree cm^2 dmol^{-1}.

The calculated helicity of the structures of apomyoglobin protein as it unfolds in the simulation is depicted in Fig. 9. The different methods give quite similar results, although the inclusion of 3_{10} and π-helical conformations in the hydrogen bond calculation leads to an over-estimate of helicity compared to the consensus. The helicity computed by the empirical approaches (based on dihedral

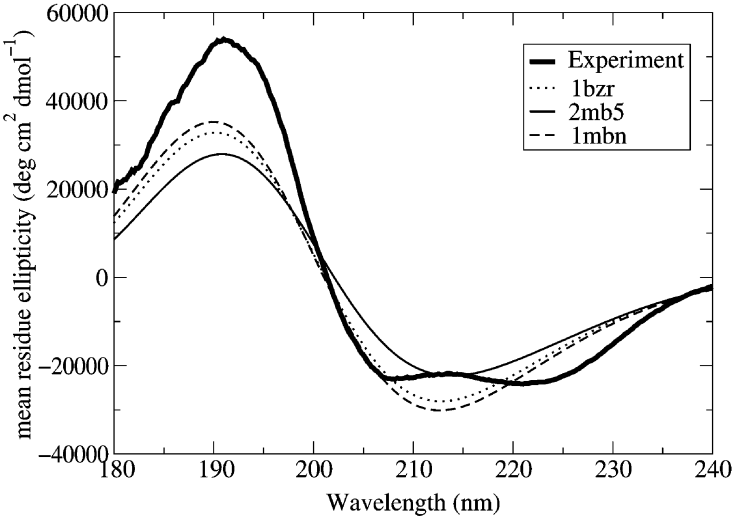

Fig. 7 Experimental and computed CD spectra of myoglobin.

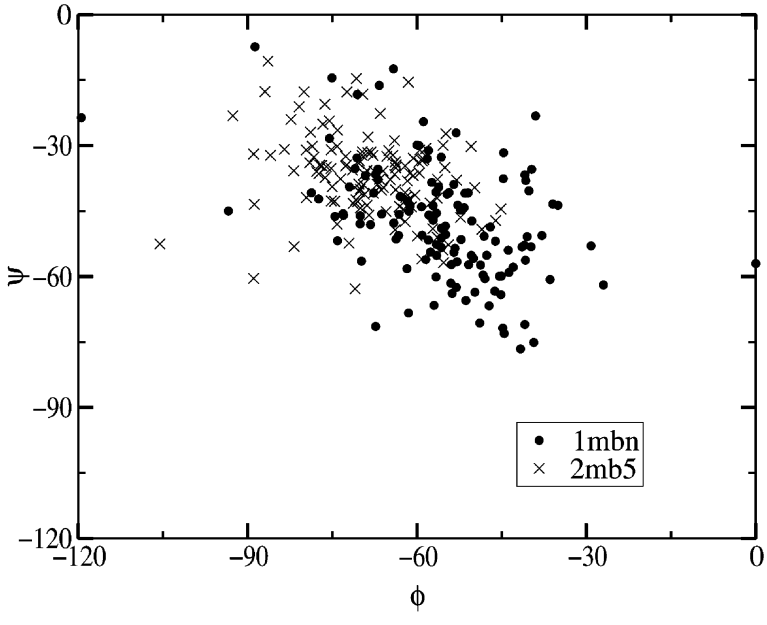

Fig. 8 Ramanchandran plot for the structures 1mbn and 2mb5.

angles or hydrogen bonds) exhibits greater fluctuations compared to the matrix method *ab initio* calculations. After about 300 ps the matrix method helicity remains stable around 0.3, whereas the α-helical hydrogen bond helicity and the helicity computed from the dihedral angles fluctuate between 0.4 and 0.2.

Discussion

We have demonstrated the feasibility of using first principles CD calculations and molecular dynamic simulations to analyse the unfolded and partially unfolded helical states. Earlier work

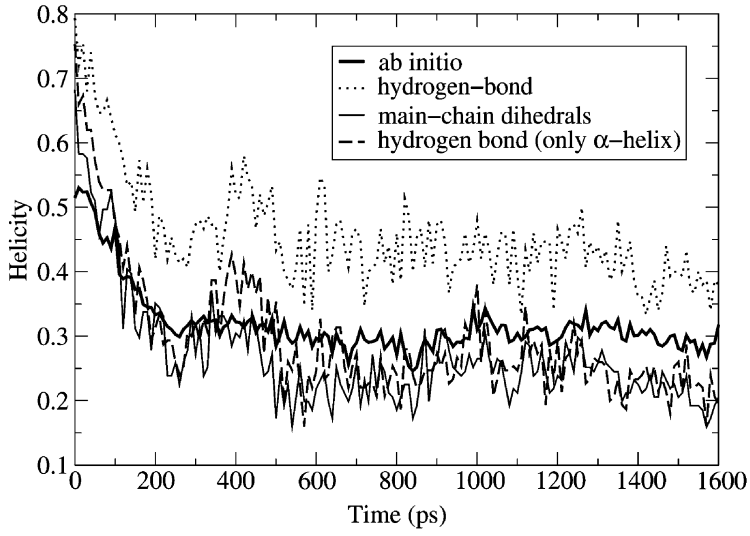

Fig. 9 Time evolution of helicity computed (by several methods) from the unfolding simulation of apo-myoglobin.

adopted this general strategy to study the conformations of cyclic-L-Tyr-L-Tyr.[60] The complexity of the systems that we have investigated using a combined approach of molecular dynamics simulation, to generate statistically meaningful ensembles, and CD calculations is significantly greater. Such calculations allow us to assess precisely when protein conformations are no longer sufficiently helical to be observed by CD spectroscopy. Thus, the theoretical calculation of time resolved spectroscopy of protein folding will provide a direct connection to experimental data.

Many definitions of helicity have been applied in the literature to analyse simulations of conformational transitions in helical peptides and proteins. Examples of the use of hydrogen-bonding patterns include: the DSSP secondary structure assignment[61] and the presence of i, $i+4$ hydrogen bonds.[62] Definitions based on main-chain dihedral angles include: a 15° range centred on (−58°, −47°),[63] a 30° range centred on (−57°, −47°)[64] and ($-100° \leqslant \phi \leqslant -30°$ and $-80° \leqslant \psi \leqslant -5°$).[65,66] In addition to the ambiguity over what constitutes a helical residue, even the definition of a hydrogen bond often involves the stipulation of arbitrary thresholds relating to distances and angles.[67] At one level, most of these definitions appear to be satisfactory. Table 1 and Fig. 9 show that empirical measures reflect the helicity of both native proteins and partially unfolded states in reasonable agreement with either experiment or the first principles calculations. However, counting all helical types seems to over-estimate the helicity of the partially unfolded states, despite strong indications that 3_{10} and π helices do contribute significantly to the intensity at 220 nm.[15,68] Nevertheless, the empirical definitions do involve somewhat arbitrary specifications. The greater fluctuations of the empirical helicities in Fig. 9 are due to the artificially discontinuous nature of the empirical definitions. A change of one degree in a dihedral angle presumably does not lead to an all or nothing change in the CD in reality. Most of the analyses of helicity based on dihedral angles define a square or rectangular window on the Ramanchandran plot. More complex regions have been investigated and the weighting of different helical regions has also been explored.[13] Whilst such an empirical model might be more appropriate, it lacks the appeal of simplicity. Our findings for the β-sheet proteins and myoglobin both suggest, however, that the CD of proteins is more sensitive to the dihedral angles of the backbone than the empirical models allow for.

In the comparison of concanavalin A and elastase, we observed that a shift in the dihedral angle population from (−85°, 95°) to (−70°, 140°) significantly influences the computed CD, bringing the calculated CD spectrum for elastase closer to the experimentally observed spectrum. This result is in accord with earlier work on ideal, regular β-strands where a change of 30° in the ψ dihedral angle from 140° to 170° led to much less intense calculated CD spectrum.[15] There are still obvious deficiencies in the calculated CD spectra, but conformational dynamics in solution at least seems to

be one important factor that needs to be accounted for. As discussed earlier, the matrix method performs best around 220 nm, where the amide $n\pi^*$ transitions and helical structures are the prime determinants of the spectrum.

Our results on the α-helical protein myoglobin show that the CD spectra of helical conformations also are sensitive to main chain dihedral angles. The analysis of different structures from the PDB and the simulation of the native state of holo-myoglobin coupled with the CD calculations indicate that accounting for conformational dynamics in solution leads to better agreement with the experimental data. In an earlier study[15] on over 100 helical fragments excised from the 29 proteins used in the benchmark in Table 1, the calculated intensity at 220 nm was shown to depend on the precise conformation of the helix. A weak, but significant, correlation was observed between $[\theta]_{220}$ and the coordinate $0.74\phi - 0.67\psi$, where the dihedral angles were the mean values of all the residues in the helix. A sensitivity to dihedral angles had also been noted elsewhere[68] and has been related to the hydrophobic or hydrophilic nature of the helix, with the latter corresponding to more negative ϕ and less negative ψ.[69] Yet other work[70] has suggested that the CD spectra of helical peptides are also quite sensitive to the length of the main-chain hydrogen bonds.

The first principles calculations of CD have many caveats, as has been discussed elsewhere.[15] Higher energy excitations of the amide chromophore are not included in the calculations. Transitions from side-chain chromophores are not considered. The influence of solvent on the CD itself, as distinct from its influence on the conformation of the protein, has not been accounted for. Whilst these and other factors warrant attention, in this study we have examined the role of conformational dynamics and we have found it to be important. Perhaps the most basic conclusion of the study is to emphasize that moderate differences in CD spectra cannot be interpreted uniquely in terms of conformational changes. In the absence of complementary data, a change in a CD spectrum could be due to several different causes, including the loss of secondary structure, subtle re-distributions of dihedral angles or the re-organization of patterns of secondary structures. To a large extent the detailed structural characterisation of unfolded and partially unfolded states is uncharted terrain and more work is needed in this area. A better understanding of the theoretical basis of protein CD combined with detailed simulations of conformational transitions should help to improve the interpretation of CD experiments on protein folding.

Acknowledgements

We thank BBSRC for financial support through grant 42/B15240. A.V.O. thanks NIH for funding (GM 575513) and Professor David Case for support and encouragement. We thank Professor Robert Woody for his comments on the manuscript.

References

1 T. E. Creighton, *Protein Folding*, Freeman, New York, 1992.
2 C. M. Dobson, A. Sali and M. Karplus, *Angew. Chem. Int. Ed. Engl.*, 1998, **37**, 868.
3 C. M. Dobson and M. Karplus, *Curr. Opin. Struct. Biol.*, 1999, **9**, 92.
4 K. Nakanishi, N. Berova and R. W. Woody, *Circular Dichroism Principles and Applications*, VCH, New York, 1994.
5 G. D. Fasman, *Circular Dichroism and the Conformational Analysis of Biomolecules*, Plenum Press, New York, 1996.
6 G. Holzwarth and P. Doty, *Proc. Natl. Acad. Sci. USA*, 1965, **87**, 218.
7 K. W. Plaxco and C. M. Dobson, *Curr. Opin. Struct. Biol.*, 1996, **6**, 630.
8 C.-F. Zhang, J. W. Lewis, R. Cerpa, I. D. Kuntz and D. S. Kliger, *J. Phys. Chem.*, 1993, **97**, 5499.
9 W. A. Eaton, V. Munoz, P. A. Thompson, E. R. Henry and J. Hofrichter, *Acc. Chem. Res.*, 1998, **31**, 745.
10 J. R. Telford, P. Wittung-Stafshede, H. B. Gray and J. R. Winkler, *Acc. Chem. Res.*, 1998, **31**, 755.
11 Y. Duan and P. A. Kollman, *Science*, 1998, **282**, 740.
12 B. Zagrovic, E. J. Sorin and V. Pande, *J. Mol. Biol.*, 2001, **313**, 151.
13 J. D. Hirst and C. L. Brooks III, *J. Mol. Biol.*, 1994, **243**, 173.
14 W. C. Johnson, *Proteins*, 1990, **7**, 205.
15 N. A. Besley and J. D. Hirst, *J. Am. Chem. Soc.*, 1999, **121**, 9636.
16 J. D. Hirst and N. A. Besley, *J. Chem. Phys.*, 1999, **111**, 2846.
17 R. W. Woody and N. Sreerama, *J. Chem. Phys.*, 1999, **111**, 2844.
18 K. A. Bode and J. Applequist, *J. Am. Chem. Soc.*, 1998, **120**, 10938.

19 P. Manavalan and W. C. Johnson, Jr., *Nature*, 1983, **305**, 831.
20 J. Wu, J. T. Yang and C.-S. C. Wu, *Anal. Biochem.*, 1992, **200**, 359.
21 R. W. Woody and A. K. Dunker, in *Circular Dichroism and the Conformational Analysis of Biomolecules*, ed. G. D. Fasman, Plenum Press, New York, 1996, p. 109.
22 V. Daggett, *Curr. Opin. Struct. Biol.*, 2000, **10**, 160.
23 P. M. Bayley, E. B. Nielsen and J. A. Schellman, *J. Phys. Chem.*, 1969, **73**, 228.
24 W. J. Goux and T. M. Hooker, Jr., *J. Am. Chem. Soc.*, 1980, **102**, 7080.
25 L. Rosenfeld, *Z. Phys.*, 1928, **52**, 161.
26 G. Karlström, *J. Phys. Chem.*, 1988, **92**, 1315.
27 G. Karlström, *J. Phys. Chem.*, 1989, **93**, 4952.
28 A. Bernhardsson, R. Lindh, G. Karlstrom and B. O. Roos, *Chem. Phys. Lett.*, 1996, **151**, 141.
29 L. Serrano-Andrés, M. P. Fülscher and G. Karlstrom, *Int. J. Quantum Chem.*, 1997, **65**, 167.
30 N. A. Besley and J. D. Hirst, *J. Phys. Chem. A*, 1998, **102**, 10 791.
31 G. N. Ramanchandran, C. Ramakrishnan and V. Sasisekharan, *J. Mol. Biol.*, 1963, **7**, 95.
32 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577.
33 F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.*, 1977, **112**, 535.
34 P. Pancoska, E. Bitto, V. Janota, M. Urbanova, V. P. Gupta and T. A. Keiderling, *Protein Sci.*, 1995, **4**, 1384.
35 S. Brahms and J. Brahms, *J. Mol. Biol.*, 1980, **138**, 149.
36 T. E. Dahms and A. G. Szabo, *Biophys. J.*, 1995, **69**, 569.
37 J. E. Draheim, G. P. Anderson, J. W. Duane and E. L. Gross, *Biophys. J.*, 1986, **49**, 891.
38 W. C. Still, A. L. Tempczyk, R. C. Hawley and T. Hendrickson, *J. Am. Chem. Soc.*, 1990, **112**, 6127.
39 W. Cornell, R. Abseher, M. Nilges and D. A. Case, *J. Mol. Graphics Mod.*, 2001, **19**, 136.
40 B. D. Bursulaya and C. L. Brooks III, *J. Phys. Chem. B*, 2000, **104**, 12 378.
41 B. N. Dominy and C. L. Brooks III, *J. Phys. Chem. B*, 1999, **103**, 3765.
42 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187.
43 I. Chandrasekar, G. M. Clore, A. Szabo, A. M. Gronenborn and B. R. Brooks, *J. Mol. Biol.*, 1992, **226**, 239.
44 P. E. Wright and R. L. Baldwin, in *Frontiers in Molecular Biology: Mechanisms of Protein Folding*, ed. R. Pain, Oxford University Press, London, 2000, p. 309.
45 J. T. Lecomte, S. F. Sukits, S. Bhattacharaya and C. J. Falzone, *Protein Sci.*, 1999, **8**, 1484.
46 D. Eliezer, J. Yao, H. J. Dyson and P. E. Wright, *Nature Struct. Biol.*, 1998, **5**, 148.
47 M. Jamin and R. L. Baldwin, *J. Mol. Biol.*, 1998, **276**, 491.
48 V. Tsui, C. Garcia, S. Cavangero, G Sizudak, H. J. Dyson and P. E. Wright, *Protein Sci.*, 1999, **8**, 45.
49 D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel and P. A. Kollman, *Comput. Phys. Commun.*, 1995, **91**, 1.
50 W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, 1995, **117**, 5179.
51 J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327.
52 X. Cheng and B. Schoenborn, *Acta Crystallogr., Sect. B.*, 1990, **46**, 195.
53 W. L. Jorgensen, J. Chandrasekhar, J. Madura and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926.
54 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684.
55 L. J. Smith, C. M. Dobson and W. F. van Gunsteren, *Proteins*, 1999, **36**, 77.
56 O. Tcherkasskaya and O. B. Ptitsyn, *FEBS Lett.*, 1999, **455**, 325.
57 J. T. Yang, C. S. C. Wu and H. M. Martinez, *Methods Enzymol.*, 1986, **130**, 208.
58 H. C. Watson, *Prog. Stereochem.*, 1969, **4**, 299.
59 G. S. Kachalova, A. N. Popov and H. D. Bartunik, *Science*, 1999, **284**, 473.
60 J. Fleischhauer, J. Grotzinger, B. Kramer, P. Kruger, A. Wollmer, R. W. Woody and E. Zobel, *Biophys. Chem.*, 1994, **49**, 141.
61 A. R. van Buuren and H. J. C. Berendsen, *Biopolymers*, 1993, **33**, 1159.
62 J. D. Hirst and C. L. Brooks III, *Biochemistry*, 1995, **34**, 7614.
63 M. J. Bodkin and J. M. Goodfellow, *Protein Sci.*, 1995, **4**, 603.
64 S.-S. Sung and X.-W. Wu, *Biopolymers*, 1997, **42**, 633.
65 V. Daggett and M. Levitt, *J. Mol. Biol.*, 1992, **223**, 1121.
66 J. Tirado-Rives and W. L. Jorgensen, *Biochemistry*, 1993, **32**, 4175.
67 G. Ravishanker, S. Vijakumar, and D. L. Beveridge, in *Modeling the Hydrogen Bond*, American Chemical Society, Washington DC, 1994, p. 209.
68 M. C. Manning and R. W. Woody, *Biopolymers*, 1991, **31**, 569.
69 T. Blundell, D. Barlow, N. Borkakorti and J. Thornton, *Nature*, 1983, **306**, 281.
70 Z. Dang and J. D. Hirst, *Angew. Chem. Int. Ed. Engl.*, 2001, **40**, 3619.