Optimization of Model Parameters for Describing the Amide I Spectrum of a Large Set of Proteins

Eeva-Liisa Karjalainen, Tore Ersmark, and Andreas Barth*

Department of Biochemistry and Biophysics, Arrhenius Laboratories of Natural Sciences, Stockholm University, SE-106 91, Sweden

E-mail: andreas.barth@dbb.su.se

Phone: +46 (0)8 162452. Fax: +46 (0)8 155597

*To whom correspondence should be addressed

Abstract

A new simulation protocol for the prediction of the infrared absorption of the amide I vibration of proteins was developed. The method incorporates known effects on the intrinsic frequencies (backbone conformation, inter-peptide and peptide-solvent hydrogen bonding) and couplings (nearest neighbor coupling, transition dipole coupling) of amide I oscillators in a parametrized manner. Model parameters for the simulation of amide I spectra were determined through fitting and optimization of simulated spectra to experimentally measured infrared spectra of 44 proteins that represent maximum structural variation in terms of different folds and secondary structure contents. Prediction of protein spectra using the optimized parameters resulted in good agreement with experimental spectra and in a considerable improvement compared to a description involving only transition dipole coupling.

Introduction

The value of infrared (IR) spectroscopy as a tool for investigation of biological systems has been clearly demonstrated over the past decades.[1–5] Due to the multifaceted relationship between molecular structure and IR spectrum, band assignments and the interpretation of IR spectra benefit from theoretical structure-spectra correlations. The size of biological systems is a great challenge for the simulation of spectra. Ab initio calculations on a full quantum mechanical level are still computationally too demanding for proteins, even without regard to dynamics or solvation.[6–8] Therefore, approximate methods are required to be able to simulate larger systems. For protein simulations, the selective simulation of only the amide I mode arising from the polypeptide backbone is an excellent option for reducing the complexity and computational time of the problem. The amide I vibration is mainly composed of C=O stretching and NH bending of a peptide unit.

Finding a theoretical description for the amide I band of polypeptides and proteins has been subject to intense research efforts since the beginning of the 1960's.[1,2,9] The efforts to date can be subdivided into empirical models aimed at predicting the amide I band of proteins, or more recent approaches that utilize small molecule ab initio data in a parametrized fashion to construct a

Hamiltonian for the system. These so called building block approaches have proven comparatively successful for simulation of polypeptides with regular secondary structures,[7] but have been found less successful for simulation of real proteins such as ubiquitin.[10]

As a crucial element of these approaches, the concept of the "floating oscillator model" was established.[11] This refers to a simplified representation of the protein vibrations that only considers one vibrating amide I oscillator per peptide unit that interacts with other oscillators via transition dipole coupling (TDC).[1,12,13] This interaction makes the amide I mode sensitive to the conformation of the polypeptide backbone conformation. In the original approach all other effects on the vibrational frequencies were neglected, such as hydrogen bonding and through-bond coupling, however some systematic as well as non-systematic changes were made to the intrinsic frequencies in order to improve agreement with experimental results. Watson and Hirst[14] later attempted to systematize the selective parameter modifications made by Torii and Tasumi. Mendelsohn and coworkers[15] further extended this approach to include several other effects influencing the amide I band such as hydrogen bonding, covalent bonds and π-π interactions. They presented results for four proteins and two synthetic peptides using this model. Common for all of these approaches is that solvation is not considered, despite that protein measurements are generally performed in aqueous environment and that the additional, non-uniformly distributed, hydrogen bonding should consequently have a drastic effect on the vibrational frequencies.

Simulation of the effects that solvent water has on the protein frequencies is complicated because of the intrinsic properties of water. Water solvent is highly dynamic, causing fluctuations of the IR frequencies of the peptides that are solvent accessible. A computationally inexpensive way of including solvent effects on the amide I frequencies is to parametrize ab initio calculations of N-methylacetamide (NMA) and water clusters.[16–30] This method exploits the linear correlation found between the C=O bond length, stretching frequency and electrostatic potential at the atoms of NMA,[18,31,32] which provides a link between perturbation of molecular (and electronic) structure by the electric field of the solvent and the resulting shift of the vibrational frequency. Parametrized electrostatic calculations using the building block approach have been used in protein amide I sim-

ulations of ubiquitin,[10,33] other proteins[34] as well as polypeptides in membrane environment.[29,35] These calculations are generally combined with MD simulations that provide coordinate trajectories, which are then utilized for calculating instantaneous normal modes and the resulting spectral line shapes.[36] It has been clearly demonstrated that heterogeneity in the site frequencies needs to be well reproduced in order to obtain agreement with spectra, as shown for ubiquitin by Cho and co-workers.[10] Solvent interactions were recently accounted for in a similar way in a calculation of the amide I absorption of a small helix-turn-helix protein[37] using otherwise a conceptually different building block approach - the Cartesian coordinate transfer method - developed by Keiderling, Bour and co-workers.[38,39]

Here, the challenge of amide I simulations is addressed by attempting to find optimized parameters for the physical models used to describe the intrinsic frequencies and couplings between amide I oscillators. A large protein set,[40] selected specifically to encompass maximum variation when it comes to different folds, secondary structure contents, etc., and for which experimental FTIR spectra were available, was utilized for the calculations. The parameters were found through fitting of simulated spectra to the experimentally measured FTIR spectra using a subset of the proteins, selected on the basis of structural variation. These optimized parameter values were then used for prediction of the amide I band of the remaining protein subset, enabling an assessment of the prediction quality of the optimized parameters. The approach results in a considerable improvement as compared to a description with solely TDC.

Theoretical Methods

Amide I band calculation

For the purpose of simulating only the amide I band of polypeptides, the protein can be approximated as consisting of N harmonic amide I oscillators, where N is the total number of peptide units in the polypeptide chain. This reduced representation of the protein vibrations, the amide I subspace, is spanned by the local amide I vibrations of each peptide unit.[11]

The Hamiltonian for a system of amide I oscillators can be described using the exciton model,[41–45] which is analogous to the "floating oscillator model" utilized to describe the amide I band of proteins by Torii and Tasumi.[11,44,46]

The diagonal elements of the Hamiltonian matrix are the intrinsic frequencies of the uncoupled, local amide I vibrations. The off-diagonal elements are the coupling terms, which describe the interaction between local oscillators and give rise to delocalization of the normal modes over several amide groups. Diagonalization of the Hamiltonian matrix then yields the N normal modes of the protein, the coupled amide I frequencies. The intensities of the respective modes can be calculated from the eigenvectors of the diagonalized Hamiltonian and the local mode transition dipoles.[36,46]

Intrinsic peptide frequencies

The intrinsic frequencies of the amide I oscillators depend on all factors that alter the electron density in the relevant bonds of the oscillator. This includes the local conformation around the peptide unit as well as hydrogen bonds to other parts of the protein and to solvent. These effects are here assumed to yield independent and additive shifts of the amide I frequency from the gas phase value $\omega^{\mathrm{NMA}} = 1723$ cm^{-1} of unperturbed, uncoupled NMA,[47–49] which has been extensively studied as a model compound of a peptide unit.[31,50–57] The intrinsic frequency of the amide I oscillator in residue i can thus be described as

$$\omega_i = \omega^{\mathrm{NMA}} + \delta\omega_i^{\mathrm{L}} + \delta\omega_i^{\mathrm{H}} + \delta\omega_i^{\mathrm{S}}$$

where $\delta\omega_i^{\mathrm{L}}$ is the shift due to local conformation, $\delta\omega_i^{\mathrm{H}}$ is the shift induced by interpeptide hydrogen bonding and $\delta\omega_i^{\mathrm{S}}$ is the shift caused by solvation effects. A peptide unit is defined as the unit between two consecutive C_α's, and is numbered by the same index as the residue containing the carbonyl group of the peptide unit. The peptide unit i thus contains the N atom of residue $i+1$. The index of a local amide I oscillator is the same as the index of the peptide unit.

The effect from the local environment was modeled as the influence exerted by the nearest

neighbors in the chain using ab initio calculations of dipeptides.[24,58–60] The shift of the intrinsic frequency is then described using two maps parametrized with respect to the dihedral angles, which are used to evaluate the influence of the local conformation before and after the oscillator

$$\delta\omega_i^{\mathrm{L}} = \delta\omega_i^{\mathrm{L,before}}(\phi_i, \psi_i) + \delta\omega_i^{\mathrm{L,after}}(\phi_{i+1}, \psi_{i+1})$$

Here, new versions of such maps were constructed by optimizing the fit to experimental protein spectra. Published maps from density functional theory (DFT) calculations of dipeptides[60] were used as a starting point. Data points in sparsely populated regions of the Ramachandran space were optimized using first a common global shift and then applying a multiplicative factor per map. Data points in significantly populated regions in Ramachandran space, mainly corresponding to the α-helix and β-sheet regions, were assigned parameters that were optimized independently (see Figure 4). The published map values were used to initially populate these parameters. The maps were read out using cubic spline interpolation. This hybrid approach was used in order to avoid constraining the optimization to small molecule DFT maps for widespread psi/phi angles on the one hand and on the other hand to reduce the dimensionality of the optimization problem for rare psi/phi angles.

NMA has three potential hydrogen bonding sites, the carbonyl oxygen can accept two hydrogen bonds that redshift the amide I frequency by 15–20 cm^{-1} each, whereas the amide nitrogen can act as a donor for one hydrogen bond that causes a 10–15 cm^{-1} shift.[52,53,55] These effects are approximately additive,[52,61]

$$\delta\omega_i^{\mathrm{H}} = \delta\omega_i^{\mathrm{H,O1}} + \delta\omega_i^{\mathrm{H,O2}} + \delta\omega_i^{\mathrm{H,H}}$$

The shift of the intrinsic frequency due to interpeptide hydrogen bonds was modeled according to the method proposed by Ge and co-workers,[62] who found that the shifts of the amide I oscillator frequencies are well described by proportional relations to the Kabsch-Sander bond energies.[63] The electrostatic Kabsch-Sander energy of a hydrogen bond between C=O and NH groups is

$$E_{\mathrm{KS}} = f q_1 q_2 (r_{\mathrm{ON}}^{-1} + r_{\mathrm{CH}}^{-1} - r_{\mathrm{OH}}^{-1} - r_{\mathrm{CN}}^{-1})$$

with $q_1 = 0.42$ e being the partial charge of C and the absolute value of the partial charge of O, and $q_2 = 0.2$ e describing the partial charges of N and H, $f = 332$ e^{-2}Å kcal/mol and all inter-atomic distances r in Ångstroms. The signs of the charges are taken care of in the summation of the reciprocal distances, where reciprocal distances between atoms with partial charges of different signs are multiplied with -1. If the hydrogen bond energy value E_{KS} is less than the threshold value -0.5 kcal/mol, a hydrogen bond is predicted to exist. If multiple potential hydrogen bonds were found, the two with the lowest energy were retained for hydrogen bonds to oxygen. For the amide hydrogen only the strongest bond was retained. The Kabsch-Sander energy was subsequently utilized to calculate the frequency shift using correlation coefficients ξ^{O} and ξ^{N}

$$\delta\omega_i^{\mathrm{H,O1/2}} = \xi^{\mathrm{O}} E_{\mathrm{KS}},\ \delta\omega_i^{\mathrm{H,H}} = \xi^{\mathrm{H}} E_{\mathrm{KS}}$$

where $\xi^{\mathrm{O}} = 2.4$ cm^{-1}/kcal and $\xi^{\mathrm{N}} = 1.0$ cm^{-1}/kcal as determined from parametrization of ab initio data for small peptides. This model has been found to yield results comparable to more sophisticated electrostatic models.[23] For the present protein simulation, the values of the coefficients were optimized to yield the best fit to the experimental data in the test set.

Hydrogen bonds to solvent were simulated by relating the solvent accessible surface (SAS) (determined as described in Computational Details) of the carbonyl oxygen and amide hydrogen to a shift of the intrinsic frequency of the amide I oscillator under consideration. The SAS determines the likelihood of making a hydrogen bond to solvent.[64] The contributions are assumed to be additive and the resulting shift of the amide I frequency due to solvation is

$$\delta\omega_i^{\mathrm{S}} = \delta\omega_i^{\mathrm{S,O1}} + \delta\omega_i^{\mathrm{S,O2}} + \delta\omega_i^{\mathrm{S,H}}$$

where O1 indicates the first hydrogen bond to the carbonyl oxygen, O2, the second and NH a hydrogen bond to the amide hydrogen. If hydrogen bond contacts exist between main-chain atoms, these are considered stronger and take priority over hydrogen bonds to solvent. I.e., a shift as a result of non-zero SAS is only included when there is room for making additional hydrogen bonds to solvent.

The frequency downshift as a function of SAS for the three possible hydrogen bonds to solvent are computed from two linear functions of SAS, calculated between two SAS ranges, below and above which the functions are zero and constant, respectively. For the amide hydrogen and the first bond to the carbonyl oxygen, the functions are zero up to a SAS of 1 Å^2, decrease for larger SAS values and reach their minimum values (i.e. maximum downshift) at SAS values that were found through optimization. The shift for the second carbonyl hydrogen bond starts to decrease at 25 Å^2[64] and becomes constant at 35 Å^2. The maximum absolute frequency shifts of all three functions were subject to optimization and resulted in different values. **The intrinsic frequency of tertiary amides, as found in X-Pro linkages, are lower than for secondary amides.[65] Therefore, the intrinsic frequencies of the oscillators before proline residues were modified by a constant value of $-10\ \text{cm}^{-1}$, as determined in initial test runs of the simulation where this parameter was optimized. With the final parameters, we checked the effect of the Pro-shift parameter for a range between -10 and $-40\,\text{cm}^{-1}$ and found that it has in most cases almost no influence on the final spectral shape and in a few cases only a minor impact.**

Interpeptide coupling

Coupling between amide I oscillators is represented by the non-diagonal terms of the exciton Hamiltonian matrix. Long-range coupling between distant oscillators is described by TDC.[1] To calculate the TDC between two amide I oscillators, each oscillator is assigned a transition dipole moment with a point location, magnitude and direction. The coupling β_{ij} is calculated as

$$\beta_{ij} = \frac{0.1}{\varepsilon} \frac{\delta\boldsymbol{\mu}_i \cdot \delta\boldsymbol{\mu}_j - 3(\delta\boldsymbol{\mu}_i \cdot \hat{\boldsymbol{n}}_{ij})(\delta\boldsymbol{\mu}_j \cdot \hat{\boldsymbol{n}}_{ij})}{r_{ij}^3}$$

where $\delta\boldsymbol{\mu}_i$ and $\delta\boldsymbol{\mu}_j$ are the transition dipole moments for peptides i and j. r_{ij} and $\hat{\boldsymbol{n}}_{ij}$ are the distance and unit vector between the transition dipoles, respectively. ε is the dielectric constant, here assumed to be unity. The magnitude and direction were here subject to optimization within ranges found in literature,[1,46] and different parameters were calculated depending on secondary structure classification.

For very short distances the point dipole approximation fails to describe the coupling between peptide units.[66,67] Also, for nearest neighbors mechanical coupling needs to be considered as the C_α connecting the peptides has been shown to be non-stationary during the vibration.[68] The coupling between covalently bonded neighbors in the polypeptide chain is therefore often described using parametrized values from ab initio coupling maps for dipeptides.[24,60,67,69–74] Here, spline interpolation in the DFT/ B3LYP map of Stock and co-workers was used.[73]

Computational Details

Protein set

The rationally selected proteins (RaSP) set developed by Goormaghtigh and co-workers[40] was used for the simulations described in this manuscript. The RaSP set was constructed to represent the maximum range of structural variation exhibited in proteins by including as many different protein folds as defined in CATH[75] and SCOP[76,77] as possible, as well as structures representing different α-helix and β-sheet contents. Also, only proteins for which high quality crystal structures were available and that also were generally commercially available qualified for the set. The set is described in detail in the original publication and the included proteins are listed in Table Table 1 More detailed information on the main characteristics of these proteins can be found in the Supporting Information.

The experimental IR spectra of the proteins were all recorded under the same conditions in the Goormaghtigh laboratory. Measurements were made in ^1H_2O using 3% protein solutions. The contributions from water vapor and buffer were subtracted as described in the original publica-

tion.[40]

To resolve overlapping bands, the resolution of the experimental spectra was here enhanced using fine-structure enhancement (FSE),[78] a technique similar to Fourier self-deconvolution. A weighting factor of 0.82 and a smoothing constant of 17 cm^{-1} were used. The amide I band was isolated by only retaining the range 1610–1720 cm^{-1} of the original spectra[40] and then linearly extrapolating to zero intensity at 1600 cm^{-1} and 1730 cm^{-1}. Finally, the processed experimental spectra were normalized to unit integral. The simulated protein spectra were subjected to the same procedure.

The PDB files corresponding to the proteins in the RaSP set were processed as described in the next section. The biological unit structure specified in the PDB file was used for calculations, with the exception of 1SCS which is in fact present as a dimer under the experimental conditions specified in the original publication,[40] rather than as a tetramer which is stated as the biological unit according to PDB. The proteins were further subdivided into an optimization set (30 proteins) and a prediction set (14 proteins) for the purpose of this paper. The protein selection for each subset aimed to retain a balanced distribution between folds and secondary structure contents.

Simulation protocol

All computations presented here are the result of an in-house developed MATLAB (R2010b, The MathWorks, Inc.) program, except where noted otherwise. Before the actual simulation, a data preparation step was performed for each protein. The PDB file of each protein was first modified to correspond to the biological unit structure by performing the symmetry operations specified in the PDB file. Subsequently, a file with secondary structure classification data per residue was created using the DSSP program.[63] Finally, the HBPLUS software[79,80] was used for adding polar hydrogens to the protein structures. At the beginning of a simulation, the PDB, DSSP and HBPLUS data were processed and dihedral angles, normalized temperature factors[81] and solvent accessible surface (SAS) areas were computed.

The solvent accessible surface (SAS), first introduced by Lee and Richards,[82] was computed

for all backbone amide hydrogen and carbonyl oxygen atoms using a variant of the Shrake and Rupley algorithm.[83] The method is general and includes also internal cavities in the protein which enables to account for hydrogen bonding by internal water molecules. For each atomic SAS to be computed, 1000 test points were approximately evenly distributed on a sphere centered on the atom of interest and with a radius equal to the sum of the van der Waals radius and the water solvent probe radius of 1.4 Å. An even distribution of points was approximated by placing the points along a spiral and separated according to the Golden Ratio. The atom SAS was then computed according to the fraction of test points left after points covered by the SAS spheres of all other atoms in the protein had been removed.

The exciton Hamiltonian for the uncoupled oscillators was constructed as described in the Theoretical Methods section. The coupled modes were found by solving the eigenvalue problem and the spectrum was subsequently computed as a superposition of one Gaussian band (FWHM 16 cm^{-1}) per mode. Finally, the intensity was subsequently computed from the eigenvectors and transition dipole moments.

The simulated spectra displayed in the Results section were enhanced using the FSE procedure as described in the previous section, unless stated otherwise. The spectra were computed with 0.5 cm^{-1} spacing and normalized to unit integral.

Optimization procedure

An optimization procedure was performed on simulated spectra of a subset of proteins in the RaSP set (detailed in Table 1) to find parameters of the physical models that yield optimum agreement between simulated and experimental spectra. The values of the parameters were the same for all proteins. The procedure started with the calculation of an initial spectrum per protein based on initial guesses for the parameters, as detailed in Table 2. The parameters influencing the spectra were then iteratively optimized within specified ranges as listed in Table 2, with the goal of finding agreement with experimental spectra, until convergence was reached.

The parameters were optimized using a curve-fitting procedure (MATLAB `lsqnonlin` func-

tion). The optimization minimizes the objective function with respect to the physical model parameters. The objective function was here the average of the sum of squares of the deviation between simulated and experimental spectra for all optimized proteins, i.e.,

$$\min_{\text{parameters}} \sum_{\text{proteins}} \left(\sum_{i} 1000 \times \left(s_{\text{exp}}(\nu_i) - s_{\text{sim}}(\nu_i) \right)^2 \right) / N_{\text{proteins}}$$

The objective function yields for a given set of parameters the objective value and the inner sum is here the objective value for an individual protein. Presented objective values refer to those corresponding to the optimized, final set of parameters.

The goal of the optimization is to find parameters suitable for prediction of absorption spectra that have not been subject to FSE. However, as spectral features are more prominent in enhanced spectra, the parameters are optimized via comparison of spectra after FSE. The FSE procedure is applied to both experimental and simulated spectra in every step of the optimization. The objective function thus quantifies the difference between those spectra.

The time required for optimization was linearly dependent on the number of optimized parameters and was thus dominated by the intrinsic frequency shift maps. The final objective values per protein are listed in Table 1 and the resulting optimized parameters used for prediction of amide I spectra are presented in Table 2.

Results and Discussion

Overview

The amide I band of 44 proteins was simulated using empirical parameters determined from optimization of a subset of 30 simulated spectra against experimental spectra. 14 spectra were excluded from optimization and instead used for testing of the prediction quality of the models and empirical parameters. The effects taken into consideration when determining the intrinsic frequencies included interpeptide hydrogen bonding, solvation and the effects from local conformation. In-

terpeptide coupling was subdivided into interactions between nearest neighbors in sequence and those between all other. The former was modeled from DFT data and the latter by TDC.

The optimized model parameters are listed in the last column of Table 2. The allowed parameter ranges were constrained to what were considered as realistic values for the respective parameters from an inspection of the literature. Constraining the parameters was necessary in order to guide the optimization in the many-dimensional parameter space and to reduce the computational time required for optimization.

The quality of the optimization and prediction can be assessed by visual inspection of the spectra, as well as by the final objective value per protein, which measures the deviation between simulated and experimental spectra for a particular protein. The simulated spectra are compared to experiments and to spectra calculated using the floating oscillator model with the parameters of Torii and Tasumi.[11]

Agreement between optimized protein spectra and experimental results

The proteins that were used for optimization of the model parameters are indicated in Table 1. Representative optimized FSE spectra for four selected proteins are shown in Figure 1. These include the apolipoprotein E3 (1LPE, 143 oscillators) which is highly α-helical (82%) and erabutoxin b (3EBX, 61 oscillators) with 44% β-sheet content. Representatives of proteins with mixed α-helix/β-sheet content include dihydropteridine reductase (1DHR, 470 oscillators) with 37% α-helices and 25% β-sheets and penicillin amidohydrolase (1PNK, 748 oscillators) with 32% α-helices and 21% β-sheets. **The complete set of optimized protein FSE spectra can be found in Supporting Information along with a figure for the unprocessed absorbance spectra**. The spectral shape of erabutoxin b (3EBX) is very well reproduced with intensity missing in only a small wavenumber range around 1690 cm^{-1}, as reflected in the very low objective value of 0.25. Also apolipoprotein E3 (1LPE) is relatively well reproduced with slight deviations in the reproduced intensity in the lower wavenumber region and an objective value of 0.55. Note that the apparent band at 1610 cm^{-1} is due to the linear extrapolation to zero absorbance between 1610

and 1600 cm^{-1} and that the missing intensity in the simulation might be due to the neglect of side chain absorption in this region. In the optimized spectrum of dihydropteridine reductase (1DHR) the two main bands are not resolved, but instead partially overlap and result in a large deviation from the experimental spectral shape and the second highest objective value in this protein subset (0.88). Also, there is intensity missing around 1680 cm^{-1}, where there is a small band in the experimental spectrum. The penicillin amidohydrolase (1PNK) has an objective value that is 0.55 and deviates slightly in intensity from the experimental spectrum over the whole wavenumber range. The optimized spectrum appears slightly shifted toward higher wavenumbers and lacks any sharp features. **Comparing the unprocessed optimized and experimental spectra (see Supporting Information) leads to similar conclusions.**

The objective values measuring the deviation between optimized spectra and experimental spectra are in the range 0.25–0.87 (exception 3PKG that has objective value of 2.04) and there is an overall good visual agreement between optimized and experimental spectra. Particularly, cytochrome c (1HRC, 103 oscillators), thaumatin (1THW, 412 oscillators) and DD-transpeptidase (3PTE, 346 oscillators) could be optimized to excellently match experimental spectra. The agreement is reflected in the very good objective values, they are in the range 0.28-0.33. Cytochrome c is a fairly small, highly α-helical (41%) protein, whereas both DD-transpeptidase and thaumatin are mid-sized proteins of mixed secondary structure content.

For proteins with $> 40\%$ α-helical content, the central peak is reproduced in all cases. Its position was generally matched to within 2.3 cm^{-1}, except for 1ISC where the main band position deviated by 5.3 cm^{-1}. The relative intensity was reproduced within 10% of the experimental value. The amide I absorbance is well contained within the 1600–1700 cm^{-1} range. Other spectral features, such as side-bands and shoulders, are in most cases well reproduced in the optimized spectra, though with minor, and in a few cases significant, deviations in position and intensity. The agreement for some proteins that are rich in β-sheet content, such as 1SXC and 1THW, is excellent. The overall impression is, however, slightly less visually satisfying than for proteins with high α-helical content as the experimental spectra of β-rich proteins contain more features.

It can be noted that the measure of spectral agreement chosen here (sum of squared deviations) does not always provide a fair correspondence to visual inspection. This is evident for the spectrum of, e.g., α-lactalbumin (1HML), which by visual inspection can be judged as well reproduced as it displays a shape that closely matches the experimental spectrum. The entire band is, however, shifted about 5 cm^{-1} towards higher wavenumbers, thereby producing a comparatively high objective value of 1.1. While posing a certain complication when evaluating results, this deficiency, which results from an imperfect description of deviations between spectra, likely results in a suboptimal optimization. Defining an objective function better suited to describe the important differences and overall agreement between spectra would improve the result of the optimization. This issue has been noted also by Watson and Hirst,[14] who used the Pearson product as a measure of spectral agreement. A brief attempt to here use the Pearson product as an objective function yielded much less satisfactory results than the current one. This was attributed to that too few spectral details were captured by the Pearson product.

Agreement between predicted protein spectra and experimental results

The proteins that were used for the prediction test are indicated in Table 1. The resulting prediction quality for four selected proteins with diverse structural properties (see Figure 2) is commented upon in particular. The complete set of predicted protein spectra (FSE spectra and unprocessed spectra) can be found in the Supporting Information.

The four proteins in Figure 2 include myoglobin (1YMB), a 153-oscillator protein with 74% α-content and concavalin A (1SCS, 237 oscillators) with 47% β-structure. The two other proteins, papain (1PPN, 212 oscillators) and ubiquitin (1UBI, 122 oscillators) represent proteins with mixed α/β-content. All four proteins have been previously investigated in different amide I simulations.[10,11,14,33]

For ubiquitin (1UBI) the central peak is well matched within 1 cm^{-1} and the intensity is excellently matched. Part of the shoulder around 1670 cm^{-1} is reproduced but shifted and there is intensity missing at higher wavenumbers in the simulated spectrum, which could be due to the

downshifted shoulder or the narrow component band in the experimental spectrum. Resolution enhancement algorithms enhance the intensity more for narrow bands than for broader bands. Around 1630 cm^{-1} a shoulder band is present that is less obvious in the experimental spectrum. The simulated spectrum of myoglobin (1YMB) does not well reproduce the experimental spectrum, as is reflected in the high objective value of 1.81. The simulated central peak has split into two large, partially overlapping peaks, resulting in a sloping shoulder towards higher wavenumbers. **Nevertheless, when the unprocessed spectra are compared (Supporting Information), the agreement is reasonable.** The papain (1PPN) spectrum can be considered as very well predicted, with the central peak and both the shoulder to the left and the one to the right well reproduced. However, the entire spectrum has too low intensity due to some absorption outside the 1610-1690 cm^{-1} interval. The main peak of concavalin A (1SCS) in the predicted spectrum is 5 cm^{-1} too high as compared to the experimental spectrum and the overall shape is not well reproduced. The experimental spectrum of concavalin A has a low-wavenumber β-sheet component around 1620 cm^{-1} and a high wavenumber band around 1690 cm^{-1} arising from β-sheet absorption. The 1620 cm^{-1} band is present in the predicted spectrum but with too high intensity, and the distinct high wavenumber is not reproduced. Instead, the shoulder around 1675 cm^{-1} has higher intensity in the simulation. The incapability to reproduce high wavenumber β-sheet absorption is a common problem for the optimized and predicted spectra presented in this paper and is further discussed below. **Apart from the particular case mentioned, the conclusions drawn from the FSE spectra also apply to the unprocessed spectra (Supporting Information).**

Comparison of simulation quality between optimized/predicted spectra and TDC

The agreement with experiment is not as satisfying for the predicted spectra as for the optimized spectra. The objective values are in the range 0.34–1.3, except for myoglobin that has an objective value of 1.8. The average objective value for the predicted spectra is 0.88, compared to 0.53 for the optimized proteins. The variation is visualized in the top panel of Figure 3A.

The central peak position is less well reproduced than for the optimized spectra. The central peak position deviates on average 3.9 cm^{-1} from the experimental spectra for the predicted proteins, the corresponding value for the optimized protein spectra is 2.4 cm^{-1}. The deviation in peak position between experimental and simulated spectra is shown in Figure 3B. The standard deviation of the difference in FWHM between simulated and experimental spectra is almost twice as large for the predicted proteins as for the optimized proteins as shown in Figure 3C.

In Figure 1 and Figure 2, both the optimized and predicted spectra are compared to spectra calculated using only TDC. Spectra labeled "TT" were calculated using the original parameter values of Torii and Tasumi, although without making the non-systematic adjustments of diagonal elements of particular proteins as described in the original paper.[11] Figure 3 shows the deviation from experiment for the position of the amide I band maximum and FWHM. In both cases, the spread of the distributions is much larger than for the optimized or predicted spectra. In general, the wavenumber distribution in the TT spectra is narrower than that of both experimental spectra and the spectra simulated with our approach and the TT spectra do not have intensity in the full 1600–1700 cm^{-1} range. The TT spectra do in many cases exhibit an excellent agreement for the maximum peak position and usually also reproduce the position of the second major spectral feature, e.g., shoulder or side peak, although the intensity is generally not well reproduced. The main improvement of the predicted spectra based in optimized parameters over the TDC approach is that the amide I band intensity is distributed over the entire 1600-1700 cm^{-1} range and that the relative intensities of spectral features are much better reproduced. There is also less variation in prediction quality between different proteins.

Comparisons to other simulations

There are only a few simulations of protein amide I spectra in the literature and some are difficult to compare to our results because of different spectra processing. This applies to the simulations by Watson & Hirst[14] who have used a stronger resolution enhancement than we did. Brauner et al.[15] have simulated six proteins with an approach similar to ours, three of which (myoglobin 1YMB,

papain 1PPN, lysozyme 1HEL) are contained in our set of predicted proteins. Their simulated unprocessed spectra seem to match the experimental spectra equally well or slightly better than ours (see Supporting Information) and we would like to point out that the conditions for this comparison are unfavorable for our approach for the following reasons: (i) Their parameters were empirically chosen to achieve a good match between simulation and experiment for the six simulated protein spectra, which include the three spectra that are compared here. In contrast, in our work these three spectra were not included in the set of spectra used for parameter optimization. In our set of optimized spectra most of the simulations show an excellent agreement comparable to that obtained by Brauner et al. (ii) Using six proteins to optimize the simulation likely provides more consistent parameters than using 30 proteins which were chosen for maximum structural variety as in our case. (iii) The parameters in the previous study were chosen to reproduce unprocessed spectra, whereas our parameters were optimized to reproduce FSE spectra. Thus our parameters were not optimized for the spectra compared here. A good match between unprocessed experimental and simulated spectra is no guarantee for a good match of resolution enhanced spectra. For example, the unprocessed spectra of several proteins (4PEP, 1ISC, 1HDA, 1OVA, 1ARV, see Supplementary Information) show a good agreement between simulation and experiment, whereas the resolution enhanced spectra reveal clear deviations for some component bands. In an initial phase of this work we attempted also to optimize the simulation parameters against the unprocessed absorption spectra. Using the current objective function, this resulted in featureless spectra which did not reproduce the experimental spectra.

Two further studies [10,33] have simulated ubiquitin, and their results can be compared with our simulation shown in Figure 2 (1UBI). The unprocessed absorption spectra shown in both studies distinguish only two features: a main band near 1640 cm^{-1} and a shoulder near 1670 cm^{-1}. In contrast, our FSE spectra distinguish four bands that need to be simulated which illustrates the challenge in this work. The simulation by Choi et al. has a shoulder at the position of the experimental main band, the main band is simulated at higher wavenumbers and the high wavenumber shoulder is less prominent and more smeared out up to 1750 cm^{-1}. Our simulation predicts the po-

sition of the main band correctly and closely reproduces the high wavenumber component bands. Note that the ubiquitin spectrum was predicted using parameters optimized for other proteins. Common for our simulation and that by Choi is a more prominent appearance of a shoulder with lower wavenumber than the main band.

Ganim & Tokmakoff simulate ubiquitin with several simulation protocols, which differ by the implementation of electrostatic interactions. Common to those with intrinsic frequency heterogeneity is that the spectrum is broader than experimentally observed and that the high wavenumber shoulder is more prominent than in the experimental spectrum. For two of the simulations the main band position is closely reproduced and no shoulder is apparent at a wavenumber lower than that of the main band. We can conclude that our simulation is successful in predicting the position of the main band and in restricting the amide I absorption to the experimentally observed spectral range.

Interpeptide hydrogen bonding

The coefficients describing the shift of the intrinsic frequency due to interpeptide hydrogen bonding were optimized to final values of $\xi^{\mathrm{O}} = 1.8$ cm^{-1}/kcal and $\xi^{\mathrm{H}} = 1.0$ cm^{-1}/kcal, similar to the values of 2.4 cm^{-1}/kcal and 1.0 cm^{-1}/kcal presented in the original paper,[62] though the ξ^{O} value was optimized to its upper bounding value. 78% of the oscillators were found to be affected by interpeptide hydrogen bonds. For these, the average shift is about -19 cm^{-1}.

For several of the β-rich protein spectra, both the optimization and the prediction failed to reproduce the intensity at the characteristic low- and high-wavenumber β-sheet band positions. The failure to reproduce the high wavenumber band can be seen, e.g., in the case of 1MOL, 1LEN and 3EBX as well as 7AHL. Problems with reproducing the low wavenumber band are evident in e.g., 8FAB and 1SCS. Attempts were made to solve this issue by modifying the intrinsic frequencies of oscillators that belong to inner strands of β-sheets by a constant value shift found through the optimization process. The aim of this modification was to eliminate low-wavenumber bands in proteins with many-stranded β-sheets, found to originate from inner-strand residues of larger sheets as pinpointed by analysis of the eigenvector elements. The need for this type of modifica-

tion has previously been reported by others.[11,84] Inclusion of this type of parametrized shift often yielded better agreement for proteins with high β-sheet content, such as in the case of concavalin A. Using a shift value of 15 cm^{-1} for inner-strand oscillators determined using the optimization procedure, a correct central peak position and a very good match of the entire spectral shape could be achieved.

Effects of solvation

The experimental spectra were measured in solvent water H_2O and thus the effect of solvation on the amide I oscillator frequencies needed to be incorporated. The effect of solvation was taken into account by correlating the solvent accessibility of the carbonyl oxygen and of the amide hydrogen to a redshift of the intrinsic frequencies. This takes into account the heterogeneity of the local oscillator environments in a simple fashion. Here, 52% of all oscillators were affected by hydrogen bonds to solvent.

Ubiquitin (1UBI, 75 oscillators), insulin (1PBH, 49 oscillators) and erabutoxin b (3EBX, 61 oscillators) are the smallest proteins in the test set. The features in the spectra of all three proteins are excellently reproduced. As the effect of solvation on the simulated spectra is more important for the small proteins, they provide a more critical test of the solvation model. The good spectral agreement for these proteins indicates that the solvation model utilized in the simulations is reasonable. In small proteins, the termini, which are expected to be more flexible than the rest of the protein, constitute a larger percentage of the total structure. Therefore, the spectra of small proteins should also be more sensitive to the neglect of dynamics. Again, the good agreement supports the validity of this simplification.

Although absorption arising from side chains in the amide I region, was not accounted for in the simulation procedure, they implicitly effect the oscillator frequencies by providing shielding of the backbone from solvent.[37]

Local conformation maps

The optimized maps that describe the shift of the intrinsic frequency due to the local conformation, as defined by the dihedral angles (ϕ, ψ) are shown in Figure 4. These maps display the overall characteristic form of maps published in previous studies[24,58–60] by construction.

The number of parameters influencing the intrinsic frequency shift maps is 25 per map. Attempts using fewer parameters led to a degradation of the quality of both optimized and simulated spectra. Extending the number of parameters to about 350 by individual point optimization of the whole Ramachandran space did not provide any discernible improvement. The data point separation used here was 30°.[60] The effect of resolution of the maps still remains to be investigated, some efforts have been made toward improving the resolution of specific regions.[74]

The approach of utilizing dipeptide ab initio data to calculate the effect on the intrinsic frequency exerted by the nearest neighbor residues has been implemented in several simulations of short polypeptides.[24,29,60,85] Ge and co-workers have, however, shown that this approach leads to worse agreement with experimental results than using only electrostatic methods for simulation of amide I spectra of an octapeptide.[86]

In an attempt to evaluate such published maps, we used the DFT map by La Cour Jansen et al.[60] for the intrinsic frequency shifts without any modification and re-optimized all other parameters. This yielded, after re-optimization of all other parameters, an average objective value of 8.7 per protein, much worse than previously attained value of 0.53 using the optimized map. The high objective value resulted from an amide I band with approximately correct shape, but shifted to a much too high wavenumber range and little overlap between the experimental and simulated spectra.

Gorbunov et al.[59] have, however, published shift maps both for dipeptides in solvent and gas phase environment that show very different magnitudes of shift, but are very similar in shape. In a solvated large protein, the local environment is not well represented by either case, but should be somewhere in between. In an attempt to take into account the effect of protein environment on the local conformation shift, while still using the published maps as a basis, we carried out a simplified

version of the optimization procedure. This involved not using any individual point optimization as done in our simulation only for the most occupied regions of Ramachandran space, but only optimization of multiplicative factors and a global shift applied to both maps. The multiplication factors for the two maps were then lowered to about 0.3 by the optimization and the global shift was optimized to around -43 cm^{-1}. The average objective value per protein resulting from this procedure was 1.9. Visual inspection conveys that the shape of the spectra are slightly better than using plain TDC.

Coupling interactions

The results presented here were computed with different TDC parameters depending on secondary structure as classified by DSSP. Different TDC parameters were optimized for sheets, helices and other, and yielded values according to Table 2. Optimization using only one parameter set was tested (data not shown) and resulted in worse agreement with experimental spectra, the objective values were about a factor two higher than those presented here.

Use of different parameter sets can be motivated by considering, e.g., the different hydrogen bonding patterns and strengths associated with the different secondary structure types. These should result in different charge distributions along both the C-N and C=O bonds, which in turn should influence the direction and magnitude of the transition dipole moment. However, it is not clear-cut if the parametrization/classification should be based on the criteria in DSSP or if effects other than hydrogen bonding patterns used to define secondary structure are the first-order effects. The transition dipole moment magnitude has also been shown to vary slightly based on DFT calculations[58,87] and also to be influenced by, e.g., exposure to solvent.[27] Both the magnitude and direction parameters of the TDM were subject to optimization, however, not the position. Different positions have been proposed in literature,[11,66] but as the position is not readily observable, there is little guidance as to the possible range.

The nearest neighbor couplings were computed using a published ab initio map[59] without any modifications. This choice is supported by the generally accepted use of this type of maps to de-

scribe short-range coupling, although the transferability and potential effects of incorporation into longer polypeptides have not been rigorously investigated. During the model development phase, we investigated if optimization to experiment could be improved by modification by a multiplicative factor and a baseline shift of the coupling map. The map values were not significantly changed, which further motivated using the coupling map without modifications.

Discussion of the optimization approach

To solve a multidimensional spectra-fitting problem with a total of 66 parameters, where each of them influences the simulated spectra in complex ways, is a challenging task for the optimization algorithms used here. 50 out of the 66 dimensions in the parameter space were associated with the shift maps for the oscillator intrinsic frequency. The optimization procedure that resulted in the parameter set presented here required about 30 CPU core hours processing time on a current desktop computer. It is of course desirable to further reduce the number of parameters in order to improve the speed of the `lsqnonlin` optimization algorithm. The intrinsic frequency shift maps contains the majority of the parameters and would be the prime candidates for parameter reduction. The required number of parameters was however extensively investigated, and found to produce significantly worse higher objective values if the number of parameters is reduced. Map parametrization using many more (350) parameters has also been tested, yielding objective values similar to what is presented here.

It is important to note that a given spectral feature in general is influenced by several of the parameters and the optimized parameters thus may suffer from crosstalk. A possible example for this are our intrinsic frequency shift maps (Fig. 4), which are more negative than those derived from DFT calculations, probably because they include a uniform shift due to hydrogen bonding. The optimized parameters are thus, in principle and per construction, only usable together with the same physical models and parameters that were used in the optimization procedure. Generally, improved physical models with fewer parameters will aid in avoiding negative side effects such as cross talk from the optimization.

Some of the physical models/parametrization schemes evaluated during the method development phase yielded less well conditioned optimization problems. These resulted in severe optimization problems where the MATLAB `lsqnonlin` function terminated early due to getting stuck in local minima. In an attempt to mitigate this problem, the simulated annealing optimization method was evaluated (`simmulannealbnd`) using the same objective function as previously. This method yielded a very slow convergence of the problem, likely due to its inherent property to try to sample the very large many-dimensional parameter space, and was therefore discarded from further use.

We note that the parameters derived here describe the unprocessed spectrum, not the fine-structure enhanced spectrum. Thus they can readily be incorporated into other simulation programs. After generating the simulated absorption spectrum, FSE is used on both the experimental and the simulated spectra to compare them and to optimize the values of the parameters. This equal treatment of experimental and simulated spectra makes the parameters less dependent on the particularities of the resolution enhancement method. As an extreme example, if we had chosen a too strong resolution enhancement, this would have produced artifact bands in the resolution enhanced spectrum which do not correspond to component bands of the unprocessed spectrum. Directly simulating the resolution enhanced spectrum would therefore have generated a model with strong amide I modes in spectral regions where they are not observed experimentally. In contrast, simulating the unprocessed absorption spectrum gives meaningful results even in this case: since simulated and experimental spectrum are treated in the same way, the similarity of the processed spectra testifies the mathematical similarity between experimental and simulated absorption spectra, although the artifact bands in the resolution enhanced spectra do not have a physical meaning.

Model limitations and simplifications

Feasible simulation of protein amide I IR spectra requires a number simplifications and approximations to be made. The fundamental assumption of a separable amide I subspace reduces the complexity of the problem to only amide I oscillators, rather than considering all atoms in the

polypeptide chain. **Furthermore, the polypeptide is simplified to a chain of linked peptide units, which means that side chain contributions to the absorption in the amide I range and the coupling between side chain vibrations and the amide I vibrations are neglected.** Recent procedures for amide I simulations have simulated also the amino acid absorption[85] or subtracted this contribution from the experimental spectrum used for comparison.[37] Subtraction of the intensity originating from the amino acids is however not a straight-forward procedure as the absorption will be influenced by the protein environment in which they are immersed,[2] and was therefore not performed here. Our approach of optimizing against resolution enhanced spectra takes partially care of this problem, as described in the following. The side chain contribution in the amide I region stems from Arg, Asn, Gln, His^+, Lys, Trp, and Tyr. **Most of them are polar and therefore expected to be involved in dynamic interactions with varying strengths, often with water at the protein surface. Thus, their absorption bands are broader (30** cm^{-1} **and more)[88] than the amide I bands which were modeled here with a width of 16** cm^{-1}**. Therefore, the side chain contributions are suppressed by resolution enhancement methods like FSE, which instead predominantly enhances the narrow component bands arising from amide I vibrations. In the same way, FSE takes care of small errors in the subtraction of the absorption of the water bending vibration. Since the water band is broad, it will be suppressed by FSE. Small errors in water subtraction will therefore not influence our simulations. A further simplification employed here is the use of general maps and parameters for all amino acids for the description of local conformation effects, nearest neighbor interactions and hydrogen bonding. These might well depend on the particular type of amino acid in question and maps for Pro have been published recently.[89]**

As previously mentioned in relation to calculation of solvation effects, the dynamics, i.e., the structural fluctuations of the protein and the solvent and the interplay between these two is not considered in these calculations. The neglect of including dynamics is expected to have consequences mainly for the intrinsic frequencies and in particular for the solvated oscillators. The coupling elements are, however, not affected as much by structural fluctuations.[33] An attempt to

include band broadening effects due to structural flexibility was made here by correlating residue temperature factors (B-factors) to the degree of conformational flexibility. Average temperature factors for each peptide unit were computed as the mean of the normalized temperature factors[81] of the constituent C, O and N backbone atoms. The temperature factor of a mode was computed as the average of all peptide unit temperature factors weighted by the squared elements of the mode eigenvector normalized to unit sum. The function relating the mode temperature factor to the FWHM was modeled as a linear function between two temperature factors and constant below and above said temperature factors. Several variations of this scheme were attempted, however no clear improvement of the prediction quality was found. Using band narrowing techniques like FSE partially resolves also this problem. Their property of enhancing narrow bands more than broad bands gears the optimization towards describing well structured protein segments with narrow bands. This is an advantage since segments that give rise to broader bands will be more dynamic and are likely to be less well represented by the static X-ray structure. **Similarly, FSE reduces also the effect of possible multiple protein conformations in solution. Often, the difference between these conformations is restricted to small flexible linker regions, whereas the majority of the residues reside in rigid domains and are little affected. In line with this, infrared difference spectroscopy of protein reactions usually detects only small changes in the amide I region.[2] Using FSE, the contribution of the flexible linkers will be reduced and the simulation is geared towards describing the rigid domains.**

Conclusions and Outlook

Our optimized parameter set for the IR amide I band reproduced the experimental spectra very well. The spectra predicted using the optimized parameters showed somewhat less agreement to experimental spectra than the optimized ones, but the prediction quality clearly superseded that of considering only the effect of TDC on oscillators with the same intrinsic frequencies.

Important venues for improvement of the prediction quality should be focused at improving

the parametrized descriptions of the underlying physical effects. The models that are currently used are highly approximate, providing fairly realistic but not very accurate values. In some cases, only the general trend of a physical effect is captured. Also additivity of the effects is generally assumed. Future improvements with respect to the intrinsic frequency maps could include a more precise dependence on the polypeptide backbone dihedral angles by not using two separate maps, but a single multi-dimensional map. A more well-suited objective function for use in the optimization procedure, in terms of how good measure of similarity between spectra it provides, could greatly improve results. Other known effects that could be investigated for implementation in the simulation model include, e.g., cooperative hydrogen bonding effects arising in long α-helices.[90]

In terms of experimental data available to optimize against, the procedure would benefit from more small proteins as well as proteins that add data points to the less populated regions of Ramachandran space. An increased number of relatively small proteins would yield benefits in terms of clearer correspondence between the simulated absorption of individual modes to features in the experimental spectra, and thus more clear-cut optimization, due to less spectral crowding.

The capability to simulate amide I spectra of large proteins is ultimately useful when it comes to calculation of difference spectra, which allow spectrum-structure analysis on a higher level of detail. Application to the calculation of difference spectra, or other methods of selective observation such as isotope labeling, would truly assess the predictive quality of the optimized parameters. Also, it would be relevant to investigate the transferability of the current parameters in terms of prediction quality to spectra measured under other conditions. If not directly transferable, the procedure to re-optimize the parameters for the desired conditions should be straight-forward, but requires appropriate experimental data. More collections of experimental spectra measured using standardized protocols would be of great use for theoretical efforts, such as the one presented here, as well as for development of secondary structure estimation methods.

Acknowledgment

The authors would like to gratefully acknowledge E. Goormaghtigh for making available the experimental protein spectra utilized in the calculations.

Supporting Information Available

SI Table 1 is analogous to Table 1 but also includes structural information for all proteins included in the simulations. SI Figure 1 shows the full set of optimized, predicted and experimental spectra for all proteins. Also spectra simulated using only TDC are included for comparison. SI Figure 2 shows the full set of optimized, predicted and experimental spectra without fine-structure enhancement for all proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Krimm, S.; Bandekar, J. *Adv. Protein Chem.* **1986**, *38*, 181–367.

(2) Barth, A.; Zscherp, C. *Q. Rev. Biophys.* **2002**, *35*, 369–430.

(3) Siebert, F.; Hildebrandt, P. *Vibrational Spectroscopy in Life Science*, 1st ed.; Wiley-VCH GmbH & Co.: Weinheim, 2008.

(4) Barth, A., Haris, P. I., Eds. *Biological and Biomedical Infrared Spectroscopy*; IOS Press: Amsterdam, 2009.

(5) Fabian, H.; Naumann, D. In *Protein Folding and Misfolding - Shining Light by Infrared Spectroscopy*, 1st ed.; Fabian, H., Naumann, D., Eds.; Springer-Verlag: Berlin, Heidelberg, 2012.

(6) Besley, N. A.; Metcalf, K. A. *J. Chem. Phys.* **2007**, *126*, 035101–035107.

(7) Cho, M. *Chem. Rev.* **2008**, *108*, 1331–1418.

(8) Amadei, A.; Daidone, I.; Di Nola, A.; Aschi, M. *Curr. Opin. Struct. Biol.* **2010**, *20*, 155–161.

(9) Miyazawa, T. *J. Chem. Phys.* **1960**, *32*, 1647–1652.

(10) Choi, J.-h.; Lee, H.; Lee, K.-k.; Hahn, S.; Cho, M. *J. Chem. Phys.* **2007**, *126*, 045102–045114.

(11) Torii, H.; Tasumi, M. *J. Chem. Phys.* **1992**, *96*, 3379–3387.

(12) Abe, Y.; Krimm, S. *Biopolymers* **1972**, *11*, 1817–1839.

(13) Krimm, S.; Abe, Y. *Proc. Natl. Acad. Sci. U.S.A.* **1972**, *69*, 2788–2792.

(14) Watson, T. M.; Hirst, J. D. *Phys. Chem. Chem. Phys.* **2004**, *6*, 998–1005.

(15) Brauner, J. W.; Flach, C. R.; Mendelsohn, R. *J. Am. Chem. Soc.* **2005**, *127*, 100–109.

(16) Ham, S.; Cho, M. *J. Chem. Phys.* **2003**, *118*, 6915.

(17) Ham, S.; Cha, S.; Choi, J.-H.; Cho, M. *J. Chem. Phys.* **2003**, *119*, 1451–1461.

(18) Cho, M. *J. Chem. Phys.* **2003**, *118*, 3480–3490.

(19) Bour, P.; Keiderling, T. A. *J. Chem. Phys.* **2003**, *119*, 11253–11262.

(20) DeCamp, M. F.; Deflores, L. P.; McCracken, J. M.; Tokmakoff, A.; Kwac, K.; Cho, M. *J. Phys. Chem. B* **2005**, *109*, 11016–11026.

(21) Kwac, K.; Cho, M. *J. Chem. Phys.* **2003**, *119*, 2247–2255.

(22) Cai, K.; Han, C.; Wang, J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 9149–9159.

(23) Maekawa, H.; Ge, N.-H. *J. Phys. Chem. B* **2010**, *114*, 1434–1446.

(24) Watson, T. M.; Hirst, J. D. *Mol. Phys.* **2005**, *103*, 1531–1546.

(25) Schmidt, J. R.; Corcelli, S. A.; Skinner, J. L. *J. Chem. Phys.* **2004**, *121*, 8887–8896.

(26) Bour, P.; Michalík, D.; Kapitán, J. *J. Chem. Phys.* **2005**, *122*, 144501–144509.

(27) la Cour Jansen, T.; Knoester, J. *J. Chem. Phys.* **2006**, *124*, 044502.

(28) Hayashi, T.; Zhuang, W.; Mukamel, S. *J. Phys. Chem. A* **2005**, *109*, 9747–9759.

(29) Lin, Y.-S.; Shorb, J. M.; Mukherjee, P.; Zanni, M. T.; Skinner, J. L. *J. Phys. Chem. B* **2009**, *113*, 592–602.

(30) Bloem, R.; Dijkstra, A. G.; la Cour Jansen, T.; Knoester, J. *J. Chem. Phys.* **2008**, *129*, 055101.

(31) Torii, H. *J. Phys. Chem. A* **2004**, *108*, 7272–7280.

(32) Ham, S.; Kim, J.-H.; Lee, H.; Cho, M. *J. Chem. Phys.* **2003**, *118*, 3491–3498.

(33) Ganim, Z.; Tokmakoff, A. *Biophys. J.* **2006**, *91*, 2636–2646.

(34) Chung, H. S.; Tokmakoff, A. *J. Phys. Chem. B* **2006**, *110*, 2888–2898.

(35) Mukherjee, P.; Kass, I.; Arkin, I. T.; Zanni, M. T. *J. Phys. Chem. B* **2006**, *110*, 24740–24749.

(36) Choi, J.-h.; Cho, M. In *Biological and Biomedical Infrared Spectroscopy*; Barth, A., Haris, P. I., Eds.; IOS Press: Amsterdam, 2009; pp 224–260.

(37) Grahnen, J. A.; Amunson, K. E.; Kubelka, J. *J. Phys. Chem. B* **2010**, *114*, 13011–13020.

(38) Bour, P.; Sopková, J.; Bednárová, L.; Malôn, P.; Keiderling, T. A. *J. Comput. Chem.* **1997**, *18*, 646–659.

(39) Kubelka, J.; Bour, P.; Keiderling, T. A. In *Biological and Biomedical Infrared Spectroscopy*; Barth, A., Haris, P. I., Eds.; IOS Press: Amsterdam, 2009; pp 178–223.

(40) Oberg, K. A.; Ruysschaert, J.-m.; Goormaghtigh, E. *Protein Sci.* **2003**, *12*, 2015–2031.

(41) Hexter, R. M. *J. Chem Phys.* **1960**, *33*, 1833–1841.

(42) Frenkel, J. *Phys. Rev.* **1931**, *37*, 17–44.

(43) Kasha, M.; Rawls, H. R.; Ashraf El-Bayoumi, M. *Pure and Applied Chemistry* **1965**, *11*, 371–392.

(44) Hamm, P.; Lim, M.; Hochstrasser, R. M. *J. Phys. Chem. B* **1998**, *102*, 6123–6138.

(45) Hamm, P.; Zanni, M. T. *Concepts and Methods of 2D Infrared Spectrscopy*, 1st ed.; Cambridge University Press: Cambridge, 2011.

(46) Torii, H.; Tasumi, M. In *Infrared spectroscopy of biomolecules*; Mantsch, H. H., Chapman, D., Eds.; Wiley Liss: New York, 1996; pp 1–17.

(47) Kubelka, J.; Keiderling, T. A. *J. Phys. Chem. A* **2001**, *105*, 10922–10928.

(48) Venkatachalapathi, Y. V.; Mierke, D. F.; Taulane, J. P.; Goodman, M. *Biopolymers* **1987**, *26*, 763–773.

(49) Jones, L. *J. Mol. Spec.* **1963**, *11*, 411–421.

(50) Mirkin, N. G.; Krimm, S. *J. Am. Chem. Soc.* **1991**, *113*, 9742–9747.

(51) Guo, H.; Karplus, M. *J. Phys. Chem.* **1992**, *96*, 7273–7287.

(52) Torii, H.; Tatsumi, T.; Kanazawa, T.; Tasumi, M. *J. Phys. Chem. B* **1998**, *102*, 309–314.

(53) Torii, H.; Tatsumi, T.; Tasumi, M. *J. Raman Spectrosc.* **1998**, *29*, 537–546.

(54) Kubelka, J.; Keiderling, T. A. *J. Phys. Chem. A* **2001**, *105*, 10922–10928.

(55) Besley, N. A. *J. Phys. Chem. A* **2004**, *108*, 10794–10800.

(56) Mennucci, B.; Martínez, J. M. *J. Phys. Chem. B* **2005**, *109*, 9818–9829.

(57) Jeon, J.; Cho, M. *New J. Phys.* **2010**, *12*, 065001.

(58) Choi, J.-H.; Cho, M. *J. Chem. Phys.* **2004**, *120*, 4383–4392.

(59) Gorbunov, R. D.; Kosov, D. S.; Stock, G. *J. Chem. Phys.* **2005**, *122*, 224904.

(60) la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J. *J. Chem. Phys.* **2006**, *125*, 44312.

(61) Torii, H.; Tatsumi, T.; Tasumi, M. *Mikrochim. Acta Suppl.* **1997**, *14*, 531–533.

(62) Maekawa, H.; Toniolo, C.; Broxterman, Q. B.; Ge, N.-H. *J. Phys. Chem. B* **2007**, *111*, 3222–35.

(63) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

(64) Petukhov, M.; Rychkov, G.; Firsov, L.; Serrano, L. *Protein Sci.* **2004**, *13*, 2120–2129.

(65) Doyle, B. B.; Traub, W.; Lorenzi, G. P.; Blout, E. R. *Biochemistry* **1971**, *10*, 3052–60.

(66) Torii, H.; Tasumi, M. *J. Raman Spectrosc.* **1998**, *29*, 81–86.

(67) Hamm, P.; Woutersen, S. *Bull. Chem. Soc. Jpn.* **2002**, *75*, 985–988.

(68) Hamm, P.; Lim, M.; DeGrado, W. F.; Hochstrasser, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2036–2041.

(69) Torii, H.; Tasumi, M. *J. Phys. Chem. B* **1998**, *102*, 315–321.

(70) Choi, J.-H.; Ham, S.; Cho, M. *J. Chem. Phys.* **2002**, *117*, 6821–6832.

(71) Cha, S.; Ham, S.; Cho, M. *J. Chem. Phys.* **2002**, *117*, 740–750.

(72) Antony, J.; Schmidt, B.; Schütte, C. *J. Chem. Phys.* **2005**, *122*, 014309–014311.

(73) Gorbunov, R. D.; Kosov, D. S.; Stock, G. *J. Chem. Phys.* **2005**, *122*, 224904.

(74) Wang, J.; Hochstrasser, R. M. *Chem. Phys.* **2004**, *297*, 195–219.

(75) CATH database. http://www.cathdb.info/.

(76) Structural Classification of Proteins (SCOP). http://scop.mrc-lmb.cam.ac.uk/.

(77) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536–540.

(78) Barth, A. *Spectrochim. Acta A* **2000**, *56*, 1223–1232.

(79) McDonald, I. K.; Thornton, J. M. *J. Mol. Biol.* **1994**, *238*, 777–793.

(80) HBPLUS homepage. http://www.csb.yale.edu/userguides/datamanip/hbplus/.

(81) Smith, D. K.; Radivojac, P.; Obradovic, Z.; Dunker, A. K.; Zhu, G. *Protein Sci.* **2003**, *12*, 1060–1072.

(82) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379–400.

(83) Shrake, A.; Rupley, J. A. *J. Mol. Biol.* **1973**, *79*, 351–371.

(84) Watson, T. M.; Hirst, J. D. *J. Phys. Chem. A* **2003**, *107*, 6843–6849.

(85) Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L. *J. Phys. Chem. B* **2011**, *115*, 3713–3724.

(86) Maekawa, H.; De Poli, M.; Moretto, A.; Toniolo, C.; Ge, N.-H. *J. Phys. Chem. B* **2009**, *113*, 11775–11786.

(87) Choi, J.-H.; Kim, J.-S.; Cho, M. *J. Chem. Phys.* **2005**, *122*, 174903–174911.

(88) Venyaminov, S. Y.; Kalnin, N. N. *Biopolymers* **1990**, *30*, 1243–1257.

(89) Roy, S.; Lessing, J.; Meisl, G.; Ganim, Z.; Tokmakoff, A.; Roy, S.; Lessing, J.; Meisl, G.; Ganim, Z.; Tokmakoff, A.; Knoester, J.; Jansen, T. L. C. *J. Chem. Phys.* **2011**, *135*, 234507.

(90) Wu, Y.-d.; Zhao, Y.-l. *J. Am. Chem. Soc.* **2001**, *123*, 5313–5319.

Tables and Figures

Table 1: The rationally selected proteins (RaSP) basis set. The upper part (#1–30) lists proteins used for optimization and the bottom part (#31–44) lists proteins for which the spectrum was predicted. The last column lists the final objective value of the protein spectra computed using optimized physical model parameters.

#	Name	PDB ID	Obj. value
1	Insulin	1BPH	0.62
2	Erabutoxin b	3EBX	0.25
3	Monellin	1MOL	0.65
4	Cytochrome c	1HRC	0.33
5	Parvalbumin	1RTP	0.53
6	Metallothionein II	4MT2	0.46
7	Phospholipase A2	1BP2	0.44
8	Apolipoprotein E3	1LPE	0.55
9	Troponin	1TOP	0.78
10	Colicin A, C-terminal domain	1COL	0.31
11	Trypsinogen	2TGA	0.48
12	Carbonic anhydrase	1HCB	0.58
13	Superoxide dismutase (Cu,Zn)	1SXC	0.34
14	Pepsin	4PEP	0.35
15	DD-transpeptidase	3PTE	0.28
16	Superoxide dismutase (Fe)	1ISC	0.36
17	Thaumatin	1THW	0.28
18	Phosphoglyceric kinase	3PGK	2.04
19	Lectin, lentil	1LEN	0.41
20	Dihydropteridine reductase	1DHR	0.88
21	Triose phosphate isomerase	7TIM	0.63
22	Ricin	2AAI	0.50
23	Hemoglobin	1HDA	0.81
24	Glucose oxidase	1GAL	0.53
25	Pepsinogen	2PSG	0.27
26	Alcohol dehydrogenase	2OHX	0.66
27	Ovalbumin (egg albumin)	1OVA	0.33
28	Penicillin amidohydrolase	1PNK	0.49
29	Lipoxygenase-1	2SBL	0.28
30	Citrate synthetase	1CSH	0.41
31	Ubiquitin	1UBI	0.98
32	α-Lactalbumin	1HML	1.09
33	Ribonuclease A	6RAT	0.82
34	Lysozyme	1HEL	1.15
35	Myoglobin	1YMB	1.81
36	Papain	1PPN	0.56
37	α-Chymotrypsinogen A	2CGA	1.18
38	Rennin (chymosin b)	4CMS	0.54
39	Peroxidase	1ARV	0.34
40	Immunoglobulin γ	8FAB	0.57
41	Glutathione S-transferase	2GST	0.85
42	Concanavalin A	1SCS	1.26
43	Avidin	1AVD	0.52
44	α-Hemolysin (alphatoxin)	7AHL	0.68

Table 2: Physical model parameters and their initial values and bounds used in the optimization. The last column shows the final optimized parameters.

Optimized parameter	Initial value	Bounds	Optimized
TDC dipole magnitude (α, β, other) ($\text{DÅ}^{-1}\text{u}^{1/2}$)	3.3, 3.3, 3.3	[2.0, 3.8], [2.0, 3.8], [2.0, 3.8]	2.28, 2.47, 2.20
TDC dipole angle (from $\hat{\boldsymbol{n}}_{\text{OC}}$ toward N) (α, β, other) (deg)	20, 20, 20	[5, 30], [5, 30], [5, 30]	30, 30, 26.6
TDC dipole distance from C (along $\hat{\boldsymbol{n}}_{\text{CO}}$, $\hat{\boldsymbol{n}}_{\text{CN}}$) (Å)	0.8680, 0.0	Not optimized	
Local conf. map, individual points $\delta\omega^{\text{L,before}}$ (25 values) (cm^{-1})	pub. map	$[-30, 20]$	See Figure 4a
Local conf. map, individual points $\delta\omega^{\text{L,after}}$ (25 values) (cm^{-1})	pub. map	$[-30, 20]$	See Figure 4b
Local conf. maps, multiplicative factor	1, 1	[0.1, 10], [0.1, 10]	1.5, 0.9
Local conf. maps, global shift (cm^{-1})	-36	[-46, 4]	-46
Solvent hydrogen bond, maximum (O1, O2, H) (cm^{-1})	$-15, -15, -10$	$[-30, 0]$, $[-30, 0]$, $[-30, 0]$	$-23.1, -1.3, -2.9$
Solvent hydrogen bond, maximum SAS (O1,H) (Å^2)	5, 5	[0.1, 7.0], [0.1, 7.0]	2.4, 7.0
Interpeptide hydrogen bond, coefficients (ξ^{O}, ξ^{H}) (cm^{-1}/kcal)	2.4, 1.0	[1.8, 3.0], [0.7, 1.3]	1.8, 1.0

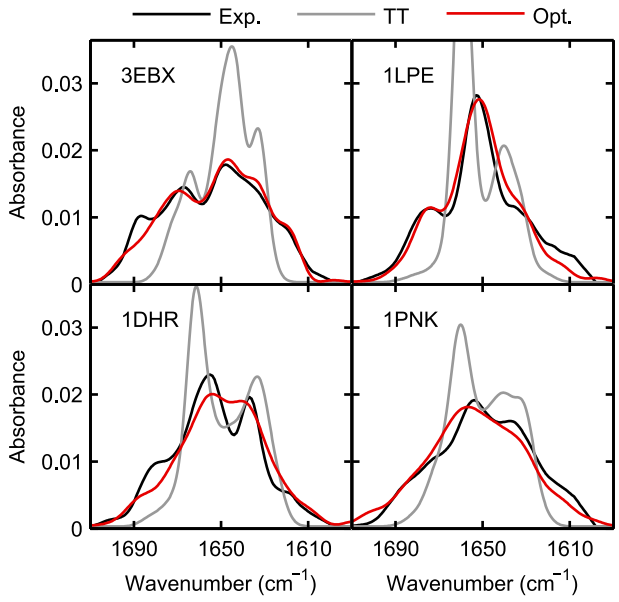

Figure 1: Optimized and experimental spectra of apolipoprotein E3 (1COL), erabutoxin b (3EBX), dihydropteridine reductase (1DHR) and penicillin amidohydrolase (1PNK). Spectra (TT) calculated with the floating oscillator model[11] are also shown.

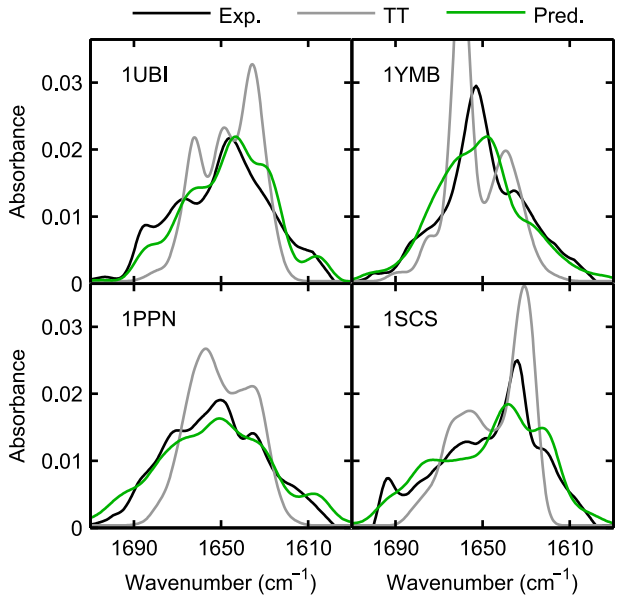

Figure 2: Predicted and experimental spectra of myoglobin (1YMB), concavalin A (1SCS), papain (1PPN) and ubiquitin (1UBI). TDC spectra (TT) using Torii and Tasumi's parameters[11] are also shown.

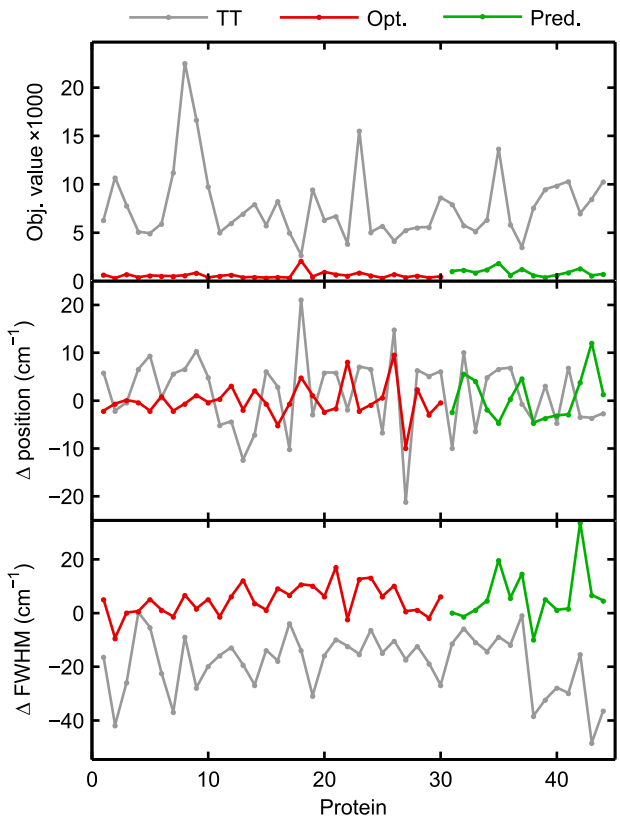

Figure 3: Deviation from experiment for optimized, predicted and TDC spectra (TT). The actual data values are indicated by points and the lines are only intended to guide the eye. The horizontal axis shows the protein index as detailed in Table 1. (Top panel) Sum of the squared deviations from experimental spectra (the objective value). (Middle panel) Deviation of maximum peak position. (Bottom panel) Deviation of amide I band FWHM.

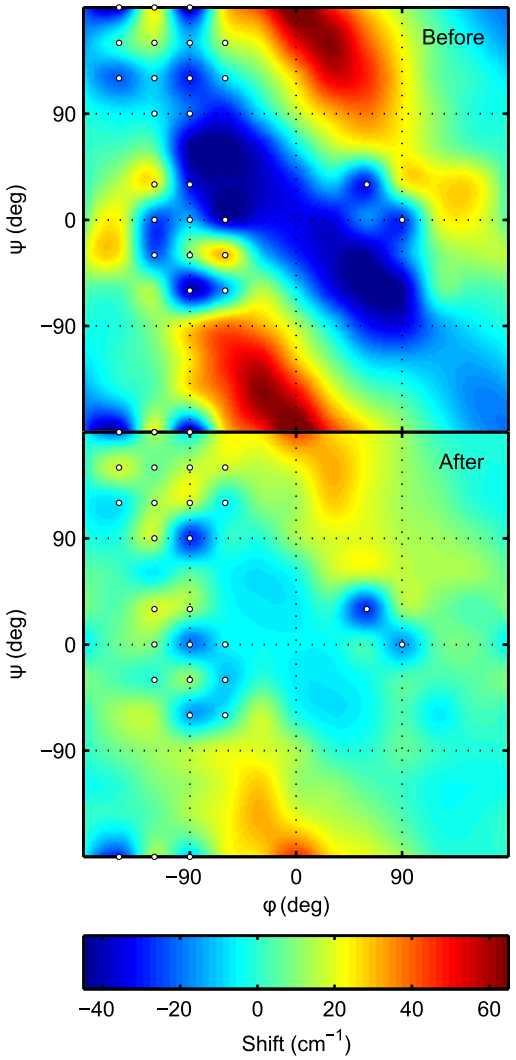

Figure 4: Optimized maps of the shift of the amide I oscillator intrinsic frequency with respect to the local conformation. The white circles indicate the data points subject to individual optimization. The maps do not include the global shift. (Top panel) $\delta\omega_i^{\mathrm{L,before}}(\phi_i,\psi_i)$. (Bottom panel) $\delta\omega_i^{\mathrm{L,after}}(\phi_{i+1},\psi_{i+1})$.

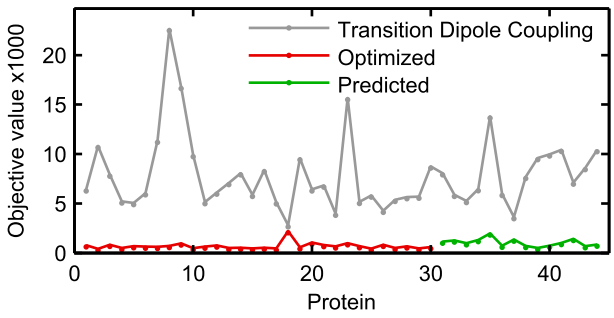

Figure 5: Table of Contents image.