Eur Biophys J (2015) 44:103–112
DOI 10.1007/s00249-014-1005-6

ORIGINAL PAPER

Biophysics in Europe

Calculation of the infrared spectra of proteins

Adam J. Mott · Peter Rez

Received: 2 August 2014 / Revised: 28 October 2014 / Accepted: 5 December 2014 / Published online: 24 December 2014

Abstract The CHARMM22 force field with associated partial charges is used to calculate the infrared spectra of a number of small proteins and some larger biothreat proteins. The calculated high-frequency region, from about 2,500 to 3,500 cm^{-1}, is dominated by stretching modes of hydrogen bonded to other atoms, and is very similar in all proteins. There is a peak at 3,430 cm^{-1} whose intensity is predicted by these calculations to be a direct measure of arginine content. The calculated low-frequency THz region, up to 300 cm^{-1}, is also very similar in all the proteins and just reflects the vibrational density of states in agreement with experimental results. Calculations show that the intermediate-frequency region between 500 and 1,200 cm^{-1} shows the greatest difference between individual proteins and is also the least affected by water absorption. However, to match experimental measurements in the amide region, it was necessary to reduce the hydrogen partial charges.

Keywords Normal modes · Molecular vibrations · Infrared spectroscopy · Terahertz spectroscopy · Proteins · Lysozyme · Myoglobin

Introduction

There is a long history of using infrared (IR) and terahertz (THz) spectroscopy to study proteins. The spectra can be divided into four regions. The high-frequency region from 2,800 to 3,800 cm^{-1} can be attributed to bond stretches involving hydrogen attached to other atoms. The amide region from about 1,300 to 1,800 cm^{-1} has been extensively studied. Recently, there has been much interest in very low frequencies in the THz region up to about 300 cm^{-1}. The less-studied "intermediate region" of the spectrum from 500 to 1,300 cm^{-1} starts from the highest frequencies of the THz range and goes up to just below the amide region.

In their pioneering work, Byler and Susi (1986) empirically showed that spectral features in the amide region from about 1,400 to 1,800 cm^{-1} provide a signature characteristic of ratios of secondary structures. Since then, there have been numerous applications, though strong water absorption in the amide region means that deuterated water has to be used for buffer solutions, leading to questions about deuterium exchange. The THz region has also received a lot of attention (Markelz 2008; Plusquellic et al. 2007), since these lower-frequency vibrations could relate to biologically significant motions in large macromolecules. There have been studies of both dried and hydrated proteins (He et al. 2011; Knab et al. 2006; Markelz et al. 2000; Vinh et al. 2011).

Theoretical modeling of IR and THz spectra has been based on normal mode analysis (Zhang et al. 2009) or Fourier analysis of the time autocorrelation of the net dipole moment (Gaigeot et al. 2005). The first step in determining normal modes is to calculate the Hessian matrix, the second derivative of the potential energy with respect to atomic displacements. The number of independent displacements is 3N, where N is the number of atoms, though six are zero-frequency translations or rotations of the complete molecule. The mass-weighted Hessian matrix then has to be diagonalized to determine the square of the eigenfrequencies and the eigenvectors. In the original work introducing normal mode analysis of proteins, Brooks and Karplus (1983) limited themselves to bovine pancreatic trypsin

A. J. Mott · P. Rez (✉)
Department of Physics, Arizona State University, Tempe,
AZ 85287, USA
e-mail: Peter.Rez@asu.edu

inhibitor (BPTI), a small protein with only 580 atoms, since they only included the polar hydrogen atoms. The Hessian was calculated from the CHARMM force fields, and Brooks and Karplus's normal mode tools became part of the CHARMM package (Brooks et al. 2009). Later developments by Tama and Sanejouand (2001) stressed the biologically significant low-frequency modes. The Lanczos (1950) method was used to extract just those low-frequency modes of potential interest, and no attempt was made to calculate the complete eigenspectrum. Various coarse-graining schemes, such as grouping large parts of the molecule as a block (Tama et al. 2000) or reducing all interactions to springs connecting C_α atoms (Bahar et al. 1997; Tama 2003; Tirion 1996), have been used in the application of low-frequency normal mode analysis to very large systems, such as the complete ribosome (Tama et al. 2003) and the cowpea chlorotic mottle virus (CCMV) (Tama and Brooks 2002).

Specialized techniques have been used for the well-studied amide region that can loosely be divided into three main bands. The amide I band between 1,600 and 1,700 cm^{-1} has been attributed to stretching of the C=O bond in the peptide link, while the amide II (1,480–1,530 cm^{-1}) and amide III (1,210–1,350 cm^{-1}) bands arise from bending and torsional motions about the peptide bond. Changes in the hydrogen-bonding network in the protein can change both frequencies and relative intensities, though it should be remembered that each of these bands has contributions from many eigenfrequencies. Early theoretical work was based on the transition dipole coupling (TDC) scheme where each peptide bond was represented by a dipole oscillator with coupling between them treated in perturbation theory. This method was applied to the calculation of the amide region spectra of a number of globular proteins (Torii and Tasumi 1992). In another approach pioneered by Bour et al. (1997), effective force fields were built up from quantum-chemistry calculations on small oligopeptides, such as *N*-methylacetamide, that were chosen to represent different possible environments in the polypeptide chain. These were pieced together in a procedure named Cartesian coordinate transfer to make up a complete protein. Choi and Cho (2009) and colleagues used quantum-chemistry codes to generate the Hessian for small molecules. They then selected those parts of the eigenvector for a given amide frequency that corresponded to atoms that were part of a particular peptide bond, and used this to construct a reduced Hessian. Again, large proteins were modeled by combining reduced Hessians from different parts of the polypeptide chain that best matched structures in small peptides. They also took account of solvation effects by introducing a frequency shift dependent on local partial charges.

Our focus is on the full range of frequencies from THz to 4,000 cm^{-1}. We used the CHARMM22 force fields to calculate the Hessian matrix for a range of small proteins and biothreat proteins. The eigenfrequencies and eigenvalues were obtained by matrix diagonalization, and an IR spectrum was calculated by convolution with a Lorentzian function. In some cases, direct integration of the harmonic equation of motion (Mott et al. 2013) was used, which is more efficient when there are more than 4,000 atoms.

The high-frequency region from about 2,800 to 3,800 cm^{-1} is dominated by stretching modes along C–H, N–H, and O–H bonds. Given that all proteins have similar proportions of C, N, O, and H, it is not surprising that the features in this region are the same for all the proteins we studied. There is a distinct peak from an antisymmetric stretch in the NH_2 group of arginine whose height is directly proportional to the number of arginine residues present.

Studies on small molecules (Mott 2012) have shown that the amide II and amide III peaks are particularly sensitive to the partial charge of hydrogen. Good agreement with experiment for the amide region between 1,300 and 1,800 cm^{-1} could be achieved by reducing the hydrogen charges from the CHARMM values by a factor of 9. The uncertainty in effective hydrogen partial charge does not affect the high-frequency region, though it does change the relative heights of peaks in the high-frequency region with respect to the low-frequency and amide regions of the spectrum.

Theory

When atoms in a molecule are displaced, they are subject to forces from neighboring atoms. Equations of motion can be written whose solutions give the vibrational modes of the molecule. The forces due to displacements of atoms are best expressed in terms of the Hessian matrix, the second derivative of the potential energy U with respect to atomic displacements.

$$H_{K,L} = \left(\frac{\partial^2 U}{\partial x_{i,k} \partial x_{j,l}} \right)_0, \tag{1}$$

in which the indices k and l refer to the atom (1, 2,…, N), and the indices i and j refer to the Cartesian direction ($1 = x$, $2 = y$, $3 = z$). The indices K and L are labels combining atom and Cartesian direction such that $K = 3(k - 1) + i$ and $L = 3(l - 1) + j$.

In terms of the Hessian matrix, the equations of motion are

$$m_k \frac{\mathrm{d}^2 \Delta x_{i,k}}{\mathrm{d}t^2} = -\sum_{l=1}^{N} \sum_{j=1}^{3} H_{K,L} \Delta x_{j,l}, \tag{2}$$

Table 1 Small proteins and biothreat proteins investigated

Protein	PDB ID	Atoms	Chain studied
Lysozyme	6LYZ	1,960	A (1 of 1)
BPTI	1BTI	882	A (1 of 1)
Myoglobin	1YMB	2,411	A (1 of 1)
Ebola GP2	2EBO	1,203	A (1 of 3)
Dengue type 3	1UZG	6,050	A (1 of 2)
Dengue type 2	1OAN	6,129	A (1 of 2)
Dengue envelope	1K4R	6,012	A (1 of 3)
Vaccinia L1	1YPY	2,698	A (1 of 2)
Vaccinia N1L	2I39	1,943	A (1 of 6)
Anthrax protective antigen	1T6B	11 352	X (1 of 2)

where m_k is the mass of the kth atom, $\Delta x_{i,k}$ is the ith component of the displacement from equilibrium of the kth atom, and N is the number of atoms in the molecule. To calculate the frequencies of vibration, one assumes harmonic motion and the equation becomes

$$m_k \omega^2 \Delta x_{i,k} = \sum_{l=1}^{N} \sum_{j=1}^{3} H_{K,L} \Delta x_{j,l}, \tag{3}$$

which can be rewritten as a matrix eigenvalue equation,

$$|\omega^2 - m_k^{-1/2} H_{K,L} m_l^{-1/2}| = 0. \tag{4}$$

In addition to the interatomic forces described in the right-hand side of Eq. 2, other forces on the atoms include hydrodynamic drag forces and forces due to the time-varying electric field of the incident electromagnetic radiation acting on charged atoms. With these additional forces, the new equations of motion are

$$m_k \frac{\mathrm{d}^2 \Delta x_{i,k}}{\mathrm{d}t^2} = -\sum_{l=1}^{N} \sum_{j=1}^{3} H_{K,L} \Delta x_{j,l} - \eta_k m_k \frac{\mathrm{d}\Delta x_{i,k}}{\mathrm{d}t} + q_k E_i \mathrm{e}^{-\mathrm{i}\omega t}, \tag{5}$$

where E_i is the ith Cartesian component of the electric field and η_k is a damping coefficient. By analogy with the single-oscillator case, a formal solution can be found for the displacements at a given frequency ω. An expression for the dipole moment and hence the polarizability, which is now a tensor, can then be derived as

$$\alpha(\omega) = \frac{1}{V\varepsilon_0} [\mathbf{q}(\mathbf{H} - \mathbf{m}\omega^2 - \mathrm{i}\mathbf{m}\eta\omega)^{-1}\mathbf{q}]. \tag{6}$$

Here, bold symbols represent matrices or vectors and V is the volume of the molecule. The polarizability tensor can be calculated from the normal modes of the molecule obtained by diagonalizing the Hessian matrix:

$$\alpha_{i,j}(\omega) = \frac{1}{V\varepsilon_0} \sum_{\text{mode } n=7}^{3N} \sum_{\text{atoms } k,l=1}^{N} \frac{q_k C_{i,k;n} C_{j,l;n} q_l}{m_k^{1/2}(\omega_n^2 - \omega^2 - \mathrm{i}\gamma_n \omega) m_l^{1/2}}, \tag{7}$$

where m_k are the masses, ω_n and γ_n are the real and imaginary parts of the eigenfrequencies of the Hessian matrix, and $C_{i,k;n}$ are elements of the eigenvector matrix. The labels i,j refer to directions, k, l to atoms, and n to eigenfrequencies. The charges used to calculate the polarizability tensor in Eq. 7 need not be the same as those used in the force field that was used to calculate the Hessian matrix or the molecular dynamics trajectories.

The frequency-dependent dielectric function can be derived from the trace of the polarizability tensor as

$$\varepsilon(\omega) = \varepsilon_0 \left(1 + N_\mathrm{m} \frac{\sum_i \alpha_{ii}}{3} \right), \tag{8}$$

where N_m is the number of molecules. The attenuation is given by

$$\mu(\omega) = \omega \mathrm{Im}(\varepsilon(\omega)). \tag{9}$$

After inserting Eq. 7, the attenuation becomes

$$\mu(\omega) = \rho \frac{1}{3} \sum_{i=1,3} \sum_{k,l,n} \frac{q_k C_{i,k;n} \gamma_n \omega^2 C_{i,l;n} q_l}{m_k^{\frac{1}{2}} \left((\omega_n^2 - \omega^2)^2 + (\gamma_n \omega)^2 \right) m_l^{\frac{1}{2}}}, \tag{10}$$

where ρ is the density of molecules.

Each eigenfrequency is convoluted by a Lorentzian whose full-width at half-maximum is the damping constant γ_n.

$$\mu(\omega) = \frac{\rho}{3} \sum_{k,l,n} \frac{q_k C_{i,k;n} \gamma_n C_{i,l;n} q_l}{m_k^{\frac{1}{2}} (4\delta^2 + \gamma_n^2) m_l^{\frac{1}{2}}}, \tag{11}$$

when the damping constant γ_n is small compared with the eigenfrequency ω_n and

$$\omega = \omega_n + \delta. \tag{12}$$

For frequencies in the THz range that are comparable to the damping constant, the motion becomes critically damped ($\gamma_n = \omega_n$) or overdamped ($\gamma_n > \omega_n$). This transition has been extensively studied by Kitao et al. (1991) and Hayward et al. (1994).

Computational methods

We obtained structures for the ten proteins in Table 1 from the Protein Data Bank (PDB). These structures were input into the CHARMM program. Coordinates for each protein's hydrogen atoms were determined from the energy-minimization procedure, since PDB files do not specify coordinates for hydrogen atoms. Since the damping is due to internal friction, except in the THz region (Hinsen and Kneller 2008), the proteins were not solvated; no explicit or implicit solvation model was used. Prior to calculation of the Hessian matrix and normal modes, the structure needed to be optimized to minimize the potential energy. To prevent too much distortion of the structure during minimization, harmonic constraints of the form $K_i\left|\vec{r}_i - \vec{r}_{i,\mathrm{orig}}\right|^2$ were temporarily added to the potential energy for each atom i except hydrogens, which had the effect of attracting each nonhydrogen atom to its original position $\vec{r}_{i,\mathrm{orig}}$. The strengths of the harmonic constraints K_i were taken to be proportional to each atom's mass. At first, the minimization was done with stiff constraints, $K_i/m_i = 10^4$ kcal mol^{-1} Å^{-2} (with m_i in atomic masses), which hardly allowed for any movement of the atoms except for the hydrogens. In successive iterations, the strength of the constraints was lowered by a factor of ten to 10^3, 10^2, …, 10^{-4} kcal mol^{-1} Å^{-2}. For each of the powers of 10, minimization was done with 2,000 steps of the steepest-descent method followed by 5,000 steps of the adopted basis Newton–Raphson method. After this, the harmonic restraints were completely removed, and an additional 5,000 steepest-descent steps and 10,000 adopted basis Newton–Raphson steps were taken to arrive at the final minimized structure. This rigorous minimization procedure ensured that the atomic coordinates were such that the total potential energy was at a minimum prior to normal mode analysis, so that no normal modes would be found to have nonsensical negative eigenvalues (imaginary frequencies).

We used CHARMM to calculate each protein's mass-weighted Hessian matrix. We then diagonalized this matrix using our own Fortran code with the LAPACK routine DSYEV, since we found this to be significantly faster than CHARMM's built-in diagonalization routine when dealing with the larger proteins. Lastly, we used our own Fortran code to calculate the protein's infrared absorption spectrum based on Eq. 11.

Results and discussion

Vibrational spectra were calculated for the small proteins and biothreat or biohazard proteins shown in Table 1. The small proteins have been extensively studied previously, and spectra for BPTI have been published by Goossens et al. (1996); spectra for lysozyme and myoglobin have been published by Ruegg et al. (1975), Dong et al. (1990), and Torii and Tasumi (1992).

The damping parameter most strongly affects the spectra in the THz region. We took the damping parameter to be 20 cm^{-1}, derived from experimental GHz optical measurements on hydrated lysozyme by Vinh et al. (2011). Recent THz measurements on crystallized hen egg white lysozyme (Acbas et al. 2014) estimated the damping parameter to be as low as 5 cm^{-1}, though it is conceivable that this is only appropriate for crystals, not proteins in solution. Clearly, more spectral detail would become apparent if this lower value were used in the calculations.

Although the well-studied amide region is not our focus, there are many published spectra for proteins in this region. Comparisons of our calculations using the CHARMM partial charges with experimental spectra are shown as Fig. 1a for lysozyme and Fig. 1c for myoglobin. The experimental IR spectrum for myoglobin in a solution of 10 mM potassium phosphate/H_2O was obtained from the Protein Infrared Database (Dong et al. 1995; Dong and Caughey 1994). A common feature in our calculated spectra is that the amide II peak is too intense compared with the amide I peak. Hydrogen atoms have large displacements, not just for stretching modes due to the lower mass, but also for bond bending and torsional rotations of bonds between atoms with attached hydrogen atoms. Displacements of the amide hydrogen atom make a significant contribution to the amide II peak. Calculations for a selection of small molecules show much improved agreement with experiment when the hydrogen partial charges used in the calculation of polarizability given by Eq. 7 are reduced (Mott 2012). The charges assigned to the H atoms of lysozyme range from $+0.05e$ to $+0.46e$, with the most common charge being $+0.09e$. In reducing the H charges, it is preferable to scale them all down by the same factor rather than assigning an equal charge to all H atoms, since the latter would ignore the different charges given to different H atoms based on their location in the molecule. The IR intensities using H charges that were 1/9 times their original values are shown as Fig. 1b for lysozyme and Fig. 1d for myoglobin. Not only is there much improved agreement with the relative heights of the amide I and II peaks, but there is also improved agreement with the portion of the spectrum below ~1,525 cm^{-1}. For this reason, we used these reduced partial charges in subsequent calculations of the IR spectra.

The choice of the scaling factor of 9 by which to reduce the hydrogen partial charges was somewhat arbitrary and was motivated by the desire to reduce the most common hydrogen partial charge from a value of $+0.09e$ down to $+0.01e$. Our aim was not to identify the ideal value for the scale factor, but rather to demonstrate that scaling down the hydrogen charges in the calculation of the IR intensities of the normal modes results in improved agreement between the calculated and experimental spectra. In principle, one could obtain a best-fit value for the scale factor by minimizing the differences between calculated and experimental spectra.

The calculated amide I and II peaks are also shifted 40 cm^{-1} higher in frequency than the experimental measurement. In this spectral region, closer agreement with

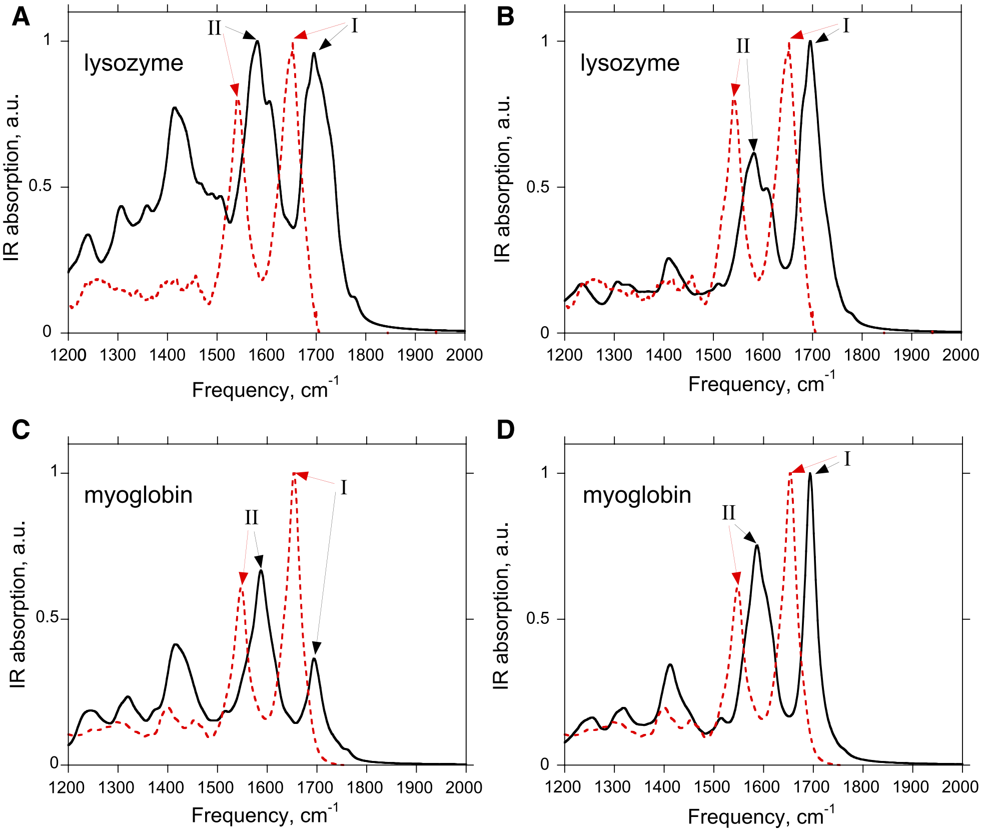


Fig. 1 Comparison of calculated amide peaks (*solid lines*) in lysozyme (**a**, **b**) and myoglobin (**c**, **d**) with experimental measurements (*dashed lines*). Myoglobin measurement obtained from the Protein Infrared Database (Dong et al. 1995). The calculated IR spectra in (**a**) and (**c**) made use of the CHARMM22 partial charges; in (**b**) and (**d**), these spectra were recalculated with hydrogen partial charges reduced by a factor of 9

observed frequencies in the works of Choi et al. (2007) and Grahnen et al. (2010) can be attributed to their use of effective force fields that are matched to the amide region. It would not be expected that the CHARMM22 force field designed for general application to a wide variety of protein properties would give as good agreement for a particular spectral region.

The curve-fitting method of Byler and Susi (1986) assumes that the amide I (or I′) band can be decomposed as a sum of six to nine Gaussians whose center frequencies are the characteristic vibration frequencies of the C=O modes associated with different secondary structures. While this simple picture has been successfully used for prediction of secondary-structure content, there are far more than nine normal mode frequencies within the amide I band. If one considers the calculated amide I band for lysozyme (Fig. 1a, b) to go from 1,670 to 1,740 cm^{-1} (approximately the interval corresponding to the band's full-width at half-maximum), then there are 174 normal modes in this band. Similarly, if one counts modes in the amide II band for lysozyme in the range of 1,540–1,630 cm^{-1}, there are 171 normal modes.

Examination of animations of the most IR-active normal modes near the centers of the amide I and II bands shows that these normal modes often involve vibrations of side chains in addition to the expected amide I and II vibrations in the protein backbone. For example, the animation of a mode near the center of the amide I band of lysozyme showed the expected C=O stretching vibration in a leucine residue, but also included in this mode was a vibration involving stretching and bending of several C–N bonds in an arginine residue.

The low-frequency THz region shows very little structure up to a frequency of 300 cm^{-1}, in agreement with experimental measurements on myoglobin by Zhang et al. (2004) and lysozyme by Knab et al. (2006). With the exception of BPTI, the intensity varies as $\omega^{0.5}$ up to about 100 cm^{-1} as would be expected from the density of vibrational states as shown in Fig. 2a for the small proteins and Fig. 2b for the biothreat proteins. Fine structures unique to a given protein are seen above 300 cm^{-1} (almost 10 THz). One could say that these structures are more characteristic of the intermediate-frequency region discussed below.

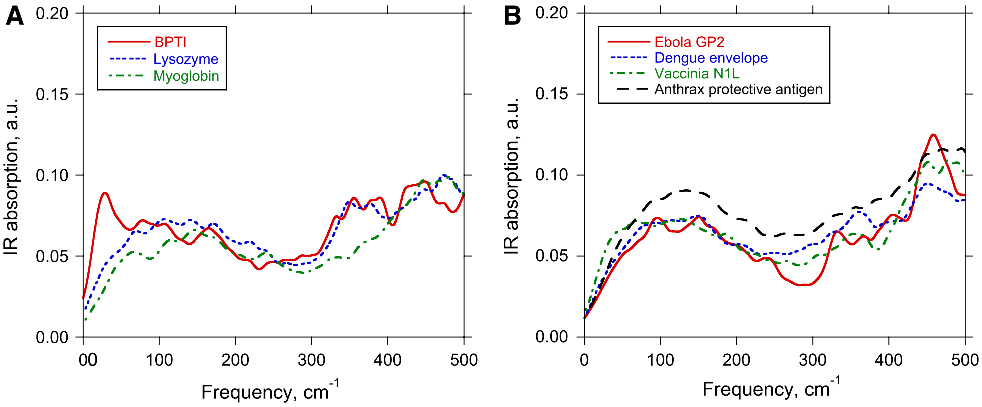


Fig. 2 Calculated spectra for the THz region for **a** small proteins and **b** biothreat proteins

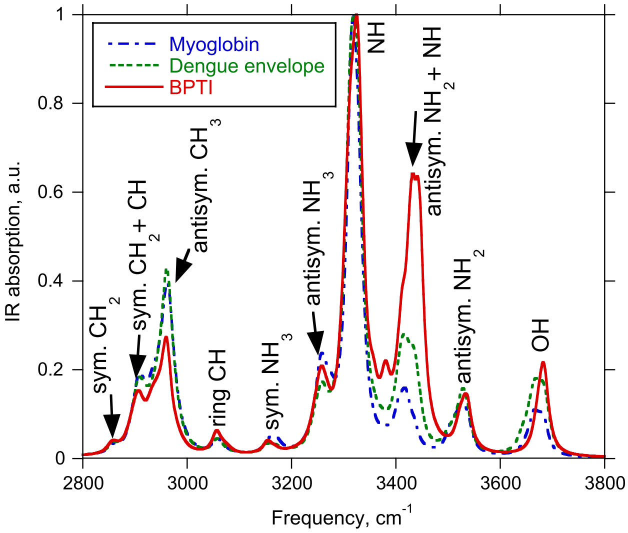


Fig. 3 High-frequency region (2,800–3,800 cm^{-1}) of the calculated IR spectra of three proteins. The protein with the lowest abundance of arginine residues (myoglobin) is in *blue*, while the one with the highest arginine abundance (BPTI) is in *red*

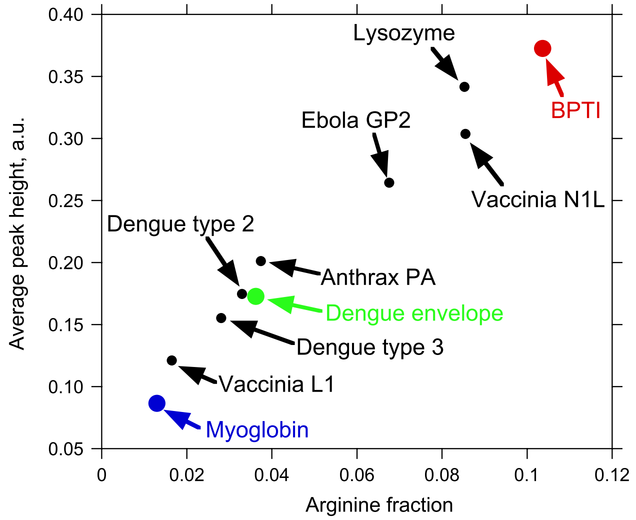


Fig. 4 Height of the peak at ~3,430 cm^{-1} (Fig. 3) averaged over 3,390–3,475 cm^{-1} versus the protein's arginine abundance (i.e., the ratio of the number of arginine residues to the total number of residues) for each of the 10 proteins in Table 1

When the calculated IR spectra of all ten proteins from Table 1 are compared in the high-frequency region (Fig. 3), they are found to be quite similar. The relative intensities of the various peaks are comparable from one protein to the next, which is an indication that the relative numbers of the different chemical groups (CH, CH_2, CH_3, NH, NH_2, NH_3, and OH) giving rise to these peaks are similar from one protein to the next. The exception to this rule is the peak at ~3,430 cm^{-1}, whose predicted intensity varies drastically among the proteins according to our calculations. From examining the corresponding normal modes in detail, we find that this peak arises from asymmetric stretching in arginine's two NH_2 groups coupled with stretching of the adjacent N–H bond in the arginine residue chain. Typically, the motion is localized in a single Arg residue for a given mode, with different modes characterizing vibrations in different Arg residues. The average intensity of the peak at ~3,430 cm^{-1} was calculated over the range of 3,390–3,475 cm^{-1}. The intensity of this peak depends on the relative abundance of Arg in the protein sequence, as demonstrated by Fig. 4. The protein with the highest Arg abundance, bovine pancreatic trypsin inhibitor (6 out of 58 residues are Arg), also had the highest average intensity of the peak at ~3,430 cm^{-1}. The protein with the lowest Arg abundance, myoglobin (2 out of 153 residues are Arg), had the lowest average peak intensity. We are unaware of any experimental work in the literature that has related the intensity of this peak to the amount of arginine in a protein; this is a result that emerges from our computations and remains to be tested experimentally.

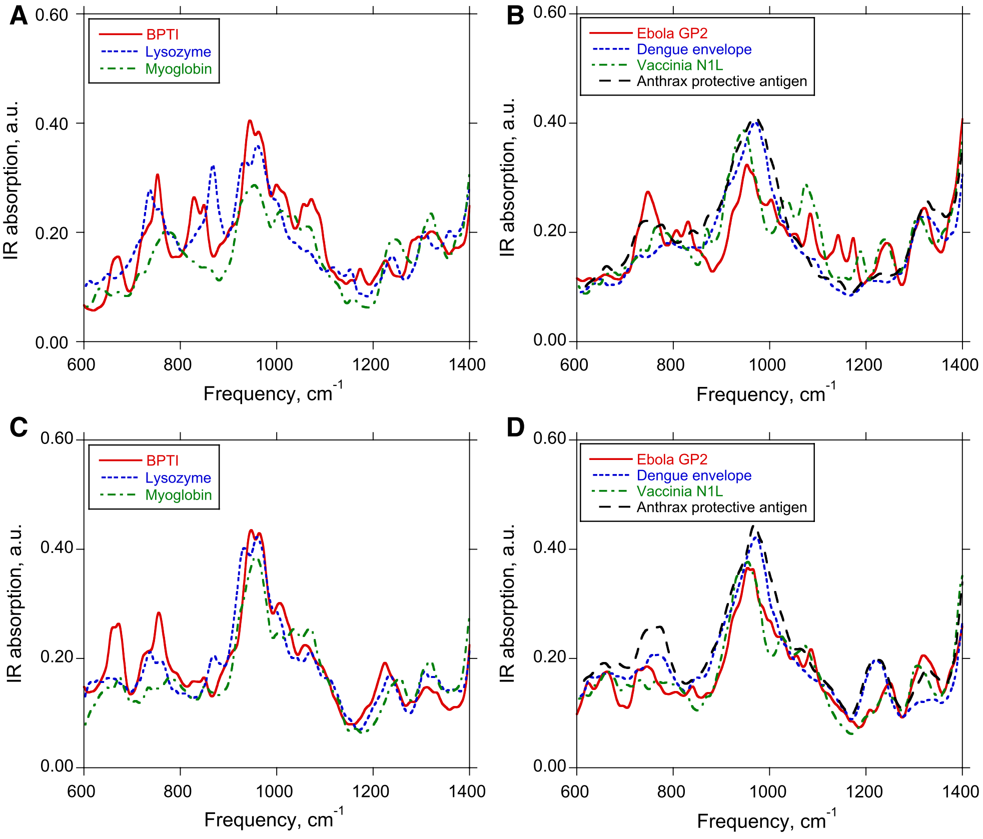


Fig. 5 Calculated spectra for the intermediate-frequency region for small proteins (**a**, **c**) and biothreat proteins (**b**, **d**). The spectra in the *top row* (**a**, **b**) use the CHARMM22 partial charges. The spectra in the *bottom row* (**c**, **d**) were calculated using the reduced hydrogen partial charges

In general, the peaks attributed to stretching of N–H bonds are stronger than those from C–H stretches. This is a consequence of the stronger charge polarization in the N–H bonds as expressed by the CHARMM partial charges. The peak at 2,908 cm^{-1}, due to symmetric stretching of CH_2 groups combined with stretching of an adjacent C–H bond, has many more normal modes than the amide A peak at 3,326 cm^{-1}, due to N–H bond stretching in the peptide bond of the protein backbone. Since these peaks all arise from stretching of bonds involving hydrogen and a heavier atom, the relative heights of peaks in this region are unaffected by the value of the hydrogen partial charge used in Eq. 7; all that changes is the relative magnitude of this part of the spectrum compared with the features at lower frequency.

The final region, between approximately 500 and 1,300 cm^{-1}, has many features that are distinctive of the particular protein. These features are most prominent with the CHARMM partial charges as shown in Fig. 5a for the small proteins and Fig. 5b for the biothreat proteins. Although there are broad peaks at approximately 780 and 1,000 cm^{-1} common to all three proteins, there are many fine structures that give an individual signature, which is why this region has been called the fingerprint region by Khajehpour et al. (2006).

Since many of the peaks arise from modes involving displacements of H atoms, the differences between spectra calculated with the reduced hydrogen partial charges are much smaller, as shown in Fig. 5c for the smaller proteins and Fig. 5d for the biothreat proteins. We quantified differences between pairs of spectra using

$$\delta = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - g_i)^2}, \tag{13}$$

where f_i and g_i are the IR intensities of two spectra $f(\omega)$ and $g(\omega)$ at $N = 801$ discrete frequency values 600, 601,…, 1,400 cm^{-1}. Prior to making this comparison, each spectrum was normalized to have a maximum value of 1 over this frequency range. The pair of small proteins whose spectra differed the most in Fig. 5a, c were lysozyme and myoglobin, which differed by $\delta = 0.142$ when the standard CHARMM charges were used to calculate the spectrum in Fig. 5a. This difference was reduced to $\delta = 0.081$ (43 % reduction) when the spectra were recalculated assuming reduced hydrogen charges (Fig. 5c). Among the biothreat proteins' spectra (Fig. 5b, d), the two that differed the most were Ebola GP2 and anthrax protective antigen; their δ was reduced by 39 % from 0.132 (Fig. 5b) to 0.081 (Fig. 5d) after switching to the reduced hydrogen charges. Comparing the spectra from all seven proteins in Fig. 5 with each other (21 pair comparisons), we found that the reduction of hydrogen charges in the spectrum calculation has the effect of reducing the difference δ between spectra by an average of 28 %, with a minimum decrease of 10 % and a maximum decrease of 45 %.

It is still the case that the intermediate or fingerprint region shows some promise for distinguishing one biothreat protein from another, though in these larger molecules the peak at 900 cm^{-1}, probably due to rocking of the CH_2 group in the side chains, is more prominent. Only the spectrum from the anthrax membrane protein appears to lack strong distinct fine structures.

Proteins with similar arrangements of secondary structures would therefore be expected to have similar spectra in this region. The three dengue proteins that were studied—dengue envelope, type 2, and type 3—are very similar in structure, and the intermediate-region spectra from these three proteins are almost identical, as shown in Fig. 6a, even when using the CHARMM partial charges. The two vaccinia proteins studied, N1L and N1, have very different tertiary structures, and this is reflected in the distinctive spectra in the intermediate-frequency region when the CHARMM partial charges are used, as shown in Fig. 6b. The differences are still present though much less apparent with the reduced hydrogen partial charges, as shown in Fig. 6c.

So far we have only considered spectra from a single chain in the case of dimers and trimers. To investigate how interactions between chains would change the spectrum in the intermediate region, the calculated spectrum for a single chain of Ebola GP2 (1,203 atoms) is compared with the result for all three chains (3,609 atoms) in Fig. 7. In Fig. 7a, we have used the CHARMM partial charges to emphasize possible differences. The spectra of the monomer and the trimer are very similar; noticeable differences are the peaks at 750 and 1,140 cm^{-1} in the monomer spectrum that are suppressed in the trimer spectrum. Also, the peak at 820 cm^{-1} in the trimer spectrum is enhanced compared with the monomer spectrum. Figure 7b shows that these differences largely disappear when the reduced hydrogen charges are used. The root-mean-square (RMS) difference δ (Eq. 13) between the trimer and monomer spectra is 0.120 in Fig. 7a, and 0.047 in Fig. 7b. (As before, these RMS differences were calculated over the frequency range of 600–1,400 cm^{-1} with each spectrum normalized to have a maximum value of 1 in that frequency range.) The overall similarity of the monomer and trimer spectra demonstrates that signatures in the fingerprint region arise mainly from interactions within a given chain rather than between chains.

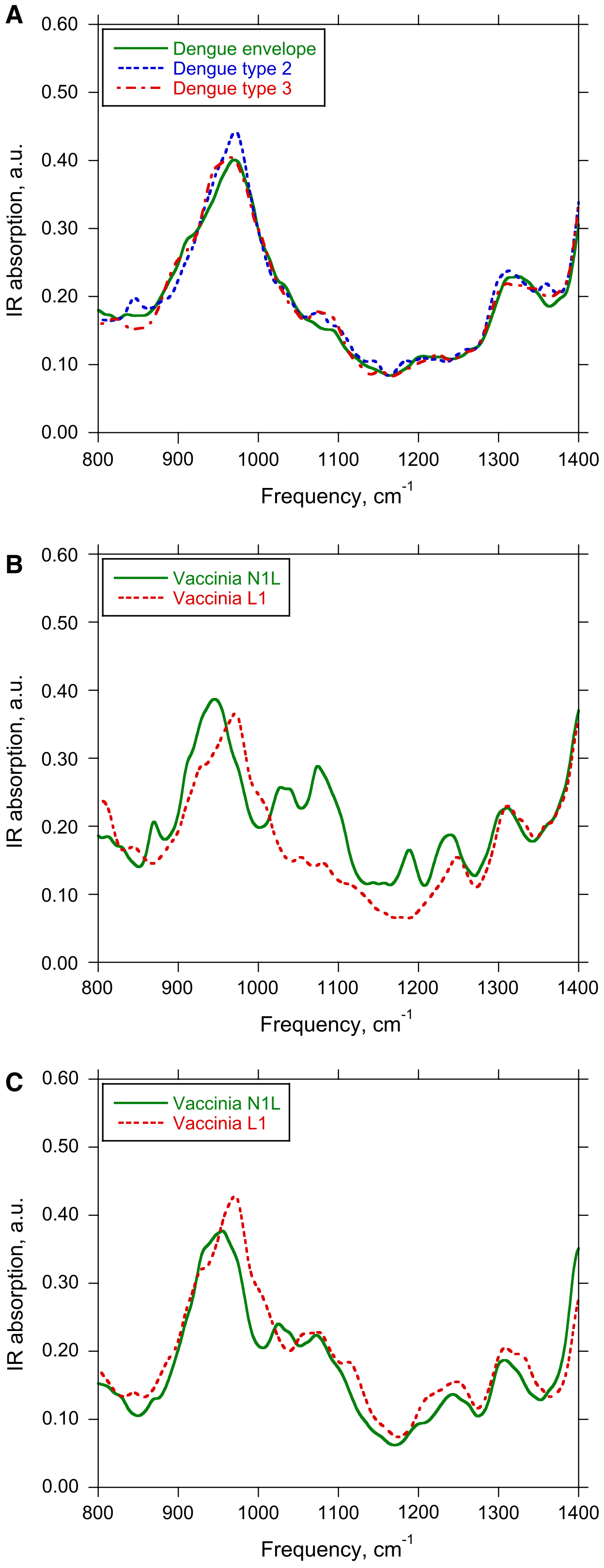
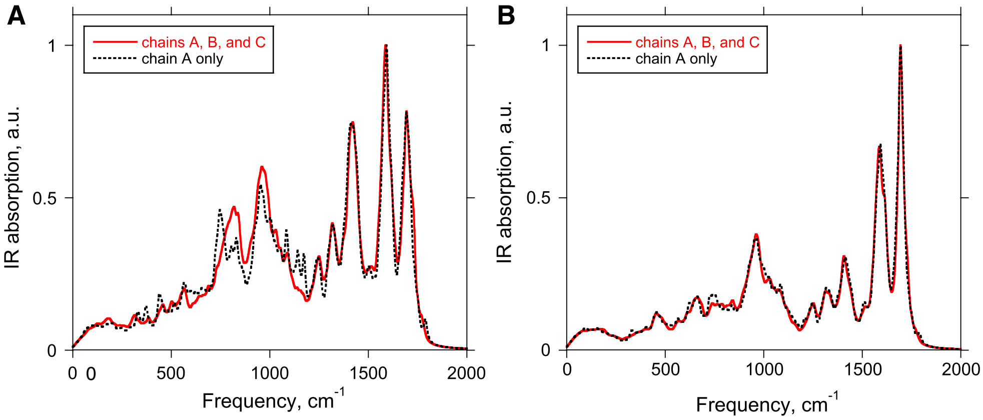


Fig. 6 Intermediate-frequency region for dengue envelope, type 2, and type 3 proteins using CHARMM partial charges (**a**), and vaccinia L1 and N1L proteins using CHARMM partial charges (**b**) and reduced hydrogen charges (**c**)

Fig. 7 Comparison of the calculated IR spectrum of the A chain (1,203 atoms) of envelope glycoprotein GP2 from Ebola virus versus that of the complete trimer consisting of chains A, B, and C (3,609 atoms). Spectra in the left graph (**a**) use the CHARMM22 partial charges; spectra in the right graph (**b**) use reduced hydrogen charges



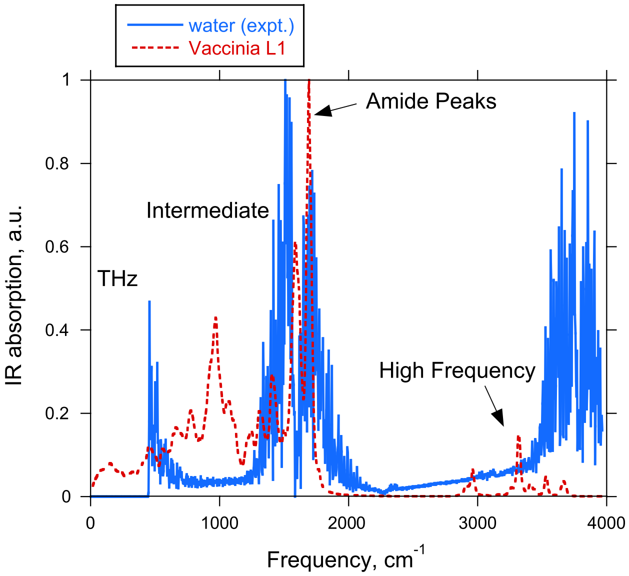


Fig. 8 Calculated vaccinia L1 IR spectrum compared with experimental water IR absorption from Stein (1992). The cutoff in the THz region is due to experimental limitations; there is substantial water absorption in the THz region

Finally, the effects of water absorption have to be considered. The main features in the water IR spectrum are the broadened symmetric and asymmetric stretch peaks between 3,600 and 3,800 cm^{-1}, the broadened peak due to bending at 1,600 cm^{-1}, and the peak in the THz region due to rotations. By superimposing an experimental water absorption spectrum from the NIST/EPA Gas Phase Infrared Library (Stein 1992) on our calculated spectrum from vaccinia L1, it can clearly be seen that the intermediate-frequency region is the least affected by water absorption, as shown in Fig. 8. The peak at 3,430 cm^{-1}, sensitive to the presence of arginine, is at the edge of the strong water absorption band.

In practice, water might still be a serious problem in any measurement that relies on absorption of IR passing through a cell, since the effects of small differences in path length may be confused with absorption in the protein.

Conclusions

Infrared absorption can be calculated from normal modes of a Hessian based on the CHARMM22 force field and partial charges. The spectral intensities are sensitive to atomic partial charges, especially the hydrogen partial charge, since the light hydrogen atoms have the greatest displacements. There is good agreement with experimental measurements for the well-studied amide peak region when hydrogen partial charges reduced by a factor of 9 are used for the calculation of spectral intensities. The vibrational spectrum can be separated into four regions: the THz region up to about 500 cm^{-1}, the intermediate region from about 500 to 1,300 cm^{-1}, the amide-peak region from 1,300 to 1,800 cm^{-1}, and the high-frequency stretch region from 2,800 to 3,800 cm^{-1}. Apart from the peak at 3,430 cm^{-1}, which is related to an antisymmetric stretch in the arginine NH_2 group, all other peaks have similar magnitudes across a range of proteins. The intermediate-frequency region shows some differences between spectra from individual proteins. These differences are greatly suppressed if the hydrogen partial charges are reduced, as was necessary to obtain agreement with experiment in the amide region.

Acknowledgments We thank the Defense Threat Reduction Agency Grant HDTRA-08-0040 for financial support. We thank Saravana Prakash Thirumuruganandham for discussions and Stephanie Cope for her experimental measurement of the IR spectrum of lysozyme (Fig. 1).

References

Acbas G, Niessen KA, Snell EH, Markelz AG (2014) Optical measurements of long-range protein vibrations. Nat Commun 5:3076

Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 2:173–181

Bour P, Sopkova J, Bednarova L, Malon P, Keiderling TA (1997) Transfer of molecular property tensors in Cartesian coordinates: a new algorithm for simulation of vibrational spectra. J Comput Chem 18:646–659

Brooks B, Karplus M (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 80:6571–6575

Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30:1545–1614

Byler DM, Susi H (1986) Examination of the secondary structure of proteins by deconvolved FTIR spectra. Biopolymers 25:469–487

Choi J-H, Cho M (2009) Computational linear and nonlinear IR spectroscopy of amide I vibrations in proteins. In: Barth A, Haris PI (eds) Biological and biomedical infrared spectroscopy. IOS Press, Amsterdam, pp 224–260

Choi JH, Lee H, Lee KK, Hahn S, Cho M (2007) Computational spectroscopy of ubiquitin: comparison between theory and experiments. J Chem Phys 126:045102

Dong AC, Caughey WS (1994) Infrared methods for study of hemoglobin reactions and structures. Method Enzymol 232:139–175

Dong A, Huang P, Caughey WS (1990) Protein secondary structures in water from second-derivative amide I infrared spectra. Biochemistry 29:3303–3308

Dong A, Carpenter JF, Caughey WS (1995) Protein Infrared Database. http://www.unco.edu/nhs/chemistry/faculty/dong/irdata.htm

Gaigeot MP, Vuilleumier R, Sprik M, Borgis D (2005) Infrared spectroscopy of *N*-methylacetamide revisited by ab initio molecular dynamics simulations. J Chem Theory Comput 1:772–789

Goossens K, Smeller L, Frank J, Heremans K (1996) Pressure-tuning the conformation of bovine pancreatic trypsin inhibitor studied by Fourier-transform infrared spectroscopy. Eur J Biochem 236:254–262

Grahnen JA, Amunson KE, Kubelka J (2010) DFT-based simulations of IR amide I′ spectra for a small protein in solution: comparison of explicit and empirical solvent models. J Phys Chem B 114:13011–13020

Hayward S, Kitao A, Go N (1994) Harmonic and anharmonic aspects in the dynamics of BPTI—a normal-mode analysis and principal component analysis. Protein Sci 3:936–943

He YF, Chen JY, Knab JR, Zheng WJ, Markelz AG (2011) Evidence of protein collective motions on the picosecond timescale. Biophys J 100:1058–1065

Hinsen K, Kneller GR (2008) Solvent effects in the slow dynamics of proteins. Protein Struct Funct Bioinform 70:1235–1242

Khajehpour M, Dashnau JL, Vanderkooi JM (2006) Infrared spectroscopy used to evaluate glycosylation of proteins. Anal Biochem 348:40–48

Kitao A, Hirata F, Go N (1991) The effects of solvent on the conformation and the collective motions of protein - normal mode analysis and molecular-dynamics simulations of melittin in water and in vacuum. Chem Phys 158:447–472

Knab J, Chen JY, Markelz A (2006) Hydration dependence of conformational dielectric relaxation of lysozyme. Biophys J 90:2576–2581

Lanczos C (1950) An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J Res Natl Bur Stand 45:255–282

Markelz AG (2008) Terahertz dielectric sensitivity to biomolecular structure and function. IEEE J Sel Top Quantum Electron 14:180–190

Markelz AG, Roitberg A, Heilweil EJ (2000) Pulsed terahertz spectroscopy of DNA, bovine serum albumin and collagen between 0.1 and 2.0 THz. Chem Phys Lett 320:42–48

Mott AJ (2012) Calculating infrared spectra of proteins and other organic molecules based on normal modes, Chap 4. ProQuest Dissertations and Theses. Arizona State University, AZ. http://login.ezproxy1.lib.asu.edu/login?; http://search.proquest.com/docview/1040706787?accountid=4485

Mott AJ, Thirumuruganandham SP, Thorpe MF, Rez P (2013) Fast calculation of the infrared spectra of large biomolecules. Eur Biophys J Biophys Lett 42:795–801

Plusquellic DF, Siegrist K, Heilweil EJ, Esenturk O (2007) Applications of terahertz spectroscopy in biosystems. Chem Phys Chem 8:2412–2431

Ruegg M, Metzger V, Susi H (1975) Computer analyses of characteristic infrared bands of globular proteins. Biopolymers 14:1465–1471

Stein SE (1992) NIST 35. NIST/EPA Gas-Phase Infrared Database. JCAMP Format

Tama F (2003) Normal mode analysis with simplified models to investigate the global dynamics of biological systems. Protein Pept Lett 10:119–132

Tama F, Brooks CL (2002) The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. J Mol Biol 318:733–747

Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. Protein Eng 14:1–6

Tama F, Gadea FX, Marques O, Sanejouand YH (2000) Building-block approach for determining low-frequency normal modes of macromolecules. Proteins Struct Funct Genet 41:1–7

Tama F, Valle M, Frank J, Brooks CL (2003) Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. Proc Natl Acad Sci USA 100:9319–9323

Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 77:1905–1908

Torii H, Tasumi M (1992) Model calculations on the amide-I infrared bands of globular proteins. J Chem Phys 96:3379–3387

Vinh NQ, Allen SJ, Plaxco KW (2011) Dielectric spectroscopy of proteins as a quantitative experimental test of computational models of their low-frequency harmonic motions. J Am Chem Soc 133:8942–8947

Zhang CF, Tarhan E, Ramdas AK, Weiner AM, Durbin SM (2004) Broadened far-infrared absorption spectra for hydrated and dehydrated myoglobin. J Phys Chem B 108:10077–10082

Zhang HL, Zukowski E, Balu R, Gregurick SK (2009) A dynamics study of the A-chain of ricin by terahertz vibrational calculation and normal modes analysis. J Mol Graph Model 27:655–663