Statistical method for predicting protein absorption peaks in terahertz region (postprint)

Authors: WU Yuting, ZHANG Wenmei, ZHAO Hongwei, SHAO Zhifeng, LI Xiaowei

Date: 2023-06-18T00:00:00+00:00

Abstract

Terahertz vibrational spectroscopy has recently been demonstrated as a novel noninvasive technique for the characterization of biological molecules. But the interpretation of the experimentally measured terahertz absorption bands requires robust computational method. In this paper, we present a statistical method for predicting the absorption peak positions of a macromolecule in the terahertz region. The essence of this method is to calculate the absorption spectra of a biological molecule based on multiple short scale molecular dynamics trajectories instead of using a long time scale trajectory. The method was employed to calculate the absorption peak positions of the protein, thioredoxin from Escherichia coli (E.coli), in the range of 10–25 cm–1 to verify the reliability of this statistical method. The predicted absorption peak positions of thioredoxin show good correlation with measured results demonstrating that the proposed method is effective in terahertz absorption spectra modeling. Such approach can be applied to predict characteristic spectral features of biomolecules in the terahertz region.

Full Text

Preamble

Statistical method for predicting protein absorption peaks in terahertz region

WU Yuting[1], ZHANG Wenmei[1], ZHAO Hongwei[2], SHAO Zhifeng[3], LI Xiaowei[3],*

[1]Institute of Modern Communication Technology, School of Physics and Electronics Engineering, Shanxi University, Taiyuan 030006, China
[2]Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Jiading Campus, Shanghai 201800, China

[3]Shanghai Center for Systems Biomedicine, Key Laboratory of Systems Biomedicine of Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract

Terahertz vibrational spectroscopy has recently emerged as a novel noninvasive technique for characterizing biological molecules. However, interpreting experimentally measured terahertz absorption bands requires robust computational methods. In this paper, we present a statistical method for predicting the absorption peak positions of macromolecules in the terahertz region. The essence of this method is to calculate absorption spectra based on multiple short-scale molecular dynamics trajectories rather than a single long-time-scale trajectory. We employed this method to calculate the absorption peak positions of thioredoxin from *Escherichia coli* (E. coli) in the range of 10–25 cm^{-1} to verify its reliability.

The predicted absorption peak positions of thioredoxin show good correlation with measured results, demonstrating that the proposed method is effective for terahertz absorption spectra modeling. Such an approach can be applied to predict characteristic spectral features of biomolecules in the terahertz region.

Key words: Terahertz, Molecular dynamics, Protein, Absorption spectrum

Introduction

Terahertz spectroscopy is an emerging technique for noninvasive characterization of low-frequency collective motions in biological molecules. Computational modeling plays an important role in predicting and interpreting these vibrational modes in the terahertz region. Several computational methods based on molecular dynamics simulations have been proposed for predicting far-IR spectra of macromolecules [**?**]. In these methods, vibrational frequencies associated with biomolecules were obtained either by normal mode analysis or quasi-harmonic analysis of trajectories from molecular dynamics simulations. Compared to normal mode analysis, which relies on the harmonic potential approximation, molecular dynamics simulations employ full potential energy functions that take anharmonicity into account. It has been shown that anharmonicity is important for low-frequency motions (lower than 80 cm^{-1}) [**?**]. Furthermore, ions and solvent, which are crucial for the stabilization of molecular conformation, can be included explicitly in the simulation. Recent calculations of absorption spectra for DNA and protein in the terahertz region show reasonable agreement with experimental results, demonstrating the capability of predicting spectral features in this region [**?**, **?**, **?**].

However, calculated absorption spectra are sensitively dependent on the convergence of molecular dynamics simulations [**?**]. In computational methods based on molecular dynamics simulations, the mass-weighted covariance matrix of atomic positional fluctuations is constructed and diagonalized to extract eigenvalues (or eigenfrequencies) and their corresponding eigenvectors. The eigenvectors define the configurational subspace where most conformational fluctuations occur. Noise in the definition of eigenvectors originates from insufficient sampling of simulation trajectories [?]. One way to solve the convergence issue is to run long production simulations to ensure sufficient thermodynamic sampling. Recent advances in computing technology have significantly increased our ability to conduct such long simulations.

Unfortunately, the conformational fluctuations of some proteins are not fully converged even at the 100 ns time scale [?, ?]. To improve conformational sampling in molecular dynamics simulations of biological molecules, it has been suggested that multiple short trajectories might be used rather than a single long trajectory [?, ?]. A recent study on optimal conditions for simulation convergence of proteins demonstrated that a production run length of about 100 ps was sufficient [?].

In this paper, we present a method using multiple short time-scale trajectories (1–2 ns) from molecular dynamics simulations to obtain absorption spectral features of biomolecules in the far-IR region. Instead of predicting absorption peaks based on a single short trajectory, our method obtains terahertz absorption features from statistical properties (i.e., histograms) of peaks calculated from multiple trajectories of relatively short simulation time. Using thioredoxin protein—a well-characterized system studied by various techniques including terahertz spectroscopy [?]—as an example, we demonstrate the reliability of the proposed method in reproducing experimentally measured low-frequency absorption bands. These results provide necessary support for applying this method to interpret characteristic absorption features of biomolecules in the far-IR region.

2 Modeling Methodology

The molecular dynamics simulations of thioredoxin were performed using the AMBER suite (version 10) [?]. The detailed simulation procedure was similar to that in Ref. [?]. The protein consists of 822 atoms in total. Four sodium ions were added to neutralize the protein using the ADDIONS routine implemented in AMBER. The thioredoxin and sodium ions were then solvated in a truncated octahedral water box using the TIP3P water model.

The preparation for MD simulations consisted of an initial energy minimization on the solvent and ions while the protein was fixed with a harmonic restraint of 2.1×10^{5} kJ/mol·nm^2. Following energy minimization of the whole system including the protein, water, and counterions, the molecular system was slowly heated to 300 K from 0 K under constant volume in 100 ps while the protein was restrained by a penalty energy of 1.04×10^{4} kJ/mol·nm^2. These restraints were gradually relaxed during a series of five segments of 1000 steps of energy minimization and 50-ps equilibration with constant temperature (300 K) and constant pressure (1 bar) via the Berendsen algorithm with a coupling constant of 0.2 ps for both parameters.

At the final stage of equilibration, we carried out a 50-ps simulation with a restraint of 2.1$\times10^{2}$ kJ/mol·nm^2 and a 50-ps unrestrained simulation. The production simulation with constant volume and temperature (NVT) ensemble was carried out for 20 ns. Electrostatic interactions were treated using the particle mesh Ewald algorithm with a real-space cutoff at 1 nm and cubic B-spline interpolation onto the charge grid with a spacing of about 0.1 nm. SHAKE constraints were employed for all bonds involving hydrogen atoms. The integration time step was 1 fs, and molecular trajectories were saved every 0.1 ps.

The molecular trajectories were first aligned against a reference structure (i.e., the initial thioredoxin conformation) to remove global translational and rotational differences between snapshots, which was implemented using the rms command of ptraj, the analysis program of the AMBER suite. Then, the 20-ns molecular trajectory was divided into multiple short trajectories with length, for example, 1 ns. For each short trajectory, the mass-weighted covariance matrix, which describes correlations between atomic positional fluctuations in the protein, was built and diagonalized to yield vibrational frequencies and corresponding amplitudes. Based on the vibrational frequencies and amplitudes, the absorption spectrum was calculated using methods presented previously [**?**]. For each absorption spectrum, absorption peak positions were measured. Histograms of the peak positions from all short trajectories were made using a bin size of 0.25 cm^{-1} for comparison with previous FTIR spectroscopic results of thioredoxin protein [**?**].

3 Results and Discussion

The overall conformational stability of thioredoxin in the molecular dynamics simulations was examined. The root-mean-square deviation (RMSD) of the 20-ns simulation with respect to the initial structure from X-ray diffraction was calculated and presented in Fig. 1. The averaged RMSD is about 0.14 nm, indicating that the protein structure remained stable over the course of the simulation.

[Figure 1: see original paper]

Figure 1 shows the RMSD of thioredoxin for the 20-ns simulation length.

Figure 2 shows the calculated absorption spectra of thioredoxin from multiple trajectories of 1-ns simulation, which were extracted from the 20-ns trajectories in the NVT ensemble. These absorption spectra clearly show similarities in terms of absorption peak positions. For example, absorption peaks at 10.5, 13.2, 15.0, 16.3, 19.0, 20.5, 23.6, and 24.8 cm^{-1} are shared across these five absorption spectra. On the other hand, significant differences in spectral features can be observed among these calculated spectra. For example, peaks at 12.4, 16.0, 18.2, 20.5, and 22.1 cm^{-1} are not found in every calculated spectrum. Furthermore, the absorption intensities associated with absorption peaks show significant differences. These differences make it difficult to compare theoretical results with experimental measurements.

[Figure 2: see original paper]

These differences arise from protein fluctuations during molecular dynamics simulation. We propose a method to statistically obtain absorption peak positions. In this method, we recorded the peak positions in the calculated spectra from 39 trajectories of 1-ns simulation. The distribution histogram of these peaks is shown in Fig. 3. The peaks in the histogram represent frequently observed absorption peak positions found in the absorption spectra. The histogram peaks are compared with experimentally measured absorption peaks in Table 1.

[Figure 3: see original paper]

Table 1 shows the measured and calculated absorption peaks associated with the protein (Unit: cm^{-1}) [**?**].

It can be seen that there is close correlation between the histogram peaks and the experimental absorption peaks.

To confirm the reliability of this method, we further investigated the sensitivity of the distribution histogram to key parameters including trajectory length and simulation ensemble. In Fig. 4, we compared the distribution histogram of absorption peaks calculated using trajectories of 1-ns length with that based on trajectories of 2-ns length. The two histograms are similar to each other, implying that the absorption peak distribution is not sensitive to trajectory length. Fig. 5 shows the statistical histograms obtained from trajectories based on NVT and NPT ensembles. It can be seen that the change of simulation ensemble does not vary the peak histogram significantly. Therefore, these results demonstrate that our method based on short trajectories can reliably predict the absorption peak positions of proteins in the terahertz region.

[Figure 4: see original paper]

[Figure 5: see original paper]

4 Conclusion

Using thioredoxin as an example, we have established that the absorption peak positions of protein molecules in the terahertz region can be successfully predicted using a statistics-based approach. This approach, which records the frequency distribution of peaks from different short trajectories, has the advantage of not needing to assess whether the simulations have converged, largely due to fluctuations in a single trajectory. This work can be generally applied to predict spectral features of other biological molecules (DNA and proteins) in molecular dynamics models and to analyze simulation trajectories involving peptides. Furthermore, we believe that it is important to consider the advantages of multiple trajectories, as this method not only makes the solution of convergence possible to a certain degree but also provides the means to measure it.

Acknowledgements

The authors thank Dr. Tatiana Globus and Dr. Boris Gelmont, Department of Electrical and Computer Engineering, University of Virginia, for critical reading of the manuscript.

References

1. Li X W, Globus T, Gelmont B, et al. J Phys Chem A, 2008, 112: 12090–12096.
2. Lee M S, Baletto F, Kanhere D G, et al. J Chem Phys, 2008, 128: 214506–214506.
3. Li X, Bykhovski A, Gelmont B, et al. IEEE Conf Nanotechnol, 2005, 1: 221–224.
4. Globus T R, Woolard D L, Khromova T, et al. J Biol Phys, 2003, 29: 89–100.
5. Bykhovskaia M, Gelmont B, Globus T, et al. Theor Chim Acta, 2001, 106: 22–27.
6. Hayward S, Kitao A, Gō N. Proteins, 1995, 23: 177–191.
7. Plusquellic D F, Siegrist K, Heilweil E J, et al. Chem Phys Chem, 2007, 8: 2412–2431.
8. Amadei A, Ceruso M A, Di Nola A. Proteins, 1999, 36: 419–424.
9. Faraldo-Gómez J D, Forrest L R, Baaden M, et al. Proteins, 2004, 57: 783–791.
10. Grossfield A, Feller S E, Pitman M C. Proteins, 2007, 67: 31–40.
11. Zhou Z, Joos B. Model Simul Mater Sc, 1999, 7: 383–395.
12. Schafer H, Mark A E, W FV Gunsteren. J Chem Phys, 2000, 113: 7809–7817.
13. Caves L, Evanseck J D, Karplus M. Protein Sci, 1998, 7: 649–666.
14. Alijabbari N, Chen Y, Sizov I, Globus T, Gelmont B. J Mol Model, 2012, 18: 2209–2218.
15. Bykhovski A, Gelmont B. J Phys Chem B, 2010, 114: 16095–16102.
16. Globus T, Bykhovskaia M, Woolard D L, et al. J Phys D Appl Phys, 2003, 36: 1314–1322.
17. Bykhovski A, Li X, Globus T, et al. Proc SPIE, 2005, 5995, 59950N.
18. Bykhovski A, Globus T, Thromova T, et al. IJHSES, 2008, 18: 109–117.
19. Case D A, Cheatham T E, Darden T, et al. AMBER 10, 2010, University of California, San Francisco.

Note: Figure translations are in progress. See original paper for figures.