Communication: Quantitative multi-site frequency maps for amide I vibrational spectroscopy

Mike Reppert and Andrei Tokmakoff

Citation: The Journal of Chemical Physics **143**, 061102 (2015); doi: 10.1063/1.4928637
View online: http://dx.doi.org/10.1063/1.4928637
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/143/6?ver=pdfcov
Published by the AIP Publishing

Communication: Quantitative multi-site frequency maps for amide I vibrational spectroscopy

Mike Reppert[1,2] and Andrei Tokmakoff[2,a)]
[1]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2]*Department of Chemistry, University of Chicago, Chicago, Illinois 60637, USA*

(Received 10 June 2015; accepted 4 August 2015; published online 14 August 2015)

An accurate method for predicting the amide I vibrational spectrum of a given protein structure has been sought for many years. Significant progress has been made recently by sampling structures from molecular dynamics simulations and mapping local electrostatic variables onto the frequencies of individual amide bonds. Agreement with experiment, however, has remained largely qualitative. Previously, we used dipeptide fragments and isotope-labeled constructs of the protein G mimic NuG2b as experimental standards for developing and testing amide I frequency maps. Here, we combine these datasets to test different frequency-map models and develop a novel method to produce an optimized four-site potential (4P) map based on the CHARMM27 force field. Together with a charge correction for glycine residues, the optimized map accurately describes both experimental datasets, with average frequency errors of 2–3 cm^{-1}. This 4P map is shown to be convertible to a three-site field map which provides equivalent performance, highlighting the viability of both field- and potential-based maps for amide I spectral modeling. The use of multiple sampling points for local electrostatics is found to be essential for accurate map performance. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4928637]

Amide I vibrational spectroscopy is an important tool for studying protein structure and dynamics in solution.[1] In addition to shedding light on bulk secondary structure content, amide I spectroscopy probes local protein structure and dynamics through the sensitivity of individual carbonyl frequencies to local electrostatics.[2] Experimentally, this local structural information is accessed by embedding ^{13}C or $^{13}C^{18}O$ isotope labels into specific amide groups along the peptide backbone, spectrally isolating the absorption bands of the labeled sites.[2]

This link between local structure and spectroscopic observables has long made the development of an accurate amide I spectral "map" a highly desirable goal.[3–11] Ideally, such a map translates molecular dynamics (MD) trajectories into spectroscopic observables such as frequencies and line shapes. In practice, the situation is complicated by inaccuracies in the map.[12] For example, a recent study on hydrogen bonding of water to N-methylacetamide (NMA) compared 14 different amide I literature maps against *ab initio* calculations, finding wide variation in map predictions (often with opposing trends) and no clear agreement with the *ab initio* data.[13]

To overcome this uncertainty, we have been developing experimental standards for amide I frequency maps. Our first dataset consisted of experimental absorption spectra from 28 dipeptide fragments in various protonation states in solution for a total of 73 reference frequencies.[14] Experimental frequencies were found to correlate strongly with the projection of the electric field along the carbonyl bond evaluated at the O atom (E_x^O), reminiscent of a Stark-like mechanism. However, this dataset was unable to distinguish between this one-site field (1F) map and several multi-site potential models which provided similar results.

This ambiguity reflects a long-standing debate as to which electrostatic variables are most suitable for predicting vibrational frequencies.[3,5] Most maps take amide I frequencies to be proportional to either the electrostatic potential Φ,

$$\omega = \omega_o + \sum_{i=1}^{N_{sites}} C_i \Phi_i, \tag{1}$$

or the field E_α,

$$\omega = \omega_o + \sum_{i=1}^{N_{sites}} \sum_{\alpha=1}^{3} C_{i,\alpha} E_\alpha^{(i)}, \tag{2}$$

evaluated at specified sites i around the amide unit (typically the location of the C, O, N, or H atoms). In the field-map expression, the index α sums over Cartesian coordinates (x, y, and z, with the x-axis pointing along the C=O bond). These electrostatic variables are obtained from the MD coordinates and atomic charges by summing over the Coulomb contributions from each atom in the simulation, making explicit water solvation critical. Arguments have been made in favor of both field and potential maps, but neither has demonstrated a clear advantage.[3,5,13]

To provide a more rigorous test of such models, we turned to isotope-enriched protein expression to produce five isotope-labeled variants of the protein G mutant NuG2b.[12] Each construct displays uniform $^{13}C^{18}O$ labeling of all amide bonds immediately following a particular residue: Gly (3 sites), Ala (8 sites), Val (6 sites), Leu (2 sites), or Phe (3 sites). In contrast to the floppy, highly solvated dipeptides in our

a)Author to whom correspondence should be addressed. Electronic mail: tokmakoff@uchicago.edu. Telephone: (773) 834-7696.

original study, the NuG2b labels probe mid-chain amide bonds in a variety of conformations in a protein of extreme structural stability. Comparison of the experimental data with simulated spectra for the maps of la Cour Jansen[4,15] and Skinner[6] and for our dipeptide 1F map[14] revealed significant discrepancies for all maps tested.

In this work, we combine our dipeptide and NuG2b datasets with new MD simulations to identify optimal map parameters. The results strongly favor multi-site maps, although the optimized map may be either field- or potential-based. Optimized 4P (four-site potential) and 3F (three-site field) maps provide excellent agreement with both the dipeptide and NuG2b datasets.

To begin our discussion, Figure 1 presents a comparison of experimental absorption spectra (shaded curves) for our five isotope-labeled NuG2b constructs against simulated absorption spectra (black lines) using the four CHARMM27 maps presented in our earlier dipeptide work (Table I of Ref. 14). MD parameters were identical to those in our previous work (CHARMM27 force field, SPC/E water).[12] Spectral simulations were calculated for the entire protein (isotope labeled sites shifted by -65 cm^{-1}) using Torii's numerical wavefunction propagation scheme[16] and averaged over ten independent 1 ns runs following a 1 ns equilibration period. Coupling constants were evaluated using la Cour Jansen's transition charge and nearest-neighbor dihedral models, while transition dipoles were assigned using the zero-field values of la Cour Jansen's electrostatic map.[15] For the dipeptide dataset alone, the 1F (E_x^O), 2P (two-site potential, Φ_C/Φ_O), and 4P maps give similar fit qualities against experiment (2-3 cm^{-1} sample standard error), with the 2F (two-site field, E_x^C/E_x^N) model showing poorer performance (~5 cm^{-1} error).[14] In contrast, performance for the NuG2b dataset is quite varied. The 2F map in particular produces serious errors in both frequencies and line shapes, effectively eliminating it (in this parameterization) as a useable spectral map. The three remaining maps (1F, 2P, and 4P) show similar performance for the Ala, Val, and Leu sites, matching experimental frequencies to within a few cm^{-1}. The more varied performance of the Gly and Phe labels is worth considering in detail.

In the Gly label spectra, performance deteriorates for all four maps, echoing similar discrepancies observed for dipeptides with N-terminal glycines.[14] Upon examining the atomic charges assigned by the CHARMM27 force field,[17] the cause of this discrepancy is easily identified. Atomic charges for CHARMM27 are defined so that for all charge-neutral amino acids *except* glycine, the net side chain charge is zero. For glycine, in contrast, the side chain consists of a single hydrogen atom onto which a partial charge of $+0.09\ e_o$ (units of elementary charge) is shifted from the α-carbon. Empirically, we find that by shifting this charge back to the α-carbon in spectral simulations (not during the MD run) and instead splitting a charge of $+0.045\ e_o$ between the two hydrogen atoms, glycine residues can be made to perform equivalently with other amino acids in the dipeptide calculations.[18] The same correction improves the agreement with the calculated NuG2b Gly label spectra presented in Figure 1 (dashed lines).

The Phe label spectra present a more complex situation. The experimental spectra show an intense peak near 1595 cm^{-1} and a weaker shoulder near 1584 cm^{-1}, likely corresponding to the two NuG2b β-sheet Phe labels and the single α-helix label, respectively.[12] As observed in our preliminary study,[12] the 1F map incorrectly inverts the order of these two peaks, putting the β-sheet peak at lower frequency. In contrast, the 2P map assigns roughly the same frequency to all three labels; the correct asymmetry appears only with the 4P map, improving further with the 2F map.

This cursory comparison reveals that frequency predictions for the Phe label improve in the order 1F < 2P < 4P < 2F. It is significant that this is also the order of increasing sensitivity to electrostatics near the NH bond: the 1F model parameterizes only the O atom, the 2P model parameterizes both C and O, the 4P model parameterizes all four sites, and the 2F model parameterizes only C and N. This correlation suggests that the 4P and 2F maps predict the correct frequency ordering by sampling electrostatics near the NH group. Physically, this is unsurprising as it has long been known that amide I vibrations are sensitive to NH-group hydrogen bonding.[19] From a map-development standpoint, these results are important for two reasons. First, the favorable performance of the 4P map strongly recommends it as a starting point for future work. Second, the more varied local environments present in the NuG2b dataset evidently provide a more restrictive test of map parameters than the uniformly solvent-exposed dipeptide data.

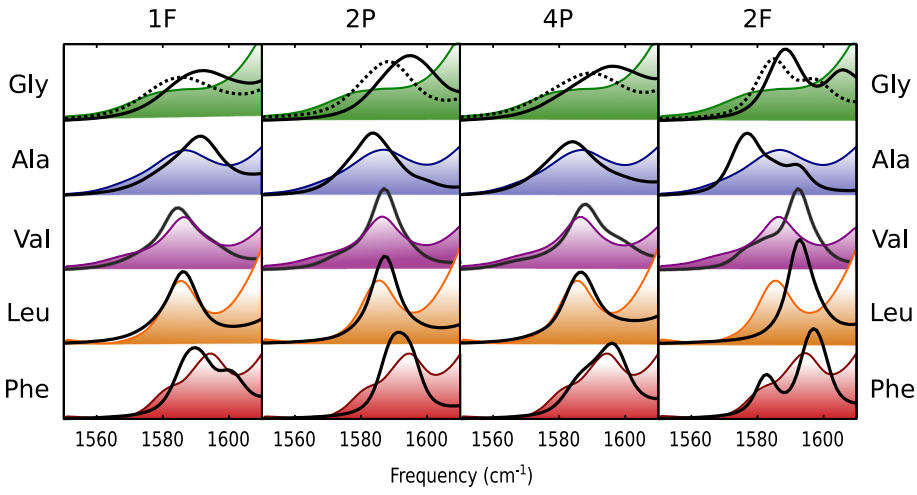

FIG. 1. Experimental (shaded curves) and simulated (black solid lines) amide I absorption spectra for isotope-labeled NuG2b for four different electrostatic maps. The dotted lines show Gly label spectra after introducing the charge correction described in the text.

These observations suggest a straightforward approach to map optimization. Using the 4P model described above, the computationally accessible dipeptide dataset can be used to create a set of partially optimized maps which are further refined against the NuG2b dataset. In this scheme, the dipeptide data effectively reduce the dimensionality of the 4P parameter space. A general 4P map contains four free parameters: a zero-field frequency ω_o and four map coefficients constrained by the charge-neutrality condition that $C_C + C_O + C_N + C_H = 0$.[3] By scanning over the full range of possibilities for two of these parameters, C_N and C_H, and determining the remaining parameters by least-squares optimization against the dipeptide dataset, we obtain a lower-dimensional family of maps which can be directly tested against NuG2b data, a search that would be both prohibitively expensive and of questionable reliability without the constraints afforded by the dipeptide dataset.

The performance of this family of restricted maps over the dipeptide dataset is presented in Figure 2 (left). Here, the standard frequency error for each map against the dipeptide experimental data is plotted as a function of the coefficients C_N and C_H. All simulations use the glycine charge correction described above; local electrostatic variables are calculated by summing over the Coulomb interaction with all other atoms, excluding only the C, O, N, and H atoms, and the two covalently bonded C_α atoms. A global minimum of 2 cm^{-1} in error is attained near $C_N = -4$ and $C_H = 187$ cm^{-1} e_o/E_H, meaning that for this map, ~68% (~95%) of all predicted dipeptide frequencies is within ±2 cm^{-1} (±4 cm^{-1}) of their experimental value. Full parameters for the optimized map are provided in Table I. As expected for a predominantly C=O stretch vibration, the largest coefficients appear on the C and O atoms, with smaller values for the N and H atoms. This is in contrast to several density functional theory (DFT)-based maps which assign a large coefficient—sometimes positive[8] and sometimes negative[9]—to the N atom itself.

In our results, a strong correlation is evident between C_N and C_H for low-error maps, with minimum error obtained

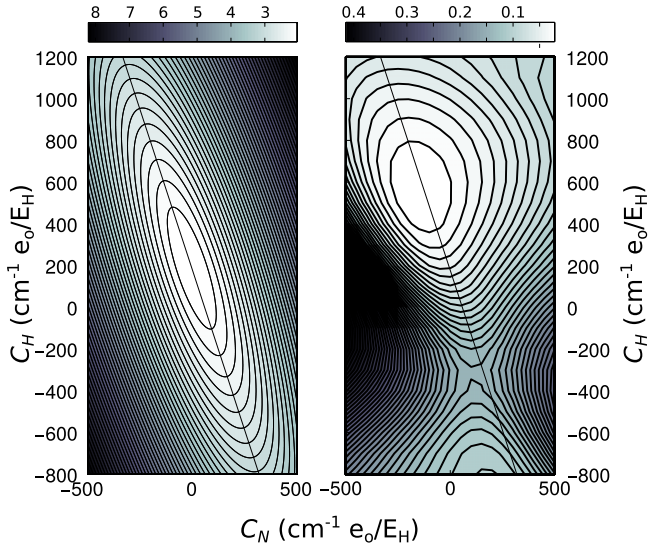

FIG. 2. Error surfaces for 4P maps against dipeptide and NuG2b datasets as a function of the coefficients C_N and C_H. Left: Standard error between predicted and experimental dipeptide mean frequencies. Right: MSD between simulated and experimental spectra averaged across all NuG2b isotope labels.

TABLE I. Dipeptide standard error (σ_{dip}), NuG2b MSD, and frequency map coefficients for the dipeptide-optimal 4P map (4PN-4), NuG2b-optimal 4P map (4PN-150), and 3F translation of the 4PN-150 map.

Map	σ_{dip} (cm^{-1})	MSD	ω_o (cm^{-1})	Coefficients
4PN-4	2.00	0.033	1716.6	$1076.8 \cdot \Phi_C$ $-1259.8 \cdot \Phi_O$ $-4.2 \cdot \Phi_N$ $187.2 \cdot \Phi_H$
4PN-150	2.25	0.022	1746.4	$1121.7 \cdot \Phi_C$ $-1571.7 \cdot \Phi_O$ $-150 \cdot \Phi_N$ $600 \cdot \Phi_H$
3F	2.16	0.016	1746.4	$3684.2 \cdot E_x^{CO}$ $600.43 \cdot E_x^{CN}$ $-957.9 \cdot E_y^{CN}$ $1161.8 \cdot E_x^{NH}$

along the line,

$$C_H \approx -3C_N + 175\frac{\text{cm}^{-1}e_o}{E_H}. \quad (3)$$

Moreover, the dipeptide-optimized values for the coefficients C_C and C_O are found to be nearly perfect linear functions of C_N and C_H so that specification of C_N and C_H effectively determines the complete map.[18] Along the minimum-error line defined by Eq. (3), then, these relations define a family of best-fit maps cast as a function of C_N to emphasize the contribution of the N and H sites. This low-dimensional family of maps should be useful for future optimization studies. It is important to note, however, that these parameters are valid only when using CHARMM27 atomic charges and are dependent on the dipeptide structural ensemble used in parameterization. Although the self-consistency of our results is encouraging, ensemble reliability remains an important question, and independent experimental analysis of the dipeptide structures sampled would do much to confirm or refine these parameters.[20–22]

For comparison, the right panel of Figure 2 shows the mean square deviation[12] (MSD) of the simulated NuG2b isotope-label spectra against experiment over the same range of parameters. For each spectrum, the MSD is calculated as

$$\Delta^2 = \left\| A_{exp}(\omega) - A_{sim}(\omega) \right\|^2, \quad (4)$$

where $A_{exp}(\omega)$ and $A_{sim}(\omega)$ are, respectively, the experimental and simulated absorption spectra normalized such that $\|A(\omega)\|^2 = \int_{1525}^{1600} (A(\omega))^2 d\omega = 1$. The reported values are averaged across the five isotope-label spectra. For these calculations, step sizes of 50 cm^{-1} e_o/E_H and 100 cm^{-1} e_o/E_H (where e_o is the elementary charge and E_H the Hartree energy) were used for C_N and C_H, respectively, within the ranges displayed in the figure. The minimum MSD is located at $C_N = -150$ cm^{-1} e_o/E_H and $C_H = 600$ cm^{-1} e_o/E_H, not far from the dipeptide minimum and along the minimum-error line defined by Eq. (3). Full map parameters are provided in Table I. The standard error for this map applied to the dipeptide dataset is 2.25 cm^{-1}, a minor increase from the minimum error of 2 cm^{-1}.

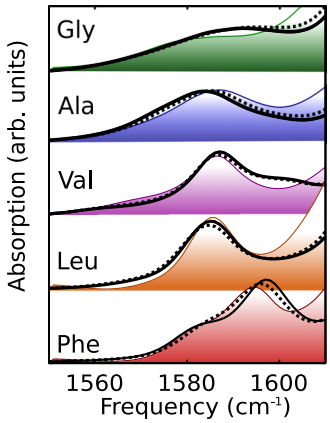

FIG. 3. Comparison of experimental (shaded) and simulated (black lines) absorption spectra for NuG2b isotope labels using the optimized 4P (solid) and 3F (dashed) maps.

Simulated spectra for the NuG2b-optimized map are presented in Figure 3 (solid black lines). Agreement with experiment (shaded curves) is significantly improved over Figure 1, with all peak frequencies matching within 3 cm^{-1} for all sites except Gly; an exact number for Gly is difficult to assign due to overlap with the main band. Agreement in the main band region (not shown) is more modest, with deviations of up to ~5 cm^{-1} between simulated and experimental peaks. Interestingly, the isotope-label and main-band regions appear to be optimized under different map parameters, with the main band favoring the broader spectra obtained with more negative values of C_N. This apparent conflict suggests coupling model refinement as an important next step since the choice of coupling model influences the main band absorption spectra much more strongly than the isotope-labeled peaks. In fact, the 4P model itself points the way to at least a partial scheme for coupling constant optimization since the frequency shift coefficients C_i may be associated with effective transition charges for use in an electrostatic coupling model.

In closing, it is worth commenting on the distinction between field- and potential-based frequency maps. Although our best results were obtained for a multi-site potential map, multi-site field maps can provide similar results. In fact, any given four-site potential map may be converted into an approximately equivalent three-site field map since for typical bond lengths, the difference in potential between two bonded atoms (e.g., $\Phi_C - \Phi_O$) is well approximated by the electric field evaluated at the middle of the bond, projected along the bond vector (E_{CO}), and scaled by its length (r_{CO}). The charge-neutrality constraint $\sum_{i=1}^{N} C_i = 0$ implies that an N-site potential map may always be rewritten in terms of $N-1$ potential differences, giving a direct translation from potential to field,

$$\omega = \omega_o + \sum_{i=1}^{N_{sites}-1} (C_1 + \cdots + C_i)(\Phi_i - \phi_{i+1})$$
$$\approx \omega_o + \sum_{i=1}^{N_{sites}-1} (C_1 + \cdots + C_i)\, r_{i,i+1} E_{i,i+1}.$$

The results of such a "translation" for our optimized 4P map are presented in Table I and Figure 3. Although slight variations can be distinguished between spectra simulated using the 4P (solid) and 3F (dashed) maps, performance is essentially similar. Importantly, this mapping between 4P and 3F maps is not generally invertible; any 3F map including electric field terms that are *not* projected along a bond has no corresponding 4P map. As noted by Torii, such contributions can be induced by nuclear motion off the bond axis; in such cases, field-based (or mixed field/potential) maps may be superior to potential maps alone.[13] For amide I vibrations, however, our results suggest that field- and potential-based maps may be equally effective, making the choice between them a matter of convenience.

The primary conclusion to be drawn from these results is that inclusion of multiple sampling points around the amide bond is essential for accurate amide I frequency prediction. By combining our dipeptide and NuG2b datasets and introducing a dipeptide-based charge correction for glycine residues, we parameterize a 4P map that predicts all dataset frequencies to within a few cm^{-1} of experiment. The agreement is particularly satisfying in that it is obtained without the introduction of any "nearest neighbor" dihedral shift for through-bond interactions,[8,11,15] suggesting that simple electrostatic maps may adequately describe amide I frequency shifts even in protein systems. Finally, our results produce a single-parameter family of 4P maps, providing a low-dimensional basis for future work in amide I map optimization.

We thank the National Science Foundation (NSF) (Grant Nos. CHE-1212557 and CHE-1414486) for support of this research. M.R. thanks the NSF for a Graduate Research Fellowship.

[1]A. Barth and C. Zscherp, Q. Rev. Biophys. **35**, 369 (2002).
[2]C. Baiz, M. Reppert, and A. Tokmakoff, in *Ultrafast Infrared Vibrational Spectroscopy*, edited by M. D. Fayer (Taylor & Francis, Boca Raton, 2013).
[3]S. Ham, J.-H. Kim, H. Lee, and M. Cho, J. Chem. Phys. **118**, 3491 (2003).
[4]T. la Cour Jansen and J. Knoester, J. Chem. Phys. **124**, 044502 (2006).
[5]J. R. Schmidt, S. A. Corcelli, and J. L. Skinner, J. Chem. Phys. **121**, 8887 (2004).
[6]L. Wang, C. T. Middleton, M. T. Zanni, and J. L. Skinner, J. Phys. Chem. B **115**, 3713 (2011).
[7]P. Bouř and T. A. Keiderling, J. Chem. Phys. **119**, 11253 (2003).
[8]T. M. Watson and J. D. Hirst, Mol. Phys. **103**, 1531 (2005).
[9]H. Maekawa and N.-H. Ge, J. Phys. Chem. B **114**, 1434 (2010).
[10]T. Hayashi, W. Zhuang, and S. Mukamel, J. Phys. Chem. A **109**, 9747 (2005).
[11]R. D. Gorbunov, D. S. Kosov, and G. Stock, J. Chem. Phys. **122**, 224904 (2005).
[12]M. Reppert, A. R. Roy, and A. Tokmakoff, J. Chem. Phys. **142**, 125104 (2015).
[13]H. Torii, J. Phys. Chem. Lett. **6**, 727 (2015).
[14]M. Reppert and A. Tokmakoff, J. Chem. Phys. **138**, 134116 (2013).
[15]T. la Cour Jansen, A. G. Dijkstra, T. M. Watson, J. D. Hirst, and J. Knoester, J. Chem. Phys. **125**, 44312 (2006).
[16]H. Torii, J. Phys. Chem. A **110**, 4822 (2006).
[17]P. Bjelkmar, P. Larsson, M. A. Cuendet, B. Hess, and E. Lindahl, J. Chem. Theory Comput. **6**, 459 (2010).
[18]See supplementary material at http://dx.doi.org/10.1063/1.4928637 for details on the glycine charge correction (Figures S1) and complete linear relations between frequency shift coefficients.
[19]H. Torii, T. Tatsumi, and M. Tasumi, J. Raman Spectrosc. **29**, 537 (1998).
[20]D. S. Kosov and G. Stock, J. Phys. Chem. B **107**, 5064 (2003).
[21]J. Grdadolnik, V. Mohacek-Grosev, R. L. Baldwin, and F. Avbelj, Proc. Natl. Acad. Sci. U. S. A. **108**, 1794 (2011).
[22]R. Schweitzer-Stenner, Mol. BioSyst. **8**, 122 (2012).