The Partition Function of Large Biomolecules, and its relevance to Terahertz and Infrared Spectroscopy.

H N Rutt, *Senior Member, IEEE*

Abstract—**Molecules of biological interest such as proteins and enzymes are typically very large compared to those traditionally studied by infrared and terahertz spectroscopy. The 'average' protein has some 5000 atoms and 15,000 vibrational modes. We show that this leads to extreme values of the partition function, essentially the probability of finding a molecule in the ground state, at room temperature. In fact for a practical sample at 300K the probability of finding a molecule in the ground state (or any other specific state) is vanishingly small since the partition function exceeds the number of molecules present by many orders of magnitude. Some implications of this fact for spectroscopy of these molecules, such as the impact of 'anharmonic broadening', sum and difference bands, are discussed.**

Index Terms—**infrared, partition functions, proteins, spectroscopy, terahertz**

I. Introduction

In recent years there has been a rapid rise in interest in the terahertz (far-infrared, FIR) spectroscopy of biological materials, in particular proteins, peptides, DNA and their constituent building blocks, amino acids. In 2005 some 15 papers in refereed Journals mention a combination of these terms, with 19 in 2006 up October, whilst a decade ago only some three papers per year were published. In contrast mid infrared spectroscopy of biomolecules has been well established for many years. The various amide bands of proteins in particular have proved useful for structural estimates, in particular of the degree of beta sheet as opposed to alpha helix structure present. [1] and [2] provide excellent reviews.

The rise of interest in THz spectroscopy is driven by the availability of time-domain techniques in the terahertz region using electro-optic detection, combined with the increasing interest in the 'life science' aspects of physical measurement. It has driven a parallel interest in the theoretical prediction of the absorption spectrum in the THz region and hence in the frequencies of the vibrational modes. To date however no large biomolecule has a complete assignment of its vibrational spectrum

For a molecule of point group C_1, which is typical of the species considered here containing N atoms there are 3N-6 vibrational modes, the six typically being negligible. All modes are formally infrared allowed, and non degenerate. Whilst the arguments in this paper apply to all large biomolecules, we use proteins as a convenient example.

From the residue frequencies for the standard twenty amino acids (EMBL-EBI database, [3]), the conversion from residues to number of atoms is N~15.2* number of residues. [3] shows that proteins of biological importance have an average size of 324 residues and some 15,000 modes. In fact the largest protein described appears to be 'Titanic Titin' of 27,000 residues and over one million modes [4].

No fully assigned experimental spectrum is available for a substantial biomolecule. Considerable effort has been expended on modeling, and can be used to estimate mode densities in the critical low frequency region. The small 58 amino acid protein BPTI [5] shows mode densities of 2 to 3 per wavenumber from ~3cm^{-1} to ~100cm^{-1}. [6] quotes lowest mode frequencies from 2.3 to 4.6cm^{-1} (the latter for BPTI) with from 19 to 62 modes below 20cm^{-1}. For DNA, [7] shows an observed 'peak' density of ~1/cm^{-1} from 10 to 200cm^{-1}.

II. Partition Functions

Protein partition functions have received little attention in the literature, and none in the context of THz spectroscopy to the author's knowledge. We thus briefly review their definition.

The vibrational partition function, generally denoted $Q_v(T)$ is defined such that the population in a given vibrational state, *s*, with degeneracy g_s, and energy E_s, is given by

$$N_s(T) = \frac{g_s N_T}{Q_v(T)} \exp\left(\frac{-E_s}{kT}\right) \quad (1)$$

When many vibrational states are present, again in the harmonic approximation the individual Q_v for each state are *multiplied* together to obtain the overall partition function. This arises because the introduction of each extra mode introduces not only that mode and its harmonics, but the additional combination states with all other modes. The overall partition function of a polyatomic molecule with non degenerate modes is thus

$$Q_v(T) = \prod_s \frac{1}{1-\exp\left(\frac{-E_s}{kT}\right)} \quad (4)$$

where the product runs over all vibrational states.

In molecular dynamics simulations mean-square-displacements and Debye-Waller factors are routinely calculated. These are closely related to partition functions but their relevance to infrared/THz spectroscopy has received scant attention.

III. The Values of Partition Functions in Proteins

Partition functions are most commonly encountered with regard to relatively small molecules. For many small molecules studied by high resolution infrared spectroscopy the value of Q_v at 300K is less than two, whilst for relatively large species with a few low lying modes in the 100-200cm^{-1} region it might reach tens. At 77K, where kT is ~54cm^{-1}, the value is close to unity for almost all molecules with detailed resolved and assigned spectra. The very large number of low frequency modes in biomolecules leads to radically different values.

Stretching modes lying in the region of 3000cm^{-1} contribute negligible amounts to the partition function. At 300K, even 10^5 modes at 3000cm^{-1} gives a Q_v of only 1.052. Modes in the so-called 'fingerprint' region ~900-1300cm^{-1} contribute significantly; Q_v reaches 3000 for 1000 modes at 1000cm^{-1}. However the low frequency modes dominate; a single mode at 3.3cm^{-1} (0.1THz) contributes a factor of 69.5 to Q_v.

For illustration (Figure 1) we consider a biomolecule with 100 modes distributed uniformly in the range from a cut-off up to 100cm^{-1} (3.3THz). This is consistent with the values quoted above for very small proteins, and extremely conservative for larger molecules. High frequency modes are ignored.

IV. Implications of Large Partition Function Values

The extremely high values of partition function at 300K, in excess of 10^{60}, grossly exceeds the number of molecules in a physical sample, ~$6*10^{20}$ for a millimole (29g of the 'average' protein.) This remains true even at 77K. ***No molecules can be expected in the ground state, or indeed in any other specific quantum vibrational state, at temperatures above ~77K!*** At 4.2K for a 3cm^{-1} cut off Q_v is still 2.64 and less than half the molecules are in the ground state. For larger molecules Q_v at 4.2K can reach substantial values; with 200 modes from 1.5 to 100 cm^{-1}, very conservative for larger proteins, Q_v=45.3.

Calculations of the infrared and THz absorption spectrum thus need to incorporate averaging over the statistically occupied states. Essentially all observed absorptions at above 77K are 'hot bands'. This contributes a form of 'anharmonic broadening'; assuming a Morse potential it can be shown that the *fractional* (as opposed to absolute) width of bands in the THz region is expected to be similar to that in the mid-IR.

Detailed calculations of the spectrum will also require the inclusion of sum and difference bands; whilst individually

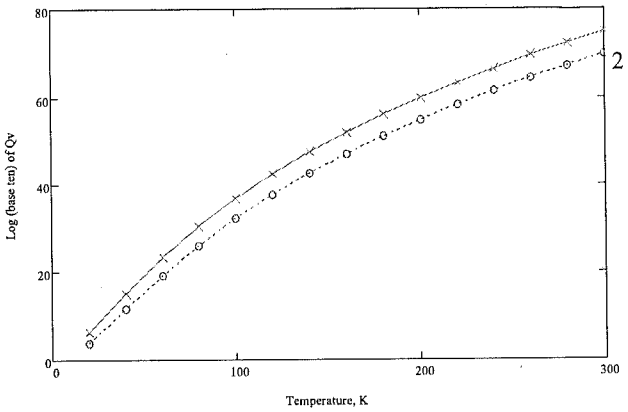

Fig. 1. $\text{Log}_{10}(Q_v)$ as a function of temperature for 100 modes uniformly distributed between 3 (continuous, crosses) or 9 (dotted, circles) cm^{-1} and 100cm^{-1}.

weak, the very large number contributing are likely to considerably alter the observed spectrum, and strongly contribute to the ubiquitous broad background absorption seen in most THz region biomolecule spectra. Realistic modeling of these anharmonic effects, required up to quantum numbers as high as over one hundred at 300K, is extremely challenging and may require new approaches.

V. Conclusions

We have shown that the partition function for large biomolecules, takes on extreme values at 300K. It remains extremely large even at 77K, and for larger molecules remains significant at 4.2K. The values are so high that a typical sample contains no molecules in the ground state – or indeed any other specific quantum state. Hence an accurate model of infrared and terahertz spectra must include 'anharmonic broadening' effects, and contributions from sum and difference bands. As a first step in this direction, we have shown that under the assumption of a Morse potential, fractional line widths in the terahertz region due to 'anharmonic broadening' should be comparable to those in the mid-IR.

References

[1] A. Barth and C Zscherp. "What vibrations tell us about proteins," Quarterly Reviews of Biophysics, vol 35, no 4, pp369-430 2002

[2] H.H. Mantsch and D. Chapman, Eds. *Infrared Scpectroscopy of Biomolecules*.Wiley-Liss 1996

[3] The UniProtKB/TrEMBL protein database available online at http://www.ebi.ac.uk/swissprot/sptr_stats/index.html

[4] S. Labeit and B. Kolmerer Titins – Giant Proteins in charge of muscle ultrastructure and elasticity. Science vol 270 (5234) pp293-296 October 1995

[5] S. Hayward and N Go. Collective variable description of native protein dynamics. Annu Rev Phys Chem vol46 pp223-250 1995

[6] M. Levitt, C Sander, P. S. Stern. Protein Normal-mode Dynamics: Trypsin Inhibitor, Crambin, Ribonuclease and Lysozyme. J Mol Biol vol181 pp423-447 1985

[7] T. R. Globus et al. THz-Spectroscopy of Biological Molecules. J of Biological Physics. Vol29 pp89-100 2003