

Quantitation of the Area of Overlap between Second-Derivative Amide I Infrared Spectra To Determine the Structural Similarity of a Protein in Different States

BRENT S. KENDRICK^x, AICHUN DONG, S. DEAN ALLISON, MARK C. MANNING, AND JOHN F. CARPENTER

Received August 7, 1995, from the *Department of Pharmaceutical Sciences, School of Pharmacy, Campus Box C-238, University of Colorado Health Sciences Center, Denver, CO 80262.* Final revised manuscript received November 8, 1995. Accepted for publication November 9, 1995[®].

Abstract □ Maintaining a nativelike structure of protein pharmaceuticals during lyophilization is an important aspect of formulation. Infrared spectroscopy can be used to evaluate the effectiveness of formulations in protecting the secondary structural integrity of proteins in the dried solid. This necessitates making quantitative comparisons of the overall similarity of infrared spectra in the conformationally sensitive amide I region. We initially used the correlation coefficient r , as defined by Prestrelski et al. (*Biophys. J.* **1993**, *65*, 661–671), for this quantitation. Occasionally, we noticed that the r value did not agree with a visual assessment of the spectral similarity. In some cases this was due to an offset in baselines, which led artifactually to an unreasonably low r value. Conversely, if the spectra were baseline corrected and there existed a large similarity between peak positions, but differences in relative peak heights, the r value would be unreasonably high. Our approach to avoiding these problems is to use area-normalized second-derivative spectra. We have found that quantitating the area of overlap between area-normalized spectra provides a reliable, objective method to compare overall spectral similarity. In the current report, we demonstrate this method with selected protein spectra, which were taken from experiments where unfolding was induced by lyophilization or guanidine hydrochloride, and artificial data sets. With this analysis, we document how problems associated with calculation of the correlation coefficient, r , are avoided.

Lyophilization is often used to prepare stable formulations of protein therapeutics. It has recently been shown that inhibition of protein unfolding during lyophilization is important for acute recovery of activity after rehydration and for long-term storage stability in the dried solid.^{1–5} Infrared spectroscopy provides the most powerful and convenient means to compare the secondary structure in the dried solid to that of the native aqueous protein. In practice, it is most useful to determine the overall or “global” similarity between the second-derivative amide I spectra for the protein in the liquid and dried states. To quantitate the overall similarity, we initially used the correlation coefficient (r value), as described by Prestrelski et al.^{1,2} This analysis proved useful. For example, we found that stabilizing additives led to an increase in the r value for the dried versus native aqueous protein.^{1,2} Also, the recovery of activity of labile enzymes after rehydration and the long-term storage stability of proteins in the dried solid correlated directly with the r values for dried versus native proteins; i.e., retention of native structure in the dried solid is necessary for activity recovery after rehydration and for storage stability.^{1–5}

However, during subsequent analysis of several proteins, we noticed that occasionally the r value did not agree with the visual impression of spectral similarity. For example, pairs of second-derivative spectra for a given protein that visually appeared very similar, except for an offset in base-

lines, had relatively low r values. Conversely, if the spectra were baseline corrected and there existed a large similarity between peak positions, but differences in relative peak heights, the r value would be unreasonably high, relative to the visual impression of similarity.

In this report, we show examples of these problems and describe a new and more reliable method to compare the overall similarity between two second-derivative infrared spectra of proteins. The foundation for this global comparison comes from the use of second-derivative spectra to quantitate protein secondary structure.^{3,6–8} By curve-fitting the component bands, their relative contributions to the total area of the amide I region can be determined. Combining this analysis with an assignment of individual bands to a secondary structure (e.g., β -sheet or α -helix), the secondary structural composition of a protein can be estimated. Similarly, to determine if relative band areas have been altered by a treatment, which is indicative of an alteration in secondary structure (e.g., lyophilization-induced intermolecular β -sheet in protein aggregates), the area-normalized second-derivative spectra can be overlaid and compared visually. To quantitate the overall similarity between spectra in this analysis, we have developed a method to calculate the area of overlap between two area-normalized spectra. In this report, we demonstrate this method with selected protein spectra and artificial data sets, and document how the problems associated with calculation of the correlation coefficient, r , are avoided.

Experimental Section

Infrared Spectra—The spectra used in this study were previously published by Dong et al.³ and Bowler et al.⁹ The methods for obtaining the spectra are described in earlier studies.^{4–8}

Calculations for Comparison of Overall Spectral Similarity—Second-derivative spectra were obtained with the derivative function of OMNIC software (Nicolet). The r value calculations were performed on second-derivative spectra by two equivalent methods. For uncorrected spectra, a custom OMNIC subroutine provided by Ben Garland of Nicolet Instruments was used. For baseline-corrected and area-normalized spectra, a custom Microsoft Excel spreadsheet function was used. Both algorithms input OMNIC spectral data from a reference spectrum and a comparison spectrum, within a user-defined wavenumber range, and output the fraction similarity (i.e., correlation) by the following formula:

$$r = \frac{\sum (x_i y_i)}{\sqrt{\sum x_i^2 \sum y_i^2}} \quad (1)$$

where x_i , y_i = intensity of (reference spectra, comparison spectra) at wavenumber (i).

For area of overlap calculations, second-derivative infrared spectra are imported into GRAMS/386 software (Galactic). This software is used to baseline correct for the desired wavenumber range (typically from 1705 to 1600 cm^{-1}), to integrate the area under the curve in this region, and to normalize the area to a value of 1.0. Using a custom program written for GRAMS/386, a column of data is created

[®] Abstract published in *Advance ACS Abstracts*, December 15, 1995.

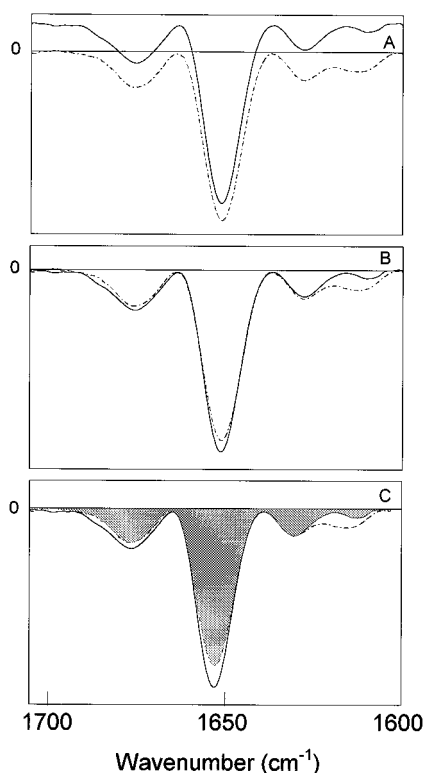


Figure 1—Second derivative amide I spectra of γ -interferon: (A) raw spectra, $r = 0.80$; (B) baseline offset corrected, $r = 0.995$; and (C) area-normalized spectra, area of overlap = 0.92. Solid lines indicate the native aqueous state, dashed lines indicate rehydrated aqueous state, and the gray fill indicates the area of overlap. Spectra modified from Dong et al.³

with a logical test function which takes the lowest intensity value of a comparison spectrum and a reference spectrum, at a given wavenumber. The wavenumber array and the lowest intensity value array now constitute a spectrum that represents the overlapping area between the two original spectra. This data set is integrated to determine the area of overlap. Since the reference and the comparison spectra were normalized to an area of 1.0, the area of the overlap gives a direct indication of the overall similarity between the two spectra, in the form of a fraction. Comparison of identical spectra gives a value of 1.0, and the less similar the spectra, the lower the value.

Results and Discussion

The problems with the correlation coefficient described in the Introduction arise in an example comparing spectra for γ -interferon after rehydration versus that for the native aqueous state. We previously reported a r value of 0.802 (Figure 1A).² However, visually the two overlaid spectra, which have been normalized for total area, appeared to have much greater than 80% similarity (Figure 1B). This disparity can be explained by observing the effect on the r value of offset between spectra that have not been baseline corrected (Figure 1A). In this example, the spectrum for the native protein contains a significant population of positive intensity values, whereas the spectrum for the rehydrated protein contains all negative intensity values. The consequences of this circumstance are apparent by examining eq 1. Having both negative and positive products within the sum in the numerator will result in a lower absolute value than if the products were all positive.

Thus, the correlation coefficient between the two spectra was artificially lowered because the spectra had different baseline positions. The nonzero baseline of a second-derivative

spectrum is a line connecting the two (or more for certain proteins' spectra^{3,6,7}) maxima within the amide I region.^{3,6,7} Two independent spectra seldom have identical baselines due to differences in the physical state of proteins, as well as differences in protein concentration and cell path length. These factors give rise to different absorbance intensity maxima (and minima), which manifest as a different nonzero baseline in the second-derivative spectra. Also, this offset is not detected if the spectra are viewed with the spectral analysis software with automatic maximum expansion of data across the abscissa (i.e., absorbance), as is commonly the case when comparing second-derivative protein spectra.

We initially deduced that for a proper correlation, a simple normalization of the spectral baselines is needed by correcting baseline slope and absorbance to zero (Figure 1B). The r value for the baseline-corrected spectra of γ -interferon is 0.995. However, a visual examination of the area-normalized, overlaid spectra indicates that there is much more than a 0.5% difference between these spectra (Figure 1B); an apparently much larger fraction of the spectral areas do not overlap. This visual impression is due to the redistribution of area among the component bands. Relative to the spectrum for the aqueous control protein, that for the rehydrated protein has less area for the α -helix (1658 cm^{-1}) and turn (1680 cm^{-1}) structures. This loss of area is compensated by the relative increase in area for the band at 1621 cm^{-1} , which is due to the intermolecular β -sheet structure of protein aggregates.³

On the basis of these observations, which reflect redistribution in protein secondary structure, it appears advantageous to compare protein structural similarity by simply determining the overall percentages of secondary structural contents. However, we have previously shown that such comparison can be misleading.³ Often major shifts in bands, which are indicative of large structural changes, do not lead to an overall change in secondary structural content. This is because bands can shift to new positions that represent the same structural types as the original positions. This situation is discussed in detail in our recent review.³

In contrast, determining the redistribution of area, due to band shifts, across the entire amide I range allows a reliable comparison of the overall similarity of second-derivative spectra. Calculating the amount of overlapping area between area-normalized spectra gives a quantitation of the visual impression of spectral similarity. This analysis for the γ -interferon spectra (Figure 1C), gave a value of 0.92, which agrees with the visual impression of similarity.

For other examples of this analysis we compared spectra for chymotrypsinogen A in the aqueous and dried states (Figures 2 and 3). Relative to the aqueous control, the spectrum for dried protein shows significant band shifting. However, as noted in an earlier review,³ there is virtually no difference in relative amounts of secondary structure. Again, the r coefficient returns a value (0.65 for raw spectra) which differs significantly from the area of overlap value of 0.79 (Figure 2). Correcting the spectral baselines leads to a r coefficient of 0.89. When chymotrypsinogen A is lyophilized in the presence of 1 M sodium thiocyanate, two new peaks appear at the edges of the amide I range at 1625 and 1695 cm^{-1} (Figure 3). These bands can be assigned to an intermolecular β -sheet structure found in protein aggregates.³ In this case, both methods of spectral comparison return similar values. An r value of 0.89 is returned, whereas the area of overlap is 0.83.

To gain further insight into the bases for the differences between the two methods of spectral comparisons, we created two artificial data sets. Figure 4 shows two identical data sets offset from each other. As noted above, such offset is a common occurrence with second-derivative infrared protein spectra. These data sets obviously have a relative area of

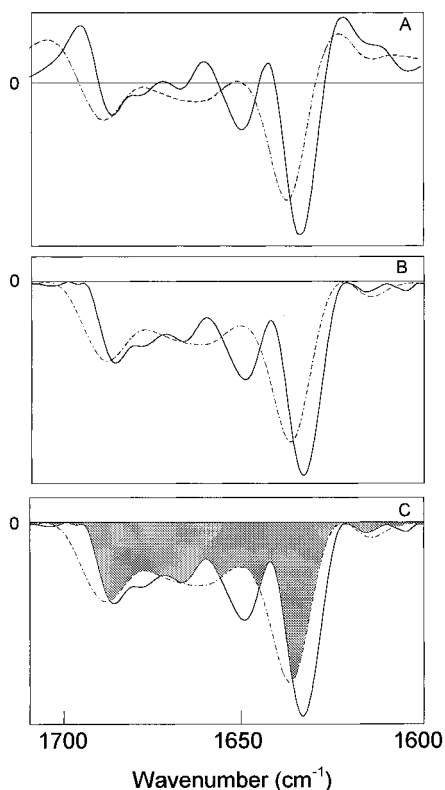


Figure 2—Second derivative amide I spectra of chymotrypsinogen A: (A) raw spectra, $r = 0.65$; (B) baseline offset corrected, $r = 0.89$; and (C) area-normalized spectra, area of overlap = 0.79. Solid lines indicate the native aqueous state, dashed lines indicate the dried state, and the gray fill indicates the area of overlap. Spectra modified from Dong et al.³

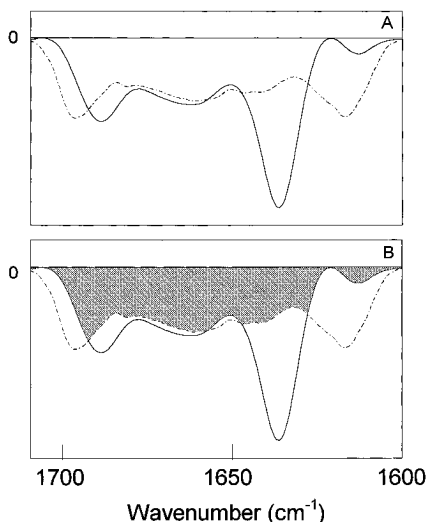


Figure 3—Second derivative amide I spectra of chymotrypsinogen A in the dried state with or without 1 M sodium thiocyanate: (A) baseline offset corrected and area normalized, $r = 0.89$; (B) baseline offset corrected and area normalized, area of overlap = 0.83. Solid lines indicate zero sodium thiocyanate, dashed lines indicate 1 M sodium thiocyanate, and the gray fill indicates the area of overlap. Spectra modified from Dong et al.³

overlap equal to 1.0, but what about the r value? Having both negative and positive products within the sum in the numerator of the r coefficient formula, while generating a sum of the squares of intensities in the denominator, returns an artificially low correlation of 0.47.

As noted in the example of γ -interferon, the r coefficient can also indicate an unrealistically high similarity, when the

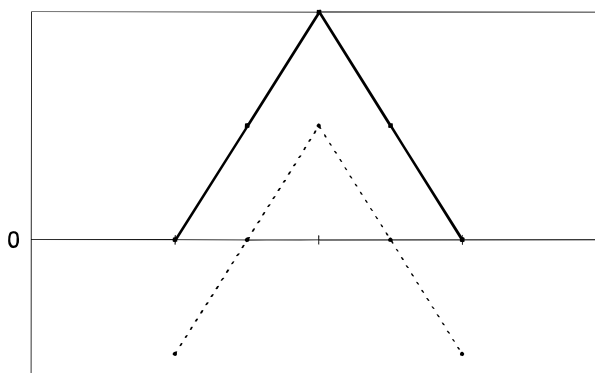


Figure 4—Two identical model "spectra" demonstrating the effect of baseline offset on the r coefficient. The area of overlap gives a value of 1.0, whereas the r coefficient returns a value of 0.47. The solid line indicates the reference, and the dashed line indicates the offset spectrum.

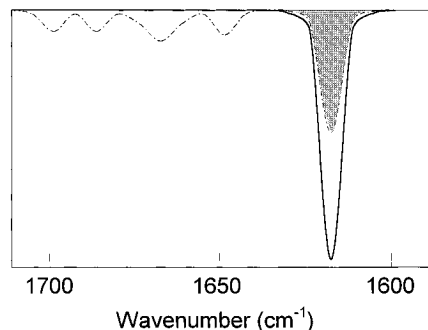


Figure 5—Model "spectra" with a common region of high intensity and symmetry. The area of overlap gives a value of 0.50, whereas the r coefficient returns a value of 0.92.

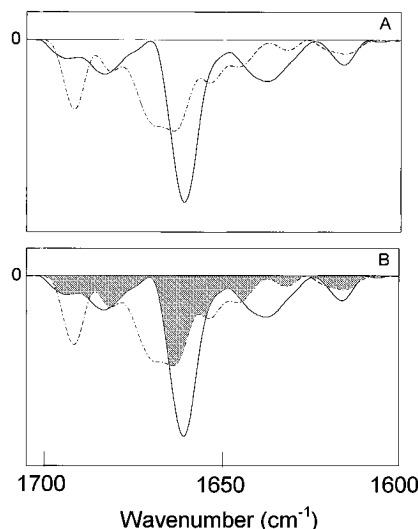


Figure 6—Second derivative amide I spectra of wild type iso-1-cytochrome *c* in aqueous solution in the presence and absence of guanidine hydrochloride: (A) baseline offset corrected and area normalized, $r = 0.72$; (B) baseline offset corrected and area normalized, area of overlap = 0.63. Solid lines indicate zero guanidine hydrochloride, dashed lines indicate 2.5 M guanidine hydrochloride, and the gray fill indicates the area of overlap. Spectra taken from Bowler et al.⁹

offset between spectra is corrected. We have seen this problem most often in spectra with a high degree of symmetry and in which a single band is predominant, as in the cases of proteins with a predominantly α -helix structure. Figure 5 provides a model of this phenomenon. The r value is 0.92 for the two data sets, which is much greater than their actual

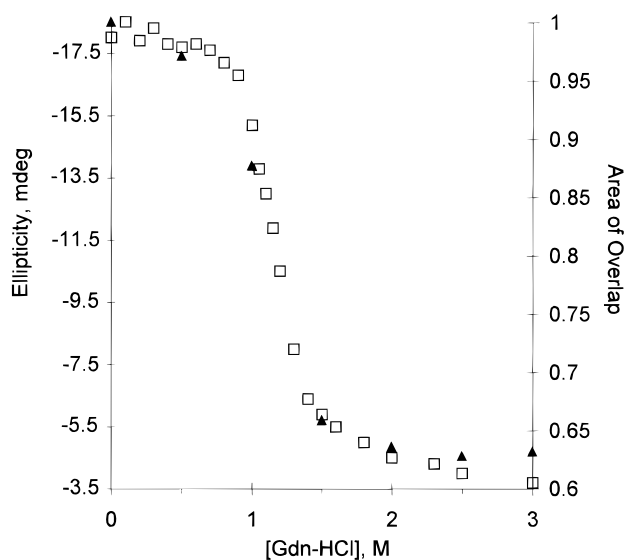


Figure 7—Guanidine hydrochloride-induced unfolding of wild type iso-1-cytochrome *c*. The unfolding transition is presented as a plot of ellipticity at 220 nm in millidegrees (open squares) and infrared spectral area of overlap (closed triangles) as a function of guanidine hydrochloride concentration. Circular dichroism spectroscopic data were taken from Bowler et al.¹⁰ and the infrared spectral area of overlap values were calculated from second-derivative spectra taken from Bowler et al.⁹

similarity. In contrast, the area of overlap provides a more accurate value for similarity of 0.50.

Finally, to ascertain the general utility of the spectral area of overlap parameter for determining relative alterations in protein secondary structure, we examined the second-derivative spectra for iso-1-cytochrome *c*, as a function of guanidine hydrochloride concentration. Previous work, using circular dichroism spectroscopy, defined a guanidine hydrochloride denaturation curve for this protein.¹⁰ In addition, qualitative alterations in the protein secondary structure induced by the denaturant have been characterized with infrared spectroscopy.⁹ In Figure 6A, the second-derivative spectrum for the native protein is compared to that for the protein unfolded in 2.5 M guanidine hydrochloride, a concentration at which the denaturation curve indicated that the protein was fully unfolded.¹⁰ As is often seen with lyophilization-induced unfolding, chemically-induced unfolding leads to large shifts in band positions, widths and relative intensities.⁹ To quantify these alterations, we calculated the area of overlap, which was 0.63 (Figure 6B). Next we calculated this value for the full range of guanidine hydrochloride concentrations at which infrared spectra of the protein had previously been acquired. This analysis resulted in a denaturation curve, which can be superimposed over the previously generated

curve (Figure 7). The similarity between the data sets indicates that infrared spectroscopy, combined with the quantitation of spectral alterations with the area of overlap calculation, is a useful method for quantitating stress-induced alterations in protein secondary structure.

Conclusions

Studies to date have documented that maintenance of natively like structure is required for maximum acute and storage stability of lyophilized protein formulations. Infrared spectroscopy is invaluable in this area, providing a means to investigate protein structure in liquid, frozen, and dried states. Comparisons of the overall similarity of second-derivative spectra in the amide I region, which is sensitive to perturbations in protein secondary structure, is needed to evaluate which formulation produces the most "natively like" protein in the dried solid. The area of overlap method for this comparison offers a straightforward, objective way to quantitate the similarity in the second-derivative amide I spectra and, hence, in overall protein structure.

References and Notes

1. Prestrelski, S. J.; Arakawa, T.; Carpenter, J. F. *Arch. Biochem. Biophys.* **1993**, *303*, 465–473.
2. Prestrelski, S. J.; Tedeschi, N.; Arakawa, T.; Carpenter, J. F. *Biophys. J.* **1993**, *65*, 661–671.
3. Dong, A.; Prestrelski, S. J.; Allison, S. D.; Carpenter, J. F. *J. Pharm. Sci.* **1995**, *84*, 415–424.
4. Prestrelski, S. J.; Pikal, K. A.; Arakawa, T. *Pharm. Res.* **1995**, *12*, 1250–1259.
5. Carpenter, J. F.; Chang, B. S. In *Biotechnology Issues in Pharmaceutical Process Engineering*; Avis, K. E., Wu, V. L., Eds.; Interpharm Press, Buffalo Grove, IL, in press.
6. Dong, A.; Caughey, W. S. *Methods Enzymol.* **1994**, *232*, 139–175.
7. Dong, A.; Huang, P.; Caughey, W. S. *Biochemistry* **1990**, *29*, 3303–3308.
8. Dong, A.; Huang, P.; Caughey, W. S. *Biochemistry* **1992**, *31*, 182–189.
9. Bowler, B. E.; Dong, A.; Caughey, W. S. *Biochemistry* **1994**, *33*, 2402–2408.
10. Bowler, B. E.; May, K.; Zaragoza, T.; York, P.; Dong, A.; Caughey, W. S. *Biochemistry* **1993**, *32*, 183–190.

Acknowledgments

We thank two anonymous reviewers for their many helpful suggestions. We gratefully acknowledge support from the Office of Naval Research (Grant N00014-94-1-0402), the National Science Foundation (Grants BES9505301 and BES9529288), the Whitaker Foundation and the American Foundation for Pharmaceutical Education. We thank Amgen, Inc., for the gift of γ -interferon and Mr. Ben Garland of Nicolet Instruments for the custom OMNIC subroutine for calculation of the protein correlation coefficient.

JS950332F