Separation of Deep UV Resonance Raman Spectra for Pure Protein Secondary Structures based on D-H Exchange Data

V. Shashilov
Department of Chemistry
University at Albany

Abstract. This project elaborates a Bayesian-based approach for extracting resonance Raman spectra of highly-ordered β-sheet structure of amyloid fibrils. The proposed algorithm incorporates prior information about characteristic spectral bands using the signal dictionary approach and information about the concentration matrix by searching over the space of template mixing matrices. Upon further improvement, the algorithm can be specifically used for extracting spectra of species present in small fractions in IR, Raman, NMR, and MS spectral mixtures.

Introduction. A chain of protein molecules can adopt three different conformations or secondary structures called random coil, α-helix, and β-sheet. Each protein can be thought of as a combination of these three secondary structures. Moreover, the dataset of deep UV resonance Raman spectra of hundreds different proteins can be fitted with three component spectra. This suggests that three pure secondary structure spectra are the same for most proteins. Knowledge of these three component spectra would allow for finding the percentage of the secondary structures in a protein by the least-squares fitting of its Raman spectrum with those three. The pure secondary spectra are not observable experimentally since no protein consists of 100 % of particular secondary structure. To tackle this difficulty we attempted to extract the latent pure component spectra using the Bayesian curve resolution approach.

Experimental Part. Amyloid fibrils are the specific form of protein composed of the highly ordered β-sheet core surrounded by random coil part. The contribution of the third component, i.e. α-helix is negligible and can be disregarded. **Figure 1** shows the structure of amyloid fibrils. Random coil parts are exposed to water while β-sheet of fibrils is knows to be buried and inaccessible to the aqueous solution. If one replaces water by deuterium water H-s of the random coil will be substituted by D-s resulting in down-shift of all Raman bands involving N-H vibrations. The N-H vibration bands of buried β-sheet will not affected by deuterium exchange and therefore will retain their positions. This allows separating β-sheet and random coil structures gradually changing H_2O/D_2O ratio in solution. Fifty samples of fibrils were prepared starting with fibrils in 100% H_2O, 98% H_2O plus 2% D_2O, 96% H_2O plus 4% D_2O, …. and 100% D_2O to acquire 50 Raman spectra. Spectra of H_20, D_20, 50/50 (H_2O + D_2O) mixtures were recorded separately.

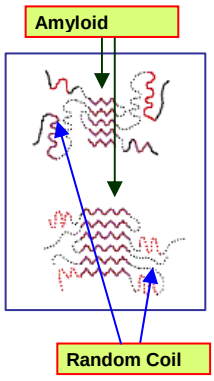


Figure 1. Structure of amyloid fibrils**.** Core β-sheet is shown with saw-shaped purple lines

Theoretical Part.

The chosen experimental procedure permitted the prediction of the contribution of species contributing to spectra of fibril.

(a) *Anticipating H_2O, D_2O, HOD contributions in each sample.*

H-s and D-s from H_2O and D_2O molecules readily interchange to form mixed HOD molecules. If the total fraction of protons in the H_2O-D_2O is q then the probabilities of forming H_2O, D_2O, HOD are as follows:

Probability	Possible Combinations	Statistical Weight
$P(H_2O \mid q, I) \sim q^2$	H-O-H	1
$P(D_2O \mid q, I) \sim (1-q)^2$	D-O-D	1
$P(DOH \mid q,I) \sim 2*(1-q)*q$	D-O-H, H-O-D	2

Simulated concentration fractions of H_2O, D_2O, HOD versus the fraction of added D_2O are shown in **Figure 2.**

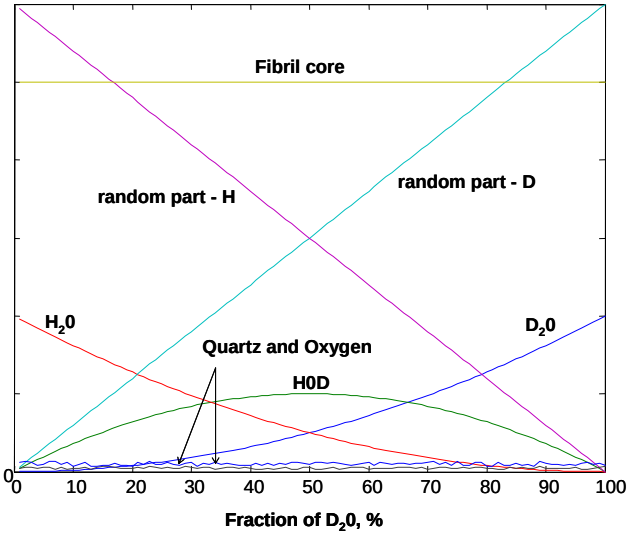


Figure 2. Anticipated concentrations all components.

(b) *Anticipating random coil(H) and random coil(D) contributions in each sample.*

Fractions of N-H bonds and N-H bonds in the random coil part are proportion to the total concentration of H-s and D-s, respectively, i.e. they follow linearly the fraction of added D_2O (**Figure 2**).

(*c) Anticipating contribution of the other components.* Each fibril sample was prepared from the same stock fibril material. This implies that the fraction of the β-sheet core with respect to the random coil part is constant across all samples, *i.e.*

$$Frac(\beta\text{-sheet}) / (Frac(coil(H)) + Frac(coil(D))) \sim const$$

All spectra exhibit the admixture of quartz signal (spectra were recorded in a quartz tube) and atmospheric oxygen. Both fractions are random (**Figure 2**).

Table 1 below summarizes available prior information about all the components

Table 1. Prior information on the spectra and concentrations of components.

Pure component	**Spectrum**	**Concentration profile**
H_20	Known	Shape is known
H0D	Known	Shape is known
D20	Known	Shape is known
Unordered part, H substituted	Unknown	Shape is known
Unordered part, H substituted	Unknown	Shape is known
Fibril core	Unknown	Shape is known
Quartz	Known	Small random contribution
Oxygen (molecular)	Known	Small random contribution

Bayesian Approach. The source separation problem is as follows

$$Data = C \cdot S + E \tag{1}$$

Where C is the concentration matrix, S is the matrix of pure component spectra and E is error where random or systematic. The matrix *Data* is known while the matrices C and S are to be estimated. The Bayes theorem for problem (1) is written as

$$P(C,S \mid Data, I) \sim P(Data \mid C,S,I) \cdot P(C \mid I) \cdot P(S \mid I) \tag{2}$$

where $P(Data \mid C,S,I)$ is the likelihood controlling the quality of fitting and $P(S \mid I)$ and $P(C \mid I)$ are prior probabilities for spectra and concentrations. Because finding either matrix C or S alone is enough for solving problem (1) the concentration matrix C is normally sought since it contains by far fewer elements. It was shown[1] that in the case of uniform prior for concentration matrix and independent sources the probability of the concentration matrix is given by

$$P(C \mid Data, I) \sim \int ds \cdot \prod_i \delta(Data_i - C_{ik} \cdot S_k) \cdot \prod_l p_l(s_l) \tag{3}$$

which in the case of noise-free data reduces to the logarithmic probability

$$P(C \mid Data, I) = \log(\det(W)) + \sum_l \log(p_l(s_l)) \tag{4}$$

where W is the separation matrix such that $S = W \cdot Data$.

Incorporating prior information about the C matrix. The actual concentration matrix should have columns proportional to the columns of the matrix sketched in **Figure 2.** Eight hyper-parameters α_i then need to be estimated so that $Cj = \alpha_i \cdot T_{ij}$ with j=1:8 and T is the template matrix shown in **Figure 1**. The parameters α_i account for the spectral fraction of each component in the mixtures (they are proportional but not equal to the physical concentrations and therefore must be found). For example, e.g. α_1 / α_6 gives the spectral fraction of fibril core with respect to that of water and is proportional to the

concentration of protein in samples. The concentration of protein is assumed to be equal in all samples since they were prepared based on the same stock solution.
As seen from **Figure 2**, contributions of quartz and oxygen are random and to be rigorous, 50 additional parameters need to be assigned for fraction of quartz in each sample and another 50 for oxygen. These are the nuisance parameters in the model. At this stage, however, we assume constant but unknown fractions of quartz and oxygen in each. After all relevant parameters are found matrix least-squares will refine those small quartz and oxygen contribution in each sample. The posterior for the concentration matrix then takes the form

$$P(C \mid Data, I) = \log(\det(W)) + \sum_l \log(p_l(s_l)) - \frac{m \cdot n}{2} \cdot \log\{\| (C - T \cdot \alpha) \|\} \qquad (5)$$

where α is diagonal matrix with parameters α_i on its diagonal and $\| \ \|$ stands for Frobenius norm, *m* is a number of experimental (m=50) spectra and *n* is a number of pure components (n=8).

Incorporating prior information about the S matrix.

In our model we have 5 pure spectra known from the experiment. It is straightforward to assign inner product of the known spectrum and the resolved spectrum for known components as their prior probabilities. Indeed, inner product equals 1 if spectra completely overlap and 0 if they have no overlapping regions. For the other three spectra $P(s_{ij})$ was set proportional to the reciprocal of s_{ij}.
$P(s_{ij}) \sim 1/s_{ij}$. The total posterior probability then transforms into **(6)**

$$P(C \mid Data, I) = \log(\det(W)) + \sum_l \log(p_l(s_l)) - \frac{m \cdot n}{2} \cdot \log\{\|(C - T \cdot \alpha)\|\} + T \cdot (\sum_{i=1}^{k1} \log(inner(s_i, ref\_s_i)) - \sum_{k1}^{m} \log(s_i)$$

As seen from (6) assumption $P(s_{ij}) \sim 1/s_{ij}$ for unknown spectra resulted in addition of

$-\sum \log(s_i)$ known as a sparsity constraint. It controls the area of resolved spectra and eliminates extraneous bands appearing as admixtures from the other spectra.

It turned out in the course of optimization that resolved spectra had characteristic bands whose shapes were close to what we expected for these components. The space in between those bands contained admixtures from other components and /or noise. Alternatively, the characteristic bands in the sought spectra were obtained using pure variable approach[2]. The latter analyzes second derivative or fourth derivative spectra where even very overlapping bands are seen as distinct sharp peaks. The purest variable is such a wavenumber at which the contribution of an individual component to the Raman intensity is maximal while the contributions from the other components are minimal. For a Raman spectrum of each sample, the intensity at a particular purest variable is approximated to be proportional to the concentration of a corresponding individual component in the sample. Consequently, the matrix of the Raman intensities at all purest variables C_{int} can be used as a concentration matrix C of the components. The shapes of spectral band in normal (not second derivative) space are then found as

$$S = Data^T C_{int} \, (C_{int}^T C_{int})^{-1} \qquad (7)$$

Knowledge of characteristic bands allows modeling the unknown spectrum as a linear combination of the non-overlapping bands which are referred to as a dictionary bands[3].

A model dictionary spectrum is used in a similar fashion as known experimental spectra. The spectrum resolved at each iteration of the algorithm is compared to the dictionary spectrum by means of inner product. The space in between known bands in a dictionary spectrum is set equal to 0. This means that any band of the resolved spectrum in the region one is not sure about will give zero contribution to the probability for this spectrum (they multiply by zeros in a dictionary spectrum) thus assigning equal probabilities to any spectral features in doubtful regions. Coefficients for contributions of various dictionary bands to model spectra were sought as additional parameters over the course of optimization.

Sampling method. The floating point genetic algorithm (GA)[4] was used as a sampling methods. The GA thoroughly explores the parameter space thus getting around the local maxima and allows strict setting the prior range for optimized parameters

Accelerating the convergence. Augmented *Data* matrix and *C* matrix were constructed at the initial stage of fitting. Namely, five known pure component spectra were put on the top of the *Data* matrix and five rows were added to the *C* matrix with units on the diagonal and zeros otherwise. Thus the algorithm was forced to fit the augmented matrix using five augmenting spectra as five of eight components.

Algorithm outline.

(1) Parameters α_i are initialized, ranges of the parameters are constrained to the expected range.
(2) Reduce dimension of *Data* from *m*=50 to *n*=8 by SVD to produce matrix D_{trunc}.
(3) For *i=1: itmax* do

I. Sample parameters α

(a) Calculate trial *C* matrix as $C_{ij}=T_{ij}*diag(\alpha_j^{conc})$, i.e. by multiplying columns of T by respective parameters, *T* is a template matrix.
(b) Calculate dictionary spectra as $S_{dic}= \alpha_j^{dict} \cdot Band_j$
(c) Add dictionary spectra multiplied by the small parameter λ on the top of the augmented matrix (to speed up convergence), add rows with units on the diagonal and zeros otherwise to the augmented concentration matrix.
(d) Calculate resolved spectra by matrix least-squares using ' \ ' Matlab operator

$$S = Data_{aug}{}^{T} C\,(\,C_{aug}{}^{T} C_{aug}\,)^{-1}$$ set $S(S<0)$=tol, tol>0 & tol<<1

(e) Given spectra *S* calculate the refined concentration matrix *Conc*

$$Conc = \left(S^T \cdot S\right)^{-1} \cdot S^T \cdot Data^T$$

using ' \ ' Matlab operator for matrix pseudo-inverse

(e) Calculate posterior as P(C,S | Data, I)=P(C | Data, I) (eq. 6) $\frac{(m \cdot T)}{2} \cdot \log \| Data - Conc \cdot S \|$

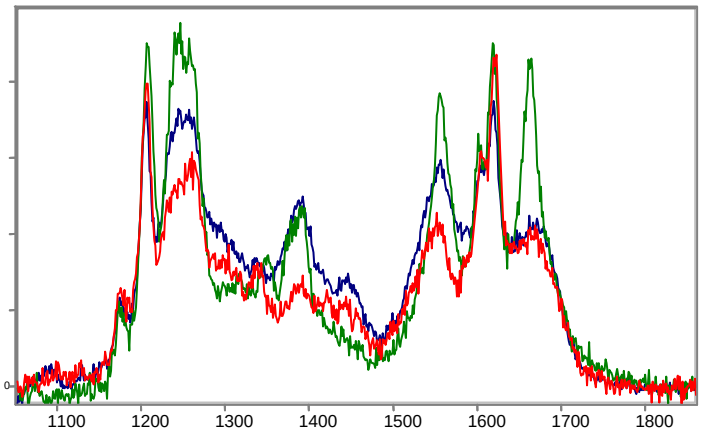
Separation matrix *W* is calculated as $W = \left(Data_{trunc}{}^{T} \cdot Data_{trunc}\right)^{-1} \cdot Data_{trunc}{}^{T} \cdot S$

II. Sample a new set of parameters α and go through (a)->(e)

(4) End of the algorithm.

What made the problem difficult.

(1) The spectra of pure secondary structures of protein are highly overlapping

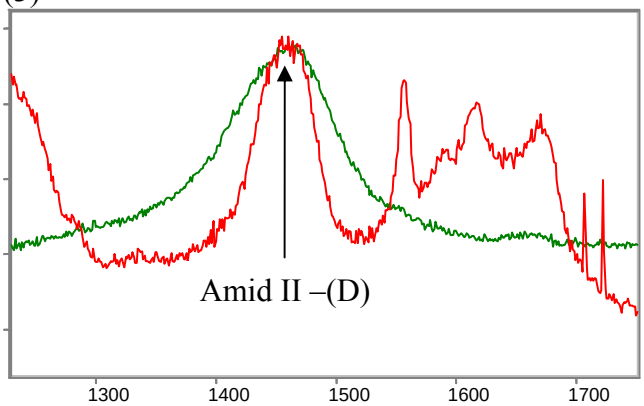


Raman shift / wavenumbers

This figure shows tentative pure secondary structure spectra (α-helix in red, β-sheet in green and random coil in blue). These spectra were measured on model polypeptides containing mostly single secondary structure.

(2) Spectra of H_20,D_20 and HOD are highly overlap with the most characteristic bands of protein.

(3)

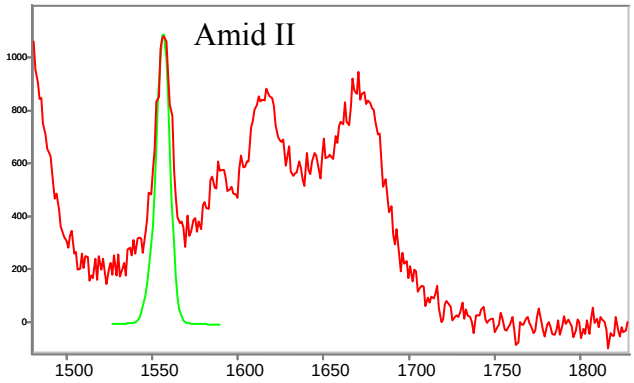


Raman shift / wavenumbers

This figure shows the overlap of Amid II –(D) band of protein (red) and peak of HOD molecule (green). Red spectrum is taken in 100 % D_20 so it has no HOD contribution. The problem is aggravated by correlation in appearance of HOD and Amid II-(D) bands across the data set.

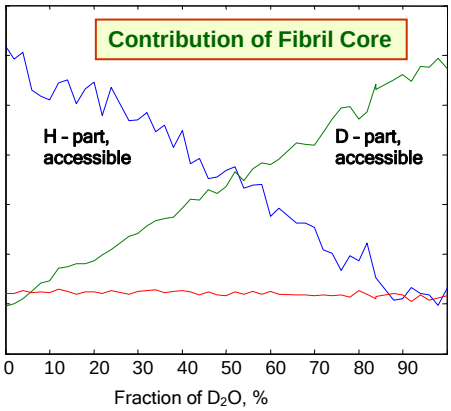
(4) Contribution of the oxygen line is the major problem. It is sharp and completely overlaps with the most informative Amide II band.

The figure below shows overlap of Amide II band (red) of β-sheet with the oxygen band (green). Relative contribution of oxygen is chaotic and starts dominating the β-sheet spectrum as β-sheet concentration drops.

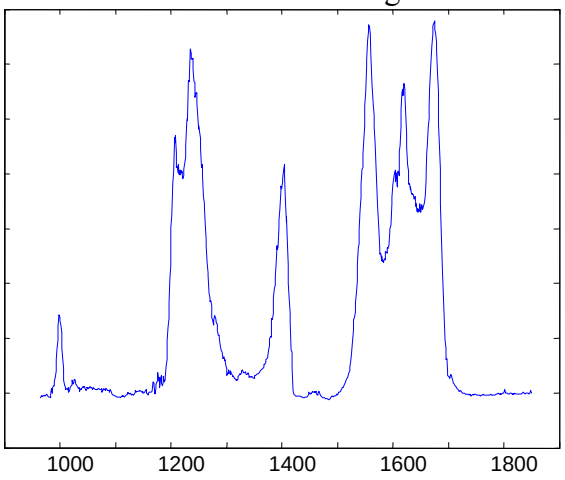


Raman shift / wavenumbers

(5) Small contribution of the compound under study (β-sheet fibrils, shown in red).



Results and Conclusions. The proposed Bayes curve resolution method allowed extracting pure spectra of the highly ordered β-sheet core although its concentration fraction was about 5 %. The figure below shows the reconstructed spectrum



On the contrary to what we expected, majority of core β-sheet was affected by deuterium exchange. Therefore, the other two resolved spectra of random coil(H) and random coil(D) were in fact mixtures of random coil and β-sheet because contributions of both changed synchronously as more D_2O was added.

Plans. A more rigorous mathematical expression for posterior probability is needed. Specifically, the likelihood and prior probability parts of the objective function must be properly weighted.

References.

[1] Knuth K. Bayesian Source Separation and Localization. SPIE'98 Proceedings: Bayesian Inference for Inverse Problems, SPIE San Diego, July, 1998. pp. 147-58.

[2] Windig W, Gallagher NB, Shaver JM, Wise BM. A new approach for interactive self-modeling mixture analysis. Chemometrics and Intelligent Laboratory Systems 2005;77:85-96.

[3] Zibulevsky M, Pearlmutter BA. Blind Source Separation by Sparse Decomposition in a Signal Dictionary Neural Computation 2001;13:863-82.

[4] Elliott L, Ingham DB, Kyneb AG, Mera NS, Pourkashanian M, Wilson CW. Genetic algorithms for optimisation of chemical kinetics reaction mechanisms. Progress in Energy and Combustion Science 2004;30:297–328.