Bayesian DOSY:
a New Approach to Diffusion Data Processing

Stanislav Sykora, Extra Byte, Italy
www.ebyte.it

Juan Carlos Cobas Gómez, Mestrelab, Spain
www.mestrelab.com

BayDOSY

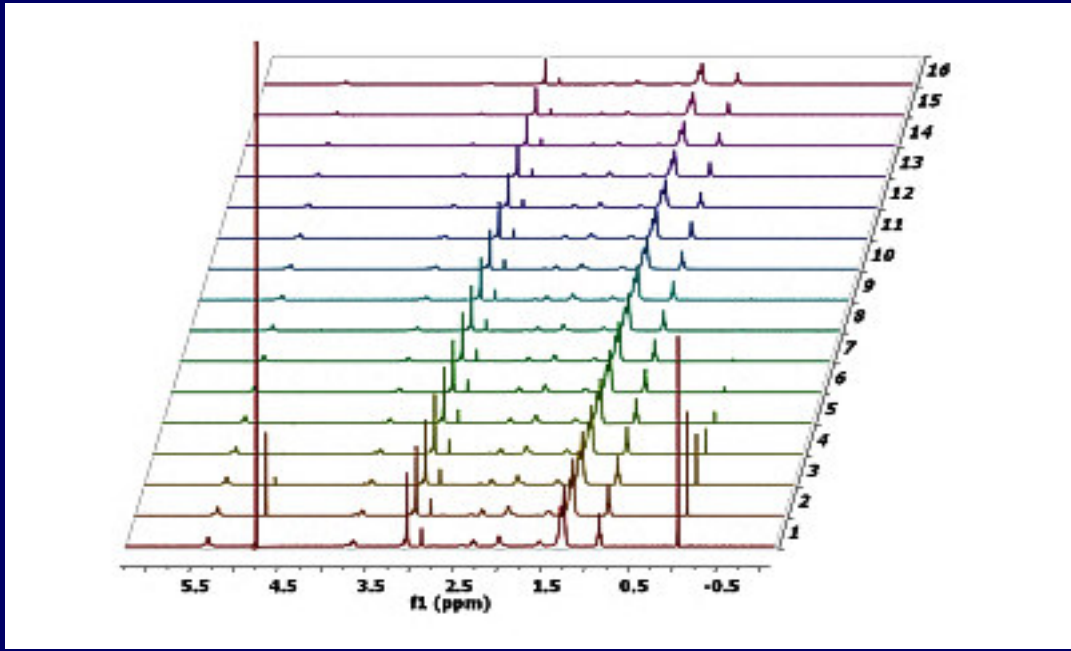
The abstract and slides of this talk are available at
www.ebyte.it/stan/Talk_Valtice_2008.html

Presented at XXXVIII **GIDRM**, September 10-13, 2008, Bressanone/Brixen, Italy

DOSY: Diffusion Ordered Spectroscopy

You certainly all know the principle of these arrayed-parameter experiments.

The raw data look somewhat like this:



The arrayed-parameter is always a field-gradient pulse property, such as amplitude (G) or duration (δ) and the number of settings is user-defined

Each setting gives one 1D spectrum

First transformation: from sequence-specific parameter settings to a universal decay variable

For simple Stejskal-Tanner sequence:

- Original decay formula:

$$S_i(d,T_{2i},\Delta,\delta,g) = S_{i0}E(\Delta,T_{2i})\exp[-d(\gamma\delta g)^2(\Delta-\delta/3)]$$

- Decay variable definition:

$$z = (\gamma\delta g)^2(\Delta-\delta/3)$$

- New decay formula in terms of d and z:

$$S_i(d,z) = A_i \exp(-dz)$$

The transformed data look the same as before, only the labels along the vertical axis change.

The advantage is that the new data set S(f,z) is independent of acquisition details, such as which pulse sequence was used or which parameter was arrayed.

All this is today elementary and the formulas are well-known.

Every DOSY software does this step in the same way

Presented at XXXVIII **GIDRM**, September 10-13, 2008, Bressanone/Brixen, Italy

The goals of DOSY

Different application areas may have different goals.

- *In Low-Resolution NMR (single line), the final goal may be simply just the* measurement of a unique diffusion coefficient D.
- *Or, still in LR-NMR, it could be a continuous Inverse Laplace Transform (ILT) providing* the distribution of D's *in a complex sample. This requires many different z-value settings (typically >128) and long measurement times.*
- *In High-Resolution NMR (spectroscopy), the reasearcher is usually a chemist and his system is composed of a* limited number of molecules *with a* discrete set of D's *(prior knowledge!). What he wants is a* fast separation *of the spectra of individual molecules according to their* D *values. He often does not care much about what the* D *values are but he is eternally concerned with* speed. *32 z-value settings are often the affordable maximum, so forget about ILT. And he confides that spreading the spectra along the frequency axis he will* avoid much of the overlap *between signals belonging to different molecules.*

What we are going to takle here is the latter context.

An overly simple-minded approach to DOSY data evaluation

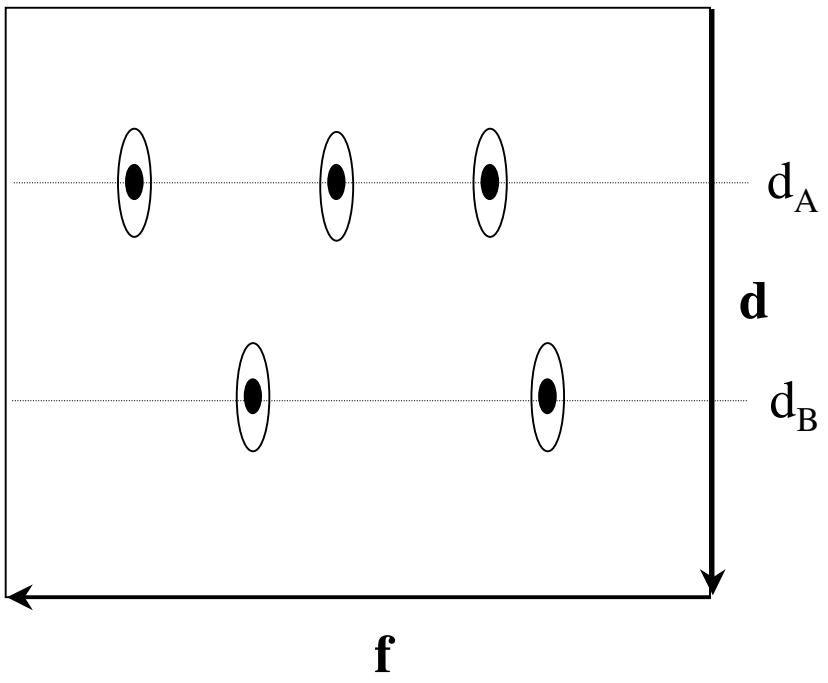
One could proceed as follows:

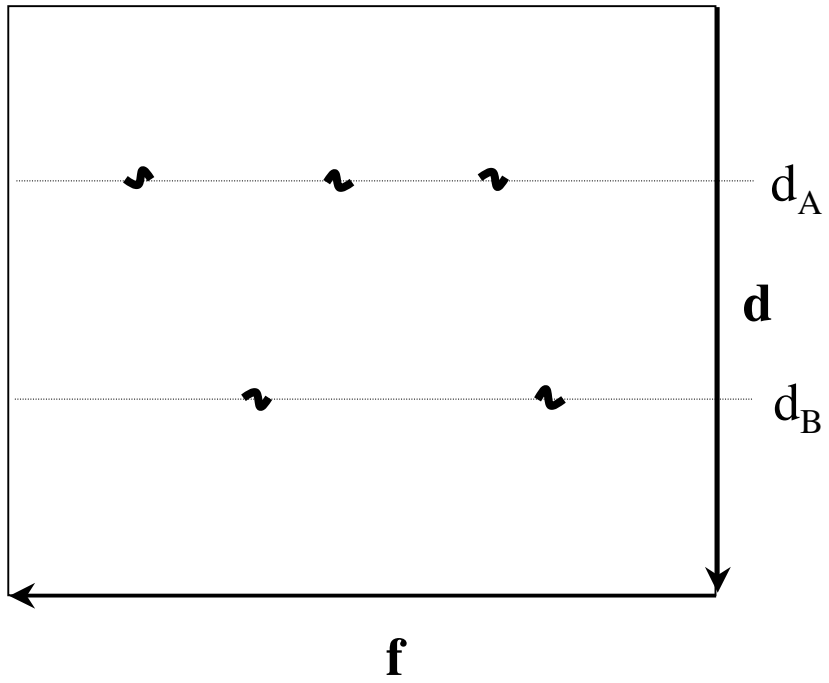
- Apply a peak-picking routine
- Select one peak at a time
- Fit its decay to an exponential to estimate the value of its diffusion coefficient and its confidence interval
- Tabulate the result for each peak
- Group the peaks into classes with similar diffusion coefficients
- Identify the classes with molecules

What is wrong with this:

- Some points along the f-axis are treated differently then others; actually, most are not treated at all.
- Unresolved and overlapping peaks represent a big problem
- The tabulated data are difficult to sort and handle manually
- The separation into classes is itself a non-trivial statistical problem
- The approach does not provide the single-glance grasp of the results

Various enhancements of the simple approach

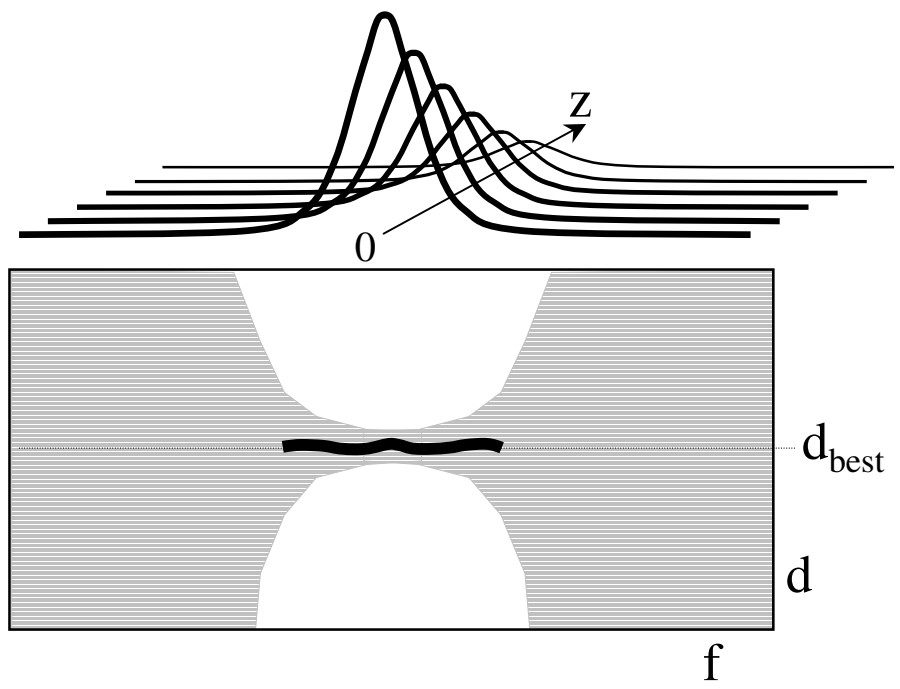




As a step towards the single-glance grasp concept, the tabulated D's can be plotted in a kind of 2D DOSY map with the spectral frequency f along the horizontal axis and the diffusion coefficient d along the vertical axis – complete with error bars/circles and other artifices.

Actually, these are no 2D plots at all – just maps representing the tabulated data.
Most points (f,d) in these graphs have no or misleading physical meaning!

The impass of all fitting methods



Points (f,d) above and below the fitted d_{best} are used to show error bars or artificial Gaussian peaks (like in DISCRETE).

But all those properties actually regard the point (f,d_{best}), not (f,d)!

Confidence intervals (the gray area) in the d-dimension are drastically dependent upon signal intensity and it is not clear whwther and how to represent them.

The butterfly artifact

Are there no physically sound methods ?

YES, there are!

A very good one is the Maximum Entropy approach of

Marc-Andrè Delsuc

It is only a pity that it is very, very slow.

An alternative: the Bayesian approach

Bayesian methods are often conceptually similar to MEM but much faster

- **Basic principle:** We do not fit the parameters of a model to the experimental data, but rather estimate for each point in the parameters space its statistical compatibility with all available raw data points.
- In our case, the parameters space is composed of all pairs (f,d) which constitute the points of a DOSY map. For numerical reasons, of course, we restrict the exploration to a pre-defined grid of points.
- We treat each point of the DOSY map and each point of the raw data in the same way, with no preference for peaks or other special spectral features.

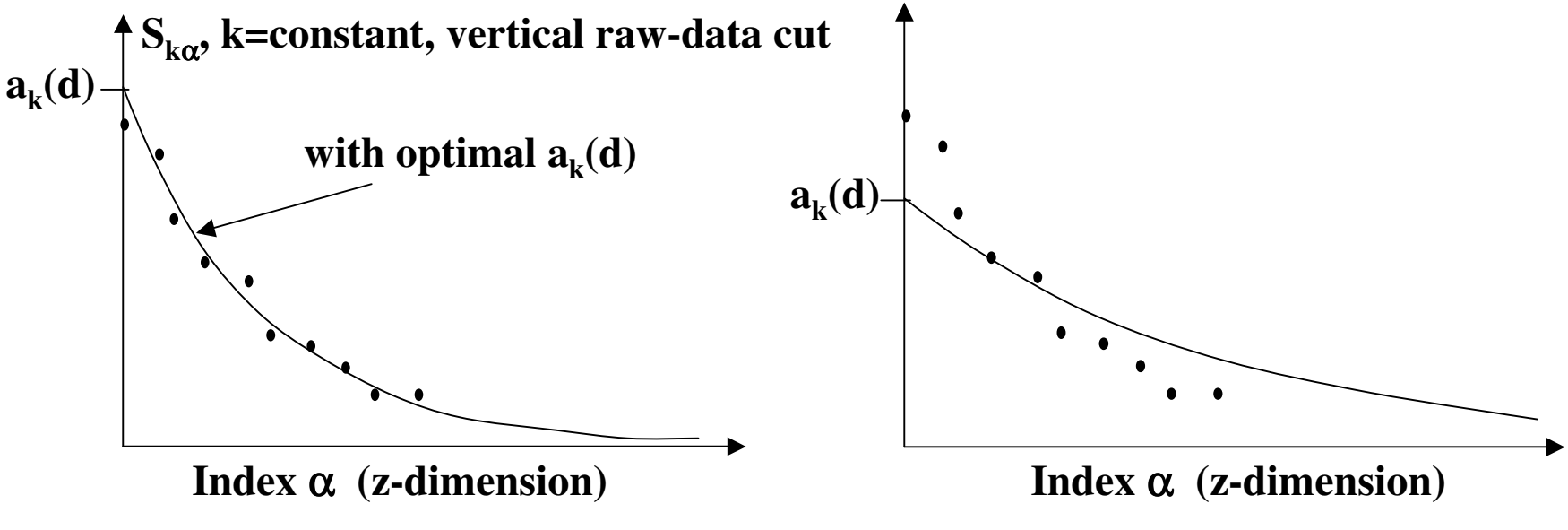
How is it done in practice

Having fixed f and d, one has only one parameter to optimize, namely the decay-curve amplitude $a_k(d)$, k being the data index corresponding to the frequency f. This, however, is known explicitely. Assuming the decay curve

$$y_{k\alpha} = a_k(d) \exp(-dz_\alpha), \text{ where } \alpha \text{ is the index within the z-set,}$$

the value of a_k which minimizes the total square deviation $\Sigma_\alpha (y_{k\alpha} - S_{k\alpha})^2$ is

$$a_k(d) = [\Sigma_\alpha S_{k\alpha} \exp(-dz_\alpha)] / [\Sigma_\alpha \exp(-2dz_\alpha)].$$



Assigning the Bayesian weight (probability) to each point in the [f,d] map

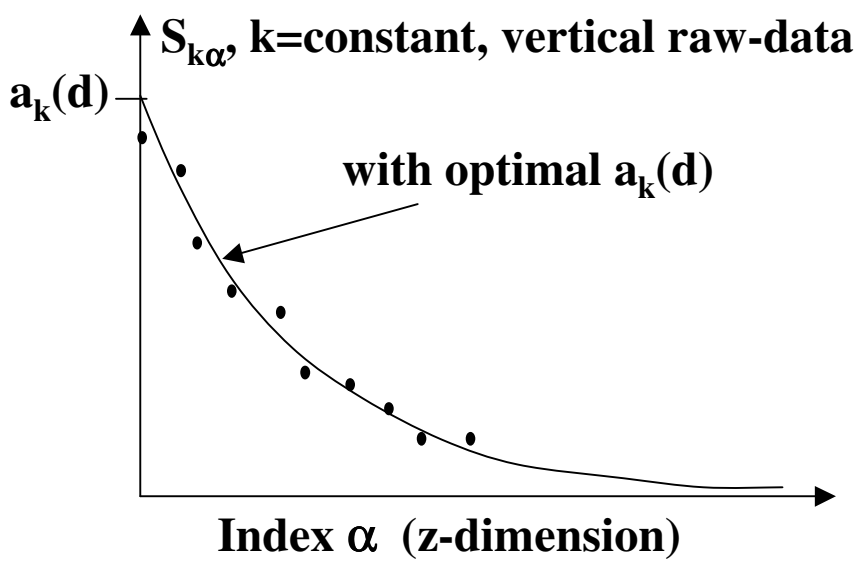
If σ is the standard deviation of the noise, the multiplicative contribution of each raw data point $S_{k\alpha}$ to w_k is the simple Gaussian $\exp(-(y_{k\alpha} - S_{k\alpha})^2/(2\sigma^2))$. Considering all points along the vertical raw data cut at frequency f, we have

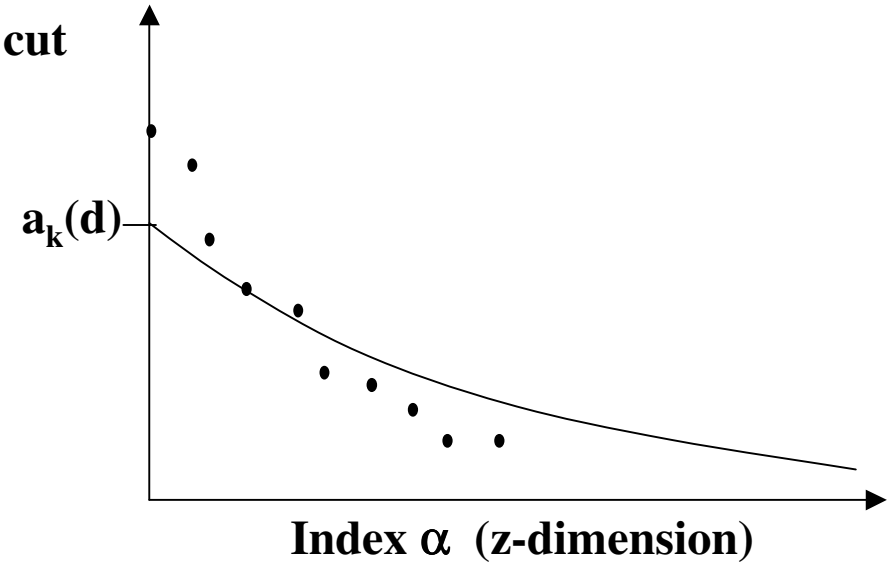
$$w_k = \Pi_\alpha \exp(-(y_{k\alpha} - S_{k\alpha})^2/\sigma^2) = \exp(-[\Sigma_\alpha (y_{k\alpha} - S_{k\alpha})^2]/\sigma^2),$$

whose maximum $w_k(d)$ coincides with the optimal $a_k(d)$. Notice that, while any d is legitimate, the value of $w_k(d)$ is appreciable only when d is close to the 'correct' value, or when there is no signal (zero $a_k(d)$).

The value of $w_k(d)$ is the proper 'vertical' value to be assigned to the (f,d) point in the 2D DOSY plot, pending a final normalization.

… in other words



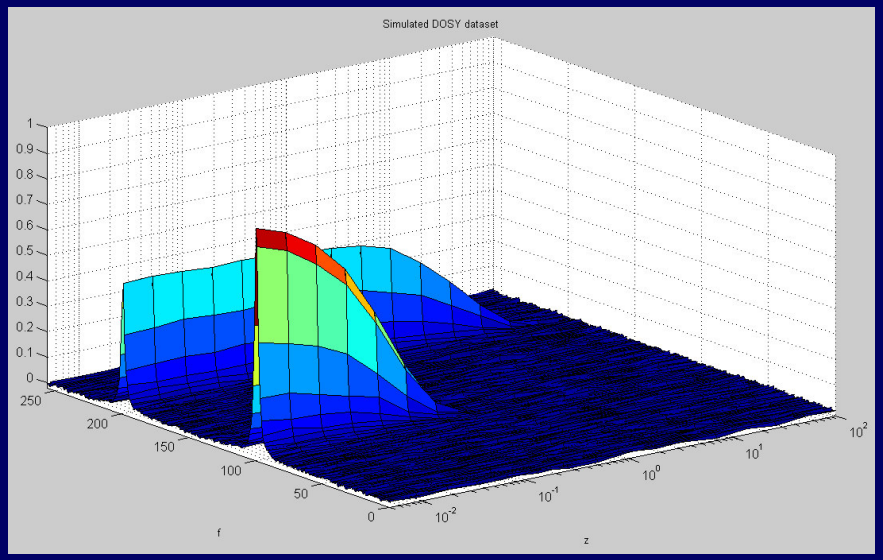


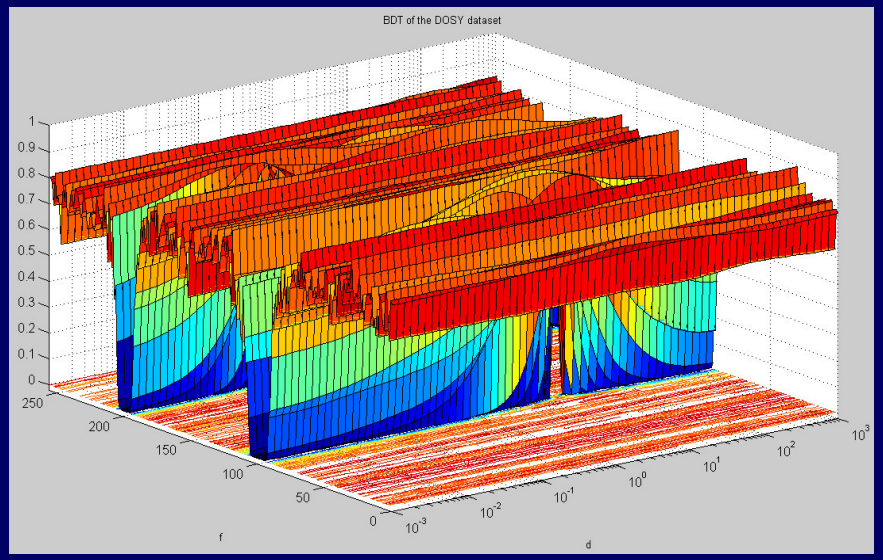
$w_k(d) \equiv w(f,d)$ large

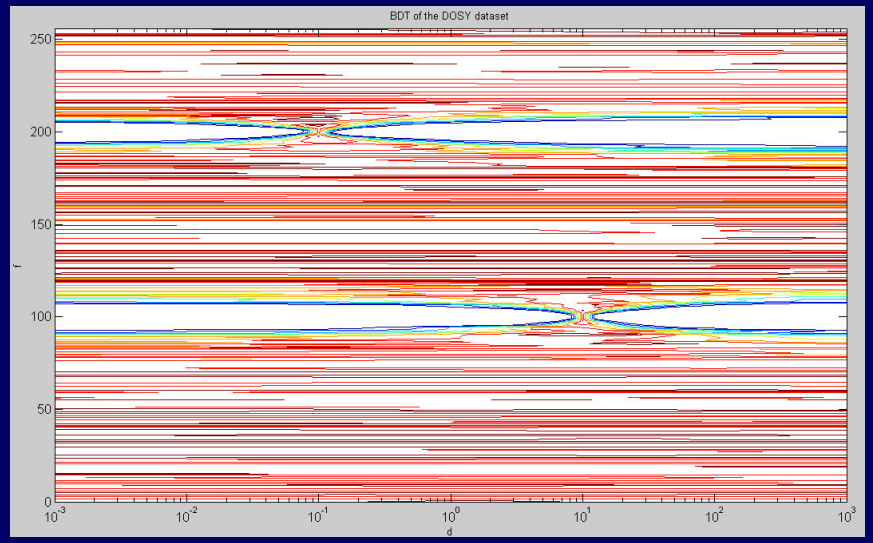
$w_k(d) \equiv w(f,d)$ tiny

The value of $w_k(d)$ is the proper 'vertical' value to be assigned to the (f,d) point in the 2D DOSY plot, pending a final normalization.

Matlab tests of the algorithm: the w(f,d) maps

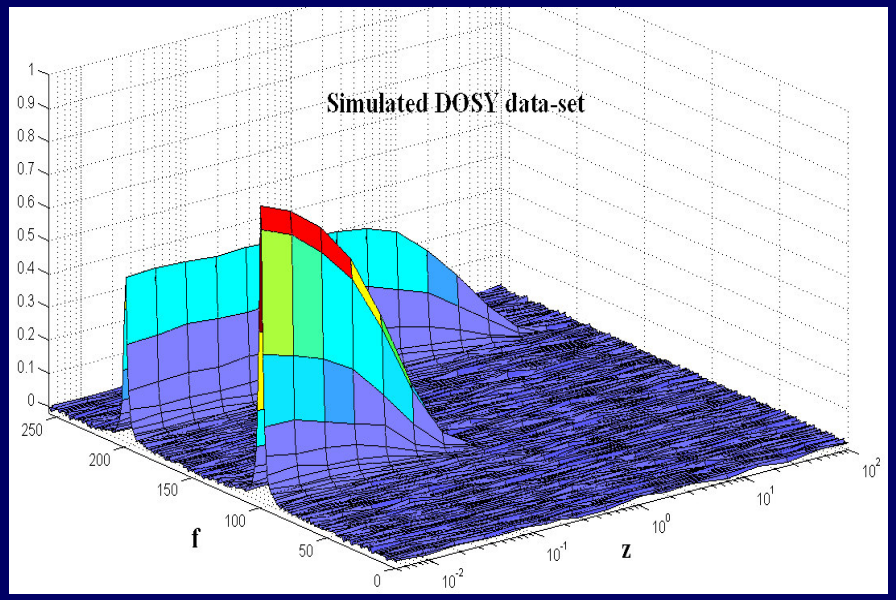


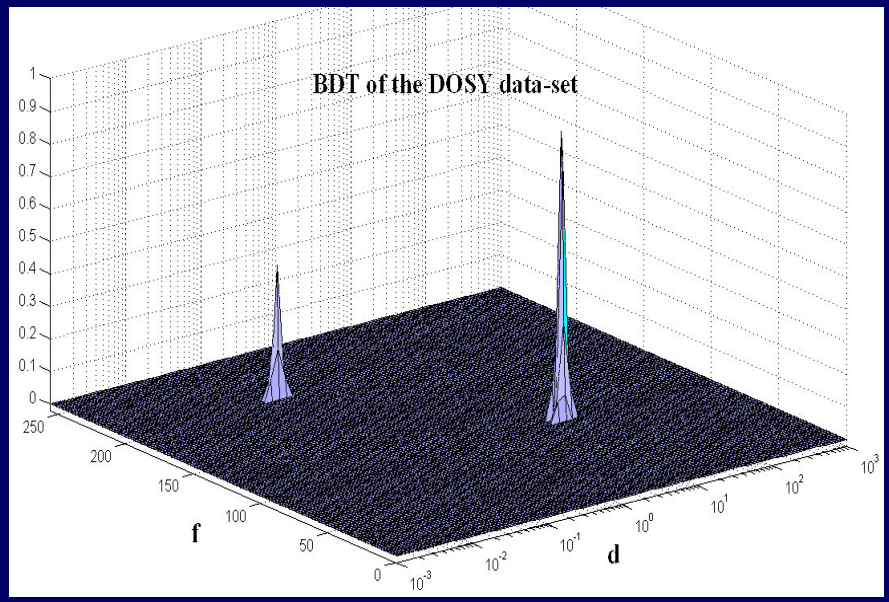




Normalization according to the total signal intensity

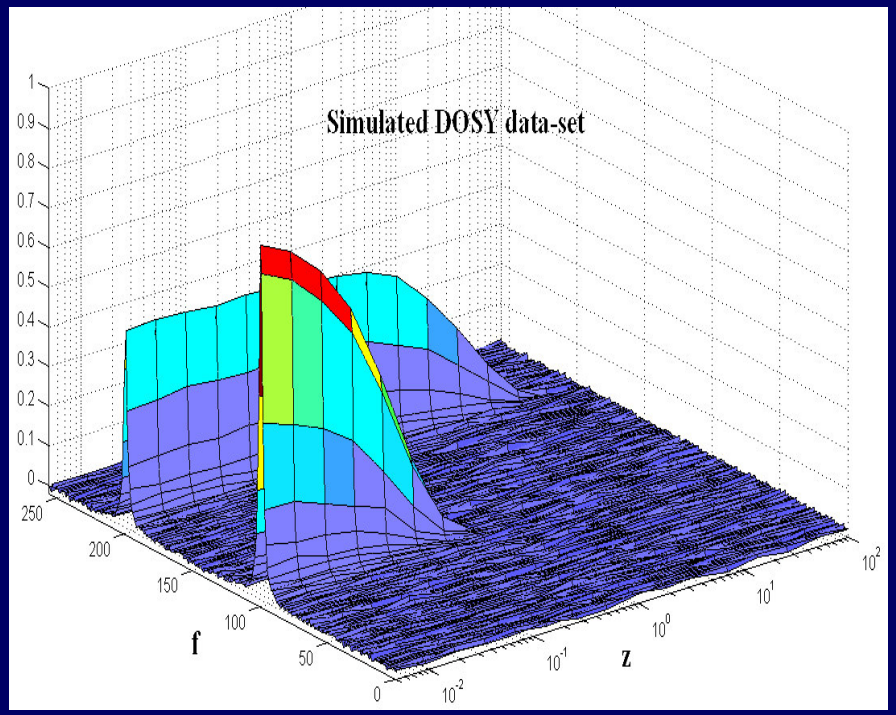
We have not yet exploited the signal intensity, interpretable as an a-priori knowledge, and a transition probability. When we do so:



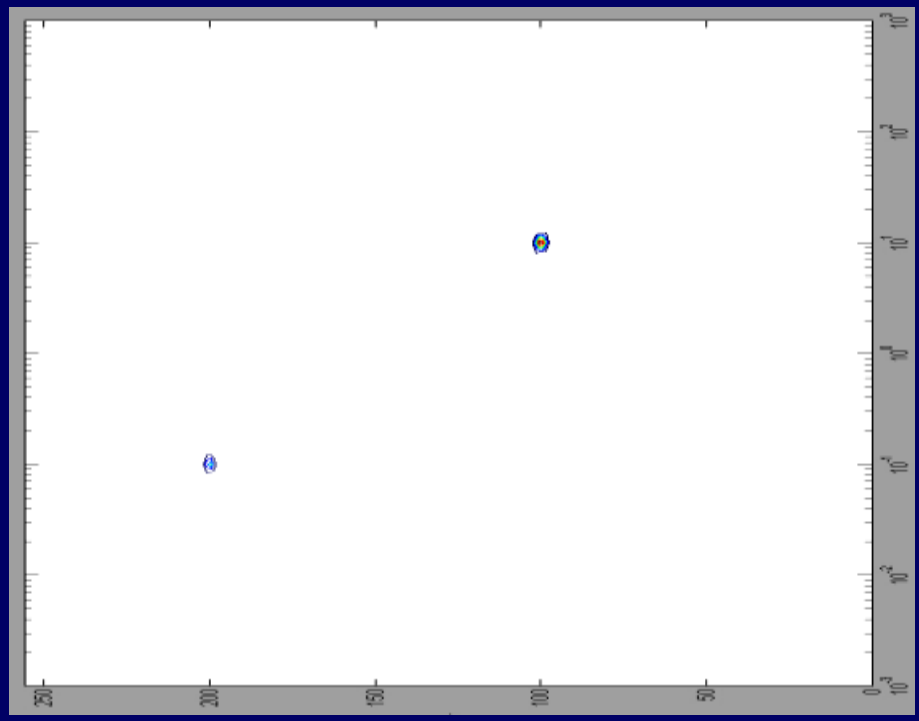


Bayesian DOSY Transform is born

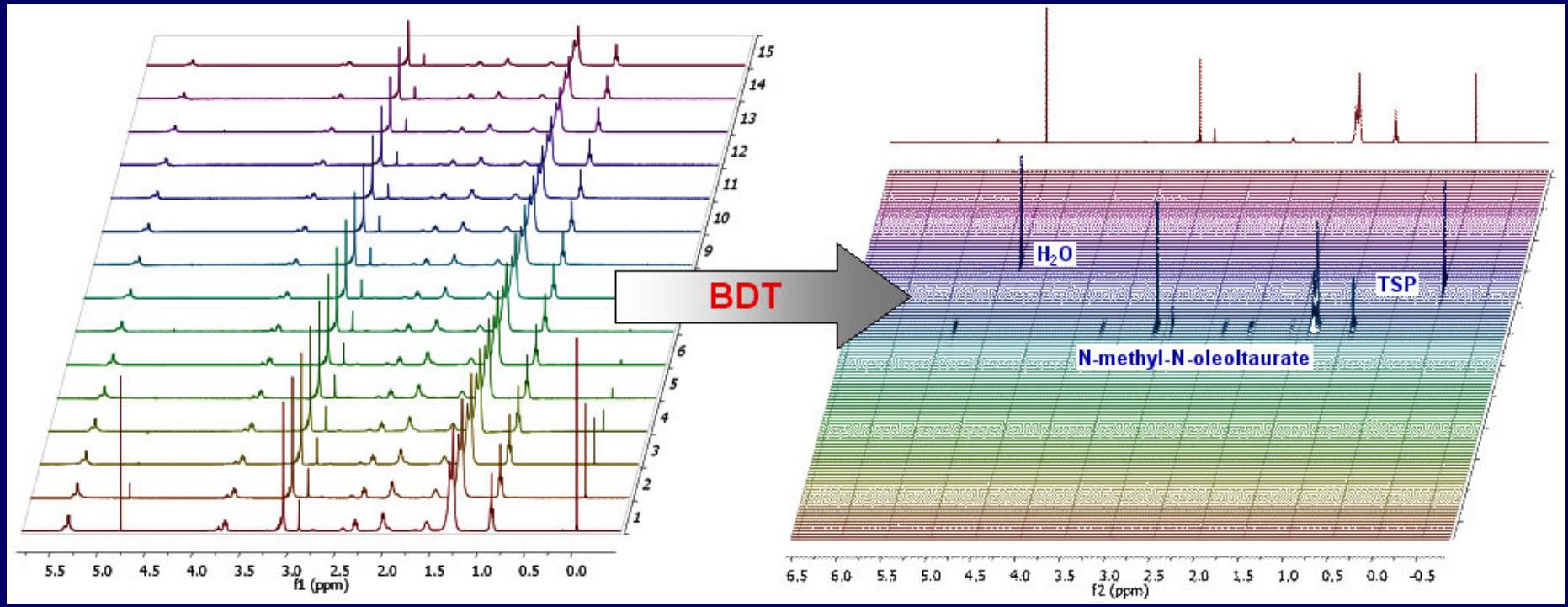
Raw data: [f,z] map



BDT: [f,d] map



A practical example of the Bayesian DOSY Transform (BDT)



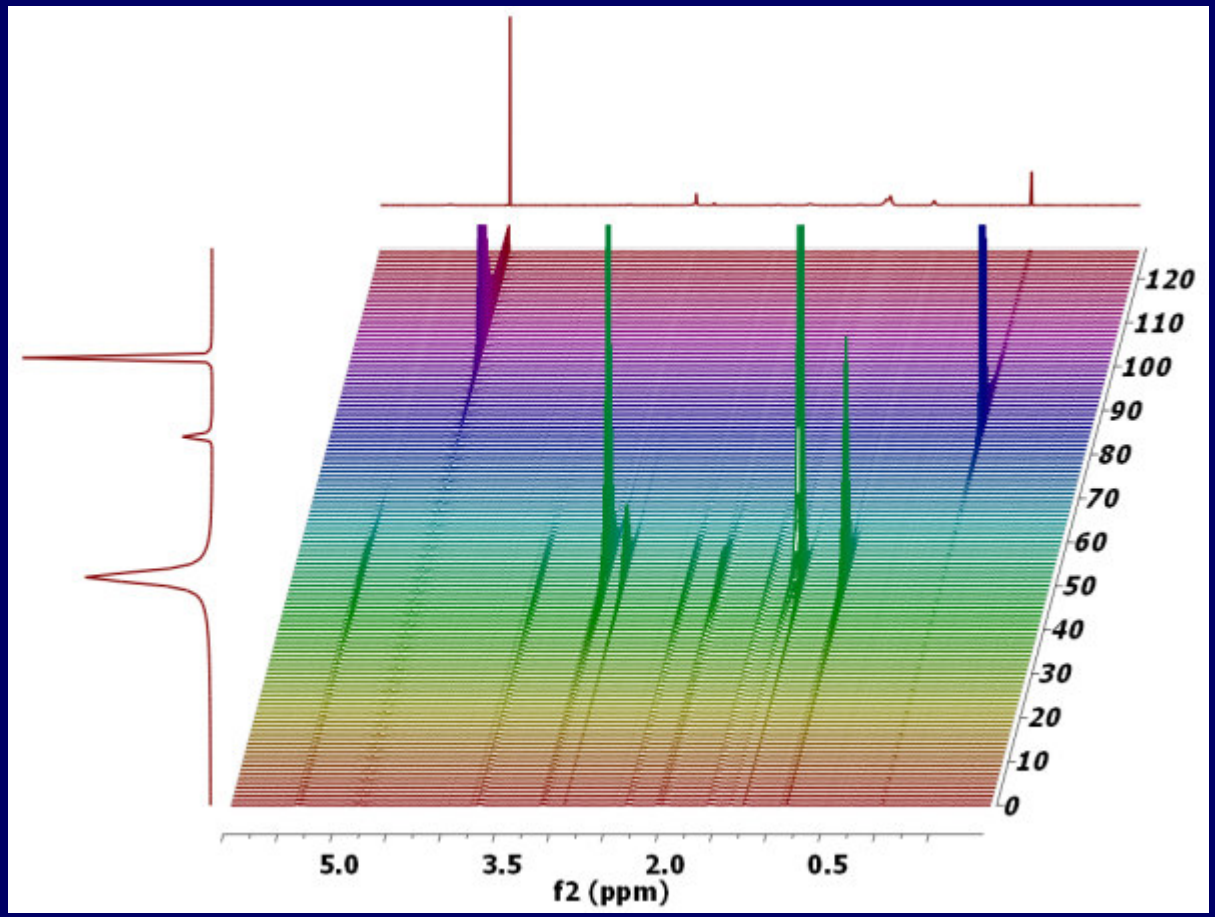
BDT of an aqueous solution of potassium N-methyl-N-oleoltaurate (a surfactant) with TSP at 23 C

The original Varian FID file has been obtained from the *VARIAN NMR USER GROUP LIBRARY* (submitted by Brian Antalek as a sample for this DECRA algorithm)

Presented at XXXVIII **GIDRM**, September 10-13, 2008, Bressanone/Brixen, Italy

Every method has its artifacts

Ridges along the vertical d-dimension in correspondence of strong peaks



The LineSNAP refinement

Bayesian methods are eminently suitable for incorporating all kinds of

a-priori knowledge.

In this case, the a-priori knowledge is:

All signals of a particular molecule must have the same value of d,

and therefore align along a horizontal line.

Every chemist knows this instinctively, but a math algorithm does not – it must be taught to take it into account.

This is what LineSNAP does.

(sorry, but there is no time to tell you how)

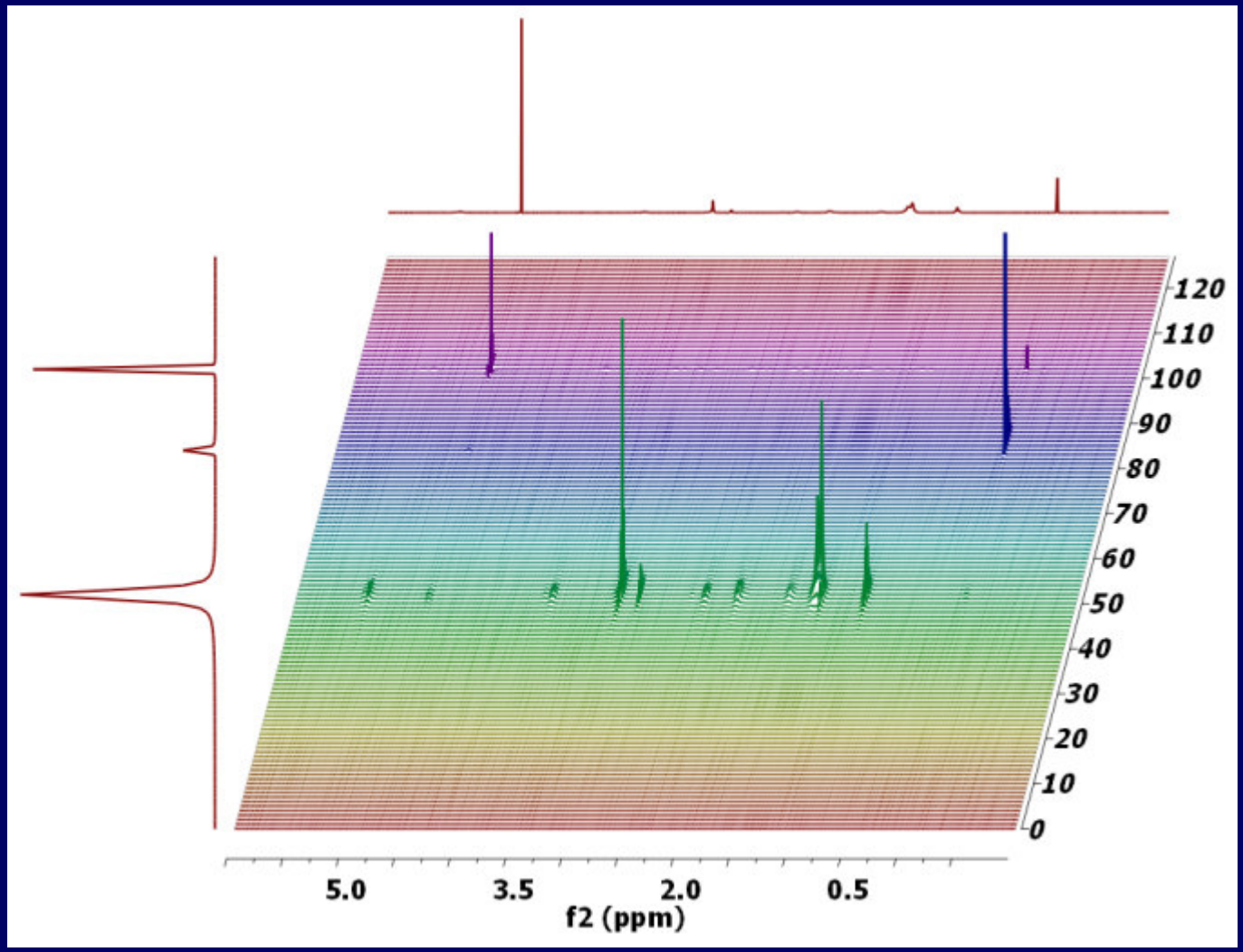
BDT + LineSNAP = BayDOSY

Now part of the Mnova software package

Presented at XXXVIII **GIDRM**, September 10-13, 2008, Bressanone/Brixen, Italy

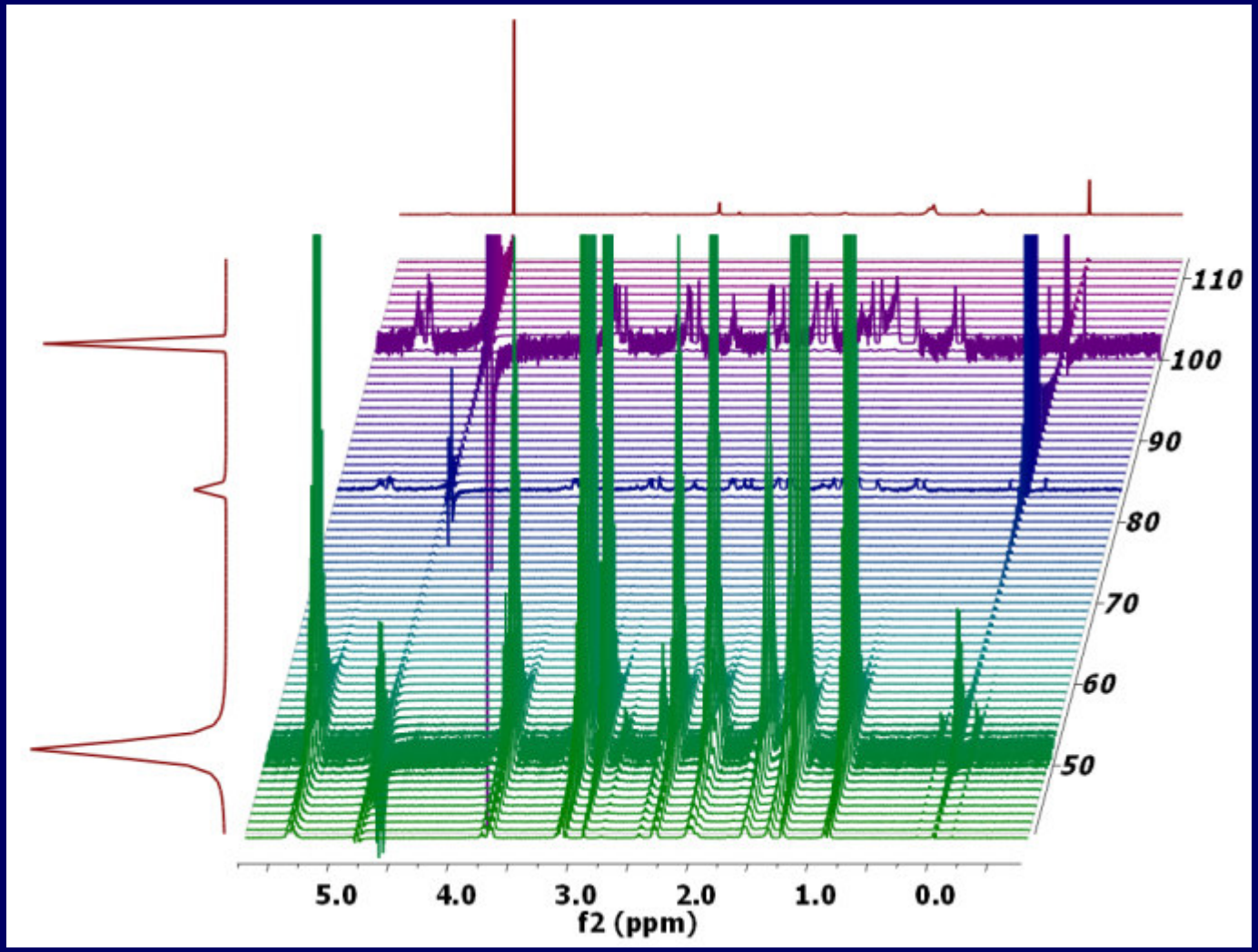
LineSNAP performance

Combined effect of LineSNAP and a probability discrimination enhancement



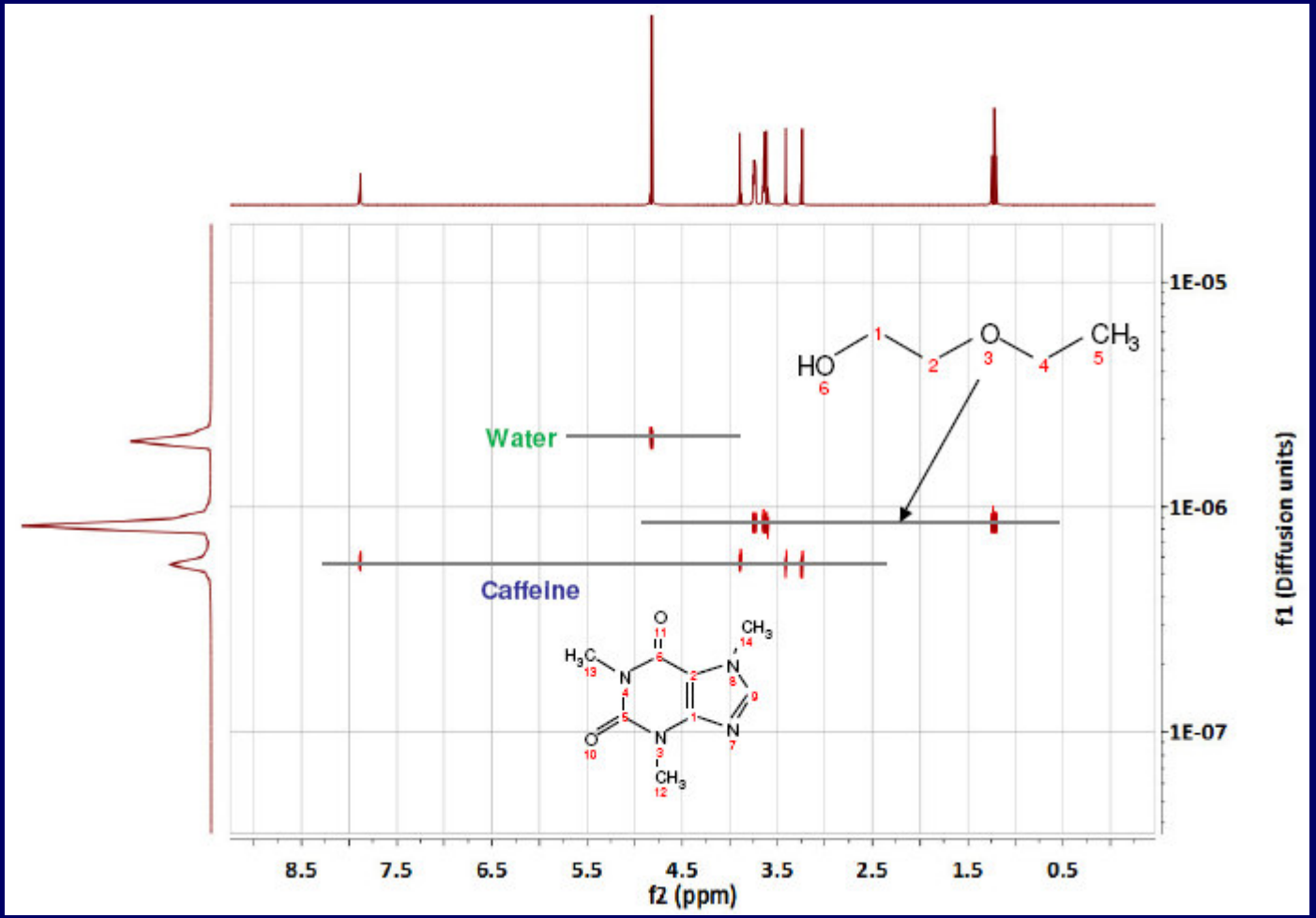
Presented at XXXVIII **GIDRM**, September 10-13, 2008, Bressanone/Brixen, Italy

Nothing is perfect! If you really blow it up …



Presented at XXXVIII **GIDRM**, September 10-13, 2008, Bressanone/Brixen, Italy

A final example (BayDOSY)



Presented at XXXVIII **GIDRM**, September 10-13, 2008, Bressanone/Brixen, Italy

Thank you for your patience

and, please, keep visiting

Stan's Hub

www.ebyte.it

You will find there all this - and much more