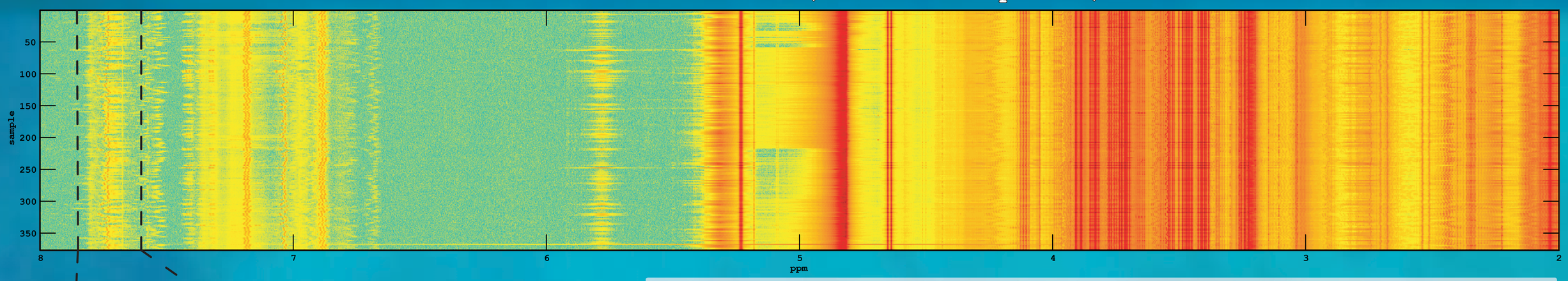
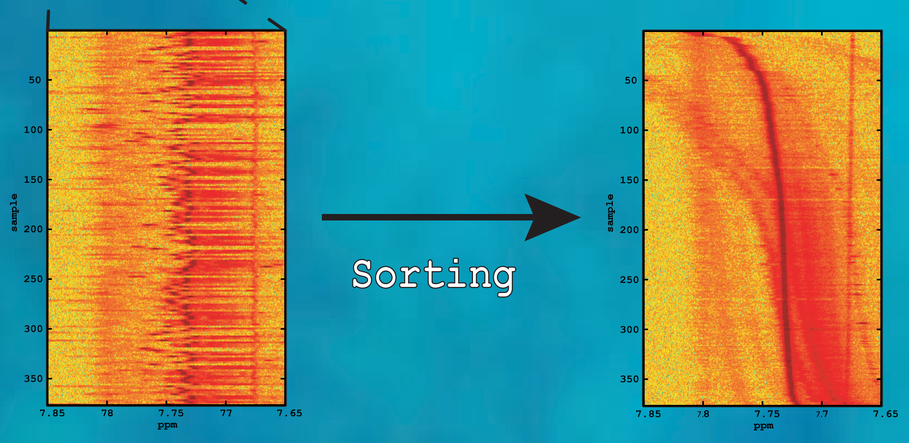
Alignment of ^{1}H-NMR data using a Generalized Fuzzy Hough Transform

Erik Alm[1)], Leonard Csenki[1)], Ralf J.O. Torgrip[1,2)], K. Magnus Åberg[1,2)]
Lars I. Nord[2)], Ina Schuppe-Koistinen[2)], Johan Lindberg[2)]

[1)] *Stockholm University, Dept. of Analytical Chemistry, BioSysteMetrics Group, SE-106 91, Stockholm, Sweden.*
[2)] *AstraZeneca R&D Södertälje, Safety Assessment, Molecular Toxicology, SE-151 85, Södertälje, Sweden.*

Unsorted ^{1}H-NMR data (human blood plasma)





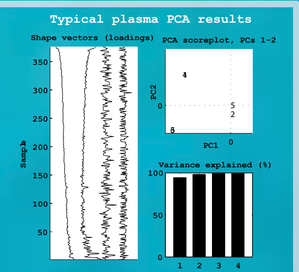
Sorting

In metabolic profiling, multivariate data analysis techniques are used to interpret 1D ^{1}H-NMR data. Multivariate data analysis techniques require that peaks are located in the *same variables* in every spectrum. This location constraint is essential for correct comparison of the intensities of several NMR spectra. However, variations in physico-chemical parameters can cause the locations of the peaks to shift. The location prerequisite may thus not be met, and so, to solve this problem alignment methods have been developed. However, current state-of-the-art algorithms for data alignment cannot resolve the inherent problems encountered when analysing NMR data of biological origin, because they are unable to align peaks when the spatial order of the peaks changes – a common phenomenon.

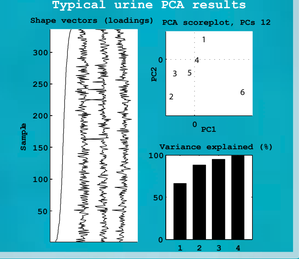
The first step towards alignment using the Fuzzy Generalized Hough Transform is to realize that there are reoccuring 2-dimensional patterns in NMR data of biological origin. These patterns are shift phenomena that arise from varying physico-chemical parameters (*e.g.* pH) in the samples.

A number of manually assignable peaks in the dataset are selected and PCA is performed to assess the number of underlying shift phenomena present. For NMR data of plasma samples, typically 1-2 principal components are enough to describe the shift phenomena while for urine samples, 3 PCs are usually required. A parameterized linear combination of selected loadings aquired from this analysis plus an offset parameter is used as a model to describe inter-sample peak behaviour in the dataset.

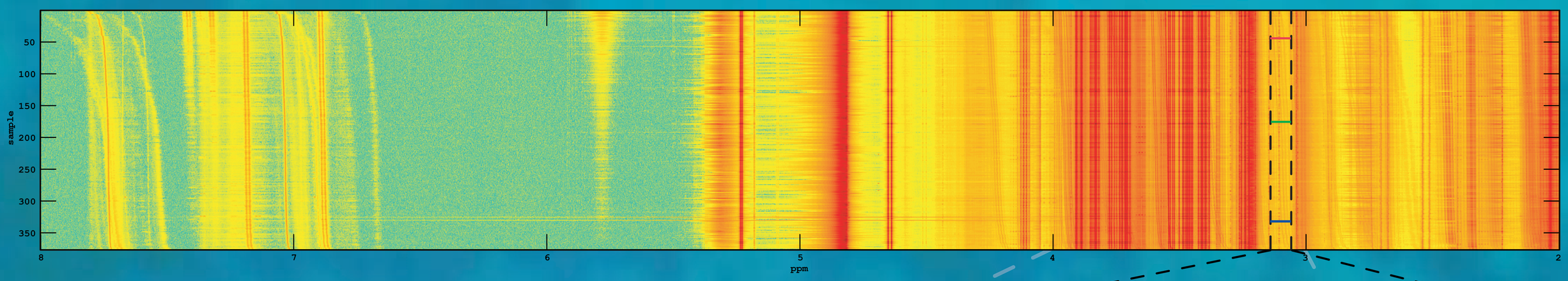
Typical plasma PCA results



Typical urine PCA results



Sorted ^{1}H-NMR data



The Generalized Fuzzy Hough Transform is a method originally used in the field of image analysis to find parameterized shapes in images. In the case of NMR data analysis, we are looking for peaks that behave according to a model described by linear combinations of loadings acquired from PCA of selected model peak shifts in the data. The actual transform from image space to Hough space is described by the equation:

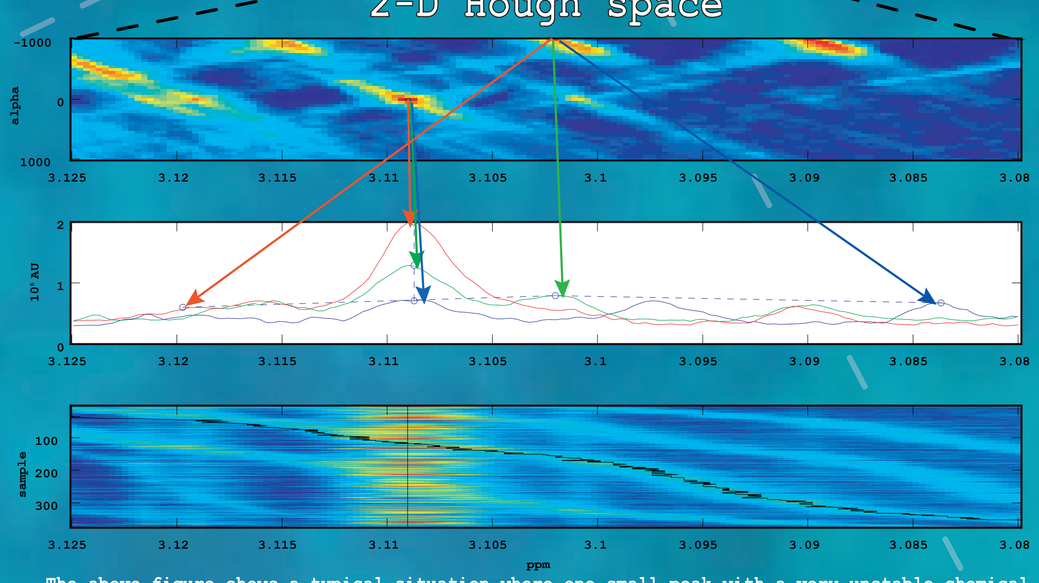
$$h(\boldsymbol{\alpha},k) = H(\mathbf{X},\mathbf{s},\sigma) = \sum_i \sum_j x_{ij}\, p(i,j,\mathbf{s},\sigma), \quad \text{where} \quad p(i,j,\mathbf{s},\sigma) = \exp\left(-\frac{1}{2}\left(\frac{(j-k)-f(i,\boldsymbol{\alpha},\mathbf{s})}{\sigma}\right)^2\right)$$

$$f(i,\boldsymbol{\alpha},\mathbf{S}) = \sum_{l=1}^{K} s_{il}\alpha_l \quad \text{is the model describing the peak shift patterns.}$$

The resulting Hough space has one dimension per parameter used in the model, this results in a space with anywhere from two dimensions and up depending on the number of PCs required to describe the shift phenomena. Each significant local maximum in this space corresponds to a set of parameters that describes a potential peak. These parameters are used to assign the peak correctly throughout the whole dataset. There are usually some false positive hits and some very small peaks that the method fails to detect. Methods for a more accurate intepretation of the Hough space are currently being developed.

The GFHT method is able to align peaks even when they change spatial order between samples, it is also able to align very small peaks in the prescence of other overlapping peaks. The only requirement is that the shift patterns can be described by a model with relatively few parameters, this criterium is generally fullfilled in metabonomic ^{1}H-NMR datasets. The GFHT method is currently in an intermediate state of development and still has a few limitations. The results are dependent on good underlying feature detection and automated peak integration routines and manual selection of model peaks. Further efforts in this area will be concentrated on making the method more robust and fully automated.

2-D Hough space



The above figure shows a typical situation where one small peak with a very unstable chemical shift crosses a larger, more stable peak. Traditional alignment methods would not be able to assign these peaks.

^{1}H-NMR data with predicted peak locations (2-D model)