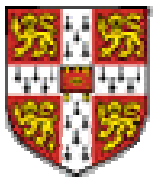
CCPN

Application of Bayesian methods for spectrum analysis



University of Cambridge

Preliminaries…

- Data harvesting
  - All information from raw data to structures deposited
- Large scale analysis of data
  - E.g. how do the structures compare to the raw data?
- Need automatic methods to quantify original data
- Databank for Experimental NMR data (DEN)
  - Using the CCPN data model to store all information

EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Methods for spectrum analysis

- Ultimately need good peak lists
  - This is the main information we want from the spectra
- Most good current methods are empirically based
  - With host of parameters to 'tune' the peak picking
- Need an 'objective' approach
  - Aim is to get signal out of data!
  - Bayesian approach best method?

Bayes' theorem

- Object *x* we wish to know about in hypothesis space *H*
- Data *D* from function *D = R(x) + noise*

$$\Pr(x \mid D, H) = \frac{\Pr(D \mid x, H) \times \Pr(x \mid H)}{\Pr(D \mid H)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Bayes' theorem

- *Prior*: Model for information on object *x*
- *Likelihood*: Information about experiment
- *Posterior*: Reconstructed object + errors
- *Evidence*: Allows comparison between models

EMBL-EBI
European Bioinformatics Institute

University of Cambridge

How does it work for NMR?

- NMR data is ‘dirty’
- Currently have to phase spectrum to get information

EMBL-EBI
European Bioinformatics Institute

University of Cambridge

What does it mean for NMR?

Prior

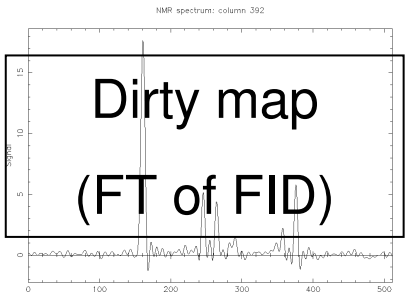
Sample point ('atom')

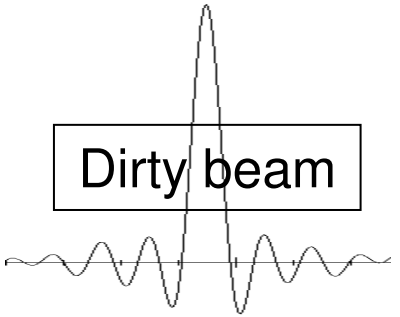
x coordinate

flux (intensity)

} 2 dimensional

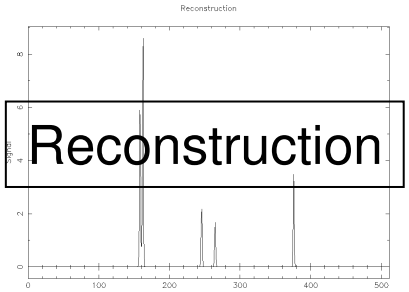
Likelihood





Posterior

(inference)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Random sampling…

- Posterior (and evidence) are determined by sampling
  - Using Markov Chain Monte Carlo (MCMC)
- Calculate likelihood for each random sample point
- Methods slowly increase sampling in areas of interest
- Finally only posterior distribution is explored

Why Bayesian?

- Separates the ‘objective’ and ‘subjective’
  - The ‘prior’ describes the parameters for the sample point
  - The ‘subjective’ likelihood part (‘*what is a peak*’) is contained within a (set of) mathematical formula(s).
  - In implementation here, only about 10 lines of code…
- Parameter settings are related to sampling
  - No ‘fudge’ factors

Why Bayesian now?

- Only recently practically applicable
  - Very computationally intensive
  - New algorithms
  - Faster computers

Under the hood…

- Implementation in the BayeSys3 program
- Run ‘ensemble’ with a number of members
  - Each member samples independently
  - Each member contains one or more sample points
  - Each set of sample points from each member is equally probable
- Can set the *rate* of ‘cooling’ to posterior

Under the hood…

- Uses Hilbert curves to reduce dimensionality
  - Allows high dimension approach
  - Can easily add more attributes (e.g. ‘y coordinate’)

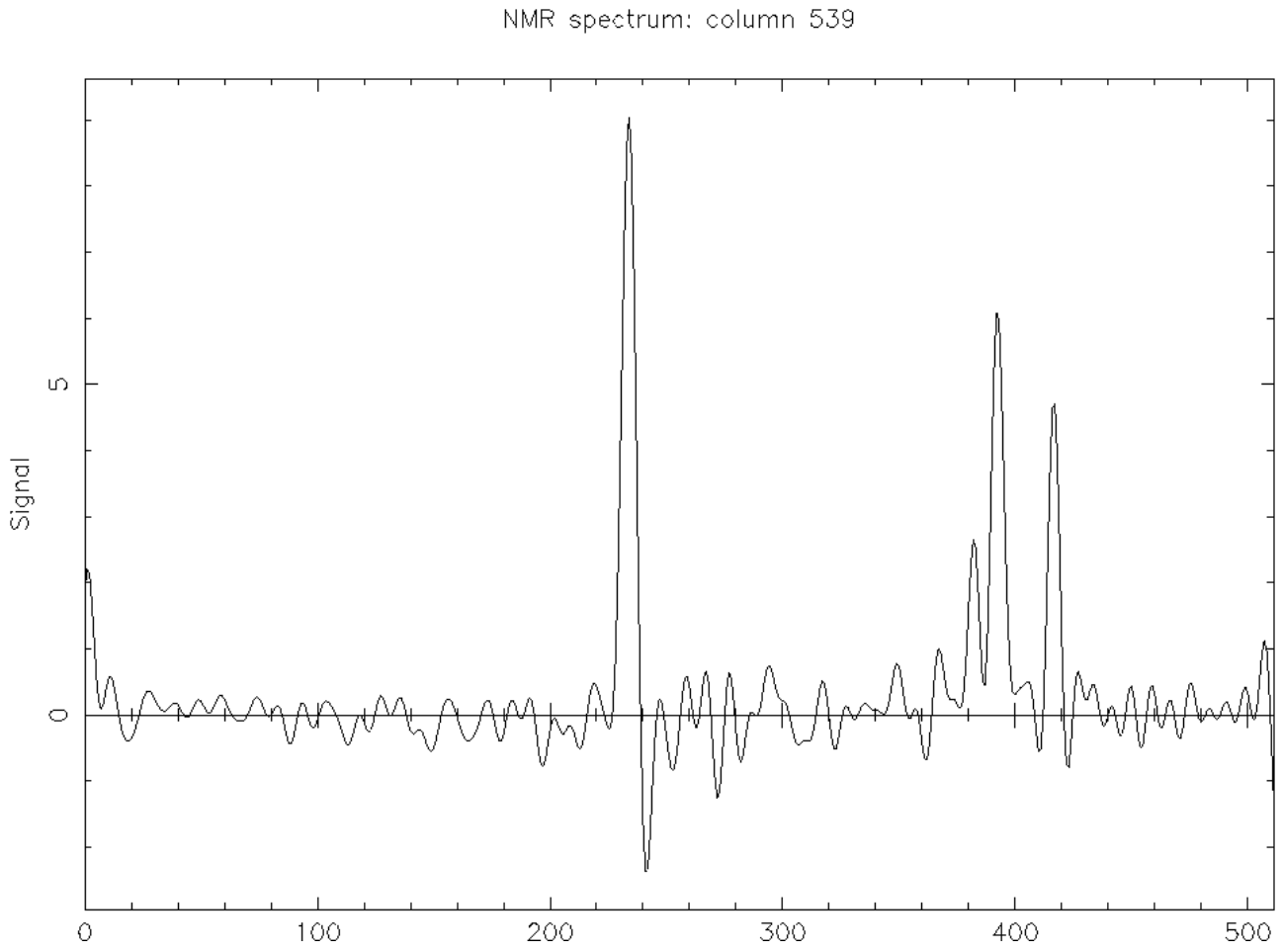
- Many exploration procedures for sampling

Running a 1D example…

Reduced and non-uniform sampling

- Bayesian methods deal very well with this
- Do not need uniformly sampled data
  - Just assume infinite error for missing points
- Example… 1D trace from 2D HSQC
  - From 100% to 20% of data

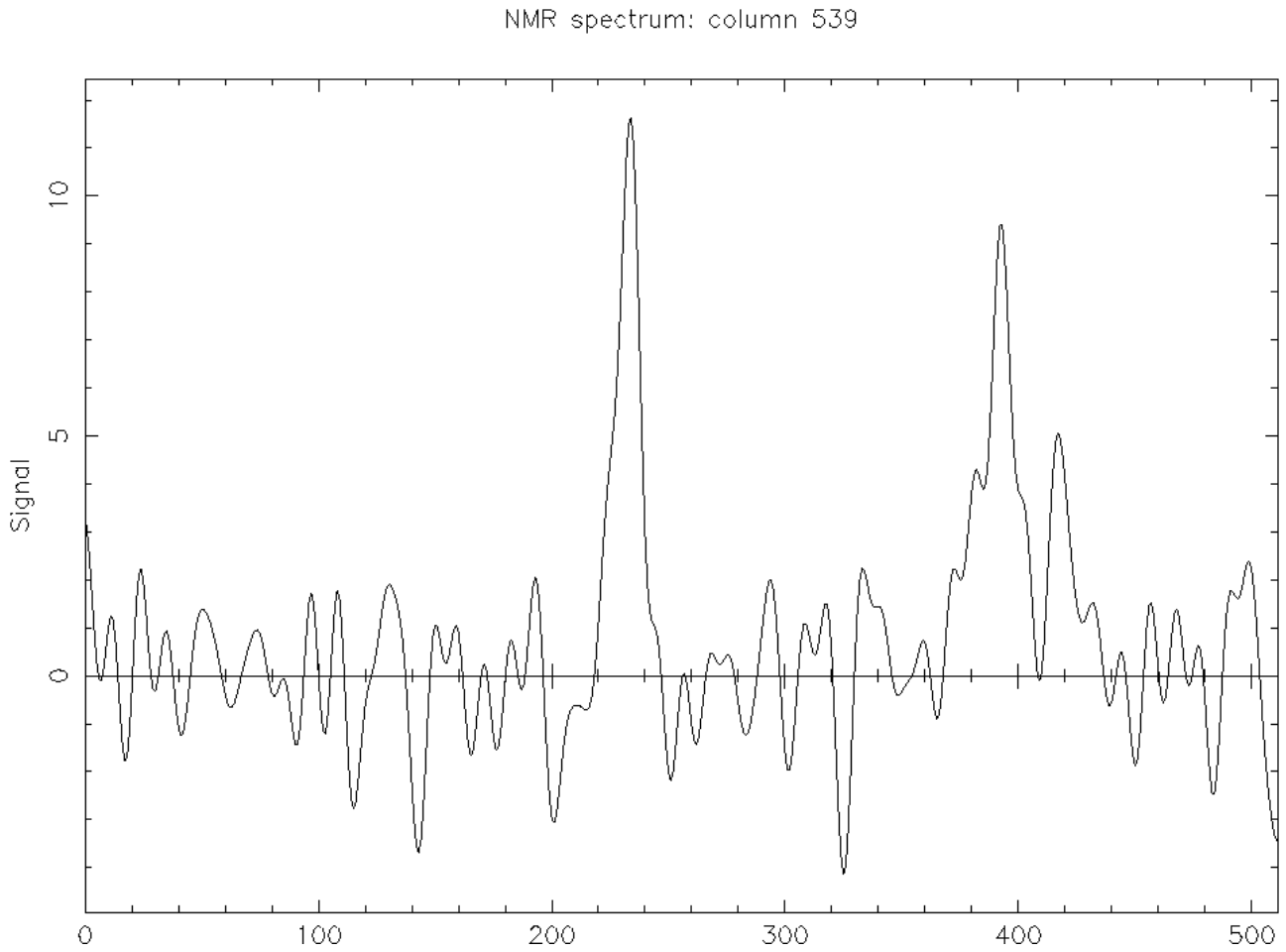
Original slice (100% data)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

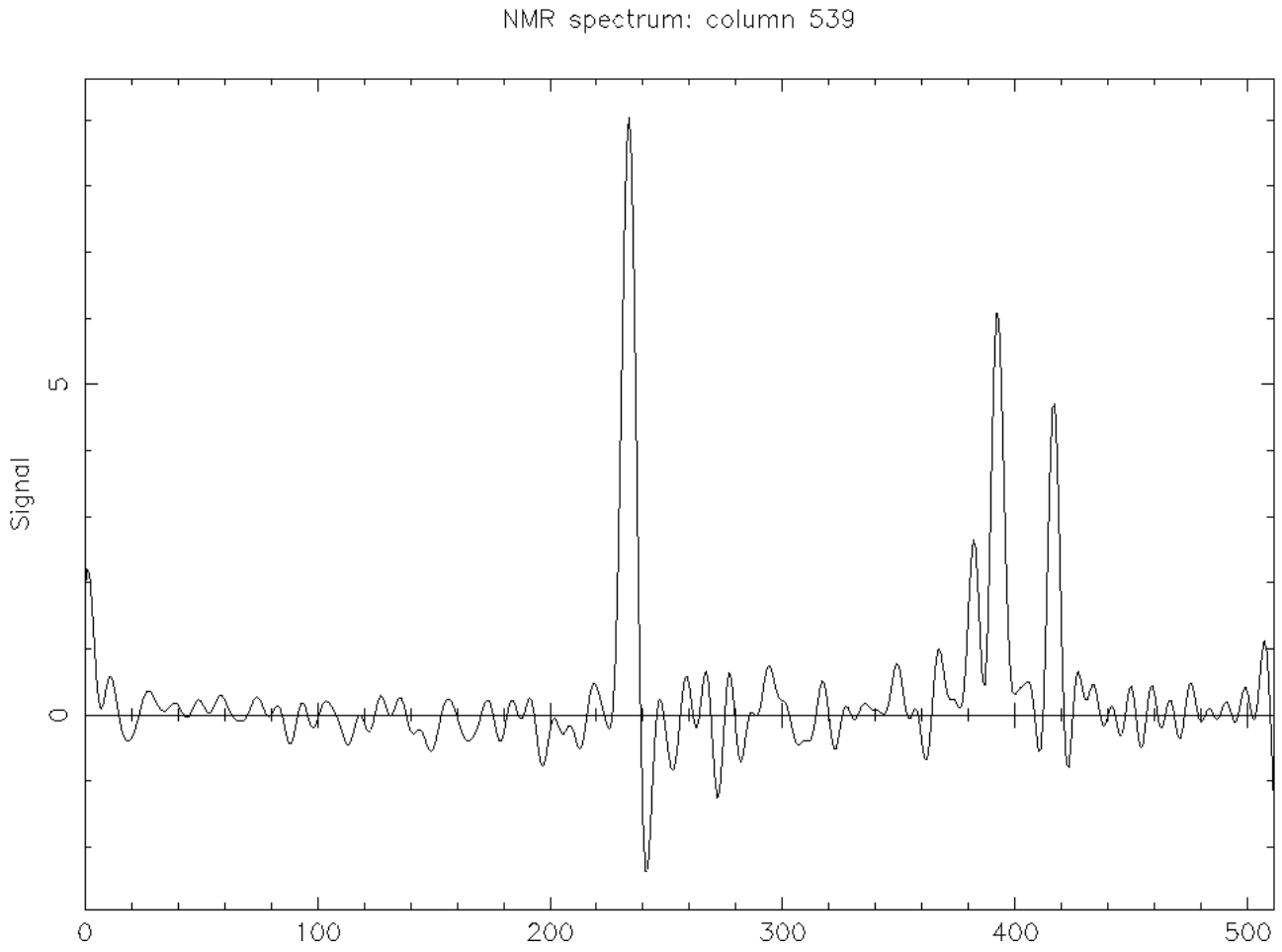
Final slice (20% data)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

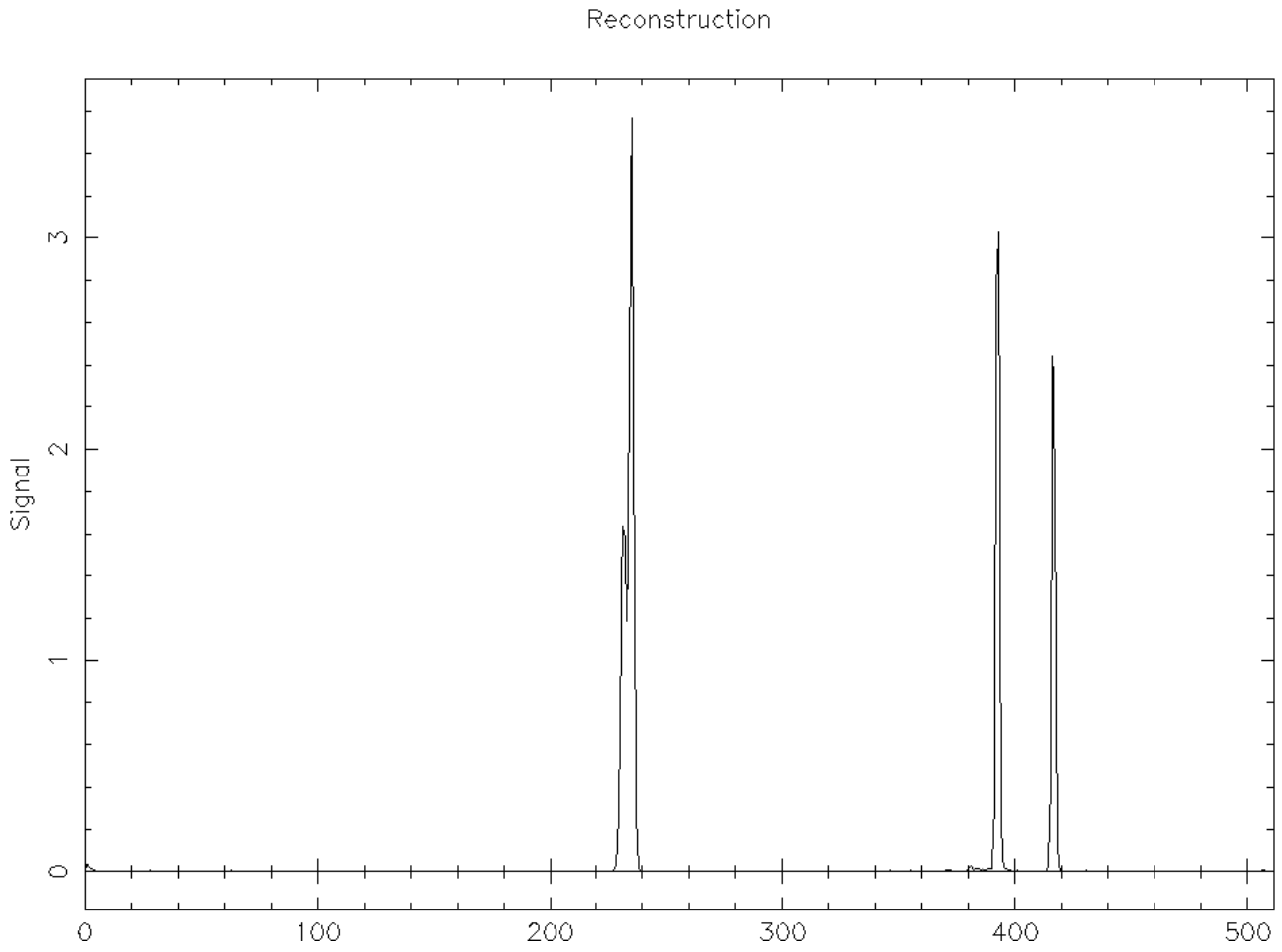
Original slice (100% data)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

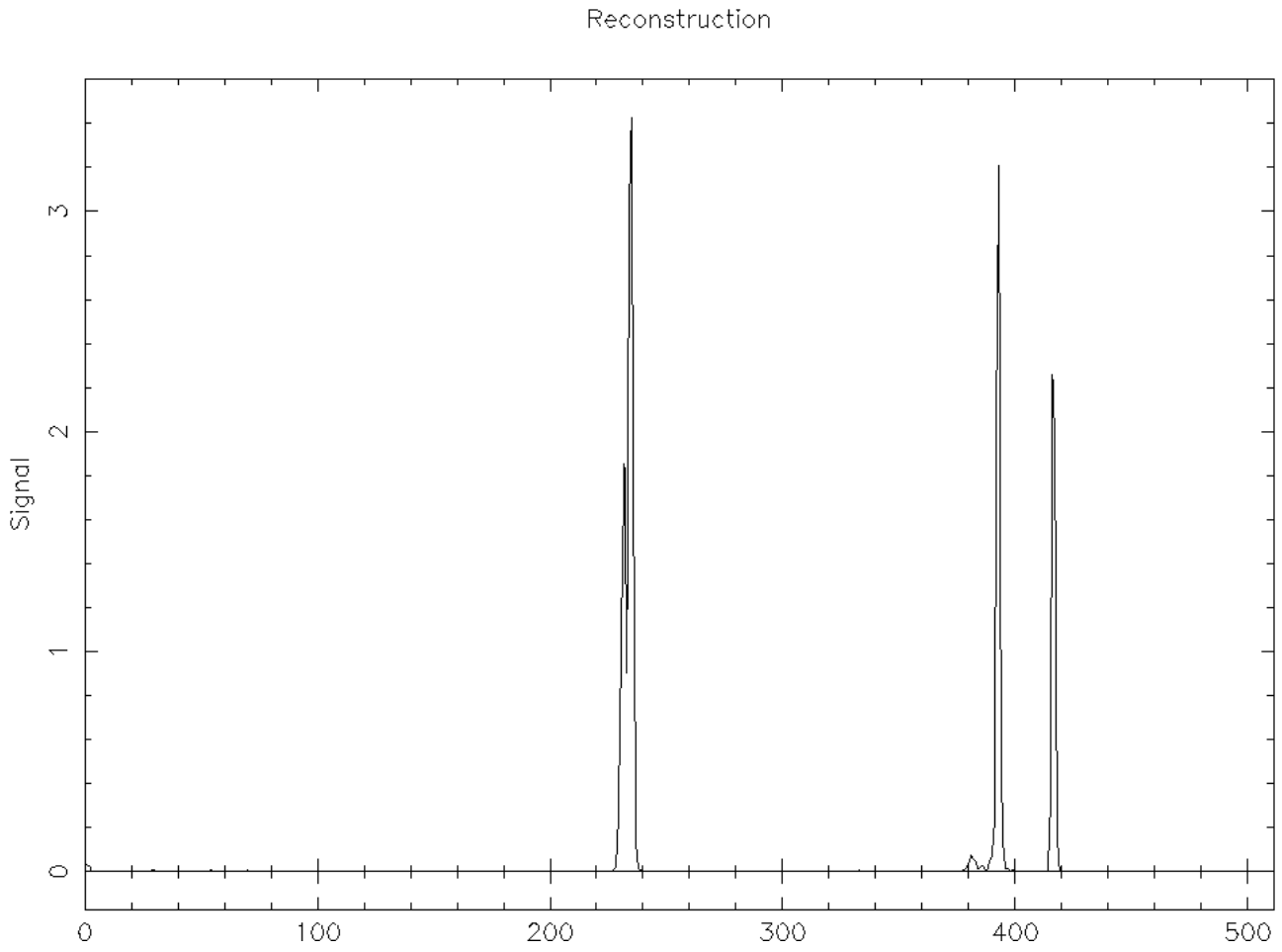
Reconstruction (100% data)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

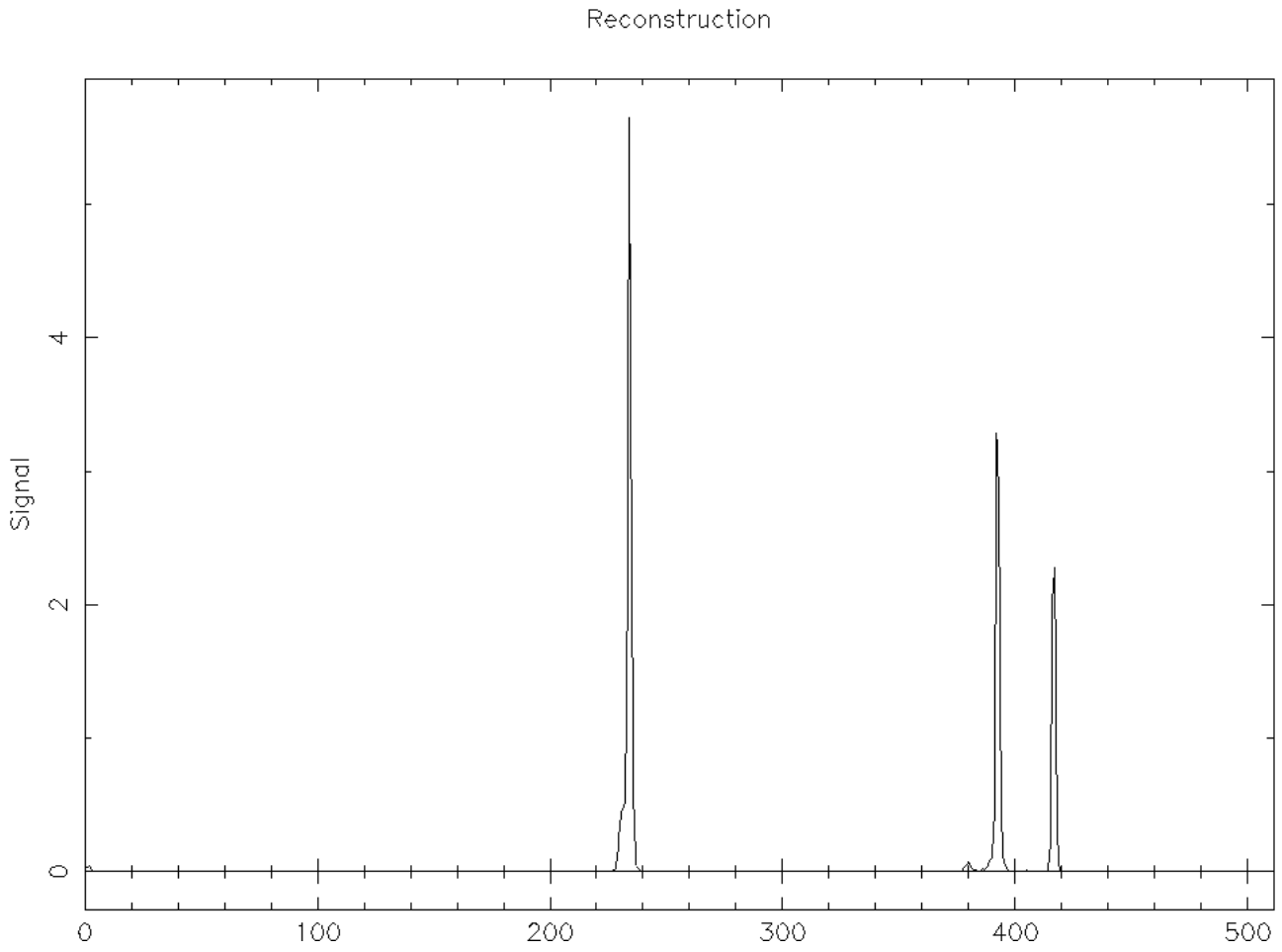
Reconstruction (80% data)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

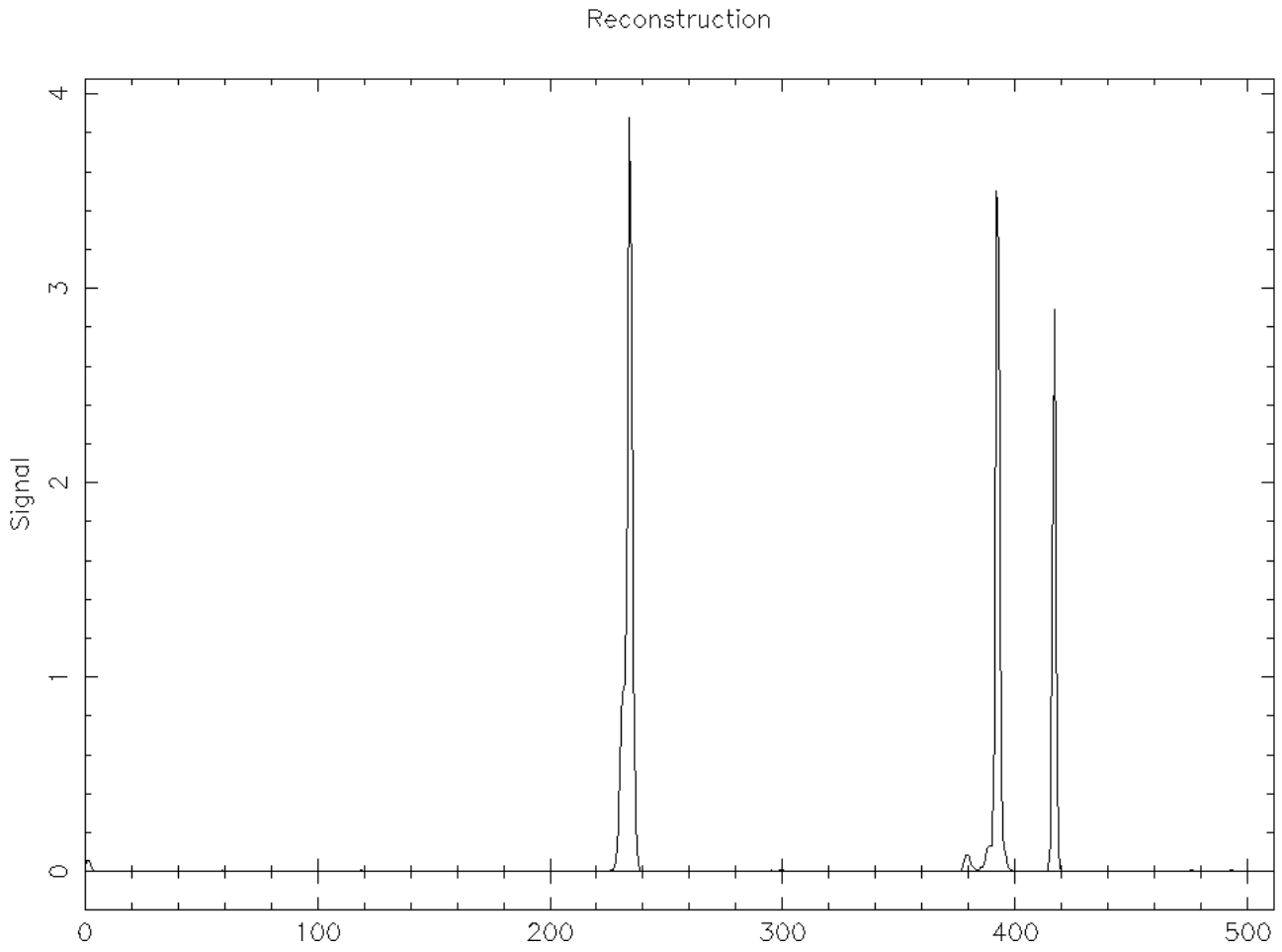
Reconstruction (60% data)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

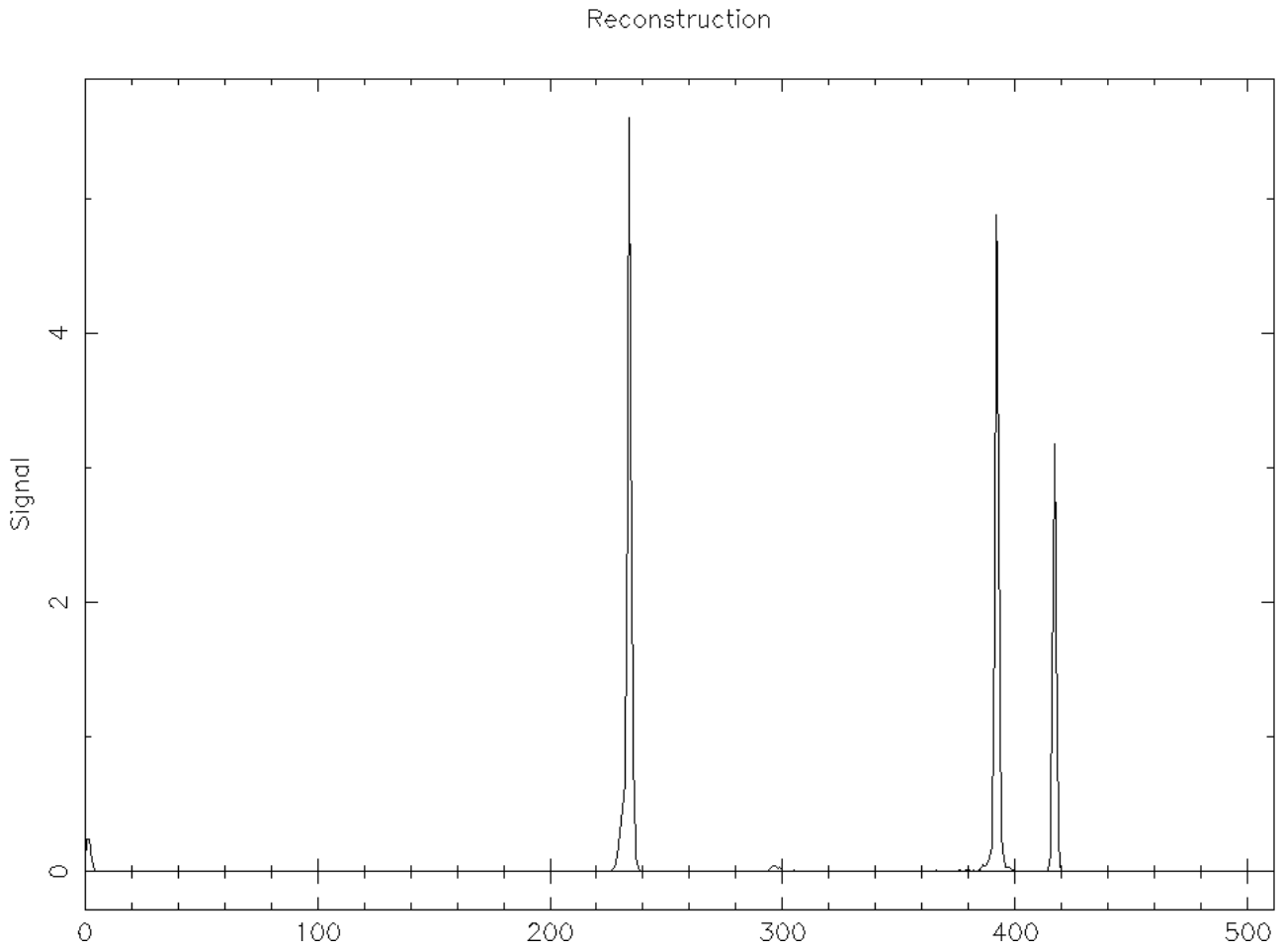
Reconstruction (40% data)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

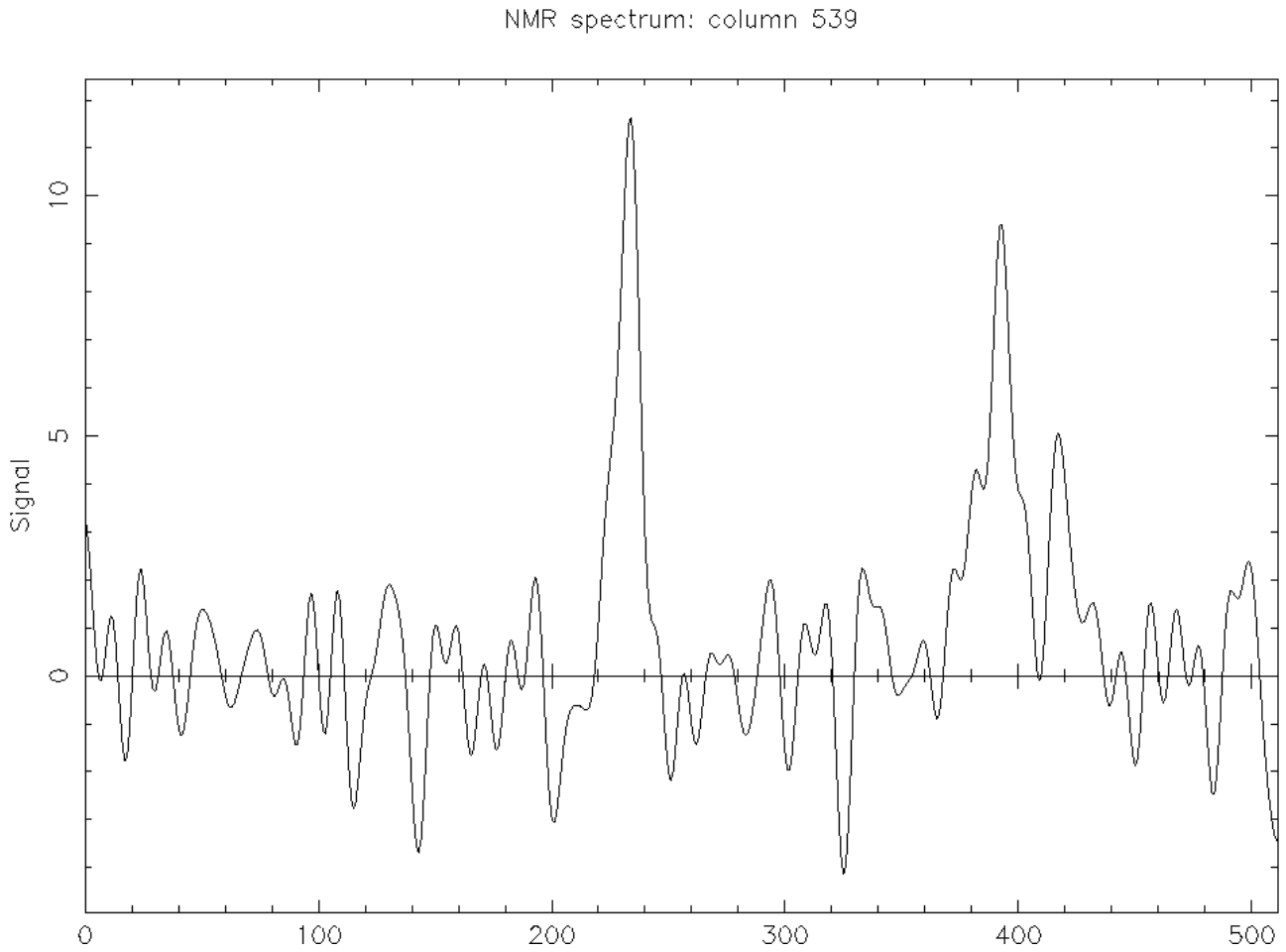
Reconstruction (20% data)



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Final slice (20% data)



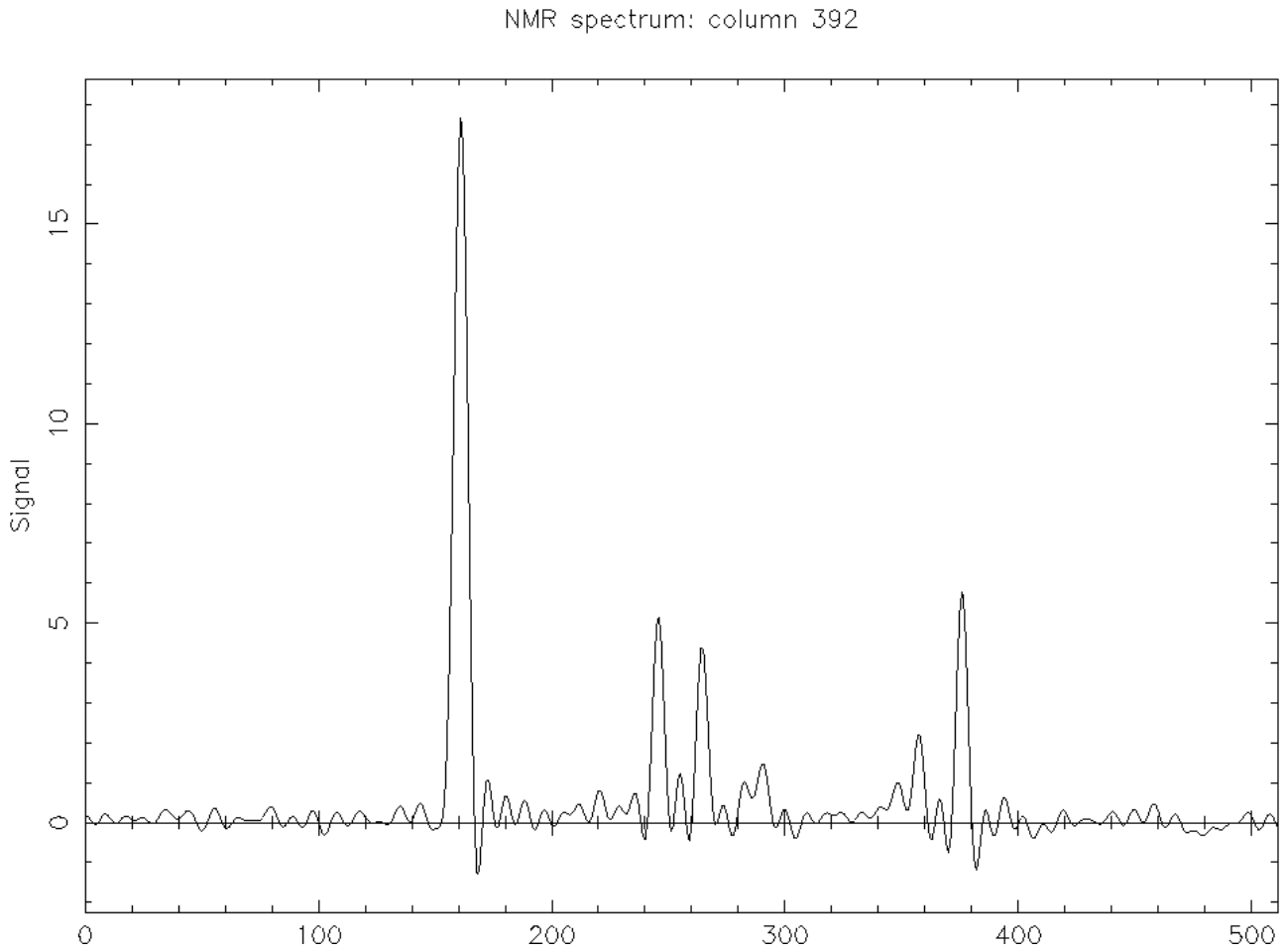
EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Comparison with MaxEnt

- Maximum entropy reconstructions using same input
- How do the methods compare?
  - Maximum entropy
    - Assumes ‘flux’ everywhere
    - Larger error bars
    - Looks more like normal peaks, but is less precise
  - Bayesian
    - Does not assume flux everywhere (no signal if there is noise)
    - Looks for ‘point source’
    - Very precise (atomic)

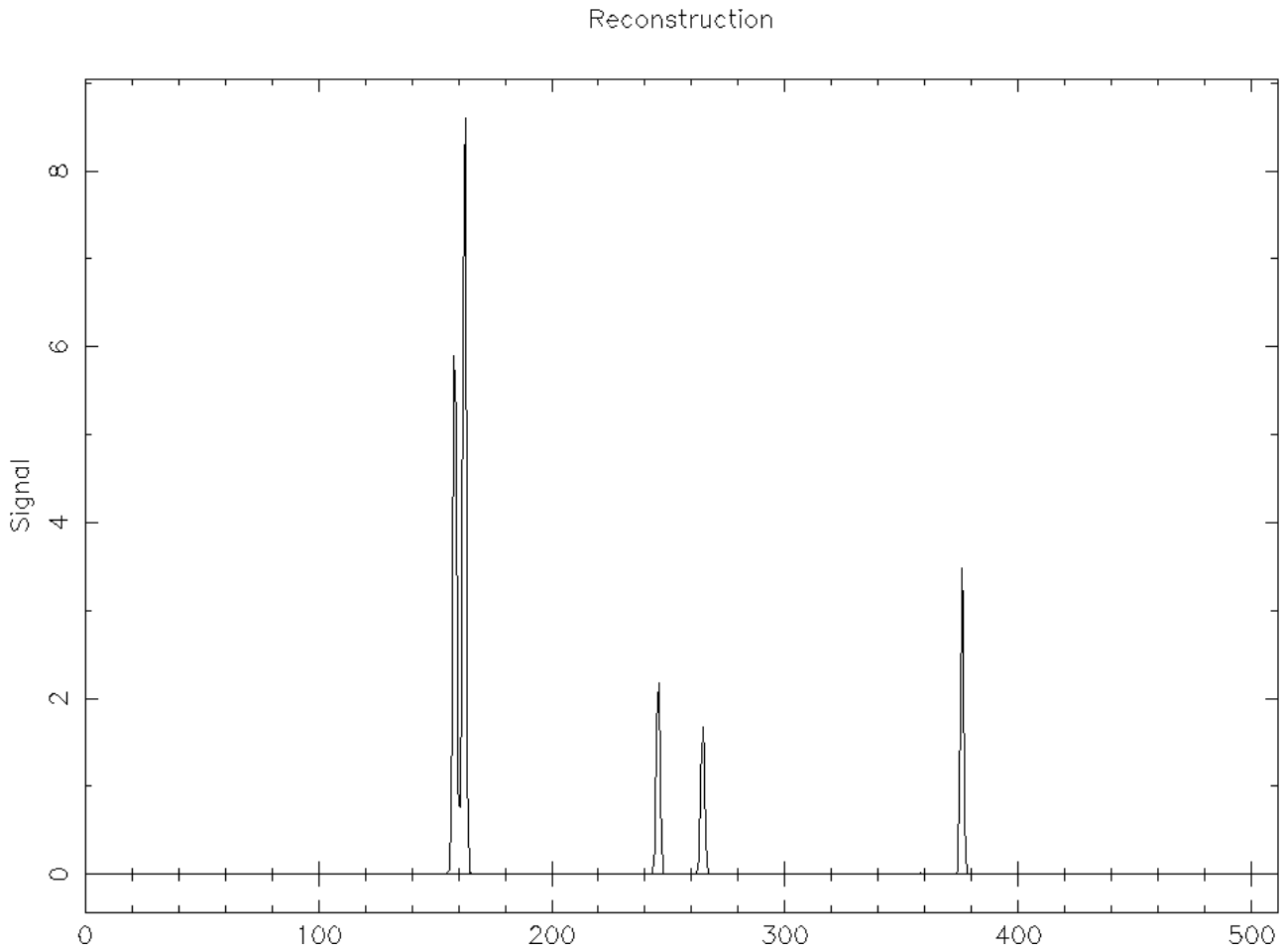
Signal



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

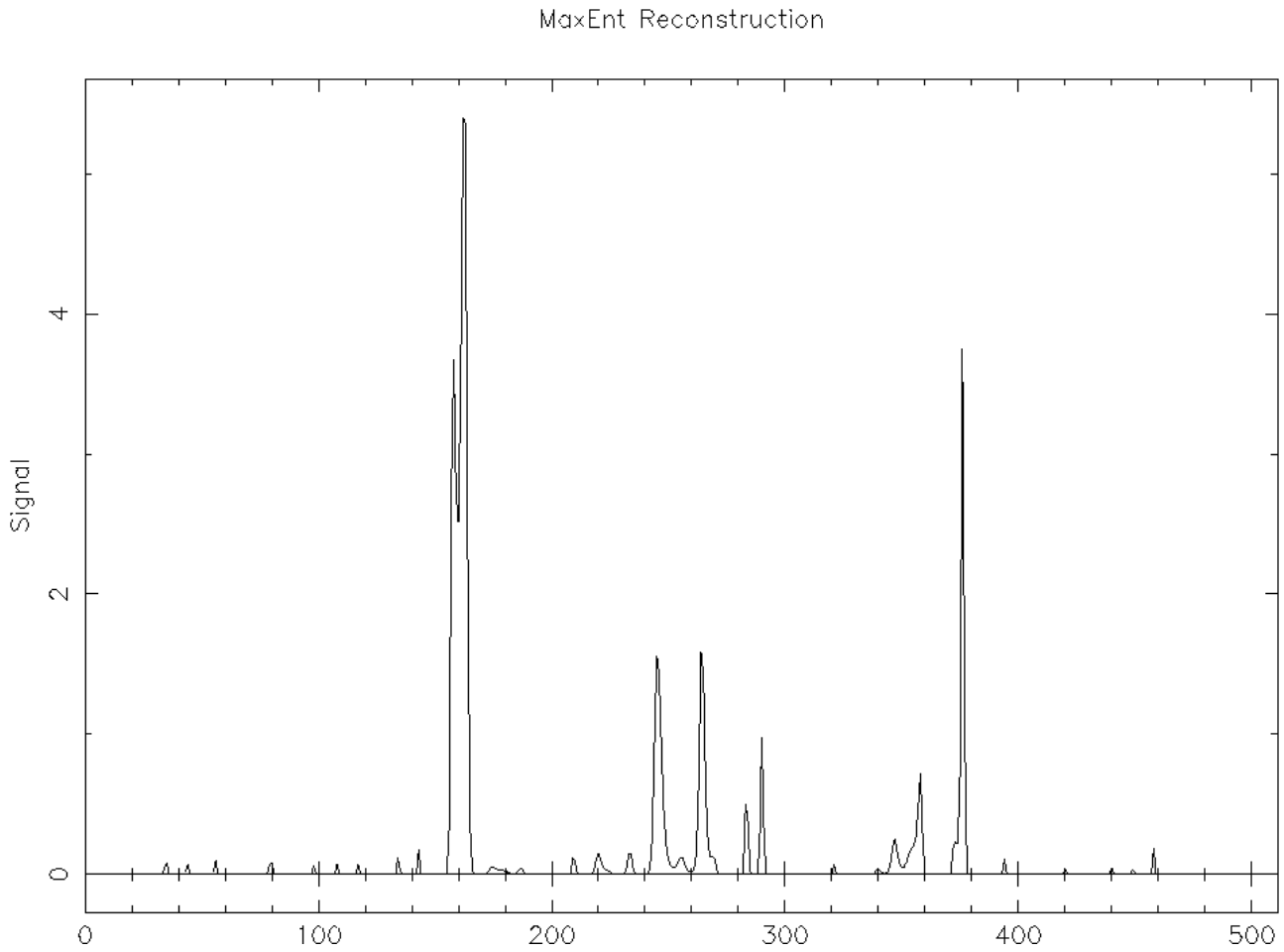
Bayesian reconstruction



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

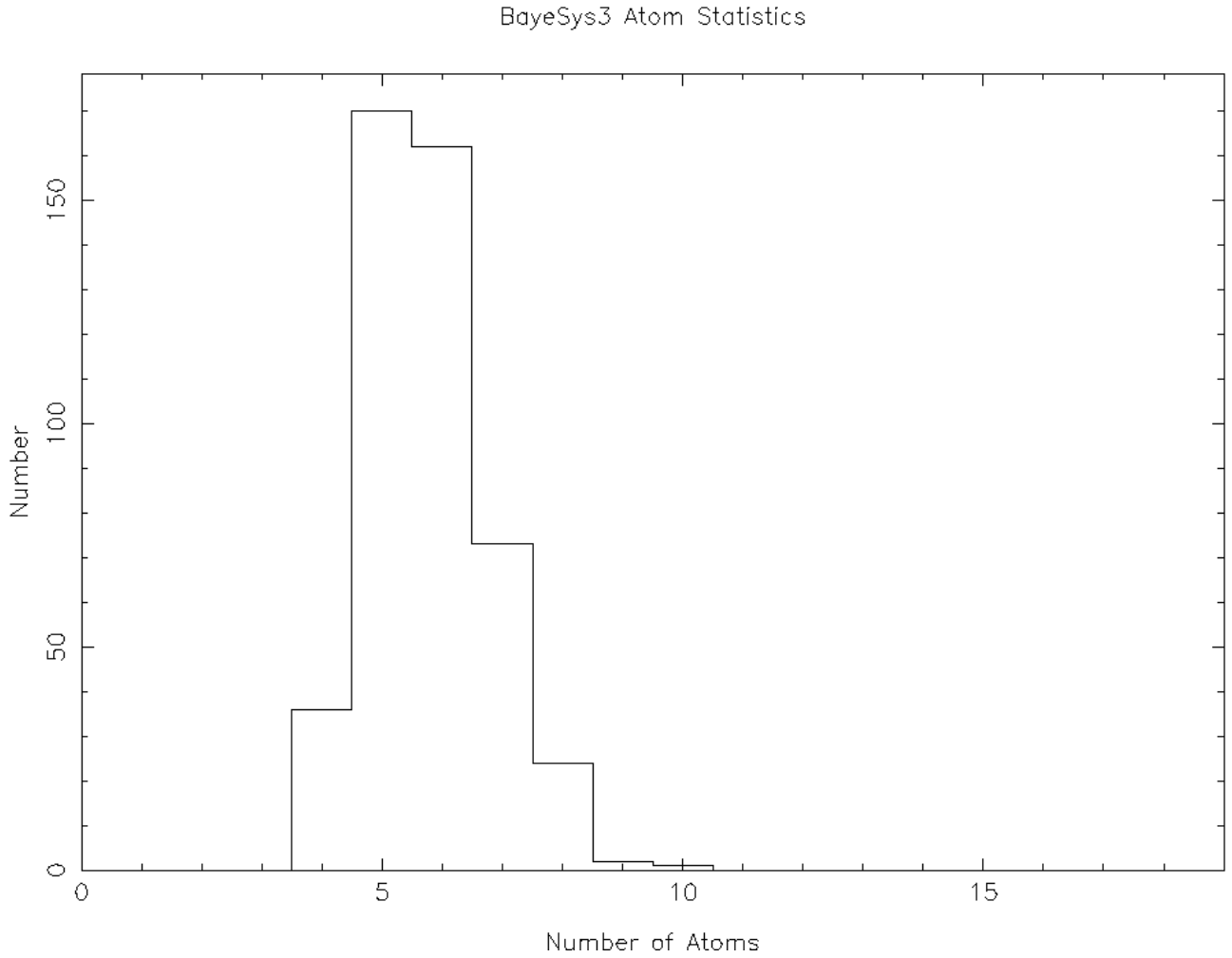
Maximum entropy reconstruction



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Bayesian reconstruction atom histogram



EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Final analysis

- Can use the posterior distribution in whatever way required

- The ‘evidence’ should be reported
  - Quantifies how well the sample points predict the data
- If does not have good results: algorithm failure
  - Not enough sampling
  - Should have reproducible results!

Problems…

- Parameter settings are always a bit of a black box
  - Should be able to find robust solutions based on spectrum type
- Random sampling, so very slow
  - Need cluster for realistic implementation
- Have to make sure enough sampling is performed
  - Reproducible results!
- Extraction of peak information from posterior

Plans…

- Add signal decay as extra dimension
- Do multiple dimensions simultaneously
- Different approaches
  - E.g. analyze specific regions of the spectrum separately
- Determine robust parameter settings for sampling based on spectrum types
- Will link the code to the Data Model
  - Can do Bayesian analysis from within the CCPN framework

Acknowledgments

- Steve Gull
- John Skilling
  - MaxEnt data consultants Ltd.
  - Bayesys3 can be downloaded from:
    - http://www.inference.phy.cam.ac.uk/bayesys/
- Wayne Boucher
- Ernest Laue

EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Formulas…

$$f \xrightarrow{R} D \xrightarrow[R^T]{} f$$

$$pr(D \mid f) \approx \alpha . e^{-\frac{x^2}{2}}$$

$$x^2 = \sum \frac{(D - Rf)^2}{\sigma^2}$$

$$f \frac{(R^T R f)}{\sigma^2} - 2 f^T \frac{(R^T D)}{\sigma^2} + \frac{D^T D}{\sigma^2}$$

(dirty beam) (dirty map) (constant)

- This is situation at one point
- Equation is refactored
- If no information for a point: infinite error, point is ignored
- Dirty map is error weighted!
- Constant part is not used – can get negative chi

EMBL-EBI
European Bioinformatics Institute

University of Cambridge

Formulas

$pr(N_a)$

$pr(A, x_0 \mid N_a)$

$$A = \frac{5.r}{1-r} \text{ (flux)}$$

N_a is number of atoms

x_0 is uniform

What does it mean for NMR?

- *Likelihood*: Calculated from Fourier transformed spectrum ('*dirty map*') and model of non-decaying peak ('*dirty beam*')
- *Prior*: Object *x* is a peak with '*x coordinate*' and '*flux*' attributes (for 1D spectrum)
- *Posterior*: Reconstructed 'clean' spectrum
- *Evidence*: How well does the reconstruction fit the data?