Results of models and binning can introduce artefacts and limit immediate biological insights by consuming and is subject to human error. Conversely, alignment steps, such as peak alignment and binning. Peak fitting is very time consuming and is subject to human error. Conversely, alignment and binning can introduce artefacts and limit immediate biological interpretation of models.

**Results:** We present the Bayesian AuTomated Metabolite Analyser for NMR spectra (BATMAN), an R package which deconvolutes peaks from 1-dimensional NMR spectra, automatically assigns them to specific metabolites from a target list and obtains concentration estimates. The Bayesian model incorporates information on characteristic peak patterns of metabolites and is able to account for shifts in the position of peaks commonly seen in NMR spectra of biological samples. It applies a Markov Chain Monte Carlo (MCMC) algorithm to sample from a joint posterior distribution of the model parameters and obtains concentration estimates with reduced error compared with conventional numerical integration and comparable to manual deconvolution by experienced spectroscopists.

**Methods**

An NMR spectrum, \( y \), can be considered as a linear combination of metabolite peaks plus noise. Metabolite peaks can be labelled catalogued or uncatalogued, according to whether their characteristic peak patterns are known to the user. We propose a two-component joint model for the catalogued \( y^c \) and uncatalogued \( y^u \) metabolites as follows:

\[
y = y^c + y^u + \epsilon, \quad \epsilon \sim N(0, \Lambda \lambda)
\]  

where \( \Lambda \) is the identity matrix and \( \lambda \) is a scalar precision parameter.

The signal from the catalogued metabolites, \( y^c \), can be modelled as a weighted summation of \( M \) metabolite templates,

\[
y^c = \sum_{m=1}^{M} \sum_{u} t_{mu} (\sigma_{mu}) \beta_{mu} \]  

where vector \( t_{mu} \) is a template spectrum of the \( m^u \) multiplet belonging to the \( m^u \) metabolite, \( \sigma_{mu} \) is the chemical shift parameter for the corresponding multiplet, and \( \beta_{mu} \) is proportional to the concentration of the \( m^u \) metabolite. The priors for \( \sigma_{mu} \) and \( \beta_{mu} \) are truncated normal distributions.

The template spectrum, \( t_{mu} \), in turn is generated as a weighted summation of Lorentzian peaks. By default BATMAN obtains this

To whom correspondence should be addressed.
A Markov Chain Monte Carlo algorithm is proposed to sample prior information on the peak pattern from the Human Metabolome Database (HMDB) (Wishart, et al., 2009). We model the Lorentzian peak width $\gamma_m$ as,

$$\ln(\gamma_m) = \mu + \nu_m$$  \hspace{1cm} (3)

where $\mu$ represents the spectrum wide average and $\nu_m$ is a random effect, representing metabolite specific deviations from $\mu$. Our priors for $\mu$ and $\nu_m$ are both Gaussian.

Signal generated by uncatalogued metabolites, $y^r$, is modelled as a linear combination of wavelet basis functions $(W)$. This alleviates difficulties associated with the incompleteness of the spectral library that can influence other methods (Mercier, et al., 2011). A probability density model is specified in the wavelet domain of the data, i.e. $\theta^rW y^r$. We use symlet-6 wavelet basis functions as they are well suited to modelling Lorentzian peaks (Astle, et al., 2012).

The distribution of $\theta$ is modelled by a truncated Gaussian, whose parameters are specified by the parameters $\lambda$, $\psi$, $\tau$. The hyper-parameter vector $\psi\sim\text{Gamma}(a,b/2)$. The hyperprecision vector $\lambda\sim\Gamma(a,b/2)\text{Gamma}(c,d/2)$ allows the prior precision associated with each wavelet deviate from the global precision $\lambda\sim\Gamma(a,b/2)$. $\tau$ is a truncation limit vector where each $\tau_i$ has a Gaussian distribution left truncated at a small negative value $h$.

3 IMPLEMENTATION

A Markov Chain Monte Carlo algorithm is proposed to sample from a joint posterior distribution of the model parameters. We temper the likelihood and penalize the wavelet component of the prior via the scale parameter $\delta$ and adapt $\nu_m$ according to a joint posterior distribution of the model parameters. We use block updates for $\beta$ jointly with $\theta$, and for $\sigma_m$ jointly with $\theta$ to allow the chain to make global moves. The details are described in (Astle, et al., 2012). Adaptation techniques are used to update the peak shift $\sigma_m$ and width parameters $\gamma_m$. The core algorithm is implemented in C++ with support for parallel processing between spectra to improve the processing speed. $R$ version 2.12.1 (or higher), together with the $R$ packages described in documentation, is required.

BATMAN runs on Windows, Mac OSX and Linux/Unix operating systems and supports text file, R data format and ID Bruker spectral data files as input. Processing 8 spectra on an Intel 3.0GHz Quad-Core 2 processor machine with 11 catalogued metabolites and taking approximately 22 minutes. The MCMC procedure ran for 2000 iterations following a 4000 iteration burn in. The time required scales linearly with the number both of metabolites and spectra (on a multicore machine, multiple spectra can be analysed in parallel with the same run time as a single spectrum analysis).

4 RESULTS

Figure 1 shows an example result of a BATMAN fit of a $^1$H NMR spectrum of normal rat urine. The catalogued metabolite peaks are correctly fit by the algorithm, with uncatalogued peaks absorbed by the wavelet component. The unimodality of peak position and concentration distributions indicate no ambiguity in the assignments. The BATMAN deconvolution has been shown to have reduced mean estimation error compared with conventional numerical integration methods and its results on an example data set fitting 26 metabolites are comparable with the manual deconvolution by five experienced NMR spectroscopists as described by (Astle, et al., 2012).

ACKNOWLEDGEMENTS

This work was funded by grant BB/E20372/1 from the UK Biotechnology and Biological Sciences Research Council (BBSRC).

REFERENCES

Astle, W., et al. (2012) A Bayesian Model of NMR Spectra for the Deconvolution and Quantification of Metabolites in Complex Biological Mixtures, Journal of the American Statistical Association, in press.

Mercier, P., et al. (2011) Towards automatic metabolic profiling of high-resolution one-dimensional proton NMR spectra, J Biomol Nmr, 49, 307-323.

Tredwell, G.D., et al. (2011) Between-Person Comparison of Metabolite Fitting for NMR-Based Quantitative Metabolomics, Anal Chem, 83, 8683-8687.

Weljie, A.M., et al. (2006) Targeted profiling: Quantitative analysis of H-1 NMR metabolomics data, Anal Chem, 78, 4430-4442.

Wishart, D.S., et al. (2009) HMDB: a knowledgebase for the human metabolome, Nucleic Acids Res, 37, D603-D610.

Zeng, C., et al. (2011) Identification and quantification of metabolites in (1)H NMR spectra by Bayesian model selection, Bioinformatics, 27, 1637-1644.