Improving the model-free analysis of protein dynamics

Edward d'Auvergne and Paul Gooley

Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Victoria 3010, Australia

Data analysis

Knowledge of protein structure is fundamental for understanding protein function. However it is becoming clear that knowledge of motion may be just as important. Experimentally, the study of NMR relaxation data is the only technique which can reveal the motions on an atomic level. Relaxation data by itself is complex to interpret but using model-free theory, amplitudes and timescales of motion can be extracted.

Model-free parameter types:

S^2 – amplitude of motion (ps to ns).

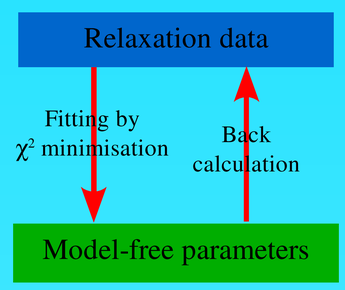
τ_e – timescale of motion (ps to ns).

R_{ex} – chemical exchange (indicator of µs to ms motions)

Model-free models

Using different combinations of the model-free parameters, multiple (usually five) model-free models are constructed.

Model-free analysis



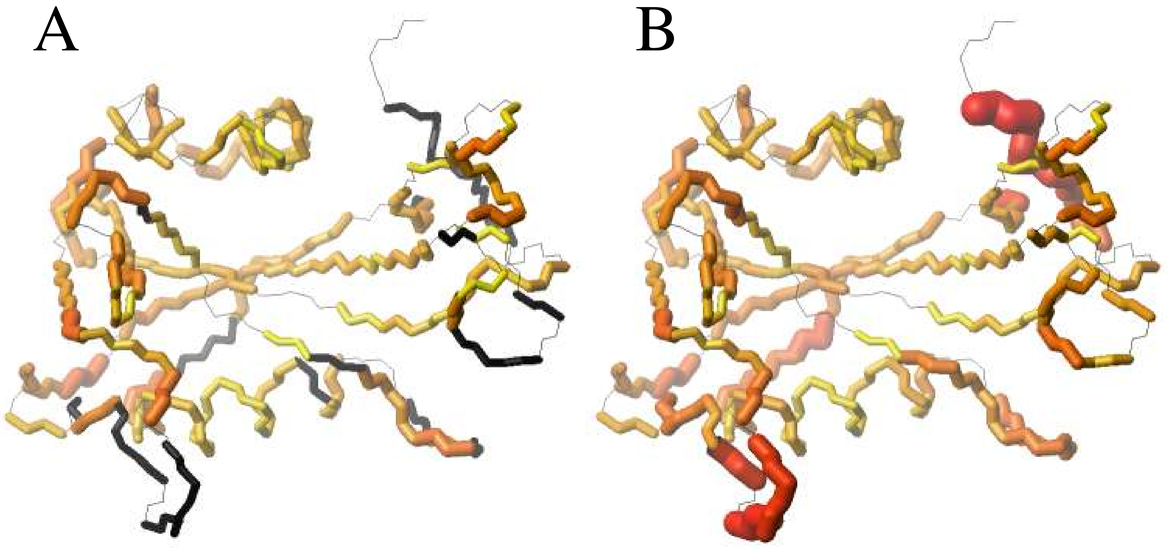
The data is fitted five times, once for each model-free model. The best model is then selected.

We have identified two issues within model-free data analysis which significantly effect the final results. These problems are in model selection and χ^2 minimisation.

Model-free model selection[1]

Current model selection consists of hypothesis testing, a flow diagram constructed from χ^2 and F-tests, to select between five model-free models. This technique hides certain motions because of two reasons, under-fitting and not selecting a model when one ought to be selected.

Numerous established techniques from the statistical field of model selection were studied including Akaike's Information Criteria (AIC), small sample size corrected AIC (AICc), Bayesian Information Criteria (BIC), cross-validation, and bootstrap model selection. The conclusion is that AIC model selection is best for model-free analysis. AIC increases the accuracy, simplicity, and speed of model-free analysis.



Model-free dynamics of Lupin Ap_4A hydrolase. A and B, S^2 values extracted using current and AIC model selection respectively. Thin black bonds indicate residues for which there is no data, thick black bonds indicate where no model is selected. Both colour and bond thickness are proportional to S^2 values. Relaxation data consists of R_1, R_2, and NOE values collected at both 500 and 600 MHz.

Model-free minimisation

Some of the programs which minimise the χ^2 value to fit the model-free parameters include Modelfree, Dasha, DYNAMICS, and Tensor. Most published analyses use the program Modelfree.

Issues identified in Modelfree:

1. Serious bug – incorrect results.
2. Bad constraints algorithm – stuck at the limits.
3. Minimisation terminated early – inaccurate results.
4. Singular matrix – Levenberg-Marquardt algorithm problem. No minimisation.

Relax

To fix the minimisation problem, the program 'relax' which implements multiple minimisation algorithms is currently being written. The best minimiser for model-free analysis appears to be Newton minimisation. The constraint algorithm used is the Augmented Lagrangian method.

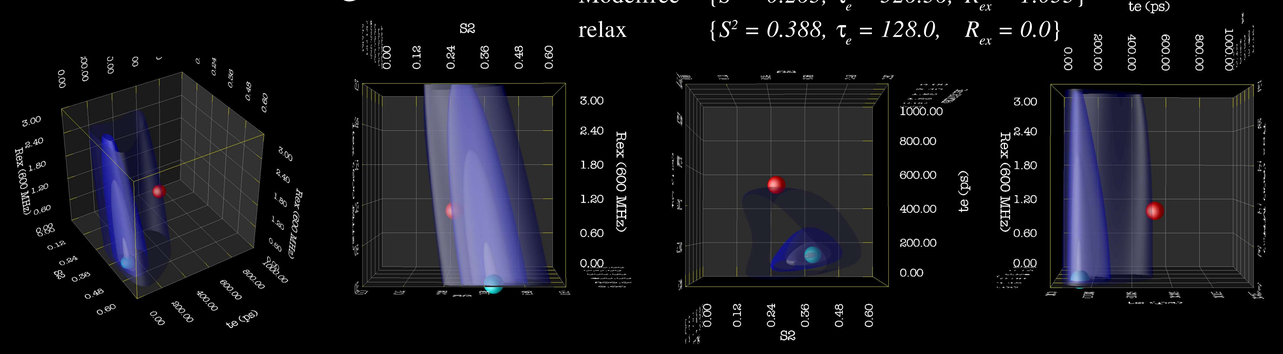
1. d'Auvergne, E. J. and Gooley, P. R. (2003) The use of model selection in the model-free analysis of protein dynamics, *Journal of Biomolecular NMR*, **25**, 25-39.

The Modelfree bug

True values $\{S^2 = 0.388, \tau_e = 128, R_{ex} = 0\}$

Modelfree $\{S^2 = 0.263, \tau_e = 526.36, R_{ex} = 1.053\}$

relax $\{S^2 = 0.388, \tau_e = 128.0, R_{ex} = 0.0\}$

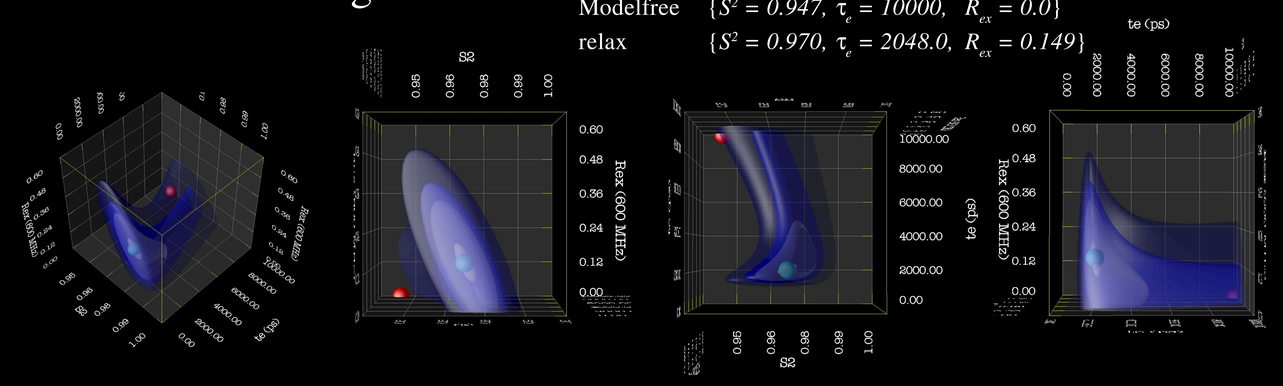


Failure of limits algorithm

True values $\{S^2 = 0.97, \tau_e = 2048, R_{ex} = 0.149\}$

Modelfree $\{S^2 = 0.947, \tau_e = 10000, R_{ex} = 0.0\}$

relax $\{S^2 = 0.970, \tau_e = 2048.0, R_{ex} = 0.149\}$

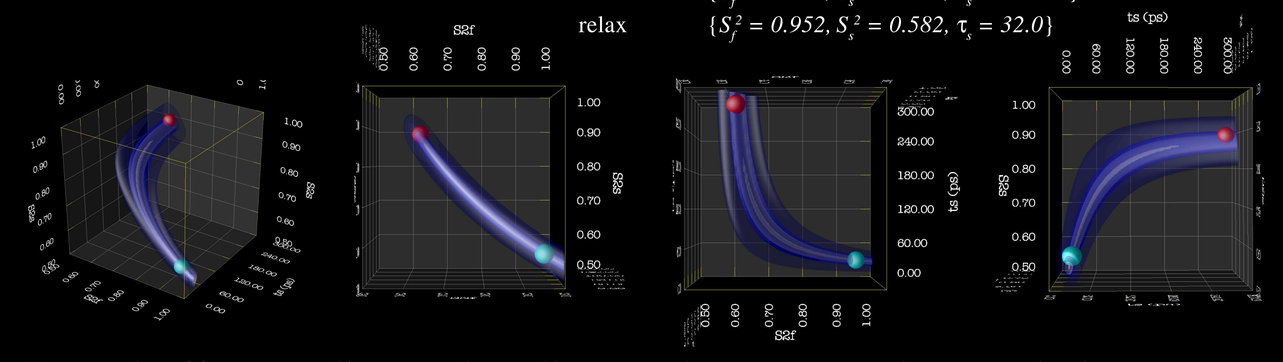


Incomplete minimisation

True values $\{S_f^2 = 0.952, S_s^2 = 0.582, \tau_s = 32\}$

Modelfree $\{S_f^2 = 0.622, S_s^2 = 0.893, \tau_s = 281.74\}$

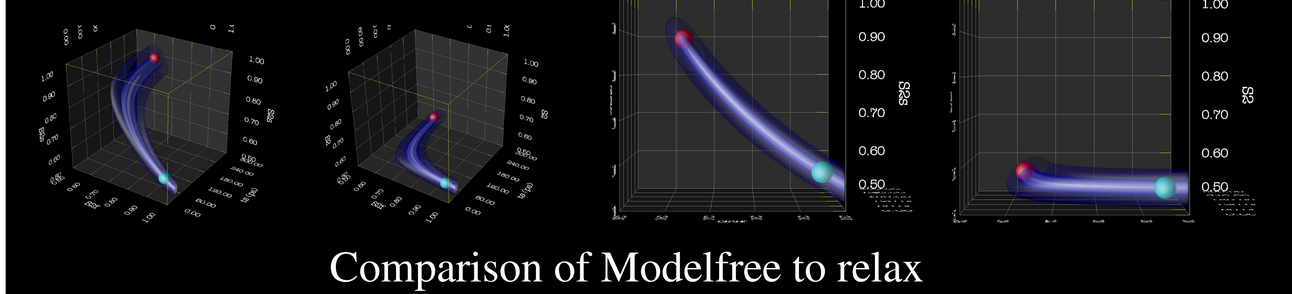
relax $\{S_f^2 = 0.952, S_s^2 = 0.582, \tau_s = 32.0\}$



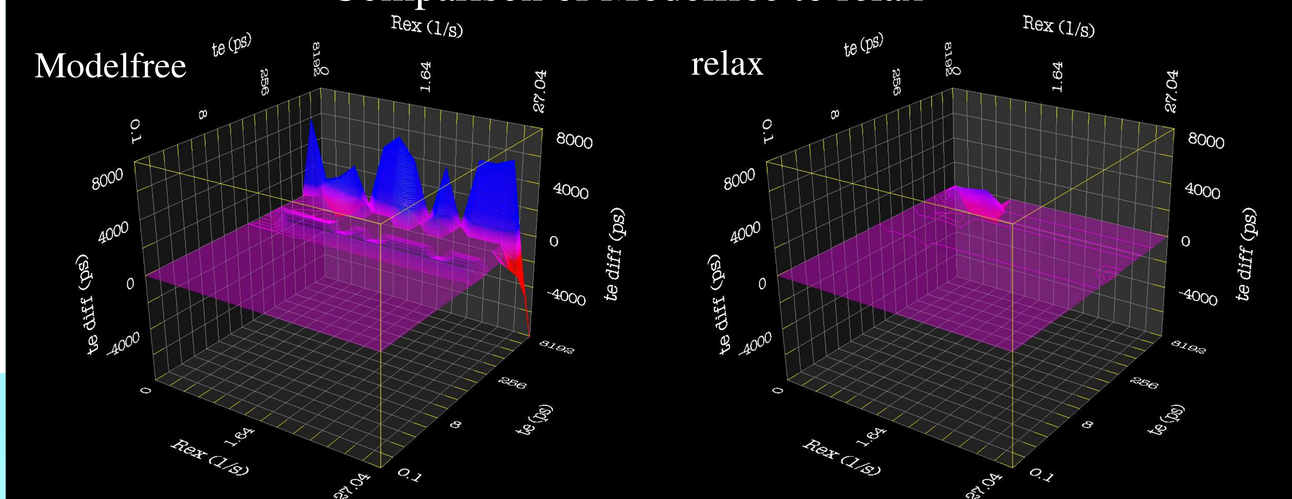
Round-off error, diagonal scaling, and parameter deconvolution

The efficiency and accuracy of minimisation was improved by eliminating round-off error, using diagonal scaling, and using parameter deconvolution.

$S^2 = S_f^2.S_s^2$



Comparison of Modelfree to relax



Difference surface for the model-free parameter τ_e. Relaxation data for 500 and 600 MHz were back calculated for R_{ex} and τ_e values specified by the axes and an S^2 value of 0.8. Modelfree and relax were used for model-free analysis, the difference value being between the final minimised τ_e value and the original value. For perfect results the surface should be flat with a value of zero at all points.

Conclusions

When analysing relaxation data, special care must be taken, for using the wrong tools will result in misleading, inaccurate, and maybe even an incorrect picture of dynamics. For model selection, AIC will significantly improve accuracy and will give a much clearer view of the dynamics of a protein. For minimisation it should be noted that for a large proportion of residues, the end results may not be fully minimised and therefore could be far from the true values.