Comments on Scoring Weights for NMR Peak Assignments – 09/03/18

We are using a variety of data to help assign NMR crosspeaks from sparsely labeled samples; NOEs, RDCs, Chemical shifts, PCSs. These data re compared to predictions based on structure to find the best assignment. Basically, a difference between crosspeak associated data and site associated predictions are turned into a "score" and a minimum in total score is sought to identify the best assignment. The total score is the sum of scores from individual data types. When different data types are combined in this way, it is important to "weight" scores so they make appropriate contributions to the final score.

There are several things to consider when weighting scores: An initial normalization to give them equal weights, the precision of the various measurements and the potential information content of the data type. To begin we choose to define most scores as a root mean square deviation between experiment and prediction, and then divide this by the possible range of observables to get a normalized score for each contribution.

$$S_{norm} = \sqrt{\sum_i (exp_i - pred_i)^2 \Big/ (n * range^2)} = S_{raw}/range$$

Here the sum is over the number of sites to be assigned, n is the number of sites, and the range can be estimated theoretically or simply by finding the maximum and minimum from the set of data. This works well for single value measurements such as chemical shifts, RDCs and PCSs. For NOEs this is not possible because for each site there are multiple NOE crosspeaks and it is not only their intensity that counts, but their position on a chemical shift scale. They are best represented as vectors that might be taken from columns rising from crosspeaks in a 3D NOESY-HSQC. We choose to use a Pearson correlation coefficient (R) that as a means of assessing the similarity of experimental and normalized vectors. R is a vector product of experimental and predicted vectors that runs from 0 to 1 with 1 for an identical pair. To get a score that minimizes at a perfect fit, we use (1-R) and combine values over sites as follows, where the range is 1.0.

$$S_{norm} = \sqrt{\sum_i (1 - R)^2 \Big/ (n * range^2)}$$

A factor has now to be introduced to account for precision of measurement. This is easily done for chemical shifts, RDCs and PCSs where error estimates are available. The precision factors in these cases are simply:

$$P = range/error$$

This too is a little more complicated for NOEs. We have developed a MATLAB script to estimate error in two steps; one for amplitude errors and one for peak position errors. Our NOE vectors all have an "autopeak" added at one extreme that has an intensity equal to the average maximum intensity for all NOE vectors and a width equal to that used for all other NOE peaks, which are usually added at appropriate chemical shifts from a peak list in the case of an experimental vector or a list of simulated NOE intensities and chemical shift predictions in the case of a predicted vector. If we assume one experimental vector is mostly noise (or we can construct a vector to simulate this) we can compare a

zero amplitude plus autopeak vector to all experimental vectors and take the minimum (1-R) value as an estimate of amplitude error. For ST6Gal1 one vector (peak 10) is mostly noise (see Figure 1A). If R values are found for all experimental vectors, this one will have the lowest R and one can just search for a minimum. The (1-R) value for crosspeak 10 is 0.15. For peak position errors we can simulate this by moving peaks in a representative spectrum by the estimated chemical shift error (0.2) and calculating the R value for the shifted in comparison to the unshifted vector. The difference from a perfect fit (1-R) gives an estimate of peak position error. The shifted and unshifted vectors for F29 are shown in Figure 1B. The (1-R) value for F29 is 0.28 and one for F171 is 0.27. Adding the amplitude and position estimates and taking the reciprocal give a *P* value of 2.4. Interestingly, this is very near the scaling value (5) used in our first publication using NOE data which came from comparing correct to incorrect scores for a test protein with known assignments.

Information content is also an important consideration. For example, we may measure NOEs or RDCs with very high precision, but all RDC vectors happen to point in roughly the same direction as they might for N-H vectors in an alpha helix, or NOEs might be very similar, as they might be in a cluster of methyl groups. We can assess these possibilities by looking at the actual distribution of scores (as squared differences) we might get by pairing every measurement with every prediction and calculating a standard deviation (std) for this distribution. Information content (I) for each measurement type can then be given as follows:

$$I = \sqrt{\frac{std}{range} * \frac{\#data}{\#sites}}$$

Note that for RDCs #data should be the number of actual data minus 5 for the number of order parameters that need to be determined. If I is taken inside the root used in scoring it becomes I^2 and the squared measurements in the scoring function are each weighted by the information content. We have written a MATLAB scripts to do this (see supplement). Figures 2a-2d show plots and give I values for 1H shifts, 15N shifts, Peg RDCs, and NOEs for data on ST6Gal1 in the legend.

Given the above considerations of normalization, precision and information content we can define a weighting factor for our various data types as follows:

$$S = S_{norm} * P * I = S_{raw} * \frac{I}{error} = S_{raw} * W$$

Note that the range factors in S_{norm} and *P* will cancel resulting in the simple scaling by error used in the initial version of ASSIGN_SPL. Only the addition of the I factor is new for chemical shifts and RDCs. For NOEs we now have a rationally developed error estimate (incorporated in P: range is 1.0) that turns out to be similar to one used previously. Assuming the square root of the mean square deviation (or 1-R) was used in all cases, only the I factors will make a difference and these don't vary much (0.42. 0.42, 0.34 and 0.44 for ^{1}H shifts, ^{15}N shifts, RDCs and NOEs, respectively).

The program ASSIGN_SLP_MD has now been revised to calculate S_{raw} without scaling by error. Error estimates are incorporated in the weighting factors, *W=I/error*, and these are entered along with data in input files. For ST6Gal1 the weights for ^{1}H shifts, ^{15}N shifts, Peg RDCs and NOEs are 1.05, 0.21, 0.034, and 1.06 respectively.

A B

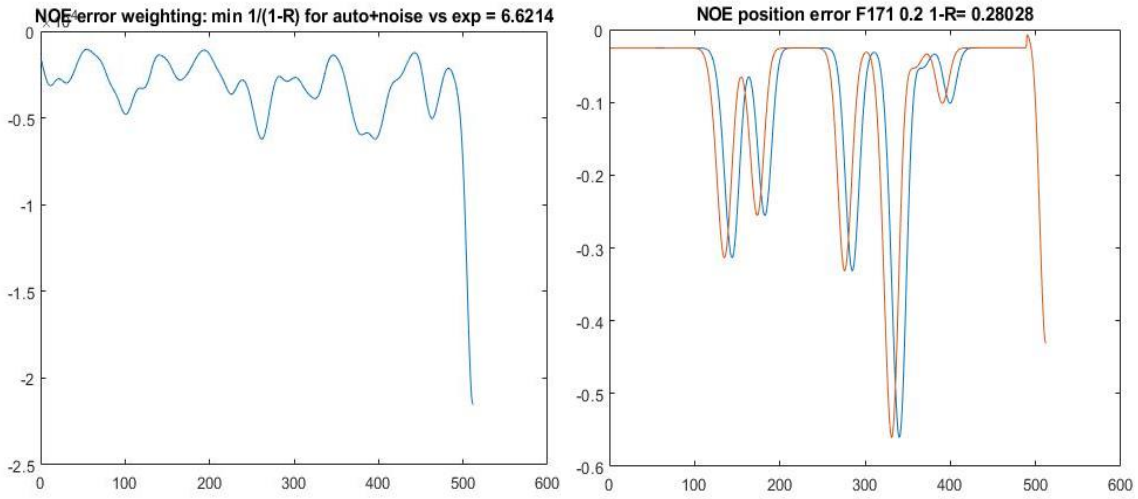
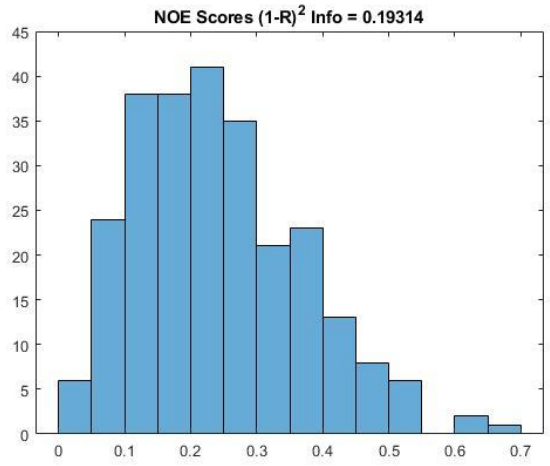
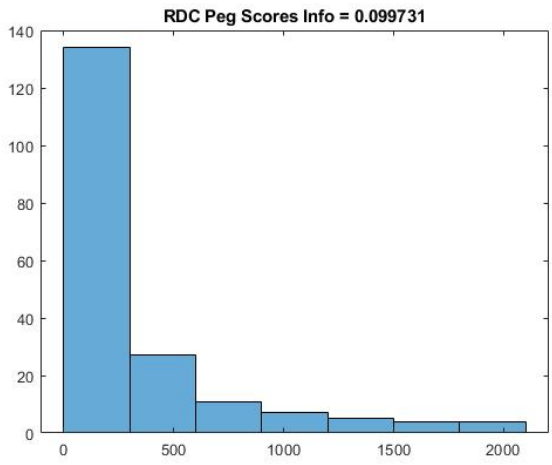


Figure 1. Extraction of NOE precision factor for ST6Gal1. (A) The minimum (1-R) value when comparing experiment to a vector with only an autopeak was for crosspeak 10. Its NOE vector is shown (1-R) = 0.15. (B) Predicted NOE vector for F29 and one shifted by 0.2 ppm. The 1-R value is 0.28. The combined P value is 2.3.

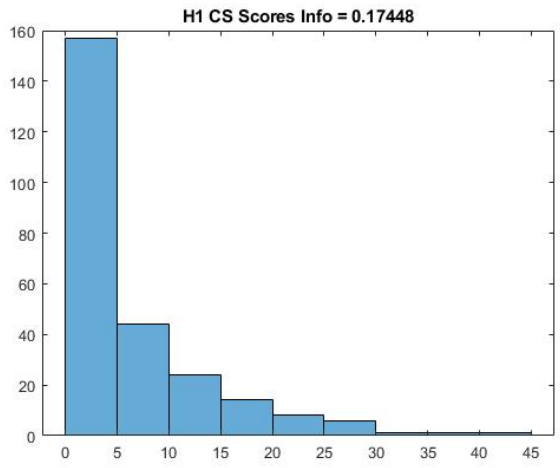
A



B



C



D

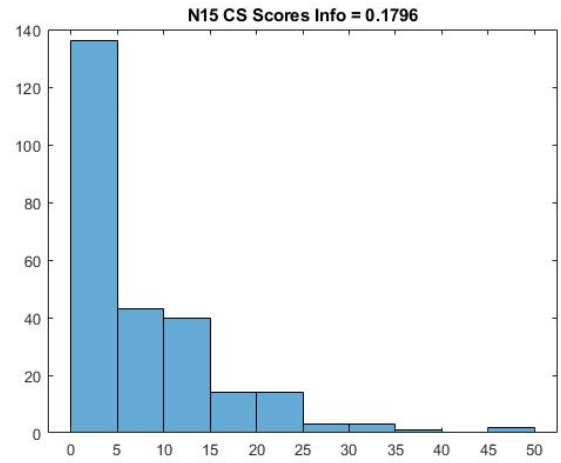


Figure 2. Estimation of information content for various data types. (A) 1H chemical shifts, (B) 15N chemical shifts, (C) NOEs and (D) Peg RDCs. Note that the plots are for I^2, not I. So, I factors are 0.42, 0.42, 0.44 and 0.32 respectively. Data ranges can also be extracted from these plots. They are: 6.7, 7.1, 1.0, 57.