Journal of Biomolecular NMR (2006) 36:189–198
DOI 10.1007/s10858-006-9082-1

Article

An automated assignment-free Bayesian approach for accurately identifying proton contacts from NOESY data

Ling-Hong Hung† & Ram Samudrala*
Department of Microbiology, University of Washington, Rosen Building,
960 Republican, Seattle, WA 98109, USA

Received 28 March 2006; Accepted 11 August 2006

Key words: automation, Bayes, bioinformatics, computation, probabilistic, protein structure, Protinfo, software, webserver

Abstract

The identification of proton contacts from NOE spectra remains the major bottleneck in NMR protein structure calculations. We describe an automated assignment-free system for deriving proton contact probabilities from NOESY peak lists that can be viewed as a quantitative extension of manual assignment techniques. Rather than assigning contacts to NOESY crosspeaks, a rigorous Bayesian methodology is used to transform initial proton contact probabilities derived from a set of 2992 protein structures into posterior probabilities using the observed crosspeaks as evidence. Given a target protein, the Bayesian approach is used to derive probabilities for all possible proton contacts. We evaluated the accuracy of this approach at predicting proton contacts on 60 ^{15}N separated NOESY and ^{13}C separated NOESY datasets simulated from experimentally determined NMR structures and compared it to CYANA, an established method for proton constraint assignment. On average, at the highest confidence level, our method accurately identifies 3.16/3.17 long range contacts per residue and 12.11/12.18 interresidue proton contacts per residue. These accuracies represent a significant increase over the performance of CYANA on the same data set. On a difficult real dataset that is publicly available, the coverage is lower but our method retains its advantage in accuracy over CANDID/CYANA. The algorithm is publicly available via the Protinfo NMR webserver http://protinfo.compbio.washington.edu/protinfo_nmr.

Introduction

The traditional manual approach to NOESY interpretation is to assign each crosspeak to one or more proton pairs. Sequential assignments (Wuthrich, 1986) rely upon contacts that are known to be consistently less than 5 Å apart. Main chain assignment strategies (Wand et al., 1991; Bailey-Kellogg et al., 2000) use similarly conserved contacts in secondary structure elements. For long range contacts, initial assignments rely heavily upon crosspeaks with chemical shifts that uniquely identify the protons involved. Assignments are confirmed through the presence of symmetry peaks and through the presence of previously identified correlated contacts, such as those between other proton pairs belonging to the same residue pair. Structural simulations based upon initial assignments can be used to resolve ambiguities in an iterative fashion (Mumenthaler et al., 1997; Nilges et al., 1997; Linge et al., 2001) until a high-resolution structure is obtained. Programs such as AUTOSTRUCTURE (Huang et al., 2005), and CANDID (Herrmann et al., 2002) incorporate these heuristic rules to assign NOE

*To whom correspondence should be addressed. E-mail: ram@compbio.washington.edu
†E-mail: lhhung@compbio.washington.edu

spectra and CYANA (Guntert 2004) integrates the iterative structural refinement process with the assignment process.

Rule-based deterministic methods are by nature simplifications of the underlying relationships. For example, sequential assignment rules are limited to predicting the three contacts per residue that are most highly conserved in experimentally determined structures even though there are many other proton pairs with distances that are only slightly less conserved (see Figure 1). However, extending the rules to predict more contacts would decrease accuracy and rule-based systems have to make compromises between accuracy and coverage. In contrast, Bayesian systems accept an initial estimate of the likelihood of any contact as input and rigorously transform it into a posterior contact probability that takes into account the information present in the crosspeaks and in other contact probabilities. The probabilistic output can be then be converted into a deterministic output by using a posterior probability cutoff to predict the contacts. The same high confidence predictions made by rule-based methods can also be made by such an approach. However, since all the inputs and all the spectral evidence can contribute to the posterior probability, more predictions with greater accuracy can be obtained using a Bayesian approach.

We describe a new Bayesian approach for the interpretation of NOESY data. We initially start with the contact probabilities for all possible types of proton pairs estimated from a set of 2992 structures solved by X-ray diffraction (Wang and Dunbrack, 2003). Our approach is assignment free: Instead of assigning each crosspeak to a particular contact (or to a small set of contacts), crosspeaks modify the contact probabilities of all proton pairs that could possibly give rise to the crosspeak. Thus, the procedure not only uses all the proton contact probabilities in the database of experimentally determined structures as input, but also utilizes the crosspeak information in an optimal fashion to generate final contact probabilities. Contact probabilities are calculated for all possible proton pairs present in the target protein, and not just for a small subset of assigned contacts. We demonstrate the accuracy of our approach with a set of 60 simulated ^{15}N separated NOESY and ^{13}C separated NOESY datasets, and a difficult real test case, and compare the performance to CANDID/CYANA. We also discuss some planned extensions of the methodology to allow the use of probabilistic chemical shift data and also allow the iterative refinement of the contact probability estimates using structural data.

Materials and methods

Overview of Bayesian methodology

The Bayesian approach consists of several stages which are described in detail in the subsections below. Briefly, we first obtain initial estimates of contact probabilities (prior probabilities) for all proton pairs from experimentally determined structures. Then the crosspeaks in the spectra are treated as evidence in a Bayesian formulation to generate posterior probabilities using the simple Bayesian relationship shown in (5). However, because symmetry crosspeaks are highly correlated, they must be treated as a unit and the Bayesian equation is modified as shown in (10). Similarly, the high correlation between contact probabilities between related protons must be accounted for. Finally, the first round of output posterior probabilities is used as input for a second round of Bayesian interpretation to take full advantage of the increase in accuracy of the contact estimates.

Generation of prior probabilities

A set of 2992 non-redundant structures solved using X-ray diffraction (Berman and Dunbrack, 2003) having less than 30% pairwise sequence identity to each other were obtained using curated lists from the culled PDB resource (Wang et al., 2003). Missing protons were generated using the program REDUCE (Word et al., 1999). The protons were grouped by residue, type (related methyl protons being treated as single type), secondary structure, and chain separation (≥5 residues apart being treated as one class) and the distances between all pairs were enumerated. The distances were binned using 0.5 Å intervals to derive a distance probability distribution for each proton pair. We then assumed that only proton distances ≤5 Å are observable and calculated the probabilities by summing the counts in the bins for distances ≤5 Å and dividing by the total number of counts for all distance bins. Initial contact probabilities were generated for all nine

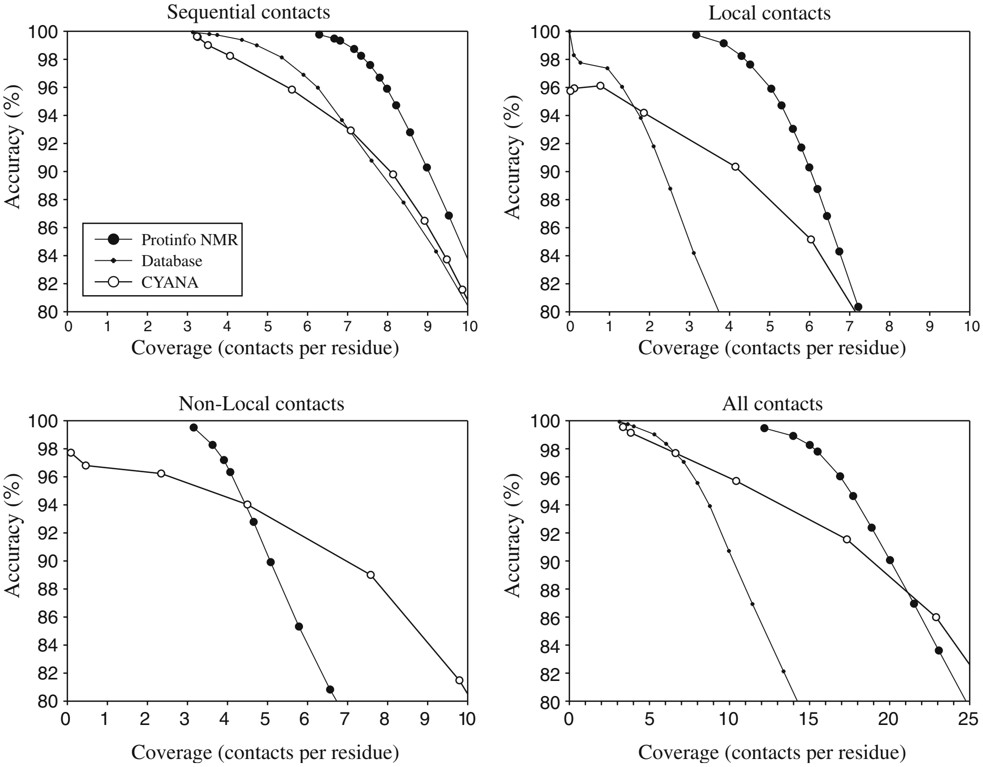

Figure 1. Comparison of the accuracy of contact predictions for 60 simulated NOESY datasets made using Protinfo NMR posterior probabilities, database prior probabilities and CYANA confidences. For each type of contact, the accuracy is shown as a function of the number of contacts (coverage) predicted at a particular confidence cutoff. Contacts were considered to be correctly predicted when they were less than 5 Å apart in the source structure. Overall, our Protinfo NMR methodology is extremely accurate, with 12.17 interresidue contacts per residue predicted at 99.5% accuracy. For sequential and local contacts, the prior probabilities predict contacts more accurately than CYANA at the highest confidences and the posterior probabilities are more accurate than CYANA regardless of the coverage. For non-local contacts, CYANA provides greater coverage but with a broad reduction in accuracy compared to Protinfo NMR, which peaks at 99.5% accuracy when identifying 3.17 contacts per residue.

combinations of secondary structure states for each protein pair. The likelihoods of the secondary structure states were then estimated using our secondary structure prediction program PsiCSI (Hung et al., 2003a), which predicts secondary structure from chemical shift and sequence with an average accuracy of nearly 90% on a large benchmark set. These likelihoods were used to weight the initial probabilities to generate a final database derived prior contact probability.

Bayesian generation of posterior probabilities

Bayes theorem can be expressed as:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B})} \tag{1}$$

where $P(\mathrm{A}|\mathrm{B})$ is the probability of an event A given the observation B, $P(\mathrm{B}|\mathrm{A})$ is the likelihood of B given A, $P(\mathrm{A})$ and $P(\mathrm{B})$ are the prior probabilities of event A and observation B. Essentially, this states that if we can narrow the set of outcomes to those that result in the observation B, the probability of event A is increased by contraction of the probability space as given by the ratio:

$$\frac{P(\mathrm{B}|\mathrm{A})}{P(\mathrm{B})} \tag{2}$$

For NOE interpretation, event A is a proton contact, and observation B is a crosspeak with coordinates consistent with the chemical shifts of the protons. We assume that the crosspeak will always form if there is a contact, so the term $P(\mathrm{B}|\ \mathrm{A})$ is unity. This gives us:

$$P_{\text{posterior}} = \frac{P_{\text{prior}}}{P_{\text{crosspeak}}} \tag{3}$$

where $P_{\text{posterior}}$ is the posterior probability of the contact (the estimate of contact probability given the crosspeak), P_{prior} is the prior probability (the estimate of contact probability obtained from the database), and $P_{\text{crosspeak}}$ is the prior estimate of the probability of the crosspeak (the normalization factor accounting for the restriction of probability space). Assuming that the crosspeak arises from contacts from at least one of the proton pairs with consistent chemical shifts we obtain:

$$P_{\text{posterior}} = \frac{P_{\text{prior}}}{1 - \prod_{i=1}^{n}(1 - P_i)} \tag{4}$$

where $\{P_i\}$ is the set of prior probabilities of all the possible contacts that could result in the crosspeak. Equation (4) handles the idealized situation, but the crosspeak can also be a spectral artifact. Even when this cannot be estimated, a low value for the artifact probability prevents absolute unity probabilities which can otherwise cause some numerical instabilities. Thus the final equation we use is:

$$P_{\text{posterior}} = \frac{P_{\text{prior}}}{1 - (1 - P_{\text{artifact}})\prod_{i=1}^{n}(1 - P_i)} \tag{5}$$

The information present in the crosspeak is manifested in the restriction of the contact possibilities to proton pairs within a narrow range of chemical shifts. The resulting restriction in probability space is what amplifies the signal present in the prior probabilities to generate the posterior probabilities.

Posterior probabilities for related crosspeaks

Generally, a proton contact will result in more than one crosspeak, with the most common case being that of symmetry peaks that are found either within a single experiment or between different experiments. It is possible due to differences in the magnetization transfer and relaxation pathways that one of the symmetry related peaks will not be observed. In this case, the probability of a contact is multiplied by the probability that a symmetry crosspeak will be missing. This probability is provided by the user or estimated from the frequency of missing crosspeaks from conserved contacts. When both symmetry crosspeaks are present, one can derive two separate posterior probabilities by treating them separately and then combine them assuming that the observation of the crosspeaks are independent events. However, since the observation of symmetry pairs is highly correlated this approach is inaccurate.

A more accurate method is to treat both crosspeaks as a single entity. For a pair of symmetry related peaks A and B, the protons pairs are separated into three sets: pairs that could contribute to both peaks A and B, pairs that contribute to only peak A, and pairs that contribute to only peak B. To calculate the Bayesian posterior probabilities for a pair of symmetry peaks, we need to calculate the prior probability of observing both peaks. We first calculate the probability that both peaks will be observed due to a contact that would form both peaks simultaneously. This is the denominator in Eq. (5):

$$1 - (1 - P_{\text{artifact}})\prod_{i=1}^{n}(1 - P_i) \tag{6}$$

where $\{Pi\}$ is the set of prior probabilities for contacts that could contribute to both peaks. The other scenario is that both peaks are formed by separate contacts that could form peak A or peak B, which is the product of Eqs. (7) and (8)

$$1 - (1 - P_{\text{artifact}})\prod_{\text{A}=1}^{j}(1 - P_{\text{A}}) \tag{7}$$

$$1 - (1 - P_{\text{artifact}})\prod_{\text{B}=1}^{k}(1 - P_{\text{B}}) \tag{8}$$

The overall probability is the probability that at least one contact forms both peaks simultaneously (given in Eq. (6)) plus the probability that this does not happen (1–6) multiplied by the probability that two separate contacts to peaks A and B form (the product of Eqs. (7) and (8)). If we assume that the artifact probabilities are zero, then the probability of observing both peaks simplifies to:

$$1+\left[\prod_{i=1}^{n}(1-P_i)\right]\cdot\left[\left[\prod_{\mathrm{A}=1}^{j}(1-P_{\mathrm{A}})\right]\cdot\left[\left[\prod_{\mathrm{B}=1}^{k}(1-P_{\mathrm{B}})\right]-1\right]-\prod_{\mathrm{B}=1}^{j}(1-P_{\mathrm{B}})\right] \tag{9}$$

The Bayesian posterior probabilities can be calculated as:

$$P_{\text{posterior}}=\frac{P_{\text{prior}}}{1+\left[\prod_{i=1}^{n}(1-P_i)\right]\cdot\left[\left[\prod_{\mathrm{A}=1}^{j}(1-P_{\mathrm{A}})\right]\cdot\left[\left[\prod_{\mathrm{B}=1}^{k}(1-P_{\mathrm{B}})\right]-1\right]-\prod_{\mathrm{B}=1}^{j}(1-P_{\mathrm{B}})\right]} \tag{10}$$

In the case of symmetry peaks, the signal is further amplified by the greater restriction of probability space since the probability of both peaks being observed is always less or equal to the probability of either peak being observed. In terms of the Bayesian equations, the denominator for the symmetry peak in Eq. (10) is always less or equal to the denominators of the equations for the individual peaks (given in Eq. (5)).

Posterior probabilities for correlated proton contacts (vicinal pairs)

In our Bayesian equations, we make the assumption that proton contact probabilities are independent. This is not always true. The most important exception is that of vicinal protons (protons that share the same heavy atom). Not only are these contacts highly (but not absolutely) correlated but they also tend to have similar chemical shifts and are often involved in the interpretation of the same crosspeaks. When this is the case, we need to treat them as a single entity (i.e. a contact to either vicinal proton constitutes a contact), for the purposes of the denominator of Eq. (5). The database derived probability that either vicinal proton is in contact with a given proton can be calculated directly from the observed proton contact frequencies from experimentally determined structures. However, it is often convenient and more accurate to treat the proton contact probabilities separately, such as when using structural consensus to generate priors or when converting posterior contact probabilities to constraints. By estimating the correlation coefficient for vicinal proton contacts from the observed correlation in our protein database, we can calculate the joint probability from the individual vicinal contact probabilities and vice versa.

Iterative refinement of posterior probabilities

The quality of the posterior probabilities obtained is affected by the accuracy of the related priors. Our initial estimates of contact probabilities were based upon probabilities observed in known proteins. The posterior probabilities are more accurate estimates of contact probabilities than these original database-derived probabilities since they reflect the information present in the crosspeaks. This is especially true for non-local prior contact probabilities, since non-local contacts vary more from protein to protein than local and sequential ones. Using the more accurate posterior probabilities from the first round of predictions as prior probabilities for input into a second round of predictions, we can increase the accuracy of the contact predictions made using the posterior probabilities.

Simulation of NOESY spectra

The first models of 60 NMR structures were used to generate simulated ^{15}N separated NOESY and ^{13}C separated NOESY peak lists. Proton pairs less than 5 Å apart in the structures were enumerated and initial crosspeak lists generated from the chemical shift assignments for these contacts. Crosspeaks that were closer than the overlap cutoff (0.05 ppm for protons, 0.4 ppm for heteroatoms) were grouped together and the new coordinates of the overlapped crosspeak were calculated from the average coordinates of the group. This process was repeated until no two crosspeaks were closer than the overlap cutoff. This resulted in differences (up to 0.06 ppm in the proton dimensions) between the final crosspeak chemical shifts and the protons used to generate the crosspeaks.

Determination and comparisons of accuracy

A proton contact prediction is considered correct when the corresponding distance in the experimental structure is ≤5 Å. Methyl and vicinal protons are considered as a single group and a prediction is considered correct when the distance to any proton of the group is ≤5 Å. For real datasets, a prediction is considered correct when the contact is present in at least 90% of the models.

For comparative evaluation of our approach on 60 datasets, we performed simulations using CYANA v2.1, an established tool for NOESY assignment and structure calculation. The default settings for the 7-cycle assignment/simulation protocol were used. Cycles after cycle 1 use structural ensembles to iteratively resolve ambiguities in the NOE assignments we compared so the assignment confidence values were read from the cycle1.noa file.

For comparison on a real test case we used the dataset deposited for 1se9 in the Biological Magnetic Resonance Bank (BMRB) (Seavey et al., 1991; Vinarov et al. 2004), http://www.bmrb.wisc.edu/data_library/timedomain/1/bmr612. Assignment confidence values for CANDID (from CYANA v1.06) were obtained from the cycle1.ass file, and were used to obtain the results shown in Figures 1 and 2.

Results

Accuracy evaluation and comparison to CYANA on simulated datasets

Similar to the posterior contact probabilities generated by our Bayesian method, CYANA provides confidences for each contact prediction. The accuracy of predictions made by CYANA with confidence greater then a given cutoff can be compared to percentage of observed contacts with a posterior contact probability greater than a given cutoff for the Bayesian method. The results will be heavily dependent on the cutoffs chosen. For both methods, there is a tradeoff between accuracy and number of contacts predicted (coverage) that is a function of the cutoff confidence or probability. Plotting the accuracy versus coverage as a function of different confidences allows a more comprehensive comparison of the methods.

Figure 1 compares the accuracy and coverage for contact predictions made using the initial database prior probabilities, the Protinfo NMR Bayesian posterior probabilities, and the CYANA assignment confidences for 60 simulated 15N separated NOESY and 13C separated NOESY peak lists. A comparison of predictions made using only database information and those made using our Bayesian approach shows that there is a significant enrichment of signal, i.e. our approach is effective at transforming the prior probabilities using the information present in the peak lists. Regardless of the accuracy of prior probabilities, the resulting posterior probabilities are superior predictors of proton contacts.

For sequential and local contacts, the prior probabilities predict contacts more accurately than CYANA at the highest confidences and the posterior probabilities are more accurate than CYANA regardless of the coverage. For non-local contacts, CYANA provides greater coverage but with a broad reduction in accuracy compared to Protinfo NMR, which peaks at 99.5% accuracy when identifying 3.17 contacts per residue. At the same coverage level, CYANA is approximately 95% accurate. When all types of contacts are considered, Protinfo NMR predicts 12.17 interresidue contacts per residue at 99.5% accuracy.

Accuracy evaluation and comparison with CANDID/CYANA on 1se9

1se9 is a 101 residue ubiquitin-fold protein (Vinarov et al., 2004) solved in a semi-automated manner using CYANA. It is unique in that the time-domain data, complete peak lists and intermediate assignments and structures are all publicly available in the BMRB http://www.bmrb.wisc.edu/data_library/timedomain/1/bmr6128. Furthermore, this is a difficult case since automated methods such as CANDID/CYANA did not converge when used in an automated manner (average RMSDs of 14.2 Å for CANDID/CYANA v1.06 and 15.2 Å for CYANA v2.1 for residues 10–99) and the structure was solved only after months of manual intervention.

Figure 2 compares the accuracy and coverage for contact predictions made for 1se9 using the Protinfo NMR Bayesian posterior probabilities, and the CYANA and CANDID assignment confidences. Even though the absolute coverage is

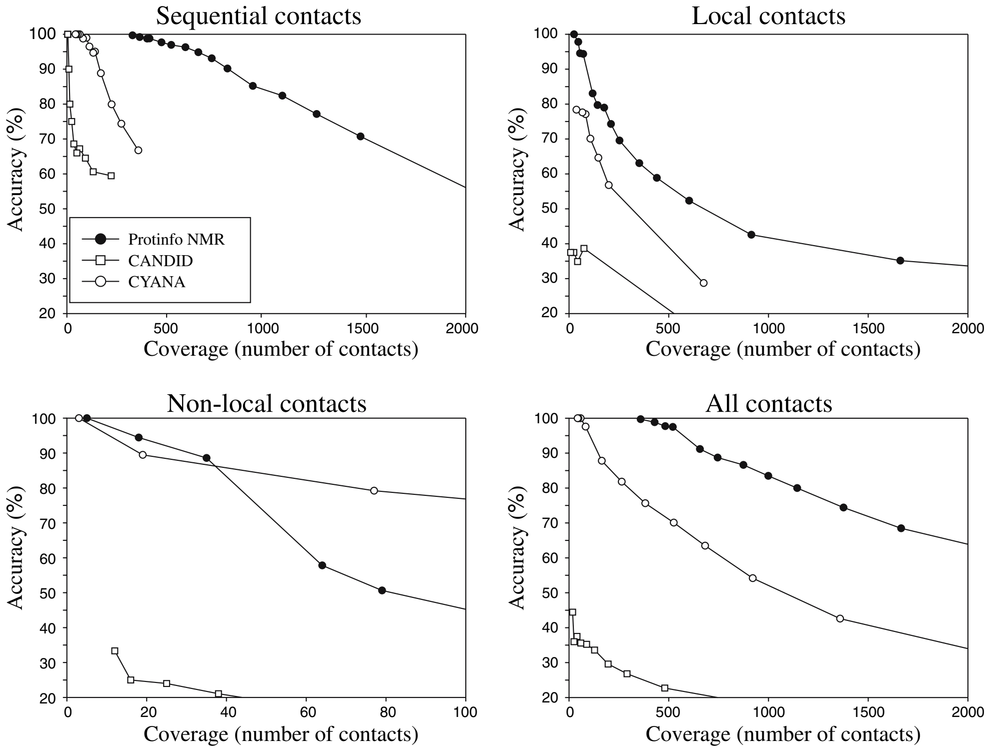

Figure 2. Comparison of the accuracy of contact predictions for a real dataset (1se9) made using Protinfo NMR, CANDID and CYANA. For each type of contact, the accuracy is shown as a function of the number of contacts (coverage) predicted at a particular confidence cutoff. Contacts were considered to be correctly predicted when they were less than 5 Å apart in 90% of the models in the experimental structure ensemble. Even though the number of contacts predicted on this difficult dataset is lower compared to the number predicted for the simulated datasets, the accuracy of Protinfo NMR remains very high, with 359 and 520 interresidue contacts predicted at 99.7% and 97.5% accuracy. In contrast, the accuracy for CYANA is 80% and 70% at the same coverage levels. The retention of accuracy indicates that noisiness of the real data is properly reflected in the posterior probabilities so that smaller numbers of very accurate predictions can still be made on ambiguous and incomplete datasets, highlighting an advantage of using our Bayesian approach.

lower, the trends observed in the accuracy-coverage analyses for 1se9 data are similar to those observed for the simulated data. For identification of sequential and local proton contacts, the Protinfo NMR Bayesian approach identifies more contacts with greater accuracy than CYANA and CANDID. The improvement in local contact prediction is significant since 1se9 is mostly a sheet protein and lacks many of the traditional canonical helical specific local contacts. For the non-local contacts, the Bayesian method is superior to CANDID and is more accurate than CYANA in the predicting contacts at high cutoffs. As was the case with the simulated data, CYANA is able to identify more contacts at lower confidences though these were not accurate enough for the structures to converge. Although the number of predictions made is much lower, the overall accuracy of the Bayesian method remains very high, with 359 and 520 interresidue contacts predicted at 99.7% and 97.5% accuracy. In contrast, the accuracy for CYANA is 80% and 70% at the same coverage levels. The retention of accuracy indicates that noisiness of the real data is properly reflected in the posterior probabilities so that smaller numbers of very accurate predictions can still be made on ambiguous and incomplete datasets, highlighting an advantage of using our Bayesian approach.

Discussion

Improvements to the Bayesian NOE interpretation approach

Our Protinfo NMR Bayesian methodology is an effective and accurate approach to interpreting

Table 1. Example of the conversion of prior to posterior probabilities

(A) Chemical shifts

	H	HN	N
Peak	1.235	9.738	130.97
43 VAL QG1	1.256	–	–
59 LEU HB3	1.242	–	–
8 ILE H	–	9.748	130.95

(B) Contact probabilities

Contact	Contact probability estimates		
	True distance	Prior	Posterior
43 VAL QG1 to 8 ILE H	4.586 Å	0.038	0.748
59 LEU HB3 to 8 ILE H	7.959 Å	0.002	0.046

An example taken from a real 1H-1H 15N separated NOESY is shown. For the given peak there are two possible protons with shifts (A) that match the aliphatic shift and one that matches the amide shift giving rise to two possible contacts consistent with the peak. Our method uses observed contact frequencies in known proteins to derive initial prior estimates (B). Because both possible contacts are non-local the absolute values of the prior contact probabilities are low. However, the VAL contact almost 20 times more likely than the LEU contact. Given the evidence that there is a peak that can only be formed by these two contacts – the posterior probability of the (correct) VAL contact becomes much higher than the initial estimate. If we had used differences between peak chemical shift between proton shift, as an estimator as most other methods do, the incorrect LEU contact would have been favored. Probabilities do not add up to unity because they are not mutually exclusive and an estimate of the likelihood that the peak is an artifact is included in the calculations (see Eq. (5)). Note that the method only deals with the contact probabilities and their transformation and makes no statement about which contact should be assigned to the NOE.

NOESY spectra and outperforms CANDID/CYANA in accurately identifying proton contacts at high confidence levels. Our approach, while rigorous and accurate, still has much room for improvement. We are refining the implementation to take into account peak volumes, chemical shift differences between assignments and crosspeaks and variations in relaxation and exchange. We are also exploring expanding the use of contact probabilities from correlated proton contacts similar to what is done by CYANA, JIGSAW (Bailey-Kellogg et al., 2000), and BACUS (Grishaev and Llinas, 2004). We are experimenting with this approach not only for vicinal proton pairs but for other significant correlations, such as between proton contacts from the same residue, and proton contacts in canonical patterns used to identify secondary structure. This type of inference transfer through contact correlations allows predictions to be made from indirect evidence when the crosspeak information is limited due to overlap and thus expands the ability of the Bayesian approach to make inferences with noisy, ambiguous, or sparse data.

Extensions and applications to aid structure determination

Currently, our interpretation engine returns posterior probabilities that can be converted to constraints and contact predictions, which can be used to aid manual assignments or to provide more accurate input for automated assignment/structure refinement software such as CYANA and ARIA (Nilges et al., 1997). The greatly increased number of contact predictions that can be made and the high accuracy of predictions will be especially useful for difficult cases.

The Bayesian framework can easily be extended to encompass and integrate the other stages of NMR data processing. Prior probabilities can be derived from any source including manual assignments and structural consensus. We are integrating our approach with structure simulation for iterative interpretation of NOESY peak lists where the posterior contact probabilities are converted to constraints used to generate a set of structures. The observed proton contact frequencies of the simulated structures then provide the input prior probabilities for the next round of refinement. In contrast to methods such as ARIA which obtains a list of assignments to refine and otherwise never revisit the original assignment process, using structure based contact frequencies as input priors for Bayesian refinement involves re-interpretation of the entire peak list. Checking against the entire dataset reduces the likelihood of simulation artifacts being passed onto subsequent rounds and should increase the accuracy of the final structures. Furthermore, the initial rounds of iteration can use structures calculated using hybrid NMR/structure prediction methods. These methods take advantage of knowledge-based energy functions and efficient sampling techniques to greatly reduce the number of contacts required to calculate an initial set of medium resolution structures suitable for further refinement (Rohl and Baker, 2002; Li et al., 2003; Hung and Samudrala, 2006). These methods require a few but highly accurate contacts, which can be provided by our Bayesian approach even with noisy or ambiguous data.

Applications of the Bayesian approach generally involve converting the probabilistic output to a deterministic form such as contact predictions or constraints. The most interesting applications take advantage of the fact that both inputs and outputs are probabilistic. For example, incorporating probabilistic chemical shift assignments from programs such as SPI (Grishaev and Llinas, 2002) can be accomplished by multiplying the prior contact probabilities in the Bayesian equations by the assignment probability and summing over all assignment scenarios. This rigorously propagates the uncertainty only to the affected posterior contact probabilities allowing accurate inferences to be made from those contact probabilities unaffected by the chemical shift ambiguities. The ability of different Bayesian approaches to be linked enables the entire data processing pipeline to be integrated with iterative structure refinement, allowing the uncertainties in the data to be iteratively resolved or reflected quantitatively as uncertainty in the final structure (Rieping et al., 2005).

Computational details and availability of software

The software is accessible through the Protinfo NMR server http://protinfo.compbio.washington.edu/protinfo_nmr (Hung and Samudrala 2003b; Hung et al. 2005). Chemical shifts, peak lists can be submitted and the posterior contact probabilities returned along with assignments in XEASY format suitable for use with CYANA. The webserver also accepts structures as optional input which may be used as an alternative source of prior contact probabilities. This allows the use of the output contact probabilities and assignments for iterative refinement. Results are typically returned within 5–10 minutes.

Acknowledgements

This work was supported in part by a Searle Scholar Award, a NSF CAREER award (IIS-0448502), NSF grant DBI-0217241, and NIH grant GM068152-01.

References

Bailey-Kellogg, C., Widge, A., Kelley, J.J., Berardi, M.J., Bushweller, J.H. and Donald, B.R. (2000) *J. Comput. Biol.*, **7**, 537–558.

Berman, H., Henrick, K. and Nakamura, H. (2003) *Nat. Struct. Biol.*, **10**, 980.

Grishaev, A. and Llinas, M. (2002) *J. Biomol. NMR*, **24**, 203–213.

Grishaev, A. and Llinas, M. (2004) *J. Biomol. NMR*, **28**, 1–10.

Guntert, P. (2004) *Methods Mol. Biol.*, **278**, 353–378.

Herrmann, T., Guntert, P. and Wuthrich, K. (2002) *J. Mol. Biol.*, **319**, 209–227.

Huang, Y.J., Moseley, H.N., Baran, M.C., Arrowsmith, C., Powers, R., Tejero, R., Szyperski, T. and Montelione, G.T. (2005) *Methods Enzymol.*, **394**, 111–141.

Hung, L.H., Ngan, S.C., Liu, T. and Samudrala, R. (2005) *Nucl. Acids Res.*, **33**, W77–80.

Hung L.H.. and Samudrala R. (2006) submitted.

Hung, L.H. and Samudrala, R. (2003a) *Protein Sci.*, **12**, 288–295.

Hung, L.H. and Samudrala, R. (2003b) *Nucl. Acids Res.*, **31**, 3296–3299.

Li, W., Zhang, Y., Kihara, D., Huang, Y.J., Zheng, D., Montelione, G.T., Kolinski, A. and Skolnick, J. (2003) *Proteins*, **53**, 290–306.

Linge, J.P., O'Donoghue, S.I. and Nilges, M. (2001) *Methods Enzymol.*, **339**, 71–90.

Mumenthaler, C., Guntert, P., Braun, W. and Wuthrich, K. (1997) *J. Biomol. NMR*, **10**, 351–362.

Nilges, M., Macias, M.J., O'Donoghue, S.I. and Oschkinat, H. (1997) *J. Mol. Biol.*, **269**, 408–422.

Rieping, W., Habeck, M. and Nilges, M. (2005) *Science*, **309**, 303–306.

Rohl, C.A. and Baker, D. (2002) *J. Am. Chem. Soc.*, **124**, 2723–2729.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.

Vinarov, D.A., Lytle, B.L., Peterson, F.C., Tyler, E.M., Volkman, B.F. and Markley, J.L. (2004) *Nat. Methods*, **1**, 149–153.

Wand, A.J. and Nelson, S.J. (1991) *Biophys. J.*, **59**, 1101–1112.

Wang, G. and Dunbrack, R.L. Jr. (2003) *Bioinformatics*, **19**, 1589–1591.

Word, J.M., Lovell, S.C., Richardson, J.S. and Richardson, D.C. (1999) *J. Mol. Biol.*, **285**, 1735–1747.

Wuthrich K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley and Sons, Inc.