Improved Bayesian Scoring Schemes for Protein NMR Backbone Resonance Sequential Assignment

James Wagner, Theodore Tegos, Xiang Wan and Guohui Lin*

Department of Computing Science, University of Alberta. Edmonton, Alberta T6G 2E8, Canada

Email: James Wagner - wagner@cs.ualberta.ca; Theodore Tegos - tegos@cs.ualberta.ca; Xiang Wan - xiangwan@cs.ualberta.ca; Guohui Lin*- ghlin@cs.ualberta.ca;

*Corresponding author

Abstract

Background: Accurately quantifying the signature information of chemical shifts provides a foundation for accurate and complete sequential resonance assignment in protein NMR spectroscopy. A nearly complete assignment is a prerequisite for three dimensional protein structure calculation.

Methods: A number of filtering steps are applied to construct two training datasets using known protein NMR data for learning scoring schemes to quantify the signature information. The scoring schemes are learned through a naive Bayesian method to use both intra-residue and inter-residue chemical shifts and to use the intermediate neural network output from the secondary structure predictor PsiPred.

Results: Two training datasets ALL and HOMO for scoring scheme learning were carefully constructed. Based on these two datasets, a total of 16 scoring schemes were proposed and examined. An extensive simulation study was set up to validate these scoring schemes and the one that performed the best was implemented into a web server, which is publicly accessible.

Conclusions: Through the extensive simulation study we found that the currently known protein NMR data is quite evenly distributed in terms of protein homology, and therefore homology removal in training dataset construction wouldn't gain a lot in the overall performance of the resultant scoring schemes. Also, we conclude that in general a naive Bayesian learning is better than a trivial distribution assumption. We believe this conclusion holds not just in our care but also for similar applications where the training data size is large. Another conclusion is that in applications where PsiPred prediction results are used as intermediate input, using its intermediate neural network output could be a better choice than using its the final prediction result.

Background

Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography are two key technologies that experimentally determine protein three dimensional structure. In protein NMR, structural restraints recorded in the NMR spectra have to be mapped to the target amino acid sequence and corresponding neighboring protons (via NOEs) in order to calculate the three dimensional structure. Such a mapping could theoretically be done directly, but in practice, due to the low accuracy and redundant of NOE data, the mapping is done through guidance from the backbone resonance sequential assignment, whose goal is to associate multiple spectral peaks of backbone chemical shifts to their corresponding amino acid residues. It is recognized that a nearly complete assignment is a prerequisite to the three dimensional structure calculation, where “complete” means that all true spectral resonance peaks must be identified and must find their corresponding amino acid residues.

A spectral resonance peak may be recorded as a two dimensional vector of chemical shifts, such as an HSQC peak, or a three dimensional vector of chemical shifts, such as a CBCA(CO)NH peak, where the chemical shift values have a one-to-one correspondence to the nuclei being measured in the NMR experiment. Typically, these interacting nuclei are at most three bonds away and reside in either a common amino acid residue or two adjacent residues. Ideally, it is expected that each peak corresponds to a unique nucleus or a unique amino acid. In practice, however, because some nuclei have very close chemical shift values and because of varying experimental conditions and instrumental errors, matching specific peaks to specific nuclei becomes non-trivial. Sequential assignment is formally defined as the process of mapping the resonance peaks to their corresponding nuclei, mostly through using individual chemical shift *signature information*. For each individual chemical shift value, whose corresponding nucleus type is easily known, the signature information it contains generally refers to the chemical and structural environment that its corresponding nucleus is in.

Besides the amino acid residue types inferred by the chemical shift values, there are two pieces of signature information that can be useful in an accurate (heteronuclear) sequential assignment. One is to correlate the chemical shift values to the secondary structure types that their corresponding residues are in. The observation underlying this correlation is that, for a common type of nuclei, their chemical shift values are affected not only by the corresponding amino acid residue types, but also by the types of secondary structure that the residues are in [1]. In other words, both residue type and secondary structure type are structural factors that affect the chemical shift values. Such an observation has also been validated by the new training datasets ALL and HOMO constructed in this work, to be detailed in the "Training Datasets" section, in which, for example, the means of carbon alpha (CA) chemical shifts in Alanines differ significantly across three different types of secondary structure. In dataset ALL, the means in coils (C), alpha helices (H), and beta sheets (E) are 52.506ppm, 54.806ppm, and 50.811ppm, respectively. The second fundamental part is to collectively use all available individual chemical shift signature information, an idea similar to *boosting* in machine learning that combines the individual chemical shift signature information to make better inferences. To this purpose, the common practice in (heteronuclear) sequential assignment is to map multiple spectral peaks to the HSQC peaks, such that peaks sharing common hydrogen (HN) and nitrogen (N) chemical shifts are grouped together to form super-vectors of chemical shifts. These super-vectors are generally referred to as *spin systems*. Note that a spin system contains some chemical shifts for nuclei in the same residue to which the hydrogen and the nitrogen nuclei belong, as well as other chemical shifts for nuclei in the preceding residue. For convenience, in this paper we refer to them as *intra-residue* and *inter-residue* chemical shifts, respectively [2]. Through the identification of spin systems, resonance sequential assignment becomes a mapping of spin systems to their corresponding residues in the target amino acid sequence. This paper is on effective mining the signature information for spin systems, where the signature information is the collective sum of the signature information of the chemical shifts in the spin system and it contains both the residue type information and the secondary structure type information.

There are a number of existing studies on how to group multiple spectral peaks into spin systems. Interested readers might refer to [3, 4] for more detailed descriptions. This paper is centered around designing scoring schemes to effectively quantify the chemical shift signature information for mapping spin systems to amino acid residues. Therefore, we do not intend to get into the issues of detailed peak grouping. Nonetheless, we will use three NMR spectra HSQC, CBCA(CO)NH, and HNCACB to briefly address how peak grouping is done. Subsequently, we will use the spin systems formed out of these three spectra to demonstrate the scoring scheme design. We point out that the stated scoring scheme design procedure is not limited to this typical combination of NMR spectra (or equivalently, chemical shifts), but is applicable to any combinations as long as they are theoretically sufficient for sequential assignment.

We assume the ideal case, in which an HSQC spectrum contains one peak (HN, N) for every pair of amide nitrogen (N) and its directly attached hydrogen (HN). That is, there is one peak for every amino acid residue, except Prolines (which don't have HN). The CBCA(CO)NH spectrum contains resonance peaks (HN, N, CA) and (HN, N, CB), where N is the amide nitrogen and HN is the directly attached hydrogen, CA and CB are the carbon alpha and the carbon beta residing in the residue that precedes the residue to which HN and N belong. For convenience, they are denoted as (HN_i, N_i, CA_{i-1}) and (HN_i, N_i, CB_{i-1}) to reflect the fact that if HN and N are in the i-th residue, then CA and CB are in the $(i-1)$-th residue. Using similar notations, the HNCACB spectrum contains four types of resonance peaks (HN_i, N_i, CA_{i-1}), (HN_i, N_i, CA_i), (HN_i, N_i, CB_{i-1}), and (HN_i, N_i, CB_i). It should be noted that there are no CBCA(CO)NH or HNCACB peaks for Prolines and there are no (HN, N, CB) peaks for Glycines (which don't have CB). After peak grouping, a typical spin system would be in the form of (HN_i, N_i, CA_i, CB_i; CA_{i-1}, CB_{i-1}), while some might be missing the CB chemical shifts, either the intra-residue or the inter-residue one or both.

In some existing works, the observed chemical shift ranges for different amino acid residue types are used to infer the possible residue types that could correspond to a spin system, a process referred to as *residue typing*. These works include TATAPRO [5], Mapper [6], PACES [3], and RIBRA [7]. In TATAPRO, for example, when CB chemical shift value falls into range [24ppm, 36ppm] and CA chemical shift value is less than 64ppm, then it restricts the residue to be one of Lys, Arg, Gln, Glu, His, Trp, $\mathrm{Cys}^{\mathrm{red}}$, Val, and Met. Usually, the number of candidate residue types inferred this way for a spin system is large (> 6), and consequently the subsequent assignment involves a very large search space and requires extra knowledge so as to be performed efficiently.

Realizing that the above use of chemical shift signature information is rough since only several chemical shift value cut-offs are used and only residue types are determined, several efforts seek to quantify the signature information by assuming that for one residue type, the chemical shift values of a nucleus follow a normal (Gaussian) distribution. BioMagResBank (BMRB, http://www.bmrb.wisc.edu/), which is a repository for known protein NMR data, collects the means and standard deviations for HN, N, CA, CB, C, and HA (and more) chemical shifts in all 20 types of amino acid residues. With these means and standard deviations at hand, a typical procedure is to use the density functions of the corresponding normal distributions to estimate a probability for mapping a specific spin system to a specific residue, using the intra-residue chemical shifts in the spin system. Taking this one step further, since secondary structure is another important structural factor that affects the chemical shift values, the means and standard deviations of chemical shifts can be collected for every combination of a residue type and a secondary structure type. Subsequently, one can predict the secondary structures for the target protein, and then estimate a probability for mapping a specific spin system to a specific residue coupled with its predicted secondary structure. Mathematically, for every intra-residue chemical shift cs in a spin system, we use the density function of the corresponding normal distribution to estimate a probability $p(cs \mid aa, ss)$ that the corresponding nucleus is in residue aa residing in secondary structure ss, where

$$p(cs \mid aa, ss) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(cs-\mu)^2}{2\sigma^2}}, \tag{1}$$

$\mu(aa, ss)$ is the mean, and $\sigma(aa, ss)$ is the standard deviation of the normal distribution for the (aa, ss) couple. The product of the probabilities for all the intra-residue chemical shifts in the spin system is taken as the probability that the (aa, ss) couple corresponds to the spin system. In our experiments, we have also implemented this method and used the logarithm of the probability to be the score. In more details, for spin system (HN$_i$, N$_i$, CA$_i$, CB$_i$; CA$_{i-1}$, CB$_{i-1}$), the score of mapping it to couple (aa, ss) is

$$-100 \times \frac{1}{4} \sum_{cs \in \{\mathrm{HN}_i, \mathrm{N}_i, \mathrm{CA}_i, \mathrm{CB}_i\}} \log\Big(p(cs \mid aa, ss)\Big), \tag{2}$$

where the factor 100 is solely for computational precision purposes and taking the average is for score normalization purposes. Clearly, a smaller score indicates a higher probability mapping. Such an estimation that quantifies the mappings between spin systems and amino acid residues coupled with secondary structures, is generally referred to as a *scoring scheme*. Existing studies adopting the above type of scoring schemes include QUASI [8] and MARS [2], where MARS does slightly differently by using z-scores rather than probabilities of the normal distributions.

Assuming normal distributions is one way to the signature information quantification, and has been widely adopted. However, we have performed substantial experiments to verify such an assumption and found that more than one third of couples do not convincingly follow normal distributions (cf. Supplementary Materials). In fact, this assumption was suspected in [1], where the authors proposed a CBM model for sequential assignment, to be detailed in the "Results" section, and a log odd like scoring scheme. Essentially, the log odd like scoring scheme 1) partitions the chemical shift range for a nucleus in every couple (aa, ss) into exactly 5 bins of equal size, 2) counts for the observed chemical shift cs the number of chemical shifts in the collected dataset that fall in the same bin to which cs belongs, and 3) uses the ratio of this number divided by the total number of chemical shifts associated with couple (aa, ss) in the collected dataset, to be the probability of mapping cs to couple (aa, ss). A refinement on this log odd scoring scheme, called a histogram-based scoring scheme, is proposed in [9]. In the histogram-based scoring scheme, the bin size is set to be one-tenth of the chemical shift range, and the bins are no longer fixed but centered around the observed chemical shift cs. In both scoring schemes, the estimated probability is again transformed into a score by taking the logarithm, similarly done as in Equation (2).

From machine learning point of view, the histogram-based scoring scheme is trained through a naive Bayesian learning, and it is thus called a *Bayesian scoring scheme* in this paper. Nevertheless, Bayesian is regarded as a learning method in this paper, and there are in total 8 Bayesian scoring schemes, among which one is identical to the above histogram-based scoring scheme and the others take in some more factors into the probability estimation and the estimation is based on two better constructed datasets than the ones in previous studies. These two datasets are ALL and HOMO, to be detailed in the "Training Datasets" section. Recall that, assuming A and B are two events, the Bayes rule says

$$p(A \mid B) = p(B \mid A)p(A)/p(B).$$

Therefore, to estimate a probability for mapping a specific chemical shift cs to a specific couple $(aa,\ ss)$,

- let N denote the total number of the same type of chemical shifts to cs in the training dataset;
- let $N(aa, ss)$ denote the number of chemical shifts associated with couple (aa, ss) (which is typically in the thousands) within N;
- let $N(cs)$ denote the number of chemical shifts in N that are close to cs, using a pre-learned threshold ϵ (cf. the "Methods" section);
- let $N(cs \mid aa, ss)$ denote the number of chemical shifts in $N(aa, ss)$ that are close to cs, using the same threshold ϵ.

Then, the Bayes rule says that the probability of mapping is

$$p(cs \mid aa, ss) = \frac{N(cs \mid aa, ss)}{N(cs)} \times \frac{N(cs)}{N} / \left(\frac{N(aa, ss)}{N}\right) = \frac{N(cs \mid aa, ss)}{N(aa, ss)}. \tag{3}$$

Such an estimation integrates the distribution assumption by replacing the assumed distribution with the actual counting and is expected to perform better than the log odd scoring schemes since the chemical shift window is dynamic rather than static. For simplicity, we call the scoring schemes assuming normal distributions *Normal* scoring schemes, while the others *Bayes* scoring schemes.

Note that in the residue typing schemes such as in TATAPRO, only intra-residue chemical shifts are used to restrict the corresponding residue types, while inter-residue chemical shifts are used mostly for *connectivity/adjacency* determination, that is, which two spin systems should match to two adjacent residues. Inter-residue chemical shifts can be used to restrict the preceding residue types following the same principles. Therefore, according to the same boosting idea in collectively using all available chemical shifts' signature information, it would be beneficial to also use these inter-residue chemical shifts' signature information in the scoring schemes. For this purpose, suppose the residue preceding aa is aa' and aa' is in secondary structure ss', then mapping spin system (HN$_i$, N$_i$, CA$_i$, CB$_i$; CA$_{i-1}$, CB$_{i-1}$) to couple (aa, ss) gets a score of

$$-100 \times \frac{1}{6}\Big(\sum_{cs \in \{\mathrm{HN}_i, \mathrm{N}_i, \mathrm{CA}_i, \mathrm{CB}_i\}} \log\Big(p(cs \mid aa, ss)\Big) + \sum_{cs' \in \{\mathrm{CA}_{i-1}, \mathrm{CB}_{i-1}\}} \log\Big(p(cs' \mid aa', ss')\Big)\Big). \tag{4}$$

Depending on whether or not the inter-residue chemical shifts are used, scoring schemes are classified into *Intra* where only intra-residue chemical shifts are used, and *Both* where both intra-residue and inter-residue chemical shifts are used. In this work, we adopted PsiPred [10] as our secondary structure predictor to get the secondary structure ss for each residue aa in the target protein. If the PsiPred final prediction results are used, then the scoring schemes are classified into category 1. Note that PsiPred uses a neural network for prediction and the last layer in the network contains three nodes each corresponds to a secondary structure type. The intermediate neural network output is a triple of predictions, which are used for final decision. In half many of the scoring schemes we chose to use this intermediate output rather than the final prediction, and subsequently these scoring schemes are classified in category 2. For example, suppose the intermediate prediction results for aa is (q_1, q_2, q_3), which basically mean that aa gets a probability q_1 (q_2, q_3, respectively) in α-helix (β-sheet, coil, respectively). Then for an observed chemical shift cs, the probability of its corresponding residue being aa should be the expected probability calculated as follows:

$$p(cs \mid aa) = \frac{1}{q_1 + q_2 + q_3}\Big(q_1 \times p(cs \mid aa, \mathrm{H}) + q_2 \times p(cs \mid aa, \mathrm{E}) + q_3 \times p(cs \mid aa, \mathrm{C})\Big), \tag{5}$$

which may replace $p(cs \mid aa, ss)$ in the above score calculations (Equations (2) and (4)). Using the above set of notations, a scoring scheme HOMO-Bayes-Both-2 is the one that is learned based on dataset HOMO, uses both intra-residue and inter-residue chemical shifts, and uses the intermediate neural network output from PsiPred to score the mappings.

Results

All the above scoring schemes are built on top of known protein NMR data. For example, for the log odd like scoring scheme and its improved version Bayesian scoring schemes, the entire distributions of known chemical shift values have to be ready; and for the scoring schemes assuming normal distributions, the means and standard deviations have to be collected beforehand. In the previous studies, several datasets have been assembled for these purposes. In our work, we re-composed two datasets to train the scoring schemes, to take the most advantage of all known protein NMR data. Essentially, one training dataset, called ALL, contains all known protein NMR data that have associated secondary structure information in Protein Data Bank (PDB), the other one is a subset of ALL, called HOMO, with homologous proteins removed. The detailed descriptions of these two training datasets ALL and HOMO, including a number of data filtering processes, are provided in the "Training Datasets" section. In each of these two datasets, the chemical shift values for a nucleus nu (in { HN, N, CA, CB }) in each combination of an amino acid aa and a secondary structure ss are organized into a separate text file for scoring scheme training purpose.

Since we have two training datasets ALL and HOMO, and there are options in the scoring schemes, i.e., to assume normal distributions or to apply the Bayesian learning, to use only intra-residue chemical shifts for residue typing or to use both, to use PsiPred final predictions or to use its intermediate neural network outputs, a total of 16 scoring schemes have been trained and examined, which are denoted as ALL/HOMO-Normal/Bayes-Intra/Both-1/2. To test the quality of each training dataset and the performance of each scoring scheme, we employed in our experiments an NMR backbone resonance sequential assignment model called *Constrained Bipartite Matching* (CBM) [1]. We remark that there are a number of other assignment models can be used for this purpose, such as AutoAssign [4], Mapper [6], MARS [2], and QUASI [8]. Nonetheless, since our focus is the quality of scoring schemes, we chose the

CBM model because CBM matches to our focus the best. In fact, after the scores for the mappings between spin systems and residues are calculated, the CBM model targets at optimal matchings, which completely depend on the scoring scheme. In this sense, the accuracy of the assignments output by CBM directly measures the quality of the scoring scheme.

An instance of CBM consists of an edge-weighted bipartite graph $G = (A, S, E)$, where A consists of the amino acid residues linearly ordered as they show up in the target protein, S consists of the spin systems, and every edge (a_i, s_j) indicates a mapping between residue a_i and spin system s_j, with its weight recording the mapping score. Given A and S, every one of the above 16 scoring schemes can be bound to the CBM model to score the edges. Without any extra information for spin systems, the above CBM instance expects a minimum-weight perfect matching, which can be computed efficiently [11]. In the output matching, i.e. assignment, the number of correctly assigned spin systems divided by the total number of assigned spin systems is defined as the *assignment accuracy*. Clearly, if the scoring scheme quantifies the signature information effectively, then the assignment accuracy should be high. Therefore, we use the assignment accuracy to measure the quality of the corresponding scoring scheme.

We have included a total of 14 protein NMR data in our experiments. These 14 proteins were not included in either of the training datasets and thus did not bias the scoring schemes. The detailed information on these 14 proteins are summarized in Table 1 [1,9]. We remark that these proteins do not have solved structures except three of them have related PDB entries. We chose to use only HN, N, CA, and CB chemical shifts in the current study such that only CA and CB chemical shifts are used as inter-residue chemical shifts. (This is similar to residue typing schemes as in TATAPRO and the scoring scheme in MARS [2]. But, note that we have also collected statistics for C and HA chemical shifts and therefore C and HA chemical shifts can be included in the experiments too.) Two of them, 4309 and 4393, do not contain CB chemical shifts and thus only CA chemical shifts are used as inter-residue ones (as we will see later, that these two proteins are harder than the others). The detailed CBM instance generating procedure is as follows: For every protein, the primary sequence was retrieved, and the secondary structure was predicted using PsiPred. Note that both the final prediction and the intermediate neural network output were saved. For every amino acid residue, the chemical shifts for HN, N, CA, and CB were retrieved from the BMRB entry, which formed an initial spin system containing only intra-residue chemical shifts. Subsequently, the chemical shifts for CA and CB in the preceding residue were appended to form the second spin system, now containing both intra-residue and inter-residue chemical shifts. Note that simulations of Proline and Glycine spin systems were a bit special since one doesn't have an HN nucleus and the other doesn't have a CB nucleus. We chose to set the corresponding chemical shifts to 0 (a similar treatment for CB has been done in [7], while most other programs including [3] choose not to simulate these values). We remark that such a simulation doesn't quite map to what is being done in practice, since for example there wouldn't be HSQC peaks for Prolines. Nonetheless, since the current work is centered at the evaluation and comparison of the scoring schemes, the adopted simulation procedure still provides a common foundation for fair comparison. In fact, one of our on-going projects is to place the scoring schemes in existing automated sequential assignment tools, including CBM, and to evaluate their performance on real protein NMR datasets.

For every chemical shift in a second spin system, we perturbed it by adding to it an "error" that follows a zero-mean normal distribution, for which the standard deviation was set to the standard deviation collected in the dataset. Note that such a perturbation step is to make the resultant spin systems more like real data, in which true chemical shifts are slightly altered by errors and noises. The result is a third spin system, which was finalized by randomly throwing away some CA and CB chemical shifts. The probability of throwing away chemical shifts was set to a very small value, i.e. 5% in our case. After that, using each of the 16 scoring schemes to score the mappings between the thus created spin systems and the residue and secondary structure couples in the target protein, a CBM instance was created and the assignment accuracy of its optimal solution was collected. The second column in Table 2 records the average assignment accuracies over all 14 proteins.

Note that there are 20 types of amino acid residues and 3 types of secondary structures, and therefore in total 60 distinct (aa, ss) couples. The tested proteins have length from 66 to 215 (cf. Table 1). It follows that there are multiple copies of an (aa, ss) couple in one protein. All the above scoring schemes, and those residue typing schemes too, signify the residue and secondary structure couple, but not the sequential position of the couple in the target protein. In this sense, the CBM model would not be effective without extra information for the spin systems — since there would be too many equivalent optimal assignments but there is only one correct assignment. Indeed, without extra information, the assignment accuracies are low (cf. the second column in Table 2) and the assignments are hardly useful for subsequent structure calculation. The extra information that makes CBM an effective sequential assignment model is the *connectivity*, or *adjacency*, between spin systems [1]. Recall that in the "Background" section where peak grouping was introduced, a resonance peak such as (HN_i, N_i, CA_{i-1}) has the CA entry appearing as an inter-residue chemical shift in one spin system s_j = (HN_i, N_i, CA_i, CB_i; CA_{i-1}, CB_{i-1}) and as an intra-residue chemical shift in another spin system s_k = (HN_{i-1}, N_{i-1}, CA_{i-1}, CB_{i-1}; CA_{i-2}, CB_{i-2}), though its values could differ slightly. This indicates that spin systems s_j and s_k must be mapped to adjacent residues in the target protein, that is, if s_j is mapped to a_i then s_k must be mapped to a_{i-1}. Depending on the quality of the spectral data, varied abundance of connectivity can be inferred using the inter-residue chemical shifts, and the connectivity connects the spin systems into strings (see for example, a step in TATAPRO [5]). The CBM model uses the connectivity as hard constraints on the feasible matchings and asks for a minimum-weight perfect *constrained* matching to respect all the connectivity. Clearly, "without extra information" is exactly the extreme case where there is no connectivity inferred from the spectral data. In another extreme case where all connectivity is achieved, all the spin systems are chained together into a single string and the assignment can be trivially done. In the general case, however, the CBM problem is NP-hard [1].

In our simulation study, since we have all connectivity for every protein, we randomly removed some portion to generate a few instances for every protein, which have different levels of connectivity abundance. In more details, an instance of k% connectivity is obtained by removing $(100 - k)$% connectivity. We have set k in tens and we are interested in reasonable amounts of connectivity, namely, $k = 50, 60, 70, 80, 90$, since in practice about 70% connectivity can be obtained. To solve the CBM instances, we adopted an exact algorithm based on IDA* search, so as to compute a minimum-weight perfect constrained matching. Note that the CBM problem is NP-hard and this IDA* based algorithm, though it is the currently fastest exact algorithm, could run in exponential time in the worst case [12]. We didn't record the running time of the algorithm as it is not the current focus, but the algorithm is expected to run very fast when the quality of the scoring scheme is high enough to make the optimal matching stand out. Depending on the scoring schemes, the actual running time for IDA* varied dramatically from less than a second to minutes to days, but it follows the general tendencies that instances with more connectivity take less time and instances by Bayes scoring schemes take less time than the corresponding instances by Normal scoring schemes. Typically, for example, using Bayes scoring schemes, the running times on instances with 70% connectivity and up were seconds, but using Normal scoring schemes it took hours to days. In our experiments, we have set a 2-day limit for running IDA*. We note that when applying the IDA* algorithm on real instances, normally a number of heuristics can be set up to speedup dramatically the search process. The 2-day limit is set for collecting as much statistics as possible to evaluate the scoring schemes. The numbers in the parentheses in Table 2 are the numbers of time-out instances, where we see that ALL/HOMO-Bayes-Both-2 has the least time-out instances (three 50% instances and two 60% instances, none of which was solved using any other scoring scheme). Our experiments were done on computers with 2.2GHz processors and a 2.5GB main memory (though in fact some easier ones were done on computers with lower specifications — recall that we have more than 1, 120 instances). About the running time, we should remark that, in some sense, the more time-out instances there are, the lower the quality of the scoring scheme is since a low quality scoring scheme makes the solution space for the instance too large to be searched over by the IDA* algorithm. In the reported average assignment accuracies, the time-out instances were set to have their assignment accuracies equal to the least assignment accuracy in the same category. Table 2 summarizes the average assignment accuracies of the 16 scoring schemes, where the average was taken over 14 instances in the same category.

We believe that setting time-out instances to have the least assignment accuracy in the same category is a reasonable treatment since time-out doesn't necessarily mean lower (than the least) assignment accuracy. Nonetheless, this is only a heuristic treatment and the true assignment accuracy might differ. Consequently, as the tables show, some scoring schemes that are expected to be better have slightly worse average assignment accuracies. For example, in Table 3, among the eleven solved HOMO-Normal-Both-2 60% instances, the least assignment accuracy is 0.859 and therefore those 3 unsolved instances were set to have an assignment accuracy of 0.859; however, twelve HOMO-Bayes-Both-2 60% instances were solved and the least assignment accuracy was 0.731. Setting the assignment accuracy to 0.731 for the two time-out instances pulls the average assignment accuracy of HOMO-Bayes-Both-2 below that of HOMO-Normal-Both-2, though HOMO-Bayes-Both-2 performs better than HOMO-Normal-Both-2 on all solved instances except 4144.

We have also fully examined the effectiveness of the HOMO-Bayes-Both-2 scoring scheme through testing it on all levels of adjacency on three proteins 4316, 4752, and 4929. These three proteins have the best NMR data quality. From Table 4 we can see that without any "forced" adjacency, the assignment accuracies have already reached 77%, and with a typical amount of adjacency, 60%, the assignment accuracies reach 100%. The IDA* algorithm took seconds on each of these instances.

Discussions

We mentioned inter-residue chemical shifts can be used to infer the connectivity among the spin systems. This is done in many sequential assignment programs. However, when typing the amino acid residues or quantifying the mapping between spin systems and residues, usually only intra-residue chemical shifts are used. For example, TATAPRO [5], Mapper [6], PACES [3], and RIBRA [7] use intra-residue chemical shifts to do the residue typing; and CBM [1], QUASI [8], and MARS [2] use them to quantify the mapping scores. We have designed scoring schemes that explicitly use both intra-residue and inter-residue chemical shifts to quantify the mapping scores. From the experimental results in Table 2, we have seen that inter-residue chemical shifts can play a significant role to improve the scoring scheme performance. To quantify its significance, we took the average over the assignment accuracies of eight scoring schemes that do not use inter-residue chemical shifts and the average over those do. The differences between these two average assignment accuracies are 12.1%, 9.5%, 1%, 0.2%, and 0% on 50%, 60%, 70%, 80%, and 90% instances, respectively. In the extreme case where no connectivity is used, using inter-residue chemical shifts can improve the average assignment accuracy by as much as 34.5%. These differences indicate that using inter-residue chemical shifts in the scoring scheme (or equivalently residue typing), besides using them in the connectivity determination, can boost the assignment accuracy significantly, typically when the amount of connectivity is small. When the connectivity is abundant, which says that inter-residue chemical shifts have already been fully taken advantage of, then they provide only little extra information for residue typing. Figure 1 plots the average assignment accuracies of scoring schemes using both intra-residue and inter-residue chemical shifts and scoring schemes using only intra-residue chemical shifts, respectively.

Regarding training dataset construction, theoretically, a good dataset should not contain bias on any typical fraction of the known protein NMR data and thus the NMR data for homologous proteins should be removed. Our training datasets ALL and HOMO both contain good quality protein NMR data (cf. the "Training Datasets" section), before and after the homology removal. We have detected that a large portion of the proteins in ALL have homologous ones and to remove the homologous ones to obtain HOMO. Consequently, the sizes of ALL and HOMO vary a lot in both the numbers of proteins and the numbers of chemical shift values. Nonetheless, the percentage of each type of chemical shift values, corresponding to a triple of a nucleus, an amino acid residue, and a secondary structure, doesn't vary much from ALL to HOMO (cf. Supplementary Materials). As a result, the effectiveness of a scoring scheme trained on ALL and its counterpart trained on HOMO do not seem to differ (cf. Table 2). Figure 2 plots their average assignment accuracies of the 8 scoring schemes trained on ALL and the 8 scoring schemes trained on HOMO, respectively. The differences between them are only 1.1%, 0.1%, 1.2%, 0.4%, 0%, and 0.5% on 50%, 60%, 70%, 80%, 90%, and the extreme 0% instances, respectively. With NMR spectroscopy becoming an increasingly employed high-throughput technology for protein structure determination, we foresee many more structures determined via NMR. If this is the case and if the balance among chemical shifts is significantly altered, we believe that homology removal would be a necessary process in good quality training dataset construction.

Results in Table 2 also tell us that Bayesian scoring schemes uniformly performed significantly better than Normal scoring schemes. The average assignment accuracies of Bayesian scoring schemes and Normal scoring schemes are plotted in Figure 3, where each average is taken over 8 scoring schemes. The differences between them are 5.1%, 5.3%, 6.1%, 2.8%, 0%, and 3.9% on 50%, 60%, 70%, 80%, 90%, and the extreme 0% instances, respectively. We interpret these differences as no surprise, for at least three reasons: one reason is that the assumption of normal distributions for chemical shifts is probably rough, though commonly adopted, for example, only two thirds of them can pass the normality testing (cf. Supplementary Materials); secondly, there might be other structural factors that affect the chemical shift values, for example, the residue solvent accessibility; and thirdly, even if the normal distribution assumption makes sense, the estimate of means and standard deviations could differ from the true values. Therefore, we believe in similar applications that involve empirical parameter estimations, a naive Bayesian learning could perform better than naive distribution assumptions on the involved parameters.

Regarding the way to use predicted secondary structures, since we know ahead of time that the secondary structures predicted by PsiPred come from a neural network where the secondary structures with only the largest “probability” are reported, using the final prediction results naively might introduce extra errors to the sequential assignment. We conjectured that using the accompanied “probabilities” by PsiPred for all three secondary structures for each residue might be helpful in reducing the prediction errors. We have implemented a scheme to take advantage of the probabilities and the experimental results demonstrated that using them indeed can improve the assignment accuracy significantly. We again calculated the average assignment accuracies of the 8 scoring schemes that take advantage of the accompanied “probabilities” and of the other 8 scoring schemes that use the final prediction results. The differences between them are 5.9%, 2.9%, 1.6%, 0.6%, 0.1%, and 3.3% on 50%, 60%, 70%, 80%, 90%, and the extreme 0% instances, respectively. Figure 4 plots these average assignment accuracies, where we can see that using the accompanied probabilities is always a better choice.

To summarize, from the results in Table 2, we are able to claim that using the known protein NMR data to learn a scoring scheme for quantifying the spin system signature information, probably homology removal doesn't matter. However, a naive Bayesian learning definitely outperforms the assumption of normal distributions; Secondly, using inter-residue chemical shifts in the scoring scheme will boost the quality, besides using them in the spin system connectivity determination; Thirdly, using the intermediate PsiPred neural network outputs on all three types of secondary structure is always a better choice than using the final prediction results naively.

The two scoring schemes ALL-Bayes-Both-2 and HOMO-Bayes-Both-2 perform equally the best among all 16 scoring schemes. Note that ALL-Normal-Intra-1 and HOMO-Bayes-Intra-1 are exactly the scoring function used in QUASI [8] and the histogram-based scoring scheme in [9], except that the training datasets in our work differ from the training datasets used in QUASI [8] and [9], respectively. Figure 5 plots the assignment accuracies of these 4 scoring schemes, ALL-Bayes-Both-2, HOMO-Bayes-Both-2, ALL-Normal-Intra-1, and HOMO-Bayes-Intra-1. It can be seen that (cf. Table 4) scoring schemes ALL-Bayes-Both-2 and and HOMO-Bayes-Both-2 are so effective that their assignment accuracies can reach as high as 80% without any given connectivity and the accuracies easily go beyond 90% with the help of a typical amount of (that is, 70%) connectivity. The HOMO-Bayes-Both-2 scoring scheme is provided freely as a web server that is accessible through http://www.cs.ualberta.ca/~ghlin/src/WebTools/score.php, where the HOMO training dataset is also available. The web server contains two main functions: one is "single testing" that returns a score for mapping an input spin system to an amino acid residue and a secondary structure couple, and the other is "batch function" that accepts a protein sequence together with its secondary structures in PsiPred format and a file containing spin systems, and returns an edge-weighted bipartite graph file, which can be readily fed to the IDA* algorithm, or any other algorithms designed for the CBM problem, together with some (or empty) connectivity information. Figure 6 shows a snapshot of the web server.

Finally, we want to remark that our current work focuses mainly on scoring scheme training for backbone resonance assignment. This is only a step towards one of our objectives to develop a fully automated tool for protein NMR backbone resonance assignment that will be both robust and efficient. The scoring schemes we have developed here can be adopted in any existing assignment frameworks besides the CBM model, such as AutoAssign [4], Mapper [6], PACES [3], MARS [2], QUASI [8], and RIBRA [7]. We expect the automated assignment tool to considerably speed up the protein structure determination process via NMR spectroscopy and to transform it from a time-consuming method to a high-throughput technology.

As far as the scoring scheme itself is concerned, it could be extended into a more general framework, oriented more towards full protein structure determination, by including side-chain nuclei into the backbone assignment, as well as J-coupling constants and residual dipolar coupling constants. We note that such an integration not only fulfills the assignment of other structural factors, but could also improve the assignment accuracy altogether as they can be used to cross validate each other.

Methods

Training Datasets

The initial set of protein NMR data was obtained from the BMRB and included all protein entries present in the database as of May 30, 2005. We applied several steps of filtering to remove potential noise and bias from the dataset to make it as clean as possible. Firstly, proteins containing less than 50 amino acids or containing amino acids not part of the standard twenty were eliminated. Secondly, corrected NMR protein entries were obtained from the RefDB and these proteins overwrote any BMRB proteins present in the dataset. In the resultant dataset, every protein entry (which is a single file) was parsed in order to obtain the primary amino acid sequence, the chemical shift value for each nucleus, as well as the PDB accession number(s). The PDB accession number was used to retrieve sequence and secondary structure information related to that protein and to apply the third filtering step. To this purpose, the proteins that made into the final dataset were those that contain PDB accession numbers where the corresponding PDB protein sequence is a subsequence of the BMRB protein sequence or the other way around. Next, the secondary structure information from the PDB protein entry was obtained for that protein. The PDB secondary structure notation has eight different letters, which we translated into a notation of three letters to match up with the PsiPred secondary structure format (namely, G, H, and I from PDB became H in PsiPred, E from PDB remained as E, and S, T, B, and non-annotated positions in PDB became C in PsiPred). We note that such a translation is necessary since we will be using PsiPred as the secondary structure predictor in our testing. Nonetheless, suitable adjustment can always be made if other secondary structure predictors are employed.

A total of $1,493$ protein entries and $165,122$ amino acid residues were obtained in the final dataset, denoted as ALL (an excel sheet containing the detailed statistics on the numbers of residues in different types of secondary structure is provided in Supplementary Materials); 456 of these proteins and $45,964$ amino acid residues were from the RefDB corrected data. A total of 6 files were created each corresponding to a nucleus from HN, N, CA, CB, C, and HA. For those protein entries in the final dataset, chemical shifts were placed into these 6 files. Each chemical shift is represented as a triplet of an amino acid residue type, a secondary structure type, and the chemical shift value.

In order to apply the scoring schemes effectively, we examined carefully the chemical shifts in every triple combination: nucleus, amino acid residue, and secondary structure. We observed that there are a tiny amount of chemical shift values should be treated as outliers since they diverse significantly from the main stream. Since the abnormal behavior of a single outlier could disrupt the scoring scheme, an efficient statistical method, namely "`boxplot`" [13] (with the relevant parameter set at 1.5), was applied to remove the outliers — this constitutes the fourth and the last filtering in dataset ALL construction.

In order to reduce the potential bias that could be caused by multiple homologous protein entries, a second dataset was generated out of dataset ALL. "BLAST 2 sequences (bl2seq)" [14] was run between every pair of sequences. Any protein having greater than 50% homology against another protein already included was removed (note that this is order dependent). The resulting dataset, denoted as HOMO, contained 822 proteins and $91,382$ residues, among which 336 proteins and $34,225$ residues were from the RefDB (the excel sheet containing the detailed statistics for dataset ALL in Supplementary Materials also contains the statistics for dataset HOMO). "`boxplot`" was also applied on HOMO to get rid of chemical shift outliers. For example, the chemical shift values for nucleus CA in Alanines residing in α helices range from 52.947ppm to 56.900ppm. Subsequently, the corresponding threshold used in Bayesian scoring schemes is set as $(56.900 - 52.947)/20 = 0.19765$ppm.

Score Generation

In each dataset (ALL or HOMO), every one of the six files that corresponds to a nucleus in {HN, N, CA, CB, C, HA} (note that we focus on backbone resonance sequential assignment) was further partitioned into 60 subfiles, each of which corresponds to an amino acid residue type and a secondary structure type couple. Exceptions are: the HN-file was only partitioned into 57 subfiles since Proline doesn't have the HN nucleus and the CB-file was only partitioned into 57 subfiles because of Glycine.

For every triplet of nucleus nu, amino acid residue aa, and secondary structure ss, let $N(aa, ss)$ denote the total number of chemical shift values collected in the corresponding subfile. Let $\mu(aa, ss)$ denote the chemical shift mean and $\sigma(aa, ss)$ denote the chemical shift standard deviation. These means and standard deviations were used in the scoring schemes assuming normal distributions on chemical shifts.

The Bayesian scoring schemes using the collected chemical shift values directly, as described in the "Background" section. In these scoring schemes, chemical shift thresholds have to be learned in order to estimate probabilities. They were set as follows: For triplet (nu, aa, ss), let ϵ_{nu} denote the window-size associated with this triplet such that exactly 20 intervals of length ϵ_{nu} cover the whole range of the chemical shifts. The value 20 was set in such a way that these window-sizes map closely to the standard deviations collected in the dataset. For every observed chemical shift value cs for nucleus nu, using cs as the midpoint, the number of chemical shifts in the (nu, aa, ss)-subfile that fall into the window of size ϵ_{nu} is $N(cs \mid aa, ss)$, which is used in Equation (3) to calculate the probability.

We adopted PsiPred [10] to predict the secondary structures for the target protein. The PsiPred secondary structure format consists of three notations, H for alpha helix, E for beta sheet, and C for coil. Besides the predicted secondary structure for each residue, PsiPred also provides a confidence value, which is a single digit in the range of 0 to 9. We note that such a confidence value is a post-treatment of the neural network output, which consists of three values associated with the three output units (helix, sheet, and coil). These three values for a residue in the target protein are stored in an intermediate PsiPred output file with suffix "ss2". These values can be regarded as the "prediction probabilities" for individual secondary structures and can be integrated into scoring schemes. In more details, when one residue aa is predicted to be in helix with probability 0.55, to be in sheet with probability 0.25, and to be in coil with probability 0.40, then $\frac{0.55}{0.55+0.25+0.40} = 45.8\%$ of the final score comes from mapping the spin system to (aa, H), $\frac{0.25}{0.55+0.25+0.40} = 20.8\%$ from mapping the spin system to (aa, E), and $\frac{0.40}{0.55+0.25+0.40} = 33.4\%$ from mapping the spin system to (aa, C) (cf. Equation (5)).

With the probability $p(cs \mid aa, ss)$ estimated for the chemical shift value cs associated with the nucleus in (aa, ss) couple, the absolute logarithm of $p(cs \mid aa, ss)$ was taken as a score in our scoring schemes. Furthermore, for a spin system, the average of all the individual chemical shift scores multiplied by 100, was taken to be the score for mapping the spin system to an (aa, ss) couple. For example, for spin system (HN_i, N_i, CA_i, CB_i; CA_{i-1}, CB_{i-1}), when only intra-residue chemical shifts were used in the scoring scheme, the score of mapping it to couple (aa, ss) is calculated by Equation (2), and when both intra-residue and inter-residue chemical shifts were used, the score of mapping it to couple (aa, ss) is calculated by Equation (4). We remark that the factor 100 is solely for computational precision purposes and taking the average is for score normalization purposes. Clearly, the smaller the score, the higher the confidence we have for the mapping.

Finally, there are a few special features of the chemical shifts that were utilized in all the scoring schemes developed above. To name a few, since there is no CB nucleus in Glycine, no CB chemical shift can be observed for a Glycine spin system. Consequently, when a spin system does contain a non-zero CB chemical shift value, it should not be mapped to Glycine. In this case, we associated with the mapping a score of `maximum`, which was set to 9999.99 and tells the assignment algorithm that such a mapping is *illegal*. Similarly, since Proline doesn't have an HN nucleus, a spin system containing a non-zero HN chemical shift value gets a score of `maximum` when mapping to Proline.

Conclusions

We have constructed two training datasets from known protein NMR data with and without homology removal, for scoring scheme training purpose. Through the extensive simulation study we found that the scoring schemes trained using them only differ marginally. Therefore, we might be able to conclude that currently known protein NMR data is quite evenly distributed in terms of protein homology, and therefore removing homology to construct a smaller training dataset wouldn't gain a lot in term of overall performance. We may also be able to conclude that in general a naive assumption on data distribution is inferior to a naive Bayesian learning. This is typical when the size of the known dataset is large. The inter-residue chemical shifts are mainly used for spin system connectivity determination in most of the existing works. We have demonstrated that using them in the scoring scheme explicitly, or equivalently residue typing, could improve the assignment accuracy significantly, typically when the amount of inferred connectivity is small. Regarding the secondary structure predictor PsiPred, since the prediction is used as an intermediate step, we believe that using its intermediate neural network output is a better choice than using its final prediction result naively.

Supplementary Materials

The web server implementing one of the best scoring schemes HOMO-Bayes-Both-2 can be accessed through http://www.cs.ualberta.ca/~ghlin/src/WebTools/score.php. The webpage also contains dataset HOMO and an excel sheet containing the detailed statistics on the numbers of amino acid residues in different secondary structures in both datasets HOMO and ALL. The normality testing results are also included. The reader may contact the correspondence author (ghlin@cs.ualberta.ca) for standalone web application and the above supplementary materials.

Authors Contributions

JW implemented all the NMR data filtering steps to construct those two training datasets ALL and HOMO, and coded all 16 scoring schemes. TT and JW worked together to run the IDA* algorithm on all generated instances and collected the statistics. XW worked with JW on designing those 16 scoring schemes. GL designed the overall framework, supervised every step of computation and data collection, and composed the manuscript.

Acknowledgments

The authors are grateful to the support from CFI and NSERC.

References

1. Xu Y, Xu D, Kim D, Olman V, Razumovskaya J, Jiang T: **Automated assignment of backbone NMR peaks using constrained bipartite matching**. *IEEE Computing in Science & Engineering* 2002, **4**:50–62.
2. Jung YS, Zweckstetter M: **Mars – robust automatic backbone assignment of proteins**. *Journal of Biomolecular NMR* 2004, **30**:11–23.
3. Coggins BE, Zhou P: **PACES: Protein sequential assignment by computer-assisted exhaustive search**. *Journal of Biomolecular NMR* 2003, **26**:93–111.
4. Zimmerman DE, Kulikowski CA, Huang Y, Tashiro WFM, Shimotakahara S, Chien C, Powers R, Montelione GT: **Automated analysis of protein NMR assignments using methods from artificial intelligence**. *Journal of Molecular Biology* 1997, **269**:592–610.
5. Atreya HS, Sahu SC, Chary KVR, Govil G: **A tracked approach for automated NMR assignments in proteins (TATAPRO)**. *Journal of Biomolecular NMR* 2000, **17**:125–136.
6. Güntert P, Salzmann M, Braun D, Wüthrich K: **Sequence-specific NMR assignment of proteins by global fragment mapping with the program Mapper**. *Journal of Biomolecular NMR* 2000, **18**:129–137.

7. Wu KP, Chang JM, Chen JB, Chang CF, Wu WJ, Huang TH, Sung TY, Hsu WL: **RIBRA – An Error-Tolerant Algorithm for the NMR Backbone Assignment Problem**. In *Proceedings of the 9th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005)* 2005:103–117.
8. Coutouly MA, Kieffer B, Atkinson RA: **QUASI: Quick Access to Spectral Interpretation**. *Comptes Rendus Chimie* 2004, **7**:335–341.
9. Wan X, Tegos T, Lin GH: **Histogram-Based Scoring Schemes for Protein NMR Resonance Assignment**. *Journal of Bioinformatics and Computational Biology* 2004, **2**:747–764.
10. Jones DT: **Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices**. *Journal of Molecular Biology* 1999, **292**:195–202.
11. Cormen TH, Leiserson CE, Rivest RL, Stein C: *Introduction to Algorithms.* Cambridge, Massachusetts: The MIT Press 2001. [Second Edition].
12. Tegos T, Chen ZZ, Lin GH: **Heuristic Search in NMR Resonance Peak Assignment**. *Journal of Bioinformatics and Computational Biology* 2005. in press.
13. Devore JL: *Probability and Statistics for Engineering and the Science.* Duxbury Press 1999. [Fifth Edition].
14. Tatusova TA, Madden TL: **Blast** 2 **sequences - a new tool for comparing protein and nucleotide sequences**. *FEMS Microbiology Letters* 1999, **174**:247–250.

Figures

Figure 1 - Both versus Intra

A comparison between using both intra-residue and inter-residue chemical shifts and using only intra-residue chemical shifts: each assignment accuracy is taken as the average of 8 scoring schemes, namely, ALL/HOMO-Normal/Bayes-1/2, respectively.

Figure 2 - ALL versus HOMO

A comparison between the two training datasets, HOMO and ALL: each assignment accuracy is taken as the average of 8 scoring schemes, namely, Normal/Bayes-Intra/Both-1/2, respectively.

Figure 3 - Bayes versus Normal

A comparison between the Bayesian scoring schemes and the scoring schemes based on normal distribution assumptions: each assignment accuracy is taken as the average of 8 scoring schemes, namely, ALL/HOMO-Intra/Both-1/2, respectively.

Figure 4 - Predicted Secondary Structures 1 versus 2

A comparison between using the intermediate neural network prediction results by PsiPred and using the final prediction: each assignment accuracy is taken as the average of 8 scoring schemes, namely, ALL/HOMO-Normal/Bayes-Intra/Both, respectively.

Figure 5 - Comparison among 4 Scoring Schemes

A plot of the assignment accuracies of 4 scoring schemes ALL-Bayes-Both-2, HOMO-Bayes-Both-2, ALL-Normal-Intra-1 (which is used as the score function in QUASI [8]), and HOMO-Bayes-Intra-1 (which is the histogram-based scoring scheme in [9]).

Figure 6 - A Snapshot of Score Webserver

A snapshot of the Score web server using “batch function”. Top left: two windows expecting a file of protein sequence together with secondary structures in PsiPred format and a file of spin systems. Bottom left: a window showing the score matrix (the complete bipartite graph). Top right: a bipartite graph with one side containing the spin systems and the other containing the linearly ordered amino acid residues in the target protein, where edges indicate the best mappings for the residues found in a greedy way (not IDA* output). Bottom right: a graphical view of the score matrix (VRML viewer http://www.parallelgraphics.com/products/cortona/), where the heights of the colored bars are proportional to the inverse of the scores.

Tables

Table 1 - Proteins Used in the Experiments

Detailed information of the 14 proteins included in the experiments. ‘Related pdbID’ records the PDB IDs of the related PDB entries; ‘Length’ records the number of amino acid residues in the protein; ‘#Proline/#Glycine’ records the number of Proline/glycine residues in the protein; ‘Accuracies’ refer to the assignment accuracies under the CBM model using scoring schemes HOMO-Normal/Bayes-Both-2. Note that entries 4309 and 4393 do not contain CB chemical shifts and thus their simulated spin systems contain only three intra-residue chemical shifts and one inter-residue chemical shifts. This might explain partially why timeout happened.

bmrbID	Related pdbID	Length	#Proline	#Glycine	Accuracies
4027	—	158	8	11	0.962/0.987
4144	1hmj	78	5	3	0.859/0.731
4288	—	105	10	5	0.924/0.933
4302	1mek	115	5	9	0.904/0.922
4309*	1dk0, 1dkh, 1b2v	178	3	27	timeout/timeout
4316	—	89	3	13	1.000/1.000
4318	—	215	12	9	timeout/0.963
4353	—	126	10	8	1.000/1.000
4391	—	66	1	6	0.879/0.924
4393*	—	156	3	6	timeout/timeout
4579	—	86	2	5	1.000/1.000
4670	—	120	2	10	0.967/1.000
4752	—	68	1	6	1.000/1.000
4929	—	114	2	5	1.000/1.000

Table 2 - Assignment Accuracies of 16 Scoring Schemes

Assignment accuracies of all the 16 scoring schemes, where the numbers in parentheses record the number of time-out instances.

	0%	50%	60%	70%	80%	90%
ALL-Normal-Intra-1	0.125	0.641(7)	0.742(5)	0.855(3)	0.949	0.997
ALL-Normal-Intra-2	0.150	0.736(5)	0.769(4)	0.943(3)	0.962	1.000
ALL-Normal-Both-1	0.455	0.783(3)	0.875(2)	0.893(1)	0.958	0.998
ALL-Normal-Both-2	0.484	0.864(3)	0.941(2)	0.900(1)	0.957	1.000
ALL-Bayes-Intra-1	0.151	0.731(5)	0.864(3)	0.962(3)	0.985	0.999
ALL-Bayes-Intra-2	0.189	0.809(5)	0.891(3)	0.955(3)	0.985	0.999
ALL-Bayes-Both-1	0.510	0.845(3)	0.898(3)	0.964(1)	0.988	0.998
ALL-Bayes-Both-2	**0.541**	**0.873**(3)	**0.926**(2)	**0.967**	**0.989**	**0.998**
HOMO-Normal-Intra-1	0.124	0.648(7)	0.758(5)	0.888(3)	0.949	0.997
HOMO-Normal-Intra-2	0.165	0.734(5)	0.783(4)	0.870(3)	0.978	1.000
HOMO-Normal-Both-1	0.456	0.781(3)	0.893(3)	0.899(1)	0.965	0.998
HOMO-Normal-Both-2	0.474	0.844(3)	0.934(3)	0.901(1)	0.963	1.000
HOMO-Bayes-Intra-1	0.139	0.711(5)	0.862(4)	0.917(3)	0.985	0.999
HOMO-Bayes-Intra-2	0.166	0.740(5)	0.858(3)	0.965(3)	0.989	0.999
HOMO-Bayes-Both-1	0.495	0.857(3)	0.898(3)	0.948(1)	0.988	0.998
HOMO-Bayes-Both-2	**0.550**	**0.874**(3)	**0.923**(2)	**0.958**	**0.991**	**0.998**

Table 3 - Assignment Accuracies of HOMO-Bayes-Both-2 on 60% Instances

Assignment accuracies of HOMO-Normal/Bayes-Both-2 on fourteen 60% instances, where '—' indicates time-out after the 2-day limit.

Instance	4027	4144	4288	4302	4309	4316	4318	4353	4391	4393	4579	4670	4752	4929
Normal-Both-2	0.962	**0.859**	0.924	0.904	—	1.000	—	1.000	0.879	—	1.000	0.967	1.000	1.000
Bayes-Both-2	0.987	**0.731**	0.933	0.922	—	1.000	0.963	1.000	0.924	—	1.000	1.000	1.000	1.000

Table 4 - Assignment Accuracies of HOMO-Bayes-Both-2 on 3 Proteins

Assignment accuracies of HOMO-Bayes-Both-2 on all levels of adjacency for three proteins 4316, 4752, and 4929.

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
4316	0.775	0.865	0.933	1.000	0.911	1.000	1.000	1.000	1.000	1.000
4752	0.809	0.838	0.912	0.971	1.000	1.000	1.000	1.000	1.000	1.000
4929	0.781	0.798	0.860	0.965	0.877	0.965	1.000	1.000	1.000	1.000

Figure 1

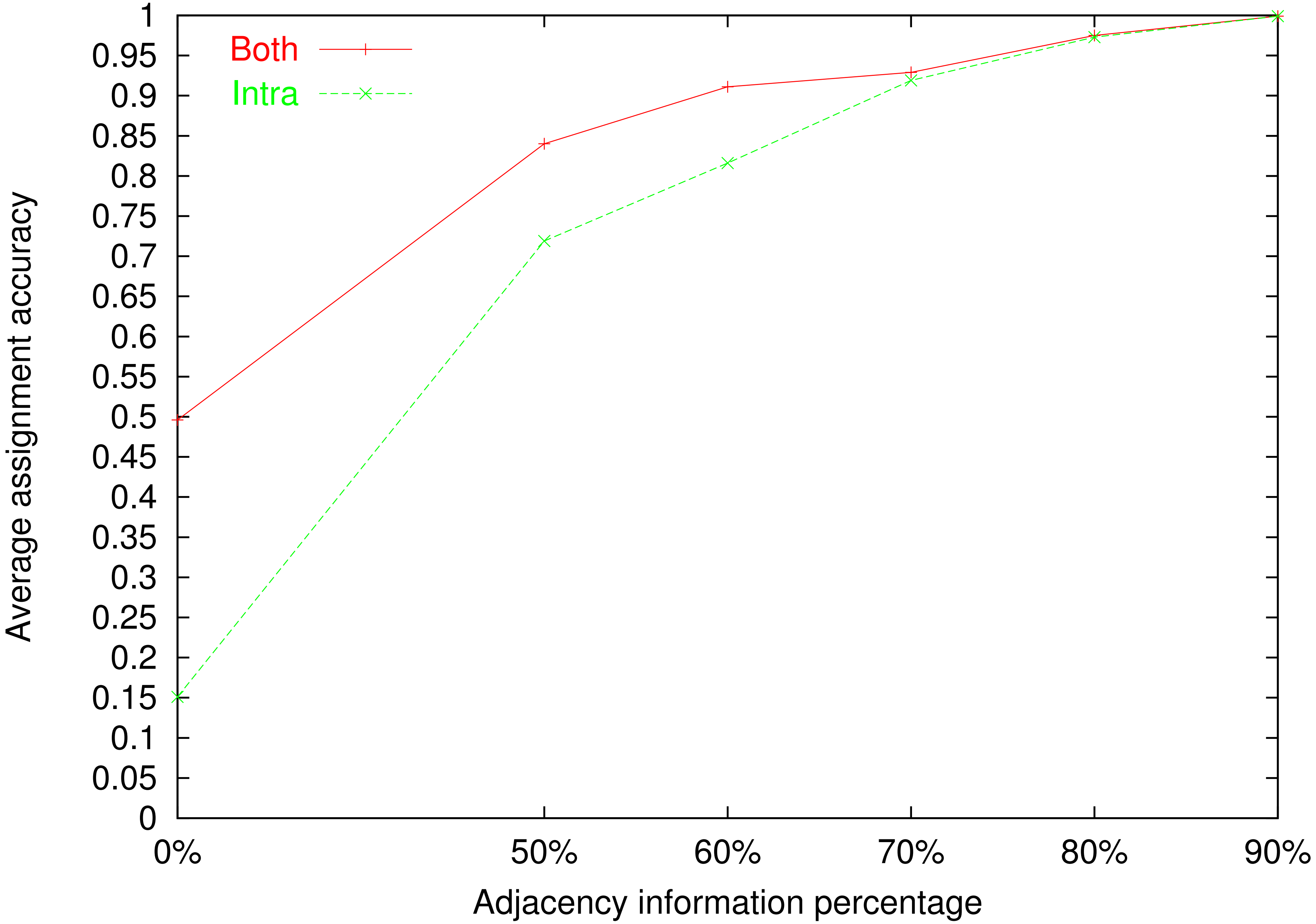

Both
Intra
Average assignment accuracy
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0%
50%
60%
70%
80%
90%
Adjacency information percentage

Figure 2

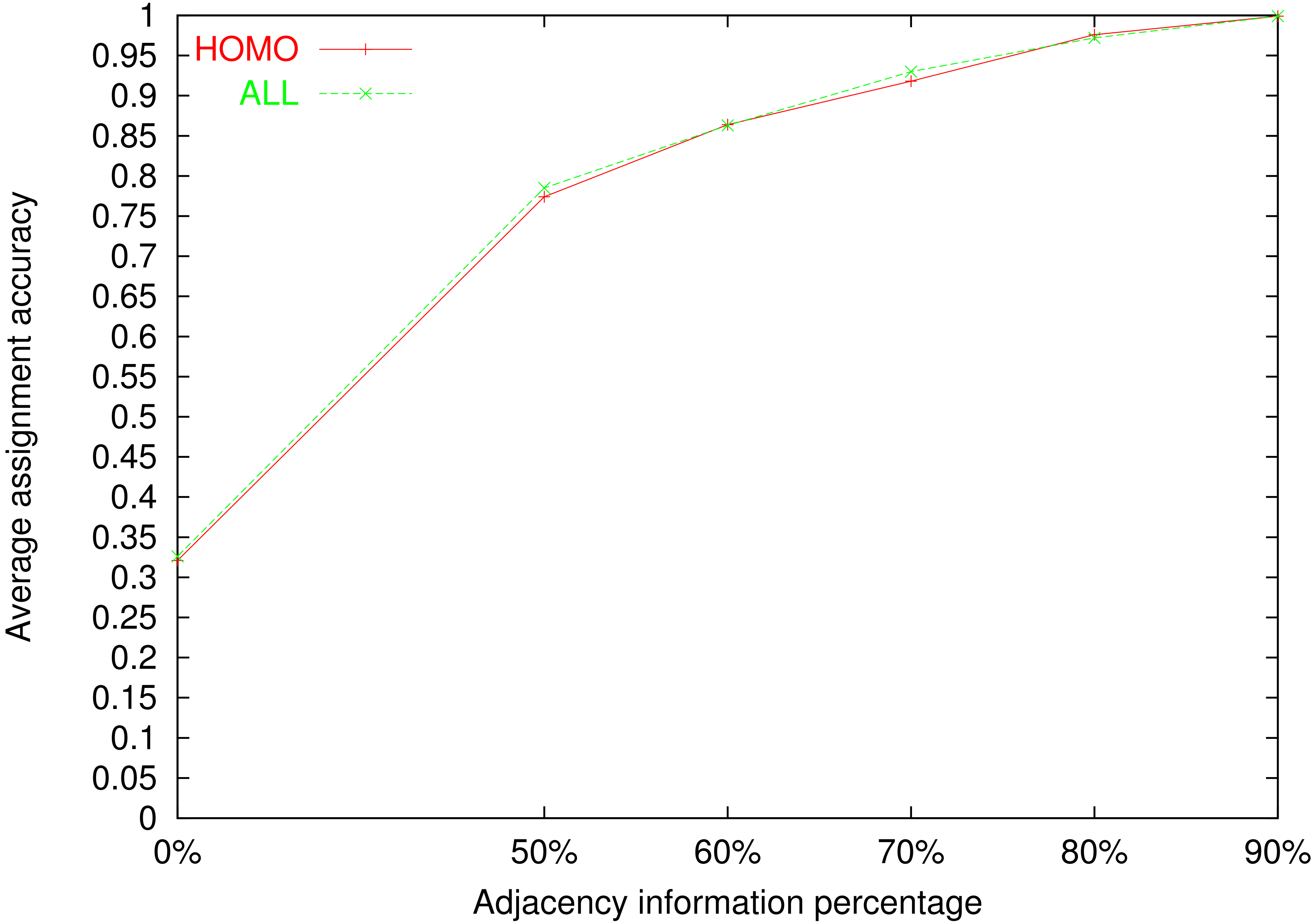

HOMO
ALL
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Average assignment accuracy
0%
50%
60%
70%
80%
90%
Adjacency information percentage

Figure 3

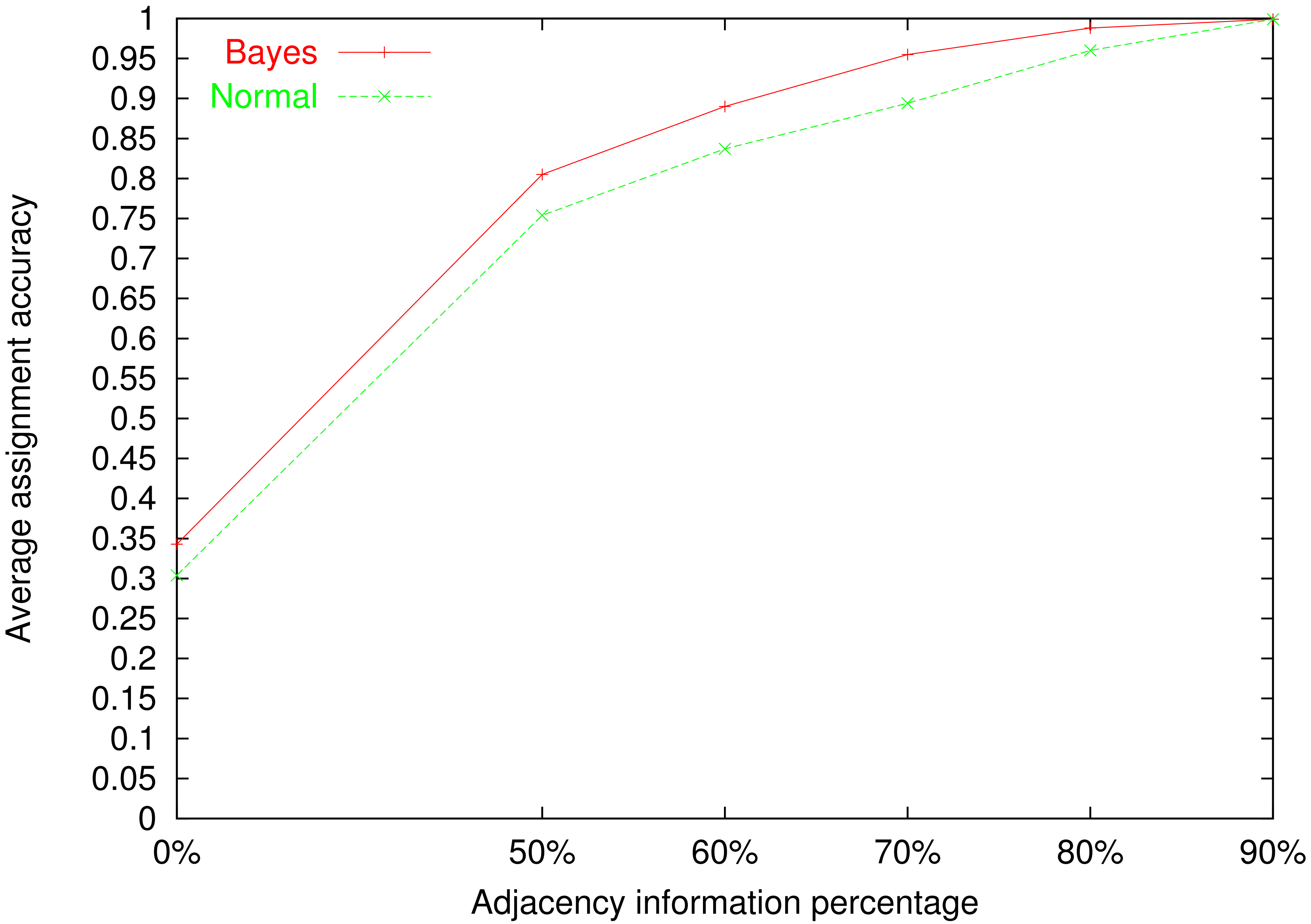

Bayes
Normal
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Average assignment accuracy
0%
50%
60%
70%
80%
90%
Adjacency information percentage

Figure 4

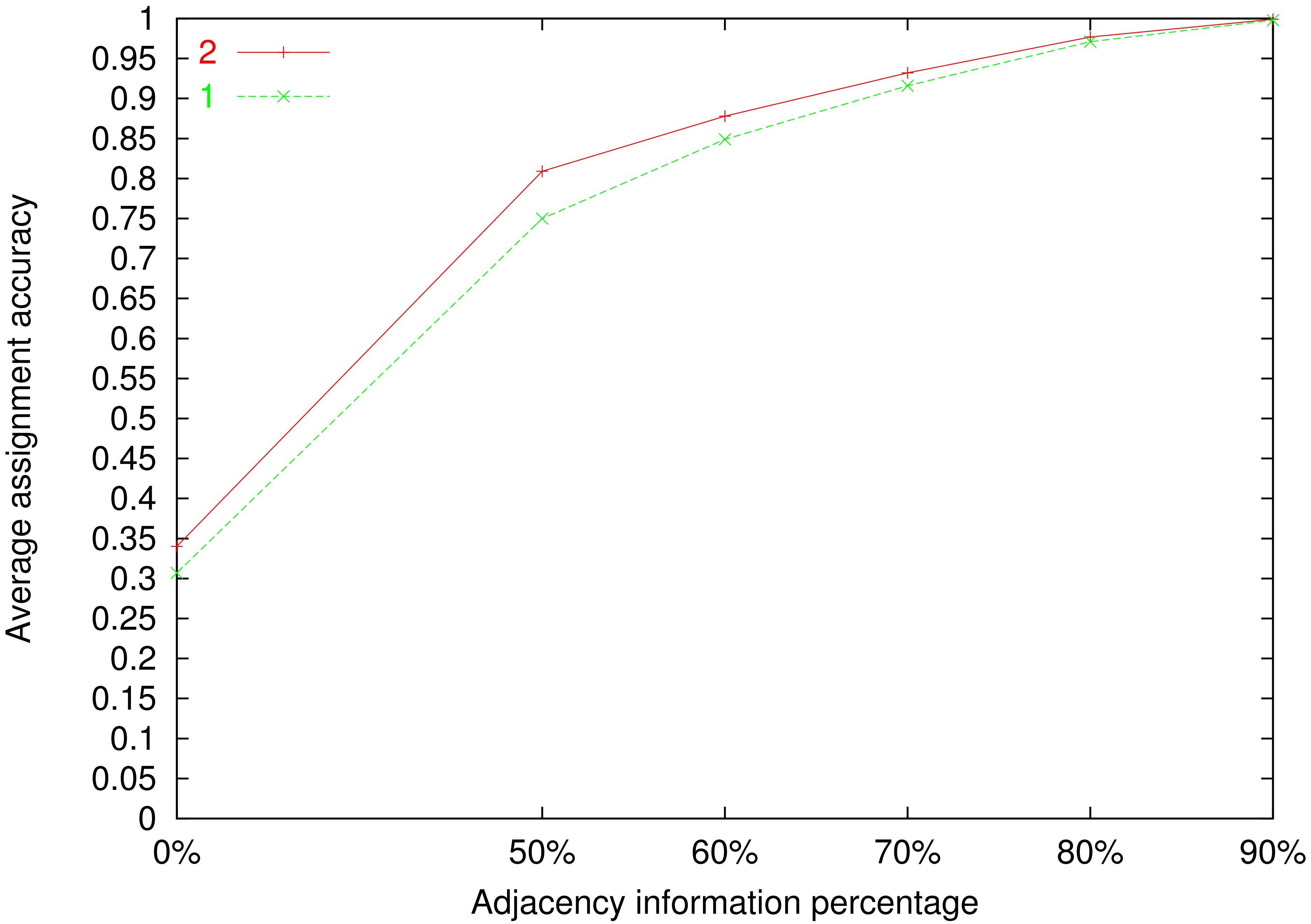
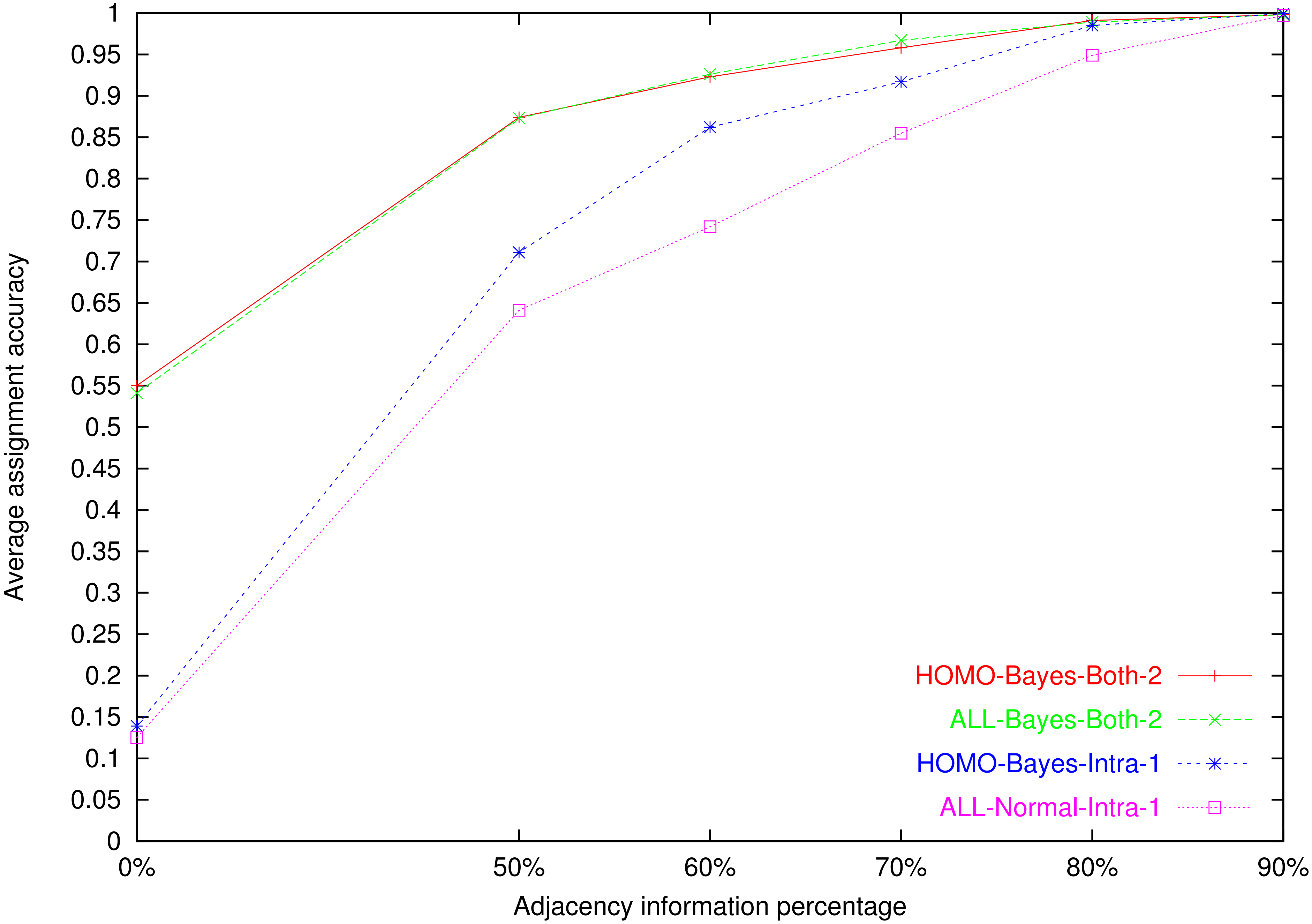
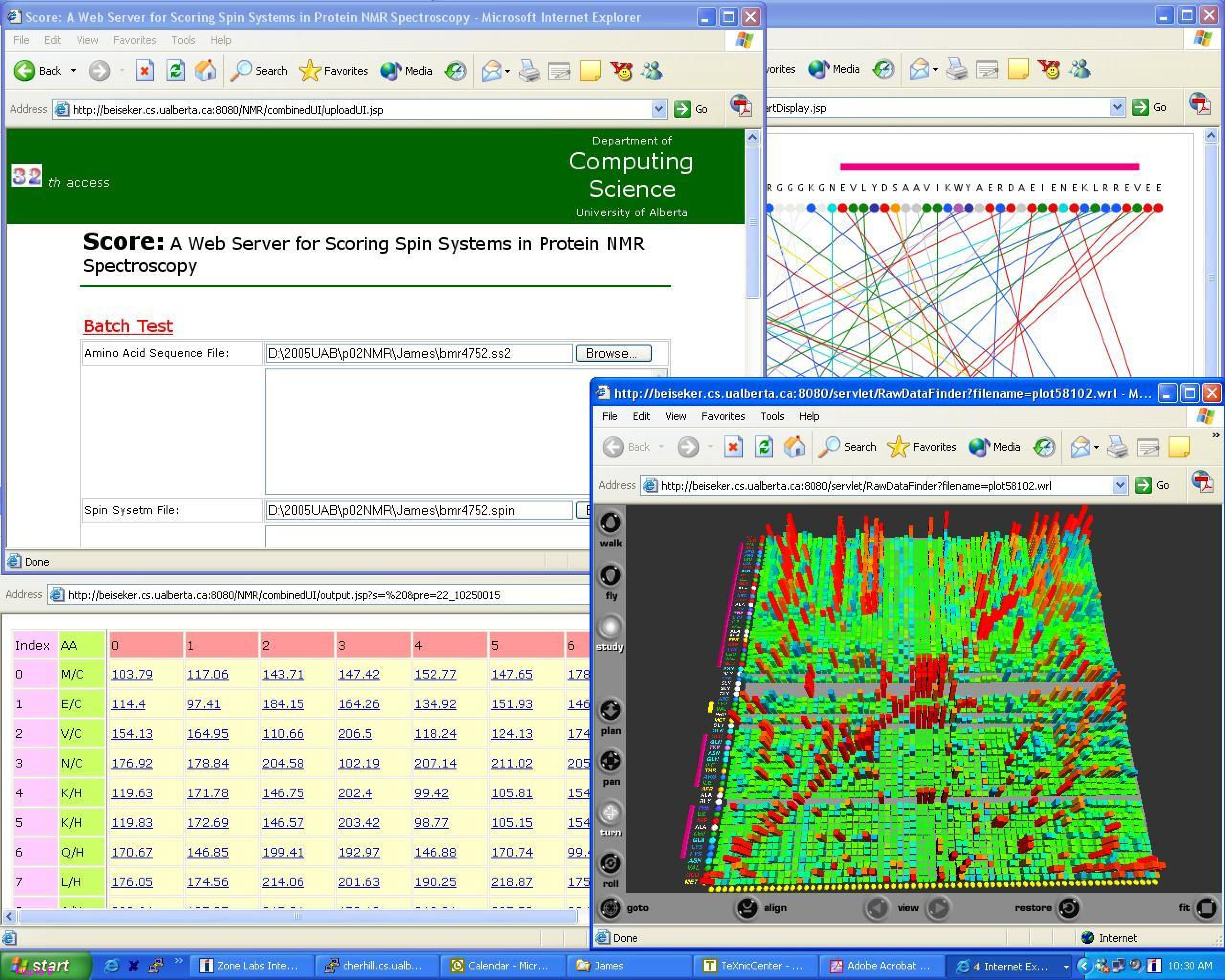

2
1
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0%
50%
60%
70%
80%
90%
Average assignment accuracy
Adjacency information percentage

Figure 5



1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Average assignment accuracy
0%
50%
60%
70%
80%
90%
Adjacency information percentage
HOMO-Bayes-Both-2
ALL-Bayes-Both-2
HOMO-Bayes-Intra-1
ALL-Normal-Intra-1



Score: A Web Server for Scoring Spin Systems in Protein NMR Spectroscopy - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
Department of
Computing Science
University of Alberta
32 th access
Done
artDisplay.jsp
orites
R G G G K G N E V L Y D S A A V I K W Y A E R D A E I E N E K L R R E V E E
http://beiseker.cs.ualberta.ca:8080/servlet/RawDataFinder?filename=plot58102.wrl - M...
walk
fly
study
plan
pan
turn
roll
goto
align
view
restore
fit
Internet
start
Zone Labs Inte...
cherhill.cs.ualb...
Calendar - Micr...
James
TeXnicCenter - ...
Adobe Acrobat ...
4 Internet Ex...
10:30 AM

http://beiseker.cs.ualberta.ca:8080/NMR/combinedUI/uploadUI.jsp

Score: A Web Server for Scoring Spin Systems in Protein NMR Spectroscopy

Batch Test

Amino Acid Sequence File:	D:\2005UAB\p02NMR\James\bmr4752.ss2 Browse...
Spin Sysetm File:	D:\2005UAB\p02NMR\James\bmr4752.spin

http://beiseker.cs.ualberta.ca:8080/NMR/combinedUI/output.jsp?s=%20&pre=22_10250015

http://beiseker.cs.ualberta.ca:8080/servlet/RawDataFinder?filename=plot58102.wrl

Index	AA	0	1	2	3	4	5	6
0	M/C	103.79	117.06	143.71	147.42	152.77	147.65	178
1	E/C	114.4	97.41	184.15	164.26	134.92	151.93	146
2	V/C	154.13	164.95	110.66	206.5	118.24	124.13	174
3	N/C	176.92	178.84	204.58	102.19	207.14	211.02	205
4	K/H	119.63	171.78	146.75	202.4	99.42	105.81	154
5	K/H	119.83	172.69	146.57	203.42	98.77	105.15	154
6	Q/H	170.67	146.85	199.41	192.97	146.88	170.74	99.4
7	L/H	176.05	174.56	214.06	201.63	190.25	218.87	175

Figure 6