A Slice-based ^{13}C -detected NMR Spin System Forming and Resonance Assignment Method

Meshari Alazmi, Ahmed Abbas, Xianrong Guo, Ming Fan, Lihua Li, and Xin Gao

Abstract—Nuclear magnetic resonance (NMR) spectroscopy is attracting more attention in the field of computational structural biology. Till recently, ^1H -detected experiments are the dominant NMR technique used due to the high sensitivity of ^1H nuclei. However, the current availability of high magnetic fields and cryogenically cooled probe heads allow researchers to overcome the low sensitivity of ^{13}C nuclei. Consequently, ^{13}C -detected experiments have become a popular technique in different NMR applications especially resonance assignment and structure determination of large proteins. In this paper, we propose the first spin system forming method for ^{13}C -detected NMR spectra. Our method is able to accurately form spin systems based on as few as two ^{13}C -detected spectra, CBCACON and CBCANCO. Our method picks slices from the more trusted spectrum and uses them as feedback to direct the slice picking in the less trusted one. This feedback leads to picking the accurate slices that consequently helps to form better spin systems. We tested our method on a real dataset of 'Ubiquitin' and a benchmark simulated dataset consisting of 12 proteins. We fed our spin systems as inputs to a genetic algorithm to generate the chemical shift assignment, and obtained 92% correct chemical shift assignment for Ubiquitin. For the simulated dataset, we obtained an average recall of 86% and an average precision of 88%. Finally, our chemical shift assignment of Ubiquitin was given as an input to CS-ROSETTA server that generated structures close to the experimentally determined structure.

Index Terms— ^{13}C -detected experiments, Spin systems, Chemical shift assignment, Denoising spectra, Protein structure determination

1 INTRODUCTION

ACCORDING to the PDB statistics, X-ray crystallography is the dominant method of protein structure determination. However, nuclear magnetic resonance (NMR) is playing more and more important roles in solution studies of proteins [1]. Unlike X-ray crystallography that requires crystallization of proteins, NMR can be used to determine structures of proteins that are difficult or impossible to be crystallized. Moreover, NMR techniques have many other applications in the area of protein research. These applications include, but are not limited to, dynamics of protein conformation, and intermolecular interactions between proteins and other molecules, such as DNA, RNA and lipids [2].

The main property of the atomic nuclei involved in the NMR experiments is the nuclear spin I . This nuclear spin results from the motion of charges in the nuclei. The nuclear spin I may take the values of 0, $1/2$, 1, $3/2$, etc. It is relying on the atomic number and the mass number of the nucleus. If the nucleus has zero spin like ^{12}C , ^{16}O and ^{32}S , then it cannot be used for NMR experiments. On the other hand, nuclei with spin $I = 1/2$, like ^1H and ^{13}C can be used in

NMR experiments [3].

Spin $1/2$ nuclei can take place in two different spin states α and β . If no external magnetic field is applied, the two spin states have the same energy and equally populated with nuclei. However, in an NMR experiment, the sample is put in a magnetic field B_0 . Consequently, α and β spins will have different energies with the α level slightly more populated than the β level. This difference in population leads to the magnetization of the sample. When the sample is irradiated with an electromagnetic field of correct frequency, a net energy is absorbed. The frequency at which absorption happens is called the resonance frequency [3].

The resonance frequency varies with very slight changes in the chemical structure, and conformation of the molecules and their environment. This very sensitive variation is usually called the chemical shift. Practically, the chemical shift measurement is done by measuring the relative ratio to a standard reference compound. Thus its value is independent of the value of the external magnetic field B_0 [3].

The large proton dipolar interactions implicit in large or paramagnetic proteins play a central role in their fast proton transverse relaxation, which dramatically broaden ^1H signals and, consequently, make ^1H signals weak and hard to be detected in ^1H related NMR experiments. ^{13}C -detected signal is less affected since ^{13}C has lower gyromagnetic ratio compared with ^1H . As the development of high ^{13}C sensitivity NMR probe, a series of ^{13}C direct detective multidimensional NMR experiments have been adopted, which demonstrate the advantage of ^{13}C -detected experiments over ^1H -detected experiments in large or paramagnetic protein signal assignment and structure determination [4], [5], [6], [7].

- M. Alazmi, A. Abbas, and X. Gao are with King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal, 23955-6900, Saudi Arabia.
- X. Guo is with King Abdullah University of Science and Technology (KAUST), Imaging and Characterization Core Lab, Thuwal, 23955-6900, Saudi Arabia.
- M. Fan and L. Li are with Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou, 310018, China.
- The first two authors contributed equally to this work. All correspondence should be addressed to Xin Gao. E-mail: xin.gao@kaust.edu.sa. Tel: +966-12-8080323.

Manuscript received March 19, 2017; revised August 26, 2017.

Another advantage of ^{13}C -detected experiments lies in the fact that ^{13}C chemical shift ranges over 200 ppm. On the contrary, ^1H chemical shift ranges over 12 ppm only. This means less overlapping of peaks in ^{13}C -detected spectra.

Moreover, all types of amino acids including Proline give signals in ^{13}C -detected experiments [4]. In contrast, Proline does not give signals in ^1H NMR experiments. Consequently, there is a better connectivity between signals in ^{13}C -detected spectra. These advantages motivate us to explore the field of ^{13}C -detected experiments in NMR spectroscopy.

Three-dimensional structure determination of a protein is a crucial application of NMR spectroscopy. Determining the 3D structure of a protein is crucial to understand its function which is imperative in developing new drugs. NMR protein structure determination is a complex process [8]. It consists of a number of time-consuming steps [9]. First, the protein sample is prepared. Second, NMR experiments are performed, and NMR spectra are generated. Third, peak picking is performed to determine chemical shifts of each peak in the spectra. Fourth, chemical shift assignment is a step that considers mapping peaks' chemical shifts to protein amino acids. Fifth, nuclear overhauser effect (NOE) assignment is generated which provide distance constraints between proton nuclei of the protein's amino acids. Finally, structure calculation according to the distance constraints is accomplished.

Peak picking is the first step after collecting the NMR spectra. The inputs of this step are the raw NMR spectra, and the outputs are the peak lists of these spectra. It is a tricky, time-consuming step that needs a lot of expert knowledge. This complexity motivated the computational community to make significant efforts to automate it [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. A missing true peak in the output peak list may lead to a lot of errors in the following steps. On the other hand, a few number of false positive peaks may be eliminated by the subsequent steps without affecting their accuracy. Therefore, the purpose of automatic peak picking methods is to generate peak lists with high recall and acceptable precision.

Chemical shift assignment is the cornerstone of the NMR protein structure determination process. Due to the clear formulation of this problem, it has attracted a great attention of the computational society [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39]. Although most of the automatic chemical shift assignment methods perform well on simulated datasets, they usually perform poorly on real datasets. This is due to various sources of errors in real NMR spectra, including noise, true peaks having weak intensities, and missing peaks. In addition to that, most of the assignment methods do not work directly on raw NMR spectra. Rather, they start their experiments from picked peak lists or pre-formed spin systems.

In our recent work [40], we developed a novel, automatic chemical shift assignment method that works directly on the raw NMR spectra. We also proposed, for the first time, to use slice picking instead of peak picking for the sake of a more accurate chemical shift assignment. A slice is a one-dimensional vector in a three-dimensional spectrum. The idea of using slices was to save the coupling information between peaks present in the same slice. The spin system

forming method based on slices proved effectiveness and efficiency, which was shown in [40].

In this paper, we extend the idea in [40] to develop the first automated spin system forming and resonance assignment method for ^{13}C -detected spectra. To benchmark our method, we use two real three-dimensional ^{13}C -detected spectra for a 76AA protein 'Ubiquitin', as well as a simulated dataset consisting of 12 proteins. Experimental results on both the real and the simulated datasets demonstrate that our method can generate accurate spin systems and resonance assignments, which can lead to accurately determined 3D structures. The workflow of our method is shown in Figure 1. The inputs of the pipeline are CBCACON and CBCANCO spectra, followed by multiple steps, including spectra denoising, slice picking, spin system forming, chemical shift assignment, and finally 3D structure determination. All these steps are explained in details in the following section.

Fig. 1: Workflow of the method.

2 METHODS

2.1 Dataset

In our experiments, we used a real dataset and a benchmark simulated dataset. Our real dataset consists of two ^{13}C -detected spectra for a protein 'Ubiquitin'. The spectra were collected on the 700MHz Bruker magnet in the NMR core lab at King Abdullah University of Science and Technology (KAUST). The benchmark simulated dataset consists of 12 proteins. For each of them, we created two ^{13}C -detected spectra. The types of the two spectra in both the real and the simulated datasets are CBCACON and CBCANCO. CBCACON is a 3D spectrum that captures the correlation between C_α named as CA_{i-1} or C_β named as CB_{i-1} of the previous residue, nitrogen N_i of the current residue, and carbonyl carbon C' of the previous residue named as CO_{i-1} . The other spectrum is CBCANCO which is also a 3D spectrum. However, this one captures the correlation between C_α or C_β of the current residue named as CA_i and CB_i , respectively, nitrogen N_i of the current residue, and carbonyl carbon C' of the previous residue named as CO_{i-1} . Each of the two spectra gives two peaks for each residue (except for Glycine which gives only one peak). Peaks of CBCACON are of the form $(N_i, CA_{i-1}, CO_{i-1})$ or $(N_i, CB_{i-1}, CO_{i-1})$. All peaks of CBCACON are positive peaks. Nuclei involved in a CBCACON peak are shown in Figure 2. On the other hand, peaks of CBCANCO are of the form (N_i, CA_i, CO_{i-1}) or (N_i, CB_i, CO_{i-1}) . All peaks of CBCANCO are positive peaks except for Glycine peaks

that are negative. Nuclei involved in a CBCANCO peak are shown in Figure 3. From the above description of these two spectra, these two spectra share fixed chemical shifts in nitrogen dimension and carbonyl carbon dimension. Our goal is to use these two spectra to form spin systems. These spin systems are vectors of chemical shifts of the form $(N_i, CO_{i-1}, CA_{i-1}, CB_{i-1}, CA_i, CB_i)$. We are aware that our real dataset is small. However, as the ^{13}C -detected experiments have only been recently used and applied in the NMR field, these two spectra are the only ^{13}C -detected spectra available to us. Here we use them as a proof-of-concept to show that using as few as two ^{13}C -detected spectra, we are able to develop fully automated methods to accurately assign the chemical shifts and consequently solve the structure of the protein. We also used the simulated dataset to give more evidence of the efficiency of our slice-based spin system forming method.

Fig. 2: Nuclei involved in CBCACON spectrum. Colored nuclei are the ones detected by this type of spectrum

Fig. 3: Nuclei involved in CBCANCO spectrum. Colored nuclei are the ones detected by this type of spectrum

2.2 Denoising NMR Spectra

NMR spectra noise sources include random noise, water bands, sample impurities and artifacts [41]. Thus, it is important to remove as much noise as possible while preserving all the real peaks in the spectrum including the weak peaks.

Recently, Cannistraci et al. [23] proposed a novel non-linear spatial adaptive filter called MMWF* which is a variation of MMWF (Median Modified Wiener Filter) [42]. This filter is an easy-to-use spatial filter that has only one parameter to tune (i.e., the window size). Similar to the Wiener filter, MMWF* is applied to the signal adaptively, with the amount of smoothing it performs differently according to the local variance of the signal (variance in the region enclosed by the filter window). If the local variance is high (like the case in the regions of real peaks), less smoothing is performed. On the other hand, if the local variance is small (like the case in the noisy background regions), more smoothing is performed.

MMWF* merges the merits of both median and Wiener filters, and nullifies their defects. It can get rid of spike noise (which is a property of the median filter). Besides, it preserves the edges and the morphology of the peaks (a property that is partially offered by the Wiener filter). Furthermore, MMWF* is even better than the original MMWF filter in that it is much less sensitive to noise. Thus, it better preserves the very weak peaks. MMWF* is also very robust to the change in the window size. Furthermore, it was shown in [23] that MMWF* performs better than wavelet smoothing in terms of recall, precision, and F-score of peak picking. Thus, in this paper, we chose to use MMWF* to denoise our two ^{13}C -detected spectra.

2.3 Slice Picking

Slice picking is a novel idea proposed in our previous work [40]. In this paper, we use it for the first time with ^{13}C -detected spectra to prove how general and efficient the idea is. A slice is a one-dimensional vector in a three-dimensional spectrum. Our two spectra have three dimensions: N (Nitrogen), CO (Carbon atom of carbonyl group) and C (either CA or CB atom). If we fix CO and C, we will obtain an N slice. If we fix N and C, we will get a CO slice. And if we fix N and CO, we will obtain a C slice. Our goal is to obtain C (carbon) slices that contain CA and CB peaks. To get these carbon slices, we need to select which N and CO to fix. The N value to be set is a local maximum in an N slice. Similarly, the CO value to be set is a local maximum in a CO slice. For every possible value of CO and C (CO_1, C_1), we get an N slice. We look for the position of the highest peak in this N slice. Let this position be N_1 . Now, we fix (N_1, C_1) to obtain a CO slice. Again, we look for the position of the highest peak in the CO slice. Let this position be CO' . If $CO' = CO_1$, we keep the two values N_1 and CO_1 . The two values (N_1, CO_1) will correspond to a carbon slice.

According to the previously described method, we extract all possible carbon slices from the CBCACON spectrum. We start with this spectrum because it is less noisy and more trustable than CBCANCO. We then keep the highest $1.5N_p$ slices for the next step, where N_p is the number of amino acids in the protein. By highest slices, we mean slices that have peaks with the highest intensities. As mentioned earlier, each carbon slice is represented by (N, CO) values. It is expected that if a slice with particular (N, CO) values is found in the CBCACON spectrum, it is highly likely to find a slice with approximately equal (N, CO) values in the CBCANCO spectrum. Accordingly, for each (N, CO) slice

extracted from the CBCACON spectrum, we search in the CBCANCO spectrum for a slice in the neighborhood ($N \pm 0.2$, $CO \pm 0.1$). Figure 4 shows two sets of matching slices from the two spectra. All possible slices are extracted from the CBCANCO spectrum, and again, only the highest $1.5N_p$ slices are kept for the subsequent steps. For each picked slice from the two spectra, we extract at most four peaks provided that the smallest peak's intensity is greater than 10% of the highest peak's intensity. We extract four peaks although we only expect two peaks to be correct because there may be false peaks with high intensities in our slices.

(a) A set of CBCACON slices.

(b) A set of CBCANCO slices.

Fig. 4: An example of two matching sets of CBCACON and CBCANCO slices.

2.4 Spin System Forming

In this section, we present our method for forming spin systems using slices for ^{13}C -detected spectra. To our knowledge, this is the first attempt to automatically form spin systems for ^{13}C -detected spectra.

For each slice picked from the CBCACON spectrum, we look for a matching slice in the list of slices picked from the

CBCANCO spectrum. The two matching slices must have approximately equal N and CO values, with tolerances 0.2 and 0.1 ppm respectively. If we find a pair of matching slices, then we fix N_i and CO_{i-1} of the spin system to be equal to the values of N and CO of the matching slices.

Now we need to detect CA_{i-1} and CB_{i-1} of the spin system. For this, we need only the CBCACON slice of the matching pair. If the slice has only one significant peak, this implies that CA_{i-1} and CB_{i-1} for this spin system are equal, and we set their values to the chemical shift value of our significant peak. However, this is a very rare case because the CA and CB ranges do not overlap for most of the amino acids. The general case is that the values of CA_{i-1} and CB_{i-1} are different. Since all peaks are positive in this spectrum, any peak in the slice can be either CA_{i-1} or CB_{i-1} . Thus, we try to cover all possibilities by taking all possible values of CA_{i-1} and CB_{i-1} for a spin system. We then filter these possibilities using the theoretical ranges of CA and CB of amino acids. We get rid of any spin system having CA_{i-1} or CB_{i-1} out of the theoretical ranges of CA and CB of amino acids respectively.

Now, we have a set of partial spin systems. Each partial spin system has the values of ($N_i, CO_{i-1}, CA_{i-1}, CB_{i-1}$) and is made up of two matching CBCACON and CBCANCO slices. To obtain the values of CA_i and CB_i for each spin system, we will use the CBCANCO slice corresponding to this spin system. If the slice has only one significant peak, which is a very rare case as mentioned above, this implies that both CA_i and CB_i have equal values. Thus, we set the values of CA_i and CB_i to the chemical shift value of that significant peak. However, the general case is that CA_i and CB_i have different values. Thus, we cover all possibilities by taking all possible values of CA_i and CB_i . Therefore, the same peak may be assigned to CA_i in a candidate spin system, and assigned to CB_i in another candidate spin system. As long as we have more than one significant peaks in the slice, CA_i and CB_i are not allowed to take similar values.

After we obtain all candidate spin systems, we perform a two-step refinement procedure, in order to decrease the number of false positive spin systems. In the first refinement step, we remove spin systems whose CA_i or CB_i values do not fall within the theoretical chemical shift ranges of CA or CB respectively. In the second refinement step, we group all the spin systems that have similar N_i and CO_{i-1} values into clusters. Thus, all spin systems that come from the same matching pair of slices are grouped into one cluster. For every spin system in a particular cluster, we check if its values of CA_i and CB_i exist as CA_{j-1} and CB_{j-1} in another spin system located in another cluster. If we find a match, then, this spin system will remain in the final list of spin systems. One cluster may have more than one spin system having their CA_i and CB_i values present in other spin systems as CA_{j-1} and CB_{j-1} values. In this case, all these spin systems are kept in the final list. We remove all other spin systems in the cluster that do not find matches. However, some clusters may not have any matches between CA_i and CB_i values of their spin systems and CA_{j-1} and CB_{j-1} of the spin systems in other clusters. This case may happen due to the problem of missing peaks or shifts in the chemical shift values of the peaks. In such cases, we keep all

spin systems of these clusters in the list of final spin systems.

2.5 Chemical Shift Assignment Using Genetic Algorithm

In the assignment problem, we have a protein consisting of N_p residues and we also have a set of M spin systems. We need to find the correct mapping between residues and spin systems such that every residue is assigned to the right spin system. Since the number of spin systems is usually much larger than the number of residues of the protein, some spin systems are not assigned to any residue. This problem can be solved using search algorithms. In this paper, we developed a genetic algorithm to solve the problem.

A general genetic algorithm produces a new population from one population of chromosomes (e.g. ones and zeros concatenated as strings) [43]. This movement between populations simulates or mimics the natural selection including crossover and mutation. The selection operator picks the fittest chromosomes in the population and allows them to reproduce. Crossover exchanges subparts of two chromosomes. Mutation changes randomly one or more locations in the chromosome.

The genetic algorithm used in this paper was inspired by the one proposed in GANA [39]. GANA fitness function of the genetic algorithm mainly cares about the information of the connectivity between the spin systems. Thus, the fitness score of a population increases as the connectivity between two spin systems assigned to two successive residues increases. In our previous work [40], we introduced a new component to the fitness function of the genetic algorithm. This new component takes into account the amino acid typing information. Thus, the fitness score increases if the spin system assigned to a particular residue preceded by another residue fits the values of the expected carbon chemical shifts of these two subsequent residues. The expected carbon chemical shift values for two subsequent residues are deducted from the BMRB database [44]. It was shown in [40] that the new fitness function component improves the assignment results when being applied to real proteins. This improvement is because of the non-perfect connectivity information between spin systems of real proteins, which make the amino acid typing information highly important. In the following section, we give a summarized description of the genetic algorithm used in our method [40].

2.6 Summary of Genetic Algorithm Used

Genetic algorithm has five main steps. These are initialization, selection, crossover, mutation, and stopping criteria. These steps have been used in our previous work [40].

The first step in the genetic algorithm is the initialization step. In this step, vector representation needs to be clearly defined. Each individual is represented in a vector. In our problem, the protein sequence is represented as an individual. The individual length is the protein sequence length. The values of the vector represent the assigned spin system to each amino acid. For the first initialization, we randomly choose a residue and assign a spin system to it randomly. One constraint here is that the randomly chosen spin system must have CA and CB chemical shifts fall within the chemical shift ranges for that picked residue.

Then, we find spin systems that their CA_i and CB_i chemical shifts fall within 0.4 ppm away from the chemical shifts of CA_{i-1} and CB_{i-1} of the previously assigned spin system. From this subset of the spin systems, we randomly pick an unassigned spin system to assign to the left residue. To assign a spin system to the right residue, we first find spin systems that have their CA_{i-1} and CB_{i-1} chemical shifts fall within 0.4 ppm away from CA_i and CB_i chemical shifts of the spin system that have been assigned to the middle residue. This is called an extension approach on both sides. The algorithm proceeds with the extension approach unless there are no more extensions. Then, we randomly pick an unassigned residue if there is any left unassigned residues. We repeat the same procedure mentioned above. This approach will finish when all the residues are assigned. This is the end of an individual initialization. We repeat this process 600 times to generate 600 individuals that represent our initial population.

The next step is the selection step. In this step, the individuals are assessed by a fitness function. This function is based on two combined scores. One is the connectivity score. The other one is the amino acid typing score. It assigns a score to each individual. An individual with high score compared with other individual scores has higher chance to be selected in the following generation.

The third step is the crossover operation. This step is the same step proposed by GANA [39]. Half of the individuals that have better fitness scores are used for crossover procedure to generate new individuals. The percentage of the number of the new individuals generated in this procedure is 70% of the chosen 50% individuals. We randomly select two individuals and randomly choose an entry (residue) from any of these two chosen individuals. The spin system assigned to that residue is copied to the new individual. Then, in the new individual, extension procedure is done on this newly assigned residue. The idea here is to copy the connected spin systems from the chosen individual to the new individual. This procedure continues until no more spin systems are connected or the chosen spin system from the chosen individual has been already assigned in the new individual. Then, we repeat this process again by randomly picking a residue from any of the chosen parents. This ends when all the residues are assigned. The other remaining 30% of the 50% chosen individuals are generated randomly as has been done in the initialization step after the crossover operation.

The fourth step is the mutation operation. The mutation rate is 0.2% that is 2 in 1000. For each residue in the new individual, we randomly pick a number from 1 to 1000. If the randomly chosen number is less than or equal to 2, then the mutation operation is performed on the chosen residue. An unassigned spin system is picked randomly and copied to the chosen residue to change the current spin system. This occurs with one constraint that is the randomly chosen spin system CA and CB chemical shifts fall within the theoretical ranges of that residue chemical shifts. Then, we apply the extension approach on the right side. After the extension is done, we check for the following residue and repeat the same process for all unvisited residues.

The last step is the stopping criteria. The stopping of the algorithm is controlled by either of the two conditions.

The First condition is that the fitness scores of the chosen individuals do not improve for 100 generations. The second condition is that if the number of generations reaches 500.

2.7 Evaluation Criteria

Each assigned residue has four chemical shifts (N_i, CO_i, CA_i, CB_i). The residue is correctly assigned if at most one of the four chemical shifts is out of the error range. This error range is 0.5 ppm for both carbon and nitrogen.

To evaluate the quality of our chemical shift assignment, we measure precision, recall and F-score. Suppose we have N_p residues of a protein and N_m of them are manually assigned, suppose that our method assigns N_o residues and suppose that T_c residues are correctly assigned, precision is then defined as T_c/N_o , recall is defined as T_c/N_m , and F-score = $(2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$.

3 RESULTS

3.1 Chemical Shift Assignment Results for Ubiquitin

We tested the performance of our method on a real protein (Ubiquitin). As mentioned earlier, we are aware that only one protein represents a very small dataset. However, we do not have available ^{13}C -detected spectra for more proteins. Thus, we tried to prove the validity of our method using the available spectra of our protein. Table 1 shows the performance of our method compared with the state-of-art method, PINE, that can take our formed spin systems as inputs and conduct resonance assignment based on such spin systems. Although Ubiquitin is a small protein, it is well known that automatic chemical shift assignment for real proteins is a very challenging task [35], [40]. Thus, recall and precision of both 92% implies quite good performance for automatic spin system forming and resonance assignment methods that start directly from raw spectra.

Analyzing the results, we found that we have six wrongly assigned residues. For those six residues, the correct spin systems for them were not generated by our program. One of the reasons for not generating a correct spin system is that the N or CO chemical shift of the correct slice of the CBCANCO spectrum is shifted by slightly more than 0.2 or 0.1 ppm respectively than the N or CO chemical shift of the correct slice of the CBCACON spectrum. Thus, the two slices did not match according to the constraints of our method. Another reason may be because there is a shift in the values of the chemical shifts of CA_i or CB_i in the slice of CBCANCO. They are different in a few cases by values greater than 0.5 ppm than the corresponding CA_{i-1} or CB_{i-1} peaks in the slice of CBCACON spectrum. The third possible reason is that the peak that should be assigned to either CA or CB in the CBCANCO slice is missing. The last possible reason is that the CBCANCO correct slice is not from the top $1.5N_p$ slices that we used to create the spin systems.

To our knowledge, existing chemical shift assignment methods like AUTOASSIGN [45], PINE [46] and RIBRA [28] take peak lists of 2D and 3D ^1H -detected spectra as inputs. These methods do not have the option of taking the peak lists of our two ^{13}C -detected spectra as inputs.

TABLE 1 Performance comparison between PINE when using our spin systems as inputs and our method on Ubiquitin. The second column provides protein length. The third column provides the total number of the manually assigned amino acids. The fourth and fifth columns give the results of both methods in the format “number of correctly assigned amino acids/total number of assigned amino acids”. *PRC*, *REC* and *F-score* stand for precision, recall and F-score, respectively.

Protein	Length	Manually assigned	PINE fed by our spin systems	Our method
Ubiquitin	76	76	63/74	70/76
<i>PRC</i>	-	-	85%	92%
<i>REC</i>	-	-	83%	92%
<i>F-score</i>	-	-	84%	92%

However, PINE has the option of taking formed spin systems as inputs. Thus, we used our spin systems with PINE to demonstrate that our formed spin systems are not only useful for the genetic algorithm-based assignment method, but also for other state-of-the-art methods. The chemical shift assignment results for PINE are shown in Table 1 as well to compare PINE’s results to ours.

3.2 Chemical Shift Assignment Results on Simulated Dataset

We also tested the performance of our method on a simulated dataset that consists of 12 proteins. The sizes of these proteins range from 68 AA to 260 AA. For each of the 12 proteins, we created two simulated spectra of the types CBCACON and CBCANCO. We created those simulated spectra using the manual chemical shift assignment available for those proteins in the Biological Magnetic Resonance Bank [44]. In addition to the spin systems obtained from the manual chemical shift assignments, we randomly generated 20% extra spin systems. These new spin systems simulate the peaks that are false positives in Ubiquitin spectra. Each spin system consists of two CBCACON peaks (except if the previous residue is Glycine, the current residue has only one peak) and two CBCANCO peaks (except if the current residue is Glycine, it will have a single peak). The chemical shift value of one of the three coordinates of each peak was shifted randomly by a probability of 10%. This shifting was introduced to simulate the linking errors in the real spectra. The N and C coordinates were shifted by +0.1 or -0.1 ppm, and the CO coordinate is shifted by +0.02 or -0.02 ppm. We chose such shifting values because the resolution of our simulated spectra is relatively high (the dimensions of our simulated spectra are 512 (N dimension) $\times$ 256 (C dimension) $\times$ 2048 (CO dimension)). To simulate the false negatives in real spectra, each peak in our simulated spectra had a probability of 1% to be removed. In addition to that, for every slice in CBCACON and CBCANCO spectra, a new peak was randomly added with a probability of 10%. Adding these new peaks introduces ambiguity in creating the spin systems using our method. After constructing all the peaks, white Gaussian noise was added to each point in the spectra. The mean of the white Gaussian noise is zero.

TABLE 2 Resonance assignment performance over the simulated dataset. The first column presents the protein ID in the BMRB database. The second column lists the protein length. The third and fourth columns list the chemical shift assignment results in terms of recall and precision, respectively.

Protein	Length	Recall	Precision
4752	68	0.92	0.94
4144	78	0.87	0.92
4579	86	0.94	0.95
4288	105	0.90	0.92
4929	114	0.93	0.95
4302	115	0.89	0.90
4670	120	0.81	0.81
4353	126	0.91	0.96
4027	158	0.89	0.90
4318	215	0.76	0.78
4836	217	0.79	0.80
4022	260	0.76	0.78
Average	-	0.86	0.88
F1-score	-	0.87	

The standard deviation is the root of the peaks intensity. Thus, our simulated spectra contain ambiguous spin systems and noise because of false positives, false negatives, and linking errors. We ran our spin system forming method on the simulated dataset to create the spin systems. The assignment results using our genetic algorithm-based method for the 12 proteins are shown in Table 2. Analyzing the results we obtained for the simulated dataset, it appeared that the spin systems that should have been assigned to the wrongly assigned residues were not generated by our method. The slices responsible for creating those spin systems were extracted successfully. However, the chemical shift values of some of the peaks in those slices are slightly out of the theoretical ranges that we used for CA or CB in the step of creating the spin systems. Thus, according to our method, those spin systems were not created. Due to the nature of the genetic algorithm-based method that we use for the assignment step, a wrongly assigned residue may result in a wrong assignment of a fragment of few residues neighboring to it. This is the main reason for the unsatisfactory results we obtained for some of the proteins in the simulated dataset.

3.3 Structure Calculation Using CS-ROSETTA

NMR protein structure determination mainly consists of two stages. The first stage is the backbone chemical shift assignment. The second stage involves the assignment of side chain resonances and a collection of inter proton distance constraints from multi-dimensional NOE spectra. However, the sensitivity of these NOE experiments tends to be relatively low. Moreover, NOE spectra have extra resonance overlap that makes the analysis more difficult. A lot of efforts have been done to automate the analysis and explanation of the huge number of cross peaks in this type of spectra. Nevertheless, this step with side chain assignment is the bottleneck in the process of protein structure determination. A group of NMR researchers started to look at different ways to skip the step of NOE assignment and to determine the 3D structure directly from the stage of chemical shift assignment. Several papers suggested that

TABLE 3 Comparison between the quality of structures obtained for 'Ubiquitin'. The column 'Man' stands for the structures obtained when feeding CS-ROSETTA with the manual assignments as input. The column 'Ours' stands for the structures obtained when feeding CS-ROSETTA with the assignments predicted by our method as input. The 'No_{CS}' stands for the structures obtained when feeding CS-ROSETTA with the amino acid sequence only as input. '*RMSD_{TopScore}*' and '*RMSD_{BestOfTop10}*' stand for the structural model RMSD that has predicted to be top one in CS-ROSETTA, and the lowest RMSD within the top 10 predicted structural models via CS-ROSETTA, respectively. The values are in Å.

	Ubiquitin		
	Man	Ours	No _{CS}
<i>RMSD_{TopScore}</i>	1.57	1.96	15.87
<i>RMSD_{BestOfTop10}</i>	1.33	1.96	12.73

determination of protein structures can be directly done from chemical shifts [47], [48], [49], [50], [51], [52].

In this paper, we skipped the step of collecting and analyzing NOE spectra. Instead, we used CS-ROSETTA [51], [52] to determine the structure of Ubiquitin. CS-ROSETTA is a program that is based on ROSETTA [53], which predicts protein structures solely from the amino acid sequences of the proteins. However, in CS-ROSETTA, the structural information presented in the NMR chemical shifts are also introduced, which significantly improves ROSETTA performance. The output of CS-ROSETTA is an ensemble of three-dimensional structures. Each of the output structures has its energy score such that the lowest energy structure is most probably the nearest to the native structure of the target protein.

For Ubiquitin, we fed CS-ROSETTA with three types of inputs. First, we used the manual chemical shift assignment that we obtained from the NMR lab that carried out the experiments. Second, we used the predicted chemical shift assignment from our method. Finally, we used the empty assignment file to see how good CS-ROSETTA can predict structures solely from the amino acid sequence. Table 3 shows the obtained results for the three cases. Figure 5 shows how well our calculated structure aligns with the experimentally determined structure of Ubiquitin.

3.4 Implementation Details

Our spin system forming method is implemented in Matlab. The time-consuming parts in the method are implemented as sequential C functions (MEX files) that are called from the Matlab code. To form the spin systems using our method, it takes around 17 minutes on a 2.5 GHz CPU. This duration is reasonable considering that the size of each of the two spectra is around 1 GB.

3.5 Importance of Denoising NMR Spectra

In this section, we evaluate our method in terms of the tolerance to the noise level. We added 1%, 2%, 5%, and 10% noise to the real protein (Ubiquitin) spectra. We used Gaussian white noise with zero mean and squared root of the point intensity as the variance. We first randomly picked

Fig. 5: Structural alignment between the top scored structure produced by CS-ROSETTA with our assignments and the experimentally determined native structure of Ubiquitin. The structure in magenta is the calculated structure while the structure in cyan is the native structure.

TABLE 4 shows the effect of the noise with different levels on Ubiquitin in the final results. Second column is the manually assigned spin systems. Third column is the correctly assigned spin systems.

Added noise level	Manually assigned	Correctly assigned
1%	76	70
2%	76	70
5%	76	70
10%	76	61

1% of the 3D spectrum. In each point, we added the noise. We did this step for both spectra. Then, the approach was repeated for other different noise levels 2%, 5%, and 10%. The results are shown in Table 4. It shows that our method is robust to reasonable levels of noise, but for very high noise levels, the performance will be affected.

4 DISCUSSION

It is worth noting that the CBCANCO spectrum is a noisy spectrum with a low signal to noise ratio. We cannot just extract the highest slices in this spectrum without prior knowledge about the approximate locations of the correct slices. Otherwise, many necessary slices with relatively low-intensity peaks will not be picked, and many false positive slices with relatively high-intensity peaks will be picked instead. These missing correct slices will adversely affect all

the subsequent steps. Thus, using the locations of slices in the CBCACON spectrum to direct the slice picking from the CBCANCO spectrum is necessary.

1.5 N_p without feedback	52/68
3 N_p without feedback	55/70
5 N_p without feedback	57/72
7 N_p without feedback	59/75
9 N_p without feedback	63/75
1.5 N_p with feedback	70/76

the subsequent steps. Thus, using the locations of slices in the CBCACON spectrum to direct the slice picking from the CBCANCO spectrum is necessary.

In order to validate the importance of the feedback from the CBCACON spectrum to direct the slice picking in the CBCANCO spectrum, we tried to see the performance without using the feedback. We picked the highest 1.5 N_p , 3 N_p , 5 N_p , 7 N_p and 9 N_p slices from the CBCANCO spectrum. We used these picked slices to form spin systems. Subsequently, we used the obtained spin systems to get the chemical shift assignment using genetic algorithms.

As we expected, the lowest performance was obtained when we used the highest 1.5 N_p slices. The performance started to get better as the number of used slices increased. The peak performance was obtained when we used 9 N_p slices. The reason for the poor performance in cases of 1.5 N_p and 3 N_p is due to the large number of missing true slices that have relatively low intensities. This led to a large number of missing correct spin systems (22 missing correct spin systems when taking top 1.5 N_p and 14 missing correct spin systems when taking 3 N_p). In case of 5 N_p , 7 N_p , and 9 N_p , the number of missing true slices is smaller; thus we obtained relatively better results. Yet, the performance when using 9 N_p slices is still not as good as when we use 1.5 N_p slices that we pick using feedback information (see Table 5).

From the previous comparison, it is clear that we need an automatic method to select the number of slices in the CBCANCO noisy spectrum. It turned out that the most reliable method for selecting the slices is to use feedback from the more trustable CBCACON spectrum to direct the slice picking in the CBCANCO spectrum.

5 CONCLUSION

In this paper, we proposed for the first time slice picking idea in a spin system forming method that works with ^{13}C -detected spectra. Our method needs only two through-bond spectra to create the spin systems. We also proposed

for the first time using the feedback idea with the ^{13}C -detected spectra. We demonstrated that using this feedback from the more trustable spectrum to direct slice picking in the less trustable one improves the quality of spin system forming significantly. We developed a genetic algorithm-based method to assign our spin systems to the residues of the protein. Assignments produced by our method are proven to be good enough and can lead to accurate structures when given as inputs to the CS-ROSETTA server. The source code of our method is freely available at <http://sfb.kaust.edu.sa/Pages/Software.aspx>.

6 ACKNOWLEDGMENTS

The ^{13}C -detected spectra for Ubiquitin were provided by the Imaging and Characterization Core Lab, King Abdullah University of Science and Technology (KAUST). The research reported in this paper is supported by the funding from King Abdullah University of Science and Technology (KAUST).

REFERENCES

- [1] A. Shekhtman and D. S. Burz, Eds., *Protein NMR Techniques*, ser. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2012, vol. 831.
- [2] K. Wüthrich, "Nuclear magnetic resonance in protein research," *Experientia*, vol. 30, no. 6, pp. 577–585, Jun. 1974.
- [3] R. J. Abraham, J. Fisher, and P. Loftus, *Introduction to NMR Spectroscopy*. Wiley, 1988.
- [4] W. Bermel, I. Bertini, I. Felli, M. Piccioli, and R. Pierattelli, " ^{13}C -detected protonless NMR spectroscopy of proteins in solution," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 48, no. 1, pp. 25–45, Mar. 2006.
- [5] W. Bermel, I. Bertini, I. C. Felli, M. Matzapetakis, R. Pierattelli, E. C. Theil, and P. Turano, "A method for C^α direct-detection in protonless NMR," *Journal of Magnetic Resonance*, vol. 188, no. 2, pp. 301–310, Oct. 2007.
- [6] W. Bermel, I. C. Felli, R. Kmmmerle, and R. Pierattelli, " ^{13}C direct-detection biomolecular NMR," *Concepts in Magnetic Resonance Part A*, vol. 32A, no. 3, pp. 183–200, May 2008.
- [7] H. Kovacs, D. Moskau, and M. Spraul, "Cryogenically cooled probes: a leap in NMR technology," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 46, no. 2–3, pp. 131–155, May 2005.
- [8] X. Gao, "Recent advances in computational methods for nuclear magnetic resonance data processing," *Genomics, proteomics & bioinformatics*, vol. 11, pp. 29–33, Feb. 2013.
- [9] K. Wüthrich, *NMR with Proteins and Nucleic Acids*. Wiley, 1986, vol. 17.
- [10] A. Rouh, A. Louis-Joseph, and J.-Y. Lallemand, "Bayesian signal extraction from noisy FT NMR spectra," *Journal of Biomolecular NMR*, vol. 4, no. 4, pp. 505–518, Jul. 1994.
- [11] C. Antz, K.-P. Neidig, and H. Kalbitzer, "A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis," *Journal of Biomolecular NMR*, vol. 5, no. 3, Apr. 1995.
- [12] Y. Cheng, X. Gao, and F. Liang, "Bayesian peak picking for NMR spectra," *Genomics, Proteomics & Bioinformatics*, vol. 12, no. 1, pp. 39–47, Feb. 2014.
- [13] R. Koradi, M. Billeter, M. Engeli, P. Güntert, and K. Wüthrich, "Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY," *Journal of Magnetic Resonance*, vol. 135, no. 2, pp. 288–297, Dec. 1998.
- [14] G. J. Kleweg, R. Boelens, and R. Kaptein, "A versatile approach toward the partially automatic recognition of cross peaks in 2D ^1H NMR spectra," *Journal of Magnetic Resonance*, vol. 88, no. 3, pp. 601–608, Jul. 1990.
- [15] D. S. Garrett, R. Powers, A. M. Gronenborn, and G. M. Clore, "A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams," *Journal of Magnetic Resonance*, vol. 213, no. 2, pp. 357–363, Dec. 1991.
- [16] E. A. Carrara, F. Pagliari, and C. Nicolini, "Neural networks for the peak-picking of nuclear magnetic resonance spectra," *Neural Networks*, vol. 6, no. 7, pp. 1023–1032, Jan. 1993.
- [17] S. A. Corne, A. Johnson, and J. Fisher, "An artificial neural network for classifying cross peaks in two-dimensional NMR spectra," *Journal of Magnetic Resonance*, vol. 100, no. 2, pp. 256–266, Nov. 1992.
- [18] V. Y. Orekhov, I. V. Ibraghimov, and M. Billeter, "MUNIN: a new approach to multi-dimensional NMR spectra interpretation," *Journal of Biomolecular NMR*, vol. 20, no. 1, pp. 49–60, May 2001.
- [19] D. M. Korzhnev, I. V. Ibraghimov, M. Billeter, and V. Y. Orekhov, "MUNIN: application of three-way decomposition to the analysis of heteronuclear NMR relaxation data," *Journal of Biomolecular NMR*, vol. 21, no. 3, pp. 263–268, Nov. 2001.
- [20] B. Alipanahi, X. Gao, E. Karakoc, L. Donaldson, and M. Li, "PICKY: a novel SVD-based NMR spectra peak picking method," *Bioinformatics*, vol. 25, no. 12, pp. 268–275, Jun. 2009.
- [21] Z. Liu, A. Abbas, B.-Y. Jing, and X. Gao, "WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering," *Bioinformatics*, vol. 28, no. 7, pp. 914–920, Apr. 2012.
- [22] A. Abbas, X.-B. Kong, Z. Liu, B.-Y. Jing, and X. Gao, "Automatic peak selection by a Benjamini-Hochberg-based algorithm," *PLoS ONE*, vol. 8, no. 1, p. 53112, Jan. 2013.
- [23] C. V. Cannistraci, A. Abbas, and X. Gao, "Median modified Wiener filter for nonlinear adaptive spatial denoising of protein NMR multidimensional spectra," *Scientific Reports*, vol. 5, no. 1, Jul. 2015.
- [24] R. Tycko and K.-N. Hu, "A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning," *Journal of Magnetic Resonance*, vol. 205, no. 2, pp. 304–314, Aug. 2010.
- [25] Xiang Wan and Guohui Lin, "CISA: combined NMR resonance connectivity information determination and sequential assignment," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 336–348, Jul. 2007.
- [26] P. Güntert, M. Salzmann, D. Braun, and K. Wüthrich, "Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER," *Journal of Biomolecular NMR*, vol. 18, no. 2, pp. 129–137, Oct. 2000.
- [27] Y.-S. Jung and M. Zweckstetter, "MARS - robust automatic backbone assignment of proteins," *Journal of Biomolecular NMR*, vol. 30, no. 1, pp. 11–23, Sep. 2004.
- [28] K.-P. Wu, J.-M. Chang, J.-B. Chen, C.-F. Chang, W.-J. Wu, T.-H. Huang, T.-Y. Sung, and W.-L. Hsu, "RIBRA—an error-tolerant algorithm for the nmr backbone assignment problem," *Journal of Computational Biology*, vol. 13, no. 2, pp. 229–244, Apr. 2006.
- [29] J. E. Masse and R. Keller, "AutoLink: automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic," *Journal of Magnetic Resonance*, vol. 174, no. 1, pp. 133–151, May 2005.
- [30] B. E. Coggins and P. Zhou, "PACES: protein sequential assignment by computer-assisted exhaustive search," *Journal of Biomolecular NMR*, vol. 26, no. 2, pp. 93–111, Jun. 2003.
- [31] A. Lemak, C. A. Steren, C. H. Arrowsmith, and M. Llins, "Sequence specific resonance assignment via Multicanonical Monte Carlo search using an ABACUS approach," *Journal of Biomolecular NMR*, vol. 41, no. 1, pp. 29–41, May 2008.
- [32] J. Volk, T. Herrmann, and K. Wüthrich, "Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH," *Journal of Biomolecular NMR*, vol. 41, no. 3, pp. 127–138, Jul. 2008.
- [33] C. Bartels, M. Billeter, P. Güntert, and K. Wüthrich, "Automated sequence-specific NMR assignment of homologous proteins using the program GARANT," *Journal of Biomolecular NMR*, vol. 7, no. 3, May 1996.
- [34] C. Bartels, P. Güntert, M. Billeter, and K. Wüthrich, "GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra," *Journal of Computational Chemistry*, vol. 18, no. 1, pp. 139–149, Jan. 1997.
- [35] B. Alipanahi, X. Gao, E. Karakoc, S. C. Li, F. Balbach, G. Feng, L. Donaldson, and M. Li, "Error tolerant NMR backbone resonance assignment and automated structure generation," *Journal of Bioinformatics and Computational Biology*, vol. 09, no. 01, pp. 15–41, Feb. 2011.
- [36] R. Jang, X. Gao, and M. Li, "Towards automated structure-based NMR resonance assignment," *Lecture Notes in Computer Science*, vol. 6044, pp. 189–207, 2010.

- [37] —, “Combining automated peak tracking in SAR by NMR with structure-based backbone assignment from 15N-NOESY,” *BMC Bioinformatics*, vol. 13, no. Suppl 3, p. S4, 2012.
- [38] —, “Towards fully automated structure-based NMR resonance assignment of 15N-labeled proteins from automatically picked peaks,” *Journal of Computational Biology*, vol. 18, no. 3, pp. 347–363, Mar. 2011.
- [39] H.-N. Lin, K.-P. Wu, J.-M. Chang, T.-Y. Sung, and W.-L. Hsu, “GANA—a genetic algorithm for NMR backbone resonance assignment,” *Nucleic Acids Research*, vol. 33, no. 14, pp. 4593–4601, Aug. 2005.
- [40] A. Abbas, X. Guo, B.-Y. Jing, and X. Gao, “An automated framework for NMR resonance assignment through simultaneous slice picking and spin system forming,” *Journal of Biomolecular NMR*, vol. 59, no. 2, pp. 75–86, Jun. 2014.
- [41] X. Gao, “Mathematical approaches to the NMR peak-picking problem,” *Journal of Applied & Computational Mathematics*, vol. 01, no. 01, 2012.
- [42] C. V. Cannistraci, F. M. Montecvecchi, and M. Alessio, “Median-modified Wiener filter provides efficient denoising, preserving spot edge and morphology in 2-DE image processing,” *PROTEOMICS*, vol. 9, no. 21, pp. 4908–4919, Nov. 2009.
- [43] J. H. Holland, *Adaptation in Natural and Artificial Systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT Press, 1992.
- [44] B. R. Seavey, E. A. Farr, W. M. Westler, and J. L. Markley, “A relational database for sequence-specific protein NMR data,” *Journal of Biomolecular NMR*, vol. 1, no. 3, pp. 217–236, Sep. 1991.
- [45] D. E. Zimmerman, C. A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C.-y. Chien, R. Powers, and G. T. Montelione, “Automated analysis of protein NMR assignments using methods from artificial intelligence,” *Journal of Molecular Biology*, vol. 269, no. 4, pp. 592–610, Jun. 1997.
- [46] A. Bahrami, A. H. Assadi, J. L. Markley, and H. R. Eghbalian, “Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy,” *PLoS Computational Biology*, vol. 5, no. 3, p. 1000307, Mar. 2009.
- [47] D. S. Wishart and D. A. Case, “Use of chemical shifts in macromolecular structure determination,” *Methods in Enzymology*, vol. 338, pp. 3–34, Jan. 2001.
- [48] H. Gong, Y. Shen, and G. D. Rose, “Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly,” *Protein Science*, vol. 16, no. 8, pp. 1515–1521, Aug. 2007.
- [49] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo, “Protein structure determination from NMR chemical shifts,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9615–9620, Jun. 2007.
- [50] D. S. Wishart, D. Arndt, M. Berjanskii, P. Tang, J. Zhou, and G. Lin, “CS23D : a web server for rapid protein structure generation using NMR chemical shifts and sequence data,” *Nucleic Acids Research*, vol. 36, no. Web Server, pp. W496–W502, May 2008.
- [51] Y. Shen, R. Vernon, D. Baker, and A. Bax, “De novo protein structure generation from incomplete chemical shift assignments,” *Journal of Biomolecular NMR*, vol. 43, no. 2, pp. 63–78, Feb. 2009.
- [52] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, and A. Bax, “Consistent blind protein structure generation from NMR chemical shift data,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 12, pp. 4685–4690, Mar. 2008.
- [53] P. Bradley, “Toward high-resolution de novo structure prediction for small proteins,” *Science*, vol. 309, no. 5742, pp. 1868–1871, Sep. 2005.

Meshari Alazmi received the BS and MS degrees in computer science from University of Hail and University of Missouri, respectively. He is a PhD student in Computational Bioscience Research Center and in Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. His main areas of research include bioinformatics and machine learning.

Ahmed Abbas is an assistant professor in Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Egypt. He received his Master's degree in Computer Science from Ain Shams University in May 2010, and his PhD degree in Computer Science from King Abdullah University of Science and Technology (KAUST) in December 2015. His main research interests are in developing new methods to solve key open problems in automated NMR protein structure determination.

Xianrong Guo received the BS and PhD degrees in Chemistry from Peking University, China. He is a senior scientist in Imaging and Characterization Core Lab, King Abdullah University of Science and Technology (KAUST). His main areas of research include biomolecular structure and dynamics, NMR spectroscopy and imaging.

and systems biology.

Ming Fan is an Associate Professor at Department of Biomedical Engineering, College of Life Information Science and Instrument Engineering, Hangzhou Dianzi University. He finished his doctoral studies at the National Laboratory of Pattern Recognition, Chinese Academy of Sciences in 2011. He was Postdoc researcher in Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST). His research interests include medical image analysis

tion, computer aided diagnosis and bioinformatics for cancer research.

Lihua Li is currently the Director and Professor of the Institute of Biomedical Engineering and Instrumentation in Hangzhou Dianzi University. He received his B.S. and M.S. degrees from Xidian University in 1985 and 1988, respectively, and his Ph.D. from Southeast University in 1990. He was an Associate Professor at University of South Florida and a Member-in-Residence at the H. Lee Moffitt Cancer Center and Research Institute in USA. His current research interests include medical image processing and percep-

Xin Gao is an associate professor, and group leader of Structural and Functional Bioinformatics (SFB) group in Computational Bioscience Research Center, and in Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST). He received the BS and PhD degrees in computer science from Tsinghua University, China, and University of Waterloo, Canada, respectively. Before joining KAUST, he was a Lane Fellow in Lane Center

for Computational Biology at Carnegie Mellon University, US. His research interests include bioinformatics, computational biology, machine learning and algorithms. His group works on building computational models, developing machine learning techniques, and designing efficient and effective algorithms to solve key open problems along the path from genome-scale sequence analysis, to protein structure prediction/determination, to function annotation, and to understanding and controlling molecule behaviors in complex biological systems.