RAIRO-Oper. Res. 50 (2016) 341–349
DOI: 10.1051/ro/2015038

RAIRO Operations Research
www.rairo-ro.org

PROGRESS IN NUCLEAR VECTOR REPLACEMENT FOR NMR PROTEIN STRUCTURE-BASED ASSIGNMENTS *

Şeyma Çetinkaya[1], Şeyma Nur Ekren[1] and Mehmet Serkan Apaydin[2]

Abstract. Nuclear Magnetic Resonance (NMR) Spectroscopy is an important technique to obtain structural information of a protein. In this technique, an essential step is the backbone resonance assignment and Structure Based Assignment (SBA) aims to solve this problem with the help of a template structure. Nuclear Vector Replacement (NVR) is an NMR protein SBA program, that takes as input ^{15}N and H^N chemical shifts and unambiguous NOEs, as well as RDCs, HD-exchange and TOCSY data. NVR does not utilize ^{13}C chemical shifts although this data is widely available for many proteins. In addition, NVR is a proof-of-principle approach and has been run with specific and manually set parameters for some proteins. NA-NVR-ACO [M. Akhmedov, B. Çatay and M.S. Apaydın, *J. Bioinform. Comput. Biol.* **13** (2015) 1550020.] remedies this problem for the NOE data and standardizes NOE usage, while using an ant colony optimization based algorithm. In this paper, we standardize NA-NVR-ACO's scoring function by using the same parameters for all the proteins and incorporating $^{13}C_\alpha$ chemical shifts. We also use a larger protein database and state-of-the-art chemical shift prediction tools, SHIFTX2 [B. Han, Y. Liu, S.W. Ginzinger and D.S. Wishart, *J. Biomol. NMR* **50** (2011) 43–57.] and SPARTA+ [Y. Shen and A. Bax, *J. Biomol. NMR* **48** (2010) 13–22], to extract the chemical shift statistics. Other practical improvements include automatizing data file preparation and obtaining a degree of reliability for individual peak-amino acid assignments. Our results show that our improvements bring NA-NVR-ACO closer to a practical tool, able to handle a variety of different data types.

Mathematics Subject Classification. 90c27.

Received September 8, 2015. Accepted September 21, 2015.

1. Introduction

Proteins are macromolecules in living systems that serve crucial functions. It is important to determine the structure of a protein for drug design, understanding protein-protein and protein-ligand interactions and understanding the relationship between structure and function of the protein. To determine a protein's structure

Keywords. NMR structure based protein assignment, NVR; score function, triple resonance experiments, reliability of assignments.

* *This work was supported by following grants to M.S.A.: The Scientific and Technical Research Council of Turkey-CNRS research collaboration grant (2505) [113e085].*

[1] Department of Graduate School of Natural and Applied Sciences, İstanbul Şehir University, 34662 Üsküdar, Istanbul, Turkey. seymacetinkaya@std.sehir.edu.tr; seymaekren@std.sehir.edu.tr

[2] College of Engineering and Natural Sciences, İstanbul Şehir University, 34662 Üsküdar, Istanbul, Turkey. apaydin@sehir.edu.tr

Article published by EDP Sciences

© EDP Sciences, ROADEF, SMAI 2016

there are two main techniques, X-Ray Crystallography and NMR spectroscopy. Most of the protein structures in the Protein Data Bank are solved by X-Ray Crystallography. In this technique the protein is crystallized. However, the crystalline form of the protein may be different than the form in solution and some proteins cannot be crystallized. For such proteins, NMR provides a good alternative. In NMR, the protein is often examined in solution. However, NMR is limited by protein size, larger proteins usually result in more missing and overlapping signals. One of the important steps in determining the protein structure starting from NMR data is to assign these peaks to the corresponding amino acids.

The assignment of proteins in NMR laboratories is a laborious process. Software programs exist to help with assignments semi-automatically (using *e.g.* Analysis [19]) or automatically (using *e.g.* Flya [15] or Mars [11]), and fully automated assignment of small proteins is possible [15]. Although there exist software to automate the assignment process, manual analysis of NMR spectra is the most reliable method. Manual verification of assignments are almost always done to handle possible errors, since automation is not trustworthy [10]. Moreover, for large proteins the assignment step can take weeks and even months [10], since the available data is incomplete and ambiguous [13]. Other challenges in obtaining the assignments is that the spectra can be crowded, noisy and there may be extra and missing peaks. Automatizing the assignment process with high accuracy is important in order to expedite NMR protein structure determination. Structure Based Assignment (SBA) achieves this objective with the help of a template structure that is homologous to the target protein. The knowledge of template provides prior information about the structure of the target protein and allows to obtain more reliable assignment results.

Nuclear Vector Replacement (NVR) is an approach that solves the protein SBA problem. NVR-EM [12] uses expectation maximization algorithm and finds a local optimum solution. NVR-BIP [4] uses binary integer programming to obtain the global solution for the problem. However, it is unable to obtain the assignments for larger proteins due to the considerable resources such an exact solution requires. For such proteins, metaheuristic approaches such as NVR-TS [7] and NVR-ACO [5] have been developed. NVR-TS uses tabu search and NVR-ACO uses ant colony optimization to arrive at a solution.

In addition to protein NMR resonance assignment, algorithms from operations research have been applied to the RNA resonance assignment problem as well. These include tabu search [6], finding the longest orderly colored longest path [18] or Hamiltonian path finding [1], and beam search [17]. In this paper, our contributions are:

1. Automatization of input data preparation.
2. Providing a measure of reliability of assignments.
3. Extending NVR to use alpha Carbon chemical shifts.
4. Using novel chemical shift prediction tools.

In the following section we describe NVR and the problem formulation. In Section 3, we describe our contributions. We give our test results in Section 4. Then, we conclude the paper and discuss future work.

2. NVR FRAMEWORK

NVR is a framework for the NMR structure based assignment problem that tries to find the optimal matching between the set of amino acids and the set of peaks using only backbone amide proton and nitrogen chemical shifts, and backbone NOEs. It can also use RDCs, TOCSY and Hydrogen-Deuterium exchange data if available.

NVR-BIP is a binary integer programming based approach that computes the assignments using CPLEX. It minimizes the score of the assignment subject to NOE constraints and finds the optimal solution for small proteins (less than approximately 150 amino acids) and has high assignment accuracies. However, NVR-BIP is unable to compute a solution for large proteins due to the large number of constraints. For such proteins, metaheuristic based approaches within the NVR framework, such as NVR-TS and NVR-ACO have been developed. NVR-TS is a tabu search based approach to the problem. Instead of applying hard constraints and disallowing NOE violations, NVR-TS uses a penalty term for NOE violations and can find a solution for large proteins. NVR-ACO is the first application of ant colony optimization to the problem and is based on the observation of

the behavior of real ant colonies searching for food sources. It finds the optimal solution for small proteins and can find solutions for large proteins with high accuracies. NVR-ACO uses backbone NOEs, however does not differentiate between HN-HA and HN-HN NOEs, and sets NOE distance thresholds (UB value in the formulation below) manually. NOE-aware version NA-NVR-ACO differentiates the type of backbone NOE and uses the appropriate coordinates from the template structure, and also obtains the NOE upperbound information directly from the NOE intensities in the data.

NA-NVR-ACO mathematical model is as follows:

Notation:

P: set of peaks
A: set of amino acids
T: set of distance types, $T = \{HN - HN, HN - HA, HA - HN\}$
s_{ij}: score associated with assigning peak i to amino acid j
N: number of peaks to be assigned ($N \leq |P|$)
d_{jlt}: distance between amide protons of amino acids j and l by using distance type t
$NOE(i)$: set of peaks that have an NOE with peak i
UB_{ik}: NOE upper bound distance limit between peaks i and k

$$b_{ijklt} = \begin{cases} 1, & \text{if } d_{jlt} \leq UB_{ik} \\ 0, & \text{otherwise} \end{cases}$$

Decision variables:

$$x_{ij} = \begin{cases} 1, & \text{if peak } i \text{ is assigned to amino acid } j \\ 0, & \text{otherwise} \end{cases}$$

Mathematical model:

$$\text{Minimize} \sum_{i \in P} \sum_{j \in A} s_{ij} x_{ij} \tag{2.1}$$

$$\text{s.t.} \sum_{i \in P} x_{ij} \leq 1, \ \forall j \in A \tag{2.2}$$

$$\sum_{i \in A} x_{ij} \leq 1, \ \forall j \in P \tag{2.3}$$

$$\sum_{i \in P} \sum_{j \in A} x_{ij} = N \tag{2.4}$$

$$x_{ij} + x_{kl} - 1 \leq b_{ijklt} \ \forall j, l \in A, \ \forall i \in P, \ \forall t \in T, \ \forall k \in NOE(i) \tag{2.5}$$

$$x_{ij} \in \{0, 1\}, \ \forall i \in P, \ \forall j \in A. \tag{2.6}$$

In this model, the objective function (1) minimizes the total score of assigning peaks to amino acids. Constraints (2) ensure that each amino acid is assigned to at most one peak and constraints (3) guarantee that each peak is mapped to at most one amino acid. Constraint (4) determines the number of peaks that are going to be assigned. This allows us to obtain a partial assignment. Constraint set (5) requires peaks i and k which have an NOE between them of type t to be assigned to amino acids j and l if the distance between these amino acids (d_{jlt}) is less than UB_{ik} Constraint set (6) forces the decision variables to be binary.

2.1. NA-NVR-ACO scoring function

In all versions of NVR, a scoring function that computes the assignment probabilities of the set of peaks to the set of amino acids plays an essential role. This scoring function is developed in reference [12] and is

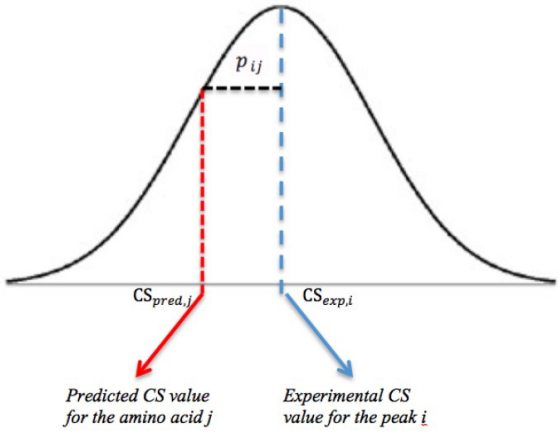

FIGURE 1. Computing CS score.

presented here for completeness. It takes chemical shift (CS), residual dipolar coupling (RDC), ^{15}N-TOCSY and Hydrogen-Deuterium exchange data and based on the difference between observed and predicted values it computes a probability. It uses BMRB statistics as well as predictions from CS prediction tools SHIFTS [20] and SHIFTX [14] in order to compute the CS score into the score function.

The score of CS data is calculated as follows:

$$S_{CS,ij} = -\log(p_{ij}). \tag{2.7}$$

Here, p_{ij} is the assignment probability of peak i and amino acid j according to the CS value and is computed by converting the difference between the experimental CS value of peak i and the expected CS value (obtained from BMRB statistics or from SHIFTS/SHIFTX predictions) of amino acid j to a probability using a Gaussian distribution (Fig. 1). Next, this probability is converted into a score for each peak-amino acid pair by taking its negative logarithm.

The scores for each data type and prediction tool are added and a total score is computed for each peak-residue pair. Finding an assignment for each peak that minimizes this total score subject to NOE constraints is the objective of the assignment problem.

$$S_{ij} = S_{RDC,ij} + S_{CS,ij} + S_{TOCSY,ij} + S_{HD-exchange,ij} + S_{SHIFTX,ij} + S_{SHIFTS,ij}. \tag{2.8}$$

In this equation, S_{ij} is the total assignment score of peak i to the amino acid j and is computed by adding the scores obtained from different data types.

In more detail, p_{ij} is computed as follows:

$$p_{ij} = \begin{cases} p_{\text{gauss}(ij)}, & \frac{(e_{ij}-\mu)}{\sigma} < \text{max std} + OR \\ & \frac{(\mu-e_{ij})}{\sigma} < \text{max std} - \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

Here, for a representative sample of BMRB, for every amino acid and secondary structure type, chemical shift values of H^N and ^{15}N atoms are extracted. Then, the mean, standard deviation, maximum standard deviations above and below the mean (max std + and max std −, respectively) of these chemical shift values are computed.

This procedure is repeated for predicted chemical shift values taken from SHIFTX and SHIFTS programs. Using these statistical values, the assignment probability of peak i to amino acid j (p_{ij}) is computed. In NVR-EM, chemical shift values of H^N and ^{15}N atoms in every amino acid of the target protein are compared with

these statistical values. If the chemical shift values of H^N and ^{15}N atoms are further apart from the mean than max std+ or max std−, then p_{ij} is assigned to zero. Otherwise, p_{ij} is computed according to the Gaussian density function.

The main problem encountered while computing the probability is that for some proteins for the correct assignment the experimental chemical shift is more distant than the max std +/− value. In order to allow the correct assignment for such cases, NVR-BIP, NVR-TS and NVR-ACO use special coefficients for some test proteins which enlarge the max std value. While this is adequate for a proof-of-principle, in order to enable NVR to work on a larger set of proteins it is necessary to use the same parameters for all the proteins. This standardization would improve the use of NA-NVR-ACO in practical applications.

2.2. Template selection

NA-NVR-ACO requires a template structure in order to compute the scoring matrix and to obtain the distance constraints to be used with NOE data. This template structure can be an X-ray structure corresponding to the same protein, or could be a structural homolog. In this paper we have used the X-ray structure as the template. Previous work [3] involved using more distant templates and improving the assignment accuracy of NVR-EM.

3. Methods

3.1. Providing a measure of reliability of assignments

NA-NVR-ACO can find the optimal solution for small proteins. For large proteins, the assignment results are distinct in different runs due to a lack of convergence to a global minimum in a very large search space. In that case, the individual result of a single assignment run is unreliable. We hypothesized that in the lack of convergence, the assignments that are more likely to be correct will occur many times in multiple runs whereas the incorrect assignments will differ. Therefore, for such large proteins, rather than computing a single assignment, we computed an ensemble of assignments and we calculated how many times a peak is assigned to the same amino acid. We determined the assignment of a peak as strong, if it is assigned to the same amino acid more than 60% of the time in all the runs. By this way, we also derived information about the reliability of our assignments, as the ratio of the number of times a peak has been assigned to a given residue over the total number of runs. This is similar to [3], but instead of using multiple templates to obtain different assignments, we use the assignment result of multiple runs.

Moreover, by using an ensemble of assignment results, we combined all assignments by obtaining a bipartite graph where a set of nodes corresponds to the peaks and the other set corresponds to the residues. The edges between peaks and residues are associated with a score corresponding to the number of times the peak is assigned to the corresponding amino acid in the assignment ensemble. We obtained a maximum bipartite matching using Hungarian algorithm. With this method, we obtain a final assignment that aggregates the results from all of the assignments.

3.2. Automatization of the data preparation

In order to run NA-NVR-ACO, a sequence of steps should be followed to prepare input files from the NMR data and the template structure. This procedure includes computing distances between protons in the PDB structure, obtaining the scoring matrix by using chemical shift prediction programs such as SHIFTS and SHIFTX and combining the NMR data coming from different sources corresponding to the same peak. We simplified this process by automating these steps and enabled running NA-NVR-ACO on novel proteins faster. The pseudo-code is as follows:

```
/* Parsing steps */
parsedResonancesFile ← parseResonanceFile(experimentalShiftFileName)
parsedPDBfile ← parsePDB_File(PDBbaseName)
```

$secondaryStructureFile \leftarrow parseSSE\_Info(parsedPDBfile)$
$NHvectorsFile \leftarrow parseVectors\_N-H(parsedPDBfile)$
$SHIFTXFile \leftarrow shiftx(PDBbaseName)$
$parsedSHIFTXFile \leftarrow parseSHIFTX\_File(SHIFTXFile)$
$SHIFTSFile \leftarrow shifts(PDBbaseName)$
$parsedSHIFTSFile \leftarrow parseSHIFTS\_File(SHIFTSFile)$

/* *Assembly step* */
$InputFileOfNVR \leftarrow assembleInput(parsedResonancesFile, NHvectorsFile,$
$secondaryStructureFile, parsedSHIFTXFile, parsedSHIFTSFile)$

parseResonanceFile parses the resonance file and extracts H^N, N chemical shifts. $parsePDB\_File$ parses the template PDB file to extract H^N, N, H_α and C_α coords. $parseSSE\_Info$ parses the secondary structure information of the template protein. $parseVectors\_N-H$ calculates N-H bond vectors from PDB file.

SHIFTX [14] and SHIFTS [20] are chemical shift prediction tools, $parseSHIFTX\_File$ and $parseSHIFTS\_File$ read the output of these and extract N, H^N and C_α chemical shifts.

Finally, *assembleInput* combines all of the files that are extracted.

3.3. Incorporating ^{13}C labeled data into NA-NVR-ACO

The different NVR approaches only use H^N and ^{15}N chemical shifts and do not utilize $^{13}C_\alpha$ chemical shifts. However, triple resonance experiments are widely available and many proteins solved using NMR have $^{13}C_\alpha$ chemical shift data. Besides, using $^{13}C_\alpha$ chemical shift information gives us extra information that potentially provides higher accuracies. Therefore, in this work, we incorporated $^{13}C_\alpha$ chemical shifts into NA-NVR-ACO program. We updated the scoring function of NVR according to this new data type. Similar to ^{15}N and H^N chemical shifts, the $^{13}C_\alpha$ chemical shifts were assumed to have a Gaussian distribution, and independent of the ^{15}N and H^N chemical shifts.

3.4. Extending the subset that is used for the statistical values

In NVR-EM, the statistical values used for computing the assignment probability of a peak to an amino acid have been obtained from a protein dataset of 457 proteins from BMRB.

In this paper, we use new statistical values that we obtain from a larger protein database. The new protein database includes 805 proteins chosen from BMRB and there is up to 40% similarity between the proteins according to their primary structures.

Note that our template structures are X-ray structures corresponding to the same protein. Our database provides the statistics required for computing the scoring function, and is not related to the selection of the template structure.

3.5. Using novel chemical shift prediction tools

For every protein in the new protein database, chemical shift statistics have been obtained using SHIFTX2 [9] and SPARTA+ [16] and these predictions have been used in the scoring function instead of the predicted values obtained by SHIFTX and SHIFTS.

The new scoring function is calculated as follows:

$$S_{ij} = S_{RDC,ij} + S_{CS,ij} + S_{TOCSY,ij} + S_{\mathrm{HD-exchange},ij} + S_{\mathrm{SHIFTX2},ij} + S_{\mathrm{SPARTA+},ij} \tag{3.1}$$

4. Results

4.1. Reliability results

We took 25 different assignment results of MBP using NA-NVR-ACO. Among these ensemble of assignment results, the assignment with minimum score has 59% accuracy. The individual assignment accuracies range between 53% and 70% and the average assignment accuracy is 64%.

MBP has 335 peaks that are all assigned. By using our reliability measure, we found that 202 peaks (60% of the peaks) were assigned to the same amino acid in 25 runs in at least 60% of the runs, and these peaks had 92% accuracy. This information could be used to partially assign the peaks with high accuracy. Additional experiments could be done for the remaining peaks to assign them correctly. Furthermore, by using Hungarian algorithm we combined the assignment results of 25 runs and obtained an assignment accuracy of 73% for all the peaks.

4.2. Automatization of input data preparation results

We study a new protein molecular-weight-protein tyrosine phosphatase A (MptpA, 150 amino acids) that is not in the set of test proteins of NA-NVR-ACO. The process of extracting the input data of NVR required almost a week to complete. Then, we automatized this process using a combination of bash, perl and matlab scripts. With the automatization, we can obtain our datafiles in a few minutes. We simulated unambiguous NOEs of MptpA and computed its assignments using NA-NVR-ACO. We obtained an assignment accuracy of 90%. We also tested with Carbon chemical shifts and achieved 100% assignment accuracy.

4.3. Novel scoring function incorporating $^{13}C_\alpha$ chemical shifts

In this section, we compare our results with those obtained by NA-NVR-ACO that uses the previous scoring function. Our new scoring function uses a larger database of proteins and $^{13}C_\alpha$ chemical shifts, as well as SHIFTX2 and SPARTA+ instead of SHIFTS and SHIFTX. We define the assignment accuracy as the ratio of the number of correctly assigned peaks to the total number of assigned peaks.

Note that the assignment accuracies are already high for most proteins using NA-NVR-ACO and they increase modestly for five proteins and decrease slighty for one protein (1LYZ). For three peaks of 1LYZ, experimental values are far from the predicted values. Since we use standard parameters, the assignment probabilities of these three peaks are assigned to zero. Therefore, the accuracy result is lower for this protein. It is remarkable that with the standard scoring function we obtain about the same or better accuracies as the previous scoring function which used different parameters for different proteins.

5. Conclusion and future work

In this paper, we have performed the following steps to improve and automate NA-NVR-ACO.

- We generated a method to determine the reliability of the assignments and we developed an ensemble based method to enhance the assignment accuracy. We tested our method on MBP and improved the assignment accuracy and provided a degree of reliability of assignments.
- To facilitate the study on new proteins, we simplified input data preparation process.
- We standardized NVR's scoring function, used a larger database and SHIFTX2 and SPARTA+ chemical shift prediction tools to extract statistics, and included carbon chemical shifts.

TABLE 1. (a) Accuracy results obtained by NA-NVR-ACO with the previous scoring function (b) Accuracy results with the new scoring function and $^{13}C_\alpha$ chemical shifts.

Protein Family	No of Residues	PDB ID	Accuracy of NA-NVR-ACO	Accuracy of NA-NVR-ACO with new scoring function
Lysozyme	126	1AKI	100%	100%
		1AZF	98%	98%
		1BGI	100%	100%
		1H87	100%	100%
		1LSC	100%	100%
		1LSE	100%	100%
		1LYZ	96%	94%
		2LYZ	98%	98%
		3LYZ	98%	100%
		4LYZ	96%	98%
		5LYZ	96%	98%
		6LYZ	100%	100%
		193L	100%	100%
Ubiquitin	72	1AAR	97%	97%
		1UBQ	100%	100%
		1G6J	97%	100%
		1UBI	100%	100%
The Rest	55	3GB1	100%	100%
	96	2A7O	89%	93%
	80	ff2	91%	91%
	243	EIN	100%	100%

With these improvements, NVR becomes closer to being a practical tool to be useful in an NMR laboratory. The time it takes to obtain the assignments for a novel protein using NVR is significantly reduced and NVR is able to handle different types of data, such as Carbon chemical shifts. Its scoring function no longer has special parameters for different proteins and can work without any manual adjustments for new test cases. It can work with more noise in the data. It must be mentioned that the reliability information for peaks is available for large proteins for which the global optimal solution is not found. For such proteins, the assignment results differ from run to run. Note that even though it seems that the assignment accuracy changes slightly with the new scoring function, the results with the new scoring function have been obtained using the same parameters for all the proteins and therefore these results are more robust. One step that remains to increase the usability of NVR is to enable it to handle ambiguous NOEs. Obtaining enough unambiguous NOEs from raw data is a challenge and may require performing 4D NOESY experiments which are not always available. While handling ambiguous NOEs, we will distinguish aromatic and aliphatic protons which have similar chemical shifts using the template structure information. In addition, we plan to distinguish the proton atoms of NH_2 in amino acids *Gln* and *Asn* from amide protons in the HSQC spectra, a step commonly performed manually. Even though these protons provide little information about the structure, they still need to be removed from the spectra before the assignment stage. Finally, we plan to assign larger proteins based on methyl group NOEs [8].

Acknowledgements. We thank Ewen Lescop for discussions and Ulrich Schieborr for NMR data of MptpA.

References

[1] R.W. Adamiak, J. Blazewicz, P. Formanowicz, Z. Gdaniec, M. Kasprzak, M. Popenda and M. Szachniuk, An Algorithm for an Automatic NOE Pathways Analysis of 2D NMR Spectra of RNA Duplexes. *J. Comp. Biol.* **11** (2004) 163–180.

[2] M. Akhmedov, B. Çatay and M.S. Apaydın, Automating unambiguous NOE data usage in NVR for NMR protein structure-based assignments. *J. Bioinform. Comput. Biol.* **13** (2015) 1550020.

[3] M.S. Apaydın, V. Conitzer and B.R. Donald, Structure-based protein NMR assignments using native structural ensembles. *J. Biomol. NMR* **40** (2008) 263–276.

[4] M.S. Apaydın, B. Çatay, N. Patrick and B.R. Donald, NVR-BIP: Nuclear vector replacement using binary integer programming for NMR structure-based assignments. *Comput. J.* **54** (2011) 708–716.

[5] J. Aslanov, B. Çatay and M.S. Apaydın, An Ant Colony Optimization Approach for Solving the Nuclear Magnetic Resonance Structure Based Assignment Problem. GECCO (2013).

[6] J. Blazewicz, M. Szachniuk and A. Wojtowicz, RNA tertiary structure determination: NOE pathways construction by tabu search. *Bioinform.* **21** (2005) 2356–2361.

[7] G. Cavuslar, B. Çatay and M.S. Apaydın, A Tabu search approach for the NMR protein structure-based assignment problem. *IEEE/ACM Trans. Comput. Biology Bioinform.* **9** (2012) 1621–1628.

[8] F.A. Chao, J. Kim, Y. Xia, M. Milligan, N. Rowe and G. Veglia, FLAMEnGO 2.0: An enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *J. Magn. Reson.* **245** (2014) 17–23.

[9] B. Han, Y. Liu, S.W. Ginzinger and D.S. Wishart, SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* **50** (2011) 43–57.

[10] R. Jang, *Fast and Robust Mathematical Modeling of NMR Assignment Problems*. Ph.D. thesis, University of Waterloo, Canada (2012).

[11] Y.S. Jung and M. Zweckstetter, Backbone assignment of proteins with known structure using residual dipolar couplings. *J. Biomol. NMR* **30** (2004) 25–35.

[12] C.J. Langmead and B.R. Donald, An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J. Biomol. NMR* **29** (2004) 111–138.

[13] C.A. MacRaild and R.S. Norton, RASP: rapid and robust backbone chemical shift assignments from protein structure. *J. Biomol. NMR* **58** (2014) 155–163.

[14] S. Neal, A.M. Nip, H. Zhang and D.S. Wishart, Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR* **26** (2003) 215–240.

[15] E. Schmidt and P. Güntert, A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.* **134** (2012) 12817–12829.

[16] Y. Shen and A. Bax, SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48** (2010) 13–22.

[17] M. Szachniuk, M. Malaczynski, E. Pesch, E.K. Burke and J. Blazewich, MLP accompanied beam search for the resonance assignment problem. *J. Heuristics* **19** (2013) 443–464.

[18] M. Szachniuk, M.C. De Cola, G. Felici and J. Blazewich, The orderly colored longest path problem - A survey of applications and new algorithms. *RAIRO: OR* **48** (2014) 25–51.

[19] W.F. Vranken, W. Boucher, T.J. Stevens, R.H. Fogh, A. Pajon, M. Llinas, E.L. Ulrich, J.L. Markley, J. Ionides and E.D. Laue, The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins: Structure, Function, and Bioinform.* **59** (2005) 687–696.

[20] X.P. Xu and D.A. Case, Automated prediction of 15N, 13Cα, 13Cβ and 13C chemical shifts in protein using a density functional database. *J. Biomol. NMR* **21** (2001) 321–333.