GANA–A Genetic Algorithm for NMR Backbone Resonance Assignment

Hsin-Nan Lin, Kuen-Pin Wu, Jia-Ming Chang, Ting-Yi Sung and Wen-Lian Hsu[1]
Institute of Information Science, Academia Sinica, Taipei, Taiwan
{arith, kpw, jmchang, tsung, hsu}@iis.sinica.edu.tw

Abstract

Automated backbone resonance assignment is very challenging because NMR experimental data from different experiments often contain errors. We developed a method, called GANA, which uses a genetic algorithm to perform backbone resonance assignment with high precision and recall. GANA takes spin systems as input data, and assigns spin systems to each amino acid of a target protein. We use the BMRB dataset (901 proteins) to test the performance of GANA. We also generate four datasets from the BMRB dataset to simulate data errors of false positive, false negative, linking error, and a mixture of the above three cases to examine the fault tolerance of our method. The average precision and recall rates of GANA on BMRB and the four simulated test cases are above 95%. Furthermore, we test GANA on two real wet-lab datasets: hbSBD and hbLBD. The precision and recall rates of GANA on these two datasets are 95.12% and 92.86% for hbSBD and 100% and 97.40% for hbLBD.

1. Introduction

Nuclear Magnetic Resonance spectroscopy (NMR) provides an alternative method to X-ray diffraction for determining the three-dimensional structures of proteins in atomic resolution. Researchers usually conduct several 3-D NMR experiments such as CBCANH, CBCA(CO)NH, or HN(CO)CA on $^{13}C/^{15}N/^{1}HN$-labeled proteins, and 2-D NMR experiments such as HSQC on $^{15}N/^{1}HN$-labeled proteins. The first requirement for these studies is sequential resonance assignment on backbone structures.

In the past, biologists had to make backbone assignments manually or semi-manually during the process of spectra analysis. Therefore, many automated tools using computational technologies have been developed to deal with the problem. However, NMR experimental data often contain the following four types of errors: noise (false positives), missing peaks (false negatives), clustered peaks, and inconsistent results among different experiments. As these four types of data errors appear in the NMR spectra, the process of automated backbone resonance assignment is very challenging.

2. Methods

We use HSQC, CBCANH, and CBCA(CO)NH spectral data to assign chemical shifts to N, H^N, C^α and C^β atoms on the backbone structure of a target protein. Cross-referencing the HSQC, CBCANH, and CBCA(CO)NH peaks for the i-th residue, we can generate two consecutive spin systems, i.e., an *inter-spin system*, denoted by $SS_{inter}(i)$, and an *intra-spin system*, denoted by $SS_{intra}(i)$. $SS_{inter}(i)$ contains the chemical shifts of C^α_{i-1}, C^β_{i-1}, and H^N_i, N_i, and $SS_{intra}(i)$ contains the chemical shifts of C^α_i, C^β_i and H^N_i, N_i.

We use *SSGroup* to denote a set of two consecutive spin systems in which all systems have identical chemical shifts of HN and N. We construct two data structures: (1) *candidate list*s to record potential *SSGroup*s for each residue in a target sequence; (2) *adjacency lists* for each $SSGroup_i$ to express the connectivity relationships between $SSGroup_i$ and all other $SSGroup_j$.

Resonance assignment is to assign *SSGroup*s to the protein sequence one by one such that the intra-spin system of (i-1)-th residue is similar with the inter-spin system of i-th residue. The fitness score of an assignment is given by $\sum_{i=1}^{l} LS(i)$, where $LS(i)$ is the linking score of i-th residue with (i-1)-th and (i+1)-th residues and l is the length of the protein (details omitted).

We use *candidate lists* and *adjacency lists* in basic operations of our genetic algorithm to find the assignment with largest fitness score among large different orderings of *SSGroup*s (details omitted).

[1] The corresponding author. Email: hsu@iis.sinica.edu.tw

3. Experimental results

The parameters used in each single *round* of GANA are as follows: the number of chromosomes in each generation = 600, the number of generations for evolution in a single round = 500. Because GAs may fall into a local maximum, we perform multiple rounds to select the chromosome with the highest fitness score as the final assignment for each protein. We test GANA on different datasets including BMRB, real wet-lab datasets and synthetic datasets.

After downloading the full BMRB dataset containing 3,129 proteins on September 10, 2004, we chose proteins of length 50 to 400 that have at least 50% residues with known answers as our dataset. The resulting dataset contains 901 proteins. For each test protein, we generate simulated *SSGroup*s according to the chemical shifts assigned to each residue.

The single round precision and recall of GANA on the BMRB dataset are 99.27% and 98.88%, respectively; and those after ten rounds are 99.40% and 99.08%, respectively.

We also use two real wet-lab datasets: the substrate binding domain of BCKD (hbSBD) and the lipoic acid bearing domain of BCKD (hbLBD, [2]). Each one contains more than 50% false positives and false negatives. The single round precision and recall of GANA on hbSBD are 95.12% and 92.86%, respectively; and those of hbLBD are 100% and 97.40%, respectively.

We regard the data from the raw BMRB dataset as the perfect case. To simulate real-world noises, we modify the original BMRB data to generate four kinds of synthetic cases: false positive, false negative, linking errors, and a mixture of these three test cases. Since the error type of clustered peaks is primitive in the original BMRB dataset, we don't simulate it.

For false positive cases, we add synthetic intra- and inter-spin systems into the *SSGroup*s. To create false negative cases, we assume that some CBCA(CO)NH peaks are missing. In this situation, we cannot recognize which C^{α} or C^{β} peak in CBCANH experiments belongs to the inter-residue. Thus, we generate all possible combinations of spin systems to solve the problem. For linking error cases, we modify the C^{α} and C^{β} chemical shifts of the inter-spin systems for all *SSGroup*s.

The experiment results of GANA on these synthetic datasets are listed in Table 1.

Table 1. Experiment results of GANA on different cases and rounds

	1 round		10 round	
Testing Cases	P	R	P	R
False Positive	99.22	98.85	99.36	99.00
False Negative	98.89	98.37	99.11	98.69
Linking Errors	97.85	97.03	98.30	97.53
Mixture	95.81	94.89	97.12	96.36

P and R denote the precision and recall, respectively. All values are percentages (%).

We compared GANA with PACES [3] and Mars [4] on the datasets specified in their papers. GANA and PACES perform well on the BMRB datasets. In the synthetic dataset of linking errors, the precision and recall rates of this data set are 95.24% and 85.74% for PACES and 97.82% and 97.12% for GANA, where GANA has a much better recall. Furthermore, the recall rates of MARS and GANA are 93.65% and 96.55%, respectively.

4. Conclusion

In this paper, we present a genetic algorithm GANA for backbone resonance assignment which is fully automatic. Under the same testing condition and data set, GAGA outperforms PACES and Mars especially in the recall rates. Recall indeed represents the accuracy of an assignment system. The higher recall of GANA can be attributed to its two data structures: candidate lists and adjacency lists. GANA takes spin systems as input data, and uses these two data structures to assign the spin systems to amino acids of a target protein. This design enables GANA to correctly map nearly all spin systems onto a target protein. Thus, the recall rates of GANA are generally high.

5. Acknowledgement

This work is supported in part by the thematic program of Academia Sinica under grant AS91IIS1PP and 94B003.

6. References

[1] Moseley, H. and Montelione, G. (1999) Automated analysis of nmr assignments and structures for proteins. *Curr. Opin. Struc. Biol.,* **9**, 635–642.

[2] Chang, C., Chou, H., Chuang, J., Chuang, D., and Huang, T. (2002) Solution structure and dynamics of the lipoic acid-bearing domain of human mitochondrial branched-chain alpha-keto acid dehydrogenase complex. *J. Biol. Chem.,* **277**, 15865–15873

[3] Coggins, B. and Zhou, P. (2003) Paces: Protein sequential assignment by computerassisted exhaustive search. J. Biomol. Nmr., 26, 93–111.

[4] Jung, Y.S. and Zweckstetter, M. (2004) Mars - robust automatic backbone assignment of proteins. Journal of Biomolecular Nmr, 30, 11-23.