

# NMR Protein Resonance Assignment Problem

Guohui Lin

Department of Computing Science

University of Alberta

[ghlin@cs.ualberta.ca](mailto:ghlin@cs.ualberta.ca)

# NMR Protein Resonance Assignment Problem

Outline:

- [Introduction](#) — protein structure and function determination and prediction

# NMR Protein Resonance Assignment Problem

Outline:

- [Introduction](#) — protein structure and function determination and prediction
- NMR protein structure determination

# NMR Protein Resonance Assignment Problem

## Outline:

- **Introduction** — protein structure and function determination and prediction
- NMR protein structure determination
- **NMR resonance assignment problem**

# NMR Protein Resonance Assignment Problem

## Outline:

- **Introduction** — protein structure and function determination and prediction
- NMR protein structure determination
- **NMR resonance assignment problem**
  - Complexity and approximations
  - Branch-and-bound
  - IDA\* search
  - Experimental results
  - Concluding remarks

When a new protein is sequenced, ...

It usually goes through the following processes:

1. The new protein is double checked if it **has been sequenced** and stored in some existing databanks; If so, its functions are reported.

When a new protein is sequenced, ...

It usually goes through the following processes:

1. The new protein is double checked if it **has been sequenced** and stored in some existing databanks; If so, its functions are reported.
2. Otherwise, it goes through a process called '**homolog search**': other proteins sharing high sequence similarity are reported and their functions are candidate functions for the new protein.

When a new protein is sequenced, ...

It usually goes through the following processes:

1. The new protein is double checked if it **has been sequenced** and stored in some existing databanks; If so, its functions are reported.
2. Otherwise, it goes through a process called '**homolog search**': other proteins sharing high sequence similarity are reported and their functions are candidate functions for the new protein.
3. If failed to get homologs, **threading** is implemented to check for possible known folds for this new protein; and the functions for the folds are candidate functions.

When a new protein is sequenced, ...

It usually goes through the following processes:

1. The new protein is double checked if it **has been sequenced** and stored in some existing databanks; If so, its functions are reported.
2. Otherwise, it goes through a process called '**homolog search**': other proteins sharing high sequence similarity are reported and their functions are candidate functions for the new protein.
3. If failed to get homologs, **threading** is implemented to check for possible known folds for this new protein; and the functions for the folds are candidate functions.
4. If again failed, the new protein is really new in the sense that it might have new functions known proteins don't have.

The **structure** need to be known — currently two key techniques

- (a) X-ray crystallography
- (b) [NMR spectroscopy](#)

### Main steps:

1. Spectral data generation
  - every NMR experiment is designed such that chemical shifts for a group of interacting atoms are observed.
2. Resonance peak identification (data filtering)
3. Resonance peak grouping and adjacency determination/prediction
  - chemical shifts for atoms from a common residue are formed into a spin system.
4. Resonance peak assignment
  - assign spin systems to their host residues ...
5. Structural information extraction
6. Structure modeling

### Main steps:

1. Spectral data generation
    - every NMR experiment is designed such that chemical shifts for a group of interacting atoms are observed.
  2. Resonance peak identification (data filtering)
  3. Resonance peak grouping and adjacency determination/prediction
    - chemical shifts for atoms from a common residue are formed into a spin system.
  4. Resonance peak assignment
    - assign spin systems to their host residues ...
- 
5. Structural information extraction
  6. Structure modeling

## NMR Protein Structure Determination

Chemical shift spread sheets:

HSQC		HNCACB			CBCACONH		
10.470	114.714	8.751	128.073	61.143	9.377	128.088	55.068
9.444	121.954	8.752	128.073	55.048	9.378	128.089	34.164
9.377	128.083	8.751	128.073	34.129	8.750	128.048	61.209
9.264	117.625	8.254	102.150	39.841	8.750	128.043	33.576
9.140	111.983	8.253	102.150	55.408	8.256	102.150	55.379
9.138	119.234	8.255	102.150	45.684	8.256	102.150	39.789
9.077	118.947	8.579	125.267	61.173	8.577	125.094	61.238
9.078	120.321	8.580	125.267	52.036	8.577	125.102	34.696
9.009	110.844	8.580	125.267	42.034	8.225	125.446	38.707
8.984	117.037	8.579	125.267	34.756	8.271	124.833	59.022
8.949	124.253	8.272	124.866	59.010	8.272	124.835	30.606
8.902	112.054	8.271	124.866	54.573	7.709	124.541	52.586
8.794	108.399	8.274	124.866	30.734	7.710	124.535	19.056
8.787	112.986	8.272	124.866	17.864	8.947	124.191	59.349
8.703	113.814	8.946	124.064	59.300	8.947	124.204	35.824
8.775	114.744	8.944	124.064	52.976	8.177	123.955	63.105
...	...	...	...	...	...	...	...

## NMR Protein Structure Determination

Chemical shift spread sheets:

HSQC		HNCACB			CBCACONH		
10.470	114.714	8.751	128.073	61.143	9.377	128.088	55.068
9.444	121.954	8.752	128.073	55.048	9.378	128.089	34.164
9.377	128.083	8.751	128.073	34.129	8.750	128.048	61.209
9.264	117.625	8.254	102.150	39.841	8.750	128.043	33.576
9.140	111.983	8.253	102.150	55.408	8.256	102.150	55.379
9.138	119.234	8.255	102.150	45.684	8.256	102.150	39.789
9.077	118.947	8.579	125.267	61.173	8.577	125.094	61.238
9.078	120.321	8.580	125.267	52.036	8.577	125.102	34.696
9.009	110.844	8.580	125.267	42.034	8.225	125.446	38.707
8.984	117.037	8.579	125.267	34.756	8.271	124.833	59.022
8.949	124.253	8.272	124.866	59.010	8.272	124.835	30.606
8.902	112.054	8.271	124.866	54.573	7.709	124.541	52.586
8.794	108.399	8.274	124.866	30.734	7.710	124.535	19.056
8.787	112.986	8.272	124.866	17.864	8.947	124.191	59.349
8.703	113.814	8.946	124.064	59.300	8.947	124.204	35.824
8.775	114.744	8.944	124.064	52.976	8.177	123.955	63.105
...	...	...	...	...	...	...	...

Assuming noise peaks are all removed, ...

## Chemical shift grouping & adjacency determination

1. Assumption: the CS value for an atom is fixed across all spectra ...

However, the observed value is usually different from one spectrum to another

- observation error
- different experimental conditions

2. Solution:

Assuming chemical shift values for every atom follow normal distribution, equivalently, **assuming observation error follows normal distribution.**

Assuming noise peaks are all removed, ...

## Chemical shift grouping & adjacency determination

1. Assumption: the CS value for an atom is fixed across all spectra ...

However, the observed value is usually different from one spectrum to another

- observation error
- different experimental conditions

2. Solution:

Assuming chemical shift values for every atom follow normal distribution, equivalently, **assuming observation error follows normal distribution.**

3. Procedure:

- (a) for every (H, N) pair in HSQC, look for 2 triples (H, N,  $C_{\alpha}^i$ ) and (H, N,  $C_{\beta}^i$ ) in HNCACB
- (b) for every (H, N) pair in HSQC, look for 4 triples (H, N,  $C_{\alpha}^i$ ) (H, N,  $C_{\alpha}^{i-1}$ ), (H, N,  $C_{\beta}^i$ ), and (H, N,  $C_{\beta}^{i-1}$ ) in CBCACONH
- (c) some double-checking can be done in the above two stages
- (d) some exceptions, for example Glycine doesn't have a  $C_{\beta}$
- (e) grouping and adjacency done simultaneously

The knowledge used in the assignment:

- Ask: what extent of assignment accuracy should be in the structure calculation?  
— nearly complete
- Signature information of spin systems  
— every type of residue has its distinct spin systems, theoretically ...  
for example, Glycine has the lowest  $C_\alpha$  about 43~47, while all other residues have  $C_\alpha$  greater than 52; ...
- Using signature information alone, we have — the Maximum weight bipartite graph matching

The knowledge used in the assignment:

- Ask: what extent of assignment accuracy should be in the structure calculation?  
— nearly complete
- Signature information of spin systems  
— every type of residue has its distinct spin systems, theoretically ...  
for example, Glycine has the lowest  $C_\alpha$  about 43~47, while all other residues have  $C_\alpha$  greater than 52; ...
- Using signature information alone, we have — the Maximum weight bipartite graph matching
- Taking in adjacency information — overcoming the shortages:
  1. not-discerning-enough signature information
  2. multiple copies of a type of residues
- The Constrained Bipartite Graph Matching

## Constrained Bipartite Graph Matching (CBM)

The problem description:

- Input:
  1. One side contains the amino acids in **linear order** as they appear in the protein sequence
  2. The other side contains **chains** of spin systems representing adjacency
  3. They have equal sizes — the length of the protein sequence
  4. Edge weights are the negative log probabilities of a spin system generated from a residue
- Directly, we search for a minimum-weight perfect matching;  
We can also use log probabilities and shift to all positives, and then search for a maximum-weight perfect matching;  
  
Relax the “perfect” in the maximization goal to allow noise; etc.
- All three problems are considered.

## Constrained Bipartite Graph Matching (CBM)

The complexities and the approximabilities:

- **Complexities:**
  1. Min-CBM is NP-hard;
  2. Max-perfect-CBM is NP-hard;
  3. Max-CBM is NP-hard and MAX SNP-hard.

## Constrained Bipartite Graph Matching (CBM)

The complexities and the approximabilities:

- **Complexities:**

1. Min-CBM is NP-hard;
2. Max-perfect-CBM is NP-hard;
3. Max-CBM is NP-hard and MAX SNP-hard.

- **Approximabilities:**

1. Min-CBM is not approximable within any constant;
2. Max-perfect-CBM is open;
3. Max-CBM is 2-approximable;  
Max-2CBM is  $\frac{13}{7}$ -approximable;  
Max-2CBM in the unweighted case is  $\frac{3}{2}$ -approximable.

## Solving Real NMR Resonance Assignment Problem

Approximation is not good enough. (Near) Optimal solutions are sought . . . :

- What is done manually in NMR labs?  
Heuristic assignment process — greedy + backtracking
- Solving near optimally
  1. Heuristics such as 2-layer algorithm
  2. Approximation algorithms such as 2-approximation and  $3 \log D$ -approximation
  3. Other heuristics using approximations as intermediate steps
- Solving optimally
  - branch-and-bound
  - search: IDA\*

Description:

- **Big idea:** taking advantage of the discerning power of the score scheme
- First layer — greedy filtering:
  1. sort the spin system chains in non-increasing length order
  2. iteratively,  
for the longest chain, bind it to every one of the  $k$  combinations, to find  $k$  best non-conflicting mapping positions  
keep only  $k$  best combinations (among  $k^2$  of them)
- Second layer — Maximum-Weight Matching:
  1. for every resultant combination, solve the remaining Maximum-Weight Matching problem
  2. report the one having maximum weight as the solution

Description:

- Def: Innermost edge
- Def: Leading innermost edge

## Description:

- Def: Innermost edge
- Def: Leading innermost edge
- Recursive algorithm — Local Ratio Technique:
  1. If  $E$  is empty, output  $\emptyset$  and halt;
  2. Otherwise,
    - (a) find a leading innermost edge  $e \in E$
    - (b)  $\Gamma = \{\text{edges conflicting } e\}$
    - (c)  $e_0 \in \Gamma$  has the minimum weight
    - (d) subtract weight  $w(e_0)$  from every edge in  $\Gamma$
    - (e)  $F = \{\text{edges in } \Gamma \text{ having weight } 0\}$
    - (f)  $E' = E - F$
    - (g) recursively call on  $E'$  to output  $M'$
    - (h) find a maximal  $M \subseteq F$ , s.t.  $M \cup M'$  is feasible
    - (i) output  $M \cup M'$

## Description:

- **Big idea** — taking advantage of the discerning power of score scheme and approximations
- Suppose we have a feasible matching  $M$  with weight  $B$ 
  - A partial assignment  $M_0$  with weight  $B_0$
  - Call 2-approx on the remaining graph to compute a matching  $M_1$  with weight  $B_1$
  - Then,
    - If  $B_0 + 2B_1 \leq B$ ,  $M_0$  is NOT a good partial assignment, and is cut off;
    - If  $B_0 + B_1 > B$ ,  $M_0 + M_1$  is a better matching, and replaces  $M$ .
- Implementation:
  - At every node, bipartition the (longest) region that a chain can map to — two child nodes
  - Leaf nodes — complete assignments for chains, call Maximum Weight Matching to complete
  - First feasible matching

## Description:

- For Minimum-Weight Perfect Matching
- [Idea \(A\\*-like search algorithm\)](#) — at every partial assignment of weight  $f$ , estimate a lower bound  $g$  to the goal  
Then, choose a node with minimum  $f + g$  to expand
- Guarantee — output an optimal matching (in fact, all optimal matchings)
- Details:
  1. How to get lower bounds? — evaluation functions
    - (a) Min-Weight — minimum edges out of chains and singletons
    - (b) UBM — discard the adjacency
    - (c) Combinations of the above two functions
  2. Open list and closed list
  3. Why IDA\*?

## Experimental results:

- Assignment accuracy comparison

	length	$R_{\text{new}}^*$	$R_{\text{old}}^*$	$R_{\text{old}}^{\text{GF-GK}}$	$R_{\text{old}}^{\text{GF-BnB}}$	$R_{\text{old}}^{\text{GF}-\frac{13}{7}}$	$R_{\text{new}}^{\text{GF-GK}}$
bmr4027_5	158	-	-	38	33	40	49
bmr4027_6		-	-	59	37	64	68
bmr4027_7		128	106	88	74	89	116
bmr4027_8		151	146	149	128	151	156
bmr4027_9		158	156	156	156	156	156
bmr4144_5	78	55	36	15	16	11	28
bmr4144_6		63	45	19	11	21	37
bmr4144_7		47	51	55	64	68	51
bmr4144_8		59	74	61	67	69	65
bmr4144_9		78	75	75	75	75	78
bmr4288_5	105	46	39	33	12	36	40
bmr4288_6		66	45	42	26	49	47
bmr4288_7		83	77	61	57	65	82
bmr4288_8		105	97	66	66	68	105
bmr4288_9		105	105	105	105	105	105
bmr4302_5	115	80	31	31	16	31	38
bmr4302_6		87	58	58	43	51	78
bmr4302_7		104	86	81	62	79	80
bmr4302_8		112	112	112	103	112	112
bmr4302_9		115	115	110	110	111	115
bmr4316_5	89	82	65	48	30	43	60
bmr4316_6		83	83	65	35	59	72
bmr4316_7		87	89	79	79	75	85
bmr4316_8		89	89	89	89	75	89
bmr4316_9		89	89	89	89	89	89
bmr4318_5	215	-	-	23	20	19	82
bmr4318_6		-	-	35	35	34	82
bmr4318_7		156	80	72	52	73	131
bmr4318_8		189	167	91	62	92	156
bmr4318_9		211	209	201	201	201	211

## Experimental results:

- Adjacency determination + assignment accuracy

	length	simulated adjacency	$R^*$
bmr4027	158	88%	158
bmr4144	78	77%	75
bmr4288	105	75%	102
bmr4302	115	72%	113
bmr4309	178	56%	93
bmr4316	89	73%	89
bmr4318	215	83%	209
bmr4353	126	75%	119
bmr4391	66	85%	64
bmr4393	156	67%	122
bmr4579	86	76%	81
bmr4670	120	69%	106
bmr4752	68	31%	58
bmr4929	114	66%	98
		71%	

What next?

Things need to be done better:

- Improving adjacency information extraction (and peak grouping)
- Improving score schemes

What next?

Things need to be done better:

- Improving adjacency information extraction (and peak grouping)
- Improving score schemes
- Improving signature information and adjacency information usage

What next?

Things need to be done better:

- Improving adjacency information extraction (and peak grouping)
- Improving score schemes
- Improving signature information and adjacency information usage
- Assignment evaluation models — maximum likelihood?

So far we are using the score of the matching and the assignment accuracy, better model?

What next?

Things need to be done better:

- Improving adjacency information extraction (and peak grouping)
- Improving score schemes
- Improving signature information and adjacency information usage
- Assignment evaluation models — maximum likelihood?  
So far we are using the score of the matching and the assignment accuracy, better model?
- Better approximation algorithms

What next?

Things need to be done better:

- Improving adjacency information extraction (and peak grouping)
- Improving score schemes
- Improving signature information and adjacency information usage
- Assignment evaluation models — maximum likelihood?  
So far we are using the score of the matching and the assignment accuracy, better model?
- Better approximation algorithms
- Faster exact algorithms
  - better evaluation functions in IDA\* search
  - taking more advantage of signature information & adjacency

NMR Protein Resonance Assignment Problem

Thanks you, and *Questions* ???