

# Accelerating Biomolecular Nuclear Magnetic Resonance Assignment with A\*

Joel Venzke, Paxten Johnson, Rachel Davis, John Emmons,  
Katherine Roth, David Mascharka, Leah Robison,  
Timothy Urness and Adina Kilpatrick

Department of Mathematics and Computer Science  
Drake University

*joel.venzke@drake.edu*

April 10, 2014

# Overview

- 1 Introduction
  - Motivation
  - Nuclear Magnetic Resonance Spectroscopy
- 2 NMR Assignment Overview
  - Data Collection and Manual Assignment
- 3 Automation Algorithm
  - Preprocessing
  - Assignment
  - Goal State
- 4 Conclusion
  - Results
  - Outlook



# Motivation

- Nuclear Magnetic Resonance Spectroscopy
  - Gain knowledge about protein structure
  - Study how mutations lead to diseases
- Problems
  - Generates large amounts of data
  - Data analysis is slow and error prone
- Goal
  - Automate the assignment process
  - Decrease human error
  - Increase productivity

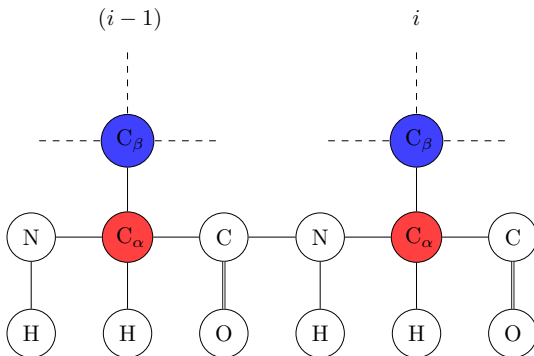


# Nuclear Magnetic Resonance (NMR)

- Used to obtain structural information
  - Chemical shift values
- HNCACB experiment
  - Generates  $C_\alpha$  and  $C_\beta$  residue  $i$  and  $i - 1$
- CBCA(CO) NH experiment
  - Generates  $C_\alpha$  and  $C_\beta$  for residue  $i$
  - Confirms residue data

# Chemical Shift Values

## HNCACB



# Timeline

Protein  
Production  
at least 5 days

Data Assignment  
20 days to 9 months

NMR  
Experiments  
1-2 days

[1]

# Manual Methods

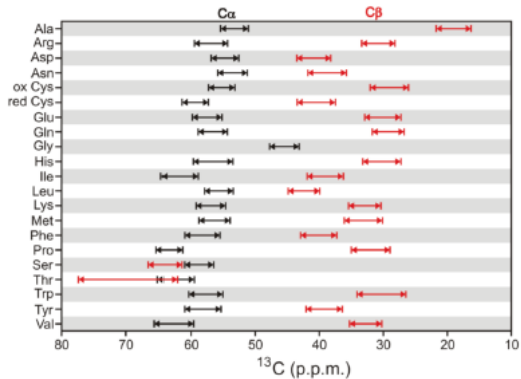
- Most time consuming part
- Prone to human error
- Missing and ambiguous data forces chunks to be skipped

# Initialization

- Input
  - Expected amino acid sequence
    - Covered to expectation chemical shift values
    - Stored as the protein chain
  - NMR chemical shift data
    - $C_{\alpha}$  and  $C_{\beta}$  for residue  $i$  and  $i - 1$
    - Stored in a tile
- Missing data
  - Place holder tile generation
- Grouping



# Grouping



[2]

# Starting the assignment

**Tiles to assign:**

13  
9

11  
13

9  
12

Protein Chain	Tiles		
11.5	13 9	11 13	9 12
12.5			
9.6			

# Cost Calculation

- Accuracy matching the protein chain residue
- Accuracy matching the tile above current tile
- Cost of all tiles place before current tile

# Generating child nodes

**Tiles to assign:**

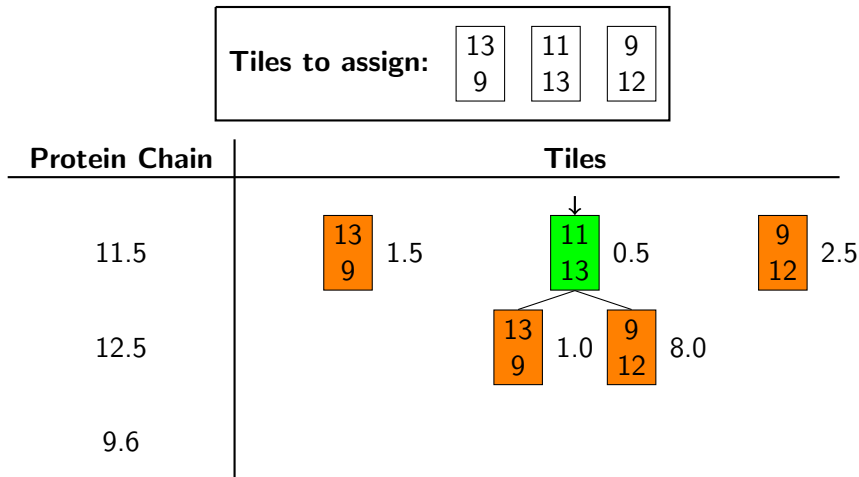
13  
9

11  
13

9  
12

Protein Chain	Tiles		
11.5	<div>13 9</div> 1.5	<div>11 13</div> 0.5	<div>9 12</div> 2.5
12.5			
9.6			

# Generating child nodes



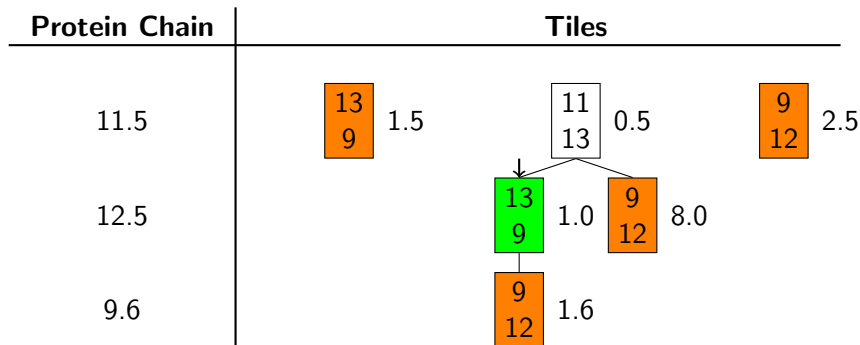
# Goal State

Tiles to assign:

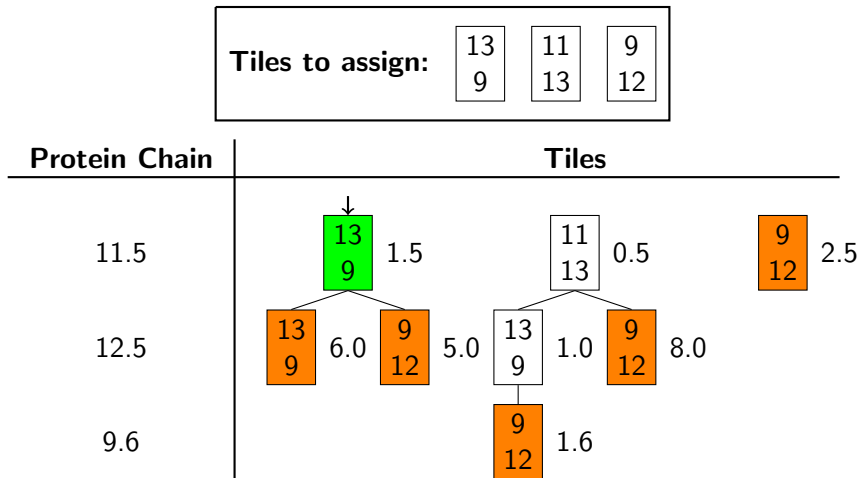
13  
9

11  
13

9  
12



# Goal State



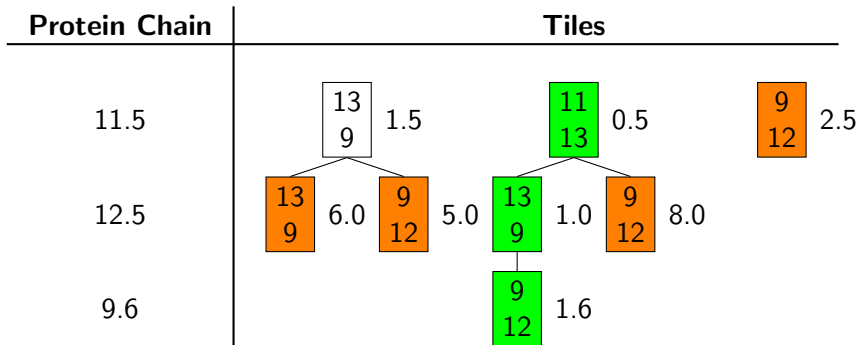
# Solution State

Tiles to assign:

13  
9

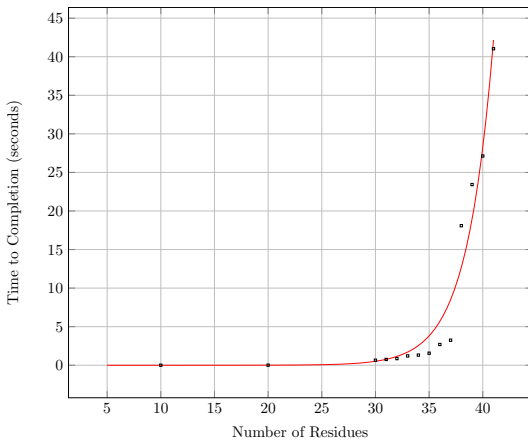
11  
13

9  
12





# Time of Assignment



# Future Goals

- Parallelization
  - Decrease assignment time
  - Allow for larger data sets
- Machine learning
  - Increase accuracy of assignment
  - Optimize cost calculation

# Acknowledgments

- Dr. Tim Urness (Mathematics and Computer Science)
- Dr. Adina Kilpatrick (Physics)
- Rachel Davis (research colleague)
- John Emmons (research colleague)
- Katherine Roth (research colleague)
- David Mascharka (research colleague)
- Leah Robison (research colleague)

# Bibliography



Babak Alipanahi, Xin Gao, Emre Karakoc, Frank Balbach, Shuai Cheng Li, Guangyu Feng, Logan Donaldson and Ming Li, *Error tolerant NMR backbone resonance assignment and automated structure generation.*, Journal of bioinformatics and computational biology, **9** (2011), 15–41.



Sean Cahill and Mark Girvin.  
*Introduction to 3d triple resonance experiments.*  
2012.

# Thank You

