# 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 Background
   - Data Collection and Manual Assignment
3. Automation Algorithm
   - Preprocessing
   - Assignment
   - Goal State
4. Conclusion
   - Results
   - Outlook

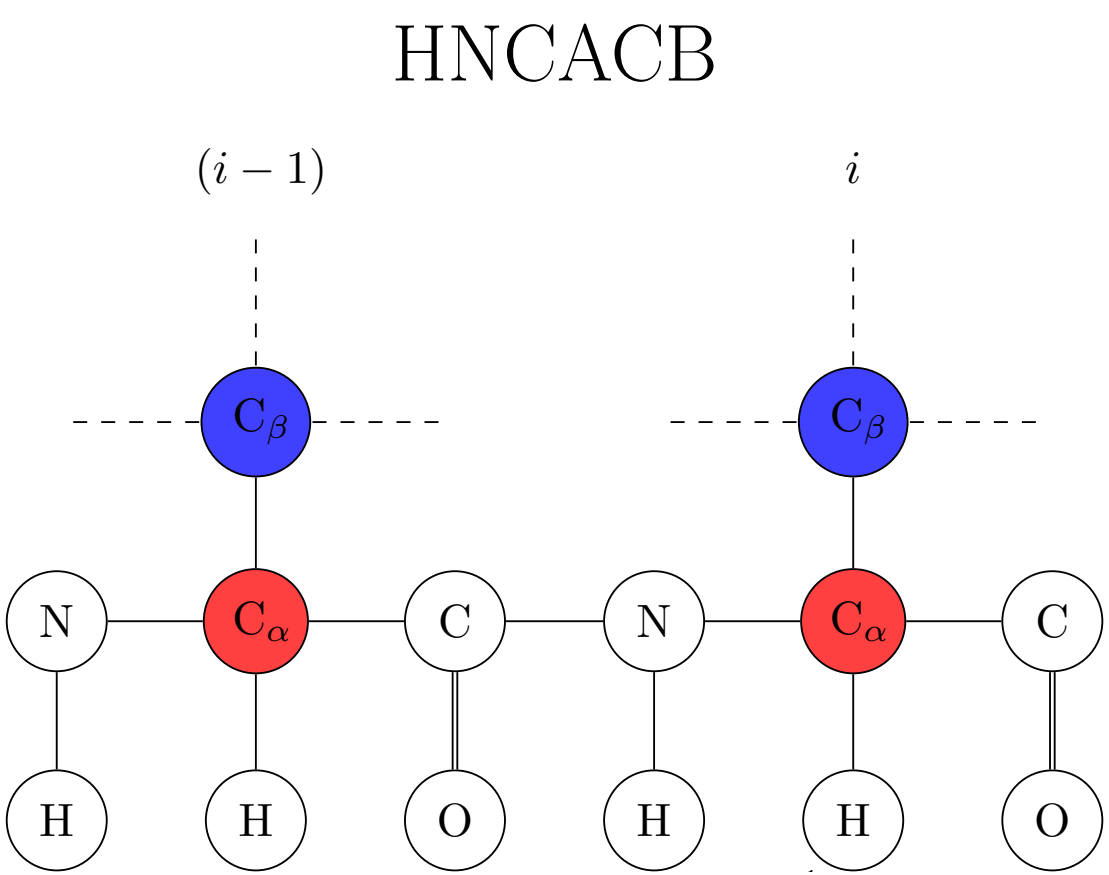
# Motivation

- 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

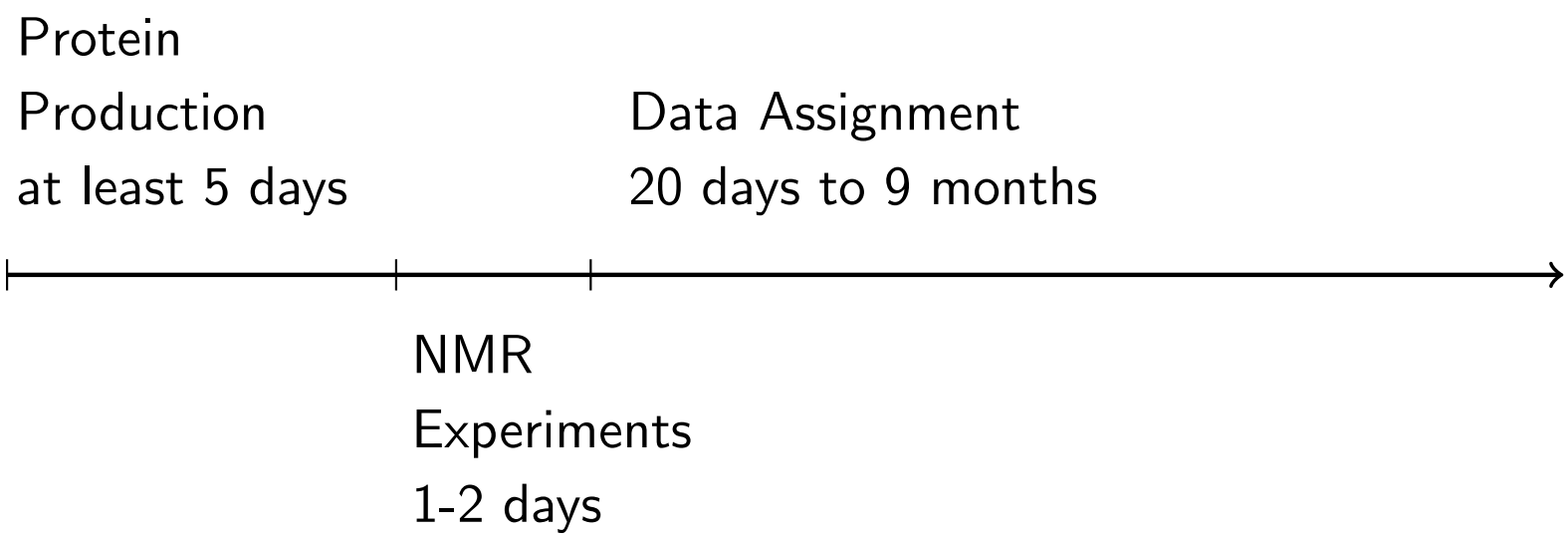
# Chemical Shift Values



# Manual Methods

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

# Timeline


Protein
Production
at least 5 days
Data Assignment
20 days to 9 months
NMR
Experiments
1-2 days

[1]

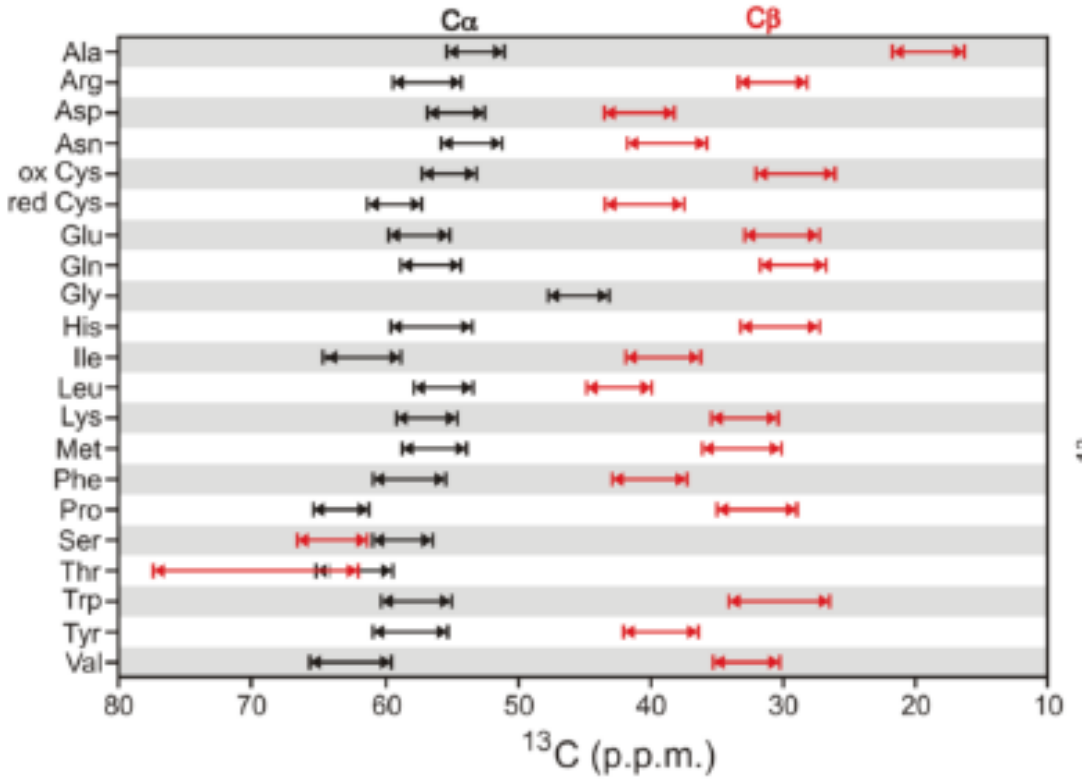
# Automating Assingment

- Initialization
- Generating child nodes
- Goal State
- Solution State

# Initialization

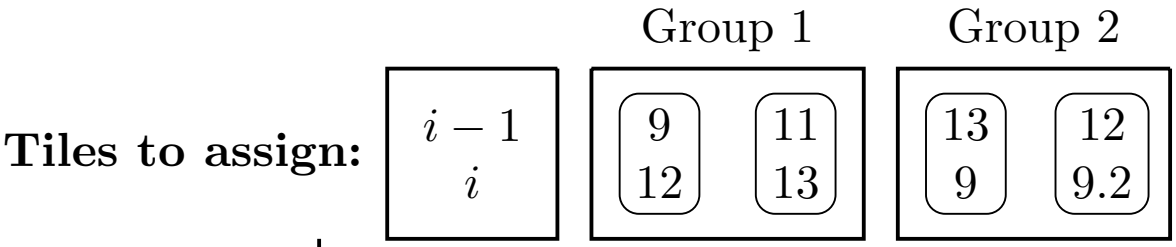
- Expected amino acid sequence
  - Converted to expected chemical shift values
  - Stored as the reference protein chain
- NMR experiment's chemical shift data
  - $C_\alpha$ and $C_\beta$ for residue $i$ and $i-1$
  - Stored in a tile
- Missing data
  - Place holder tile generation
- Grouping

# Grouping


Cα
Cβ
Ala
Arg
Asp
Asn
ox Cys
red Cys
Glu
Gln
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val
80
70
60
50
40
30
20
10
13C (p.p.m.)

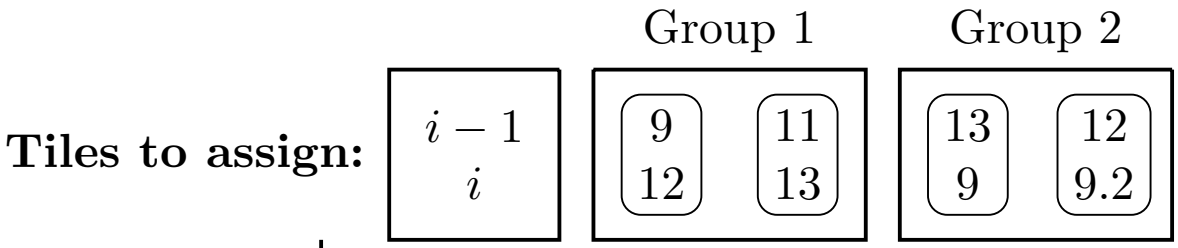
[2]

# Starting the assignment




| Reference Protein Chain | | Nodes |
|---|---|---|
| Chemical Shift | Group | |
| 13.5 | 1 | |
| 9.5 | 2 | |
| 11.4 | 1 | |
| 8.8 | 2 | |

# Starting the assignment



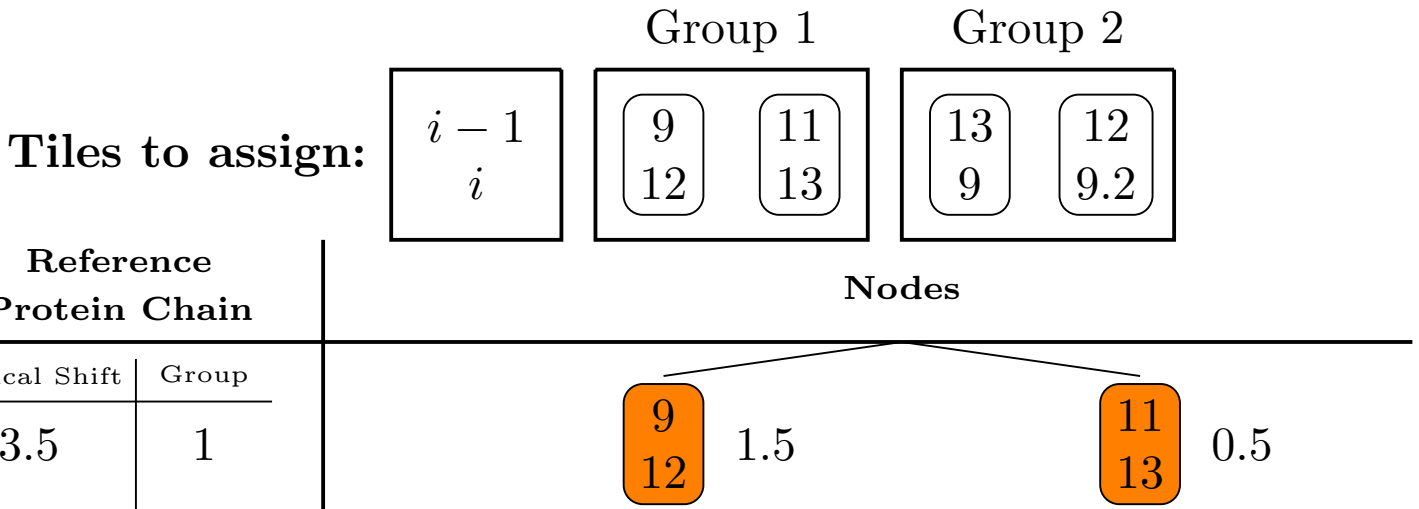

| Reference Protein Chain | |
|---|---|
| Chemical Shift | Group |
| 13.5 | 1 |
| 9.5 | 2 |
| 11.4 | 1 |
| 8.8 | 2 |

**Nodes**

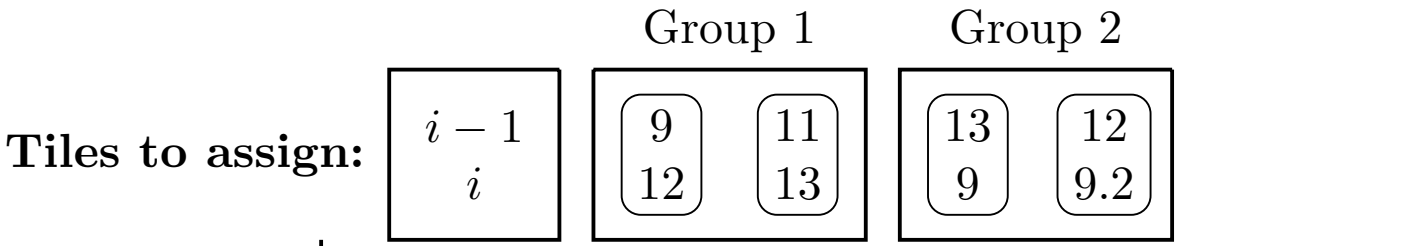
# Cost Calculation

- Accuracy matching the protein chain residue
- Accuracy matching the tile above current tile
- Cost of placing all previous tiles

# Generating child nodes

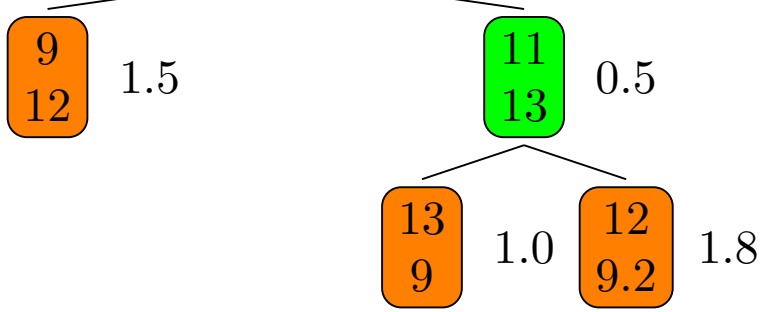


# Generating child nodes

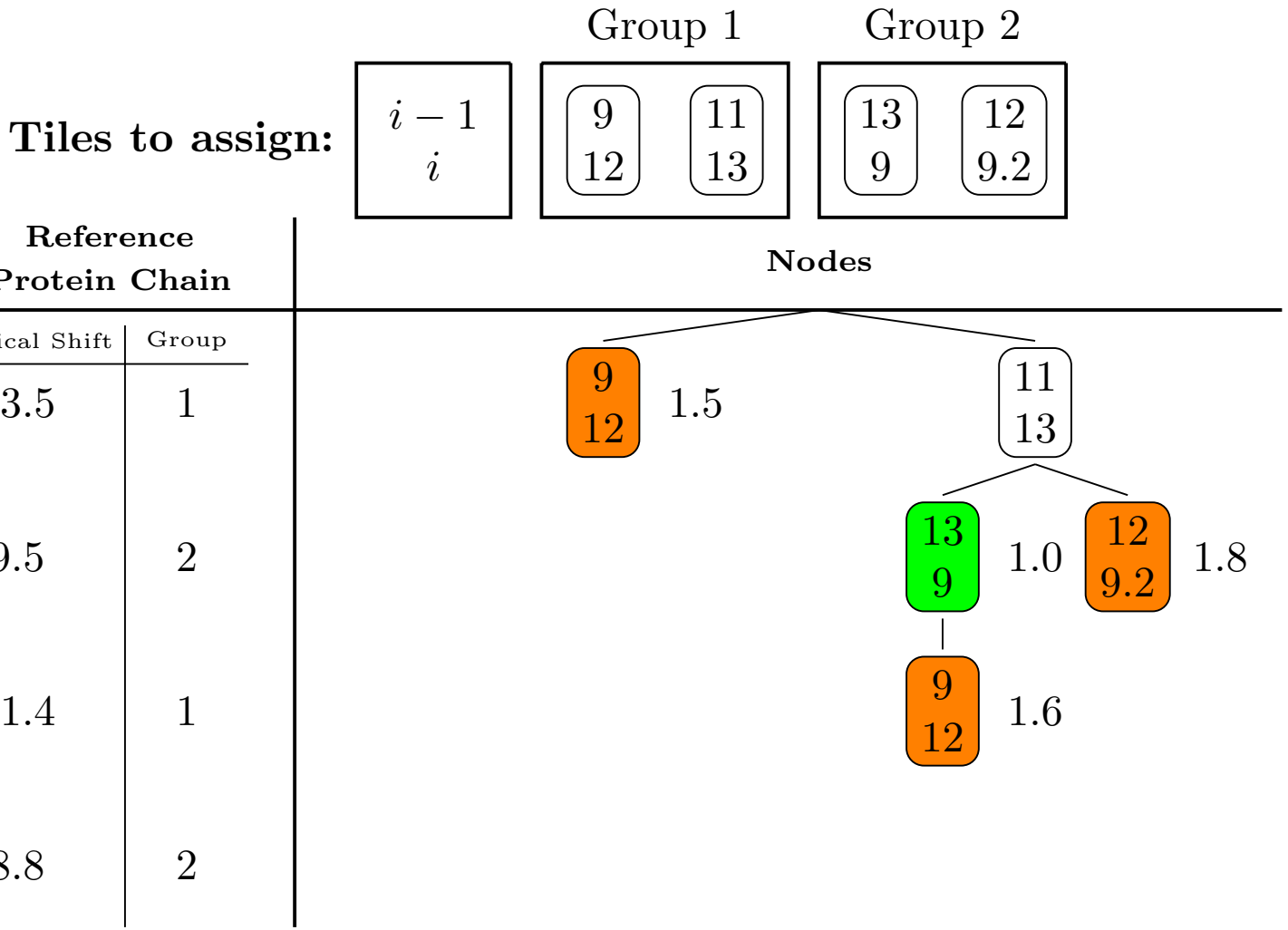



| Reference Protein Chain | |
|---|---|
| Chemical Shift | Group |
| 13.5 | 1 |
| 9.5 | 2 |
| 11.4 | 1 |
| 8.8 | 2 |

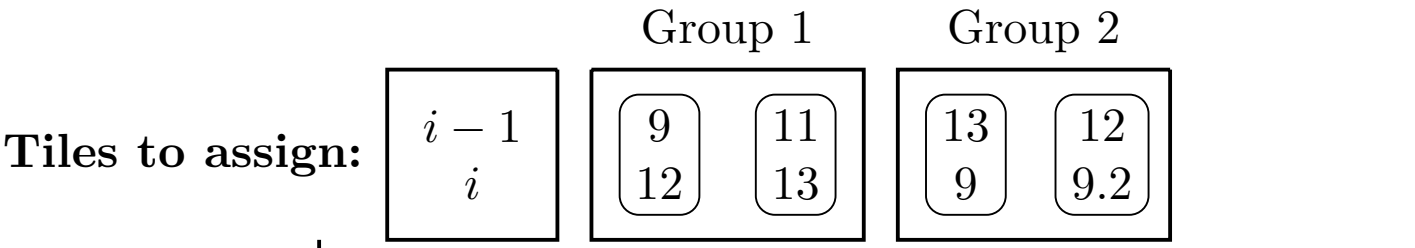
**Nodes**



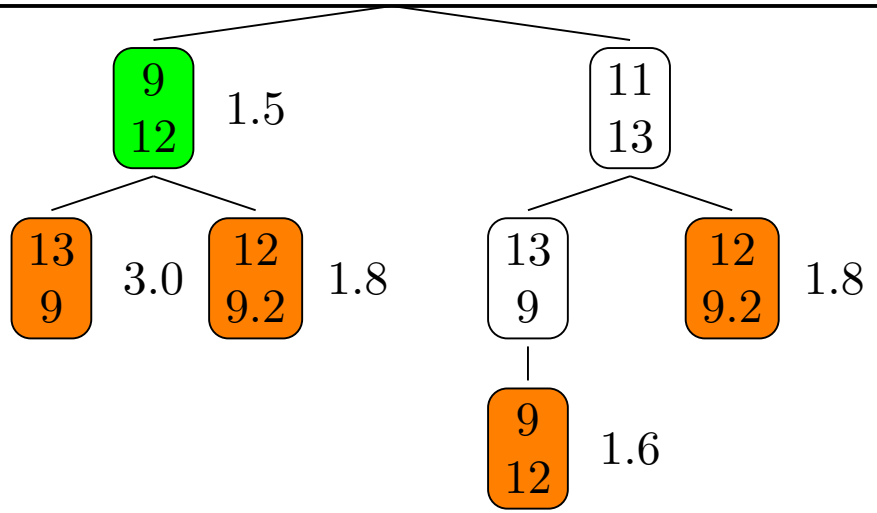
# Generating child nodes



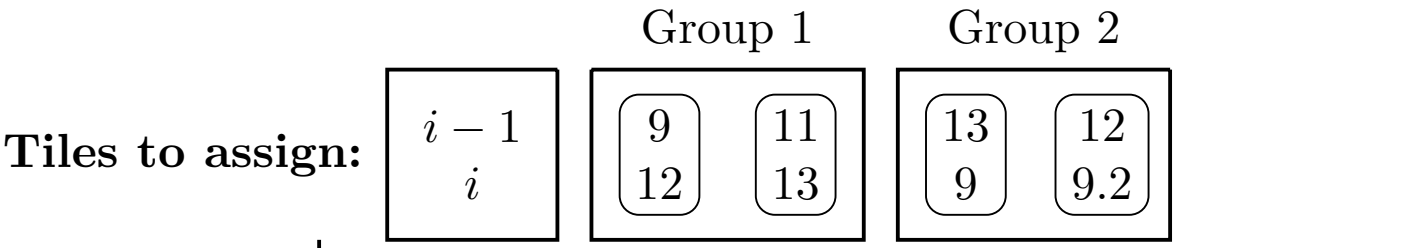
# Generating child nodes




| Reference Protein Chain | |
|---|---|
| Chemical Shift | Group |
| 13.5 | 1 |
| 9.5 | 2 |
| 11.4 | 1 |
| 8.8 | 2 |



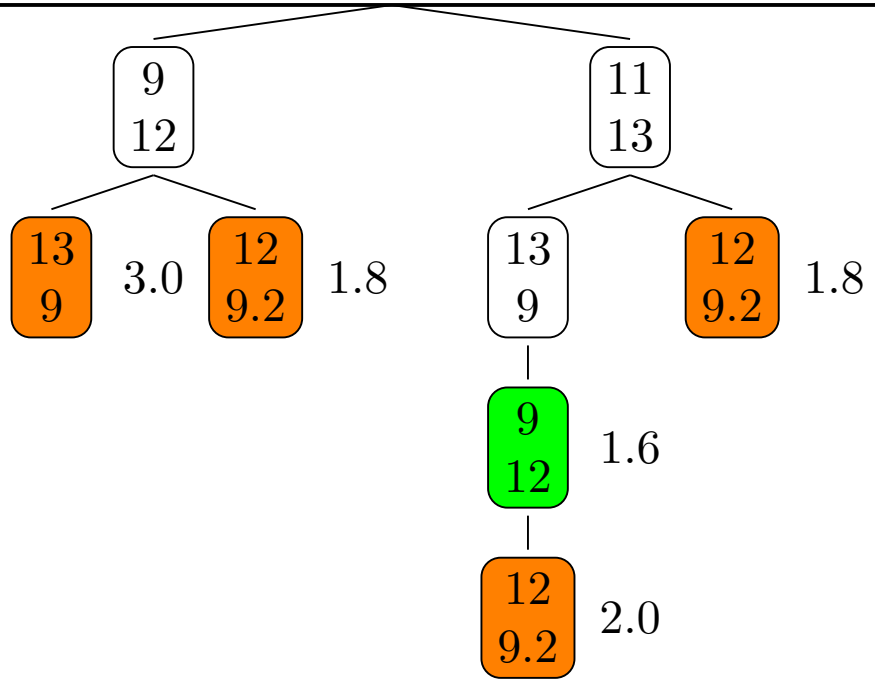
# Goal State



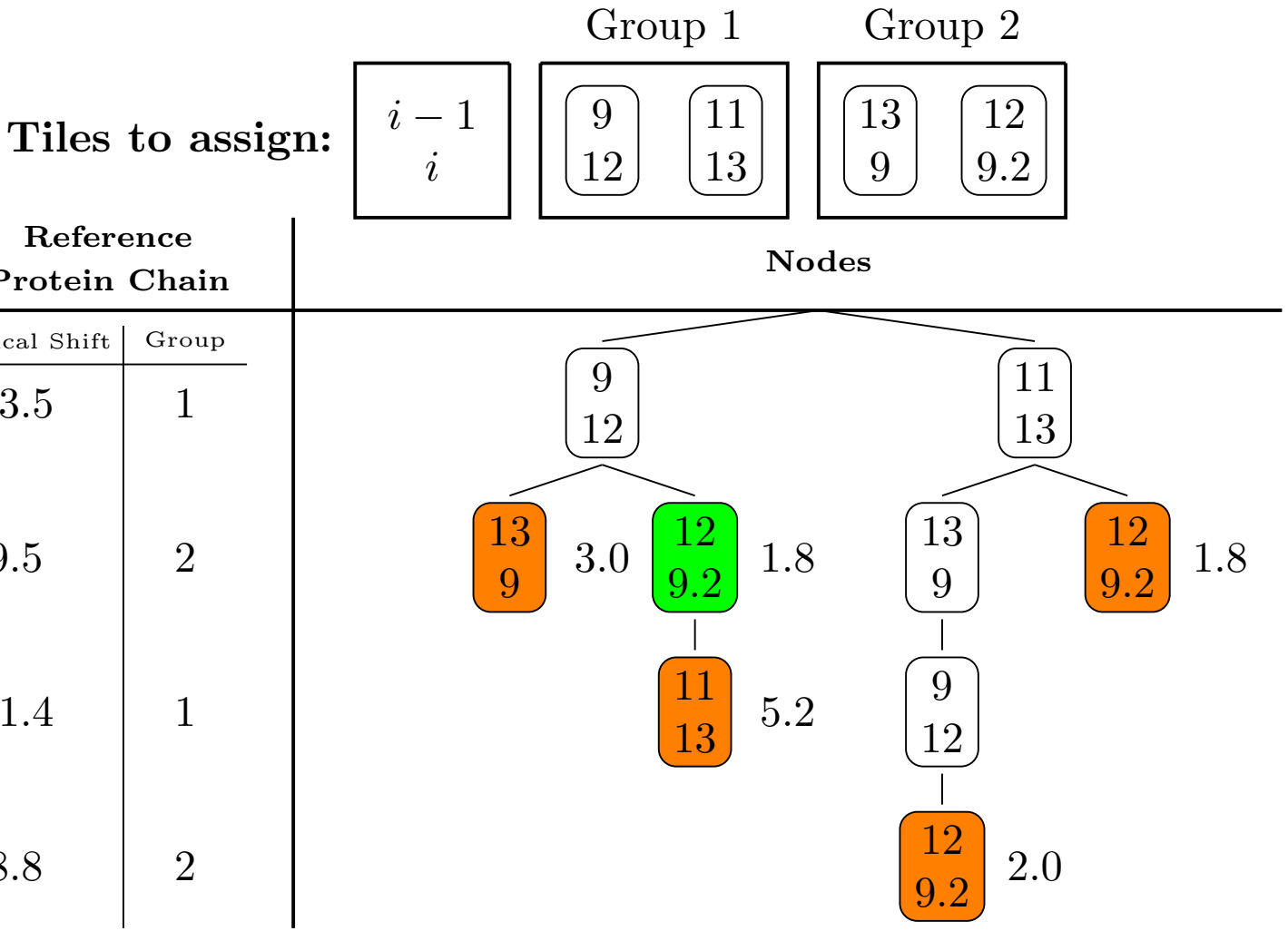

**Reference Protein Chain**

| Chemical Shift | Group |
|---|---|
| 13.5 | 1 |
| 9.5 | 2 |
| 11.4 | 1 |
| 8.8 | 2 |

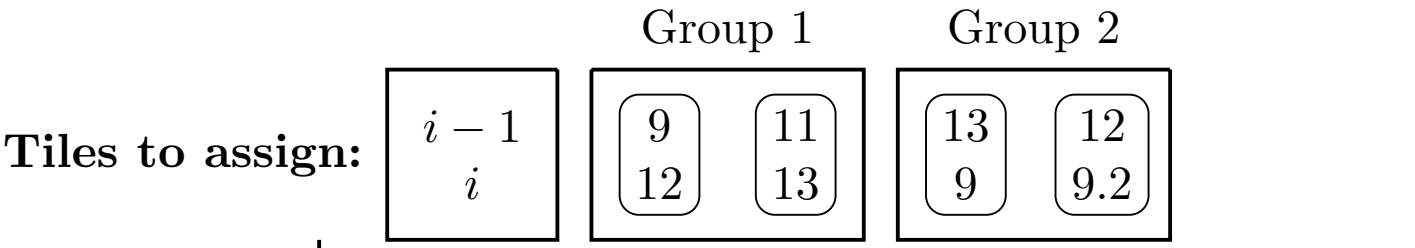
**Nodes**



# Goal State


Group 1
Group 2
Tiles to assign:
$i-1$
$i$
9
12
11
13
13
9
12
9.2
Reference Protein Chain
Chemical Shift
Group
13.5
1
9.5
2
11.4
1
8.8
2
Nodes
9
12
11
13
13
9
3.0
12
9.2
1.8
11
13
5.2
13
9
12
9.2
1.8
9
12
12
9.2
2.0

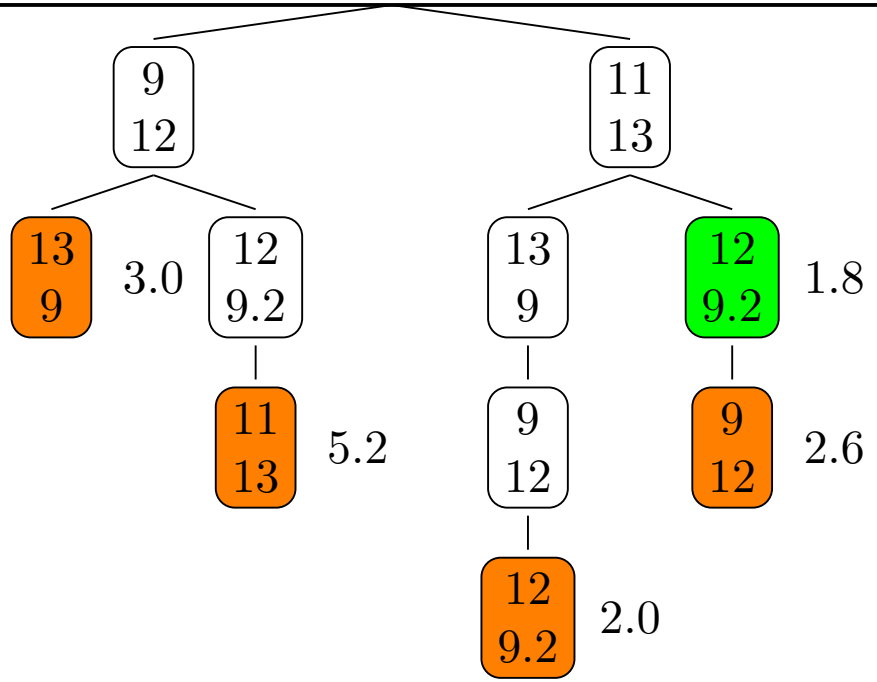
# Goal State



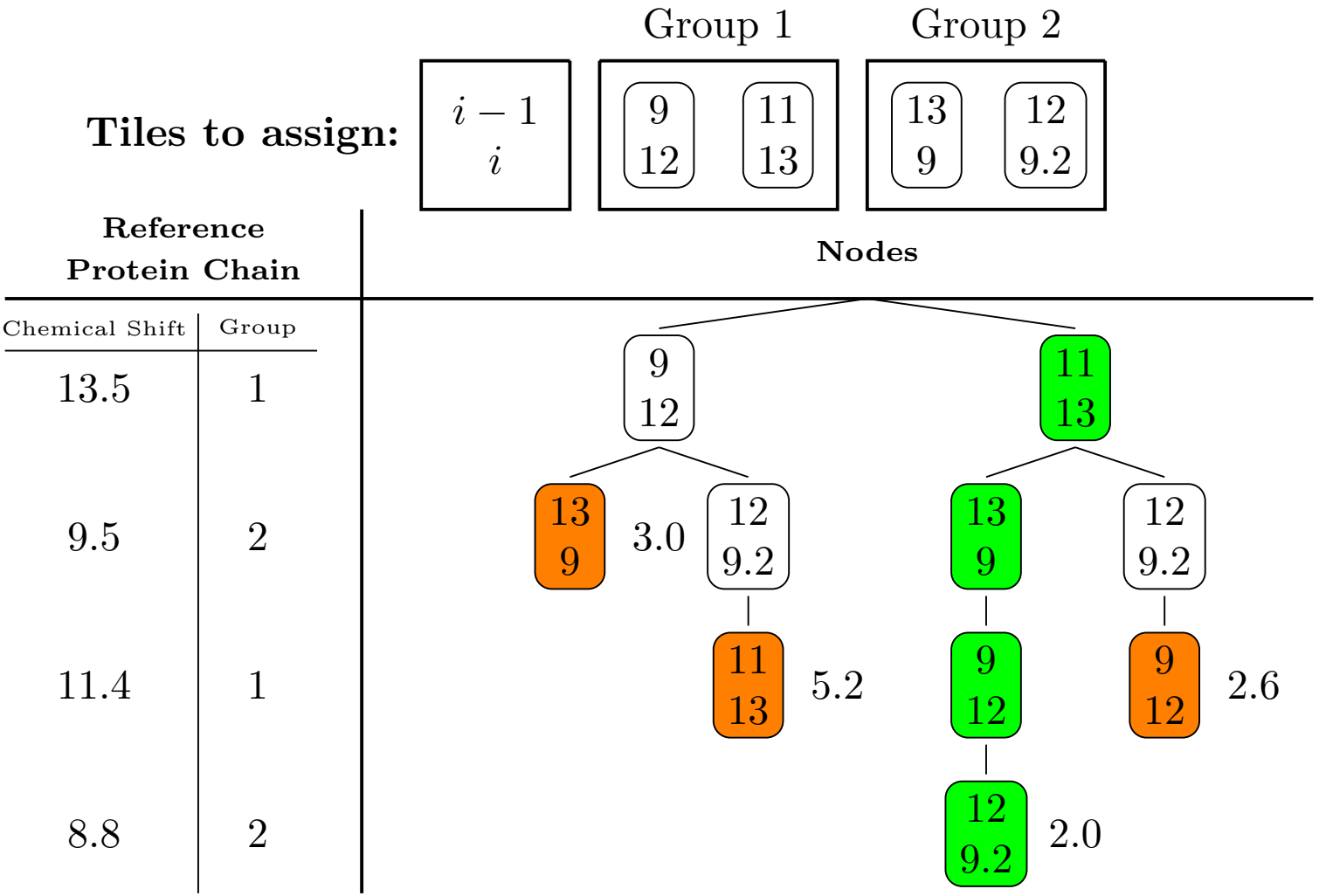

**Reference Protein Chain**

| Chemical Shift | Group |
|---|---|
| 13.5 | 1 |
| 9.5 | 2 |
| 11.4 | 1 |
| 8.8 | 2 |

**Nodes**

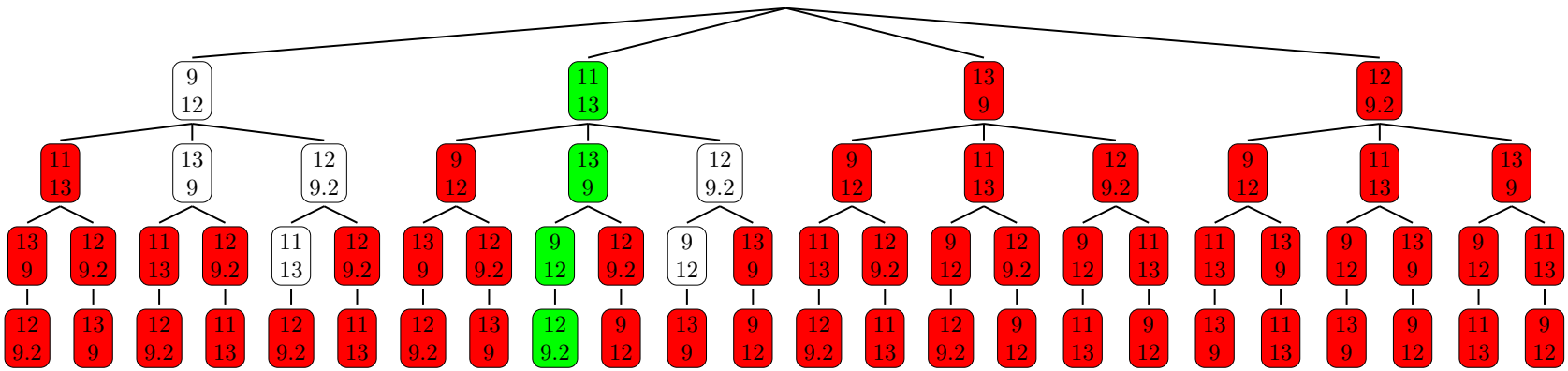


# Solution State

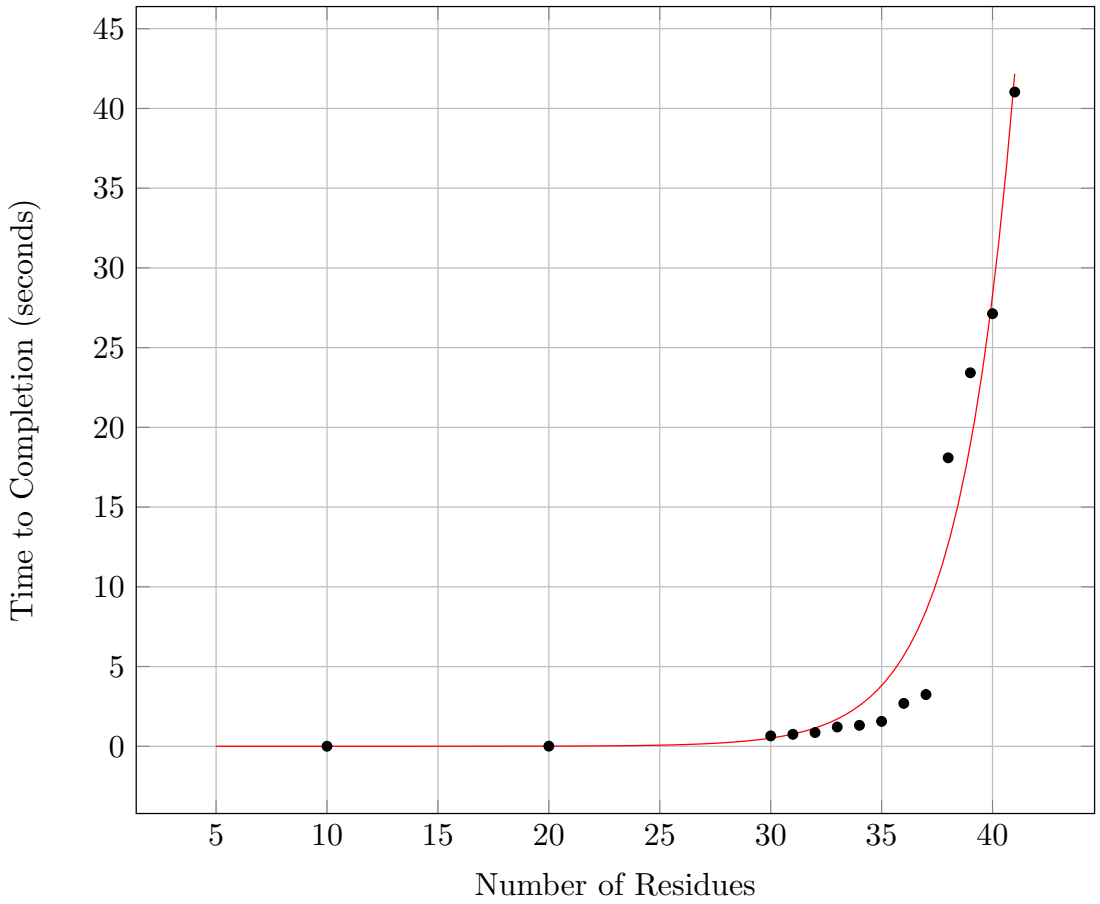


# Compared to Naive Approach

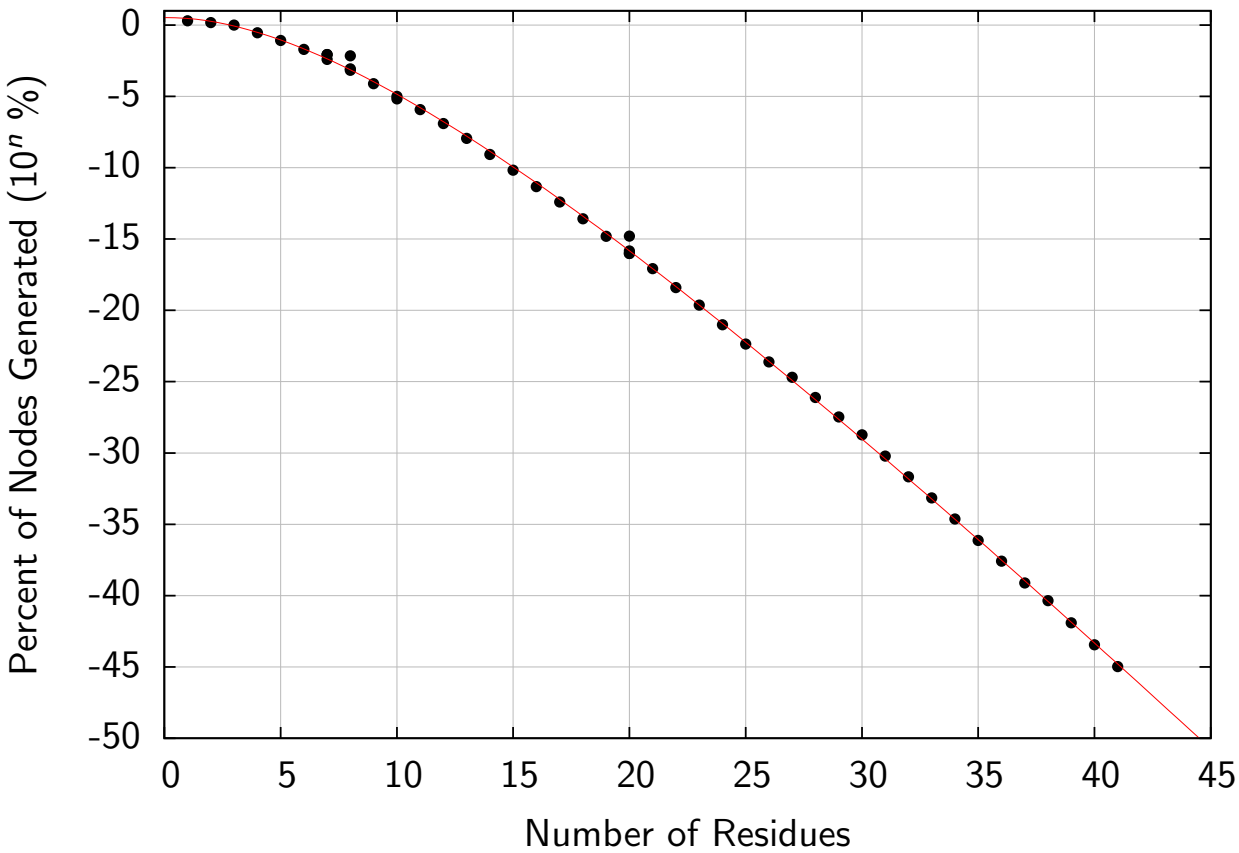
Only 14.1% of the possible combinations



# Time of Assignment



# Child Nodes Generated



# Future Goals

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

# 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

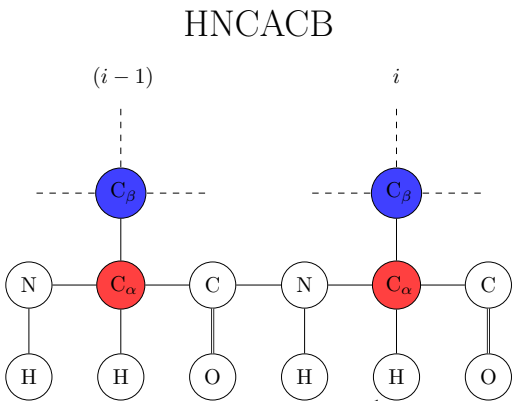
Babak Alipanahi, Xin Gao, Emre Karakoc, Frank Balbach, Shuai Cheng Li, Guangyu Feng, Logan Donaldson and Ming Li, *Error tolerant NMR backbone resonance assignment and automated structure generation.*, Journal of bioinformatics and computational biology, **9** (2011), 15–41.

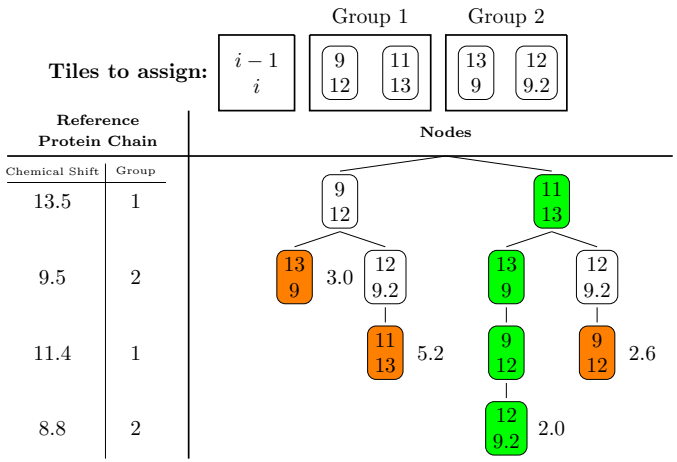
Sean Cahill and Mark Girvin.
*Introduction to 3d triple resonance experiments*.
2012.

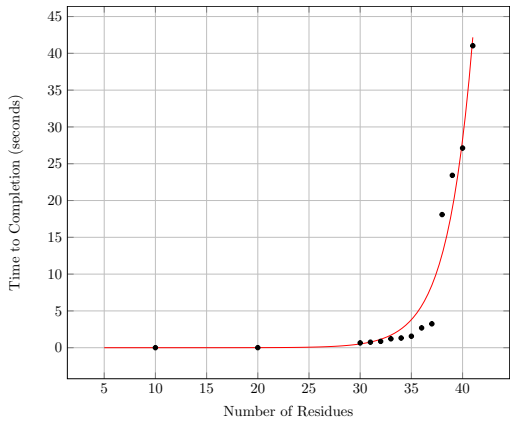
Peter Guntert, *Automated structure determination from NMR spectra*, European Biophysics Journal, **38** (2009), 129–143.

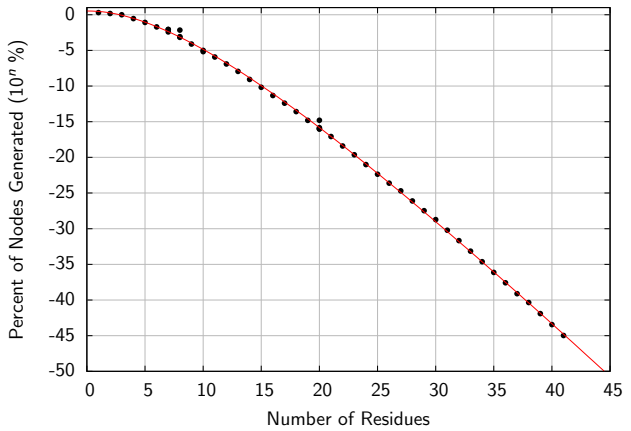
Flemming M. Poulsen.
*A brief introduction to nmr spectroscopy of proteins*.

# Thank You


HNCACB
$(i-1)$
$i$
$C_\beta$
$C_\alpha$
N
C
H
O


Tiles to assign:
Group 1
Group 2
$i-1$
$i$
Reference Protein Chain
Chemical Shift
Group
13.5 1
9.5 2
11.4 1
8.8 2
Nodes
3.0
5.2
2.6
2.0


Time to Completion (seconds)
Number of Residues


Percent of Nodes Generated ($10^n$ %)
Number of Residues