Accelerating Biomolecular Nuclear Magnetic Resonance Assignment

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

September 12, 2014

Overview

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

Applications

- Studying Mutations
  - Protein folding
  - Biological effects
- Pharmaceuticals
  - Block proteins leading to mutations
  - Prevent onset of disease

Mutation Example

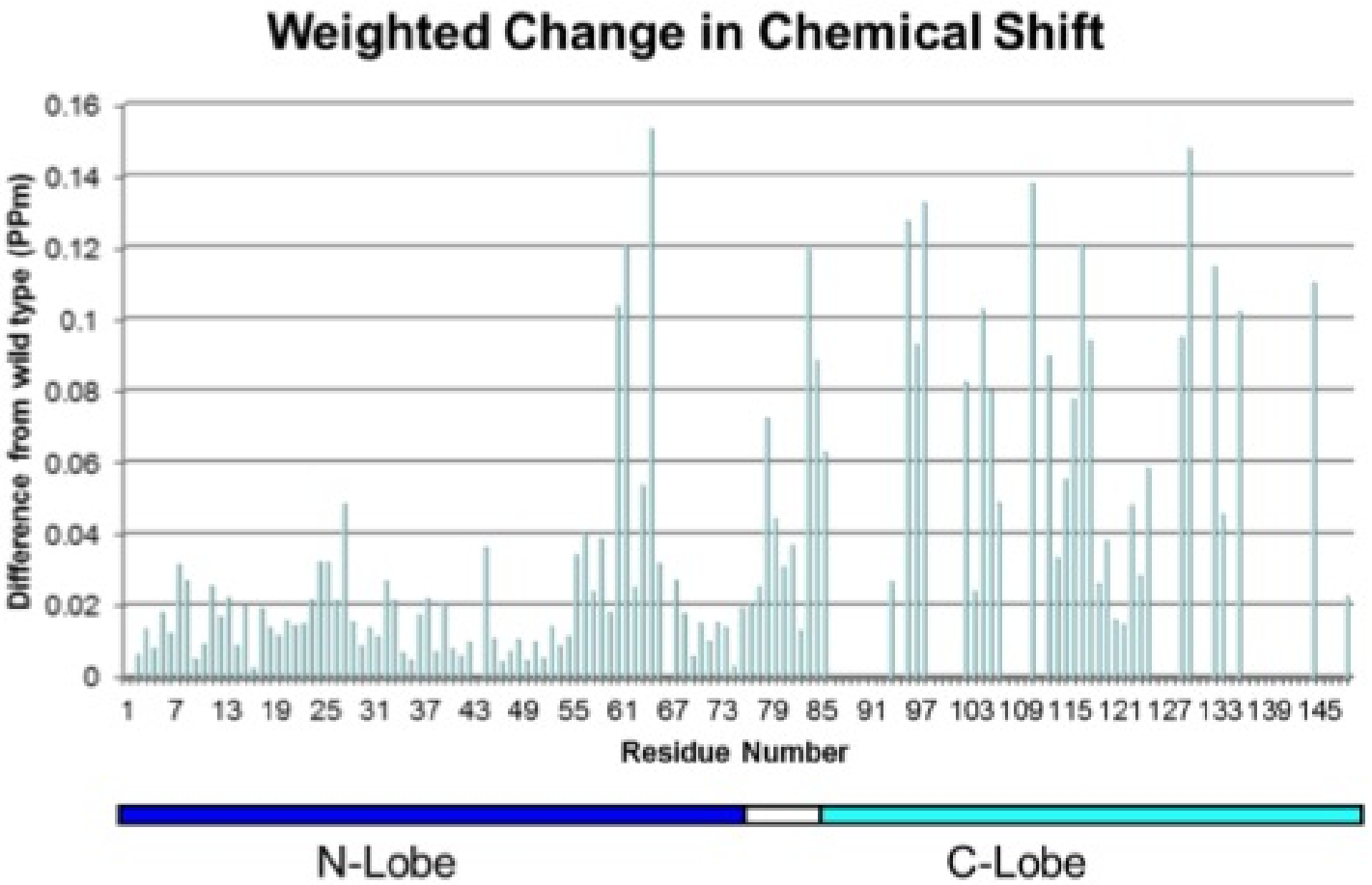
Nuclear Magnetic Resonance (NMR)

- NMR: phenomenon in which atomic nuclei absorb and re-emit EM radiation

1. Alignment of nuclear magnetic spins in constant magnetic field
2. EM radiation pulses disturb alignment

- Chemical Shift Values Generated
  - Resonant frequency depends on environment
  - Deviation from the Larmor Frequency creates chemical shift values

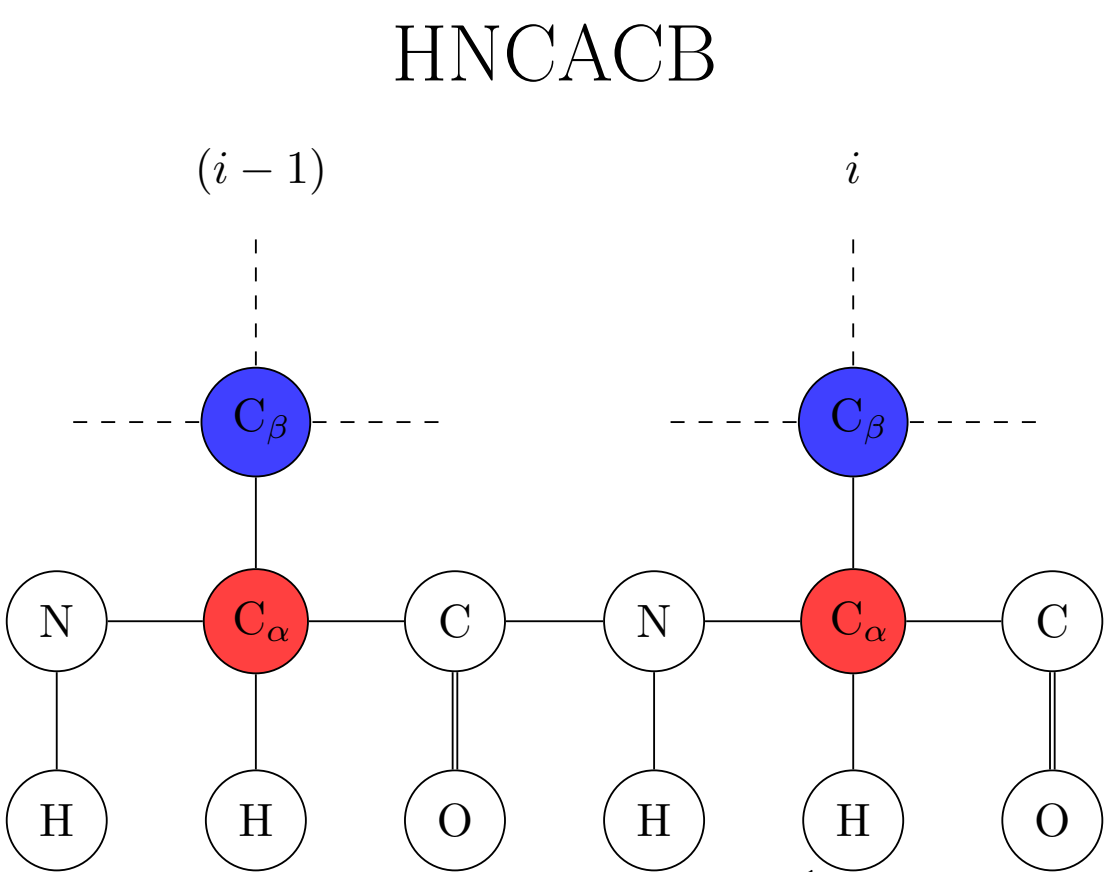
Assigned Chemical Shift Values



NMR Experiments

- HNCACB experiment
  - Generates C_α and C_β residue i and $i-1$
- CBCA(CO) NH experiment
  - Generates C_α and C_β for residue i
  - Confirms residue data

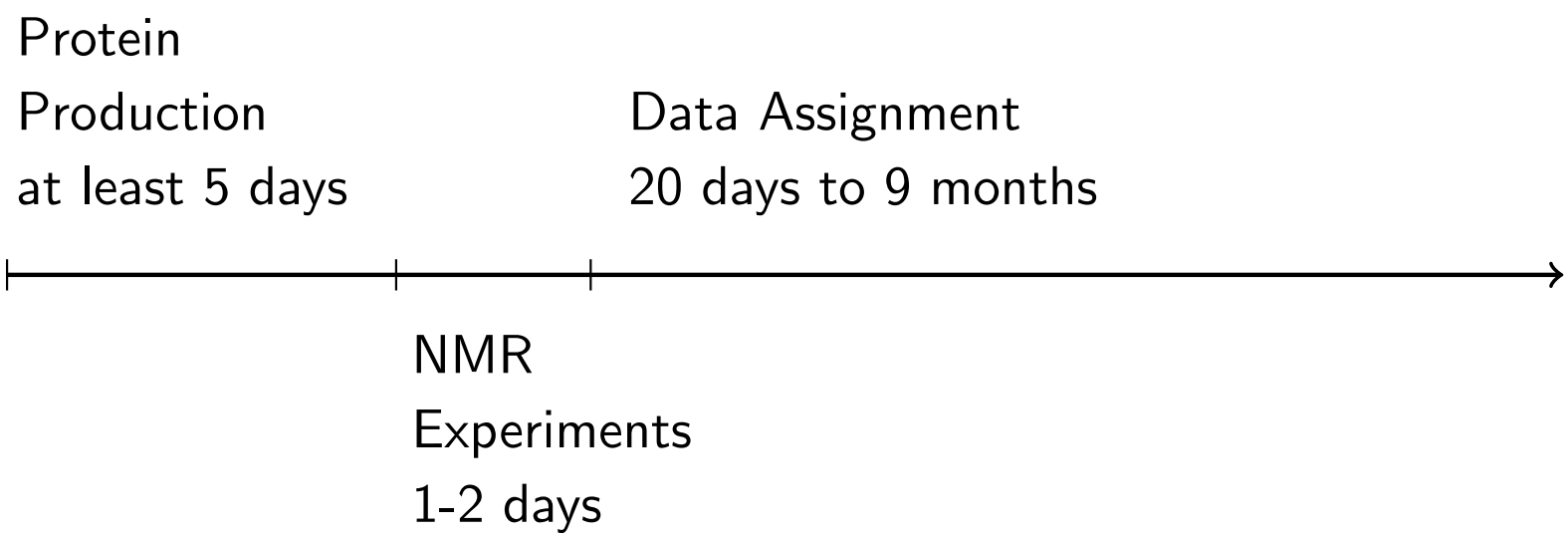
Backbone Structure



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]

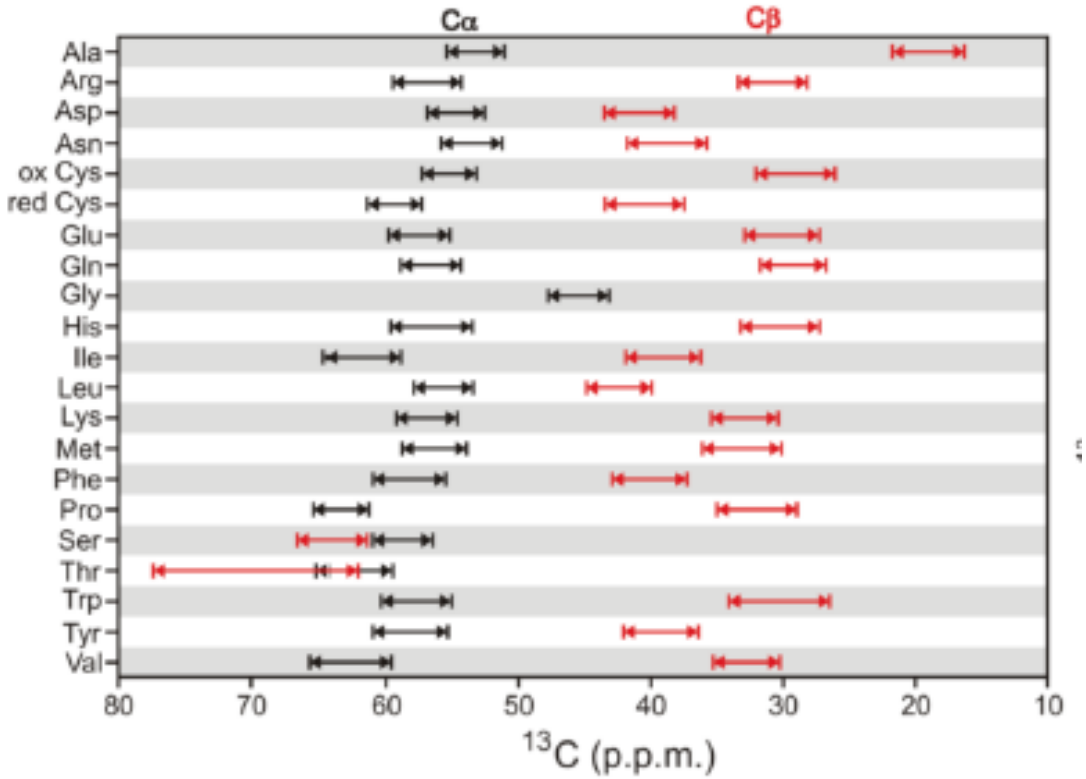
Automating Assignment

1. Initialization
2. Generating child nodes
3. Goal State
4. Solution State

Initialization

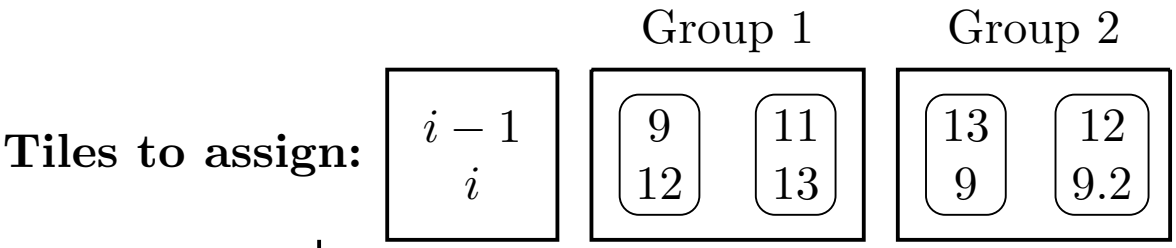
- Expected amino acid sequence
  - Converted to expected chemical shift values
  - Stored as the reference protein chain
- NMR experiment's chemical shift data
  - C_α and C_β for residue i and $i-1$
  - Stored in a tile
- Missing data
  - Place holder tile generation
- Grouping

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.)

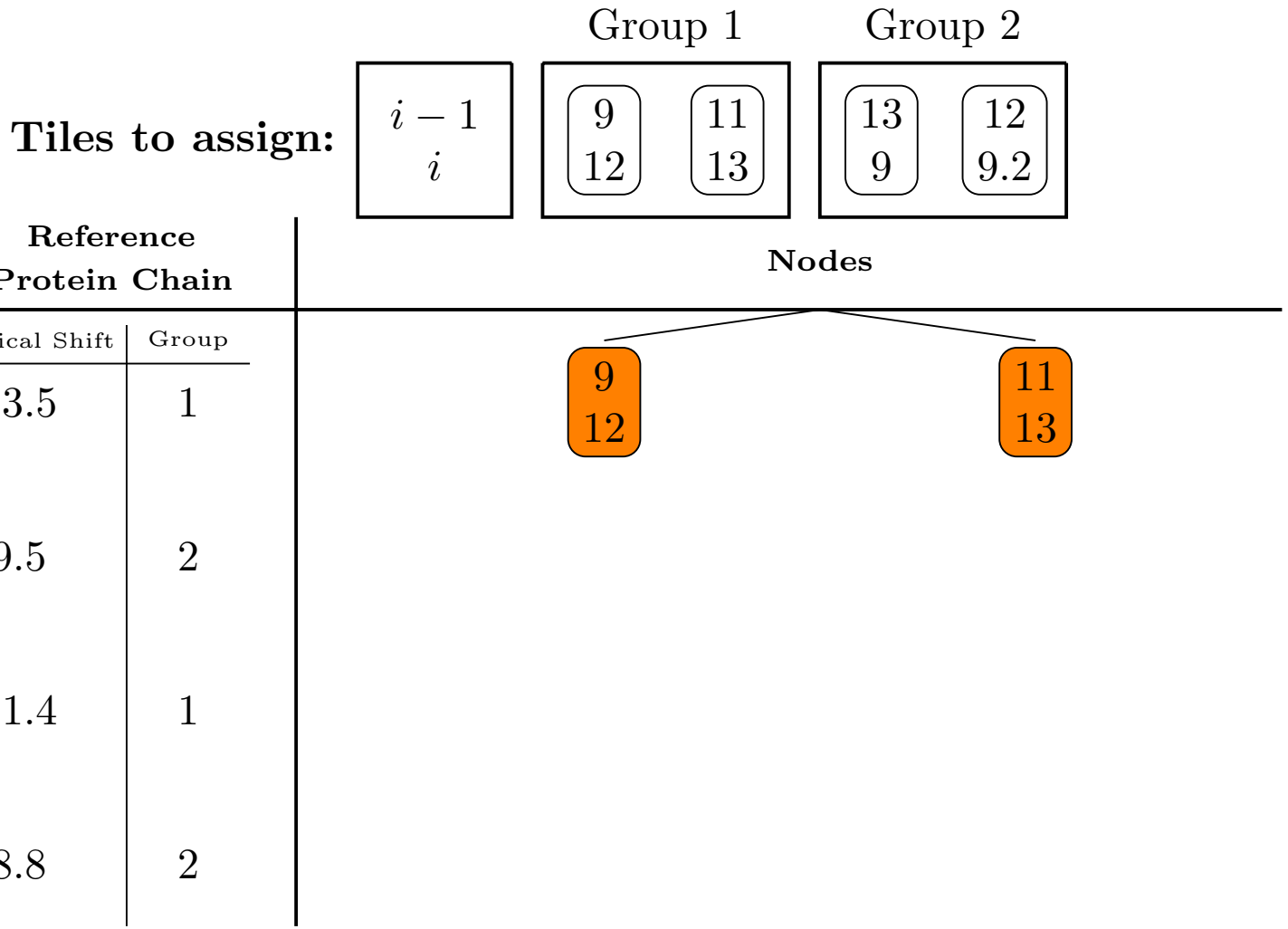
[2]

Starting the assignment



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

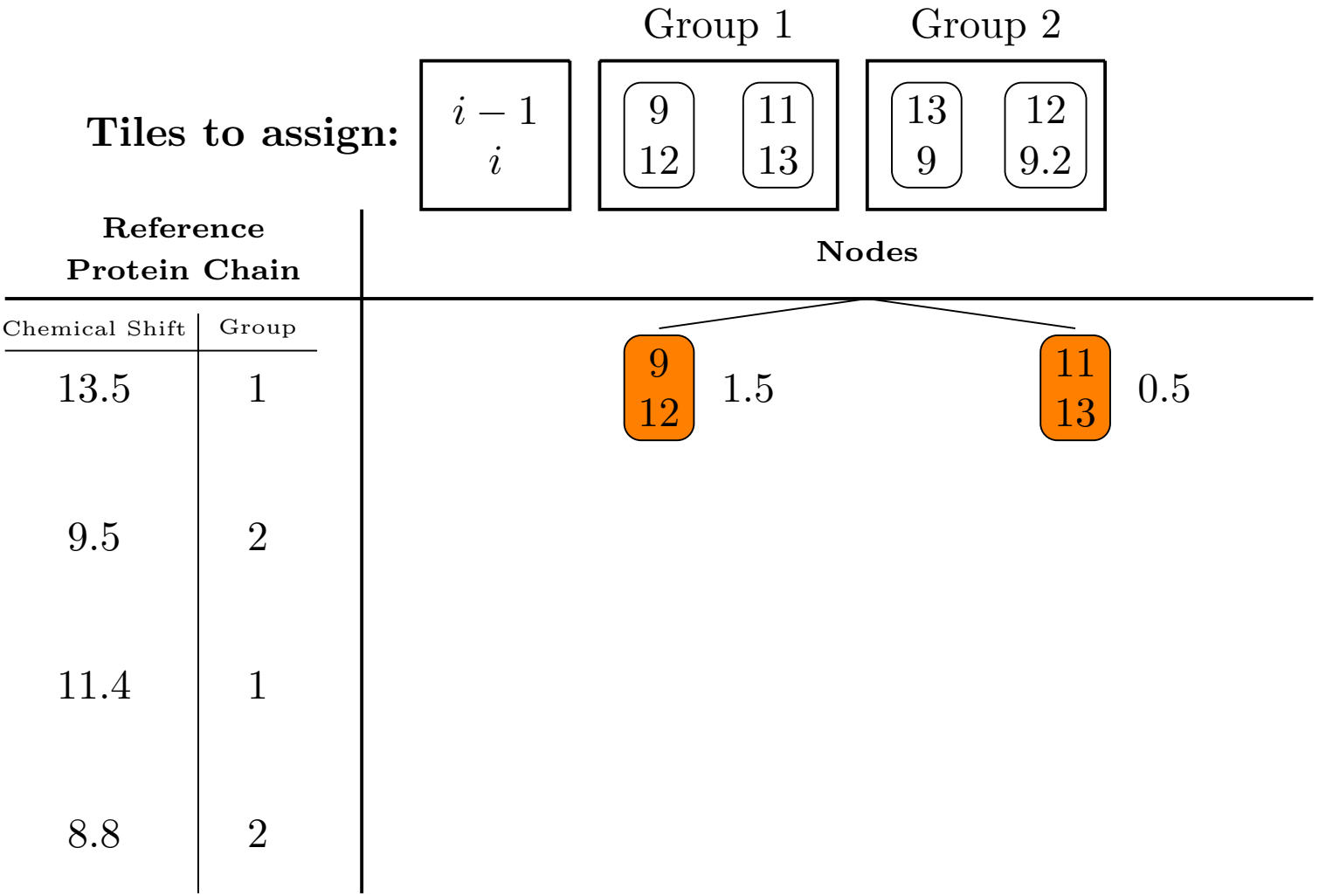
Starting the assignment



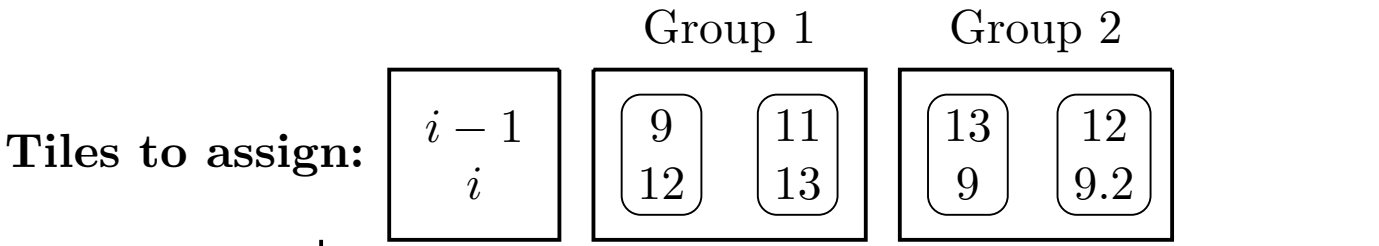
Cost Calculation

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

Generating child nodes



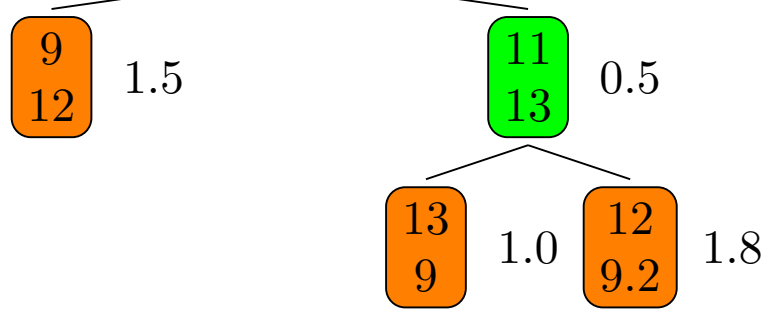
Generating child nodes



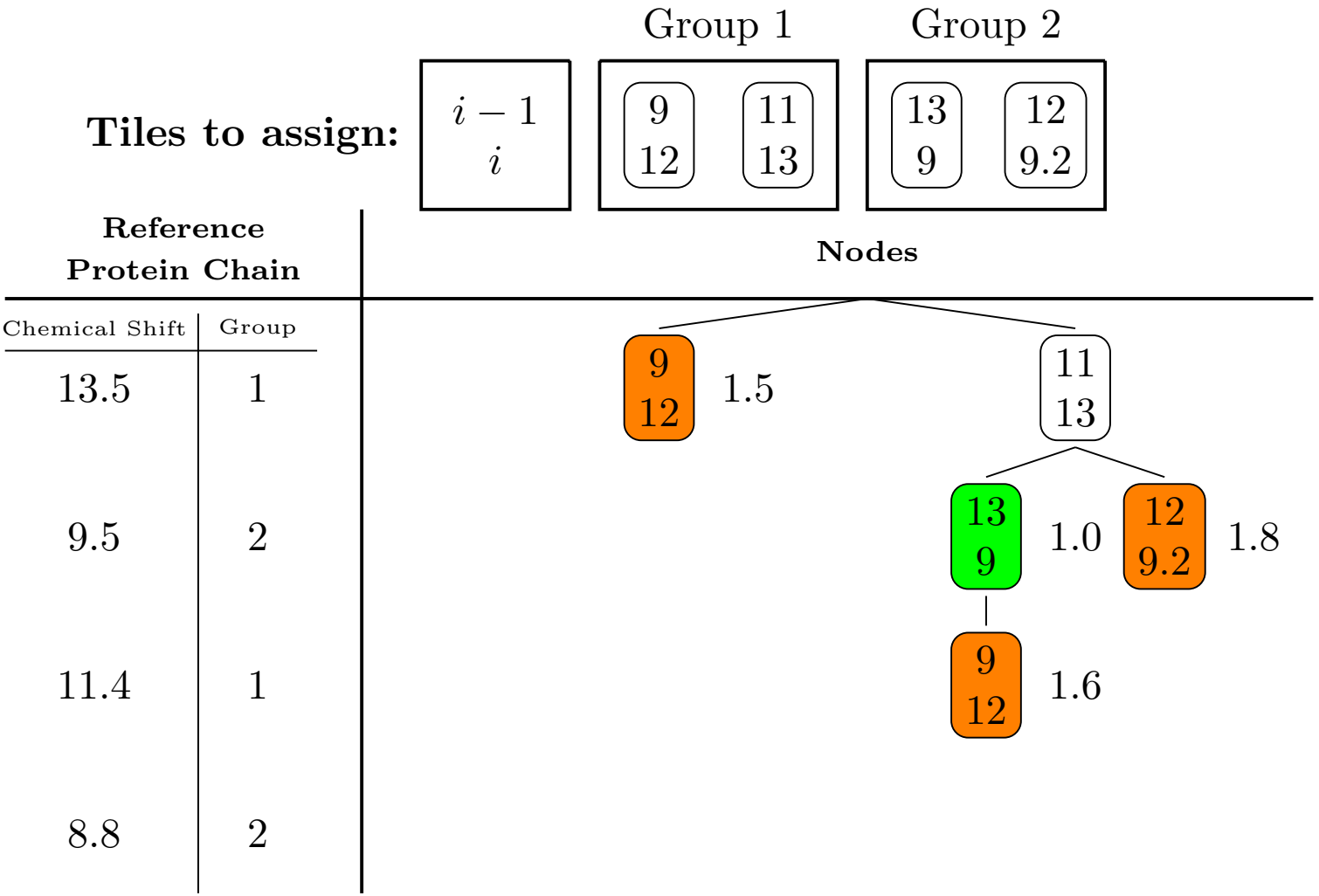
Reference Protein Chain

Chemical Shift	Group
13.5	1
9.5	2
11.4	1
8.8	2

Nodes

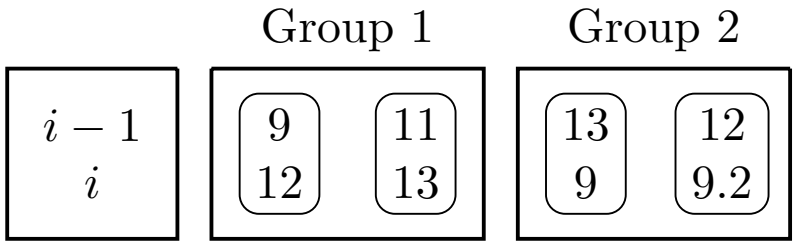


Generating child nodes



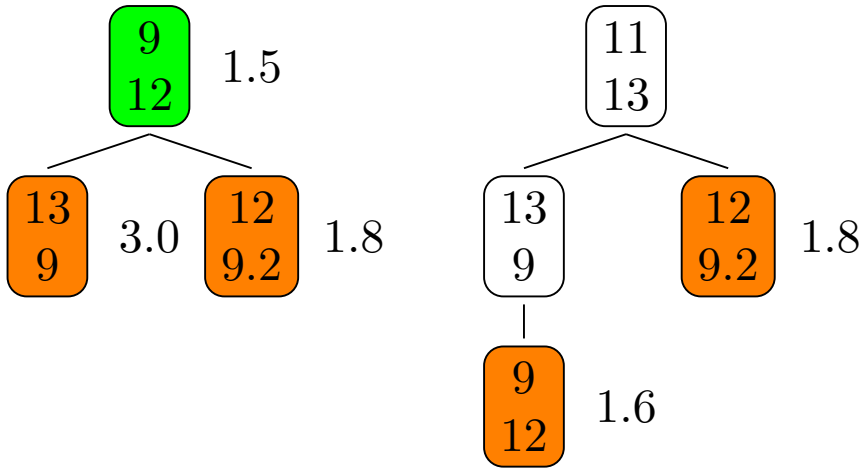
Generating child nodes

Tiles to assign:

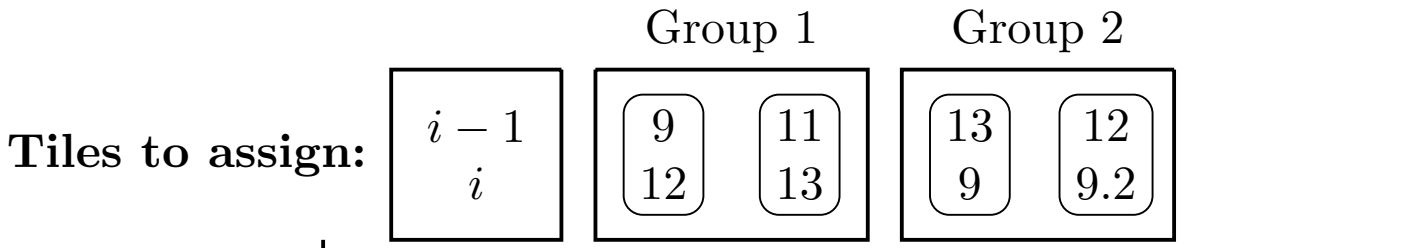


Reference Protein Chain	
Chemical Shift	Group
13.5	1
9.5	2
11.4	1
8.8	2

Nodes



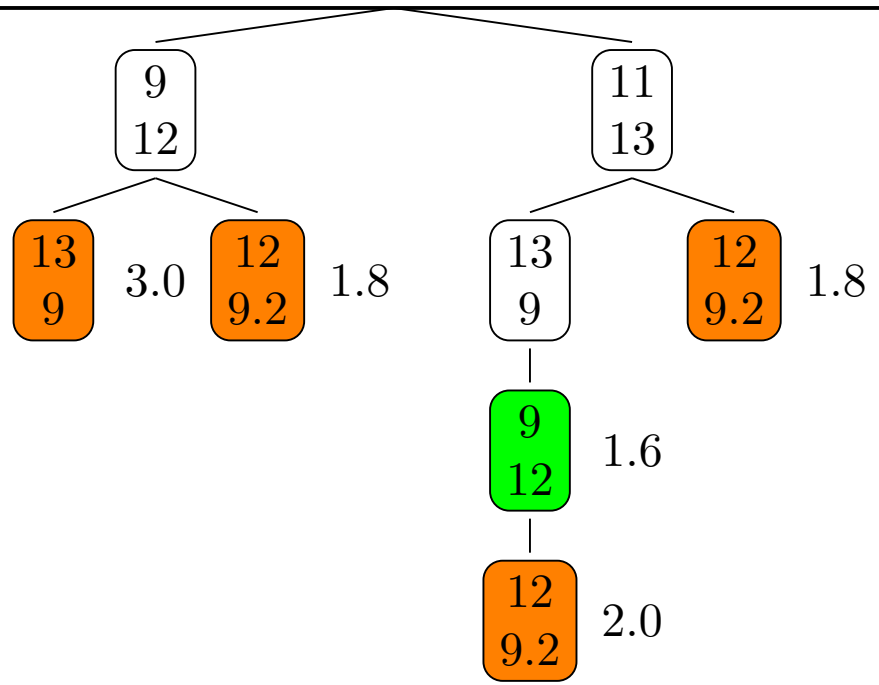
Goal State



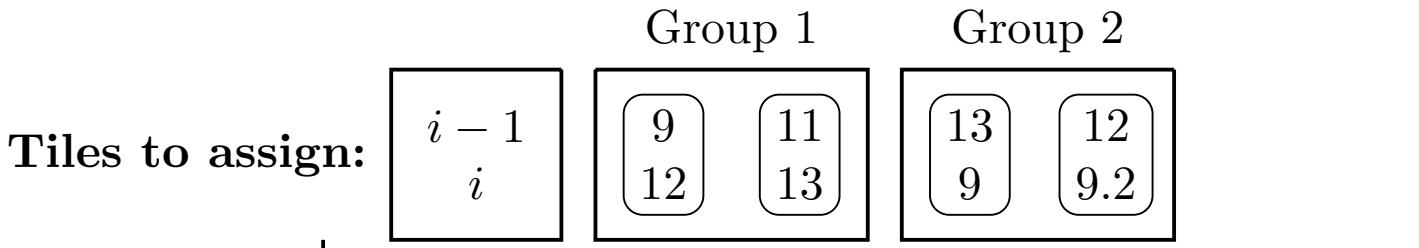
Reference Protein Chain

Chemical Shift	Group
13.5	1
9.5	2
11.4	1
8.8	2

Nodes



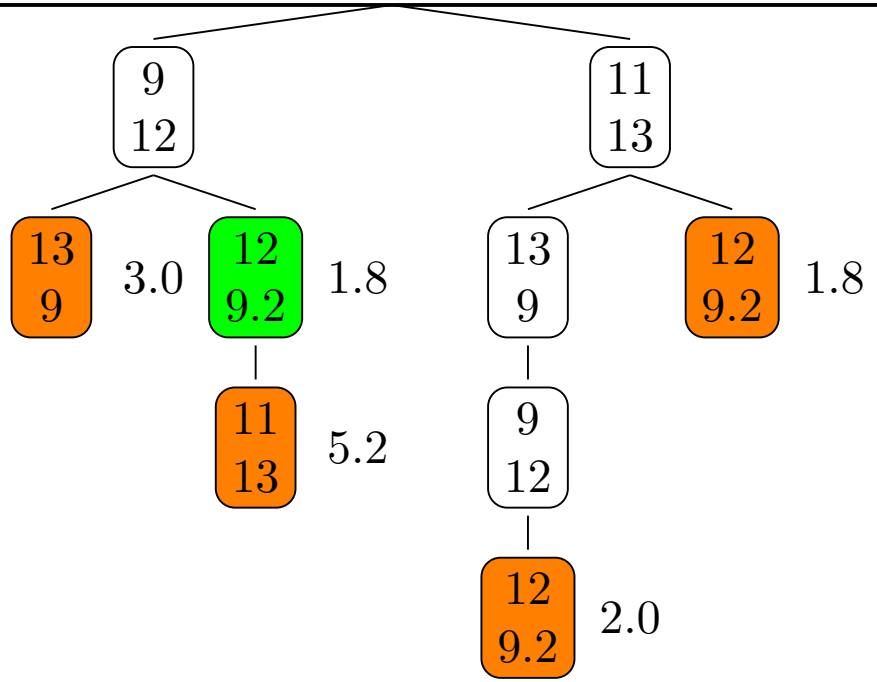
Goal State



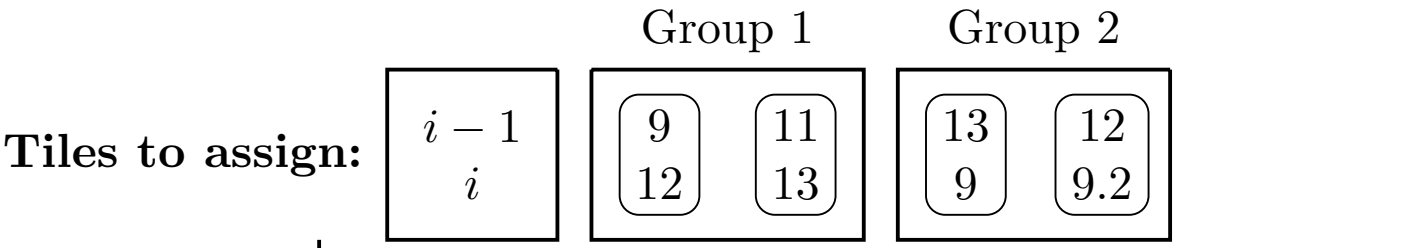
Reference Protein Chain

Chemical Shift	Group
13.5	1
9.5	2
11.4	1
8.8	2

Nodes



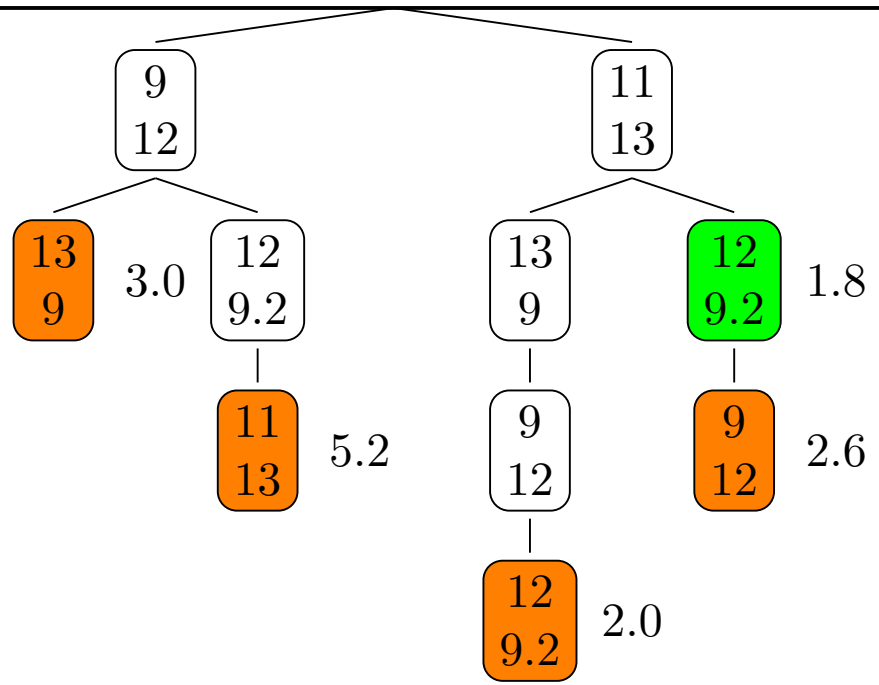
Goal State



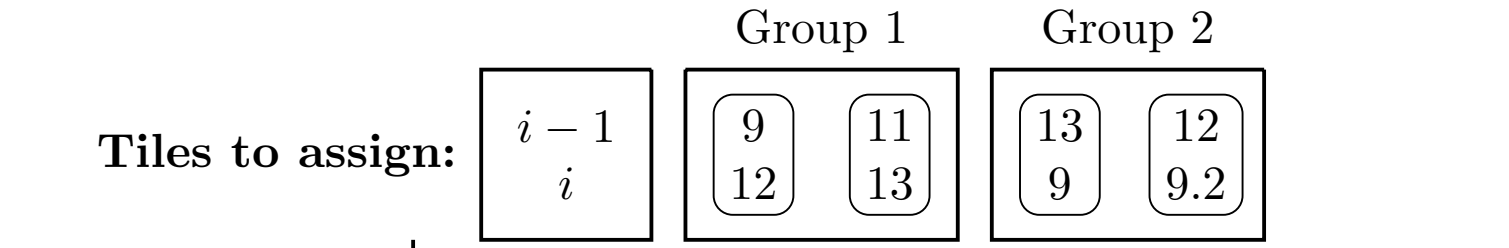
Reference Protein Chain

Chemical Shift	Group
13.5	1
9.5	2
11.4	1
8.8	2

Nodes



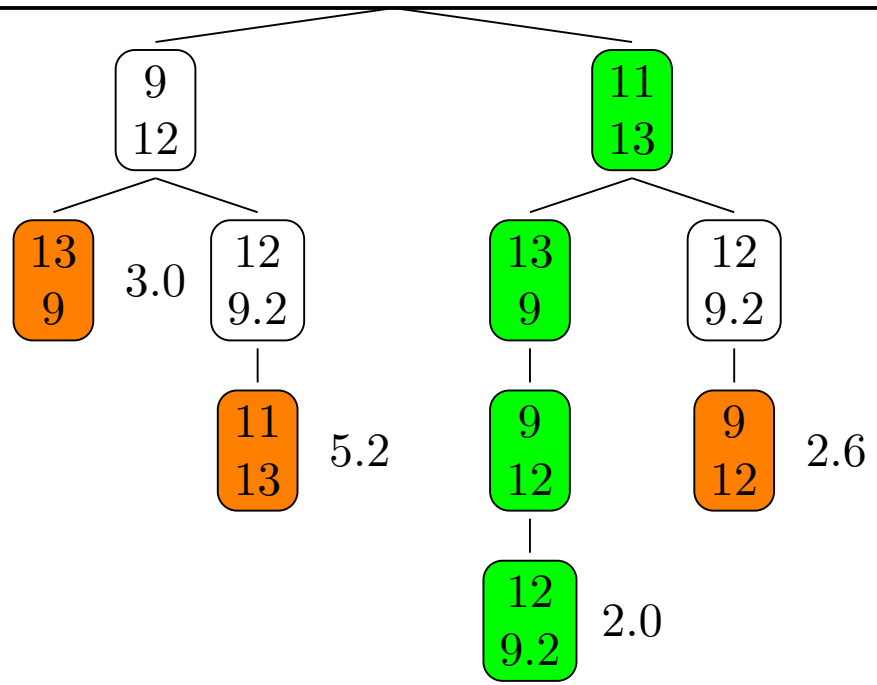
Solution State



Reference Protein Chain

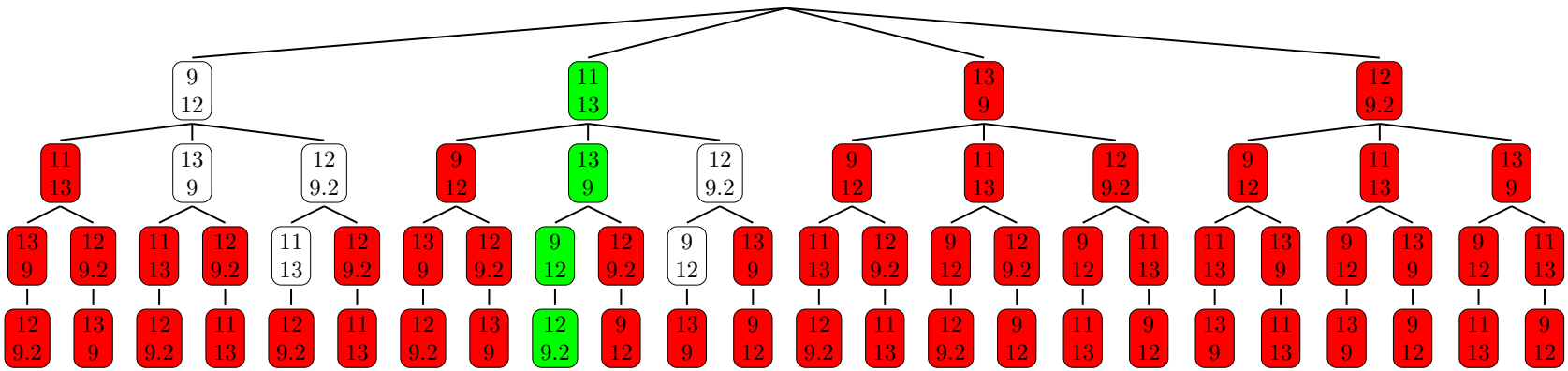
Chemical Shift	Group
13.5	1
9.5	2
11.4	1
8.8	2

Nodes

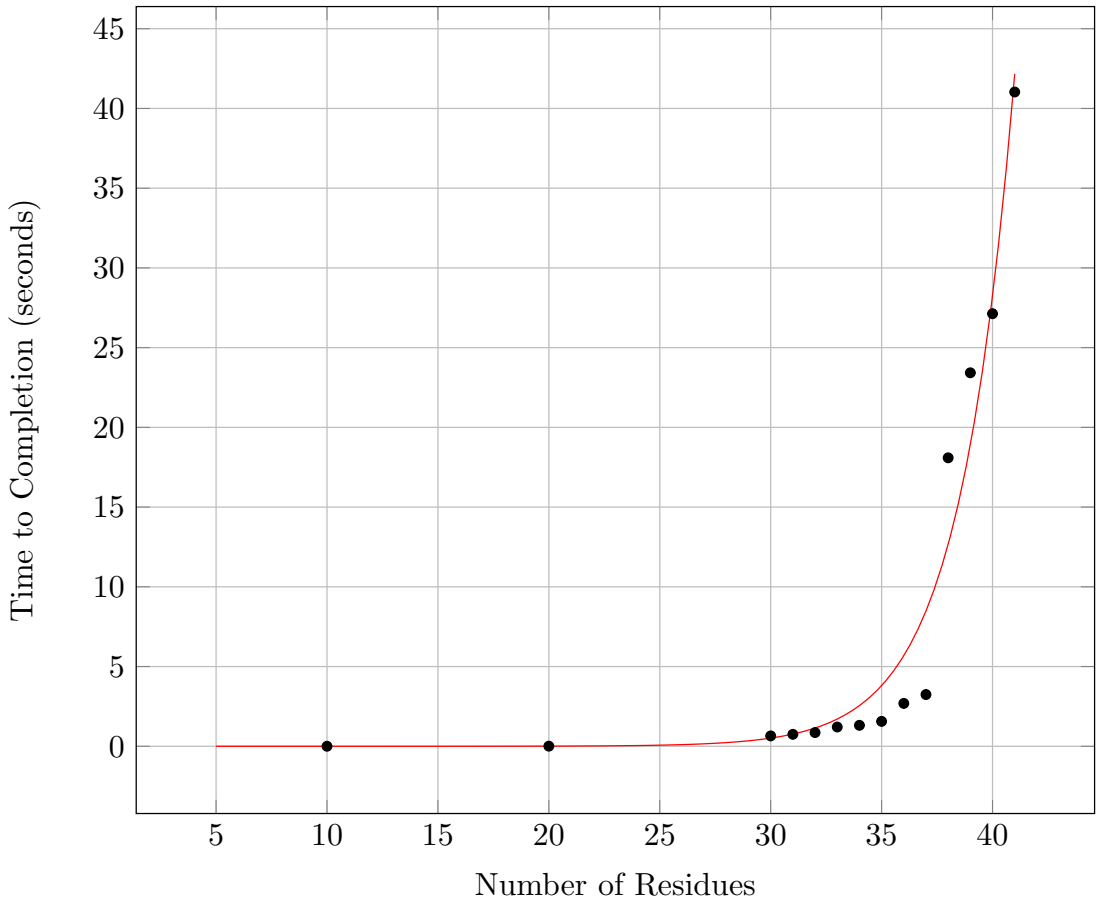


Compared to Naive Approach

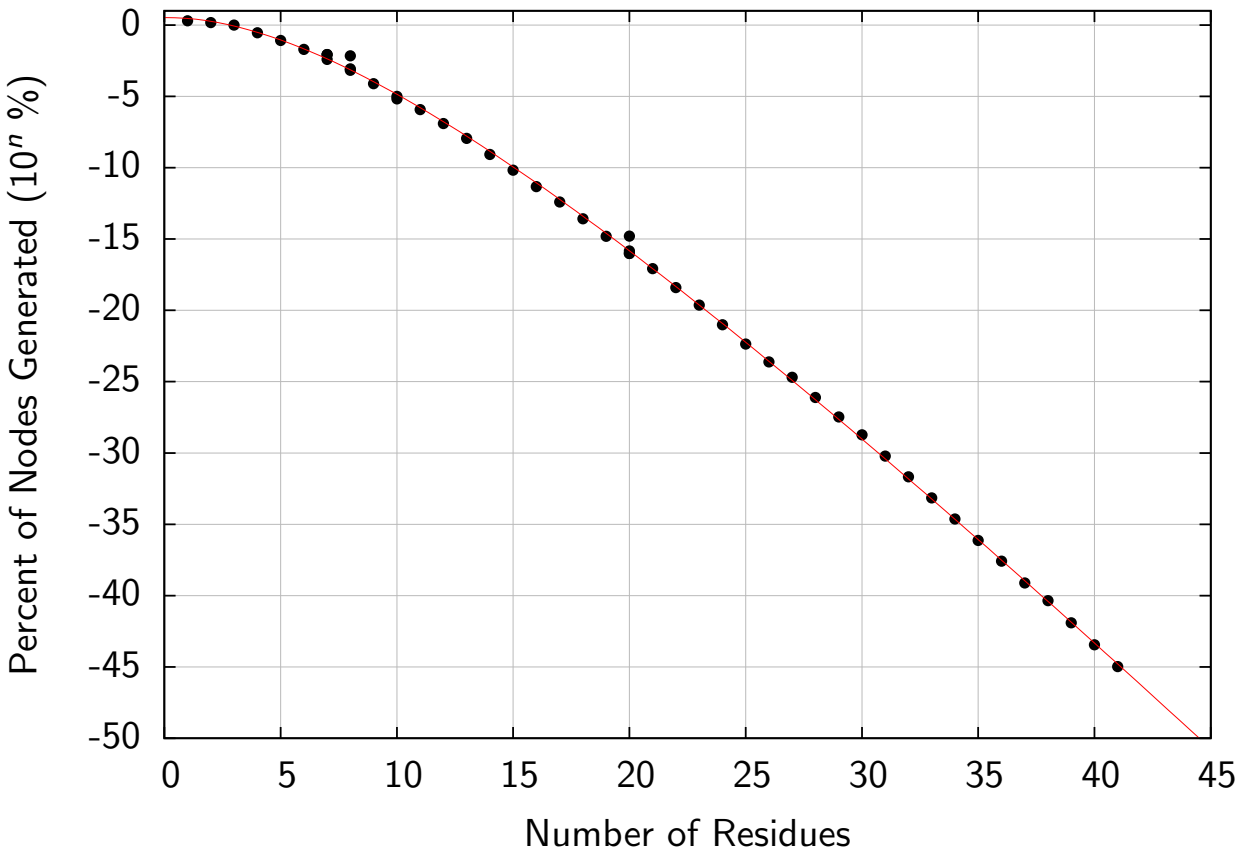
14.1% of the possible combinations



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
  - Decrease assignment time
- Custom data structure
  - Limit storing repetitive data
  - Faster node selection and generation

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

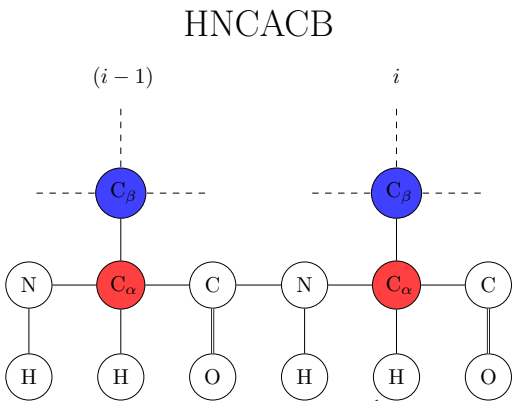

Babak Alipanahi, Xin Gao, Emre Karakoc, Frank Balbach, Shuai Cheng Li, Guangyu Feng, Logan Donaldson and Ming Li, *Error tolerant NMR backbone resonance assignment and automated structure generation.*, Journal of bioinformatics and computational biology, **9** (2011), 15–41.

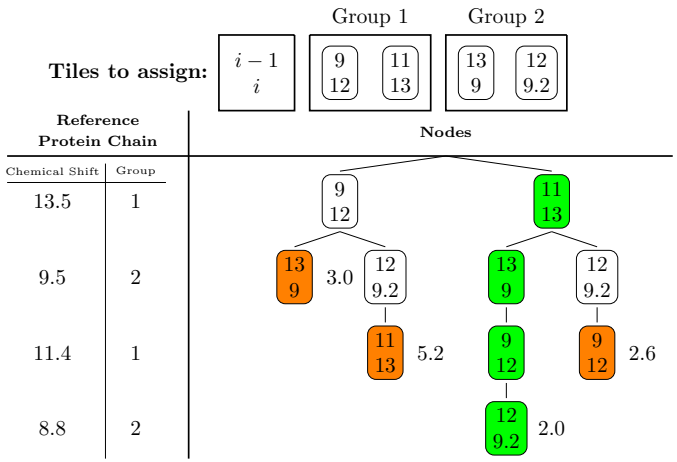

Sean Cahill and Mark Girvin.
Introduction to 3d triple resonance experiments.
2012.

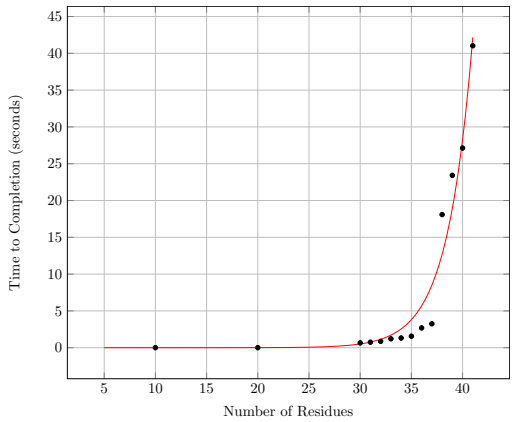

Peter Guntert, *Automated structure determination from NMR spectra*, European Biophysics Journal, **38** (2009), 129–143.

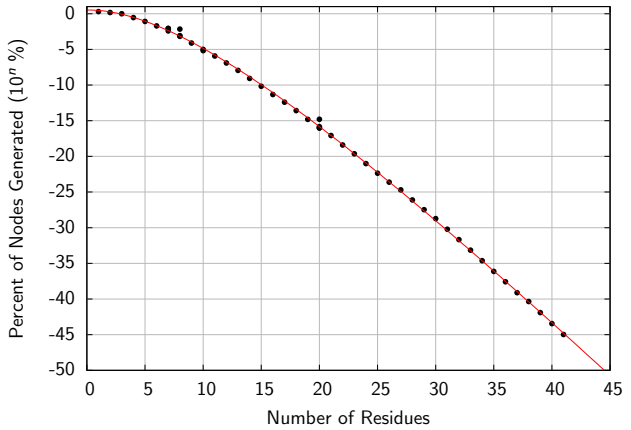

Flemming M. Poulsen.
A brief introduction to nmr spectroscopy of proteins.

Thank You


HNCACB
$(i-1)$
i
C_β
C_α
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