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

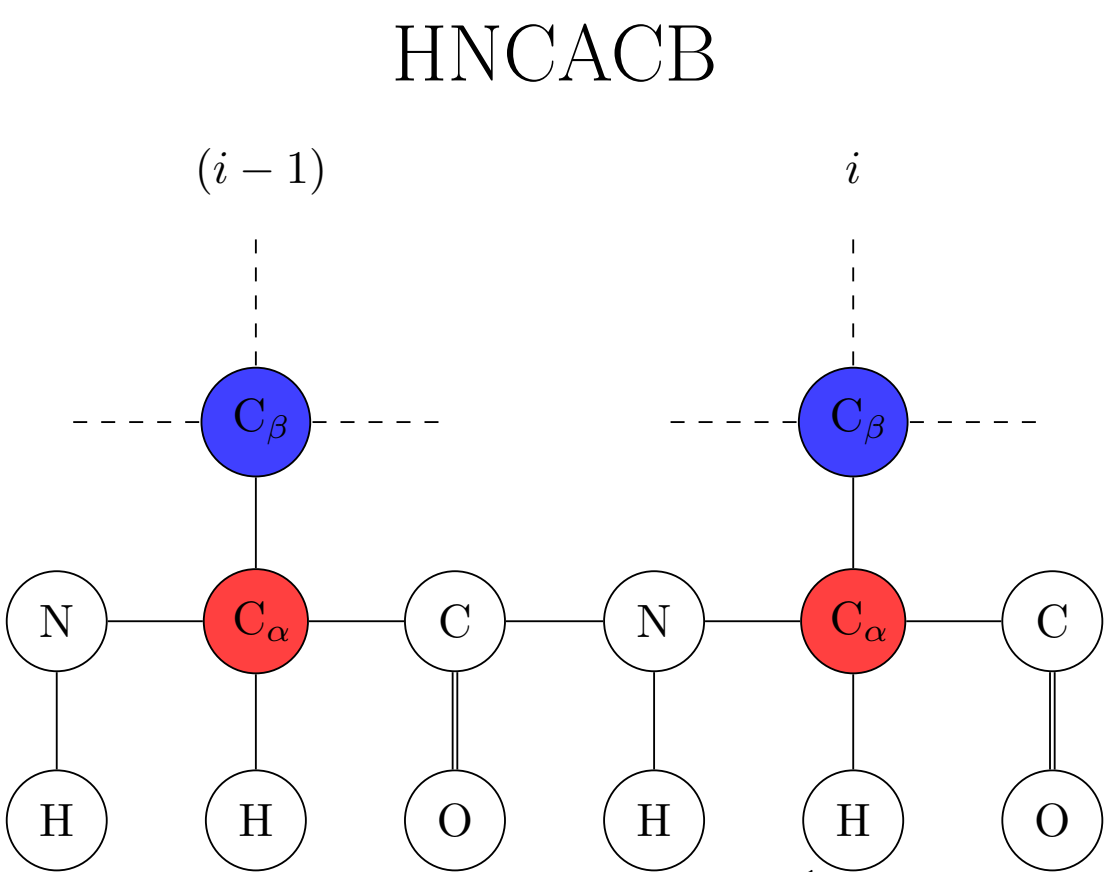
Motivation

- 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_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

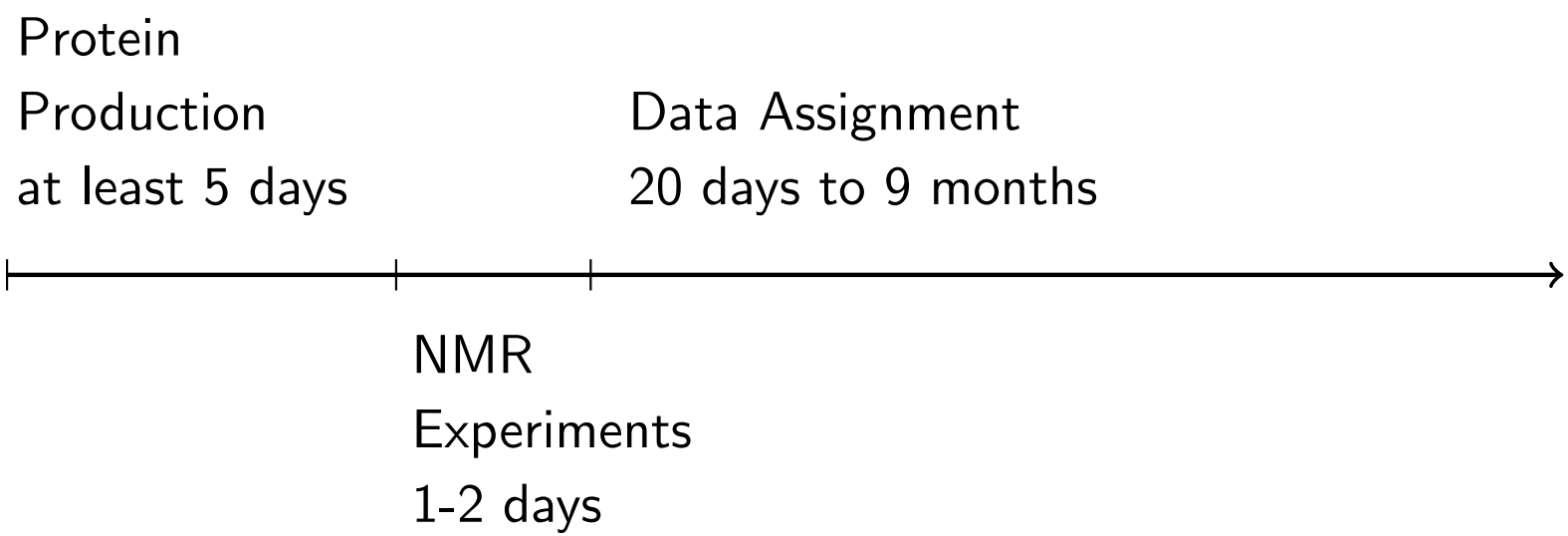
Chemical Shift Values



Manual Methods

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

Timeline

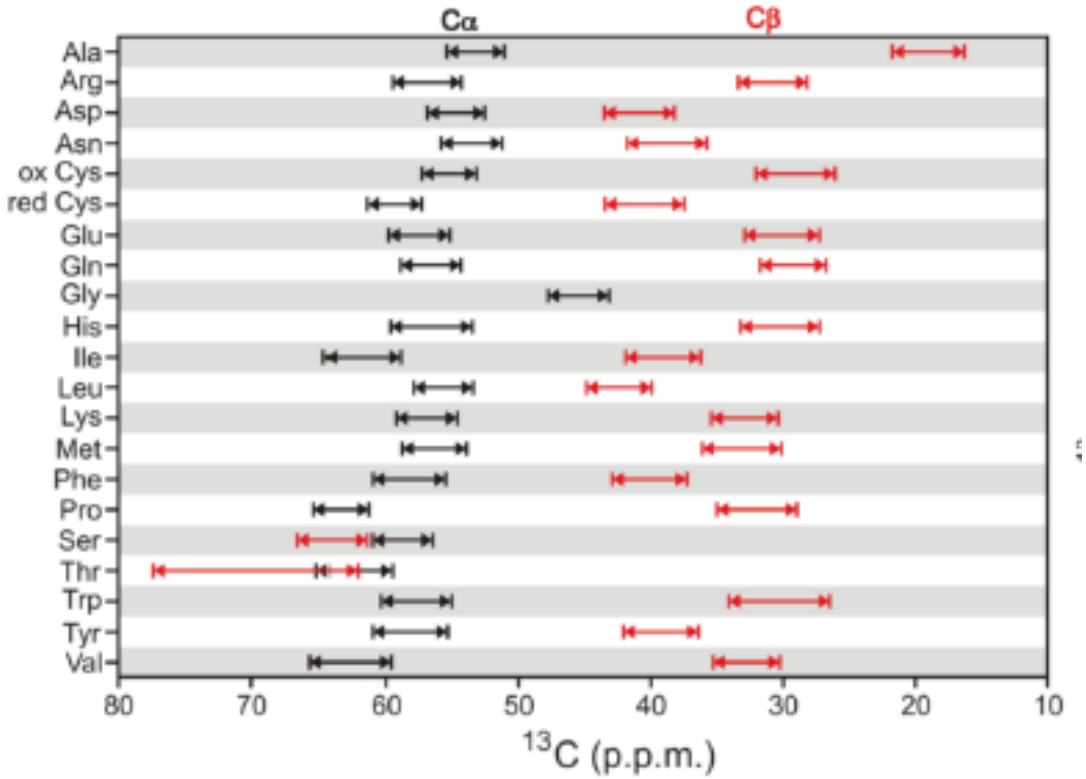

Protein
Production
at least 5 days
Data Assignment
20 days to 9 months
NMR
Experiments
1-2 days

[?]

Initialization

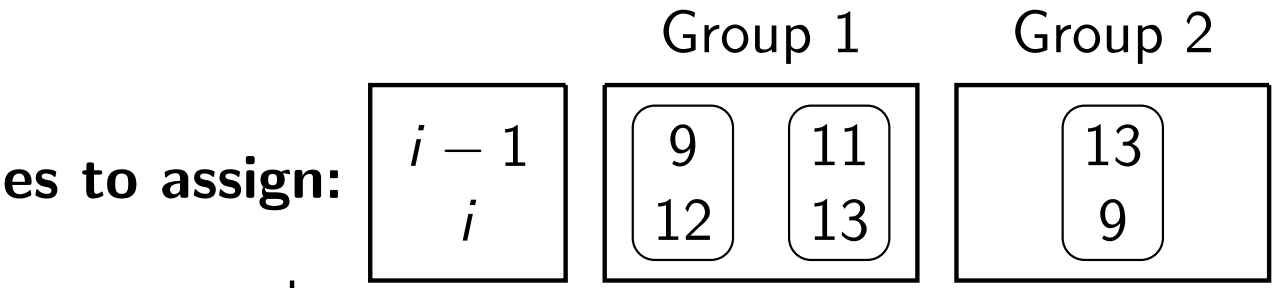
- Expected amino acid sequence
  - Converted to expected chemical shift values
  - Stored as the reference protein chain
- NMR experiment's chemical shift data
  - C_α and C_β for residue i and $i-1$
  - Stored in a tile
- Missing data
  - Place holder tile generation
- Grouping

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.)

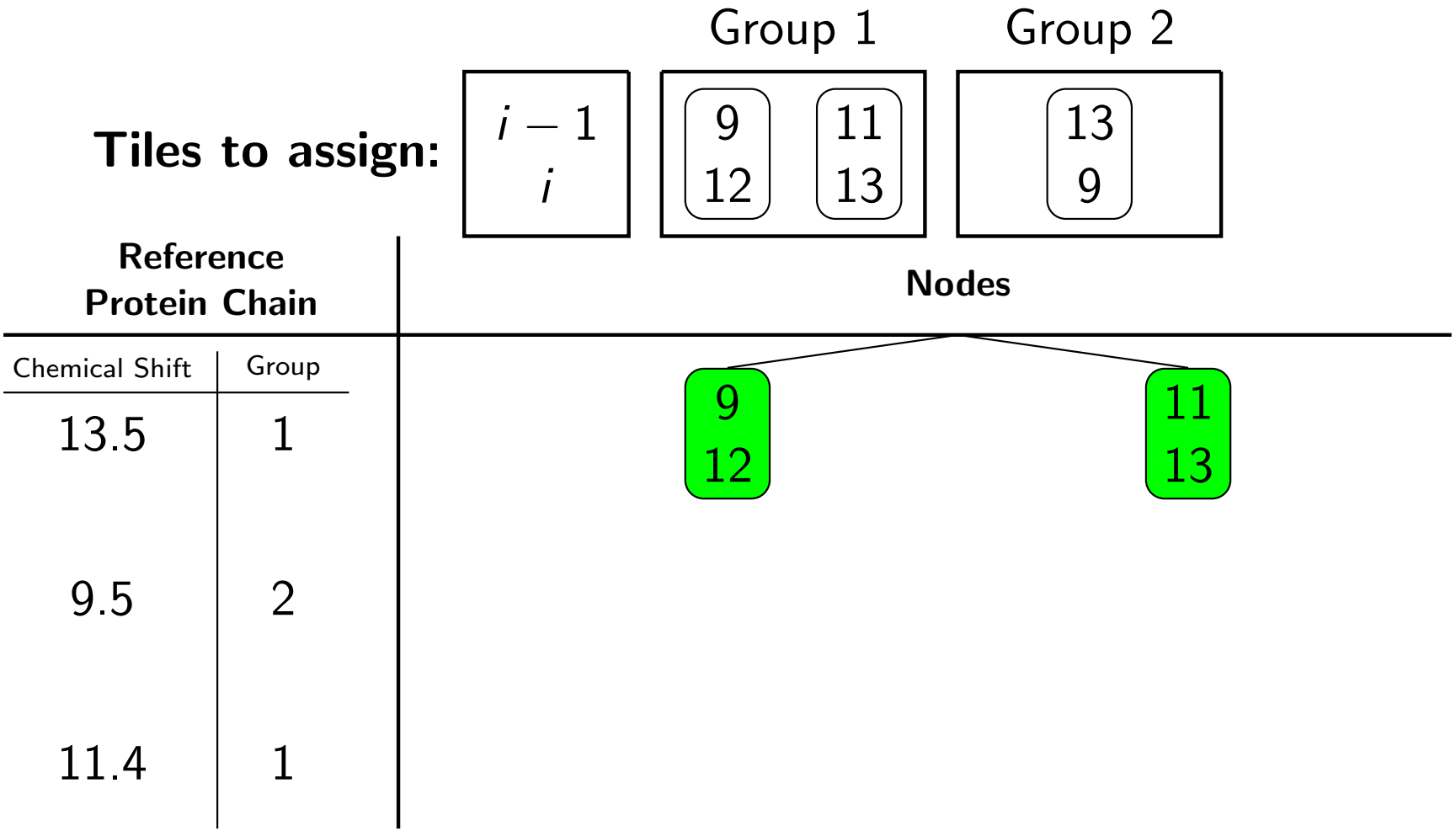
[?]

Starting the assignment



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

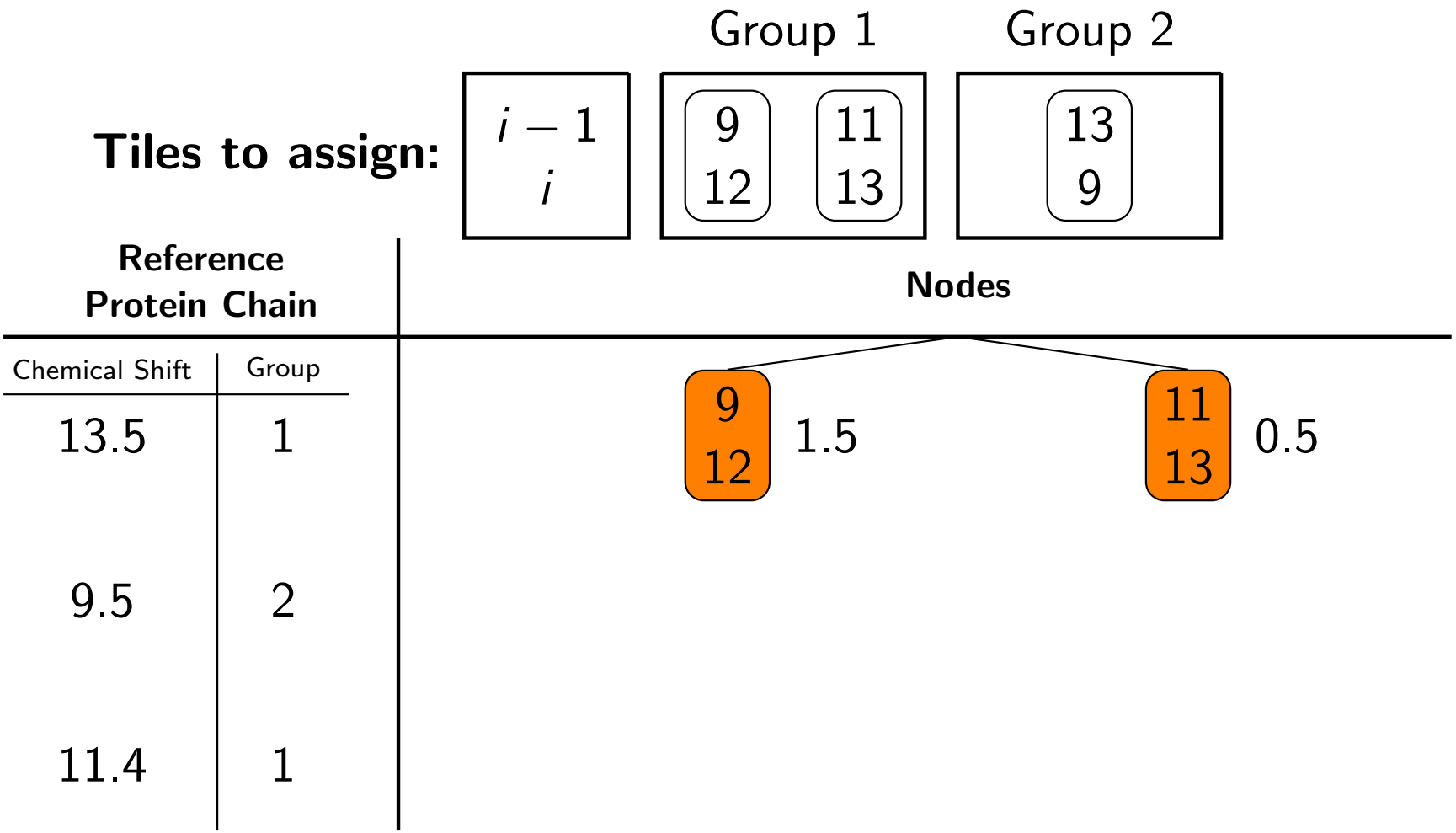
Starting the assignment



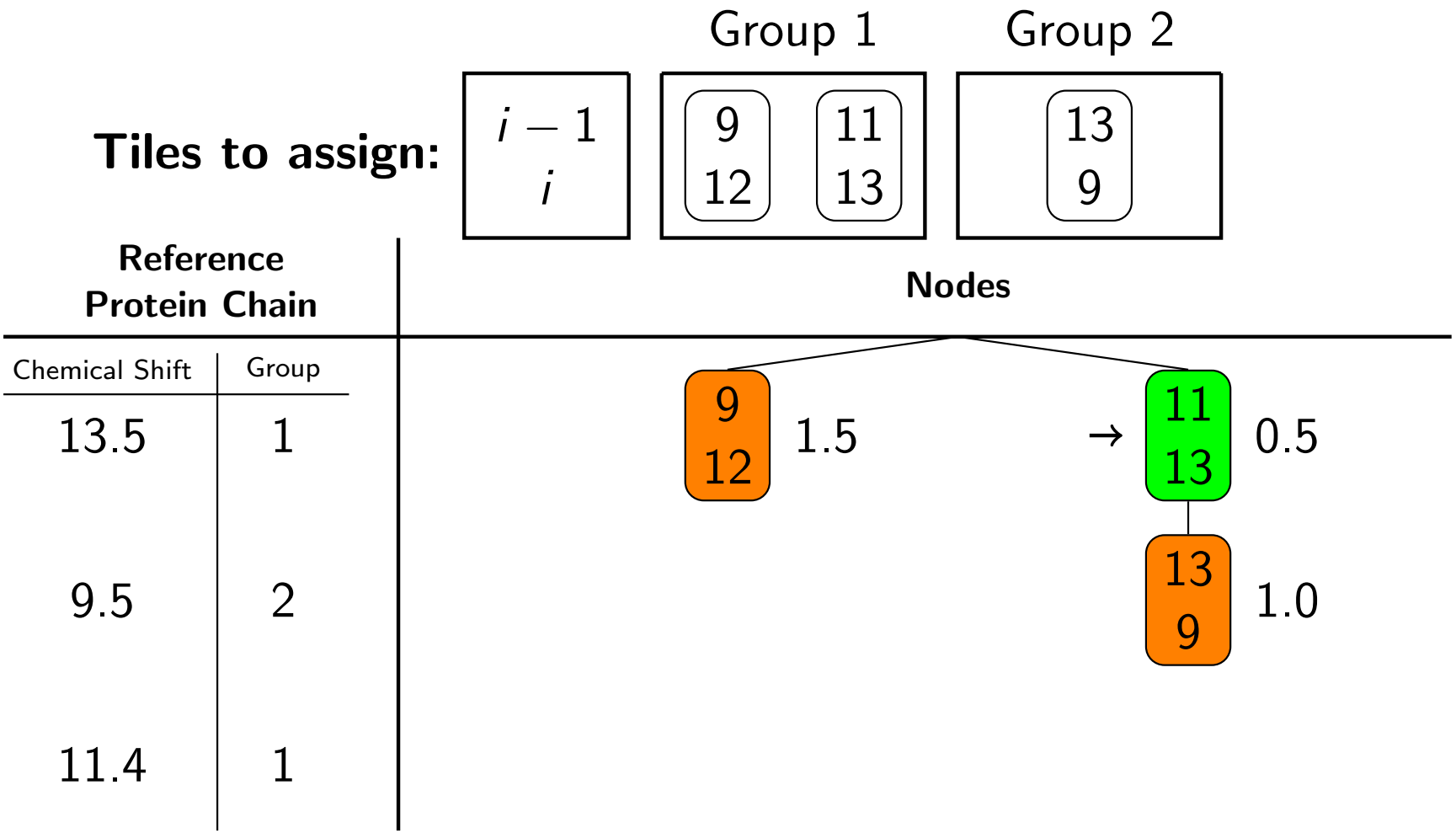
Cost Calculation

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

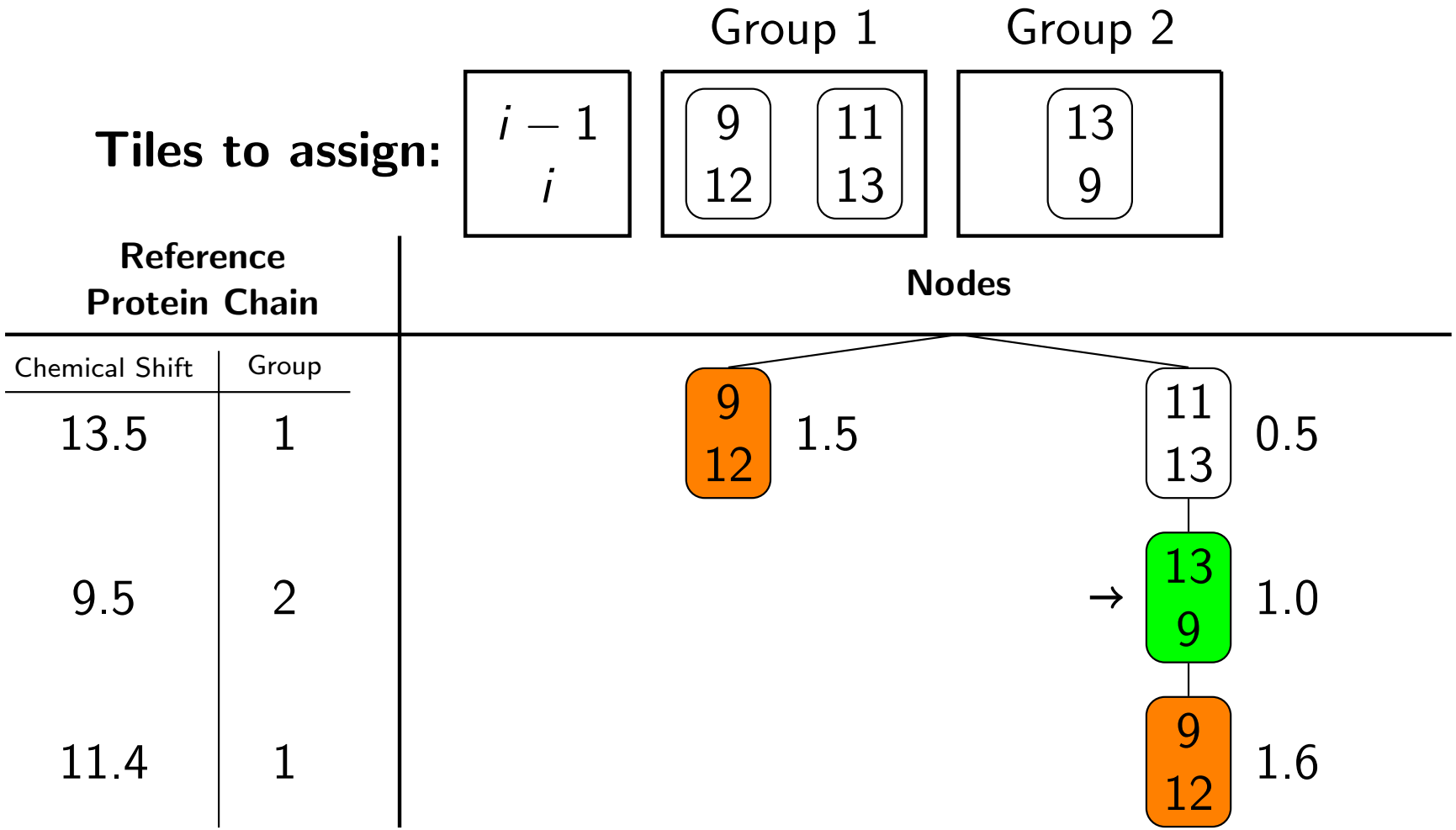
Generating child nodes



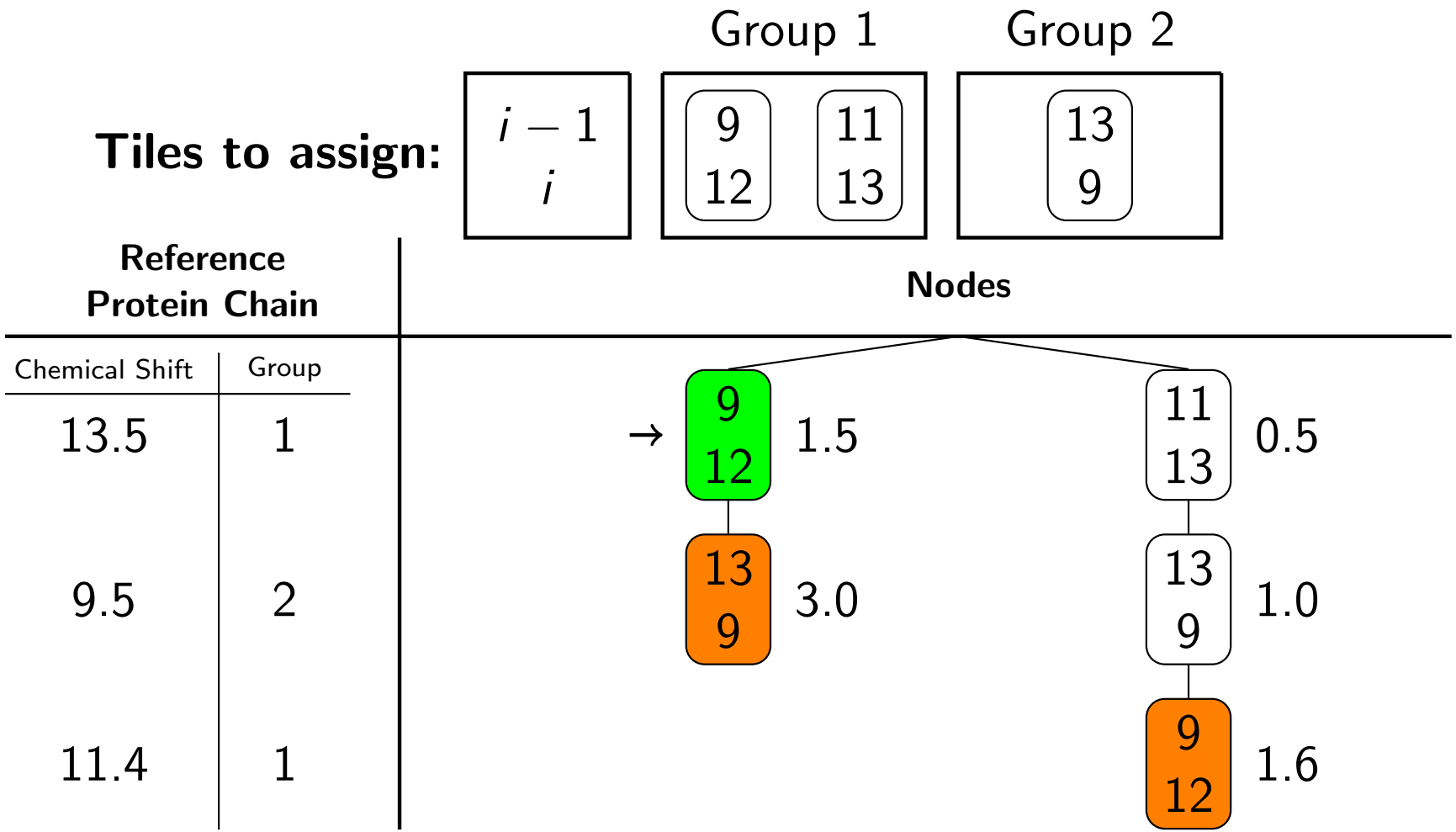
Generating child nodes



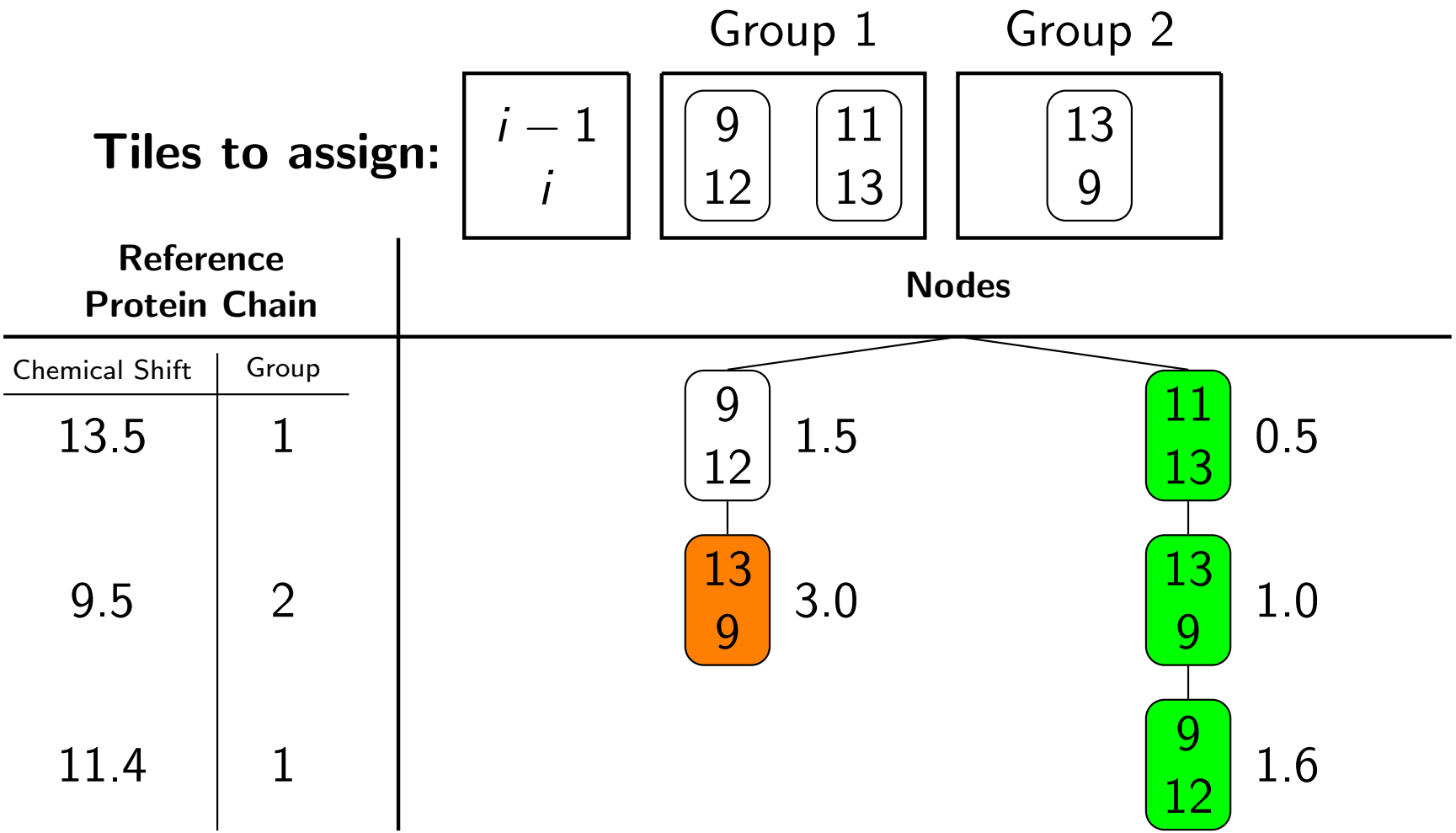
Goal State


Group 1
Group 2
Tiles to assign:
$i-1$
i
9
12
11
13
13
9
Reference Protein Chain
Chemical Shift
Group
13.5
1
9.5
2
11.4
1
Nodes
9
12
1.5
11
13
0.5
13
9
1.0
9
12
1.6

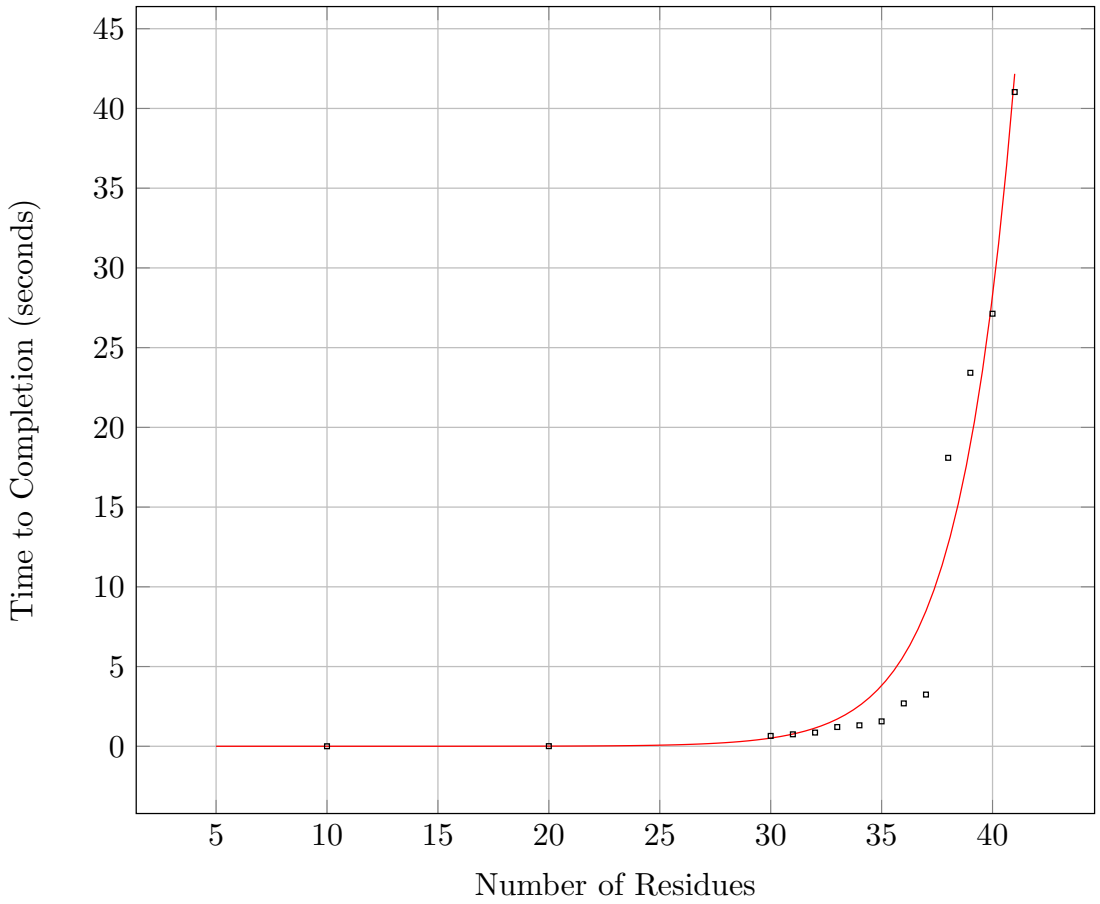
Goal State


Group 1
Group 2
Tiles to assign:
$i-1$
i
9
12
11
13
13
9
Reference Protein Chain
Chemical Shift
Group
13.5
1
9.5
2
11.4
1
Nodes
→
9
12
1.5
13
9
3.0
11
13
0.5
13
9
1.0
9
12
1.6

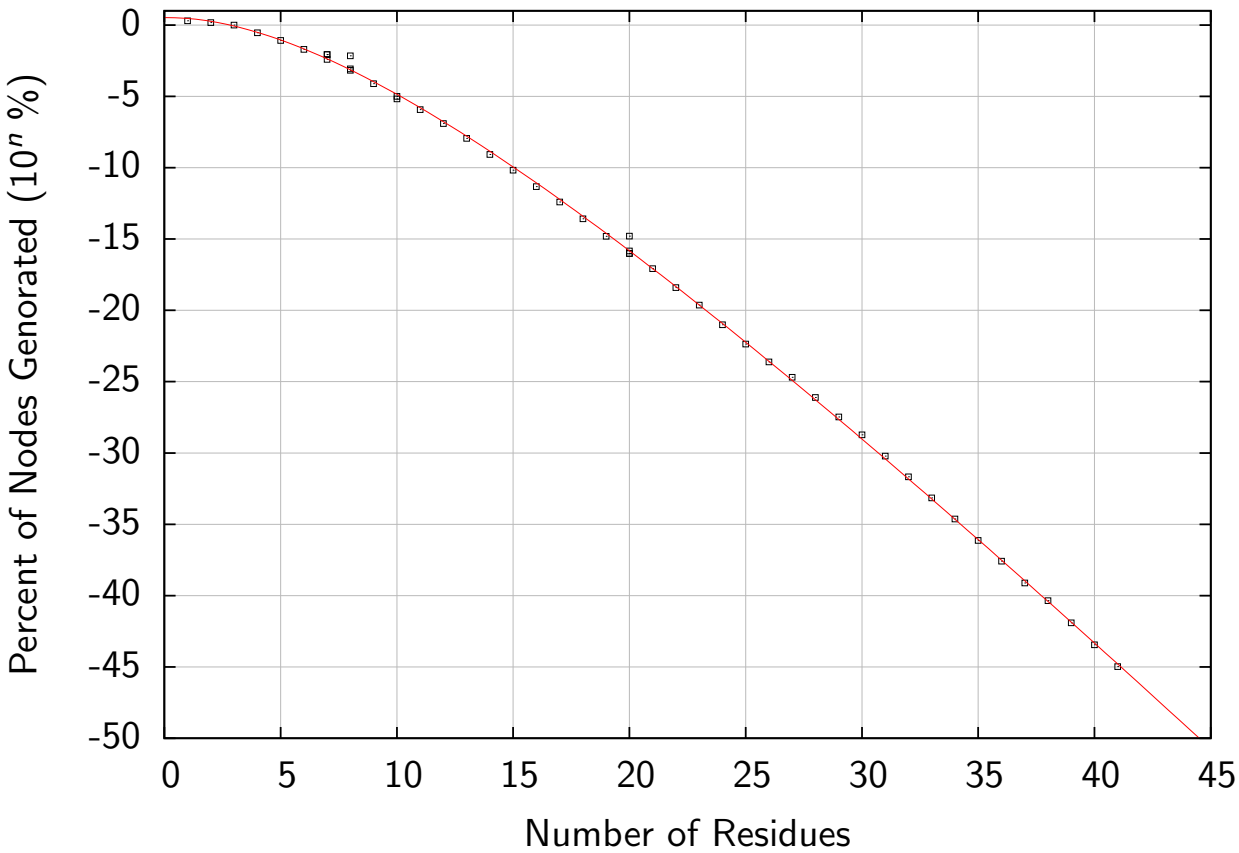
Solution State



Time of Assignment



Child Nodes Genorated



Future Goals

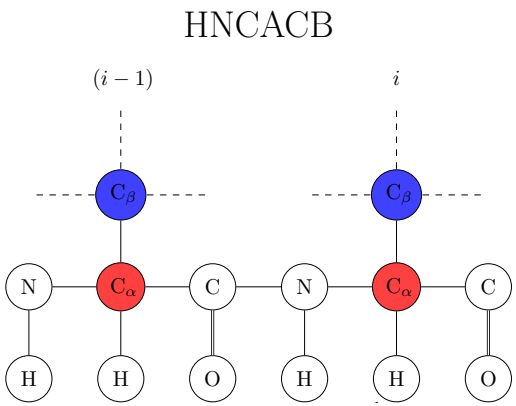
- Parallelization
  - Decrease assignment time
  - Allow for larger data sets
- Machine learning
  - Optimize cost calculation
  - Increase accuracy of assignment

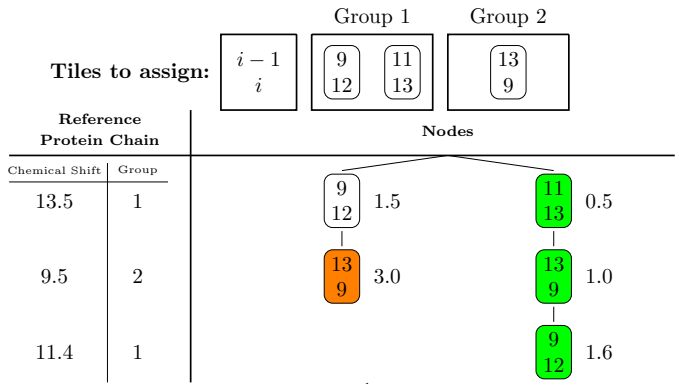
Acknowledgments

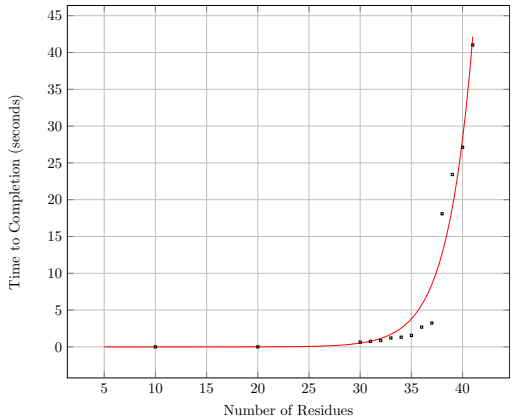
- Dr. Tim Urness (Mathematics and Computer Science)
- Dr. Adina Kilpatrick (Physics)
- Rachel Davis (research colleague)
- John Emmons (research colleague)
- Katherine Roth (research colleague)
- David Mascharka (research colleague)
- Leah Robison (research colleague)

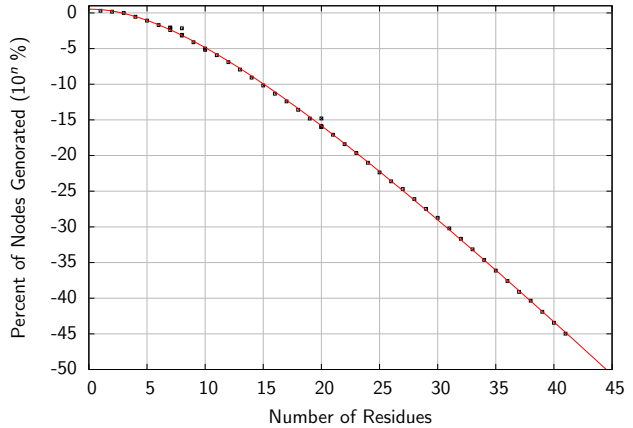
Bibliography

Thank You


HNCACB
$(i-1)$
i
C_β
C_α
N
C
H
O


Group 1
Group 2
Tiles to assign:
$i-1$
i
Reference Protein Chain
Nodes
Chemical Shift
Group
13.5 1
9.5 2
11.4 1
1.5
3.0
0.5
1.0
1.6


Time to Completion (seconds)
Number of Residues


Percent of Nodes Genorated (10" %)
Number of Residues