Utilizing Machine Learning to Accelerate Automated Assignment of Backbone NMR Data

Joel Venzke[12], David Mascharka[1], Paxten Johnson[12], Rachel Davis[1], Katherine Roth[1], Leah Robison[1], Timothy Urness[1] and Adina Kilpatrick[2]

[1]Department of Mathematics and Computer Science
[2]Department of Physics and Astronomy
Drake University

joel.venzke@drake.edu

April 16, 2015

Overview

Motivations

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
- Takes a few days to months to assign manually[1]

Goal

- Automate the assignment process
- Decrease human error
- Increase productivity

---

1. Jens P. Linge et al., "ARIA: Automated NOE assignment and NMR structure calculation," *Bioinformatics*:2003.

NMR Experiments

NMR Data Sets

- Produces data corresponding to structure

HNCACB experiment

- Generates C_α and C_β residue i and $i-1$

CBCA(CO) NH experiment[2]

- Generates C_α and C_β for residue i
- Confirms residue data

2. Linge et al., "ARIA: Automated NOE assignment and NMR structure calculation."

Machine Learning

Overview

- Generalize large amounts of data
- Predicts a label based on attributes
- Many different algorithms exist

Supervised vs Unsupervised Learning

- Supervised
  - Given large amounts of labeled data
  - Classifies data based on attributes
- Unsupervised
  - Given large amounts of unlabeled data
  - Looks for patterns

Machine Learning Algorithms

J4.8[3]

- Decision tree model
- Splits by a single attribute

Logistic Model Tree (LMT)[4]

- Decision tree model
- Splits by linear regression of attribute

Decision Table[5]

- Set of labeled data is searched
- Majority match is used

---

3. Ross Quinlan, *C4.5: Programs for Machine Learning* (San Mateo, CA: Morgan Kaufmann Publishers, 1993).

4. Niels Landwehr, Mark Hall, and Eibe Frank, "Logistic Model Trees," *Machine Learning* 95, nos. 1-2 (2005): 161–205.

5. Ron Kohavi, "The Power of Decision Tables," in *8th European Conference on Machine Learning* (Springer, 1995), 174–189.

Algorithmic Overview

Model Training

Preprocessing

The Search

Model Training

Biological Magnetic Resonance Bank (BMRB)[6]

- 9,736 datasets containing chemical shifts for the C_α and C_β resonances of 689,977 residues
- Removing outliers leaves 681,363 pairs of C_α and C_α
  - 3 standard deviations from the mean
  - Avoids over-fitting
  - Improves algorithmic performance

6. Eldon L. Ulrich et al., "BioMagResBank.," *Nucleic Acids Research* 36, no. Database-Issue (January 23, 2008): 402–408, http://dblp.uni-trier.de/db/journals/nar/nar36.html#UlrichADHILLMMMNSTWYM08.

Model Training

Training the Model

Performed Once

- Time consuming task
- Trained once, used many times

Models Trained

- DecisionTable, J4.8, LMT

Reading Data

Protein Sequence

- Read in as letters
- Converted to BMRB average values
- Used for comparison in the search

NMR Data Set

- Read in C_α, C_β for Residue i and $i-1$
- Stored in Tile

Tile

Residue $i-1$

C_α, C_β

Residue i

C_α, C_β

Confidence Level Calculation

Machine Learning

- Input
  - C_α, C_β values for residue i
- Output
  - Confidence levels for each of the 20 amino acids
  - $P_1, P_2, \cdots, P_{19}, P_{20}$
  - Confidence levels are on a scale from 0.0-1.0
  - 1.0 being a prefect match

Tile

Residue $i-1$
C_α, C_β
Residue i
C_α, C_β
Confidence Levels
P_1, P_2, $\cdots$, P_{19}, P_{20}

Missing Data

Blank Tile Creation

- Compare length of protein sequence to NMR Data set
- Blank tiles are created to make up the gap

Proline

- Lacks H-N spin system
- Does not produce C_α, C_β values
- Protein sequence is examined
- Special flags are set

Blank Tile

Residue $i-1$

\- , -

Residue i

\- , -

Confidence Levels

$1.0, 1.0, \cdots, 1.0, 1.0$

Proline

yes/no

First Tile

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

Reference Protein Chain		**Nodes**
Chemical Shift	P_n	
13.5	1	
9.5	2	
11.4	1	
-	Proline	
8.8	2	

First Tile

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

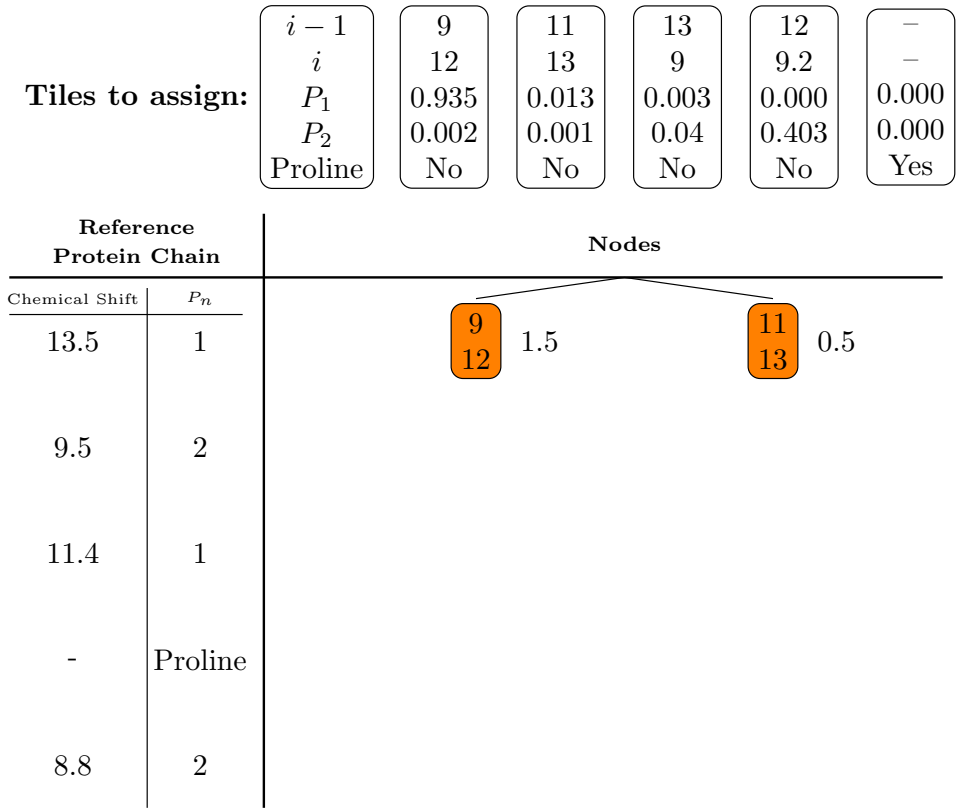
Reference Protein Chain

Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
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



Node Generation

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

Reference Protein Chain	
Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2

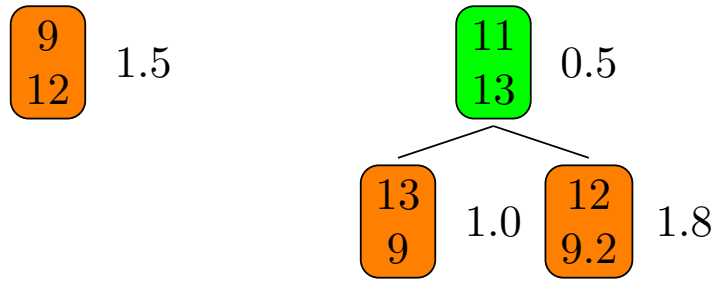
Nodes

Node Generation

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

Reference Protein Chain	
Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2

Nodes

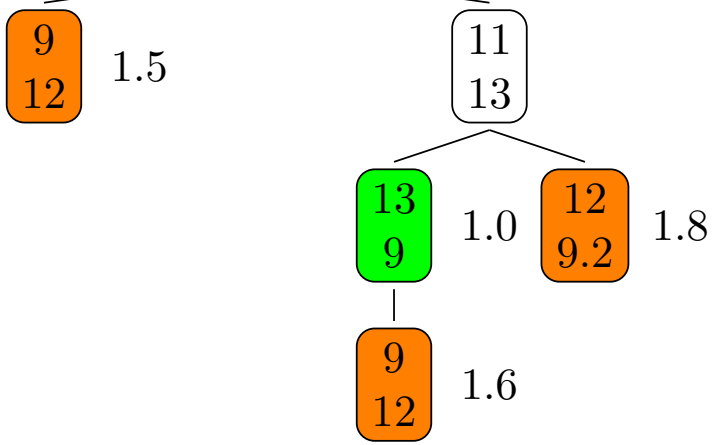


Node Generation

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

Reference Protein Chain	
Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2

Nodes

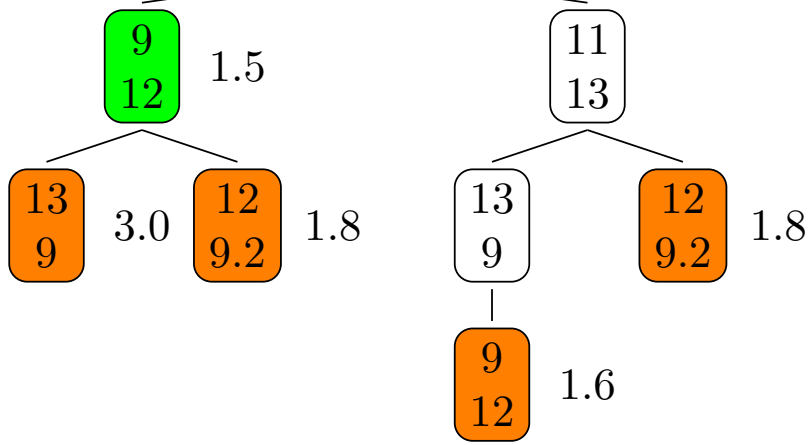


Node Generation

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

Reference Protein Chain	
Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2

Nodes



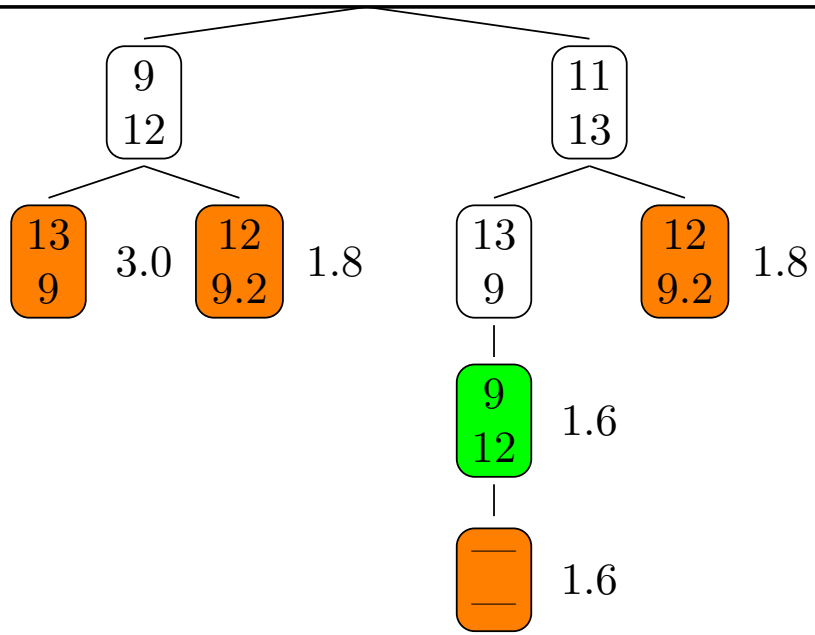
Proline Checking

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

Reference Protein Chain

Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2

Nodes

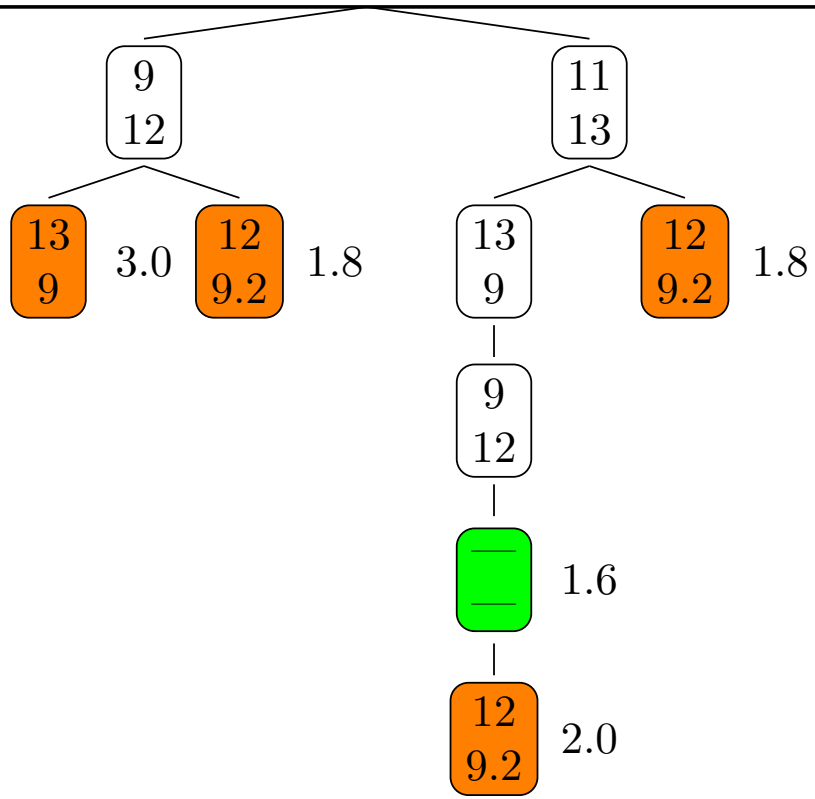


Proline Checking

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

Reference Protein Chain	
Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2

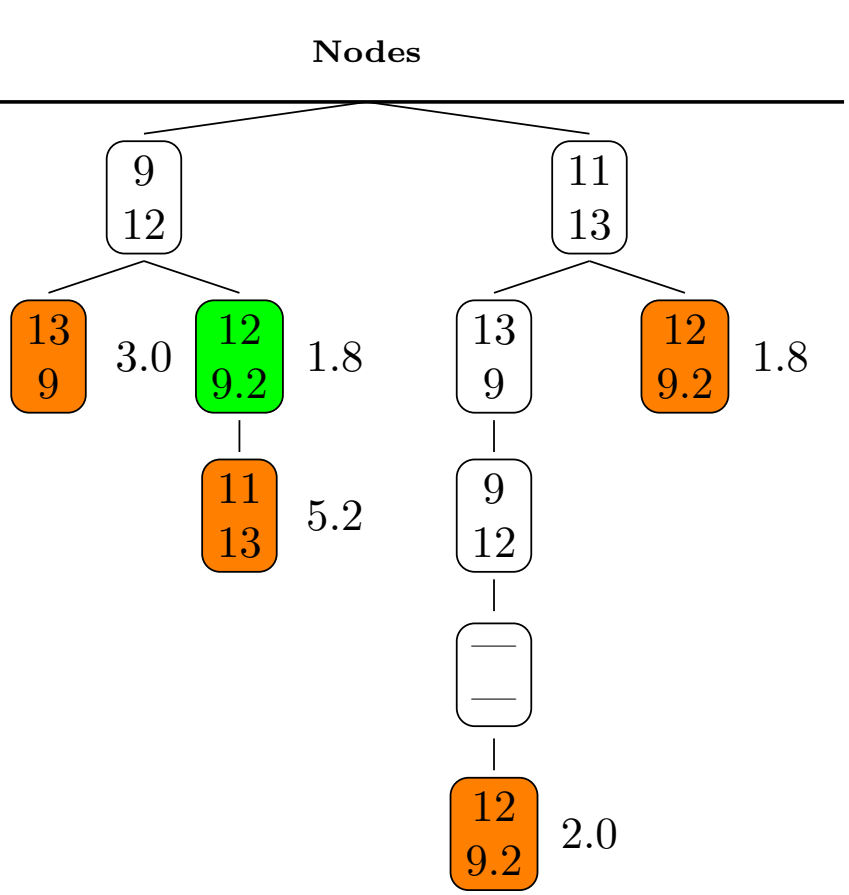
Nodes



Node Generation

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

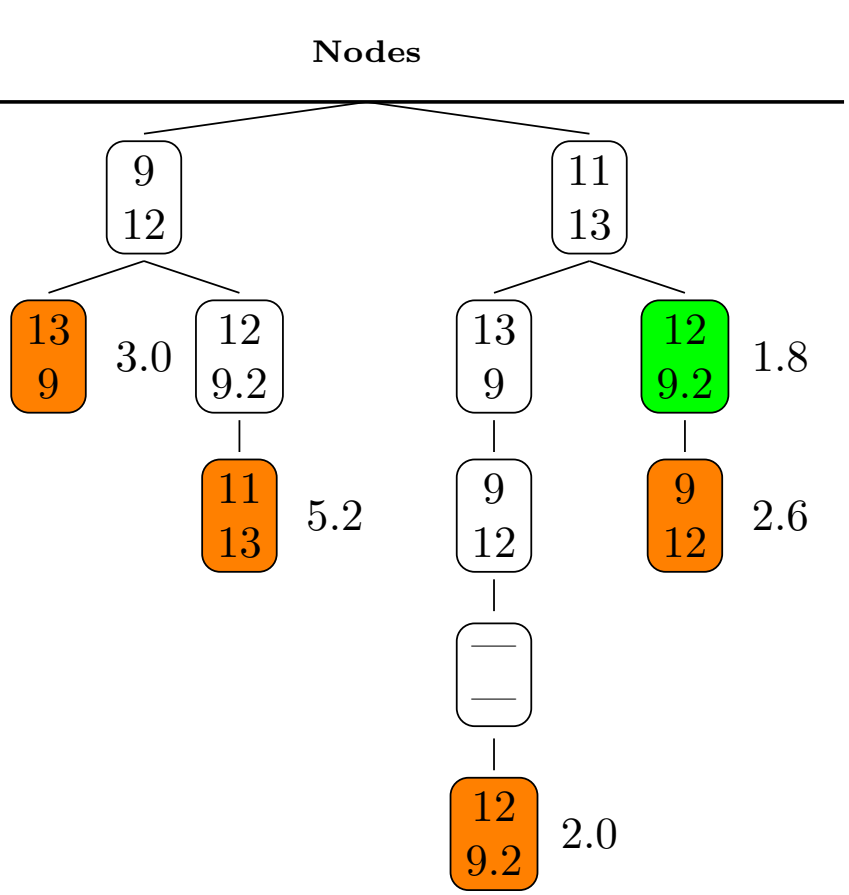
Reference Protein Chain	
Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2



Node Generation

Tiles to assign:	$i-1$	9	11	13	12	–
	i	12	13	9	9.2	–
	P_1	0.935	0.013	0.003	0.000	0.000
	P_2	0.002	0.001	0.04	0.403	0.000
	Proline	No	No	No	No	Yes

Reference Protein Chain	
Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2



Finished Assignment

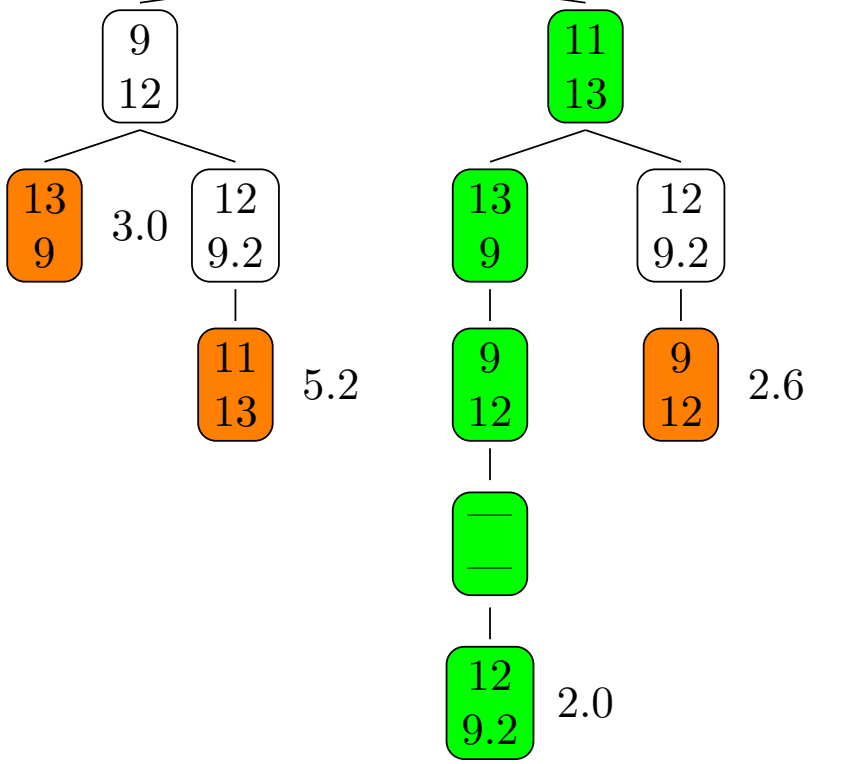
Tiles to assign:

$i-1$	9	11	13	12	–
i	12	13	9	9.2	–
P_1	0.935	0.013	0.003	0.000	0.000
P_2	0.002	0.001	0.04	0.403	0.000
Proline	No	No	No	No	Yes

Reference Protein Chain

Chemical Shift	P_n
13.5	1
9.5	2
11.4	1
-	Proline
8.8	2

Nodes



Algorithm Performance

Correct Assignments

- Assigned a protein sequence of length 62 in approximately 40 minutes
- Major progress in time of assignment

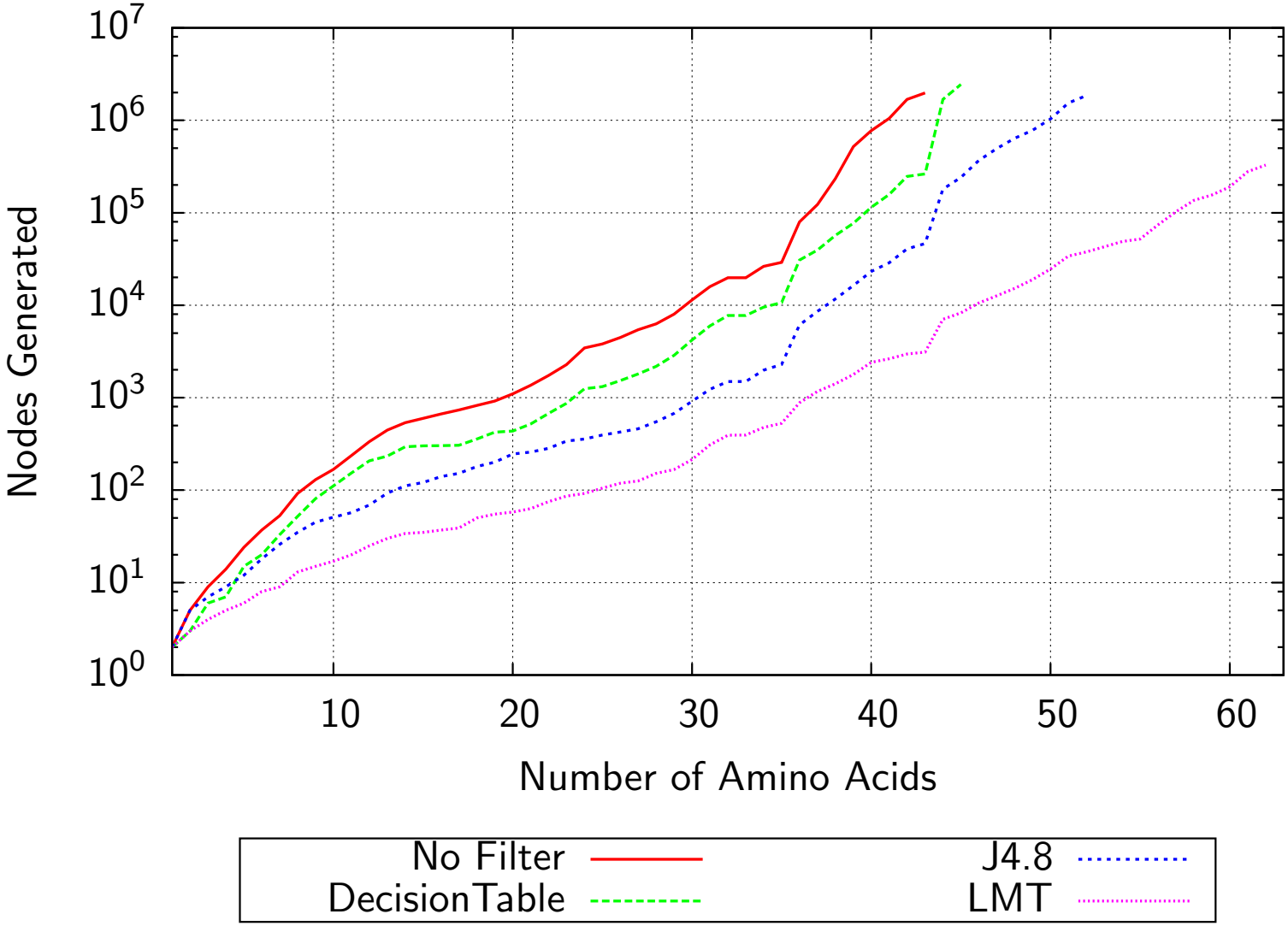
Protein used

- C-terminal domain of the Tfg1 subunit of the yeast transcription factor TFIIF[7]

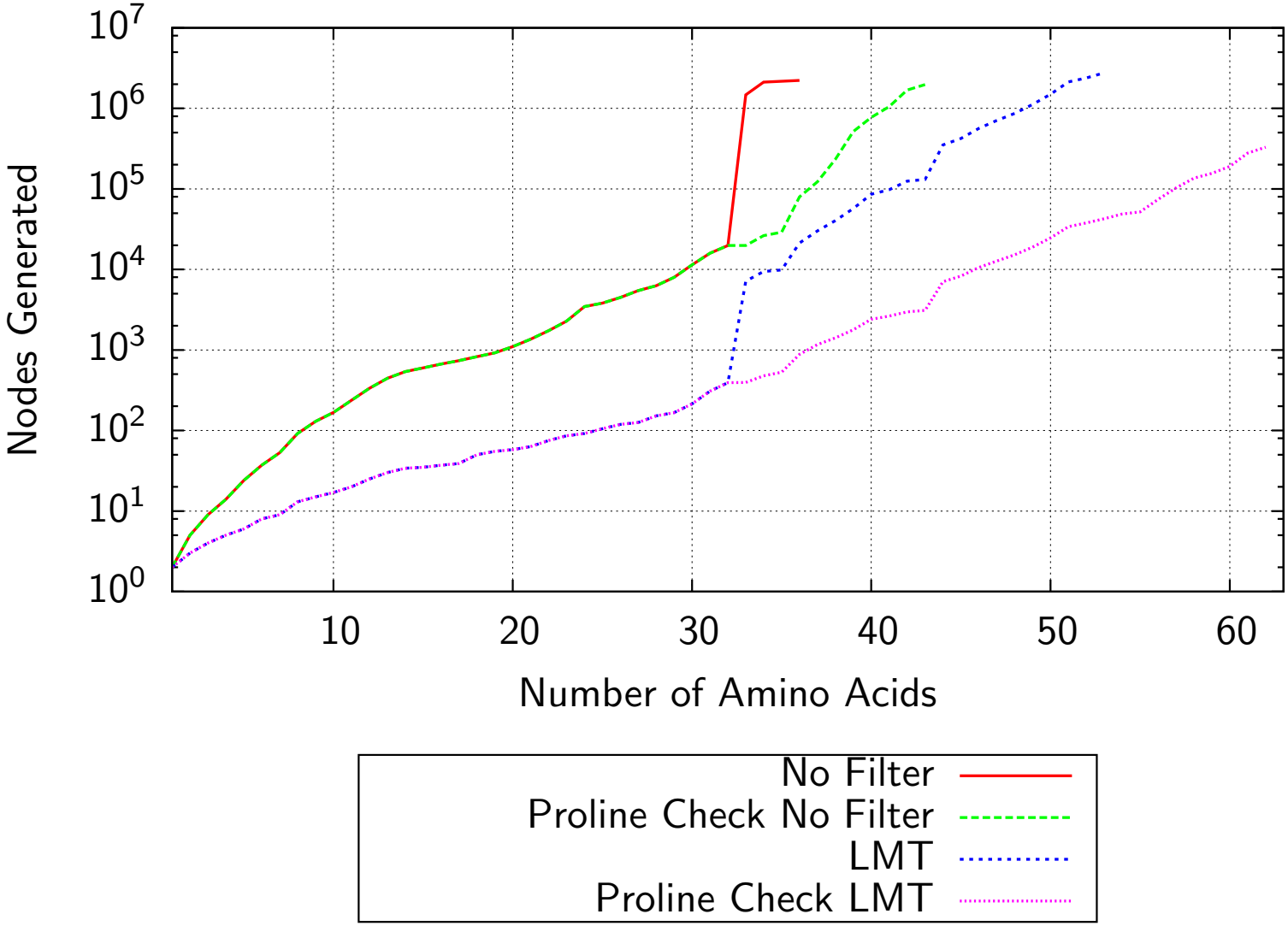
7. Adina M. Kilpatrick et al., "Structural and binding studies of the C-terminal domains of yeast TFIIF subunits Tfg1 and Tfg2," *Proteins: Structure, Function, and Bioinformatics* 80, no. 2 (2012): 519–529, doi:10.1002/prot.23217, http://dx.doi.org/10.1002/prot.23217.

J. Venzke
D. Mascharka
P. Johnson
R. Davis
K. Roth
L. Robison
T. Urness
A. Kilpatrick

Machine Learning Algorithms



Proline Checking



Future Research

Extend the Proline checking to other amino acids

Include a heuristic for assignment cost prediction

Assign subsets and combine to generate full assignments

Acknowledgments

John Emmons

Mentors

- Adina Kilpatrick
- Timothy Urness

Research team

- David Mascharka
- Paxten Johnson
- Rachel Davis
- Katherine Roth
- Leah Robison

Drake University

DUCURS

References I

Kilpatrick, Adina M., Leonardus M.I. Koharudin, Guillermo A. Calero, and Angela M. Gronenborn. "Structural and binding studies of the C-terminal domains of yeast TFIIF subunits Tfg1 and Tfg2." *Proteins: Structure, Function, and Bioinformatics* 80, no. 2 (2012): 519–529. doi:10.1002/prot.23217. http://dx.doi.org/10.1002/prot.23217.

Kohavi, Ron. "The Power of Decision Tables." In *8th European Conference on Machine Learning*, 174–189. Springer, 1995.

Landwehr, Niels, Mark Hall, and Eibe Frank. "Logistic Model Trees." *Machine Learning* 95, nos. 1-2 (2005): 161–205.

References II

Linge, Jens P., Michael Habeck, Wolfgang Rieping, and Michael Nilges. "ARIA: Automated NOE assignment and NMR structure calculation." *Bioinformatics*:2003.

Quinlan, Ross. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

Ulrich, Eldon L., Hideo Akutsu, Jurgen F. Doreleijers, Yoko Harano, Yannis E. Ioannidis, Jundong Lin, Miron Livny, et al. "BioMagResBank." *Nucleic Acids Research* 36, no. Database-Issue (January 23, 2008): 402–408. `http://dblp.uni-trier.de/db/journals/nar/nar36.html#UlrichADHILLMMMNSTWYM08`.

Thank You

Questions?