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

Background

Machin Learnin

Algorith

Overview Model Training Preprocessing The Search

Results

Outlook

References

Utilizing Machine Learning to Accelerate Automated Assignment of Backbone NMR Data

Joel Venzke¹², David Mascharka¹, Paxten Johnson¹², Rachel Davis¹, Katherine Roth¹, Leah Robison¹, Timothy Urness¹ and Adina Kilpatrick²

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

> > joel.venzke@drake.edu

April 16, 2015

NMR

Assignment with Machine Learning

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

Background

Machin Learnin

Algorith

Overview Model Training Preprocessing

Results

Outlook

Reference

Overview

Background

Nuclear Magnetic Resonance (NMR) Machine Learning

2 Algorithm

Overview Model Training Preprocessing The Search

3 Results

4 Outlook

NMR Assignment

with Machine Learning

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

NMR

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 manually¹

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 Assignme

Assignment with Machine Learning

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

Background

Machin

Overview Model Trainin

Model Training Preprocessing The Search

Results

Outloo

Reference

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 experiment²

- Generates C_{α} and C_{β} for residue i
- Confirms residue data

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

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

Background NMR Machine

Learning

Algorith

Overview Model Trainin Preprocessing The Search

Results

Outloo

Reference:

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

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

A. Kilpatrick

NMR Machine

Learning

Overview Model Training

Results

Outloo

Reference

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 Table⁵

- 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.

J. Venzke D. Mascharka P. Johnson R. Davis K. Roth L. Robison

T. Urness

A. Kilpatrick Background

NMR

Machine Learning

Δlgorith

Overview

Model Training Preprocessing

. .

Outloo

References

Algorithmic Overview

Model Training

Preprocessing

The Search

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

Background

Machin Learnin

Algorith

Overview Model Training

Preprocessing

Results

Outloo

Reference

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.

J. Venzke

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

Background

Machin

Algorith

Overview

Model Training

Preprocessing

Results

Outloo

Reference

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}

NMR Assignment

with Machine Learning

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

Preprocessing

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 $P_1, P_2, \cdots, P_{19}, P_{20}$ NMR

Assignment with Machine Learning

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

Background NMR

Machine Learning

Algorith

Overview Model Trainin

Preprocessing

Results

Outlool

References

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

 $1.0, 1.0, \dots, 1.0, 1.0$ **Proline**

yes/no

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

A. Kilpatrick

NMR Machine

Learning

Overview

Model Training

The Search

First Tile

Tiles to assign:

i-1
i
P_1
P_2
Proline

9	\bigcap
12	
0.935	(
0.002	(
No	

11		13
13		9
0.013		0.003
0.001		0.04
No	l	No

12
9.2
0.000
0.403
No
$\overline{}$

10

_
_
0.000
0.000
Yes

Refer Protein		Nodes
Chemical Shift	P_n	
13.5	1	
9.5	2	
11.4	1	
-	Proline	
8.8	2	

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

A. Kilpatrick

NMR Machine Learning

Overview Model Training

The Search

First Tile

i-111 12 13 Tiles to assign: 0.9350.013 0.0020.001Proline No

9 No

13 9 0.003 0.04No

9.2 0.0000.403No

12

0.0000.000Yes

	,	(110 (110) (110)		
Refer Protein		Nodes	Nodes	
Chemical Shift	P_n			
13.5	1	12	11 13	
9.5	2			
11.4	1			
-	Proline			
8.8	2			

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

Background

Machine

Algorith

Overview Model Training

Preprocessing
The Search

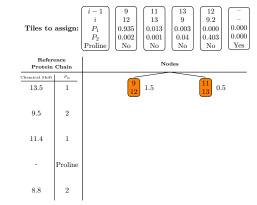
Recults

Outloo

Reference

Cost Calculation

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



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

A. Kilpatrick

Background NMR Machine

Learning

Algorith

Overview Model Training Preprocessing

The Search

Results

Outioo

References

Node Generation

Refer Protein			Nodes
Chemical Shift	P_n		
13.5	1	9 1.5	$\begin{bmatrix} 11 \\ 13 \end{bmatrix} 0.5$
9.5	2		
11.4	1		
-	Proline		
8.8	2		

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

A. Kilpatrick

Machine Learning

Overview Model Training

The Search

Node Generation

i-112 13 Tiles to assign: 0.9350.013 0.0020.001Proline

11 9 No No

13 9 0.003 0.04No

12 9.2 0.0000.403No

0.0000.000Yes

	,			
Refere Protein		Nodes		es
Chemical Shift	P_n			
13.5	1	9	1.5	$\begin{array}{c} 11 \\ 13 \\ \end{array}$ 0.5
9.5	2			$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$
11.4	1			
-	Proline			
8.8	2			

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

A. Kilpatrick

Background NMR Machine

Learning

Overview

Model Training Preprocessing The Search

_ .

Outloo

Reference

Node Generation

Refere Protein		Nodes	
Chemical Shift	P_n		
13.5	1	$\frac{9}{12}$ 1.5	$\underbrace{\begin{bmatrix} 11 \\ 13 \end{bmatrix}}_{}$
9.5	2		$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$
11.4	1		$\frac{9}{12}$ 1.6
-	Proline		
8.8	2		

J. Venzke D. Mascharka P. Johnson R. Davis K. Roth L. Robison

T. Urness A. Kilpatrick

Background NMR

Machine Learning

Algorithm

Model Training
Preprocessing

The Search

Results

Reference

Node Generation

	rence n Chain	Nodes
Chemical Shift	P_n	
13.5	1	$ \begin{array}{c c} 9 \\ 12 \end{array} 1.5 \qquad \begin{array}{c} 11 \\ 13 \end{array} $
9.5	2	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$
11.4	1	9 1.6
-	Proline	
8.8	2	

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

A. Kilpatrick

Background NMR

Machine Learning

Overview

Model Training
Preprocessing

The Search

Outloo

Reference

Proline Checking

Reference Protein Chain		Nodes	
Chemical Shift	P_n		
13.5	1	$ \begin{array}{ c c } \hline 9\\12\\ \hline \end{array} $	$ \begin{array}{c} 11 \\ 13 \end{array} $
9.5	2	$ \begin{array}{c c} 13 \\ 9 \end{array} 3.0 \begin{array}{c} 12 \\ 9.2 \end{array} 1.8 $	$ \begin{array}{c c} \hline 13 \\ 9 \\ \hline \end{array} $ 12 1.8
11.4	1		9 1.6
-	Proline		1.6
8.8	2		

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

A. Kilpatrick

Background NMR Machine

Learning

Overview

Model Training
Preprocessing

The Search

Outloo

Reference

Proline Checking

Reference Protein Chain		Nodes	
Chemical Shift	P_n		
13.5	1	$ \begin{array}{ c c } \hline 9\\12\\ \hline \end{array} $	$\underbrace{\begin{bmatrix} 11 \\ 13 \end{bmatrix}}$
9.5	2	$ \begin{array}{c c} \hline 13\\ 9 \end{array} 3.0 \begin{array}{c} 12\\ 9.2 \end{array} 1.8 $	$ \begin{array}{c c} \hline 13\\ 9\\ \hline 9.2 \end{array} $ 1.8
11.4	1		$\begin{pmatrix} 9\\12 \end{pmatrix}$
-	Proline		1.6
8.8	2		$\begin{bmatrix} 12 \\ 9.2 \end{bmatrix} 2.0$

J. Venzke D. Mascharka P. Johnson R. Davis K. Roth L. Robison

T. Urness A. Kilpatrick

Background NMR

Machine Learning

Algorith

Overview Model Training Preprocessing

The Search

Reference

Node Generation

Reference Protein Chain		Nodes	
Chemical Shift	P_n		
13.5	1	9 12	$ \begin{array}{c} 11 \\ 13 \end{array} $
9.5	2	$ \begin{array}{c c} \hline 13\\ 9 \end{array} 3.0 \begin{array}{c} 12\\ 9.2\\ \end{array} 1.8 $	$ \begin{array}{c c} \hline 13\\ 9\\ \hline 9.2 \end{array} $ 1.8
11.4	1	11 5.2	9
-	Proline		
8.8	2		$\begin{bmatrix} 12 \\ 9.2 \end{bmatrix} 2.0$

J. Venzke D. Mascharka P. Johnson R. Davis K. Roth L. Robison

T. Urness A. Kilpatrick

Background

Machine Learning

Algorith

Overview Model Training Preprocessing

The Search

0....

Reference

Node Generation

Reference Protein Chain		Nodes	
Chemical Shift	P_n		
13.5	1	$\begin{bmatrix} 9 \\ 12 \end{bmatrix}$	$\underbrace{\begin{bmatrix} 11 \\ 13 \end{bmatrix}}$
9.5	2	$ \begin{array}{c c} \hline 13\\9 \end{array} 3.0 \begin{array}{c} 12\\9.2 \end{array} $	$ \begin{array}{c c} \hline 13 \\ 9 \\ \hline \end{array} $ 1.8
11.4	1	11 5.2	$ \begin{array}{c} 9 \\ 12 \end{array} $ 2.6
-	Proline		
8.8	2		$\begin{bmatrix} 12 \\ 9.2 \end{bmatrix} 2.0$

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

A. Kilpatrick

Background NMR Machine

Learning

Overview

Model Training

The Search

0....

Reference

Finished Assignment

Reference Protein Chain		Nodes	
Chemical Shift	P_n		
13.5	1	$\underbrace{\begin{bmatrix} 9 \\ 12 \end{bmatrix}}$	11 13
9.5	2	$ \begin{array}{c c} \hline 13\\9 \end{array} 3.0 \begin{array}{c} 12\\9.2 \end{array} $	13 12 9.2
11.4	1	$\begin{bmatrix} 11 \\ 13 \end{bmatrix} 5.2$	9 9 12 2.6
-	Proline		
8.8	2		$\begin{bmatrix} 12 \\ 9.2 \end{bmatrix} 2.0$

NMR

Assignment with Machine Learning

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

Background

Machin

Algorith

Overview Model Trainin

Preprocessing The Search

Results

Outloo

References

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 TFIIF⁷

^{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

Background

Learning

Algorith

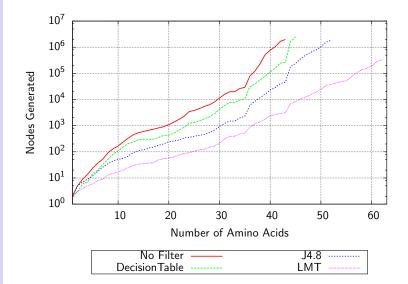
Overview Model Training Preprocessing

Results

Outloo

References

Machine Learning Algorithms



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

Background NMR

Machine Learning

Algorith

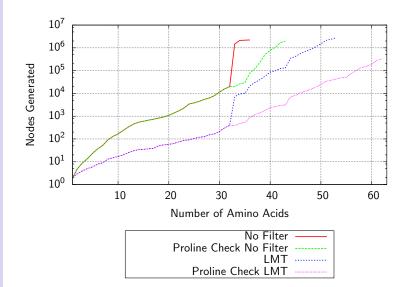
Overview Model Trainin Preprocessing

Results

Outloo

References

Proline Checking



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

A. Kilpatrick

Background

Machin

Learning

Algorith

Overview Model Training Preprocessing

Results

Outlook

References

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

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

A. Kilpatrick

Background

Machi

Algorith

Overview Model Training Preprocessing

Results

Outloo

References

Acknowledgments

John Emmons

Mentors

Adina Kilpatrick

Timothy Urness

Research team

- David Mascharka
- Paxten Johnson
- Rachel Davis

Katherine Roth

Leah Robison

Drake University

DUCURS

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

Background

Machin Learnin

Algorith

Overview Model Trainin Preprocessing

Results

Outlool

References

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.

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

Background NMR

A Laurentinia

Overview Model Training Preprocessing

Result

Outlook

References

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#UlrichADHILLMMNSTWYM08.

NMR Assignment

Assignment with Machine Learning

J. Venzke D. Mascharka P. Johnson R. Davis K. Roth

L. Robison T. Urness A. Kilpatrick

NMR

Machine Learning

Algorith

Overview Model Training

The Sea

Results

Outloo

References

Thank You

Questions?