NMR Assignment with Machine

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

K. Roth L. Robison T. Urness A. Kilpatrick

Background

Machin

Algorith

Overview Model Training Preprocessing

Results

Outlook

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

Algorith

Overview Model Training Preprocessing

Results

Outlook

Overview

Background
 NMR
 Machine Learning

2 Algorithm

Overview
Model Training
Preprocessing
The Search

- 3 Results
- 4 Outlook

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

A. Kilpatrick

Background

NMR

Machine Learning

Algorit

Overview

Model Training
Preprocessing
The Search

Results

Outlook

Harvard University Conference

- 1898
- Say some things

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

Background

Machine

Learning

Algorit

Algorit

Overview
Model Training
Preprocessing
The Search

Result

Outlook

Harvard University Conference

- 1898
- Say some things

I Venzke

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

Background

Machin

Algorith

Overview

Model Trainin

Results

Outloo

Algorithmic Overview

Model Training

- Preformed once during algorithm development
- Provides model used in Preprocessing

Preprocessing

- Imports NMR data set
- Filters NMR data using machine learning model

The Search

- Uses results from Preprocessing
- Assigns NMR data set
- Records results

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

Model Training

Training Data Set

Biological Magnetic Resonance Bank (BMRB)

- 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

Training the Model

Preformed Once

- Time consuming task
- Trained once, used many times

Models Trained

DecisionTable, j.48, LMT

Background NMR Machine

Algorith

Overview Model Training Preprocessing

Results

Outloo

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 $C_{lpha},\,C_{eta}$ Residue i $C_{lpha},\,C_{eta}$ Confidence Levels $P_1,\,P_2,\,\cdots,\,P_{19},\,P_{20}$

NMR Assignment with Machine

J. Venzke D. Mascharka P. Johnson

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

Backgroun

Machine Learning

Algorith

Overview Model Trainin

Preprocessing The Search

Results

Outlool

Confidence Level Calculation

Tile

Machine Learning

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

NMR

Learning

Overview

Overview Model Trainin Preprocessing

The Search

Results

Outlool

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

Backgroun

Machine

Learning

Algorith

Overview Model Training Preprocessing

The Search

Results

Outlook

Harvard University Conference

- 1898
- Say some things

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

Background

Machine Learning

Algorith

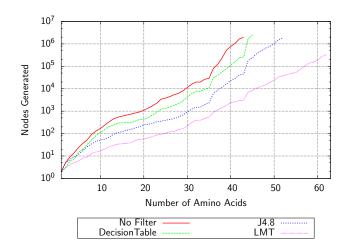
Overview

Model Trainin
Preprocessing

Results

Outlook

Machine Learning Algorithms



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

Background NMR Machine

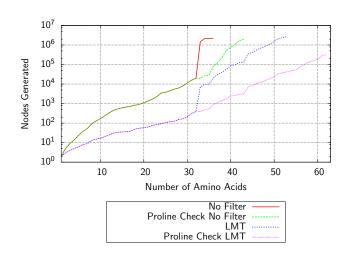
Learning

Overview Model Trainin Preprocessing

Results

Outlook

Proline Checking



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

A. Kilpatrick

Background

Machine Learning

Algorith

Aigoriti

Overview Model Trainin Preprocessing

Results

Outlook

Future Research

Extend the Proline checking to other amino acids

Include a hysteric for assignment cost prediction

Assign subsets and combine to generate full assignments

NMR Assignment

with Machine Learning

J. Venzke D. Mascharka

P. Johnson

R. Davis

K. Roth

L. Robison T. Urness

A. Kilpatrick

Backgrou

NMR

Machine

Learning

Algorithr

Overview Model Training

Preprocessing

The Searc

Results

Outlook

Thank You