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

Harvard University Conference

- 1898
- Say some things

Harvard University Conference

- 1898
- Say some things

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

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

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

Confidence Level Calculation

Machine Learning

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

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

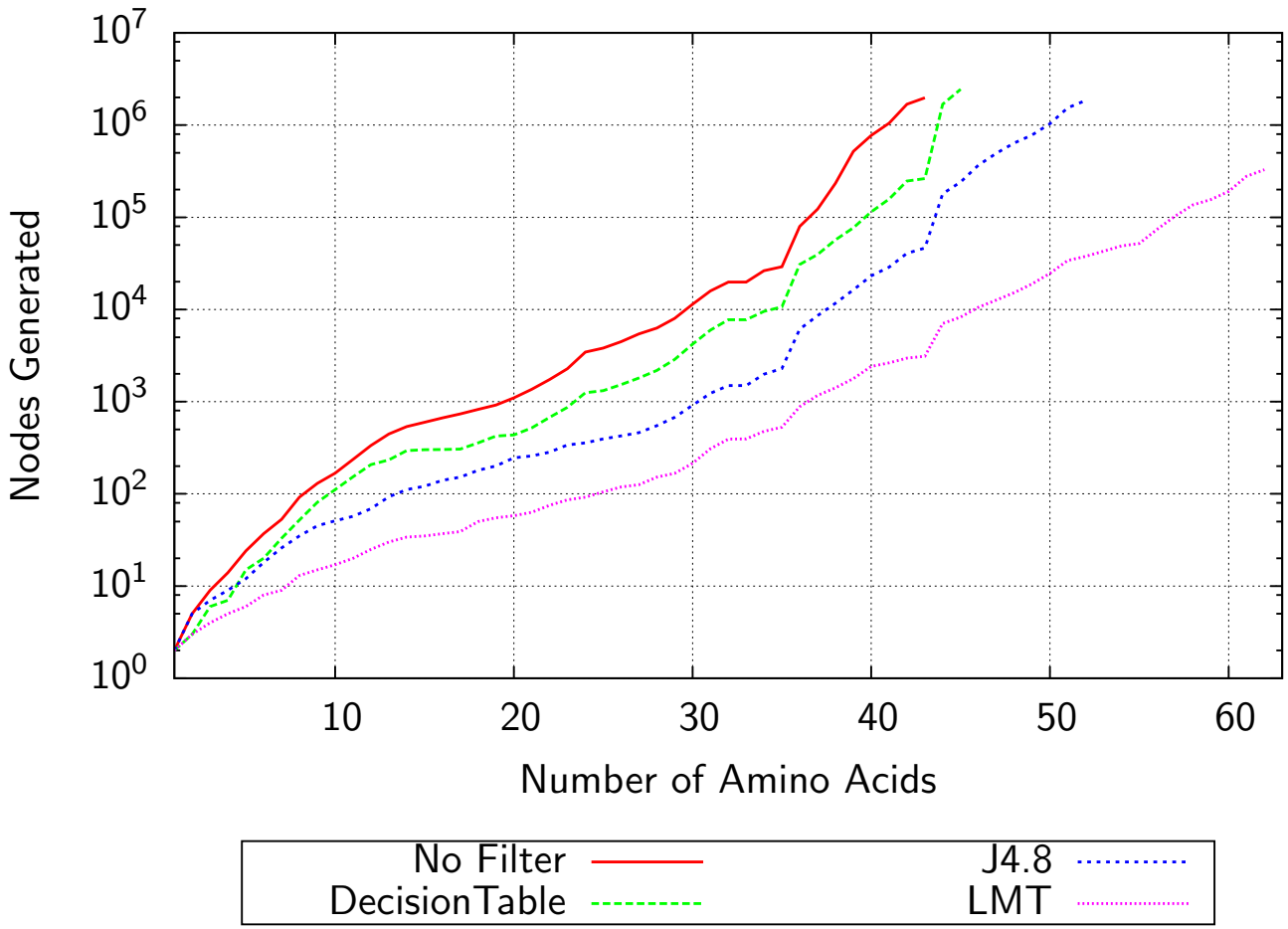
Proline

yes/no

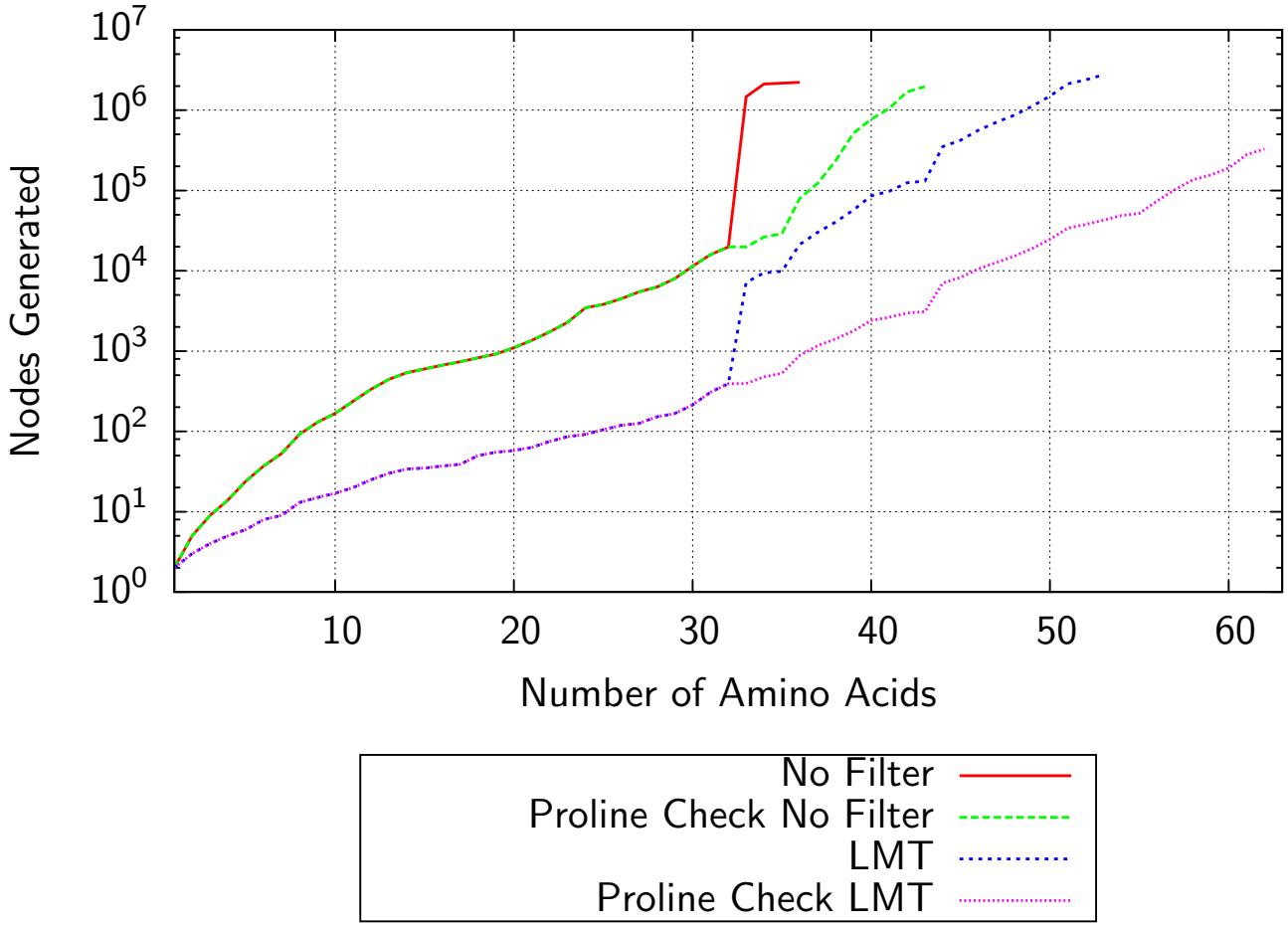
Harvard University Conference

- 1898
- Say some things

Machine Learning Algorithms



Proline Checking



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

Thank You