Protein Backbone Angle Prediction with Machine Learning Approaches

by R Kang, C Leslie, & A Yang

in

Bioinformatics, 1 July 2004, vol 20 nbr 10
pp 1612-1621

with additional background information
from Brad Morantz

Presented by
Brad Morantz PhD
15 November 2004

Index

- Introduction
- Brief background of authors
- Why is this paper important? (So what!)
- Introduction and background
- Quick overview of artificial neural networks
- Quick overview of support vector machines
- Hypothesis
- Methodology
- Results
- Conclusions
- Additional references

Brief Background of Authors

- Kuang is in Computer Science
- Leslie is from Pharmacology
- Yang is in Pharmacology, Columbia Genome Center, and Computational Biology and Bioinformatics
- Research Funded by NIH and PhRMA

Importance of this Paper

- Protein backbone torsion angle provides more information than alpha, beta, & coil (conventional 3 state predictions)
- More information will contribute to better modeling of local structure of protein sequence segments
- Structure and function are highly correlated
- Hence, will allow better prediction of protein function

Introduction & Background

- Protein backbone torsion angles are highly correlated to protein secondary structures
- Loop residues in protein chain structurally determine regular secondary structure elements which leads to specific protein folding topology
- Involve enzymatic activities and protein to protein interactions such as antibody and antigen

The Key Concept

- The analysis of protein sequence-structure function relationship is facilitated significantly by local structure information from predictive algorithms

More Background

- Conformational variability is high, causing problems in molecular modeling.
- Three state (alpha, beta, & coil) structure modeling do not distinguish loop structure
- Backbone torsion angle modeling helps in modeling loop regions
- Little attention has been paid to this area.

Literature Review

- DeBrevern et al (2000): study of predictability
- Bystroff et al (2000): first backbone torsion angle work using HMM
- Karchin (2003): fold recognition, not prediction
- Yang & Wang (2003): database prediction using RMS residual

Overview of ANNs
(Artificial Neural Networks)

- In Supervised mode
  - Data driven general function approximator
  - Teacher shows it examples and what each one is. The ANN is trained on this information.
  - Now show it something and it will tell you what it is.
  - It also functions like regression, only not constrained to linear functions
  - Black box performance
  - No understanding available from internal parameters

Non Linear Regression

Training	
1, 2, 4	21
3, 5, 2	38
4, 4, 6	68
5, 7, 7	99
7, 7, 7	99.3
10, 10, 2	99.6

Testing	
1, 2, 5	30
4, 4, 4	48
5, 6, 4	77
8, 8, 8	99.4
9, 9, 6	99.5

Note: not tested on same as training

Concept: sum of squares of inputs but total never quite reaches 100

Pattern Recognition

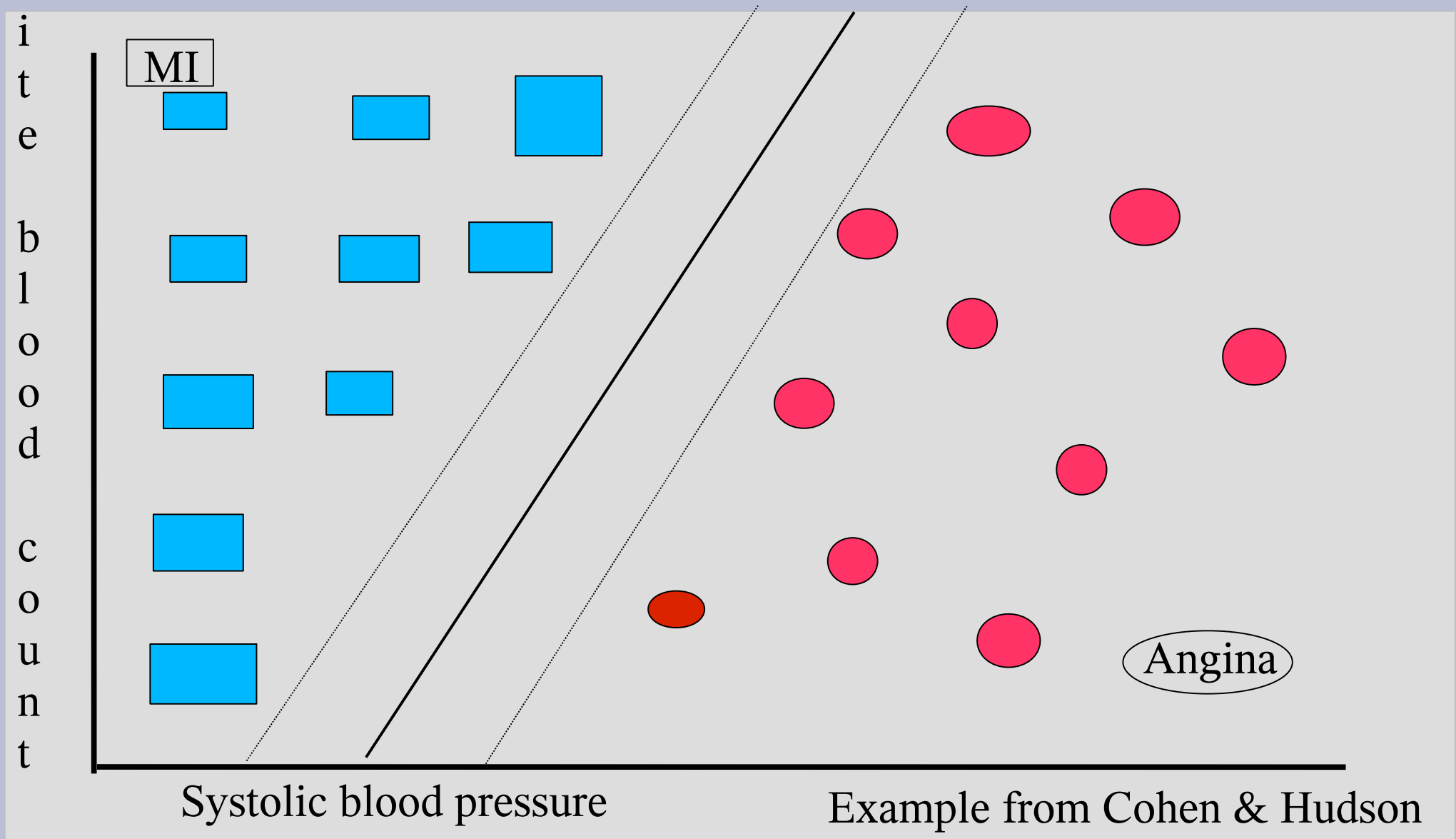
Apple

Flower

Support Vector Machine

- Linear Separator
- Draws line to divide groups
- Attempts to maximize width of line
  - Create fences to maximize separation
- Trade off is trying to maximize width but then there are more violations
- Concept expanded for N space
- PSVM has polynomial kernel
  - Allows curved line to separate

Example SVM


White blood count
MI
Angina
Systolic blood pressure
Example from Cohen & Hudson

Goal

- Predict backbone conformational state of each residue in protein chains
- Based on 4 (A, B, G, E) or 3 (A, B, G/E) conformational states

Hypothesis

- The two methods (SVM & ANN) are more accurate at predicting backbone torsion angle than previously published methods

Methodology

- Protein backbone torsion angles mapped onto Φ-Ψ plot.
- Divide the map into the 4 conformational states (A, B, G, & E)
- Use PSSM (position specific score matrix)
- Use nine-sequence segments in non-redundant protein structures

The ANN Predictor

- 216 input nodes
  - 9 groups of 24 (1 group for each residue)
    - Categorical inputs of the 20 amino acids
    - Flag for residue position outside the C or N terminus of the protein chain
    - 3 for backbone torsion angle prediction from LSBPS1 database
- 50 nodes in hidden layer
- 3 output nodes
  - A, B, & G/E
    - E has only 1.7% of training cases so grouped with G

Methodology Cont'd

- Output from ANN converted to PSSM (position specific score matrix) by using long formula
- ANN trained on line
  - I suspect back propagation
  - Not related to the testing set
  - Terminated training when accuracy attenuated
  - Wrong way to do it!
  - Indicative of other mistakes
- 10 fold Jack-knife cross validation process
  - 10 runs, each with a different 10% testing set
  - Average of runs used for predictive accuracy

Accuracy Calculation

- Compare true backbone conformational state with predicted
- Predicted by using the output node with the largest value
- Then run trained ANN on entirely new set of proteins
- Cross validation produced average accuracy of 78.2%

Comparing Input Contribution

- Amino acids only gave 61.5%
- Torsion angle prediction from database gave 67.8%
- Both together gave 78.2%
- Not surprising
  - Standard method as more information supplied then better result expected

SVM Prediction

- Classification of 3 or 4 conformational states
- Inputs:
  - Amino acid sequences
  - Profiles from PSI Blast
  - secondary structures from PSI-Pred
- Input vector 189 dimensions
  - 9 protein sequences
  - 21 to code the categorical data for each
- Choose prediction to be class that gives the biggest margin for each example
- Use publicly available SVM light package

Testing the SVM

- Dunbrack-culled PDB dataset
  - Benchmark testing to compare to literature
- LSBSP1 dataset
  - 10 fold cross validation
- Results about the same for both methods
- Dunbrack-in scop dataset
  - 3 fold cross validation to match other test

SVM

- Binary mapping worse
- Profile mapping 6% better
- Secondary feature mapping 3% better
- Profile and secondary good in alpha and beta regions, but less in the loops
- No mention of repeatability (precision), ANOVA, or t tests (which means these close margins prove absolutely nothing)
- Polynomial kernel tried but only 1% improvement (see above note)

ANN and LSBSP1

- 81.5% of A Backbone correctly predicted
- 76.6% of B backbone correctly predicted
- 46.5% of G/E backbone correctly predicted
- 77% correct overall
- Large enough sample size to be valid test Results

SVM vs ANN

- SVM better than ANN
  - On all residues 78.7% vs 78.2%
  - On loop residues 65.1% vs 63.5%
- No proper statistics supplied in support of this conclusion
  - 1% difference would easily be statistically insignificant with a large enough variance

Conclusions

- Optimally combining information to improve prediction is standard in decision science, but is “a difficult challenge in knowledge-based protein structure prediction procedures”
- Nearing limit of prediction accuracy of protein structures.
- Notice that they did not address the original hypothesis, nor do statistical testing to compare against other published work

Additional References

- IEEE Computational Intelligence Society
  - www.cis-ieee.org
- Author's web page
  - www.cs.columbia.edu/compbio/backbone
  - www.columbia.edu/~ayl
- Brad's web page
  - www.machine-cognition.com
- AAAI- American Association for Artificial Intelligence
  - www.aaai.org