Protein secondary structure prediction with machine learning methods

Daniel S. Standage

BCB 569

November 14, 2011

Project focus

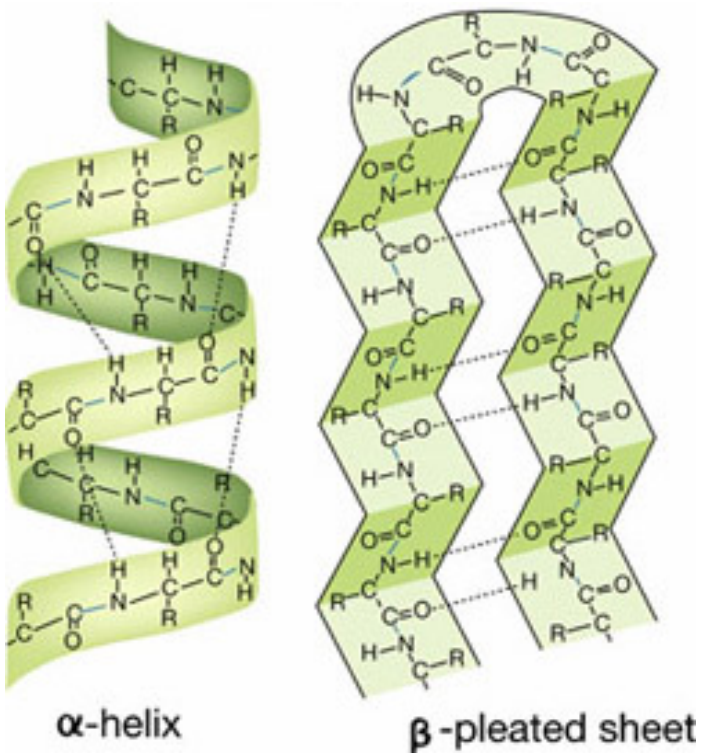
Problem

Protein secondary structure prediction

Approach

Machine learning methods

Secondary structure



Secondary structure

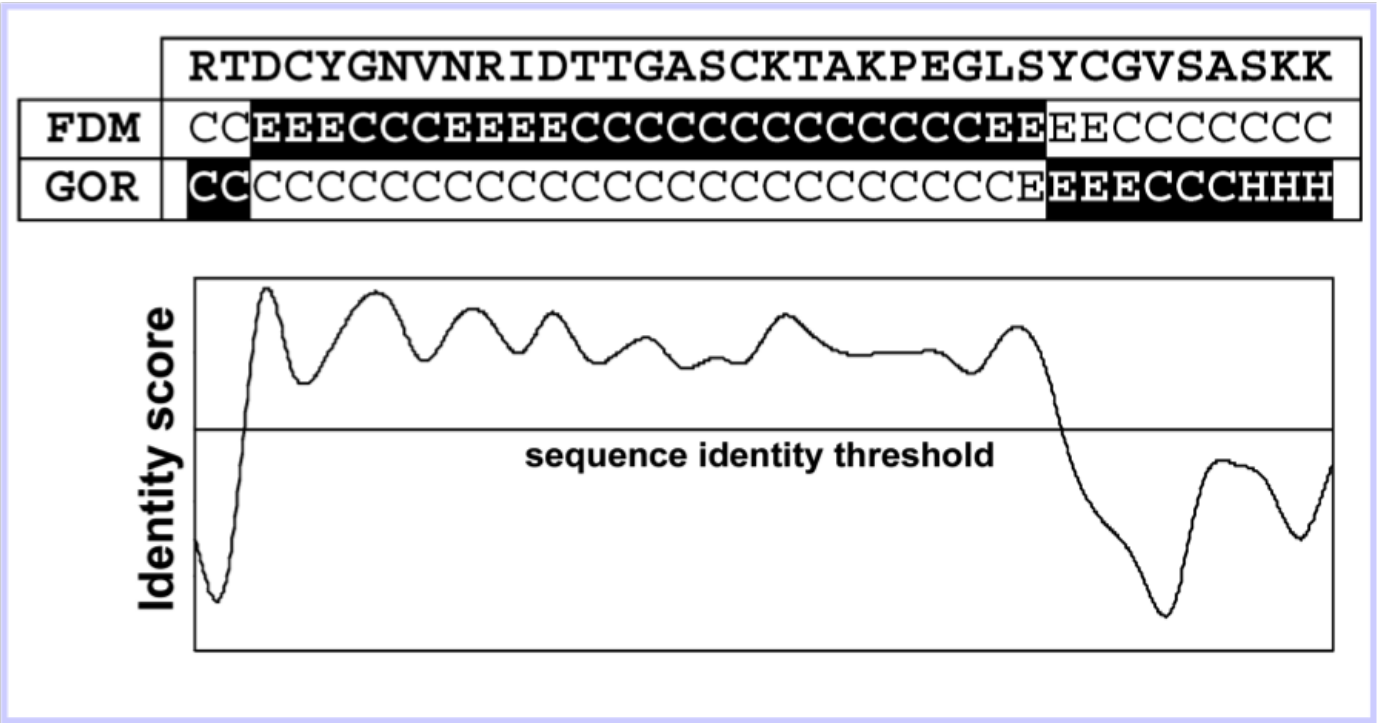
```
  1 GSHSMRYFYT AMSRPGRGEP RFIAVGYVDD TQFVRFDSDA ASPRTEPRPP
      EEEEEEEE EE  TTSSS  EEEEEEEETT EE EEEETTT TT   EE SG

 51 WIEQEGPEYW DRNTQIFKTN TQTYRENLRI ALRYYNQSEA GSHIIQRMYG
    GGTTTTHHHH HHHHHHHHHH HHHHHHHHHH HHHHTT  TT S  EEEEEEE

101 CDLGPDGRLL RGHDQSAYDG KDYIALNEDL SSWTAADTAA QITQRKWEAA
    EEEETTTEE  EEEEEEEETT EE EEE TTS S EEESSHHH HHHHHHHHTT

151 RVAEQLRAYL EGLCVEWLRR YLENGKETLQ RADPPKTHVT HHPVSDHEAT
    THHHHHHHHH HTHHHHHHHH HHHHTHHHHT   B  EEEEE EEE SSSEEE

201 LRCWALGFYP AEITLTWQRD GEDQTQDTEL VETRPAGDRT FQKWAAVVVP
    EEEEEESEES S  EEEEEET TEE TTT EE    EE SSS  EEEEEEEEEE

251 SGEEQRYTCH VQHEGLPKPL TLRWEPHH
    TTTGGGEEEE EE TTSSS E EE
```

Statistical methods

- Chou-Fasman
- Garnier, Osguthorpe, Robson (GOR)
- Consensus data mining (CDM)

Statistical methods

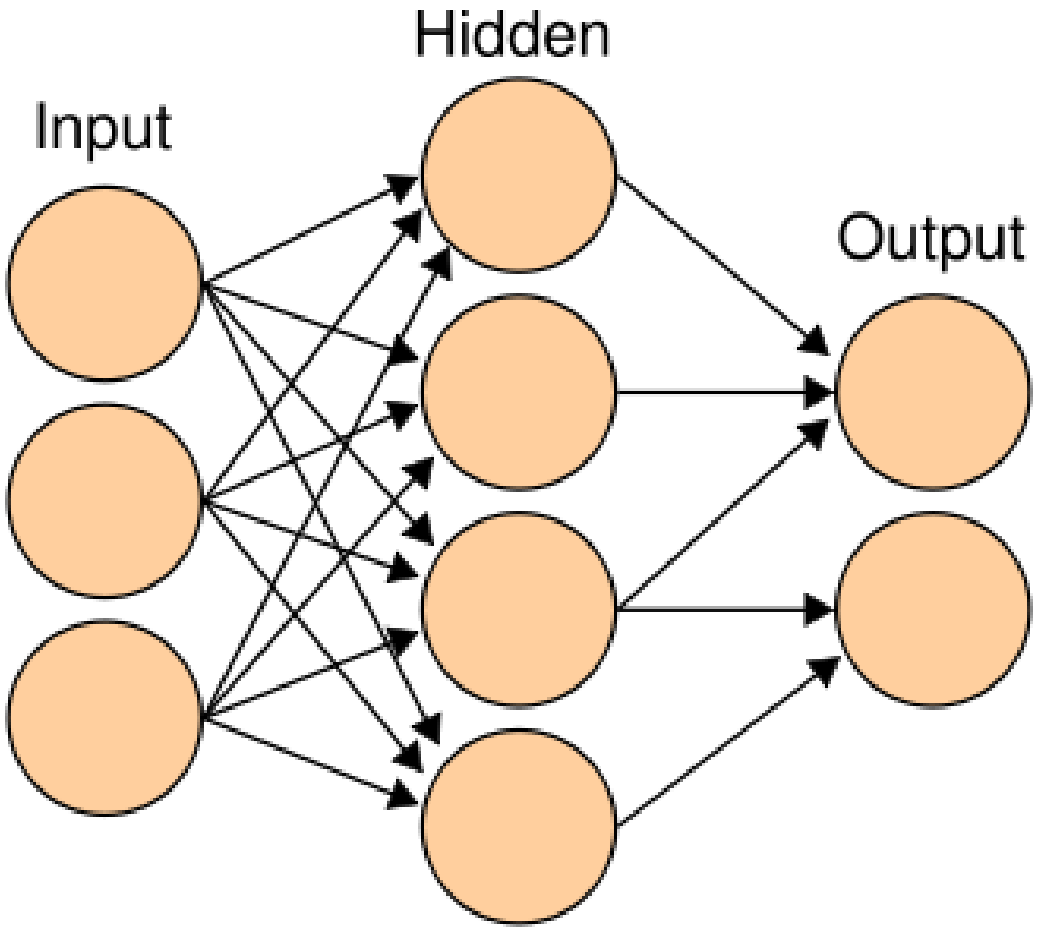
Consensus data mining



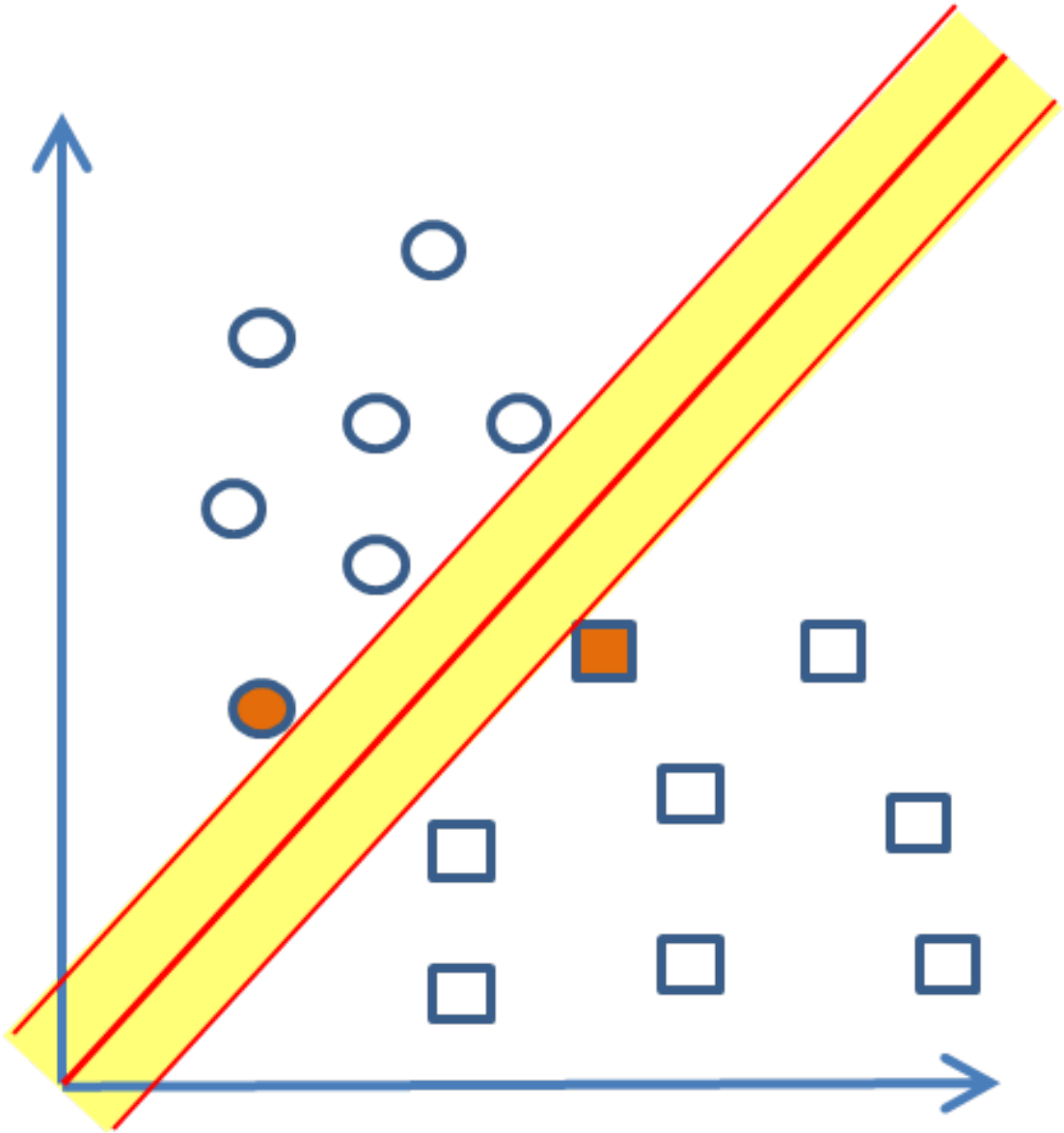
Machine learning methods

- Neural network
- Support vector machine (SVM)

Neural network



Support vector machine



Proposed project

- Explore additional machine learning approaches (Bayesian networks, naive Bayes, etc)
- Explore a consensus approach (combine multiple machine learning methods)
- Explore additional features to improve structure prediction