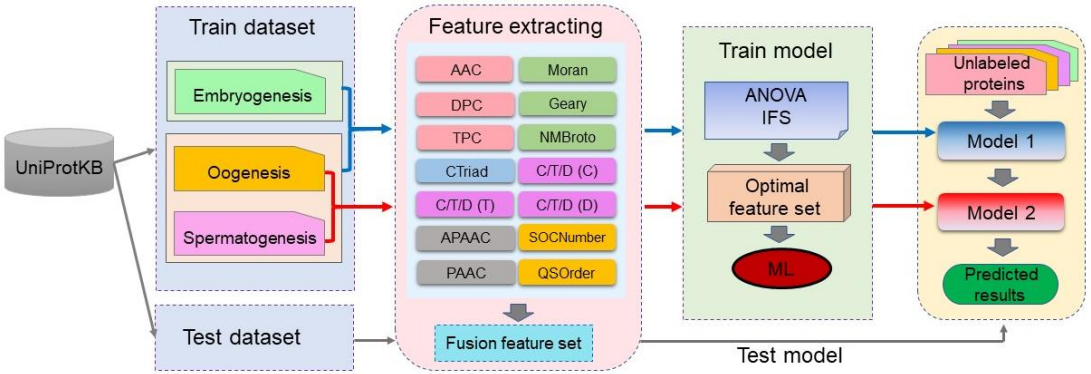
A computational model to identify fertility-related proteins using sequence information

Yan LIN, Jiashu WANG , Xiaowei LIU, Xueqin XIE , De WU, Junjie ZHANG, Hui DING

Frontiers of Computer Science, DOI: 10.1007/s11704-022-2559-6

Problems & Ideas

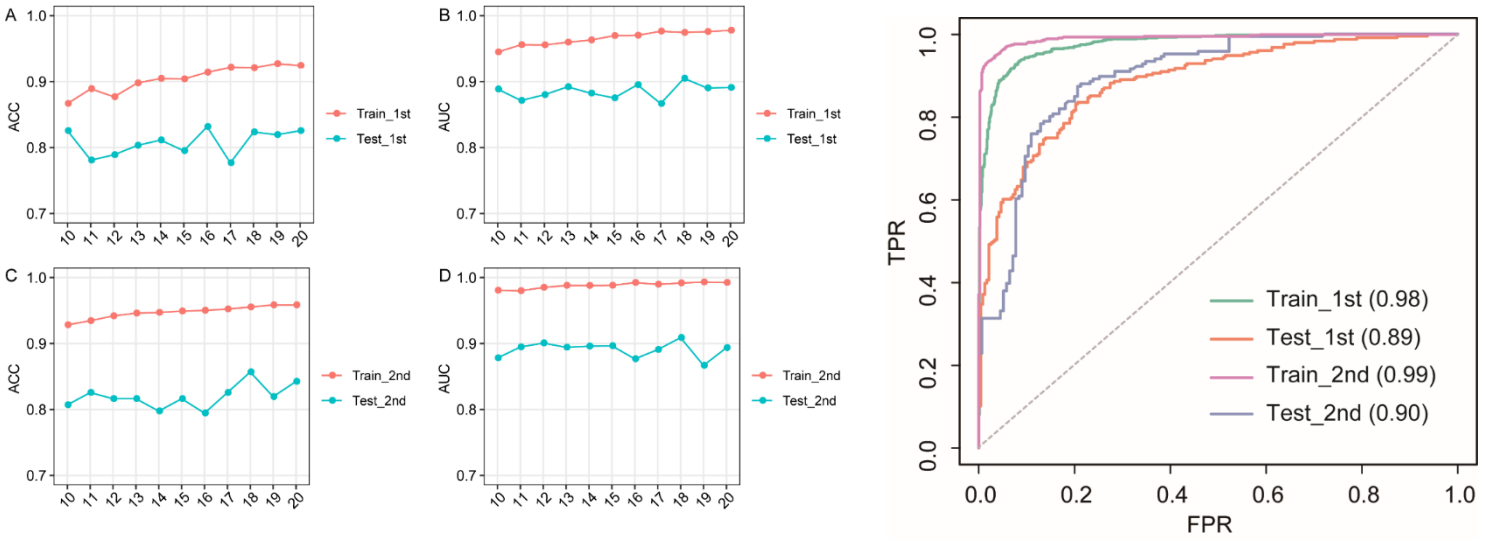
- Problems of conventional stereo matching approaches:
  - There is no classifier that can directly classify the three fertility-related proteins, including oogenesis-, spermatogenesis- and embryogenesis-related proteins.
- Ideas:
  - A two-layer classifier that directly predict and classify the three fertility-related proteins. The first classifier is constructed to distinguish embryogenesis-related proteins from oogenesis-related and spermatogenesis-related proteins. The second classifier is trained to discriminate the oogenesis-related proteins from spermatogenesis-related proteins.



Flowchart for the identification of fertility-related proteins.

Main Contributions

- Contributions:
  - A novel two-layer SVM classifier to identify proteins related to spermatogenesis, oogenesis and embryogenesis based on 14 feature coding schemes and feature optimization strategy;
  - The classifier shows good performance, and the average ACC and AUC are 83.35% and 0.89, respectively, which indicates this method can well identify fertility related proteins.



ACC and AUC results for SVM models under optimal feature subsets. Left: the performance of models trained by features between top 10% and 20%; Right: the ROC curves of SVM models