Choose Ohio First

James Munyon*, advised by Drs. Andy Chang* and Jack Min**

*Department of Mathematics and Statistics, **Department of Biological Sciences; Youngstown State University

Youngstown STATE UNIVERSITY

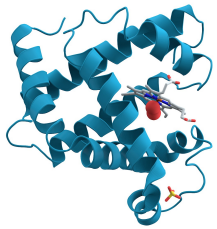
Overview

Proteins must be in proper cell locations to "do their jobs".
Expensive/difficult to measure directly.
Protein data enters databases faster than locations can be recorded.
Is a need for models which can predict locations of unknown proteins, based on proteins with known locations.

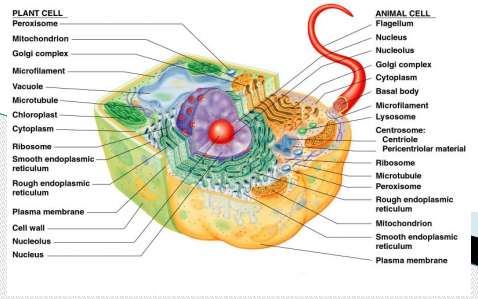
Machine learning approaches:
Predict unknown class labels (locations) based on known ones using different algorithms and methods.
Decision Tree Ensembles, K Nearest Neighbors, Support Vector Machines, etc.

Definitions

Proteins are large biological molecules, consisting of one or more long chains of amino acid residues. Proteins differ from one another primarily in their sequence of amino acids, which usually results in folding of the protein into a specific three-dimensional structure that determines its activity.



Subcellular locations are generally cell organelles, subunits, or sub-structures that hold the cell together and work in the cell.



Protein Data

An individual protein:

P = MALEPIDYTTHSREIDAEY…, the protein's amino acid sequence.

Transformation: Chou's Pseudo Amino Acid Composition:

P = [f1f2…f20+λ], where the first 20 values are the normalized occurrence frequencies of the 20 standard amino acids, and the last λ values are "interaction terms" designed to take some of the sequence order information into account. λ = 15 (by

	subcellular	PAC.R	PAC.K	PAC.E
A1XIH3	Secreted	0.015576841	0.020769121	0.01661530
A4QUT2	Secreted	0.020178363	0.025631975	0.01963300
B0YLW7	Cytoplasm	0.009188170	0.045940850	0.01225089
B0YLW8	Cytoplasm	0.023681875	0.039469791	0.04341677
B3A0S5	Secreted	0.025128965	0.019838656	0.01851608
B6EY09	Secreted	0.015175498	0.008129731	0.01951135
B7ZFT1	Cytoplasm	0.008483881	0.050903287	0.02969358
C0IW58	Secreted	0.021629250	0.007633853	0.02035694

UniProt

R

Obtaining Data / Subsetting

UniProt (Universal Protein Resource). Good benchmark dataset comes from following criteria. Proteins must:

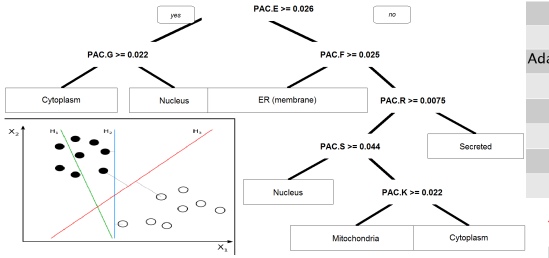
(1) be reviewed and annotated
(2) have annotation for just one location
(3) have experimental location evidence
(4) be non-fragments
(5) have no unknown amino acids
(6) be "sufficiently" long (>100 amino acids)

"50/50" BLASTClust algorithm: cluster similar proteins, keep only one from each cluster.

Benchmark dataset: 3002 fungal proteins from 16 locations: cytoplasm, cytoskeleton, ER, ER*, Golgi apparatus, Golgi apparatus*, mitochondria, mitochondria*, nucleus, nucleus*, [illegible]xisome, peroxisome*, plasma me[illegible]creted, vacuole, vacuole*; * = membra[illegible]

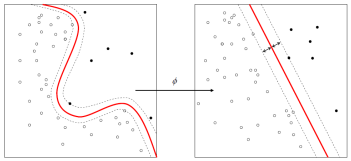
Decision Tree Methods

Predictive models that perform statistical classification.
Algorithms create decision structure as shown below.
Ensemble: collection of many trees.
Final location prediction is the one that occurs most often for a protein.
Validation strategies: 70%/30% training/testing split, OR ten-fold cross-validation (all data used for training and testing at different times).



Support Vector Machines

Find the line above that best separates the black and white dots (the red one!).
Extend this idea to a 35-D space that the protein data lives in.
Split points from two locations with a hyperplane? Probably not so simple.
Project data to a >35-D space, will probably do better (but not perfect).
Kernel function projects the points.



16(16-1)/2 = 120 models made, each just considering proteins from two locations at a time.
Jackknife validation: create 120 models using all but one protein to predict for the left-out protein (by literature).

Covariant Discriminant Algorithm

Each location has an "average" protein.
Compare a protein to each of the 16 averages.
35-D space means complicated distance/(dis)similarity measure.
Protein's prediction = location associated with the smallest of the 16 distances.
Repeat process for each protein.

Results/Conclusions

Method	Error Rate of Classification
Random Forests	53.16%
Adaptive Boosting	57.49%
Adaptive Boosting (ten-fold cross-validation)	58.39%
SAMME	58.49%
Bagging	58.82%
Bagging (ten-fold cross-validation)	58.46%
Support Vector Machines	49.3 %
Covariant Discriminant Algorithm*	39.17%

*Couldn't calculate distances for five locations, thus, can't consider those.
Error rate arguably 37.27% after fixing.
Here, the Covariant Discriminant Algorithm gives the best results, and could expect them to improve if/when the mentioned issue is remedied.
Specifically: SVMs good for locations with many proteins, CDA good for locations with few proteins.

Future Work

Different transformation of data using PSI-BLAST and Position-Specific Scoring Matrices (use more biological information to represent proteins differently than with Pseudo Amino Acid Composition). Methods: same and new.

Selected References

Chou, K.C. & Shen, H.B. (2007). Recent progress in protein subcellular location prediction. Analytical Biochemistry, 370, 1-16.

Hua, S. & Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 17(8), 721-728.

Neizer-Ashun, K.A., Yu, F., Meinken, J., Min, X. & Chang, G.H. Prediction of Plant Protein Subcellular Locations. Unpublished manuscript.

Chou, K.C. & Elrod, D.W. (1999). Protein subcellular location prediction. Protein Engineering, 12, 107-108.

Meinken, J. & Min, X.J. (2012). Computational Prediction of Protein Subcellular Locations in Eukaryotes: an Experience Report. Computational Molecular Biology, 2(1), 1-7.