Model Selection & Performance Evaluation

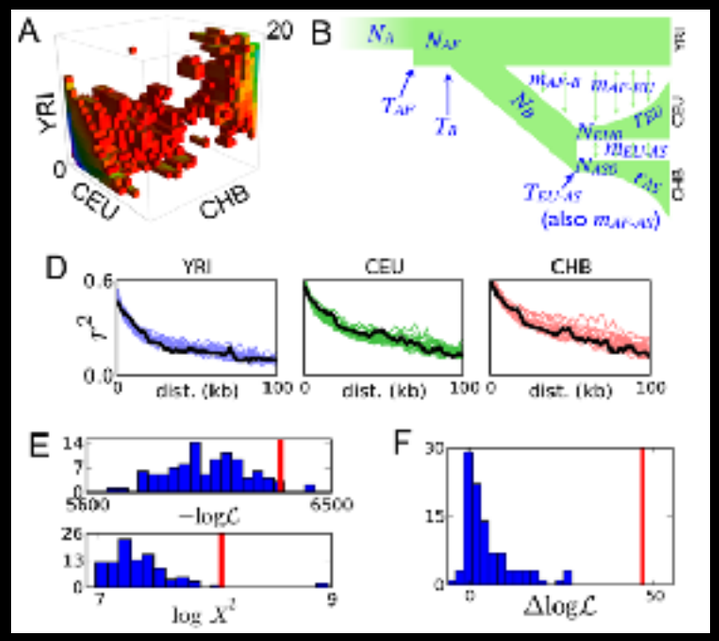
Katie Pollard

BMI 206

In this unit we will learn …

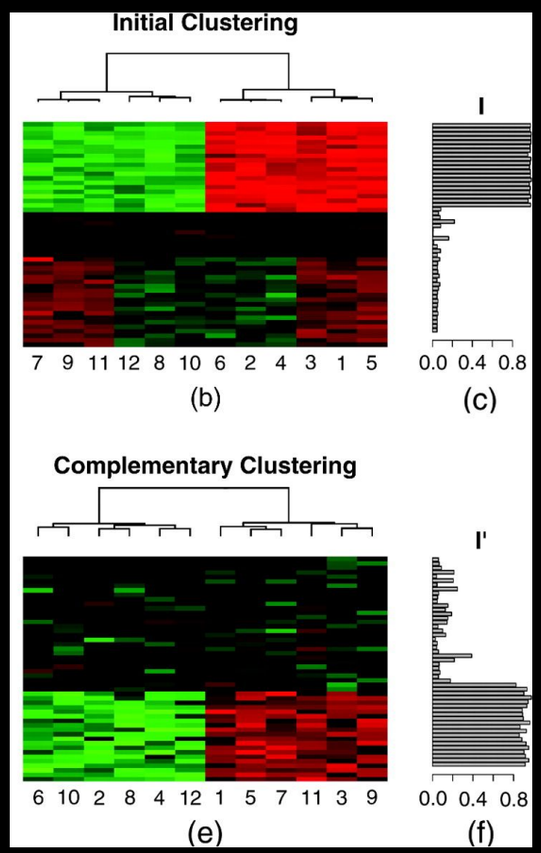
- What "big data" means in bioinformatics: many, typically correlated variables but relatively small sample sizes
- Why bioinformatics data is a challenge for modeling
- How to choose which variables to include in a statistical model
- How to evaluate model fit / performance visually and quantitatively
- Techniques for identifying the most important variables in a fitted model

Model selection in bioinformatics



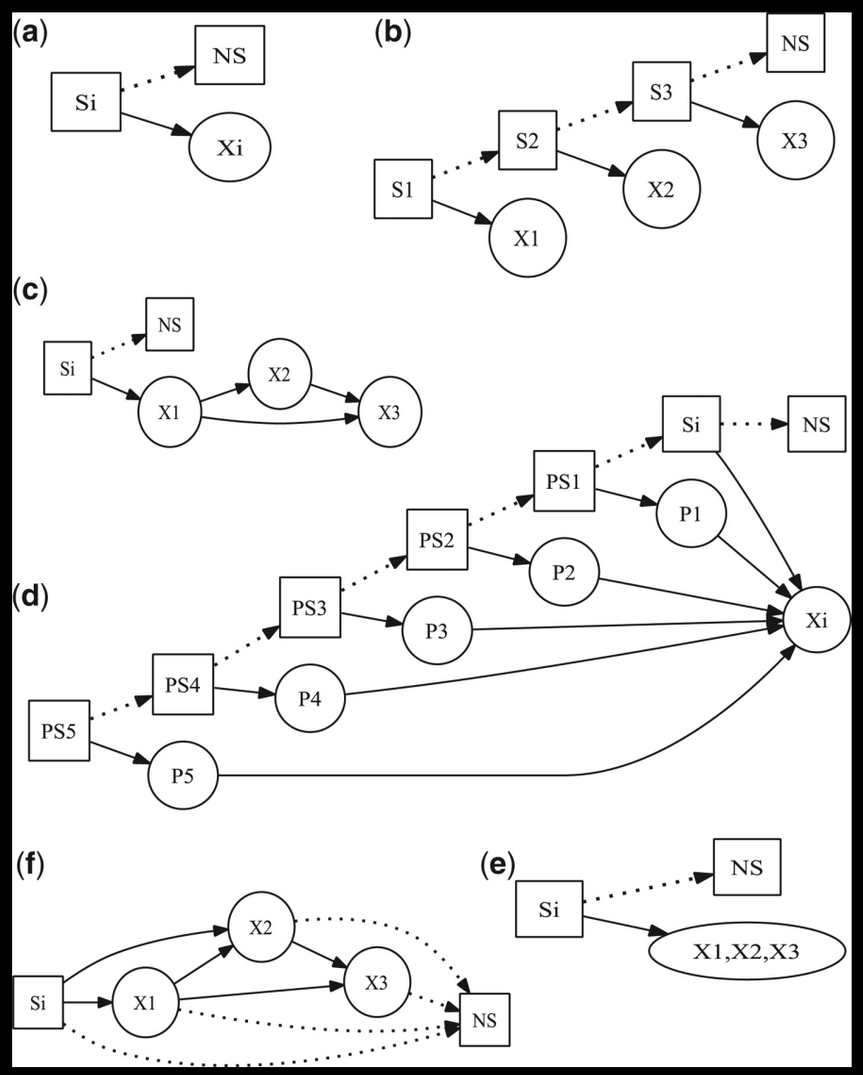
Gutenkunst et al. (2009) PLoS Genetics

A model of human expansion out of Africa that includes contemporary migration fits polymorphism data from Africa, Europe, and China better than a model without migration.



Nowak & Tibshirani (2007) Biostatistics

Clustering results and variable importance are very different after down-weighting highly expressed genes. The groups identified have different associations with survival and prognostic variables.



Mork & Holmes (2012) Bioinformatics

A collection of HMMs for modeling bacterial protein-coding gene potential.

Model selection in bioinformatics

In bioinformatics, typically many covariates are measured:

- Expression of thousands of genes in each tumor
- Genotypes at millions of variants in each cell line
- Evolutionary signatures at hundreds of pairs of residues in each protein structure
- Abundance of hundreds of thousands of proteins in each metagenome

Should all of these variables be in a model for an outcome of interest? Can they even feasibly be included?

With so many variables, over-fitting the observed data is a serious risk. This reduces generalizability of the results.

Modeling with Many Variables

Goal: a model that explains variation in the outcome without overfitting.

- Training: use sampled data to build a model of the outcome as a function of a *subset* of the large set of measured variables.
- Evaluation: quantify how well the model performs. Best if done using held out or independent "test" data.
- Prediction: model predicts value of the unknown outcome given observations of the covariates in future samples.

Modeling with Many Variables

Goal: a model that explains variation in the outcome without overfitting.

- Training: use sampled data to build a model of the outcome as a function of a *subset* of the large set of measured variables.
- Evaluation: quantify how well the model performs. Best if done using held out or independent "test" data.
- Prediction: model predicts value of the unknown outcome given observations of the covariates in future samples.

Modeling with Many Variables

How to pick the best model given the data?

- best subset of variables (variable selection),
- best functional form,
- best fitted parameters for this function.

Visualizing Linear Model Fit

Visualization techniques give a subjective evaluation of fit and can highlight reasons for poor fit.

- Scatter plot of observed Y versus predicted Y from the fitted model
- Scatter plot of residuals versus X: no trends if linear relationship
- QQ Plot of residuals versus normal quantiles
- Influential points: big effect on estimates, usually small residual, does the fit change if you drop?

Modeling with Many Variables

How to pick the best model given the data?

- best subset of variables (variable selection),
- best functional form,
- best fitted parameters for this function.

Visualization gives a qualitative measure of fit.

To quantitatively evaluate model fit and compare different models for variable selection requires a criterion for assessing model performance.

Quantifying Linear Model Fit

- Quantitative model diagnostics quantify fit:

$$r^2 = \frac{SS_{total} - SS_{resid}}{SS_{total}} = 1 - \frac{SS_{resid}}{SS_{total}} \qquad s_e = \sqrt{\frac{SS_{resid}}{n-2}}$$

  - Coefficient of determination (r^2) is the amount of variation in Y that can be explained by the linear relationship (i.e., model) between X and Y.
  - Pearson's correlation coefficient (r) is the square root of the coefficient of determination.
  - Standard deviation about the least squares line (s_e) or residual stand error is average distance points are from the line.

More Model Selection Criteria

- General criteria for evaluating "best"
  - Likelihood ratio statistic (compare to chi-square for test)
  - Change in residual sum of squares
  - Adjusted R^2
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
- The last three account for the number of parameters with penalties to avoid over-fitting or too complex models.

Algorithms for searching large space of possible models

All subsets selection involves enumerating all possible models and picking the best one. Some times this is computationally infeasible. Alternatives include

- Forward selection: Start with a small model and build up
- Backward selection: Start with the full model and remove terms
- Forward-backward selection: After building up, try removing terms to see if fit improves
- Deletion-substitution-addition and others: Algorithms for searching in a less linear fashion

Additional issue: Include interactions without main effects?

Correlated variables

In bioinformatics, covariates are often highly correlated:

- Co-expressed genes
- Genotypes in haplotype blocks
- Evolutionary signatures at adjacent residues
- Proteins in the same pathway or macromolecule

Consequently, only one (or a few) correlated variables will typically be selected for the model.

What determines which variable is selected?

Is the selected variable more important biologically?

Modeling with Many Variables

Goal: a model that explains variation in the outcome without overfitting.

- Training: use sampled data to build a model of the outcome as a function of a *subset* of the large set of measured variables.
- Evaluation: quantify how well the model performs. Best if done using held out or independent "test" data.
- Prediction: model predicts value of the unknown outcome given observations of the covariates in future samples.

Test Set Validation

Even with penalties for large models, observed data can be overfit, reducing generalizability and repeatability of results. Some solutions:

- Cross-validation involves holding out a random subset of the data and assessing model fit / performance on the held out data, repeatedly.

- External validation involves assessing model fit / performance on a totally independent data set (e.g., from a replication study or another population)

These can be used to pick variables, functional form, and compare different choices of fitted functions.

Classification Model Performance

TRUTH \ PREDICT	1	0
0	FP	TN
1	TP	FN

Specificity = true negative rate (TNR) = TN/(FP+TN)
Fall out = false positive rate (FPR) = FP/(FP+TN)
Sensitivity = true positive rate (TPR) = recall
= power = TP/(TP+FN)
Miss rate = false negative rate (FNR) = FN/(TP+FN)

Precision = positive predicted value = TP/(FP+TP)
False discovery rate = FP/(FP+TP)
Negative predicted value = TN/(TN+FN)
False omission rate = FN/(TN+FN)

Summary Performance Measures

Accuracy = (TP+TN)/(FP+TN+TP+FN)

F-score = (1+b^2)(precision x recall)/
(b^2 precision + recall)

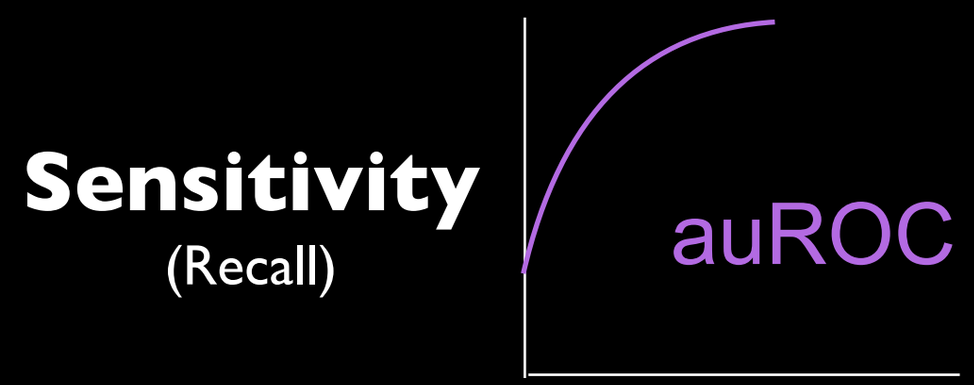
F1 = 2TP/(2TP+FP+FN)

F1 is harmonic mean of precision and recall

F2 weights recall higher

F0.5 weight precision higher

Area under the curve:



Variable importance

The importance of each covariate towards model fit can be measured with various statistics, e.g.:

- Estimated coefficient divided by its standard error (linear models - this fails in many other models)
- Sign of coefficient
- Fit of model with and without the variable included
- Average error minus error after permuting the covariate values, divided by its standard error (in cross-validation)
- Decrease in node impurity after adding variable (random forests)

For classification, assess importance for each class.