Statistics and Comparative Genome Analysis

Leopold Parts
leopold@mit.edu
29/09/2006

Theory Days and Biology

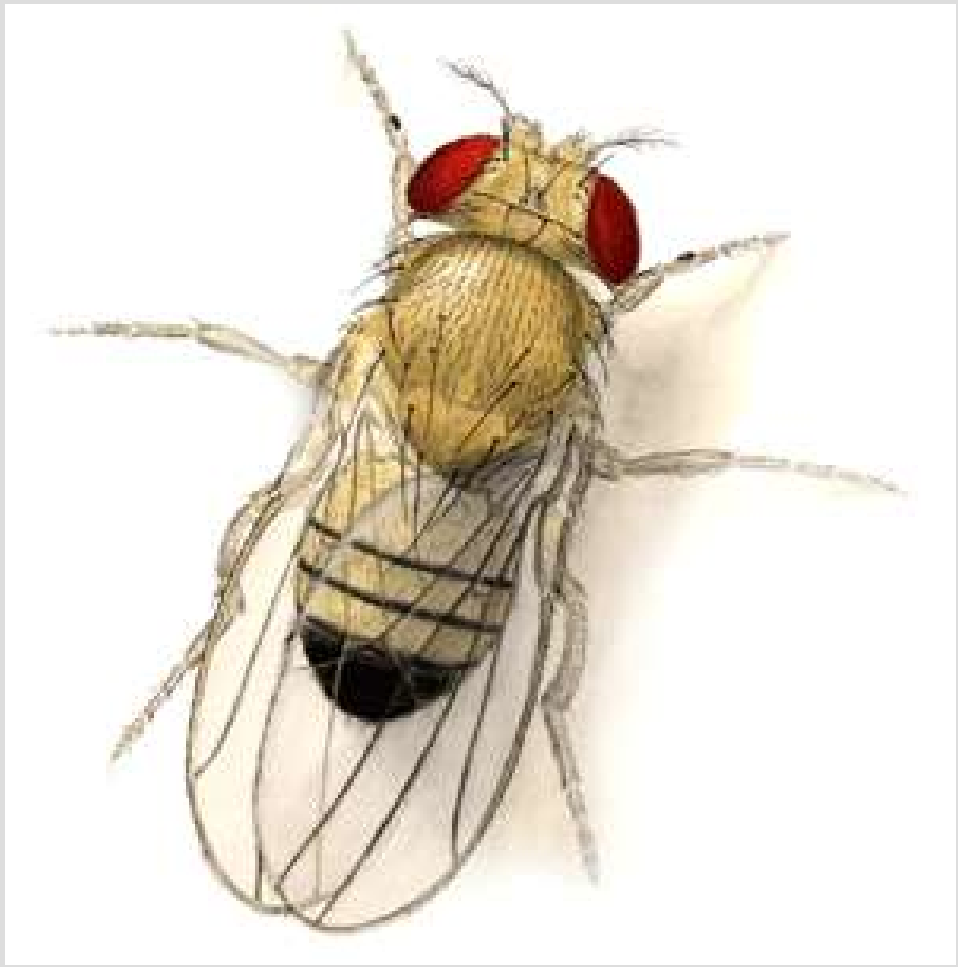
Today

- Comparative Genomics
  - Model organisms
  - Footprints of evolution
  - How many genomes to choose?
- Machine Learning for Computational Biology
  - Support Vector Machines
  - Decision Trees, Random Forests
  - Conditional Random Fields
  - Application: calling microRNA genes

Comparative Genomics

- Interdisciplinary science
- Understanding life
- Using model organisms

Fruit Fly – *Drosophila melanogaster*



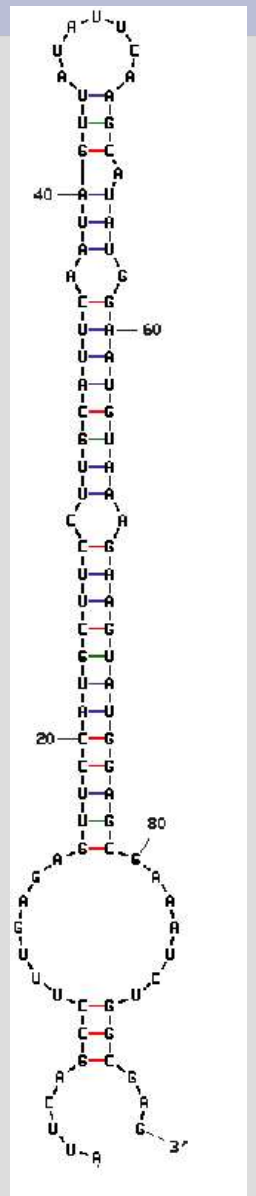
Footprints of Evolution

```
human TGCCAGCCGCGCGAGGTGGCCGCCTCGGCAGCCGCAGCTAAGAAGGAGCTCAAGTAC
mouse TGCCAGCCGCGCGAGGTGGCCGCCTCGGCAGCCGCAGCTAAGAAGGAGCTCAAGTAC
  rat TGCCAGCCGCGCGAGGTGGCCGCCTCGGCAGCCGCAGCTAAGAAGGAGCTCAAGTAC
  dog TGCCAGCCGCGCGAGGTGGCCGCCTCGGCAGCCGCAGCTAAGAAGGAGCTCAAGTAC
      ************************************************************
```

```
human TGC---CCGCGCGAGGTGGCCGCCTCGGCAGCCGCAGCTAAGAAGGAGCTCAAGTAC
mouse TGCCAGCCACGTGACGTGGCTG---TGGCAGCGGCAGCTAAAAAAGAGCTTAAGTAT
  rat TGCCAGCCACGCGACGTGGCCG---TGGCAGCAGCCGCTAAAAAGGAACTTAAGTAC
  dog TGCCAGCCACGCGAGGTGGCGG----------CTGCGGCCAAGAAAGAGCTCAAGTAC
      ***   ** ** ** ***** *          * ** ** ** ** ** ** *****
```

Footprints of Evolution

```
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGT-T-ATATTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGACAT dmel
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGT-T-ATACTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGACAT dsim
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGT-T-ATACTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGACAT dsec
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-ATACTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGTCAT dyakrec
GCCT-GTGAGAGTTCCATGCTTCCTTGCATTCAATAGT-T-ATACTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGACAT dere
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-AATATCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCAAGACTT dana
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-AACATAAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCAAGACAG dpserec
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-AACATAAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCAAGACAT dper
GCCTCTTGAGAGTTCCATGCTTCCTTGCATTCAATAGTATATTAAATAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGTCCGACAT dwil
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-TTATATAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAAGCAT dmoj
GCCT-TTGAGAGTTCCATGCTTCCTTACATTCAATAGTAT-TCAAATAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAAACAT dvir
GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-ATAAATAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAAACAT dgri
**********************************************************************************************
********************************************** **************************************************
********************************************** **************************************************
************************************** ***** **************************************************
************************************** ***** **************************************************
************************************** ***** * ************************************************
************************************** ***** * ************************************************
************************************** *** *   ****************************************** ******
****************************************** **    ****************************************** * ****
****************************************** **    ****************************************** *  ***
****  ********************* *********** *       ****************************************     *
```



Conservation, not abundance

- Quantify patterns of conservation

Defining 'conservation'

- 4-letter alphabet
- Single letter – probability α of changing
- Markov matrix:

$$\begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix}$$

Defining 'conservation'

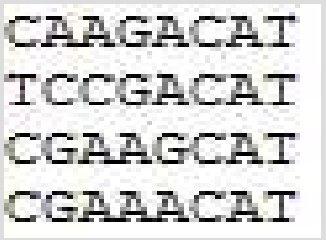
- After *t* time steps:

$$\begin{pmatrix} \frac{1}{4}+\frac{3}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \dots \\ \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}+\frac{3}{4}(1-\frac{4}{3}\alpha)^t & \vdots & \\ \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \vdots & \\ \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & & \end{pmatrix}$$

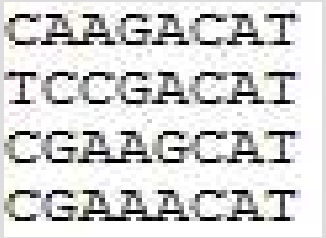
- P(no change) ~ 0.25 + 0.75 exp(-1.33 α)

Extending 'conservation'

- 'Feature': L independent letters (8)
- 'Alignment': N aligned genomes (3)
- 'Feature Conservation'
  - Background mutation rate: α (higher)
  - Conserved mutation rate: $\alpha\omega$ (lower, $\omega < 1$)
  - Observed mutation rate: α' = % conserved (16/24)
  - 'Conserved': α'LN >= C (some threshold)

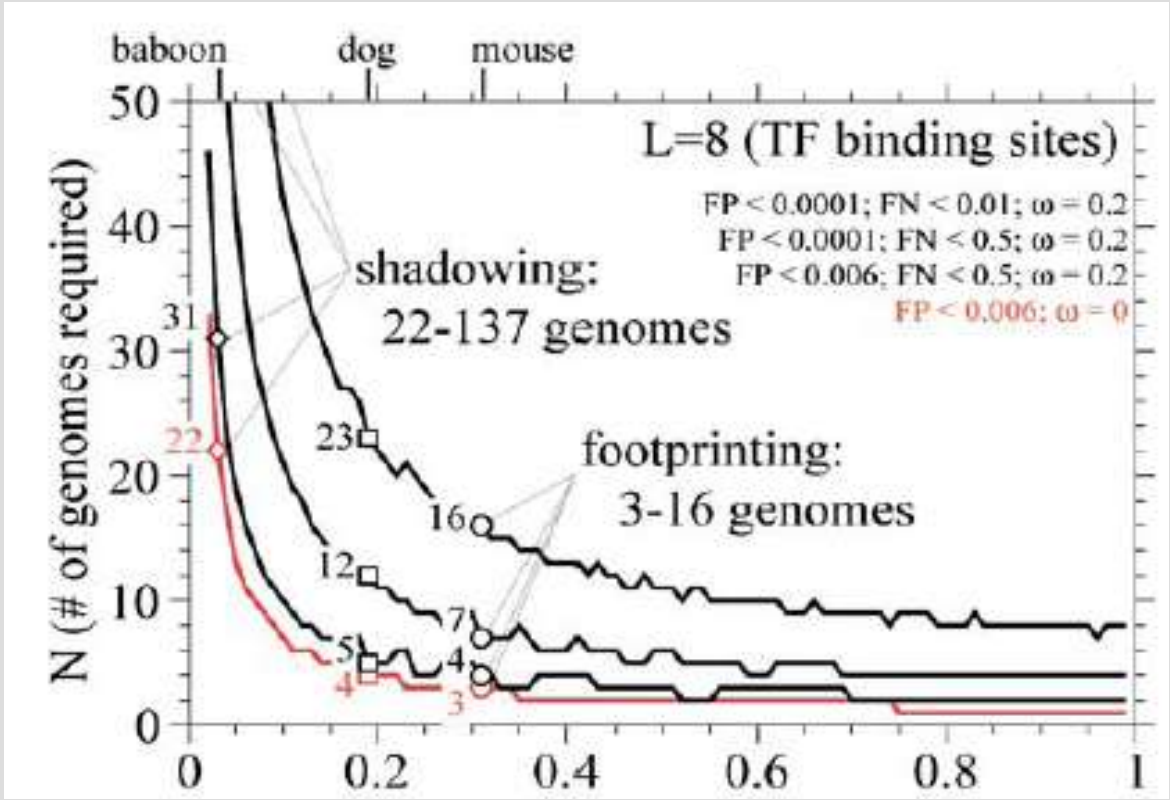


Error rate



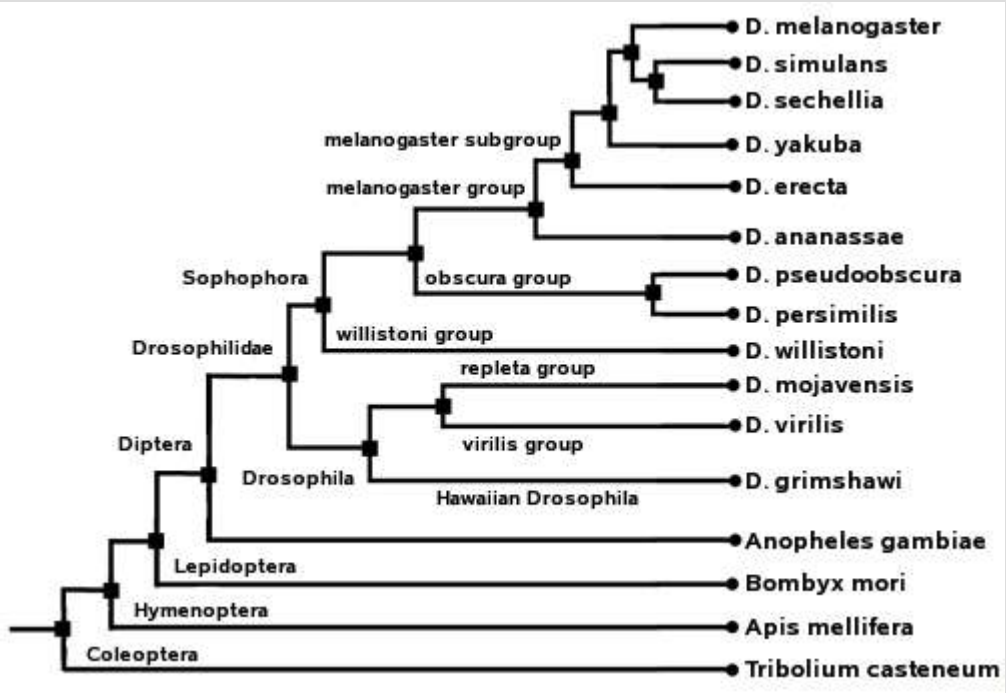
- Binomial distribution
- False positive: P(>= C conserved | neutral)
- False negative: P(< C conserved | conserved)

Error rate and # of genomes



Our project

- 12 genomes
- Mean # of changes = 0.5 (min 0.1, max 1.9)



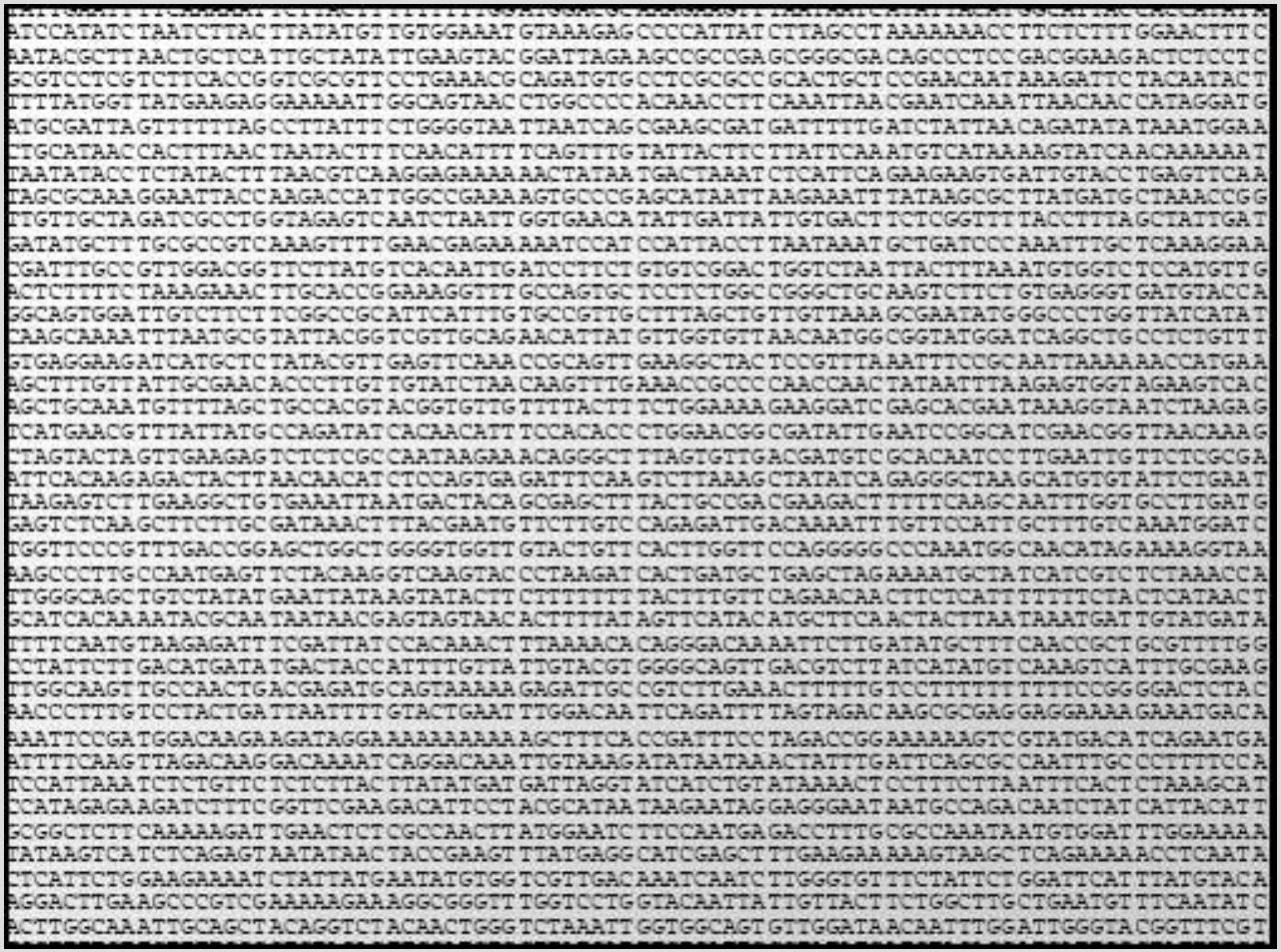
How many genomes to choose?

- Detectable feature size ~ 1 / # of genomes
- The more genomes the better resolution
- Close distances disfavored

Today

- Comparative Genomics
  - Model organisms
  - Footprints of evolution
  - How many genomes to choose?
- **Machine Learning for Computational Biology**
  - Support Vector Machines
  - Decision Trees, Random Forests
  - Conditional Random Fields
  - Calling microRNA genes

Finding needles in haystack

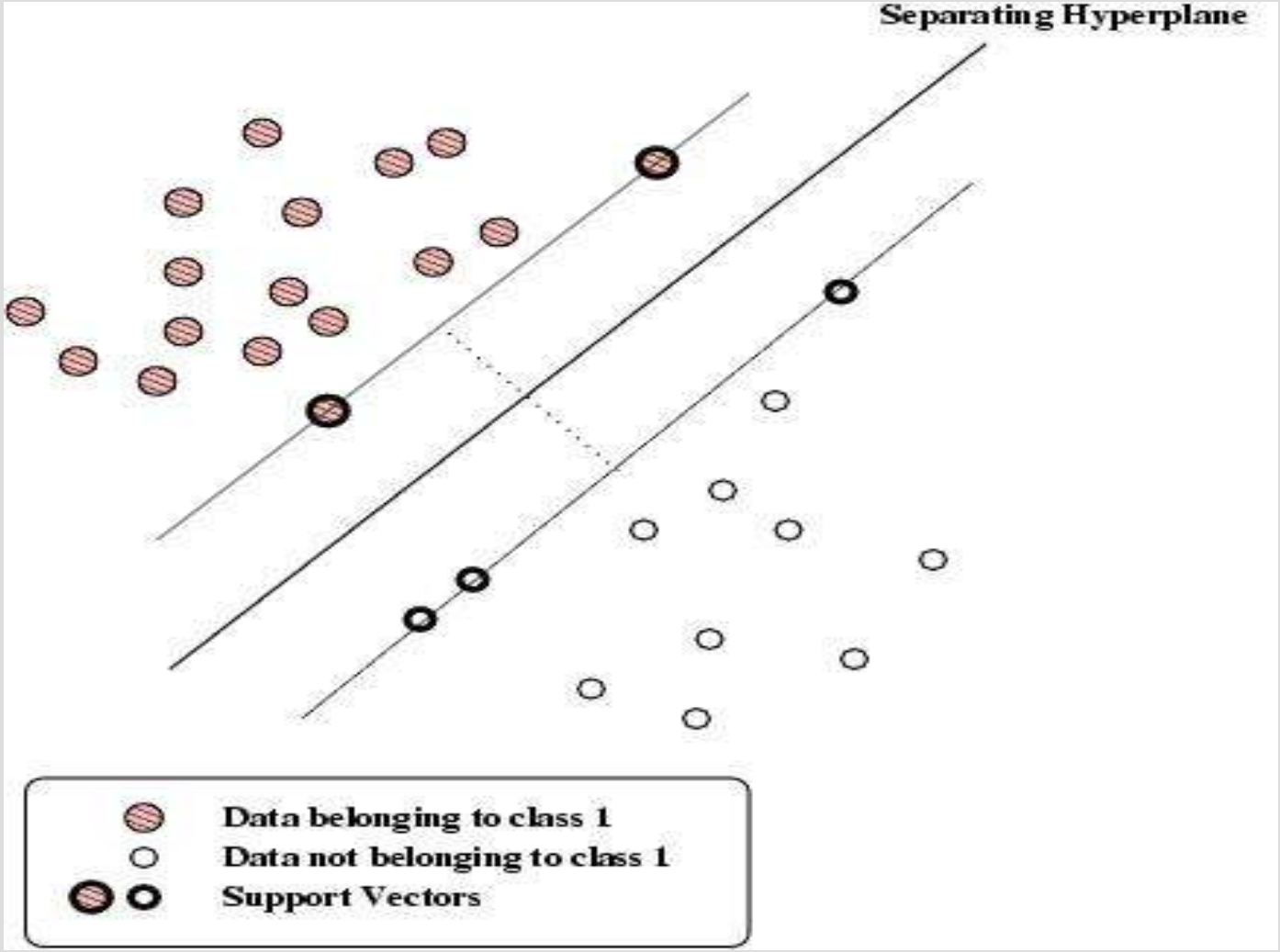


Finding needles in haystack

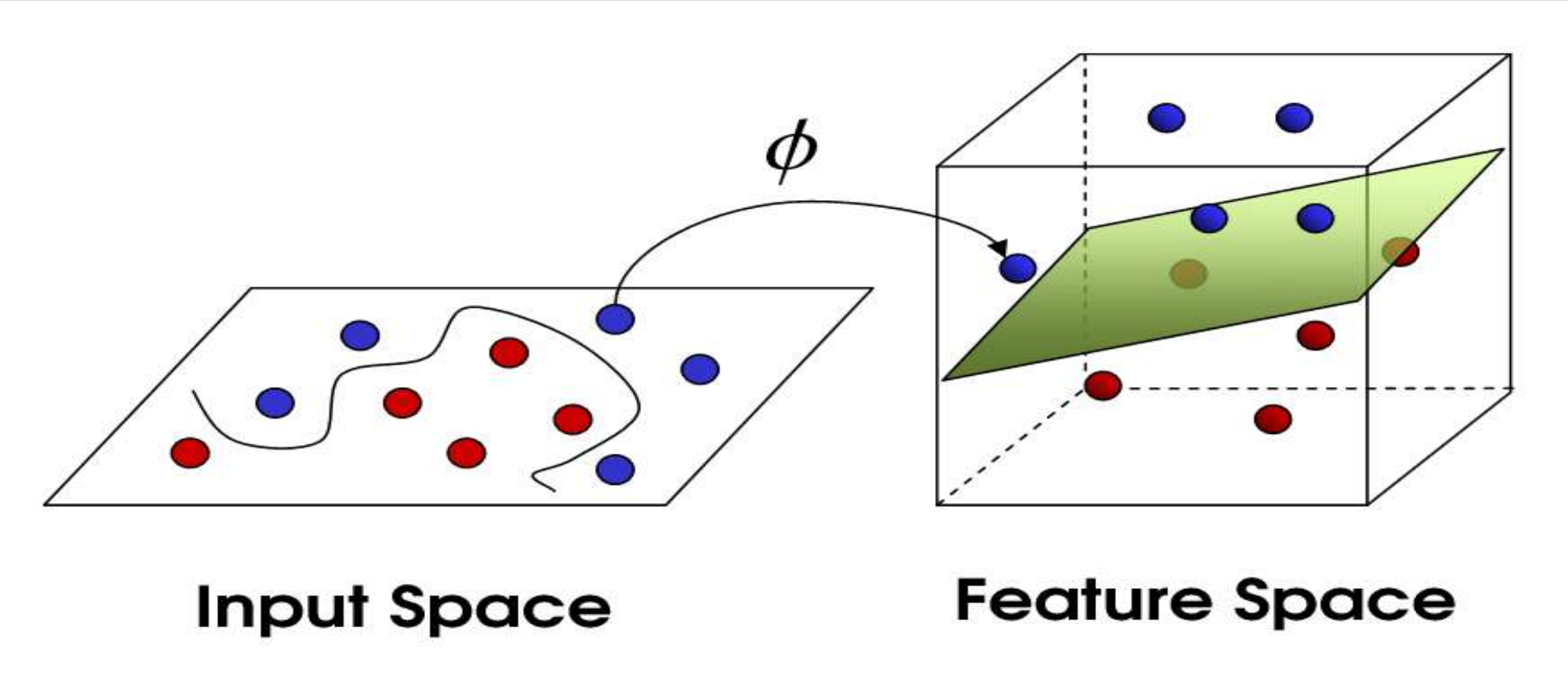
Finding needles in haystack

- Define a set of features
  - Biologically meaningful (hopefully)
  - Conservation, structure, size, etc.
- **Classify**

Support Vector Machine



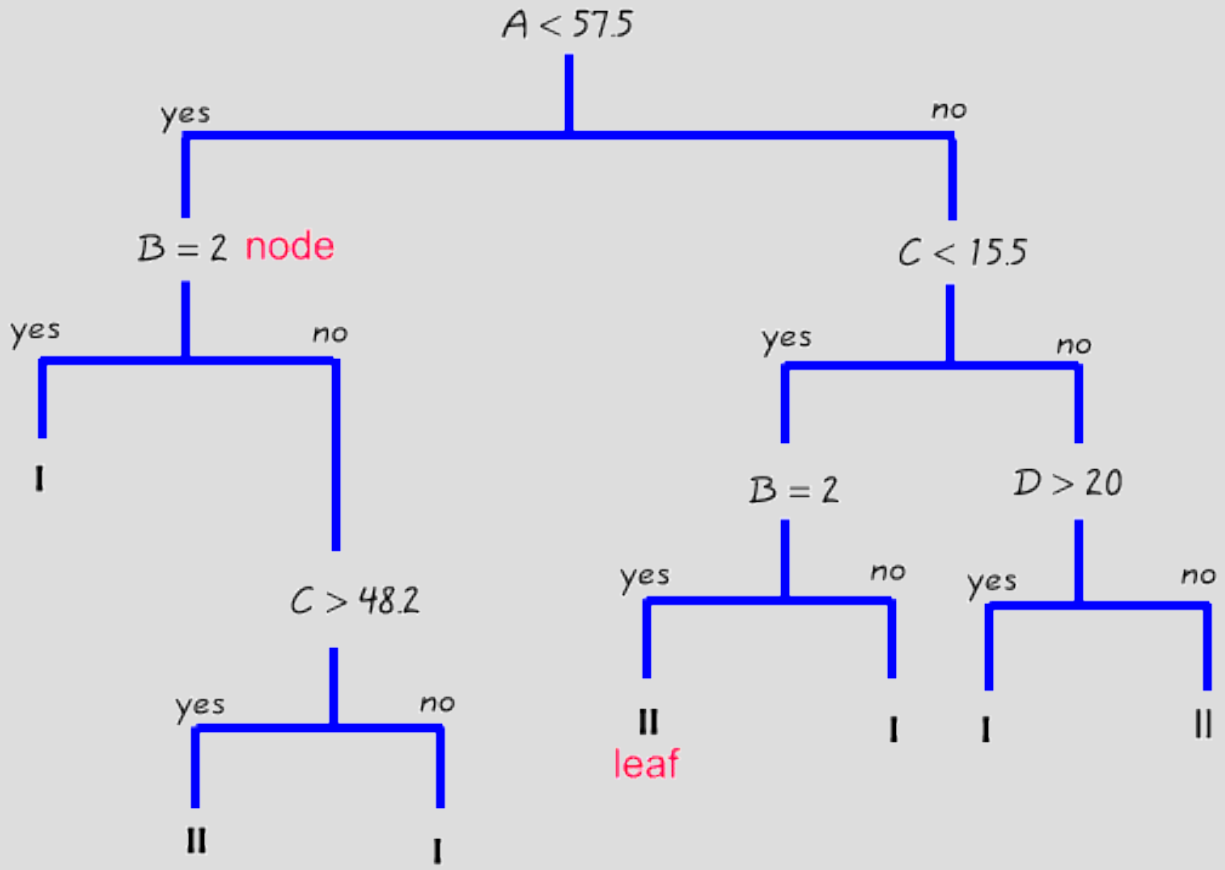
Support Vector Machine



Support Vector Machine

- Performs well
- Lots of useful kernel functions
- SVMLight software

Decision Tree



Decision Tree

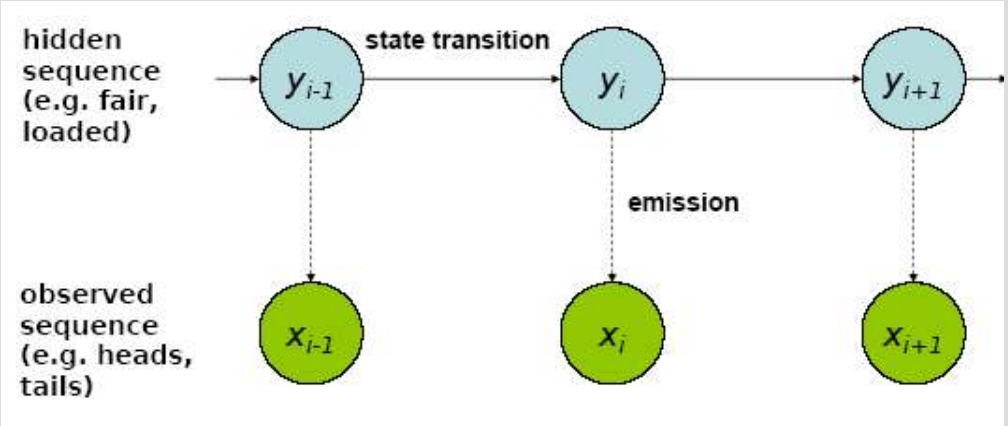
- Choose split based on entropy/error/...
- Good:
  - Don't care for scaling/normalization
  - Does feature selection
  - Easily interpretable
- But, high variance

Random Forest

- Forest = lots of trees
- Choose *m* out of *M* variables at each split
- Sample data with replacement, test on rest
- Majority voting scheme
- Does not overfit (empirically)

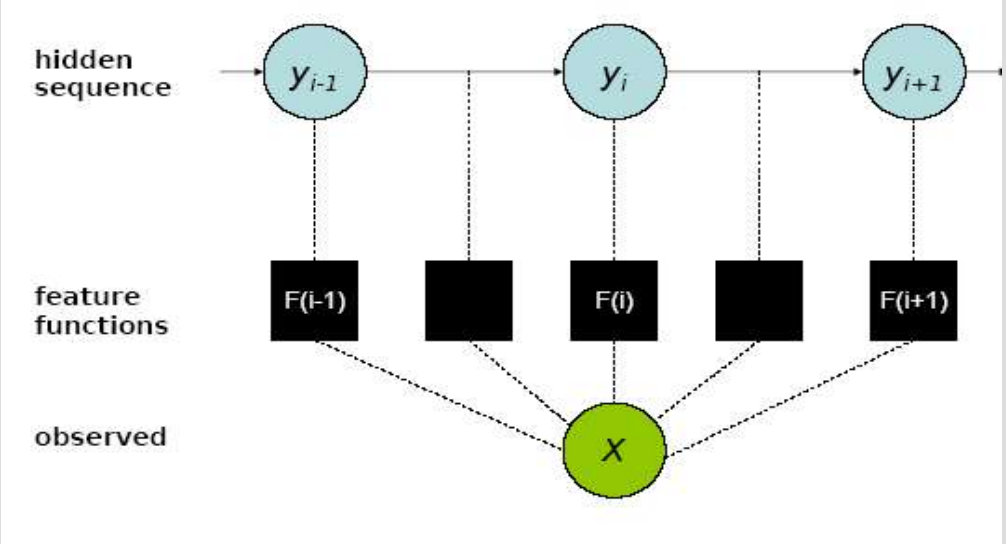
Conditional Random Field

- Extension of Hidden Markov Model



Conditional Random Field

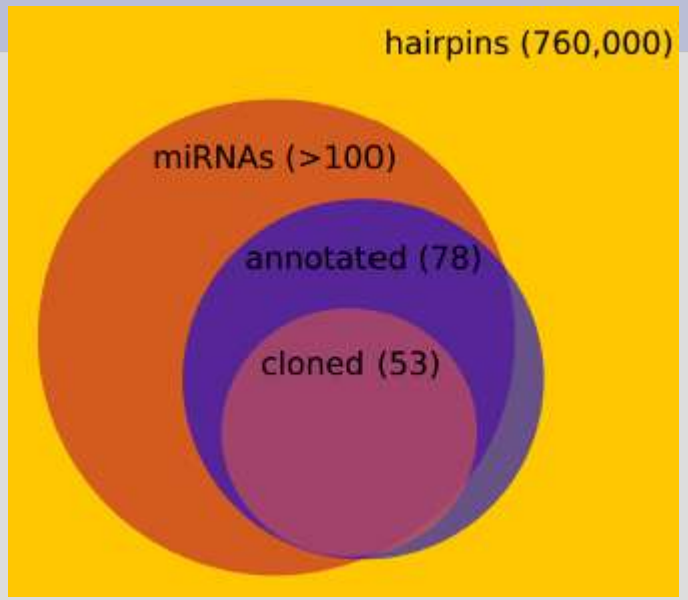
- Extension of Hidden Markov Model
  - Emission probabilities => linear combination of feature functions
  - Train feature weights



Conditional Random Field

- Very flexible, powerful
- Feature functions
  - Need not be probabilistic
  - May perform any computation on inputs
  - May use entire observed sequence
- Hard to implement, no ready software

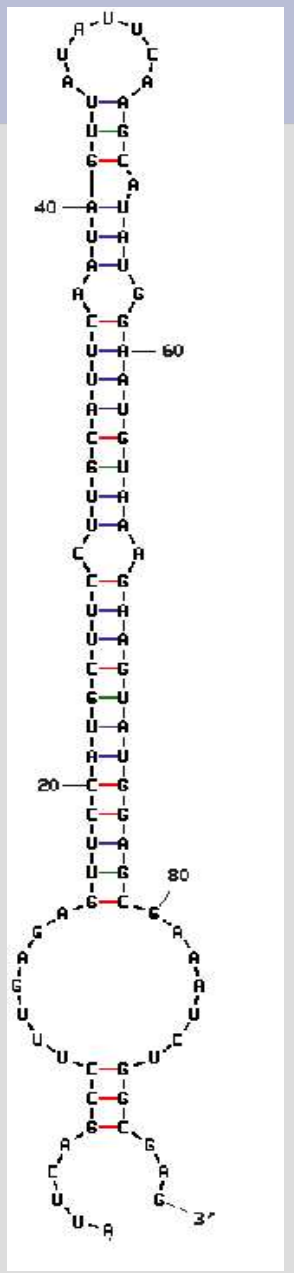
Calling microRNA genes



- Mistakes in known positives
- Many unknowns
- No clear negatives
- Many unknowns clearly wrong

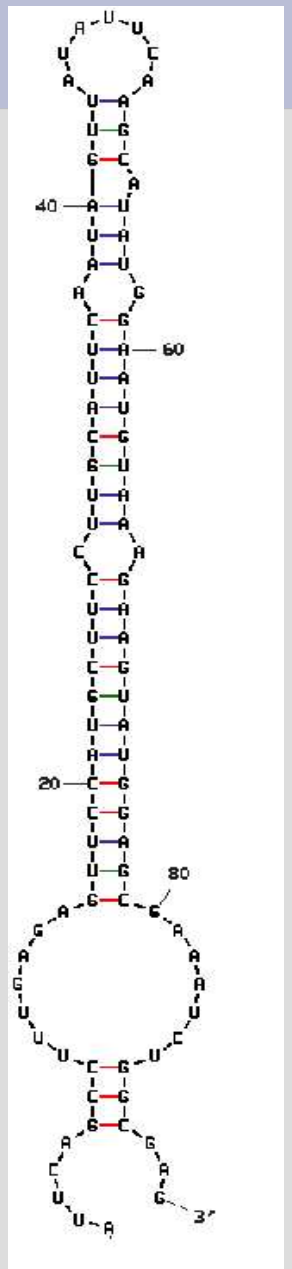
Calling microRNA genes

- ~800 000 test data, ~500 features
- Measure of goodness – LOOCV
- Classify:
  - SVM
  - Decision tree
  - Random forest



Calling microRNA genes

- Calling start of mature microRNA
- ~90 000 test data, 89 features
- Highly variable training data
  - Combine classifiers trained separately?
  - Normalizing features



Results

- MicroRNAs
  - 96% sensitivity
  - 22 of top 25 predictions biologically verified
  - Previously- 53/78, now 75/100!
- Mature microRNAs
  - 83% correct
  - bad predictive ability for new finds

Finally

- Quantifying patterns of change
- Explaining life
- Next up – 63 yeast genomes

- Thank you's:
  - Alex Stark, Manolis Kellis, Pouya Kheradpour, Jaak Vilo