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 TGC---CCGCGCGAGGTGGCCGCCTCGGCAGCCGCAGCTAAGAAGGAGCTCAAGTAC mouse TGCCAGCCACGTGACGTGGCTG---TGGCAGCGGCAGCTAAAAAGAGCTTAAGTAT rat TGCCAGCCACGCGACGTGGCCG---TGGCAGCAGCCGCTAAAAAGGAACTTAAGTAC dog TGCCAGCCACGCGAGGTGGCGG----CTGCGGCCCAAGAAAGAGCTCAAGTAC *** ** ** ** ** ** *****

Footprints of Evolution

GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGT-T-ATATTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGACAT dmel GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGT-T-ATACTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGACAT dsim GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGT-T-ATACTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGACAT dsec GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-ATACTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGTCAT dvakrec GCCT-GTGAGAGTTCCATGCTTCCTTGCATTCAATAGT-T-ATACTCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAGACAT dere GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-AATATCAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCAAGACTT dana GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-AACATAAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCAAGACAG doserec GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-AACATAAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCAAGACAT dper GCCTCTTGAGAGTTCCATGCTTCCTTGCATTCAATAGTATATTAAATAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGTCCGACAT dwil. GCCT-TTGAGAGTTCCATGCTTCCTTGCATTCAATAGTAT-TTATATAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAAGCAT dmoi GCCT-TTGAGAGTTCCATGCTTCCATTCAATAGTAT-TCAAATAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAAACAT_dvir GCCT CCTTGCATTCAATAGTAT-ATAAATAAGCATATGGAATGTAAAGAAGTATGGAGCGAAATCTGGCGAAACAT dori



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:

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

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: $\alpha' = \%$ conserved (16/24)
 - 'Conserved': $\alpha' LN \ge C$ (some threshold)

CAAGACAT TCCGACAT CGAAGCAT CGAAACAT

Error rate

CAAGACAT TCCGACAT CGAAGCAT CGAAACAT

- 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

	<u></u>
ATCCATATC TAATCTTAC TTATATGT TGTGGARAT GTARAGAG CCCCATTAT CTTAGCCT AAAAAAAACC TTCTCTTT GGRACT	TTC
ratrosott ractsotor tisotata tisaastac ssattasrascosocor sossecsr casooctoc grossas actoto	CTC
SCGTCCTCG TCTTCACCG GTCGCGTT CCTGAAACG CAGATGTG CCTCGCGCC GCACTGCT CCGAACAAT AAAGATTC TACAATA	ACT
etteatget tatgaagag gaaaaatt ggcagtaac ctggcccc acaaaccet caaattaa cgaatcaaa ttaacaac cataggi	ATG
ATGCGATTA GTTTTTTAG CCTTATTT CTGGGGTAA TTAATCAG CGAAGCGAT GATTTTTG ATCTATTAA CAGATATA TAAATG(GAA
CTGCATAAC CACTTTAAC TAATACTT TCAACATTT TCAGTTTG TATTACTTC TTATTCAA ATGTCATAA AAGTATCA ACAAAAA	AAT
FAATATACC TCTATACTT TAACGTCA AGGAGAAAA AACTATAA TGACTAAAT CTCATTCA GAAGAAGTG ATTGTACC TGAGTTC	CAA
FAGCGCRAA GGRATTACC AAGACCAT TGGCCGARA AGTGCCCG AGCATAATT AAGAAATT TATAAGCGC TTATGATG CTAAACC	CGG
FIGTIGCTA GATEGEETG GTAGAGTE AATETAATT GGTGAACA TATTGATTA TIGTGAET TETEGGTIT TAEETTTA GETATTE	GAT
GATATGCTT TGCGCCGTC AAAGTTTT GAACGAGAA AAATCCAT CCATTACCT TAATAAAT GCTGATCCC AAATTTGC TCAAAG	GAA
CGATTTGCC GTTGGACGG TTCTTATG TCACAATTG ATCCTTCT GTGTCGGAC TGGTCTAA TTACTTTAA ATGTGGTC TCCATG1	TTG
RCTCTTTTC TARAGAARC TTGCRCCG GARAGGTTT GCCRGTGC TCCTCTGGC CGGGCTGC RAGTCTTCT GTGRGGGT GATGTRC	CCA
bgcagtgga tigtcitict toggoogcaticatitg tgoogttg cittagcig tigttaaa gogaatatg ggoootgg tiatoai	TAT
CAAGCAARA TTTAATGCG TATTACGG TCGTTGCAG AACATTAT GTTGGTGTT AACAATGG CGGTATGGA TCAGGCTG CCTCTG	TTT
STGRGGRAG AT CATGCTC TATACGTT GAGTTCARA COGCAGTT GARGCTAC TOCGTTTA RATTTCOGC ARTTARARA ACCATO	GAA
AGCTTTGTT ATTGCGAAC ACCCTTGT TGTATCTAA CAAGTTTG AAACCGCCC CAACCAAC TATAATTTA AGAGTGGT AGAAGTC	CAC
rgctgcaaa tgttttagc tgccacgt acggtgttg ttttactt tctggaaaa gaaggatc gagcacgaa taaaggta atctaac	GAG
PCATGARCG TTTATTATG CCRGRTAT CACAACATT TCCRCRCC CTGGRRCGG CGATRTTG RATCCGGCA TCGRRCGG TTRACAA	AAG
CTAGTACTA GTTGAAGAG TCTCTCGC CAATAAGAA ACAGGGCT TTAGTGTTG ACGATGTC GCACAATCC TTGAATTG TTCTCGC	CGA
ATTCACAAG AGACTACTT AACAACAT CTCCAGTGA GATTTCAA GTCTTAAAG CTATATCA GAGGGCTAA GCATGTGT ATTCTGI	AAT
FAAGAGTET TGAAGGETG TGAAATTA ATGAETACA GEGAGETT TAETGEEGA EGAAGAET TTTTEAAGEAATTTGGT GEETTGJ	ATG
GAGTETEAA GETTETTGE GATAAACT TTAEGAATG TTETTGTE EAGAGATTG AEAAATT TGTTEEATT GETTTGTE AAATGGI	ATC
rggttcccgtttgaccggagctggctgggtggtt gtactgtt cacttggtt claggggg cccaaatgg caacatag aaaagg	TAA
AAGCCCTTG CCAATGAGT TCTACAAGGTCAAGTAC CCTAAGAT CACTGATGC TGAGCTAG AAAATGCTA TCATCGTC TCTAAA	CCA
ITGGGCAGC TGTCTATAT GAATTATA AGTATACIT CITITITT TACITIGIT CAGAACAA CITCICATITITITICIA CICATA	ACT
SCATCACAA AATACGCAA TAATAACGAGTAGTAAC ACTTTTAT AGTTCATAC ATGCTTCA ACTACTTAA TAAATGAT TGTATGA	ATA
FTTTCAATG TAAGAGATT TCGATTAT CCACAAACT TTAAAACA CAGGGACAA AATTCTTG ATATGCTTT CAACCGCT GCGTTT	IGG
CCTATTCTT GACATGATA TGACTACCATTTTGTTA TTGTACGT GGGGCAGTT GACGTCTT ATCATATGT CAAAGTCA TTTGCG	AAG
ITGGCAAGT IGCCAACTG ACGAGATG CAGTAAAAA GAGATIGC CGICITGAA ACIITIIG ICCIIITIII IIIICOGG GGACICI	TAC
RACCCTTTG TCCTACTGA TTAATTTT GTACTGAAT TTGGACAA TTCAGATIT TAGTAGAC RAGCGCGAG GAGGAAAA GAAATGI	ACA
AAATTCCGA TGGACAAGA AGATAGGA AAAAAAAA AGCTTTCA CCGATTTCC TAGACCGG AAAAAAGTC GTATGACA TCAGAAI	FGA
ATTTTCAAG TTAGACAAG GACAAAAT CAGGACAAA TTGTAAAG ATATAATAA ACTATTTG ATTCAGCGC CAATTTGC CCTTTTC	CCA
recattaaa tetetgite tetettae tratatgat gattaggt ateatetgt ataaaact cetttetta attteaet etaaage	CAT
CCATAGAGAAGATCTTTC GGTTCGAA GACATTCCT ACGCATAA TAAGAATAG GAGGGAAT AATGCCAGA CAATCTAT CATTACI	ATT
SCGGCTCTT CARAAAGAT TGAACTCT CGCCAACTT ATGGAATC TTCCAATGA GACCTTTG CGCCAAATAATGTGGAT TTGGAAJ	AAA
fataagtca teteagagt aatataac taeegaagt ttatgagg eategaget ttgaagaa aaagtaage teagaaaa aceteaj	ATA
CTCATTCTG GAAGAAAAT CTATTATGAATATGTGG TCGTTGAC AAATCAATC TTGGGTGT TTCTATTCT GGATTCAT TTATGTA	ACA
rggacttgragcccgtcgraragrarggcgggfttggtcctggtrcartirttgttacttctggcttgctgratgtttcrati	ATC
rcttggcaa attgcagct acaggtet acaactggg tetaaatt ggtggcagt gttggata acaatttgg attgggta eggttte	CGT

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