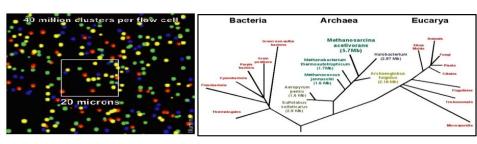


TAACCCTAACCCTAACCCTAACCCTA CCTAACCCTAACCCTAACCCTAACCC CCCTAACCCTAACCCTAACCCTAAC AACCCTAACCCTAACCCTAACCCTA ACCCTAACCCCAACCCCAACCCCAAC CTACCCTAACCCTAACCCTAACCCTA ACCCTAACCCTAACCCTAACCCTAA

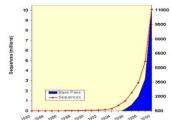


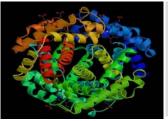
Transcriptome Analysis with noncoding RNAs

北京大学生物信息学中心 高歌 Ge Gao, Ph.D.

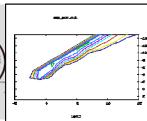
Center for Bioinformatics, Peking University





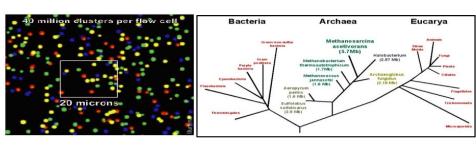








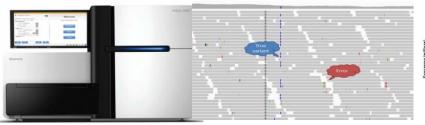
TAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAA

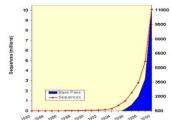


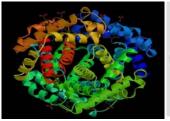
Unit 2: Data Mining: Identify long ncRNAs

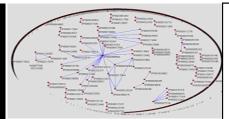
北京大学生物信息学中心 高歌 Ge Gao, Ph.D.

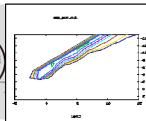
Center for Bioinformatics, Peking University



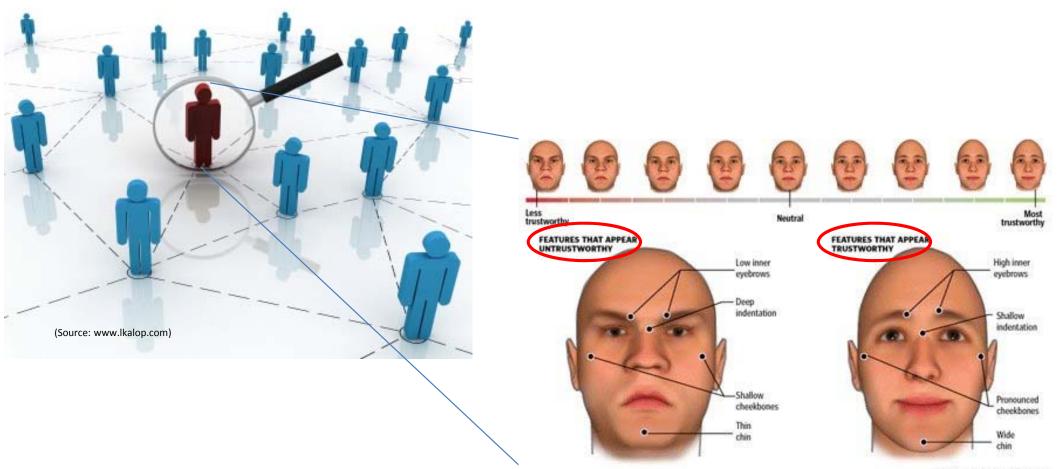








Identification



 $(Source: www.lemondrop.com/2009/01/22/certain-facial-features-found-to-create-a-feeling-of-trust/) \ \ The \ Boston \ \ Globe$

Features ~ property of an entity

Structural features

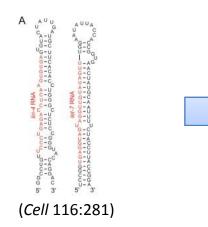
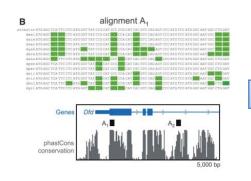
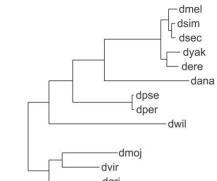


Table 1. Comparison of some filter-based approaches to miRNA gene finding in animals

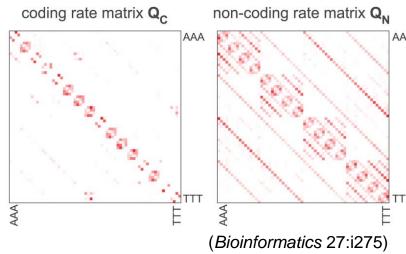
| | Initial set | Structural criteria | Conservation criteria | Additional filters Hairpins containing short repeats or with low quality structure are eliminated | | |
|--------------------------|---|--|---|--|--|--|
| Grad et al. (50) | Stem-loop structures in repeats- masked intergenic regions | MFE, GC content, matches, mismatches, gaps and occur- rence of multi-loops | Homologous stem-loops transitively identified in two additional genomes | | | |
| MirScan (8) | Folded structures identified sliding a 110-nt window along the genome | Number of bp, MFE, no overlap with repeats, no skewed base composition | Homologous stem-loops identified in an addi- tional genome | Log-odds score for severa features of the miRNA region of the stem-loop | | |
| Berezikov et al. (54) | Regions exhibiting a typical conservation pattern identified using phylogenetic shadowing | Only highly probable stable stem-loops are retained | Implicitly considered in the initial set | | | |
| MiRSeeker (9) | Aligned non-coding non-annotated regions from two species | Metrics involving length of longest stem-arm, MFE, internal loops, | Typical divergence pattern | | | |
| | | asymmetric loops and bulges applied to predicted structures in aligned regions | (Nucl. Acids Re | es, 37(8):2419) | | |

Evolutionary features





max. likelihood tree



Sequence features only

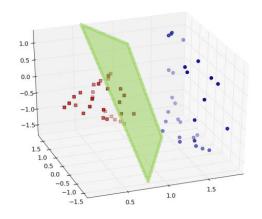
Mechanism neutral: works for both long and small ncRNAs

Accurate and Fast

SVM classifier

SVM – support vector machine

Separate transformed data with a hyper plane in a high-dimensional space



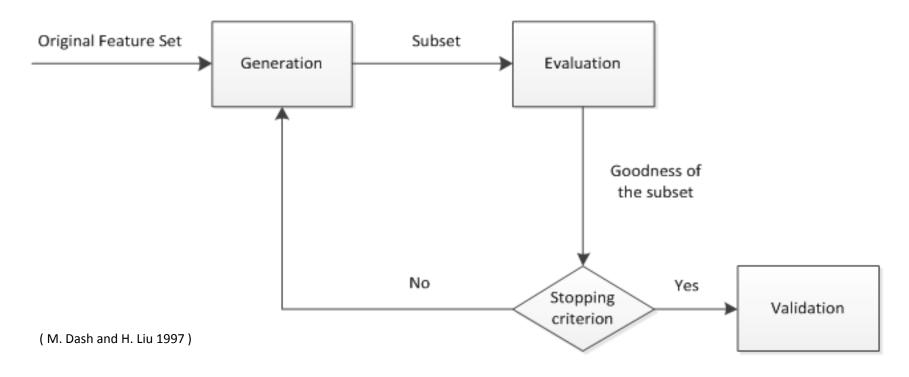
- Kernel function Radial Basis Function(RBF)
- Grid-search to select proper values of parameter

Sequence features Sequence Compositions (Predicted) Homology Secondary Structure Feature Selection (Conceptually) Sequence Translational domains Product

Copyright © Peking University

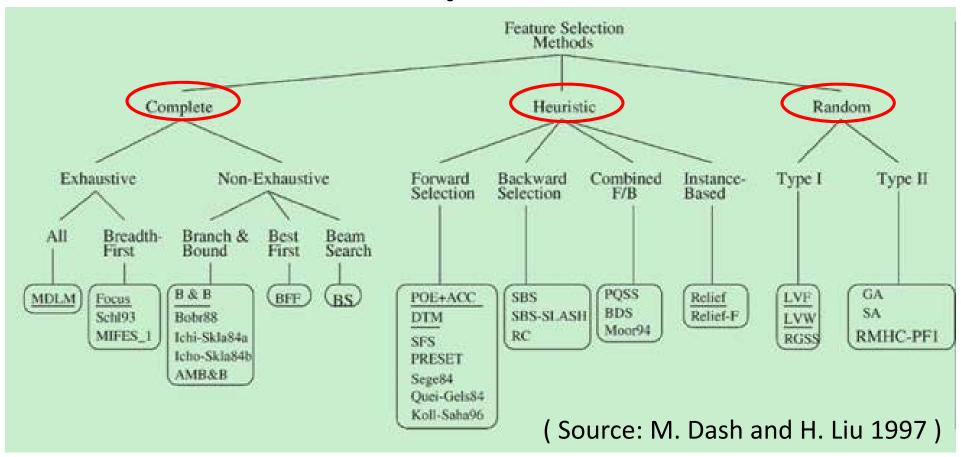
Feature Selection

Purpose: Choose the best feature set in term of accuracy, speed, and computing space

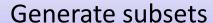


Copyright © Peking University

Find The Optimal Subset



Original *n* features



training

evaluation and select optimal solution

Stop training, use the test set to test result

Complete Search: Breadth First

The breadth-first traversal of all variables

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{n-1} + \binom{n}{n}$$

empty set

Add the most significant feature

training

evaluation

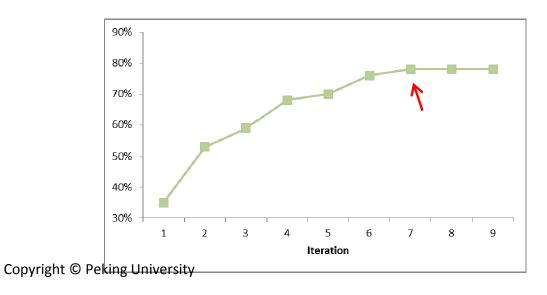
get the optimal solution

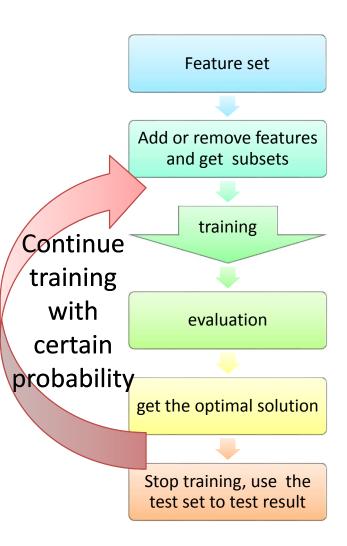
Stop training, use the test set to test result

Heuristic Search: Sequential Forward Selection

The overall performance increase?

Features added greedily until the addition of further features does not increase the overall performance.

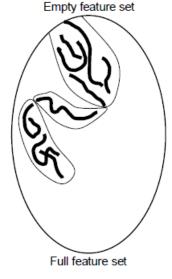




Random Search: Simulated Annealing

reach the optimal solution

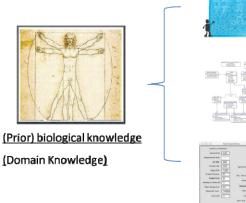
adding or removing features based on an "annealing-like" probability



(Source: http://courses.cs.tamu.edu/rgutier/cs790_w02/l17.pdf)

Initialized feature set

- Properties of entity
- Speculate based on existed knowledge
- Certain statistic established by predecessors
- The data that is thought to be relevant



Data

Model/Algorithm

Parameters

Sequence Compositions

e.g. frequency of k-mer

Sequence domains

e.g. known binding motif

(Predicted) Secondary Structure

e.g. folding energy (MFE)

(Conceptually)
Translational
Product

e.g. ORF length

Homologous

e.g. # of BLASTX hits

60+ features

Fine-tune with Breadth First Searching

11 features

Sequential Forward Selection

Coverage

ORF Integrity

LOG-ODD score

Copyright © Peking University

of BLASTX hits

Hit Score

Frame Score

(Conceptually) Translated Product

Coverage

$$Coverage(S) = \frac{L_{ORF} - (L_{mismatch} + 2 * L_{frameshift})}{\text{Total Length}}$$

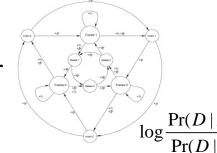
ORF Integrity

start codon and ends with an in-frame stop codon

LOG-ODD score

(Nucleic Acids Res. 35:W345)

indicator of the quality of a predicted ORF. The higher the score, the better the quality of the ORF



Homologous

of BLASTX hits

A true protein-coding transcript is likely to have more hits with known proteins than a non-coding transcript does

Hit Score

For a true protein-coding transcript, the hits are also likely to have higher quality $S_i = \max_i \{-\log_{10} E_{ij}\}, i \in [0,1,2]$

HIT SCORE =
$$\underset{i \in \{0,1,2\}}{mean} \{S_i\} = \frac{\sum_{i=0}^{2} S_i}{3}$$
,

Frame Score

For a true protein-coding transcript, most of the hits are likely to reside within one frame, whereas for a true non-coding transcript, even if it matches certain known protein sequence segments by chance, these chance hits are likely to scatter in any of the three frames $\text{FRAME SCORE} = \underbrace{variance}_{i \in [0,1,2]} \{S_i\} = \underbrace{\sum_{i=0}^2 \left(S_i - \bar{S}\right)^2}_{2}$

(Nucleic Acids Res. 35:W345)

Coverage

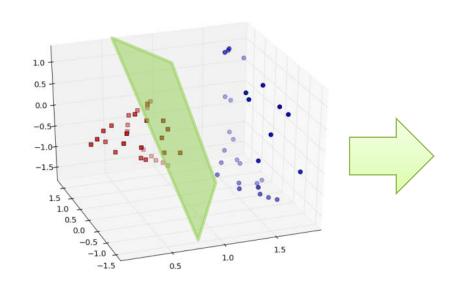
of BLASTX hits

ORF Integrity

Hit Score

LOG-ODD score

Frame Score



http://cpc.cbi.pku.edu.cn



Run CPC Get Results Quick Guide Recent transcriptome studies have revealed that large number of transcripts in mammals and other organisms do not encode proteins but function as noncoding RNAs (ncRNAs) instead. As millions of transcripts are generated by large-scale cDNA and EST sequencing projects every year, there is a need for automatic methods to distinguish protein-coding RNAs from noncoding RNAs accurately and quickly. We developed a Support Vector Machine-based classifier, named Coding Potential Calculator (CPC), to assess the protein-coding potential of a transcript based on six biologically meaningful sequence features. 10-fold cross-validation on the training dataset and independent testing on three large standalone datasets showed that CPC can discriminate coding from noncoding transcripts with high accuracy. Furthermore, CPC also runs an order-of-magnitude faster than a previous state-of-the-art tool and has higher accuracy. We developed a user-friendly webbased interface of CPC at http://cpc.cbi.pku.edu.cn. In addition to predicting the coding potential of the input transcripts, the CPC web server also graphically displays detailed sequence features and additional annotations of the transcript that may facilitate users' further investigation.

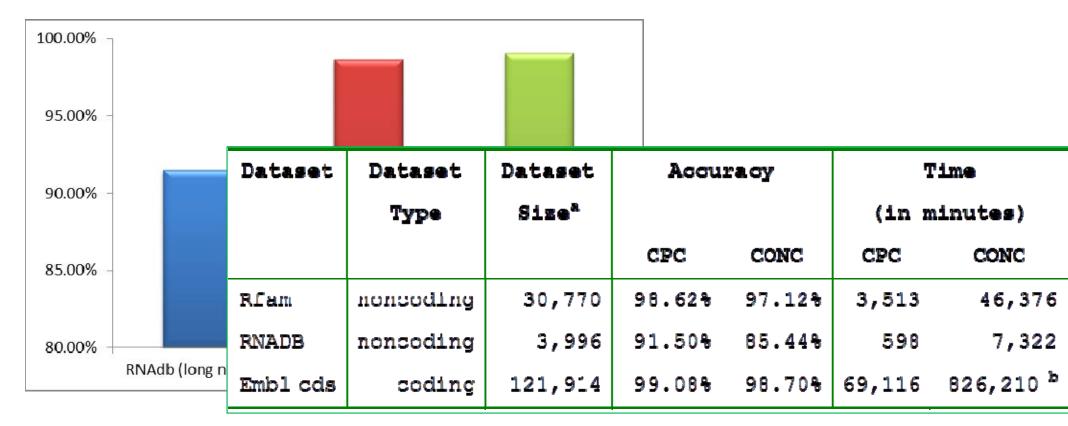


페 The coding potential calculator tool reads FASTA data format as input. "A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column.(ncbi) "

If you still do not very clear with what we are talking about, please refer to lemma FASTA at CPC Glossary.

To start your calculate task, dick HERE. And there is a step by setp guide to teach users how to use our CPC online. After user input sequences and run, the calculator will assign user a Task ID which is unique. You can use it to access your results at our Data Retrival Page





(Nucleic Acids Res. 35:W345)



| Gene Regulation | Function of ncRNA | H Van Bakel <i>et al.</i> PLoS Biology , 2010 | | | | | | | |
|--------------------|---|--|------|------|---------|--------|-------|------|------|
| | Long ncRNA | H Jia et al., RNA, 2010 TG Belard et al., Neuron, 2011 I Ulitsky et al. Cell, 2011 RS Young et al. Genome Biol Evol, 2012 | | | | | | | |
| | Short Peptide | X Yang et al., Genome Res, 2011 | | | | | | | |
| Stem Cell | Self-Renewal | JS Mohamed et al., RNA, 2010 | | | | | | | |
| | Neuron development | SY Ng et al., EMBO Journal, 2011 | 12 - | 32 m | nillior | ı seau | ences | | |
| Disease | Heart diseases | JH Lee et al., Circ Res, 2011 | 10 - | | | 00+ u: | | | |
| | Cancer Marker | BP Mello et al., Nucleic Acid Res, 2009 | 8 - | arou | ınd th | ie wo | rld | | |
| | Tumor mechanism | AC Tahira <i>et al.</i> , Molecular Cancer , 2011 RJ Flockhart <i>et al.</i> , Genome Res , 2012 | 6 - | | | | _ | | |
| Evolution | New genes | D Rose <i>et al.</i> , J Bioinform Compt Bio. , 2008 JF Sousa <i>et al.</i> , PLoS One , 2010 | 2 - | | | | | | |
| | Function divergence of duplicated genes | JT Wang <i>et al.</i> , BMC Genomics , 2012 Copyright © Peking Universit | 0 + | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |

Summary Question

• It could be argued that the feature selection is not necessary since the SVM can just work with hundreds of features. What do you think? Explain.

生物信息学:导论与方法 Bioinformatics: Introduction and Methods

