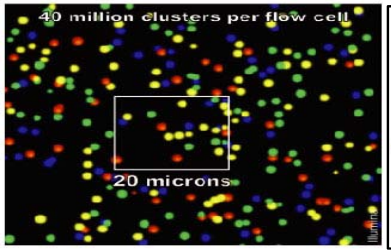
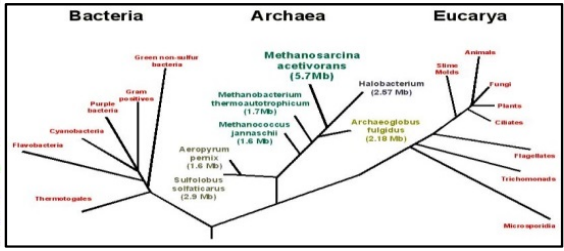
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



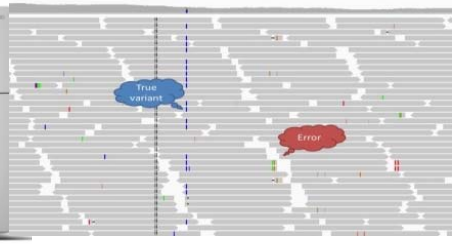


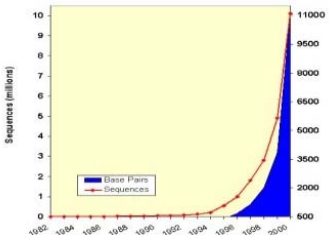
Transcriptome Analysis with noncoding RNAs

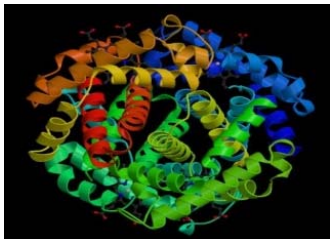
北京大学生物信息学中心 高歌

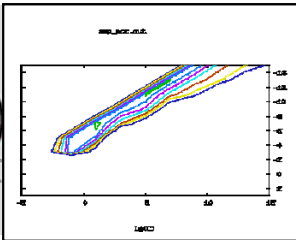
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

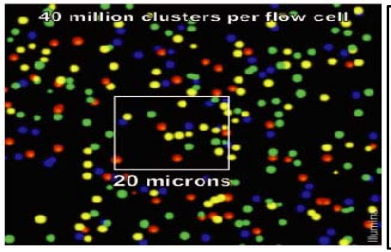


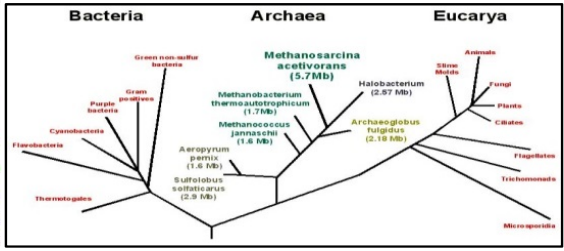






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



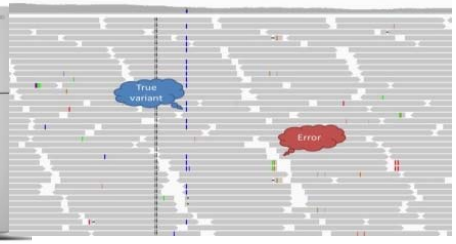


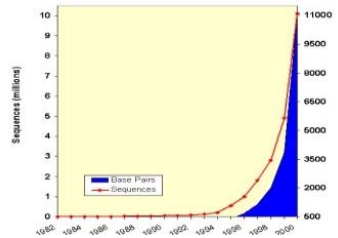
Unit 2:
Data Mining: Identify long ncRNAs

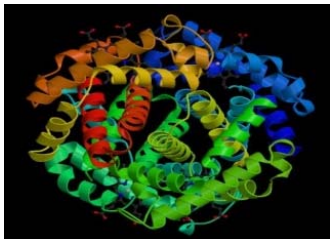
北京大学生物信息学中心 高歌

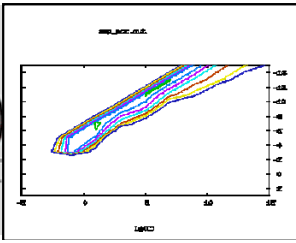
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University



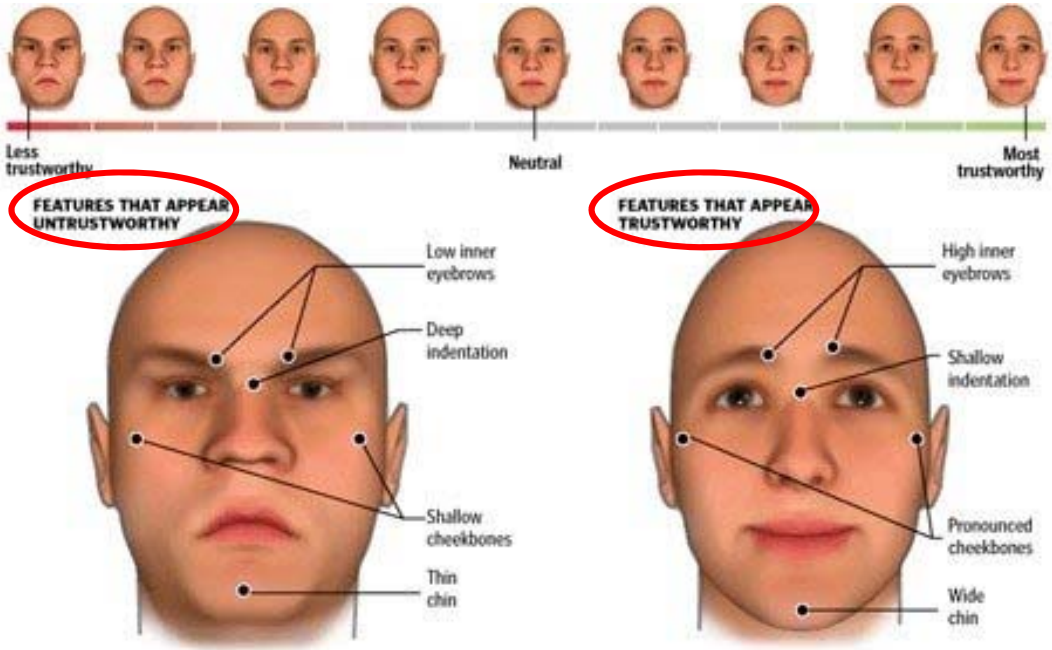






Identification

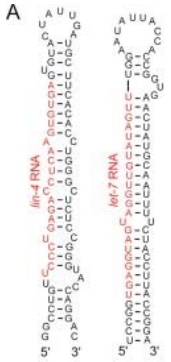
(Source: www.lkalop.com)



(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



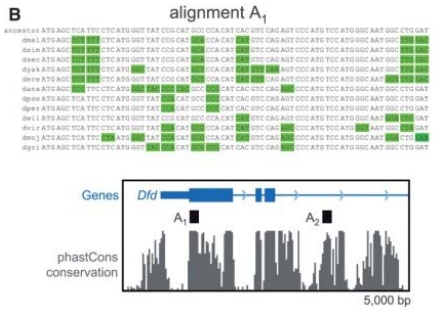
(*Cell* 116:281)

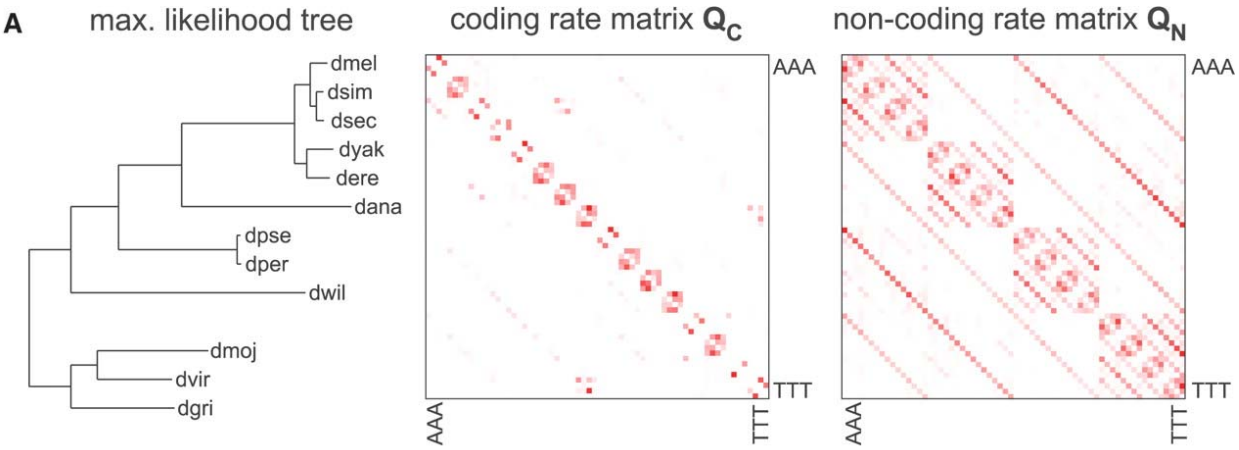
Table 1. Comparison of some filter-based approaches to miRNA gene finding in animals

	Initial set	Structural criteria	Conservation criteria	Additional filters
Grad *et al.* (50)	Stem–loop structures in repeats-masked intergenic regions	MFE, GC content, matches, mismatches, gaps and occurrence of multi-loops	Homologous stem–loops transitively identified in two additional genomes	Hairpins containing short repeats or with low quality structure are eliminated
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 additional genome	Log-odds score for several 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, asymmetric loops and bulges applied to predicted structures in aligned regions	Typical divergence pattern	

(*Nucl. Acids Res*, 37(8):2419)

Evolutionary features





(*Bioinformatics* 27:i275)

Sequence features only

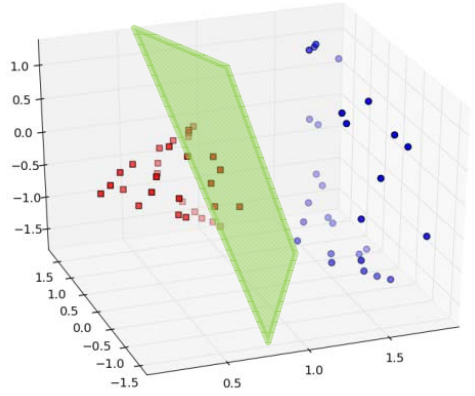
Mechanism neutral: works for both long and small ncRNAs

Accurate and Fast

SVM classifier

- **SVM – support vector machine**

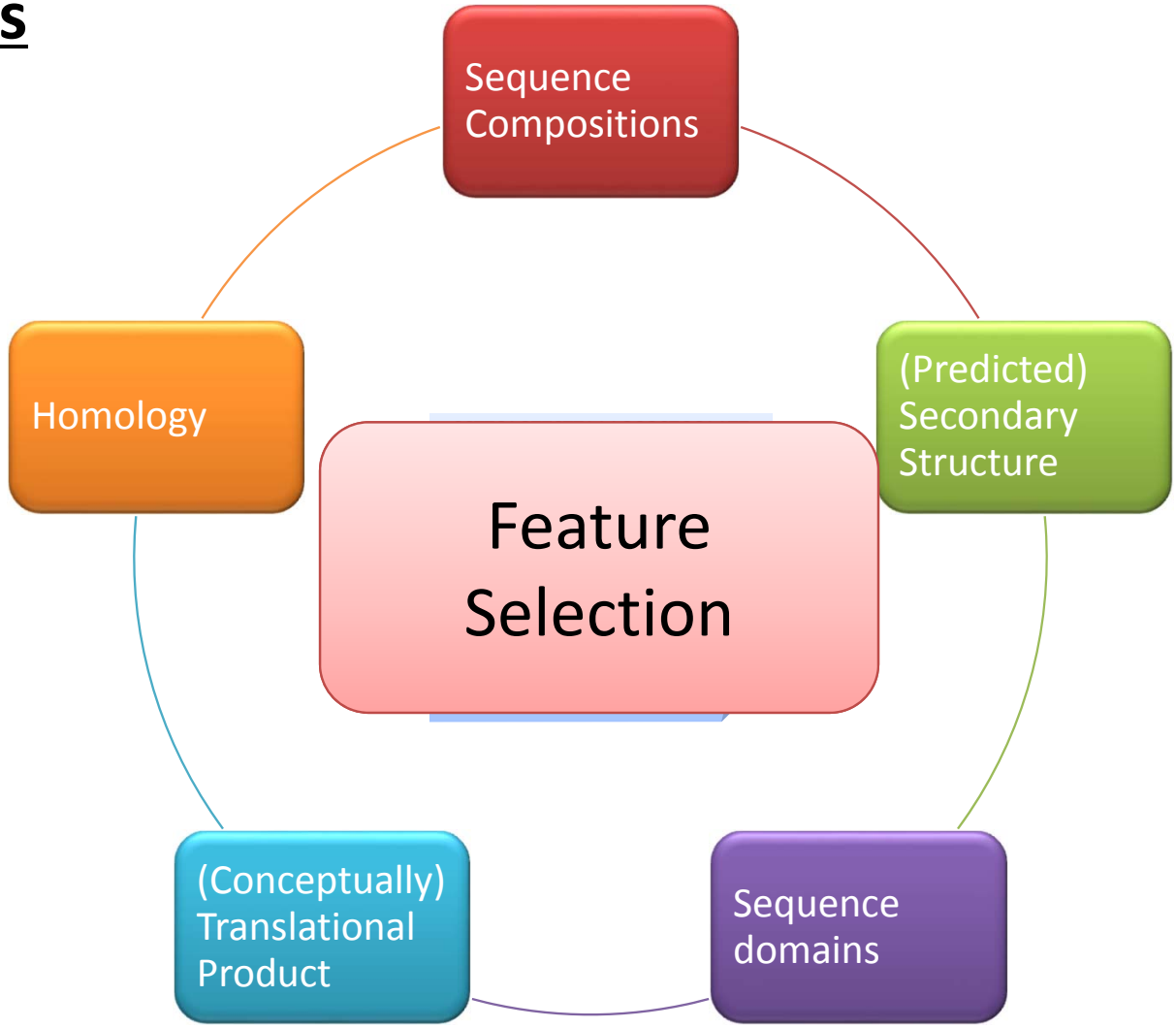
  Separate transformed data with a hyper plane in a high-dimensional space



- **Kernel function – Radial Basis Function(RBF)**

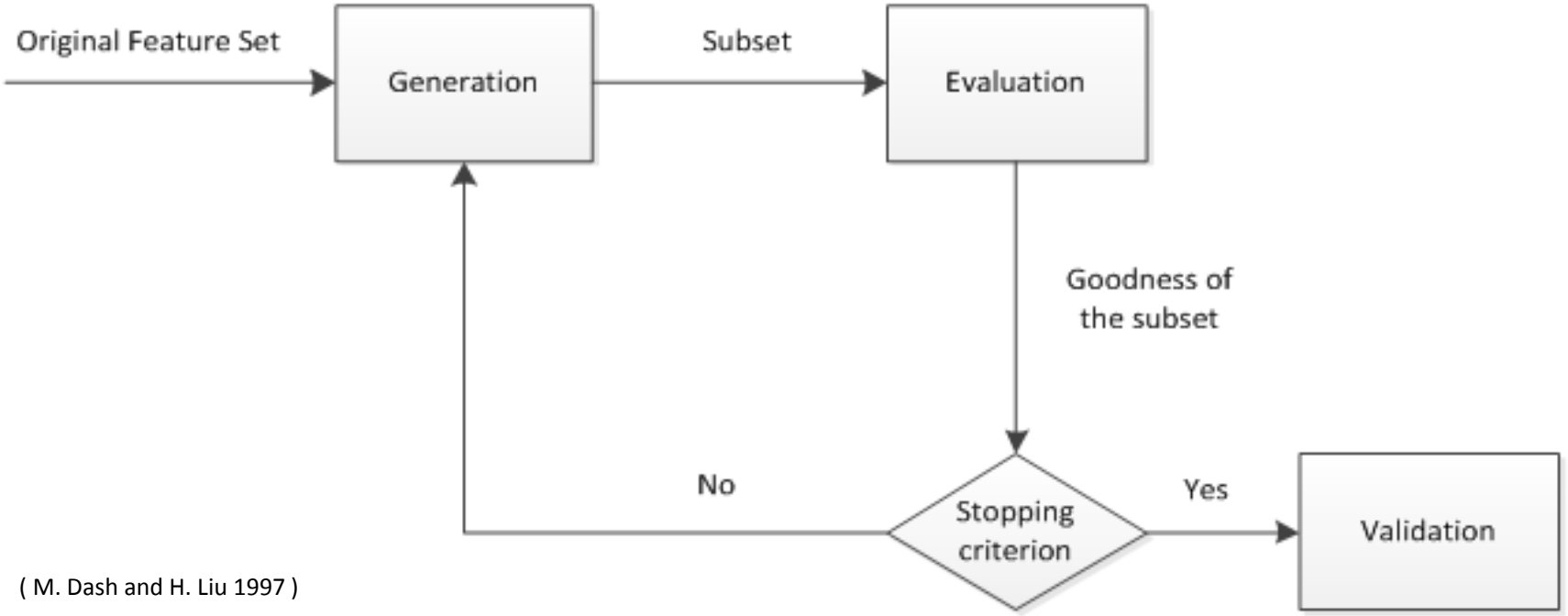
- **Grid-search to select proper values of parameter**

Sequence features



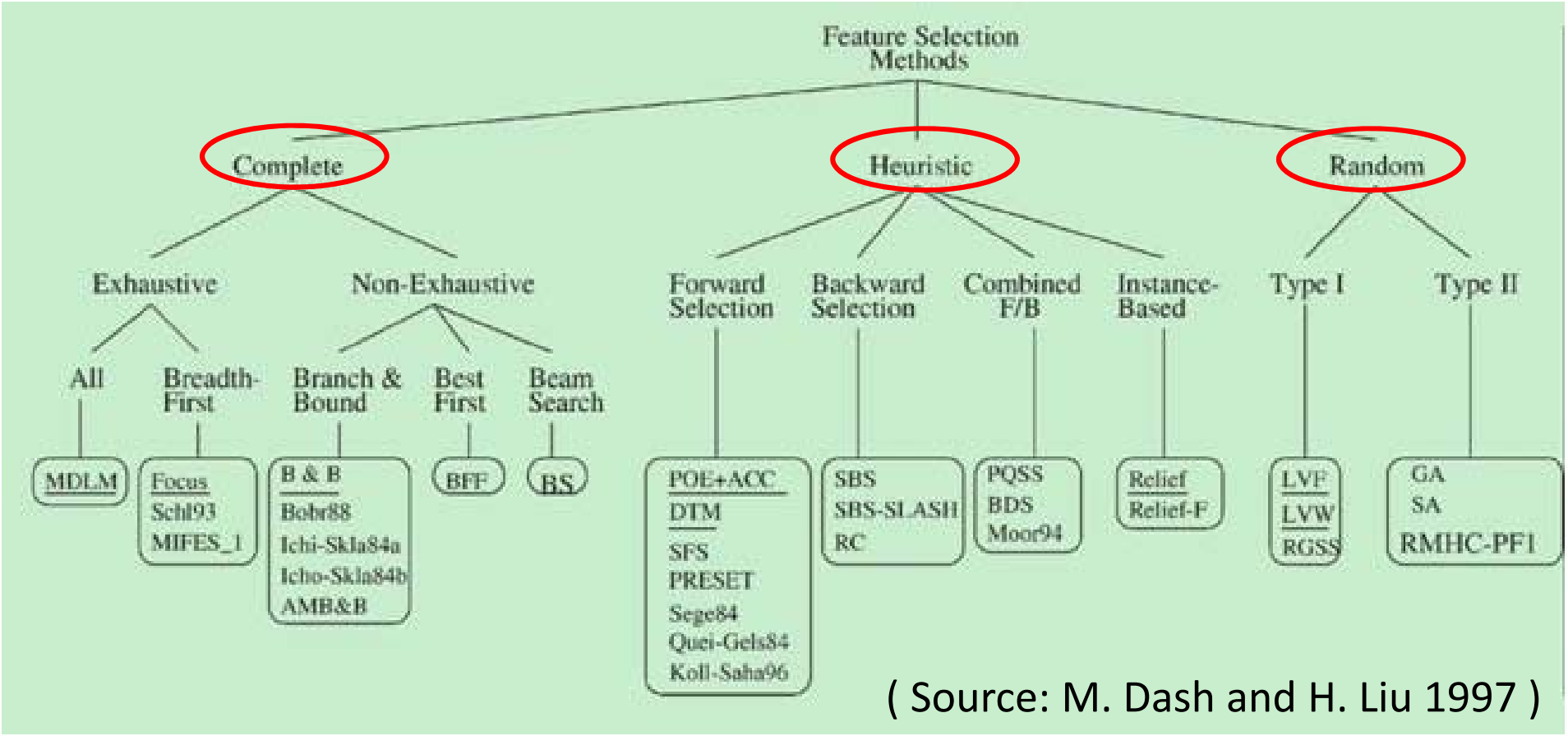
Feature Selection

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

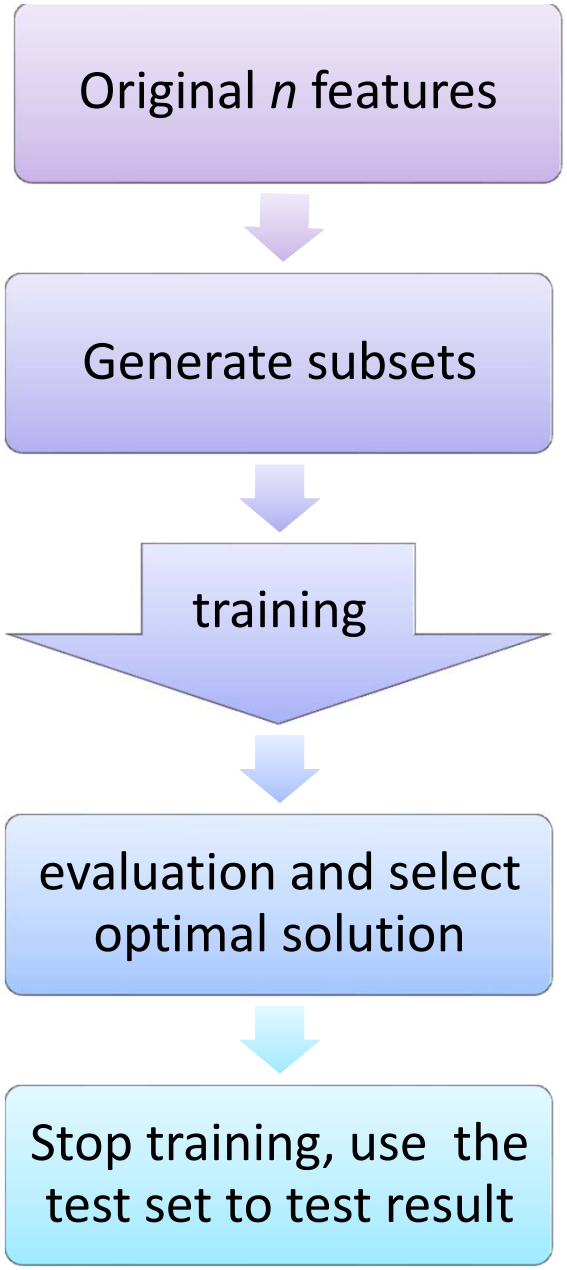


(M. Dash and H. Liu 1997)

Find The Optimal Subset



(Source: M. Dash and H. Liu 1997)

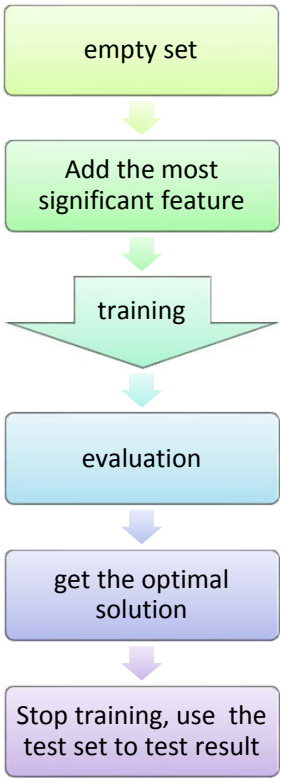


Complete Search: Breadth First

The breadth-first traversal of all variables

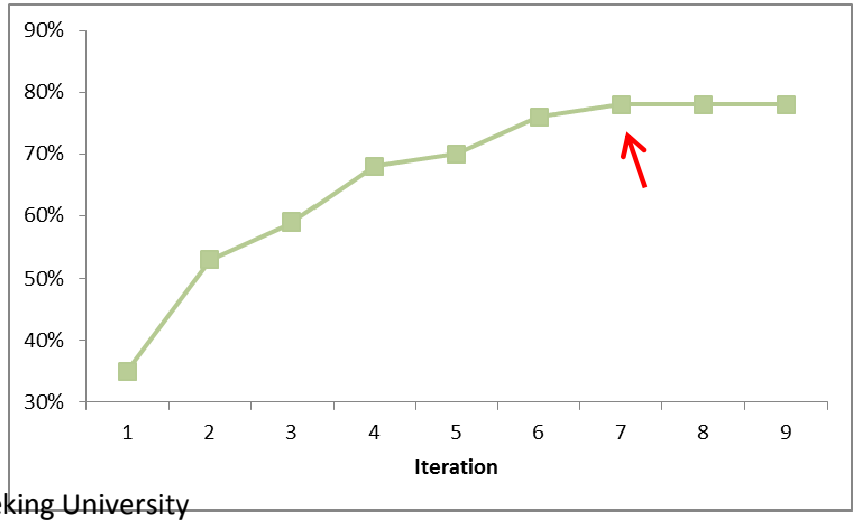
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

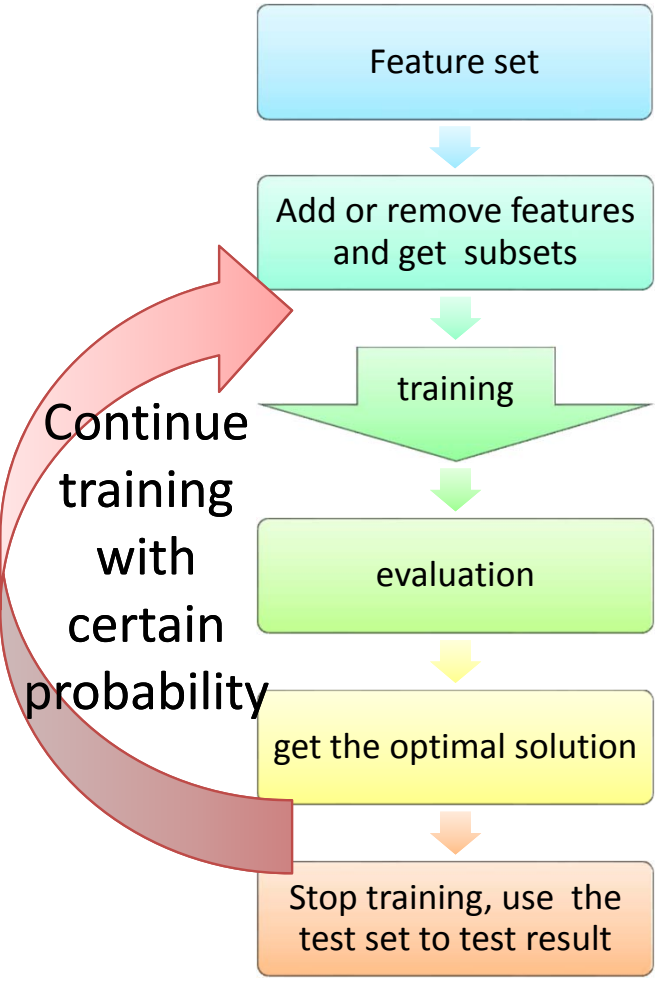
$$\binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{n-1} + \binom{n}{n}$$



Heuristic Search: Sequential Forward Selection

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

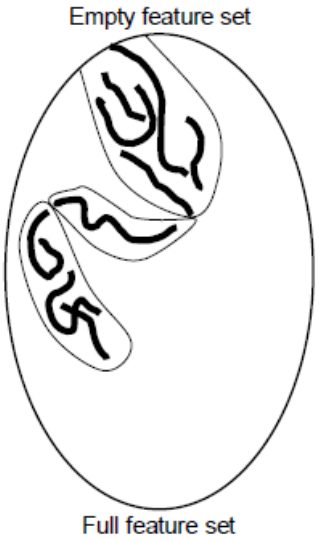




Random Search: Simulated Annealing

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

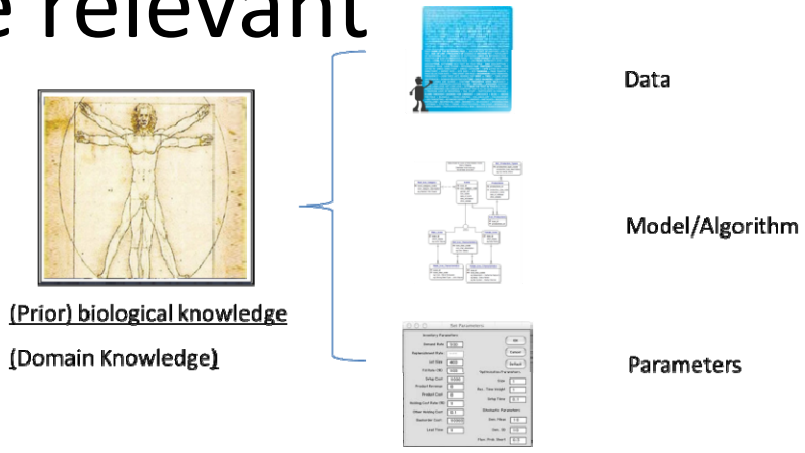
1. Determine an annealing schedule T(i)
2. Create an initial solution Y(0)
3. While $T(i)>T_{MIN}$
   3a. Generate a new solution Y(i+1) which is a neighbor of Y(i)
   3b. Compute $\Delta E= - [J(Y(i+1)) - J(Y(i))]$
   3b. If $\Delta E<0$
       then
           always accept the move from Y(i) to Y(i+1)
       else
           accept the move with probability $P=\exp(-\Delta E/T(i))$

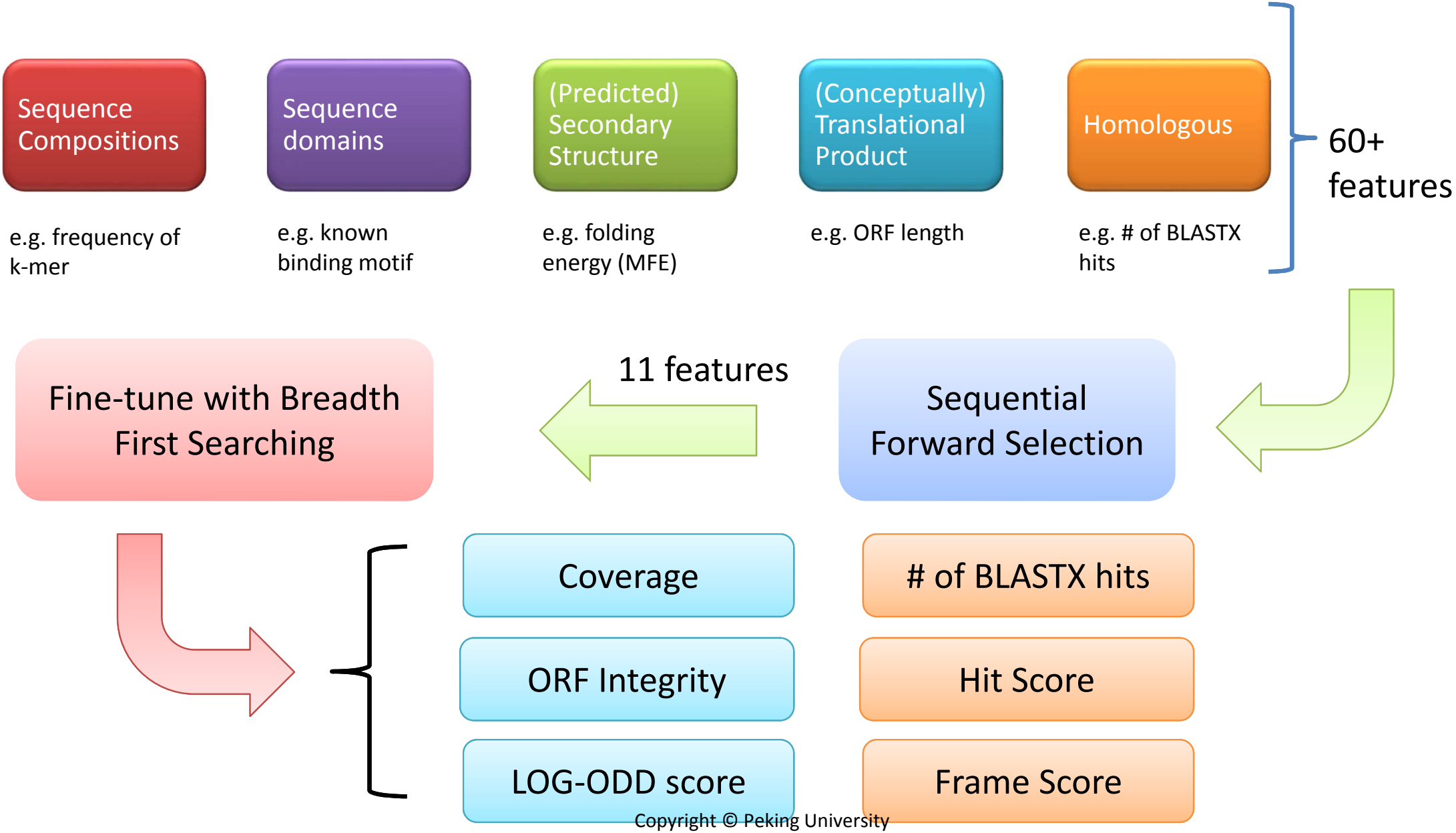


(Source: http://courses.cs.tamu.edu/rgutier/cs790_w02/l17.pdf)

Initialized feature set

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





(Conceptually) Translated Product

Coverage

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

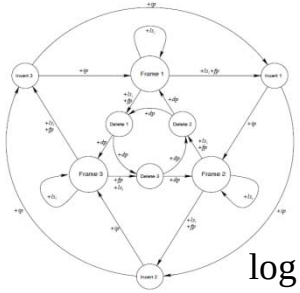
ORF Integrity

indicates whether the predicted ORF begins with a start codon and ends with an in-frame stop codon

```
ATG CCG GCT TAC CAC TCT TCT CTC ATG GAT CCT GAT ACC AAA TAG
 M   P   A   Y   H   S   S   L   M   D   P   D   T   K   *
```

LOG-ODD score

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



$$\log \frac{\Pr(D \mid M)}{\Pr(D \mid R)}$$

(*Nucleic Acids Res.* 35:W345)

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 $S_i = \underset{j}{\text{mean}}\{-\log_{10} E_{ij}\}, \quad i \in [0,1,2]$

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

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

$$\text{FRAME SCORE} = \underset{i\in\{0,1,2\}}{variance}\{S_i\} = \frac{\sum_{i=0}^{2}(S_i - \bar{S})^2}{2}$$

(*Nucleic Acids Res.* 35:W345)

Coverage

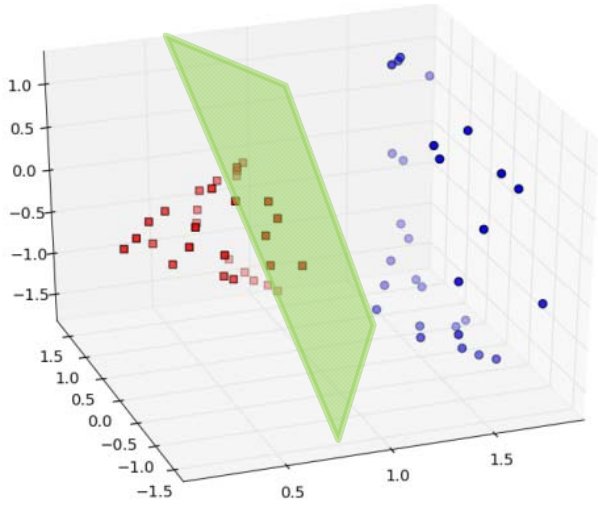
# of BLASTX hits

ORF Integrity

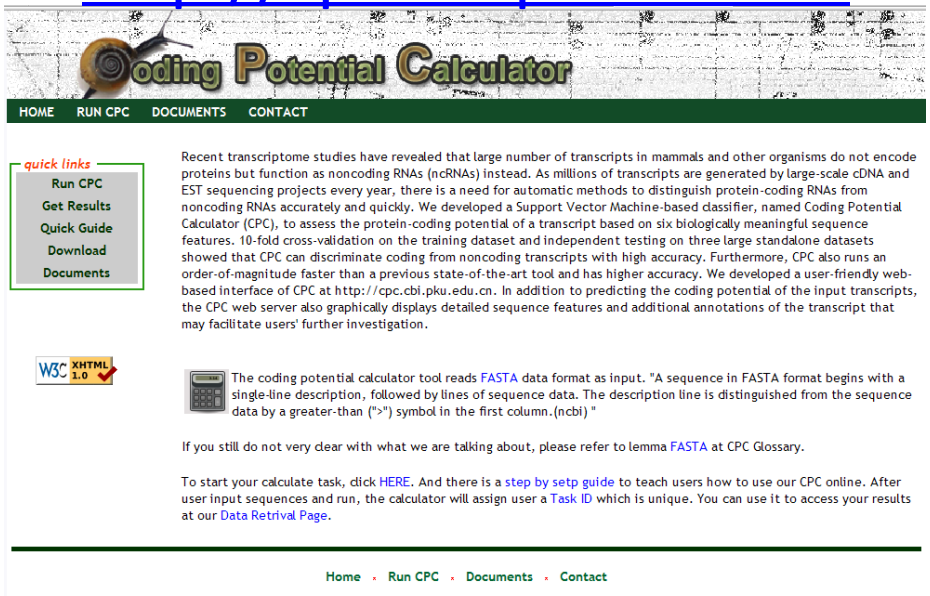
Hit Score

LOG-ODD score

Frame Score

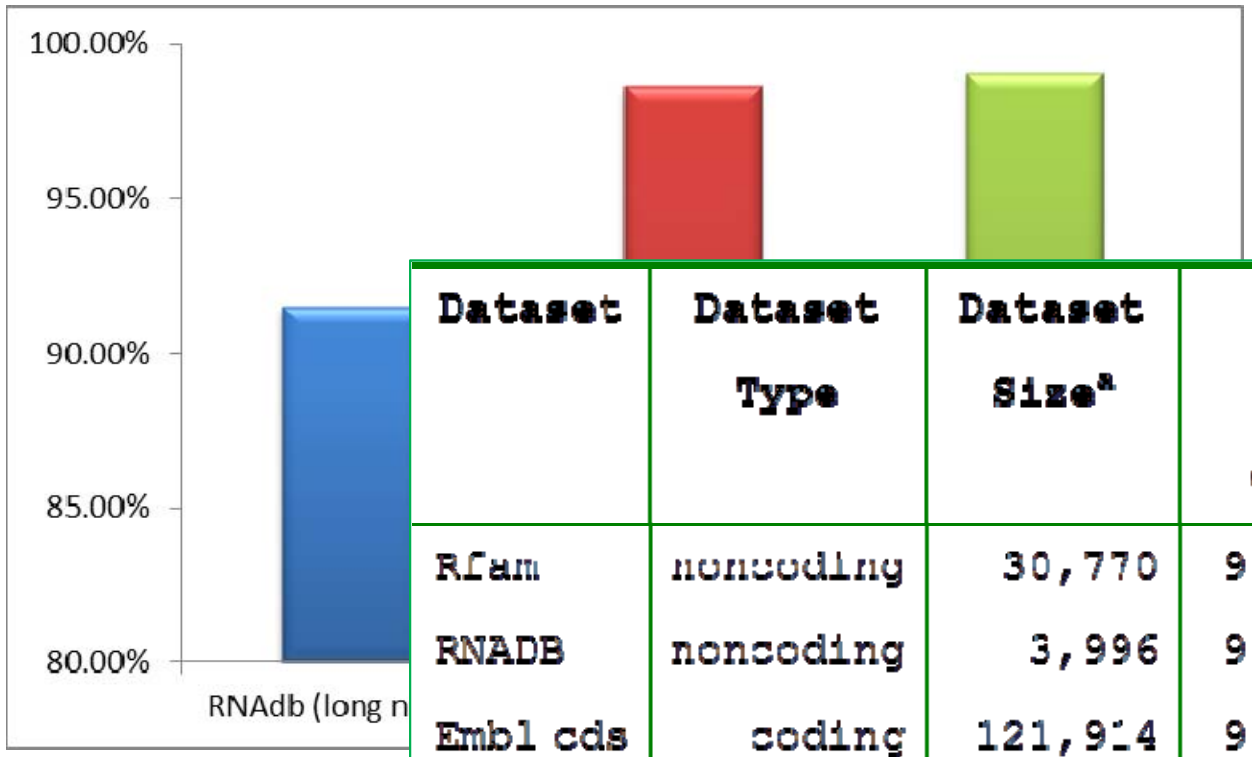


http://cpc.cbi.pku.edu.cn



Coding Potential Calculator

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



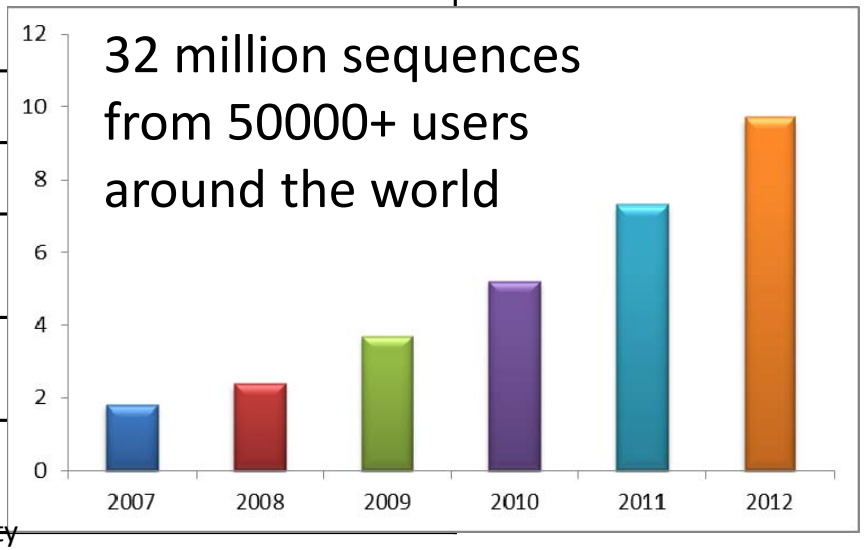
Dataset	Dataset Type	Dataset Size[a]	Accuracy		Time (in minutes)	
			CPC	CONC	CPC	CONC
Rfam	noncoding	30,770	98.62%	97.12%	3,513	46,376
RNADB	noncoding	3,996	91.50%	85.44%	598	7,322
Embl cds	coding	121,914	99.08%	98.70%	69,116	826,210 [b]

(*Nucleic Acids Res.* 35:W345)

Coding Potential Calculator

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

Gene Regulation	Function of ncRNA	H Van Bakel *et al.* **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
Disease	Heart diseases	JH Lee *et al*., **Circ Res**, 2011
	Cancer Marker	BP Mello *et al.*, **Nucleic Acid Res**, 2009
	Tumor mechanism	AC Tahira *et al.*, **Molecular Cancer**, 2011 RJ Flockhart *et al.*, **Genome Res**, 2012
Evolution	New genes	D Rose *et al*., **J Bioinform Compt Bio.**, 2008 JF Sousa *et al.*, **PLoS One**, 2010
	Function divergence of duplicated genes	JT Wang *et al.*, **BMC Genomics**, 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