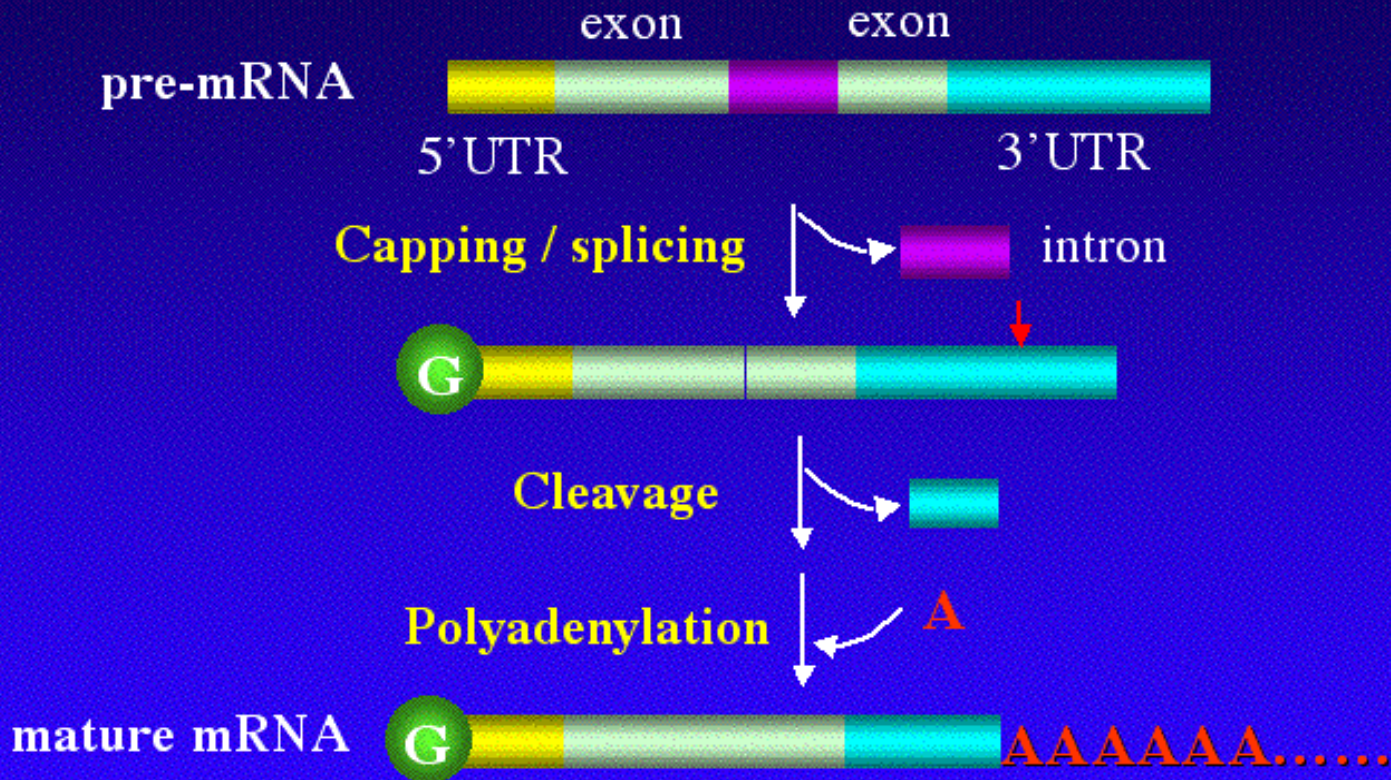


Recognition Of Polyadenylation Sites From Arabidopsis Genomic Sequences

Chuan Hock KOH
Limsoon WONG

Eukaryotic pre-mRNA processing



Other Approach

- PASS (Polyadenylation Site Sleuth)
 - based on Generalized Hidden Markov Model
 - Available for download at www.polya.org
 - Published in BMC Bioinformatics Feb 2007

My Approach

- Overview
 1. Feature Generation
 2. Feature Selection
 3. Feature Integration
 4. Cascade Classification
- Uses the first 3 step
 - TIS of Human (ATG)
 - PolyA site of Human (AATAAA or its slight variant)

1) Feature Generation

- 1-gram (A, C, G, U) - 4
 - 2-gram (AA, AC, ..., GU, UU) - 16
 - 3-gram (AAA, AAC, ..., UGU, UUU) - 64
 - 4U/1N (NUUUU, UNUUU, ..., UUUUN) - 1
 - 4A/1N (NAAAA, ANAAA, ..., AAAAN) - 1
 - G/U*7 (A stretch of G or U for 7 bp) - 1
- Total: 87
- Windows: (-110/+5), (-35/+15) and (-50/+30)
 - Total # in 3 window: $87 \times 3 = 261$

2) Feature Selection

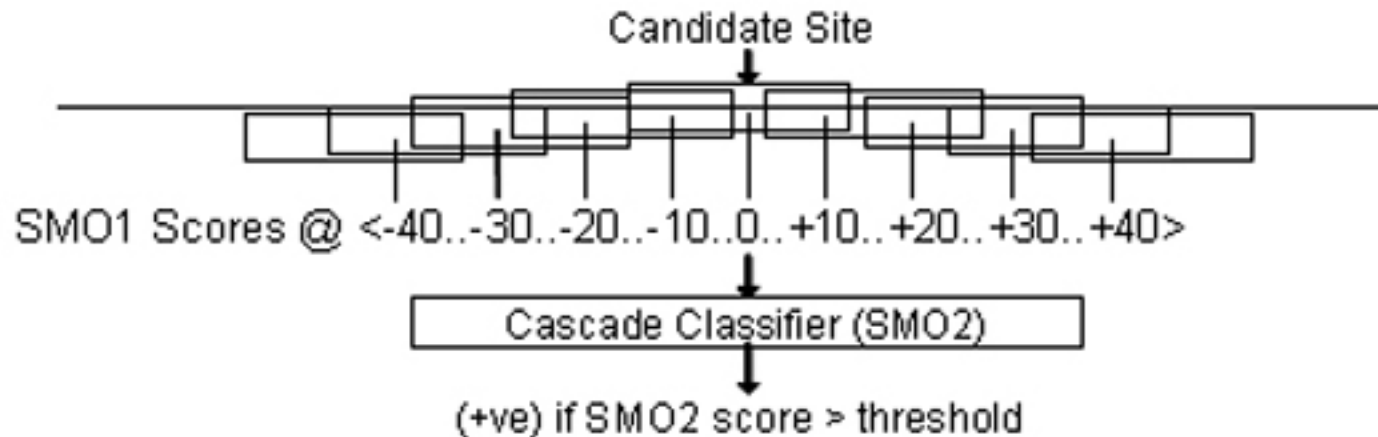
- Chi-squared with threshold 0 in WEKA
- WEKA – University of Waikato
- 228 out of 261 selected
- UP => (-110/+5)
- CLOSE => (-35/+15)
- DOWN => (-50/+30)
- Top 10 ranked features
 - 1) UP_T
 - 2) DOWN_G
 - 3) UP_G
 - 4) UP_TA
 - 5) CLOSE_G
 - 6) DOWN_TA
 - 7) UP_TGT
 - 8) UP_TT
 - 9) UP_GG
 - 10) DOWN_GG

3) Feature Integration

- SVM from WEKA
 - WEKA implementation of support vector machine using John Platt's sequential minimal optimization algorithm
- SMO1

4) Cascade Classification

- SMO2 – SVM from WEKA
- Uses 81 features
 - (-40/+40) SMO1 scores of a candidate site



Results

Table 1. Equal-error-rate points of SMOA, SMO2, and PASS 1.0 for SN_0.

SN_0	SMO A		SMO 2		PASS 1.0	
	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
Control Sequences						
CDS	91.1%	0.33	94.3%	0.24	95.3%	3.76
5'UTR	79.3%	0.50	84.9%	0.48	77.7%	5.53
Intron	63.9%	0.68	71.1%	0.68	62.8%	6.36

Table 2. Equal-error-rate points of SMOA, SMO2, and PASS 1.0 for SN_10.

SN_10	SMO A		SMO 2		PASS 1.0	
	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
Control Sequences						
CDS	94.8%	0.42	96.5%	0.31	96.5%	4.02
5'UTR	85.8%	0.61	89.2%	0.60	80.7%	5.81
Intron	72.5%	0.75	78.8%	0.76	67.7%	6.62

Table 3. Equal-error-rate points of SMOA, SMO2, and PASS 1.0 for SN_30.

SN_30	SMO A		SMO 2		PASS 1.0	
	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
Control Sequences						
CDS	97.1%	0.50	97.5%	0.37	97.5%	4.29
5'UTR	89.8%	0.69	91.5%	0.67	84.0%	6.13
Intron	79.2%	0.81	83.0%	0.81	71.7%	6.85

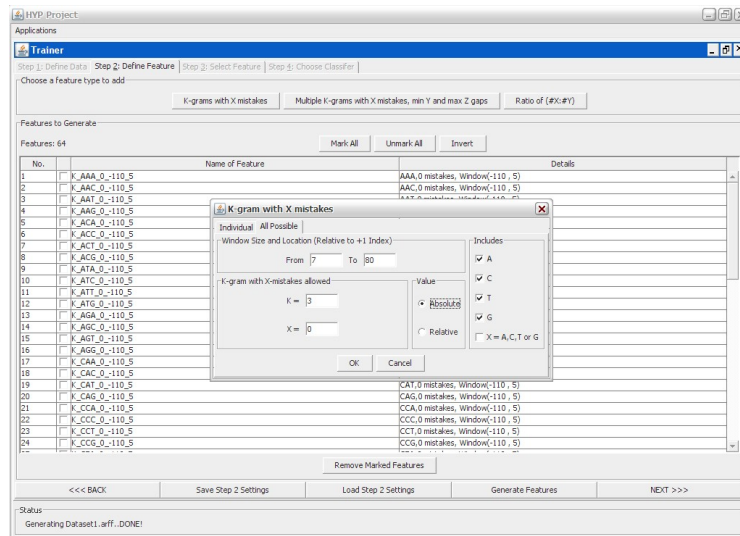
Conclusion

- Outperformed PASS 1.0 by 7 – 11% in most cases
- Cascade Classification step helped increase sensitivity and specificity

<http://www.comp.nus.edu.sg/~wongls/projects/dnafeatures/giw07-supplement/>

Currently..

- Embed the 4-step methodology into software package
- Ability to build prediction system for any functional site of any organism with just a few mouse clicks



HYP Project
Applications
Trainer
Step 1: Define Data | Step 2: Define Feature | **Step 3: Select Feature** | Step 4: Choose Classifier

Choose a Feature type to add
K-grams with X mistakes Multiple K-grams with X mistakes, min Y and max Z gaps Ratio of (#X:#Y)

Features to Generate
Features: 64

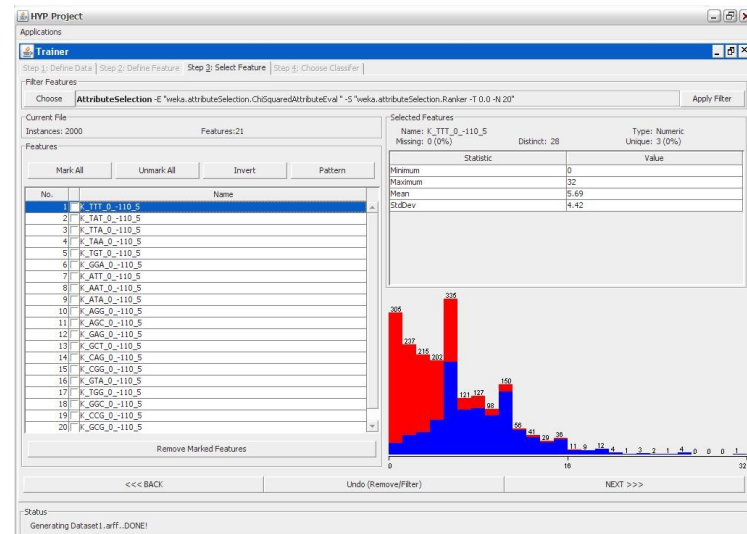
No.	Name of Feature	Details
1	K_AAA_0_110_5	AAA,0 mistakes, Window(110, 5)
2	K_AAC_0_110_5	AAC,0 mistakes, Window(110, 5)
3	K_AAT_0_110_5	AAT,0 mistakes, Window(110, 5)
4	K_AAG_0_110_5	AAG,0 mistakes, Window(110, 5)
5	K_ACA_0_110_5	AAC,0 mistakes, Window(110, 5)
6	K_ACC_0_110_5	ACC,0 mistakes, Window(110, 5)
7	K_ACT_0_110_5	ACT,0 mistakes, Window(110, 5)
8	K_ACS_0_110_5	ACS,0 mistakes, Window(110, 5)
9	K_ATA_0_110_5	ATA,0 mistakes, Window(110, 5)
10	K_ATC_0_110_5	ATC,0 mistakes, Window(110, 5)
11	K_ATT_0_110_5	ATT,0 mistakes, Window(110, 5)
12	K_ATG_0_110_5	ATG,0 mistakes, Window(110, 5)
13	K_AGA_0_110_5	AGA,0 mistakes, Window(110, 5)
14	K_AGC_0_110_5	AGC,0 mistakes, Window(110, 5)
15	K_AGT_0_110_5	AGT,0 mistakes, Window(110, 5)
16	K_AGG_0_110_5	AGG,0 mistakes, Window(110, 5)
17	K_CAA_0_110_5	CAA,0 mistakes, Window(110, 5)
18	K_CAC_0_110_5	CAC,0 mistakes, Window(110, 5)
19	K_CAT_0_110_5	CAT,0 mistakes, Window(110, 5)
20	K_CAG_0_110_5	CAG,0 mistakes, Window(110, 5)
21	K_CCA_0_110_5	CCA,0 mistakes, Window(110, 5)
22	K_CCC_0_110_5	CCC,0 mistakes, Window(110, 5)
23	K_CCT_0_110_5	CCT,0 mistakes, Window(110, 5)
24	K_CCG_0_110_5	CCG,0 mistakes, Window(110, 5)

K-gram with X mistakes
Individual All Possible
Window Size and Location (Relative to +1 Index)
From 7 To 80
K = 3
X = 0
Includes: A, C, T, G
X = A,C,T or G

Remove Marked Features

<<< BACK Save Step 2 Settings Load Step 2 Settings Generate Features NEXT >>>

Status
Generating Dataset1.arff..DONE!



HYP Project
Applications
Trainer
Step 1: Define Data | Step 2: Define Feature | Step 3: Select Feature | **Step 4: Choose Classifier**

Filter Features
Choose **AttributesSelection** "wela.attributeSelection.ChSquaredAttributeEval"-S"wela.attributeSelection.Ranker-T 0.0-N 20"
Apply Filter

Current File: Instances: 2000 Features: 21 Selected Features: Name: K_TTT_0_110_5 Type: Numeric
Missing: 0 (0%) Distinct: 28 Unique: 3 (0%)

Statistic	Value
Minimum	0
Maximum	32
Mean	5.69
StdDev	4.42

Mark All Unmark All Invert Pattern

No.	Name
1	K_TTT_0_110_5
2	K_TAT_0_110_5
3	K_TTA_0_110_5
4	K_TAA_0_110_5
5	K_TGT_0_110_5
6	K_TGA_0_110_5
7	K_TTT_0_110_5
8	K_AAT_0_110_5
9	K_ATA_0_110_5
10	K_AGC_0_110_5
11	K_AGC_0_110_5
12	K_GAG_0_110_5
13	K_GAT_0_110_5
14	K_CAG_0_110_5
15	K_CGG_0_110_5
16	K_GTA_0_110_5
17	K_TGG_0_110_5
18	K_GGC_0_110_5
19	K_CCG_0_110_5
20	K_GCG_0_110_5

Remove Marked Features

<<< BACK Undo (RemoveFilter) NEXT >>>

Status
Generating Dataset1.arff..DONE!

- Dataset A (used to set parameters):
 - 804 (+ve) sequences with EST-supported polyadenylation sites
 - 9742 (-ve) coding sequences
- Dataset B (used for SMO1 training):
 - 2640 (+ve) sequences with EST-supported polyadenylation sites
 - 900 (-ve) coding sequences
 - 476 (-ve) 5'UTR sequences
 - 954 (-ve) intronic sequences
- Dataset C (used for SMO2 training):
 - 1500 (+ve) sequences with EST-supported polyadenylation sites
 - 100 (-ve) coding sequences
 - 100 (-ve) 5'UTR sequences
 - 100 (-ve) intronic sequences
- Dataset D (used for SMOA and SMO2 testing):
 - 2069 (+ve) sequences with EST-supported polyadenylation sites
 - 501 (-ve) coding sequences
 - 288 (-ve) 5'UTR sequences
 - 527 (-ve) intronic sequences
- Dataset E (used for SMOA training):
 - 4140 (+ve) sequences with EST-supported polyadenylation sites
 - 1000 (-ve) coding sequences
 - 576 (-ve) 5'UTR sequences
 - 1054 (-ve) intronic sequences