NUS National University of Singapore

Recognition Of Polyadenylation Sites From Arabidopsis Genomic Sequences

Chuan Hock KOH
Limsoon WONG

Eukaryotic pre-mRNA processing

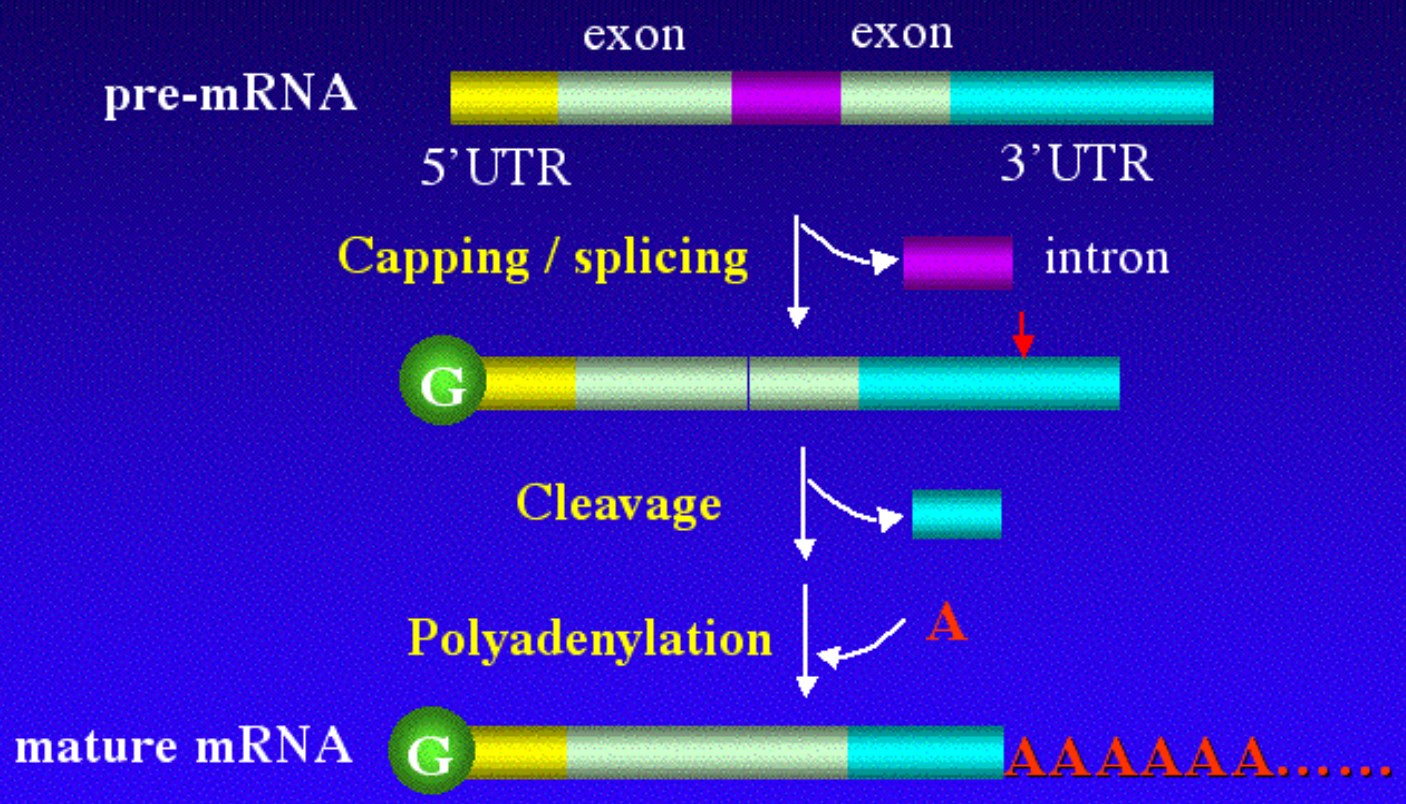


Image From: www.polya.org

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 x 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

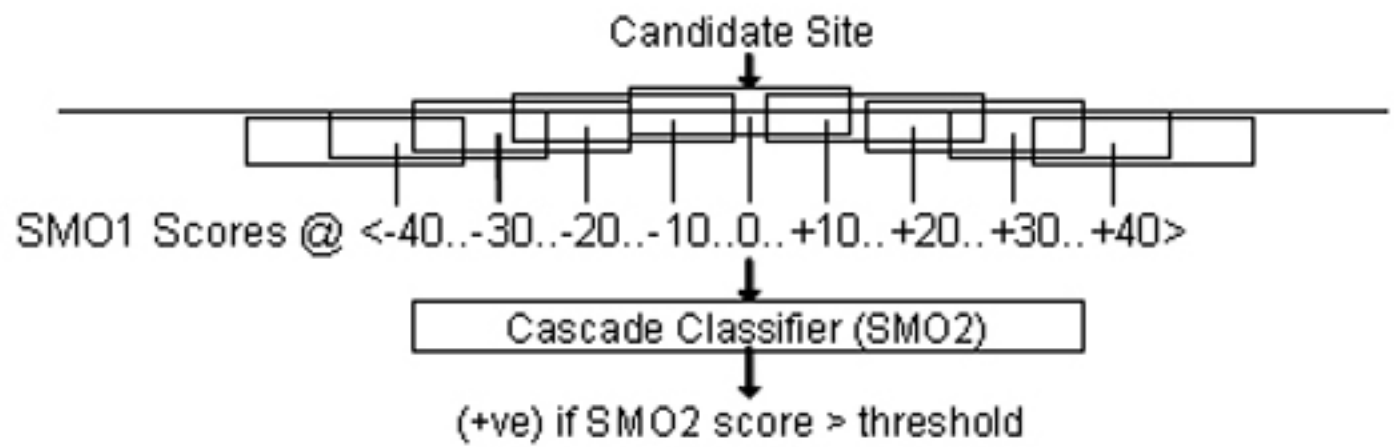
1) UP_T
2) DOWN_G
3) UP_G
4) UP_TA
5) CLOSE_G
6) DOWN_TA
7) UP_TGT
8) UP_TT
9) UP_GG
10) DOWN_GG

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	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
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	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
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	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
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

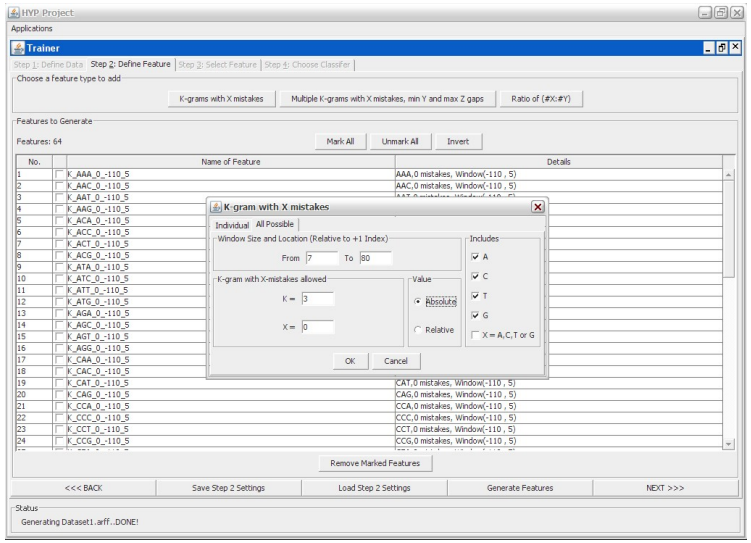
Conclusion

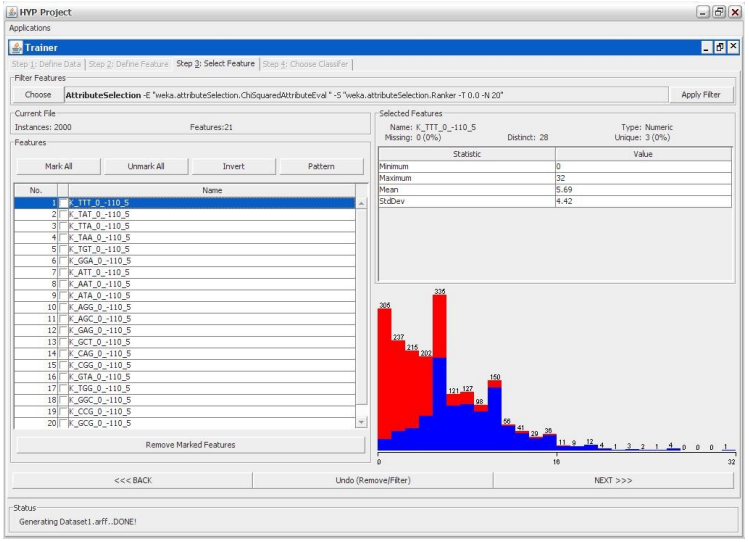
- Outperformed PASS 1.0 by 7 – 11% in most cases
- Cascade Classification step helped increase sensitivity and specificity

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





- 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