

#### Recognition of Polyadenylation Sites from Genomic Arabidopsis Sequences

Chuan Hock KOH Limsoon WONG



# Human Genome Project

- Completed in 2003
- Whats next?
  - Genomes to Life (GTL) Project
    - Understanding the detail mechanism of the cell
  - ENCODE Project
    - Annotating the whole Genome





# **Research Project**

- Phase 1
  - Develop a computational method to analyze (annotate) DNA sequences
- Phase 2
  - Make this method easy to use, especially for biologists with no or little computer expertise



## Phase 1

- Develop a 4 step methodology
  - 1. Feature Generation
  - 2. Feature Selection
  - 3. Feature Integration
  - 4. Cascade Classification
- 18<sup>th</sup> International Conference on Genome Informatics







#### 1. Feature Generation

- Thickness of Eyebrows, Hair Length, Hair Color, Jaw Line, Eye Color etc
- 1. Feature Selection
  - Reject Hair Length and Hair Color
- 1. Feature Integration
  - Use all selected features to judge



















- 1. Cascade Classifier
  - Looking at its surroundings



#### Phase 2

 To enable the Method to be deployed with ease

- Developed a software

pplications							
🔓 Trainer							- 0
itep <u>1</u> : Define Data   Step	2: Define Feature Step	p 3: Select Feature Ste	o <u>4</u> : Choose Level One Clas	ssifer			
Filter Features							
Choose Attributes	Selection -E "weka.attr	ibuteSelection.ChiSquare	dAttributeEval " -S "weka.a	attributeSelection.Ranker -T -1.7	976931348623157E308 -N -1		Apply Filter
Current File				Selected Features			
nstances: 1300 Features:65				Name: K_TTT_0110_5 Type: Numeric			
Features				Missing: 0 (0%)	Distinct: 28	Unique: 3 (0%)	
				Statistic		Value	
Mark Ali	Unmark All	Invert	Pattern	Minimum	0		
				Maximum	32		
No.		Name		Mean	6.935		
1 K_TTT_0	110_5		<u>.</u>	StdDev	4,596		
8 K_AGS_0.   9 K_AGC_0.   10 K_AGC_0.   11 K_GCT_0.   12 K_ATA_0.   13 K_CAG_0.   13 K_CAG_0.   14 K_CGG_0.   15 K_GGG_0.   16 K_TGG_0.   17 K_GGC_0.   18 K_GGC_0.   19 K_CCGC_0.   20 K_GCC_0.	-110_5 -110_5 -110_5 -110_5 -110_5 -110_5 -110_5 -110_5 -110_5 -110_5 -110_5 -110_5 -110_5 -110_5			216 220 118 109 122 104	60 00 00		
	Remove Mar	ked Features				16 1 <u>3 2 3</u>	
<<< BACK Undo (Ret				⊥ 0 move/Filter)	16 3		
SSS DAUN UNDO (Rer				movernicery		NEAT 222	



## Conclusion

- Personal Genome Project
  - Individual genome for < USD 1000</li>
- Develop more efficient ways to store, retrieve and 'crunch' data