

Gene Classification by Decision Tree Data Mining
Methods: Identification of Genes Likely to be
Involved in Human Genetic Disease

Julius Goth

ClarLynda Williams-Devane

Jihye Kim

North Carolina State University

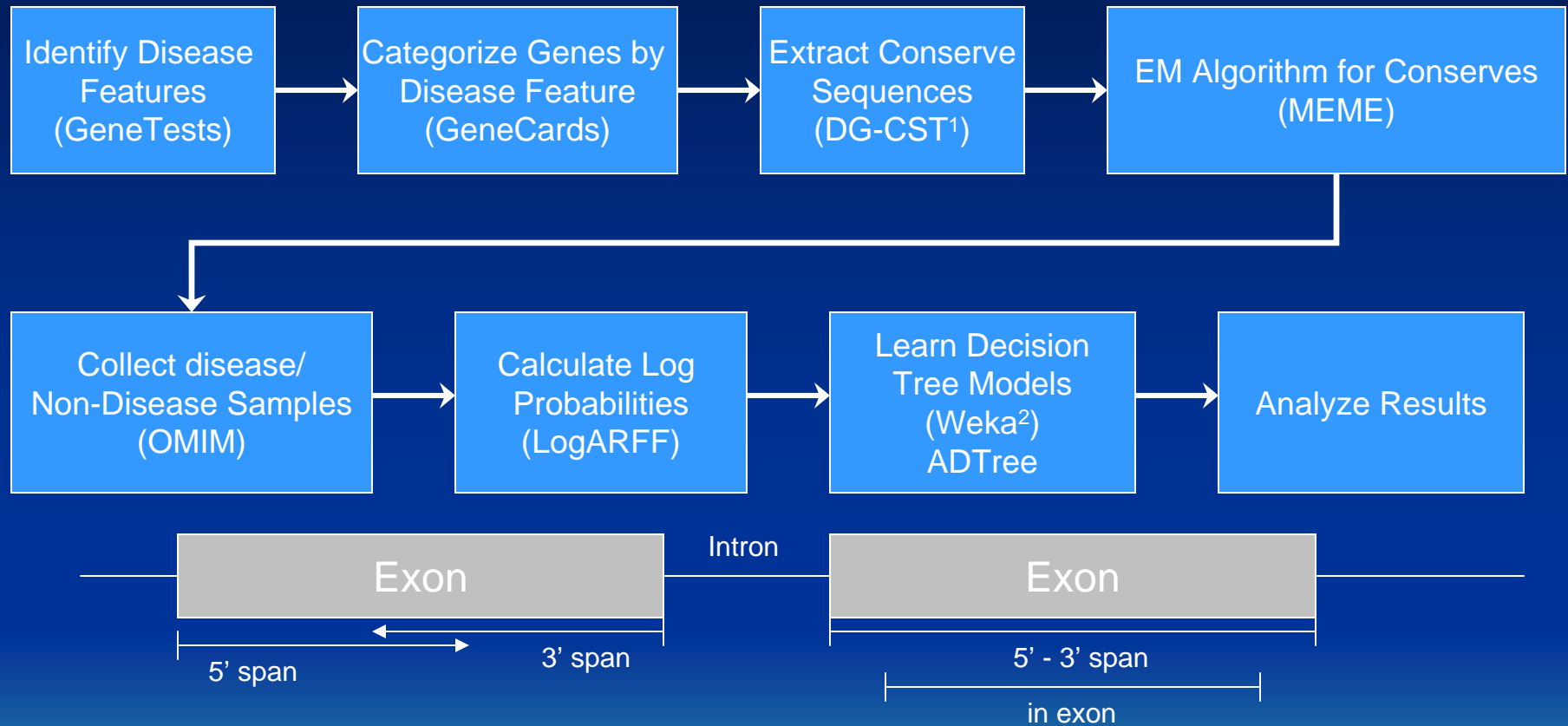


Introduction

- Task
 - Learn patterns from known disease features and Conserve Sequence Tags
- Goal
 - Classify newly sequenced genes as probable disease/non-disease related
- Strategy
 - Learn 69 decision trees
 - 23 “forests” (1 per chromosome)
 - Each consisting of 3 trees (1 per gene region)
 - Disease features serve as attributes (variable)
 - Outcome results from majority voting



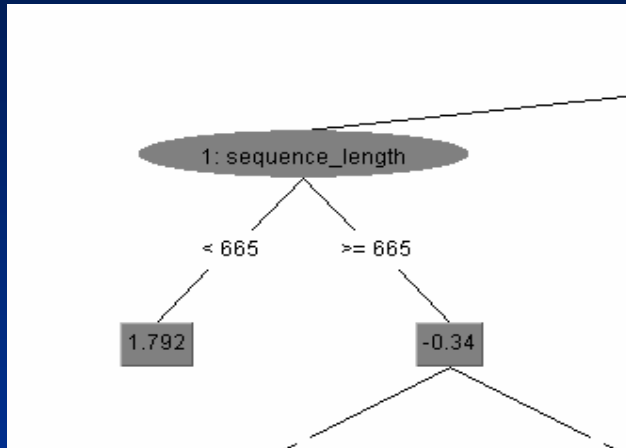
Method



¹ <http://dgcst.ceinge.unina.it>

² Weka: University of Waikato Environment for Knowledge Analysis

Conclusion



Average Sequence Length
 Diseased=60,824bp (std 69,889)
 Non-Diseased=36,774bp (std 108,135)

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.714	0.343	0.676	0.714	0.694	true
0.657	0.286	0.697	0.657	0.676	false

=== Confusion Matrix ===

```

a b <-- classified as
50 20 | a = true
24 46 | b = false
  
```

