

Gene Prediction: Accurate Prediction of Translation Initiation Sites

George Tzanis and Ioannis Vlahavas



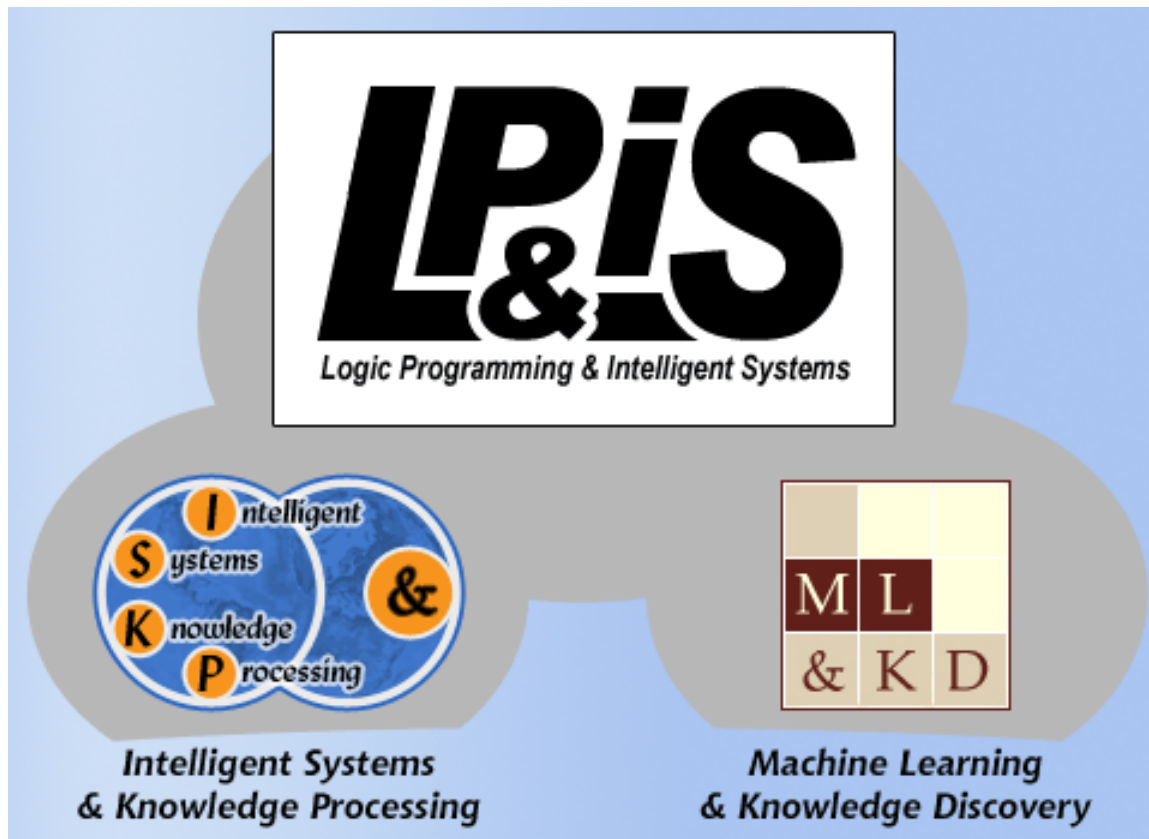
Machine Learning and Knowledge Discovery Group
Department of Informatics
Aristotle University of Thessaloniki



Outline

- Who We Are
- Background
- Our Work
- Results
- Conclusions and Future Directions

Who We Are



MLKD Group (1/2)

□ People

- Ioannis Vlahavas, Professor, Head of Group
- Grigorios Tsoumakas, Lecturer
- Christos Berberidis, PhD Student
- Ioannis Katakis, PhD Student
- Ioannis Partalas, PhD Student
- George Tzanis, PhD Student

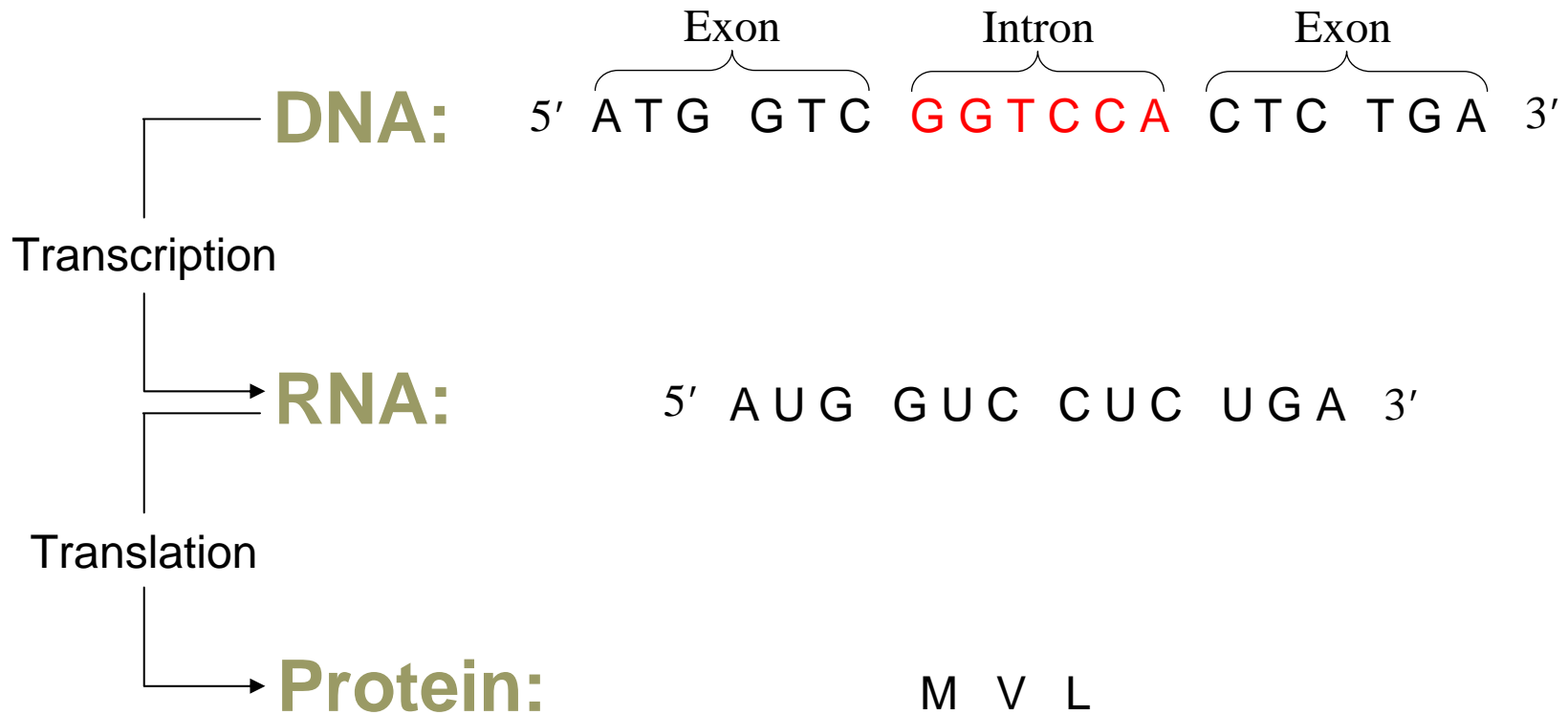
□ URL: <http://mlkd.csd.auth.gr>

MLKD Group (2/2)

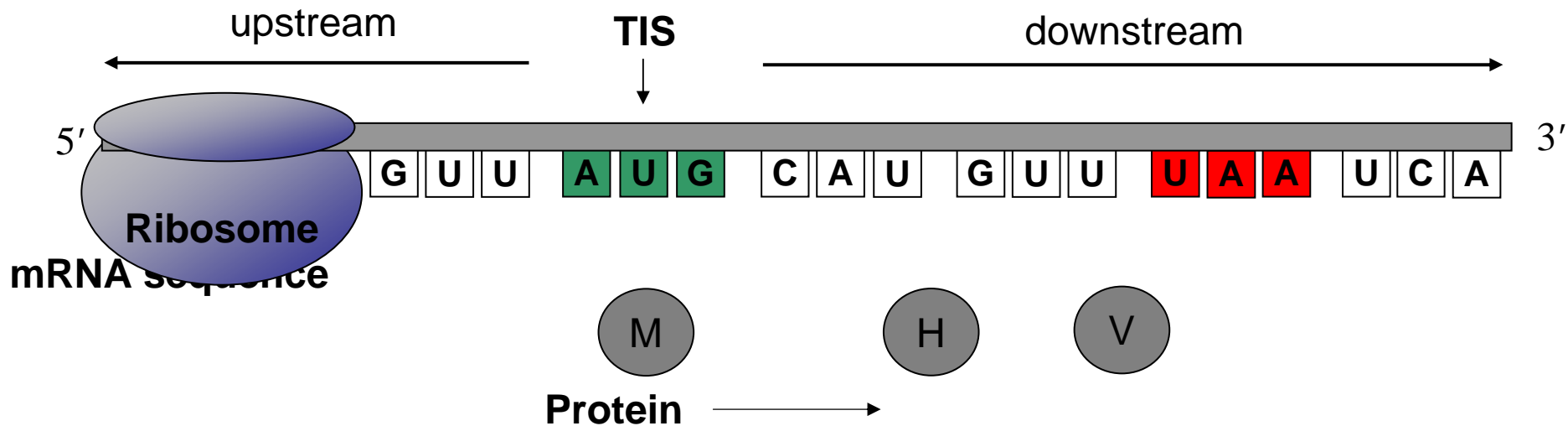
□ Research

- Knowledge Discovery from Biological Data
- Periodicity Detection in Temporal Sequences
- Learning for Planning
- Ensemble Methods - Ensemble Pruning
- Text Mining
- Reinforcement Learning
- Distributed Data Mining
- Multi-Label Classification

Central Dogma of Molecular Biology



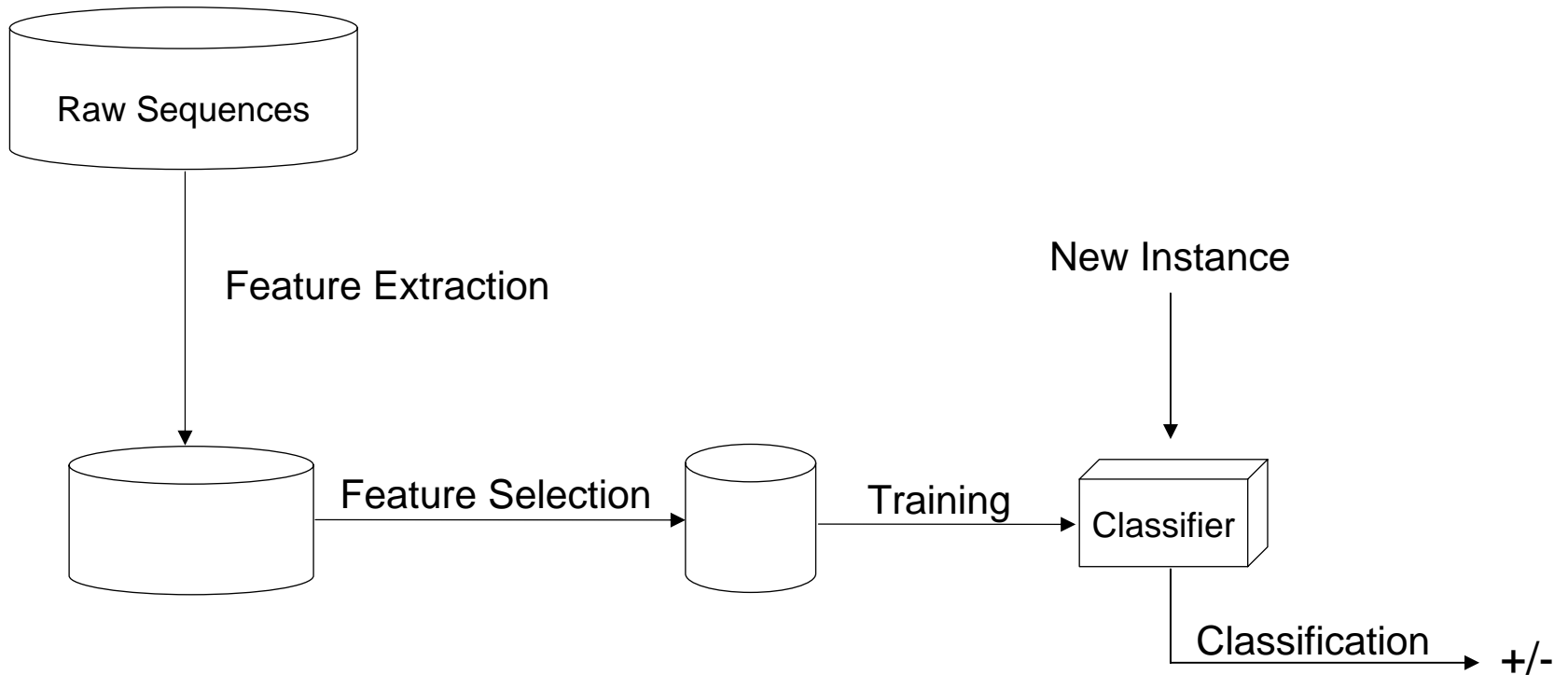
Translation



Problem Formulation

- Each cDNA sequence contains a TIS
- The TIS is almost always an ATG codon
- Problem
 - Which of the ATG codons is the TIS?
- Solution
 - Build classifiers that can learn to discriminate the TIS from the other ATG codons

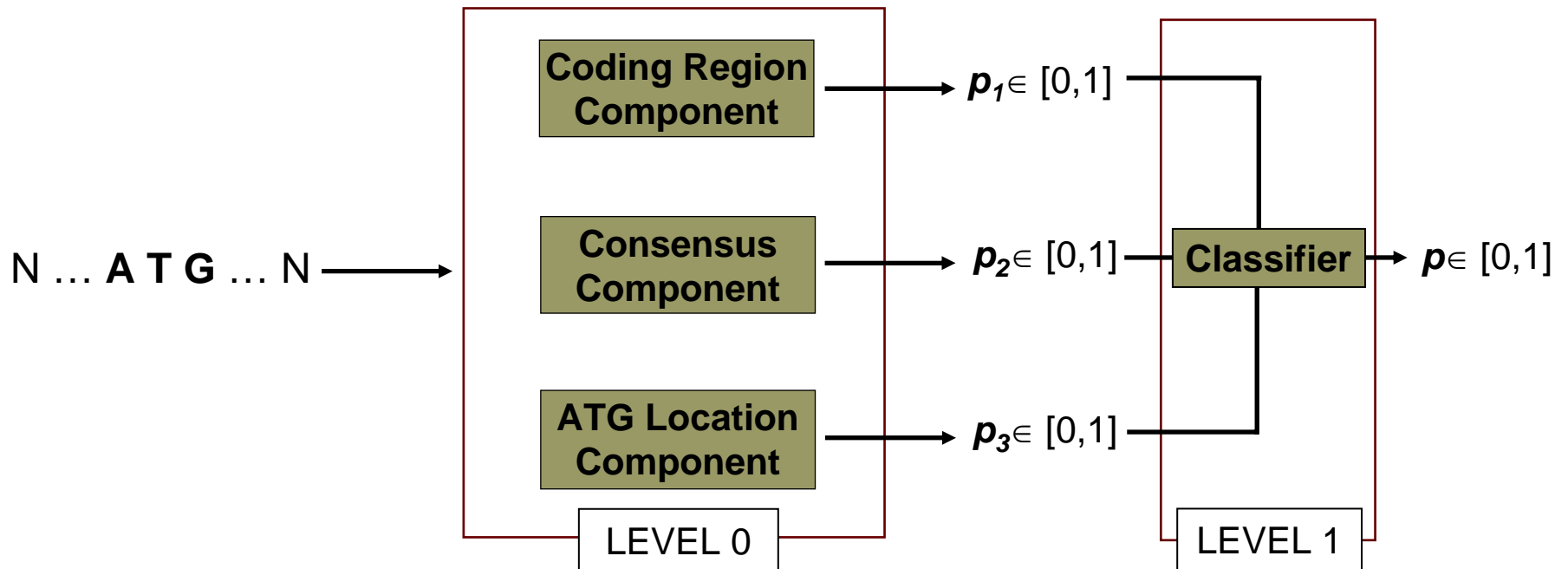
Basic Methodology



Classification Algorithms

- Artificial Neural Networks
- Support Vector Machines
- Gaussian Mixture Models
- Bayesian Methods
- Decision Trees

The MANTIS Methodology



The Coding Region Component (2/2)

□ Extracted Features

- Counts of amino acids upstream or downstream the ATG codon
- Chemical properties of amino acids
- Number of upstream ATGs
- Presence of downstream stop codon
- Difference between upstream and downstream counts
- Periodic occurrence of nucleotides inside each codon

The Consensus Component

5' N ... N $\overset{-7}{\boxed{\text{N N N N N N N N A T G N N}}}$ N ... N 3'

↓
**Subsequence
Extraction**

↓
**Model
Construction**

1st order homogeneous Markov chain
2nd order homogeneous Markov chain
1st order non-homogeneous Markov chain

↓
 $p_2 \in [0,1]$

The ATG Location Component

Distance Based

Distance of the ATG
from the 5' end

**ATG Location
Component**

$p_3 \in [0,1]$

Order Based

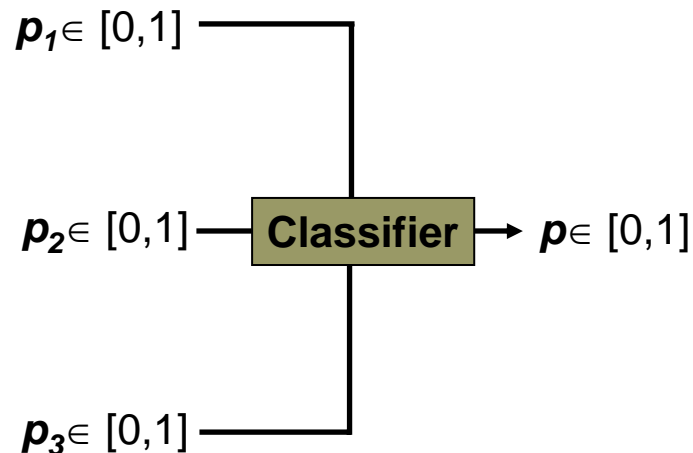
Order of the ATG
from the 5' end (i.e. 1st, 2nd, etc.)

**ATG Location
Component**

$p_3 \in [0,1]$

The Level-1 Classifier

- M5, a model tree classifier
- Multi-response Linear Regression (MLR) classifier



Experimental Setup

□ Datasets

- *A.aegypti* (262 cDNAs – 6453 ATGs)
- *A.thaliana* (523 cDNAs – 2048 ATGs)
- *H.sapiens* (480 cDNAs – 14108 ATGs)
- Vertebrates (3312 cDNAs – 13503 ATGs)

□ Evaluation Method

- Stratified 10-fold Cross Validation

Evaluation Metrics

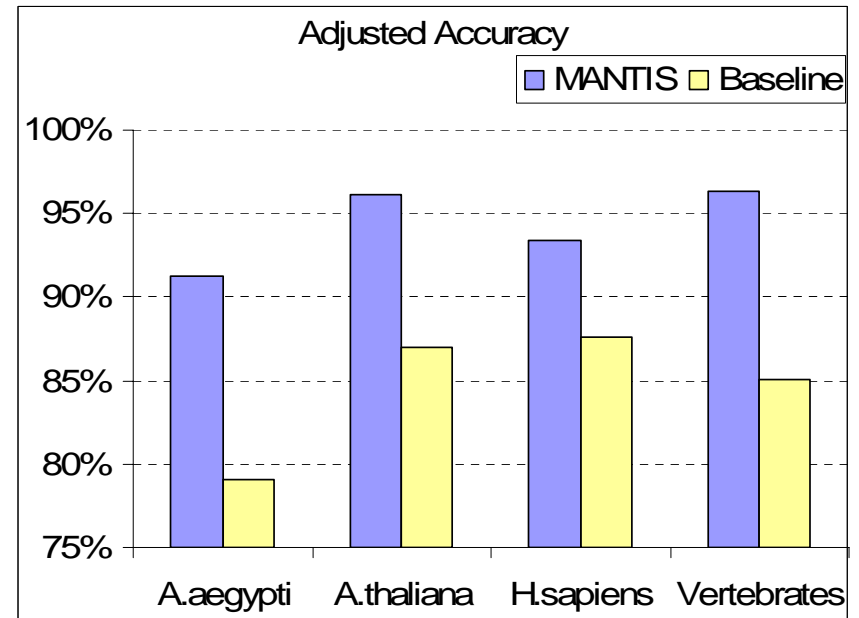
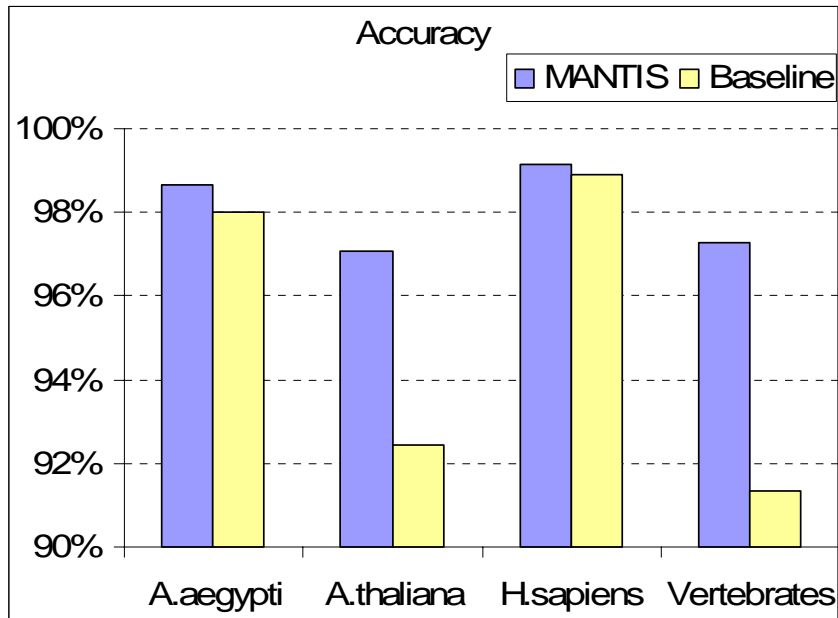
- Accuracy = $\frac{\text{Correctly Predicted}}{\text{Total Instances}}$

- Adjusted Accuracy = $\frac{\text{Sensitivity} + \text{Specificity}}{2}$
 - Sensitivity = $\frac{\text{True Positives}}{\text{Total Positives}}$

 - Specificity = $\frac{\text{True Negatives}}{\text{Total Negatives}}$

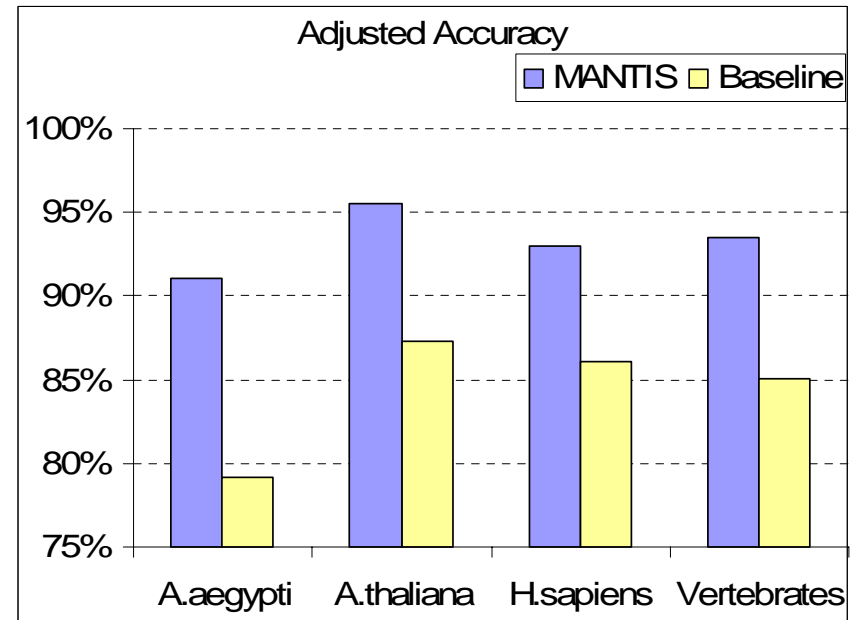
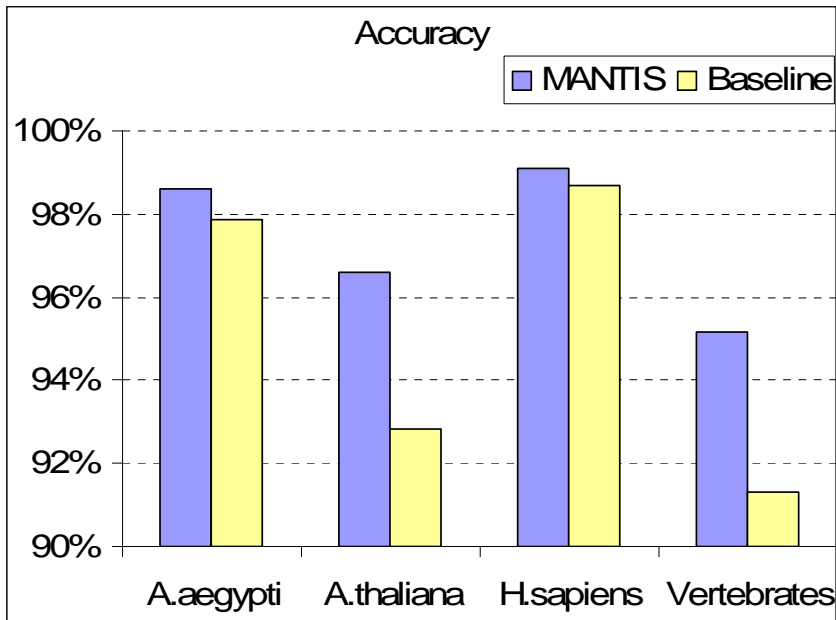
Results (1/2)

- M5 is used as level-1 classifier



Results (2/2)

- MLR is used as level-1 classifier



Conclusions

- We have proposed approaches for predicting Translation Initiation Sites in cDNA or mRNA sequences (i.e. MANTIS)
- Improvements in terms of accuracy and adjusted accuracy have been achieved
 - Adjusted accuracy is more suitable than accuracy for skewed datasets

Future Directions

- Prediction of TIS and coding region in EST (Expressed Sequence Tags) sequences
- Incorporation of frame shift error correction
- Development of a web based version of MANTIS for public use

THANK YOU!

Gene Prediction: Accurate Prediction of Translation Initiation Sites

George Tzanis and Ioannis Vlahavas



Machine Learning and Knowledge Discovery Group

<http://mlkd.csd.auth.gr>