Data Mining for Transcriptomics and Proteomics

Dr. Jaume Bacardit
School of Computer Science and School of Biosciences
Jaume.bacardit@nottingham.ac.uk

About me
• I obtained my PhD in Machine Learning in 2004 (Ramon Llull University, Spain)
• Joined UoN in 2005 to work as a postdoc in a EPSRC-funded project on Protein Structure Prediction
• Appointed as Lecturer in Bioinformatics in 2008
  – Joint appointment between the schools of Computer Science and Biosciences
  – Aim to encourage research at the interface between both disciplines

What can I do?
• My main expertise lays in the Artificial Intelligence area called Machine Learning
• Machine learning: How to construct programs that automatically learn from experience [Mitchell, 1997]
• ML looks at data, in many cases without any domains knowledge at all to
  – Cluster samples
  – Find patterns within the samples
  – Of, if the samples are labelled (e.g. case vs control), construct an explanation of the labelling, which can be used afterwards to predict the outcome of new samples

How to construct explanations
• For instance, by rule learning

What can the ML methods do for biological domains?
• They can provide high accuracy predictions for many domains
• They can identify key variables (e.g. genes, proteins) that are important for the prediction
  – That is, biomarkers or regulators
• They can identify interactions between variables
  – Which can be used to construct networks

Mining transcriptomics data
• Microarray data obtained from seed tissue of Arabidopsis Thaliana
• 122 samples represented by the expression level of almost 14000 genes
• It had been experimentally determined whether each of the seeds had germinated or not
• Can we learn to predict germination/dormancy from the microarray data?
Generating rule sets

- We used our own ML method called BioHEL [Bacardit et al., 09], that was able to predict the outcome of the samples with 93.5% accuracy
- Example of a rule set

<table>
<thead>
<tr>
<th>Rule</th>
<th>Predict</th>
<th>germination</th>
</tr>
</thead>
<tbody>
<tr>
<td>If At1g27595&gt;100.87 and At3g49000&gt;68.13 and At2g40475&gt;55.96</td>
<td>Predict</td>
<td>germination</td>
</tr>
<tr>
<td>If At4g34710&gt;349.67 and At4g37760&gt;150.75 and At1g30135&gt;17.66</td>
<td>Predict</td>
<td>germination</td>
</tr>
<tr>
<td>If At3g03050&gt;37.90 and At2g20630&gt;96.01 and At3g02885&gt;9.66</td>
<td>Predict</td>
<td>germination</td>
</tr>
<tr>
<td>If At5g54910&gt;45.03 and At4g18975&gt;16.74 and At3g28910&gt;52.76 and At1g48320&gt;56.80</td>
<td>Predict</td>
<td>dormancy</td>
</tr>
</tbody>
</table>

Identifying regulators

- Rule building process is stochastic
  - Generates different rule sets each time the system is run
  - But if we run the system many times, we can see some patterns in the rule sets
    - Genes appearing quite more frequent than the rest
      - Some associated to dormancy
      - Some associated to germination
  - We have experimentally verified this analysis
    - By ordering and planting knockouts for the highly ranked genes
    - We have been able to identify four new regulators of germination, with different phenotype from the wild type

Generating networks of interactions

- For each of the rules shown before to be true, all of the conditions in it need to be true at the same time
  - Each rule is expressing an interaction between certain genes
- From a high number of rule sets we can generate regulatory networks
- The network shows different topology when compared to other type of network construction methods (e.g. by gene co-expression)
- Different regions in the network contain the germination and dormancy genes

Other domains

- Initial experiments on proteomics samples, using data generated with Mass Spectrometry
- ML extensively applied also to Protein Structure Prediction
  - Some of our predictors have been highly ranked in the last two editions of CASP
  - Useful to inform site-directed mutagenesis

Conclusions

- Mined fairly large biological datasets
  - To generate robust predictions
  - Identify key regulators/biomarkers
  - Identify interactions between genes/proteins
- Always extracting explanations from our models
- Providing machine learning and optimisation techniques that can be applied to a very broad set of problems

Acknowledgements

- Prof. Natalio Krasnogor
- Prof. Michael Holdsworth
- Dr. George Bassel
- Dr. Enrico Glaab
- Dr. Pawel Widera
- Maria Franco
- Anna Swan

- EPSRC GR/T07534/01 & EP/H016597/1