ECBDL'14: Evolutionary Computation for Big Data and Big Learning Workshop July 13th, 2014 Big Data Competition

Jaume Bacardit – jaume.bacardit@ncl.ac.uk

The Interdisciplinary Computing and Complex BioSystems (ICOS) research group

Newcastle University

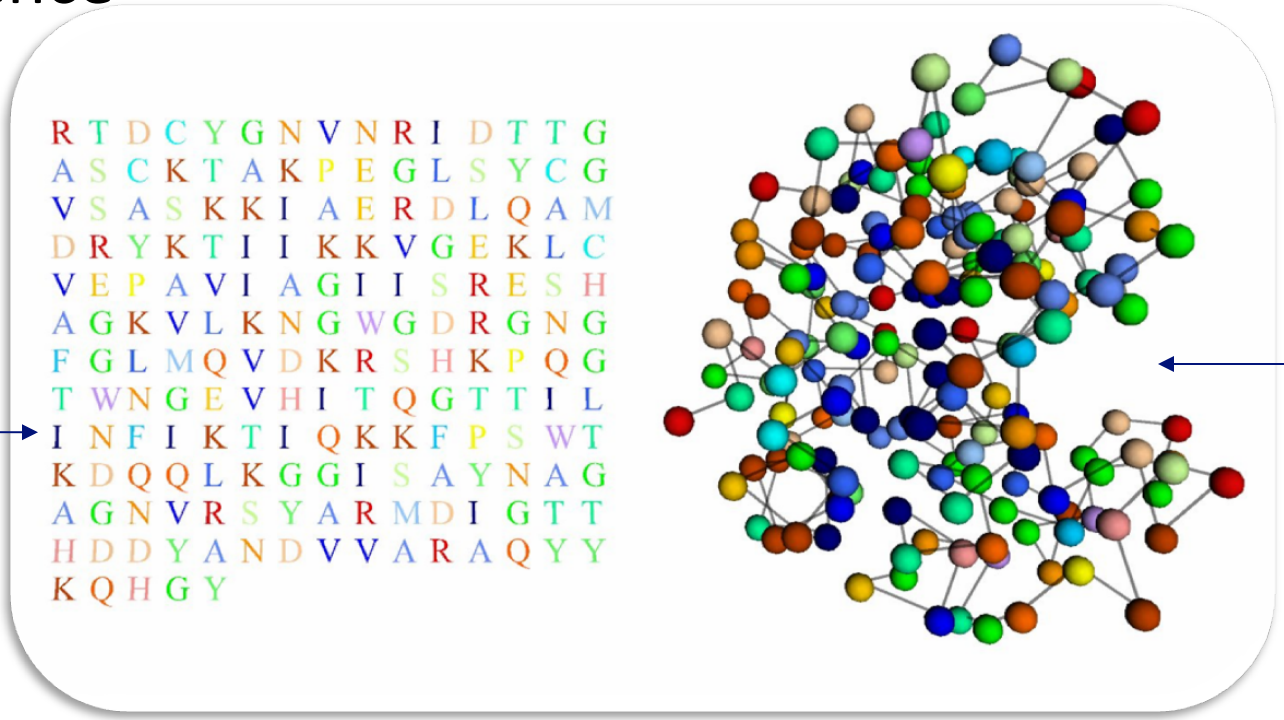
Outline

- Description of the dataset and evaluation
- Final ranking and presentations of participants
- Overall analysis of the competition

DATASET DESCRIPTION AND EVALUATION

Source of dataset: Protein Structure Prediction

- Protein Structure Prediction (PSP) aims to predict the 3D structure of a protein based on its primary sequence

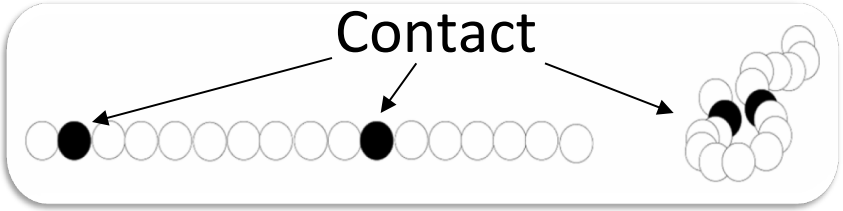


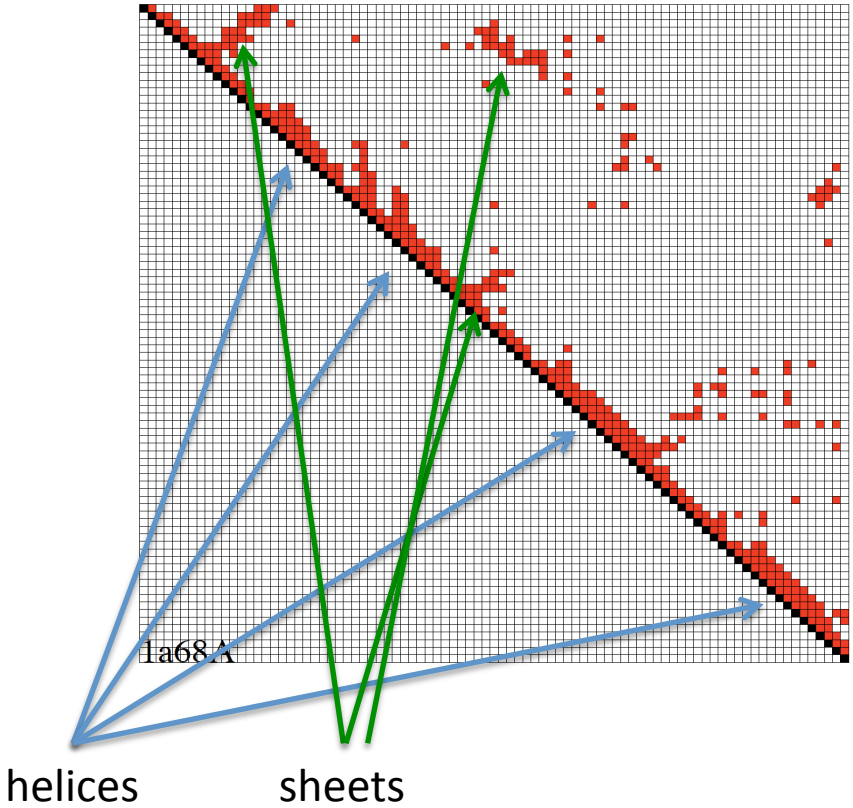
Protein Structure Prediction

- Beside the overall 3D PSP (an optimization problem), several structural aspects can be predicted for each protein residue
  - Coordination number
  - Solvent accessibility
  - Etc.
- These problems can be modelled in may ways:
  - Regression or classification problems
  - Low/high number of classes
  - Balanced/unbalanced classes
  - Adjustable number of attributes
- Ideal benchmarks
  - http://ico2s.org/datasets/psp_benchmark.html

Contact Map

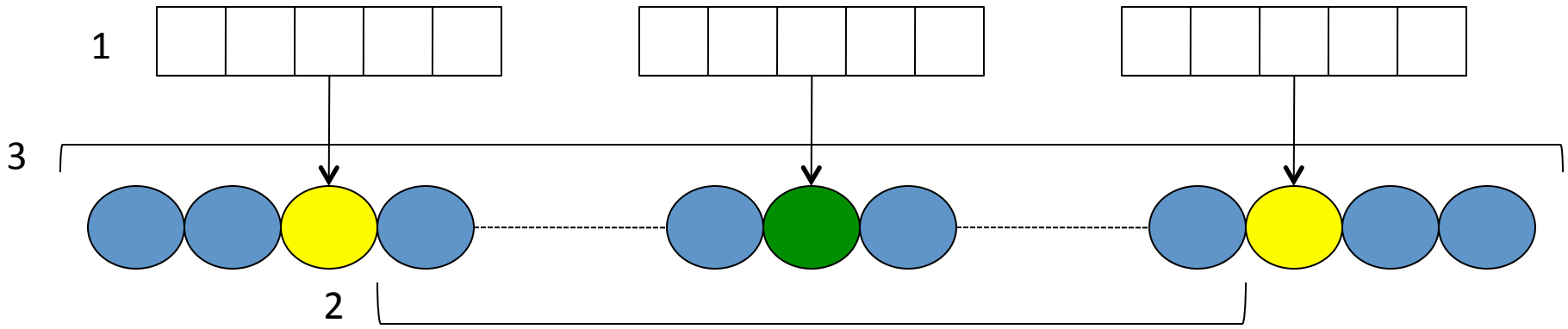
- Two residues of a chain are said to be in contact if their distance is less than a certain threshold
- Contact Map (CM): binary matrix that contains a 1 for a cell if the residues at the row & column are in contact, 0 otherwise
- This matrix is very sparse, in real proteins there are less than 2% of contacts
- Highly unbalanced dataset





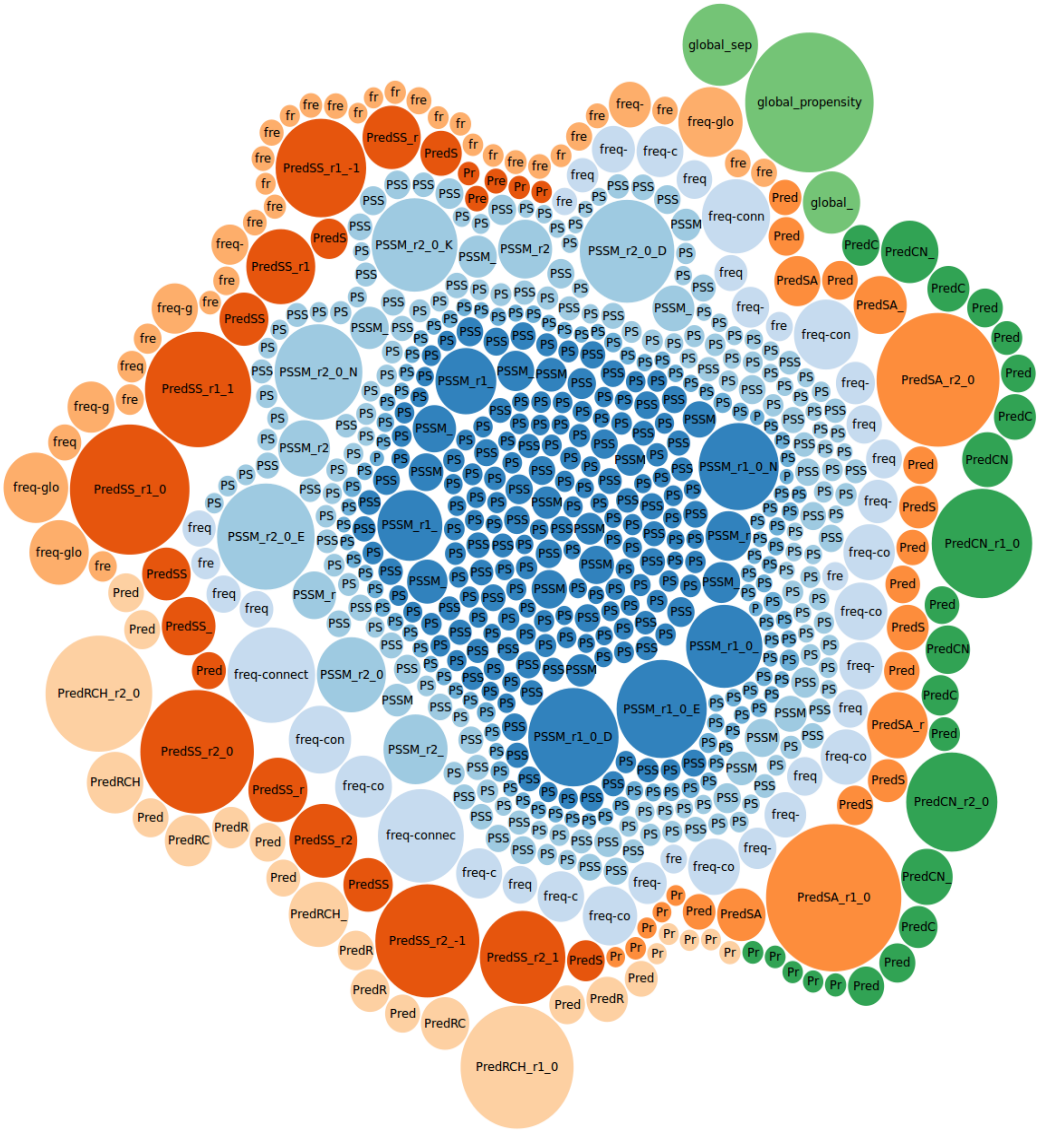
Characterisation of the contact map problem (631 variables)

- Three types of input information were used
  1. Detailed information of three different windows of residues centered around
     - The two target residues (2x)
     - The middle point between them
     - 552 variables
  2. Information about the connecting segment between the two target residues (38 variables)
  3. Information about the whole chain (41 variables)



Attribute relevance (estimated from the ICOS contact map predictor)

- Colour = group of features
- Bubble size = attribute relevance in our rule-based contact map predictor
- All attributes were used in our models (but some rarely)



Contact Map dataset

- A diverse set of 3262 proteins with known structure were selected
  - 90% of this set was used for training
  - 10% for test
- Instances were generated for pairs of AAs at least 6 positions apart in the chain
- The resulting training set contained 32 million pairs of AA and 631 attributes
- Less than 2% of those are actual contacts
- +60GB of disk space
- Test set of 2.89M instances

Evaluation

- Four metrics are computed for each submission
  - True Positive Rate (TP/P)
  - True Negative Rate (TN/N)
  - Accuracy (TP+TN)/(P+N)
  - Final score of TPR x TNR
- The final score was selected to promote predicting the minority class (P) of the problem

Submission of predictions

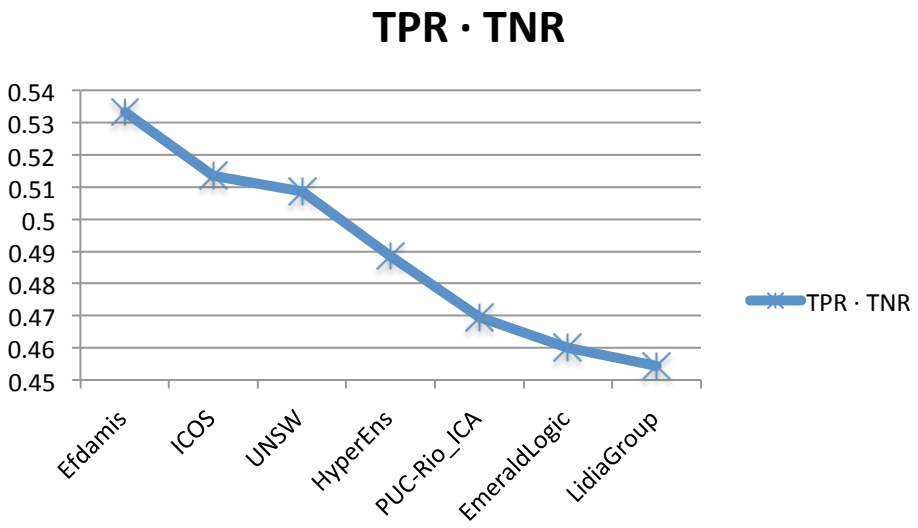
- After registration each team was given a submission code
- To submit predictions teams had to specify
  - The team name
  - The submission code
  - Upload a file with the predictions (one predicted class per row)
  - A brief description of method and resources

OVERALL SCORE AND PARTICIPANT'S PRESENTATIONS

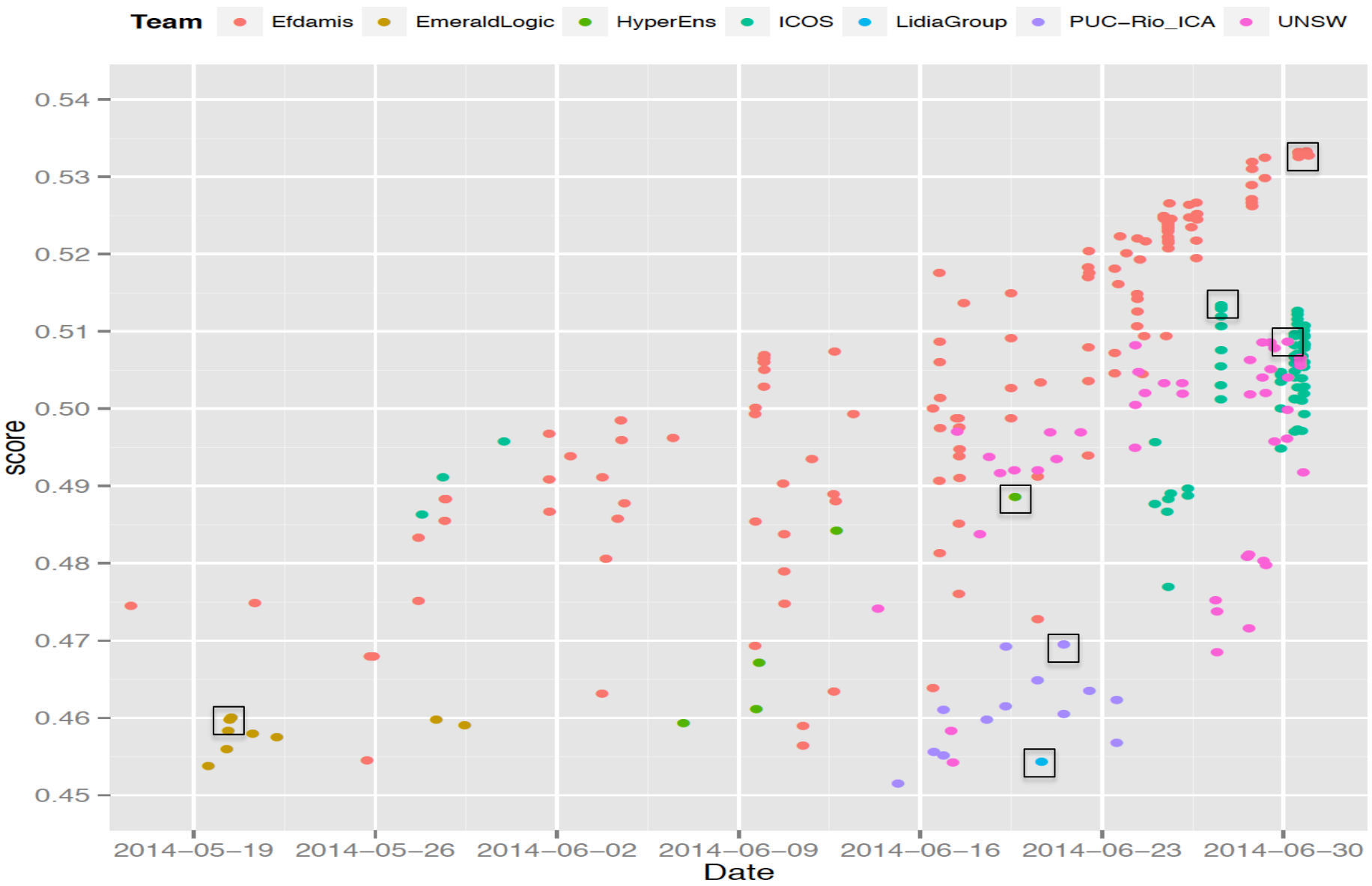
Overall scores

Team Name	Predictions	TPR	TNR	Acc	TPR · TNR
Efdamis	156	0.730432	0.730183	0.730188	0.533349
ICOS	63	0.703210	0.730155	0.729703	0.513452
UNSW	51	0.699159	0.727631	0.727153	0.508730
HyperEns	10	0.640027	0.763378	0.761308	0.488583
PUC-Rio_ICA	40	0.657092	0.714599	0.713634	0.469558
EmeraldLogic	17	0.686926	0.669737	0.670025	0.460059
LidiaGroup	27	0.653042	0.695753	0.695036	0.454356

Total: 364 predictions submitted

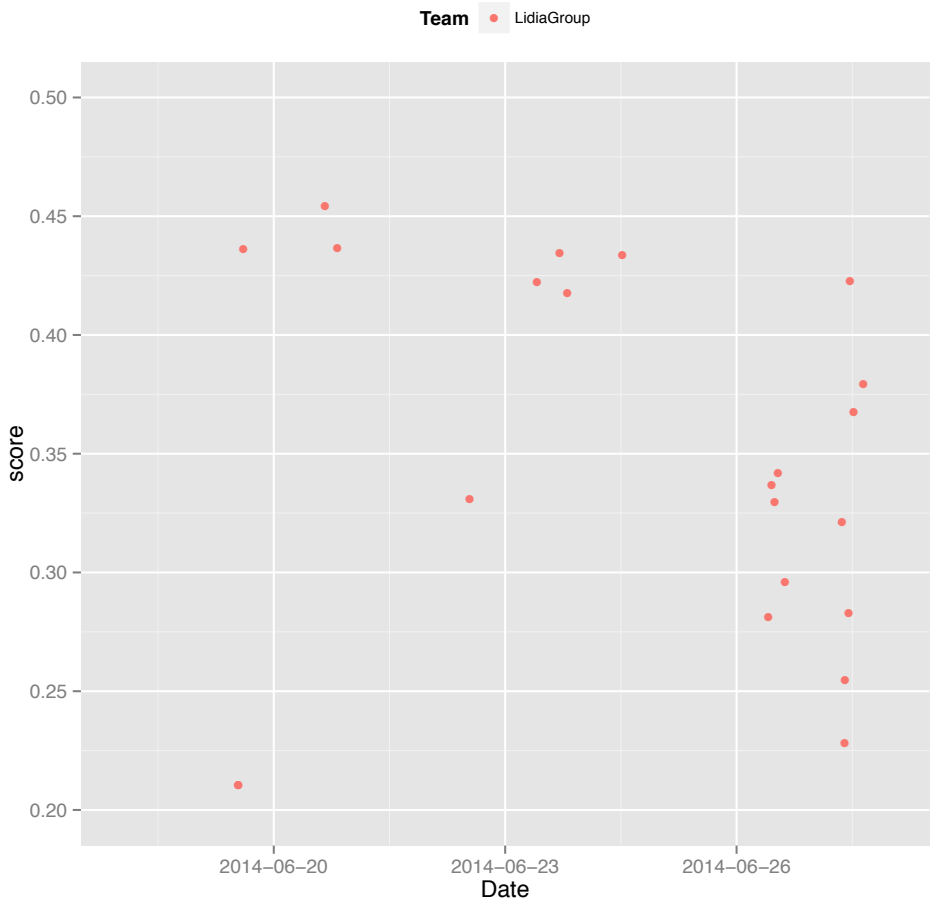


Timeline


Team
Efdamis
EmeraldLogic
HyperEns
ICOS
LidiaGroup
PUC–Rio_ICA
UNSW
score
0.54
0.53
0.52
0.51
0.50
0.49
0.48
0.47
0.46
0.45
2014–05–19
2014–05–26
2014–06–02
2014–06–09
2014–06–16
2014–06–23
2014–06–30
Date

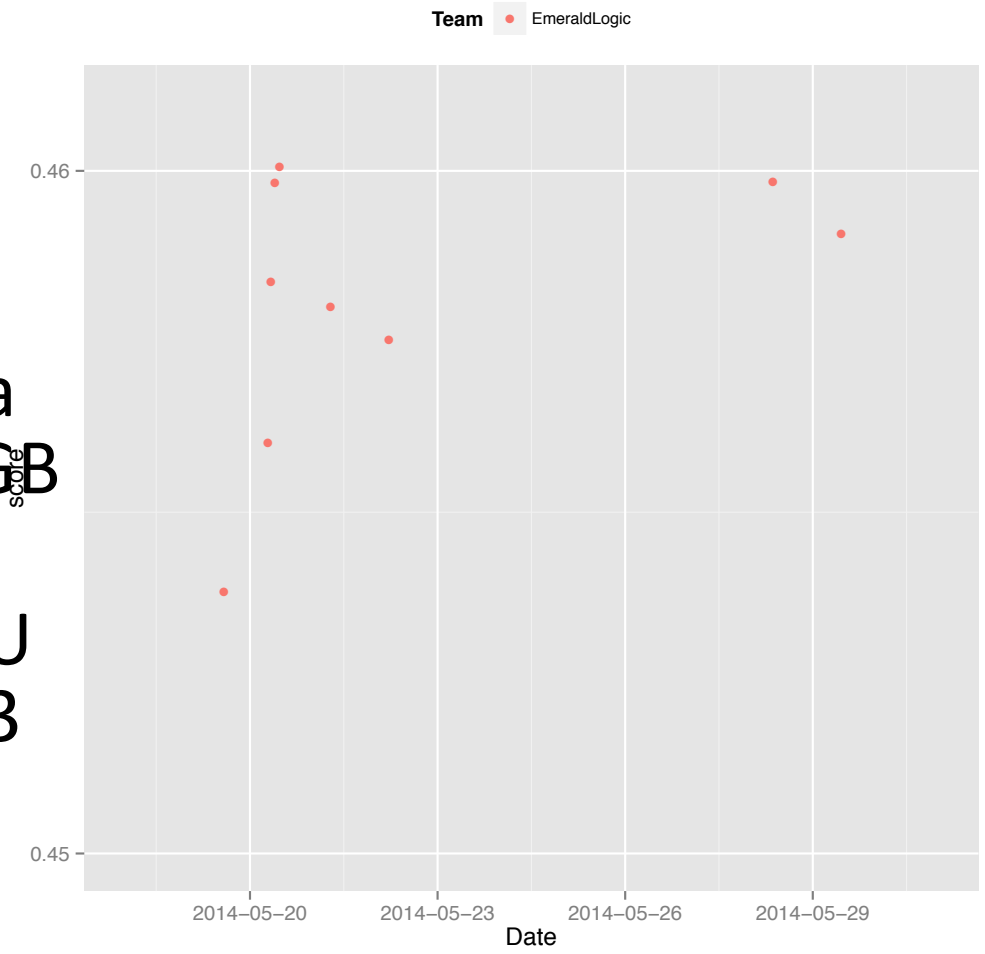
LidiaGroup, Universidade da Coruña

- One-layer neural network using all the data (with oversampling) and only 100 features.
- Resources not specified
- Also tested one-class classification



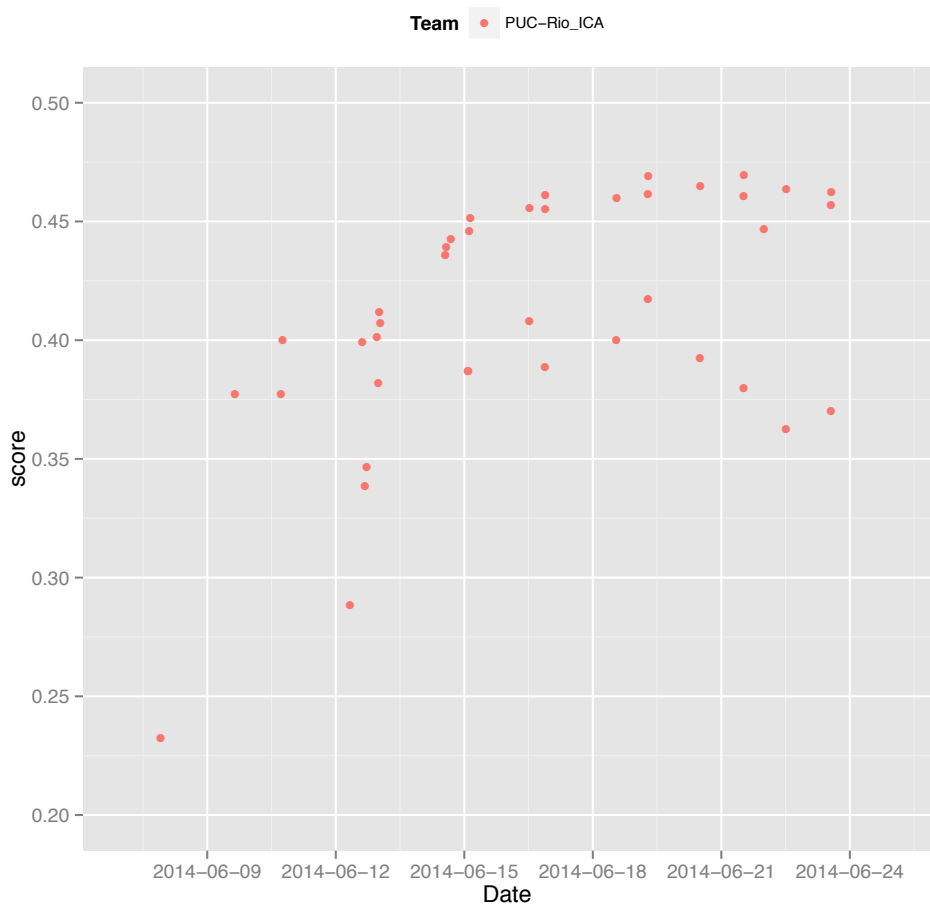
EmeraldLogic

- Linear Genetic Programming-style learning algorithm
- Hardware –Intel i7-3930K CPU, Nvidia Titan Black GPU, 32GB RAM
- 70 minutes CPU+GPU training time plus 0.3 man hours post-analysis



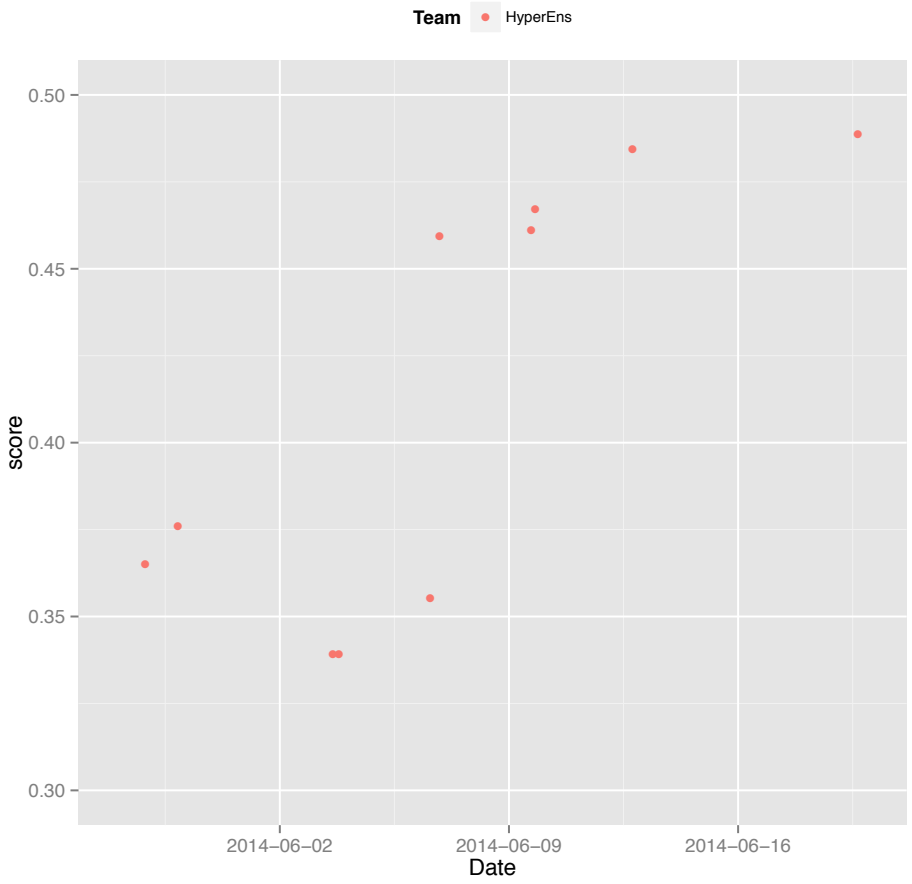
PUC-Rio_ICA

- GeForce GTX Titan
- Linear Genetic Programming
- Resources not specificed



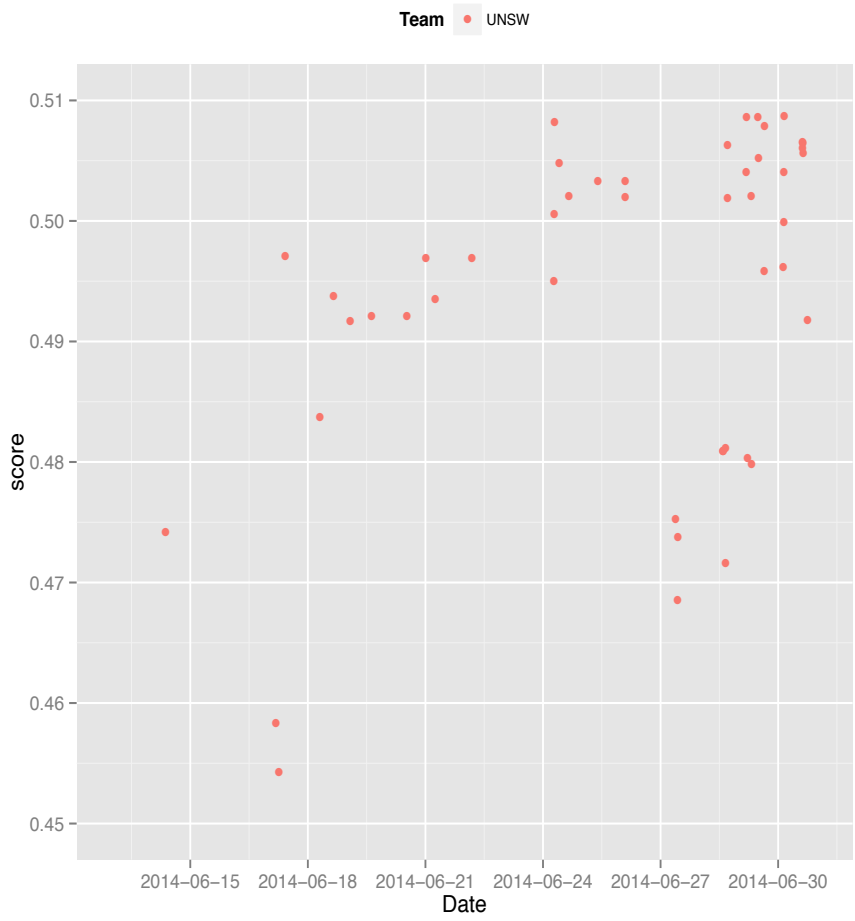
HyperEns

- Standard and Budgeted SVMs with bayesian optimisation of parameters (C, γ, cost-sensitive class errors)
- Best model: 4.7 days of parameter optimisation in a 16-core machine



UNSW

- Hybrid Learning Classifier System/ Deep Learning Architecture
- Each run used 8250 CPU hours in a 24-core machine

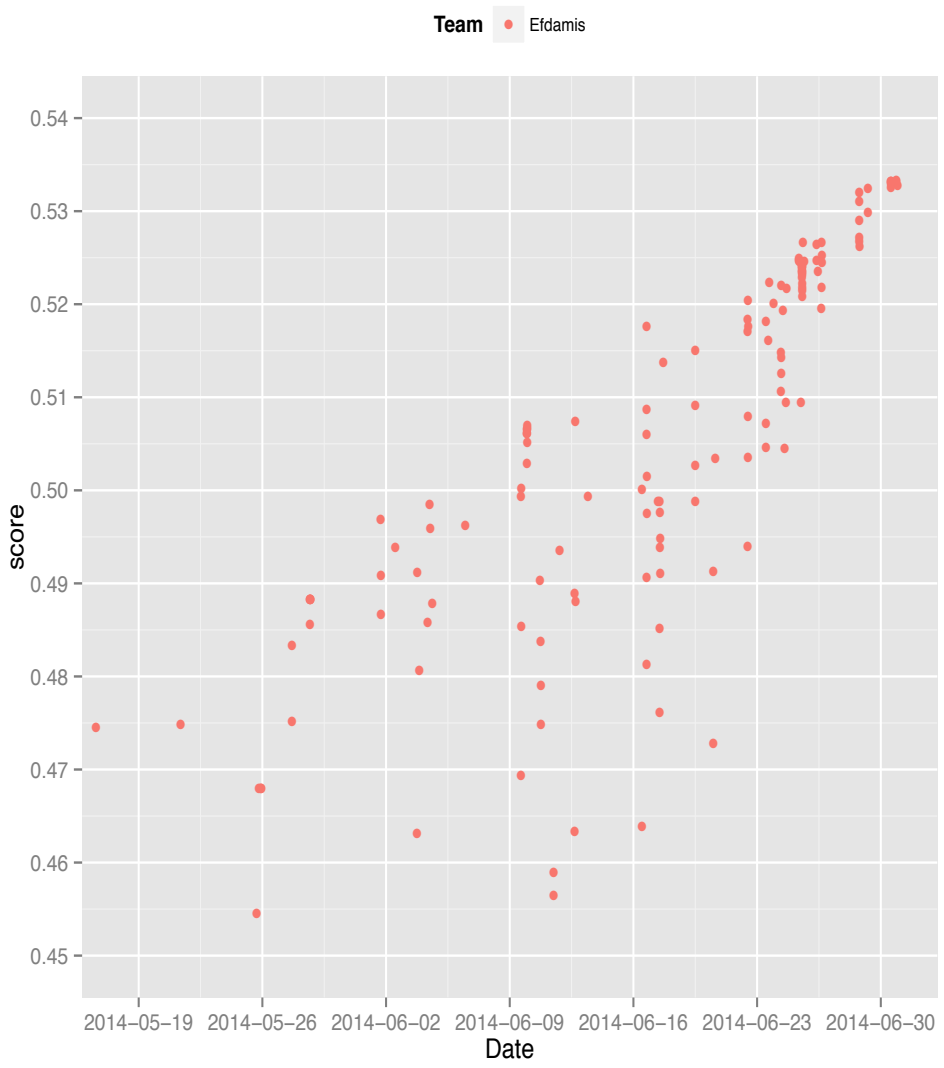


ICOS

- Ensemble of rule sets learnt from samples with 1:1 ratio of positive-negative examples
- Using the BioHEL evolutionary machine learning algorithm to generate rule sets
- Training time for best solution ~3000 CPU hours

Efdamis

- Hadoop-based solution
- Pipeline of random over-sampling, evolutionary feature weighting and Random Forests
- Best model: 39h of Wall-clock time in a 144-core cluster (not used exclusively)



Computational effort vs score

Team Name	Wall-clock hours	CPU hours	Score
Efdamis	39h	---	0.533349
ICOS	Best case: 12h	2998h	0.513452
UNSW	70h	8250h	0.508730
HyperEns	21.1h	337.6h	0.488583
PUC-Rio_ICA	??	??	0.469558
EmeraldLogic	70' (CPU+GPU)	---	0.460059
LidiaGroup	??	??	0.454356

Learning paradigm vs score

Team Name	Learning paradigm	Score
Efdamis	Random Forest	0.533349
ICOS	Ensemble of Rule sets	0.513452
UNSW	LCS of Deep Learning classifiers	0.508730
HyperEns	SVM	0.488583
PUC-Rio_ICA	Linear GP	0.469558
EmeraldLogic	~Linear GP	0.460059
LidiaGroup	1-layer NN	0.454356

Evolutionary vs Non Evolutionary

- Totally evolutionary:
  - ICOS (2nd), UNSW (3rd), PUC-Rio_ICA (5th), EmeraldLogic (6th)
- Partially evolutionary:
  - EFDAMIS (1st)
- Non-evolutionary:
  - HyperEns (4th), LidiaGroup (7th)

What went well

- Diversity of
  - Learning paradigms
  - Computation framework
  - Contribution of evolutionary computation
- Total flexibility of strategy for participants, lightweight submission system
- Teams enjoyed it, learnt a lot from experience

What could be better

- Analysis of computational effort is quite qualitative
- Competition platform
  - More sophisticated engine giving richer information
  - Separate validation (for leaderboard) and test (for final score) sets

Next one?

- Data donors!
  - This challenge was possible because the dataset, until the competition, was private
  - We need data sources that can remain non-public while the competition runs
  - Other types of datasets
- Prize?
- Challenge with uniform computational budget but at the same time flexibility of strategy

ECBDL'14: Evolutionary Computation for Big Data and Big Learning Workshop
July 13th, 2014
Big Data Competition

Thanks to all participants!