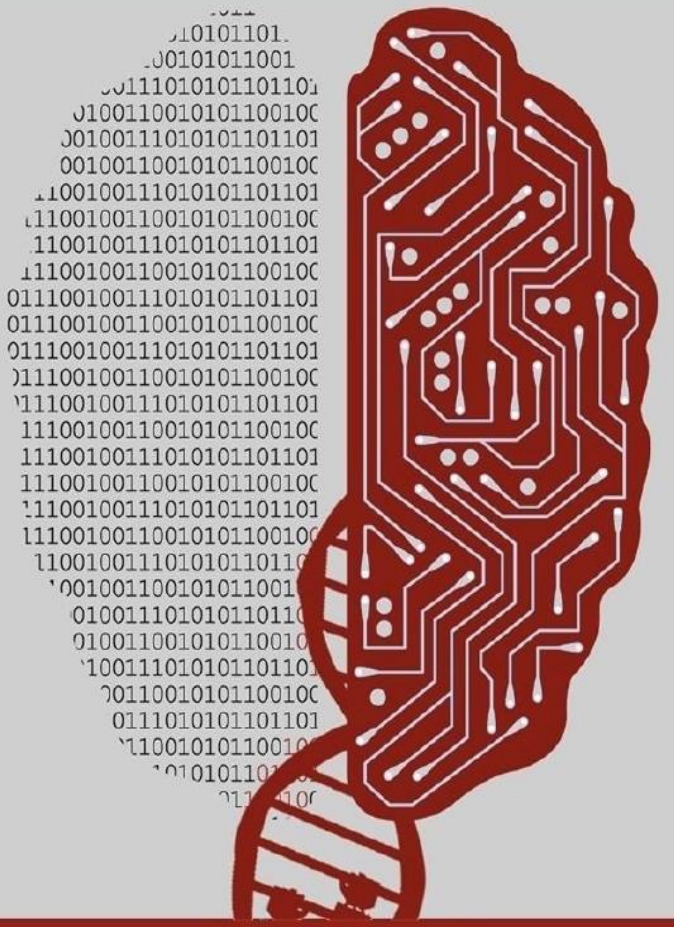
Bioinformatics Open Days

www.bioinformaticsopendays.org



BOOK OF ABSTRACTS

Ciências ULisboa

NEBIUM

BIOINFORMATICS OPEN DAYS

Campus de Gualtar, Universidade do Minho

18th and 19th of February 2016

General chairman

Miguel Rocha, U. Minho

Scientific committee

Isabel Rocha, U. Minho
Rui Mendes, U. Minho
Francisco Couto, U. Lisboa
Francisco Dionísio, U. Lisboa
Octávio Paulo, U. Lisboa

Organizing committee

André Santiago, U.Minho
Bárbara Barbosa, U.Minho
Bruno Silva, U.Minho
Bruno Veloso, U.Minho
Camila Ramos, U.Lisboa
Catarina Lemos, U.Minho
César Catarina, U.Minho
Daniel Braga, U.Minho
Daniel Varzim, U.Minho
Diana Lemos, U.Minho
Joana Ferreira, U.Minho
João Silva, U.Minho
Jorge Ferreira, U.Minho
Jorge Reis, U.Minho
Marco Louro, U.Lisboa
Miguel Santos, U.Lisboa
Nuno Osório, U.Minho
Pedro Martins, U.Lisboa
Raquel Silva, U.Minho
Sara Martins, U.Minho
Susana Gomes, U.Minho
Tiago Alves, U.Minho

Development of a machine learning framework for biomedical text mining

Rodrigues R[1,2], Costa H[2], Rocha M[1]

[1]School of Engineering, University of Minho; [2]Silicolife, Lda

The biomedical literature contains non-structured data, written in natural language, which makes the extraction of high-quality information a challenging task. Biomedical text mining is a scientific field dedicated to create methodologies and tools concerning the challenges of searching and structuring information in biomedical literature.

Named entity recognition and relationship extraction are two of the main biomedical text mining tasks, with the purpose of identifying textual mentions to entities with biological meaning and the identification of possible relations between those entities taking into account the context present in the text stream. Dictionaries, regular expressions, natural language processing approaches and machine learning algorithms are used to address the tasks.

The development of a framework, BioTML, which includes a number of machine learning-based approaches to address named entity recognition and relation extraction tasks, was proposed to fill the gap

between @Note2's operations and state-of-art machine learning approaches. The framework was integrated in @Note2, an open-source computational framework for biomedical text mining based on the model-view-controller paradigm, in the form of a novel plug-in, which allows users to run the methods through a user friendly interface.

ML algorithms like *Hidden Markov Models, Conditional Random Fields* and *Support Vector Machines* were implemented to address named entity recognition and relation extraction tasks, working with a set of more than 60 feature types that can be used to create machine learning models. Both the implementation of machine learning algorithms and natural language processing methods for feature generation were supported in open-source software frameworks, such as *MALLET, LibSVM, ClearNLP* or *OpenNLP.*

Several manually annotated document sets (evaluation corpora) were used to enable the validation of BioTML, in terms of its performance and capability to extract information from unannotated documents, encompassing both entity and relation annotations. The results show promising results, while there is definitely room for much improvement in the future.
