Universidade do Minho
Escola de Engenharia

DSPT DATA SCIENCE PORTUGAL

Machine learning for cancer research: Experiences in the Dream challenges

Miguel Rocha

mrocha@di.uminho.pt

Associate Professor, University of Minho, Dept. Informatics

Senior Researcher, Centre Biological Engineering, U.Minho

CTO at SilicoLife

CENTRE OF BIOLOGICAL ENGINEERING

SILICOLIFE

Our group

BIOINFORMATICS AND SYSTEMS BIOLOGY TEAM: CENTRE OF BIOLOGICAL ENGINEERING

CENTRE OF BIOLOGICAL ENGINEERING

Collaboration between two departments since 2004:

Computer Science

Biological Engineering

Team:

7 PhD - faculty / post docs

Around 15-20 PhD students

~ 10 MSc students + grants

Funding:

Portuguese national agency (FCT)

European Union

Companies

Main areas:

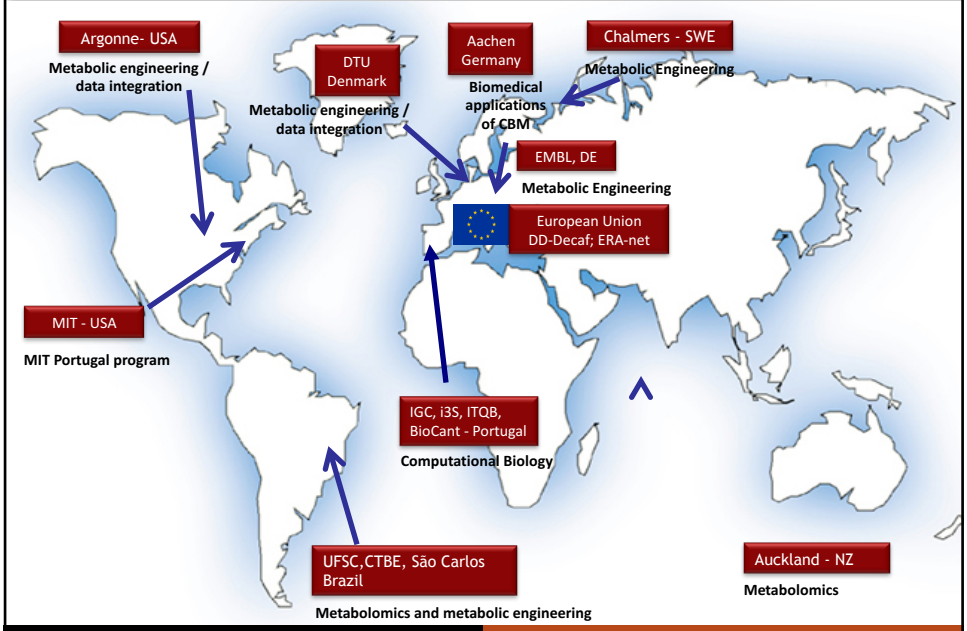
Constraint based modeling/ model reconstruction: metabolic engineering and health

Omics data science/ machine learning

Biomedical Text Mining

http://www.ceb.uminho.pt/biosystems/Labs?lab=1

Our group - collaborations



DREAM challenges

DREAM CHALLENGES

- Challenges that invite participants to propose solutions for biomedical problems proposed by diverse organizations
- Typically, in the form of a **competition**, but fostering collaboration and **community-building**
- Vision: allowing individuals and groups to collaborate openly so that the ***"wisdom of the crowd"*** provides the greatest impact on science and human health
- Expertise and institutional support are provided by Sage Bionetworks

4

DREAM challenges

- Many challenges consist of predicting clinical outcomes from available data
- Training sets are provided, and teams need to develop predictive models, where Machine Learning takes an important role
- But, "external" data may be used – knowledge of domain is important
- Evaluation on unseen data to rank the teams
- Many recent challenges devoted to cancer, focusing on **precision medicine**
- **U. Minho** team has participated in a few competitions since 2014

5

Gene Essentiality Prediction- 2014

The translation of cancer genomic data into cancer therapies remains a challenge

- **Targeted cancer therapy** needs effective treatments and good biomarkers to identify sets of patients likely to respond to those treatments -> need to accurately **predict essential genes** across cancer subtypes
- Essential genes - lead to loss of cell viability when suppressed; genes required for the survival of tumor cells, but not normal ones, provide opportunity as drug targets
- Goal 1: develop **predictive models** that can **infer gene essentiality** in cancer cell lines from their features (gene expression, copy number, and mutation)
- Goal 2: find a small set of **biomarker** features that can best predict gene essentiality

6

Gene Essentiality Prediction- 2014

- Data from large-scale screening of **cancer cell lines** - characterize the **molecular alterations** (mutations, copy number alterations, basal gene expression, etc.) of primary tumors (from Cancer Cell Line Encyclopedia)
- Project Achilles (Broad Institute) provided data on **gene essentiality** for 149 cell lines; over 98000 experiments (RNAi) for individual genes

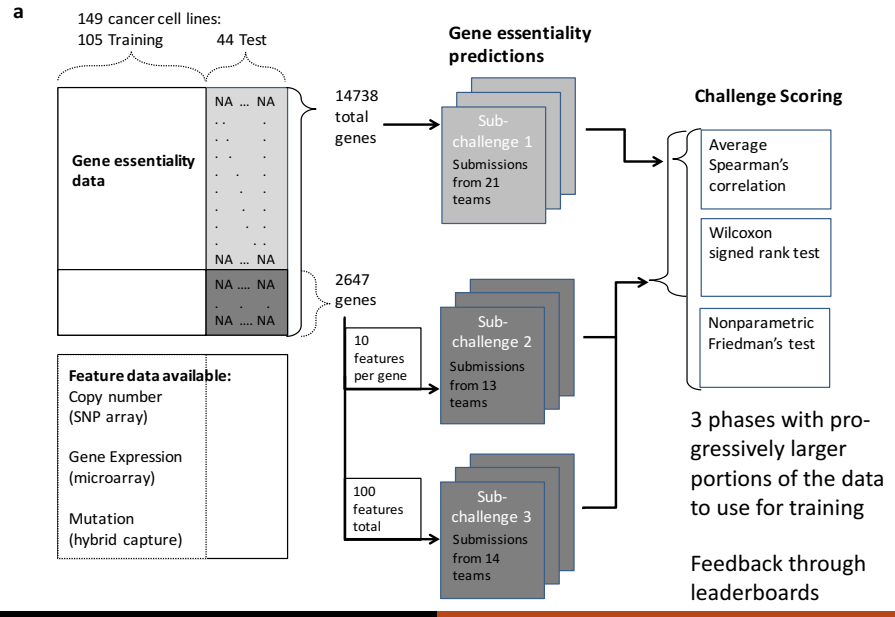
7

Gene Essentiality Prediction- 2014

- Sub-challenge 1: Build a **model** that best **predicts the gene essentiality** values of thousands of genes, using the molecular features of the cancer cell lines.
- Sub-challenge 2: Identify the **most predictive features for each gene essentiality** for a prioritized list of genes. For each gene, select at most 10 features (gene expression, copy number, and mutation) and predict gene essentiality using only these features.
- Sub-challenge 3: Identify the most predictive features for **all** essentiality values of a prioritized list of genes. The aim is to identify a single list of at most 100 features and predict essentiality using only these features for all prioritized genes.

8

Gene Essentiality Prediction- 2014



Gene Essentiality Prediction- 2014

- Our approach
  - Designed a machine learning pipeline based on Python: pandas, scikit-learn
  - Steps: data filtering, feature selection and model fitting
  - Datasets: best results with gene expression data, but we also used CNV (SC1)
  - Models tried: KNN (for regression), SVR, Linear regression, Passive-Aggressive regressor, Ridge, Lasso, ElasticNet, etc.
  - Best results: ensemble: SVR + PAR + Ridge
  - Feature selection: variance threshold, remove highly correlated, K-best (univariate), RFE
  - Hyper-parameter optimization with cross-validation
  - SC3: multitask learning

Gene Essentiality Prediction- 2014

- Main conclusions (from publication):
  - algorithms combining essentiality data across multiple genes demonstrated increased accuracy
  - gene expression was the most informative molecular data type
  - the identity of the gene being predicted was more important than the modeling strategy in influencing model accuracy
  - well-predicted genes and inferred molecular predictors demonstrated clear enrichment in functional categories
  - frequently selected gene expression features correlate with survival in primary tumors
  - wining teams used kernel ridge regression, linear regression and Greedy Regularized Least-Squares for Multi-task Learning for sc1, sc2 and sc3, respectively

11

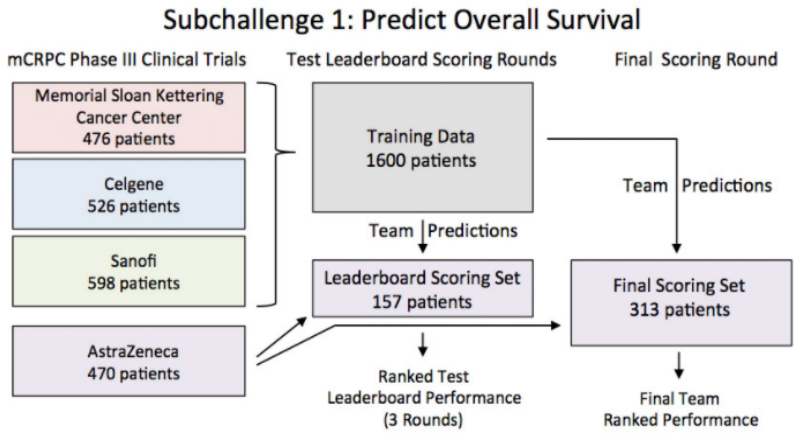
Prostate cancer - 2015

- Aims:
  - Improve the **prediction** of survival and toxicity of docetaxel treatment in patients with metastatic castrate resistant **prostate cancer** (mCRPC)
  - Establish new quantitative benchmarks for **prognostic modeling** in mCRPC, with a potential impact for clinical decision making and ultimately understanding the mechanism of disease progression
- Participating teams asked to submit **predictive models** based on over 150 clinical variables from clinical trials with over 2,000 mCRPC patients treated with first-line docetaxel

12

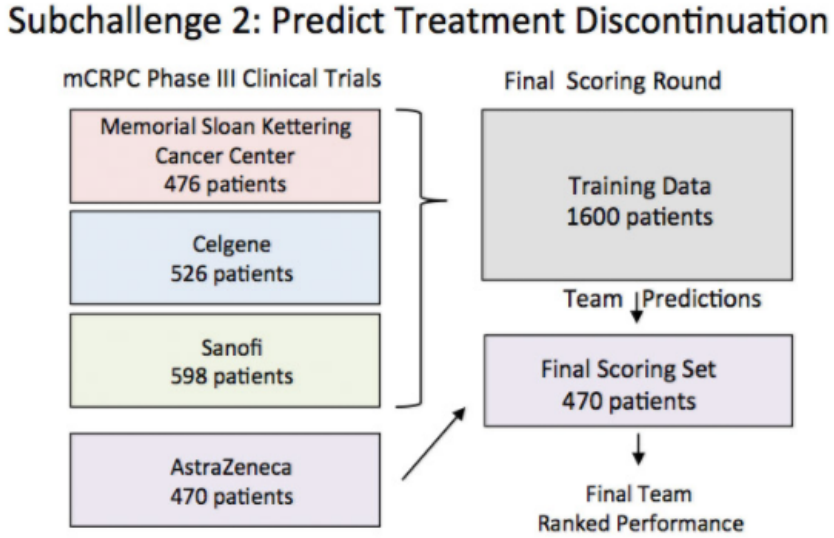
Prostate cancer - 2015

- Sub-challenges:
  - Predict overall survival of mCRPC patients (1a – risk of death; 1b - time to event in days)



13

Prostate cancer - 2015



14

Prostate cancer - 2015

- Data providers: AstraZeneca, Celgene, Sanofi, Memorial Sloan Kettering Cancer Center, Project DataSphere

StudyID	Data Provider	# of patients
ASCENT2	Memorial Sloan Kettering Cancer Center	476
CELGENE	Celgene	526
EFC6546	Sanofi	598
AZ	AstraZeneca	470

Clinical trials

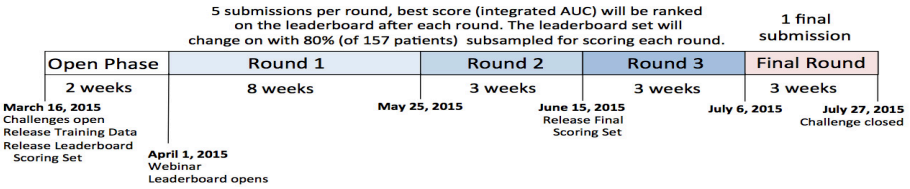
Table Name	Level	Table Description
CoreTable	Patient level	Subject level summary table including dependent variables for the two Subchallenges, and clinical covariates
PriorMed	Patient-event level	Prior Medication table records medication patients took or had taken before 1st treatment date of the trial.
MedHistory	Patient-event level	Medical History table records patient reported diagnoses (co-existing disease) at time of patient screening to participate in the
LesionMeasure	Patient-event level	Lesion table records target and non-target lesion measurement.
LabValue	Patient-event level	Lab test table includes all lab data (hematology and urinary lab)
VitalSign	Patient-event level	Vital Sign table records patient vital sign (height, weight, etc.)

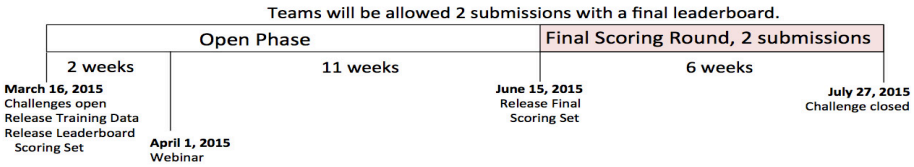
Data tables

15

Prostate cancer - 2015

- Timeline





16

Prostate cancer - 2015

- Our approach
  - Machine learning pipeline in R
  - Hybrid two-phase approach
  - Initial filtering of the dataset: filter redundant/non variable features; Lasso and elastic-net regularized generalized linear models are used to define the importance of each variable and conduct **feature selection**
  - Cox proportional hazards **regression models** to predict the risk scores for each patient using the features selected
  - Predictions obtained using different sets of features and methods to fit models, namely using "random forests", "support vector machines", "decision trees", etc. Submissions based on the best in cross-validation or, in the last round, in the leaderboard examples

17

Prostate cancer - 2015

- Results (from publication)
  - reference model, based on eight clinical variables and a penalised Cox proportional-hazards model
  - top performer was based on an ensemble of penalised Cox regression models (ePCR), which uniquely identified predictive interaction effects with immune biomarkers and markers of hepatic and renal function
  - top performer obtained results significantly better than reference model
  - meta-analysis across all methods confirmed previously identified predictive clinical variables and revealed an important, albeit previously under-reported, prognostic biomarker.

18

Drug combination effects- 2015/16

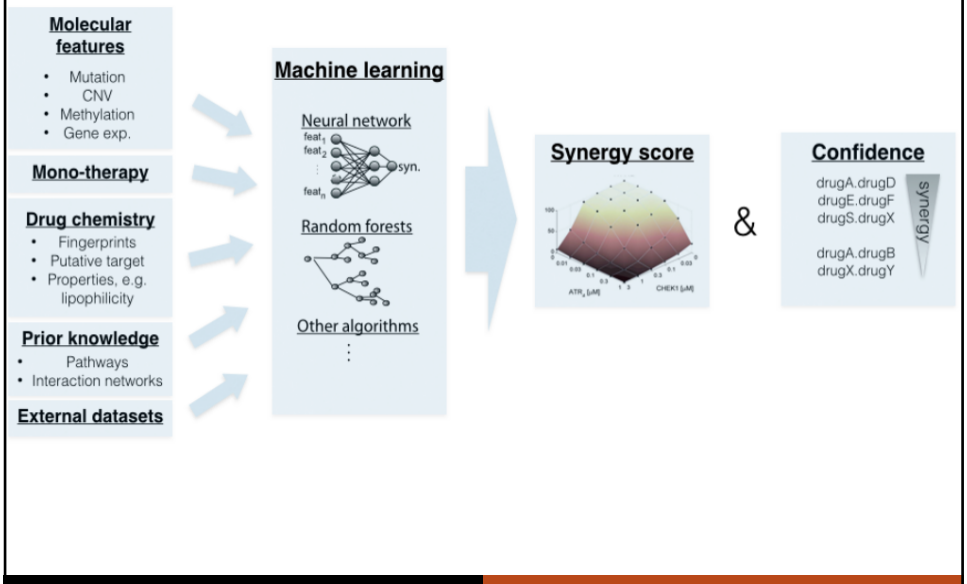
- Rationale: Tumors' innate/acquired **resistance** to therapies may render treatments ineffective. To overcome resistance, **drug combinations** can be a solution.
- Targeting multiple mechanisms simultaneously, the potency of the treatment may be increased and tumor cells are less likely to develop resistance
- Aim: explore traits that underlie effective combination treatments and synergistic drug behavior using genomic data

Drug combination effects- 2015/16

- Sub-challenges:
  1. Predict synergy value of pair of drugs (1A- all datasets; 1B – limited inputs)
  2. Predict synergy without training data
- Data provided: 118 drugs, combined at varying concentrations, tested on 85 cancer cell lines (primarily colon, lung, and breast)
  - Drug data (AstraZeneca): monotherapy data; drug targets, chemical & structural information
  - Molecular data (Sanger Institute): mutations; copy number variation (CNVs); methylation; gene Expression
  - Cell line origin

20

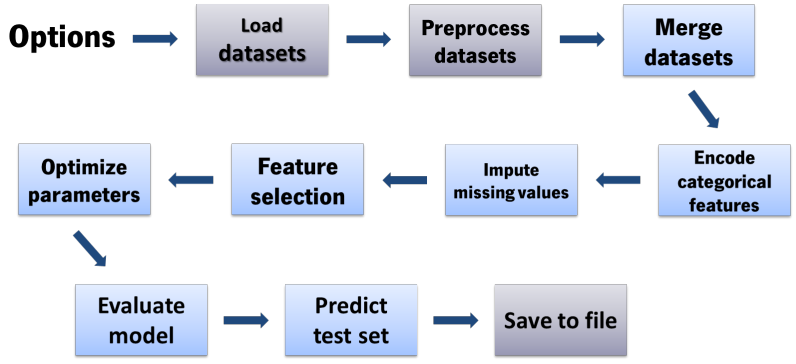
Drug combination effects- 2015/16



21

Drug combination effects- 2015/16

Our approach: Machine learning pipeline



Models for regression:
Linear regression (Ridge), SVR, Random forests,
Gradient boosting, PLS, "Wisdom of crowds", Stacking

load_save_data.py

machine_learning.py

22

Drug combination effects- 2015/16

- Most important datasets: monotherapy data, mutations and CNVs
- Gene expression or methylation data did not improve performance
- Drug data greatly decreased performance
- Best results with ensembles (wisdom of the crowds); best single classifier: gradient boosting

23

Multiple Myeloma - 2017

- Since risk-adapted therapy is becoming standard of care in MM, there is an urgent need for a risk stratification model
- Aim: **Predict high-risk patients** with multiple myeloma using genomics, transcriptomics and clinical information
- Over $150,000 in prize money for the top teams!
- Sub-challenges:
  1. Models based on genomics
  2. Models based on transcriptomics
  3. Combined models

Multiple Myeloma - 2017

Submission model:

A model is trained using supplied data and packaged into a Docker image

Training data

Trained model

Genomics data (mutations)
Gene expression
Clinical data

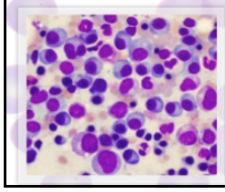
The image is sent to the server and predicts high-risk patients using validation data not accessible to us

Trained model

Validation data

Predictions: Risk scores for test data

Team submission evaluation



Multiple Myeloma DREAM Challenge

Celgene · MMRF · Sage · MOFFITT CANCER CENTER · DREAM · Myeloma Genome Project · DANA-FARBER CANCER INSTITUTE · Heidelberg University Hospital · M2GEN · UAMS

Multiple Myeloma - 2017

- Data reading pipeline – gene expression

RNA-Seq

Patient information

Micro-array

RNA-Seq matrix

Data handler

Microarray matrix

Read patient information

Get RNA-Seq filenames

Get microarray filenames

Read files and merge by type of data

Deal with missing values and **return the combined data**

Implemented using Python (with *pandas* for data reading/processing)
Similar scheme to extract genomic features from VCF files (mutations)

26

Multiple Myeloma - 2017

- Model training pipeline – gene expression

Data preprocessing	• Normalization and scaling along rows/columns • Fill missing values (flat value or median)
Feature engineering	• Create discrete features • Thresholds for gene expression
Dimensionality reduction	• Univariate feature selection • Add/remove based on literature
Estimator selection	• Automatic selection with parameter optimization • Select best estimators based on F1-score
Cross-validation	• 10-fold cross-validation with stratification • Each fold has the same sample proportions
Serialization	• Package data transformers and estimator(s) in a file for future use with validation data

Implemented using Python (with *scikit-learn* as the machine-learning framework)

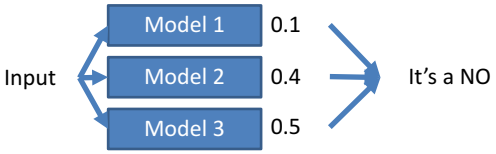
27

Multiple Myeloma - 2017

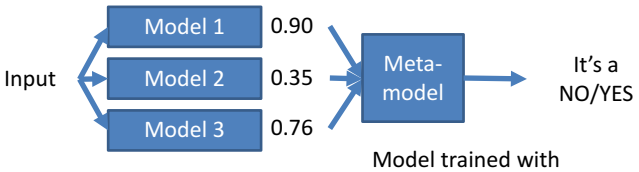
Standard Classifier

Naïve Bayes, MLPs, Logistic regression, SVMs

Voting Classifier



Model Stacking



Model trained with scores from other models

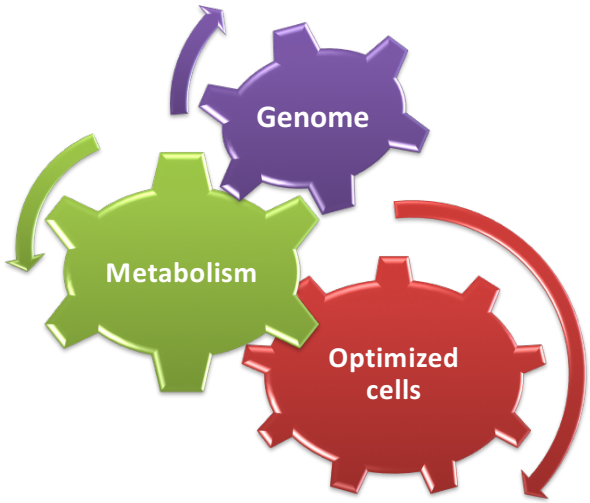
Metabolic engineering

In biotechnology, to produce **desired compounds** (e.g. antibiotics, fuels, vitamins) from **microbial cell factories** it is generally necessary to **retrofit the metabolism**

Metabolic Engineering envisages the introduction of **directed genetic modifications** leading to desirable phenotypes, as opposed to traditional methods

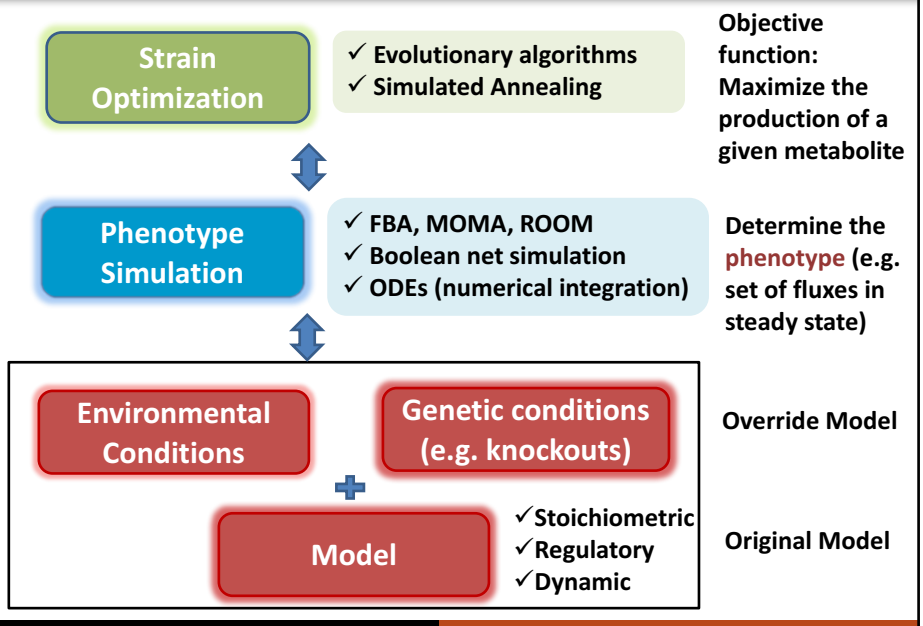
29

Cell optimization: computational architecture



30

In silico strain optimization



@Note

@note Biomedical Text Mining

www.anote-project.org

@Note2 is a multi-platform user-friendly Biomedical Text Mining platform

Allows to perform text mining pipelines using state of the art **Information Retrieval** and **Information Extraction** methods

@Note overview

@note Biomedical Text Mining

Modules

Publication Manager

Information Retrieval
Query Management
Document Labeling
PDF Retrieval

Resources

Lexical Resources Management
- Dictionaries
- Lookup Tables
- Rules
- Ontologies
- Lexical Words

Corpora

Corpora Management
Named Entity Recognition
Relation Extraction
Manual Curation

Pubmed Search
Patent Search
Springer Search

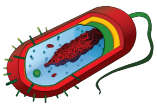
Standard csv files import
Loaders to databases flat files
OBO Ontology Loader

Import Standard Corpora
Linnaeus Tagger
Rel@tioN
BioTML – machine learning

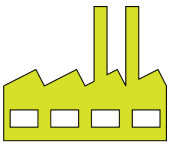
Methods & Operations

Silico Life

Artificial Inteligence + Biological knowledge



= in silico design of **optimal cell factories**



- **Founded in 2010** and privately owned by its founders
- Enabling the design of **optimized microbial strains** for the production of biofuels, chemicals or biopolymers
- Bridging computer sciences, life sciences and bioengineering

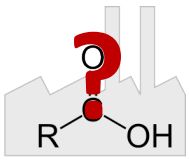
www.**silicolife**.com

SILICOLIFE

A complete pipeline from the **problem** to the **solution**

SILICOLIFE

Problem from client



Strain optimization for desired end product

Using computational biology to **tailor metabolic pathways**

Development Partners

Network of partners:
- ✓ Wet lab
- ✓ Genome sequencing
- ✓ State-of-the-art from Academia

Acknowledgments

Thanks to the BiSBII @ UMinho and SLife teams!!

Funding/ support:
FCT; European Commission; FEDER/ Norte2020

36