Machine Learning algorithms for making inferences on networks and answering questions in Biology and Medicine

Alberto Paccanaro
Department of Computer Science
Centre for Systems and Synthetic Biology
Royal Holloway, University of London
Why ML and biology

We need to analyse the cell at systems level
Biological networks

Cell as webs of interactions between biomolecules
Experimental data have a natural representation as networks

Many problems in biology and medicine can be formulated as problems of inference on biological networks

[Horak, Genes & Dev.; DeRisi, Science; Qian, J. Mol. Bio; Jeong, Nature; Tong, Science; Goh, PNAS]
In my lab, we develop Machine Learning methods for answering questions in biology and medicine *focus on biological networks*

- At the heart of our research is the biological question, not the methodology – *different areas of ML*
- **Diverse problems**
- Collaborate with *experimentalists*
- We implement *software tools* that allow biologists and clinicians to easily use the methods that we develop
# Diverse problems, diverse approaches

<table>
<thead>
<tr>
<th>Problem</th>
<th>Type of ML Approach/Technique</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantifying the functional similarity between genes (ontology-based)</td>
<td>Random Walks over ontology structures (DAGs)</td>
<td>Caniza et al, <em>Bioinf.</em>, 2014; Yang, Nepusz, Paccanaro <em>Bioinformatics</em>, 2012</td>
</tr>
<tr>
<td>Selecting transcriptomics experiments for a given functional category</td>
<td>Supervised learning</td>
<td>Bhat, Yang, Paccanaro <em>PLoS ONE</em>, 2017</td>
</tr>
<tr>
<td>Selection of representative gene in a co-expression network</td>
<td>Function maximization (greedy, but global)</td>
<td>Yang, Paccanaro in preparation</td>
</tr>
</tbody>
</table>
### Diverse problems, diverse approaches

<table>
<thead>
<tr>
<th>Problem</th>
<th>Type of ML Approach/Technique</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction of patients phenotype/outcome</td>
<td>Semi-supervised learning</td>
<td>Gliozzo et al, <em>PLoS Comp. Biol.</em> (under review)</td>
</tr>
<tr>
<td>Prediction of drug cocktails against Chagas disease</td>
<td>Supervised learning</td>
<td>Jimenez et al, <em>in preparation</em></td>
</tr>
<tr>
<td>Predicting the frequency of drug side effects</td>
<td>Collaborative filtering (matrix factorization)</td>
<td>Galeano, Paccanaro bioRxiv 594465; doi: 10.1101/594465</td>
</tr>
</tbody>
</table>
1. Quantifying the distance between disease modules on the interactome

[Caniza, Romero, Paccanaro, Nature Scientific Reports, 2015]
**Network Medicine: Disease as perturbations of molecular networks**

Genes associated with a specific disease tend to cluster in the same neighbourhood – the disease module.

The disease modules of diseases that are phenotypically similar tend to be located in closeby regions of the interactome.
Question

Define a “distance” between diseases using the disease phenotypes such that it is related to the distance between disease modules.
The problem

calculate a distance here which is...

Phenotype

Genotype

...related to a distance here
Outline of the method

[Caniza et al, Scientific Reports, 2015]

STEP 1: Translate a genetic disease into a set of MeSH terms
STEP 2: quantify a distance between two sets of terms on an ontology

Luckily, we had developed a measure for that!

Host Similarity Measure (HSM)

- HSM between every pair of leaves weighted by their probabilities
- Existence of common descendants affect the random walk
- Uncertainty

Random Walk Contribution (RWC)

- HSM between every pair of leaves weighted by their probabilities
- Existence of common descendants affect the random walk

Yang et al, Bioinformatics, 2012
Caniza et al, Bioinformatics 2014
http://www.paccanarolab.org/gosstoweb/
STEP 2: quantify a distance between two sets of terms on an ontology

Does our distance reflects the distance between disease modules?

Luckily 😊, we had developed a measure for that!
(Yang et al, Bioinformatics, 2012; Caniza et al, Bioinformatics, 2014)
1. Evaluation as a prediction problem

A. Diseases related by physical interactions (PPI) of diseases proteins

\[(D_i, D_j) \rightarrow 1\]
iff \(\exists \alpha \in D_i\) and \(\beta \in D_j\)
s.t. \(\alpha\) interacts with \(\beta\)

Our similarity measure

<table>
<thead>
<tr>
<th>(D_1)</th>
<th>(D_2)</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D_1)</td>
<td>(D_3)</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>(D_i)</td>
<td>(D_j)</td>
<td>1</td>
</tr>
</tbody>
</table>

How well does column A predict column B?

B. Diseases related by sequence similarity of disease proteins

C. Diseases related by evolutionary relatedness of disease proteins (Pfam)

D. Coverage (% of OMIM diseases)
Results of AUC analysis

**Robinson**: builds an ad-hoc diseases ontology (Human Phenotype Ontology) and then calculates a distance on it (Köhler et al, NAR, 2013)

**Park**: similarity between two diseases is determined by an association score based on the cellular co-localisation of their disease proteins (Park et al, Mol. Sys. Bio. 2011)
2. Embedding diseases in low dimensional space

1) Calculate the distance between every pair of OMIM diseases

2) Embed diseases as points in a low dimensional space based on our distance.

Goh et al, PNAS (2007)
Embedding diseases in 3D using t-SNE

MIM:180550 - Ring Dermoid of Cornea – cancer/dermatological/ophthalmological
MIM:609528 - Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome – neurol./dermatol.
MIM:308240 - Lymphoproliferative syndrome – cancer/immunological
Landis – the Landscape of Disease Similarities
http://www.paccanarolab.org/landis

Differential diagnoses

It provides explanations
2. Using disease distances to predict disease genes for Uncharted Diseases

[Caceres, Paccanaro, PLoS Comp. Biology, to appear]
Disease gene prediction

- **Charted** diseases: some disease genes are known
- **Uncharted** diseases: no known disease genes

Disease gene prediction for charted diseases: search in a neighbourhood of known disease genes

Can we use our disease similarity measure for predicting disease genes for uncharted diseases?

Data from Online Mendelian Inheritance in Man (OMIM), Sept 2018
Predicting genes for *uncharted* diseases – the idea

**Triangulation**: a mobile phone is detected within a radius from each of the towers.
A new disease gene prediction algorithm

**Soft labels + diffusion**

1. Calculate the similarity between our uncharted disease and each charted disease
2. Place known genes in the interactome.
3. Learn a similarity-to-label mapping
4. Assign a “soft” label to the disease genes
5. Diffuse the soft labels
Diffusing soft labels (semi-supervised learning)

For a given disease, the soft label is related to the probability for that gene to be a disease gene for that disease.

\[
F^* = \arg \min_F Q(F)
\]

\[
Q(F) = \frac{1}{2} \left( \sum_{i,j=1}^{n} W_{ij} \left\| \frac{1}{\sqrt{D_i}} F_i - \frac{1}{\sqrt{D_j}} F_j \right\|^2 + \mu \sum_{i=1}^{n} \| F_i - Y_i \|^2 \right)
\]

Interacting nodes have similar labels
Preserve initial labelling

\[ F^* = (1 - \alpha)\left( I - \alpha D^{-1/2} WD^{-1/2} \right)^{-1} Y \]

(Zhou et al, NIPS 2004, "Consistency" method)
Prospective evaluations
Using information from 2013, predict new disease genes known in 2018

Leave-one-out
Using data from 2018, a single association is removed and is predicted back
Performance – uncharted diseases

Prospective evaluations

Leave-one-out
Performance – charted diseases

Prospective evaluations

Leave-one-out

DIAMOnD -- Ghiassian, Menche, Barabasi, PLoS Comp Bio 2015
Prodige1,4 -- Mordelet, Vert, BMC Bioinformatics, 2011
Prince -- Vanunu, Magger, Ruppin, Shlomi, Sharan, PLoS Comp Bio 2010
### Prospective evaluation -- Examples

<table>
<thead>
<tr>
<th>Disease</th>
<th>2013 Status</th>
<th>Gene</th>
<th>Our Ranking</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schimmelpenning-Feuerstein-Mims syndrome (MIM:163200)</td>
<td>Charted</td>
<td>NRAS</td>
<td>1</td>
<td>Lim et al., Human molecular genetics 23, 2014</td>
</tr>
</tbody>
</table>
Conclusions of Part 1 & 2

✓ A distance between disease modules on the interactome which uses exclusively disease phenotype information.

✓ How diffusion methods + our disease similarity measure can be used to infer disease genes for uncharted diseases.

✓ These methods can provide explanations
3. A collaborative model for predicting the frequency of drug side effects

Drugs side effects

A drug-side effect association in humans can be:

- **Very rare**: < 0.01%
- **Rare**: < 0.1%
- **Infrequent**: < 1%
- **Frequent**: < 10%
- **Very frequent**: > 10%

**Clinical Trials**
Phase I-III
(Premarketing)

**Post-marketing Surveillance Systems**
(FAERS-FDA)

*FDA-approved (In-market)*

**Placebo-controlled study**
One disease
Limited size

**Observational study**
Multiple diseases
Multiple medications
Question

Can we predict the frequency of drug side effects?

Few methods exist which are aimed at predicting the presence/absence of side effects. These exploit molecular or cellular features.
The data

The Side Effect Resource (SIDER) 4.1 [Khun et al., 2015]

- 996 side effect terms
- 760 drugs
- 760 drugs
- 996 side effect terms

<table>
<thead>
<tr>
<th>Density</th>
<th>Very rare = 1</th>
<th>Rare = 2</th>
<th>Infrequent = 3</th>
<th>Frequent = 4</th>
<th>Very Frequent = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

density ~ 5% (sparse)
Let’s look at the data...
How do we predict (recommend) movies?

Matrix decomposition models are useful for very sparse datasets with potential latent features.

\[ Y \]

\[ Y_{i,j} \approx p_i^T \cdot q_j \]

\[ Y_{n \times m} \approx P_{n \times k} \cdot Q_{k \times m} \]
Our idea: recommending side effects to drugs

Very rare = 1
Rare = 2
Infrequent = 3
Frequent = 4
Very Frequent = 5

\[ Y_{n \times m} \]

\[ Y_{i,j} \approx p_i^T \cdot q_j \]
\[ Y_{n \times m} \approx P_{n \times k} \cdot Q_{k \times m} \]
Learning the latent representations

\[
\min_{P,Q} J(P, Q) = \frac{1}{2} \| Y - PQ \|_F^2 + \frac{\lambda}{2} (\| P \|_F^2 + \| Q \|_F^2)
\]

subject in order to increase interpretability

We learn this with a multiplicative rule (similar to NMF) or with Conjugate Gradient Descent + projections

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 2 & 0 & 3 & 0 & 4 & 0 \\
4 & 0 & 0 & 3 & 0 & 4 & 0 & 0 & 0 & 1 \\
0 & 0 & 4 & 0 & 0 & 0 & 1 & 0 & 3 & 0 \\
5 & 0 & 0 & 0 & 5 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 4 & 0 & 3 & 0 & 2 & 0 & 0 & 0
\end{bmatrix}
\]

…it does not work 😞
Our new cost function

\[
\min_{W,H \geq 0} J(W,H) = \frac{1}{2} \sum_{Y_{i,j} \in \{1,2,3,4,5\}} (Y_{i,j} - (WH)_{i,j})^2 + \frac{\alpha}{2} \sum_{Y_{i,j} = 0} ((WH)_{i,j})^2
\]

Fits clinical trials frequency data

Fits unobserved associations with confidence \(\alpha_{null}\)

\(Y_{n \times m}\) of \(n\) drugs and \(m\) side effects
\(W_{n \times k}\): drug signatures
\(H_{k \times m}\): side effect signatures
\(0 \leq \alpha \leq 1\)

We are confident on clinical trials data (values 1-5) but only \(\alpha\)-confident on the unobserved associations (0s)

Our model uses the large amount of zeros as a regularization
- Small \(\alpha\) allows the weights in \(W\) and \(H\) to grow
- Large \(\alpha\) keeps the weights in \(W\) and \(H\) small and induces sparsity.
Our cost function *converges to a local optimum* using the update rules (satisfy the Karush-Kuhn-Tucker conditions):

\[
W \leftarrow W \circ \frac{P_{\Omega}(Y)H^T}{(P_{\Omega}(WH) + \alpha P_{\Omega}^-(WH))H^T}
\]

\[
H \leftarrow H \circ \frac{W^TP_{\Omega}(Y)}{W^T(P_{\Omega}(WH) + \alpha P_{\Omega}^-(WH))}
\]

- \(P_{\Omega}\): selection function for entries \{1,2,3,4,5\}
- \(P_{\Omega}^-\): selection function for entries \{0\}
- \(\circ\) is the Hadamard product

Inspired by non-negative matrix factorization (NMF) [Lee, Seung, Nature, 1999]

**Multiplicative learning rule – no learning rate, no projection function**
Prediction on Test Set

Higher predicted values correspond to higher side effect frequencies

No significant differences between the predicted scores for the very rare side effects and the post-marketing side effects
Examples

Gabapentin  
(anticonvulsant drug)

Arrhythmia  
(cardiovascular side effect)
Percentage of accuracy at predicting the frequency class of drug side effects

Question: can we “explain” how the prediction works?

Very rare: <0.01%
Rare: <0.1%
Infrequent: <1%
Frequent: <10%
Very frequent: >10%

Sensitivity of 0.97
Specificity of 0.57
Predictions can be explained in terms of the latent features

**Example:** Atorvastatin is known to cause frequent respiratory and thoracic-related side effects

**Question:** do the latent representations tell us something about the biology of the problem?
Drug signature are related to clinical activity of the drug

Hierarchical categorization of drugs according to ATC (from WHO):

1. Anatomical
2. Therapeutic
3. Pharmacological
4. Chemical
Drug signature similarity predicts drug clinical activity

Predicting if 2 drugs share the same category using the drug signature similarity.

Hierarchical categorization of drugs according to ATC (from WHO):

1. Anatomical class
2. Therapeutic subclass
3. Pharmacological subclass
4. Chemical subclass

Question: can we exploit the latent representations for predictions in pharmacology?
Drug latent representations predict shared targets

There is a significant difference in the cosine similarity between drug signatures for pairs that share targets

Prediction of whether 2 drugs share molecular targets using similarity between drug signatures
Conclusions of Part 3

✓ A method for predicting the frequency of side-effects in the population.

✓ It tells us something about the biology of the problem

✓ It can be used for directing clinical trials.

✓ It can provide explanations
Acknowledgements

Horacio Caniza
Juan Caceres
Mateo Torres
Diego Galeano
Ruben Jimenez
Alfonso E. Romero
Haixuan Yang
Santiago Noto
Cheng Ye
Jessica Gliozzo

One postdoc position available
shared position between Royal Holloway and Yale University

http://www.paccanarolab.org