The inference of ground truth from biological literature

Alexander Belikov, James Evans, Andrey Rzhetsky
The University of Chicago

Introduction

It is known that in the process of scientific discovery is not straightforward. The intrinsic properties of certain scientific domains, such as

- the complexity of context, surrounding a statement
- the sociological bias towards positive results
- the complexity of the required experimental setup, prevent rapid convergence or even render certain statements undecidable

We try to find out if it is possible to infer:

- assess the correctness of a claim made by a scientific team relative to the "objective" biological interaction
- given a set of claims made by scientific teams, what the likely sign of the interaction is

A typical record in the Geneways database has the form: "abg" *prevents* "tert". We aggregate all positive and negative lexical actions (with the simplifying assumption that a claim with a positive action and negative flag set to $true$ translates to a negative action) and project these expression onto the boolean space $\{true, false\}$.

Datasets

- Geneways: 130K unique interactions, 198K unique claims, from 109K publications
- Literome: 144K unique interactions, 259K unique claims from 220K unique publications
- LINCS L1000: 77 cell-lines, 5K perturbagens, 11K downstream genes
- Geneways-Lincs projection: 23K unique interactions, 44K unique claims from 33K unique publications.
- Literome-Lincs projection: 36K unique interactions, 75K unique claims from 69K unique publications.

The overlap between Geneways and Literome (aligned with LINCS) is 89 claims (0.57 corr). Before alignment the overlap is 0.9K claims (0.35 corr). Correlation between LINCS and Geneways or Literome ≈ 0.

Features

Popularity
C_α - the set of claims pertaining to interaction α. The number of claims made before time t is the *observed popularity* $\nu_\alpha(t) = |\{c_i(t') : c_i(t') \in C_\alpha, t' < t\}|$. Also density. Windowed density. Log-normal works for a small fraction (power-law distribution for interactions). WIP: capture the fluctuations.

Network measures
The size of community in the weighted network of claims (calculated separately for Literome and Geneways). Using infomap. WIP: LINCS-induced network. Normalize popularity by community.

Time
Time passed from the first publication regarding the interaction: $\Delta_i = t_i - t_0^\alpha$

Journal/Affiliation Quality
Using the Article Influence (proxy to PageRank) to rank journals (using Web of Science, more than 50M publications). Affiliation quality taken from QS World University ranking. Using hierarchical clustering to disambiguate affiliation (cf. App.).

Citation metrics
Citation data was fit with a lognormal distribution (alternatively we use the window of w years.)

Herfindahl index
We compute "the share of the market" $s(t)$ for each interaction for institution and authors and from there derive the normalized HI: $H(t) = \sum s_i^2(t)$; NHI tells us how uniform or monopolististic the current domain is. Social independence.

Setup

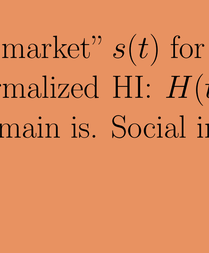
Index α refers to a specific interaction. π^α is the continuous strength of the interaction (z-score of over-expression from LINCS, which is also aggregated and tested for consistency). We discretize it, splitting the cdf of π^α into $(0, \epsilon]$, $(\epsilon, 1-\epsilon)$ and $[1-\epsilon, 1.0)$ and assign to the intervals "negative", "indifferent" and "positive". We fix $\epsilon = 0.2$ and define in this manner $\hat{\pi}^\alpha$ taking discreet values 0, 0.5 and 1.
After projecting GW and Literome on LINCS, $\hat{\pi}^\alpha$ acquires a bias (mostly positive interactions in literature) for Literome we have 50.4K claims, 4.8% pos, 1.5% neg, 93.7% ind; for Geneways 15.5K, 10% pos, 1.5% neg, 88.5% ind.
TP53 (7157) on CDKN1A (1026) interaction excluded from Literome dataset.
We define the distance $a_i^\alpha = |c_i^\alpha - \hat{\pi}^\alpha|$ and train classifiers $P\{a_i^\alpha(t)|f_{ik}^\alpha(t')\}$, where $t' < t$ assuming conditional independence at time t. NB: some of the features explicitly depend on c_i^α at earlier times. By symmetrizing, we reduce the number of parameters (rare classes are not numerous).
Having defined a_i we can infer the experimental strength from a sequence of claims $\{(c_i, f_{ij})\}$: invert the conditional probabilities using Bayes law under the simplifying assumption (index α omitted):

$$\begin{aligned} P(\pi|\{c_i, f_i\}) \propto & \quad P(\{c_i, f_i\}|e)P(\pi) \\ \propto & \quad \prod_i P(c_i, f_i|\pi)P(\pi) = \prod_i P(f_i)P(\pi)\sum P(c_i|a_i\pi)P(a_i|f_i) \end{aligned}$$

Appendix

- **Disambiguation problem - hierarchical string distance:** Affiliations are provided as ordered sets of phrases. e.g. "'Dept. Anim. Sci., Rutgers University, New Brunswick, NJ 08903". Solution: given a string metric, compute word to word distance, define the alphabet of words, compute the distance on metawords, etc. Cluster on the set of phrases distances.
- **Stratified random forest (Forest of Forests):** Cross-validation is performed on the observed frequencies of classes, however, for the training (depending on the metrics we are optimizing) we want to sample more the underrepresented classes. Solution: grow many forests sampled from training set with less extreme proportions. Improves AUROC by 30-40%.
- **Unsupervised approach, graphical models:**

Assume there are latent states which have Markovian or coarse Markovian structure and which generate the observable variables.



- **Generalized bin-packing (applied to GM inference):** Split the dataset in batches, so that the running times are similar, while sampling in stratas for feature pdfs to facilitate the convergence of the posteriors.

Results

Geneways

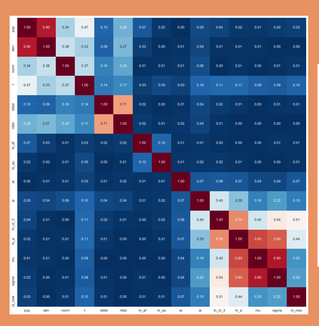


Fig. 2: Correlation matrix

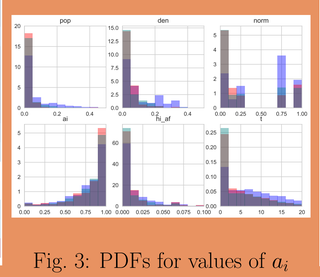


Fig. 3: PDFs for values of a_i

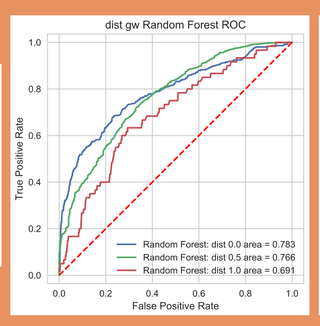


Fig. 4: AUC for a_i

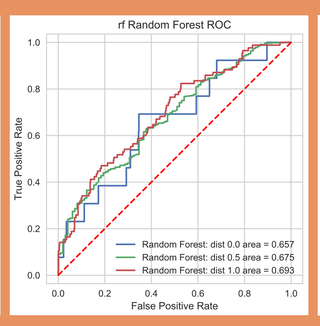


Fig. 5: AUC for $\hat{\pi}^\alpha$

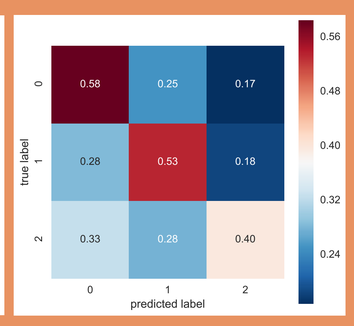


Fig. 6: Confusion matrix for $\hat{\pi}^\alpha$

Literome

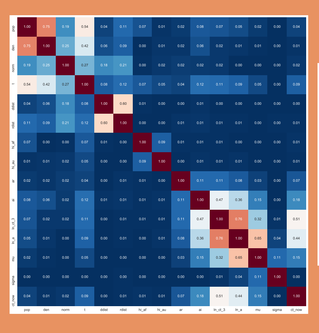


Fig. 7: Correlation matrix

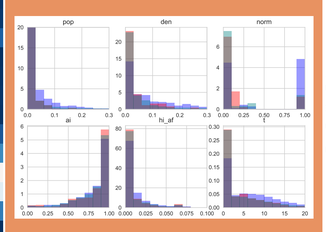


Fig. 8: PDFs for values of a_i

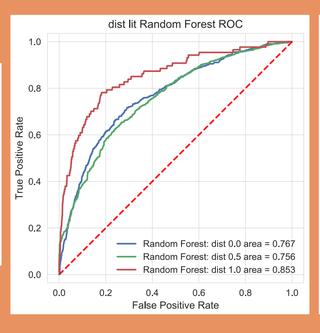


Fig. 9: AUC for a_i

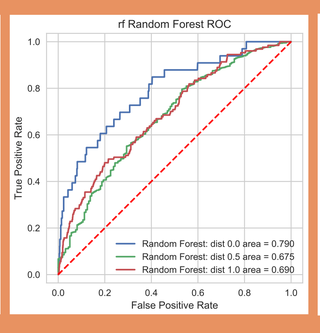


Fig. 10: AUC for $\hat{\pi}^\alpha$

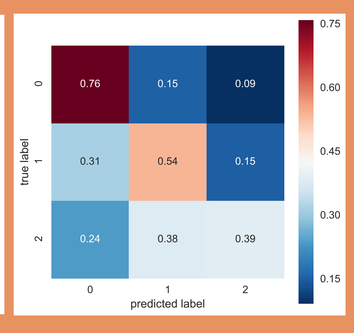


Fig. 11: Confusion matrix for $\hat{\pi}^\alpha$

Conclusions

a) positive features are the size of the community, the social attention, the latency
b) publication metrics have no effect on *the correctness* the claim
c) the correlations are zero between Geneways, Literome and Lincs
d) Geneways and Literome independently select the same features in the same order