Systematic explanation of drug effects and gene dependency correlations using mechanistic networks assembled from literature mining and databases

John A. Bachman[1], Benjamin M. Gyori[1], Klas Karis[1], and Peter K. Sorger[1]

[1]Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA

Availability: https://github.com/sorgerlab/indra

Funding: DARPA ASDF program, ARO contract W911NF018-1-0124

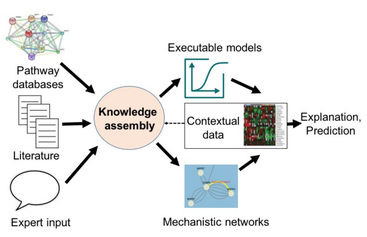
INTRODUCTION

Biology is currently grappling with the challenge of integrating large datasets with a body of scientific knowledge that has grown too large and complex for any single scientist to read or understand. The size and scope of large-scale data compendia ("big data") has created analytical challenges that require new solutions, ideally ones that make use of aggregated scientific knowledge. The capacity of modern experimental methods to generate data about biological processes has surpassed the ability of existing informatics approaches to generate meaningful mechanistic explanations. Mechanistic systems biology models could potentially address this gap, but model construction remains a labor-intensive process requiring both biological knowledge and modeling expertise. As a result, modeling studies remain fairly small in scope and are disconnected from genome-scale research. For mechanistic models to attain the necessary scope, methods for the automated assembly and analysis of large models from available knowledge sources will be required. Here we describe the use of the Integrated Network and Dynamical Reasoning Assembler (INDRA)[1] to assemble mechanistic facts from databases and literature into different types of models for explanation of large datasets.

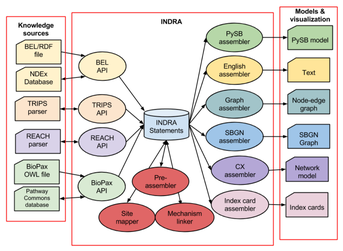
RESULTS

System architecture and approach

The Integrated Network and Dynamical Reasoning Assembler (INDRA)[1] automatically assembles mechanistic models from pathway databases, literature, and expert knowledge expressed in natural language. INDRA draws on three existing natural language processing systems[4,5,6] and uses a modular architecture to build different types of models from a variety of sources.



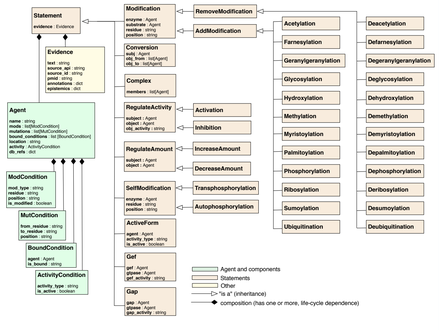
Conceptual overview of automated assembly



INDRA software architecture

Mechanisms are normalized into *Statements*

Mechanisms extracted from each source format are normalized into Statements, an SBO-compatible internal representation, where they are processed to remove errors, identify overlaps, and estimate reliability. Statements are designed to correspond in both specificity and ambiguity to descriptions of biochemistry as found in text (e.g., "MEK1 phosphorylates ERK2", rather than a detailed reaction mechanism). The representation currently encompasses post-translational modifications, chemical conversions, protein expression and degradation, and generic activation/inhibition relationships.

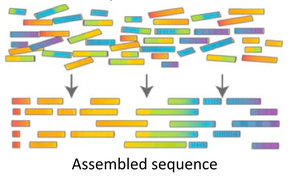


The assembly challenge

Assembly of a large number of mechanistic facts is analogous to genome assembly: databases and literature yield a large number of redundant, partially overlapping facts that may contain errors. Mechanisms must be corrected and "aligned" in order to produce a set of facts suitable for generating a non-redundant, non-degenerate model.

Genome assembly

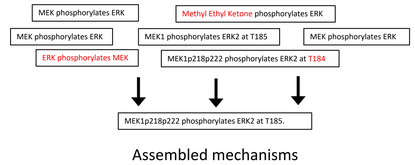
Sequence reads



Assembled sequence

Knowledge assembly

"Raw" mechanisms

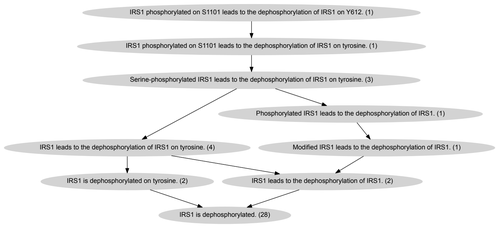


Assembled mechanisms

Identifying relationships between mechanisms

A key challenge in assembling detailed mechanistic networks is that a single mechanism may be described at different levels of specificity among the literature and various databases. Reconciling these overlapping mechanisms is essential to eliminate spuriously distinct edges in the assembled model. Using hierarchical ontologies of protein modification types, activity types, and the protein family information provided in Bioentities, INDRA implements duplicate removal, hierarchy-based redundancy resolution, and other forms of error correction and mechanism linking.

Relations can be organized into a hierarchy based on their specificity



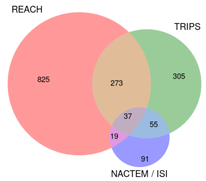
Estimating the reliability of extracted mechanisms

Even state-of-the-art NLP and text mining algorithms have limited accuracy, with roughly 20-30% of extracted relations representing a misinterpretation of the corresponding sentence ("reader error"). Given empirical estimates of the per-sentence error rate for different readers, INDRA's BeliefEngine component aggregates results to estimate the overall probability that a relation is the result of reader error. It accomplishes this by:

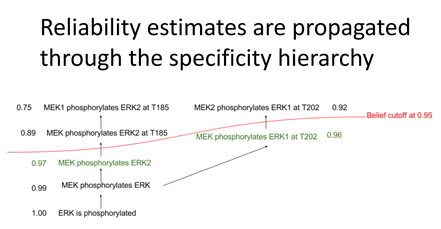
1) aggregating evidence from multiple sentences read by the *same reader*
2) aggregating results from different reading algorithms on the *same sentence*
3) propagating error estimates through the network of related statements

Mechanisms can then be filtered with a precision threshold (e.g., 95% confidence).

Reading systems produce partially overlapping extractions

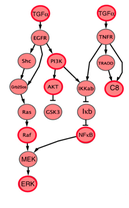


Reliability estimates are propagated through the specificity hierarchy



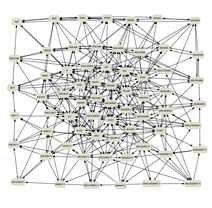
Model representations for statically identifying causal paths

Directed protein interaction graph



Boolean network

Kappa rule influence map[2]



Chemical reaction network

Mechanistic detail/causal context

More false positive paths (less stringent context)

More false negative paths (more stringent context)

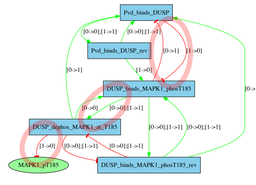
In directed interaction graphs, the relatively limited causal context leads to an explosion of paths between any two proteins. This leads to many false positive paths and makes identification of long causal chains difficult (or even intractable) in large networks.

Generating explanations from the Kappa[2] rule influence map

The Kappa influence map captures detailed context while avoiding the combinatorial explosion of chemical species. Paths are obtained by:

- identifying rules whose activity is increased by the abundance of the subject (e.g., drug)
- searching for a path to an observable representing the object (e.g., a measured protein) with the appropriate **overall polarity**
- scoring paths by whether the signs of measured intermediate nodes are correctly predicted

Causal path for "Pervanadate increases MAPK1 phosphorylation"



Use case 1: interpreting phosphoproteomic data

To evaluate the ability of INDRA to systematically generate explanations of high-throughput data, we assembled a rule-based executable model to explain a previously published dataset of the phospho-proteomic response of a melanoma cell line to 12 different drugs.[3] A rule-based model containing 221 proteins and 1451 rules was assembled from mechanisms extracted from databases and ~95,000 publications (abstracts and full texts). Static analysis of the rule influence map provided by Kappa identified possible mechanistic paths linking drug targets to experimentally observed effects on phosphoprotein abundances.

http://www.sanderlab.org/pertbio/

Performance: For the largest effects in the data (>50% fold-change) the model generated biochemically plausible explanations for 20 of the 22 effects (91%). For effects at the 20% fold-change level, the model explained 95/135 (70%) of effects. Notably, performance was biased toward drug targets well-represented in the literature corpus: the model explained 94/106 (89%) of effects due to PI3K, PKC, SRC, MTOR, MEK, AKT, RAF, and JAK inhibition, but only 1/29 (3%) of effects due to CDK, STAT or MDM2 inhibition.

Drug Target	Antibody	Fold-change	Path ?
MEK	MAPK pT202	0.47	
SRC	CHK2 pT68	1.75	
SRC	4EBP1 pT37	0.44	
AKT	AKT pT308	0.25	
AKT	GSK3A/B pS21	0.44	
AKT	AKT pS473	0.17	
AKT	S6 pS235	0.36	
CDK4	4EBP1 pS65	0.44	
CDK4	YBI pS102	2.13	
MTOR	AKT pT308	2.19	
MTOR	S6 pS240	0.05	
MTOR	AKT pS473	3.19	
MTOR	p70S6K pT389	0.33	
MTOR	S6 pS235	0.05	
PKC	GSK3A/B pS21	1.59	
PKC	S6 pS240	0.47	
PKC	S6 pS235	0.3	
PI3K	p70S6K pT389	0.5	
PI3K	S6 pS240	0.44	
PI3K	AKT pS473	0.2	
PI3K	S6 pS235	0.27	

Example explanation: *How does Src inhibition increase CHK2 pT68?*

SRC phosphorylated on Y418
phosphorylates PAK2 on S20. PAK2
phosphorylated on S20
phosphorylates RAF1 on S338. RAF1
phosphorylated on S338, T269 and
S471 phosphorylates MAPK1 on T185.
MAPK1 phosphorylated on T185 and
Y187 phosphorylates TP53 on S15.
TP53 phosphorylated on S20 and S15
decreases the amount of PLK1. PLK1
phosphorylates CHEK2 on T68, which
is measured by CHK2_pT68.

Use case 2: explaining gene dependency correlations

We are evaluating the effectiveness of our text-mined network of ~7.5 million unique interactions mined from the literature to identify mechanisms mediating correlations between gene knockouts in the Broad Cancer Dependency Map dataset (http://depmap.org)

Of the 1.7 million correlations (absolute value > 0.3) we found that while only 6,043 (0.3%) could be explained by a known mechanistic link, 22% could be explained with a single intermediate linking gene, either as a common downstream target or part of a pathway.

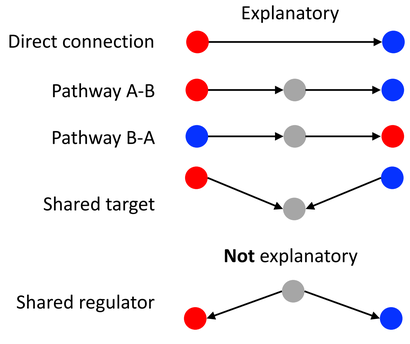
Type	Count	Percent
(All correlations)	155,417,265	
Correlations > 0.3	1,763,551	100.0
Direct link	6,043	0.3
Pathway or shared target	380,404	21.6
Correlations explained	386,447	21.9

For a subset of 48,000 correlations involving 340 cancer-related genes, we found that a higher proportion (75%) of the correlations can be explained with a single intermediate linking gene.

We are building a browser, the INDRA DepMap Explainer, to explore these explanations.

Example: correlation of PTPN11 (Shp2) with EXTL3: 0.42

Network configurations



Example Explanation: EXTL3 ⟶ FGF2 ⟶ PTPN11

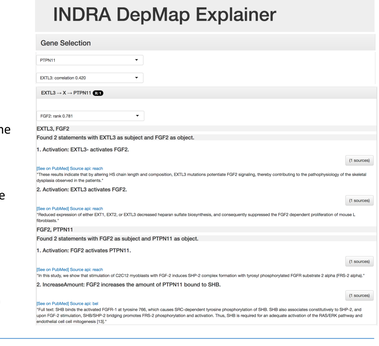
Support for EXTL3->FGF2:

"These results indicate that by altering HS chain length and composition, EXTL3 mutations potentiate FGF2 signaling, thereby contributing to the pathophysiology of the skeletal dysplasia observed in the patients." (PMID 28148688)

"Reduced expression of either EXT1, EXT2, or EXTL3 decreased heparan sulfate biosynthesis, and consequently suppressed the FGF2 dependent proliferation of mouse L fibroblasts." (PMID 29305908)

Support for FGF2 -> PTPN11:

"In this study, we show that stimulation of C2C12 myoblasts with FGF-2 induces SHP-2 complex formation with tyrosyl phosphorylated FGFR substrate 2 alpha (FRS-2 alpha)."



REFERENCES

1. B. M. Gyori*, J. A. Bachman*, K. Subramanian, J. L. Muhlich, L. Galescu, and P. K. Sorger. "From word models to executable models of signaling networks using automated assembly." *bioRxiv*, 2017.
2. V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. "Rule-Based Modeling of Cellular Signaling." *Concurrency Theory (CONCUR) 2007, Lecture Notes in Computer Science,* 4703:17–41, 2007.
3. E. J. Molinelli, A. Korkut, et al., "Perturbation biology: Inferring signaling networks in cellular systems." *PLoS Computational Biology,* 9(12):e1003290, Dec 2013.
4. J. Allen, W. de Beaumont, L. Galescu, and C. M. Teng. "Complex event extraction using DRUM." 2015.
5. M. A. Valenzuela-Escarcega, G. Hahn-Powell, T. Hicks, and M. Surdeanu. "A domain-independent rule-based framework for event extraction." In *Proc. 53rd Annual Meeting of the ACL-IJCNLP,* 2015.
6. D. McDonald et al., "Extending Biology Models with Deep NLP *over* Scientific Articles." *Workshops at the 30th AAAI Conference on Artificial Intelligence,* 2016.
7. C. F. Lopez*, J. L. Muhlich*, J. A. Bachman*, and P. K. Sorger. "Programming biological models in python using PySB." *Molecular Systems Biology,* 9(1):646–646, Apr 2014.