Datanator: an integrated database of molecular data for quantitatively modeling cellular behavior

Yosef D. Roth†, Zhouyang Lian†, Saahith Pochiraju, Bilal Shaikh & Jonathan R. Karr*

August 6, 2020

Icahn Institute for Data Science and Genomic Technology and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

†These authors contributed equally to this work

*Corresponding author: Jonathan Karr (karr@mssm.edu)

Abstract

Integrative research about multiple biochemical subsystems has significant potential to help advance biology, bioengineering, and medicine. However, it is difficult to obtain the diverse data needed for integrative research. To facilitate biochemical research, we developed Datanator (https://datanator.info), an integrated database and set of tools for finding *clouds* of multiple types of molecular data about specific molecules and reactions in specific organisms and environments, as well as data about chemically-similar molecules and reactions in phylogenetically-similar organisms in similar environments. Currently, Datanator includes metabolite concentrations, RNA modifications and half-lives, protein abundances and modifications, and reaction rate constants about a broad range of organisms. Going forward, we aim to launch a community initiative to curate additional data. Datanator also provides tools for filtering, visualizing, and exporting these data clouds. We believe that Datanator can facilitate a wide range of research from integrative mechanistic models, such as whole-cell models, to comparative data-driven analyses of multiple organisms.

1. Introduction

Integrative research about multiple biochemical subsystems has significant potential to help advance biology, bioengineering, and medicine. For example, whole-cell models that account for all of the biochemistry in cells could help scientists conduct experiments in silico, help physicians personalize medicine, and help engineers design microbes.[1,2]

Historically, it has been challenging to obtain the diverse and extensive data needed to analyze multiple subsystems. For example, Resource Balance Analysis (RBA) requires data about the composition of cells and macromolecular complexes, the abundances of enzymes, and the maximum rates of reactions.[3] Whole-cell models require even more data, such as the concentration of each metabolite, the abundance and turnover of each RNA and protein, and the rate law of each reaction.

New experimental technologies have begun to produce the data needed for integrative research. For example, RNA-seq has enabled transcriptome-wide profiles of RNA abundances and half-lives, tandem mass spectrometry has enabled proteome-scale profiles of protein abundances and half-lives, and liquid chromatography-mass spectrometry has enabled metabolome-scale profiles of metabolite concentrations. Furthermore, this data is increasingly accessible. For example, ArrayExpress contains over 400,000 RNA expression profiles.

Several investigators have begun to use this data to conduct more integrative analyses of biochemical networks. For example, Sanchez et al.[4] used proteomic data and reaction rate parameters to develop a constraint-based model of the gene expression and metabolism of yeast; Thiele et al.[5] used genetic information and other data to build a constraint-based model of the transcription, RNA modification, translation, complexation, and metabolism of *Escherichia coli*; Goelzer et al. used metabolomic and proteomic data to develop an RBA model of 72 subsystems of *Bacillus subtilis*;[3] and we and others used a wide range of data[6] to develop a hybrid model of 28 subsystems of *Mycoplasma genitalium*.[7] However, even the most extensive studies have only used a fraction of the available data. Encouraged by these successes, we believe that more comprehensive analysis is possible with more data.

Although extensive data is available, a recent community survey revealed that obtaining data is one of the biggest bottlenecks to integrative research.[8] One fundamental barrier is our limited ability to experimentally characterize biochemistry. For example, the most comprehensive metabolomics datasets only capture a small fraction of the metabolome. Long-term, the community must develop new measurement technologies.

It is also difficult to utilize the existing data: the existing data is siloed in different databases and articles for different types of data and organisms; the existing databases and articles use different formats, identifiers, and units (e.g., PAXdb provides data in TSV format, whereas SABIO-RK provides data in SBML format); the APIs to the existing databases have different interfaces; and the existing databases provide limited tools for finding data for similar biology for imputing missing information. This heterogeneity makes it difficult to compare and integrate data from multiple sources.

We believe these secondary challenges can be addressed with practical computational solutions. To address these challenges, we developed Datanator (https://datanator.info), an integrated database of molecular data and a single web interface and API for finding a broad range of quantitative and categorical data about a wide range of organisms.

To start, we have aggregated several key types of data, including metabolite concentrations, RNA modifications and half-lives, protein abundances and modifications, and reaction rate parameters from several databases and articles. Going forward, we aim to launch a community initiative to curate additional data. To help researchers search, compare, and integrate this heterogeneous data, we have normalized the annotation, units, and representation of the data.

On top of this database, Datanator provides tools for finding relevant data for research. Like other databases, Datanator provides tools for finding data about specific molecules and reactions in particular organisms and environments. In the absence of direct data, Datanator also provides unique tools for building *clouds* of data that include chemically-similar molecules and reactions in phylogenetically-similar organisms and similar environments. For example, Datanator can find data about the concentration of ATP and similar metabolites in *E. coli* or other enterobacteria in M9 or other minimal media. These clouds can help researchers impute missing information. Datanator also provides tools for filtering these clouds, charting their distribution, and exporting them to files.

We believe that Datanator can facilitate a wide range of research. For example, Datanator can help investigators find information for comparative analyses of multiple organisms and tissues, multi-dimensional analyses of relationships within biochemical networks, and integrative mechanistic models of multiple subsystems such as whole-cell models.

Here, we describe the content of Datanator and the search and visualization tools. In addition, we summarize how we implemented Datanator, compare Datanator to several existing databases, outline the types of research that Datanator can facilitate, and discuss how we plan to enhance Datanator.

2. Integrated database of molecular data

Because research often requires many types of data, Datanator includes several types of quantitative and categorical data. Currently, this includes 3,841 measurements of the concentrations of 1,621 metabolites aggregated from ECMDB,[9] YMDB,[10] and several articles; 589 reconstructed RNA modification profiles aggregated from MODOMICS;[11] 75,836 measured RNA half-lives aggregated from several articles; 2,634,941 measurements of the abundance of 846,970 proteins aggregated from PaxDB;[12] 4,083 reconstructed protein modification profiles aggregated from PRO;[13] and 61,734 measurements of the rate parameters of 37,858 reactions aggregated from SABIO-RK[14] (Figure 1a). In addition, we are working to include nearly 1,000,000 measurements of RNA localizations from RNALocate,[15] lncATLAS,[16] and one article; over 100,000 measured and predicted protein localizations from eSLDB,[17] the Human Protein Atlas,[18] PSORTdb,[19] and SubCell;[20] more than 50,000 measurements of protein half-lives from nine publications; more than 200,000 additional measurements of reaction parameters from BRENDA;[21] and reaction flux measurements for 36 organisms from CeCaFDB.[22]

In total, Datanator currently includes data for 1,030 organisms (Figure 1b) across a wide range of environmental conditions (Figure 1c,d) from over 8,000 articles (Figure 1e).

3. Tools for searching the sea of data

Datanator also provides unique tools for searching for relevant data for research. Like other databases, Datanator provides basic functionality for finding data about specific molecules and reactions in particular organisms and environments. On top of this, Datanator provides tools for assembling *clouds* of related data – ensembles that also encompass measurements of similar molecules and reactions in similar organisms and environments. In the absence of direct measurements, these clouds can help researchers impute missing information.

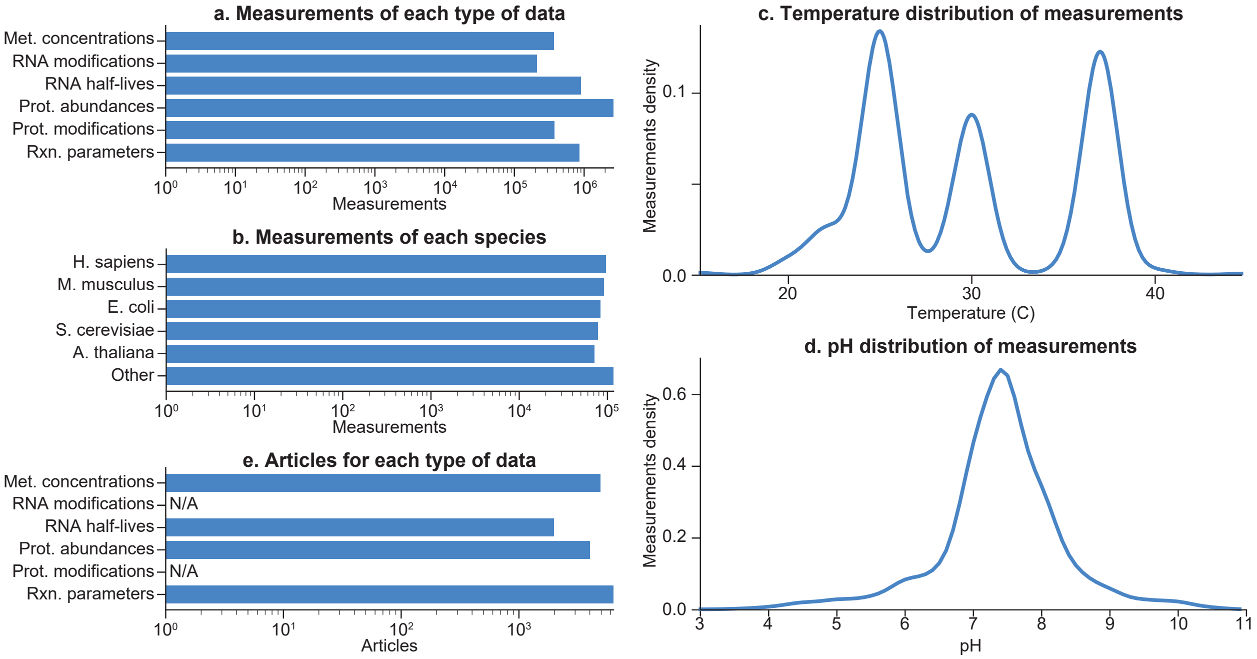


Figure 1. Datanator encompasses several key types of molecular data about a wide range of organisms in a broad range of environments. (a) Number of measurements of each type of data. **(b–d)** Distributions of the genotype and environment of the measurements in the database. **(e)** Number of articles that Datanator integrates data from for each type of data.

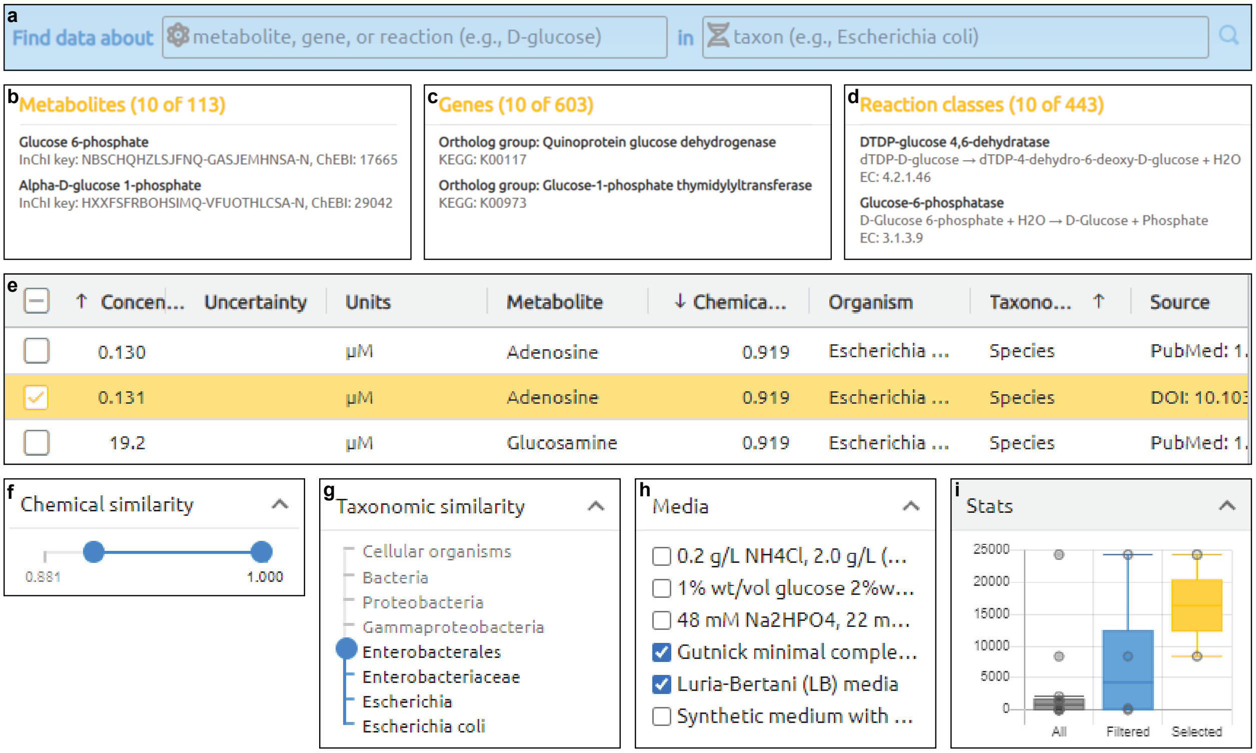
Basic searching for specific measurements

Users can use the search form (Figure 2a) to query for data about a specific metabolite, gene (RNA or protein), or reaction in a specific organism. Users can search for data about molecules and reactions by their names, identifiers (e.g., ChEBI id), chemical structures (e.g., InChI representation), and descriptions. Datanator recognizes all organisms in the NCBI Taxonomy database. After submitting the form, Datanator presents three lists of search results for metabolites, genes, and reactions, each ranked by their relevance to the query (Figure 2b–d). Users can click on each result to obtain tables of measured properties (e.g., concentration, half-life) of the selected molecule or reaction (Figure 2e).

Advanced searching for clouds of similar measurements

While the amount of data has exploded over the last decade, we still do not have data about every molecule and reaction in every organism and environment. Instead, researchers often must impute missing information from data about similar molecules and reactions in related organisms and environments. However, it is difficult to find data about similar biology. For example, searching Google for data about ATP synthase or its orthologs in *E. coli* or other Enterobacteria such as *Klebsiellas* requires combinatorially-many searches for each ortholog in each Enterobacterial species.

To help researchers conduct research even without direct measurements, Datanator provides tools for assembling *clouds* of data around a specific molecule or reaction in a particular organism and environment. For metabolites, Datanator identifies measurements of chemically-similar metabolites (Tanimoto coefficient $\geq$ 0.65). For genes, Datanator identifies measurements of the same KEGG Orthology group. For reactions, Datanator identifies measurements of the same Enzyme


a
Find data about
metabolite, gene, or reaction (e.g., D-glucose)
in
taxon (e.g., Escherichia coli)
b
c
d
e
f
g
h
i

Metabolites (10 of 113)

Glucose 6-phosphate
InChI key: NBSCHQHZLSJFNQ-GASJEMHNSA-N, ChEBI: 17665

Alpha-D-glucose 1-phosphate
InChI key: HXXFSFRBOHSIMQ-VFUOTHLCSA-N, ChEBI: 29042

Genes (10 of 603)

Ortholog group: Quinoprotein glucose dehydrogenase
KEGG: K00117

Ortholog group: Glucose-1-phosphate thymidylyltransferase
KEGG: K00973

Reaction classes (10 of 443)

DTDP-glucose 4,6-dehydratase
dTDP-D-glucose → dTDP-4-dehydro-6-deoxy-D-glucose + H2O
EC: 4.2.1.46

Glucose-6-phosphatase
D-Glucose 6-phosphate + H2O → D-Glucose + Phosphate
EC: 3.1.3.9

⊟	↑ Concen...	Uncertainty	Units	Metabolite	↓ Chemica...	Organism	Taxono... ↑	Source
□	0.130		μM	Adenosine	0.919	Escherichia ...	Species	PubMed: 1
☑	0.131		μM	Adenosine	0.919	Escherichia ...	Species	DOI: 10.10
□	19.2		μM	Glucosamine	0.919	Escherichia ...	Species	PubMed: 1

Chemical similarity
0.881 – 1.000

Taxonomic similarity
Cellular organisms
Bacteria
Proteobacteria
Gammaproteobacteria
● Enterobacterales
Enterobacteriaceae
Escherichia
Escherichia coli

Media
□ 0.2 g/L NH4Cl, 2.0 g/L (...
□ 1% wt/vol glucose 2%w...
□ 48 mM Na2HPO4, 22 m...
☑ Gutnick minimal comple...
☑ Luria-Bertani (LB) media
□ Synthetic medium with ...

Figure 2. Datanator provides interactive tools for creating and visualizing clouds of data centered on specific molecules and reactions in specific organisms and environments. (**a**) Search form. (**b–d**) Search results grouped by class (metabolites, genes, or reaction). (**e**) Tables of clouds of potentially relevant measurements about a molecule or reaction. (**f–h**) Tools for filtering the data clouds. (**i**) Charts for visualizing the distributions of the data clouds.

Commission (EC) class. Datanator identifies measurements from all organisms and environments.

Datanator displays these clouds using the data tables described above (Figure 2e). To help users understand these clouds, the data tables include columns that describe the similarity of each measurement to the search query. For metabolites, Datanator displays the Tanimoto coefficient between the target and measured metabolites. Datanator indicates the phylogenetic relevance of each measurement by displaying the taxonomic rank of the most recent canonical-ranked common ancestor (e.g., species, genus) between the target and measured organisms.

To help users narrow these data clouds to the most relevant measurements, Datanator provides sliders and select menus for filtering measurements by their metadata (Figure 2f–h). This includes a slider for filtering metabolite concentration measurements by their Tanimoto coefficient, a slider for filtering measurements by their taxonomic distance, sliders for filtering measurements by their temperature and pH, and a select menu for filtering measurements by their growth media.

In addition, Datanator provides users checkboxes for creating data clouds manually (Figure 2e). These fine-grained controls can help users create clouds that only include the data that they believe are relevant to their research.

Finally, Datanator provides buttons for exporting data clouds to CSV and JSON files.

Programmatic searching via a REST API

Users can also programmatically query Datanator via a REST API. The API makes it easy for users to obtain data about multiple molecules or reactions in multiple organisms.

4. Tools for visualizing data clouds

To help researchers interpret their data clouds, Datanator provides charts of the distributions of all of the potentially relevant data, the filtered measurements, and the selected measurements (Figure 2i). Users can mouse over each measurement to highlight it in the corresponding data table.

5. Implementation

Datanator consists of a database of molecular data, an OpenAPI-compliant REST API for querying the database, and a web application for graphically exploring the database.

Database construction

We assembled the Datanator database in several steps. First, we used Google, Google Scholar, and Microsoft Academic to find potential sources for each type of data. Second, for each kind of data, we decided to focus on the largest databases. In the absence of an existing database, we decided to focus on articles that report genome-scale measurements.

Third, we downloaded and parsed each source. Where available, we used file downloads and REST APIs. For example, we downloaded PaxDB as an archive of TSV files, and we downloaded the Protein Ontology as an OBO file. For databases that do not provide a download or complete API, we wrote scripts to scrape and parse their websites.

Fourth, we normalized each measurement. We normalized the subject of each measurement (molecule or reaction) by mapping it to one of several namespaces. We used Open Babel[23] to map each metabolite to the InChI representation of its structure, mapped each mRNA and protein to its UniProt id, and mapped each reaction to the InChI representations of its participants. We used BpForms[24] to normalize the representation of each RNA and protein modification profile. We normalized the taxonomic context of each measurement by using the ETE toolkit[25] to map each organism to its NCBI Taxonomy id. We used the pint Python package to normalize the units of each measurement. Lastly, we combined the measurements into a single MongoDB database.

To enable us to pull updates to each source into Datanator going forward, we implemented the database construction as a repeatable script.

REST API

We implemented the API using Python, connexion, pymongo, and Elasticsearch.

Web application

We implemented the web application using the React framework and several packages, including ag-Grid (data tables), Blueprint (search form), and ChartJS (charts).

Testing

We tested the database construction and REST API using unittest. We tested the web application using Cypress and Jest.

6. Comparison to existing databases

Numerous resources help researchers find data, including databases for individual types of data and model organism, genome, pathway, and federated databases. While several of these resources overlap with Datanator, Datanator uniquely contains multiple types of quantitative and categorical molecular data for multiple organisms. When direct measurements are unavailable, Datanator also provides unique tools for finding data about similar biology.

Data type-specific databases

Although most of Datanator's data is available from other databases such as ECMDB and PaxDB, it is challenging to use these databases together because they provide data via different interfaces in different formats, and they annotate their data with different identifiers and ontologies. In addition, these databases provide few tools for obtaining data about similar biology for imputing missing information. Datanator makes their data more accessible by providing a central, consistent portal to the data and a complete API. Datanator also provides tools for building clouds of their data. Additionally, Datanator has more metabolite concentrations than ECMDB or YMDB because it combines ECMDB, YMDB, and several articles.

Model organism databases

Several model organism databases have overlapping content with Datanator. EcoCyc contains reaction kinetic parameters for *E. coli*.[26] SubtiWiki contains RNA and protein expression data for *B. subtilis*.[27] However, these databases focus on genetic and relational data, and they each only contain data for one organism. The CCDB (*E. coli*, 28), MyMPN (*Mycoplasma pneumoniae*, 29), and WholeCellKB (*M. genitalium*, 6) contain multiple types of data. Datanator improves upon these databases by incorporating data for multiple organisms and providing more powerful search tools.

Genome and pathway databases

Several genome and pathway databases also have overlapping content with Datanator. For example, Reactome contains RNA and protein expression data[30] and UniProt contains data about protein modifications.[31] However, these databases focus on genetic and relational information, such as the locations of promoters and their regulators. These databases have limited quantitative information, such as concentrations.

Federated databases

Like Datanator, federated databases such as BioMart[32] and Omics DI[33] are also central portals for discovering multiple types of data. While both databases contain more data than Datanator, their content has little overlap with Datanator. Furthermore, the content of Datanator is more integrated than that of BioMart and Omics DI. This integration makes it easier for researchers to search Datanator and use its data.

7. Use cases

Through helping researchers obtain data, we believe that Datanator can assist a wide range of tasks such as comparative analyses of multiple tissues and organisms, multi-dimensional analyses of molecular networks, calibrating mechanistic models, and validating conclusions with independent data.

Multi-dimensional analyses of biochemical networks

Because Datanator contains multiple types of data, Datanator is a good source of data for analyzing relationships among molecules, reactions, and their chemical properties. For example, Datanator could provide data for analyzing the correlation between the abundances of enzymes and their maximum reaction velocities, classifying reactions as bottlenecks by comparing metabolite concentrations and enzyme-metabolite affinities, or learning a regression model of the half-lives of RNAs from their sequences and modifications.

Calibrating mechanistic models

For similar reasons, Datanator is also a good source of data for calibrating mechanistic models. For example, Datanator could provide protein abundances to constrain missing initial conditions; provide metabolite concentrations to construct a more comprehensive biomass equation for a flux balance analysis (FBA) model; or provide organism, tissue, or cell-type-specific protein abundances for recalibrating a model to create variants for different organisms or cell types. In particular, Datanator is ideal for integrative modeling of multiple biochemical subsystems such as whole-cell models. For example, Datanator could provide protein abundances and maximum reaction velocities to expand an FBA model to capture gene expression and further constrain the model.[4]

Validating computationally-generated conclusions

In the absence of an experimental collaboration, computational scientists often validate their conclusions with independent data from articles. However, this is often time-consuming because this requires finding data about specific molecules or reactions in particular organisms. By making it easier to find data about specific biology, Datanator empowers computational scientists to validate their own conclusions.

8. Discussion

In summary, Datanator is a unique portal for obtaining several key types of quantitative and categorical molecular data. The foundation of Datanator is an integrated database of several types of data about numerous organisms. This includes the first collection of multiple studies of RNA half-lives. Datanator is also the first database to contain the concentrations of metabolites in multiple organisms. To help researchers use this data, Datanator provides unique tools for assembling clouds of data centered on specific molecules and reactions in particular genotypes and environments. In addition, Datanator provides charts for visualizing the distribution of these clouds. These clouds can help users obtain informative data even in the absence of direct measurements. We anticipate that Datanator will help researchers build integrative models of multiple biochemical subsystems, including models of entire cells. We believe that Datanator is also a valuable source of data for multi-dimensional analyses of biochemical networks and comparative studies of multiple organisms.

Community initiative to curate additional data

Going forward, we aim to launch a community initiative to curate additional data. To facilitate community contributions, we aim to make it easy to submit data. To start, we have developed a single format for representing any type of data and a tool for validating data sets. This format outlines how investigators can submit data. Already, this is enabling several high school students to contribute data about protein localizations and half-lives. Ultimately, we plan to accept data through a simple form. We welcome suggestions via GitHub issues and pull requests or by joining our team.

Additional search and analysis tools

To further help users use the Datanator data, we plan to develop additional search and analysis tools. For example, we aim to create charts of the global distribution of each type of data; implement a tool for extracting multi-dimensional data clouds such as clouds of the expression and half-lives of mRNA and their proteins; and develop finer-grained filters for data about similar RNAs, proteins, and reactions based on the sequence similarity of RNAs and proteins and the structural similarity of the participants of reactions.

Higher-level tools for integrative modeling

In addition, we aim to develop higher-level tools for leveraging Datanator for integrative modeling. For example, we aim to build tools that further automate the process of using data from Datanator to constrain the values of missing model parameters, recalibrate models to represent different cell types, and further constrain FBA models with enzyme abundances and reaction parameters.

9. Availability

The application and API are available at https://datanator.info and https://api.datanator.info along with a tutorial and documentation. The data and software are openly available under the CC BY-NC-ND and MIT licenses, respectively.

10. Acknowledgements

We thank Yin Hoon Chew, Paul Lang, Wolfram Liebermeister, and the Center for Reproducible Biomedical Modeling for feedback.

11. Funding

This work was supported by National Institutes of Health awards P41EB023912 and R35GM119771, National Science Foundation award 1548123, and the Icahn Institute of Data Science and Genomic Technology.

Conflict of interest statement.

None declared.

References

1. Goldberg,A. P., Szigeti,B., Chew,Y. H., Sekar,J. A., Roth,Y. D. and Karr,J. R. (2018) Emerging whole-cell modeling principles and methods. *Curr. Opin. Biotechnol.,* **51**, 97–102.

2. Karr,J. R., Takahashi,K. and Funahashi,A. (2015) The principles of whole-cell modeling. *Curr. Opin. Microbiol.,* **27**, 18–24.

3. Goelzer,A., Muntel,J., Chubukov,V., Jules,M., Prestel,E., Nölker,R., Mariadassou,M., Aymerich,S., Hecker,M., Noirot,P. et al. (2015) Quantitative prediction of genome-wide resource allocation in bacteria. *Metab. Eng.,* **32**, 232–243.

4. Sánchez,B. J., Zhang,C., Nilsson,A., Lahtvee,P. J., Kerkhoven,E. J. and Nielsen,J. (2017) Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.,* **13**, 935.

5. Thiele,I., Jamshidi,N., Fleming,R. M. and Palsson,B. Ø. (2009) Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.,* **5**, e1000312.

6. Karr,J. R., Sanghvi,J. C., Macklin,D. N., Arora,A. and Covert,M. W. (2013) WholeCellKB: model organism databases for comprehensive whole-cell models. *Nucleic Acids Res.,* **41**, D787–D792.

7. Karr,J. R., Sanghvi,J. C., Macklin,D. N., Gutschow,M. V., Jacobs,J. M., Bolival Jr,B., Assad-Garcia,N., Glass,J. I. and Covert,M. W. (2012) A whole-cell computational model predicts phenotype from genotype. *Cell,* **150**, 389–401.

8. Szigeti,B., Roth,Y. D., Sekar,J. A., Goldberg,A. P., Pochiraju,S. C. and Karr,J. R. (2018) A blueprint for human whole-cell modeling. *Curr. Opin. Syst. Biol.,* **7**, 8–15.

9. Sajed,T., Marcu,A., Ramirez,M., Pon,A., Guo,A. C., Knox,C., Wilson,M., Grant,J. R., Djoumbou,Y. and Wishart,D. S. (2016) ECMDB 2.0: a richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Res.,* **44**, D495–D501.

10. Ramirez-Gaona,M., Marcu,A., Pon,A., Guo,A. C., Sajed,T., Wishart,N. A., Karu,N., Djoumbou Feunang,Y., Arndt,D. and Wishart,D. S. (2017) YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res.,* **45**, D440–D445.

11. Boccaletto,P., Machnicka,M. A., Purta,E., Piątkowski,P., Bagiński,B., Wirecki,T. K., de Crécy-Lagard,V., Ross,R., Limbach,P. A., Kotter,A. et al. (2018) MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.,* **46**, D303–D307.

12. Wang,M., Herrmann,C. J., Simonovic,M., Szklarczyk,D. and von Mering,C. (2015) Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics,* **15**, 3163–3168.

13. Natale,D. A., Arighi,C. N., Blake,J. A., Bona,J., Chen,C., Chen,S.-C., Christie,K. R., Cowart,J., D'Eustachio,P., Diehl,A. D. et al. (2017) Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic Acids Res.,* **45**, D339–D346.

14. Wittig,U., Rey,M., Weidemann,A., Kania,R. and Müller,W. (2018) SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Res.,* **46**, D656–D660.

15. Zhang,T., Tan,P., Wang,L., Jin,N., Li,Y., Zhang,L., Yang,H., Hu,Z., Zhang,L., Hu,C. et al. (2017) RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.,* **45**, D135–D138.

16. Mas-Ponte,D., Carlevaro-Fita,J., Palumbo,E., Pulido,T. H., Guigo,R. and Johnson,R. (2017) LncATLAS database for subcellular localization of long noncoding RNAs. *RNA,* **23**, 1080–1087.

17. Pierleoni,A., Martelli,P. L., Fariselli,P. and Casadio,R. (2007) eSLDB: eukaryotic subcellular localization database. *Nucleic Acids Res.,* **35**, D208–D212.

18. Thul,P. J. and Lindskog,C. (2018) The Human Protein Atlas: a spatial map of the human proteome. *Protein Sci.,* **27**, 233–244.

19. Peabody,M. A., Laird,M. R., Vlasschaert,C., Lo,R. and Brinkman,F. S. (2016) PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res.,* **44**, D663–D668.

20. Fazal,F. M., Han,S., Parker,K. R., Kaewsapsak,P., Xu,J., Boettiger,A. N., Chang,H. Y. and Ting,A. Y. (2019) Atlas of subcellular RNA localization revealed by APEX-seq. *Cell,* **178**, 473–490.

21. Jeske,L., Placzek,S., Schomburg,I., Chang,A. and Schomburg,D. (2019) BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.,* **47**, D542–D549.

22. Zhang,Z., Shen,T., Rui,B., Zhou,W., Zhou,X., Shang,C., Xin,C., Liu,X., Li,G., Jiang,J. et al. (2015) CeCaFDB: a curated database for the documentation, visualization and comparative analysis of central carbon metabolic flux distributions explored by 13C-fluxomics. *Nucleic Acids Res.,* **43**, D549–D557.

23. O'Boyle,N. M., Banck,M., James,C. A., Morley,C., Vandermeersch,T. and Hutchison,G. R. (2011) Open Babel: an open chemical toolbox. *J. Cheminform.,* **3**, 33.

24. Lang,P. F., Chebaro,Y., Zheng,X., Sekar,J. A., Shaikh,B., Natale,D. A. and Karr,J. R. (2020) BpForms and BcForms: a toolkit for concretely describing non-canonical polymers and complexes to facilitate global biochemical networks. *Genome Biol.,* **21**, 1–21.

25. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.,* **33**, 1635–1638.

26. Keseler,I. M., Mackie,A., Santos-Zavaleta,A., Billington,R., Bonavides-Martínez,C., Caspi,R., Fulcher,C., Gama-Castro,S., Kothari,A., Krummenacker,M., Latendresse,M., Muñiz Rascado,L., Ong,Q., Paley,S., Peralta-Gil,M., Subhraveti,P., Velázquez-Ramírez,D. A., Weaver,D., Collado-Vides,J., Paulsen,I. and Karp,P. D. (2017) The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res.,* **45**, D543–D550.

27. Zhu,B. and Stülke,J. (2018) SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis. *Nucleic Acids Res.,* **46**, D743–D748.

28. Sundararaj,S., Guo,A., Habibi-Nazhad,B., Rouani,M., Stothard,P., Ellison,M. and Wishart,D. S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. *Nucleic Acids Res.,* **32**, D293–D295.

29. Wodke,J. A., Alibés,A., Cozzuto,L., Hermoso,A., Yus,E., Lluch-Senar,M., Serrano,L. and Roma,G. (2015) MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae. *Nucleic Acids Res.,* **43**, D618–D623.

30. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. et al. (2020) The Reactome pathway knowledgebase. *Nucleic Acids Res.,* **48**, D498–D503.

31. Consortium,U. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.,* **47**, D506–D515.

32. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M. H., Baldock,R., Barbiera,G. et al. (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.,* **43**, W589–W598.

33. Perez-Riverol,Y., Bai,M., da Veiga Leprevost,F., Squizzato,S., Park,Y. M., Haug,K., Carroll,A. J., Spalding,D., Paschall,J., Wang,M. et al. (2017) Discovering and linking public omics data sets using the Omics Discovery Index. *Nat. Biotechnol.,* **35**, 406–409.