**Introduction**

WholeCellKB is a collection of free model organism databases designed specifically to enable comprehensive, dynamic simulations of entire cells and organisms. WholeCellKB provides comprehensive, quantitative descriptions of individual species including:

- Their subcellular organization,
- Their chromosome sequences,
- The essentiality, location, length, direction, and homologs of each gene,
- The organization and promoter of each transcription unit,
- The expression and degradation rate of each RNA gene product,
- The specific folding and maturation pathway of each RNA and protein species including the localization, N-terminal cleavage, signal sequence, prosthetic groups, disulfide bonds, and chaperone interactions of each protein species,
- The subunit composition of each macromolecular complex,
- Their genetic code,
- The binding sites and footprint of every DNA-binding protein,
- The structure, charge, and hydrophobicity of every metabolite,
- The stoichiometry, catalysis, coenzymes, energetics, and kinetics of every chemical reaction,
- The regulatory strength of each transcription factor on each promoter,
- Their chemical composition, and
- The composition of its typical SP-4 laboratory growth medium.

WholeCellKB currently contains a single database of *Mycoplasma genitalium*, an extremely small gram-positive bacterium and common human pathogen. This database is the most comprehensive description of any single organism to date, and was used to develop the first whole-cell computational model.

**Comparison to existing databases**

WholeCellKB has several major differences from previous databases such as BioCyc and BiGG. First, WholeCellKB was designed to represent every cellular process including well-studied processes such as metabolism as well as less well-understood processes such as DNA damage and repair and RNA and protein degradation.

Second, whole-cell modeling requires model organism databases which explicitly define the participants of each molecular interaction and chemical reaction. WholeCellKB addresses this need by representing the specific molecules involved in every molecular interaction and by requiring structures for each molecule. For example, WholeCellKB represents the specific RNA and protein species involved in every RNA and protein modification reaction and the specific proteins folded by each molecular chaperone. In comparison, most previous databases only explicitly represent metabolic and transcriptional regulatory interactions.

Third, where available WholeCellKB contains not only structural, but also quantitative functional descriptions of each molecule and molecular interaction. For example, WholeCellKB contains chemical reaction rate laws and kinetic parameters, RNA transcript expressions and half-lives, and cellular and growth medium chemical compositions.

**Curation**

The *M. genitalium* database was curated according to six principals:

- *Species specificity*: where possible we based the database on research conducted in *M. genitalium*. 

• **Comprehensiveness:** to develop a fully comprehensive database, where necessary we curated *M. genitalium* based on studies of close relatives such as *E. coli* and *Bacillus subtilis*. We determined homology by bi-direction best BLAST.

• **Accuracy:** where possible we catalogued multiple relevant research reports and based our annotation on the consensus of these reports.

• **Molecular specificity:** each piece of information was annotated to a specific molecule or set of molecules, each with a defined chemical structure.

• **Quantitativeness:** where possible we curated the numerical properties of each molecular and molecular interaction including for example, the thermodynamics and kinetics of each chemical reaction.

• **Transparency:** we catalogued the source of every piece of information as well as the researcher who curated it and when.

The *M. genitalium* database was curated from over 900 primary research articles, reviews, books, and databases over four years by a team of researchers at Stanford University.

First, we curated the overall structure of *M. genitalium* including its size, shape, subcellular organization, and chemical composition based on several experimental studies including Morowitz et al., 1962. We also curated the chemical composition of mycoplasma laboratory growth medium to the level of specific metabolite species based on analyses provided by Solabia.

Second, we curated the structure of the *M. genitalium* chromosome including its sequence, the location, length, and direction of each gene and its transcription unit organization based on the Comprehensive Microbial Resource Güell et al., 2009. We curated the location of each promoter and the expression, degradation rate, and essentiality of each gene product from studies by Weiner et al., 2000, Weiner et al., 2003, Bernstein et al., 2002, and Glass et al., 2006, respectively. We curated DNA-binding protein binding sites and transcriptional regulatory interactions from several sources including DBTBS.

Third, we curated the structure of each RNA and protein gene product. We curated the post-transcriptional processing and modification of each RNA transcript from several sources including Peil, 2009. We curated the signal sequence, localization, chaperone-mediated folding, post-translational modification, disulfide bonds, subunit composition, and DNA footprint of each protein and macromolecular complex from a large number of primary research articles, computational models, and databases. We curated the chemical regulation of each gene product from several sources including DrugBank. We used ExPASy ProtParam to calculate the pi, extinction coefficient, half-life, instability index, aliphatic index, and grand average of hydropathy of every protein species.

Fourth, we curated the specific chemical reactions catalyzed by each gene product starting from the CMR, GenBank, KEGG, and UniProt genome annotations and the RNA and protein maturation pathways we had already curated. To maximize the scope of the database and to fill gaps in the genome annotation we expanded each gene product's annotation based on primary research articles we identified by systematically searching PubMed and Google Scholar for each gene and its homologs. We consulted BioCyc, KEGG, two flux-balance analysis models of bacterial metabolism, hundreds of additional primary research articles to curate the stoichiometry of each chemical reaction. We curated the thermodynamics and kinetics of each chemical reaction from several databases including BRENDA, SABIO-RK, and UniProt.

Lastly, we curated the *M. genitalium* metabolome. We included any metabolite involved in any curated reaction or molecular interaction as well as any metabolite present in *M. genitalium* biomass or in typical mycoplasma growth medium. We curated the empirical formula, structure, and charge of each metabolite based on several databases including BioCyc and PubChem. We used ChemAxon Marvin to calculate the molecular weight, volume, pi, log_d and log_p of each metabolite.
The supplementary information to Karr et al., 2012 provides a detailed discussion of the curation process and sources.

**Versioning**

WholeCellKB separately versions each entry. Each entry's page displays two properties: *created* and *last updated* which indicate each entry's version history. Similarly, entries exported from WholeCellKB contain two properties *created_date* and *last_updated_date*.

**Implementation**

WholeCellKB was specifically designed to enable modelers to organize and structure all of the molecular data required for whole-cell dynamical models. WholeCellKB was designed with five major principals:

- **Rapid curation**: enable modelers to rapidly curate data by providing user-friendly interfaces for single as well as batch editing,
- **Collaboration**: enables modelers to collaboratively curate data through a user-friendly web interface,
- **Computability**: enable modelers to make data readily computable by making the data model easy to expand and edit,
- **Traceability**: enable modelers to annotate the primary source of each piece of data,
- **Transparency**: enable modelers to rapidly develop user-friendly interfaces similar to that of other model organism databases for themselves as well as other scientists to easily browse, search, and export the data, and
- **Customization**: enable modelers to rapidly develop custom data models and views.

As described above, the goal of WholeCellKB is to enable modelers to rapidly develop the databases needed to build whole-cell models. Consequently, we chose to implement WholeCellKB in Python using the Django web framework. This design enables modelers to rapidly develop custom data models and validation without any knowledge of relational database design or SQL and minimal programming. This design also allows WholeCellKB to provide modelers generic views for editing, batch editing, and exporting WholeCellKB entries, further enabling modelers to rapidly develop databases. Because WholeCellKB is built on top of Python, modelers can also easily perform scientific calculations on WholeCellKB entries using SciPy, NumPy, and Biopython for data validation and simulation. Furthermore, these theoretical calculations can easily be displayed alongside curated data in the WholeCellKB user interface.

WholeCellKB was implemented entirely with free, open-source software. WholeCellKB runs on Apache, an open-source web server using the open-source module *mod_wsgi*. WholeCellKB is stored using the open-source relational database *MySQL*. Full text search was implemented using the open-source libraries *Haystack* and *Xapian*. Batch Excel import and export were implemented using *OpenPyXL*. The RESTful JSON and XML interfaces were implemented using *simplejson* and *xml.dom*. PDF export was implemented using *xhtml2pdf*. Several bioinformatics calculations were implemented using *Biopython*.

Documentation for the source code is available [here](#). Additional documentation for the data model is available [below](#).

**Data model**

The WholeCellKB data model is composed of 18 types of entries:
Each type of entry is implemented by a distinct Python class by the same name derived from the `Entry` class. The `Entry` class provides five properties:

- `wid`: unicode identifier for entry. wid's are unique within a species.
- `name`: unicode representing the name of each entry.
- `comments`: unicode representing comments for each entry.
- `created_date`: datetime when the entry was created.
- `last_updated_date`: datetime when the entry was last updated.

Through the `CrossReference`, `Synonym`, and `User` classes, all entries also contain lists of cross-references to other databases, synonyms, and the identity of the users who created and last updated each entry. The Python classes for each entry type add additional properties.

In addition to the `Entry` subclasses which represent WholeCellKB entries, the WholeCellKB data model contains a second type of class derived from the `EntryData` superclass. These classes do not correspond to WholeCellKB entries, but rather are used to encode relationships among WholeCellKB entries.

The diagram below shows the WholeCellKB data model. Click the diagram of the link [here](link) for a high-resolution version. See also the [source code documentation](link).

**RESTful API**

WholeCellKB is also accessible via a RESTful API which mirrors the content of every HTML page in Bib, JSON, PDF, and XML formats. See [below](link) for further information about the available export formats. Please use the following URL patterns to access the API with the `format` query argument set to one of "bib", "json", "pdf", or "xml".

- Individual entries:
  - [https://www.wholecellkb.org/detail/SpeciesWID/EntryWID?format=Format](https://www.wholecellkb.org/detail/SpeciesWID/EntryWID?format=Format)
  - [https://www.wholecellkb.org/detail/Mgenitalium/ATP?format=json](https://www.wholecellkb.org/detail/Mgenitalium/ATP?format=json)
• Lists of entries:
  https://www.wholecellkb.org/list/SpeciesWID/EntryType?format=Format
  https://www.wholecellkb.org/list/Mgenitalium/Reaction?format=json

• Complete knowledge bases:
  https://www.wholecellkb.org/export/SpeciesWID?all_model_types=True&format=Format
  https://www.wholecellkb.org/export/Mgenitalium/?all_model_types=True&format=json

• Full-text searches:
  https://www.wholecellkb.org/search/SpeciesWID?q=QueryString&format=Format
  https://www.wholecellkb.org/search/Mgenitalium/?q=ATP&format=json

Allowed values for the URL fragment `SpeciesWID` include:

- Mgenitalium

The allowed values for the URL fragment `EntryWID` are marked at the top of each entry page with the label "WID". Below are several examples of valid `EntryWIDs`:

- Gene: MG_001
- Metabolite: ATP
- Reaction: AtpA
- Reference: PUB_0001
- Transcription unit: TU_001

Allowed values for the URL fragment `EntryType` include:

- Chromosome
- ChromosomeFeature
- Compartment
- Gene
- Metabolite
- Note
- Parameter
- Pathway
- Process
- ProteinComplex
- ProteinMonomer
- Reaction
- Reference
- State
- Stimulus
- TranscriptionUnit
- TranscriptionalRegulation
- Type

Finally, the URL fragment `QueryString` can be set to any string.

**Editing WholeCellKB**

We welcome additions and corrections to the version of WholeCellKB at https://www.wholecellkb.org. Click the icon located at the bottom-right corner of most pages to email us additions and corrections to WholeCellKB.
Alternatively, you can download the WholeCellKB source code and content to create and customize – including the content, data model, and user interface – your own model organism database. After you’ve installed and logged into your own version of WholeCellKB, the icons will direct you to web-forms which will enable you to edit your model organism databases. Additionally, after you’ve logged in you will have access to a batch upload page which will enable you to use the WholeCellKB Excel interface to more rapidly edit your model organism databases.

See below for more information on creating your own model organism databases with WholeCellKB.

**Downloading WholeCellKB content**
The content of WholeCellKB can be downloaded [here](#).

The content of WholeCellKB is available in six formats:

- **BibTex**
  This format provides all of the references cited by the selected entries in BibTex format. See [Wikipedia](#) for more information about the BibTex citation format.

- **Excel 2007**
  This format provides the selected entries in a tabular format, with separate tables on separate worksheets for each entry type (eg. Metabolite, Reaction, etc). Entries correspond to rows in the exported tables. Properties correspond to columns. Properties which represent many-to-many relationships among entries are presented as comma-separated lists and JSON-formatted strings.

- **HyperText Markup Language (HTML)**
  This format returns an HTML-formatted table containing the WID and name of the selected entries.

- **JavaScript Object Notation (JSON)**
  This format returns a JSON-encoded JavaScript object with four properties:
  
  - **title**: String which describes exported species.
  - **comments**: String which describes how and when the data was exported.
  - **copyright**: String which describes how and when the data was curated.
  - **data**: Array which contains the selected entries. Each element of the array is a JSON-serialized version of a single WholeCellKB entry, and contains several properties including `wid`, `name`, `created_date`, and `last_updated_date`. See the [data model section](#) above for more information about the properties of each entry type.
  
  See [Wikipedia](#) for more information about the JSON data format.

- **Portable Document Format (PDF)**
  This format returns a PDF-formatted table containing the WID and name of the selected entries.

- **Extensible Markup Language (XML)**
  This format returns an XML object with one property, `objects`, which contains a list of the selected entries, each represented by an XML object of type `object` containing XML objects of type `field` which represent their attributes including the WID, name, creation date, and last updated date. See the [data model section](#) above for more information about the properties of each entry type. See [Wikipedia](#) for more information about the XML data format.

Entries exported in BibTex, Excel, JSON, or XML format contains two properties `created_date` and `last_updated_date` which indicate each entry's revision history. See [above](#) for further information.

**Downloading WholeCellKB source code**
The WholeCellKB source code is freely available at SimTK. Documentation for the source code is available [here](#).

Additional documentation for the data model is available [above](#).
Want to develop your own model organism database?
The WholeCellKB software was designed to enable modelers to quickly develop the databases needed for whole-cell models. In particular, the WholeCellKB software enables modelers to:

- Collaboratively create and edit model organism databases via web-based and Excel interfaces
- Easily customize its data model without any knowledge of SQL and only minimal programming
- Easily customize its web-based user interface including the layout and information displayed on entry page

To create and customize your own model organism database download and install the WholeCellKB source code at SimTK. See the Developer's Guide for detailed instructions on installing the WholeCellKB software, creating new model organism databases, and customizing the data model and user interface. See also the source code and data model documentation.

Mathematical modeling
WholeCell was specifically designed to enable comprehensive, dynamic whole-cell simulations. Please see Karr et al., 2012 Data S1 for information about how WholeCellKB can be used to develop whole-cell simulations.

Getting started
The best ways to get started are to browse or search WholeCellKB using the menu or the search box at the top of this page. See the tutorial for additional help getting started.

Citing WholeCellKB
Please use the following references to cite WholeCellKB:


Development Team
WholeCellKB was developed by a team of researchers at Stanford University:

- Jonathan Karr, Graduate Student in Biophysics
- Jayodita Sanghvi, Graduate Student in Bioengineering
- Derek Macklin, Graduate Student in Bioengineering
- Abhishek Arora, Graduate Student in Electrical Engineering
- Markus Covert, Assistant Professor of Bioengineering

Questions & comments
Please use the icon at the bottom right of each data page to suggest edits to the content of WholeCellKB.

Please contact us at wholecell@lists.stanford.edu with any questions and/or comments about WholeCellKB.