ReadME for the AlloRep Database

Filipa L. Sousa, Daniel J. Parente, David L. Shis, Jacob A. Hessman, Allen Chazelle, Matthew R. Bennett, Sarah A. Teichmann, and Liskin Swint-Kruse

AlloRep is divided into five modules. Explanations of relevant tables and abbreviation used in each section are below, as well as example queries that can be used to link the information between the various modules. Alternatively, tables can be browsed and sorted by clicking on various column headings.

Module 1: Mutagenesis data

This module contains information collected from an exhaustive literature search; citations are contained in module 5. Module 1 entails two tables: "mut1_single" and "mut2_combinatorial". For variants in "mut1_single", all outcomes can be attributed to a single mutation, either by comparing the properties of a single mutation to those of the wild-type protein, or, for example, by comparing a double mutant to a variant that contains the relevant single mutation. Variants in the "mut2_combinatorial" table contain multiple mutations that have not yet been parsed into their component contributions.

Both tables contain: a unique internal_id for each variant, subfamily classification, species of origin, position number in the parent protein, LacI numbering, one-letter codes for the original amino acid and the mutational variant, and PMIDs of the original publications. The mut1_single also contains the parent protein that provides the basis for comparison of experimental results.

In both tables, additional columns contain all available experimental information for the variant. Information regarding the effect on protein secondary structure and/or oligomerization state (where "D" stands for dimer, "T" for tetramer and "M" for monomer) are stored in columns with those names. Effects on urea stability, thermal denaturation, trypsin digestion assays, and temperature sensitivity are stored in other columns. The phenotypic and biochemical characterizations are provided in the "phenotype", "allostery" and "reverse phenotype" columns. When possible, the relative differences are indicated with the symbols: [0] or [---] for total loss, [- -] for a significant decrease, [-] small decrease, [=] or ~ if comparable with wild type, [+] for small increase and [+ +] for a significant increase. Any additional information is provided in the "observation" column.

Module 2: Sequence Data

This module contains three tables with: (1) the manually-curated alignment of

representative sequences for the entire LacI/GalR family (each homolog is contained in a separate row); (2) the separate alignments for all subfamilies (each subfamily alignment is contained in one row); and (3) a table containing unaligned "orphan" sequences (one per row) that do not match any of the current subfamilies. All data are stored in fasta format. After selecting a table of interest, it can be downloaded using the export button at the bottom of the page and selecting the desired format. Note that the output options can be customized for a better compatibility with the user's operating system. The subfamily alignments can be matched to the spacing of the manually-curated, whole-family alignment using the program MARS_PROT (https://github.com/djparente/MARS).

Module 3: Structural Data

All available structures for LacI/GalR homologs were retrieved from the Protein Data Bank database; citations are in Module 5. Module 3 contains all the information regarding the PDB description (struct1_pdb_overview table), available ligand information (struct2_ligand_description table), and four tables with different types of contacts.

For each LacI/GalR structure, non-covalent contacts were defined when any two residues had at least one non-hydrogen atom within 5 Å of the other. For all structures, the full set of contacts is stored in the table "struct3_contacts_monomers" where contacts were grouped according to their protein subfamily, inter or intramonomeric nature and ligand. . Next, for the table "struct4_contacts_heatmap", equivalent structures (those for the same protein and liganded state) were combined into one column and used to calculate the frequency of each contact pair .. For example, apo LacI has two structures (1lbi and 3edc) each of which contains four monomers. In two of the 8 chains (25%), LacI residues E100 and C107 are within 5Å of each other; thus the occupancy score for this contact is 25%. The table "struct5_contacts_macromol" contains information regarding the contacts between the LacI/GalR proteins and macromolecular ligands such as DNA or heteroproteins. Contacts between LacI/GalR proteins and small-molecule ligands are stored in the table "struct6_contacts_ligand table", which also includes information on the total contact surface area and the number of contacts.

Module 4: Translation Tables

This section contains two tables – "translate_numbering_table" and "translate_numbers_to_laci" – that allow the conversion between the numbering system of *E.coli* LacI and those of other LacI/GalR homologs. "Translate_numbers_to_laci" contains the necessary information for connecting both structural or mutagenesis data to the "translate_numbering_table". The "translate_numbering_table" contains the structural alignment of all crystallographic structures as well as representative sequences for each protein subfamily that has available mutagenesis data.

Using either the PDB identifier and residue numbering as input (from tables in the structural module) or information regarding the LacI/GalR subfamily and residue numbering as input (from tables in the mutation module), the user can obtain the code to be used in the translation_numbers_to_laci and retrieve the original sequence numbering.

Module 5: Citations

A final table (x_data_sources_cited) contains all bibliographic information and can be queried using the PMID or the citation code provided in the structural and mutagenesis tables.

Example Queries: These queries are also available to download as separate * .sql files. Note that the semi-colon at the end of each query is required syntax.

Translation queries: AlloRep can be used to translate information between the different modules of the database, using the following queries.

Query 1: Translate a given LacI position to analogous positions in homologs that represent other LacI/GalR subfamilies. The example given is for LacI position 35. This query also allows the user to analyze the amino acid composition conservation of particular positions using the whole-family sequence alignment comprising representative sequences from each subfamily.

SELECT * FROM translate_numbering_table WHERE (lacI_num LIKE '**35**');

Query 2: Retrieve the original numbering for position pairs of LacI/GalR homologs using the table "struct3_contacts_monomers". The output will show the translation of the selected positions (in sequence numbering) as well as the amino acid composition of the selected LacI/GalR protein subfamily. Note that in the generic query 2 provided separately on the AlloRep website as a downloadable .sql file, the fields in bold are left empty. These should to be completed with the information regarding the residue pairs from the table "struct3_contacts_monomers".

SELECT * FROM translate_numbering_table WHERE (lacI_num LIKE **88** OR lacI_num LIKE **106**) and subfam like '**TreR**';

Query 3: Retrieve the original numbering from LacI/GalR proteins from the table "struct5_contacts_macromol" or "struct6_contacts_ligand".

SELECT * FROM translate_numbering_table WHERE (lacI_num LIKE **188**) and subfam like '**PurR**';

Other example queries:

Query 4: Conservation of intra- and inter-monomeric. non-covalent contacts among the LacI/GalR subfamilies, grouped by residue pair and type of contact.

```
SELECT Contact_type, Position1, Position2, count(*) FROM struct3_contacts_monomers
GROUP BY Contact_type, Position1, Position2
ORDER BY count(*) DESC;
```

Query 5: Retrieve all non-covalent residue-residue contacts for a given position. This query will allow the user to retrieve all non-covalent contacts for analogous amino acid positions, regardless of subfamily or primary sequence. This query also allows the user to retrieve all residue-residue contacts for a position selected from the mutagenesis data ("mut1_single"). Additional filters to narrow outputs are created by replacing * with the subfamily, ligand and/or contact_type of interest.

```
SELECT * FROM struct3_contacts_monomers WHERE Position1 LIKE '35' or Position2 LIKE '35';
```

Query 6- Search for all variants with a specific phenotype (in the example, those with abolished activity).

```
SELECT * FROM mut1_single
WHERE phenotype LIKE '%0%' or phenotype LIKE '%----%';
```

Query 7- Retrieve all non-covalent residue-residue contacts involving positions with the same phenotype.

```
SELECT * FROM struct3_contacts_monomers
WHERE Position1 IN (SELECT LacI_numbering from mut1_single WHERE phenotype LIKE '%---%' or phenotype LIKE '%0%');
```