HHS Public Access

Author manuscript

J Mol Biol. Author manuscript; available in PMC 2017 February 22.

Published in final edited form as:

J Mol Biol. 2016 February 22; 428(4): 671–678. doi:10.1016/j.jmb.2015.09.015.

AlloRep: a repository of sequence, structural and mutagenesis data for the LacI/GalR transcription regulators

Filipa L. Sousa[a], **Daniel J. Parente**[b,1], **David L. Shis**[c], **Jacob A. Hessman**[b,1], **Allen Chazelle**[b], **Matthew R. Bennett**[c,d], **Sarah A. Teichmann**[e,f], and **Liskin Swint-Kruse**[b,*]

[a]Institute of Molecular Evolution, Heinrich-Heine Universität Düsseldorf, Universitätstrasse 1, 40225 Düsseldorf, Germany

[b]The Department of Biochemistry and Molecular Biology, The University of Kansas Medical Center, Kansas City, KS 66160, USA

[c]The Department of Biosciences, Rice University, Houston TX, 77005, USA

[d]The Department of Bioengineering, Rice University, Houston TX, 77005, USA

[e]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[f]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

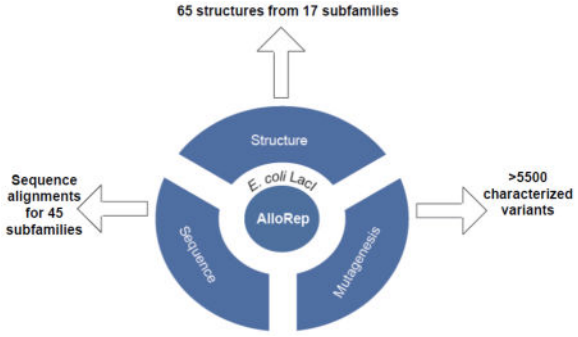
Abstract

Protein families evolve functional variation by accumulating point-mutations at functionally-important amino acid positions. Homologs in the LacI/GalR family of transcription regulators have evolved to bind diverse DNA sequences and allosteric regulatory molecules. In addition to playing key roles in bacterial metabolism, these proteins have been widely used as a model family for benchmarking structural and functional prediction algorithms. We have collected manually-curated sequence alignments for >3000 sequences, *in vivo* phenotypic and biochemical data for >5750 LacI/GalR mutational variants, and non-covalent residue contact networks for 65 LacI/GalR homolog structures. Using this rich data resource, we compared the non-covalent residue contact networks of the LacI/GalR subfamilies to design and experimentally validate an allosteric mutant of a synthetic LacI/GalR repressor for use in biotechnology. The AlloRep database (freely available at www.AlloRep.org) is a key resource for future evolutionary studies of LacI/GalR homologs and for benchmarking computational predictions of functional change.

Graphical Abstract

*To whom correspondence should be addressed: lswint-kruse@kumc.edu, 913-588-0399.

[1]Present address: The University of Kansas School of Medicine



Keywords

transcription regulation; variant database; sequence alignment; contact map; allostery

Introduction

Sequence- and structure-based comparisons of protein homologs have been frequently used to predict amino acids critical to function. With advances in high-throughput sequence and structure determination, the amount of data available has exploded. To translate these data into meaningful information, myriad computational tools have been developed (i) to detect patterns of amino acid change, and (ii) to make predictions about homolog function and mutational outcomes. Development and validation of these programs requires experimental datasets against which to test predictions.

One commonly-used (*e.g.* [1–12]) dataset comprises *in vivo* characterization of ~4100 mutational variants of the lactose repressor protein (LacI) [13–16]. In addition, scores of mutational variants for LacI and its paralogs have been the subject of detailed biochemical and expanded phenotypic studies over the last three decades. However, these additional experimental results have been under-utilized by the computational community, due to the challenge of curating the relevant information scattered throughout the literature. Nevertheless, these studies provide in-depth insights that would be extremely valuable for assessing computational predictions. Further, structures for numerous LacI/GalR homologs have become available, mainly through the Protein Structure Initiative [17,18].

Here, we present AlloRep, a repository of published experimental information for homologs of the LacI/GalR family. AlloRep contains (i) manually-curated sequence alignments for >3100 sequences, (ii) experimental results for >5750 LacI/GalR mutational variants, and (iii) residue-residue contact networks that were derived from 65 crystallographic structures available for full-length homologs and/or their regulatory domains of 17 LacI/GalR subfamilies (Fig 1). This database is freely available at www.AlloRep.org and can be queried using MySQL. Information contained in the AlloRep database complements information about predicted regulons for >1300 LacI/GalR homologs, which was recently added to the RegPrecise database [19].

The data in AlloRep also have important applications in protein design: These data can be used to test robustness of protein engineering approaches and to hypothesize novel ideas for

engineering synthetic transcription repressors. As proof-of-principle, we used AlloRep to merge structural, mutational, and sequence data to identify a position that can be mutated to alter allosteric regulation. Most LacI/GalR homologs are allosterically regulated: The DNA-binding domains of the apo-proteins bind to their cognate DNA sequences with high affinity, and DNA binding is modulated when a distant site on the regulatory domain is occupied by a small molecule effector (or in some cases, a heteroprotein). The LacI/GalR paralogs have evolved specificities for different DNA sequences and allosteric effectors [20]. Although domain recombination shows that the allosteric mechanism may largely be the same, the magnitude and direction of allosteric response can be modulated [20,21]. In general, predicting the locations of allosteric positions has been challenging. Our prediction was successfully tested in a synthetic, chimeric repressor that was previously constructed from LacI and the cellobiose repressor (CelR).

Results and Discussion

Overview of the AlloRep database

The AlloRep database comprises 14 tables and can be queried using MySQL. Example queries as well as a database scheme are supplied in the accompanying “Data in Brief” publication [22]. A key advantage of AlloRep is that all entries have been mapped to the analogous position of a single homolog – the full-length *E. coli* LacI protein. This is a powerful way to compare different homologs, as well as different structural conformations of the same protein. This mapping allows a single query to extract information about a given position from all available homologs (Fig 1). In addition, a translation table (“translate_numbering_table”) is included to easily convert the LacI position number to those of the other homologs. For example, a user can identify residues involved in interdomain contacts, analyze their respective conservation across families, and integrate the available mutagenesis information for that particular position (Queries 1 to 3 in [22]). Both user-queried information and the entire database can be exported in several formats for offline use.

Sequence information

The AlloRep database contains the manually-curated sequence alignments for 45 LacI/GalR subfamilies (Fig 1 in [22]) as well as >95 sequences that appear to be single representatives of otherwise unclassified protein subfamilies. Separate tables contain whole-family and subfamily alignments, so that the user can search for conservation patterns at these distinct phylogenetic levels [23].

More specifically, a central table (“seq1_family_ref_align”) contains the manually-curated alignment for representative sequences from each of the LacI/GalR subfamilies. In addition to providing information about family-wide properties, this table can be used to convert the numbering of each subfamily alignment into that of the family alignment, using the program MARS-Prot (https://github.com/djparente/MARS) [24]. The alignments for each subfamily are stored as rows in the table “seq2_subfam_alignments” (Figure 1 in [22]). The first column contains the subfamily name; the second column contains all subfamily homolog sequences.

For 34 subfamilies, the sequence alignment construction was described in [23,25]; where possible, these sequence alignments were benchmarked against structure-based sequence alignments. The AlloRep database also includes additional sequences for (a) the AscG, CytR, FruR, and LacI subfamilies and (b) 11 new LacI/GalR subfamilies that were nucleated from 11 unpublished structures deposited in the Protein Data Bank by the Protein Structure Initiative (PDB: 3bil, 3cs3, 3d8u, 3e3m, 3gv0, 3h5t, 3hs3, 3jvd, 3jy6, 3k4h and 3kjx) [17,18]. Finally, the LacI/GalR sequences that do not yet fall into subfamilies are stored in the "seq3_unaligned_orphan_seqs" table. This table also contains information regarding the presence or absence of the "YPAL" motif that appears to play an important role in inter-domain allosteric communication.

Mutagenesis data

A collection of mutagenesis data for LacI/GalR homologs was constructed from a comprehensive literature search. Data for >5750 single mutants with functional and/or structural information are stored in the table "mut1_single". For these variants, all mutational 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. Information about the single amino acid change and the comparative protein are listed in separate columns of this table. In addition to the position number in the parent homolog, all mutation entries have been translated to the analogous *E.coli* LacI position number. This allows the user to compare the mutational outcomes of a selected position across many subfamilies, by using the LacI numbering in the translation table "translate_numbering_table" (Query 1 in [22]).

The data for the single mutant table include ~4100 mutations from the Miller lab *in vivo* studies of LacI [13–16]. Another ~1100 variants are from *in vivo* and biochemical studies of LacI/GalR chimeras [21,26–29]. The remaining variants are from biochemical studies of LacI, PurR, GalR, CytR, FruR and CcpA. Citations specific to each variant are included as PubMed ID numbers (PMID) in the table "mut1_single"; the full citation is listed in the table "x_data_sources_cited". When available, information about each mutation's effect on secondary structure, oligomerization state, stability to urea denaturation, trypsin digestion, DNA binding, allosteric ligand binding, sensitivity to pH changes, and allosteric phenotype was included. A more complete description of table and column content are presented in [22].

AlloRep includes a separate table ("mut2_combinatorial") containing variants with multiple mutations that have not yet been parsed into their component contributions. For example, a protein with three mutations was placed in the "mut2_combinatorial" table if its function could only be compared to that of its wild-type parent protein. In contrast, as noted above, other proteins with triple mutations were placed in the "mut1_single" table if their functions could be compared to those of a relevant double mutant variant. The combinatorial table also contains information regarding the original and the LacI translated numbering, as well as the PMID for all original studies.

Structural data and comparisons *via* residue contact networks

The three-dimensional structure of a protein or protein-protein complex can be represented as a residue-residue contact network [30–34]. In this network, amino acid positions are represented as nodes and their non-covalent interactions with other positions comprise the network edges (Fig 1); the latter are delimited by distance thresholds (here, 5 Å between the two closest atoms in the PDB structure). This representation retains information about individual atoms, reduces the visual complexity, links protein structural interpretation with graph theory [35], and allows a parallel analysis of multiple structures [34,36]. We previously used this approach for comparing protein-protein interfaces [32,33,36,37].

Here we expand this comparative approach to the more complex networks required to represent all intra- and inter-molecular noncovalent interactions. This approach was successfully used in GPCRs to identify positions key to their conserved allosteric mechanism for Gα activation [38]. A second example comes from the PyrR family, for which this approach was used to analyze amino acid positons far from the dimer interface that altered oligomerization when mutated; results showed that these variants effectively "re-wired" the contact networks, similar to rearrangements observed for various allosteric states of these proteins [39].

AlloRep includes the contact maps calculated from all structures that are available for full-length and/or the regulatory domains of the LacI/GalR homologs (see Figure 2 in [22] for an illustration of the calculation). Individual structures are listed by their PDB identifier and PMID in the table "struct1_pdb_overview"; definitions and abbreviations for the various ligands are described in the table "struct2_ligand_description"; and full citations are listed in the table "x_data_sources_cited". These LacI/GalR structures were used to construct three groups of noncovalent contacts: Those between the monomers of LacI/GalR homodimers (stored in the "struct3_contacts_monomers" table); those between the LacI/GalR proteins and macromolecular ligands – operator DNA and, for the CcpA homolog the heteroproteins hpr and crh ("struct5_contacts_macromol" table); and those between protein and small allosteric ligands that act as inducers, co-repressors, anti-inducers, and neutral effectors ("struct6_contacts_ligand" table).

As with the mutation tables described above, the database is constructed so that interactions can be easily compared among homologs and/or different conformations. A global view of the contacts conservation across subfamilies is provided in the table "struct4_contacts_heatmap", which is grouped by contact type (intra- or inter-monomeric), subfamily, and ligand (see Figure 2 in [22]). In this table, the contact maps for equivalent structures (those for the same protein and liganded state) were (i) combined to create one column and (ii) used to determine the contact frequency of a particular amino acid for that state. 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%. This normalization accounts for natural fluctuations within a particular liganded state, and for other effects (*e.g.* packing effects) that might not have any functional/biological role. In the cases where a group is represented by a single monomeric chain (*e.g.* the structures of orphan LacI/GalR

homologs), occupancy is calculated as 100% by default. With the determination of new crystal structures, the averages are expected to change.

An example structural comparison is shown in Fig 2A–C for LacI and PurR. This pair of proteins is interesting to compare because high affinity DNA binding occurs *via* opposite "open" and "closed" conformations of the regulatory domains [36]. Most pairs of interacting residues are common to all conformations of both proteins (Fig 2A). In addition, one group of contacts is conserved at the inter-monomeric interface for the "open" structures of the two proteins (LacI with DNA and apo-PurR; Fig 2B), and a second group in the ligand binding region are conserved between the "closed" structures (LacI without DNA and PurR with DNA; Fig 2C). Previously, we identified (i) a different subset of interface contacts in common to the DNA-free forms of both LacI (closed) and PurR (open) and (ii) one subset of interface contacts unique to the closed structures of LacI and PurR [36]. Thus, it appears that features of a common geometric arrangement were differentially co-opted to evolve varied allosteric responses.

An experimental case study

In addition to benchmarking computational tools, another use of AlloRep is to facilitate hypotheses about functionally important amino acid positions in the LacI/GalR homologs. LacI is widely used in biotechnology, and synthetic LacI/GalR homologs are being developed to carry out novel, complex, biosynthetic circuitry [40].

To that end, AlloRep simplifies the amalgamation of non-covalent structural information with mutagenesis and sequence conservation data. For example, comparative structural analyses highlighted the importance of the position analogous to LacI R118, which is located on the regulatory domain. In the structural contact network, this position is a hub position involved in multiple inter- and intra-monomeric interactions with: DNA, the inter-domain linker sequence, and the regulatory-domain (Fig 3A). Highly connected nodes within residue-residue contact networks are frequently found to be functionally relevant [41,42]. A search of the table "mut1_single" showed that, in *E. coli* LacI, 12 mutations of position 118 had impaired or abolished DNA binding [13,15]. Sequence alignments showed that position 118 is conserved as arginine in both the LacI and PurR subfamilies [25]. However, in other subfamilies, position 118 is conserved as different residues. For example, in the cellobiose repressor of *Thermomonospora fusca* (CelR; [43]), position 118 is occupied by a histidine. Together, these data suggest that position 118 can be mutated to evolve new variations on the common LacI/GalR function.

Since position 118 is between the DNA and inducer binding sites, its central position in the network suggested that mutations would alter allosteric regulation. To verify this hypothesis, we used a chimeric protein comprising the DNA binding domain and linker of LacI and the ligand binding domain of CelR; this chimera also contained three mutations required for measureable repression ("LLhE-3mut"; see Methods for more details). This synthetic repressor is useful for testing our hypothesis because the starting chimera has both modest repression and allosteric response [21]: This should allow us to detect either positive or negative functional changes that arise from mutating position 118. Our specific hypothesis was that mutating H118 to arginine on the CelR regulatory domain would alter allosteric

communication with the LacI DNA-binding domain. Indeed, results showed that the LLhE-3mut/H118R mutation required >10-fold more inducer for allosteric response (Fig 3B). This indicates the role of position 118 in evolving varied allosteric communication and illustrates the utility of the aggregate data contained in the AlloRep database.

Conclusion

The AlloRep database (www.AlloRep.org) organizes available sequence, structural, and experimental data for the LacI/GalR protein family. This dataset will be useful for the development and validation of computational analyses of protein families. We are committed to the continued integration of mutagenesis, structural and sequence information as they become available for LacI/GalR homologs. We invite the scientific community to send their mutagenesis data to AlloRep, so that this experimental resource remains up-to-date.

Methods

Sequence retrieval and alignments

The sequence identity boundaries of the new LacI/GalR subfamilies were defined as described in [25]. For the subfamilies represented by the new PSI pdb structures, a structure-based reference alignment was constructed with PROMALS3D [44] and integrated into the whole family alignment with the program MARS-Prot (https://github.com/djparente/MARS) [24]. For all new homologs, subfamily alignments were constructed using MUSCLE [45] and representative sequences were integrated into the whole family alignment with MARS-Prot.

Contact maps

Crystallographic structures belonging to the LacI/GalR family were downloaded from the PDB database [46] and manually classified according to their different allosteric states and bound ligands. When only one monomer was present in the crystallographic data, coordinates for the homodimer were generated using the symmetry operation in Pymol if possible [47]. If multiple dimers were present in the unit cell, they were treated as separate structures. In total, 65 structures were retrieved and used for analyses. Intra- and intermolecular pairs of interacting residues were determined with the Ncont program available in the CCP4 program suite version 6.2 [48] using the distance of 5Å between any two non-hydrogen atoms as threshold. Angles and other geometries were not considered. Intramolecular contacts were restricted to residue pairs at least 5 amino acids apart in the polypeptide sequence: closer contacts are expected to occur in all polypeptide chains and thus are not very informative. To directly compare the different structures, a structural alignment was performed and each contact pair was translated according to the *E. coli* LacI sequence numbering. All contact maps relate to LacI numbering and the original structure numbering can be traced back by using the “translate_numbers_to_laci” and “translate_numbering_table” tables. Example queries are shown in the accompanying “Data in Brief” publication [22].

Experiments with the LacI:CelR chimera

Construction of the LacI:CelR chimera was reported in [21]. Here, we used the "LLhE-3mut" variant, which contains three mutations in the linker sequence (I48V, Q55A, and Q60R) that are necessary to convey detectable repression of the *lac* operon. LLhE-3mut also has a modest allosteric response upon binding cellobiose [21]. For this work, the H118R variant was created by overlapping primer mutagenesis [49]. Primers used were 5′-CGGACAGCGCGTCGACGGGGTCCTCCTGC and 5′-CGTCGACGCGCTGTCCGGCCAGGTAGCCG. Repression and induction of the LLhE-3mut variants were assayed as described in Shis *et al.* [40].

Acknowledgments

This work was supported by Fundação para a Ciência e Tecnologia, SFRH/BPD/73058/2010 (FLS), NIH GM 079423 (LSK), the University of Kansas Medical Center Biomedical Research Training Program (DJP), the joint NSF/NIGMS Mathematical Biology Program R01GM104974 (MRB), and the Robert A. Welch Foundation C-1729 (MRB) and private funds. We thank Tina Perica for many stimulating discussions about this project.

References

1. Pei J, Cai W, Kinch LN, Grishin NV. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. Bioinformatics. 2006; 22:164–71. [PubMed: 16278237]
2. Bharatham K, Zhang ZH, Mihalek I. Determinants, discriminants, conserved residues--a heuristic approach to detection of functional divergence in protein families. PLoS ONE. 2011; 6:e24382. [PubMed: 21931701]
3. Mazin PV, Gelfand MS, Mironov AA, Rakhmaninova AB, Rubinov AR, Russell RB, et al. An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. Algorithms Mol Biol. 2010; 5:29. [PubMed: 20633297]
4. Marini NJ, Thomas PD, Rine J. The use of orthologous sequences to predict the impact of amino acid substitutions on protein function. PLoS Genet. 2010; 6:e1000968. [PubMed: 20523748]
5. Lee W, Zhang Y, Mukhyala K, Lazarus RA, Zhang Z. Bi-directional SIFT predicts a subset of activating mutations. PLoS ONE. 2009; 4:e8311. [PubMed: 20011534]
6. Ye K, Vriend G, IJzerman AP. Tracing evolutionary pressure. Bioinformatics. 2008; 24:908–15. [PubMed: 18304936]
7. Cooper GM, Brown CD. Qualifying the relationship between sequence conservation and molecular function. Genome Res. 2008; 18:201–5. [PubMed: 18245453]
8. Ye K, Feenstra KA, Heringa J, IJzerman AP, Marchiori E. Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. Bioinformatics. 2008; 24:18–25. [PubMed: 18024975]
9. Needham CJ, Bradford JR, Bulpitt AJ, Care MA, Westhead DR. Predicting the effect of missense mutations on protein function: analysis with Bayesian networks. BMC Bioinformatics. 2006; 7:405. [PubMed: 16956412]
10. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005; 15:978–86. [PubMed: 15965030]
11. Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics. 2003; 19:2199–209. [PubMed: 14630648]
12. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001; 11:863–74. [PubMed: 11337480]
13. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential

residues, as well as “spacers” which do not require a specific sequence. J Mol Biol. 1994; 240:421–33. [PubMed: 8046748]

14. Miller JH. Genetic studies of the lac repressor. XII. Amino acid replacements in the DNA binding domain of the *Escherichia coli* lac repressor. J Mol Biol. 1984; 180:205–12. [PubMed: 6392567]
15. Kleina LG, Miller JH. Genetic studies of the lac repressor. XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. J Mol Biol. 1990; 212:295–318. [PubMed: 2157024]
16. Suckow J, Markiewicz P, Kleina LG, Miller J, Kisters-Woike B, Müller-Hill B. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J Mol Biol. 1996; 261:509–23. [PubMed: 8794873]
17. Lee D, de Beer TAP, Laskowski RA, Thornton JM, Orengo CA. 1,000 structures and more from the MCSG. BMC Struct Biol. 2011; 11:2. [PubMed: 21219649]
18. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci U S A. 2014; 111:3733–8. [PubMed: 24567391]
19. Ravcheev DA, Khoroshkin MS, Laikova ON, Tsoy OV, Sernova NV, Petrova SA, et al. Comparative genomics and evolution of regulons of the LacI-family transcription factors. Front Microbiol. 2014; 5:294. [PubMed: 24966856]
20. Swint-Kruse L, Matthews KS. Allostery in the LacI/GalR family: variations on a theme. Curr Opin Microbiol. 2009; 12:129–37. [PubMed: 19269243]
21. Meinhardt S, Manley MWJ, Becker NA, Hessman JA, Maher LJ III, Swint-Kruse L. Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression. Nucleic Acids Res. 2012; 40:11139–54. [PubMed: 22965134]
22. Sousa FL, Parente DJ, Hessman JA, Teichmann SA, Swint-Kruse L. Publications, structural analyses, and queries used to build and utilize the AlloRep database. n.d Data in Brief. submitted.
23. Parente DJ, Swint-Kruse L. Multiple co-evolutionary networks are supported by the common tertiary scaffold of the LacI/GalR proteins. PLoS ONE. 2013; 8:e84398. [PubMed: 24391951]
24. Parente DJ, Ray JCJ, Swint-Kruse L. Amino acids positions subject to multiple co-evolutionary constraints can be robustly identified by their eigenvector network centrality scores. Proteins. n.d in press.
25. Tungtur S, Parente DJ, Swint-Kruse L. Functionally important positions can comprise the majority of a protein’s architecture. Proteins. 2011; 79:1589–608. [PubMed: 21374721]
26. Meinhardt S, Swint-Kruse L. Experimental identification of specificity determinants in the domain linker of a LacI/GalR protein: bioinformatics-based predictions generate true positives and false negatives. Proteins. 2008; 73:941–57. [PubMed: 18536016]
27. Tungtur S, Meinhardt S, Swint-Kruse L. Comparing the functional roles of nonconserved sequence positions in homologous transcription repressors: implications for sequence/function analyses. J Mol Biol. 2010; 395:785–802. [PubMed: 19818797]
28. Tungtur S, Egan SM, Swint-Kruse L. Functional consequences of exchanging domains between LacI and PurR are mediated by the intervening linker sequence. Proteins. 2007; 68:375–88. [PubMed: 17436321]
29. Meinhardt S, Manley MW, Parente DJ, Swint-Kruse L. Rheostats and toggle switches for modulating protein function. PLoS ONE. 2013; 8:e83502. [PubMed: 24386217]
30. Soundararajan V, Raman R, Raguram S, Sasisekharan V, Sasisekharan R. Atomic interaction networks in the core of protein domains and their native folds. PLoS ONE. 2010; 5:e9391. [PubMed: 20186337]
31. Zhang X, Perica T, Teichmann SA. Evolution of protein structures and interactions from the perspective of residue contact networks. Curr Opin Struct Biol. 2013; 23:954–63. [PubMed: 23890840]
32. Swint-Kruse L. Using networks to identify fine structural differences between functionally distinct protein states. Biochemistry. 2004; 43:10886–95. [PubMed: 15323549]

33. Swint-Kruse L, Brown CS. Resmap: automated representation of macromolecular interfaces as two-dimensional networks. Bioinformatics. 2005; 21:3327–8. [PubMed: 15914544]
34. Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM. Molecular signatures of G-protein-coupled receptors. Nature. 2013; 494:185–94. [PubMed: 23407534]
35. Krishnan A, Zbilut JP, Tomita M, Giuliani A. Proteins as networks: usefulness of graph theory in protein science. Curr Protein Pept Sci. 2008; 9:28–38. [PubMed: 18336321]
36. Swint-Kruse L, Elam CR, Lin JW, Wycuff DR, Shive Matthews K. Plasticity of quaternary structure: twenty-two ways to form a LacI dimer. Protein Sci. 2001; 10:262–76. [PubMed: 11266612]
37. Swint-Kruse L, Larson C, Pettitt BM, Matthews KS. Fine-tuning function: correlation of hinge domain interactions with functional distinctions between LacI and PurR. Protein Sci. 2002; 11:778–94. [PubMed: 11910022]
38. Flock T, Ravarani CNJ, Sun D, Venkatakrishnan AJ, Kayikci M, Tate CG, et al. Universal allosteric mechanism for Gα activation by GPCRs. Nature. 2015; 524:173–9. [PubMed: 26147082]
39. Perica T, Kondo Y, Tiwari SP, McLaughlin SH, Kemplen KR, Zhang X, et al. Evolution of oligomeric state through allosteric pathways that mimic ligand binding. Science. 2014; 346:1254346–6. [PubMed: 25525255]
40. Shis DL, Hussain F, Meinhardt S, Swint-Kruse L, Bennett MR. Modular, multi-input transcriptional logic gating with orthogonal LacI/GalR family chimeras. ACS Synth Biol. 2014; 3:645–51. [PubMed: 25035932]
41. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, et al. Network analysis of protein structures identifies functional residues. J Mol Biol. 2004; 344:1135–46. [PubMed: 15544817]
42. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Sci. 2006; 15:2120–8. [PubMed: 16882992]
43. Spiridonov NA, Wilson DB. Characterization and cloning of celR, a transcriptional regulator of cellulase genes from Thermomonospora fusca. J Biol Chem. 1999; 274:13127–32. [PubMed: 10224066]
44. Pei J, Kim B-H, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008; 36:2295–300. [PubMed: 18287115]
45. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004; 5:113.10.1186/1471-2105-5-113 [PubMed: 15318951]
46. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–42. [PubMed: 10592235]
47. Schrödinger LLC. The PyMOL Molecular Graphics System, Version~1.3r1. J Mol Biol. 2010; 219:623–34.
48. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr. 2011; 67:235–42. [PubMed: 21460441]
49. Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene. 1989; 77:51–9. [PubMed: 2744487]
50. Bell CE, Lewis M. A closer view of the conformation of the Lac repressor bound to operator. Nat Struct Biol. 2000; 7:209–14. [PubMed: 10700279]

Highlights

- AlloRep compiles sequence, mutagenesis, and structural data for LacI/GalR proteins.
- Alignments for >3000 sequences are grouped by subfamily and sampled in the whole family alignment.
- AlloRep includes detailed phenotypic and biochemical data on almost 6000 variants.
- Structural data for 65 proteins are available as residue-contact networks.
- A predicted allosteric position was validated by altering a synthetic repressor.

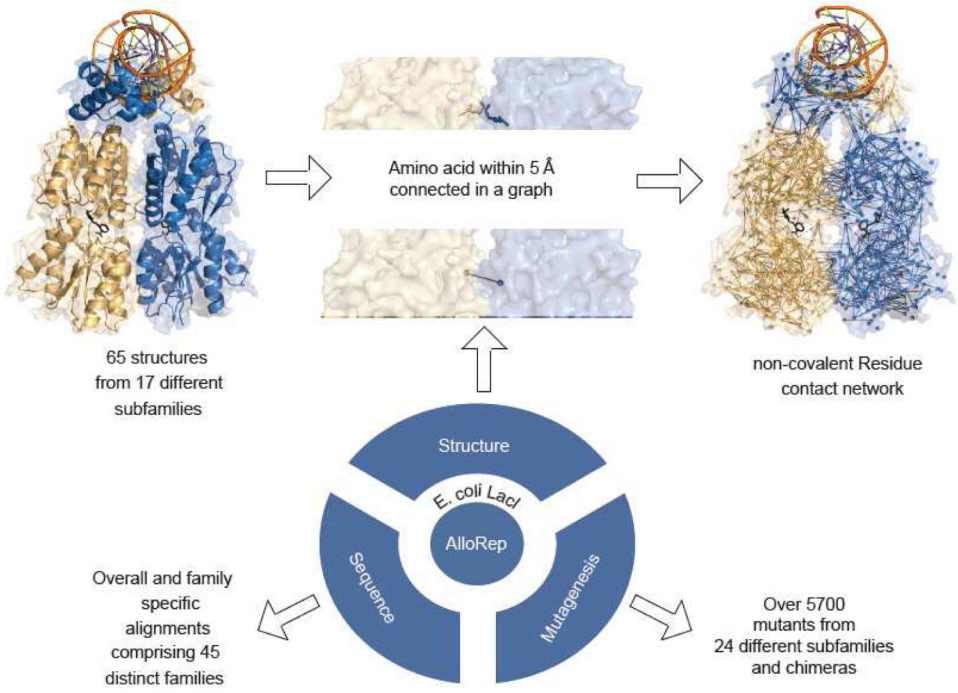


Figure 1.

Overview of the AlloRep database. Each structure at the top shows the two monomers of a representative LacI/GalR homodimer (pdb: 1efa; [50]) in blue and yellow, respectively. DNA is shown as a ribbon at the top of each structure. The allosteric binding sites on the regulatory domains are indicated by the presence of a black ligand. Contacts are defined when any non-hydrogen atoms of two residues come within 5 Å of each other.

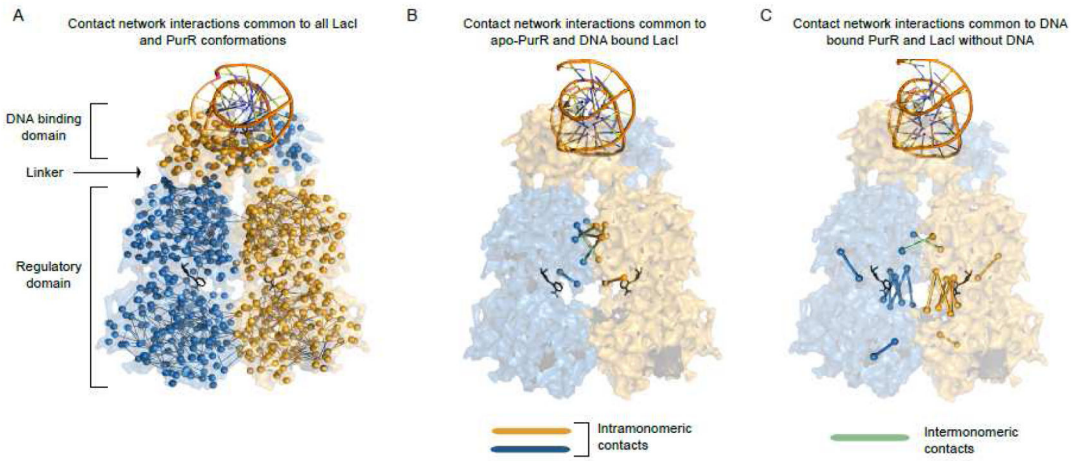


Figure 2.

Examples of residue contact-networks. Structures are represented as in Fig 1. For each panel, the network models are super-imposed over the space-filling model of the structure. For these contact networks, nodes are represented as the Cα atoms and the atomic contacts between two residues are represented as edges. Residues are considered to be in contact if any of their non-hydrogen atoms are within 5 Å of each other. (A) Network representation of the common noncovalent contacts that are present in all available PurR and LacI structures. (B) Network representation of the noncovalent contacts common to apo-PurR structures (no DNA bound) and LacI structures with DNA-bound. (C) Network representation of the noncovalent contacts common to LacI structures without DNA and DNA-bound-PurR conformations. Figures were prepared using Pymol [47].

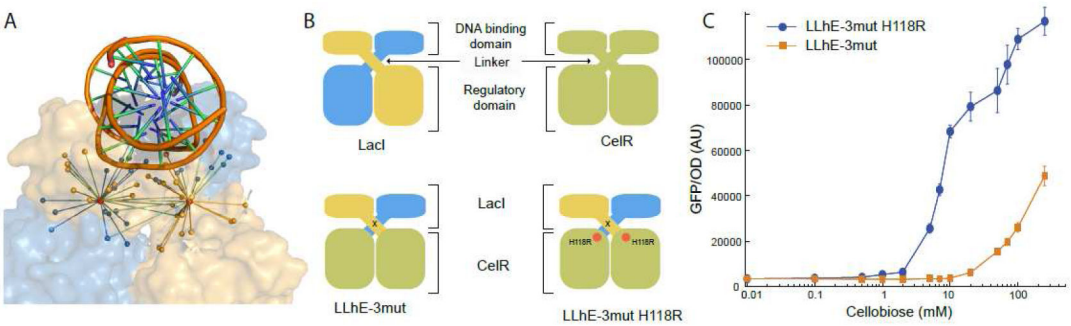


Figure 3.

Experimental validation of an allosteric prediction. A) Network representation of residue contacts for position 118. These networks were derived using all available, full-length structures of LacI/GalR homologs. A region of the LacI homodimer is shown, with monomers colored as in Fig 1; DNA is shown as a ribbon at the top of the structure. Position 118 (one per monomer) is located at the centers of the two networks that are super-imposed on the structure. Other contacting residues are shown as connecting nodes. B) Construction of the LLhE-3mut synthetic repressor [21]. Monomers of this chimera comprise (i) the DNA binding domain and linker of *E.coli* LacI, and (ii) the regulatory domain from *T. fusca* CelR. For measurable repression, an additional three mutations were required in the linker: I48V, Q55A, and Q60R (represented as "x"). In this study, the H118R mutation was added to LLhE-3mut to alter allosteric regulation (represented as a red dot). C) Repression of LLhE-3mut and LLhE-3mut/H118R variants as a function of inducer concentration. The repressor variants were used to control expression of green fluorescent protein (GFP). At low cellobiose concentrations, both variants repressed the *gfp* coding region, and GFP expression could not be detected. Increasing cellobiose concentrations induced the LLhE-3mut variants, allowing expression of GFP in a dose-dependent manner. Note that the H118R mutation required >10-fold more cellobiose, which indicates that position 118 participates in allosteric regulation. Data represent the averages of 3 or 4 independent measurements; error bars (which are usually smaller than the data symbols) show the standard deviation. Lines are to aid visual inspection of the data.